Bioinformatics, 31(19), 2015, 3198–3206
doi: 10.1093/bioinformatics/btv356
Advance Access Publication Date: 10 June 2015
Original Paper

OXFORD

Bioimage informatics

EXIMS: an improved data analysis pipeline based on a new peak picking method for EXploring Imaging Mass Spectrometry data

Chalini D. Wijetunge[1,*], Isaam Saeed[1], Berin A. Boughton[2], Jeffrey M. Spraggins[3,4], Richard M. Caprioli[3,4,5,6], Antony Bacic[2,7,8], Ute Roessner[2] and Saman K. Halgamuge[1]

[1]Department of Mechanical Engineering, [2]Metabolomics Australia, University of Melbourne, Parkville, VIC 3010, Australia, [3]Department of Biochemistry, [4]Mass Spectrometry Research Centre, [5]Department of Chemistry and [6]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA, [7]ARC Centre of Excellence, School of Biosciences and [8]Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, VIC 3010, Australia

*To whom correspondence should be addressed.
Associate Editor: Robert Murphy

Received on October 20, 2014; revised on June 2, 2015; accepted on June 4, 2015

Abstract

Motivation: Matrix Assisted Laser Desorption Ionization-Imaging Mass Spectrometry (MALDI-IMS) in 'omics' data acquisition generates detailed information about the spatial distribution of molecules in a given biological sample. Various data processing methods have been developed for exploring the resultant high volume data. However, most of these methods process data in the spectral domain and do not make the most of the important spatial information available through this technology. Therefore, we propose a novel streamlined data analysis pipeline specifically developed for MALDI-IMS data utilizing significant spatial information for identifying hidden significant molecular distribution patterns in these complex datasets.

Methods: The proposed unsupervised algorithm uses Sliding Window Normalization (SWN) and a new spatial distribution based peak picking method developed based on Gray level Co-Occurrence (GCO) matrices followed by clustering of biomolecules. We also use gist descriptors and an improved version of GCO matrices to extract features from molecular images and minimum medoid distance to automatically estimate the number of possible groups.

Results: We evaluated our algorithm using a new MALDI-IMS metabolomics dataset of a plant (Eucalypt) leaf. The algorithm revealed hidden significant molecular distribution patterns in the dataset, which the current Component Analysis and Segmentation Map based approaches failed to extract. We further demonstrate the performance of our peak picking method over other traditional approaches by using a publicly available MALDI-IMS proteomics dataset of a rat brain. Although SWN did not show any significant improvement as compared with using no normalization, the visual assessment showed an improvement as compared to using the median normalization.

Availability and implementation: The source code and sample data are freely available at http://exims.sourceforge.net/.

Contact: awgcdw@student.unimelb.edu.au or chalini_w@live.com

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

1 Introduction

Matrix Assisted Laser Desorption Ionization-Imaging Mass Spectrometry (MALDI-IMS) (Caprioli *et al.*, 1997) has gained popularity in 'omic' sciences because of its ability to produce detailed spatial information of molecules in a given biological sample. By using a section of a tissue, IMS can visualize a wide variety of biomolecules such as proteins, lipids and metabolites (Sugiura and Setou, 2010). Although this technique has been well established in proteomics (Groseclose *et al.*, 2008; Jardin-Mathé *et al.*, 2008; Lagarrigue *et al.*, 2012), the scope of the biological questions that can be addressed through this technology is still limited mainly due to the shortage of streamlined data analysis pipelines.

Figure 1 illustrates the typical MALDI-IMS data acquisition workflow. In general, MALDI-IMS experiments are performed by thaw mounting a frozen tissue section onto MALDI target, a matrix is applied that promotes molecular desorption and ionization, and a spectrum is acquired at each of the defined pixels. That is, for each spatial coordinate of the optical image a mass spectrum is obtained which represents the intensities of ionizable molecules with various mass/charge (*m/z*) values. Therefore, a MALDI-IMS dataset can be considered as a collection of spectra. Moreover, by plotting the intensity values corresponding to one particular *m/z* value, the intensity image represents the spatial distribution of the corresponding molecule(s). Therefore, a MALDI-IMS dataset can also be considered as a collection of intensity images. Biologists are mainly interested in identifying molecules that are differentially distributed (show different abundances) in diverse anatomical regions thus providing valuable information about the living organism under study.

Usually the MALDI-IMS datasets are complex and extremely large in size. For example, the Eucalypt leaf MALDI-Fourier Transform Ion Cyclotron Resonance-Imaging Mass Spectrometer (FTICR-IMS) data set, which is explained later in the sub section 2.1.1, includes 6545 pixels (spectra) each of which has ~720 peaks resulting in a data set that is ~38 GB. Therefore, it is difficult and time-consuming to manually extract meaningful information from these datasets. Currently, some visualization software packages like BioMap, FlexImaging, ImageQuest, Datacube Explorer, HDI (high-definition MALDI MS imaging), Metabolite Imager, MSiReader, SCiLS Lab and TissueView have been mainly used on continuous type data sets, where the set of *m/z* values is the same for all spectra (Gustafsson *et al.*, 2011). However, some of them do not support visualization of processed type datasets, where the set of *m/z* values is different from one spectrum to another (Norris *et al.*, 2007). Not only these software packages, but also there are many other published supervised and unsupervised methods which have not been implemented as software tools.

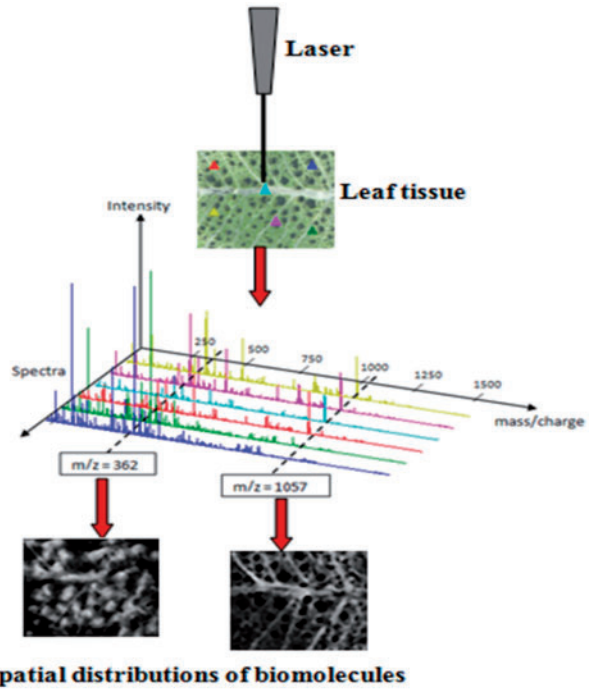

Fig. 1. MALDI-IMS data acquisition workflow. Mass spectra are measured at discrete spatial coordinates of the tissue section. An intensity image is obtained for each *m/z* value representing the spatial distribution of the corresponding biomolecule

