Published in final edited form as:

J Chromatogr A. 2008 February 1; 1179(2): 205–215. doi:10.1016/j.chroma.2007.11.101.

Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm

Cheolhwan Oh^{a,#}, Xiaodong Huang^{b,#}, Fred E. Regnier^{a,b}, Charles Buck^a, and Xiang Zhang^{a,*}

^aBindley Bioscience Center, Purdue University, West Lafayette, IN 47907, USA

^bDepartment of Chemistry, Purdue University, West Lafayette, IN 47907, USA

Abstract

We report a novel peak sorting method for the two-dimensional gas chromatography/time-of-flight mass spectrometry (GC×GC/TOF-MS) system. The objective of peak sorting is to recognize peaks from the same metabolite occurring in different samples from thousands of peaks detected in the analytical procedure. The developed algorithm is based on the fact that the chromatographic peaks for a given analyte have similar retention times in all of the chromatograms. Raw instrument data are first processed by ChromaTOF (Leco) software to provide the peak tables. Our algorithm achieves peak sorting by utilizing the first and second dimension retention times in the peak tables and the mass spectra generated during the process of electron impact ionization. The algorithm searches the peak tables for the peaks generated by the same type of metabolite using several search criteria. Our software also includes options to eliminate non-target peaks from the sorting results, e.g., peaks of contaminants. The developed software package has been tested using a mixture of standard metabolites and another mixture of standard metabolites spiked into human serum. Manual validation demonstrates high accuracy of peak sorting with this algorithm.

Keywords

Two-dimensional gas chromatography/time-of-flight mass spectrometry; peak sorting; algorithm; metabolite; human serum

1. Introduction

Metabolomics is the study of the low molecular mass endogenous metabolites in tissues and biofluids that provides nontargeted, global, qualitative and quantitative information of these small molecules in biological systems. This approach is gaining increasing interest in discovery of biomarkers, elucidation of disease mechanisms and description of new metabolic pathways. Effective application of this approach will provide better understanding of disease processes, improved disease prognoses and diagnoses, as well as more efficient drug discovery [1–4]. Although the number of known metabolites in many organisms is fewer than the number of genes or proteins, development of profiling methods and analytical

© 2007 Elsevier B.V. All rights reserved.

*Corresponding author, Tel: (765) 496-1153. zhang100@purdue.edu (X. Zhang).

#These authors made equal contribution to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

techniques for metabolite detection and identification is a challenge due to the chemical diversity of metabolites and large concentration ranges. Many analytical techniques have been applied in metabolomics including liquid chromatography (LC), capillary electrophoresis (CE), gas chromatography (GC), and each of these coupled with mass spectrometry (LC-MS, CE-MS and GC-MS) [5–9]. Single dimension methods however, often provide discriminating capability that is too limiting. An emerging technology, comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC/TOF-MS), brings increased separation capacity, chemical selectivity and sensitivity for metabolomics analyses [10, 11]. Compared with conventional one-dimensional GC, two-dimensional gas chromatography offers significant advantages for analysis of complex samples: an order-of-magnitude increase in separation capacity (discrimination of over one thousand peaks), significant increase in signal-to-noise ratio and dynamic range, and improvement of mass spectral deconvolution and similarity matches. Since GC×GC/TOF-MS can provide more analytical information, it represents an alternative analytical tool for analysis of low-molecular-mass metabolites in complex biological systems.

Although GC×GC/TOF-MS systems show high reproducibility of metabolite retention times in highly controlled applications, retention times may still shift in two dimensions within certain variation windows as a result of several, sometimes uncontrollable factors such as temperature and pressure fluctuations, matrix effects on samples and stationary phase degradation. Retention time shifts introduce difficulty in data processing for chemometric differential metabolomics analyses because of the requirement that chromatographic peaks for a given analyte have the same retention time in all chromatograms analyzed. Therefore, the retention time variation must be adjusted, i.e., aligned, before applying chemometric methods.

Many sorting and/or alignment algorithms have been suggested for one-dimensional retention time alignment but few studies have addressed two-dimensional gas chromatography separations. Fraga *et al.* developed a rank-based algorithm for GC×GC retention time alignment based on the previous work of Prazen *et al.* [12]. Mispelaar *et al.* also developed a correlation-optimized shifting-based algorithm for alignment of a local region of a GC×GC chromatogram [13]. However, these two methods only can be used to align small regions of interest in the two-dimensional data set and they both require a standard as a point of reference. Recently, Pierce *et al.* proposed an alignment algorithm that can correct the entire chromatogram in both GC dimensions [14]. The algorithm used a novel indexing scheme together with a piecewise retention time alignment algorithm [15–17], in which one chromatogram is used as the target and other chromatograms aligned to it. A chromatogram is defined as an $M \times N$ matrix where there are N units along the first column dimension and M units along the second column dimension. A sample chromatogram can be firstly indexed along the first GC column dimension into M vectors, each N unit long. One of the vectors is then projected into the target chromatogram indexed by the same means and aligned using the piecewise alignment algorithm. In this scheme all of the M vectors in the sample chromatogram can be aligned along the first column dimension. Subsequently, the sample chromatogram is indexed along the second GC column dimension into N vectors, each M unit long and these N vectors are also aligned by the above method. As a result, the entire of GC×GC chromatogram can be aligned in both retention time axes.

