BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–10

Unsupervised Spatial Tissue Exploration via Imaging Mass Spectrometry: Preprocessing and Clustering

Raf Van de Plas [a c], *Fabian Ojeda* [a c], *Maarten Dewil* [e], *Ludo Van Den Bosch* [e], *Bart De Moor* [a c], *Etienne Waelkens* [b c d]

[a] *Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium.*
[b] *Katholieke Universiteit Leuven, Department of Molecular Cell Biology, Afd. Biochemie, O & N, Herestraat 49 - bus 901, B-3000 Leuven, Belgium.*
[c] *Katholieke Universiteit Leuven, ProMeta, Interfacultary Center for Proteomics and Metabolomics, O & N 2, Herestraat 49, B-3000 Leuven, Belgium.*
[d] *Katholieke Universiteit Leuven, BioMacS, Interfacultary Centre for Biomacromolecular Structure, p/a IRC KULeuven Campus Kortrijk, E. Sabbelaan 53, B-8500 Kortrijk, Belgium.*
[e] *Katholieke Universiteit Leuven, Department of Neurosciences, Neurobiology, O & N 2, Herestraat 49, B-3000 Leuven, Belgium.*

ABSTRACT

Motivation: Many studies that discuss organic imaging via mass spectrometry, employ a univariate approach for data exploration. This approach is not necessarily well suited to the large amounts of data that these techniques tend to produce. This paper reports on an initial approach towards performing an *in silico* unsupervised multivariate exploration of MALDI-based imaging mass spectrometry (IMS) data. The goal is to detect in an unsupervised way zones of similar chemical composition in organic tissue slices, using data acquired via IMS measurement. Once such a tissue slice has been segmented into zones, the representative mass spectra are used to investigate the main characteristics that differentiate these areas.

Results: The described computational analysis is applied to the IMS measurement of a nerve tissue section from a standard control rat. The main tools that are used for the segmentation are *k*-means clustering and the *silhouette* cluster validation metric. In addition, the influence on the clustering results of two preprocessing steps, *null spot removal* and *principal component analysis*, is investigated. The results imply that MALDI-based imaging mass spectrometry, combined with appropriate *in silico* preprocessing and clustering, is capable of successfully partitioning a tissue slice into biologically relevant chemical zones without supervised guidance. In our rat spinal cord example, this resulted in the differentiation between gray and white matter in the nerve tissue.

Contact: raf.vandeplas@esat.kuleuven.be

1 INTRODUCTION

Proteomics and peptidomics are the direct qualitative and quantitative analysis of respectively the proteine complement and the peptide complement in an organism, tissue, or cell under a set of known physiological and environmental conditions. Mass spectrometry, aided by fractionation techniques such as liquid chromatography and gel electrophoresis, has become the central tool for proteomics/peptidomics research. It has been at the basis of a considerable set of successes in fields such as experimental biochemistry (Aebersold *et al*., 2003) and the elucidation of pathologies such as Alzheimer's Disease (Kanninen *et al*., 2004).

Most proteomics and peptidomics studies, however, have focused solely on the identification and the quantitative analysis of proteins and peptides in a sample, without taking the absolute spatial location or exact origin of the sample within tissue into account. The few studies where chemical sample analyses were linked to their spatial origin in a tissue section, demonstrated that this type of information could provide vital clues in gathering further insight into biological processes (Huh *et al*., 2003; Nygren *et al*., 2004).

In this study we want to preserve and examine the explicit link between peptidomics-oriented mass spectral measurements and the absolute spatial locations of these measurements within an organic tissue slice. The aim is to use the combination of mass spectrometry and spatial information to acquire some insight into the more spatially-oriented biomolecular processes, such as peptide location and distribution.

© Oxford University Press 2006.

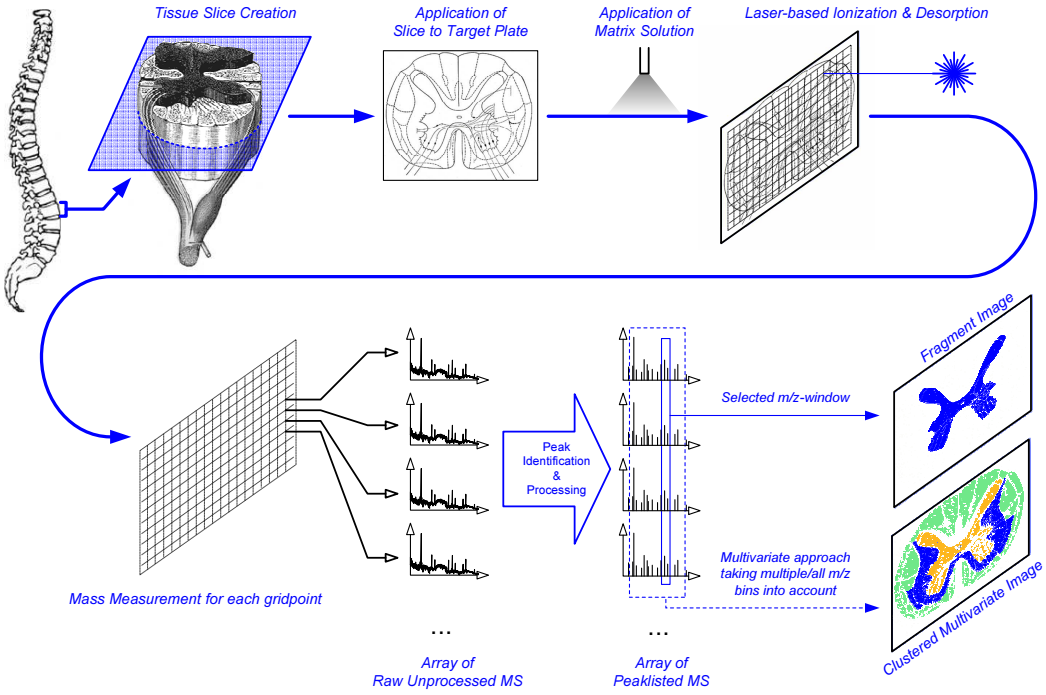


Fig. 1. Overview of the imaging mass spectrometry experiment. The tissue slice creation, the mounting on the target plate, and the application of an appropriate matrix solution are wet-lab steps. The mass spectral measurements, the data collection, and the translation into peaklisted mass spectra take place inside the mass spectrometer. The array of peaklisted mass spectra forms the starting point for an *in silico* analysis.

