Spatial Querying of Imaging Mass Spectrometry Data for the Biochemical Characterization of Anatomical Regions in Tissue

Raf Van de Plas[1,3], Kristiaan Pelckmans[1], Bart De Moor[1,3], and Etienne Waelkens[2,3]

[1] Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium.
{raf.vandeplas, kristiaan.pelckmans, bart.demoor}@esat.kuleuven.be

[2] Katholieke Universiteit Leuven, Department of Molecular Cell Biology, Sec. Biochemistry, O & N, Herestraat 49 - bus 901, B-3000 Leuven, Belgium.
etienne.waelkens@med.kuleuven.be

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

Abstract. Imaging mass spectrometry or mass spectral imaging (MSI) is a technology that adds spatial information to mass spectral biochemical analysis. It delivers insight into the spatial distribution of biomolecules such as proteins, peptides, and metabolites throughout an organic tissue section. In this paper we develop methods that enable spatial querying of MSI data. The objective is to retrieve the ions (or masses) that are specific to a certain spatial area of interest. Such questions arise for example in pathomechanisms that show location-specific behavior (e.g. Parkinson's and Huntington's disease), the search for anatomical region-specific biomarkers, the study of local biochemical phenomena, and the addition of spatial information into biological models. We focus specifically on a multivariate approach in which the search for biochemical distributions that resemble the query image is formulated as a least squares optimization problem. By exploiting the positivity of ion counts, we are able to solve the problem efficiently using convex programming. Interpretable results are obtained by the typical occurrence of zeros in the resulting mass contributions. As a case study, we apply these methods to the MSI measurement of a sagittal section of mouse brain, and demonstrate how both synthetic area specifications as well as microscopic pictures can be used for spatial queries.

Key words: proteomics, bioinformatics, mass spectrometry, imaging, spatial query, least squares, convex optimization

1 Introduction

High-throughput proteomics, peptidomics and metabolomics provide a powerful approach to study the complex interactions of biomolecules such as proteins, peptides, and metabolites in biological systems. These analyses are primarily facilitated by the analytical technique of mass spectrometry [1], delivering a very accurate measurement of the molecular masses present in a given sample. Most mass spectrometry studies however disregard the exact spatial origin of a sample within tissue. This makes it difficult, if not impossible, to incorporate spatial aspects of biochemical phenomena into the models under study. A growing body of research [8, 11, 13] shows that adding spatial information to a biochemical analysis can significantly deepen the insight into biological pathways and mechanisms. Imaging mass spectrometry or mass spectral imaging (MSI) is a technique that preserves the link between a spatial tissue location and the biochemical characterization of what is found there. It delivers a view on the spatial behavior of molecular mass markers, which reflects its use in diagnostic settings and it can steer further investigation by exploiting MSI's high-throughput nature. Additionally, the mass markers can be further identified to known molecules using tandem mass spectrometry, enabling the incorporation of spatial aspects into network-type studies for systems biology. This paper focuses specifically on MALDI[4]-based MSI [10] as it offers the capability to study biomacromolecules such as proteins.

1.1 MALDI Imaging Mass Spectrometry

MALDI-based imaging mass spectrometry [13] uses the molecular specificity and sensitivity of normal organic mass spectrometry to collect a direct spatial mapping of biomolecules (or rather their ions[5]) from a tissue section. It can simultaneously track any molecules that show up within its mass range and does not require labeling or a prior target molecule hypothesis as with complementary technologies such as fluorescence microscopy. Meistermann *et al.* [11] gives a good example of its use in biomarker discovery.

A quick overview of a MSI experiment is shown in Fig. 1. A more thorough treatment is available from Stoeckli *et al.* [13] and from Van de Plas *et al.* [15]. The result of a MSI experiment consists of a grid of measurement locations or 'pixels' covering the tissue section, with an individual mass spectrum connected to each pixel. The data structure can be considered as a three-mode array or tensor with two spatial modes (x and y) and one mass-over-charge mode (m/z).

1.2 Exploring MSI Data

A massive amount of measurements can be extracted from a single tissue section by using MSI as an instrument to study biochemical distributions in tissue. The size and dimensionality of the data set is a direct result of the spatial resolution that was used and the extent of the mass spectral range that was scanned. This high dimensionality prohibits direct interpretation by a human expert. This subsection describes a number of tools for presenting in an expert-friendly way useful information hidden in the data.

Unsupervised Data Mining A family of unsupervised multivariate data mining approaches [15, 9, 14] is available as an appropriate tool for cases where (i) no hypothesis is available regarding which biomolecules, ions, or masses are relevant for the pathway or pathomechanism under study, or (ii) in prospective studies where one wants to avoid biasing the analysis towards such a hypothesis. These methods decompose a MSI data tensor into a reduced set of major biochemical trends found within the tissue by grouping together masses that show, for example, correlated behavior. The resulting trends are each characterized by a spatial distribution (showing which tissue areas differ significantly from other areas) and a mass spectral signature (identifying the masses responsible for these differences). An example of such a decomposition can be found in Van de Plas *et al.* [15]. As these unsupervised methods require no user-specified query

[4] MALDI or 'matrix-assisted laser desorption ionization' is a mass spectrometry ionization method that is well suited for the study of larger biomolecules such as proteins. It ionizes molecules by firing a laser at the sample embedded in a crystalline chemical matrix solution on the target plate.

