Journal of Chromatography A, 1218 (2011) 7031–7038

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Comparison of two algorithmic data processing strategies for metabolic fingerprinting by comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry

Martin F. Almstetter[a,*,1], Inka J. Appel[a,1], Katja Dettmer[a], Michael A. Gruber[b], Peter J. Oefner[a]

[a] Institute of Functional Genomics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany
[b] Department of Anesthesiology, University Hospital Regensburg, Franz-Josef-Strauss-Allee 11, 93053 Regensburg, Germany

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form 15 July 2011
Accepted 5 August 2011
Available online 12 August 2011

Keywords:
Alignment
Data mining
GC × GC–TOFMS
Metabolic fingerprinting
Metabolomics
Spike-in experiment

ABSTRACT

The alignment algorithm Statistical Compare (SC) developed by LECO Corporation for the processing of comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC–TOFMS) data was validated and compared to the in-house developed retention time correction and data alignment tool INCA (Integrative Normalization and Comparative Analysis) by a spike-in experiment and the comparative metabolic fingerprinting of a wild type versus a double mutant strain of *Escherichia coli* (*E. coli*). Starting with the same peak lists generated by LECO's ChromaTOF software, the accuracy of peak alignment and detection of 1.1- to 4-fold changes in metabolite concentration was assessed by spiking 20 standard compounds into an aqueous methanol extract of *E. coli*. To provide the same quality input signals for both alignment routines, the universal *m/z* 73 trace of the trimethylsilyl (TMS) group was used as a quantitative measure for all features. The performance of data processing and alignment was evaluated and illustrated by ROC curves. Statistical Compare performed marginally better at the lower fold changes, while INCA did so at the higher fold changes. Using SC, quantitative precision could be improved substantially by exploiting the signal intensities of metabolite-specific unique (U) *m/z* ion traces rather than the universal *m/z* 73 trace. A list of 56 features that distinguished the two *E. coli* strains was obtained by the SC alignment using *m/z* U with an estimated false discovery rate (FDR) of <0.05. Ultimately, 23 metabolites could be identified, one additional and five less than with INCA due to the failure of SC to extract unitized *m/z* U's across all fingerprints with suitable spectral intensities for the latter metabolites.

1. Introduction

Metabolomics aims at the comprehensive quantitative analysis of metabolites in a biological system as influenced by environmental, nutritional, genetic, and other factors. However, realization of this goal in a single analysis is impeded by the large number of chemically diverse metabolites present over a wide range of concentrations [1,2].

Comprehensive two-dimensional gas chromatography (GC × GC) is highly suited to the separation of low-molecular weight components of complex samples [3,4]. Its greatest benefits are enhanced resolution, a multiplicative increase in peak capacity by combining two columns with orthogonal separation properties via thermal modulation, increased height of 2nd dimension peaks caused by thermal modulation resulting in lower limits of detection, and a structured separation space [5]. Coupled to an electron ionization (EI) fast acquisition time-of-flight mass spectrometer (TOFMS) for the identification and quantification of analytes, GC × GC has been applied successfully to metabolite fingerprinting [6–9].

GC × GC–TOFMS metabolic fingerprints enable the comparative characterization of samples as the entire available (non-targeted) information from all experiments is taken into account for subsequent statistical analysis, thereby facilitating the detection of novel biomarkers. This requires the reliable and preferably automated recognition of metabolites over many GC × GC–TOFMS analyses.

Different solutions have been presented in recent years for the alignment and processing of GC × GC–TOFMS spectra. Shellie et al. [10] applied a combination of chromatogram subtraction, averaging routines, weighting factors and a Student's *t*-test to directly compare GC × GC–TOFMS metabolite fingerprints of mouse tissue extracts against a reference chromatogram making use of the

* Corresponding author. Tel.: +49 941 9435001; fax: +49 941 9435020.
E-mail address: martin.almstetter@klinik.uni-regensburg.de (M.F. Almstetter).
[1] These authors contributed equally to this work and therefore should be considered equal first authors.

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.08.006

compare function in the ChromaTOF software by LECO Corp. (St. Joseph, MI). As chromatograms differ, every sample has to serve as a reference, which leads to computer-intensive comparisons. The parallel factor analysis (PARAFAC) algorithms [11,12] use raw higher order data as input, which can be reduced by a fully automated Fisher ratio method [13], for deconvolution, alignment, and quantification. Fraga et al. [14] and van Mispelaar et al. [15] proposed algorithms to correct retention time (RT) variations in comprehensive two-dimensional separations, but both methods can only be applied to small regions of interest in the 2D data sets. An RT correction over the entire chromatogram in both separation dimensions was achieved by Pierce et al. [16] by means of an indexing scheme and Zhang et al. [17] via a correlation optimized warping (COW) algorithm. While the piecewise alignment algorithm of Pierce et al. only allows simple scalar shifts, the COW algorithm interpolatively stretches and compresses local regions to maximize the correlation between the warped and the reference chromatogram. Alternatively, Schmarr and Bernhardt [18] applied an image processing approach that is routinely used in the comparison of two-dimensional protein gels for the compensation of run-to-run variations and peak merging. These methods can align GC × GC–TOFMS data based on first- and second-dimension retention times, albeit a metabolite is additionally characterized by its specific mass spectrum, which is easily accessible.