The technology used to gather this type of data is known as laser-based or MALDI-based Imaging Mass Spectrometry (IMS).

1.1 Imaging Mass Spectrometry

Imaging mass spectrometry is a relatively new technology that uses the molecular specificity and sensitivity of normal mass spectrometry to collect a direct spatial mapping of biomolecules (or their molecular fragments) in tissue sections. It has been successfully used in a set of pioneering studies that have mainly focused on mammalian tissue (Stoeckli *et al.*, 2001, 1999).

The wet-lab procedure consists of cutting and preparing a tissue section, fixing the section onto a MALDI target plate, applying an appropriate chemical matrix solution, and performing a MALDI mass spectral measurement on each grid point of a virtual array that has been superimposed on the tissue section. The result is an array of spots or 'pixels' covering the tissue section, with a mass spectrum linked to each individual pixel. Figure 1 gives an overview of the steps involved in performing an IMS experiment on a spinal cord tissue section and the subsequent *in silico* paths that were explored. After performing single MS measurement of each spot, the mass spectrometer firmware translates the raw mass spectra into peaklisted mass spectra, which are collected into an IMS raster by an IMS-enabling software module of the mass spectrometer operator software (in our case the *4000 Series Imaging* module).

Once the raster of peaklisted mass spectra is available in electronic form, the *in silico* processing and exploration can start. The data generated by an IMS experiment has the form of a mathematical space that has two spatial dimensions (the x and y-dimension) and as many feature dimensions as the mass range settings of the mass spectrometer allow (typically, the m/z-dimension). Normally, this means a grid of pixels in which each element encodes a peaklisted mass spectrum as a vector of intensities for discretized mass-over-charge (m/z) values.

1.2 Fragment Images

A first step, often taken in IMS-oriented studies, is generating false color images representing the spatial mapping on the tissue section of peak intensities at a selected m/z-window that shows an interesting variation or spread. As these types of images depict the spread of a certain peak over the tissue slice, and a mass spectral peak corresponds to a fragment of a biomolecule that was detected by the mass spectrometer, these images can be called 'fragment images'. In essence, a fragment image is the result of a cross-cut at a certain m/z-window of the three-way array that holds the IMS data. Figure 2 shows a fragment image taken at $5484.29\ m/z$ from the spinal cord data set studied in this paper. The fragment image shows that the rat spinal cord tissue section was slightly off center with

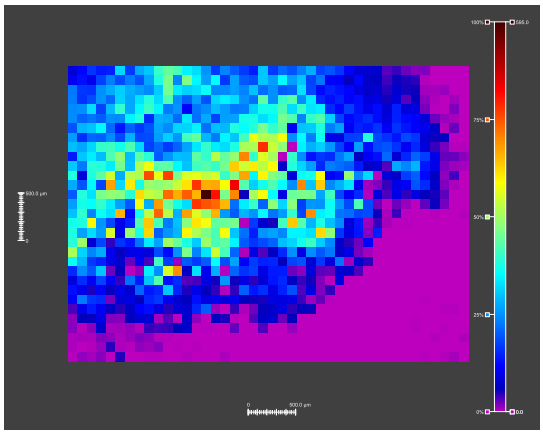


Fig. 2. Fragment image taken at $5484.29\ m/z$ from the rat spinal cord data set.

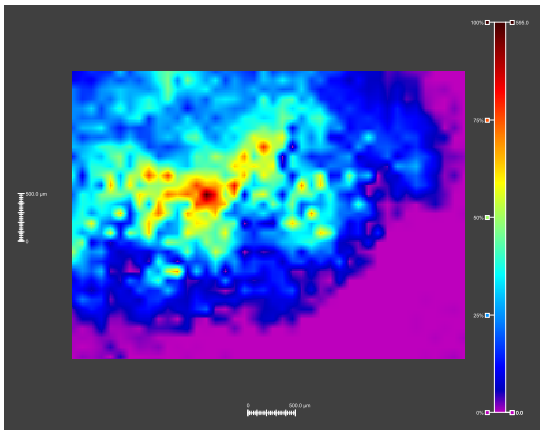


Fig. 3. Bilinearly interpolated version of the fragment image in Fig. 2, using $7\times$ subsampling.

regards to the IMS raster, with the lower right corner area being clear of any tissue material.

As the spatial resolution of IMS is still rather coarse (e.g. $100\ \mu m$ pixelwidth) compared to imaging techniques such as light microscopy (e.g. $1\ \mu m$ pixelwidth), it can be hard for a human observer to recognize structures. For this reason, we calculated a bilinearly interpolated version of the raw fragment image to facilitate easier human interpretation. Figure 3 shows the bilinearly interpolated version, using $7\times$ subsampling, of the fragment image shown in Fig. 2.

