DEVELOPMENT OF A COMPLETE ADVANCED COMPUTATIONAL WORKFLOW FOR HIGH-RESOLUTION LDI-MS METABOLOMICS IMAGING DATA PROCESSING AND VISUALIZATION

Pere Ràfols Soler

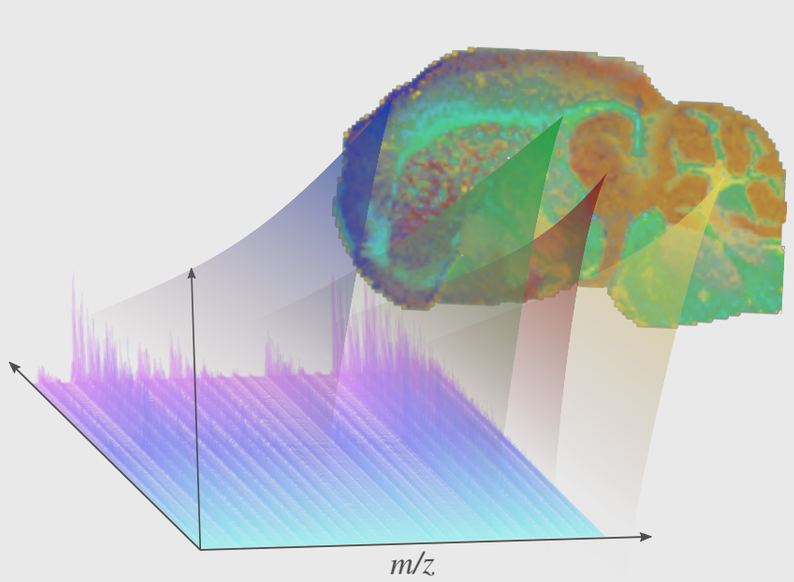
UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF A COMPLETE ADVANCED COMPUTATIONAL WORKFLOW FOR HIGH-RESOLUTION LDI-MS METABOLOMICS IMAGING DATA PROCESSING AND VISUALIZATION
Pere Ràfols Soler

UNIVERSITAT
ROVIRA i VIRGILI

Pere Ràfols Soler

Development of a complete advanced computational workflow for high-resolution LDI-MS metabolomics imaging data processing and visualization

Pere Ràfols Soler



Doctoral Thesis
2017

Development of a complete advanced computational workflow for high-resolution LDI-MS metabolomics imaging data processing and visualization

Doctoral
Thesis
2017

UNIVERSITAT
ROVIRA i VIRGILI

Pere Ràfols Soler

Development of a complete advanced computational workflow for high-resolution LDI-MS metabolomics imaging data processing and visualization

DOCTORAL THESIS

Supervised by Dr. Xavier Correig Blanchar and Dr. Jesús Brezmes Llecha

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica
(DEEEA)

UNIVERSITAT
ROVIRA i VIRGILI

Tarragona
2017

UNIVERSITAT
ROVIRA i VIRGILI

Escola Tècnica Superior d'Enginyeria

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica

Av. Països Catalans 26

Campus Sescelades

43007 Tarragona

We STATE that the present study, entitled: *Development of a complete advanced computational workflow for high-resolution LDI-MS metabolomics imaging data processing and visualization*, presented by Pere Ràfols Soler for the award of the degree of Doctor, has been carried out under our supervision at the Department of Electronic, Electric and Automatic Engineering (DEEEA) of this university and meets the requirements to qualify for International Mention.

Tarragona, November 2017

Doctoral thesis supervisors

Dr. Francesc Xavier Correig Blanchar

Dr. Jesús Brezmes Llecha

Acknowledgments

El desenvolupament d'aquesta tesis ha estat possible gràcies a la col·laboració de varies persones a les quals m'agradaria agrair el seu suport. En primer lloc, gràcies als meus directors de tesis. Xavier, valoro molt la confiança que des d'un primer moment vas dipositar en mi. També vull agrair-te haver-me descobert uns mons nous i completament desconeguts per a mi: la metabolòmica i la imatge per espectrometria de masses. Ha estat un plaer aprendre que un enginyer no està obligat a viure confinat en un món de metall i silici. Jesús, gràcies per proposar-me aquest projecte i per totes les idees que m'has aportat en l'àmbit del processat de dades.

A continuació, agrair al Dídac i al Raül les hores invertides en formar-me en l'ús de l'equip de "sputtering", la sala blanca i els coneixements en nano-tecnologies. Al Dídac també li vull donar gràcies per ajudar-me a fer els meus primers passos amb l'instrument MALDI. Als companys de despatx més antics: Xavi, Josep, Rubén i Dani per les converses inspiradores que hem tingut centrades en el processat de senyals. Aquest intercanvi de idees ha fet possible en bona mesura aquesta tesis. I també als companys de despatx més recents: Sònia, Alex, Carla i Lluc per els moments compartits. També vull donar gràcies a l'Esteban per totes les converses sobre programació.

El fet d'estar vinculat a un projecte multidisciplinari fa que freqüentment els coneixements d'un enginyer siguin del tot insuficients. Per aquest motiu, vull agrair molt especialment a l'Oscar, la Noelia, la Sònia i la Sara per el suport que m'han donat amb la interpretació i enteniment del conceptes propis de la química i la biologia. A tota la gent de Yanes Lab i Biosfer Teslab per les estones que hem passat junts i també per compartir els vostres coneixements. Amb vosaltres, he aprés a sortir del meu món immers en codi i m'heu fet començar a intuir un mica el funcionament del sistema biològic macro-complex que al fi i al cap és la vida.

A la Lorena per ensenyar-me a tallar els teixits amb el criòstat. A la Serena per tota l'ajuda i suport que m'ha donat en la part experimental i més recentment la responsabilitat en les tasques de processat de dades. Als responsables de l'ordinador de càlcul: al Josep M. i al Jordi, per resoldre ràpidament els problemes i fer que la màquina continues treballant durant tot aquest temps. A la Rita i al Lukas de la unitat de microscòpia i sala blanca per l'ajuda que m'han donat.

I would like to acknowledge all the people from LUMC (Leiden, Netherlands)

for the warm welcome I received to develop my PhD stay. In special, to Bram for share all his knowledge with me and for all the fruitful discussions we had. I also would like to acknowledge Liam who gave me the great opportunity of making my PhD stay at LUMC.

Un agraïment molt especial a tot el conjunt de "hackers" anònims que impulsen les tecnologies de codi obert i que segurament mai sabran res d'aquesta tesis. Sense aquest esplèndids desenvolupadors res d'això hagués estat possible. Des del propi llenguatge R, el compilador de C++, tot un llarg llistat de llibreries fins el propi sistema operatiu Linux; han estat la pedra angular sobre la qual s'ha articulat tot el software desenvolupat en aquesta tesis.

Gràcies a tota la meva família: als meus pares, al meu germà i als meus sogres per el suport i la paciència. Al meu difunt avi, per la seva confiança en les meves decisions tot i no entendre gran cosa del que jo em dedicava. I per acabar, un agraïment molt especial a la meva dona, la Mar. Per donar-me la força necessària per abandonar la meva antiga vida d'enginyer industrial i començar un altre camí: el de la recerca. Un camí que realment m'ha omplert.

"Sharing knowledge is the most fundamental act of friendship. Because it is a way you can give something without loosing something"

Richard Stallman

Abstract

Mass spectrometry imaging (MSI) is a technique that maps the spatial distribution of an analyte directly onto a tissue section. This allows representing specific molecule distributions directly from a tissue section. MSI is becoming a valuable technology for histopathology since rich chemical information is recorded in a single experimental run. Three main ionization methods have been developed for MSI: MALDI, SIMS and DESI. In this work, MALDI (matrix assisted laser desorption/ionization) is used due to its advantages compared to other ionizations techniques. MALDI provides a high spatial resolution with good ionization characteristics of low-weight compounds. In MALDI, a laser scans the sample surface and promotes the ionization of each pixel in the image. MALDI is an established technique for the acquisition of the high mass range of the MS spectrum. However, it is still not widely adopted for metabolomics studies. The MALDI acquisition of low molecular weight compounds is a challenging task mainly due to the MS signal interferences introduced by the organic matrices compounds used to promote the ionization. We have developed an alternative laser desorption/ionization (LDI) method to improve the MSI detection of metabolites. Our LDI method consists in coating the tissue with a gold nano-layer. This nano-layer is deposited by means of the sputtering technique which is a very robust and repetitive process. In contrast to classic MALDI, no solvent is used to deposit the gold nano-layer since sputtering is a dry deposition procedure. This overcomes the problem of compound lateral diffusion of sprayed MALDI matrices enabling the MSI acquisition at ultra-high lateral resolution. The sputtered gold nano-layer also provides a reliable method for obtaining low mass range MSI datasets because very few background MS signals are generated from the sputtered layer. Moreover, the MS peaks corresponding to gold clusters appear homogeneously distributed throughout the MS spectrum at every image pixel. This enables an accurate mass calibration by using the gold MS peaks as mass references.

The following step after the MSI acquisition is the data processing. MSI generates a large quantity of complex spectral data. Translating the MSI raw data into relevant chemical information is still a challenging task because of such factors as the experimental variation and the huge size of the MSI data. This requires implementing computationally efficient routines to process the raw MSI data. To address this, we developed two software packages for the R platform: rMSI and

rMSIproc. These software packages establish a novel and flexible platform for MSI data analysis, completely free and open-source.

The rMSI package is focused on providing an efficient way to manage MSI data together with a graphical user interface (GUI) integrated in R environment. MS data is loaded in rMSI custom format optimized to minimize the memory footprint yet maintaining a fast spectra access. The data format is designed to place all the data in the hard disk drive following a matrix-like structure. Then, only the data chunks needed at each time are automatically loaded in the computer memory. This allows an appropriate management of larger than memory MSI datasets. The rMSI GUI is designed for simple and effective data exploration and visualization. Moreover, rMSI is designed to be integrated in the R environment through a library of functions that can be used to share MS data across other R packages.

The rMSI package provided us with a solution to manage and visualize MSI large datasets. However, it is necessary to assign MS peaks to chemical entities in order to extract relevant biological information from the MSI experiment. This analyte annotation process is intrinsically linked to the mass accuracy of the data. Mass accuracy and analyte identification are determined by such factors as the experimental set up and the data processing workflow. We present an MSI data processing workflow that uses a label-free approach to compensate for mass shifts. The algorithms developed were designed to perform efficiently even for large datasets generated from an FTICR mass spectrometer. We assessed the overall mass accuracy in the range m/z 400 to 1200 using silver and gold sputtered nanolayers. With our novel processing workflow we were able to obtain mass errors as low as 5 ppm using a TOF instrument. This mass accuracy enhanced workflow is implemented in the rMSIproc package. Besides, rMSIproc also includes a complete pre-processing pipeline able to produce a reduced peak matrix from an MSI experiment performed with TOF or FTICR spectrometers. The generated peak matrix is a data reduced but accurate representation of the whole MSI dataset. Moreover, the peak matrix is also small enough to fit in computer memory. Thus, this enables the use of previously developed statistical analysis algorithms to be easily applied to MSI datasets. rMSIproc takes advantage of rMSI data model to work with files larger than the computer memory capacity. Most of the rMSIproc internal routines are implemented in C++ using a multi-threading strategy. This allows to take profit from modern multi-core processors thus provides a better processing performance to the open-source MSI data analysis.

We believe the developed experimental workflow together with the developed software packages will have a positive impact on MSI for spatial metabolomics applications. In our opinion, this work will contribute to a future better understanding of modern molecular histopathology from the point of view of metabolomics.

List of Abbreviations

m/z	Mass-to-charge ratio
$AMASS$	Algorithm for MSI Analysis by semi-supervised Segmentation
ANN	Artificial neural network
$ASCII$	American Standard Code for Information Interchange
AUC	Area under the curve
CPU	Central processing unit
DA	Discriminant analysis
$DESI$	Desorption electros spray ionization
$DESI-MSI$	Desorption electrospray ionization mass spectrometry imaging
DWT	Dscrete wavelet transform
FT	Fourier transform
$FTICR$	Fourier transform icon cyclotron resonance
GA	Genetic algorithm
GB	Gigabyte
GPU	Graphical processing unit
GUI	Graphical user interface
$H\&E$	Hematoxylin and eosin
HC	Hierarchical clustering
HDD	Hard disk drive
$HDDC$	High dimensional discriminant clustering
$HPLC-MS$	High-performance liquid chromatography mass spectrometry
ICA	Independent Component Analysis
IR	Infrared
$ISODAT$	Iterative Self-Organizing Data Analysis Technique Algorithm
ITO	Indium-tin oxide-coated
LDI	Laser desorption ionization
$MALDI$	Matrix assisted laser desorption ionization
MRI	Magnetic resonance imaging
MS	Mass spectrometry
MSI	Mass spectrometry imaging
$NIMS$	Nanostructure initiator mass spectrometry
$NMFA$	Nonnegative Matrix Factorization Analysis
NMR	Nuclear magnetic resonance spectroscopy

OMP	Orthogonal matching pursuit
$PALDI$	Particle assisted laser/desorption ionization
$PARAFAC$	Parallel Factor Analysis
PCA	Principal component analysis
$PCA-DA$	Principal component analysis discriminant analysis
$PCA-SDA$	Principal component analysis-symbolic discriminant analysis
$PLS-DA$	Partial least squares discriminant analysis
$pLSA$	Probabilistic Latent Semantic Analysis
PQN	Probabilistic quotient normalization
RAM	Random access memory
RGB	Red green blue
$RMMC$	Recursive maximum margin criterion
ROI	Region of interest
$SALDI$	Surface assisted laser/desorption ionization
$SIMS$	Secondary ion mass spectrometry
SMA	Simple moving average
SMM	Simple moving median
$SMST$	Sorted Mass Spectrum Transform
SNR	Signal to Noise Ratio
$SOFM$	Self-organizing feature map
SOM	Self-Organizing Maps
$SQM1$	Simple moving first quartile
SVM	Support vector machine
TEC	Tissue excitation coefficient
TIC	Total ion count
TOF	Time of flight
TV	Total variation
VE	Variance explained
VSN	Variance stabilizing normalization

List of Publications

Pere Ràfols, Dídac Vilalta, Sònia Torres, Raul Calavia, Bram Heijs, Liam A. McDonnell, Jesús Brezmes, Esteban del Castillo, Oscar Yanes, Noelia Ramírez and Xavier Correig. *"Assessing the potentiality of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications"*.[Submitted].

