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

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

\textsuperscript{a} TNO Research Group Microbiology & Systems Biology, Utrechseweg 48, 3700 AJ Zeist, The Netherlands
\textsuperscript{b} TNO Research Group Quality & Safety, Utrechseweg 48, 3700 AJ Zeist, The Netherlands
\textsuperscript{c} TNO Triskelion B.V., P.O. Box 844, 3700 AV Zeist, The Netherlands
\textsuperscript{d} Department of Metabolic and Endocrine Diseases, University Medical Centre Utrecht, Utrecht, The Netherlands
\textsuperscript{e} Leiden/Amsterdam Center for Drug Research, Analytical BioSciences, Gorlaeus Laboratories, University of Leiden, Einsteinweg 55, 2333 CC Leiden, The Netherlands
\textsuperscript{f} 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.

**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 chromatographic–mass spectrometry (LC–MS) measurements where the gradient of the solvent and the washing solvent 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 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.

© 2012 Elsevier B.V. All rights reserved.

1. **Introduction**

Hyphenated mass chromatogram data obtained by gas chromatography–mass spectrometry (GC–MS) or liquid chromatography–mass spectrometry (LC–MS), herein termed...
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 S4, 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
with ion chromatogram intensities over time are the binned mass chromatogram dataset.

The binning results from the individual dataset are combined to yield the result for multiple datasets. Firstly, the binning results (the representative \(m/z\) for all channels) from the individual dataset are pooled. Secondly, the pooled \(m/z\) values are clustered and the centroid of each cluster is calculated. The representative \(m/z\) values are then the centroids of the clusters. The representative \(m/z\) values together with the respective ion chromatogram intensities from each dataset form the binned mass chromatogram datasets.

2.1. Example for mass binning

As an example, mass binning of single LC–HR-MS data using only one unit \(m/z\) (\(M^{316}\), i.e. unit \(m/z = 316\)) will be discussed. Table 1 shows a subset of the reordered matrix \(M_{316}\) for unit \(m/z = 316\) (\(M^{316}_{1}\) represented as \(M_{316}\_i\)), where the \(i\) is the index of the read accurate mass (i.e. \(0 < M^{316}_{1} < 316 < 1\)). The rows contain the read \(m/z\) values for contiguous scans from 1 to 3200, while each column contains all \(m/z\) read at a specific scan. The reordered matrix \(M_{316}\) was created with scan number 417 as the pivot column that contained the maximum number of \(m/z\) values. Fig. 2 shows the different \(m/z\) channels (4 channels) and the corresponding \(m/z\) values for all scans (1–3200 scans) after reordering the matrix, \(M_{316}\). The location of the pivot at scan number 417, where the maximum number of \(m/z\) values was read, is also indicated. A blank location within a \(m/z\) channel indicates that a \(m/z\) value belonging to that specific class of \(m/z\) channel was not read. Fig. 3 shows a subset of scan contiguous \(m/z\) values from a row, \(M^{316}\) (corresponding to the third row of Table 1 and the asterisks in Fig. 2) of the reordered matrix, where the corresponding ion chromatogram has the maximum peak (Fig. 4). The median value, 316.2842, is calculated from a subset of \(m/z\) values that were selected based on the lower and upper limit, which is 10% and 90% of the maximum ion chromatogram intensity respectively is then the representative \(m/z\) value for \(M^{316}_{3}\).

![Fig. 1. Reordering procedure: a table with four scan columns: S1, S2, S3, and S3 are filled with \(m/z\) readings: m1, m2, m3, and m4, at locations depicted with markers. The scan column S3 is selected as the pivot column because it has the maximum number of \(m/z\) measurements. 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).](Image)

![Fig. 2. The four \(m/z\) channels and the corresponding \(m/z\) values for all scans from 1 to 3200 scans after reordering the matrix, \(M_{316}\). (For interpretation of the references to color in Fig. 4 legend, the reader is referred to the web version of this article.)](Image)

![Fig. 3. A subset of scan contiguous \(m/z\) values from a row, \(M^{316}\) (corresponding to the third row of Table 1 and the asterisks in Fig. 2) of the reordered matrix, where the corresponding ion chromatogram has the maximum peak (Fig. 4). The median value, 316.2842, is calculated from a subset of \(m/z\) values that were selected based on the lower and upper limit, which is 10% and 90% of the maximum ion chromatogram intensity respectively is then the representative \(m/z\) value for \(M^{316}_{3}\).](Image)

![Fig. 4. Normalized ion chromatogram that corresponds to the subset of scan contiguous \(m/z\) values from the reordered matrix \(M^{316}\) (third row of Table 1 and the red asterisks of Fig. 2).](Image)

<table>
<thead>
<tr>
<th>(M)</th>
<th>S.No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M^{316}_{1})</td>
<td>1</td>
</tr>
<tr>
<td>(M^{316}_{2})</td>
<td>0</td>
</tr>
<tr>
<td>(M^{316}_{3})</td>
<td>0</td>
</tr>
<tr>
<td>(M^{316}_{4})</td>
<td>316.2844</td>
</tr>
<tr>
<td>(M^{316}_{5})</td>
<td>316.9158</td>
</tr>
</tbody>
</table>
3. Method for baseline correction

In this paper we propose to use the statistical entropy as a metric to distinguish true metabolite signals from (systematic) noise. Firstly, the entropy is defined, and secondly, how this can be used for baseline correction is described. This is followed by an example to illustrate the approach.