1.3 Multivariate Images

Fragment images can be considered as a univariate approach towards IMS imaging, as they are the result of picking one feature variable per pixel to investigate. This means that an IMS data set can deliver as many different fragment images as there are m/z-values available. The number of features per pixel can easily run into the thousands and selecting the features that

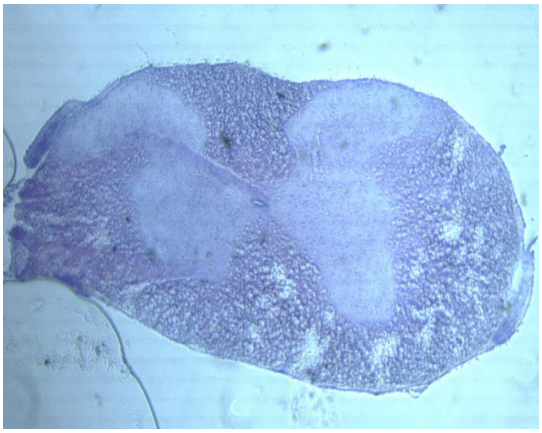


Fig. 4. Microscopic image of a transversal section of a rat spinal cord. Although it is not a picture from the section that was used for the IMS imaging, it was produced using the same protocols in the same lab. It illustrates, using conventional immunocoloring and light microscopy, the presence of the butterfly-shaped *Substantia grisea* and the surrounding white matter in the nerve tissue. A comparison can be made with the fragment image of Fig. 2.

will be most relevant is a nontrivial problem. For example, in our rat spinal cord data set there are $7451\ m/z$-values per pixel available. Essentially, one could argue that picking a certain m/z-window and studying the fragment image it produces, is a good course of action once it is known which m/z-range is relevant to the particular problem under study. However, for unguided, prospective exploration, fragment images are much less suited.

Therefore, we suggest taking a multivariate approach to IMS imaging and producing an image that tells the investigator which trends and zones are found when all - instead of just one - m/z-values per pixel are taken into account. The objective is to construct a multivariate overview image for the tissue slice that segments the measured area into zones of similar chemical composition, taking the entire mass spectral range per pixel as input and without external guidance. The multivariate approach produces a higher level description of the spatial mapping (areas instead of pixels) and can be used to identify interesting mass spectral ranges or spatial areas that allow for further exploration using fragment images.

2 APPROACH

Our goal is to group together - in an unsupervised fashion - spots that show a similar chemical composition on the basis of their mass spectrum. As each of these spots is linked to a spatial location in the tissue slice, this grouping can be seen as a segmentation of the image area into zones[1]. Segmentation in image space (the x and y-dimensions) can be

[1] In contrast to segmentation in a common computer vision context, we do not impose the constraint that segmentation should produce contiguous zones.

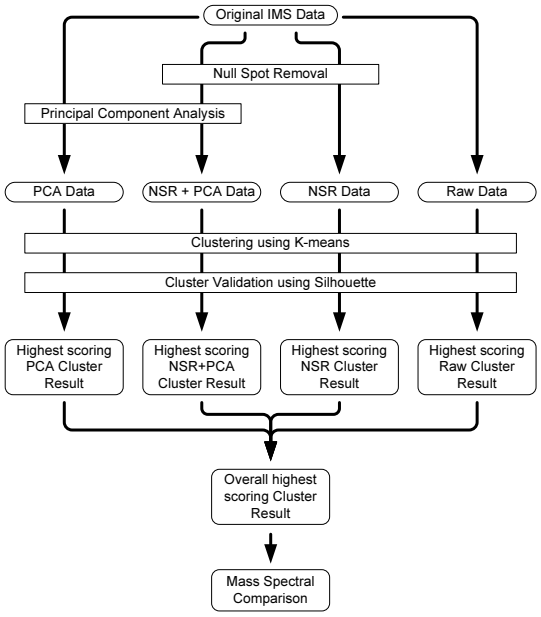


Fig. 5. Overview of the multivariate approach. First, the creation of four data classes by applying different combinations of preprocessing steps. Then, clustering of each of the parameter combinations, followed by cluster validation using *silhouette*. Finally, a comparison between the representative mass spectra of the zones found in one the highest scoring segmentation results.

achieved by clustering in feature space (the m/z-dimension). As a clustering method for the mass spectra, we use the *k*-means clustering algorithm with a set of different parameters and distance measures. The use of *k*-means clustering in an IMS context has been described by McCombie *et al*., 2005, but the validation of the found clusters was done by visual inspection. In this study, the intent is to have an as unbiased approach to the data as possible, meaning that we use the *k*-means algorithm with various different parameter combinations. The large number of cluster partitionings resulting from the many parameter combinations, demands an unsupervised, computational method for grading the clustering performance. To this end, we employ the *silhouette* metric (Rousseeuw, 1987), as a means of determining relative quality of clustering.

In addition to the clustering procedure, we describe two preprocessing steps and their influence on the segmentation results. One of these steps is Null Spot Removal (NSR). Fig. 2 indicates how the IMS raster covers most of the tissue section, but there remains an empty section where there is no tissue to be measured. This area contains spots for which the mass spectrum is practically flat-lined and with NSR we lift these 'empty spots' from the analysis prior to performing heavy calculations. This improves resource usage of the algorithms and avoids that the clustering process is influenced by spots that carry no information anyway. The other preprocessing step is the application of Principal Component Analysis (PCA) to reduce the dimensionality of the clustering problem, while retaining from the data as much of the original variance and information as possible.

In a final phase we take a closer look at the best clustering results according to our validation metric and investigate the differences between the representative mass spectra that characterize each of the zones. Figure 5 gives an overview of the various routes under study.

