Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA

Raf Van de Plas[1,3,*] , 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, Sec. Biochemistry
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.

***Abstract*— Imaging mass spectrometry or mass spectral imaging (MSI) is a technology that provides us with the opportunity to study the spatial distribution of biomolecules such as proteins, peptides, and metabolites throughout organic tissue sections. MSI adds a spatial dimension to mass spectrometry and biomarker-oriented studies without the requirement for labels, as is the case with more traditional techniques such as fluorescense microscopy. It has particular merit for studies where no prior hypothesis of target molecules is available, as it can simultaneously track a wide range of molecules within its mass range. This makes MSI a potent exploratory tool for elucidating the spatiobiochemical topology in tissue.**

This paper elaborates on the principal component analysis (PCA)-based unsupervised decomposition of an MSI-measured organic tissue section into its underlying biochemical trends. We introduce a method to control the weight that particular peak intensity ranges are allowed to exert on the final decomposition model. The extension provides a way for peak intensity-based scaling to be incorporated directly into the decomposition process, for the purpose of denoising or contrast enhancement. The method makes use of peak height transformations that are conceptually equivalent to what is known in digital image processing as gray level transformations, but rather than aiming to enhance contrast for human interpretation they are used to influence the unsupervised decomposition process. As an example, we apply a combined denoising/contrast stretching measure to the MSI-measurement of a section of rat spinal cord.

I. INTRODUCTION

Mass spectrometry has become the primary analytical method for most proteomics, peptidomics, and metabolomics-oriented research [1]. In most studies, however, the spatial origin of a sample within the tissue is not taken into account. A growing body of research [3], [9], [10] is demonstrating that adding spatial information to the analysis can provide deeper insight into the biological processes under study.

In order to study the spatial distribution of biomolecules in organic tissue, an explicit link has to be preserved between the mass spectral measurements holding the biochemical information, and their exact spatial origin within the tissue. For this purpose we employ a technology termed imaging mass spectrometry or mass spectral imaging (MSI). MSI entails imaging based on secondary ion mass spectrometry (SIMS) as well as matrix-assisted laser desorption/ionization (MALDI[1]) mass spectrometry [8]. As we are primarily interested in the localization of biomacromolecules, this paper will focus specifically on MALDI-based MSI.

A. MALDI-based Imaging Mass Spectrometry

MALDI-based imaging mass spectrometry [10] uses the molecular specificity and sensitivity of normal mass spectrometry to collect a direct spatial mapping of biomolecules (or rather their ions) from a tissue section. No complex chemistry or an *a priori* target molecule is required as with complementary technologies such as immunochemistry and fluorescence microscopy. A good example of its use for biomarker discovery is found in Meistermann *et al.* [9].

Fig. 1 gives a quick overview of the wet-lab steps involved in performing an MSI experiment and a more thorough treatment is also available in Van de Plas *et al.* [12]. The result of an MSI experiment consists of an array of spots or 'pixels' covering the tissue section, where every pixel has an individual mass spectrum connected to it. The measurements are typically encoded as a 3-mode array with two spatial modes (x and y) and one mass-over-charge mode[2] (m/z).

B. Ion Images and Multivariate Images

A common use of MSI data is to study the spatial spread of one particular mass (or m/z bin) in the form of an ion image. An example ion image can be found in Fig. 3 which shows the distribution of ion m/z 5490.52 across the rat spinal cord tissue section discussed in section II. However, due to the massive amount of simultaneous measurements available in an MSI data set, more elaborate exploratory data analysis techniques show considerable promise as well. These multivariate techniques [7], [5], [12] use clustering and matrix decomposition to learn the major biochemical trends underlying the data, grouping masses and spatial areas on the basis of similar behavior. These trends allow for a higher-level model of the tissue composition to be constructed, which in turn allows for more elaborate biological questions to be asked (e.g. biomarkers for certain areas, testing co-location-related hypotheses,...).

* Corresponding author: raf.vandeplas@esat.kuleuven.be

[1]MALDI 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.

[2]When MALDI ionization is used the charge z of an ion is usually $+1$.

