# Multivariate Analysis of Imaging Mass Spectrometry Data

E. R. Muir[1,*], I. J. Ndiour[2,*], N. A. Le Goasduff[2,*], R. A. Moffitt[1], Y. Liu[3],
M.C. Sullards[3,4], A.H. Merrill, JR.[3], Y. Chen[4,#], and M. D. Wang[1,2,#]

*[1]Department of Biomedical Engineering, [2]Department of Electrical and Computer Engineering, [3]School of Biology, [4]Department of Chemistry & Biochemistry. Georgia Institute of Technology[1,2,3,4] and Emory University[1], Atlanta, GA. [*]denotes equal contribution. [#]To whom correspondence should be addressed.*

## Abstract

*Imaging mass spectrometry can be used to reveal spatial distributions of multiple molecular species in a 2D biological sample. Because of the large amount of data produced by this technology, it is difficult and time-consuming to manually extract meaningful results from imaging mass spectrometry experimentation. We have developed and implemented an original approach to easily and consistently process mass spectrometry imaging data with the goal of automatically identifying interesting regions of molecule expression. Based on multivariate analysis techniques such as principal component analysis, the system allows researchers to conveniently define and visualize spatial regions based on spectral similarity. Features of our system are demonstrated on mouse cerebellum data.*

## 1. Introduction

Recent improvements in imaging mass spectrometry (IMS) are due in a large part to the introduction of matrix-assisted laser desorption ionization (MALDI) [1]. MALDI mass spectrometry analysis can be applied to biological tissues to produce ion density maps, or *molecular images*, of the sample surface. In this way, MALDI-IMS) reveals the expression levels and spatial distributions of known and unknown molecules in complex tissues, permitting sensitive, rapid and molecularly specific analyses of biomolecules in 2D tissue sections. MALDI-IMS has been recognized as a promising tool for numerous applications including biomarker discovery, drug bio-distribution monitoring, and molecular mechanism investigation, as well as many other types of peptide, lipid, or metabolite analysis [2].

Computation of IMS data allows the construction of ion density maps for each signal detected by the MALDI device. Ultimately, thousands of molecular images, with thousands of pixels, may be obtained for each specific mass to charge ratio (*m/z*).

One of the most popular software used to study IMS data is BioMap. It allows the user to visualize the data as thousands of ion images. However it does not provide any tools to facilitate data processing. The user must manually select the ion images and analyze them by eye to find the molecular distribution of specific ions. It usually takes much time for the user to analyze the data to obtain meaningful data.

Due to the high dimension of data created by mass spectrometry imaging, it is extremely difficult to manually distinguish between regions of interest in the image. Typically the data is analyzed by serial viewing of molecular images at individual *m/z* values using systems such as BioMap. Although there may be some spatial information at a single *m/z* value, it is nearly impossible to obtain an overview of the spatial distribution of every detected molecule from thousands of individual ion images. Therefore, an unsupervised data processing method for quickly observing masses of interest or spatial patterns in large collection of IMS data is urgently needed.

The contribution of this paper is to provide methods to visualize IMS data and distinguish spatial regions and tissue types. Multivariate analysis is ideal for this task, and can be used to analyze the correlations within the entire imaging data set in order to provide valuable guides for the unveiling and understanding of related biochemical process. These methods are based on multivariate techniques such as principal components analysis (PCA), linear discriminant analysis (LDA), multivariate analysis of variance (MANOVA), and clustering.

This paper is organized as follows. In section 2, we present the multivariate techniques. In section 3, we demonstrate the analysis of sulfatide in mouse brain tissue using multivariate techniques described in section 2. The results obtained by applying the methods proposed are presented in section 4. In section 5, we discuss the results, and future work.

## 2. Methodology

This section is devoted to the rationale and methodology behind the system.

### 2.1. Principal component analysis (PCA)

Principal component analysis is a popular unsupervised multivariate analysis technique which may be used to perform data reduction on high dimensional data sets such as IMS. PCA has been previously used in IMS to help find patterns and to differentiate tissues [3-6]. Here, we use PCA in a dual role: 1) to identify MS peaks of interest; and 2) as a basis for further region clustering and classification analysis. The results obtained by applying PCA on experimental data are described in section 4.1. PCA, also called discrete Karhunen-Loeve transform (KLT) or the Hotelling transform, is a linear transformation that defines a new orthogonal basis in which the greatest variance of the data lies in the direction of the first (principal component) basis vector, the second greatest variance along the second basis vector, and so on. PCA relies implicitly on two assumptions: 1) that large variances have important dynamics; and 2) that data has a high SNR (signal-to-noise ratio). Fortunately, high SNR is a well-studied property of MS, and our results suggest that the first assumption is safe as well. Figure 1 illustrates the principle of principal component analysis in 2D.

