PROSPECTIVE EXPLORATION OF BIOCHEMICAL TISSUE COMPOSITION VIA IMAGING MASS SPECTROMETRY GUIDED BY PRINCIPAL COMPONENT ANALYSIS

Raf Van de Plas[1,3] †, Fabian Ojeda[1,3], Maarten Dewil[5], Ludo Van Den Bosch[5], Bart De Moor[1,3] and Etienne Waelkens[2,3,4]

[1] *Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium.*

[2] *Katholieke Universiteit Leuven, Department of Molecular Cell Biology, Afd. Biochemie, O & N, Herestraat 49 - bus 901, B-3000 Leuven, Belgium.*

[3] *Katholieke Universiteit Leuven, ProMeta, Interfaculty Centre for Proteomics and Metabolomics, O & N 2, Herestraat 49, B-3000 Leuven, Belgium.*

[4] *Katholieke Universiteit Leuven, BioMacS, Interfaculty Centre for Biomacromolecular Structure, IRC KUL Campus Kortrijk, E. Sabbelaan 53, B-8500 Kortrijk, Belgium.*

[5] *Katholieke Universiteit Leuven, Department of Neurosciences, Neurobiology, O & N 2, Herestraat 49, B-3000 Leuven, Belgium.*

MALDI-based Imaging Mass Spectrometry (IMS) is an analytical technique that provides the opportunity to study the spatial distribution of biomolecules including proteins and peptides in organic tissue. IMS measures a large collection of mass spectra spread out over an organic tissue section and retains the absolute spatial location of these measurements for analysis and imaging. The classical approach to IMS imaging, producing univariate ion images, is not well suited as a first step in a prospective study where no *a priori* molecular target mass can be formulated. The main reasons for this are the size and the multivariate nature of IMS data. In this paper we describe the use of principal component analysis as a multivariate pre-analysis tool, to identify the major spatial and mass-related trends in the data and to guide further analysis downstream. First, a conceptual overview of principal component analysis for IMS is given. Then, we demonstrate the approach on an IMS data set collected from a transversal section of the spinal cord of a standard control rat.

Keywords: principal component analysis; imaging mass spectrometry; proteomics; bioinformatics; rat nerve tissue.

†To whom correspondence should be addressed: `raf.vandeplas@esat.kuleuven.be`

1. Introduction

Mass spectrometry allows one to very accurately measure the molecular masses found in an unknown sample. It has become one of the primary analytical instruments in proteomics and peptidomics research, which is the study of respectively proteins and peptides within the scope of an organism, tissue, cell, or organel and under a set of known physiological and environmental conditions.[1] Most proteomics and peptidomics studies, however, disregard the exact spatial origin of a sample within tissue, focusing solely on identification and quantitation. A number of studies[2–6] have demonstrated that incorporating spatial information into the analysis can provide further insight into biological processes.

The study of the spatial distribution of biomolecules in organic tissue requires that an explicit link is preserved between proteomics/peptidomics-oriented mass spectral measurements and their exact spatial origin within an organic tissue section. For this purpose we employ a relatively new technology, termed laser-based or MALDI-based imaging mass spectrometry.

1.1. *MALDI-based Imaging Mass Spectrometry*

MALDI-based Imaging Mass Spectrometry[‡‡] (IMS) is a technology that uses the molecular specificity and sensitivity of normal mass spectrometry to collect a direct spatial mapping of biomolecules (or rather their ions) in tissue sections. It allows for massive multiplexing of followed molecules (covering an entire mass range) and does not require complex chemistry or an *a priori* target molecule as is the case with complementary technologies such as immunochemistry and fluorescence microscopy. IMS has been succesfully used in a number of pioneering studies that mainly focused on mammalian tissue.[3,4]

The wet-lab side of the procedure consists of cutting an organic tissue section, mounting it on a MALDI target plate, applying an appropriate chemical matrix solution, and performing a MALDI mass spectral measurement at 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 a schematic overview of the wet-lab and *in silico* steps involved with performing IMS on the cross-section of spinal cord nerve tissue. Typically, the data generated by an IMS experiment populates a mathematical space that has two spatial dimensions (the x and y-dimension) and the mass-over-charge dimension (m/z).

