BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2488–2489
doi:10.1093/bioinformatics/btq441

Data and text mining

Advance Access publication July 29, 2010

Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data

H. Paul Benton*, Elizabeth J. Want and Timothy M. D. Ebbels*
Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Sir Alexander Fleming Building, Imperial College London, London, SW7 2AZ, UK
Associate Editor: Martin Bishop

ABSTRACT

Motivation: High mass accuracy is an important goal in liquid chromatography–mass spectrometry experiments. Some manufacturers employ a mass calibration system that regularly switches between the analyte and a standard reference compound, and leads to gaps in the analyte data. We present a method for correction of such gaps in global molecular profiling applications such as metabolomics. We demonstrate that it improves peak detection and quantification, successfully recovering the expected number of peaks and intensity distribution in an example metabolomics dataset.

Availability and implementation: Available in XCMS versions 1.23.3 and higher. Distributed via Bioconductor under GNU General Public License. (http://www.bioconductor.org/packages/2.7/bioc/html/xcms.html)

Contact: hpaul.benton08@imperial.ac.uk; hpbenton@gmail.com; t.ebbels@imperial.ac.uk

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

Received on June 2, 2010; revised on July 20, 2010; accepted on July 27, 2010

1 INTRODUCTION

Liquid chromatography–mass spectrometry (LC–MS) has become a widely used methodology in metabolomics. Profiles generated by this method are highly complex and many software packages have been designed to analyse this data [e.g. mzMine, XCMS (Katajamaa *et al*., 2006; Smith *et al*., 2006)]. These typically employ a peak-picking algorithm for selection of peaks and removal of noise. Peaks are then matched between samples, correcting small drifts in retention time (RT) and the final result is a table of the integrated intensities of each peak in each sample.

An important goal in LC–MS experiments is the accurate measurement of the mass to charge ratio (m/z). Usually, a mass calibration method is used throughout the experiment to keep the m/z deviation within the desired range. MS manufacturers employ different methods to calibrate their instruments. Some manufacturers use a switching method, which regularly changes the MS input between the analyte and a standard reference (Balogh, 2009). Switching the flow in this way can avoid the problem of ionization suppression encountered with continuous flow methods. However, changing the flow to the lock mass spray temporarily stops the analyte from being detected, resulting in gaps in the analyte data. These lock mass gaps can cause a single peak to be split into several smaller peaks (Fig. 1), confusing the process of peak matching. Even if the peak is detected correctly, the true intensity will be underestimated due to the gap. The situation is further exacerbated by RT drifts across samples, which will cause the position of the gap within the peak to shift. The peak detection, multiplicity and loss in intensity will thus vary depending on where the gap is in the peak. This variability can lead to both false positives and false negatives in peak detection, and inaccurate intensity estimation. While most vendor software is presumed to take account of such elements of MS design, this is not true of widely used and community supported open source software for LC–MS data processing. We have developed a method that recovers the lost intensity by filling in the gap. The method greatly reduces errors in peak detection, matching and intensity estimation and is freely available as a built-in function 'stitch', for the XCMS package.

*To whom correspondence should be addressed.

2 METHODS

2.1 Analytical data and processing

A 10 μl of *Caenorhabditis elegans* extract (see Supplementary Material) was separated on a Waters ACQUITY UPLC system using an HSS T3 column (2.1 × 100 mm, 1.7 μm Waters Corporation, Milford, MA, USA). A 30-min water to acetonitrile linear gradient was used with a flow rate of 500 μl/min. All solvents were HPLC grade (ACN: Fluka, H_20: Romil Ltd). Detection was performed on a Waters LCT Premier ToF mass spectrometer. The sample was run twice, once with a lock mass interval (LMI) of 50 scans and once with an LMI of 1000 scans. The scan range was set to 50–1000 m/z, scan time to 0.2 s and inter-scan delay to 0.01 s. The majority of peak widths were in the range 2–20 s.

Data were converted to mzXML using MassWolf (version 4.3.1). Each dataset was processed using XCMS (version 1.23.1). A third dataset was generated from the LMI 50, by correcting the lock mass gaps using the new 'stitch' method. These were then processed using the centWave (Tautenhahn *et al*., 2008) peak picker on all three datasets. Parameters were 'peakwidth' of 2–20 s and 'snthresh' of 5. Data were then grouped using mzwid = 0.05; other parameters were set at default values.

2.2 Gap filling 'stitch' algorithm

The basic idea of the algorithm is to use the nearest available analyte scan to fill the gap. More precisely, if the location of the gap is from scan n to scan $n+k$, then the algorithm uses scan $n-1$ to fill the first half of the gap and scan $n+k+1$ to fill the second half of the gap. If the gap length is odd, the middle scan is filled using scan $n-1$. The algorithm requires three input parameters, which are all readily available in the vendor software. These are: the scan number of the first gap, the LMI (the time between gaps) and the length of each gap (default two scans). Despite its simplicity, this algorithm is very effective at correcting lock mass gaps and requires minimal user input.