[5] MALDI-based ionization typically produces ions with a charge z equal to one, which is why we equate and use interchangeably the concepts of mass-over-charge and molecular mass from this point onwards.

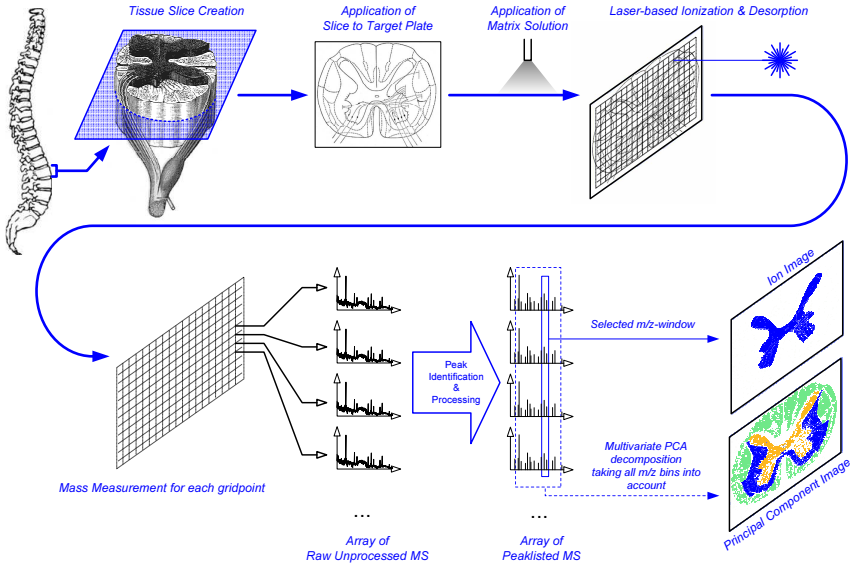


Fig. 1. Overview of an MSI experiment on spinal cord. An identical procedure was followed for the mouse brain sections in this paper. (wet-lab) A tissue section is cut using a microtome, mounted on a target plate, and covered with an appropriate chemical matrix to enable ionization. (mass spec) Individual mass spectra are collected from the tissue area of interest, while their spatial relationships are retained. (in silico) The data is collected into a three-mode array for analysis.

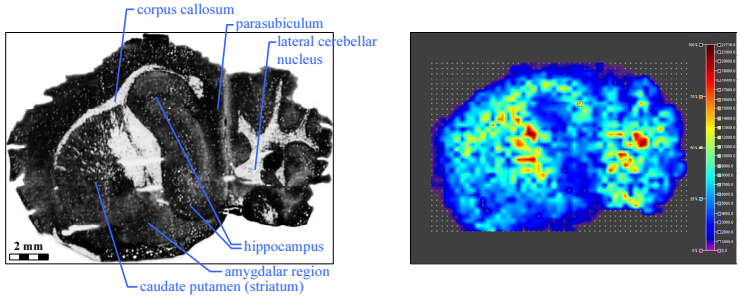


Fig. 2. (left) Picture of the sagittal mouse brain section imaged in section 3. (right) Ion image showing the presence of m/z 14148 in anatomical regions such as the corpus callosum and the lateral cerebellar nucleus.

and perform their decomposition according to a predefined similarity measure, they have particular merit in exploratory or lead discovery settings.

Supervised Mass Query Unsupervised methods are much less suited when e.g. a scientific hypothesis requires validation based on the MSI measurements. As the hypothesis specifics are not taken into account by the decomposition procedure, one has to examine all resulting trends for information on the particular masses or spatial areas one is interested in. Let us refer to such an examination as to a *query*. An unsupervised approach runs the risk that interesting masses or areas become overshadowed by dominating effects since the employed measures might not be appropriate for the task. If however the query can be formulated purely in the mass domain as a certain mass or ion of interest, a straightforward method exists in prompting a specific set ion images. An ion image shows the spatial distribution throughout the tissue of a particular mass (or m/z), and is a common approach for interrogating MSI data, albeit with the *caveat* of being a univariate method that ignores the between-mass relationships. An example of an ion image is shown in Fig. 2 (on the right). It shows the presence of m/z 14148 in the sagittal section of mouse brain used in the case study of section 3. Figure 2 (left) shows a microscopic picture of the tissue slice, with a number of anatomical regions specified.