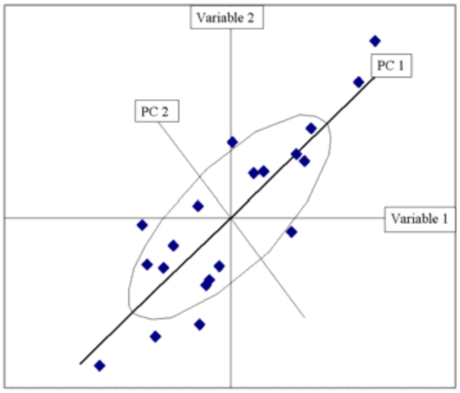


**Figure 1.** Principal component analysis.

To perform PCA on multi-spectral images, such as those produced by MALDI-IMS, the data must first be ordered into a 2D array. In this case, the original data are a 3D array of size $M \times X \times Y$, where $X$ and $Y$ are the familiar spatial dimensions and M is the number of *m/z* values detected by the MS machine. This transformation is done by a simple reshaping of the $M \times X \times Y$ data into an $M \times N$ matrix, where $N = X \cdot Y$. The purpose of this transformation is to achieve a form with $N$ observations (pixels) of $M$ variables (*m/z* peaks) for determining signatures (sets of molecules) which vary together strongly.

For any data set of the form $N$ observations of $M$ variables, PCA can be completed using the classical covariance method:

- Organize the data set in a matrix $X$ of size $M \times N$ such that each column $X_i$ represents an observation.
- Compute the mean-subtracted data $B$ by subtracting the mean observation $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ from each column.
- Estimate the covariance matrix
$$C = E[B \cdot B^T] = \frac{1}{N} B \cdot B^T$$
- Find the eigenvectors and eigenvalues of the covariance matrix. Since $C$ is a symmetric matrix, the eigenvectors are orthogonal and $C$ can be diagonalized using a matrix $V$ of eigenvectors: $V^{-1} \cdot C \cdot V = \Delta$. The diagonal matrix $\Delta$ consists of eigenvalues of $C$ arranged in descending order.
- Compute the cumulative energy content for each eigenvector:
$$e[m] = \sum_{k=1}^{m} \Delta[k,k]$$
- Select a subset of the eigenvectors as basis vectors. The first L columns of $V$ are to form an $M \times L$ matrix denoted $W$. L can be determined using the cumulative energy $e$. This is typically done by choosing the smallest L such that the cumulative energy $e$ is above a desired threshold.
- Compute the scores: $Z = \frac{B}{S}$ (element wise division) using the matrix S of size $M \times N$ constructed such that all $N$ columns are the same column, formed by squared roots of eigenvalues of $C$.
- Project the original data set into the new coordinate system to obtain the reduced data $Y$: $Y = W^T \cdot Z$

  The new data set is of size $L \times N$.

After PCA is applied on a data set, the reduced data obtained is of size $L \times N$ and can be reshaped into a multi-spectral image of size $L \times X \times Y$. This operation has two benefits: 1) the basis vectors themselves can be used as a method of region and peak identification; and 2) the size and complexity of the data is reduced for future classification tasks such as clustering.

## 2.2. Clustering

Multivariate and clustering analysis are widely used in many branches of engineering and science, and have previously been used in mass spectrometry to automatically group and label pixels in the image with similar spectra [3]. This can be useful when very little prior information is known about the spatial characteristics of the mass spectrometry data.

Here, a K-means clustering algorithm is applied after modest data reduction by a two step procedure of averaged downsampling along the *m/z* axis followed by PCA. Only the most significant 50-200 PCs are usually needed to account for 90% of the energy in the IMS data, a significant decrease from many thousand *m/z* values present in the raw data.