Mainly two unsupervised learning approaches, namely Component Analysis and Spatial Segmentation, have been utilized for untargeted analysis of MALDI-IMS data (Alexandrov, 2012). Several component analysis methods such as Principal Component Analysis (PCA), non-negative matrix factorization, maximum autocorrelation factorization and probabilistic latent semantic analysis have been used to uncover the variation present in MALDI-IMS data (Jones *et al.*, 2011). Among these methods, PCA is the most widely applied algorithm (Bonnel *et al.*, 2011; Klerk *et al.*, 2007) which represents spatial patterns of molecules existing in the imaging dataset in terms of a set of score images. Even though the mass spectra do not contain negative intensity values, PCA score images can contain negative values which make the interpretation a challenge (Alexandrov, 2012). A further difficulty is presented when attempting to determine co-localized ion images for each identified pattern.

The most commonly used unsupervised data mining approach in MALDI-IMS is spatial segmentation (Alexandrov and Kobarg, 2011; Bruand *et al.*, 2011; Trede *et al.*, 2011). In this method, the imaging dataset is represented as a segmentation map, created by grouping spectra based on their similarity, highlighting chemically equivalent regions of the tissue. Since the segmentation map is created solely based on mass spectral intensities, without considering their spatial relationships, it does not capture the underlying spatial structure. Alexandrov *et al.* (2010) carried out de-noising of the individual intensity images before clustering spectra in order to improve the segmentation map. However, not all of the spatial relationships between spectra can be extracted through de-noising alone. Hierarchical clustering algorithm has been extensively used for clustering mass spectra (Bonnel *et al.*, 2011; Deininger *et al.*, 2008). However, it performs poorly when dealing with large and noisy imaging datasets.

The main weakness of the above unsupervised learning approaches is that they mainly process the dataset in the spectral domain and do not make the most of the important spatial information available through this IMS technology. Recently, a spatial distribution based peak picking method and a clustering approach have been introduced and they have shown promising results (Alexandrov and Bartels, 2013; Alexandrov *et al.*, 2013). Therefore it is essential to develop a streamlined data analysis pipeline specifically for IMS data utilizing the important spatial information available through this IMS technology.

In this paper, we propose a new data analysis pipeline for revealing hidden significant molecular distribution patterns in MALDI-IMS data (Fig. 2). The proposed algorithm consists of five consecutive main steps: where the first step involves spectra preprocessing. Then, Sliding Window Normalization (SWN) is used to normalize the spectra and limit the influence of high intensity peaks.

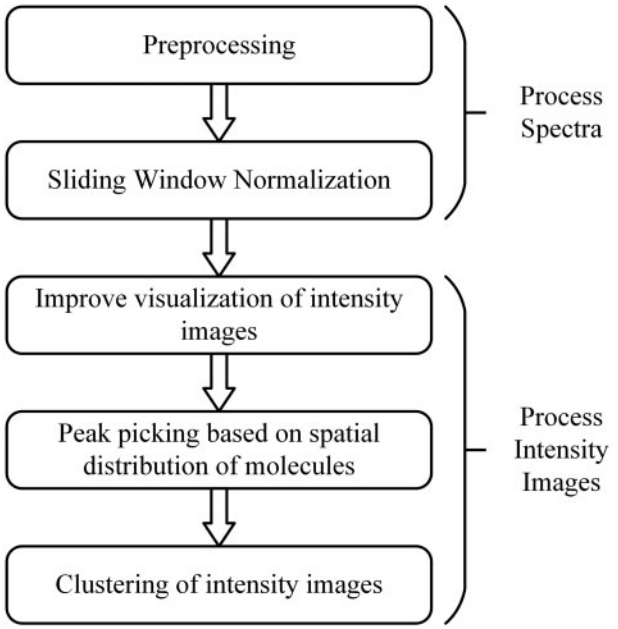

Fig. 2. The proposed MALDI-IMS data analysis pipeline

Third, image de-noising and contrast enhancement are used to improve the visualization of intensity images. Fourth, peak picking is performed by processing the individual intensity images. We introduce a new peak picking method based on an improved version of Gray level Co-Occurrence (GCO) matrix. It should be noted that, in this context, 'peak picking' denotes the process of identifying molecules with structured spatial distributions; which is different to the standard process of finding peaks in a mass spectrum. Finally, intensity images are clustered using the fuzzy c-means clustering algorithm. Prior to clustering, gist descriptors and the improved version of GCO matrices are used to extract features from individual intensity images, and the number of clusters is automatically estimated using minimum medoid distance.

The primary difference between the proposed workflow (Fig. 2) and the existing spatial segmentation based methods is that after spectra normalization the dataset is considered as a collection of intensity images and the subsequent steps are applied on individual images in order to incorporate the significant spatial information such as pixel positions and the details of the neighbourhood pixels. We used a MALDI-IMS metabolomics dataset of a Eucalypt leaf in order to assess the complete pipeline and a publicly available proteomics dataset of a rat brain to assess our peak picking method.

2 Methods

2.1 Datasets

2.1.1 MALDI-IMS metabolomics data of eucalypt leaf

A leaf from a juvenile Eucalypt tree (*Eucalyptus cladocalyx*) was flash frozen in liquid nitrogen then sectioned on a cryostat at 35 µm thickness, thaw mounted on a glass slide then dried in a desiccator. A thin layer of MALDI matrix was applied using a custom built sublimation apparatus, 2,5-Dihydroxy benzoic acid (DHB, Sigma-Aldrich) was applied by sublimation to an average density of $0.25 \pm 0.05\,mg/cm^2$.