Supervised Spatial Query A different type of queries presents itself in the spatial domain rather than the mass domain. They arise from scientific questions that focus on a particular zone in the tissue and they try to obtain insight into the chemical signatures and relationships specific to that zone. Although these types of questions arise quite frequently in biology and medicine, a method for handling such spatial queries to MSI experiments is currently lacking. This absence is in stark contrast to the potential uses for such a method, particularly for elucidating pathomechanisms that show spatially specific behavior. These include pathologies such as Parkinson's disease (where dopamine producing brain nuclei such as the *amygdala* and the *putamen* are affected [7]), Huntington's disease (with cell death in the *striatum* [17]), amyotrophic lateral sclerosis (with a striking degeneration of motor neuron regions in the spinal cord [4, 12]), and spinocerebellar ataxia affecting the Purkinje cell structure with an anatomical shrinking of the *cerebellum* (demonstrating variable degrees of neurodegeneration in the *cerebellum*, brainstem, and spinocerebellar tract [16]). Additionally, it has been shown that the affected regions can behave differently depending on the stage of the disease. An example in Parkinson's disease is the upregulation of the dopaminergic function in the *pallidal*, *amygdala*, and *cingulate* regions in the early stage and the reduction later on. The loss of *putamen* dopaminergic function has been shown to correlate with rigidity and bradykinesia [3]. For all these diseases, mouse models are currently available and in addition, mouse models demonstrating Tau accumulation at the spinal motor neuron regions have been described [6]. Concluding, it is apparent that in all these examples imaging mass spectrometry can be applied to study the alterations in the affected regions at various stages of the disease, with particular advantages if the analyses can be further focused on the known areas involved.

With these applications in mind, this paper attempts to fill the void in supervised spatial querying of MSI data by first introducing a univariate method, based on query-vs.-ion image correlations. Then we formulate a more powerful multivariate approach based on a least squares optimization problem, which delivers a more interpretable and ordered view into the masses that are active in the region of interest.

2 Methods

To provide a formal grounding for the methods developed in this section, we first define the concept of a spatial query to MSI data.

Definition 21 (Spatial Query to a MSI Experiment) *A* spatial query *posed to a MSI data array aims to retrieve molecular masses (or ions) which* 'exhibit' *a spatial pattern of interest. A spatial pattern can typically be represented as a query image. The understanding is to retrieve a small number of such masses which are most likely contributing to the pattern, helping the researcher to investigate a location specific effect or to design a specialized experiment.*

Subsection 2.1 explains a naive univariate approach to the problem, followed by a more powerful multivariate method elaborated on in subsection 2.2. Both methods are demonstrated in the case study of section 3 using a MSI measurement of mouse brain.

2.1 Univariate Correlation-based Spatial Querying

Consider a set of ion images sampled from a single tissue section and covering a certain mass range, collected into a MSI data tensor. We refer to this set as to the different *features* of a MSI data set. Let those $M \in \mathbb{N}$ features be denoted as vectors of length $K \in \mathbb{N}_0$, where K denotes the number of pixels in the image, or

$$\left\{\phi^m \in \mathbb{R}_+^K\right\}_{m=1}^M. \tag{1}$$

Important here is that the features are positive by construction since they represent ion counts. Similarly, let the image query be described by a positive vector $q = (q_1, \ldots, q_K)^T \in \mathbb{R}_+^K$ of length K. Typically, a query image is binary $q \in \{0,1\}^K$ or gray level, say $q \in [0,1]^K$.

Arguably the most simple approach to answer a spatial query is to return the masses corresponding to the top correlated features. This amounts to a univariate method were we scan each mass and rank the

different measures of correlation. The Pearson correlation coefficient provides a most convenient measure $\rho \in [-1, 1]$ defined as

$$\rho^m = \frac{q^T \phi^m}{\|q\|_2 \|\phi^m\|_2}, \quad \forall m = 1, \ldots, M. \tag{2}$$

Further, it can be argued to use a rank correlation coefficient in this case as for example Kendall's τ to rank the different features. This is the case when the underlying distributions can be expected to deviate substantially from normality. Kendall's tau taking a value between -1 and 1 is defined as

$$\tau^m = \frac{2}{(1-K)K} \sum_{1 \le k < l \le K} \operatorname{sign}\left((q_k - q_l)(\phi_k^m - \phi_l^m)\right), \quad \forall m = 1, \ldots, M. \tag{3}$$

Now, one can pick the masses according to the largest coefficient $\{\rho^m\}_{m=1}^M$ or $\{\tau^m\}_{m=1}^M$.

While the profile of those coefficients over different masses (see Fig. 3(b)) gives immediately feedback to the query, it remains up to the user how many top ranked features should be studied. This amounts to thresholding the correlations at a user-defined value. It can be hoped for to find such a threshold based on a statistical test of correlation, see e.g. [5]. However, since the test is applied uniformly over all M features, one needs to apply a statistical correction for multiple testing such as Bonferroni's. Because of the large M, this results in inferior or even non-informative results. A second disadvantage is that this approach does not take the correlation between different features into account. This can be seen as follows: if all features reflect one spatial pattern (e.g. the contour of the tissue), then the correlations of all features with a related query image will simultaneously increase. Therefore, the use of such a data-independent threshold method (e.g. based on a significance test) would not work reliably.