[‡‡]MALDI stands for 'matrix-assisted laser desorption ionization' and refers to a particular mass spectrometry ionization method which is well suited for the study of larger biomolecules such as proteins. It involves firing a controlled laser shot at the sample embedded in a crystalline chemical matrix solution on the target plate.

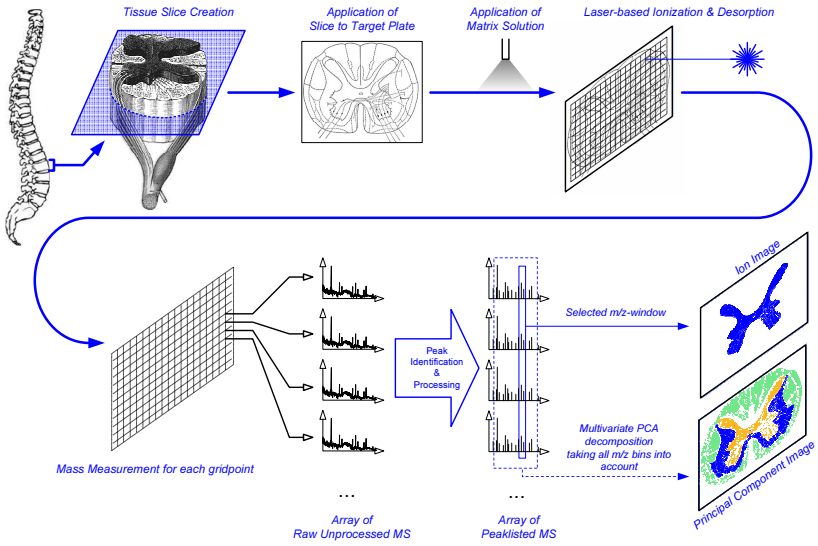

Fig. 1. Overview of the imaging mass spectrometry experiment. The tissue slice creation, the mounting on the target plate, and the application of an appropriate matrix solution are wet-lab steps. The mass spectral measurements, the data collection, and the translation into peaklisted mass spectra take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

It can be represented as a three-way array or tensor, with an x-, a y-, and a m/z-dimension.

1.2. *Ion Images*

A common approach taken in IMS-oriented studies[3–5] is to generate ion images from the IMS data tensor. These images are a false color visualisation of the spatial distribution of peak height for a particular m/z-window. They are called ion images because they show the spatial spread of a particular peak's height over the tissue, and because a mass spectral peak represents the amount of a particular ion that was measured. This ion can be the molecular ion, or a charged fragment of the original molecule. From a mathematical standpoint, an ion image can be seen as a cross-section of the data tensor at a particular mass (or m/z). Four examples of such ion images are shown in Fig. 2, 3, 4, and 5, which were generated from the data set of rat spinal cord tissue which is further discussed in section 2.3.

Ion images are a univariate approach to IMS imaging where one particular feature per pixel is picked for analysis and visualisation. This is very informative when the goal is to follow the spatial distribution of a particular molecule and you know beforehand which particular m/z-value is relevant to the study.

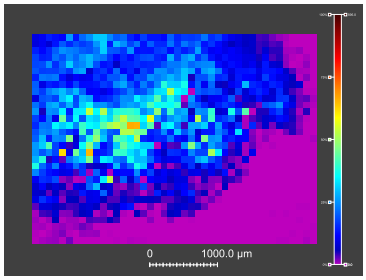

Fig. 2. Ion image at m/z 5490.52 from the rat spinal cord data set.

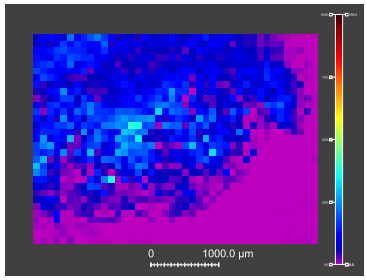

Fig. 3. Ion image at m/z 5634.79 from the rat spinal cord data set.

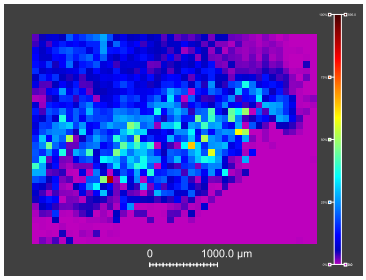

Fig. 4. Ion image at m/z 8565.30 from the rat spinal cord data set.

