SpectralAnalysis: software for the masses

Race, Alan M; Palmer, Andrew; Dexter, Alex; Steven, Rory T; Styles, Iain B; Bunch, Josephine

DOI:
10.1021/acs.analchem.6b01643

SpectralAnalysis: software for the masses

Alan M. Race,[*,†,‡] Andrew D. Palmer,[¶,‡] Alex Dexter,[†,‡] Rory T. Steven,[†] Iain B. Styles,[§,‡] and Josephine Bunch[*,†,‖]

†*National Centre of Excellence in Mass Spectrometry Imaging (NiCE-MSI), National Physical Laboratory, Teddington, UK*

‡*PSIBS Doctoral Training Centre, School of Chemistry, University of Birmingham, Birmingham, UK*

¶*European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*

§*School of Computer Science, University of Birmingham, Birmingham, UK*

‖*School of Pharmacy, University of Nottingham, Nottingham, UK*

E-mail: alan.race@npl.co.uk; josephine.bunch@npl.co.uk

Abstract

The amount of data produced by spectral imaging techniques, such as mass spectrometry imaging, is rapidly increasing as technology and instrumentation advances. This, combined with an increasingly multi-modal approach to analytical science, presents a significant challenge in the handling of large data from multiple sources. Here we present software that can be used through the entire analysis workflow, from raw data through preprocessing (including a wide range of methods for smoothing, baseline correction, normalisation, and image generation) to multivariate analysis (for example memory efficient PCA, NMF, MAF and PLSA), for datasets acquired from single experiments to large multi-instrument, multi-modality, multi-center studies. SpectralAnalysis was also developed with extensibility in mind to stimulate development, comparisons and evaluation of data analysis algorithms.

Introduction

Spectral imaging is a broad category of techniques where a spectrum is acquired at spatially resolved locations. Such techniques include mass spectrometry imaging (MSI) and Raman spectroscopy, which have also been used in combination.[1,2] Mass spectrometry imaging covers a whole suite of techniques which rely on different ionisation principles, mass analysers and detectors to measure the m/z values of gas phase ions produced at spatially resolved locations. Different ion sources affect the classes of molecules that can readily be ionised, spatial resolution that can be achieved and degree of fragmentation that occurs and so have the capability to provide complementary data about the chemical composition of a given sample.[3,4] Similarly, in matrix assisted laser desorption/ionisation (MALDI) MSI complementary data can be achieved through the use of different matrices, mass ranges and/or polarities to target different classes of molecule on the same sample.[5]

Even when a single mass spectrometer is used within a study, the number of datasets being analysed together is increasing. Recently, 9 tissue sections per time point resulting in a total of 27 sections were analysed to investigate protein digestion,[6] 63 sections from gastric cancer and 32 from breast cancer were analysed to investigate intratumour heterogeneity[7] and 96 sections taken from 32 mice were analysed to investigate the consequences of cortical spreading depression.[8]

Both the incorporation of multiple modality data, or multiple MSI modality data, and the increasing trend towards larger MSI studies present a significant challenge in the data handling, visualisation and analysis. Each MSI instrument vendor supplies software for processing and visualising data acquired on their instruments. These software cannot analyse data from other instrument manufacturers due to the use of proprietary data formats. As there are no common preprocessing methods between any of these software (a list of which are given in Table S1) there can be no guarantee that the data have been treated equally, eliminating them from consideration as the software of choice for multimodality studies. Similar challenges exist when trying to analyse data from large studies, as these will often

span multiple data acquisitions (resulting in multiple data files) which can only be analysed one at a time in most software packages.

The first vendor neutral, and often still used, tool for visualising MSI data was BioMAP.[9] This enabled a user friendly means of visualising data acquired on many mass spectrometers, even before the community agreed upon the imzML standard for sharing data.[10] A limitation of BioMAP in the processing of large MSI data is the limit to the number of m/z channels that can be loaded for any dataset (32768, due to the number format used) which has become a more significant issue as the instrumentation has improved. Since the advent of imzML, a wide number of third party software packages have been developed and released as open source software (MSiReader,[11] Cardinal[12] and OmniSpect[13]), freely available software (OpenMSI,[14] DataCubeExplorer[15] and msiQuant[16,17]), made available to collaborators only (Mirion[18]) or as a commercial product (SCiLS Lab,[19] MALDIVision[20] and Quantinetix[21]). The preprocessing methods available in each of these software tools are presented in Table S1. Conversion to imzML is possible through most vendor software tools as well as third party software such as imzMLConverter[22] (shown in Figures S2-S4 in the Supporting Information). As most of these packages support the loading of data in the imzML format (the exception being SCiLS Lab[19]) it is now possible to process data from any instrument that has a corresponding imzML converter, while simultaneously increasing the preprocessing methods available to the analyst through the choice of software. The drawback that still remains is that these software packages do not export processed data or partial results to a format readable by other software packages, meaning that the user is restricted to the functionality included within their chosen software tool.

Here we present software that can be used through the entire analysis workflow, from raw data through preprocessing to multivariate analysis, for datasets acquired from single experiments to large multi-instrument, multi-modality, multi-center studies. Such a wide collection of capabilities does not exist in any currently available software, which are variously limited by the instruments supported,[19] preprocessing capabilities[15] or support for

multivariate analysis.[11,20,21]

Experimental

All experiments were conducted in accordance with local ethical guidelines for animal care. MALDI MSI data of a sagittal section of rat brain were acquired using a QSTAR Elite (SCIEX, Ontario, Canada) as described by Carter *et al.*[23] Coronal rat brain sections were prepared using the protocols described by Steven *et al.*[5] and MALDI MSI data were acquired using either an ultrafleXtreme (Bruker, Bremen, Germany) or Synapt G2 (Waters, Manchester, UK) with a pixel size of 100 μm. DESI MSI data of a fingerprint were acquired using an LTQ Orbitrap Velos (Thermo Scientific, Bremen, Germany) as described by Bailey *et al.*,[24] in negative ion mode. Mouse lung was sectioned at 12 μm thick and thaw mounted on ITO-coated glass slides (Bruker). SIMS MSI data were acquired using a TOF-SIMS IV (IONTOF, Muenster, Germany) equipped with a 25 keV Bi_3^+ primary ion source delivering an ion dose of 1.1 x 10^{10} ions per cm^2.

