BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1786–1792
doi:10.1093/bioinformatics/btt275

Data and text mining

Advance Access publication May 11, 2013

MetPP: a computational platform for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics

Xiaoli Wei[1], Xue Shi[1], Imhoi Koo[1], Seongho Kim[2], Robin H. Schmidt[3,4], Gavin E. Arteel[3,4], Walter H. Watson[3,4,5], Craig McClain[3,4,5,6] and Xiang Zhang[1,*]

[1]Department of Chemistry, [2]Department of Bioinformatics and Biostatistics, [3]Department of Pharmacology & Toxicology, [4]Alcohol Research Center, [5]Department of Medicine, University of Louisville and [6]Robley Rex Louisville VAMC, Louisville, KY 40292, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Due to the high complexity of metabolome, the comprehensive 2D gas chromatography time-of-flight mass spectrometry (GC×GC-TOF MS) is considered as a powerful analytical platform for metabolomics study. However, the applications of GC×GC-TOF MS in metabolomics are not popular owing to the lack of bioinformatics system for data analysis.

Results: We developed a computational platform entitled metabolomics profiling pipeline (MetPP) for analysis of metabolomics data acquired on a GC×GC-TOF MS system. MetPP can process peak filtering and merging, retention index matching, peak list alignment, normalization, statistical significance tests and pattern recognition, using the peak lists deconvoluted from the instrument data as its input. The performance of MetPP software was tested with two sets of experimental data acquired in a spike-in experiment and a biomarker discovery experiment, respectively. MetPP not only correctly aligned the spiked-in metabolite standards from the experimental data, but also correctly recognized their concentration difference between sample groups. For analysis of the biomarker discovery data, 15 metabolites were recognized with significant concentration difference between the sample groups and these results agree with the literature results of histological analysis, demonstrating the effectiveness of applying MetPP software for disease biomarker discovery.

Availability: The source code of MetPP is available at http://metaopen.sourceforge.net

Contact: xiang.zhang@louisville.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 5, 2012; revised on February 28, 2013; accepted on May 8, 2013

1 INTRODUCTION

Comprehensive 2D gas chromatography time-of-flight mass spectrometry (GC×GC-TOF MS) uses two GC columns connected via a thermal modulator. Compared with the first column, the second column is usually a short column with a different stationary phase and is operated at a higher temperature. The metabolites co-eluted from the first column are further separated in the second column and are directed to a time-of-flight mass spectrometer for detection. Consequently, the GC×GC-TOF MS system brings more accurate and rich information about compound retention times and mass spectrum than a 1D GC-MS system, representing a powerful technique for analysis of metabolites in complex biological systems.

The GC×GC-TOF MS system generates a huge amount of high-dimensional data in metabolomics study that require efficient and accurate data analysis algorithms to uncover the biological information. Many data analysis algorithms have been developed to process the GC×GC-TOF MS data for peak picking (Hoggard *et al.*, 2009; Reichenbach *et al.*, 2005; Sinha *et al.*, 2004; Vivo-Truyols 2012), chromatogram alignment (Fraga *et al.*, 2001; Pierce *et al.*, 2005; van Mispelaar *et al.*, 2003; Zhang *et al.*, 2008) and peak list alignment (Almstetter *et al.*, 2009; Kim *et al.*, 2011; Wang *et al.*, 2010). Guineu (Castillo *et al.*, 2011) is the only reported tool that uses the peak lists as its input and performs retention time correction, peak list alignment, normalization and statistical significance tests. A comprehensive review of the developed methods is given by Reichenbach *et al.* (2012). Compared with liquid chromatography mass spectrometry (LC-MS), the applications of GC×GC-TOF MS to metabolomics were not fully explored during the past decade. One significant bottleneck limiting the usage of GC×GC-TOF MS in metabolomics is the lack of accurate and comprehensive data analysis tools.

We herein report a computational platform for analysis of metabolomics data generated from the GC×GC-TOF MS instrument in a form of metabolomics profiling pipeline (MetPP). MetPP uses the peak lists deconvoluted from the instrument data as its input and then renders peak filtering and merging, retention index matching, peak list alignment, normalization, statistical significance tests and pattern recognition. To test its performance, MetPP was used to analyze the GC×GC-TOF MS data of 30 metabolite extracts of mouse livers with spiked-in metabolite standards at different concentrations, and two groups of liver metabolite samples from mice fed different diets. MetPP was developed using MATLAB (The Mathworks, Natick, MA, USA), and its source code is available at http://metaopen.sourceforge.net.

*To whom correspondence should be addressed.

2 METHODS

2.1 Materials

2.1.1 Spike-in samples A total of 30 spike-in samples were analyzed on a LECO Pegasus 4D GC×GC-TOF MS instrument (LECO Corp., St. Joseph, MI, USA). The sample set was composed of five sample groups, i.e. G10, G20, G40, G50 and G80, with six samples per group. All samples in the five sample groups had the same amount of metabolite extract from mouse livers. All six samples in the same sample group had the same amount of the 28 spiked-in metabolite standards (Supplementary Table S1), but amounts were different among the sample groups with concentration ratio of C_{G10}:C_{G20}:C_{G40}:C_{G50}:C_{G80} = 1:2:4:5:8.