2.2 Multivariate Least Squares-based Spatial Querying

This subsection describes the extension to a multivariate approach based on a least squares argument. It looks for the most optimal (and smallest) combination of ion images that when multiplied by their mass contribution coefficients adds up to the the target image specified in the query. The following linear model is adopted

$$q_k = \sum_{m=1}^{M} \phi_k^m p_m + \epsilon_k, \quad \forall k = 1, \ldots, K, \tag{4}$$

where the coefficients $p = (p_1, \ldots, p_M)^T$ are restricted to positivity, encoding the assumption that the image query q is a weighted average of the features, up to the residuals $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^T \in \mathbb{R}^K$. This means that the query image is assumed to be a sum of positive contributions from a finite set of molecular masses. A classical approach to approximate linear coefficients based on a set of measurements is to minimize the squared norm of the residuals, or

$$p^* = \arg\min_p \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{m=1}^{M} \phi_k^m p_m - q_k \right)^2 \quad \text{s.t.} \quad p_m \ge 0 \ \ \forall m = 1, \ldots, M. \tag{5}$$

Remark that the choice of a Poisson regression model could be more actual as the involved data are count data. We, however, stick to a least squares approach for computational reasons. Let 0_K denote a vector of size K with all-zeros. Then one can reformulate the problem (5) as a standard convex quadratic problem as follows

$$\min_p \frac{1}{2} p^T H p - f^T p \quad \text{s.t.} \quad p \ge 0_K \tag{6}$$

where the positive semi-definite matrix $H \in \mathbb{R}^{M \times M}$ is defined as $H_{ij} = {\phi^i}^T \phi^j$ for all $i, j = 1, \ldots, M$, and $f \in \mathbb{R}^M$ is defined as $f_m = q^T \phi^m$ for all $m = 1, \ldots, M$. This estimator has the peculiar property that many mass coefficients are set to zero as a consequence of the positivity constraints. This improves the interpretability considerably when compared to the univariate method discussed in section 2.1. These convex programs can be solved efficiently (e.g. 30 minutes for the case study queries) using the MOSEK Toolbox[6].

[6] For additional information, see `http://www.mosek.com`

Moreover, if one knew *a priori* where the positive non-zero coefficients would be located, duality theory [2] learns us that the optimal non-zero coefficients can be computed by solving an ordinary least squares problem of correspondingly reduced size. This implies that the solution is fairly reliable in case of small number of non-zero coefficients. On the other hand, since usually $K < M$ the effect of ill-conditioning would deteriorate the solution if many coefficients are nonzero. A classical method to cope with those numerical issues is to impose an extra regularization term on the Hessian matrix $H' = H + \gamma I_M$ for an appropriate choice of $\gamma > 0$. The resulting quadratic program (6) solves then the regularized least squares problem

$$p^*_\gamma = \arg\min_p \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{m=1}^{M} \phi_k^m p_m - q_k \right)^2 + \gamma \sum_{m=1}^{M} p_m^2 \quad \text{s.t.} \quad p_m \geq 0, \quad \forall m = 1, \ldots, M. \tag{7}$$

This can be interpreted as follows: if the optimization problem does not know which solution p^* to prefer up to a numerical quantity, choose then the solution with the smallest norm. Here γ regulates what is meant by such a numerical quantity: if γ is large, then the norm of the solutions is more influential than the exact least squares fit, and vice versa for γ small.

Now, we illustrate how this formulation can be extended for pixels where we do not have a specific query requirement, that is, the query model (4) is extended with the occurrence of *don't care* pixels. A naive solution is to omit such pixels in the least squares formulation (5) so that we effectively answer the query based on far less information. Since the inference problem (4) is in general already ill-conditioned as $M > K$, we prefer the use of a proper weighting scheme to decrease the influence of pixels which are of minor importance to the query at hand. The interpretation is that we do not mind the values of down-weighted pixels, unless it could help the algorithm to make a choice between a set of equally good hypotheses. This is encoded by a positive vector $w = (w_1, \ldots, w_M)^T \in \mathbb{R}^M_+$ as follows

$$p^*_w = \arg\min_p \frac{1}{K} \sum_{k=1}^{K} w_k \left(\sum_{m=1}^{M} \phi_k^m p_m - q_k \right)^2 \quad \text{s.t.} \quad p_m \geq 0 \;\; \forall m = 1, \ldots, M. \tag{8}$$

These pixel weights can be interpreted as a *mask* image, focusing attention of the search strategy to the relevant spatial neighborhood. In the experiments of section 3, we use the weights $w_k = 1$ for relevant pixels, and $w_k = 0.01$ for don't care.

3 Case Study

In this section, we apply the spatial query methods developed in section 2 to the MSI measurement of a mouse brain tissue section.

3.1 Materials & Methods

The example is an MSI measurement of a tissue slice (10 μm thick) that was taken from an off-center sagittal section of the brain of a BL57/6 mouse. The recorded mass range extended from m/z 2800 to 25000 with 6490 m/z-bins. Sinapinic acid (in 75% acetonitrile and 0.2% TFA) was applied as a chemical matrix solution, using an ImagePrep station from Bruker Daltonics. A MALDI mass spectral measurement was performed on each grid point of a virtual raster of size 51×34 that was superimposed on the tissue section with an interspot distance of 300 μm in both the x and y-directions. The mass spectrometer was the Autoflex III MALDI TOF/TOF from Bruker Daltonics in linear mode, and the data collection was guided by the FlexImaging module. Data analysis and processing was done using in-house developed software, MATLAB from MathWorks Inc., and the MOSEK Optimization Toolbox from MOSEK ApS.