QSTAR Elite data were converted to mzML using MS Data Converter version 1.3 (SCIEX). Synapt G2 and LTQ Orbitrap Velos data were converted to mzML using msconvert as part of ProteoWizard.[25] ultrafleXtreme data were converted to mzML using CompassXport (Bruker). All mzML data were converted to imzML using imzMLConverter.[22] TOF-SIMS IV data were converted to GRD format using SurfaceLab 6 (IONTOF) and then to imzML using imzMLConverter.[22]

The interface for the SpectralAnalysis is shown in Figure S1 in the Supporting Information and was written primarily in MATLAB to provide an easier means of modification and custom access to and manipulation of data, with some features written in C and Java for performance improvements. The source code and an executable version (which has no additional software requirements) will be made available at https://github.com/AlanRace/SpectralAnalysis.

Discussion

The remainder of the manuscript will discuss the novel combination of features included within SpectralAnalysis. Full descriptions of included algorithms for memory efficiency and ensuring a consistent m/z axis are omitted here for brevity but can be found within the Supporting Information.

Preprocessing

The purpose of preprocessing is to remove artefacts introduced during the data acquisition stage, to make spectra comparable to one another and to improve the efficacy of peak detection routines. The common preprocessing methods applied in mass spectrometry are smoothing, baseline correction, normalisation and peak detection. Here we include commentary on each of these methods as well as an additional step which is not often discussed, methods for ensuring a consistent m/z axis across a dataset.

The suitability of certain preprocessing methods largely depends on the nature of the data to be analysed. Preprocessing methods included in each vendor's software (for which there is a publicly available description) are also included in SpectralAnalysis, making it one of the most feature complete software tools currently available with the widest applicability to process spectra acquired using any instrument. Taking this a step further and allowing the effects of the preprocessing methods to be visualised in real time, enables the user to select appropriate methods and associated parameters for optimally removing experimental artefacts and noise. The interface for performing this is shown in Figure S5.

Furthermore, it is possible to create a custom 'preprocessing workflow' that allows a sequence of preprocessing methods to be applied in a user specified order using the interface shown in Figure S6. Custom workflows can be saved, shared and reused. This not only gives the user great flexibility over the transformations applied to each spectrum, but also enables the recreation of previously published routines such as LIMPIC[26] as well as in-house

workflows. This allows rapid evaluation and incorporation of newly developed preprocessing workflows without the need for additional software and provides a route for methods to be published alongside articles or submitted as part of the review process. A dataset preprocessed in this way can also be exported to imzML, allowing preprocessing to be performed within SpectralAnalysis and enabling subsequent processing to be performed in the analyst's software package of choice, archival of preprocessed data or submission of preprocessed data to public repositories.

Recently Oetjen *et al.*[27] made a number of 3D MSI datasets publicly available to stimulate development of software capable of handling, processing and evaluating the reproducibility of such data. The preprocessing techniques included within this software can be used to help answer one of the key questions asked by Oetjen and coworkers, what method(s) increase reproducibility of the experiments?[27] In other words, which method(s) best correct for differences in sample preparation between two sections, day-to-day variation and pixel-to-pixel variation? This is enabled by the extensibility of the software (discussed below), providing a powerful tool supporting the benchmarking of new methods against any and all currently implemented methods. An example preprocessing workflow applied to one of the publicly available datasets is presented in Figure S7 in the Supporting Information, however a thorough evaluation of preprocessing methods within this context is beyond the scope of this article.

The order in which the preprocessing methods are applied has an effect on the resulting data. The widely accepted order for preprocessing time of flight (TOF) data is smoothing or denoising followed by baseline correction prior to peak detection.[26,28,29] Noise reduction or removal methods such as baseline correction and smoothing aim to improve the peak detection method of choice.

When comparing, averaging or otherwise mathematically manipulating two or more spectra it is important that they are represented by the same number of m/z bins with the same m/z intervals. This is a common requirement in data reduction routines, where one or more

summary spectra, such as the mean spectrum, are used for feature detection[29,30] or peak alignment.[31] This also has the benefit of enabling spectra to be directly stored as a matrix (a 2D matrix as required for many post processing techniques such as PCA or a 3D 'datacube' for efficient image generation and manipulation). Methods for achieving this are defined by Algorithms S1-S5 and visualised in Figures S8-S10, with detailed discussion on each method found elsewhere.[32]

Smoothing aims to remove small, local, fluctuations in intensity, often caused by noise, that prevent peak detection algorithms from functioning optimally. The *de facto* standard smoothing method used is Savitzky-Golay due to its intensity preserving properties.[33] This, along with other commonly used methods such as moving average and Gaussian, are window based techniques, meaning that they consider a set number of data points at once, the 'window', to generate a single data point in the resulting data. The window is then 'slid' along the data to the next point where a new window is considered.

Care must be taken when combining certain methods for ensuring a consistent m/z axis (a set of m/z values that is the same for every spectrum, discussed in more detail elsewhere[32]) and window based preprocessing methods as peak widths often vary across the mass range. For example, when processing TOF data and the detector based m/z axis is used, the chosen window size for the smoothing function is appropriate for a peak at m/z 826, but applying smoothing with the same window size to a peak at m/z 104 causes peak broadening and a reduction in the peak height, as shown in Figure S11a. However, when the same number of data points span both peaks, neither peak is broadened and the heights are retained, minus noise. The FWHM of the peak at m/z 104 in Figure S11a is 0.03, compared to 0.02 in Figure S11b which is equivalent to the mass resolving power being reduced to 3500 from 5200 (calculated at m/z 104.08537).

