Peak picking as a pre-processing technique for imaging time of flight secondary ion mass spectrometry

DOI:
10.1002/sia.5062

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Moore, J. D., Henderson, A., Fletcher, J. S., Lockyer, N. P., & Vickerman, J. C. (2013). Peak picking as a pre-processing technique for imaging time of flight secondary ion mass spectrometry. *Surface and Interface Analysis*, *45*(1), 461-465. https://doi.org/10.1002/sia.5062

Published in:
Surface and Interface Analysis

Peak Picking as a Pre-processing Technique for Imaging Time of Flight Secondary Ion Mass Spectrometry.

Jimmy D. Moore[1], Alex Henderson[1,*], John S. Fletcher[1,†],
Nicholas P. Lockyer[2] and John C. Vickerman[1].

[1] School of Chemical Engineering and Analytical Science, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess St., Manchester, M1 7DN, UK.
[2] School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess St., Manchester, M1 7DN, UK.
[†] Present address: Department of Chemistry, University of Gothenburg, 41296 Gothenburg, Sweden.

*Corresponding author e-mail: alex.henderson@manchester.ac.uk

Abstract

High surface sensitivity and lateral resolution imaging make ToF-SIMS a unique and powerful tool for biological analysis. However, with the leaps forward made in the capabilities of the ToF-SIMS instrumentation, the data being recorded from these instruments has dramatically increased. Unfortunately, with these large, often complex, datasets a bottleneck appears in their processing and interpretation.

Here an application of peak picking is described and applied to ToF-SIMS images allowing for large compression of data, noise removal and improved contrast, whilst retaining a high percentage of the original signal. Peak picking is performed to locate peaks within ToF-SIMS data. By using this information, signal arising from the same distribution can be summed and overlapping signals separated. As a result, the data size and complexity can be dramatically reduced. This method also acts as an effective noise filter, discarding unwanted noise from the data set. Peak picking and separation is evaluated against the conventional methods of mass binning and manually selecting regions of a peak to image on a model data set.

Introduction

The size of ToF-SIMS datasets has become an issue in recent times. Though computers are increasing in capability, the data sets being acquired from the SIMS instrumentation have also been increasing dramatically. Instrumentation with higher duty cycles produce images with more pixels more quickly [1]. 3D molecular imaging is now also possible [2] resulting in stacks of images that can easily reach tens of gigabytes. Improved mass resolution and sensitivity provide more discernable peaks [3]. This paper describes a method of compression proposed as an alternative to the classical practices of binning data, where mass resolution is lost, and manual selection of peaks to image.

Here we use the term 'peak picking' to describe the exercise of determining a 'discrete spectrum from continuous data'[‡]. This is in contrast to arbitrary peak selection based on *a priori* knowledge of the sample [4] and/or by applying a simple intensity threshold to the data.

[‡] A term suggested at the 59th IUVSTA Workshop on *Surface Chemical Analysis: Improving data interpretation by multivariate and informatics techniques*, Trinidad and Tobago, 2010

Background

Peak picking is a method of locating peaks within a spectrum. Classically peak picking has been used to find peaks to generate peak lists [5]. From this the area of each peak discovered can be used instead of its original or fitted distribution, thereby reducing the information to a single mass channel. There have been many approaches adopted to do this by the mass spectrometry community [6][7]. In this paper peak picking is applied to locate peaks, fit a peak shape and use this information to perform separation of overlapping signal and selective alignment of pixel data to achieve compression of an image or an image stack.

Individual pixels, in general, do not contain enough information to accurately identify a peak’s position and distribution, but summing the signal from all the pixels in the image usually produces a spectrum with good signal to noise: the total ion spectrum. A peak list can be formulated and each of the constituent pixel spectra can be queried as to their values under the peak's distribution.

However, if there are shoulders or overlapping peaks within the total ion spectrum this complicates analysis. These spectral features must be 'deconvoluted' or separated to estimate realistic values. By evaluating in this way, a highly sparse matrix for all spectra contained in an image can be compiled. In-house testing indicates that this method can give a compression of over 95%, *e.g.* assuming four discernable peaks per amu and a mass range of 1-1000 amu, with 100,000 mass channels, this gives a compression of 96%. This allows for much faster statistical analysis to be performed and opens the door for more computationally intensive methods of data processing to be adopted *e.g.* multivariate analysis (MVA).