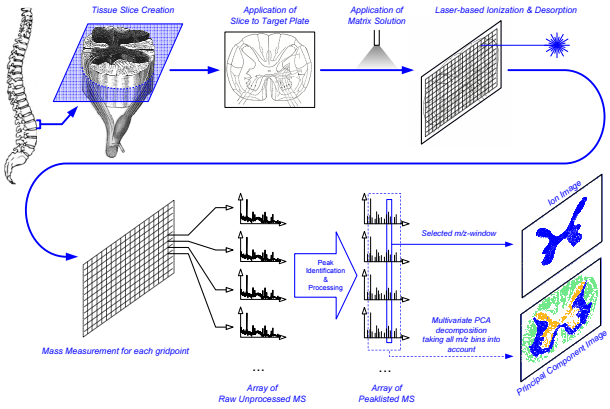

Fig. 1. Schematic overview of the imaging mass spectrometry experiment. Obtaining of a tissue section using a microtome, mounting the section on a target plate, and applying an appropriate chemical matrix solution to enable ionization take place in the wet lab. The mass spectral measurements, the data collection, and any low-level processing of the mass spectra take place inside the mass spectrometer. The resulting array of mass spectra can then be processed *in silico* by a data analysis method such as the peak intensity weighted PCA discussed in this paper.

II. CASE STUDY - MATERIALS & METHODS

The examples shown in the coming sections are based on an MSI measurement of rat spinal cord nerve tissue (Fig. 2). 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 solution. 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.

As Fig. 3 shows, 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.

III. STANDARD PCA-BASED DECOMPOSITION

Principal component analysis (PCA) is a data analysis technique that decomposes a data set into a reduced set of uncorrelated signals for the purpose of dimensionality reduction or trend detection [4]. In an imaging mass spectrometry context its use translates to an unraveling of the MSI data into a set of principal components, which can be interpreted as a set of uncorrelated biochemical trends in the tissue. The goal in applying a decomposition method such as PCA to MSI data is to identify zones of similar chemical composition in the tissue (along the x and y modes) and to get insight into the molecular masses (along the m/z mode) most responsible for these zones (correlated as well as anticorrelated). PCA is mentioned in a MALDI-MSI context by McCombie *et al.* [7] for the purpose of dimensionality reduction, and Van de Plas *et al.* [12] and Klerk *et al.* [5] describe its use for trend detection in tissue. Tyler *et al.* [11] discuss similar approaches for SIMS-based MSI.

A. PCA Applied to Imaging Mass Spectrometry

An MSI experiment typically delivers a 3-mode array or tensor $\overline{D}$ with two spatial modes (x and y) and one mass spectral mode (m/z). 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 $k = 1, \ldots, K$). As PCA requires a matrix rather than a tensor, $\overline{D}$ is refolded into a 2-mode array D of size $(I.J) \times K$ to perform the decomposition. A thorough discussion of this procedure can be found in Van de Plas *et al.* [12]. Applying PCA results in a relatively small set of principal components, each one characterized by a spatial signature indicating which zones differ markedly from other zones in the tissue. Additionally, each component is characterized by a mass spectral signature, which indicates the correlated molecular masses primarily responsible for these zones. Finally, each component has an eigenvalue assigned to it, indicating the amount of variance explained by that principal component.

B. Implicit Variable Weights

PCA is usually implemented as a singular value decomposition (SVD) of a square matrix, which represents the relationships between the variables in the data matrix D (M observations $\times$ N variables). Commonly this is the covariance matrix, but in some cases the correlation matrix is used as well. The difference lies in the relative influence that every variable is allowed to exert on the final principal component model. In the covariance matrix C each element c_{pq} represents the covariance between variables p and q (see 1). Calculating the covariance involves substraction of the mean per variable (mean-centering), but absolute deviations from the mean are left intact. As a result variables with large observations give large entries in the covariance matrix, and thus, exert more influence on the components.

$$c_{pq} = \frac{1}{M-1} \sum_{m=1}^{M} (d_{mp} - \overline{d}_p)(d_{mq} - \overline{d}_q) \qquad (1)$$

In the correlation matrix R, every element r_{pq} is further standardized by scaling all entries to a range between -1 and 1, dividing the covariance by the standard deviations of the variables p and q. The scaling has the effect of equalizing the weight that each variable contributes to the decomposition. Variations over small absolute values will become equally influential as large absolute variations. This makes sense when the data consists of heterogenous variables whose values do not have a direct physical relationship to each other. However, in the case of mass spectrometry the variables do have a physical relationship and are directly comparable in that they are all ion counts measured by the ion detector of the same mass spectrometer, only with a different m/z parameter for the mass analyzer. Therefore, it makes physical sense to assign large peaks more importance than smaller peaks, making the covariance matrix the preferred basis for decomposition (and the extension of section IV).