2.1.2 Mouse liver samples Four-week-old male C57Bl/6J mice were fed AIN-76A purified diet for 1 week before initiating feeding with either low-fat diet (LFD; 13% fat in calories) or high-fat diet (HFD; 42% fat in calories) for 10 weeks, as described in our previous study (Tan *et al.*, 2011). This study evaluated six mice fed the LFD and five mice fed the HFD. For termination, mice were anesthetized with ketamine/xylazine (100/15 mg/kg, i.m.). Portions of liver tissue were frozen immediately in liquid nitrogen. Metabolites were extracted from mouse liver using solvent methonal:water (v:v = 4:1).

Details of chemicals, sample preparation (spike-in samples and liver samples), GC×GC-TOF MS analysis and instrument data analysis using ChromaTOF are provided in the Supplementary Information.

2.2 Algorithm

Figure 1 shows the workflow of MetPP software that consists of seven functional modules: retention index matching, peak filtering and merging, peak list alignment, quant mass conversion, normalization, statistical significance tests and pattern recognition. MetPP uses the peak list deconvoluted from the instrument as its input. In case of the LECO Pegasus instrument, the instrument control software ChromaTOF reduces the instrument data into peak lists, where each peak is characterized by more than 60 values. Nine of these reported values are used by MetPP for data analysis, including first dimension time (s), second dimension time (s), chemical abstract system (CAS) number, quant masses, area, profile purity, purity, concerns and mass spectra.

2.2.1 Retention index matching Retention index is a concept in gas chromatography to convert the retention times into system-independent constants (Kovats, 1958). The linear retention index is defined as follows:

$$I^T = 100z + 100\left(\frac{t_{R(S)} - t_{R(Z)}}{t_{R(Z+1)} - t_{R(Z)}}\right) \quad (1)$$

where t_R is retention time, s refers to the target compound that elutes off the GC column between two adjacent n-alkane reference compounds with carbon numbers z and $z+1$, respectively, z refers to the n-alkane with z carbon atoms and $z+1$ represents the n-alkane with $z+1$ carbon atoms.

Using retention index information can increase the confidence of compound identification. Currently, MetPP can perform retention index matching only for the first dimensional GC. We improved the algorithms of iMatch approach (Zhang *et al.*, 2011) by enabling retention index matching of multiple top-ranked identification results, to ensure that the true metabolite can still be preserved even though it may not have the best score for mass spectral matching (Koo *et al.*, 2011). We further updated the retention index database used in iMatch to NIST11 retention index database and calculated all retention index distributions. Briefly, all retention index data recorded in the NIST11 retention index database are categorized into nine subgroups based on experimental conditions (stationary phase and temperature gradient). An empirical distribution function (DF) of the absolute retention index deviation to its mean value is generated from each of the grouped retention index data. The DF information is then used for retention index matching based on a user-defined confidence interval (P value), from which a retention index variation window is deduced.

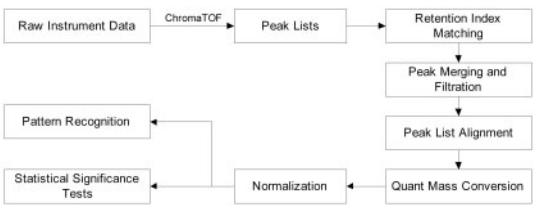

Fig. 1. Workflow of MetPP software

2.2.2 Peak merging and filtration It is possible that multiple peak entries can be falsely generated for one metabolite in the peak list during spectrum deconvolution. To minimize data variations introduced into the downstream statistical analysis, all peak entries of the same metabolite need to be recognized and merged as one entry in each peak list. MetPP uses a two-step approach to recognize and merge the multiple peak entries originated from the same metabolite. Following criteria are used to check the same metabolite appears more than once:

$$\begin{cases} \text{same CAS number} \\ \left|{}^1t_{i,j} - {}^1t_{i,j+1}\right| \le k \cdot P_M \\ \left|{}^2t_{i,j} - {}^2t_{i,j+1}\right| \le {}^2\varepsilon_R \end{cases} \quad (2)$$

where ${}^1t_{i,j}$ and ${}^1t_{i,j+1}$ are the first dimensional retention times of two adjacent j-th and $(j+1)$-th peak entries in the i-th sample S_i, respectively, k is a user-defined coefficient, P_M is modulation period, ${}^2t_{i,j}$ and ${}^2t_{i,j+1}$ are the second dimensional retention times of the two adjacent peak entries, respectively, and ${}^2\varepsilon_R$ is user-defined maximum variation of ${}^2t_R$. The default values of k and ${}^2\varepsilon_R$ are 1 and 0.15, respectively.