Imaging a single species

When imaging a peak the analyst can manually sum across a portion of a peak to generate an image. This can be representative of the signal's distribution throughout the image. However, this is very dependent on the mass limits manually chosen by the user; an inherently subjective approach. This is especially true if peaks are overlapping. The closer the chosen bound is set towards the neighbouring, overlapping peak, the more likely it is to erroneously contain signal from that neighbouring peak. However, by moving the bound away from the neighbouring peak some of the true distribution will be lost from the image resulting in a reduction in contrast. The proposed technique outlined here aims to limit the false positives and false negatives arising in an image while maintaining maximum signal levels and thus contrast. This also gives an automated, reproducible method for all peaks/images.

Binning

Since manual imaging of peaks is generally not performed for all peaks in the spectrum, a common approach adopted to perform multivariate analysis is binning. When data is binned, either to nominal mass or some other arbitrary unit, the distribution of original signal is lost and so the true peak position, the centre of that distribution is also lost. The binning method simply sums mass channels together in a regular fashion. The advantage of this is that signal is summed together, therefore the amount of data is reduced and signal distributions are also condensed. However, the latter is also one of the disadvantages of binning data. Signal distributions are not considered and as a result can be spread across two or more bins. This implies that signal arising from the same distribution can be separated and considered differently as far as multivariate methods are concerned. Conversely distributions that are independent of each other can be put into the same bin. This gives an unnecessary mixing of the data.

With recent advances in instrumentation mass resolution has increased, however mass resolution is lost when data is binned resulting in an effect contra to the developments in instrumentation. As has been previously noted [8], different pre-treatments of images can have a large effect on MVA techniques and some binning routines were noted as having a detrimental effect on certain types of MVA.

Method

Peak picking

Peak separation/deconvolution is necessary since overlapping peaks can distort the desired information in neighbouring peaks. Peak shoulders are one example of this. Here we define a peak shoulder as a peak overlapping to the extent that a clear valley between the two peaks is no longer evident, but there is an obvious deviation from the expected peak shape due to an additional signal distribution. An estimate of the actual information contained in these shoulders and overlapping peaks is essential to obtain an accurate representation of the chemistry contained within a mass spectrum.

The total ion spectrum was chosen to act as a starting point since this contains the most statistically useful data. By peak picking this total ion spectrum, irrelevant or unwanted noise peaks can be easily excluded. Peak picking begins with the identification of spectral features by finding maxima within the total ion spectrum using a continuous wavelet transform. Using the approach of Du *et al* [7], peak shoulders are then located by taking the first derivative of these spectral features and finding additional maxima.

In-house scripts written in MATLAB (version R2009a, MathWorks Inc., MA, USA), some utilising routines from the PLS_Toolbox (version 4, Eigenvector Research Inc., WA, USA), were used to fit Gaussian peak shapes to spectral features discovered in the total ion spectrum. Here the Gaussian peak shape is used in order to exemplify the procedure performed. Other mathematical functions, or peak shapes derived directly from the data, are equally applicable. Once fitted, the resulting components provide the underlying peak positions, widths etc. to be used in the analysis of the image.

Two peak model data

A data set was fabricated in order to produce two distributions of signal overlapping in the total ion spectrum, but spatially separated in the total ion image. Here the Box-Muller [9] transform was utilised to generate a randomised collection of Gaussian peaks of equal intensity, with Gaussian distribution on the mass scale. The procedure was repeated for two Gaussian peaks of different intensity and which exhibited an overlap. These separate distributions were then arranged in an image such that all pixels on the left of the image had signal arising from peak A (left peak) and those from the second, lower intensity peak B (right peak) had signal to the right of the image as shown in figures 1a and 1b.

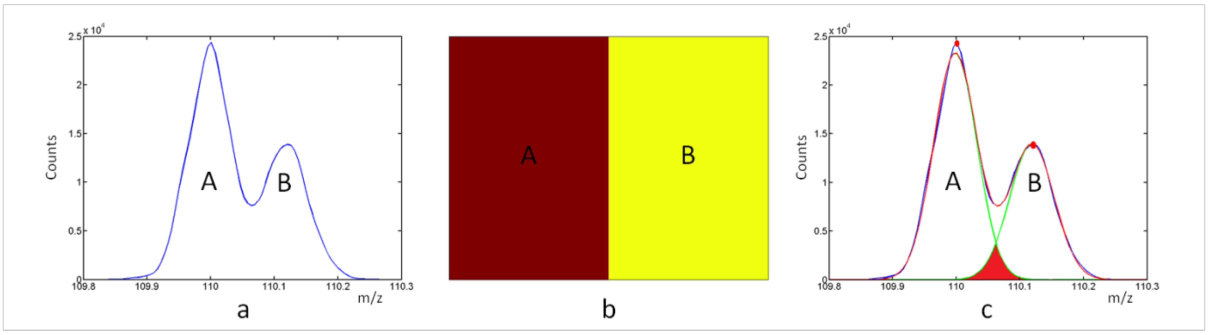

Figure 1: a: model data total ion spectrum, b: model data total ion image of 1a, c: model data total ion spectrum after peak picking and fitting (the shaded section defines the overlap region).