It is well known that thousands of metabolites may be present in a biological sample. Some of these metabolites will have similar retention times in both dimensions. Therefore, aligning metabolite peaks based on retention time alone, whether in one- or two-dimensions, may introduce false positive alignments. This situation worsens when complex samples such

as human serum or plant extracts are analyzed. Coupling two-dimensional gas chromatography with mass spectrometry not only increases the analytical peak capacity, but also provides more information for reliable data analysis. In this paper, we report a new method for directly aligning the peaks of GC×GC/TOF-MS experiments according to their retention time in both dimensions and the metabolite mass spectrum information. Raw experimental data are first processed using Leco ChromaTOF software to generate peak tables. The proposed algorithm was implemented in MSort software using MATLAB R2006b. Compared with existing sorting and/or alignment software, our method utilizes metabolite mass spectra generated during electron impact to greatly reduce the false positive peak sortings.

2. Experimental

2.1. Materials and reagents

An amino acid standard mixture, fatty acid standards, organic acid standards, anhydrous pyridine and human serum were obtained from Sigma-Aldrich (St. Louis, MO, USA). Amino acids were mixed at a concentration of 2.5 $\mu\text{mol/mL}$ in 0.1 M HCl except L-cystine at 1.25 $\mu\text{mol/mL}$. The fatty acids and organic acids were prepared by mixing the 0.5 mg/mL fatty acid standards and 0.5 mg/mL organic acid standards in pyridine, respectively. Anhydrous methanol and chloroform were purchased from Mallinkrodt Baker (Phillipsburg, NJ, USA). The solution of ethoxyamine hydrochloride was prepared by dissolving 50 mg ethoxyamine hydrochloride from Sigma-Aldrich in 1 mL pyridine. The derivatization reagent (N-methyl-N-tert-butyldimethylsilyl)-trifluoroacetamide (MTBSTFA) was obtained from Regis Technologies (Morton Grove, IL, USA). Standards used in experiment are listed in Table 1.

2.2. Derivatization

The amino acid mixture (200 μL) was dried with nitrogen flow and redissolved in 100 μL pyridine before derivatization. Two type of test mixtures were prepared by mixing 10 μL of a fatty acid mixture with 10 μL of this amino acid mixture (FA+AA), and with 10 μL organic acid mixture (FA+OA). Eight FA+AA and eight FA+OA test samples were derivatized with 20 μL MTBSTFA for 30 min at 60 °C, respectively.

2.3. Extraction and Derivatization of Serum Samples

Human serum aliquots (50 μL , 100 μL , 200 μL) were mixed with 400 μL organic solvent (water/methanol/chloroform, 2:5:2, v/v) for removal of proteins. After sitting at room temperature for 1 hour and sonication for 10 min, the mixtures were centrifuged at 15 000 rpm to pellet the proteins. The liquid phases were collected and evaporated to dryness by SpeedVac and then redissolved in 150 μL pyridine. The fatty acid mixture (4 μL) was spiked in 20 μL aliquots of the three serum samples. These mixtures were reacted with 10 μL 50 mg/mL ethoxyamine hydrochloride solution for 30 min at 60 °C and subsequently derivatized with 20 μL MTBSTFA for 1 h at 60 °C. A blank was prepared in parallel under the same conditions but without the addition of human serum and fatty acid mixture.

2.4. GC×GC/TOF-MS Analyses

Analyses of the derivatized standard mixtures and the extracted serum sample were performed using a Pegasus 4D GC×GC/TOF-MS instrument (Leco Corporation, St. Joseph, MI, USA) equipped with a cryogenic modulator. This instrument employs an Agilent 6890 gas chromatograph (Agilent Technologies, Palo Alto, CA, USA) and the injector is a CTC Combi PAL auto sampler (CTC Analytics, Zwingen, Switzerland). The first dimension chromatographic column was a 10 m DB-5 capillary column ((5%-phenyl)-methylpolysiloxane) with an internal diameter of 180 μm and a film thickness of 0.18 μm .

The second dimension chromatographic column was a 1 m DB-17 capillary column ((50%-phenyl)-methylpolysiloxane) with an internal diameter of 100 μm and a film thickness of 0.1 μm . High purity helium was used as carrier gas at a flow rate of 1.0 ml/min. The first dimension column oven temperature began at 50 $^{\circ}\text{C}$ with a hold time of 0.2 min, then increased to 300 $^{\circ}\text{C}$ at a rate of 10 $^{\circ}\text{C}/\text{min}$ and held at this temperature for 5 min. The second dimension column oven temperature was 20 $^{\circ}\text{C}$ higher than the corresponding first dimension column oven temperature with the same rate and hold time. The second dimension separation time was set for 4 s. A solvent delay of 110 s was set up for standard mixtures and 250 s for serum sample. Derivatization solutions (2 μL) were injected at a split mode with a ratio of 50:1 except for the serum samples injected at 20:1. The inlet temperature and the transfer line were set at 280 $^{\circ}\text{C}$. The ion source was held at 200 $^{\circ}\text{C}$. The detector voltage was 1600 V and the filament bias was -70 V. Electron impact (EI) mass spectra were collected from 50 to 800 m/z at 50 spectra/s. The Leco ChromaTOF software version 3.25 equipped with the US National Institutes of Standards and Technology (NIST) MS database (NIST MS Search 2.0, NIST/EPA/NIH Mass Spectral Library; NIST 2002) was used for data processing and peak matching.