MetPP merges all the multiple peak entries into a representative peak. To select a representative peak from the multiple peak entries that fulfills all the three criteria in Equation (2), the CAS number of these peak candidates is used to search the NIST11 mass spectral library to get its reference spectrum s_r. Linear regression is used to fit each of the mass spectra s_i to s_r. The peak entry with maximum fitting coefficient can be determined as follows:

$$s = \arg\max_{i=1,\ldots,m}\{Cor(s_i, s_r)\} \quad (3)$$

where $Cor(x,y)$ is the Pearson's correlation coefficient between x and y. The peak with the maximum value of $Cor(x,y)$ is selected as the representative peak with its peak area equals to the sum of the peak areas of the multiple peak entries, and its values of the first-dimension retention time and the second-dimension retention time are determined by the peak-area-weighted average values.

2.2.3 Peak list alignment MetPP uses the 2D retention times and mass spectrum of each peak for alignment. The metabolite peaks in different samples are aligned based on their similarity of these three pieces of information measured by a mixture score. The retention time value of each peak in a peak list P_i is first transformed into a modified z-score as follows (Wang *et al.*, 2010):

$$z_{i,j} = \frac{t_{i,j} - \mu}{\sigma} \quad (4)$$

where $t_{i,j}$ denotes the retention time of the j-th peak in the i-th sample S_i, μ and σ are the median values of means and standard deviations of the retention time values among peaks of all sample set S, respectively.

The mass spectral similarity between two peaks is calculated as follows (Kim *et al.*, 2012):

$$R = \frac{\alpha_w \circ \beta_w}{||\alpha_w|| \cdot ||\beta_w||} \quad (5)$$

where $w=(x,y)$ is a vector of weight factors of intensity and m/z value, respectively, $\alpha_w = (\alpha_i^w)_{i=1}^n$ and $\beta_w = (\beta_i^w)_{i=1}^n$ are weighted intensities and

$$\alpha_i^w = (\alpha_i)^x \cdot (z_i)^y \text{ and } \beta_i^w = (\beta_i)^x \cdot (z_i)^y \quad (6)$$

where $\alpha = (\alpha_i)_{i=1,\ldots,n}$ and $\beta = (\beta_i)_{i=1,\ldots,n}$ are the intensities of the two matching mass spectra, respectively, z_i is the m/z value of the i-th intensity, $i = 1, 2, \ldots, n$, and x and y are weight factors with $(x, y) = (0.53, 1.3)$.

To measure the matching quality between two peaks, a mixture similarity score C_m is defined as follows:

$$C_m(d_i, R_i|w) = w\left[1 - \frac{d_i}{d_{\max}}\right] + (1-w)R_i \quad (7)$$

where d_i is the Euclidean distance of z-score transformed retention times in the 2D chromatogram between the i-th two matched peaks, $d_{\max}$ is the maximum value of the Euclidean distance between two matched peaks in all samples, R_i is the spectral similarity between the matched peak pair and w is a coefficient ($0 \le w \le 1$).

To perform alignment, MetPP uses a two-step peak alignment approach, full alignment and partial alignment. The full alignment recognizes all potential landmark peaks, i.e. peaks generated by the same metabolite and presented in all samples. During the full alignment, a reference peak list (P_r) is first randomly selected from the sample set S. Each of the remaining peak lists is aligned to P_r, respectively. Considering two peak lists $\{P_r, S_i\}$, the mass spectrum of each peak in P_r is used to first find all corresponding peaks in S_i with a mass spectral similarity $R \ge 0.6$. It should be noted that 0.6 was empirically chosen based on our experience and that this threshold can be decided by a user in MetPP. The peak in S_i with the minimal Euclidian distance to the peak in P_r in the 2D chromatogram is then considered as a possible match to the peak in P_r. All of these matched peak pairs are considered as candidates of landmark peaks between $\{P_r, S_i\}$, and further used to optimize the value of weight factor w for calculation of C_m by maximizing

$$\arg\max_w \left(\sum_{i=1}^{L} C_m(d_i, s_i|w)\right) \quad (8)$$

where L is the number of peak pairs and w is set as 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95, respectively. This process is operated on all peak list pairs $\{P_r, S_i|i = 1, \ldots, n-1\}$, and the optimal weight factor set $\{w_1, \ldots, w_{n-1}\}$ is obtained for $\{S_i|i = 1, \ldots, n-1\}$. The outlier values in $\{C_m^i|i = 1, \ldots, n-1\}$ are further detected at a confidence level of 95%. All landmark peak candidates are removed if their C_m values are detected as outliers. The minimum mixture score $C_m^{\min}$ of the test peak lists is then used as the threshold value of C_m in the partial alignment.