3 MATERIALS AND METHODS

The tissue sections (15 micrometers thick) were taken from a transversal section of the spinal cord of a standard control rat. The recorded mass range extended from 5000 to 12000 m/z and alpha-cyano-4-hydroxy cinnamic acid was used as a matrix. A MALDI mass spectral measurement was performed on each grid point of a virtual raster that was superimposed on the tissue section with an interspot distance of 100 micrometers in both the x and y-directions. The mass spectrometer that was used is the ABI 4800 MALDI TOF/TOF Analyzer from Applied Biosystems Inc. Per spot 500 spectra were collected and averaged into a representative spot mass spectrum per 'pixel'. The data collection in the mass spectrometer was guided by the *4000 Series Imaging* module, available at `http://www.maldi-msi.org`. The resulting data cube was processed using custom developed software and representative fragment images were exported in both a raw format and a more human-interpretable, bilinearly interpolated form.

4 PREPROCESSING

4.1 Null Spot Removal

As mentioned in section 2, the tissue slice is slightly off center with regards to the IMS raster. This results in a number of spots being 'empty' measurements with essentially a flat-lined mass spectrum connected to them. These types of spots can also result from other conditions, such as holes in the tissue. The mass spectra of these spots generally consist of very small intensity peaks (e.g. due to measurement noise) and zero values. As these 'null spots' can lead to degenerative or biased results in further steps of the data analysis process, we implemented an approach to identify them and investigate their impact on the final clustering outcome.

The technique we used to identify the null spots, relies on quantifying the area under the curve (AUC) of each of the mass spectra. The AUC is computed as the sum of all intensities within a given range of m/z-values. As the intensities in this data set are always positive and the range of m/z-values is equally spaced and common for every mass spectrum, the AUC becomes equivalent to computing the average intensity per spot.

To evaluate whether a mass spectrum is 'empty', we compare the AUC of the spot under investigation to the AUC of the average mass spectrum of the entire data set. All spots

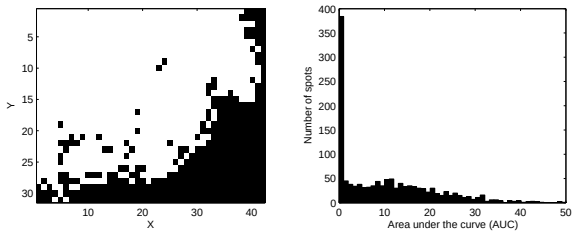


Fig. 6. (Left) Null spots represented as black dots, mostly located in the lower right corner of the image raster. These spots were found to be linked to mass spectra for which the AUC fell below the threshold (10% of average ACU). (Right) Histogram plot showing approximately 400 spots having an AUC lower than 1.031 (10% of the average AUC).

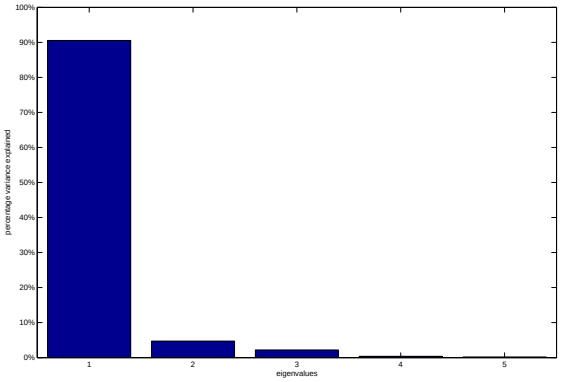


Fig. 7. Bar plot showing the variance explained by each of the first five principal components. The cut-off was set at 0.1%. Lower contributions are not shown.

having an AUC lower than 10 percent of the average AUC, are labeled null spots. For the spinal cord data set, the average AUC is 10.31, which means that spots with an AUC lower than 1.031 are filtered out. The motivation for putting this threshold at a value of 10 percent, came from the histogram plot, which is shown in the right panel of Figure 6. The histogram shows about 400 spots, mostly located in the lower right corner region of the fragment image (see black dots on the left panel of Figure 6), having AUCs lower than the imposed threshold.

4.2 Principal Component Analysis

In their unprocessed form, the spots in an IMS data cube are linked to a vector of features which can have thousands of entries (one for each m/z-value, the amount depending on the mass range settings that were used). For our rat spinal cord example, there are 7451 features per spot. Hence, every spot can be represented as a point in a 7451-dimensional space. The question now is whether the cloud of spots in this high-dimensional space can be described in a much lower-dimensional space with still a considerable amount of information retained. A representation of spots in a lower-dimensional space, where each spot is characterized by only a few features instead of the original thousands, allows for faster processing and the ability to filter out variation that can be labeled as noise. Principal component analysis (PCA) (Joliffe *et al.*, 1986) is an unsupervised multivariate technique that transforms a set of correlated features into a set of uncorrelated variables, expressed as linear combinations of the original features. The extent of PCA-based dimensionality reduction that can be achieved with a preset amount of information retained, depends on the amount of redundancy that the data set contains.

In a mass spectral setting, the dimensionality reduction turns out to be quite spectacular. We applied PCA to the spinal cord data set via singular value decomposition of the covariance matrix of the spot-feature matrix and charted, using the eigenvalues, the amount of variance in the original data set that is explained by each of the returned principal components (PCs). As we have 7451 features and 1302 spots (42×31) the number of relevant PCs is maximally 1302. However, if we calculate the amount of variance explained by each of these PCs, we notice that only the first 5 PCs contribute substantially. Figure 7 shows a bar graph of the variance explained by the first five PCs. The cut-off of this graph was set at 0.1%, which indicates that the 1297 PCs that didn't pass the threshold have a neglible contribution. On the basis of these figures, we select the first 5 PCs (which still manage to describe 98% of the variance in the original data set) as new, lower-dimensional space in which the spots are described. One explanation for this strong dimensionality reduction, going from 7541 dimensions to just 5, might be that mass spectra tend to have a lot of inherent redundancy. Many peaks in a mass spectrum are intrinsically related to other peaks in the same mass spectrum (for example, because the peaks are representations of different fragments that came from the same parent molecule).