Pere Ràfols, Esteban del Castillo, Oscar Yanes, Jesús Brezmes, and Xavier Correig. *"Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer"*. [Submitted].

Pere Ràfols, Bram Heijs, Esteban del Castillo, Oscar Yanes, Liam A. McDonnell, Jesús Brezmes and Xavier Correig. "rMSIproc: an R package that efficiently imple-ments a complete pre-processing workflow for mass spectrometry imaging". [Submitted].

Pere Ràfols, Sònia Torres, Noelia Ramírez, Esteban del Castillo, Oscar Yanes, Jesús Brezmes and Xavier Correig. *"rMSI: an R package for MS imaging data handling and visualization"*. Bioinformatics, 2017, 33, (15), pp 2427–2428. DOI: 10.1093/bioinformatics/btx182

Pere Ràfols, Dídac Vilalta, Jesús Brezmes, Nicolau Cañellas, Esteban del Castillo, Oscar Yanes, Noelia Ramírez and Xavier Correig. *"Signal* preprocessing*, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications"*. Mass Spectrometry Reviews, 2016. DOI: 10.1002/mas.21527

List of Congresses

Pere Ràfols, Esteban del Castillo, Bram Heijs, Liam McDonnell, Jesús Brezmes, Oscar Yanes and Xavier Correig. *"rMSIproc: An R package that implements a complete mass spectrometry imaging pre-processing pipeline with optimized memory footprint and performance"*. OurConV, Doorn, Nethernalds (2017). **[Best poster award]**

Pere Ràfols, Dídac Vilalta, Noelia Ramírez, Esteban del Castillo, Jesús Brezmes, Oscar Yanes and Xavier Correig. *"Exploratory analysis of sputtered metal layers for application to mass spectrometry imaging"*. OurConV, Doorn, Nethernalds (2017).

Pere Ràfols, Dídac Vilalta, Sònia Torres, Jesús Brezmes, Noelia Ramírez and Xavier Correig. "*rMSI: An R package for MSI data handling and visualization*". OurConIV, Ustron, Poland (2016). **[Oral communication]**

Sonia Torres, **Pere Ràfols**, Dídac Vilalta, Raul Calavia, Oscar Yanes, Noelia Ramírez and Xavier Correig. "*New technologies for metabolomics MS imaging*". 12th Annual Conference of the Metabolomics Society, Dublín, Ireland (2016). **[Oral communication]**

Noelia Ramírez, **Pere Ràfols**, Dídac Vilalta, S. Dhal, N. Adhami, M. Martins-Green and Xavier Correig. "*Laser Desorption/Ionization Mass Spectrometry in Exposure Science: Thirdhand tobacco smoke exposure as case study*". 25th Annual Meeting of the International Society of Exposure Science, Henderson, USA (2015).

Pere Ràfols, Dídac Vilalta, Oscar Yanes, Noelia Ramírez, Raul Calavia and Xavier Correig. "*Nanoparticle-assisted laser desorption ionization mass spectrometry imaging (NP-LDI-MSI) for neurodegenerative diseases study*". Laboratorio Ideas (CIBERSAM), Barcelona, Espanya (2015). **[Oral communication]**

Pere Ràfols, Dídac Vilalta, Noelia Ramírez, Raul Calavia, Oscar Yanes, Xavier Correig and Jesús Brezmes. *"Visualization of metabolic images of Langerhans islets by ultra-high resolution LDI-MS imaging using Au nanolayers"*. 11th Annual Conference of the Metabolomics Society, San Francisco, USA (2015).

Dídac Vilalta, Noelia Ramírez, Raul Calavia, Maria Vinaixa, **Pere Ràfols**, Oscar Yanes and Xavier Correig. "*New WO3 solid-state surfaces for laser desorption/ionization mass spectrometry for high throughput metabolomics*". 11th Annual Conference of the Metabolomics Society, San Francisco, USA (2015).

Contents

Chapter 1

Introduction

1.1 Histopathology

Histopathology is the science that studies the manifestations of disease in a tissue by means of microscopy. The word histopathology is derived from a combination of three Greek words: "histos", "pathos" and "logos" which refers to a tissue, disease or suffering and the study in this context respectively. The medical specialist who studies the tissue section to elaborate a diagnosis based on microscopic detailed observations is called histopathologist. Histopathology is mainly used as a tool for medical diagnosis where it typically involves the examination of a biopsy. After the sample has been surgically removed from the patient, various steps must be accomplished to obtain the histological images [1]. Histology refers to the study of tissue through microscopy but not the disease itself. Hence, histology is an important aspect of histopathology that comprises the necessary steps to prepare a specimen slice sample to put under the microscope. The knowledge of the morphology of a tissue section is also studied by histology.

Five stages are mainly used to prepare samples for histology: fixing, processing, embedding, sectioning and staining [1]. In the fixing stage, samples of biological tissues are treated using chemicals or snap frozen in a cryoprotective embedding medium. This stage ensures the preservation of the cells and prevents them from the postmortem decay. When snap frozen fixation is used, all further steps are omitted and the frozen sample is directly sliced using a cryostat. If chemical fixation is used the next step is the processing. Tissue processing is done to remove water from the sample and replacing such water with a medium that solidifies. This provides the sample with a robust structure that allows slicing the tissue in very thin sections. After tissues have been fixed and dehydrated, they have to be embedded in a very hard solid block to allow the optimal sectioning. In the embedding stage, several tissue samples are mounted together in a mold. Then a liquid embedding material, which is then hardened, is used to create the solid block. It is necessary to section the tissue in very thin slices to clearly observe the microstructure of cells in a microscope. In case of using an optical microscope, slices are cut with a thickness of ca. 10 μm and placed onto a glass slide. The final stage before placing the sample under the microscope is the staining. Here, an appropriate histology staining substance is used to enhance the observation of microstructures. A biological tissue has very little variation in color when it is observed using a microscope. Thus, several staining compounds exist to increase the contrast of the targeted microstructures. The most commonly used stain is a

combination of hematoxylin and eosin (H&E). Hematoxylin is used to stain the cell nucleus in blue, while eosin stains the cell cytoplasm and extracellular tissue in pink.

Although histopathology is used for medical diagnosis in a standardized basis, it is not a completely reliable science yet. In some cases, it is a challenge to obtain a thin tissue section that is representative of the complete tissue sample. In other situations, the used staining strategy may not be able to properly highlight the relevant tissue features. Nevertheless, the main challenge relies on the difficulty of interpreting histological images. The diagnosis through this technique is sometimes subjective. In some cases, different histopathologists may elaborate different conclusions with the same image [2, 3].

In the last years, various strategies have been developed to overcome the classical problems of histopathology. Histochemistry and immunohistochemistry are two techniques used to improve the staining of specific aspects of the tissue. This allows a better expression of tissue characteristics known to be relevant for a given disease. Histochemistry is based on using specific chemical compounds designed to react with target substances to be observed in a tissue [4]. In the other hand, immunohistochemistry employs antibodies to stain particular proteins, lipids and carbohydrates [5]. The addition of histochemistry and immunohistochemistry in a histopathological workflow improves the reliability of the diagnosis. However, the specificity of these two techniques requires selecting and optimizing the appropriate chemicals and antibodies used in each possible scenario. More recently, mass spectrometry imaging has emerged as a completely different and valuable technique to complement histopathology in a novel manner [2].

1.2 Mass spectrometry imaging

Mass spectrometry imaging (MSI) is a modern technique that can take histopathology one step further. MSI is able to obtain rich chemical information directly from a tissue section retaining the spatial localization of the recorded data [6]. MSI is a broad term that involves various instrumental platforms to reach the goal of acquiring mass spectrometry (MS) data spatially correlated with the sample morphology. Every MSI capable instrument must contain three main parts: a spectrometer, an ionization source and a system capable to focus each spectrometer acquisition to a pixel in a defined raster.

The most commonly used spectrometer are the time of flight (TOF) detectors and its variations based on using an ionic mirror known as reflectron to increase its accuracy [7]. TOF based detectors have demonstrated to be a very reliable platform for MSI since they provide a good balance between mass accuracy and acquisition speed. In a TOF detector, an electric field is used to accelerate the ions in a flying tube. Ions acquire different acceleration speeds according to their mass to charge ratio (m/z) (see Fig. 1.1A). Then, a detector placed at the end of the flying tube senses each ion impact recording its arrival time. This principle of operation allows TOF detectors to acquire a single spectrum in far less than one second, which enables such platform to acquire a large MSI dataset in a few hours. TOF instruments are able to provide a mass resolutions in the range of 15.000 [6] which is a high value given the time needed to complete an acquisition. Nevertheless, in some situations the mass resolution provided by a TOF detector may not be enough to properly resolve all the required chemical species. In such cases, detectors based on ionic traps, like the Fourier transform ion cyclotron resonance (FTICR) were introduced to increase mass resolution [8]. FTICR detectors are based on the cyclotron frequency of the ions in a fixed magnetic field. The ions enter the ICR cell, where they are trapped due to the effect of an applied electric field. Then, ions are accelerated to its cyclotron frequency using an oscillating electric field orthogonal to the magnetic field. These ions induce an image current on a pair of electrodes as the packets of ions pass close to them. The electronically measured signal is called free induction decay (FID) and is representative of the specific ions in the ICR cell. Finally, a Fourier transform is applied to the FID signal together with a calibration function to obtain the mass spectrum. This principle of operation is schematized in Fig. 1.1B.

FTICR instruments are able to increase by a factor of ten the mass resolving power of a TOF instrument [6]. But, this enhanced mass resolution is obtained at the expense of a much higher acquisition time and a large increment of the size of the data generated. This increase of the acquisition time is related to the principle of operation of the FTICR, where the mass resolution is related to the length of the time window of the FID used to calculate the Fourier transform. As example, to obtain a resolving power of 100.000 a time windows of ca. 1 second must be configured. The increment in the amount of data generated is also related to the high resolving power. A higher number of sample points must be recorded to properly reproduce the narrower peaks displayed in the high resolution spectra.

These two factors result in a practical limitation of high mass resolution applied to MSI. Due to that, lateral resolution is generally adjusted to lower values for high mass resolution acquisitions. Therefore, to two mainly used detectors, TOF and FTICR are complementary. With TOF systems it is possible to obtain MS images with a high lateral resolution but a limited mass resolution and vice versa for FTICR instruments.

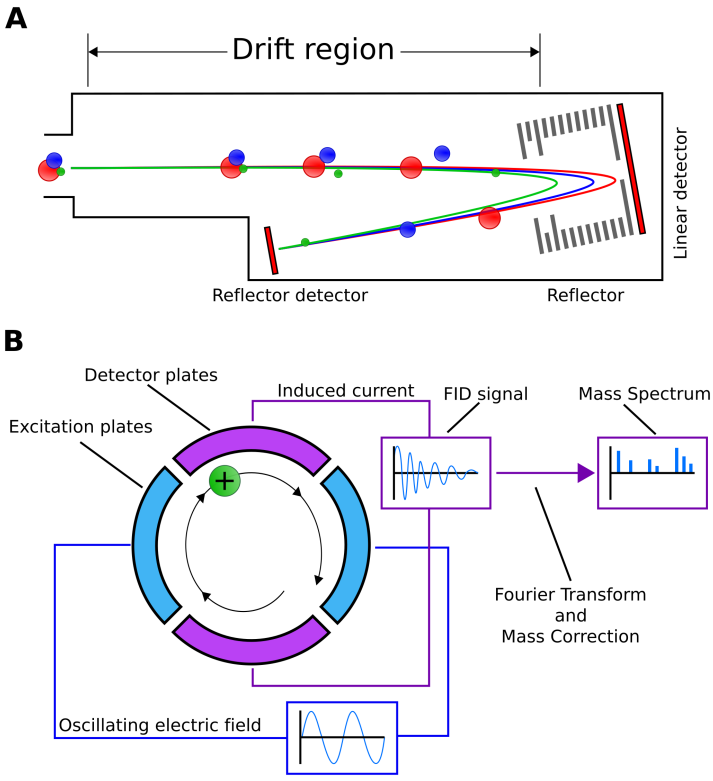


Figure 1.1: Principle of operation schematic of a TOF detector **(A)** and a FTICR detector **(B)**.