Following the procedure outlined above, the total ion spectrum data was peak picked. This provided information (location and full width at half maximum etc.) about each of the two total ion peak distributions, A and B. The major challenge with this data is the overlap region (shaded in figure 1c). Data arising left of the overlap can be assigned to peak A and conversely for the signal to the right of the overlap region which can be assigned to peak B, but signal from the overlap region cannot be assigned similarly. It is known that in this region some pixels contain signal from peak A and some contain signal from peak B.

Scaling the overlap region

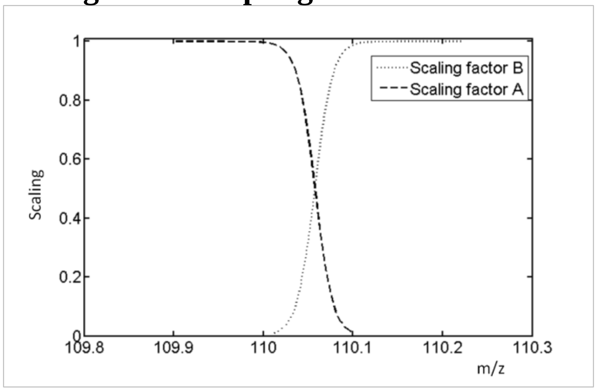

Figure 2: Scaling parameters for each peak.

As signal closer to the centre of peak A is more likely to originate from peak A than peak B and the opposite for peak B, a scaling factor which reflects this has been created to apply to signal in the overlap region in each pixel of the image. Equations 1 and 2 are created from a normalised fit of a Gaussian to the two overlapping peaks in the data (Figure 2). Since all the signal in the overlap region should be used, these scaling factors sum to one. Each pixel in the image is then scaled using this function to define the proportion of the overlap region that should be attributed to peaks A and B. The scaled signal for A and B was then summed with the signal outside the overlap region within each individual pixel. This defines a value of peak A and peak B within each pixel. By comparing this assigned signal value in each peak, the signal in the overlap region for each pixel can be assigned to the distribution with highest value, or greater than some user defined threshold: here a 20% difference was used. However, if the threshold is not exceeded and the signal associated with peak A and B is similar, some other information is needed to make the decision.

$$A_{scaling} = \frac{A_{norm}}{A_{norm} + B_{norm}}$$ Equation 1.

$$B_{scaling} = \frac{B_{norm}}{A_{norm} + B_{norm}}$$ Equation 2.

Where A_{norm} is the normalised (normalised to height 1) fit of a Gaussian to peak A, B_{norm} the normalised fit of a Gaussian to peak B, $A_{scaling}$ is the calculated scaling factor for peak A and $B_{scaling}$ is the calculated scaling factor for peak B.