After full alignment, the retention time values of remaining peaks in S_i are corrected based on the retention time difference of the landmark peaks between S_i and P_r. Local linear fitting is used to correct retention time values of metabolites eluted between two adjacent landmark peaks in the 2D retention time domain, respectively. To correct the first dimensional retention time of peaks that do not elute off the GC column between two landmark peaks, i.e. peaks eluted before the first landmark peak and the peaks eluted later than the last landmark peak, an iterative optimization method is applied to each of these two sections of peaks, respectively. In each optimization process, 30% of landmark peaks are randomly selected and a polynomial model fitting is used to correct the retention times of peaks in the section of interest. The polynomial fitting error is computed as follows:

$$\varepsilon = \sum_{j=1}^{k} \left| {}^1t_{i,j}^o - {}^1t_{i,j}^f \right| \quad (9)$$

where ${}^1t_{i,j}^o$ is the original z-score transformed first dimensional retention time of the j-th peak in the i-th sample S_i, ${}^1t_{i,j}^f$ is the fitted retention time of the j-th peak and k is the number of peaks in S_i at the section of interest. This process is repeated 1000 times and the model with minimum fitting error is selected and used for retention time correction. The same process is performed to correct the second dimensional retention time.

After the retention time correction in S_i, the partial alignment is performed to align all the non-landmark peaks in S_i to the peaks in P_r. For each peak pair in $\{P_r, S_i\}$, a mixture score C_m is calculated using Equation (6). A peak pair is considered as a match if its mixture score $C_m \ge C_m^{\min}$. If one peak in the test sample is matched to multiple peaks in P_r, or *vice versa*, the peak pair with the maximum mixture score is kept and the remaining matches are discarded. If a peak in S_i cannot be matched to any peaks in P_r, this peak is considered as a new peak to P_r and is added to P_r. The updated P_r is then used to align peaks in the next test peak list, and this process is repeated until all test peak lists are aligned.

2.2.4 Quant mass conversion After peak alignment, an alignment table is obtained, $A = [P_1, P_2, \ldots, P_m]^T$, where $P_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,n}\}$ represents all aligned peaks of the i-th compound in S, m is the number of aligned metabolite peaks, $p_{i,j}$ denotes the peak of the i-th compound in the j-th sample and n is the number of samples. If the peak list is generated by ChromaTOF, the peak area of each peak refers to the peak area of a specific fragment ion entitled quant mass. The same compound may have different quant mass in each sample. Therefore, it is critical to ensure that the peak area of each aligned peak is consistent across samples.

A reference spectrum-based peak area conversion method is developed in MetPP to convert the values of peak area to be quant mass independent. For a group of aligned peaks P_i, the quant mass with the highest frequency across all samples is selected as the representative quant mass q_i. The CAS number of the peak P_i with the largest spectral similarity in the peaks with the same q_i is then used to extract reference mass spectrum s_r from the NIST11 mass spectral database. A linear regression process is applied between s_r and each spectrum of the aligned peaks in P_i, respectively, to obtain a set of regression coefficients $\{a_{i,1}, a_{i,2}, \ldots, a_{i,n}\}$. Each peak area in P_i is then converted into a new value by

$$I'_{i,j} = \frac{a_{i,j}}{a_{i,1}} I_{i,j} \quad (10)$$

where $a_{i,1}$ is the fitting coefficient of the i-th peak in sample S_1, $a_{i,j}$ is the fitting coefficient of the peak in S_j and $I_{i,j}$ is the original peak area of the i-th compound in S_j.

2.2.5 Normalization Six normalization methods are implemented in MetPP, including quantile normalization, cyclic loess normalization (Dudoit *et al.*, 2002), contrast-based normalization (Astrand, 2001), trimmed constant mean normalization, trimmed constant median normalization and group-based quantile normalization (Wei *et al.*, 2011).

Quantile normalization makes the peak area distribution in each sample the same across all samples. Cyclic loess and contrast-based normalizations are two extensions of the difference in log expression values versus the average of the log expression values method. Trimmed constant mean and trimmed constant median are two scaling-based normalization methods. Group-based quantile normalization first performs quantile normalization for the samples that belong to the same sample group, and then uses a trimmed constant mean method to normalize all samples across the sample groups.

2.2.6 Statistical significance tests Multiple conventional statistical significance test methods are provided in MetPP, including two-tailed t-test, two-sample Kolmogorov–Smirnov test, Kruskal–Wallis test and Wilcoxon rank sum test. To increase the confidence of the statistical testing, sample permutation is also provided as an option for each conventional statistical significance test. Permutation test, also called re-randomization test, is a nonparametric test. It first randomly exchanges sample labels and then performs one of the above mentioned conventional statistical test methods.

The Bonferroni correction (Curran-Everett, 2000), the Benjamini–Hochberg method (Benjamini and Hochberg, 1995) and the q-value method (Storey, 2002) are available to correct multiple comparison issues.

2.2.7 Pattern recognition MetPP first filters data based on a user-defined frequency threshold, defined as the number of samples, in which a metabolite was detected, divided by the number of all samples. The *k*-nearest neighbor imputation algorithm is then used to estimate the missing data (Troyanskaya *et al*., 2001). Two feature selection methods, principal component analysis (PCA) (Jolliffe, 2002) and partial least squares (Rosipal, 2006), are provided as options for the user, if the user decides to reduce irrelevant variables for improving clustering efficiency. Three clustering methods were implemented, including *k*-means clustering, agglomerative hierarchical clustering and fuzzy c-means clustering (Bezdek, 1981). Two options for clustering objects are provided, users can choose all data to do clustering analysis or select only the significant molecules obtained from statistical analysis, for clustering. The clustering accuracy is calculated as the number of correctly clustered samples divided by the number of all samples.