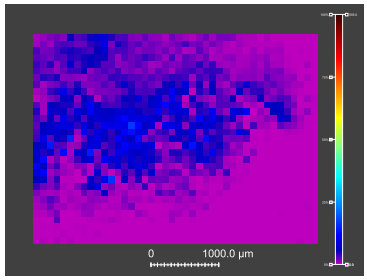

Fig. 5. Ion image at m/z 9974.26 from the rat spinal cord data set.

However, ion images are much less suited for propective studies where no *a priori* hypothesis of a target molecule or mass is formulated. In this kind of high-throughput discovery use of IMS one can potentially extract from the IMS tensor as many different ion images as there are m/z-bins available, and this number can easily run into the thousands (depending on the extent of the mass range that was scanned by the mass spectrometer). As an example, in our rat spinal cord data set this means 7451 distinct ion images of which just four are shown in Fig. 2, 3, 4, and 5. Acquiring an overview and identifying the ion images that show meaningful spatial variation from this set of thousands is a nontrivial task, and does not lend itself well to human execution. This is why we employ multivariate data analysis methods, such as the principal component analysis discussed in this paper, to perform a preliminary exploration of the data tensor in order to identify spatial and mass trends that merit further investigation. The insights delivered by such a preliminary multivariate analysis can serve as a guide for further investigation using more traditional approaches such as the ion images. As shown in section 2.3, the PCA-results can even be used to discriminate between biologically relevant chemical zones in the tissue on the basis of their mass spectral footprint.

2. Principal Component Analysis (PCA) as a prospective guide to IMS data

In this section we investigate the use of principal component analysis (PCA) as a guide for the prospective exploration of data coming out of an IMS experiment. The goal is to use the PCA results as a first stepping stone towards more elaborate multivariate analysis of IMS data. In section 2.1 we first discuss the general idea behind the PCA technique, followed by a treatise on the specific use of PCA in an IMS context in section 2.2. In the case study discussed in section 2.3, we apply the technique to real data from an IMS experiment where a rat nerve tissue section was imaged.

2.1. *Principal Component Analysis*

Principal component analysis is a latent variable* data analysis technique, widely employed in many areas for uses such as dimensionality reduction.[7] It was mentioned in a MALDI-IMS context by McCombie *et al.*[8] for the purpose of dimensionality reduction and denoising. In this study the method is used for trend detection in both the mass and the image domain.

Before formulating a definition of PCA, it is necessary to explain some aspects of the concept of the rank of a matrix. The rank of a matrix M is the maximum number of linearly independent rows (or columns) of M. This means that the rank of M is the smallest number of outer products of vectors that can be used to reproduce the matrix M exactly. Another definition is that the rank of M is equal to the number of nonzero eigenvalues of $M^T M$. A matrix of rank 1 can therefore be completely represented as the outer product of two vectors, while a matrix of rank 5 requires the sum of (at least) five such outer products for it to be completely reconstructed.

PCA is a decomposition of a matrix X, of size $N \times K$ and with a certain rank, into matrices of rank 1, designated F_a:

$$X = \sum_{a=1}^{A} F_a. \tag{1}$$

Given the definition of rank, it is evident that the smallest value of A for which this equation still holds is equal to the rank of the matrix X. The matrices F_a have the same size as X $(N \times K)$, but as they are rank 1 they can be replaced by the outer product of two vectors s_a $(N \times 1)$ and l_a $(K \times 1)$ in equation 1:

$$X = \sum_{a=1}^{A} F_a = \sum_{a=1}^{A} s_a l_a^T = SL^T. \tag{2}$$

*A latent variable is a variable which we do not observe directly, but its existence can be inferred from the observed variables.