The K-means algorithm is an Expectation-Maximization algorithm which iteratively alternates between 1) assigning pixels to the closest available cluster center and 2) recalculating cluster centers as the mean spectrum of all the pixels in the cluster. Pixels are reassigned, and centroids recalculated until the algorithm converges. Initial cluster centers are chosen by selecting random spectra from the image. To measure the distance between pixels and cluster centers, the cosine distance is used. The cosine distance measure is calculated as one minus the cosine of the angle between vectors describing the spectra. Additionally, when using the cosine distance, the new cluster centroid is the mean spectra of the pixels in that cluster after they have been normalized to unit Euclidean length. The final output is a pixel-level classification which is visualized as an image where pixels are colored by cluster membership. This unsupervised classification can then be further refined with a supervised technique such as LDA.

## 2.3. Linear discriminant analysis

Linear discriminant analysis is a supervised multivariate classification technique that is used to classify the observations of a data set into groups using regression equations which maximize between-group variance and minimize within-group variance. For IMS, LDA is used to maximally differentiate pixels of the image based on their spectra, and should provide more controlled classification than clustering [3,6].

To use LDA the number of pixels in the groups being analyzed should be larger than the amount of data in spectra, which is usually not the case with IMS data. To avoid this problem the data is reduced using the previously discussed methods of downsampling and PCA.

To use LDA with mass spectrometry imaging, multiple training pixels must be manually selected from the image so that the pixels within a group have similar spectra. The clustering procedure described earlier is used for this purpose; any combination of clusters can be grouped together to define the training groups for LDA. Once the groups have been defined, LDA classifies each pixel in the image into one of the groups, and a resulting image is output where each group is represented as a different color.

## 2.4. Multivariate analysis of variance

Multivariate Analysis of Variance (MANOVA) tests differences between the mean of groups of multivariate data. To the best of our knowledge, MANOVA has not been used to analyze mass spectrometry data. Although it is not used as commonly as other multivariate techniques, MANOVA has been occasionally used in analytical chemistry [7]. One step in MANOVA is performing canonical correlation. This is similar to PCA in that the data are projected onto sets of orthogonal axes. In canonical correlation however, the first axis provides the largest separation between groups, the second axis has the second largest separation, and so on.

The first step in applying MANOVA is to select groups. Groups for MANOVA are selected in the same manner as the groups used for training in LDA. Also like LDA, to use MANOVA the number of pixels in the groups being analyzed should be larger than the amount of data in spectra. Since this will usually not be true with raw data, the data is reduced using the previously discussed methods.

This data is mean centered and standardized to unit standard deviation and then used in canonical analysis. The within-group sum of squares and the between-group sum of squares matrices are used to calculate eigenvectors and eigenvalues [7]. Canonical correlation is very similar to PCA, except that in canonical correlation the eigenvectors are sorted to provide maximum variation between groups whereas the eigenvectors of PCA are sorted to provide maximum total variation. The entire standardized data may then be projected onto the first few eigenvectors.

The eigenvectors can also be visualized as in PCA to reveal regions of interest.

## 3. Experimentation

In this study, we use imaging mass spectrometry to analyze negatively charged lipids in the mouse cerebellum. We apply the previously described multivariate techniques to the data for ion classification and spatial pattern detection using our new GUI developed in MATLAB. The imaging data shown here were also processed using commonly used software (BioMap) for comparison.

### 3.1. Biological samples

To analyze brain samples by IMS, frozen mouse brain tissues (cerebellum) were first sectioned into 8~10 um slices at -18 ºC using a cryostat (Microm Cryo-Star HM 560MV) and thaw-mounted to MALDI plates. The neighboring section was thaw-mounted onto glass-slide for hematoxylin and eosin staining by Leica autostainer. No embedding medium was used during the whole process. DHB matrix (30 mg/mL 2,5 dihydroxybenzoic acid in 50:50 acetonitrile/0.1%TFA in $dH_2O$) was then deposited on the tissue surface by a home-built oscillating capillary nebulizer (OCN) matrix sprayer system. Mass spectra of brain tissue were acquired using a Voyager DE STR MALDI-TOF-MS (Applied Biosystems) with a 337nm $N_2$ laser under delayed extraction conditions in reflector mode. MS data sets were acquired using MMSIT (Novartis Pharma AG, Basel, Sweden) over the tissue section. Ion images were reconstituted using BioMap software package (Novartis Pharma AG, Basel, Sweden) and multivariate techniques presented in this paper.

### 3.2. Software Implementation

In order to easily visualize the link between tissue images and molecular profiles, an interface has been created in MATLAB. All the methods described in section 2 have been integrated into it.