3. Algorithm

We developed a peak sorting software package for experimental data generated on the Pegasus 4D GC $\times$ GC/TOF-MS instrument equipped with electron impact ionization (Leco Corporation, St. Joseph, MI, USA). Since metabolites are fragmented during ionization, the mass information of the metabolite is lost. The measured mass spectrum contains all fragments and potentially the parent metabolite ions. The mass spectrum therefore provides a fingerprint of the metabolite chemical structure. Raw data from the instrument are first processed by Leco ChromaTOF software to generate a peak table. This peak table contains the features of each identified peak such as metabolite name (if determined), CAS registry number, peak area, first dimension retention time, second dimension retention time, fragment spectrum, and statistical confidence of metabolite identification. This information provides the input for our sorting algorithm. Figure 1 shows the flow chart of MSort software, which contains three major functionalities: peak entry merging, sorting, and contaminant elimination.

Peak entry merging is a preprocessing task carried out before the actual sorting process. In comprehensive two-dimensional chromatography, a short sampling time in the second dimension is prerequisite to obtain high two-dimensional resolution. This means that each separated peak in the first dimension should be split into at least three or four fractions in the second dimension [18–19]. In our experiments, the short modulation time, e.g., 4 s, for the first dimension retention time causes multiple peak entries for one chromatographic peak generated by one metabolite. For example, the chromatographic peak of the same metabolite may be split into two peak entries in one sample and three peak entries in another due to retention time shift or abundance change in two samples. Unequal splitting of a single metabolite peak will prevent further data analysis in a differential metabolomics study. Therefore, peak entries from the same chromatographic peak must be merged so that each peak entry in the merged peak table represents one chromatographic peak. The peak entry merging process utilizes parameters such as the first dimension retention time, the second dimension retention time, the threshold of the spectrum correlation value (R_{min}), and the name of the identified metabolites (*SID*). Inclusion of the name of the identified metabolite (*SID*) in this process is optional because the metabolite identification may provide false positive identification results. Users are thereby reminded to exercise caution when employing the name of an ‘identified metabolite’ as a criterion for peak entry merging. The spectrum correlation is calculated using Pearson’s correlation method. Pearson correlation coefficient (R) assumes that a linear function best describes the relationship between two

variables. It can be used to evaluate data for n subjects, each of which has contributed a score on two variables designated as X and Y . R is calculated as follows:

$$R = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}} \quad (1)$$

In our sorting algorithm, X and Y are the intensities of the two spectra that are compared with each other and n is the number of data points in the spectra, i.e., the number of metabolite fragments.

The peak entry merging algorithm is as follows. First, each peak table is sorted by first dimension retention time in ascending order. We choose the first peak entry (that has the smallest first dimension retention time) as the initial reference peak entry. Then, we seek a peak entry that satisfies the following conditions: (1)

$|T_r^1 - T_c^1| \leq \Delta T^1$ and $|T_r^2 - T_c^2| \leq \Delta T^2$, where T_r^1 and T_c^1 are the first dimension retention times of the reference peak and the current peak in the search, T_r^2 and T_c^2 are the second dimension retention times, and ΔT^1 and ΔT^2 are the limits of allowable first and second dimension retention time drifts, respectively; (2) $R_{rc} \geq R_{min}$; and (3) optionally $SID_r = SID_c$, where R_{rc} is the Pearson's correlation coefficient between the mass spectra of the reference peak and the current peak in the search. R_{min} is the threshold of Pearson correlation coefficient used to determine if two fragment spectra are generated by the same metabolite. SID_r and SID_c are the metabolite names of reference metabolite and current metabolite identified from the NIST database search, respectively. In other words, we seek a peak entry of which not only the first and the second dimension retention times are close to those of the reference peak but also mass spectrum is similar to that of the reference peak (high correlation value), where R_M is the threshold for the decision. If the conditions above are satisfied, we conclude that the current peak entry is from the same chromatographic peak as a reference peak entry and put them in the same group. The current peak entry that was just found becomes a new reference peak entry, and the same procedure continues until no peak entry is found to satisfy the criteria. After a group of peak entries is set, we represent them with one representative peak entry. The values of the first dimension retention time and the second dimension retention time of the representative peak entry are determined by the peak-area-weighted average values or by the median values of the member peaks, which is much more accurate than simply using retention time of the largest peak to represent the metabolite peak location. The mass spectrum of the representative peak is calculated by the sum of all the mass spectra of the member peaks according to the m/z value of each fragment ion. After a group of peak entries is replaced with a single representative peak, we repeat the procedure for the remaining peak entries until all the peak entries in the peak table are converted to merged peaks.

