Spatial Querying of Imaging Mass Spectrometry Data:
A Nonnegative Least Squares Approach

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

This extended abstract reports on the development of an optimization-based query engine for mining spatial/biochemical data coming from imaging mass spectrometry experiments. It is shown how a high-dimensional linear query model and a non-negative least squares argument provide a practical approach for answering spatial queries. This work elaborates on the technical report [7][1] where further biological motivation and case studies for this approach were reported.

A growing body of research [2, 4, 5] shows that adding a spatial dimension to the analysis of biomolecular interactions can provide deeper insight into the biological processes under study. One of the primary tools for studying such interactions on the proteomic, peptidomic, and metabolomic level is mass spectrometry [1], which gives an accurate measurement of the molecular masses present in a given sample. However, most mass spectrometry studies disregard the exact spatial origin of a sample within tissue. Making and mining the connection between biomolecules such as proteins, peptides, and metabolites and their localized expression or distribution within organic tissue is central to the work described here. This spatial mapping can be retrieved through a developing technology that is known as MALDI[2]-based imaging mass spectrometry or mass spectral imaging (MSI) [3].

The work presented here aims to develop a method for spatial querying of massive MSI data. The objective is to retrieve the molecules (or ions) that are specific to a certain spatial area of interest in the tissue or whose expression is tied to a particular anatomical region. Such questions arise for example in pathomechanisms that show location-specific behavior (e.g. Parkinson's and Huntington's disease), in the search for anatomical region-specific biomarkers, in the study of local biochemical phenomena, and with the incorporation of spatial information into biological models.

Imaging mass spectrometry 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 explains its use in diagnostic studies, and it can steer further investigation by

[1] available at ftp://ftp.esat.kuleuven.be/pub/SISTA/rvdplas/reports/TechReport_Raf_VandePlas_msi_spatial_query.pdf

[2] 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.

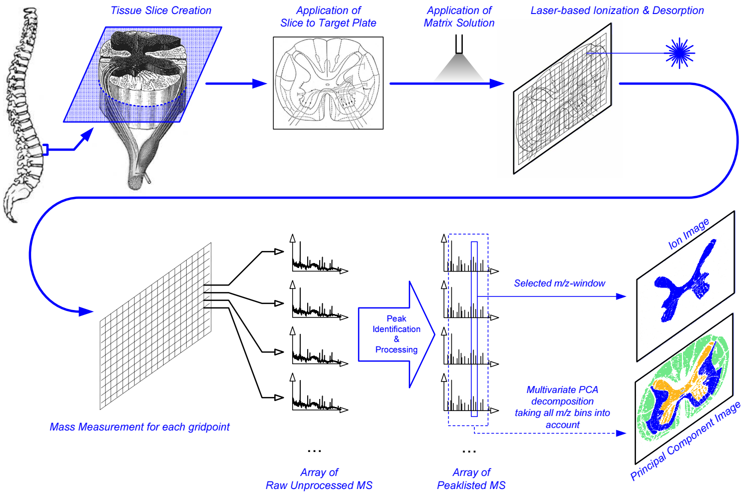
Tissue Slice Creation
Application of Slice to Target Plate
Application of Matrix Solution
Laser-based Ionization & Desorption
Ion Image
Selected m/z-window
Peak Identification & Processing
Multivariate PCA decomposition taking all m/z bins into account
Mass Measurement for each gridpoint
...
...
Array of Raw Unprocessed MS
Array of Peaklisted MS
Principal Component Image

Figure 1: Overview of an MSI experiment on spinal cord. (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.

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. Figure 1 shows an example of a MSI experiment on rat spinal cord, and a more thorough explanation is available from Stoeckli *et al.* [5] and from Van de Plas *et al.* [6]. Typically, the measurements of a MSI experiment are captured into 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).

Current methods for interrogating a MSI tensor are primarily mass-centric in the sense that they retrieve the spatial distribution of a particular ion (known as an ion image) or of a set of masses. The method developed here starts from a spatial question and retrieves answers in the mass domain instead. A schematic of this approach is shown in Fig. 2. It allows the researcher to specify a tissue area or a pattern of interest and the method will return the molecular masses or ions whose spatial presence best fits the spatial query. The biological desiderata mentioned above are tackled with a computational framework based on a nonnegative least squares argument. The following linear positive query model is adopted.