**3.2.1. Visualizing mass spectrum profiles.** By clicking on one or many points in the image, the user can display the corresponding mass spectra at each location. Points may also be selected by manually entering the (x,y) coordinates of the pixel of interest. After simultaneous display of multiple molecular profiles, it is possible to undock a given mass spectrum profile by clicking on it. This gives the user a larger visualization of the profile and the ability to zoom in to investigate particular *m/z* values.

**3.2.2. Displaying molecular images.** After determining the *m/z* ratio(s) of a peak of interest, the user may generate grayscale molecular images which show where the given molecules are present. Similar to the mass spectrum display, the user can choose to visualize one ion image or up to six *m/z* ratios side-by-side. Two methods can be used for this visualization. The first method directly displays the corresponding intensity of the given *m/z* values in grayscale.

The second method looks near the given *m/z* ratio in all the spectra to determine if the intensity at the *m/z* value is a likely peak or not. The resulting tissue image displays the magnitude of the largest peak in the neighbourhood, specified by the user, of the selected *m/z* value. The peak detection method used is discussed further in section 3.2.3.

For both methods, if multiple slices are displayed, the user can decide to display those slices side-by-side or to superimpose them. By superimposing up to three of these molecular images, the user can see how different molecules coexpress throughout the image.

Figure 2 shows ion images corresponding to four given *m/z* values in the top panel. The bottom panel of the figure shows mass spectrum profiles from six spatial locations in the image.

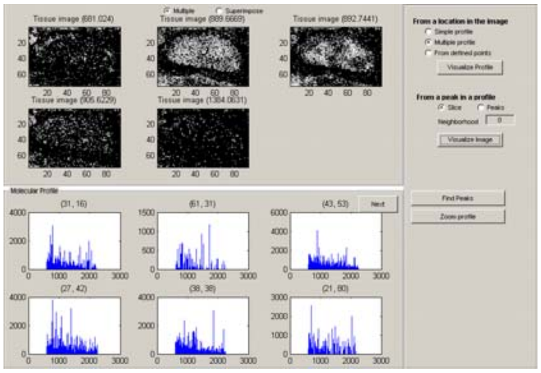


**Figure 2.** Visualization of ion images and profiles.

Figure 3 (a) shows a superimposed pseudo-fluorescence ion image for two selected *m/z* values of a mouse cerebellum IMS dataset. Figure 3 (b) shows a similar image prepared from a different anatomical sample using three *m/z* values. In each case, spatial overlap of molecular species can be observed in the blending of the colors red green and blue. This feature is particularly useful for detecting the correlation among several molecular species.

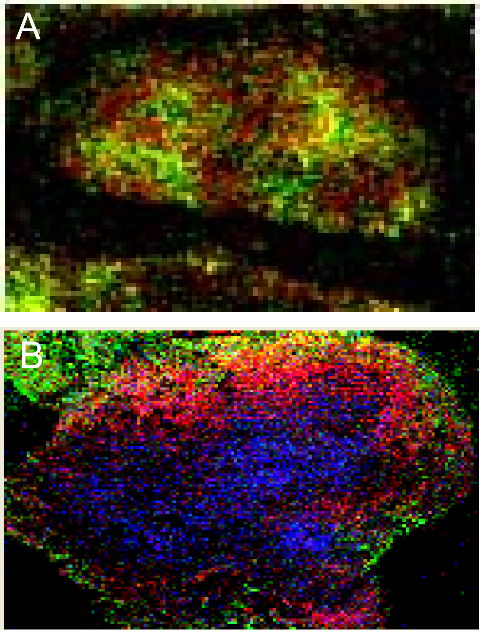
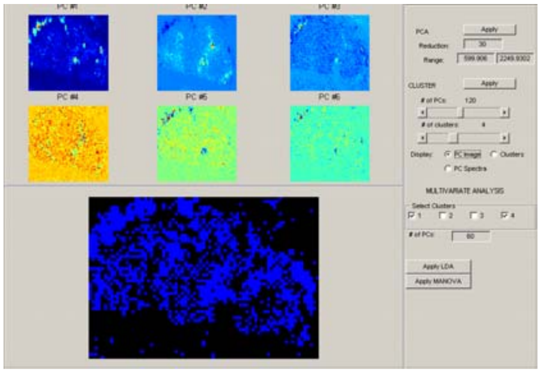


**Figure 3.** Superimposing slices from given *m/z* using pseudo-fluorescence color imaging.

**3.2.3. Peak detection.** A peak in a mass spectrum corresponds to the significant presence of a molecule with a specific *m/z* value. The signal intensity for a *m/z* value is considered a peak if it is the highest among its nearest ±N neighboring points. This definition of a peak has been given in [8]. This algorithm is implemented in C++ and interfaced with the Matlab code.



