Journal of Chromatography A, 1394 (2015) 111–117

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

# Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography–mass spectrometry data

Heather D. Bean [a], Jane E. Hill [a], Jean-Marie D. Dimandja [b,*]

[a] Dartmouth College, Thayer School of Engineering, Hanover, NH, USA
[b] Spelman College, Department of Chemistry and Biochemistry, Atlanta, GA, USA

ARTICLE INFO

*Article history:*
Received 8 December 2014
Received in revised form 24 February 2015
Accepted 1 March 2015
Available online 7 March 2015

*Keywords:*
Comprehensive two-dimensional gas chromatography
GC × GC–MS
Data alignment
Untargeted metabolomics
Biomarker

ABSTRACT

The potential of high-resolution analytical technologies like GC × GC/TOF MS in untargeted metabolomics and biomarker discovery has been limited by the development of fully automated software that can efficiently align and extract information from multiple chromatographic data sets. In this work we report the first investigation on a peak-by-peak basis of the chromatographic factors that impact GC × GC data alignment. A representative set of 16 compounds of different chromatographic characteristics were followed through the alignment of 63 GC × GC chromatograms. We found that varying the mass spectral match parameter had a significant influence on the alignment for poorly-resolved peaks, especially those at the extremes of the detector linear range, and no influence on well-chromatographed peaks. Therefore, optimized chromatography is required for proper GC × GC data alignment. Based on these observations, a workflow is presented for the conservative selection of biomarker candidates from untargeted metabolomics analyses.



## 1. Introduction

Samples of biological origin typically contain hundreds to thousands of compounds that range widely in chemical properties and concentrations, making separation and characterization of the mixtures challenging. Multidimensional chromatography (MDC) techniques, such as comprehensive two-dimensional gas chromatography (GC × GC), are powerful analytical tools that are well suited for the analysis of biological mixtures due to the enhanced peak capacity afforded by the additional chromatographic dimensions [1,2]. The utility of GC × GC for targeted analyses of biological samples as well as compound discovery in complex matrices, including untargeted metabolomics [3,4], is well recognized and applicable to many areas of inquiry. For example, Hartman et al. [5] report using GC × GC in targeted analyses to quantify 3,4-methylenedioxymethamphetamine and three of its metabolites in blood and serum, and in an untargeted metabolomics analysis, Cordero and colleagues applied GC × GC to the detection of previously unidentified volatile metabolites from the leaves of *Mentha* species [6]. Turning the large data sets that are generated by MDC methods into information (e.g., quantifiable peaks and compound identities) is becoming less time-intensive and more reliable through advancements in commercial software packages (e.g., ChromaTOF, GCImage, and ChromSquare), as well as new unsupervised data processing and statistical analysis methods that are being developed or retooled for MDC (recently reviewed in [7–12]). In addition, a number of chromatographic alignment methods for comparing data between multiple GC × GC analyses, a cornerstone of biomarker discovery, have been described and employed with great success [13–19]. However, many of these algorithms have been developed and validated for targeted analyses, which only requires local alignment of the peaks of interest [20], and importantly, commercial or public availability of these programs are lagging [21].

The aim of untargeted biomarker analysis is to identify robust and predictive differences between sample classes, e.g., between diseased and healthy individuals, without a priori knowledge of the metabolism [22]. MDC separations provide a significant advantage in the search for biomarkers by increasing the number of identifiable and quantifiable metabolites by approximately 10-fold [23,24], but the resulting large volumes of new data have now shifted the burden of biomarker discovery to the data processing and statistical analysis steps [21]. One common approach for identifying putative biomarkers is to calculate Fisher ratios – an assessment of

* Corresponding author. Tel.: +1 404 270 5743.
*E-mail address:* jdimandja@spelman.edu (J.-M.D. Dimandja).

http://dx.doi.org/10.1016/j.chroma.2015.03.001
0021-9673/

the between-class vs. within-class variance of each compound in the samples – where the compounds with higher Fisher ratios are pursued as candidate biomarkers [25]. Accurate variance calculations rely upon the alignment of each chromatographic peak across every sample [25,26], but the inherent variance within biological sample classes requires the analysis of a large number of samples, thus making alignment challenging [16,26]. The importance of good chromatography (i.e., optimized peak resolution) for targeted and untargeted MDC analyses is well understood [9]; Kim and Zhang [27] have reported on the influence of peak density on GC × GC peak-table-based alignment algorithms. However, the influence of peak intensity and resolution on the outcomes of chromatographic alignment has not been reported on a peak-by-peak basis, which has significant implications for the development of unsupervised data processing and statistical methods for biomarker discovery from MDC analyses.

In this study, we assess how chromatographic peak characteristics influence the success of peak table-based alignment. Employing a commonly-used column set [3], we performed 63 GC × GC–TOFMS analyses on the volatile metabolites of closely-related bacterial isolates, and aligned the chromatograms using ChromaTOF Statistical Compare (LECO Corp.). We varied the mass spectral match parameter in the alignment algorithm to look at the interplay between peak intensity and resolution as they affect alignment, and evaluated the success of the alignment by following 16 peaks with chemical and chromatographic characteristics (e.g., area, S/N, resolution, and retention times) that are representative of the diversity of peaks observed in these complex biological samples. Based on our observations, we propose a workflow to identify high-quality biomarker candidates from less-than-perfect separations of complex samples, the concepts of which can be applied to the alignment of MDC data collected on any software platform.