The ionization source is a crucial part of every mass spectrometer since it provides the molecules with the electrical charge necessary to be detected. In MSI, the ionization source stage has also the function of extracting the molecules from the tissue section. Three techniques are mainly employed to achieve the ionization in MSI: matrix assisted laser desorption ionization (MALDI), secondary ion mass spectrometry (SIMS) and desorption electrospray ionization (DESI) [9]. In MALDI, an organic matrix is deposited over the tissue section to promote the ionization process. Then a laser is shot at each defined raster position till the complete sample is acquired. The laser is the responsible of transmitting the ionization

energy to the sample, thus the ionization process is known as laser desorption ionization (LDI) [9]. MALDI is currently the most common ionization method used in MSI since the laser focus can be precisely adjusted to provide a high spatial resolution. Recently, alternative materials have been studied as a replacement of the organic matrix in order to improve the ionization characteristics for some specific application. Many of these organic matrix alternatives are focused on the ionization of the lowest mass range (<1500 Da). Organic matrices are known to generate abundant mass peaks in this area that can interfere with the biological sample signals. Some of these techniques are based on porous silicon surfaces [10]. Other strategies are based on the controlled deposition of metal or metal oxides nanolayers [11, 12, 13]. Nevertheless, the goal of these matrix-free methods is to promote an efficient LDI process for low weight compounds minimizing any possible interference.

SIMS uses high energy primary ions as Ar^+ , Ga^+ or In^+, to strike the sample surface. When the molecular beam hits the sample surface, a collision cascade transfers the energy of primary ions to the molecules over the surface. Besides, SIMS can perform MSI acquisitions directly on the tissue section without any sample preparation [9, 14]. The mass range of this technique is limited to >1000 Da because extensive fragmentation occurs since the energy of the primary ions must be relatively high. Nevertheless, it is possible to extend the analyzed mass range using MALDI organic matrices [9]. These primary ions can be precisely focused to a defined raster, which allows the acquisition of ultra-high lateral resolution MSI datasets. Therefore, SIMS is capable of achieving a lateral resolution higher than MALDI. Lateral resolutions in the submicron range are possible with SIMS [6, 14].

DESI is carried out by spraying solvent charged droplets directly onto the tissue section. The impact of the charged droplets with the sample is capable to trigger the desorption process of the analytes [9, 14]. In contraposition to other ionization techniques, DESI operates at ambient pressure and no organic matrix or ionization material is needed. This simplifies the sample preparation allowing the use of DESI in vivo. Nevertheless, DESI cannot achieve the high lateral resolutions which MALDI and SIMS are capable of, since the focalization of the laser or primary ion beams cannot be matched with a solvent sprayer [9, 14].

Independently of the used ionization technique, the achievable lateral resolution of any MSI experiment is far less than the resolution obtained by classical

histology through microscopy. This means that MSI cannot replace current histological techniques. MSI just provides an alternative strategy to obtain detailed chemical information from the tissue. Besides, the task of a hystopatologist is still necessary in MSI since all morphological structures in the tissue need an expert interpretation.

1.3 Spatial Metabolomics

Metabolomics is the field in the "omics" sciences that studies the interaction of small molecules in biological matrices. These small molecules ($<$ 1500 Da) are known a "metabolites" and are the intermediates and products of the metabolism. Metabolites are involved in a diversity of cellular functions, including cell energetics, inflammation, signaling, as well as building blocks of structural biopolymers such as proteins and DNA [15]. Aspects as disease, nutrition or environmental factors are able to influence the endogenous metabolites. Every living organism is capable of altering its metabolism in order to compensate for such influence. The study of these metabolites changes may provide a comprehensive understanding of the phenotype of a biological system.

The metabolome is considered to be in the lowest layer of the "omics cascade" [16]. In the upper layers, genomics, transcriptomics and proteomics are responsible of a global or holistic study of genes and proteins. These layers are subject to epigenetic regulation and post-translational modifications [16]. However, it is not possible to obtain a complete understanding of a biological system using just the upper layers because the phenotype information is always relevant. Metabolites serve as substrates and products of enzymatic reactions, and are influenced by gene and environmental factors, providing a bridge between genotype and phenotype [17]. Therefore, metabolomics is the lacking part of the big picture that will allow a comprehensive understanding of a biological system.

The two mainly used analytical platforms in metabolomics are mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). The characteristics of these two techniques are often combined to increase the number of detected metabolites. MS is commonly coupled to a gas or liquid chromatography to provide a better compound separation before the ions reach the spectrometer detector. MS is a high sensitive technique that is suitable for the detection of hundreds of metabolites. In the other hand, NMR is a quantitative, repetitive,

reproducible and very robust technique but is not capable of detecting metabolites at low concentrations.

Two different approaches are used in metabolomics studies: targeted and untargeted. In targeted metabolomics, a hypothesis is established in the experimental design. Here, a small number of metabolites are measured and quantified [18]. The targeted approach simplifies the workflow since a minimal effort and resources are required to profile these specific metabolites. On the contrary, the untargeted methods are aimed to simultaneously measure as many metabolites as possible from the biological samples without bias [16]. This metabolomics approach is used for hypothesis generation and biomarker discoveries. Here, two or more sample groups (i.e. healthy vs. diseases) are compared to obtain the metabolites that are relevant to distinguish between the experimental conditions.

Metabolomics and histopathology are able to study a tissue in great detail. However, the traditional metabolomics analytical platforms are not able to provide chemical information correlated with tissue morphology. The concept of spatial metabolomics is related to obtaining metabolomics information spatially correlated with the tissue morphological structures. Various experimental technologies are suitable to perform spatial metabolomics acquisition. Infrared (IR) spectroscopy and RAMAN spectroscopy are two well-known techniques that are able to obtain very high lateral resolution images of molecular signatures. However, in both cases the low spectral resolution makes it almost impossible to obtain enough chemical information to conduct a metabolomics study. In contrast, MSI technologies provide a suitable platform for the full development of spatial metabolomics. MSI is able to obtain rich metabolic information directly from a tissue section with high lateral resolution. As described above, three main ionization techniques exist for MSI: MALDI, SIMS and DESI. However, this work is focused on MALDI instruments because it is the available MSI platform in our laboratory and it is able to acquire low mass range spectra with a good balance between lateral resolution and fragmentation ratio. The experimental workflow to conduct a spatial metabolomics experiment using MALDI instruments starts with the sample preparation procedure. First, the tissue is cut into thin sections (ca. 10 μm) using a cryostat. Then, the tissue sections are mounted over a conductive indium-tin oxide-coated (ITO) glass slide and dried in a desiccator for a few minutes. When ready, the sample is coated with an organic matrix or some other material to promote the LDI process. Only after these steps is when the sample is placed in the

MALDI spectrometer to acquire the MS spectra directly over the tissue surface inside a defined raster. Lastly, the recorded spectral data is processed using manual ion images exploration or advanced statistical analysis tools. This workflow is explained with more detail in chapter 2 and summarized in Fig. 2.1. A standardized workflow to obtain metabolomics MS images is currently not available because there are still many experimental difficulties to be addressed. This opens up a new field of study consisting in the improvement of MSI technologies to make them suitable for direct metabolomics analysis of tissue sections.

1.4 Thesis motivation and objectives

The work presented in this thesis is the result of the research carried out in the Signal Processing for Omics Sciences (SIPOMICS) group. SIPOMICS is a research group located in the Department of Electronic, Electrical and Automation Engineering (DEEEA) at the Rovira i Virgili University (URV), and the Metabolomics Platform (www.metabolomicsplatform.com). The Metabolomics Platform is part of the Pere Virgili Health Research Institute (IISPV) and CIBER of Diabetes and Metabolic Diseases (CIBERDEM).

This thesis is the first one carried out at the SIPOMICS group in the field of mass spectrometry imaging. The main goal was to provide the knowledge to allow the Metabolomics Platform to perform metabolomics studies directly onto tissue sections. Since our research group is focused in metabolomics, the first challenge to address consists in improving the MS imaging acquisition of the lower m/z range. Therefore, it is necessary to develop and optimize an experimental workflow capable of acquiring metabolomics MSI data. Once the spectral data has been acquired the next step is to perform the data analysis to extract biologically relevant information. However, the generated data from the MSI experiments tends to be larger than most computers memory. Hence, it can be considered big data. This hinders the data analysis because most of the available software tools will not be able to process such amount of information. Therefore, advanced data analysis strategies must be developed to process the obtained MSI information. In view of the foregoing, the objectives of this thesis are:

1. Develop and optimize an experimental workflow to acquire metabolomics MS images with high lateral resolution and low background signal interferences. This, will be based on gold nano-layers deposited by sputtering.

2. Develop and implement a complete MSI pre-processing workflow able to compensate for experimental variability, enhance mass accuracy and reduce the size of the data retaining the relevant information.

3. Create a software package that includes the developed pre-processing algorithms that is computationally efficient in memory footprint and multi-threading processing.

The first objective of obtaining MSI data to perform metabolomics studies is related to the capability of acquiring the lower m/z range with a high reliability. This work is focused on the optimization of the LDI process of low weight molecules because a MALDI instrument is used. When a mass range below m/z 1000 is acquired, an abundant number of peaks from the organic matrix material are detected in classical MALDI. The matrix clusters peaks difficult the process of metabolite identification and the further data analysis [19]. Moreover, the use of a sprayer system to deposit the organic matrix over the tissue section is prone to provoke compound diffusion since matrix is mixed with an organic solvent [20]. This effect reduces the lateral resolution of the recorded MS images. Matrix sublimation methods have been proposed as an alternative deposition procedure to avoid the compound lateral diffusion. However, this strategy is not able to reduce the number of detected matrix cluster peaks. Recently, metal layers deposited using a sputtering system have been introduced as an alternative to organic matrices [12, 13]. Previous expertise already existed in the nanotechnology and nanomaterial fields in the research group were this thesis has been executed. This allowed a fast adoption of the sputtering technology as a suitable alternative to organic matrices for MSI application. The combination of metal nanoparticles with the sputtering deposition provided a reliable methodology for the acquisition of metabolomics MSI data. The MS spectra recorded using such sputter deposited layers display very few and controlled background signals plus the availability to acquire ultra-high lateral resolution MS images.

The second objective of this thesis is related with the data analysis of the huge data produced in every MSI experiment. A single MSI acquisition can easily produce various gigabytes (GB) of raw data because thousands of pixels are usually obtained and each pixel contains a complete MS spectrum with thousands of sampling points. Moreover, a complete experiment usually involves various tissue samples that must be compared. Dealing with such amount of data hinders the process of extracting relevant biological information. Most of MSI experiments

are based on a targeted approach where a software tool is used to reconstruct the images of the ions that are being studied. In some cases, data is explored manually to locate the ions that are significant between each case of study. However, this tedious procedure is not able to develop the full potential of MSI. In the other hand, untargeted data analysis strategies can be used to select statistically significant ions between regions of interest in an automated way. Nevertheless, in most MSI experiments it is not possible to apply the untargeted approach due to the size of the raw data and the difficulty to select the relevant ions in a tissue image, where the ions concentrations varies pixel-to-pixel. The second objective of this thesis is to develop and implement a data reduction strategy to reduce the size of the data preserving the relevant information. Besides the size of the MSI data, the experimental variation must be considered as a factor that hampers the data analysis. The experimental variation is reflected in the data in a manner that makes the comparison of various m/z features across various samples a challenging task. The developed data reduction strategy must provide a MS spectra pre-processing workflow able to reduce this experimental variability. Such variability is displayed in the data as pixel-to-pixel intensity variations and mass misalignment. Intensity normalization routines have been implemented to improve the first type of variability. The mass misalignment problem is complex to address since many experimental factors are involved in mass accuracy degradation. Nevertheless, a novel automated spectral mass alignment algorithm is presented in this thesis. The developed spectral alignment strategy allowed us to calibrate the complete MSI dataset with high confidence. Moreover, the metal cluster peaks of our sputtered layer used to promote the LDI process have been identified and used as mass calibration references throughout the complete m/z range [12, 13].

The third objective consists in the development of a set of two software packages for the R platform (www.r-project.org). The packages include a complete MSI data visualization and pre-processing workflow. These tools provide all the facilities to handle MSI data larger than the computer's RAM memory in a user friendly manner. The raw data management approach is based on keeping all the MS information in the computer's hard disk drive and only loading small chunks of data to memory each time it is needed. Following this procedure, a complete pre-processing pipeline is executed to construct a peak matrix that discards the noise and retains only the informative parts of each MSI dataset. The resulting peak matrix is a robust and reduced representation of the MSI data that can be easily

fitted into computer's memory. This enables the possibility of using all available algorithms for untargeted analysis to be used in MSI.