Another advantage of performing PCA is that it can provide us with an preliminary view on the trends that are underlying the data. Each PC differentiates a certain subset of spots from others, but in their vector form it is very hard to learn anything from these PCs directly. However, if we fold the PCs back to image space as is done in figure 8, we can see structures show up. The visualisation of Fig. 8 shows the first four PCs folded to image space, and uses a coloring in which each PC differentiates the blue spots from the red spots. Taking into account the eigenvalues that indicate the contribution of each of these PCs (Fig. 7), it is clear that one of the most important sources of variation in the data, described by the first PC (90.5%), is linked to the difference between spots in the center roughly following a butterfly-shape, and the spots that surround this area. The second principal component describes 4.7% of the data, and differentiates spots found in the lower yellow-to-red cloud from a rather contiguous blue region of spots found at the top. The weight of the third principal component has

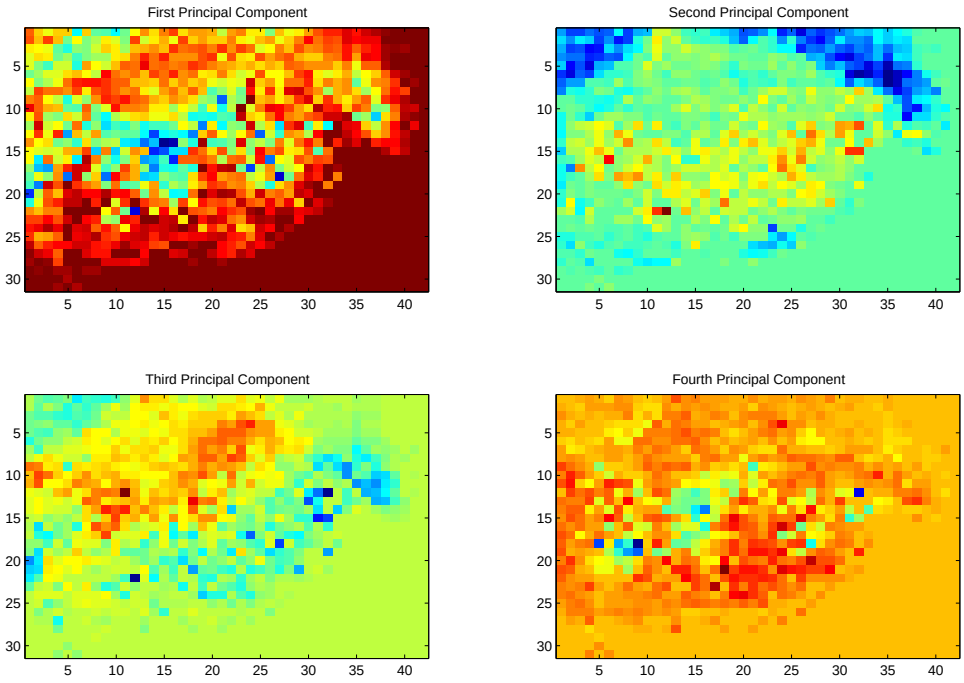


Fig. 8. Principal components folded back to image space. Keep in mind the percentages of variance explained by each of these PCs (see Fig.7).

dropped to 2.2%, and it is mainly employed to differentiate a lower grouping of spots from an upper grouping of spots. This might be related to dorsal/ventral chemical differences which are known to exist in rat spinal cord. Further details on what these areas might mean can be found in section 6.

5 CLUSTERING

The goal of clustering is to partition a given data set into groups or clusters, such that the data points belonging to the same group, will be more similar to each other than to points in different clusters. In this study we employ the widely used *k*-means clustering algorithm and various objective distance measures, such as the squared Euclidean distance (Jain *et al.*, 1988). Since there is lack of a general theoretical solution to finding the optimal number of clusters for any given data set, we compare the results of multiple runs with different values for *k* and use the *silhouette* validation metric to define a relative performance.

5.1 Distance Measures and Clustering Parameters

For the *k*-means clustering algorithm, several parameters need to be defined before the clustering process can commence. The partitioning may depend heavily on these initial settings, which is why we particularly look at the distance measure and the number of clusters. Three different distance functions are tried: squared Euclidean distance (sqEuclidean), sum of absolute differences (cityblock), and one minus the cosine of the included angle between points (cosine). The number of clusters is varied from 2 to 11. All other parameters, such as the number of randomizations (10) and the maximum number of iterations are kept fixed.

5.2 Clustering Validation

Visual inspection of the partitioning results can usually give a good impression of clustering performance. However, it remains a rather subjective indication which doesn't fit well with our goal of unsupervised processing, and in many cases, such as with the large number of parameter combinations used here, it is just not practically feasible. Therefore, we require a computational means of assessing clustering quality.

The *silhouette* index (Rousseeuw, 1987) can be used to evaluate cluster coherence, i.e. it defines how well assigned a sample is. This metric considers a sample to be well clustered if its distance to the other samples in the same cluster is small, and its distance to the samples in other clusters is large. The *silhouette* index ranges from -1 to 1, and the closer it approaches 1, the higher the quality of the clustering. In this paper, the *silhouette* index (or rather its average over all clustered spots) is used to select the 'best' clustering result

from the many combinations of distance measures, random initializations, number of clusters, and applied preprocessing steps.

