Analytica Chimica Acta 740 (2012) 12–19

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Instrument and process independent binning and baseline correction methods for liquid chromatography–high resolution-mass spectrometry deconvolution

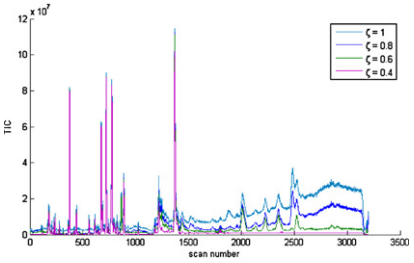
Shaji Krishnan[a,g,*], Jack T.W.E. Vogels[b], Leon Coulier[b], Richard C. Bas[c], Margriet W.B. Hendriks[d,g], Thomas Hankemeier[e,g], Uwe Thissen[f]

[a] TNO Research Group Microbiology & Systems Biology, Utrechtseweg 48, 3700 AJ Zeist, The Netherlands
[b] TNO Research Group Quality & Safety, Utrechtseweg 48, 3700 AJ Zeist, The Netherlands
[c] TNO Triskelion B.V., P.O. Box 844, 3700 AV Zeist, The Netherlands
[d] Department of Metabolic and Endocrine Diseases, University Medical Centre Utrecht, Utrecht, The Netherlands
[e] Leiden/Amsterdam Center for Drug Research, Analytical BioSciences, Gorlaeus Laboratories, University of Leiden, Einsteinweg 55, 2333 CC Leiden, The Netherlands
[f] KeyGene N.V., P.O. Box 216, 6700 AE Wageningen, The Netherlands
[g] Netherlands Metabolomics Centre, Einsteinweg 55, 2333 CC Leiden, The Netherlands

HIGHLIGHTS

- Appropriately collating the m/z values over time is an effective machine independent scheme for aggregating a metabolite profile from a mass chromatogram.
- Entropy is an effective metric for baseline correction.
- The efficacy of each was proven on two LC–HR-MS datasets.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 12 October 2011
Received in revised form 17 April 2012
Accepted 11 June 2012
Available online 28 June 2012

Keywords:
Accurate mass binning
Baseline correction
Liquid chromatography–mass spectrometry
Pre-processing
Deconvolution

ABSTRACT

Setting appropriate bin sizes to aggregate hyphenated high-resolution mass spectrometry data, belonging to similar mass over charge (m/z) channels, is vital to metabolite quantification and further identification. In a high-resolution mass spectrometer when mass accuracy (ppm) varies as a function of molecular mass, which usually is the case while reading m/z from low to high values, it becomes a challenge to determine suitable bin sizes satisfying all m/z ranges. Similarly, the chromatographic process within a hyphenated system, like any other controlled processes, introduces some process driven systematic behavior that ultimately distorts the mass chromatogram signal. This is especially seen in liquid chromatogram–mass spectrometry (LC–MS) measurements where the gradient of the solvent and the washing step cycle—part of the chromatographic process, produce a mass chromatogram with a non-uniform baseline along the retention time axis. Hence prior to any automatic signal decomposition techniques like deconvolution, it is a equally vital to perform the baseline correction step for absolute metabolite quantification. This paper will discuss an instrument and process independent solution to the binning and the baseline correction problem discussed above, seen together, as an effective pre-processing step toward liquid chromatography–high resolution-mass spectrometry (LC–HR-MS) data deconvolution.

1. Introduction

Hyphenated mass chromatogram data obtained by gas chromatography–mass spectrometry (GC–MS) or liquid chromatography–mass spectrometry (LC–MS), herein termed

* Corresponding author. Tel.: +31 888662393.
E-mail address: shaji.krishnan@tno.nl (S. Krishnan).

0003-2670/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2012.06.014

as mass chromatogram data are complex data to process, especially in high mass-resolution mode. Hence automatic processing methods like deconvolution [1], and software tools like XCMS [2], MZmine [3], TNO-DECO [4], etc., are necessary to effectively and efficiently process mass chromatogram data. The mass accuracy of high precision mass spectrometers typically varies from 0.5–1.0 ppm for a 100,000 resolution instrument to 3.0–5 ppm for a 100 resolution instrument [5]. This variation is usually non-uniform along the mass chromatogram m/z axis and among measurement samples, and it depends on various instrumental and process factors. This often leads to large amounts of data that is impractical to process automatically. For these reasons, binning is one of the critical steps in the automatic mass chromatogram data processing [6]. Nevertheless, defining a proper form of binning is a tedious task. Forcing a fixed bin size, either larger or smaller, has a disadvantage because a larger bin size may result in poor resolution of the peaks, while a smaller bin size may split a peak [2]. In both cases, the outcome is a noisy peak and the consequence is a poor deconvolution result. As an alternative choice, some authors have proposed variable bin sizes or overlapping bins [7]. One of the problems while applying these methods is the lack of information in the raw data to estimate the parameters for setting these values. Finally, there are also other approaches for analyte quantification that avoid binning [8–10]. However, these approaches are based on peak picking instead of finding metabolite entries. This suggests that additional processing is required to classify the peaks, and thereafter extract the pure mass spectra.