R language is the chosen programing platform to develop the software packages produced in this thesis. This decision was made because R is a completely open-source solution that is very popular in bioinformatics sciences. Moreover, many R packages have been released to address different statistical problems. R environment has demonstrated to be a reliable platform for MS data analysis and metabolomics. However, R language does not allow a finer control over memory nor multithreading execution. For this reason, the most computationally intense routines were written in C++ which is a well-known programming language to provide a complete control over memory and parallel (multithreated) execution. All the developed software is released under the terms of the general public license (GPL), an open-source license to facilitate the distribution and the adoption of the developed software in the MSI community.

1.5 Organization of the document

This thesis is divided in seven chapters which consist in this general introduction, the work that was published or submitted as scientific articles and a final discussion. The chapters have been ordered according to the goals of the thesis. First, the experimental aspects are described, and only then the data processing approach, techniques and their functionality is presented.

Chapter 1 contains the general introduction to MSI, its histological background, and the organization of this document. Chapter 2 follows the introduction and continues elaborating a bit more on the state of the art of the signal processing and the bioinformatics tools for MSI. In this second chapter, the MSI experiment is reviewed to then explain the commonly used processing strategies and which problems are addressed in each processing stage. Chapter 2 also includes a review of the software tools previously available to the work done in this thesis. Actually, chapter 2 is an already published review article in the journal Mass Spectrometry Reviews. Chapter 3 contains a submitted article which describes the experimental workflow we have designed to acquire metabolomics MSI data. Here, a novel procedure used to optimize a sputtered gold nano-layer for MSI is described. A nano-layer characterization and an application example using a mouse brain tissue are provided as well.

The second and the third objectives of this thesis are elaborated in chapters 4, 5 and 6 were the developed software is explained. Chapter 4 contains a published article in the Bioinformatics journal which presents the first R package developed during this thesis: rMSI. The rMSI package was the first MSI tool able to manage MSI datasets larger than computer's memory in an R session. rMSI also includes a graphical user interface which enables the data exploration in a user-friendly manner. Chapter 5 contains a submitted article that describes the strategies developed to improve the mass measurement accuracy in both, the spectral data and the processed peak matrix. Here, a novel spectral algorithm and fast peak detection methodology are presented. Moreover, a sputtered silver-gold nano-layer is introduced as a reliable approach to study mass measurement accuracy throughout the mass range m/z 400 to 1200. Chapter 6 contains a submitted article which presents the second developed R package in this thesis: rMSIproc. This package is the complement of the previously released rMSI package and is designed to take advantage of the rMSI data management strategy. The combination of these two packages have demonstrated to be a reliable platform for MSI data processing in R. FTICR acquired datasets with data sizes up to 200 GB have been tested and successfully processed using this approach.

Lastly, chapter 7 contains a final discussion which includes a general conclusion and the future work perspectives to continue improving MSI for metabolomics and the associated data analysis with an open-source philosophy.

References

[1] Hani A Alturkistani, Faris M Tashkandi, and Zuhair M Mohammedsaleh. "Histological Stains: A Literature Review and Case Study." In: *Glob. J. Health Sci.* 8.3 (June 2015), pp. 72–9.

[2] Jeremy L Norris and Richard M Caprioli. "Imaging mass spectrometry: A new tool for pathology in a molecular age". In: *PROTEOMICS - Clin. Appl.* 7.11-12 (Dec. 2013), pp. 733–738.

[3] Andreas Römpp et al. "Histology by Mass Spectrometry: Label-Free Tissue Characterization Obtained from High-Accuracy Bioanalytical Imaging". In: *Angew. Chemie Int. Ed.* 49.22 (Apr. 2010), pp. 3834–3838.

[4] Luke D Lavis. "Histochemistry: live and in color." In: *J. Histochem. Cytochem.* 59.2 (Feb. 2011), pp. 139–45.

[5] J. A. Ramos-Vara and M. A. Miller. "When Tissue Antigens and Antibodies Get Along". In: *Vet. Pathol.* 51.1 (Jan. 2014), pp. 42–87.

[6] Liam A McDonnell and Ron M A Heeren. "Imaging mass spectrometry." In: *Mass Spectrom. Rev.* 26.4 (2007), pp. 606–43.

[7] Christian Weickhardt, Friedrich Moritz, and Jürgen Grotemeyer. "Time-of-flight mass spectrometry: State-of the-art in chemical analysis and molecular science". In: *Mass Spectrom. Rev.* 15.3 (Jan. 1996), pp. 139–162.

[8] Alan G. Marshall, Christopher L. Hendrickson, and George S. Jackson. "Fourier transform ion cyclotron resonance mass spectrometry: A primer". In: *Mass Spectrom. Rev.* 17.1 (1998), pp. 1–35.

[9] Erika R Amstalden van Hove, Donald F Smith, and Ron M.A. Heeren. "A concise review of mass spectrometry imaging". In: *J. Chromatogr. A* 1217.25 (June 2010), pp. 3946–3954.

[10] Matthew P Greving, Gary J Patti, and Gary Siuzdak. "Nanostructure-initiator mass spectrometry metabolite analysis and imaging." In: *Anal. Chem.* 83.1 (Jan. 2011), pp. 2–7.

[11] Justyna Sekuła et al. "Gold nanoparticle-enhanced target (AuNPET) as universal solution for laser desorption/ionization mass spectrometry analysis and imaging of low molecular weight compounds". In: *Anal. Chim. Acta* 875 (May 2015), pp. 61–72.

[12] Martin Dufresne et al. "Silver-Assisted Laser Desorption Ionization For High Spatial Resolution Imaging Mass Spectrometry of Olefins from Thin Tissue Sections". In: *Anal. Chem.* 85.6 (Mar. 2013), pp. 3318–3324.

[13] Martin Dufresne, Jean-François Masson, and Pierre Chaurand. "Sodium-Doped Gold-Assisted Laser Desorption Ionization for Enhanced Imaging Mass Spectrometry of Triacylglycerols from Thin Tissue Sections". In: *Anal. Chem.* 88.11 (June 2016), pp. 6018–6025.

[14] Anna Bodzon-Kulakowska and Piotr Suder. "Imaging mass spectrometry: Instrumentation, applications, and combination with other visualization techniques". In: *Mass Spectrom. Rev.* 35.1 (Jan. 2016), pp. 147–169.

[15] M.S. Monteiro et al. "Metabolomics Analysis for Biomarker Discovery: Advances and Challenges". In: *Curr. Med. Chem.* 20.2 (Jan. 2013), pp. 257–271.

[16] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. "Metabolomics: the apogee of the omics trilogy". In: *Nat. Rev. Mol. Cell Biol.* 13.4 (Mar. 2012), pp. 263–269.

[17] J. Bruce German, Bruce D. Hammock, and Steven M. Watkins. "Metabolomics: building on a century of biochemistry to guide human health". In: *Metabolomics* 1.1 (Mar. 2005), pp. 3–9.

[18] William J. Griffiths et al. *Targeted metabolomics for biomarker discovery.* July 2010.

[19] Cheng-Kang Chiang, Wen-Tsen Chen, and Huan-Tsung Chang. "Nanoparticle-based mass spectrometry for the analysis of biomolecules." In: *Chem. Soc. Rev.* 40.3 (Feb. 2011), pp. 1269–1281.

[20] Satu M Puolitaival et al. "Solvent-free matrix dry-coating for MALDI imaging of phospholipids." In: *J. Am. Soc. Mass Spectrom.* 19.6 (June 2008), pp. 882–6.

Chapter 2

Signal pre-processing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications

2.1 Abstract

Mass spectrometry imaging (MSI) is a label-free analytical technique capable of molecularly characterizing biological samples, including tissues and cell lines. The constant development of analytical instrumentation and strategies over the previous decade makes MSI a key tool in clinical research. Nevertheless, most MSI studies are limited to targeted analysis or the mere visualization of a few molecular species (proteins, peptides, metabolites, or lipids) in a region of interest without fully exploiting the possibilities inherent in the MSI technique, such as tissue classification and segmentation or the identification of relevant biomarkers from an untargeted approach. MSI data processing is challenging due to several factors. The large volume of mass spectra involved in a MSI experiment makes choosing the correct computational strategies critical. Furthermore, pixel to pixel variation inherent in the technique makes choosing the correct pre-processing steps critical. The primary aim of this review was to provide an overview of the data-processing steps and tools that can be applied to an MSI experiment, from pre-processing the raw data to the more advanced strategies for image visualization and segmentation. This review is particularly aimed at researchers performing MSI experiments and who are interested in incorporating new data-processing features, improving their computational strategy, and/or desire access to data-processing tools currently available.

2.2 Introduction

In recent years, mass spectrometry imaging (MSI), also called imaging mass spectrometry (IMS), has become a key analytical technique in proteomics, lipidomics, metabolomics [1] and related research fields, such as drug discovery and toxicology [2, 3]. MSI provides molecule-specific images that enable correlation of the spatial occurrence of target molecules and their abundance by direct analysis of biological samples without labeling or staining. To date, hundreds of biological and clinical MSI applications can be found in the literature detailing tissue-based disease classification, discovery of phenotypic intra-tumor heterogeneity, therapy-response prediction and prognosis, and drug development in the fields of oncology, pathology, diagnostics, and surgery [4, 5, 6].

Among analytical strategies used for MSI, matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF) [7, 8] is the most commonly used technique due

to its simplicity, soft ionization, fast analysis, high throughput, and the versatility and selectivity ensured by a wide range of successfully used organic matrices. Furthermore, the recent developments in MALDI-TOF instrumentation allow for high-throughput acquisition of high-resolution MS images, revealing MSI as a potential tool for diagnostics and clinical applications.

Nevertheless, MSI applications are sometimes limited by the complexity of data processing due, among other factors, to the large amount of raw data generated, peak misalignment during image acquisition, or adduct formation and/or molecule fragmentation produced by the desorption/ionization processes. Therefore, the aim of this review was to provide an overview of the data processing steps necessary for MS data treatment and visualization of MS images in proteomics, lipidomics, or metabolomics, as well as the processing tools and software currently available. This review consists of seven sections that include an introduction, a brief description of the MSI workflow, data pre-processing steps, multivariate analysis, data handling strategies and considerations, currently available software packages, and concluding remarks. In 2012, two reviews concerning the data processing of MALDI-based MSI were published and mainly focused on proteomics applications [9, 10]. More recently, another review was published focusing on strategies for data mining and visualization of 3D images [11]. Our review extends the information provided in these previous works by both collecting and reporting on the most updated bibliography in this field and specifically addressing aspects not reviewed previously, such as data formats and other computational considerations, as well as the currently available software tools and the specific problems derived from the use of matrix-free methods currently employed in metabolomics applications.

Although, this review focuses on the data processing challenges of MALDI-MS and matrix-free LDI-MS in proteomics, lipidomics, and metabolomics applications, the computational and statistical strategies discussed here can generally be applied to other MSI approaches. We hope with this review to encourage researchers currently performing MSI experiments to incorporate new data processing features to either improve their computational strategies or broaden their knowledge regarding the data processing tools currently available.

2.3 MSI workflow

A typical MSI workflow has three main steps: sample preparation, MS acquisition, and data processing and visualization. As an example, Fig. 2.1 shows a typical MALDI-MS experimental workflow. In this section, we review the first two steps of the workflow and their influence on the subsequent data processing step. Further information regarding MALDI-MS experiments and their basis can be found in recent reviews [1, 8].

Sample handling and preparation is key to optimizing sensitivity and spatial resolution [12, 8], and parameters, such as tissue-section thickness (generally 3–20 μm), must be optimized for the analytical platform selected for data acquisition. Biological tissues are usually snap frozen and stored at −80°C immediately after collection. MSI measurement of tissues fixed in paraffin- or alcohol-embedding media is not straightforward, because the molecules of the fixing material interfere and can cause contamination and ion suppression [13, 12, 8]. However, it was recently demonstrated that it is possible to perform MSI experiments from formalin-fixed and paraffin-embedded clinical tissue samples [14].

In MALDI-MS-based MSI, an organic matrix is deposited over the tissue to assist in ionization. Standard matrix-deposition techniques consisting of deposition by the spraying of organic matrices (i.e., α-cyano-4-hydroxycinnamic acid, 2,5-dihydroxybenzoic acid, etc.) could lead to metabolite delocalization (compromising the spatial resolution) and the formation of heterogeneities that cause unexpected variations in signal intensities and background noise. These affect biological interpretation of the results and determine the application of specific data-processing algorithms. Matrix effects resulting from ionization of the matrix compounds are also common in MALDI-MS experiments and interfere and suppress MS signals in the m/z region <1000 Da, which is the common m/z region in metabolomics experiments. Nevertheless, several strategies were recently developed to minimize analyte delocalization and improve sensitivity and imaging spatial resolution [15, 16] and overcome interference from matrix peaks [17, 18]. Matrix-free LDI-MS platforms, such as surface-assisted laser/desorption ionization (SALDI) [18, 19, 20] or nanostructure-initiator mass spectrometry (NIMS) [21], have recently emerged as valuable alternatives, especially for the analysis of low-molecular-weight metabolites, offering minimal analyte delocalization and fewer background peaks <1000 Da. Furthermore, the recent application of metal and metal oxide nanoparticles and nanolayers to MSI (frequently called nano-PALDI-

MSI) is opening up a wide range of possible approaches in this field. The main advantages of this technique are the few interfering peaks in the low m/z area of the metal nanolayers, the high homogeneity of the surfaces, and high spatial resolution (down to 5 μm and only limited by the diameter of the laser) [22]. The main drawback is the possible formation of metal and metal oxide adducts of the metabolites with the different isotopic forms of the metals, which can make metabolite identification more difficult. Nevertheless, the characteristic metal peaks and clusters can be used for internal mass calibration throughout the various m/z regions of the obtained spectrum [22].