5.3 Clustering Strategy

As Fig. 5 indicates, we perform clustering using the various parameter combinations on four classes of data. The first class (Raw Data) is simply the raw IMS data without any preprocessing steps applied to it. This means that this data still contains null spots and that it has high dimensionality (7451 features). The second class of data (NSR Data) has only the null spot removal preprocessing step applied to it, which results in a lesser number of spots to consider in the analysis, but they are still described in a high-dimensional space. The third data class (NSR+PCA Data) has been fully processed by both the null spot removal step and PCA. This results in a low-dimensional data set (only 5 PCs or less are retained) and null spots are also kept out of the analysis. The fourth and final class (PCA Data) has only undergone dimensionality reduction via PCA.

After we perform the clustering for each parameter combination in each of the four data classes, we do an intra-class ranking of the partitioning results. Then, we take a look at the overall best results over all four the classes. In this range of high-scoring partitions, we take a closer look at the representative mass spectra of the clustered zones, to see where the differences lie in a mass spectral context.

5.4 Clustering Results

The results are first considered within the context of their data class, followed by an overall look at the best performing combinations of parameters and preprocessing steps.

5.4.1 Raw Data The results for the Raw Data class are shown in Table 1. The five best performing combinations according to the average *silhouette* index all used the squared Euclidean distance (sqEuclidean) with the highest validation score at two clusters. The index for three clusters is a close second, indicating a still considerably good partitioning. For the cityblock measure the best result was found at 2 clusters with an average *silhouette* of 0.5797, whereas for the cosine distance the null spots prevented the clustering algorithm from converging (due to their mass spectra having very small relative magnitudes). The reason is that the cosine distance first normalizes all the entries to unit Euclidean length, causing very small relative magnitudes to be effectively considered zero.

5.4.2 NSR Data The results for the NSR Data class are shown in Table 2. As for the Raw Data, the squared Euclidean distance turns out to be the dominant measure. The cosine measure gave as its best result, an average *silhouette* index of 0.4523, at two clusters. Compared with the Raw Data, we see noticably lower average *silhouette* values appear after the null spot removal has taken place. One might argue that this indicates that the null spot removal step is hurting performance of the clustering algorithm. However, this (relatively small) drop in clustering performance was to be expected and can actually be considered a correction of the *silhouette* values that turned up in the Raw Data. In Rousseeuw, 1987, the *silhouette* index gives values close to 1 when the clusters are homogeneous and distant from each other. In the Raw Data the null spots form a very homogeneous (but trivial) cluster with *silhouette* indices close to 1, which push up the average *silhouette* index. Hence, the absence of these null spots leads to relatively lower *silhouette* values, though no trivial clusters (bias) are incorporated into the final result.

Table 1. Clustering results for the original data set (Raw Data). Top five results according to the average *silhouette* index.

Distance	Number of clusters	Average silhouette
sqEuclidean	2	0.7281
sqEuclidean	3	0.6967
sqEuclidean	4	0.6440
sqEuclidean	6	0.6294
sqEuclidean	5	0.6219

Table 2. Clustering results for the data set in which null spots are not considered (NSR Data). Top five results according to the average *silhouette* index.

Distance	Number of clusters	Average silhouette
sqEuclidean	2	0.6762
sqEuclidean	3	0.5725
sqEuclidean	4	0.5190
cityblock	2	0.5131
sqEuclidean	7	0.4991

5.4.3 PCA Data The results for the PCA Data class are shown in Table 3. Performing dimensionality reduction on the original high-dimensional data results in consistently higher clustering performance, presumably due to the denoising effect of applying PCA with a limited number of principal components. Using PCA as a preprocessing step, allows us to also vary the number of PCs to include. As Table 3 shows, using two PCs gave the best performance. This is to be expected, given that most variance in the data is explained by the first two PCs. Adding additional PCs seems to add noise and influence the average *silhouette* index negatively. The highest average *silhouette* value for the cosine distance was 0.7323 at two clusters using the first two principal components. For the cityblock measure, the best average

silhouette was 0.6937 when looking at three clusters and using two principal components.

Table 3. Clustering results for the PCA projection of the original data (PCA Data). Top five results according to the average *silhouette* index.

Distance	Principal comp.	Number of clusters	Average silhouette
sqEuclidean	2	2	0.7489
sqEuclidean	3	2	0.7392
sqEuclidean	4	2	0.7376
sqEuclidean	5	2	0.7366
sqEuclidean	6	2	0.7362

5.4.4 NSR+PCA Data The results for the NSR+PCA Data class are shown in Table 4. In this data class we experience a quite large increase in clustering performance, according to the *silhouette* metric. Although we saw a decrease in average *silhouette* values when removing null spots (see Tables 1 and 2) and an increase when applying dimensionality reduction through PCA (see Tables 1 and 3)), it seems that the combination of NSR and PCA gives a larger boost in clustering performance than the increase encountered when only PCA was applied. The reason for this appears to be that as the null spots are removed prior to performing PCA, the principal components can be optimally calculated for the regions that describe actual tissue, instead of biasing the PCs to take the considerable amount of null spots into account. Essentially the relevant variance that remains after NSR, can be explored more thoroughly with the same amount of PCs, than would be the case when null spots were not removed. The cosine distance gives the highest values for this data class. The best result for the sqEuclidean measure was 0.7089 at two clusters using two principal components. For cityblock the best performance was 0.4907, with the same settings.

5.4.5 Overall Figure 9 shows the best clustering result for each of the four data classes. It is apparent that within all four classes the highest performing partitioning always consisted of two clusters. In the two cases where NSR was not performed, one of the clusters needs to take into account the null spots (seriously biasing the average representative of that cluster). In the cases where NSR was performed, we saw that the two clusters were able to segment more thoroughly the tissue section, due to not being skewed by the null spots. When comparing the two NSR cases, we also see that the denoising effect caused by PCA (in addition to the savings in time and memory resources), results in a somewhat more contiguous partitioning. The reason that two-cluster segmentations seem to perform so well, might be related to a preference of the *silhouette* metric for a small number of clusters. To determine whether this is the case, we would need to add other clustering validation metrics and reach some form of consensus on what is 'good' cluster performance. This is one of the avenues we will investigate in the near future.