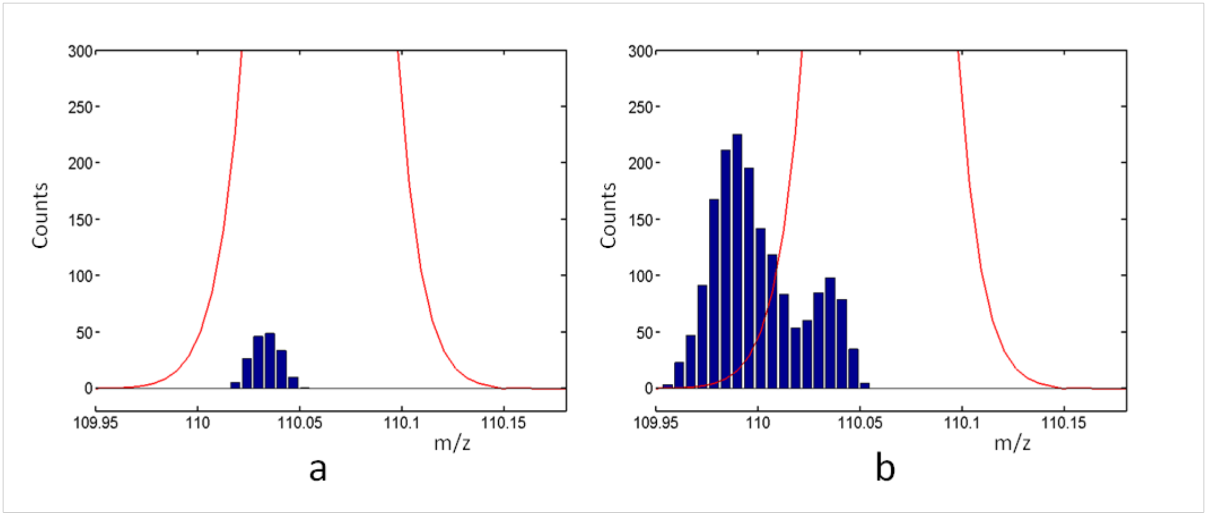

Figure 3. a: close up of overlap region (shaded area in figure 1c) with overlay of a spectrum from a single pixel, b: close up of overlap region with overlay of signal from 3a and its eight neighbouring pixels.

Consider a single pixel, pixel x, from the model data in figure 1 which only has signal within the overlap region. It is impossible to decide from which peak distribution this signal came without prior knowledge (Figure 3a). The only other information available is the spatial distribution of signal within the image. By incorporating surrounding pixels, a trend of signal in that local area can be observed. Figure 3b is a plot of the signal arising from pixel x in 3a with the eight neighbouring pixels' signal also included. After scaling and proportioning the signal in 3b a clear difference is observed; that there is more signal within peak A. Using this the signal from pixel x in 3a is assigned to peak A. If there remains not enough difference between peak A and peak B in the immediate pixels around pixel x to make a decision, then the region of interest can be expanded again to incorporate more pixels, thus widening the area of the image to aid in the classification. This method is then repeated for each pixel in the image and all signal from the overlap region assigned to their respective peaks.

However, this method assumes that the nearest neighbours of pixel x will have signal originating from the same distribution, that of peak A. If there is no localisation of signal in the immediate pixels around pixel x expanding the region of interest incorporates signal that is spatially further away from pixel x. This is more likely to bring in signal from the incorrect distribution, *i.e.* peak B. If this is the case and the signal is incorrectly assigned, that would imply that the true distribution from which the signal in pixel x originated is isolated spatially from other signal arising from that distribution. In real world data there is little information that can be gleaned from such a pixel.

Results

The approach described above allows for imaging of each peak individually as opposed to manually selecting a portion of the peak. Recalling Figure 1 where all signal from peak A arises from the left half of the image. Figure 4 shows the model data using an arbitrarily selected mass range relating to peak A (4a), and following peak picking and separation (4b). The total ion images of these mass ranges are shown in 4c and 4d respectively with a difference image between the two methods shown in 4e.