Apart from the random ionization noise introduced by the mass spectrometry instrument, the chromatography step introduces systematic noise that sometimes alters the characteristic form of the mass chromatogram data. This noise results from a number of sources, such as the mobile phases and buffers used for liquid chromatography [11]. Correcting for this systematic noise is a mandatory pre-processing step and the consequence of avoiding this step is a poor quantification result. Obviously, this has been subject of research in the past where most of the baseline correction methods relied on determining a suitable offset that is subtracted from the original data. The offsets are usually polynomials whose coefficients are estimated from the mass chromatogram along the retention time axes [12–15]. Although these methods seem to meet the objective, they do not tackle the problem in a fundamental way. Stated otherwise, the process related systematic unwanted signals (noises) have to be treated from a process perspective rather than fitting a curve from a collection of points. An example where baseline correction is performed from a process perspective, instead of curve fitting, is described in Baggerly et al. [16], where a process induced unwanted component, a sinusoidal noise from the power source, was identified as the cause for a non-uniform baseline and subsequently corrected.

In contrast to the several binning approaches discussed earlier, the approach proposed in this paper will prepare the data for deconvolution and is based on the combination of ideas from Aberg et al. [8] and Stolt et al. [9], using a so-called minimal distance approach. For the mass chromatogram baseline correction, we propose an alternative procedure in which entropy is used as a powerful metric to distinguish metabolite related signals from noise. One of the major advantages of this proposed technique is that it is process and instrument independent, and hence can be applied to any mass chromatogram data. The tool that implements the proposed mass binning, and the baseline correction methods import raw liquid chromatogram–high resolution-mass spectrometry (LC–HR-MS) data files in CDF format to the MATLAB computing environment.

2. Method for mass binning

LC–HR-MS data is a collection of ion current measurements either continuous or discrete (centroid mode) for progressive values of m/z, recorded as a function of (retention) time. Hence an element of LC–MS data is an ion current measurement at a certain m/z value and a certain retention time. As elaborately discussed by Stolt et al. [9], a unique m/z peak, along the m/z axis, can be distinguished from the others (a.k.a. noise or other peaks) based on the features like the location of the centroid and the distance to its neighbors. However, because of (limited) instrumental precision and intensity changes, the m/z value of a peak can vary from one scan to another. Hence it becomes necessary to track the m/z changes over time, and then assign a representative value for those m/z values that belong to the same metabolite. In previous studies, a Kalman filter was used by Aberg et al. [8], to trace the time trajectory of each m/z channel, while Tautenhahn et al. [10] choose a subset of m/z from the high intensity region of a peak, arguing that the m/z value in centroid mode shows lesser variation with intensity.

The method that is proposed in this paper, uses the above-mentioned ideas but in contrast to the Kalman filter approach that uses a less definitive starting point, a more robust selection of such a point is made. This, in combination with a windowing approach, provides more selectivity on the number of m/z channels considered at a time, which effectively decreases the computing load and increases the precision of the method. Another advantage of the method is that it is developed on the so-called divide and conquer strategy which means that during multiple sample mass binning, each sample file is initially treated independently and the result from each file is combined at a later stage to produce one final output ready for multiple sample deconvolution.

Binning a single dataset (sample) begins with a definition for a binning window having an m/z range and a spatial range. The m/z range is defined in terms of unit m/z, while the spatial range is defined in terms of scan numbers (or retention time). In the discussion, and the example to follow, a binning window of 1 unit mass and maximum scan number (1 to maximum scan number) has been chosen to simplify the treatment. However, choosing any window size and adapting the method accordingly is quite straightforward.

First, the m/z readings from each scan are gathered in a two dimensional table. The rows represent the m/z channels, while the columns represent the scan numbers. Since the number of m/z readings varies from one scan to another, the number of rows filled with a m/z readings vary from one column to another. Second, from the table, the scan column that has the maximum number of m/z readings is selected as a pivot (reference) column. Third, the rest of the columns are consecutively and iteratively chosen for m/z re-ordering. This is done by row-wise reordering of each column such that absolute difference (i.e. Manhattan distance or L1 norm) between the m/z readings of a reordering column and the pivot column is minimal. This is schematically shown in Fig. 1 that shows a table with four scan columns: S1, S2, S3, and S3, filled with m/z readings: m1, m2, m3, and m4, at locations depicted with markers. All m/z readings that belong to the same class are indicated with similar markers. The scan column S3 is selected as the pivot column because it has the maximum number of m/z readings. The rest of the columns (S1, S2 and S4) is iteratively chosen for m/z re-ordering (which is shown by the arrow which points to the new location in the table). The data-reordering step is complete when all m/z readings collected in the table are assigned to similar class. Fourth, a group representative value is selected as the median m/z over a subgroup of m/z in time that fall between the 10 and 90 percentage of the ion chromatogram intensity values. Fifth, steps 1–4 are repeated for each unit m/z in the data set. The group representative values from all unit m/z together