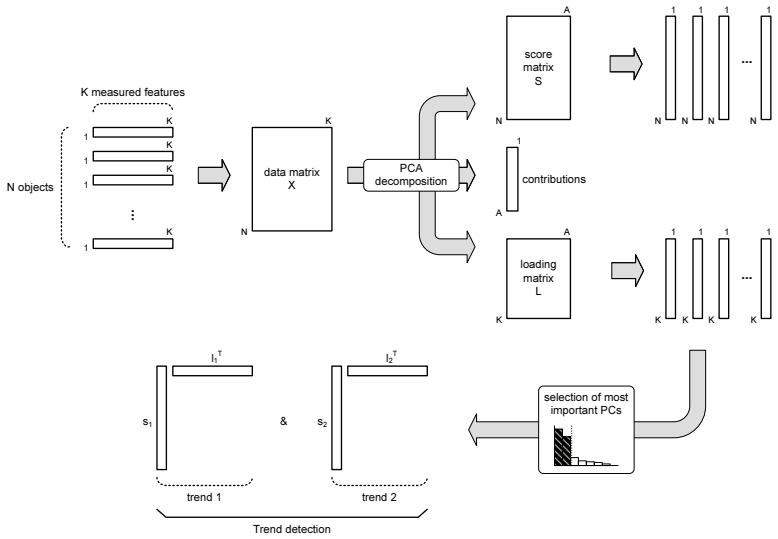

Fig. 6. Graphical representation of matrix decomposition and reconstruction using principal component analysis. The upper half shows the decomposition of data matrix X into A principal components, or the outer product of score matrix S (grouping score vectors s_1 to s_A) and loading matrix L (grouping loading vectors l_1 to l_A). The lower half depicts selecting the principal components with the largest contribution of variance (or information) in order to study the uncorrelated trends found in X.

The vectors** s_a are generally called the score vectors of the decomposition, while vectors l_a are named loading vectors. Each pair of vectors s_a and l_a can be designated a principal component of matrix X and has a particular coefficient connected to it (stemming from the eigenvalue of the underlying $X^T X$), which indicates the relative amount of variance of X explained by its particular principal component. By further utilizing matrix notation, PCA can be written more consicely as the decomposition of the data matrix X into the single product of a matrix S of size $N \times A$, holding all the score vectors, and a matrix L of size $K \times A$, holding all the loading vectors. In order to complete the decomposition with a minimum value for A, the matrices F_a and their composing vectors s_a and l_a are necessarily maximally uncorrelated. A schematic overview of these steps is available in Fig. 6. Using the coefficients (actually eigenvalues) connected to them, one can order the various principal components according to their contribution in terms of variance of X explained and information contained. A number of uses follow from this notion such as dimensionality reduction and denoising,[7] but we focus specifically on the

**In this paper we follow the convention of representing a vector as a column vector unless explicitly transposed.

use of trend detection, in which the principal components with the largest contribution characterize the major uncorrelated trends underlying the data.

2.2. *PCA Applied to IMS Data*

In applying PCA to IMS data, our primary goal is to identify the major uncorrelated trends that can be found in the spatial domain (x and y) as well as in the mass domain (m/z). These trends, which tell us which pixels or m/z-bins behave similarly or dissimilarly, can be used as a guide in exploring these often complex and very large data sets, and to avoid the proverbial 'drowning in information' that can be experienced when classical ion images are employed without a prior hypthesis of target mass.

As mentioned in section 1, the data measured during an IMS experiment can be stored as an array of order 3, or tensor, $\overline{D}$ with two spatial dimensions (x and y) and one m/z dimension (cfr. the abscissa in a mass spectrum). Each scalar value d_{ijk} in the tensor represents the absolute intensity of a particular mass peak at a certain x-position i, a certain y-position j, and measured at a certain m/z-bin k (with $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$).

One way of applying PCA to an IMS tensor $\overline{D}$ is to refold the tensor into an array of order 2, or matrix, D to fit the expression shown in equation 2. This refolding process is done by reordering all discrete spatial positions in the x- and y-dimensions, or 'pixels' if you will, into one long vector holding $I.J$ elements. The result is a matrix D of size $(I.J) \times K$, holding all information contained within the original tensor $\overline{D}$. Applying PCA in the manner discussed in section 2.1 delivers a score matrix S of size $(I.J) \times A$, a loading matrix L of size $K \times A$, and a vector of eigenvalues λ indicating each principal component's variance contribution.