MALDI-IMS experiments were performed in the positive ion mode using a Bruker SolariX 7 Tesla Hybrid-ESI/MALDI-FTICR-IMS (Bremen, Germany) operated with Compass solariXcontrol (Ver. 1.5.0, Build 103). Spectra were collected across the mass range *m/z* 200–1500, with the laser set to minimum laser spot size (approximately 20 µm in diameter), 500 shots were fired at rate of 1 kHz using 15% laser power in a 50×50 µm laser spot array (6545 total px). Compass flexImaging (Ver. 4.0, Build 4.0.32.0) was used to visualize initial imaging data. Raw data was extracted and converted into mzXML line spectra format using Bruker Daltonics' CompassXport tool for the ease of manipulation. This is one of the common file formats supported by most of the currently available visualization tools.

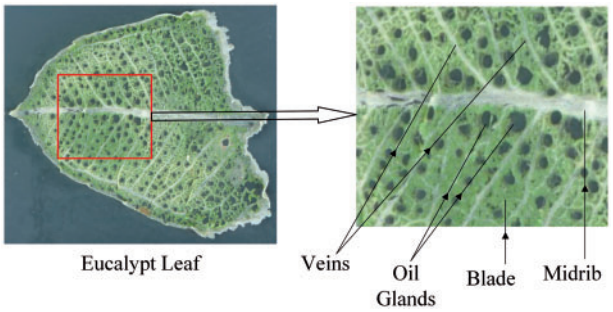

Fig. 3. Basic parts of the Eucalypt leaf namely midrib, veins, oil glands and blade

The dataset contains mass spectra for pixels within the highlighted window in Figure 3. The clearly visible anatomical structure of the Eucalypt leaf makes this an appropriate dataset to benchmark different methods. Figure 3 illustrates the basic parts of this leaf namely midrib, veins, oil glands and blade. These different regions have different molecular compositions.

2.1.2 MALDI-IMS proteomics data of a rat brain

Alexandrov and Bartels (2013) used 250 intensity images taken from a MALDI imaging proteomics dataset of a rat brain (Alexandrov *et al.*, 2010) to evaluate their peak picking algorithm. It consists of 50 unstructured and 200 structured images which can be further divided into four groups based on their spatial patterns namely 'Regions' (contain large separated regions of high intensity pixels), 'Curves' (contain curve like regions with high intensity pixels), 'Gradients' (contain large regions of high intensity pixels with outwardly decreasing intensity gradients around them) and 'Islets' (contain small regions of high intensity pixels). We used these test sets to compare our peak picking method with existing methods (Fig. 4).

2.2 Pipeline for discovering hidden profiles in MALDI-IMS data

The algorithm outlined in Figure 2 consists of five main steps namely data preprocessing, spectra normalization, improving visualization of intensity images, peak picking and clustering of biomolecules based on their spatial structure. These steps are explained in detail below:

2.2.1 Data preprocessing

Binning is used to reduce the dimensionality of the mzXML dataset. This is a frequently used method in mass spectral data analysis (Clerens *et al.*, 2006; Norris *et al.*, 2007). Although, the MALDI-FTICR-IMS data was collected at high mass resolution (>100 000 at 400 *m/z*), within this study we chose bin widths of 0.1 *m/z* units to reduce computational overhead. Then, all *m/z* values and the

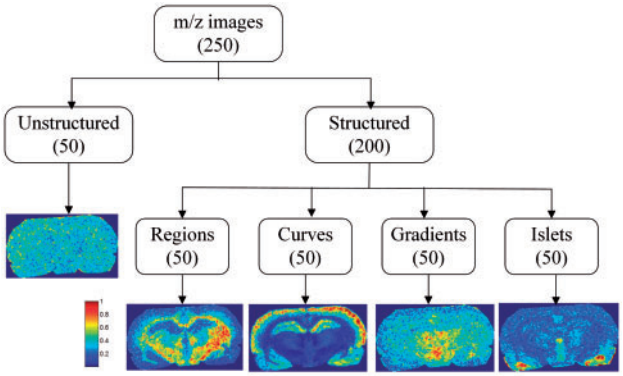

Fig. 4. Test sets of intensity images taken from the rat brain proteomics dataset (Alexandrov and Bartels, 2013). It consists of 50 unstructured and 200 structured images which are further divided into four classes namely 'Regions', 'Curves', 'Gradients' and 'Islets' based on their spatial patterns

corresponding intensity values within the bin are represented by the average *m/z* value and the maximum intensity. We assume that the peaks do not slide from a bin to the other during the acquisition causing errors in peak alignment. This is a reasonable assumption as the mass resolution of the data collected by MALDI-FTICR-IMS is significantly higher than the selected bin width. However, the *m/z* values corresponding to ions can be identified more accurately by reducing the bin size.

2.2.2 Spectra normalization

Spectra Normalization is used to facilitate direct comparison of peak intensities between different spectra. In order to overcome the limitations of the existing MALDI-IMS spectra normalization techniques, we propose Sliding Window Normalization (SWN). In SWN, the normalization factor of each peak is determined only by a set of peaks within a certain range of the spectrum, i.e. the peaks within the window defined on the *m/z* axis (Fig. 5). Each peak intensity is divided by the median intensity value of the peaks within the window. Then, the window can be slid along the spectrum to normalize all the peaks. The width of the window is set to 50 *m/z* empirically.

2.2.3 Improving visualization of intensity images

We used median filtering with a 3×3 neighbourhood in order to 'clean' the spectra from technology driven artefacts leaving only the relevant information. Also, histogram equalization was used to eliminate hotspots and to improve visualization of intensity images (Van de Plas *et al.*, 2007). It adjusts the contrast level by spreading out most frequent intensity values.

2.2.4 Peak picking

Peak picking is used to reduce the large amount of data contained in the MALDI-IMS datasets by selecting *m/z* values only for specific peaks. Biologists are interested in selecting molecules that are differentially distributed (show different abundances) in diverse regions thus creating structured molecular images. Therefore, we propose a new peak picking method based on spatial structure captured by Gray level Co-Occurrence (GCO) matrix (Gadelmawla *et al.*, 2004) to extract those structured molecular images.