3 RESULTS

MetPP has a modular design including project management and multiple data analysis components. Supplementary Figure S1 is a screenshot of the project management module, which contains all sample meta-information and information of experimental conditions, while Supplementary Figures S2–S6 are the screenshots of data analysis components.

3.1 Analysis of spike-in samples

A total of 30 samples were analyzed on GC×GC-TOF MS. Manual review of the peak lists generated by ChromaTOF shows that the 28 spiked-in metabolites were correctly identified in every sample. Of the 28 spiked-in metabolites, 23 metabolites were originally present in mouse liver extract. Therefore, the concentrations of these 23 metabolites are different from the other 5 spiked-in metabolites in each sample group.

About 300 peak entries in each sample were assigned to a metabolite name via mass spectral matching. For retention index matching, the confidence interval of the empirical DF of the absolute retention index deviation to its mean value was set as $P \leq 0.001$. In all, 20.7% of the identified metabolites were confirmed by the first-dimension retention index matching and 53.5% were also preserved due to the lack of retention index information in the NIST11 retention index database. However, 25.8% of the mass spectral matched metabolites were removed owing to the large first-dimension retention index deviation from the database values. The mean absolute deviation between the database value and experimental retention index of the removed metabolites is 308 index units (i.u.), with a standard deviation of 257 i.u. The corresponding values of the preserved metabolites are 28 ± 12 i.u. All 28 spiked-in metabolites are preserved in each of the 30 samples after the retention index matching. Metabolites butyric acid and L-tryptophan were preserved owing to the lack of retention index values in the NIST11 database, while the others were preserved by correct retention index match.

About 200–400 peaks were left in each sample for alignment. Peak picking is likely the primary cause for such variation. By manual validation, all 28 spiked-in metabolites were present in all 30 samples with a relative standard deviation (RSD) of the first and second dimensional retention times of 0.07% and 0.55%, respectively, demonstrating a good stability of the instrument during the 2D GC separation and a high accuracy of ChromaTOF in determining peak location in the 2D gas chromatograms.

Figure 2 depicts the peak area distribution of the 28 spiked-in metabolites in G10 before and after peak area conversion. A large variation in each box represents the peak area variation of the same compound between the six samples. By design, each compound should have the identical peak area among the six samples. However, ChromaTOF reports only the peak area of a quant mass (a fragment ion that has the highest quality for quantification), and it often chooses different quant mass for the same metabolite, depending on the data. Of the 28 metabolites, 10 metabolites in at least one of the six samples have quant mass different from the other samples. For instance, compound 4 (oxalic acid) in Figure 2a has 2 quant mass values and the value of peak area has a large deviation, ranging from 1 041 108 to 3 206 235. After reference spectrum-based peak area conversion, the span of peak area is greatly reduced to 1 329 043–1 935 698 (Fig. 2b), demonstrating the effectiveness of the reference spectrum-based peak area conversion. The converted peak area values not only reduced the variation in peak area, but also guarantee the consistency of the values of peak area across samples for the downstream statistical analysis.

To demonstrate whether the concentration differences of the spiked-in metabolite standards can be recognized from the alignment table, a two-tailed *t*-test was used to check the mean difference of the peak area of each compound between sample groups through different *p*-value settings. The true-positive rate (TPR) and the false-positive rate (FPR) are used as measures for metabolite relative quantification (Supplementary Information). Figure 3 depicts the receiver–operating characteristic curve (ROC) of recognizing the concentration difference of the spiked-in metabolite standards between sample groups using

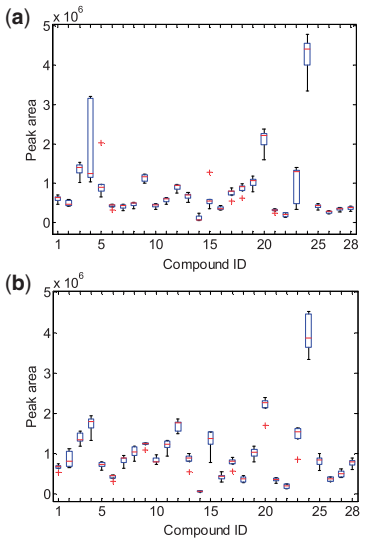

Fig. 2. Boxplot of peak area of the 28 spiked-in metabolites in G10 before reference spectrum-based peak area conversion (**a**) and after the reference spectrum-based peak area conversion (**b**). + refers to the values of peak area considered as outliers