Baseline correction aims to remove an experimental artefact, often attributed to chemical noise, to aid peak detection and increase comparability between spectra. The effect is more pronounced when acquiring data over a large mass range and is a common feature in protein

imaging by MALDI MS due to the use of a linear TOF. The type, or lack thereof, of baseline present in the data is dependent on the both the instrument and experimental parameters, such as the mass resolving power, mass range, the laser power (inducing and subsequently increasing fragmentation), the analyte and the matrix used. The choice of baseline correction method will depend on the style of baseline present, where some methods make certain assumptions about the shape of the baseline. Different methods and their corresponding assumptions are discussed in more detail elsewhere.[32]

Normalisation is a relatively controversial topic in mass spectrometry imaging with a significant amount of debate still ongoing. A detailed review of common normalisation methods is provided by Deininger *et al.*[34] Since then an additional method for normalisation has been proposed by Fonville *et al.*[35] A visual comparison of these normalisation methods applied to a sagittal section of rodent brain is shown in Figure 1.

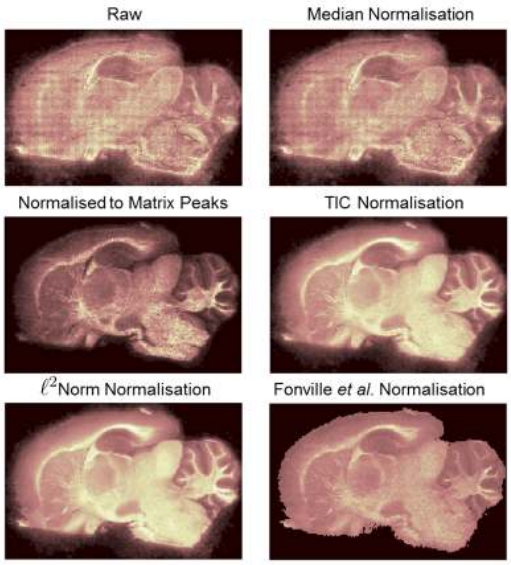


Figure 1: Comparison of normalisation techniques applied to the same ion image (m/z 810 in a sagittal section of formalin fixed rat brain).

In the raw image there are quite apparent experimental artefacts in the form of criss-cross patterns. These patterns are removed in all methods except median normalisation, which normalises to an approximate measure of the intensity of the baseline. In this dataset there

is no baseline, resulting in a similar median value for every pixel (which in this case becomes a measure of noise rather than the baseline, and is approximately 2 arb. unit). As the spectral sparsity of a given dataset increases, the median tends towards 0, at which point this method becomes inappropriate for normalisation. Alternatively, the zero values can be omitted while calculating the median value. In this case it is likely that the estimation of the baseline, or noise level, will be an overestimate and small, low intensity, spectral features may be removed as part of the baseline correction process.

The other normalisation methods considered make assumptions about the nature of the data. A frequently employed method, especially in drug quantification studies, is normalisation to the intensity of an internal standard selected as a close mimic of the compound of interest, for example a deuterated analogue.[36] In situations where an internal standard was not included, a pseudo internal standard can be selected from components that are present within the data. In MALDI MSI, normalising to matrix peaks assumes that the matrix should be constant across the image and so by normalising to the matrix peaks the aim is to compensate for any heterogeneity of the matrix distribution. When considering only matrix regions, the sum of all detected matrix ions (fragments, clusters and adducts) could potentially provide a good normalisation factor. However, once an analyte is incorporated suppression effects can cause ions to be detected differently, or not at all, and so this method becomes less suitable. As this relies on the matrix peaks this method is only applicable to MALDI data, however the matrix peaks could be replaced with other experimental constants prevalent in other techniques such as solvent peaks in desorption electrospray ionisation (DESI) or liquid extraction surface analysis (LESA), but with similar caveats.

The total ion current (TIC), and similarly the ℓ^2, normalisation method makes the assumption that at every pixel location the same number of ions should be detected. In homogeneous single compound samples this would hold true, however any form of heterogeneity renders this assumption inappropriate. It could be argued that within a given area there is only a given amount of charge present required for the formation of ions and so

despite the heterogeneity this method is applicable. Due to varying proton affinities of molecules present, suppression effects and reactions that may occur within the plume (for example charge transfer or metastable fragmentation) this assumption is unlikely to hold true. Fonville *et al.*[35] attempt to provide a more robust method of normalisation that does not suffer from the issues listed above, by only considering signal from the analyte when constructing the scaling factor for each pixel. However, given the heterogeneity of the analyte this is still not an ideal solution.

Until there is a consensus on which method is most applicable in which situation, it falls to the analyst to evaluate and investigate these methods and their appropriateness in the context of their data and so each of the methods described above are included within SpectralAnalysis.

The method employed for generating ion images from a MSI dataset can result in different apparent spatial distributions, demonstrated in Figure S12. Methods that simply extract a single m/z channel are more susceptible to noise in the data, and so ion images generated with such methods often appear to include high frequency fluctuations in intensity. Through the application of appropriate preprocessing prior to image generation these fluctuations can be lessened, producing a smoother image, which can help reveal patterns previously masked by noise.

A difference in the spatial distribution produced by each of the ion image generation methods in both cases, with and without preprocessing being applied, can be observed in Figure S12. The difference is primarily between the methods that extract a value at a single m/z channel and those that integrate across the peak. This can likely be explained by the effect illustrated in Figure S13 where the distribution of intensities when integrating the left side of the peak is different to that of the right side, potentially due to unresolved ions having different spatial distributions.