Following sample preparation, an ultraviolet (UV) or infrared (IR) laser is used in MA(LDI)-MS to desorb and ionize the molecules. The mechanisms involved in desorption/ionization are still not fully understood and depend upon the LDI approach [1, 19, 23, 24, 25]. The spatial resolution of the MS image is determined by the matrix-crystal size, the possible lateral compound diffusion occurring along the matrix-deposition process, and the laser-beam diameter of a specific instrument (normally between 10 μm and 250 μm [8]). One strategy to reach spatial resolutions below the beam diameter involves use of an oversampling method [8]. Although at low spatial resolutions smaller tissue regions can be molecularly characterized, the acquisition time increases and the quality of the MSI worsens due to the abundance of lower MS peaks in the acquired spectra. Furthermore, lower resolutions generate higher volumes of data and, therefore, the need of sophisticated computational strategies. As an example of acquisition time, a laser operating at 2 KHz can perform a simple pixel measurement within 1 s, enabling acquisition of a 1 cm $\times$ 1 cm tissue sample over 1 h at a lateral resolution of 100 μm. The increase of the lateral resolution by a factor of two causes a 4-fold increase in acquisition time. Nevertheless, it is worth mentioning that recent developments in MALDI instruments could significantly decrease acquisition time. The recently released Bruker RapifleX MALDI Tissuetyper spectrometer (Bruker Daltonics, Billerica, MA, USA) is capable of acquiring 50 pixels/s, resulting in <2-min data acquisition for a 1 cm $\times$ 1 cm MS image. Therefore, the spatial resolution of each experiment must be fixed as a compromise between the abovementioned factors.

The MS platform most suitable for each application depends upon the sensitivity required, dynamic range (the range of analyte concentration that can be detected), mass accuracy, and resolving power. Time of flight (TOF) analyzers are the most commonly used detectors, especially in MALDI applications for pro-

teomics [8]. The most common type of detector is the axial TOF spectrometer, which provides a mass-accuracy error between 10 ppm and 20 ppm due to the initial velocity/drift of the generated ions. The addition of an ion reflector together with delayed ion extraction helps to compensate for this effect, which can result from non-flat-sample morphology. Using this configuration, mass accuracies of 5 ppm to 10 ppm can be achieved. Modern MALDI spectrometers are equipped with an orthogonal reflector capable of deflecting the ions perpendicular to the original direction of motion, thereby eliminating the high initial axial-velocity distribution of the plume. Mass errors <10 ppm are common with this configuration. If higher mass resolution is needed, Fourier transform orbitrap (FT-orbitrap) and Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers are available, with mass errors <1 ppm at m/z 300 [26], which makes it easier to identify compounds by their exact mass. Tandem mass spectrometry (MS/MS), a feature commonly found in MS detectors, also increases selectivity and improves identification power. MS-acquisition ranges differ depending on the MSI application, from masses <1000 Da for metabolomics to thousands of Da for proteomics.

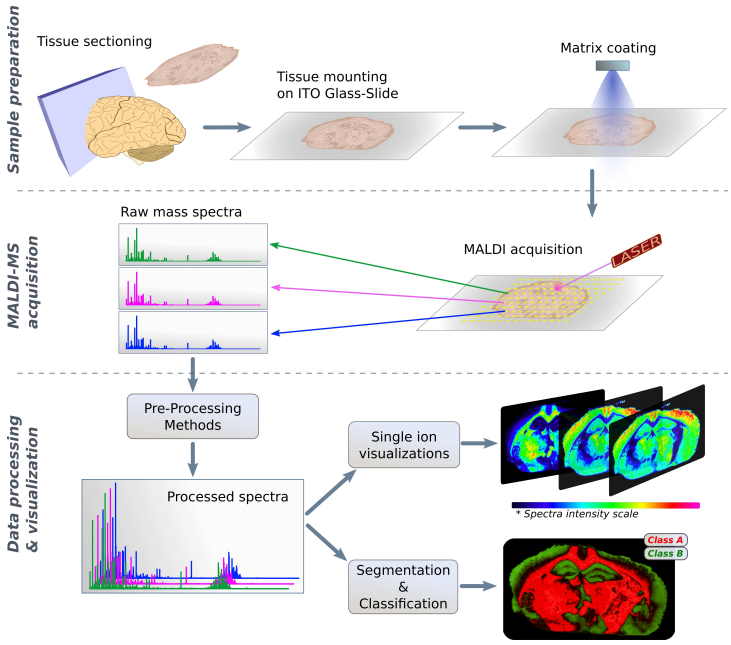


Figure 2.1: A typical MALDI-MS experiment workflow. In the sample-preparation stage, tissue is sectioned, fixed on indium tin oxide glass slides, and coated with matrix. MS spectra are then acquired using a MALDI instrument. Raw spectra are preprocessed, single-ion images are visualized, and a segmented image is displayed.

2.4 Image pre-processing

The pre-processing stage is fundamental for any MSI experiment, because the quality of the MS images depends largely upon the appropriateness of the previous pre-processing operations. The experimental variability in mass spectrometry derives from sample-preparation procedures and MAL(LDI)-MS acquisition. This variability is reflected in the raw data by introducing chemical noise, variations in the intensity and exact mass of each MS peak. In the case of large samples or high-resolution images, the overall MS spectra intensities can drift during acquisition due to instrumental reasons, such as the deposition of debris on the MS-ionization source [9]. As reference for the magnitude of this drift, we can observe ~30% intensity reduction during the acquisition of a MS image of >8000 pixels acquired at 500 shots per pixel using a commercial MALDI-TOF spectrometer.

The purpose of pre-processing is to improve image reconstruction by reducing the unwanted effects introduced by experimental variation and sample preparation. A carefully designed pre-processing workflow also helps the peak-picking procedure, the process of converting a mass spectrum into a list of relevant features for further data analysis and biological interpretation, making the statistical analysis more robust and reliable. In a typical MSI-pre-processing pipeline, the common algorithms are as follows: baseline correction, noise reduction, spectral alignment, normalization, peak picking, binning, and removal of matrix peaks. The order of the pre-processing steps is not a fixed sequence and should be adapted to accommodate the requirements of each application. Fig. 2.2 illustrates each of the pre-processing steps described below using simulated data. Some of these steps may be omitted or computed in a different order, depending on the experiment. Table 2.1 in the appendix summarizes the pre-processing methods used to date for MSI. Notably, most data-processing methods are focused on single pixel/spectrum processing and, therefore, can be also used in other MS applications.

2.4.1 Baseline correction

In MS, the baseline is the smooth curve offsetting the actual compound peaks throughout the spectrum. This signal is clearly identifiable and interferes in the base of the MS peaks, especially on those with low intensity. The effect of baseline can be observed by comparing the RAW spectra in Fig. 2.2A with the baseline-corrected spectra in Fig. 2.2B. Several baseline-correction algorithms can be used

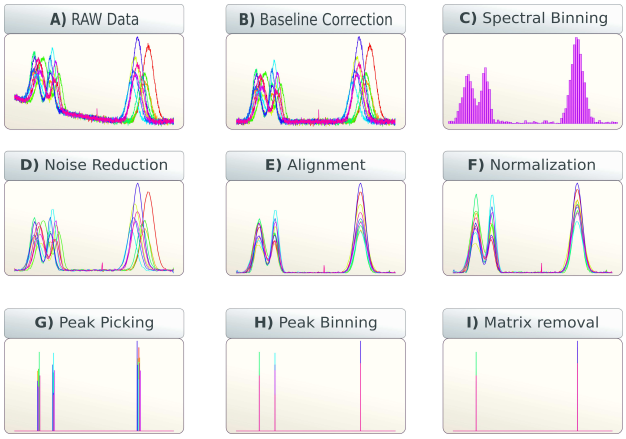


Figure 2.2: Representation of MSI-pre-processing steps (figures created with simulated data). **A)** RAW MS spectra before processing. **B)** MS spectra after baseline correction. The baseline was estimated and then subtracted using the Top-Hat method. **C)** Spectral binning used to reduce the number of data points. This use of binning is generally performed as one of the first pre-processing steps to take advantage of data reduction. Here each data point is mapped to nearest mass bin represented as a bar in the graphic. **D)** Noise reduction using a Savitzky–Golay smoothing routine. The spectral random noise is drastically reduced and the peaks shape is retained. **E)** Mass alignment and calibration that allow correction of possible mass drifts. Here, some peaks were identified as reference compounds and used to calculate m/z shifts and minimize the drift. **F)** Intensity normalization applied to reduce variability. All spectra are mapped to a similar intensity scale using TIC normalization. **G)** Peak picking that reduces each spectrum to a list of MS peaks. Each peak list is represented here as a color line pointing to the original peak location. Each detected peak retains information of: m/z, intensity and signal to noise ratio (SNR). **H)** Peak binning applied immediately after peak picking. Binning is used after peak peaking to eliminate slight mass shifts between the same detected compounds throughout all spectra. **I)** Matrix-peak removal. Peaks known as matrix peaks are removed from the binned peak list.

to correct this effect. One of the most common algorithms is Top-Hat [27], which applies a moving minimum (called an erosion filter) and subsequently a moving maximum (called a dilation filter) to the intensity values. Another generic method consists of fitting the baseline with a monotonic decay function and subtracting it from the spectra. Källback et al. [28] compared three methods for baseline estimation based on sliding windows [simple moving first quartile (SMQ1), simple moving average (SMA), and simple moving median (SMM)], concluding that SMQ1 provided a better baseline correction with minimal peak deformation. Another approach based on peak detection was introduced for baseline correction in the LIMPIC software package [29]. In this package, MS peaks obtained from sample acquisition are detected and removed from the spectrum. The resulting function is an approximation of the baseline profile. In this method, the peak-detection threshold must be adjusted to obtain accurate baseline estimations. Another strategy for baseline correction consists of standardizing the intensity of each spectrum in a defined mass range using statistical methods. This complex

and powerful method provides an estimated baseline minimally affected by higher peaks [30].

Baseline correction is generally performed prior to any other pre-processing step, because most pre-processing stages take advantage of baseline-corrected spectra; however, this could be generally avoided in metabolomics studies, because the baseline curve is very low (<1000 Da). A visual inspection of the corrected spectrum to find the flattest resulting baseline is recommended for choosing the correct baseline-compensation procedure. Therefore, the selection of a specific method depends upon the characteristics of the acquired spectra [31].

2.4.2 Noise reduction

Noise is mixed into the spectra due to the random experimental variability associated with many factors, including biological noise, matrix or surface inhomogeneity, electronic fluctuations, or ionization effects. Fig. 2.2D shows that the application of a noise-reduction algorithm attenuates the small random variations In the spectra, thereby increasing spectra quality. The application of a noise-reduction step is always recommended, because noise can interfere in most of the subsequent data-analysis steps and, therefore, must generally be performed as soon as possible in the pre-processing chain. The most common noise-reduction technique is smoothing that removes the random variations in intensity from the spectra without significant alteration of the actual signal peaks. There are numerous smoothing algorithms, each of which has their own adjustable parameters. Common smoothing techniques include moving-average windowing and low-pass filtering. As previously stated, these methods are also useful for baseline estimation, but are used for noise reduction with a different parameterization that does not completely smooth signal peaks. A more sophisticated smoothing method is the Savitzky-Golay polynomial approach, which preserves data shape [32]. Another completely different noise-reduction technique is based on a hard threshold adjusted at a noise-level estimation [30]. This method is useful when spectrum shape must be maintained for high-intensity peaks, with possible loss of low-intensity signals under the threshold.

A more robust alternative to smoothing is the application of de-noising methods using neighboring pixels instead of simply processing isolated spectra (as is done in common smoothing algorithms). These strategies may be especially useful to reduce noise in MS images reconstructed from spectra with low-intensity peaks.

Based on the assumption that peak intensity should not largely change in a local domain, vector-valued median filtering and Markov random fields are also valid strategies for noise reduction [33].

Spectra-smoothing methods are generally efficient to de-noise the signal for most applications. Single-spectrum-smoothing algorithms require less computational cost than approaches using information from neighbor pixels. Consequently, simple smoothing techniques are often preferred, except for some specific applications as mentioned.