The next step is an sorting process. The sorting algorithm is based on the fact that in an ideal case, the same metabolites have the same first dimension retention time, the same second dimension retention time, and the same mass spectrum. In realistic experimental situations, however, the first and the second dimension retention times of the same metabolite may differ from experiment to experiment. To accommodate this experimental retention time shift, we adopted the retention time variation window concept. That is, if the difference between the retention times of two metabolite peaks in different experiments is less than a specified value, the peaks are considered to belong to the same retention time window. In addition, the mass spectra obtained from the same metabolites should be similar to each other. This motivated us to use the correlation value between two fragment spectra as another search criterion. The values of the retention time variation window and the

similarity of mass spectra could be estimated from replicate analysis of the same sample. Although the metabolite name for each peak is provided by Leco ChromaTOF software, we did not adopt it as one of our basic criteria for peak sorting, but left it as optional. This software searches the NIST spectral database to find the best match of the fragmentation spectrum of a peak. For example, this search outputs the name of a metabolite whose spectrum is the best match of the testing spectrum even if the spectrum of a true metabolite is not in the database. With this approach, the reported metabolite may not be the actual metabolite that gave rise to the fragment spectrum.

Multiple samples are usually analyzed in differential metabolomics. Each of these samples will generate a peak table. Given the multiple peak tables, we first choose a table that has the maximum number of merged metabolite peaks as a reference table (*RefTbl*) to sort as many peaks as possible with the current reference table. We designate the rest of the tables as search tables (*SchTbl*). Each peak in the reference table is set as a reference peak in each sorting task. The objective of each sorting task is to find the peaks in the search tables that were generated from the same metabolite as the corresponding reference peak. As discussed above, the basic sorting parameters are the first dimension retention time variation window size (ΔT^1), the second dimension retention time variation window size (ΔT^2), and the threshold of mass spectrum correlation value (R_{min}). Once again, the reported name of the identified metabolites (*SID*) is an optional parameter. The values of the basic parameters can be determined based on the performance of the instrument while the optional parameter can be employed based on the user's confidence with the identification. Information from the basic group is used to limit the searching space in each input table.

The sorting algorithm is as follows. After the reference table (*RefTbl*) is determined, the search tables are numbered from *SchTbl*₁ to *SchTbl*_{*n*-1}, where *n* is the number of all the peak tables. Then, the algorithm searches peaks in each search table *SchTbl*_{*k*} (*k* = 1, 2, ..., *n*-1) that satisfy the sorting criteria. In other words, the algorithm searches peaks in the search table *SchTbl*_{*k*} such that

$$|T_r^1 - T_c^1| \leq \Delta T^1 \quad (2)$$

$$|T_r^2 - T_c^2| \leq \Delta T^2 \quad (3)$$

$$R_{rc} \geq R_{min} \quad (4)$$

and optionally

$$SID_r = SID_c \quad (5)$$

where T_r^1 and T_c^1 are the first dimension retention times of the reference peak in the reference table and the peak in the current search table respectively, T_r^2 and T_c^2 are their second dimension retention times, R_{rc} is the correlation value between the mass spectra of the reference peak and the peak in the current search table, R_{min} is a threshold of Pearson's correlation coefficient of mass spectra derived from sample metabolites, and SID_r and SID_c are the identified names of the metabolites corresponding to the reference peak and the peak in the search table. If no peak satisfies the sorting criteria, we conclude that there is no peak in the search table *SchTbl*_{*k*} that could be sorted to the reference peak in *RefTbl*. If there is one peak in the search table *SchTbl*_{*k*} that satisfies the sorting criteria, we choose it as an sorted peak. If there are multiple peaks in the search table *SchTbl*_{*k*} that satisfy the sorting

criteria, we choose a peak that has the highest mass spectra correlation value as an sorted peak.

If an sorted peak is found in the search table $SchTbl_k$, the reference values are updated as follows for searching the next search table $SchTbl_{k+1}$.

$$T_{rn}^1 = \frac{T_{ro}^1 \times I_{ro} + T_c^1 \times I_c}{I_{ro} + I_c} \quad (6)$$

$$T_{rn}^2 = \frac{T_{ro}^2 \times I_{ro} + T_c^2 \times I_c}{I_{ro} + I_c} \quad (7)$$

and

$$S_{rn} = S_{ro} + S_c \quad (8)$$

where I_{ro} is the sum of all the intensities of processed mass spectra, I_c is the sum of all the intensities of current mass spectrum, T_c^1 and T_c^2 are the first dimensional and second dimensional retention time, respectively. T_{ro}^1 and T_{ro}^2 are the intensity weighted first dimensional and second dimensional retention time of processed spectra. T_{rn}^1 and T_{rn}^2 are the intensity weighted first dimensional and second dimensional retention time of processed spectra and the current spectrum. In other words, the weighted averages using the total ion currents for the first dimension retention time and the second dimension retention time are used as new reference values. The sum of mass spectra of the reference peak (S_{ro}) and the sorted peak (S_c) is used as a new reference mass spectrum (S_{rn}). Note that S_{ro} , S_c , and S_{rn} are vectors whose elements are the intensities of spectra.