Multivariate Analysis

Multivariate analysis techniques have been shown to be a powerful tool for aiding interpretation of these complex datasets. Despite this, very few freely available software packages include such techniques and those that do only include one or two.[12,14] The efficacy of any single technique used in isolation has recently been brought into question.[37]

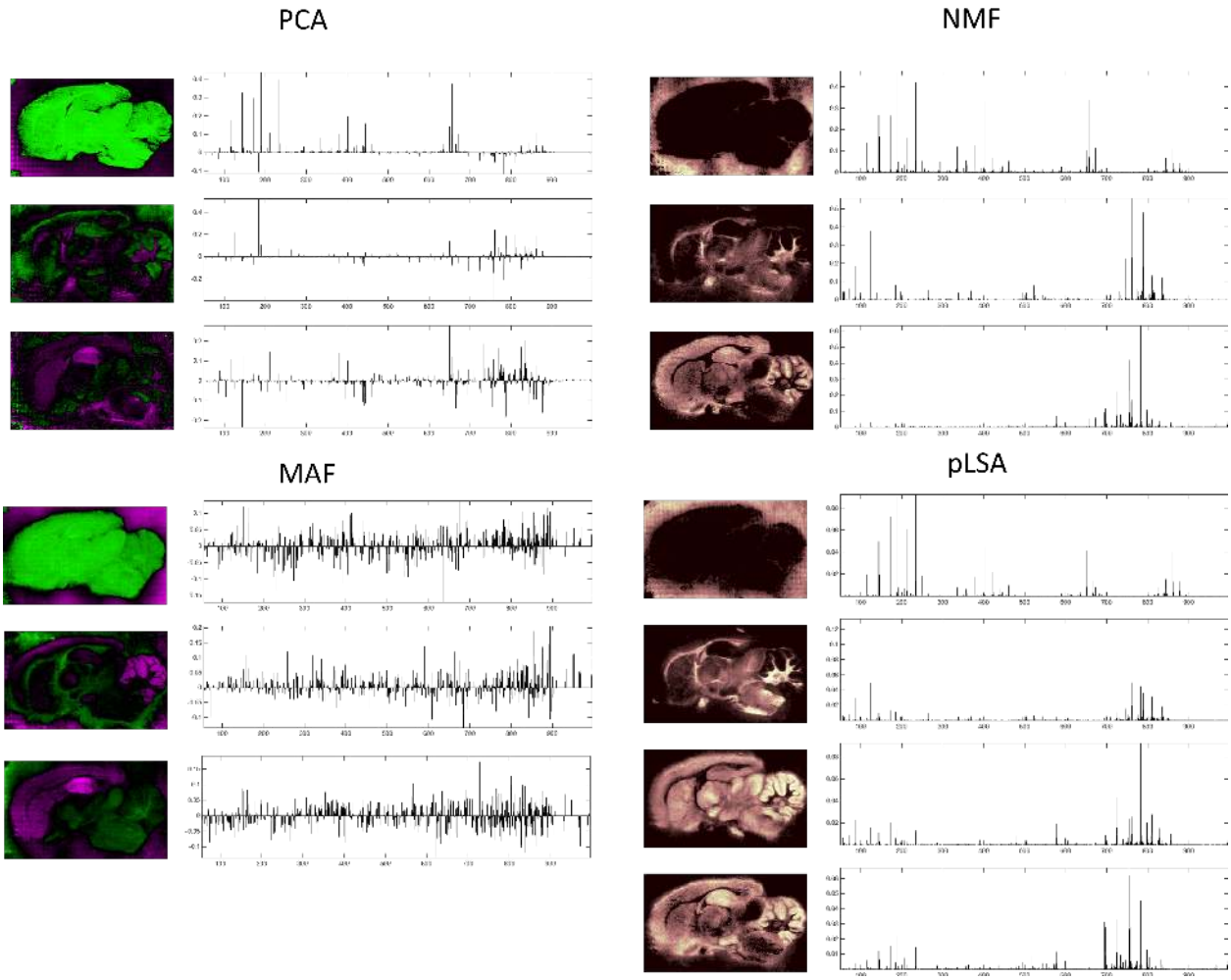


Figure 2: Selected factors from principal component analysis (PCA), non-negative matrix factorisation (NMF), maximum autocorrelation factor (MAF) and probabilistic latent semantic analysis (PLSA) applied to a MALDI MS image of a sagittal section of rat brain.

To address this, SpectralAnalysis includes principal component analysis (PCA), non-negative matrix factorisation (NMF), maximum autocorrelation factor (MAF)[38] and probabilistic latent semantic analysis (PLSA).[39] Selected factors from each of the techniques

applied to a MALDI MS image of a sagittal section of rat brain[23] are shown in Figure 2 using diverging colour schemes where appropriate.[40] Different anatomical features can be distinguished in the different techniques, for example the hippocampus region is highlighted most prominently in the third MAF factor and is visible (but could easily be overlooked) in the fourth PLSA latent variable and third PCA component, but is not prominent within any of the NMF factors. However, NMF highlights significant contrast between the gray and white matter regions in the second and third components which are less obvious in the PCA and MAF. Depending on the question at hand, the increased ability to differentiate spectrally different regions, such as anatomy, could be very powerful, especially in drug distributions studies.

Supporting Large Scale MSI

Multimodality Data

It is becoming increasingly desirable to incorporate multiple additional techniques into the analysis of mass spectrometry imaging data. This can range from simply including histology images to determine co-localisation with anatomy, through the inclusion of additional MSI data (either from the same instrument or a complementary one),[3] to the inclusion of other spectral imaging modalities such as Raman.[2] To cater for this scenario, SpectralAnalysis was written in such a way that enables any spectral data to rapidly be incorporated, allowing any of the core functionality (such as preprocessing and multivariate analysis) to be performed without alteration. A selection of data from different modalities processed using SpectralAnalysis is given in Figure 3. Although some preprocessing techniques are only suitable for specific styles of data, the end goal largely remains the same and the majority of algorithms for smoothing, baseline correction and peak detection included are technique independent providing a powerful platform for multimodality processing, investigation and visualisation.