Based on the amount of variance explained, we can now identify and take a closer look at the most important principal components. A single principal component is characterized by one score vector or one loading vector. The score vector is of size $(I.J) \times 1$ and does not easily allow for direct exploration. However, a reordering operation that reverses the effect of the operation performed to go from $\overline{D}$ to D allows us to refold this vector of size $(I.J) \times 1$ to the image space defined by the two spatial dimensions x and y, resulting in an image matrix of size $I \times J$. In their image form the score vectors deliver a more human-interpretable view on the underlying spatial correlations. This type of images gives us an idea of which pixels, or laser spots, have a similar mass spectral footprint when all m/z-bins are taken into account (note the difference with univariate ion images). The corresponding loading vector of size $K \times 1$ does not require a refolding operation as it can be expressed directly in the m/z-domain. A visualisation of the loading vector gives us an indication of which m/z-bins behave similarly within the context of one principal component.

It is necessary to mention here that the above linkup of score vectors with the image domain and loading vectors with the mass domain is based on the assumption that PCA is performed on a data matrix where the rows represent pixels and the columns represent m/z-bins. This assumption would be in line with the convention of an objects $\times$ features data matrix used in most PCA literature. However, when there are more features available than objects, which is usually the case for IMS (e.g. 7451 versus 1302 in the spinal cord data set), it is more computationally efficient to use the transpose of D instead.[7] The results of this more 'economic' PCA are identical to the ones from the procedure described earlier, with the only difference being that the loading vectors are now linked to the image domain and the score vectors to the mass domain. This economized PCA was used in the case study of section 2.3.

2.3. *Case Study: Rat Spinal Cord Nerve Tissue*

In this section we demonstrate the use of PCA in an IMS context by applying it to the IMS measurement of rat nerve tissue. For reference, Fig. 7 shows a microscopic image of a nerve tissue section taken from the same animal as the one used in this case study. Figure 7's tissue slice has undergone histological staining to bring out the gray/white tissue differentiation which is not visually apparent in untreated samples, but which does show up in the PCA analysis performed in this section.

Materials and methods The tissue section (15 micrometers thick) was taken from a transversal section of the spinal cord of a standard control rat. The recorded mass range extended from m/z 5000 to 12000 and alpha-cyano-4-hydroxy cinnamic acid (7 mg/ml, in acetonitrile 50%, 0.05% TFA) was used as a chemical matrix. A MALDI mass spectral measurement was performed on each grid point of a virtual raster of size 31×42 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 in linear mode. The data collection in the mass spectrometer was guided by the *4000 Series Imaging* module, available at `http://www.maldi-msi.org`. Processing was done using in-house developed software.

Preprocessing As Fig. 2, 3, 4, and 5 show, the IMS raster was slightly off center with regards to the tissue section, resulting in a tissue-free area in the bottom right corner (shown in purple). To avoid these empty measurements consuming variance and influencing the PCA-results, we disregarded them when their total ion current fell below a 10% threshold.

Analysis Results We applied PCA via singular value decomposition of the covariance matrix of the data matrix, using the economized version of PCA

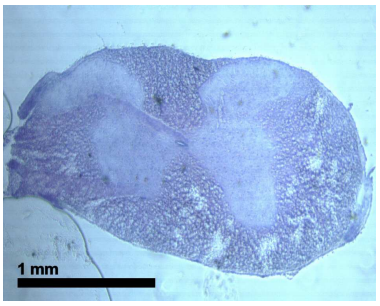

Fig. 7. Microscopic image of a transversal section of rat spinal cord, histologically stained to show the butterfly-shaped central area known as the *Substantia grisea* (grey matter), surrounded by white matter nerve tissue.

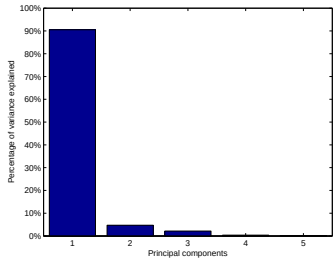

Fig. 8. Percentage of variance explained by each principal component (PC). This graph only shows PCs with a contribution larger than 0.1%. The total number of PCs is actually given by $\min(N, K)$, but most have a negligible or zero contribution.

mentioned in section 2.2. This means that the loading vectors are represented in the image domain, while the score vectors are shown in the mass domain. When interpreting these visualisations, the relative differences in value are important, not the absolute value or sign (e.g. see first score vector in Fig. 10). In the image domain (Fig. 9) low valued areas in blue are discriminated from high valued pixels in red, indicating zones within which the mass spectral footprint (and the underlying chemical composition) is similar. In the mass domain (Fig.10) m/z-bins carrying similar values correlate strongly in behavior across the tissue (the peaks vary together), and can be discriminated from bins with dissimilar values.