GCO matrix is a second order histogram that is computed from the intensities of consecutive pairs of pixels in the image. First, the intensity values in the image should be quantized, reducing the number of intensity levels to 8. Here, we suggest quantising the intensity

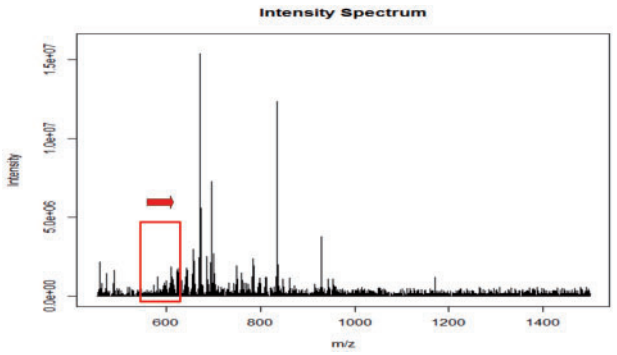

Fig. 5. Sliding window normalization

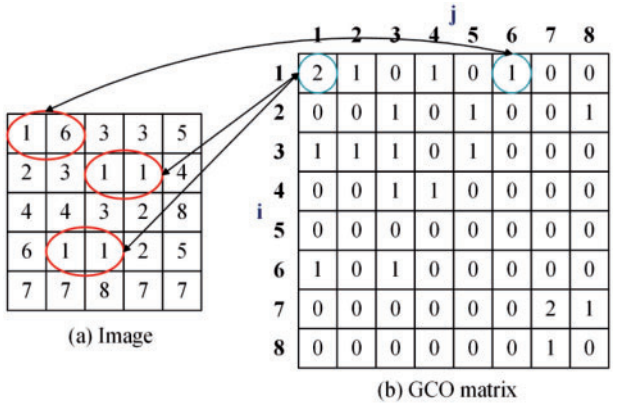

Fig. 6. (**a**) The quantized square image with intensity values {1,2…8} and (**b**) The corresponding horizontal GCO matrix. In the GCO matrix, the entry at the coordinate (i,j) is the number of pixel pairs having gray levels i and j consecutively in the horizontal direction of the quantized square image

values of the image based on its intensity histogram. We split the intensity range into 8 intervals in such a way that the area under the image histogram is same for all the intervals. An example of a quantized square image with intensity values {1, 2…8} is shown in Figure 6a. In the GCO matrix shown in Figure 6b, the entry at each coordinate (i,j) is defined as the number of pixel pairs having gray levels i and j consecutively in the horizontal direction of the quantized square image. For example, the existence of two pairs of consecutive pixels with intensity values {1,1} (meaning two co-occurrences of {1,1}) in the quantized image would mean that, the GCO matrix has the entry '2' at the coordinate (1,1). In this work, for each image, gray level co-occurrences in four directions, 0°, 45°, 90° and 135° are calculated.

However, not all the values in the GCO matrix are important for measuring the level of structure. The main idea is that, the structured images should have a pattern containing a large number of pairs of pixels with co-occurring low intensity values and co-occurring high intensity values, because the contrast between those two sets of intensity values exhibits a clear structure. Therefore, only the numbers of those pairs of pixels are of interest. We propose to use two weight vectors created logically based on the above idea, along with the GCO matrices for identifying intensity images exhibiting structured distributions. The proposed weight vectors assign higher weights to the aforementioned important values in the GCO matrices (see highlighted values in Fig. 7b) and disregard the other values.

These weights are assigned accordingly to capture two measures M1 and M2 representing regions with low intensities and high intensities respectively.

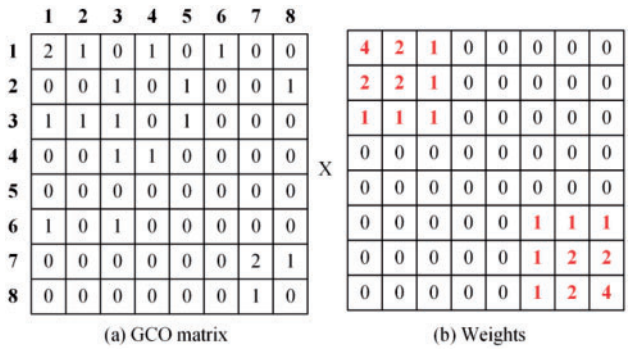

Fig. 7. (**a**) A GCO matrix and (**b**) its corresponding weight vector

M1 and M2 are calculated from each GCO matrix by multiplying its values with its corresponding weights (see Fig. 7) as follows:

$$\begin{aligned} M1 \;=\; & 4{*}gco(1,1) + 2{*}\{gco(1,2)+gco(2,1)+gco(2,2)\} \\ & +\{gco(1,3)+gco(2,3)+gco(3,1)+gco(3,2)+gco(3,3)\} \end{aligned} \tag{1}$$

$$\begin{aligned} M2 \;=\; & 4{*}gco(8,8)+2{*}\{gco(8,7)+gco(7,8)+gco(7,7)\} \\ & +\;\{gco(8,6)+gco(7,6)+gco(6,8)+gco(6,7)+gco(6,6)\} \end{aligned} \tag{2}$$

where, gco(i,j) is the entry at the coordinate (i,j) in the GCO matrix.

These two measures should be high for structured images and low for unstructured images. Therefore, all the intensity images are ranked based on these values and structured images are selected. Unlike the other methods, this technique does not completely rely on intensity values and selects molecules based on their spatial distribution.

2.2.5 Clustering

Two different methods based on gist descriptors and GCO matrices are used to extract features from intensity images prior to clustering. Human capability to recognize a scene or gather the necessary details to distinguish it from another scene is known as the gist of a scene. In our first method, the following steps are used to extract gist features related to orientation from intensity images (Supplementary Fig. S1).

First, each image is split into nine regions and then each region is filtered with four Gabor filters representing four angles (see Supplementary Fig. S1). Gabor filtering emphasizes the visual details in the considered direction and it is believed that the orientation representations of Gabor filters are equivalent to human vision (Siagian and Itti, 2007). After applying Gabor filters, averaging operation is used to find a representative value for each region.

After extracting features through gist descriptors, the improved version of the GCO matrix based method is used to further extract textural features from the intensity images. In this work, for each image, gray level co-occurrences in four directions, 0°, 45°, 90° and 135° are calculated. Moreover, in order to further improve the accuracy of the results, at the beginning, each image is divided in to 9 regions in such a way that there are overlaps between the regions and GCO matrices are calculated for each region separately.