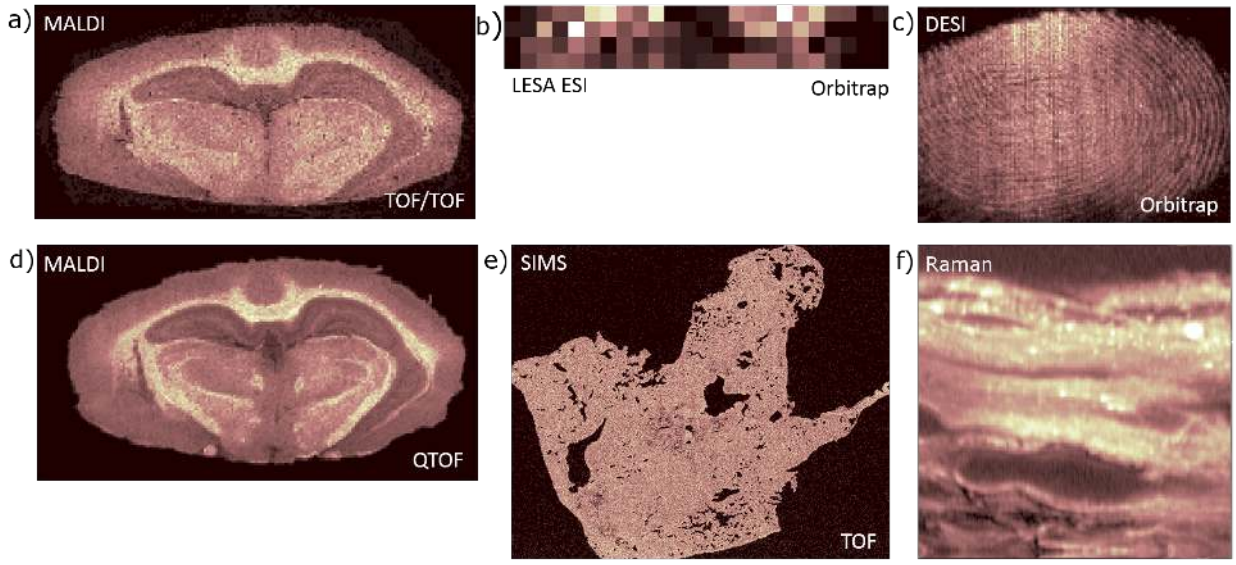


Figure 3: A selection of different modality imaging data, acquired at different length scales, processed using SpectralAnalysis. a) MALDI MSI (m/z 826.6) of coronal rat brain acquired with a pixel size of 100 μm using an ultrafleXtreme (Bruker Daltonics) b) LESA MSI of liver acquired with a pixel size of 1000 μm using an Orbitrap Elite (Thermo Scientific) c) DESI MSI (m/z 509.36) of a fingerprint acquired with a pixel size of 200 μm using an LTQ Orbitrap Velos (Thermo Scientific) d) MALDI MSI (m/z 826.6) of coronal rat brain acquired with a pixel size of 100 μm using a Synapt G2S (Waters) e) SIMS data (m/z 104.1) of murine lung acquired with a pixel size of 7 μm using a TOF-SIMS IV (ION-TOF) f) Spontaneous Raman scattering (SRS) of a living skin equivalent with a pixel size of 0.2 μm using a home built system (NPL).

Handling Extremely Large Datasets

As instruments develop and improvements are made in both the mass resolution and the lateral resolution, the data size is correspondingly increasing, with raw data easily capable of exceeding 10s to 100s of GB for a single MSI dataset. Furthermore, the move towards larger biomedical studies with increased cohort and sample numbers (including replicates) significantly increases the data handling challenge. The vast majority of MSI software loads the data to be processed into RAM before any visualisation or analysis can be performed. This then introduces a restriction on the size of the data that can be processed based on the hardware of the computer being used, and as the data size is rapidly outpacing the hardware specifications this is becoming an increasing problem and may render some software/hardware/data combinations unusable.

This problem has been addressed previously by enabling the ability to load and analyse a subsection of the dataset, limiting the number of pixels, the mass range, or both.[15,18] SpectralAnalysis also includes this option and expands upon it by allowing the user to select an arbitrarily shaped region of interest as well as an optional mass range limit to be loaded into memory, the interface for this is shown in Figure S14. In order to do this the m/z axis must be consistent and so any of the techniques discussed above can be employed to ensure this. In some cases the parameters can be specified such that this process also contributes to the reduction of data (such as rebinning) at the cost of potentially discarding information.

However, this approach does not solve all situations as it limits the analyst's view of the dataset as a whole and involves discarding data (and potentially analytically useful information) which can be detrimental to the analysis, while still being fundamentally constrained by the RAM available. This issue of datasize is compounded when multiple MS images are combined as discussed below, requiring the analyst to compromise even further to be able to visualise the data. SpectralAnalysis includes memory efficient methods that enable datasets vastly exceeding the size of the available RAM to be visualised, preprocessed and analysed using multivariate analysis.

Generation of individual ion images does not require the whole dataset to be loaded into memory. Instead only the data points that fall within the peak boundaries (m_{min} and m_{max}) are required so that one of the image generation methods shown in Figure S12 can be applied. This can also be taken a step further, and since the calculation of an intensity at a given pixel is completely independent of all other pixels, only one spectrum is required in memory at a given point in time. This significantly reduces the amount of memory required, as a single spectrum ranges from 100s kB to 1-2 MB, compared to the 10s of GBs for the whole dataset. In many cases it is desirable to preprocess the data prior to ion image generation. The algorithm presented in Algorithm S6 presents a memory efficient method of generating ion images from preprocessed data. Each spectrum is loaded in sequentially, preprocessed and then the data points within peak limits are extracted and an intensity is generated based on the image generation method of choice. The spectrum can then be removed from memory before the next is loaded in. This reduces the amount of memory required to the size of a single spectrum, plus the size of the ion image(s) to be generated, which is orders of magnitude smaller than the whole data, allowing TBs of data to be visualised on even the most memory constrained systems.