3.2 Results

We perform spatial queries on two specific anatomical regions present in the sagittal mouse brain section. The goal is to assess which molecular masses show a spatial distribution that resembles as a whole or in part the area of interest specified by the query. The first region of interest is the upper half of the *hippocampus*, visible in Fig. 2 as the central bulbous region. The second area under investigation is the elongated *corpus callosum*, shown in Fig. 2 as the narrow pale area.

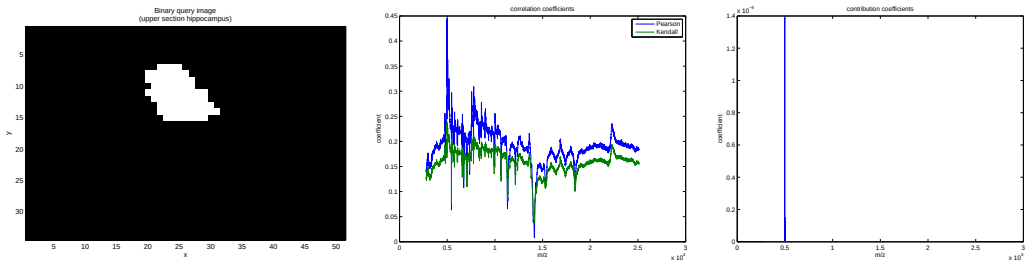


Fig. 3. (a) Binary query image examining the upper *hippocampus* area in the sagittal mouse brain section. (b) Univariate method – Correlation profile of the masses (or ions) that are spatially specific to the upper *hippocampus* area, using both the Pearson and the Kendall correlation coefficients. (c) Multivariate method – Contribution profile of the masses (or ions) whose presence corresponds spatially to the binary specification of the upper *hippocampus* area. Due to the extent of the mass range the graph seems to show only one specific mass contribution, but this peak is actually a narrow representation of two neighboring peak families at m/z 4980 and 5016.

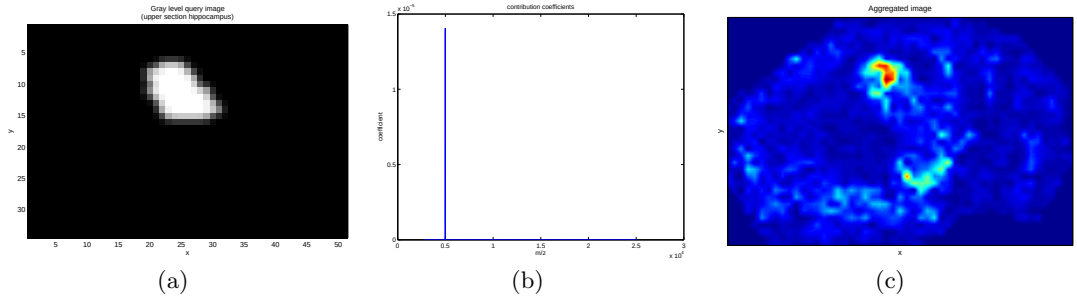
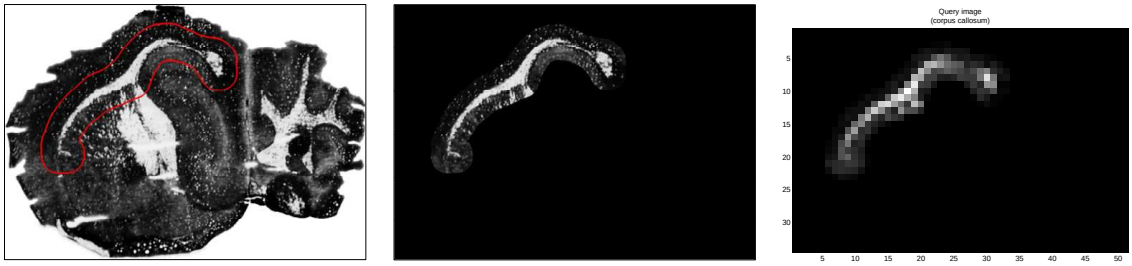


Fig. 4. (a) Gray level query image examining the upper *hippocampus* area in the sagittal mouse brain section. (b) Contribution profile of the masses (or ions) whose presence corresponds spatially to the gray level specification of the upper *hippocampus* area. Due to the extent of the mass range the graph seems to show only one specific mass contribution, but this peak is actually a narrow representation of two neighboring peak families at m/z 4980 and 5016. (c) Aggregated image calculated from the mass mixture found to be specific to the upper *hippocampus* area. The image indicates that the lower *hippocampus* area has an identical mass spectral signature to the query region.