Definition 1 (Query Model) *Consider a set of ion images measured 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 spatial query image 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 $q \in [0,1]^K$. This study describes a multivariate approach to spatial querying 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 target image specified in the query. The following linear model is adopted

$$q_k = \sum_{m=1}^{M} \phi_k^m p_m + \epsilon_k, \quad \forall k = 1, \ldots, K, \tag{2}$$

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 $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^T \in \mathbb{R}^K$.

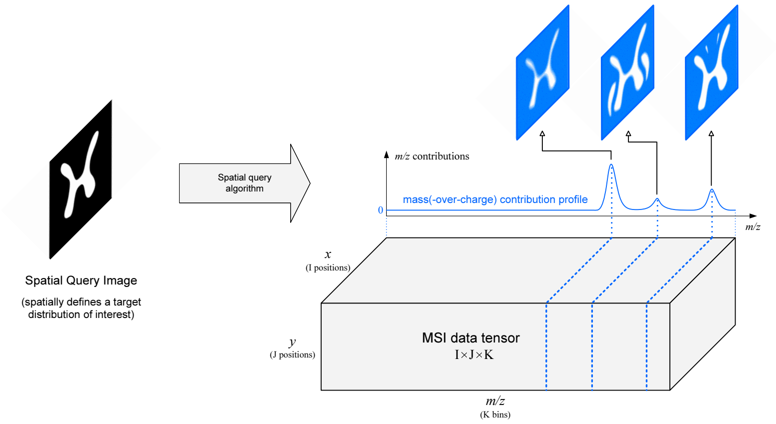

Figure 2: Overview of the spatial querying procedure. The query is formulated in the spatial domain as a region of interest in the tissue (drawn on the MSI measurement grid). The method returns a solution vector in the mass domain (marked as the mass-over-charge contribution profile). The masses that have a nonzero contribution describe a distribution in the tissue that (to some extent) mimics the shape and gray level topology specified in the spatial query.

This means that the query image is assumed to be a sum of positive contributions from a finite set of molecular masses and their spatial distribution throughout the tissue. A classical approach to approximate linear coefficients based on a set of measurements is to minimize the squared norm of the residuals, or

$$p^* = \arg\min_p \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{m=1}^{M}\phi_k^m p_m - q_k\right)^2 \quad \text{s.t.} \quad p_m \geq 0 \;\; \forall m = 1, \ldots, M. \tag{3}$$

A few consequences make this approach most convenient for the task at hand, including

Sparseness The solution vector p^* contains often many values set to zero, indicating that those features are not relevant for the query at hand. Practical experiments indicate even an elevated sparseness exceeding 90%. The curious fact is that this sparseness is independent per se of any hyper-parameter.

Tractability Such nonnegative least squares problems could be solved as a convex optimization problem (i.e. using a quadratical programming solver), and could be sped up considerably using dedicated nonnegative least squares solvers (as present in most software tools). As a consequence, one could handle queries of more than 6000 features (m/z-bins) and 2000 pixels in less than half an hour using a standard laptop PC.

Stability Duality theory learns us that the solution has the same efficiency as if the indices with zero coefficients would be omitted *a priori*. Stability could easily be further improved using classical regularization techniques.

The model allows for straightforward extension. One example is a weighted formulation that allows for *don't care* pixel areas in the spatial query where the expression level of the molecules is largely ignored. Another extension is the capability to define the spatial query on another imaging modality (usually with a higher spatial resolution) such as a microscopic image of the tissue section. This allows for domain experts to leverage their experience on the modalities they are more familiar with and provides for more conclusive deliniation of anatomical zones.

In biology and medicine questions regarding the biochemical signature specific to certain tissue areas frequently arise. The current lack of analysis methods able to answer such questions from MSI data prompted the development of the spatial query model. The basic linear model has considerable power and a number of interesting properties allow for fast and efficient searching in vast amounts of data. Additionally, a number of extensions to the model allow for more complex types of spatial queries to be formulated as well. The method is demonstrated on real MSI data in a technical report [7] using a sagittal section of mouse brain. A short summary of one particular case from this study, employing a number of the extensions mentioned earlier, is depicted in Fig. 3.