PCA Apply

Reduction: 30

Range: 599.906 2249.9302

CLUSTER Apply

# of PCs: 120

# of clusters: 4

Display: ● PC Image ○ Clusters ○ PC Spectra

MULTIVARIATE ANALYSIS

Select Clusters ☑ 1 ☐ 2 ☐ 3 ☑ 4

# of PCs: 60

Apply LDA

Apply MANOVA

**Figure 4**. GUI for data analysis.

**3.2.4. Multivariate analysis**. All of the multivariate techniques discussed in section 2, PCA, LDA, clustering and MANOVA, have been incorporated into the GUI. Figure 4 is a screenshot of the analysis module. From top to bottom, the first six principal components and resulting classification using LDA are shown.

# 4. Results

## 4.1. Principal components analysis (PCA)

PCA analysis allows us to quickly find structural information and molecules that are important in defining regions. Figures 5 and 6 demonstrate the capability of PCA to extract underlying information out of huge data sets. Instead of an exhaustive inspection through all the dimensions, PCA analysis in the *m/z* range of 750-1200 gives rich structural information using only a few principal components. For example, the second principal component shows important peaks around an *m/z* value of 890, corresponding to sulfatide molecules (ST24:0) that have been found in brain [9].

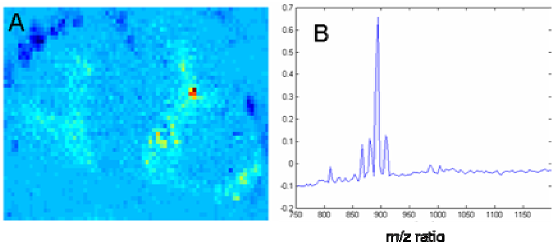


**Figure 5.** (a) 2nd PC image (b) 2nd PC spectrum on lower range *m/z* (750 - 1200).

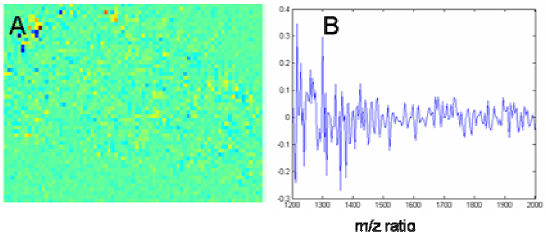


**Figure 6.** (a) 2nd PC image (b) 2nd PC spectrum on higher range *m/z* (1200 - 2000).

The PCA data reveal that the molecular distributions match the H&E stained image of the cerebellum, shown in Figure 7 (a). Figure 7 (b) shows the ion image of sulfatide (ST24:0). The spatial distributions of sulfatides indicate their abundances in white matter are much higher than in the granular layer or molecular layer. The PCA analysis also correlates two sulfatide species (ST24:0, *m/z* 890.7 and ST24:1(OH) *m/z* 904.7) to the same location in the white matter. PCA analysis in mass range of 1200~2000 does not show any significant feature (Figure 6), which is consistent with expectations.

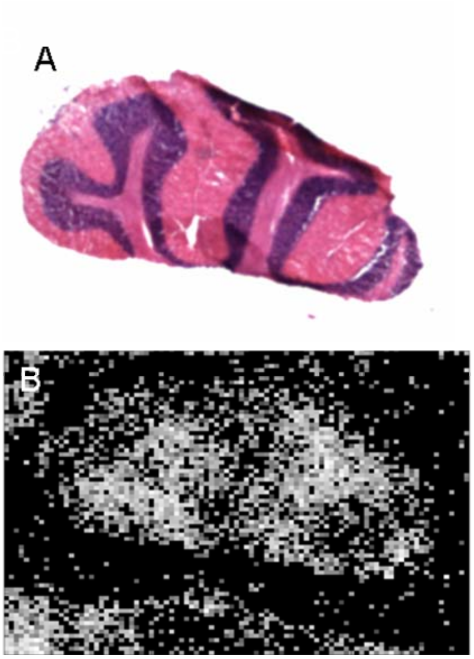


**Figure 7.** (a)Anatomical image and (b) ion image at *m/z* 890.7 of the stained cerebellum

### 4.2. Clustering

Figure 8 shows the results of applying k-means (k=4) clustering to the data collected from the same sample used in Figure 7. The lower range of the mass spectra (750-1200) was used with 80 principal components selected for clustering.