Upper Hippocampus The upper *hippocampus* area is examined using two synthetic query images, drawn as overlays on the MSI measurement grid. The first query, shown in Fig. 3(a), is binary encoded. The pixels that are considered part of the area of interest are given a value of 1. The search for masses whose presence corresponds to the drawn area, is performed using both the univariate and the multivariate methods developed in section 2. The univariate correlation-based method is run twice, once with the Pearson and once with the Kendall correlation coefficient. For each m/z-bin, the correlation between its ion image and the query image is calculated. The resulting profile of correlation coefficients spanning the mass range is shown in Fig. 3(b). The binary query is also processed using the multivariate least squares method, with no *don't care* pixels specified and with the regularization parameter γ set to 5×10^8. The resulting profile of contributing masses is given in Fig. 3(c) (although not clearly visible due to strong compression along the m/z axis). It shows two mass windows that know an increased presence in the upper *hippocampus*: one from m/z 4975 to 4984 (highest contribution at m/z 4980), and the other from m/z 5014 to 5016 (highest at m/z 5014).

A binary query image is the most basic format of a spatial question. In the search for the most optimal combination of ion images that can represent the query, all the region of interest pixels have equal target value. It is however possible to bring more nuance to the query by using more than two levels in the target image. The area of interest can be specified with a certain texture, indicating subareas with higher or lower presence or a gradient from one location to another. When viewed as a monochrome image, this



(a) (b) (c)

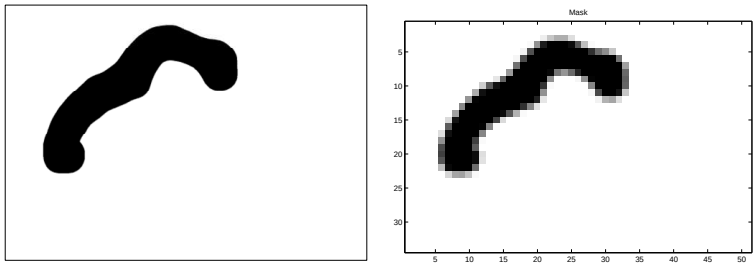
Fig. 5. (a) Microscopic image of the brain tissue section, registered to the MSI measurement grid. The borders of the selected *corpus callosum* region are indicated in red. (b) Selected area from the registered microscopic image at its original resolution. (c) Downsampled version of the selected region, which is used as the query image for examining *corpus callosum*-specific masses.

type of query takes the form of a gray level image, and it allows for a richer specification of the spatial distribution one is looking for. We apply a gray level version of the upper *hippocampus* query to the mouse brain measurements (see Fig. 4(a)). The difference with the binary query is the gradual decrease of molecular presence at the edges of the area. This effect is closer to real biochemical distributions than the crisp area edges specified in Fig. 3(a). The corresponding profile of involved masses is presented in Fig. 4(b). The retrieved masses are identical to the binary query results, but the individual coefficients differ somewhat (primarily a decrease of contribution from the window at m/z 5016.)

An interesting extension is the construction of an *aggregated* image from these masses by multiplying their ion images with their contribution coefficients and summing the products. The result is a type of mass-constrained weighted average image. Masses whose ion image mimics the spatial query more closely will have a higher contribution coefficient and thus, a higher influence on the aggregated image. However, in practice an ion image will never be an exact copy of the spatial query, which means that the aggregated image will pull in other areas as well. The aggregated image is a means of examining which other areas in the tissue show a chemical composition that is identical to the one originally queried for. In essence, one is asking a spatial question, retrieving the answers in the mass domain, and then using this mixture of masses to pull in other spatial areas that exhibit the same chemical mixture. Figure 4(c) shows the aggregated image that was calculated from the results of the gray level *hippocampus* query (Fig. 4(b)).

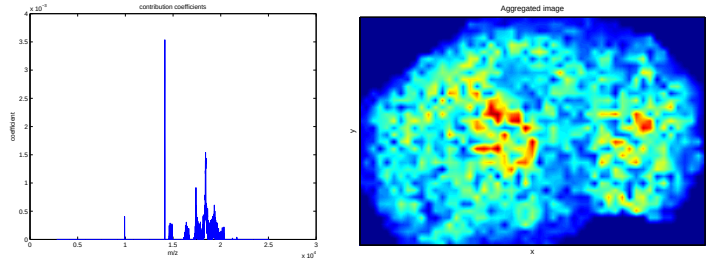
Corpus Callosum The spatial query used for the examination of the *corpus callosum* is not synthesized by specifying an area of interest on top of the MSI measurement grid. Instead the query image is the result of delineating a region on a registered microscopic image of the tissue section, and by using the visual intensity levels to specify the texture of the query area. The goal is to demonstrate that a spatial query does not have to be an abstract selection of pixels at the usually crude spatial resolution of the MSI experiment, but can also be constructed by an expert who has experience with the tissue type under study, by selecting regions in a microscopic image of the section. It is expected that this will improve accuracy and ease of use as the higher spatial resolution of the microscopic picture allows for clearer tissue navigation. Additionally, it allows experts to pose questions to MSI data working from pictures that lie much closer to the traditional imaging techniques they usually have experience with (e.g. fluorescence miscrocopy). Figure 5(a) shows the registered microscopic image with the borders of the selected *corpus callosum* region indicated in red. The area is selected sufficiently wide in order to incorporate its border gradient from pale to dark into the query texture later on. The result of the selection procedure is shown Fig. 5(b). This image is downsampled to the MSI grid size using a standard nearest-neighbor algorithm, and then used as the (gray level) spatial query for investigating the *corpus callosum* (see Fig. 5(c)).