## 2. Experimental

### 2.1. Sample preparation and volatiles collection

Thirty-five clinical isolates of *Pseudomonas aeruginosa* were used for this study. Between 1 and 3 biological replicates were analyzed for each isolate, yielding 63 samples. For bacterial volatiles analysis, 10 mL of spent media from stationary phase cultures and a stir bar were sealed into 20 mL glass vials with PTFE/silicone caps. The volatile metabolites of the bacteria were sampled from the headspace using solid-phase microextraction (SPME; divinylbenzene/carboxen/polydimethylsiloxane, 50/30 μm; Supelco/Sigma–Aldrich, St. Louis, MO), as described previously [1].

### 2.2. Chromatography and mass spectrometry

Two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC–TOFMS) was performed using a LECO Pegasus 4D (St. Joseph, MI). The instrument was fitted with a two-dimensional column set consisting of a DB-5MS (5% diphenyl/95% dimethyl polysiloxane; 30 m × 0.25 mm × 0.25 μm (length × internal diameter × film thickness); Agilent Technologies) as the first dimension ($^1$D) column, and a ZB-50 (50% diphenyl/50% dimethyl polysiloxane; 2 m × 0.1 mm × 0.1 μm; Phenomenex) as the second dimension ($^2$D) column, joined by a press-fit connection. The columns were heated independently; the $^1$D column was initiated at 35 °C (0.2 min hold), then heated at 15 °C/min to 230 °C (0.8 min hold); the $^2$D column was heated with a +5 °C offset relative to the primary oven. A quad-jet modulator was used with a 4 s modulation period (0.4 s hot, 1.6 s cold pulses) and a +25 °C temperature offset relative to the secondary oven. The helium carrier gas flow rate was 1 mL/min. A 10:1, 30 s pulsed split injection was used. The inlet and transfer line temperatures were 250 °C. Mass spectra were acquired at 200 Hz over the range of $m/z$ = 25–500. Data acquisition was performed using ChromaTOF software (LECO Corp.), v.4.22.

### 2.3. Data processing and chromatographic alignment

Data processing and chromatographic alignment were performed using the Statistical Compare package of ChromaTOF v.4.50. The baseline was drawn through the middle of the noise and the signal-to-noise (S/N) cutoff for peak finding was set to 10 for a minimum of 2 apexing masses. The $^1$D and $^2$D peak widths were set to 8 and 0.15 s, respectively, based on the observed widths in the chromatograms for non-saturated peaks. ChromaTOF combined subpeaks across multiple injections into the secondary column when the second dimension retention time ($^2t_R$) shift was ≤100 ms early for subsequent modulation periods, and the mass spectral match was ≥600. Peaks were identified by a forward search of the NIST 08 Mass Spectral Library.

Chromatographic alignment was performed using ChromaTOF Statistical Compare v.4.50. For a peak to be identified as the same compound across chromatograms, both the retention times and the mass spectra had to meet minimum match criteria. For alignment, the first dimension retention time ($^1t_R$) could not vary more than 4 s (1 modulation period) from chromatogram to chromatogram and the second dimension retention time ($^2t_R$) could not vary more than 100 ms, based on the observed maximum variability in $^1t_R$ and $^2t_R$ for quinolone, an exogenous retention time marker added to each sample ($^1t_R$ = 744 s, $^2t_R$ = 2.04–2.12 s). The mass spectrum for aligned peaks had to meet a minimum inter-chromatogram match threshold, which was varied for each analysis from 100 to 900, in increments of 100, to evaluate the influence of spectral match on alignment success. Alignment data for 16 peaks (Table 1) were evaluated for each of the nine experiments, yielding 144 experimental results. A peak was deemed to be aligned if all occurrences of the peak from the 63 individual chromatograms had been grouped into a single peak table entry in the alignment results, and the peak could be aligned using at least two mass spectral match score minima. As an additional measure of alignment, Fisher ratios were calculated for each set of aligned peaks using two sample groups. A minimum of two peaks in each group were required for the calculation, otherwise the Fisher ratio was reported as "Undefined." Chromatographic alignment for a peak across multiple mass spectral match scores yields the same Fisher ratios for each alignment result.

## 3. Results and discussion

### 3.1. Chromatography: characteristics and alignment

The headspace of *P. aeruginosa* bacterial cultures contain hundreds of volatile metabolites, which are highly varied in chemical class [1] and concentration, making good chromatography (i.e., Gaussian peak shapes, high peak resolution) of every peak challenging. Using a common non-polar/semi-polar column combination for the $^1$D and $^2$D columns, respectively, and a 10:1 split ratio to enhance the detection of trace compounds, we obtained good separations for the *P. aeruginosa* volatiles, except for the low-boiling point compounds, where the chromatographic space was congested ($^1t_R$ < 420 s, Fig. 1). Peak deconvolution algorithms, such as the one employed by ChromaTOF, are able to find many imperfectly-resolved peaks by reconstructing mass spectra from the apexes of individual ions, which significantly enhances compound detection and discovery in targeted and untargeted analyses, respectively. However, alignment of poorly-resolved peaks across