The bar plot in Fig. 8 shows us the relative amount of information contained in each principal component (PC) (above a 0.1% cut-off). It is apparent that the first PC is very prominent with more than 90% variance explained. This means that the spatial and mass-related correlations connected to the first PC can be considered as the primary structure found in the chemical composition of the tissue slice. Secondary and tertiary uncorrelated trends are also apparent as the second and third PC still hold a non-negligible amount of information. However, from the fourth PC onwards the contributions become less influential, tending towards noise in the data. Therefore, we will focus on the first three loading and score vectors shown in Fig. 9 and 10. The strong reduction in complexity indicates a large amount of correlation in the spatial domain (indicating region formation) and the mass domain (indicating ions, or m/z-bins, behaving similarly; e.g. by coming from the same parent molecule).

The primary trend, characterized by the first loading vector in Fig. 9 and the first score vector in Fig. 10, shows that a butterfly-shaped region in the cen-

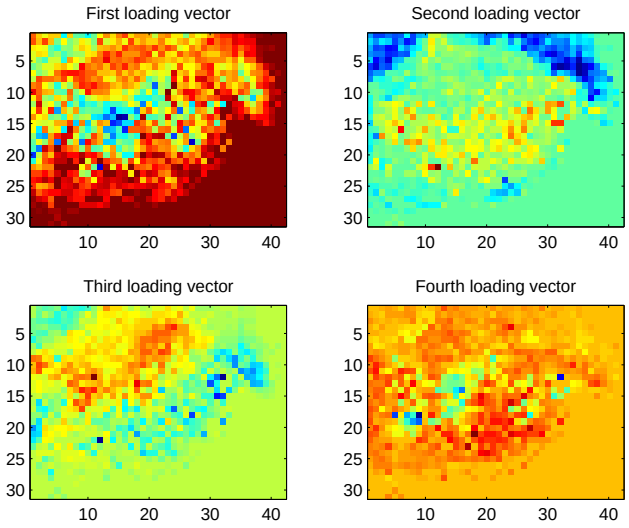

Fig. 9. The first four loading vectors. Folded back into the image domain, they show correlations at the pixel level.

ter of the tissue has a dissimilar chemical composition from the areas surrounding it (blue vs. red). This area correlates strongly in location and shape with the anatomical region called the *Substantia grisea* (grey matter), surrounded by white matter nerve tissue (also visible in Fig. 7). The first score vector shows that all m/z-bins have negative values with differing relative amounts, indicating that the spatial discrimination between the grey and white matter areas is mainly explained through quantitative differences in the chemistry, rather than qualitative ions showing up or disappearing. Also notice the two characteristic peaks at m/z 5484 and 8564, that show up consistently across the nerve tissue but whose relative quantity can be employed as a mass marker for grey matter.

The second trend differentiates the blue region of tissue at the top of the raster from the red/yellow area at the center. When studying the second score vector it becomes clear that the differences between these areas are mainly caused by the dense peak area between m/z 5000 and m/z 8000 showing up more prominently while the peaks at m/z 5484 and 8564 lose intensity. One has to bear in mind that this secondary trend only accounts for some 4% of the chemical variation across the slice.

With a contribution coefficient of 1 to 2%, the third loading vector differentiates between a ventral and dorsal area in the tissue. This third trend

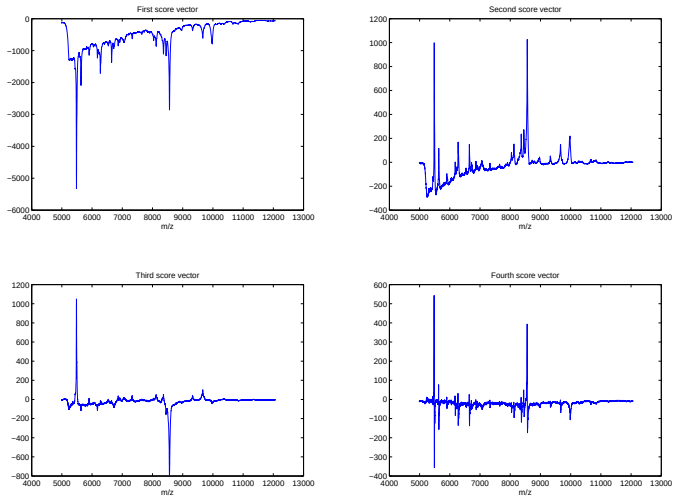

Fig. 10. The first four score vectors. Represented in the mass domain, they show correlations between m/z-bins.