After the reference values are modified, the algorithm searches the next search table $SchTbl_{k+1}$ to find the peak that satisfies the criteria with the highest correlation value. If a peak is found in $SchTbl_{k+1}$, the reference values are modified as above and the searching process is performed in the next search table. This process is continued until the algorithm finishes the searching task in the last search table. All the sorted peaks are recorded in the sorting table, and removed from the reference table and the search tables. The next peak in the reference table is set as a new reference peak and the same procedure is carried out. This process is repeated until all the peaks in the reference table are covered. Since the current reference table is empty at this point, we discard the empty table.

Using the remaining tables, we repeat all the above procedures. The peak table that contains the maximum number of peaks is chosen as a new reference table and the rest are chosen as new search tables. The subsequent steps described above are performed using the new reference table and search tables. The algorithm stops when there is no remaining table in the searching space.

As an additional feature, we implemented a process to deal with metabolites that are not of interest. 'Uninteresting' metabolites may be eliminated before or after the actual sorting operation via two methods: elimination by the exclusion table or elimination of sorted peaks using peak tables of blank samples. In Figure 1, these methods are classified as type 1, 2, and 3 eliminations. The exclusion table contains the information of the metabolites that are not of interest to the user. Type 1 is the case where an exclusion table is provided that contains only the names of the non-target metabolites; these are eliminated before the sorting process. In Type 2 elimination, the exclusion table with the names and the first and second dimension retention times of the non-target metabolites are provided. Non-target

metabolites are eliminated using the metabolite names and the first and second retention times after the sorting process. Use of the peak tables generated by blank samples is Type 3 elimination. In this case, each peak table is generated for each blank sample and the peaks included in these peak tables are considered to be contaminants. Peak tables from blank samples are considered as 'regular' tables and included in the sorting procedure. After sorting is finished, we remove the peaks that are sorted to the contaminant peaks in the blank samples.

4. Results and Discussion

Algorithms for GC×GC/TOF-MS peak sorting have been implemented in our MSort software using MATLAB R2006b. We tested the performance of MSort with experimental data sets generated from two distinct groups of samples. The first is a mixture of standard metabolites (Table 1); the second is a mixture of standard metabolites in human serum. Figure 2a shows the raw data obtained for a human serum sample, where 1535 peaks have been detected by Leco ChromaTOF software with the minimum signal to noise (S/N) ratio set at 10. Three peaks identified as L-methionine are displayed in Figure 2b. Manual validation shows that only one of these peaks is L-methionine, the other two are incorrectly identified. This is probably because the spectra of the metabolites that correspond to the mis-identified peaks were not present in the database. This illustrates the potential false positive identification of metabolites and explains why we included the identified metabolite name as an optional parameter for peak sorting in MSort.

4.1. Analysis of data generated from standard metabolite mixtures

As described above, we used the value of the correlation coefficient between two mass spectra as one of the criteria for the sorting task. To determine the threshold (ThCorr) of the mass spectra correlation coefficients, we manually picked all mass spectra of standard metabolites from 16 samples (mixture of standard metabolites) and calculated Pearson's correlation coefficients between every mass spectrum pair. By this process 116,403 correlation values were generated. It should be noted that the range of the Pearson's correlation coefficient value is from -1 to 1. The closer two spectra are to each other, the closer the correlation coefficient is to 1. We then studied the relation between true positive rate (TPR) and false positive rate (FPR) at different correlation thresholds (*ThCorr*). 'Positive' in this context means that a pair of mass spectra is from the same metabolite. Accordingly, 'negative' means that a pair of mass spectra is from the different metabolites. Since the mass spectra from the same metabolites (positive case) are usually very similar to each other, the correlation coefficient for the positive pair is high. Hence, we can set a threshold and classify a pair of mass spectra with a higher correlation value than the threshold as positive. If the correlation value of two mass spectra is lower than the threshold, the pair is classified as negative. After the calculation of the correlation value for every pair of spectra, we can obtain TPR and FPR for a given threshold using the following,

$$TPR = \frac{TP}{TP + FN} \text{ and } FPR = \frac{FP}{FP + TN}, \quad (9)$$

where TP (true positive) is the number of positive pairs that were classified as positive, FN (false negative) is the number of positive pairs that were classified as negative, FP (false positive) is the number of negative pairs that were classified as positive, and TN (true negative) is the number of negative pairs that were classified as negative. While changing the threshold value from -1 to 1, we calculated the values of FPRs and TPRs, where each pair of (FPR, TPR) corresponds to a specific threshold value. The receiver operating characteristic (ROC) curve obtained by this process is shown in Figure 3. Since every point

on the ROC curve has a corresponding threshold value, we can decide the threshold value that achieves specific TPR and FPR by choosing a point on the curve.

Using the ROC curve, we set the threshold of correlation value as 0.95, which resulted in TPR = 0.92 and FPR = 0.11. It should be noted that the actual FPR value is expected to be much less than 0.11 since metabolite retention time was not considered in calculation of the correlation coefficients. Two distinct metabolites may share a high correlation coefficient but will most likely have distinct retention times. Such metabolites will not interfere with each other during the sorting process. However, such pairs of metabolites are included in our calculation of FPR.