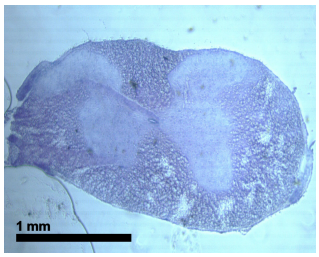

Fig. 2. Microscopic image of a transversal section of rat spinal cord (similar to the MSI section), histologically stained to show the butterfly-shaped central area known as the *Substantia grisea* (grey matter), surrounded by white matter.

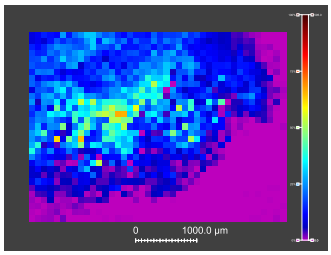

Fig. 3. Ion image at m/z 5490.52 from the rat spinal cord data set.

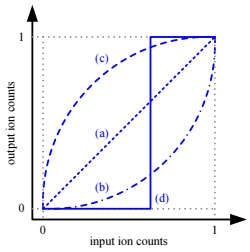

Fig. 4. Examples of ion count (or gray level) transformations. (a) Identity transformation. (b) nth Power transformation. (c) nth Root transformation. (d) Threshold transformation.

C. Case Study Results

Applying PCA based on the covariance matrix to the data set of section II results in a compact set of principal components (PCs). A full treatment of the results is given in [12], but for comparison with the peak intensity weighted PCA discussed in section IV, we show two of the most interesting trends in Fig. 6 and 7. Fig. 6 (top) shows the first PC responsible for 90% of the variance in the data, which delineates the butterfly-shaped region of gray matter in the center from the chemically different surrounding white matter (see also Fig. 2). Mass spectrally, the differences can be primarily contributed to the relative presence of m/z 5484 and 8564 that function as a marker for these regions. The second PC in Fig. 7 (top) differentiates the blue region at the top from the central area. The mass signature tell us that the difference between these areas is primarily the result of a consistently higher baseline showing up in the blue area, indicating that ionization took place more easily there.

IV. PEAK INTENSITY WEIGHTED PCA-BASED DECOMPOSITION

The method discussed in this section, aims to influence the relative importance connected to particular peak height ranges in the data. This can be useful towards, for example, dimishing the influence of noisy intensities or grouping together peak heights that fall within a certain range, as equally important.

The weighing is performed by scaling the data before the inherent weighing of the covariance matrix comes into play. The effect of the scaling will reflect in the covariance entries which will be decomposed by SVD as usual. Unlike the scaling performed to go from a covariance matrix to a correlation matrix, this scaling will take the form of a transformation of the histogram of peak heights found throughout the data set. Notice that this paper discusses weight assignment according to peak intensity, not according to spatial information or molecular mass considerations. However, an adaptation to take these considerations into account is straightforward.

A. Ion Count Transformations

An ion count histogram describes the distribution of peak heights throughout a collection of mass spectra. The type of scaling we are looking for, also uses it as the input for a transformation function that translates the original ion count d_{ijk} in tensor $\overline{D}$ to a modified ion count d'_{ijk}. Examples of such histograms can be found in Fig. 5. Output ion count d'_{ijk} is related to input ion count d_{ijk} by a transformation T such that $d'_{ijk} = T(d_{ijk})$. T can take many different forms, depending on the goal in mind.

An interesting observation is that this type of histogram manipulation is conceptually equivalent to an image enhancement technique called gray level transformations [2], often employed in digital image processing. In a similar way as with the ion counts, pixel intensities (or gray levels) are mapped to different intensity values with the intent of increasing contrast. In digital image processing this contrast enhancement is focused on improving human interpretability towards a certain purpose. In an MSI setting the contrast modification serves to make certain types of responses more 'visible' to the PCA decomposition. A link between digital image processing and MSI is also mentioned in [6].