In this example, we also make use of the *don't care* pixel extension developed in section 2.2. As earlier, we want to find the masses that show a spatial presence resembling the one specified with gray levels inside the query area. However, in the earlier *hippocampus* queries the area outside the specified region was set to zero, forcing to reconstruct also the black plateau of the query. For the *corpus callosum* query we want to make the mass search solely dependent on the selected area, mostly ignoring the ion images'



(a) (b)

Fig. 6. (a) Mask image at original microscopic resolution. The white area represents the *don't care* pixels, having a value of 1. (b) Downsampled version of the mask image at MSI resolution. It is translated to a vector of weights for the optimization problem.



(a) (b)

Fig. 7. (a) Contribution profile of the masses (or ions) whose presence corresponds spatially to the visual picture gray levels of the *corpus callosum* area. (b) Aggregated image calculated from the mass mixture found to be specific to the *corpus callosum* area.

content outside that area. This is done using a *mask* image that gets translated into a pixel weight vector for the optimization problem. The mask image for the spatial query of Fig. 5(c) is constructed from the same user-specified area as earlier, and is binary encoded using a value of 1 for all pixels that fall outside the *corpus callosum* region and that are to be ignored. This mask, shown in Fig. 6(a), is downsampled to the mask image at MSI resolution of Fig. 6(b) and fed to the multivariate query method.

The gray level query of Fig. 5(c), combined with the mask of Fig. 6(b), is processed by the multivariate spatial query method using again a regularization parameter $\gamma = 5 \times 10^8$, and produces the contribution profile shown in Fig. 7(a). The profile shows that several mass windows along the range demonstrate *corpus callosum*-presence. Particularly the masses at m/z 9918, 14140, 17400, and 18410 have ion images that follow the visual appearance that was extracted from the picture. An aggregated image is calculated as well and shown in Fig. 7(b).

3.3 Discussion

Univariate Spatial Querying The correlation profile of Fig. 3(b) provides fast and qualitative answers to the query. It indicates for example the high relevance of bins occurring around m/z 4980. However, the disadvantages of the univariate method become apparent as well. There are no zero contribution masses, defying the purpose somewhat as in practice the entire mass range is selected as part of the location-specific mass mixture. Thresholding the coefficients would help, but the profile is noisy to such an extent that developing a dynamic thresholding strategy is not a trivial problem. The disadvantages show up both using the Pearson and the Kendall correlation coefficients. An additional concern is the effect of the multiple testing problem inherent to the approach and the need to employ a suitable correction for the results affected by it.

Multivariate Spatial Querying Using Binary Image Using the same binary query, the multivariate method delivered a much more informative contribution profile (Fig. 3(c)). The solution is much sparser, with

only two mass windows having nonzero contribution coefficients. This sparseness avoids the need for a thresholding strategy as with the univariate approach. The nonzero coefficients, indicating which masses (or ion images) approximate the spatial query best, also turn out to be grouped into windows of related neighboring m/z-bins. This concurs with the notion of adjacent m/z-bins often receiving ion counts from the same or related mother molecules due to effects such as their isotopic distribution or the binning strategy of the mass spectrometer. Additionally, the size of the individual contribution coefficients implies an ordering on the selected masses according to their overall resemblance to the query image. These observations illustrate the added value of the proposed multivariate method over the univariate apporach.

Multivariate Spatial Querying Using Gray Level Image We performed the same analysis for the gray level query as for the the binary query and the contributing masses were found to be virtually identical (see Fig. 3(c) versus Fig. 4(b)). As expected, there are small differences in their individual contribution, indicating that certain masses from the set approximate the gradient edges of the gray level query better than others. In the example, m/z 5014 and 5016 show a diminished contribution when compared to the binary query results.

The aggregated image of Fig. 4(c) pulls in other spatial areas that show a chemical profile similar to the upper *hippocampus* area that was queried for. The tissue-wide spatial presence of the found mass mixture is shown from yellow to red. The primary added region is the lower area of the hippocampus. This is unsurprising as we set out using a spatial query that selected only part of the *hippocampus*, which as a known anatomical region in the brain has a more or less homogeneous composition throughout. This example indicates the capability of the aggregated image to complete anatomical regions if only part is selected. It can equally well be used in a discovery setting to identify unknown regions that show a chemical signature similar to an area known to be active in a certain pathway or mechanism. It provides an interesting added source of information, and has particular value in exploratory settings or for experiment design.

Multivariate Spatial Querying Using a Visual Picture Area The contribution profile of Fig. 7(a) demonstrates the capability for a visual picture to be used as the source for spatial querying. This type of application is very interesting for the interrogation of MSI data by non-MSI experts such as pathologists. It provides an interface closer to traditional biochemical imaging technologies such as microscopy. Also, by using the visual texture from the microscopic image as the gray level query target, we are in a sense combining information from a separate data source, a microscopic image of the tissue section, with the MSI measurements performed on that same tissue slice.