All the features extracted through these two methods (outcomes of Gabor filtering and values in each GCO matrix) are scaled to be in the range 0 to 1 and then, Principal Component Analysis (PCA) is used to reduce the dimensionality of the feature space. In this study, fuzzy c-means algorithm is used for clustering biomolecules instead of a hard clustering method. Only the data objects with membership values over 0.5 are selected for each cluster. However, when using the fuzzy c-means clustering algorithm, we need to specify the number of clusters *a priori*. We propose a measure calculated using medoid distances to estimate the number of clusters. Medoid is an element of a cluster whose average difference to all the other members of the same cluster is minimal. We expect clustering algorithms to maximize the distance between clusters. Since medoids are representative objects of clusters, a better clustering result can be obtained by minimizing the distances between cluster medoids. If the distance between the medoids of the two nearest clusters is high, then it denotes a good clustering result. Therefore, we suggest using the number of clusters which maximize the minimum medoid distance as the optimal number of clusters.

The following equation is used to estimate the number of clusters where m_i, m_j are cluster medoids.

$$A = \left(\underset{i \neq j}{min} |m_i - m_j|^2 \right) \tag{3}$$

3 Results

The important steps in the proposed data analysis pipeline are discussed in detail along with the results under Sections 3.1–3.4. We use the data sets described in Section 2.1 to further describe the pipeline and to evaluate the performance.

3.1 Spectra normalization

The most commonly used normalization method in MALDI-IMS is the Total Ion Count (TIC) normalization, where all peak intensities in the spectrum are divided by its TIC value (sum of all peak intensities in the spectrum). However, in TIC normalization, the area under a spectrum is assumed to be similar for all spectra, which is not a reasonable assumption for processed type datasets. Hence, it cannot be used to normalize the Eucalypt leaf dataset. Deininger *et al.* (2011) has proposed to use median normalization in order to overcome this limitation. Therefore, we compare the performance of our SWN method with median normalization. Figure 8 shows an intensity image (i) before normalization, (ii) after median normalization and (iii) after sliding window normalization. The distribution of the metabolites within the considered mass bin (especially in the upper section of the leaf) is clearly more visible in the third image than the first two images, which encourages the consideration of SWN for normalizing MALDI-IMS data.

3.2 Improve visualization

Figure 9 shows an intensity image (i) before histogram equalization, (ii) after histogram equalization and the corresponding image histograms (a), (b). It can be clearly seen that the hot spot in the original image has lessen the appearance of the distribution of the molecule in other regions. Since, histogram equalization adjusts the contrast level by spreading out most frequent intensity values, it can be used to eliminate hotspots and improve visualization of intensity images.

3.3 Peak picking

The new peak picking method selects molecular images completely based on their level of spatial structure. We use the dataset described in Section 2.1.2 to compare the performance of the new peak picking method with the widely used mean spectrum based approach and the spatial distribution based approach recently introduced in Alexandrov and Bartels (2013). Each structured test set ('Regions',

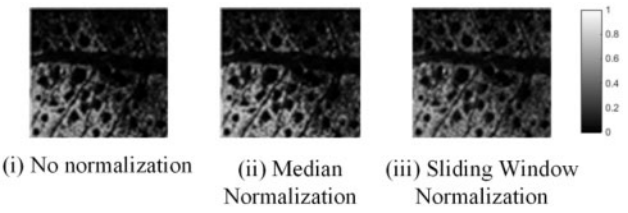

Fig. 8. Normalization results. An intensity image (i) before normalization, (ii) after median normalization and (iii) after sliding window normalization

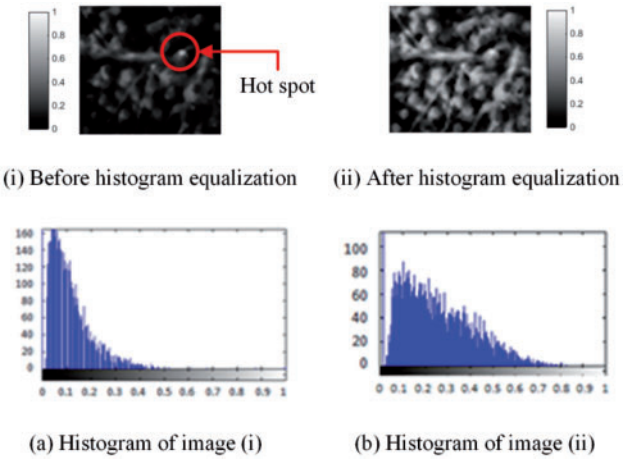

Fig. 9. Histogram equalization results. An intensity image before histogram equalization (i), after histogram equalization (ii) and the corresponding image histograms (**a**) and (**b**), respectively

'Curves', 'Gradients' and 'Islets') is mixed with the unstructured test set and the accuracy of extracting the structured images is calculated for each case. Figure 10 illustrates the accuracies of the three methods when selecting 'k' number of images from the mixed groups, where k varies from 1 to 50. Here, the accuracy is defined as the percentage of structured images out of the selected images. According to these figures, the new peak picking method performs significantly better than the widely used spectrum based approach in all four cases. Also, it outperforms the method introduced in Alexandrov and Bartels (2013) in all the instances except when selecting structured images with small regions of high intensity pixels ('Islets'). Our method extracts structured images with Regions, Curves and Gradients with 100% accuracy when $k \leq 49$, $k \leq 27$ and $k \leq 13$, respectively.

Moreover, the most structured 250 intensity images out of the 3150 images were extracted by applying this new spatial peak picking method to the Eucalypt leaf dataset.

3.4 Clustering

Both gist descriptors and GCO matrices together extracted 2340 features. Therefore, Principal Component Analysis (PCA) was used to reduce the dimensionality of the feature space. PCA reduced the number of features to 13 by preserving up to 98% of the variance. The number of clusters estimated by the proposed minimum medoid distance method is 4.

Figure 11A shows the significant metabolite distribution patterns discovered after using fuzzy c-means clustering algorithm along with SWN. Note that these are cluster average patterns. In Figure 11A, Pattern (A1) represents metabolites that are distributed in the midrib, veins and oil glands, but not in the blade. Similarly, Pattern (A2) represents metabolites distributed mainly in the midrib and veins. Both Pattern (A3) and Pattern (A4) represent metabolites distributed only in the blade, however metabolites in Pattern (A4) are only concentrated in the lower section of the image. The proposed MALDI-IMS data analysis pipeline discovered these hidden significant molecular distribution patterns which the existing methods failed to extract. These patterns are followed by 52, 50, 66 and 15 mass bins respectively. Supplementary Figure S2 shows a set of mass bins that follow Pattern (A1). Interestingly, we can see potential isotopes in this figure such as (610.3 & 611.3) which further validates the applicability of the proposed pipeline because in general, ions related to the same metabolite such as isotopes are expected to follow the same spatial pattern.