2.4.3 Spectral alignment and mass calibration

Tissue-surface irregularities (or sample topography) in conjunction with spectrometer drift originate through small dilatations/contractions of the flight tube in the case of TOF detectors, with slight variations in high-voltage power sources producing small and random mass shifts in the spectra [31, 34]. Pixel-to-pixel mass shifts can degrade the reconstructed image and reduce the performance of subsequent data-analysis methods; consequently, direct data analysis can easily lead to erroneous peak detection due to mass variations at all raster positions. To overcome these problems, a spectral-alignment algorithm is typically used in the pre-processing pipeline. The spectral alignment consists of equalizing the mass axis of each raster spot to obtain an internal coherency when peaks are compared pixel to pixel. Fig. 2.2E illustrates spectra following an alignment stage. Alignment algorithms work by comparing the peak distribution of an unaligned spectrum, known as the test spectrum $t(x)$, with a reference spectrum, $r(x)$, that contains the correct m/z information. The algorithm is designed to find the warping function, $w(x)$, that minimizes the mass error of known peaks in $t(x)$ after applying the mass-axis transformation $t(x + w(x))$. The reference spectrum, $r(x)$, can be built using two main approaches: using actual m/z values from known compounds (calibration) or calculated from the MS image itself (label-free).

Calibration involves the use of known peaks homogeneously distributed over the tissue surface or using well-known endogenous molecules. To obtain an accurate mass calibration, various known peaks must be selected as references covering the m/z range of interest, because spectra misalignment often varies in a nonlinear way. For example, in low-weight-compound studies using matrix-free approaches, substrate background peaks can also be used to align the masses. This strategy is accurate in the case of MSI using metal nanolayers, such as silver nanolayers,

where metal peaks are distributed throughout the spectrum [22, 35].

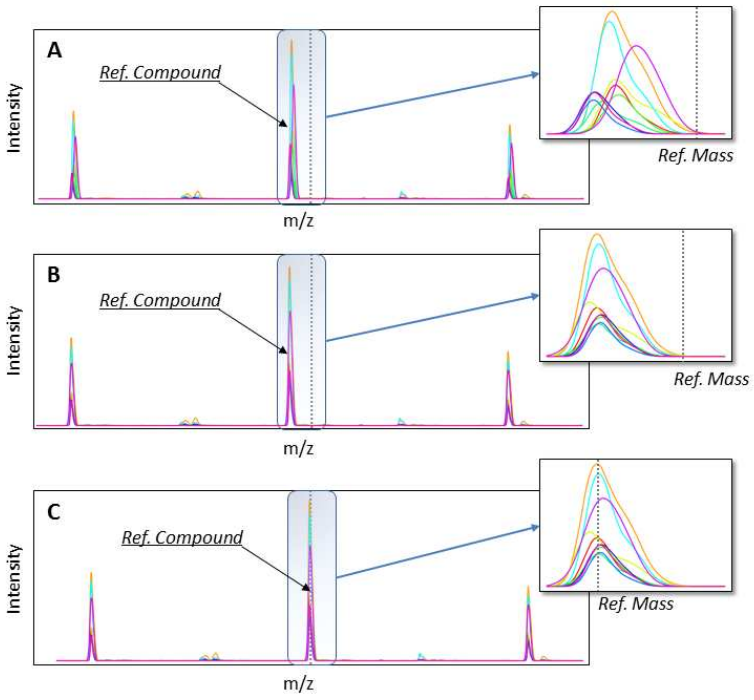


Figure 2.3: Example of two alignment approaches using simulated data. **A)** Raw spectra without alignment. **B)** Spectra aligned with a label-free technique. Here, all spectra share the same mass axis, but the peak masses are inaccurate. **C)** Spectra aligned to internal standards (calibration). Reference-compound peaks and its theoretical mass are represented as dashed lines.

Label or reference-free alignment strategies are based on the use of cross-correlations between pixels to align biologically similar spectra. Label-free methods are based on algorithms designed to detect repeated peaks throughout the dataset and use this information to minimize the mass shifts. All spectra can be aligned to a reference spectrum calculated as the average of various spectra or the spectrum with the highest correlation coefficient of all the spectra in the dataset [36]. It is also possible to align spectra without using any reference spectrum [34]. These strategies are useful in cases where it is difficult to correctly detect the calibration compounds at every raster position or where no reference compounds are used [37]. Label-free alignment can also be performed prior to mass calibration, enabling the same calibration function to be applied to all pixels independently, regardless of whether the calibrated compounds are found in a given pixel. This alignment method is represented in Fig. 2.3, where raw spectra (Fig. 2.3A) are aligned

to the same mass axis, but the masses are still not calibrated to their references (Fig. 2.3B). In Fig. 2.3C, a mass calibration method is applied to previously label-free-aligned spectra.

2.4.4 Normalization

Normalization is defined as the process of transforming the spectral intensity of every pixel to a common intensity scale [31]. Normalization is a crucial step in overcoming pixel-to-pixel intensity variability due to substrate inhomogeneity and/or experimental drifts during acquisition. Fig. 2.2f shows the changes in relative intensities of the different spectra when normalizing by the total ion count (TIC). Despite using an appropriate normalization approach, artifacts can still be introduced. Therefore, normalization might alter pixel relative-intensity distributions in an undesired way. The most common and simplest method is normalization by TIC, which assumes that an overall variation in the spectral intensity is associated with the matrix distribution throughout the sample. However, this assumption is not always true, because the concentrations of the tissue-detected compounds vary according to the biological composition of every pixel. Therefore, in tissues with clearly differentiated areas, such as brain samples, TIC normalization tends to equalize the intensities of the biological regions, which leads to inaccurate image reconstruction [38, 39]. A useful alternative could be to scale the intensities in accordance with the TIC computed using the selected peaks after peak picking or using a set of peaks relevant to the study [38]. In general, TIC normalization should be preferred for untargeted analysis due to its implementation simplicity and wide availability. However, in situations where tissue holes or "hot spots" are present [9], TIC normalization can introduce side effects for further data analysis, subsequently requiring exploration of other normalization strategies. Another normalization method consists of replacing each peak intensity based on the signal-to-noise ratio (SNR) estimated around a window [39]. This strategy assumes that the analyte concentration is proportional to the SNR of the peak and not only standardizes the intensity axis, but also compensates for the baseline noise. Nevertheless, SNR does not take into account that the ionization efficiency is not homogeneous throughout the tissue slice.

Advanced normalization algorithms based on statistical data analysis can improve the results of untargeted data analysis. Normalization based on statistics aims to compensate for the effect of experimental variance and minimize the influ-

ence of biological information on the intensity scaling factor. Here, the rationale is that the spatial distribution of variations in intensity associated with experimental variance tends to be uncorrelated with biological-sample morphology. The most simple statistical-normalization factor is probably the median of the intensities of the peaks of interest [38]. This method calculates the normalization factor using the selected peaks as input to compute the median. More complex normalization approaches include histogram matching or probabilistic quotient normalization (PQN), where spectra are scaled by a coefficient associated with the distance of the median spectrum from each TIC spectrum. These methods are reportedly more robust due to their compensation for acquisition artifacts and presentation of better noise separation when multivariate methods are used [38]. As a successful example of the application of statistics on normalization, Veselkov et al. [40] introduced variance-stabilizing normalization (VSN), a logarithmic normalization method that decreases much of the variance in high-intensity peaks and allows for hyperspectral profiling of lipid signatures in colorectal cancer tissues.

Another side effect from normalization is the change in intensity of each ion, which might exert a strong impact on reconstructed images, because different normalization strategies can produce very different results and alter final image-intensity distribution. Fig. 2.4 shows the effects of various normalization algorithms on image reconstruction of three different ions. This MS image was acquired using a sputtered-gold nanolayer over the tissue to promote ionization, but does not provide any MS signal in the absence of tissue. Fig. 2.4A shows the images associated with the data without any processing (RAW data). Fig. 2.4B shows the effect of TIC normalization calculated as the sum of all intensities of each RAW spectrum. In Fig. 2.4C, maximum-intensity normalization is used. Here, the peak with the maximum-intensity value was used as the normalization factor, and the most intense peak was assumed to be representative of the rest of the spectrum intensities. Fig. 2.4D introduces a TIC-based normalization approach designed to compensate for ionization-source degradation during MS acquisition. Here, the produced images were very similar to the RAW version (Fig. 2.4A), but exhibited a flat overall intensity variation across the entire image. Each normalization strategy presented in Fig 2.4 produced different spatial distributions of the ions, thereby confirming the relevance of selecting the appropriate normalization strategy depending on experimental design.

When target-compound concentration requires quantification, normalization is

performed relative to the peak intensity of a reference molecule deposited homogeneously over the sample. Usually an isotope-labeled compound is deposited on tissue, and normalization is performed by dividing each spectrum by the labeled-standard peak intensity [9]. Additionally, the tissue-extinction coefficient (TEC) factor was introduced as a quotient of the intensity of a standard deposited with the matrix either on or off of the tissue [41]. This coefficient evaluates the signals lost due to ionization effects for a given molecule and can be used as a normalization factor. Accurate quantification results were reported using TEC normalization without using isotope-labeled compounds.

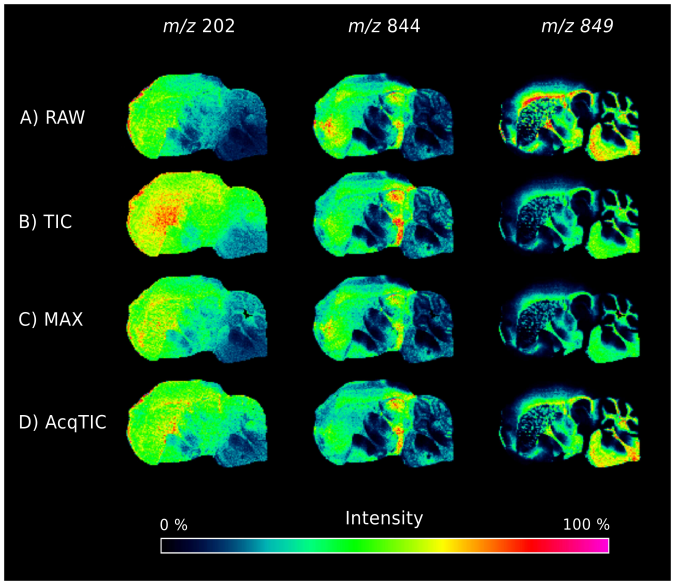


Figure 2.4: Comparison of the intensity maps for three ions (m/z 202, 844, and 849) from a sagittal mouse brain section using various normalization approaches. The MS image has been acquired using sputtered-gold nanoparticles to promote ionization and analyzed in a MALDI TOF/TOF UltrafleXtreme instrument from Bruker Daltonics in a mass range from m/z 80 to 1000 using a raster size of 80 µm. **A)** Raw data without performing any normalization. **B)** TIC normalization was computed as the sum of all intensities in each spectrum. Pixels acquired outside of the tissue are removed from the normalized image. Such pixels are detected, because they have a very low TIC ($<$ the mean TIC minus one standard deviation). **C)** Maximum normalization was calculated by dividing each spectrum by the intensity of its maximum peak. Here, pixels acquired outside of the tissue are discarded using the same criterion as that used for TIC normalization. **D)** AcqTIC is used to compensate for MALDI instrument ionization-source degradation during acquisition. AcqTIC was calculated as TIC smoothed by the TIC of neighboring pixels using a sliding window.

2.4.5 Peak picking and peak selection

Peak picking allows for the detection of peaks in a mass spectrum and provides information about peak m/z, intensity, and quality. This process reduces a mass

spectrum to a list of characteristics where only peak information is retained. Fig. 2.2G illustrates the peak-picking process, which involves retaining only peaks positions instead of the entire spectrum. The simplest approach for a peak-detection algorithm consists of locating the zero crossings in the first derivative of the spectrum. However, this will result in significant mass errors due to factors that include limitations of spectrometer resolution and noise in the data. More accurate peak m/z values can be obtained using the peak shape to predict the actual m/z instead of using only the most intense MS peak. Various methods were proposed to accurately determine the peak shape. Källback et al. [28] detected the approximate peak locations using the zero crossing of the spectrum first derivative. A cubic interpolation was also applied around each peak area to determine the peak mass more accurately. Alexandrov et al. [42] used a sequence of different algorithms. First, they modeled each mass spectrum as a sequence of Dirac delta peaks convolved with a Gaussian kernel, followed by using the orthogonal matching-pursuit (OMP) algorithm [43] to de-convolve the peaks. They then applied the maximum-likelihood method consisting of fitting the spectrum contained in a sliding window to a Gaussian shape [34].

The result of peak picking is an array-like data structure summarizing all of the relevant features of the entire MS image. In this data array, each peak can be considered as a variable and each pixel as an observation to perform further processing.

2.4.6 Binning

Binning describes the process of reducing the number of points in the spectrum by mapping neighbor m/z values into the same mass bin. Binning can be performed in two different ways. In one case, binning is performed prior to peak picking, and the mass bin size should be defined according the desired mass error. Using this method, the data size can be dramatically reduced in order to successfully execute demanding data-analysis algorithms. This binning approach is illustrated in Fig. 2.2C, where the number of points on the spectra has been reduced. Fonville et al. [38] used binning to test different normalization techniques under principal components analysis (PCA). The drawback of this method is that some close peaks derived from different compounds can be merged together, resulting in degradation of the results from further data analysis.