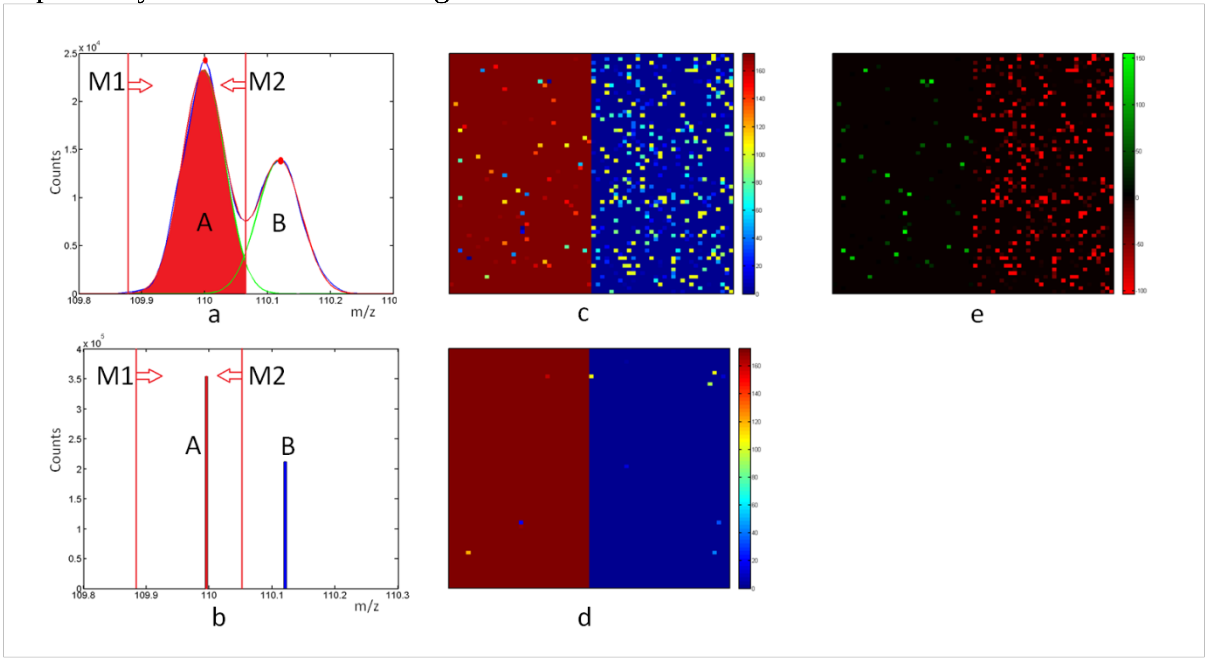

Figure 4: a: region of total ion spectrum selected to image (original data), b: region to image after peak picking and separation (same region as image 4a), c: image of region selected in 4a, d: region imaged after peak picking and separation, e: the difference between image 4c and image 4d.

On imaging the original data with the valley between the two peaks as a terminating mass limit, M2 in Figure 4a, there is clearly signal missing from the left hand side and additional unwanted signal on the right (Figure 4c). To sharpen the image the mass limits can be changed. Shifting limit M2 to a lower mass would reduce the inclusion of signal from peak B in the image but real signal from peak A would also be lost. Imaging such a peak manually is subjective and time consuming, and arbitrarily mass binning the data will suffer from similar effects depending on the width and position of the mass bin.

In comparison the peak picked and separated/deconvoluted data produces an image with improved contrast (Figure 4d). There is a reduction in signal in the right half and an increase in the left half of the image.

The difference image between the original data image and the peak picked and separated data is shown in Figure 4e. Here black pixels are those that are unchanged. Increased signal is represented by green and reduced signal is characterized by red. Examining this image it is evident that the increased signal is found in the left half of the image, the true location of peak A and there is an obvious reduction in signal in the right half where peak B originated.

Conclusions

In this paper a new method of visualising and compressing data has been outlined. This is proposed as an improved method of data compression for analysis and imaging as opposed to binning or manual selection of peak regions to image. In principle this could be extended to 3D image stacks and work is progressing in this area.

Peak picking and separation was performed on model data. Using information obtained from peak picking, separation of overlapping image signal was achieved. Large data compression was attained (greater than 95%) and visible image contrast increased. This method gives a generalised framework which is both automated and reproducible.

Though the model data used here had a Gaussian distribution and different instruments will have diverse generalised distributions in their respective total ion spectra, this model could be adopted for other peak distributions to similar effect.

Acknowledgements

The authors gratefully acknowledge the financial support of the Engineering and Physical Sciences Research Council, EPSRC, UK under grant number EP/G045623/1.

References

[1] J. S. Fletcher, S. Rabbani, A. Henderson, P. Blenkinsopp, S. P. Thompson, N. P. Lockyer, J. C. Vickerman, *Analytical Chemistry* **2008**, *80*, 9058.

[2] G. Gillen, A. Fahey, M. Wagner, C. Mahoney, *Applied Surface Science* **2006**, *252*, 6537.

[3] A. Carado, M. K. Passarelli, J. Kozole, J. E. Wingate, N. Winograd, A. V. Loboda, *Analytical Chemistry* **2008**, *80*, 7921.

[4] M. S. Wagner, T. A. Horbett, D. G. Castner, *Biomaterials* **2003**, *24*, 1897.

[5] B. Tyler, *Applied Surface Science* **2003**, *203-204*, 825.

[6] N. Nguyen, H. Huang, S. Oraintara, A. Vo, *Bioinformatics* **2010**, *26*, i659.

[7] P. Du, W. A. Kibbe, S. M. Lin, *Bioinformatics* **2006**, *22*, 2059.

[8] A. Henderson, J. S. Fletcher, J. C. Vickerman, *Surface and Interface Analysis* **2009**, *41*, 666.

[9] G.E.P. Box, M.E. Muller, *Annals of Mathematical Statistics,* **1958**, 29, 610.