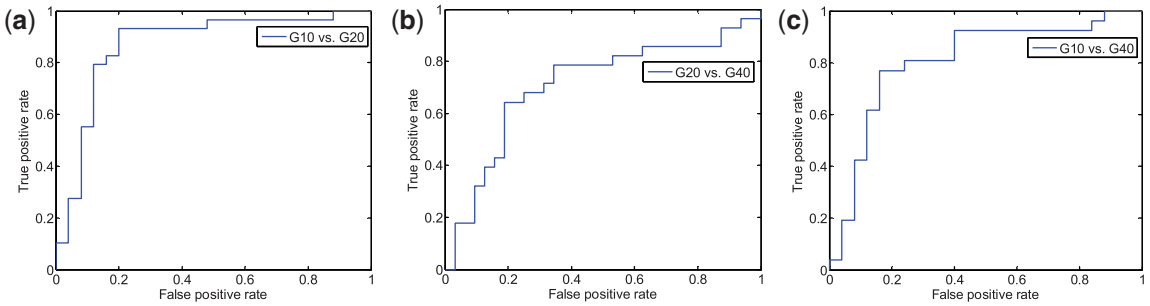

Fig. 3. The ROC curves of recognizing the concentration difference of the 28 spiked-in metabolites between sample groups (**a**) G10 and G20, (**b**) G20 and G40 and (**c**) G10 and G40 using the alignment results of MetPP software

the alignment results. As expected, the FPR increases with the increase of TPR. The TPR levels off at 1.0 when the FPR reaches 0.9 between all the comparing sample groups. The AUC of the ROC curve of G10 vs G20 is 0.87, while the AUC of the ROC curve of G20 vs G40 and G10 vs G40 is 0.71 and 0.81, respectively. A high value of AUC indicates a high accuracy of recognizing the concentration difference of the spiked-in compounds between sample groups, which is achieved on the basis of correct alignment of the spiked-in compounds. Moreover, it is worth mentioning that the spiked-in metabolites that were not recognized as metabolites with significant concentration difference between the two testing sample groups are all present in the mouse liver extract before the addition of the mixture of authentic standards. This means that the concentration differences of these metabolites between sample groups may be much smaller than 2:1 or 4:1, depending on the amount of the metabolites in each sample of mouse liver extract.

3.2 Analysis of biological samples

About 600 metabolites were identified in each mouse liver sample via mass spectral matching. By setting $P \leq 0.001$ for retention index matching, 18.6% of the identified metabolites were confirmed by the first dimension retention index matching and 51.8% were preserved owing to the lack of retention index information in the NIST11 database. However, 29.5% of the mass spectral matched metabolites were removed owing to the large first-dimension retention index deviation from the database values. The mean absolute deviation between the database value and experimental retention index of the removed metabolites is 318 i.u., with a standard deviation of 267 i.u. The corresponding values of the preserved metabolites are 21 ± 16 i.u.

After the retention time filtering, about 300–500 peaks were left in each sample for alignment. Among the aligned peaks, the maximum values of the RSD for the first dimensional retention time and the second dimensional retention time were only 3.15 and 3.30%, respectively.

Table 1 lists all the metabolites detected with significant abundance changes between sample groups LFD and HFD. Compared with the LFD group, the abundances of 10 metabolites were increased and 5 metabolites were decreased in the HFD group. Figure 4 depicts the abundance distribution of metabolite L-threonine in the samples of LFD group and HFD group.

Three free fatty acids (entries 1–3 in Table 1) were detected with significant abundance changes between the two testing sample

Table 1. Metabolites with significant abundance difference between sample groups HFD and LFD

Name	*p*-value	^1t_R(s)	^2t_R(s)	CAS	Fold change[a]
Dodecanoic acid	7.2×10^{-2}	2089.3	1.204	143-07-7	1.3
Tetradecanoic acid	1.4×10^{-2}	2323.8	1.228	544-63-8	1.5
Pentadecanoic acid	1.4×10^{-4}	2432.6	1.243	1002-84-2	2.3
L-valine	1.2×10^{-1}	1782.5	1.148	72-18-4	−1.2
L-cysteine	3.0×10^{-5}	2440.0	1.262	7048-04-6	1.7
L-threonine	2.4×10^{-2}	1900.5	1.208	72-19-5	2.1
2,3-Dihydroxypropanoic acid[b]	1.3×10^{-1}	2153.2	1.153	473-81-4	−1.6
Citric acid[b]	8.4×10^{-2}	2911.1	1.593	77-92-9	1.9
2-Hydroxybutyric acid	6.9×10^{-2}	1642.4	1.135	600-15-7	−1.4
Glycine	1.1×10^{-1}	1655.0	1.159	56-40-6	−1.2
L-lysine	5.2×10^{-4}	2625.7	1.272	56-87-1	−2.5
Pentanoic acid, 3-methyl-4-oxo-[b]	3.6×10^{-2}	2205.1	1.363	6628-79-1	1.2
L-methionine	1.3×10^{-2}	2177.5	1.280	63-68-3	1.4
L-glutamic acid	2.4×10^{-2}	2510.0	1.244	56-86-0	1.5
Taurine	3.8×10^{-6}	2117.5	1.351	107-35-7	4.5

[a]The sample group LFD is the reference group. '+' sign refers to abundance increase in sample group HFD, while '−' sign refers to abundance decrease in HFD group.
[b]Tentative identification without technical verification using authentic standards.