Moreover, in order to assess the effect of spectra normalization - in particular the proposed SWN method - on the final result, we compared the cluster average patterns generated by the data analysis pipeline utilizing SWN (Fig. 11A), median normalization (Fig. 11B) and no normalization (Fig. 11C). It can be clearly seen that the pipeline which used median normalization has failed to extract some hidden unique molecular distribution patterns (eg. Pattern A2). Further, the obtained patterns are not as distinctive as those obtained from the pipeline with SWN. This is due to the fact that, the standard normalization techniques, which use a single normalization factor calculated considering the entire spectrum, tend to suppress some low abundant molecules (Deininger *et al.*, 2011). Interestingly, the molecular distribution patterns extracted by the pipeline with no normalization are similar to those patterns extracted by the pipeline with SWN. However, in terms of visualization, a slight improvement can be seen in the patterns extracted by the latter.

Supplementary Figure S3 shows the patterns extracted from the rat brain proteomics dataset.

4 Discussion

4.1 Importance of the new algorithm

Mainly two unsupervised learning approaches namely component analysis and spatial segmentation have been used for untargeted analysis of MALDI-IMS data (Alexandrov, 2012). However, sometimes these techniques fail to uncover the general data structure of complex and high volume MALDI-IMS datasets as they do not make the most of available spatial information. We demonstrate the limitations by applying them to the MALDI-IMS Eucalypt leaf dataset described in Section 2.1.1 (see Supplementary Fig. S4, S5).

The PCA score images in Supplementary Figure S4 do not reveal the hidden differential molecular distribution patterns, which define the important anatomical regions of the biological tissue under study. Also, we cannot see a strong visual correlation between the score images and their corresponding highest loading intensity images, which makes the use of loadings for selecting intensity images co-localized with each pattern, as the case for the component analysis method questionable. Moreover, the segmentation maps of the MALDI-IMS Eucalypt leaf dataset, obtained by clustering the spectra using the widely used hierarchical clustering method, poorly capture the underlying spatial structure (see Supplementary Fig. S5).

The algorithm proposed in this paper, successfully extracts the hidden significant molecular distribution patterns existing in the dataset. The primary difference between the new workflow and the existing spatial segmentation method is that, in this new MALDI-IMS data analysis pipeline, after spectra normalization the dataset is considered as a collection of intensity images rather than a collection of spectra and the subsequent steps are applied on

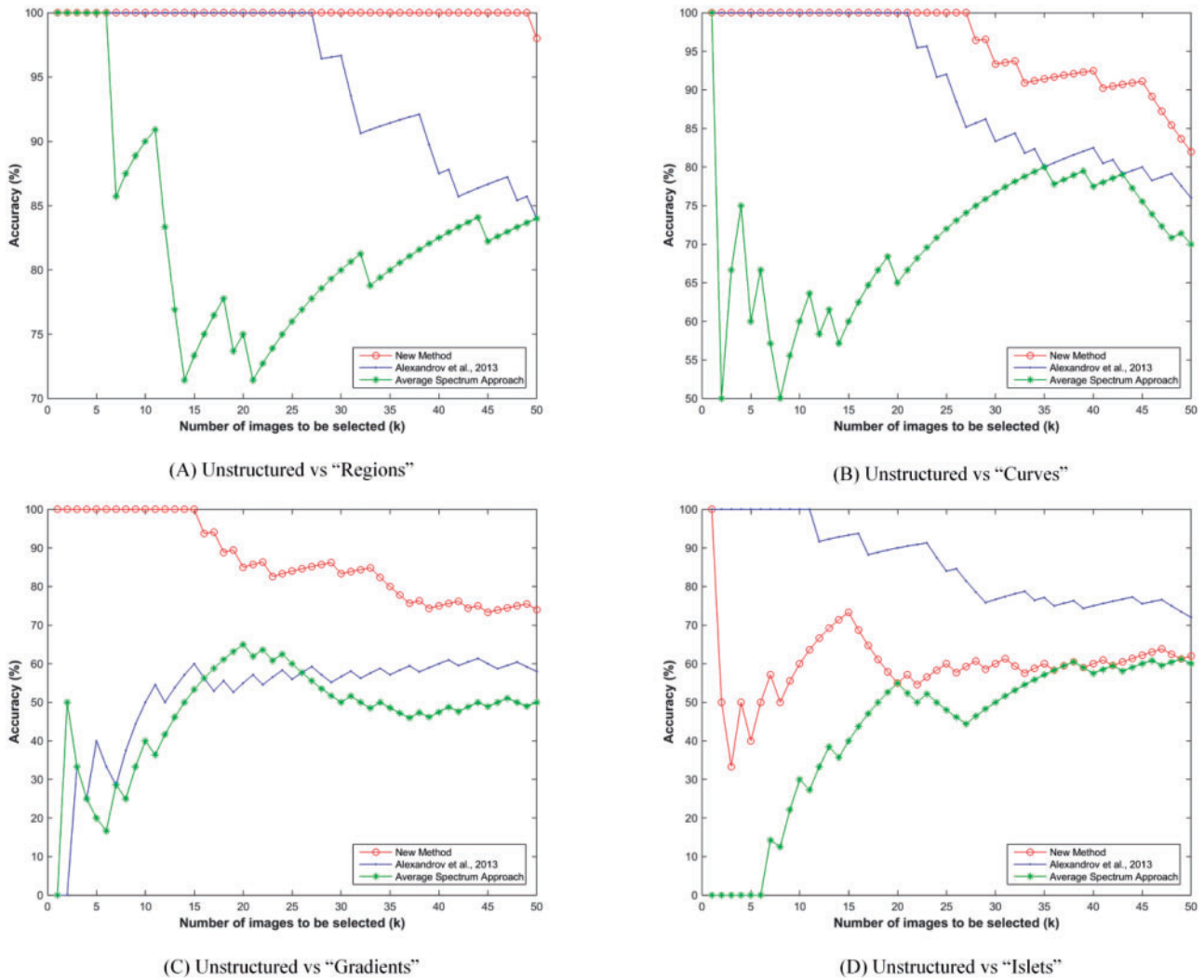

Fig. 10. Comparison of different peak picking methods. Each structured test set ('Regions', 'Curves', 'Gradients' and 'Islets') is mixed with the unstructured test set and the accuracy of extracting the structured images is calculated for each case (**A-D**). The accuracies of the three peak picking methods are plotted against k, which is the number of images to be selected. At k = 50, our method extracted structured images with Regions, Curves, Gradients and Islets with an accuracy of 98%, 84%, 72% and 62% respectively

individual images. Therefore, the new algorithm incorporates not only the peak intensities but also the spatial relationships of spectra when clustering the molecular images, which leads to the identification of hidden patterns.