Peak detection is often performed on spectral representations of the data.[29] As above, these only require a single spectrum to be loaded into memory at once and can be generated in a memory efficient manner using Algorithm S9 (in the Supporting Information). It is possible to generate multiple representations at once, requiring only a single pass through the data, by including additional update methods after line 8 of Algorithm S9 (in the Supporting Information). This provides a memory efficient method of generating all spectral representations proposed by McDonnell *et al.*[29] for optimal peak detection in a given dataset.

By combining the above methods it is possible to reduce the MS image to a 'datacube' in a memory efficient manner, using Algorithm S10 (in Supporting Information). In this case, only a single spectrum and the datacube is required to be in memory at any one point in time. This allows reduction of data to peak lists without a limitation applied to the number

of peaks retained.

The algorithm as it is presented reduces and loads the data into memory, however this can also be used to write the reduced data to disk by altering line 10 in Algorithm S10 to be a disk write instead of a matrix update. In this case only a single spectrum is required to be in memory, making this process feasible on memory constrained systems where the datacube is larger than that of the RAM. Then all methods for handling large datasets described above can be employed to visualise and further process the data.

A previously published memory efficient PCA algorithm is also fully integrated into SpectralAnalysis.[41] The capabilities of this algorithm have been expanded to include the ability to use any user defined preprocessing workflow and to allow memory efficient scaling of the data (shown in Algorithm S11) to be applied by each of the techniques investigated by Tyler *et al.*[42]

Multi-dataset Studies

The ability to combine multiple datasets acquired separately but which together form a single experiment is extremely powerful. Consider the experiments presented by Carter *et al.*[23] and Griffiths *et al.*[43] where different sample preparation methods are being compared but the data were collected as separate mass spectrometry images. To compare these data the analyst would have to load an image, perform any preprocessing necessary, search for an ion image of interest, then repeat for any other dataset being compared. Then the analyst would have to ensure that the intensity scales that the images were presented on were comparable prior to any interpretation. This is a laborious and time consuming process and would have to be repeated for each ion image that was investigated. While this is manageable for an experiment only consisting of two MS images, when trying to perform this on 14 serial sections, such as the data presented by Steven *et al.*,[44] it becomes impractical.

The imzMLConverter tool[22] provides the ability to tile and combine multiple imzML files together into a single imzML file. In combination with SpectralAnalysis, this feature

enables the analyst to rapidly compare the spatial distributions and relative abundances of ions in the visualisation software of their choice without needing to open multiple datasets and manually ensure colour schemes and intensity ranges of each ion image generated for each dataset are comparable. When considering this, and additionally the support for large data discussed above, the data presented by Steven *et al.*[44] and Oetjen *et al.*[27] become much more manageable to process, and much less error prone, when the whole study is considered as one large dataset. Ion images can be generated in seconds (an example of which is given in Figure 4) rather than minutes to hours when having to process each dataset individually and manually. This can be used to successfully analyse data from larger studies such as those published by McDonnell and co-workers.[6–8]

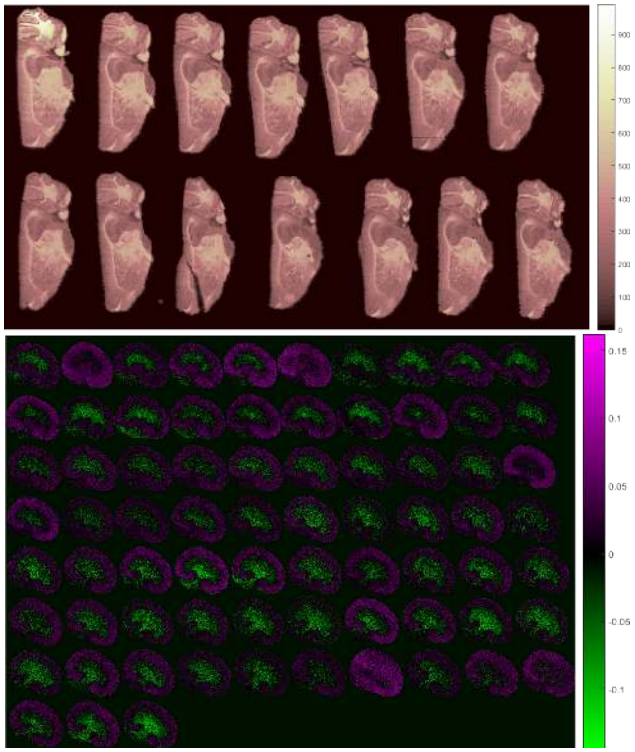


Figure 4: Visualisation and processing of large data (top, 10 GB comprised of 104,916 spectra and bottom, 42 GB comprised of 1,362,830 spectra) from multiple experiments within SpectralAnalysis. **Top:** MSI data (m/z 826.6) of 14 serial sagittal sections of mouse brain. Data were acquired from 14 separate imaging acquisitions by Steven *et al.*[44] Data were subsequently combined using imzMLConverter and processed together.[22] **Bottom:** Principal component 2 calculated using memory efficient PCA[41] on the 3D kidney dataset made publicly available by Oetjen *et al.*[27]

If a 3D imzML dataset is opened within SpectralAnalysis, such as those recently released to the community by Oetjen *et al.*,[27] then this is automatically detected and the data is presented as a 2D tile. All included algorithms and features can then be applied to the data, for example memory efficient PCA as shown in Figure 4. This provides the ability to visualise 3D data while also including the tools necessary to be able to evaluate suitability of methods for handling variations in signal intensity observed between different sections, as was noted as one of the main reasons for releasing the data. While 3D visualisation is not natively included, it could be included at a later date due to the extensible nature of the software (discussed below).