3.1. Entropy based mass chromatogram signal quantification

The gradual increase of organic solvent in the mobile phase and the simultaneous decrease of water during the chromatographic run result in a so-called background ion chromatogram signal as shown in Fig. 5. This is in contrast to an analyte ion chromatogram signal, which usually is uni-modal and less randomized as shown in Fig. 6. By suitably quantifying the ion chromatograms one can differentiate the informative signals from non-informative and noisy signals, and then removing the latter, the baseline corrected mass chromatogram data can be obtained.

A suitable function to measure the randomness of a ion chromatogram signal is entropy. The entropy, $S$, as discussed in Jansen [17], is a measure of disorder of a system, and is defined as in Eq. (1)

$$ S = -k \sum p_i \log(p_i) $$

where $p_i$ is the probability of the system in the $i$th state, and $k$ is the Boltzmann constant.

An ion chromatogram data normalized to its highest signal intensity can be equivalently treated as discrete probability states and therefore an entropy can be defined for an ion chromatogram ($S_m$) as in Eq. (2).

$$ S_m \approx \sum \frac{l_m}{l_m} \log \left( \frac{l_m}{l_m} \right) $$

where $l_m$ is the normalized intensity of the mass, $m$, at the $s$th scan time.

Because the entropy of a randomly behaving background ion chromatogram is higher than a less randomly behaving analyte ion chromatogram signal, an entropy threshold, $\zeta$, can be defined, below which all informative analyte $m/z$ channels can be retained for further processing. This is performed by first calculating the

Fig. 5. The background ion chromatogram signal resulting from the gradual increase of organic solvent in the mobile phase and the simultaneous decrease of water during a liquid chromatographic run. The signal characteristic is in contrast to the analyte ion chromatogram signal.

Fig. 6. An analyte ion chromatogram signal, which usually is uni-modal and less randomized than the background signal in Fig. 5.

Fig. 7. This is an example of applying the proposed baseline correction procedure with different values of entropy threshold $\zeta$ to mass chromatogram data, resulting in a baseline corrected mass chromatogram. The total ion current (TIC) of the mass chromatogram data for different entropy values is shown.

Fig. 8. The total ion current (TIC) of the subtracted baseline from the example mass chromatogram data shown in Fig. 7 with different entropy values.
entropy, $S_m$, for all m/z channels of the liquid chromatogram—mass spectrometry data and retaining those channels for which the values are lower than the entropy threshold. This procedure leads to baseline corrected chromatogram data. The value of the entropy threshold $\zeta$ can be determined empirically, and a good estimate from our experience with several LC–MS datasets indicated a value of $m(S) + s(S)$, where $m(S)$, and $s(S)$ are the mean and standard deviation of all entropies (one for each m/z channel), S, of the mass chromatogram.

3.2. Example for baseline correction

An example of applying this procedure with different values of entropy threshold $\zeta$ to mass chromatogram data, resulting in a baseline corrected mass chromatogram is shown in Fig. 7. The corresponding baseline signal is shown in Fig. 8. The increase of the total ion current observed at the end of the chromatogram, caused by a washing step with high percentage of organic solvent, is significantly lowered at a threshold value of $\zeta = 0.6$. Considerably lowering the threshold any further, like $\zeta = 0.4$ removes some major analyte peaks.

Fig. 9. The number of m/z channels per block available for multiple sample block-wise deconvolution after multiple binning, and baseline correcting the samples S1 and S2, and suppressing all m/z channels having signal intensity lower than 0.1% of the maximum intensity.

Fig. 10. Deconvolution results in terms of the percentage number of analytes from the target list.

Fig. 11. Detail investigation of the deconvolution results showed that 14% of the metabolites from the target list belonged to a m/z channel with an entropy higher than the entropy threshold, $\zeta = 0.6248$. This means that respective analytes did not leave a typical analyte signal, instead left a high intensity non-uniform signal all along the mass chromatogram. An example of such an analyte signal from the target list as appeared in sample S1 is shown.

4. Results and discussion

Raw LC–HR-MS data obtained from 2 plasma samples (S1 and S2) using a LC–MS method for free fatty acids and phospholipids by Ramaker et al. [18], are subjected to the two major data pre-processing steps—the dataset binning and the baseline correction, described in the previous sections. The datasets are individually binned and baseline corrected before combining them to form one large multiple binned, baseline corrected mass chromatogram dataset. The combined mass chromatogram is segmented, along the retention time axis, into a number of equally sized mass chromatogram blocks. Each of these mass chromatogram blocks are then deconvoluted as described by Jellema et al. [4]. In the following sections the mass binning, baseline correction and the overall deconvolution results will be discussed.