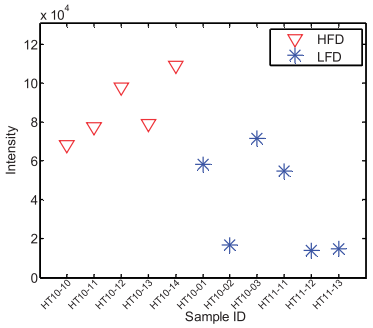

Fig. 4. Abundance distribution of metabolite L-threonine in the samples of LFD group and HFD group. The abundance test (pairwise two-tailed *t*-test) shows that the regulation of this metabolite in the HFD group is increased with a fold change of 2.10 and a *p*-value of 2.4×10^{-2} comparing with LFD group

groups. The levels of these free fatty acids are all increased in mice fed a HFD. This observation agrees with the results of literature-reported hepatic lipid level test (Tan *et al*., 2011).

4 DISCUSSION

MetPP uses the peak lists generated by peak deconvolution software, such as commercial software ChromaTOF, as its input. It is well known that the ChromaTOF software has some shortcomings, including multiple peak table entries for the same compound, limited accuracy in calculation of peak areas, etc. To minimize these shortcomings of ChromaTOF software, MetPP provides options for users to filter data from the peak lists generated by ChromaTOF using the peak quality information (profile purity, purity, concerns) provided by ChromaTOF. The multiple peak entries of the same compound are recognized based on user defined 2D retention time variation windows. These peak entries are then merged and represented by one representative peak entry. To handle peaks detected in blank sample(s), MetPP has three options for the user to select from: no action, deducting the peak area of blank sample(s) from the biological samples and removing the metabolites. The user can also remove metabolites using a user-provided exclusion list. In this study, the default values for peak quality-based filtering and merging multiple peak entries were used. The analysis of spike-in data demonstrates that the data preprocessing methods implemented in MetPP is able to remove some of the shortcomings of ChromaTOF software without affecting identification and quantification of the spiked-in standards.

To reduce the rate of false-positive identification, the current version of MetPP provides an option for users to match the first-dimension retention index of each compound with the database values. A large value of the mean absolute deviation between the database value and experimental retention index of the removed metabolites (308 ± 257 i.u. for the spike-in data, 318 ± 267 i.u. for the biomarker discovery data) and a small value of the corresponding values of the preserved metabolites (28 ± 12 i.u. for the spike-in data, 21 ± 16 i.u. for the biomarker discovery data) demonstrates the mass spectral matching-based metabolite identification can introduced a high rate of false identification, and the retention index filtering methods in MetPP can detect and remove some of these false-identified metabolites.

MetPP performs peak list alignment using a mixture score function to simultaneously evaluate the deviation of the 2D retention time and mass spectral similarity. Compared with the progressive retention time map searching method implemented in the distance and spectrum correlation optimization (DISCO) algorithm (Wang *et al*., 2010), the mixture score approach is able to align chromatographic peaks that may have a large variation in either retention time or mass spectral similarity, as long as the overall quality of two peak are similar. Supplementary Figure S7 depicts the alignment results of the spike-in experimental data by DISCO and MetPP. All 28 spiked-in metabolites were fully aligned in all 30 samples by MetPP. However, 23 metabolite standards were fully aligned by DISCO, while the rest of the spiked-in metabolites were aligned in a portion of the 30 samples. In terms of metabolite quantification, the data analysis algorithms implemented in MetPP are able to correctly identify and quantify all spiked-in metabolite standards.

Analysis of biological samples using MetPP revealed that 15 metabolites have significant abundance changes between two sample groups. The accuracy of clustering samples using all metabolite data in each sample is 0.82, while the clustering accuracy is improved to 0.91 when the abundance of the 15 metabolites in each sample are used for clustering. The improvement in sample clustering accuracy demonstrates that MetPP correctly recognized the metabolites with significant abundance difference between sample groups LFD and HFD. The abundance changes of the free fatty acids agree with the literature results of hepatic lipid level test, a different analytical method. These analysis results further demonstrate that the MetPP software can be used to recognize the metabolites with significant abundance changes between sample groups.

The performance of MetPP in analysis the spike-in data was also compared with the existing Guineu software (Castillo *et al*., 2011) in terms of accuracy of alignment and relative quantification. Supplementary Figure S8 shows that only 9 of the 28 spiked-in metabolites were fully aligned by Guineu, while all the 28 spiked-in metabolites were fully aligned by MetPP. For relative quantification, MetPP also significantly outperforms Guineu in all three statistical measures of TPR, PPV and their harmonic mean F1 for analysis of all pairwise comparison between the five sample groups (Supplementary Fig. S9).

Analysis of both the spike-in data and biomarker discovery data demonstrates that the methods implemented in MetPP for peak filtering and merging are effective in minimizing the shortcomings of ChromaTOF. However, compound 24 (nonanoic acid) in Figure 2a still has a large variation in peak area after the peak filtering and merging. For the future development, accurate peak deconvolution algorithms are needed to improve the quality of peak lists. A large variation in peak area introduced during peak deconvolution can significantly affect the downstream statistical analysis. It is also important to develop algorithms for analysis of time course data and construction of metabolite correlation networks.