correlates strongly with the biochemical differences in ventral/dorsal cellular composition of the spinal cord. In the mass domain of the third score vector we see that this ventral/dorsal difference is mainly due to the 5484-peak showing up and the 8564-peak diminishing in one area while the reverse happens in the other area. In the primary and secondary trends the differences were mainly quantitative in nature. However, in this third trend we see an example of a more qualitative difference with the presence of a particular ion characterizing a particular area in the tissue. The fourth loading and score vectors are shown for completeness, but it is evident that spatially the correlations are less localized and structured, tending towards spread-out noise. In the mass domain we see differentiation between m/z-bins which are close together. This is rather unlikely to be structured given that there are isotopic and other ties between m/z-values this close together, which further seems to indicate that from this trend onwards we are dealing with modeled noise.

In summary, the PCA-results tell us that in this particular IMS data set the chemical composition is dominated by the difference between grey matter nerve tissue and white matter, and two quantitative ion markers for these areas are observed at m/z 5484 and 8564. In addition to that, a ventral/dorsal difference was measured which can be related to known ventral/dorsal differences in the spinal cord.

3. Conclusions

We described a procedure for using PCA in an IMS context as an intrument for guiding prospective analysis of the chemical tissue composition. In the spatial domain it can show the human observer which regions have a particular mass spectral footprint, and it can differentiate these from other areas in the tissue slice without the need to perform invasive chemistry on the sample as is the case with e.g. histological staining. In the mass domain, specific molecular masses responsible for these differences (in m/z-form) are identified, and lend themselves for further downstream analysis using, for example, ion images. The case study on rat nerve tissue demonstrated these uses by delineating grey matter from white matter and by identifying two mass markers that can be used to differentiate between these zones. It also illustrates how, in addition to the visual aspect of differentiating zones in the tissue, IMS as a technology permits a direct measurement of the chemical reality responsible for these differing areas, in the form of molecular masses.

The use of PCA as described in this paper is but a first step towards a more insightful interrogation of IMS data. A thorough investigation of the influence of factors such as preprocessing of the data and robustness of the method will be required before it can be established as a firm first analysis step. Also, comparisons with other multivariate techniques, such as independent component analysis, are currently under way and will prove to be an interesting research avenue.

4. Acknowledgements

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO , several PhD/postdoc & fellow grants; Flemish Government: - FWO: PhD/postdoc grants, projects G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, G.0129.00, research communities (ICCoS, ANMMM, MLDM); - IWT: PhD Grants, GBOU-McKnow-E, GBOU-SQUAD, GBOU-ANA, TAD-BioScope-IT, Silicos; Belgian Federal Science Policy Office: IUAP P5/22; EU-RTD: ERNSI; FP6-NoE; FP6-IP, FP6-MC-EST; ProMeta, BioMacS.

References

1. R. Aebersold and M. Mann, *Nature* **422**, 198(Mar 2003).
2. W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman and E. K. O'Shea, *Nature* **425**, 686(Oct 2003).
3. M. Stoeckli, T. B. Farmer and R. M. Caprioli, *J Am Soc Mass Spectrom* **10**, 67(Jan 1999).
4. M. Stoeckli, P. Chaurand, D. E. Hallahan and R. M. Caprioli, *Nat Med* **7**, 493(Apr 2001).
5. H. Nygren, P. Malmberg, C. Kriegeskotte and H. F. Arlinghaus, *FEBS Lett* **566**, 291(May 2004).
6. R. M. A. Heeren, *Proteomics* **5**, 4316(Nov 2005).
7. I. T. Joliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
8. G. McCombie, D. Staab, M. Stoeckli and R. Knochenmuss, *Anal Chem* **77**, 6118(Oct 2005).