Fig. 12. The same analyte from the target list and discussed in Fig. 11 as appeared in sample S2 is shown.
4.1. Multiple mass binning and baseline correction results

After binning of the individual datasets the number of representative m/z channels found for samples S1 and S2 are 1458 and 1487 respectively. The independent baseline correction with an entropy threshold, $\zeta = 0.6248$, that followed the individual dataset binning reduced the number of m/z channels further to 624 and 623 respectively. The lowest and the highest read m/z values for sample S1 were 202.0078 and 999.9617 respectively. In the case of unit m/z binning, the number of m/z channels would be 707 which is the unit m/z difference between the highest and the lowest read m/z values. The number of m/z channels after applying the mass binning and baseline correction methods is not only lower than the unit m/z binning result but also accurate up to three decimals for most analytes (m/z accuracy results are addressed at a later section). The situation is similar for sample S2 whose lowest and highest measured m/z values were 202.0072 and 999.7836 respectively.

4.2. Deconvolution results

Fig. 10 shows the deconvolution results in terms of the percentage number of analytes from the target list discussed in Ramaker et al. [18]. Deconvolution found 62% percentage of the analytes from the target list. On further investigation, it was discovered that 14% of the metabolites from the target list belonged to a m/z channel with entropy higher than the entropy threshold, $\zeta = 0.6248$. This
means that respective analytes did not leave a typical analyte signal, instead left a high intensity non-uniform signal all along the mass chromatogram. For example, Figs. 11 and 12 shows an example for the same analytes (i.e. same m/z values) observed in samples S1 and S2 with high entropy. The rest of those analytes that were not found had a low concentration (more than 2 orders of magnitude lower than those found) in the samples that were analyzed. At least, 10% of the analytes produced faint, but clear analyte signals (low intensity). Their intensities were lower than the signal to noise ratio of 3 and hence could not be distinguished from the background. Another 3% and 7% of the metabolites from the target list other than having a low concentration, produced ion chromatogram signals typical to liquid chromatography column bleed [19] (see Fig. 13) and noise (see Fig. 14) respectively.

For the analytes found by deconvolution a pie chart with percentage-wise classification of the analytes based on the accuracy of the calculated m/z values compared to the theoretical values was made (see Fig. 15). The results show that the calculated m/z value is accurate up to 4, 3, 2 and 1 decimal value for 15%, 46%, 21% and 17% of the analytes respectively, and unit m/z accurate for 2% of analytes. On manual inspection of the bins, the m/z values for the other analytes, not found with deconvolution, produced similar results.

To provide a visual overview of the deconvolution results, Fig. 16 is the total ion current (TIC) of mass chromatogram, sample S1, after mass binning and baseline correction, while Fig. 17 is the TIC of the reconstructed mass chromatogram after deconvolution.

5. Conclusion

In this paper new methods for mass binning and baseline correction are demonstrated for multiple deconvolution of LC–HR–MS data. The mass binning method employed a minimal distance approach to capture the distribution of different m/z values over scans and then used the ion chromatogram intensity as a decision metric to determine a set of representative m/z values that eventually led to binned mass chromatogram data. For multiple datasets, the binned mass chromatogram data from individual datasets could be combined to yield similar results. The major advantage of this mass binning method is that it uses the distribution of m/z values over scans, and hence is a procedure independent to varying measurement accuracy of the mass spectrometry instrument. In addition, the method employs the selectivity criterion of choosing m/z values from more stable regions of the mass chromatogram, and therefore the results are more robust. The method of baseline correction, in contrast to other curve fitting methods, sought a fundamental treatment to the problem and in the process introduced a statistical metric, the entropy, capable of distinguishing m/z channels that inject systematic noise into the mass chromatogram data. Subsequent removal of the m/z channels with systematic noise signals resulted in baseline corrected mass chromatogram data. Applying the binning and baseline correction methods to prepare 2 datasets for block wise multiple dataset deconvolution, resulted in less than 10 m/z channels (average) per block (1 block = 75 scan numbers), each with a mass accuracy up to 4 decimal places. The percentage of analytes from the target list found by deconvolution was 62% of which 60% with at least 3 decimals mass accurate to the theoretical m/z. The reasons for not able to find the rest of the analytes from the target list have been attributed to lipid–polar method of chromatography that still faces challenges in chromatographic separability. Some of the future developments, in the direction of building a fully integrated web-based LC–HR–MS pre-processing tool, are to include methods that use predefined knowledge like isotopic distribution pattern of the metabolites, the m/z variation over scan, etc., to increase the quality of quantification results.

Entropy curve

In addition to identifying the systematic noise and correcting the baseline, we further observed that the entropy of the mass chromatogram in also a valuable metric to classify the different types of m/z channels below the entropy threshold. These observations, together with some hypothesis, are discussed in the supplemental part of this paper.

Acknowledgment

This research was conducted with the support of grants from the Netherlands Metabolomic Centre (NMC).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aca.2012.06.014.

References