Fig. 4 shows a number of example transformation curves. The most basic example of the family of linear transformations is the identity transformation shown as curve (a). It is a trivial transformation in that it leaves all ion counts the same. Another interesting family of curves is known as power-law transformations, which typically take the form of $d'_{ijk} = \alpha(d_{ijk})^{\beta}$ where α and β are constants. These include nth power transformations such as curve (b), which map a wide range of lower ion counts to a narrower range, while higher ion count levels are allowed to expand into a wider range of values. The inverse happens when nth root transformations (see curve (c)), which are also part of the power-law family, or a logarithmic transformation are applied. T can also be a piecewise linear function, such as the thresholding curve (d). Another good example of this type is shown in Fig. 5 (middle), where it is applied to the case study of section II. This type of mapping is known as a 'contrast stretching' transformation. The lower ion counts are compressed into a narrow ion count range close to zero, while higher ion counts that cross a certain level are compressed together at the high end of the range, equalizing their influence. The ion counts in between are spread out over a wider range, thinning that area of the histogram. In conclusion one can state that the transformation function T can take any form as long as it is monotonically increasing (in order to avoid artificially created intensity artefacts).

B. Case Study Results

As an example, we apply a contrast stretching transformation to the data set of section II (Fig. 5). It acts as a soft thresholding filter, diminishing the influence of lower ($<$

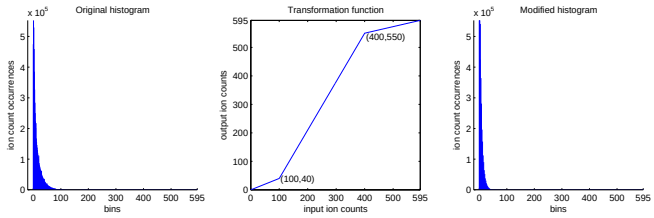

Fig. 5. Overview of the contrast stretching transformation performed on the spinal cord data set. (left) The original histogram of all ion counts in the data set. Given the predominating presence of low ion counts, this graph does not show the distribution of measurements at higher ion counts. However, the entire histogram is populated up to the maximum response in this experiment at 595 ion counts. (middle) The gray level transformation that was applied. It pushes lower ion counts (< 100) together in narrow range close to zero, reducing their impact on the decomposition. Higher ion counts (> 400) are compressed into a narrow range close to the maximum response, equalizing their importance in the decomposition. (right) The modified histogram showing the effect (only visible in the low ion count range) of the contrast stretching procedure.

100) ion counts, while ion counts above a certain threshold (> 400) get grouped together and are given a more equal weight. The chosen parameters are the result of an empirical assessment of the input ion count histogram. The automatic assignment of such parameters is an avenue of further research. An evaluation of the influence of the imposed weights for each component is given in the captions of Fig. 6 and 7. Overall, the results demonstrate a particular use of the weighing scheme as a means of denoising mass spectral data, without losing too much of the original information. The parameterized simplification demonstrated by the mass spectral signatures is an interesting example of this.

V. CONCLUSIONS

Interpretation of MSI data requires the consideration of noise and artifacts along the spatial, the mass-over-charge, and the peak intensity ranges. This paper focuses on the peak intensity range and describes a method that is complementary with measures taken along the other modes. It introduces an extension to standard PCA-based decomposition of MSI data, providing the researcher with a means of assigning weights according to signal intensity. The extension is implemented as a general ion count transformation, that can be customized towards a particular goal, such as grouping together peaks in a certain intensity range, attenuating low noise peaks, thresholding, or contrast enhancement for bringing out a particular aspect of the ion count distribution. Due to its monotonicity requirement, the transformation preserves the order of the peak heights and with it most of the information. As the transformation function can be specified by the researcher, the peak intensity weighted PCA allows for a more complex type of questions to be posed to MSI data.

ACKNOWLEDGMENTS

RVDP is a research assistant of the IWT at the Katholieke Universiteit Leuven, Belgium. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. EW is a full professor at the Katholieke Universiteit Leuven, Belgium.