Table 4. Clustering results for the PCA projection of the data with null spots left out (NSR+PCA Data). Top five results according to the average *silhouette* index

Distance	Principal comp.	Number of clusters	Average silhouette
cosine	2	2	0.8360
cosine	2	3	0.8132
cosine	2	4	0.8085
cosine	2	5	0.7944
cosine	2	6	0.7780

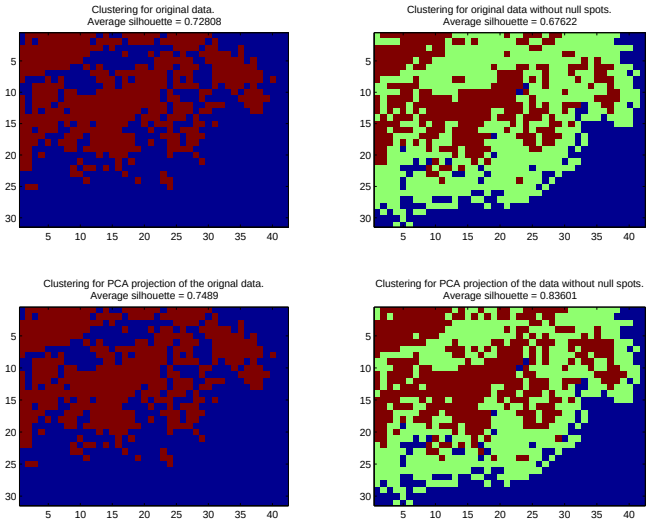


Fig. 9. Best clustering results for each of the four data classes. (upper-left) Raw Data. (upper-right) NSR Data. (lower-left) PCA Data. (lower-right) NSR+PCA Data.

Taking all the *silhouette* rankings into account, it seems clear that the best performing combination of preprocessing and clustering steps, for our spinal cord data set, is:

- Preprocessing:
  - Null Spot Removal (NSR)
  - Dimensionality Reduction (PCA) with 2 PCs
- Clustering via *k*-means
  - Distance measure: cosine
  - Number of clusters: 2

In fact, most of the NSR+PCA combinations seem to outperform any of the other data classes, and this even when a rather high number of clusters is preset (e.g. 6 clusters in Table 4 outperform the best 2-cluster segmentation in Table 3). Figure 10 shows the four best performing segmentations over all data classes (all NSR+PCA). As all these have a very high

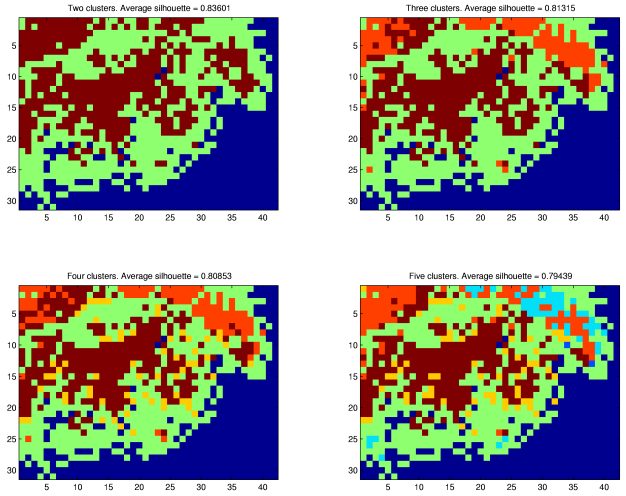


Fig. 10. Clustering results for the best overall preprocessing and parameter combination (NSR+PCA, using the cosine distance measure), varying the number of clusters. (upper-left) 2 clusters. (upper-right) 3 clusters. (lower-left) 4 clusters. (lower-right) 5 clusters.

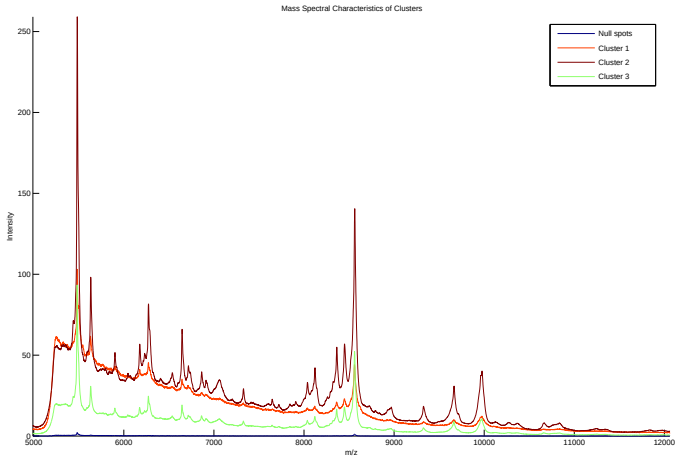


Fig. 11. Comparison plot of the average mass spectra that characterize the tree clusters (and the null spot region) of the second best overall segmentation result (NSR+PCA (2 PCs), with *k*-means using the cosine distance measure and preset to 3 clusters - avg. *silhouette* is 0.8132). The corresponding segmentation in image space is shown in Fig. 10 in the top-right corner.

average *silhouette*, each one of them can be considered a good result.