We tested the sorting algorithm with a mixture of fatty acids and amino acids (FA+AA), and a mixture of fatty acids and organic acids (FA+OA) listed in Table 1. The 16 peak tables, which were generated by eight replicate samples of the FA+AA mixture and another eight replicate samples of FA+OA mixture, were used for sorting. In our tests, an exclusion table containing 35 metabolite names was used to demonstrate the software feature to filter the metabolites listed in the exclusion list before the sorting operation (Type 1 elimination).

The sorting table generated by MSort is organized such that each column represents a sample and each row represents a set of sorted metabolites. With parameters $\Delta T^1 = 0.01 \times T_r^1$, $\Delta T^2 = 0.05 \times T_r^2$, $R_{min} = 0.95$, and $SID = \text{False}$, MSort generated an sorting table with 8683 rows, that is, 8683 sets of sorted peaks. Figure 4a indicates that many of the rows include only a small number of entries in the sorting table. More than 6900 peaks were detected in only one of the 16 samples. For each set of sorted peaks (i.e., each row of the sorting table), the average peak area is calculated. This area value is then classified into one of the 16 groups corresponding to the number of entries in the row of the sorting table. For example, the average area value obtained from a row with 10 peak entries becomes a member of group 10. Since there are 16 analyte samples, we have 16 groups, because a row in the sorting table can have 1 – 16 peak entries. This process is performed for all rows in the sorting table to obtain the distribution of average peak areas in each group. The distribution of the average areas is illustrated in a box plot in Figure 4b. The rows with a small number of entries have small peak areas. These are likely to be random peaks and should be excluded for further analysis.

We set the minimum signal to noise ratio (S/N) at 10 for peak detection. At this setting the ChromaTOF instrument software reported about 2500 peaks in each sample. Many of these peaks are clearly noise with very low confidence identifications; the other intense peaks may arise from the derivatization reagent or GC column artifacts. Just a few peaks from the derivatization reagent would appear in very many slices because of large amount of the reagent. These peaks usually elute before 225 sec. The GC column artifacts may generate many bleed peaks that usually elute during last 5 – 10 min of separation. About 700 of the reported peaks have an S/N larger than 100 while only 200 peaks have a S/N larger than 1000. From manual validation of these data, we have determined that the S/N values of most of the standard acids are in the range of 1500 – 31000. However, there are several standards with S/N less than 1500; heptanoic acid (S/N = 18), decanoic acid (S/N = 83), oxalic acid (S/N = 47) and adipic acid (S/N = 87). For the purpose of sorting, we included all peaks with S/N ≥ 10, even those of low confidence and artificial peaks, as this provides an advantage to test the robustness of our software.

In our sorting table, each peak entry indicates that a metabolite has been detected in the corresponding sample. We expect to observe rows that have 8 peak entries or 16 peak entries based on the composition of our samples. Each fatty acid should appear in all 16 samples because it is included in both FA+AA and FA+OA mixtures. Therefore, it should

be sorted over 16 samples, which gives a row with 16 peak entries. Each amino acid and organic acid should give a row with 8 peak entries because it is included only in either the FA+AA or FA+OA mixtures. However, we also obtained rows with 9, 11, 14, and 15 peak entries in the sorting table (Figure 4b). Manual validation revealed that all 46 standard metabolites assessed were correctly sorted, e.g., 16 amino acids, 17 fatty acids and 13 organic acids. The other sorted 'peaks' were either from the derivatization reagent or degradation of GC×GC columns.

Figure 5 shows the sorting results. Random samples from both groups (Sample 1: FA+AA-1 and Sample 9: FA+OA-1) were selected for evaluation. Scatter plots of the first- and second-dimension retention times are shown in (a) and (b). The correlation coefficient between the first-dimension retention times of Sample 1 and Sample 9 was 1.000. For the second-dimension retention times it was 0.996. These results show that the retention times in the sorting table have a very strong linear relationship, which demonstrates high instrument reproducibility and accurate sorting results.

To further check the data quality, we calculated the peak distribution in GC×GC chromatograms using the Kolmogorov-Smirnov (K-S) test. The K-S statistic value measures the difference between two distributions. The more similar two distributions are, the closer the K-S statistic value is to zero. The histograms of the K-S test statistic values are shown in Figure 5c and d. To obtain the histogram in Fig. 5c, the K-S statistic value for the distributions of first-dimension retention times was calculated for each pair of samples. The same procedure was taken to obtain the histogram in Fig. 5d for the second-dimension retention time. Most K-S statistical values in these analyses are lower than 0.05 in (c), and lower than 0.1 in (d). This indicates that the distributions of the first- and the second-dimension retention times are very close between samples, which suggests that the sorting results are quite accurate.

4.2. Analysis of data generated from mixtures of human serum and fatty acids

We tested our algorithm with another set of data from three mixtures of human serum and fatty acids (50 μ L, 100 μ L, and 200 μ L of human serum combined with 4 μ L of the fatty acid mixture). Each mixture was experimentally analyzed 5 times with 2 μ L injections. In this experimental design, we varied the concentration of fatty acids in human serum to test the sensitivity of our sorting algorithms.