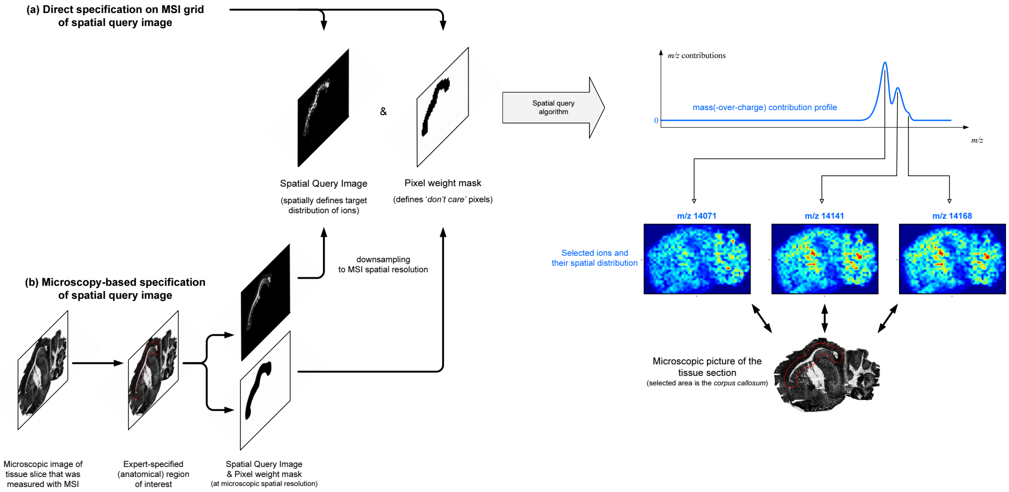
(a) Direct specification on MSI grid of spatial query image
&
Spatial Query Image
(spatially defines target distribution of ions)
Pixel weight mask
(defines 'don't care' pixels)
Spatial query algorithm
m/z contributions
mass(-over-charge) contribution profile
0
m/z
m/z 14071
m/z 14141
m/z 14168
Selected ions and their spatial distribution
downsampling to MSI spatial resolution
(b) Microscopy-based specification of spatial query image
Microscopic image of tissue slice that was measured with MSI
Expert-specified (anatomical) region of interest
Spatial Query Image & Pixel weight mask
(at microscopic spatial resolution)
Microscopic picture of the tissue section
(selected area is the corpus callosum)

Figure 3: Example of the spatial querying procedure applied to a section of sagittal mouse brain and described in [7]. Additionally, it shows the extensions to the basic linear model, allowing for *don't care* pixels and the capability to create a spatial query from another registered imaging modality such as a microscopic image. Notice that the returned ions do not show up solely in the elongated *corpus callosum* region specified in the spatial query. Their presence follows the general shape of the query area, but in addition they are shown to be present in other anatomical areas of the tissue as well. This is a result of the *don't care* pixel mask added to the query, which allows for other areas of similar chemical composition to be drawn into the result set as well. The focus lies on matching the gray level topology of the query for the area filtered by the mask.

Acknowledgements

We kindly acknowledge Dagmar Niemeyer and Sören-Oliver Deininger from Bruker Daltonics in Bremen, Germany and Justin Vijay Louis from the Katholieke Universiteit Leuven, Belgium. RVDP is a research assistant of the IWT at the Katholieke Universiteit Leuven, Belgium. KP is a postdoctoral researcher 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 Metabolomics, 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] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[2] Ron M A Heeren. Proteome imaging: a closer look at life's organization. *Proteomics*, 5(17):4316–4326, Nov 2005.

[3] Liam A McDonnell and Ron M A Heeren. Imaging mass spectrometry. *Mass Spectrom Rev*, 26(4):606–643, Jul 2007.

[4] Helene Meistermann, Jeremy L Norris, Hans-Rudolf Aerni, Dale S Cornett, Arno Friedlein, Annette R Erskine, Angelique Augustin, Maria Cristina De Vera Mudry, Stefan Ruepp, Laura Suter, Hanno Langen, Richard M Caprioli, and Axel 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.

[5] 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.

[6] Raf Van de Plas, Fabian Ojeda, Maarten Dewil, Ludo Van Den Bosch, Bart De Moor, and Etienne Waelkens. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany Murray, and Teri 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.

[7] Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens. Spatial Querying of Imaging Mass Spectrometry Data for the Biochemical Characterization of Anatomical Regions in Tissue. *Internal Report 07-171, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*, 2007.