5 CONCLUSIONS

A computational platform for analysis of metabolomics data generated on GC×GC-TOF MS instrumentation was developed in the form of MetPP using MATLAB. MetPP uses the peak lists generated by a spectral deconvolution software such as commercial software ChromaTOF as its input and then performs peak merging and filtering, retention index matching, peak list alignment, normalization, statistical significance tests and pattern recognition. Two sets of experimental data acquired on GC×GC-TOF MS were used to validate the performance of MetPP. The analysis results demonstrate that MetPP is able to recognize the concentration difference of the spiked-in metabolite standards between sample groups. The analysis results of biological samples agree with the hepatic lipid level test, further demonstrating that MetPP is able to correctly process the data of complex samples.

ACKNOWLEDGEMENTS

The authors thank Mrs Marion McClain for review of this manuscript. We thank Drs Mark Merrick, Wei Chen, Peter

Willis and Jihong Wang at LECO Corporation for their help on using ChromaTOF software. The authors also thank Dr Mark Styczynski's research group at Georgia Institute of Technology for their suggestions of improving the software.

Funding: This work was supported by National Institute of Health (NIH) grants 1RO1GM087735, R21ES021311, P01AA017103, R37AA010762 and 1RC2AA019385, and the VA.

Conflict of Interest: none declared.

REFERENCES

Almstetter,M.F. *et al.* (2009) Integrative normalization and comparative analysis for metabolic fingerprinting by comprehensive two-dimensional gas chromatography-time-of-flight aass spectrometry. *Anall. Chem.*, **81**, 5731–5739.

Astrand,M. (2001) Normalizing oligonucleotide arrays. http://www.stat.berkeley.edu/~terry/zarray/Affy/GL_Workshop/Astrand_manuscript.pdf (7 October 2011, date last accessed).

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, **57**, 289–300.

Bezdek,J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Castillo,S. *et al.* (2011) Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Anal. Chem.*, **83**, 3058–3067.

Curran-Everett,D. (2000) Multiple comparisons: philosophies and illustrations. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **279**, R1–R8.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, **12**, 29.

Fraga,C.G. *et al.* (2001) Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions. *Anal. Chem.*, **73**, 5833–5840.

Hoggard,J.C. *et al.* (2009) Toward automated peak resolution in complete GC x GC-TOFMS chromatograms by PARAFAC. *J. Chemometr.*, **23**, 421–431.

Jolliffe,I.T. (2002) *Principal Component Analysis*. Springer, New York.

Kim,S. *et al.* (2011) An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure. *Bioinformatics*, **27**, 1660–1666.

Kim,S. *et al.* (2012) A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry. *Bioinformatics*, **28**, 1158–1163.

Koo,I. *et al.* (2011) Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Anal. Chem.*, **83**, 5631–5638.

Kovats,E. (1958) Gas-chromatographische charakterisierung organischer verbindungen. Teil 1: retentionsindices aliphatischer halogenide, alkohole, aldehyde und ketone. *Helv. Chim. Acta.*, **41**, 1915–32.

Pierce,K.M. *et al.* (2005) A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data. *Anal. Chem.*, **77**, 7735–7743.

Reichenbach,S.E. *et al.* (2005) Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry. *J. Chromatogr. A*, **1071**, 263–269.

Reichenbach,S.E. *et al.* (2012) Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. *J. Chromatogr. A*, **1226**, 140–148.

Rosipal,R.K.N. (2006) Overiew and recent advances in partial least squares. In: *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS 2005)*. Springer-Verlag, Berlin, Germany, pp. 34–51.

Sinha,A.E. *et al.* (2004) Trilinear chemometric analysis of two-dimensional comprehensive gas chromatography-time-of-flight mass spectrometry data. *J. Chromatogr. A*, **1027**, 269–277.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.*, **64**, 479–498.

Tan,M. *et al.* (2011) Chronic subhepatotoxic exposure to arsenic enhances hepatic injury caused by high fat diet in mice. *Toxicol Appl. Pharmacol.*, **257**, 356–364.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

van Mispelaar,V.G. *et al.* (2003) Quantitative analysis of target components by comprehensive two-dimensional gas chromatography. *J. Chromatogr. A*, **1019**, 15–29.

Vivo-Truyols,G. (2012) Bayesian approach for peak detection in two-dimensional chromatography. *Anal. Chem.*, **84**, 2622–2630.

Wang,B. *et al.* (2010) DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Anal. Chem.*, **82**, 5069–5081.

Wei,X. *et al.* (2011) MetSign: a computational platform for high-resolution mass spectrometry-based metabolomics. *Anal. Chem.*, **83**, 7668–7675.

Zhang,D.B. *et al.* (2008) Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data. *Anal. Chem.*, **80**, 2664–2671.

Zhang,J. *et al.* (2011) iMatch: a retention index tool for analysis of gas chromatography-mass spectrometry data. *J. Chromatogr. A*, **1218**, 6522–6530.