Most of the approaches cited above use raw instrument data only and ignore available data preprocessing functions, such as automated baseline correction, mass spectral deconvolution, peak picking, integration, library search and signal/noise filtering, all of which are provided, for instance, by the ChromaTOF software. Oh et al. [19] developed the MSort software for GC × GC–TOFMS that uses data processing provided by the ChromaTOF software to generate peak tables. The algorithm utilizes first- and second-dimension retention times and mass spectra correlation for peak sorting. However, the need to define RT windows of a fixed size impairs the software's utility. The algorithm is only able to handle small and linear RT distortions and becomes prohibitively computationally intensive in the handling of large-scale metabolomic datasets. For large RT shifts, Wang et al. [20] recently proposed a distance and spectrum correlation optimization (DISCO) algorithm that uses the Euclidean distances of 2D retention times and mass spectrum correlation for peak alignment. The method allows alignment of data acquired under various experimental conditions.

In 2009, we developed retention time correction and data alignment tools (INCA) for comparative metabolite fingerprinting of *Escherichia coli* (*E. coli*) strains using peak lists generated by the ChromaTOF software [21]. To account for shifts in first- and second-dimension retention times, two independent linear models were fitted to a series of measurements using odd-numbered fatty acids as reference substances. The peak lists were then aligned in one data matrix based on retention times and EI mass spectral information for subsequent multivariate statistical analysis. The signal intensity of the trimethylsilyl ion at *m/z* 73 was used to quantify differences in metabolite abundance.

The aim of the current study was to compare the performance of INCA with the Statistical Compare (SC) alignment tool implemented in ChromaTOF software version 4. Both alignment tools used the same peak lists and signal intensities of the *m/z* 73 traces as quantifier for every feature and were validated by the same spike-in experiment. Additionally, SC was evaluated employing the signal intensities of metabolite-specific unique (U) *m/z* ions for quantification instead of the universal *m/z* 73 trace and applied to the comparative GC × GC–TOFMS metabolite fingerprinting of the two different *E. coli* strains used in Almstetter et al. [21]. Multivariate statistical analysis was performed in order to evaluate the respective enhancements and limitations of the two algorithms.

2. Experimental

2.1. Instrumentation

The evaluation of the SC alignment tool was accomplished using the raw data generated by Almstetter et al. [21] on a LECO Pegasus 4D GC × GC–TOFMS instrument that had been equipped with an MPS-2 Prepstation sample robot (Gerstel, Muehlheim, Germany) for automated sample derivatization and handling. Briefly, an Rxi®-5ms column (30 m × 0.25 mm ID × 0.25 μm film thickness) from Restek (GmbH, Bad Homburg, Germany) was used as the first-dimension column, while an Rtx-1701 (2 m × 0.1 mm ID × 0.1 μm film thickness, Restek) served as the second-dimension column. The oven temperature was held initially at 50 °C for 0.2 min, raised at 8 °C/min to 265 °C, and held for 10 min. A positive offset of 5 °C was used for the second-dimension column and a 15 °C offset relative to the first-dimension column for the modulator. The column flow (constant) was 1 mL He/min. A sample volume of 1.5 μL was injected in splitless mode. A 4-s modulation period was used and mass spectra were acquired from *m/z* 40–600 at a rate of 100 spectra/s.

2.2. Spike-in experiment

Hundred microliter of an 80% aqueous methanol extract of *E. coli* BL21 was fortified with 0.25, 0.275, 0.3125, 0.375, 0.5, 0.625, and 1.0 nmol, respectively, of a mixture of 20 metabolites (adipate, decanoate, dimethylsuccinate, dodecanoate, eicosanoate, erythritol, 2-hydroxybutyrate, 3-hydroxybutyrate, 2-hydroxy-3-methylbutyrate, malonate, mandelate, mannitol, 3-methyl-2-oxovalerate, nicotinate, phenylacetate, phenyllactate, suberate, triethanolamine, vannilate, and xylitol) [21].

2.3. Metabolite fingerprinting

The metabolic fingerprints of the *E. coli* wild-type MG1655 and the *E. coli* MG1655 *double*-mutant Δ*UdhA*–Δ*PntAB* were generated by cultivating each strain in three separate flasks and by filtrating each culture in triplicates. Cells were harvested by fast filtration [22] after reaching stationary phase growth and extracted using 80% (v/v) methanol in deionized water at −20 °C. Nine samples per strain and three medium blanks were measured in random order to avoid a systematic error using GC × GC–TOFMS.

2.4. Data processing

Raw data were processed with the ChromaTOF software version 4.32. Regions of the 2D chromatogram containing excessive noise were excluded. Baseline correction, deconvolution and peak picking were performed. Signals exceeding a predefined *S/N* of 500 and for 2nd dimension subpeaks an *S/N* of 50 were selected and combined in the second-dimension separation using a spectral matching factor of 700. The separate subpeak *S/N* is an improvement over previous versions of ChromaTOF. An override of the allowed 2nd dimension retention time shift was set at 0.150 s early and 0.050 s late to improve the separation of closely eluting peaks with an identical 1D retention time and a similar 2D RT. For compound identification, commercial standards were run individually and EI spectra were matched against both an in-house library of 150 metabolites, which included sugars, alcohols, fatty acids, organic acids, amino acids, and amino acid metabolites, and the NIST 05 library. Every sample was processed according to parameters selected above generating a corresponding peak table that listed name, retention times, area, similarity, *S/N*, quant mass and unique mass for every compound. Note that unique mass refers to a metabolite-specific unique *m/z* ion trace, which exhibits the least