Extensibility

SpectralAnalysis was developed with extensibility in mind, providing a platform for visualisation and processing that it is simple to include additional data format readers, preprocessing, multivariate analysis and clustering algorithms without the requirement to write new user interface or data visualisation code. A block diagram visualising the core components that can be extended is shown in Figure S15 in the Supporting Information. A 'Parser' handles the reading of a given file format, for example imzML, to get meta information such as the image dimensions (width, height, depth) and whether the data are stored in sparse format or dense (to determine the need to ensure a consistent m/z axis as discussed previously) as well as to read parts of the data from disk. Extension of this allows data in different formats (such as older MSI formats like Analyze 7.5) as well as file formats associated with other imaging modalities to be visualised and processed. The 'DataRepresentation' determines how the data are to be handled, either in memory or left on disk, and could be extended to include additional capabilities such as a hybrid of the two (cached data in memory, majority remain on disk). 'Preprocessing' and associated subcomponents include all of the features discussed in the 'Preprocessing' section and can be extended to include methods or algorithms that are currently omitted or to develop new algorithms and make use of the real-time visualisation

of the effects on spectral data. 'Postprocessing' includes all multivariate analysis techniques shown in Figure 2, as well as clustering algorithms not shown, and can be extended to include additional algorithms. This provides a platform for rapid testing of algorithms at every stage of the analysis process on multiple modality datasets with instant visualisation of the results.

By having this design philosophy it is hoped that SpectralAnalysis will enable the community to evaluate current methods against one another and, most importantly, evaluate current methods against newly developed ones. This is especially important for quantification studies, where normalisation plays a significant role in the data processing, but remains a heavily debated and actively researched topic.

Conclusions

SpectralAnalysis provides a unique, and currently the most exhaustive, collection of algorithms for preprocessing and subsequent multivariate analysis of spectral imaging data. This, combined with the flexibility of the extensibility to include additional algorithms, results in a platform suitable for comparisons of preprocessing methods on MSI data acquired on any instrument.

Due to the capability of handling multiple spectral imaging modalities, each of which capture different information about the sample, SpectralAnalysis would be an excellent platform on which to develop and integrate multi-modality processing techniques such as image fusion. Image fusion aims to combine data from multiple sources to gain information that was not present in each source in isolation. For example, the combination of a high spatial resolution, single channel image with a multispectral, but low spatial resolution image resulting in a high spatial multispectral image.[45] This could be extended and applied to multiple MSI datasets to combine, for example, the high spatial resolution of SIMS data with the high mass range of MALDI data.

Acknowledgements

AMR (2010 - 2014), ADP (2009 - 2013), AD (2012 - 2016) and RTS (2009 - 2013) gratefully acknowledge financial support from the EPSRC through studentships from the PSIBS Doctoral Training Centre (EP/F50053X/1). JB gratefully acknowledges funding from NPL strategic research programmes 116301 and 117194 and strategic capability programme AIMS HIGHER.

Thanks to Jocelyn Sarsby (University of Liverpool) for preparing and acquiring the LESA MSI data. Thanks to Peter Marshall (GSK) for preparing the lung tissue and Melissa Passerelli (NPL) for acquiring the corresponding SIMS data. Thanks to Epistem Ltd (Manchester) for preparing the living skin equivalent sample and to Alasdair Rae (NPL) for acquiring the SRS data.

Supporting Information Available

Detail on how to convert data into a suitable format (using imzMLConverter[22]) and algorithms describing preprocessing and memory efficient methods included within SpectralAnalysis. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Bocklitz, T.; Crecelius, A.; Matthaus, C.; Tarcea, N.; Von Eggeling, F.; Schmitt, M.; Schubert, U.; Popp, J. *Anal. Chem.* **2013**, *85*, 10829–10834.

(2) Ahlf, D. R.; Masyuko, R. N.; Hummon, A. B.; Bohn, P. W. *Analyst* **2014**, *139*, 4578–4585.

(3) Eberlin, L. S.; Liu, X.; Ferreira, C. R.; Santagata, S.; Agar, N. Y.; Cooks, R. G. *Anal. Chem.* **2011**, *83*, 8366–8371.

(4) Eijkel, G.; Kükrer Kaletaş, B.; Van der Wiel, I.; Kros, J.; Luider, T.; Heeren, R. *Surf. Interface Anal.* **2009**, *41*, 675–685.

(5) Steven, R. T.; Bunch, J. *Anal. Bioanal. Chem.* **2013**, *405*, 4719–4728.

(6) Heijs, B.; Tolner, E. A.; Bovée, J. V.; van den Maagdenberg, A. M.; McDonnell, L. A. *J. Proteome Res.* **2015**, *14*, 5348–5354.

(7) Balluff, B.; Frese, C. K.; Maier, S. K.; Schöne, C.; Kuster, B.; Schmitt, M.; Aubele, M.; Höfler, H.; Deelder, A. M.; Heck, A. J.; Hogendoorn, P. C.; Morreau, J.; Altelaar, A. M.; Walch, A.; McDonnell, L. A. *J. Pathol* **2015**, *235*, 3–13.

(8) Carreira, R. J.; Shyti, R.; Balluff, B.; Abdelmoula, W. M.; van Heiningen, S. H.; van Zeijl, R. J.; Dijkstra, J.; Ferrari, M. D.; Tolner, E. A.; McDonnell, L. A.; Maagdenberg, A. M. J. M. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 853–861.

(9) Hosokawa, N.; Sugiura, Y.; Setou, M. *Imaging Mass Spectrometry*; Springer, 2010; pp 113–126.

(10) Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M.-F.; Stoeckli, M.; Spengler, B.; Römpp, A. *J Proteomics* **2012**, *75*, 5106–5110.

(11) Robichaud, G.; Garrard, K. P.; Barry, J. A.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 718–721.

(12) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. *Bioinformatics* **2015**, 2418–20.