The aggregated image, shown in Fig. 7(b), pulls in a number of different anatomical regions that exhibit a similar mass spectral signature to what was found in the *corpus callosum*. These regions (indicated in yellow and red) include the *lateral cerebellar nucleus* and its extensions (which are also clearly visible on the picture of Fig. 2) and the *caudate putamen* or *striatum* region. These inclusions indicate a mass spectral signature common to all three regions.

4 Conclusion

The technology of imaging mass spectrometry adds spatial information to the traditional mass spectral biochemical analysis. With the methods of spatial querying presented in this paper we provide the researcher with a means of interrogating MSI experiments from the spatial viewpoint, rather than the traditional mass-centric approach. The objective is to learn which molecular masses or ions show behavior that is specific to a certain area in the tissue. This type of information has particular value e.g. in the search for anatomical region-specific biomarkers, for the study of local biochemical phenomena, and for the hypothesis-driven study of pathomechanisms. Therefore, a univariate and multivariate approach were developed. The multivariate method is based on a least squares argument and can be solved efficiently as a convex programming problem. The methods can handle purely synthetic queries in a binary or more nuanced gray-level format as well as queries based on a microscopic picture of the measured tissue section. Although the examples presented in this paper were primarily focused on neurobiological tissue, these methods have no inherent link to a certain type of sample and can be put to use in any setting where MSI has value.

Acknowledgements

We kindly acknowledge Dagmar Niemeyer and Sören-Oliver Deininger from Bruker Daltonics in Bremen, Germany. RVDP is a research assistant of the IWT at the Katholieke Universiteit Leuven, Belgium. KP is a postdoctoral researches of the FWO 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. Additionally, RVDP, BDM, and EW are affiliated with the Interfaculty Centre for Proteomics and Metabo-lomics, ProMeta at the K.U.Leuven (www.prometa.kuleuven.be).

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 & P6/28; EU-RTD: ERNSI; FP6-NoE; FP6-IP, FP6-MC-EST, FP6-STREP, ProMeta, BioMacS.

References

1. R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.
2. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
3. D. J. Brooks and P. Piccini. Imaging in Parkinson's disease: the role of monoamines in behavior. *Biol Psychiatry*, 59(10):908–918, May 2006.
4. A. M. Clement, M. D. Nguyen, E. A. Roberts, M. L. Garcia, S. Boillee, M. Rule, A. P. McMahon, W. Doucette, D. Siwek, R. J. Ferrante, R. H. J. Brown, J.-P. Julien, L. S. B. Goldstein, and D. W. Cleveland. Wild-type nonneuronal cells extend survival of SOD1 mutant motor neurons in ALS mice. *Science*, 302(5642):113–117, Oct 2003.
5. J. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. CRC Press, 2003.
6. J. Gotz, N. Deters, A. Doldissen, L. Bokhari, Y. Ke, A. Wiesner, N. Schonrock, and L. M. Ittner. A decade of tau transgenic animal models and beyond. *Brain Pathol*, 17(1):91–103, Jan 2007.
7. N. Hattori and Y. Mizuno. Pathogenetic mechanisms of parkin in Parkinson's disease. *Lancet*, 364(9435):722–724, Aug 2004.
8. R. M. A. Heeren. Proteome imaging: a closer look at life's organization. *Proteomics*, 5(17):4316–4326, Nov 2005.
9. L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere, and R. M. 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):222–236, Feb 2007.
10. L. A. McDonnell and R. M. A. Heeren. Imaging mass spectrometry. *Mass Spectrom Rev*, 26(4):606–643, Jul 2007.
11. H. Meistermann, J. L. Norris, H.-R. Aerni, D. S. Cornett, A. Friedlein, A. R. Erskine, A. Augustin, M. C. De Vera Mudry, S. Ruepp, L. Suter, H. Langen, R. M. Caprioli, and A. Ducret. Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat. *Mol Cell Proteomics*, 5(10):1876–1886, Oct 2006.
12. V. Silani, L. Cova, M. Corbo, A. Ciammola, and E. Polli. Stem-cell therapy for amyotrophic lateral sclerosis. *Lancet*, 364(9429):200–202, Jul 2004.
13. 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):493–496, Apr 2001.
14. R. Van de Plas, B. De Moor, and E. Waelkens. Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted pca. In *Proceedings of the Third IEEE-NIH Life Science Systems and Applications Workshop: 8-9 November 2007; Bethesda, Maryland (in press)*, 2007.
15. 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 R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing 12: 3-7 Jan 2007; Maui*, pages 458–469. World Scientific Publishing Co. Pte. Ltd., 2007.
16. W. M. C. van Roon-Mom, S. J. Reid, R. L. M. Faull, and R. G. Snell. TATA-binding protein in neurodegenerative disease. *Neuroscience*, 133(4):863–872, 2005.
17. F. O. Walker. Huntington's disease. *Lancet*, 369(9557):218–228, Jan 2007.