5.5 Mass Spectral Comparison

For the mass spectral comparison, we choose from the four best performing candidates overall one interesting segmentation to investigate : NSR+PCA (2 PCs), with *k*-means using the cosine distance measure and preset to 3 clusters (avg. *silhouette*: 0.8132) (top-right segmentation in Fig. 10). Fig. 11 shows a comparison plot of the representative mass spectra[2] of each zone.

Relevant differences between cluster 2 (dark red) and clusters 1 (orange) and 3 (light green) can be attributed to the variation in the peak intensities at $5483\ m/z$ and $8563\ m/z$. These peaks can act as discriminatory features that segregate the central tissue zone (grey matter, see section 6) from other zones surrounding it. The differences between clusters 1 and 3 are less blatant at the mentioned peaks. However, relative intensity differences in the m/z-region between the two main peaks for clusters 1 and 3 can still distinguish one zone from the other.

In general, for the spinal cord section under study, we notice that the differences between the segmented chemical zones is mainly of a quantitative nature, rather than of a qualitative type. This means that the tissue zones can be discriminated on the basis of the peak intensity amounts in certain m/z-windows, rather than on the basis of the presence or absence of particular peaks. This quantitative comparison of peaks is valid in the context of an IMS experiment, due to the fact that for all the spots the environmental conditions are essentially the same, they undergo the same chemical treatment, they are placed on the same target plate, and are measured using the same mass spectrometer and settings.

6 BIOLOGICAL INTERPRETATION

Although the clustering results are only meant as a first stepping stone towards further investigation using narrower mass ranges and selected spatial areas, we submitted the final segmentation and PCA results to researchers with experience in neurobiology and rat spinal cord sections (be it via immunostaining and microscopy). They informed us that, although there is no direct verification of these interpretations as the experimental setup is purely unsupervised and not a classification problem, some of the partitioning seems to differentiate between white matter (green in Fig. 10 and Fig. 11) and grey matter (dark red in Fig. 10 and Fig. 11) in the nerve tissue. One can also discern that the grey matter has a distinct butterfly-shape which corresponds with a region, labeled the *Substantia grisea*, which shows up consistently in microscopic images as well (see Fig. 4). Another interesting aspect is the fact that the third principle component, which still accounts for 2.2% of the information, seems to differentiate between the dorsal and ventral part of the spinal cord. Such differences are in line with the notion that these regions contain nerve cells involved in different functions. The ventral horn of the spinal cord is the location of the large motor neurons that innervate skeletal muscles, while the dorsal horn contains a lot of sensory neurons, which amongst others handle pain-related signals.

[2] All mass spectra that we discuss are in their discretized peaklisted form. However, for reasons of clarity we have plotted the mass spectra in Fig.11 using a continuous line connecting the peak tops, instead of the usual peak bar style.

Although the cell composition of both regions will be very similar, an immediate consequence of the functional difference is that at least some of the molecular machinery (e.g neurotransmitters) will be different . This could concur with the 2.2% difference.

7 CONCLUSION

This study has demonstrated that it is possible to build an entirely unsupervised workflow, capable of delivering relevant multivariate chemical tissue slice summaries on the basis of IMS data. These multivariate images can serve as an initial, high-throughput tool for exploring the high-dimensional data that is generated by an IMS experiment.

The computational steps performed on the rat spinal cord data use common algorithms and measures for each individual phase and they are not customized towards IMS or mass spectral data in general. However, they still are able to give satisfactory results, and the interesting links with regards to biological interpretation strenghten the idea that some form of multivariate analysis should preceed univariate approaches such as fragment images when no *a priori* information about the tissue sample is known.

ACKNOWLEDGEMENTS

Raf Van de Plas is a research assistant of the IWT at the Katholieke Universiteit Leuven, Belgium. Fabian Ojeda is a research assistant at the Katholieke Universiteit Leuven, Belgium. Maarten Dewil is a research assistant of the FWO at the Katholieke Universiteit Leuven, Belgium. Ludo Van Den Bosch is assistant professor at the Katholieke Universiteit Leuven, Belgium. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Etienne Waelkens is a full professor at the Katholieke Universiteit Leuven, Belgium.

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: - FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), research communities (ICCoS, ANMMM, MLDM); - IWT: PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope, Silicos; Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006) ; EU-RTD: FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain

REFERENCES

Aebersold,R., Mann,M. (2003) Mass spectrometry-based proteomics, *Nature*, **422(6928)**, 198-207.

Huh,W.-K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S., O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast, *Nature*, **425(6959)**, 686-691.

Jain,A.K., Dubes,R.C. (1988) Algorithms for Clustering Data, *Prentice Hall*.

Joliffe,I.T. (1986) Principal Component Analysis, *Springer-Verlag*.

Kanninen,K., Goldsteins,G., Auriola,S., Alafuzoff,I., Koistinaho,J. (2004) Glycosylation changes in Alzheimer's disease as revealed by a proteomic approach, *Neurosci Lett*, **367(2)**, 235-240.

McCombie,G., Staab,D., Stoeckli,M., Knochenmuss,R. (2005) Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis, *Anal Chem*, **77(19)**, 6118-24.

Nygren,H., Malmberg,P., Kriegeskotte,C., Arlinghaus,H.F. (2004) Bioimaging TOF-SIMS: localization of cholesterol in rat kidney sections, *FEBS Lett*, **566(1-3)**, 291-293.

Rousseeuw,P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *J Comput Appl Math*, **20**, 53–65.

Stoeckli,M., Chaurand,P., Hallahan,D.E., Caprioli,R.M. (2001) Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues, *Nat Med*, **7(4)**, 493-496.

Stoeckli,M., Farmer,T.B., Caprioli,R.M. (1999) Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument, *J Am Soc Mass Spectrom*, **10(1)**, 67-71.