3 RESULTS

To test the method, data were acquired for a single sample at two different LMIs. A first acquisition at an LMI of 1000 scans (giving

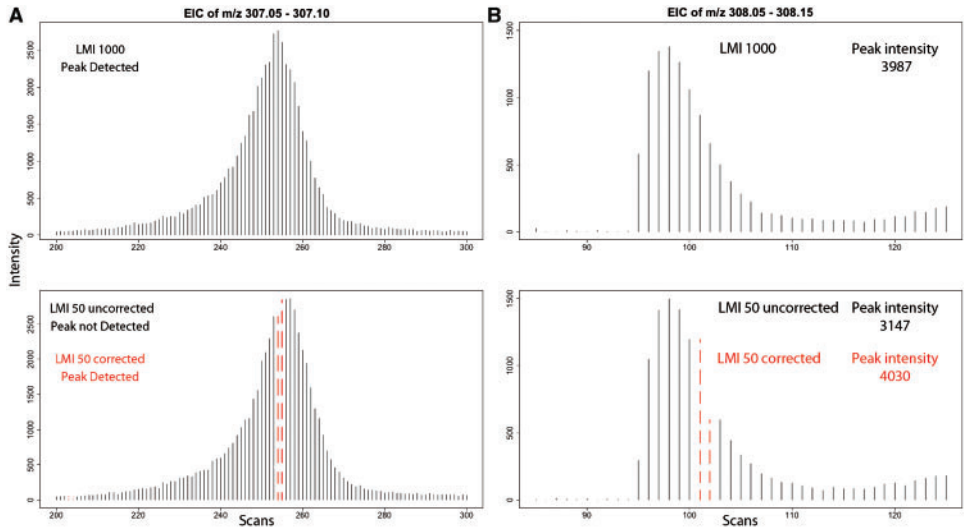

Fig. 1. Extracted ion chromatograms (EICs) for two example peaks damaged by lock mass gaps for the three datasets: LMI 1000 (top), LMI 50 (bottom, solid black) and LMI 50 corrected (bottom, dashed red line). The gap in (**A**) resulted in the peak not being detected by the peak detection software. In (**B**) the peak was detected but with reduced intensity. In both cases, the gap-filling algorithm rectified the problem.

a total of three lock mass scans over the entire acquisition) gave a dataset where few peaks would be affected by lock mass gaps. This approximates data unaffected by the lock mass gap problem. A second acquisition with an LMI of 50 scans (71 lock mass scans in total) was chosen to correspond to a typical metabolomics protocol. We then compared the XCMS output for the LMI 50 data with and without correction with the LMI 1000 data without correction.

To illustrate the success of the method, we extracted two example peaks from the data (Fig. 1). The first peak ($m/z = 307.08$, RT = 126 s) demonstrates how the correction algorithm improves peak detection. The bottom panel of the figure, plotted in black, clearly shows two scans missing due to lock mass acquisition; these are corrected (shown dashed in red) resulting in a peak with very similar shape and intensity to that in the LMI 1000 data (top panel). This peak is detected by the software in the LMI 1000 and LMI 50 corrected data, but not in the uncorrected LMI 50 data. In the second example, the peak ($m/z = 308.08$, RT = 42 s) is detected in all three datasets. However, in the uncorrected LMI 50 the peak is much shorter, 3 s compared to 5 s in the LMI 1000 and the corrected LMI 50. The gap effectively leads to a shorter peak and this reduces the intensity estimate by 21% compared to the 'undamaged' LMI 1000 data (LMI 50 intensity = 3146, LMI 1000 intensity = 3987). Applying gap correction results in an intensity estimate just 1% higher than the LMI 1000 data (LMI 50 corrected intensity = 4030).

Many peak detection algorithms use the shape of the peak to help identify it. With the correction method, this shape is recovered, and more peaks are found. Using the same parameter settings, 2784 peaks were detected in the LMI 1000 data, while only 2625 peaks were detected in the LMI 50 data, a decrease of 6%. In the LMI 50 corrected data, 2869 peaks were found, a gain of 3% over the LMI 1000 data. After grouping, 2761 peak groups were matched between the three datasets. Extracted ion chromatograms were visually inspected, confirming successful gap correction (Supplementary Material 1). The correction method brought the global intensity distribution closer to the undamaged LMI 1000 data. The median intensity dropped from 772 in the LMI 1000 data to 700 in the LMI 50 data (a 9% reduction), but was recovered with the correction method to 775 in the corrected LMI 50 data (0.4% difference from the LMI 1000 data). As a further demonstration of the success of the correction method, peaks eluting while lock mass correction occurred were manually identified. For the 300 peaks with lowest m/z, 78 were affected by gaps. The fractional difference in intensity between the LMI 50 and LMI 1000 data was computed for both uncorrected (f_u) and corrected (f_c) data for both gapped and non-gapped peaks. Gap correction resulted in a distribution for gapped peaks much closer to that for non-gapped peaks than for uncorrected data (median $f_u = -6.73$ and 3.26 and $f_c = 4.80$ and 3.13 for gapped and non-gapped peaks respectively, see Supplementary Material).

The stitch function allows peak detection algorithms to accurately find and integrate each peak. We note that the method is generic and can be applied to data derived from any LC–MS experiment, e.g. in metabolomics or proteomics. We have implemented the algorithm within XCMS, one of the most widely used freely available LC–MS data processing packages. In summary, our gap correction method allows LC–MS data acquired with lock mass gaps to be analysed with a higher degree of accuracy and confidence than currently possible.

ACKNOWLEDGEMENTS

We thank Florian Geier for assistance acquiring the LC-MS data.

Funding: Capacity building studentship from the UK Medical Research Council (to H.P.B.); Waters Corporation (to E.J.W.).

Conflict of Interest: none declared.

REFERENCES

Balogh,M. (2009) *The Mass Spectrometry Primer*. Waters Corporation, Milford Massachusetts.

Katajamaa,M. *et al.* (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, **22**, 634–636.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Tautenhahn,R. *et al.* (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, **9**, 504.