The proposed algorithm would work well for data generated from low mass resolution imaging mass spectrometers, such as MALDI Time of Flight (TOF). In this study, data was generated from MALDI-FTICR-IMS across 200 and 1500 *m/z* with a very high mass resolution. Processing of these data requires substantial computer power. Therefore, a binning technique was used to reduce the computational overhead. As a result, the algorithm extracted mass bins showing significant spatial distributions, not individual *m/z* values. Several *m/z* values can be contributing to each mass bin (see Supplementary Fig. S6). We demonstrate that our algorithm also works on real *m/z* values by using small mass ranges (see Supplementary Fig. S7).

The required computational time depends on the image size, the number of peaks within the considered mass range as well as the selected bin width. When the bin width was set to 0.1 *m/z*, our algorithm took ~6 min to produce the output considering the entire mass range (200–1500 *m/z*), on a Windows 7 (64-bit) operating system running on a Core i7-2600 CPU at 3.40 GHz with 8.0 GB Random Access Memory. When 0.0001 *m/z* bins were considered, the run time of the proposed algorithm was ~7 min for the mass range 450–460 *m/z*.

4.2 Spectra normalization

The existing spectra normalization methods may produce misleading results due to the influence of high intensity peaks. The principal difference of our SWN method from the existing methods is that, it uses the local distribution when normalizing peaks rather than using the entire spectrum, thus reducing the influence of high intense peaks. Also, it can be used for normalizing both continuous type datasets (where the set of *m/z* values is the same for all spectra) as well as processed type datasets (where the set of *m/z* values is different from one spectrum to another).

However, when using the SWN method, it is required to set the width of the window, which should be less than the length of the spectrum. In this study, it was set to 50 *m/z* by attempting several values and visually inspecting the resultant ion images to assess the

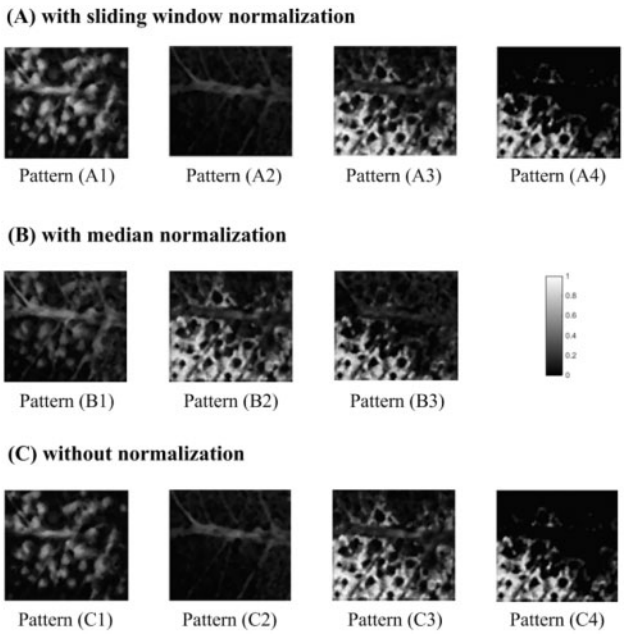

Fig. 11. Cluster average patterns of the Eucalypt leaf dataset (**A**) with sliding window normalization, (**B**) with median normalization and (**C**) without normalization. Pattern (A1) represents metabolites that are distributed in the midrib, veins and oil glands, but not in the blade. Pattern (A2) represents metabolites distributed mainly in the midrib and veins. Both Pattern (A3) and Pattern (A4) represent metabolites distributed only in the blade, however metabolites in Pattern (A4) are only concentrated in the lower section of the leaf

suitability of the width. Even though we believe that this value is suitable for most of the spectral datasets as generally the length of a spectrum is higher than 50 *m/z*, the normalization results can be further improved by optimizing this value. More work needs to be done using simulated or real datasets with some prior knowledge in this regard.

Although the importance of normalizing the low resolution IMS data like MALDI-TOF-IMS data has been validated (Deininger *et al.*, 2011), the fact that there is no significant difference among the patterns extracted from the proposed pipeline with SWN and without normalization, suggest to further assess the need of normalization for MALDI-FTICR-IMS data. We suggest that a more significant improvement in results can be obtained with SWN normalization when the anatomical structure of the biological sample under study and the corresponding molecular distributions are more complex as the SWN method tend to make the images clearer. However, more work needs to be done using multiple datasets to support this suggestion.

Also, it should be noted that even though this method shows advantages over the other normalization techniques in the context of IMS data processing, it is not suitable to be used for normalizing mass spectrometry data where the relative abundances of ions need to be preserved.

4.3 Improve visualization

Visualization of molecular distributions can be significantly improved through de-noising. However, it is important to select a de-noising method that does not eradicate edges and small important details. We use median filtering as the de-noising method as it is capable of removing noise while preserving boundaries and small details, which is more important for this application. Moreover, in MALDI-IMS datasets, some spectra may show artificial patterns that lead to hotspots in intensity images. These hotspots lessen the appearance of the distribution of the molecule in other regions and affect the subsequent steps (Watrous *et al.*, 2011). Histogram equalization eliminates these hotspots thereby improving visualization of intensity images.

4.4 Peak picking

The proposed new spatial distribution based peak picking method is the most significant contribution of this paper. Peak picking has always been a challenge when processing high volume imaging datasets as the existing spectra-based methods suffer from low sensitivity (Alexandrov and Bartels, 2013). Mostly peak picking is applied to the dataset mean spectrum, and *m/z* bin values corresponding to high intense peaks are selected (Deininger *et al.*, 2008). However, we cannot always guarantee that high intense peaks correspond to metabolites with significant molecular distributions. Also, this method may not be able to select high intense peaks that appear in a small region of the tissue (Watrous *et al.*, 2011). These spectra based approaches have significant limitations as they only consider peak intensities and do not pay attention to the important spatial relationships of spectra. To the best of our knowledge, only one paper has proposed a peak picking method by incorporating spatial relationships of spectra (Alexandrov and Bartels, 2013).