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants; Flemish Government: - FWO: PhD/postdoc grants, projects G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, research communities (ICCoS, ANMMM, MLDM); - IWT: PhD Grants, GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos; SBO-BioFrame; Belgian Federal Science Policy Office: IUAP P6/25; EU-RTD: ERNSI; FP6-NoE; FP6-IP, FP6-MC-EST, FP6-STREP, ProMeta, BioMacS.

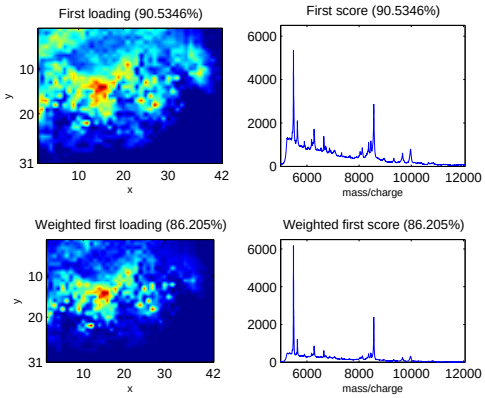

Fig. 6. (top) Unweighted first PC. (bottom) Peak intensity weighted first PC. By weighing down the smaller peaks we get a cleaner delineation of the *Substantia grisea* and a considerable simplification of the mass spectral signature. Particularly the top area, specifically selected by PC2 in Fig. 7, is prevented from causing a *bleed-through* via its many low ion count values.

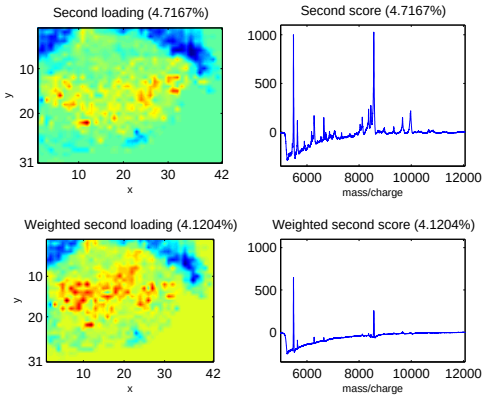

Fig. 7. (top) Unweighted second PC. (bottom) Peak intensity weighted second PC. Most apparent is the strong simplification of the mass spectral signature, which in its reduced form, according to the spatial images, still delineates the same zone quite accurately with just two of its original peaks. The smaller peaks that were attenuated in the weighted version turn out to influence the central area of the spinal cord.

REFERENCES

[1] R. Aebersold and M. Mann, Mass spectrometry-based proteomics, *Nature*, 422:6928, 2003, pp 198–207.

[2] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice Hall Inc. (2nd edition), Upper Saddle River, NJ; 2002.

[3] R.M.A. Heeren, Proteome imaging: a closer look at life's organization, *Proteomics*, 5:17, 2005, pp 4316–4326.

[4] I.T. Joliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[5] L.A. Klerk, A. Broersen, I.W. Fletcher, R. van Liere and R.M.A. Heeren, Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets, *Int J Mass Spectrom.*, 260:2–3, 2007, pp 222236.

[6] Y.C. Ling, M.T. Bernius and G.H. Morrison, SIMIPS: secondary ion mass image processing system, *J Chem Inf Comput Sci*, 27:2, 1987, pp 86–94.

[7] G. McCombie, D. Staab, M. Stoeckli and R. Knochenmuss, Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis, *Anal Chem.*, 77:19, 2005, pp 6118–6124.

[8] L.A. McDonnell and R.M.A. Heeren, Imaging mass spectrometry, *Mass Spectrom Rev*, 26:4, 2007, pp 606–643.

[9] H. Meistermann *et al.*, Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat, *Mol Cell Proteomics*, 5:10, 2006, pp 1876–1886.

[10] 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*, 7:4, 2001, pp 493–496.

[11] B.J. Tyler, G. Rayal and D.G. Castner, Multivariate analysis strategies for processing ToF-SIMS images of biomaterials, *Biomaterials*, 28:15, 2007, pp 2412–2423.

[12] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor and E. Waelkens, "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis," *in Proceedings of the Pacific Symposium on Biocomputing 12*, Maui, HI, 2007, pp. 458-469.