(13) Parry, R. M.; Galhena, A. S.; Gamage, C. M.; Bennett, R. V.; Wang, M. D.; Fernández, F. M. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 646–649.

(14) Rübel, O.; Greiner, A.; Cholia, S.; Louie, K.; Bethel, E. W.; Northen, T. R.; Bowen, B. P. *Anal. Chem.* **2013**, *85*, 10354–10361.

(15) Klinkert, I.; Chughtai, K.; Ellis, S. R.; Heeren, R. *Int. J. Mass Spectrom.* **2014**, *362*, 40–47.

(16) Källback, P.; Shariatgorji, M.; Nilsson, A.; Andrén, P. E. *J Proteomics* **2012**, *75*, 4941–4951.

(17) Källback, P.; Nilsson, A.; Shariatgorji, M.; Andrén, P. E. *Anal. Chem.* **2016**,

(18) Paschke, C.; Leisner, A.; Hester, A.; Maass, K.; Guenther, S.; Bouschen, W.; Spengler, B. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1296–1306.

(19) SCiLS / SCiLS Lab - The statistical analysis software. `http://scils.de/software/`, 2014; Accessed: 21/05/2014.

(20) Biosoft, P. MALDIVision. `http://www.premierbiosoft.com/maldi-tissue-imaging/index.html`, Accessed: 20/03/2016.

(21) imabiotech, Quantinetix. `https://www.imabiotech.com/Benefits`, Accessed: 20/03/2016.

(22) Race, A. M.; Styles, I. B.; Bunch, J. *J Proteomics* **2012**, *75*, 5111–5112.

(23) Carter, C. L.; McLeod, C. W.; Bunch, J. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1991–1998.

(24) Bailey, M. J.; Bradshaw, R.; Francese, S.; Salter, T. L.; Costa, C.; Ismail, M.; Webb, R. P.; Bosman, I.; Wolff, K.; de Puit, M. *Analyst* **2015**, *140*, 6254–6259.

(25) Chambers, M. C. et al. *Nature biotechnology* **2012**, *30*, 918–920.

(26) Mantini, D.; Petrucci, F.; Pieragostino, D.; Del Boccio, P.; Di Nicola, M.; Di Ilio, C.; Federici, G.; Sacchetta, P.; Comani, S.; Urbani, A. *BMC bioinformatics* **2007**, *8*, 101.

(27) Oetjen, J. et al. *GigaScience* **2015**, *4*, 1–8.

(28) Gibb, S.; Strimmer, K. *Bioinformatics* **2012**, *28*, 2270–2271.

(29) McDonnell, L. A.; Van Remoortere, A.; De Velde, N.; Van Zeijl, R. J.; Deelder, A. M. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1969–1978.

(30) Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A.; Kobayashi, R. *Bioinformatics* **2005**, *21*, 1764–1775.

(31) Alexandrov, T.; Kobarg, J. H. *Bioinformatics* **2011**, *27*, i230–i238.

(32) Race, A. M. Investigation and interpretation of large mass spectrometry imaging datasets. Ph.D. thesis, University of Birmingham, 2016.

(33) Savitzky, A.; Golay, M. J. *Anal. Chem.* **1964**, *36*, 1627–1639.

(34) Deininger, S.-O.; Cornett, D. S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. *Anal. Bioanal. Chem.* **2011**, *401*, 167–181.

(35) Fonville, J. M.; Carter, C.; Cloarec, O.; Nicholson, J. K.; Lindon, J. C.; Bunch, J.; Holmes, E. *Anal. Chem.* **2012**, *84*, 1310–1319.

(36) Pirman, D. A.; Reich, R. F.; Kiss, A.; Heeren, R. M.; Yost, R. A. *Anal. Chem.* **2012**, *85*, 1081–1089.

(37) Jones, E. A.; van Remoortere, A.; van Zeijl, R. J.; Hogendoorn, P. C.; Bovee, J.; Deelder, A. M.; McDonnell, L. A. *PloS one* **2011**, *6*, e24913.

(38) Nielsen, A. A. *Image Processing, IEEE Transactions on* **2011**, *20*, 612–624.

(39) Hanselmann, M.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M.; Hamprecht, F. A. *Anal. Chem.* **2008**, *80*, 9649–9658.

(40) Race, A. M.; Bunch, J. *Anal. Bioanal. Chem.* **2015**, *407*, 2047–2054.

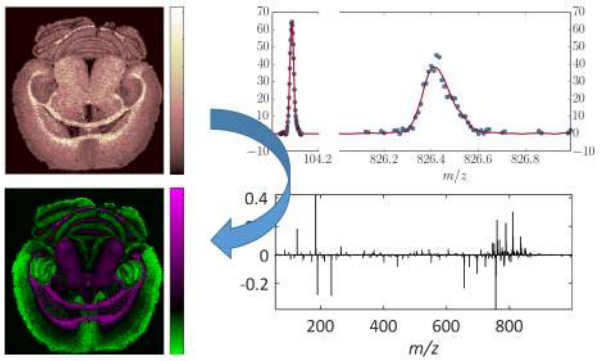
(41) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. *Anal. Chem.* **2013**, *85*, 3071–3078.

(42) Tyler, B. J.; Rayal, G.; Castner, D. G. *Biomaterials* **2007**, *28*, 2412–2423.

(43) Griffiths, R. L.; Sarsby, J.; Guggenheim, E. J.; Race, A. M.; Steven, R. T.; Fear, J.; Lalor, P. F.; Bunch, J. *Anal. Chem.* **2013**, *85*, 7146–7153.

(44) Steven, R. T.; Race, A. M.; Bunch, J. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 801–804.

(45) Van de Plas, R.; Yang, J.; Spraggins, J.; Caprioli, R. M. *Nature methods* **2015**, *12*, 366–372.


70
60
50
40
30
20
10
0
−10
104.2
826.2
826.4
826.6
826.8
m/z
0.4
0
-0.2
200
400
600
800
m/z

For TOC only.