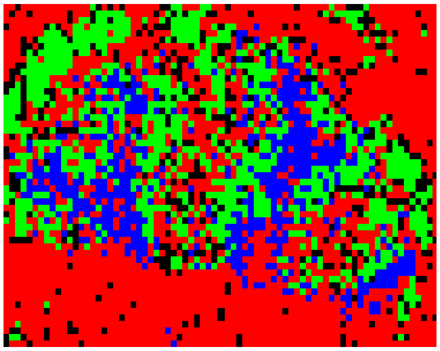


**Figure 8.** Clustering on PCA images. Each color represents a different cluster.

### 4.3. Linear discriminant analysis

Figure 9 show the results of applying LDA to the cerebellum sample shown in Figure 7. The *m/z* range from 750-2000 was used in the analysis and 50 principal components were used in LDA. Two clusters were selected for analysis in Figure 9 (a), which classifies pixels in white matter separately from the rest of the pixels. Three clusters were used in Figure 9 (b), which shows the pixels classified in three groups, background, gray matter, and white matter.

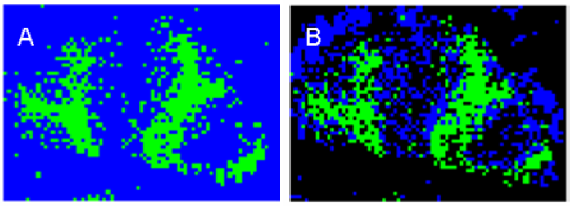


**Figure 9.** (a) LDA on PCA images with 2 clusters and (b) with 3 clusters.

### 4.4. Multivariate analysis of variance

Figure 10 shows the results of using MANOVA on the sample from Figure 7. The *m/z* range from 750-2000 was used in the analysis and 50 principal components were used in MANOVA. Two clusters were selected for analysis in Figure 10 (a), which distinguishes the gray matter of the cerebellum from the background and the white matter. Three clusters were used in Figure 10 (b), which shows the background, gray matter, and white matter all distinctly.

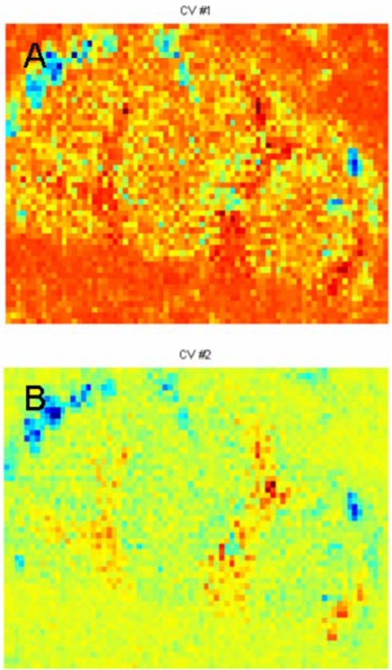


**Figure 10.** (a) MANOVA with 2 clusters and (b) 3 clusters.

## 5. Discussion

BioMap is one of the most popular software tools for the analysis of imaging mass spectrometry data. It too reconstructs the data into thousands of ion images. The ion image of sulfatide (ST24:0) obtained using BioMap is shown in Figure 11 (a), and the ion image using our program is shown in Figure 11 (b).

Although Biomap provides some functions for the reconstruction of IMS data, it does not provide any multivariate tools for the data processing. The onus is left on the user to go through all the ion images to find meaningful spatial patterns. In general, BioMap requires prior information and strong hypotheses in order to reach meaningful conclusions. It is also not convenient to classify ion images using BioMap. Because those images can not be compared directly, they need to be investigated by third-party software. In contrast, the multivariate approaches described in this paper increase the speed of the data analysis and improve the quality of ion images while maintaining high accuracy.

Our methodology provides a convenient framework to link tissue imaging with molecular profiling. The use of PCA to project the data onto principal components alone seems to be able to show different spatial regions in the mass spec data. Out of the first few PCA images, at least some of the images consistently show separate spatial regions in the sample. Additionally, the results of clustering seem to show groupings that have similar spatial distributions to those produced by PCA alone (compare Figure 8 and 5(a)). Both PCA and clustering are very useful when prior information about the sample is limited.