Figure 6 shows the sorting results from these test data. We picked up one sample from each group (50-Serum+FA, 100-Serum+FA, and 200-Serum+FA) respectively, and obtained the three dimensional scatter plots for both the first- and second-dimension retention times of all sorted entries. Sorting was performed with parameters $\Delta T^1 = 0.01 \times T_r^1$, $\Delta T^2 = 0.05 \times T_r^2$, $R_{min} = 0.95$, and $SID = \text{True}$ (Fig. 6(a) and (b)), and with parameters $\Delta T^1 = 0.01 \times T_r^1$, $\Delta T^2 = 0.05 \times T_r^2$, $R_{min} = 0.95$, and $SID = \text{False}$ (Fig. 6(c) and (d)). The correlation coefficients between the samples for the first- and the second-dimension retention times in Figure 6 are summarized in Table 2. The sorting performances for the first dimension retention time (Fig. 6(a) and (c)) were almost the same regardless of the SID option selection. However, for the second dimension retention time, the sorting performance with $SID = \text{True}$ was slightly better (Fig. 6b and d).

To further study the performance of MSort, we analyzed the similarity of the fragment spectra of all full entries with full sorting. First, we picked up the rows with a complete entry (15 peak entries) from the sorting table. Then, we calculated the correlation coefficient between each pair of fragment spectra within the same row. Since there are 105 ($=15 \times 14/2$) pairs, we could obtain 105 different values of the correlation coefficients from one row. We

repeated the same procedure for all the selected rows. Finally, we drew the box plot for the distribution of the correlation coefficients (Figure 7) demonstrating the similarity of the fragment spectra of the sorted peaks. Each box illustrates the distribution of the 105 correlation coefficients for each selected row. All the values of correlation coefficients are very close to 1.0 even though our threshold of correlation was set 0.95. This result again demonstrates high performance of our sorting algorithm. We also manually checked the sorting table and concluded that all of fatty acid standards were correctly sorted.

Most of current available GC or GC×GC alignment algorithms perform chromatogram alignment, e.g., aligning metabolites based on retention time. It should be noted that MSort algorithm does not include chromatogram alignment functionality. Based on our experience, the retention time shift of the same metabolite in the Leco GC×GC/TOF-MS instrument does not have significant variation during the entire separation and the overall retention time drift is less than 1% for the first dimensional retention time and 4% for the second dimensional retention time for replicate continuous runs. For this reason, the user predefined retention time variation windows for the first and second dimensional retention times are used to set a retention time rectangular, i.e., retention time box, in the two dimensional retention time map of the other sample, e.g., the search sample. We expect that the true corresponding metabolite peak in the search sample presents in this retention time box. Then, the metabolite fragment spectrum in the reference sample is used as the key criterion for find the right metabolite peak from peak table of the search sample, which is done by studying the linear correlation between the fragment spectrum of the reference sample with the fragment spectra of every peaks of the search sample present in the retention time box. The peak that has the best correlation score in the search sample is then selected the candidate peak that corresponding to the peak in the reference sample.

The size of user defined retention time box affects the reliability and efficiency of MSort software. If the size of retention time box is smaller than the actual retention time shift between the reference sample and the search sample, the true corresponding peak may not present in the retention time box in the search sample and therefore, will not be selected. MSort will treat the current peak as a missing peak in the search sample. If the size of retention time box is much larger than it should be, more peaks in the search table will be analyzed. This will increase the sorting space in the search sample and therefore, increase the computing time. However, the large retention time box will not introduce either false positive or false negative sorting because MSort selects a peak from the search table that has the largest spectral correlation with the current peak in the reference sample.

5. Conclusions

Metabolomics profiling has been widely used to obtain quantitative information about expression of small molecules in biological systems. The GC×GC/TOF-MS instrument empowers metabolomics profiling studies with high reproducibility, resolution, and analysis power. We have developed MSort, an informatics tool to find metabolite peaks generated from the same metabolite but detected in different biological samples. The developed software has been tested with two types of samples; a mixture of standard metabolites, and a mixture of human serum and standard metabolites. Our analysis showed that MSort is able to sort metabolite peaks generated from these two types of samples with very high accuracy.

Acknowledgments

The authors thank Dr. Mark Merrick at Leco Corporation for his advice in data analysis of using Leco ChromaTOF software. This project was supported by National Institute of Health through grant NIH DK070290.