A second type of binning is performed after peak picking. Here, the m/z of

each peak is slightly adjusted to report exactly the same m/z for each detected compound in all pixels (Fig. 2.2H). To achieve this, each peak mass is compared with its neighbors through all pixels in a defined tolerance. The most representative mass is then used for all of the peaks expected to derive from the same compound. This technique was successfully used to enhance mass resolution in MALDI experiments [34]. However, the main drawback of this binning technique is that spectra must be well aligned in order to enable selection of a tight bin tolerance, resulting in the requirement for complex implementation processes.

2.4.7 Matrix-peak removal

In MALDI-MSI or LDI-MS spectra, the organic matrix, metal ions, or surface compounds ionize with the molecules of the biological sample. Here, we discuss the different strategies used to eliminate these matrix peaks. Such non-informative peaks do not appear in all MSI applications, because these matrix signals are commonly more intense in the low m/z range, and, therefore, this step is not always required. In MSI applications where unwanted signals are strong, matrix-peak removal may be beneficial, especially in the case of untargeted data analysis and metabolomics studies where matrix signals have the most impact on low-mass ranges.

Determining which peaks correspond to a matrix or have a biological origin can be challenging. Some methods were developed for robust and automatic selection. Fonville et al. [38] described two approaches for obtaining background-related peaks. First, the signal acquired outside of the tissue (containing only matrix peaks) is correlated using the signal acquired on the tissue (containing the matrix plus biological peaks). These correlation factors are then used to retain uncorrelated variables that are defined as biologically relevant. Second, the variance explained (VE) is used to determine which peaks constitute background signals, because matrix-related signals should be homogeneously distributed over the entire surface, leading to lower VE values. In another study, [37] identified background peaks using multivariate analysis tools to manually draw regions of interest (ROIs) inside and outside of tissue regions. In this respect, algorithms, such as PCA, can determine which masses are associated with the tissue. Once matrix-related peaks have been determined, they can be removed by deleting their corresponding variables in the peak list. To illustrate this, Fig. 2.2I shows removal of one of the spectral peaks due to its origination for the matrix.

2.5 Multivariate analysis of images

The most common and direct strategy in MSI consists of spatial visualization of one ion or a small group of ions, each of which is assigned to a color code. This strategy is especially useful for targeted analysis, with most commercial and open-source programs including many functions for plotting images of the ions of interest. However, this simple visualization strategy does not exploit the full potential of MSI, such as biomarker discovery and identification, image clustering (or segmentation), histology driven image reconstruction, tissue classification, and 3D-image reconstruction. To achieve all these objectives, multivariate methods that consider the full MS spectrum of each pixel as an intrinsic multivariate problem are introduced here.

We have divided the discussion here into three sections. The first section corresponds to the multivariate analysis of images and it is also divided in to three different approaches: supervised, using histological or microscopy images as a reference; unsupervised, which does not require previous information about the samples; and unsupervised strategies with further expert evaluation that combine the information given by the histological images with unsupervised algorithms. The second section focuses on 3D-image-reconstruction strategies, and the final section describes the different uses of PCA, which is the most used multivariate algorithm in MSI.

The up-to-date bibliography discussed here is also reviewed in the appendix Tables 2.2, 2.3 and 2.4. Notably, few papers attempt to identify the key ions involved in cluster differentiation. Although this is an essential task in biomarker discovery in proteomics, lipidomics, and metabolomics applications, the problems associated with the generation of adduct ions, possible fragmentation of the molecular ions, poor mass resolution of the TOF detectors (the most commonly used), and low sensitivity of the MS/MS working mode makes this task difficult. To overcome this, a common strategy is to identify the metabolites detected in the MSI experiment by performing high-performance liquid chromatography MS analysis using the same tissue sample [44].

2.5.1 MS Image multivariate processing

Supervised strategies

Many studies have described MSI techniques involving hematoxylin and eosin (H&E) staining or immunohistochemistry images [45]. These strategies are used for tissue recognition and classification and for biomarker discovery. Many different algorithms, including random forest, support vector machines (SVM) [46], PCA-discriminant analysis [47], recursive maximum-margin criterion (RMMC), or artificial neural networks (ANN) [48], have been used to compare MSI and histological images.

McCombie et al. [37] used compression algorithms (PCA, hierarchical clustering, k-means, and iterative self-organizing data analysis technique) in combination with a DA algorithm to maximize the spectral differences between two ROIs in a brain section from an Alzheimer's disease rat model. Results showed that the multivariate methods were capable of extracting complex information from a tissue section and that it was much easier to identify contrasting regions in an image taken from a complete rat head. Genetic algorithms and SVMs were used by [48] to differentiate prostatic tissues with and without cancer. Additionally, an SVM was able to identify four distinctively overexpressed peaks, with overall cross-validation, sensitivity, and specificity >85%.

[49] presented a new methodology for analyzing MSI datasets. The (Pearson) correlation coefficient was calculated between images acquired in an experiment with rat brain tissues to determine the correlation between ions. As a result, an interesting correlation-map matrix was obtained that described distribution similarities between 28 biomolecular ions. One important problem encountered in this study was that the method was highly sensitive to background noise.

The output of random forest algorithms was used as a class-probability estimate for classifying human breast cancer in mice models [33]. Using this approach, various regions (separate necrotic tissue, viable tumor, gelatin, tumor interface, and glass/hole) were differentiated within and between samples with high sensitivity rates (˜90%) and positive predictive values (˜85%).

ANN and SVM algorithms were used to differentiate HR2+ and HR2– regions in breast cancer tissues [50]. The area under the curve calculated by receiver operating characteristic analysis exhibited high sensitivity (83%) and specificity (92%), and an overall accuracy of 89%. Furthermore, they discovered specific

changes in protein/peptide expression (ion m/z 8404, identified as cysteine-rich intestinal protein 1) that were strongly correlated with HER2 overexpression.

Recently, Veselkov et al. [40] reported interesting advances in MSI, including automated algorithms for co-registration of histology and molecular images to aid correlation of histological and biochemical features. In the same study, partial least squares-discriminant analysis (PLS-DA) was used to extract tissue-specific molecular patterns and maximize the variance between regions and minimize variance within regions. This enabled characterization of lipid signatures in tissue regions surrounding colorectal cancer tissue. Fig. 2.5 shows the regions selected from H&E-stained high-resolution optical images used to guide a segmentation process in a desorption electrospray ionization (DESI)-MSI image. This strategy is also widely used to compare case/control samples. For example, it was used to compare brain samples from an Alzheimer's disease mouse model from those of controls [51]. In this study, ROIs from equivalent histological regions in both samples were selected, and PCA combining the pixels of the ROIs was used to identify the metabolites differentially expressed between both samples as a consequence of disease progression.

Recently, Caprioli and collaborators [52] established a new paradigm consisting of the fusion of histological and MS images. They calculated a correlation function, q, between microscopy and MS-image patterns in an attempt to create new images that combined the high-spectral resolution of microscopy with the high molecular-specificity of MSI.

Unsupervised Strategies

Unsupervised strategies were introduced in MS-image processing to disclose new molecular and morphological information independently from classical histology. These unsupervised strategies do not require prior information for clustering and are capable of revealing several molecular fingerprints, making it ideal for analyzing heterogeneous tissues and discovering biomarkers. Clinical research requires independent methods for tissue evaluation beyond classical histology. In this sense, the number of clinical studies attempting to correlate MS images with biological and clinical variables increases annually, highlighting the need for clinically validated "molecular histology". Clustering techniques, such as PCA, self-organizing maps (SOMs) [51], probabilistic latent semantic analysis (pLSA), and k-means, can be used in this respect.

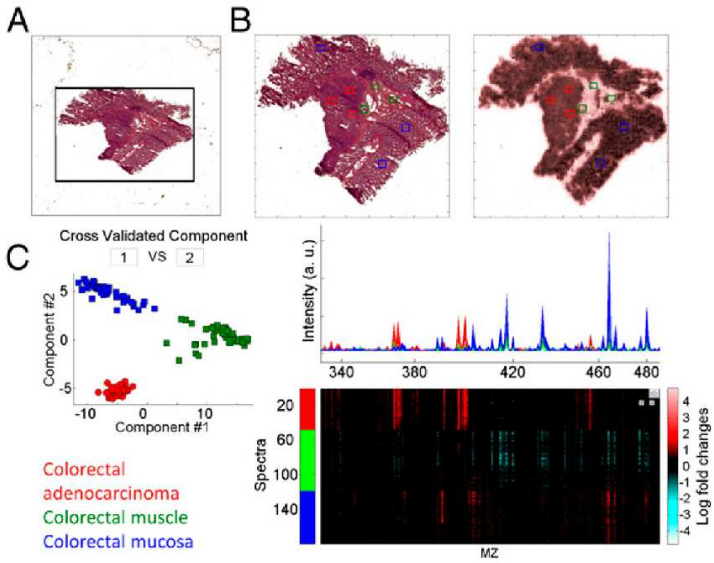


Figure 2.5: Image co-registration, feature co-selection, and multivariate analysis. **A)** Automatic image transformation for accurate co-registration of biochemical and histological features. **B)** High-resolution optical image of an H&E tissue section with regions of tumor (red boxes), muscle (green boxes), and healthy mucosa (blue boxes) selected. Shown is aligned DESI-MSI image with automated co-selection of pixels corresponding to defined regions of interest. **C)** Discriminatory analysis using the RMMC method with leave region-out cross-validation for enhanced separation of tissue classes based on biochemistry [taken from [40] and reprinted with permission of the National Academy of Sciences].

The segmentation of an image is the only technique that allows for visualization of regions with similar molecular compositions that is essential for image comparison and tissue characterization and recognition. One intrinsic problem in unsupervised clustering comprises the difficulty in the determination of the optimum number of clusters, the setting of parameters values for pixel clustering, and validation of the results [9].

Cho et al. [53] used PCA to compare the lipid, peptide, and protein profiles of various biological matrices, including MS images of tissue sections. They used PCA loadings to select the ions differentiating biological regions in a tissue sample. However, an important drawback of PCA is the negative and positive distributions of the scores, making it difficult to interpret the results when applied to MSI. To overcome this limitation, classical techniques, such as the discrete wavelet transform (DWT) [54] algorithm for data compression and de-noising, was used to solve information-technology problems based on its generating reduced sets of wavelet coefficients. DWT was used for the MSI analysis of sagittal sections of mouse brain [55]. In this study, the results of DWT application were proven to be

more compact than those obtained using PCA. Another advantage of using DWT is that it also retains mass-spectral information by means of the inverse DWT.

Another useful approach is pLSA, a statistical technique that allows for a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables. pLSA has been used to analyze MALDI-TOF images [56] and provide better physical interpretations relative to those provided by PCA, independent component analysis (ICA), and non-negative PARAFAC [57], because the decomposed components can be directly interpreted as peak-intensity lists.

A SOM is a type of ANN that is trained using unsupervised learning to produce a low-dimensional map (typically 2D) as a discretized representation of the input space of the training samples. SOMs use a neighborhood function to preserve the topological properties of the input space and are very attractive in MSI analysis. Franceschi et al. [58] used SOMs to illustrate the spatial distribution of ions associated with the regions generated for a dataset of apple slices, retaining the key ions for further analysis and metabolite identification.

PCA-symbolic discriminant analysis based on hierarchical analysis was used by [59] in a study of prostate cancer and was suitable for identifying and localizing specific markers in human prostatic tissues.

Non-negative matrix factorization analysis (NMFA) was used to resolve glial and neuronal cell-enriched brain regions [44]. Based on potassium adducts from a set of 18 selected lipids, NMFA provided six components representing spectral patterns associated with brain morphology. A method for hyperspectral visualization was recently proposed [60], consisting of a RGB color-coding based on the spectral characteristics of every pixel. The application of this strategy to various data-reduction models [PCA, SOM, and t-distributed stochastic neighbor embedding (a neural-network-based manifold-learning technique)] revealed its capability for unsupervised creation of images exhibiting good correspondence between molecular and anatomical information.

Unsupervised strategies with further expert evaluation

Unsupervised strategies with further expert evaluation strategies assess the results of unsupervised clustering by comparing them with histological images, even though these images do not take part in the clustering process. [61] used hierarchical analysis coupled with PCA to identify several gastric cancer and non-neoplastic mucosa tissues. Using this semi-supervised approach, classifications were based on

pathological information about healthy and cancerous regions, thus opening avenues for the discovery of new cancer biomarkers.

Jones et al. [62] used various statistical methods (PCA, ICA, NMFA, pLSA, k-means clustering, and hierarchical clustering) to automatically determine clusters in datasets of intermediate-grade myxofibrosarcoma [62]. Results showed that the MS images generated by the different methods exhibited similar distributions, confirming the ability to discover different nodules in identical histology tumor sections and suggesting its usefulness as a "molecular histology" technology.