The results of LDA are very similar to clustering when the clusters are used to seed the LDA. Fewer clusters can be used though to selectively differentiate one tissue type from the rest as figure 9 shows.

When the image is projected onto the canonical variables calculated using MANOVA, images similar to the PCA images are created. Unlike PCA, regions can be analyzed selectively using MANOVA, and it appears that this technique can better differentiate some regions than PCA alone. In figure 10 for example, notice the same three distinct regions that were found using LDA, and that visually the three regions appear more distinct than in the PCA image.

The spatial distributions of sulfatides reveal that their abundances in white matter are much higher than granular layer and molecular layer. The PCA analysis also correlates two sulfatide species (ST24:0 and ST24:1) with similar distributions in the white matter, which could provide more insight on the biosynthetic pathways and functions of negatively charged lipid molecules.

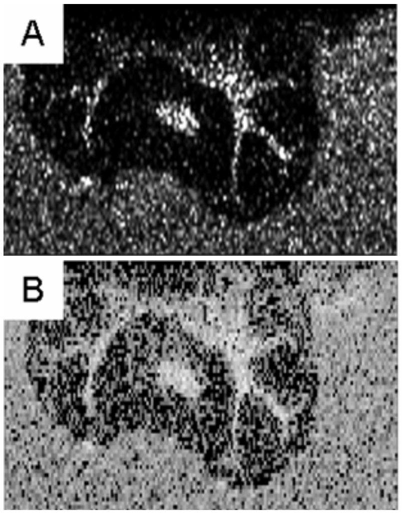


**Figure 11.** Ion images of sulfatide (ST24:0) in mouse cerebellum using (a) BioMap, and (b) our tool.

## 6. Conclusion

The multivariate tool presented here provides a universal platform for fast and effective processing of imaging mass spectrometry data using PCA, clustering, LDA and MANOVA. Many features such as peak detection, superimposing and multiple simultaneous images/spectra display have been developed to facilitate the data processing. This tool generates clear mass footprints and high-quality images of biological samples that match the histological results. Future work will involve the incorporation of other data reduction methods, such as independent components analysis (ICA). Since much of the differentiation is based upon initial clustering, a study to define optimal metrics on the space of mass spectrum profiles will also greatly improve the performances of the data analysis system.

## 7. Acknowledgment

This research has been supported by grants from Microsoft Research, National Institutes of Health (Bioengineering Research Partnership R01CA108468, P20GM072069, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to M. Wang), and National Science Foundation (Fellowship to R. Moffitt).

## 8. References

[1] M. Stoeckli, P. Chaurand, D.E. Hallahan, and R.M. Caprioli, "Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues", *Nat Med*, 7, 2001, 493-496.

[2] R. Lemair, M. Wisztorski, A. Desmons, J.C. Tabet, R. Day, M. Salzet, and I. Fournier, "MALDI-MS Direct Tissue Analysis of Proteins: Improving Signal Sensitivity Using Organic Treatments", *Anal Chem*, 78, 2006, 7145-7153.

[3] G. McCombie, D. Staab, M. Stoeckli, R. Knochenmuss, „Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis", *Anal Chem*, 77, 2005, 6118-6124.

[4] S. Aoyagi, Y. Kawashima, M. Kudo, "TOF-SIMS imaging technique with information entropy", *Nucl. Instrum & Methods Phys Res*, Sect. B 232, 2005, 146-152.

[5] R. Van de Plas, F. Ojeda, M. Dewil, L. Van den Bosch, B. De Moor, E. Waelkens, "Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis", Pacific Symposium on Biocomputing, 2007.

[6] N.P. Lockyer, J.C. Vickerman, "Progress in cellular analysis using ToF-SIMS", *App. Surf Sci*, 231, 2004 337-384.

[7] L. Shintu, S. Caldarelli, "Toward the Determination of the Geographical Origin of Emmental(er) Cheese via High Resolution MAS NMR: A Preliminary Investigation", *Agric Food Chem*, 54, 2006, 4148-4154.

[8] Yasui, McLerran, Adam, Winget and Feng, "An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers", *J Biomedicine and Biotech*, 2003, 242-248.

[9] S.N. Jackon, H.J. Wang, A.S. Woods, "In Situ Structure Characterization of Glycerophospholipids and Sulfatides in Brain Tissue Using MALDI-MS/MS", *J Am soc Mass Spectrum*, 18, 2007, 17-26.