References

1. Harrigan, GG.; Goodacre, R., editors. Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. Boston, MA: Kluwer; 2003. p. 171
2. Watkins SM, German JB. Curr. Opin. Mol. Ther. 2002; 4:224. [PubMed: 12139307]
3. Dunckley T, Coon KD, Stephan DA. Drug Discovery Today. 2005; 10:326. [PubMed: 15749281]
4. Harrigan, GG.; Goodacre, R., editors. Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. Boston, MA: Kluwer; 2003. p. 199
5. Dunn WB, Ellis DI. Trends Anal. Chem. 2005; 24:285.
6. Kopka J. J. Biotechnol. 2006; 124:312. [PubMed: 16434119]
7. Wilson ID, Plumb R, Granger J, Major H, Williams R, Lenz EM. J. Chromatogr. B. 2005; 817:67.
8. Soo EC, Aubry AJ, Logan SM, Guerry P, Kelly JF, Young NM, Thibault P. Anal. Chem. 2004; 76:619. [PubMed: 14750855]
9. Lenz EM, Wilson ID. J. Proteome Res. 2007; 6:443. [PubMed: 17269702]
10. Marriott P, Shellie R. Trends Anal. Chem. 2002; 21:573.
11. Shellie RA. Aust. J. Chem. 2005; 58:619.
12. Fraga CG, Prazen BJ, Synovec RE. Anal. Chem. 2001; 73:5833. [PubMed: 11791551]
13. Mispelaar VGV, Tas AC, Smilde AK, Schoenmakers PJ, van Asten AC. J. Chromatogr. 2003; A1019:15.
14. Pierce KM, Wood LF, Wright BW, Synovec RE. Anal. Chem. 2005; 77:7735. [PubMed: 16316183]
15. Watson NE, VanWingerden MM, Pierce KM, Wright BW, Synovec RE. J. Chromatogr. A. 2006; 1129:111. [PubMed: 16860329]
16. Pierce KM, Hope JL, Hoggard JC, Johnson KJ, Wright BW, Synovec RE. J. Chromatogr. A. 2005; 1096:101. [PubMed: 16301073]
17. Pierce KM, Wright BW, Synovec RE. J. Chromatogr. A. 2007; 1141:106. [PubMed: 17174960]
18. Murphy RE, Schure MR, Foley JP. Anal. Chem. 1998; 70:1585.
19. Schoenmakers P, Marriot P, Beens J. LC-GC Eur. 2003; 16:335.

Figure 1.
Flow chart of MSort software. The right side provides details for information flow in the sorting step.

Figure 2. Three-dimensional visualization of raw GCxGC/TOF-MS data. (a) Raw data from a serum sample. (b) Three distinct peaks identified by ChromaTOF software as L-Methionine (two of these are false positives).

Figure 3.

Receiver operating characteristic (ROC) curve (TPR: true positive rate, FPR: false positive rate). 'Positive' in this plot means that a pair of mass spectra is from the same metabolite. The TPR is the rate that pairs of mass spectra from the same metabolite are correctly assigned to that metabolite. The FPR is the rate that pairs of mass spectra from different metabolites are incorrectly classified arising from the same metabolites. Each point on the ROC plot has a corresponding threshold value (*ThCorr*). Therefore, by choosing a point on the plot, we can obtain the threshold value that achieves specific TPR and FPR.

Figure 4.

(a) The number of peak entries in a row of the sorting table versus the number of rows in the sorting table. (b) The box plot for the peak area. The upper end of the upper whisker is the maximum of the average peak area, the upper end of the box is the upper quartile, the lower end of the box is the lower quartile, the lower end of the lower whisker is the minimum, and the bar inside the box is the median.

Figure 5. Sorting results (parameter values: $\Delta RT1 = 0.01 \times RT1_{ref}$, $\Delta RT2 = 0.05 \times RT2_{ref}$, $ThCorr = 0.95$, $SID = True$).

Figure 6.

Sorting results from sample 1, 6, and 11 with sorting parameter values $\Delta RT1 = 0.01 \times RT1_{ref}$, $\Delta RT2 = 0.05 \times RT2_{ref}$, $ThCorr = 0.95$. (a) and (b) show the first and second dimension retention times with $SID = \text{True}$, resp. (c) and (d) show the first and second dimension retention times with $SID = \text{False}$, resp.

Figure 7.

Box plots of the correlation coefficients between fragment spectra. The upper end of the upper whisker is the maximum of the correlation coefficient, the upper end of the box is the upper quartile, the lower end of the box is the lower quartile, the lower end of the lower whisker is the minimum, and the bar inside the box is the median. (Parameter values: $\Delta RT1 = 0.01 \times RT1_{ref}$, $\Delta RT2 = 0.05 \times RT2_{ref}$, $ThCorr = 0.95$, and $SID = \text{True}$.)

Table 1

Standards used in experiments.

Amino acid	Fatty acid	Organic acid
L-Alanine	Hexanoic acid	Isobutyric acid
Glycine	Heptanoic acid	Butyric acid
L-Valine	Octanoic acid	L-(+)Lactic acid
L-Leucine	Nonanoic acid	Benzoic acid
L-Isoleucine	Decanoic acid	Oxalic acid
L-Methionine	Undecanoic acid	Malonic acid
L-Serine	Dodecanoic acid	Maleic acid
L-Threonine	Tridecanoic acid	Succinic acid
L-Proline	Tetradecanoic acid	Fumaric acid
L-Phenylalanine	Pentadecanoic acid	Adipic acid
L-Aspartic acid	Hexadecanoic acid	D-Malic acid
L-Glutamic acid	Heptadecanoic acid	D-Tartaric acid
L-Lysine	Octadecanoic acid	Citric acid
L-Histidine	Nonadecanoic acid	
L-Tyrosine	Docosanoic acid	
L-Cystine	Tricosanoic acid	
	Tetracosanoic acid	

Table 2

Correlation coefficients between the samples for the first and the second dimension retention times in Figure 6.

	(a)	(b)	(c)	(d)
Correlation coefficient between Samples 1 and 6	1.000	0.999	1.000	0.972
Correlation coefficient between Samples 6 and 11	1.000	0.999	1.000	0.967
Correlation coefficient between Samples 11 and 1	1.000	0.999	1.000	0.981