In this study, we use GCO matrices to extract structured intensity images. In all existing studies on GCO matrix method, intensity range of the image is divided into 8 equal intervals (Gadelmawla *et al.*, 2004). However, the intensity distributions of molecular images are not always uniform. For this reason, sometimes GCO matrices fail to represent the significant changes in pixel intensities accurately. Therefore, we quantized the intensity values of the image based on its intensity histogram. Moreover, the GCO matrices along with the weight vectors proposed in this paper can be used to identify any type of image which shows a structured distribution.

By using the MALDI-IMS dataset of a rat brain, we show that our new unsupervised peak picking method outperforms both the spectra based method and the method introduced by Alexandrov and Bartels (2013). The new method achieved an accuracy of 98% when selecting intensity images with large separated regions of high intensity pixels. For the 'gradients' also our method achieved a significant accuracy, unlike the method in Alexandrov and Bartels (2013), which performed weakly as edge detection cannot detect clear edges when the separation between the regions is not clear. Although, the method in Alexandrov and Bartels (2013) outperforms our method when selecting intensity images with small regions of high intensity pixels ('islets'), usually the occurrence of structured images with 'islets' is less when compared to that of 'regions', 'gradients' and 'curves'. It should also be noted that the method in Alexandrov and Bartels (2013) requires a parameter to be set and they have shown an improvement in the results of their method when the best possible parameter value is used. However, as they have mentioned, selecting the best possible parameter value is not feasible in practice. Therefore, those results were not used in our comparison study. In addition, the proposed new peak picking algorithm also complements the detection of important molecules in other imaging datasets.

4.5 Clustering

Prior to clustering, each intensity image was divided into nine regions and features were extracted from each region separately. This step was utilized in order to capture the dissimilarity of images showing the same distribution in different regions thus avoiding them being clustered into the same group.

Real datasets usually contain significant amounts of background noise and the clusters might have different sizes. As a result, hard clustering methods like kmeans and hierarchical clustering mostly fail to extract all hidden significant patterns that exist in the dataset and generate false clustering results. In fuzzy clustering, data objects are not directly assigned to clusters; instead each object is assigned a membership value for each cluster. This method has proven to be more efficient when dealing with noisy data (Schwämmle and Jensen, 2010) because usually data objects related to background noise have distributed membership values. With this method, the influence of such data objects to the final result can be reduced as they have low influence in estimating cluster centroids.

Moreover, minimum medoid distance is a simple and efficient method that can be used for automatically estimating the number of clusters, and to the best of our knowledge it has not been used before.

Acknowledgements

The authors thank the University of Melbourne-Vanderbilt University Partnership Grant Scheme and also the School of Botany node of Metabolomics Australia (MA) at The University of Melbourne, a member of Bioplatforms Australia Pty Ltd which is funded through the National Collaborative Research Infrastructure Strategy (NCRIS), 5.1 Biomolecular Platforms and Informatics and co-investments from the Victorian Government.

Funding

This work is partially funded by Australian Research Council (Grant DP1096296).

Conflict of Interest: none declared.

References

Alexandrov,T. *et al.* (2010) Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J. Proteome Res.*, **9**, 6535–6546.

Alexandrov,T. and Kobarg,J.H. (2011) Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, **27**, i230–i238.

Alexandrov,T. (2012) MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, **13**, S11.

Alexandrov,T. and Bartels,A. (2013) Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics*, **29**, 2335–2342.

Alexandrov, T. *et al.* (2013) Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Anal. Chem.*, **85**, 11189–11195.

Bonnel,D. *et al.* (2011) Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: application to prostate cancer. *Anal. Bioanal. Chem.*, **401**, 149–165.

Bruand,J. *et al.* (2011) AMASS: algorithm for MSI analysis by semi-supervised segmentation. *J. Proteome Res.*, **10**, 4734–4743.

Caprioli,R.M. *et al.* (1997) Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem.*, **69**, 4751–4760.

Clerens,S. *et al.* (2006) Createtarget and analyze this!: new software assisting imaging mass spectrometry on Bruker Reflex IV and Ultraflex II instruments. *Rapid Commun. Mass Spectrometry*, **20**, 3061–3066.

Deininger,Sr-O. *et al.* (2008) MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J. Proteome Res.*, **7**, 5230–5236.

Deininger,S-O. *et al.* (2011) Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Anal. Bioanal. Chem.*, **401**, 167–181.

Gadelmawla,E. (2004) A vision system for surface roughness characterization using the gray level co-occurrence matrix. *NDT & E International*, **37**, 577–588.

Groseclose,M.R. *et al.* (2008) High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics*, **8**, 3715–3724.

Gustafsson,J.O. *et al.* (2011) MALDI imaging mass spectrometry (MALDI-IMS)-application of spatial proteomics for ovarian cancer classification and diagnosis. *Int. J. Mol. Sci.*, **12**, 773–794.

Jardin-Mathé,O. *et al.* (2008) MITICS (MALDI Imaging Team Imaging Computing System): a new open source mass spectrometry imaging software. *J. Proteomics*, **71**, 332–345.

Jones,E.A. *et al.* (2011) Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PLoS One*, **6**, e24913.

Klerk,L.A. *et al.* (2007) Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets. *Int. J. Mass Spectrometry*, **260**, 222–236.

Lagarrigue,M.l. *et al.* (2012) New analysis workflow for MALDI imaging mass spectrometry: application to the discovery and identification of potential markers of childhood absence epilepsy. *J. Proteome Res.*, **11**, 5453–5463.

Norris,J.L. *et al.* (2007) Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *Int. J. Mass Spectrometry*, **260**, 212–221.

Schwämmle,V. and Jensen,O.N. (2010) A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. *Bioinformatics*, **26**, 2841–2848.

Siagian, C. and Itti, L. (2007) Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29**, 300–312.

Sugiura,Y. and Setou,M. (2010) Imaging mass spectrometry for visualization of drug and endogenous metabolite distribution: toward in situ pharmacometabolomes. *J. Neuroimmune Pharmacol.*, **5**, 31–43.

Trede,D. *et al.* (2011) On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. *J. Integrative Bioinformatics*, **9**, 189–189.

Van de Plas,R. *et al.* (2007) Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA. In: *Life Science Systems and Applications Workshop, LISA 2007, IEEE/NIH*, pp. 209–212.

Watrous,J.D. *et al.* (2011) The evolving field of imaging mass spectrometry and its impact on future biological research. *J. Mass Spectrometry*, **46**, 209–222.