In the field of multivariate approaches to clustering of MS images, it is worth mentioning the work of T. Alexandrov's research group. They used high-dimensional discriminant clustering to analyze and interpret a larynx carcinoma section and compared the automatic spatial-segmentation image obtained by MALDI-TOF with H&E-stained microscopic images [42]. The molecular image enabled exploration of tumor heterogeneity and pharmaceutical metabolism. The same research group also proposed novel strategies for spatial segmentation that incorporated spatial relationships between pixels into cluster regions, enabling pixels to be clustered together with their neighbors [63]. Additionally, they evaluated the segmentation method in a rat brain section and a neuroendocrine tumor section and identified various tumor regions by discovering the anatomical structure and identifying functionally similar regions. In 2011, they created an algorithm to increase the spatial resolution of the segmentation map by resizing the map by splitting the pixels [64].

The algorithm for MSI analysis by semi-supervised segmentation (AMASS) method was created to match pathological and segmented MS images [65] in order to determine correspondences between the two images in a semi-supervised way. The AMASS method has helped distinguish between anatomical regions in slices of rat brain and enabled the discovery of peptide masses that are differentially expressed between segmented regions. Recently, the same group published a new segmentation method where m/z images are clustered on the basis of spectral similarity in the pixels [66], enabling pixels exhibiting common ion patterns in the spectra to be clustered together. Fig. 2.6 shows a rat brain segmented in 10 regions (Fig. 2.6C), a spectrum with the ions associated with every segment (Fig. 2.6B), and the spatial pattern of every segment (Fig. 2.6A).

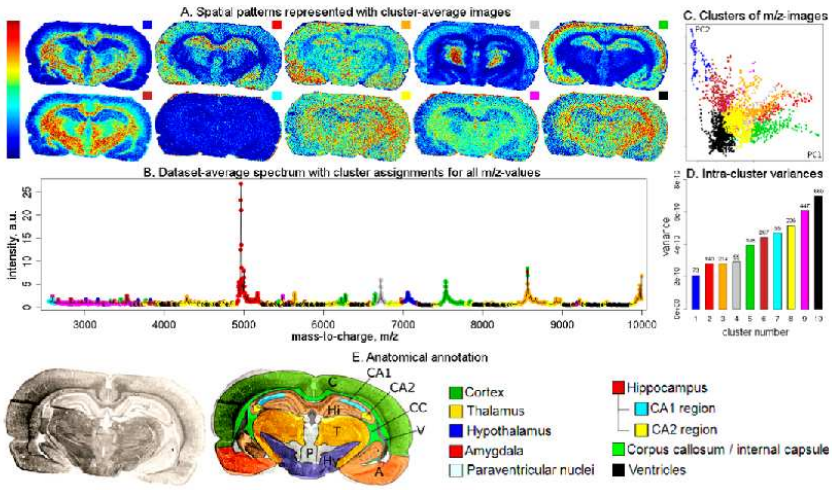


Figure 2.6: Results of the analysis of a MALDI imaging mass spectrometry dataset of a rat brain coronal section, following the proposed approach based on clustering m/z images into 10 clusters according to their spatial similarity. **A)** Cluster-averaged images represent detected spatial patterns. **B)** Dataset-averaged spectrum with assignments of m/z values to the clusters. **C)** Visualization of m/z images in the space of their two first principal components; one dot represents an m/z image, and dots are colored according to their cluster assignments. **D)** Intracluster variances, where the numbers on the top of the bars represent cluster sizes. **E)** Optical image of the section with anatomical annotation provided. Plots A-B show the variety of the spatial patterns among m/z images and help understand how each m/z image looks. Plots C and D help evaluate clustering results [taken from [66] and reprinted with permission from ACS Publications].

2.5.2 3D-image reconstruction

3D-image reconstruction is performed using combinations of images obtained from consecutive tissue slices. One of the main analytical challenges of 3D-image reconstruction is that the extended period required to acquire all tissue sections needed for constructing a full 3D image makes tissue degradation a critical issue. The challenges, approaches, and future research directions associated with 3D images obtained by serial sectioning and MALDI-MS have been extensively reviewed [67].

In 3D-image segmentation and reconstruction, the lack of efficient computational algorithms for data reduction, processing, and visualization of large 3D datasets constitutes a bottleneck. Xiong et al. [68] developed many algorithms for 3D MSI, including data reduction, 2D data alignment, 3D visualization, and statistical analysis for clustering. The morphological features of brain-tissue sections were revealed using a self-organizing feature map ANN on MS images obtained by DESI-MS. Of particular interest was the ability of this method to directly compare 3D images acquired by MALDI-MS and magnetic resonance imaging (MRI), making it possible to match information from morphological and molecular datasets.

A new data-processing pipeline for analyzing and interpreting 3D MALDI-MS images was proposed by Trede et al. [69], which was based on the edge-preserving, de-noising methods developed for 2D-image segmentation, that implements a hierarchical-clustering method called bisecting k-means. The reconstructed 3D images consisted of 33 serial sections of mouse kidney at 3.5 μm thickness acquired at a resolution of 50 μm. More than half a million spectra were acquired, representing >50 GB of data. The computational pipeline showed the anatomical structure of the kidney following correct alignment of the 2D sections, as well as molecular-mass co-localization at major anatomical regions. The same group used the PAXgene tissue container [70] and paraffin embedding to preserve tissues, with results similar to those obtained on frozen samples. The same publication compared the MRI images of a mouse kidney with 3D MS images, enabling reconstruction of the anatomical structure.

2.5.3 On the uses of PCA in MSI

PCA is likely the most often used algorithm in multivariate analysis and MSI applications. There are four primary uses of PCA in MSI analysis: exploratory analysis, data compression, clustering-performance assessment in unsupervised strategies, and biomarker identification. These four uses are illustrated in real examples in Fig. 2.7.

1. **Exploratory analysis.** Exploratory analysis of MS images constitutes the most frequent use of PCA. PCA can be used for the assessment of ionization-source drift as shown in Fig. 2.7A. Another application of PCA could be the detection of outlier pixels denoting possible hot-spots or holes in a tissue section.

2. **Data compression.** The high dimensionality of the pixel spectra results in files with large dimensionality in MSI experiments. A direct method for data compression consists of transformation of the original variables into the principal components of the PCA. In general, most of the variance can be retained in five principal components. Fig. 2.7B depicts a RGB brain MSI image considering the three principal components. The main problem associated with this data-compression approach is that the information concerning individual MS ions is lost.

3. **PCA for clustering-performance assessment in unsupervised strategies.** The validation of unsupervised-clustering results is difficult due to the lack of alternative methods for their comparison. The representation of the pixels in a PCA, labeled according to the cluster to which they belong, offers an estimation of clustering-algorithm efficacy. Generally, the higher the pixel separation between clusters, the better the performance of the clustering process. Fig. 2.7C shows an example of PCA used to validate an in-house-developed clustering method.

4. **Biomarker identification.** In targeted strategies that use references of histological (or microscopy) images, it is common to compare pixels between different regions of interest (i.e., healthy and tumorous regions). The analysis of the PCA loading is a powerful technique enabling identification of the most influential ions in the pixel separation (Fig. 2.7D).

2.6 Data handling strategies and considerations

Because MSI data consists of a large collection of mass spectra corresponding to the spatial location of each pixel of the tissue acquired, the amount of data produced in this kind of experiment tends to be very large, and, therefore, the computational strategies required to handle data processing are complex. In this section, we have divided these strategies into three sections: data formatting, processing requirements, and data-reduction strategies, including peak-picking, feature-selection, and data-compression strategies.

2.6.1 Data formatting

As previously mentioned, for a single imaging experiment, a MALDI-MS instrument generates a large amount of raw data. In most cases, the format used to store the acquired data is determined by the instrument manufacturer and is only compatible with a few supported software tools generally provided by the same manufacturer. Such proprietary formats usually force the end user to adopt the data-processing workflow defined by the software producer, limiting flexibility in the data-analysis process. Fortunately, the main MSI open-data format imzML [71], has started to become a standard used throughout all software platforms, with almost every manufacturer currently offering some level of compatibility.

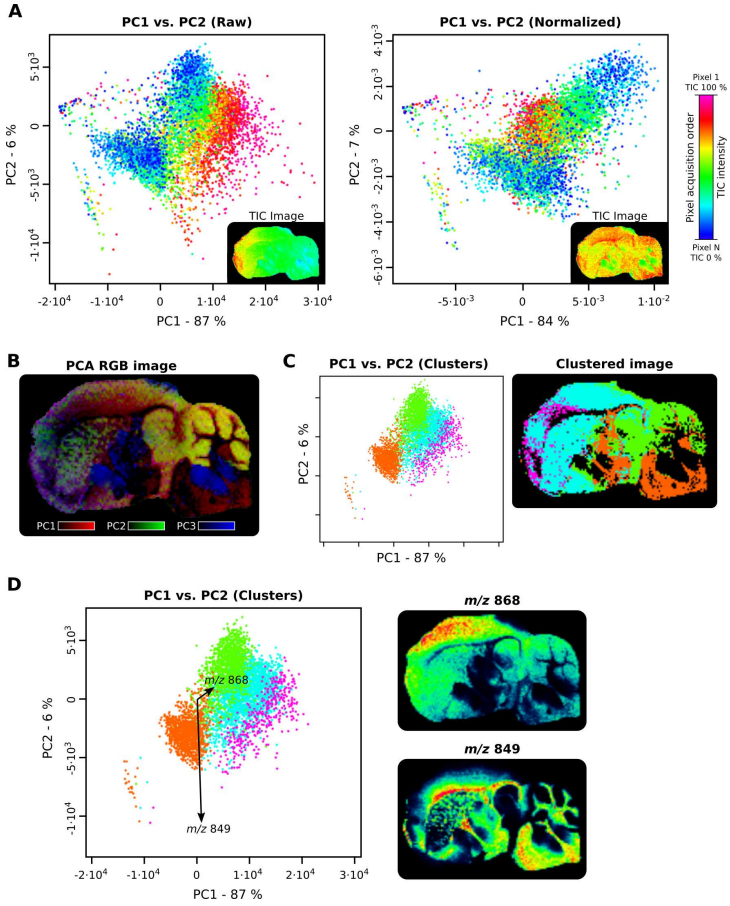


Figure 2.7: Several uses of PCA in MSI experiments. PCA is computed using data from the same experiment shown in Fig. 2.4. PCA was calculated from a peak list generated using alignment, calibration, peak picking, and peak binning as pre-processing steps. **A)** PCA as a tool for exploratory MSI analysis. Comparison of PCA performed before (left) and after (right) spectra normalization. We used AcqTIC (described in Fig. 2.4D) to compensate for ionization-source degradation. In both cases, PC1 versus PC2 is plotted, coloring each data point according to its order during acquisition. In PCA with no normalization, PC1 is affected by intensity degradation during acquisition. In PCA after normalization, data points do not follow the acquisition pattern compensating for intensity degradation effects. **B)** PCA as a compression tool. An RGB image is built using PC1, PC2, and PC3 to encode red, green, and blue colors, respectively. This RGB image shows how PCA is able to compress almost all of the information using only three components, facilitating tissue-region localization. **C)** PCA used for image-segmentation evaluation. In the clustered image (upper), the four larger clusters determined by an in-house image-segmentation method are shown. Pixels are colored according their cluster. The PCA plot shows an excellent separation between the pixels of different clusters, confirming the good performance of the image-segmentation technique. **D)** PCA for biomarker-compound discovery. Two loadings contributing to group separation are selected for illustration (ions m/z 868 and m/z 849) of the PCA plot. The corresponding ion-intensity maps reveal a high degree of complementarity between the intensity of both ions, indicating its influence in the identification of different morphological areas in brain tissue.

Public-domain data formats, such as plain-text files (ASCII), mzML [72], or Analyze7.5 (included in the BioMap software), can be used to exchange MSI data. However, these formats have not been developed specifically for MSI and, there-

fore, present some limitations. In the case of plain text files (ASCII), the MS image is converted into a collection of text files where each file contains a spectrum associated with a raster spot. This strategy is very straightforward, but requires much more hard-drive space as compared to that required for binary data formats, and spectra take longer to parse. Additionally, the portability of MSI experiments to ASCII files is not efficient, because this format does not support storage of metadata (i.e., the type of experiment) and other significant information, such as pixel coordinates associated with each spectrum. The mzML [72] format improves the situation by adding binary data formatting and enabling the storage of standardized metadata. However, mzML was not created for MSI, and some important information, such as raster positions, are not supported. Finally, the Analyze7.5 format was initially designed to store medical 3D images, but is also an export option in some MSI-software packages. Analyze7.5 stores 2D information with the raster positions, and 3D fields are populated with mass spectra. Analyze7.5 is optimal for storing MSI data in a compact way, but it presents limitations in metadata storage, because it was not designed for MS applications. Moreover, mass spectra intensities are often encoded in 16-bit integers when exporting to Analyze7.5 in some software tools, reducing the accuracy of the original intensity axis encoded in 32-bit integers by the instrument detector.

To overcome these limitations, an open standard has been developed under the name imzML [71], which aims to become the global MSI reference-exchange format. The openness refers to the data-formatting specification used, which is fully detailed and available in open-access format that enable everyone to access and implement imzML support. Recently, imzML began being incorporated as a data-export option in many proprietary software packages, with some developers creating third-party tools to facilitate data conversion to imzML. An example of this effort to promote imzML is the "imzMLConverter" tool, which converts files from mzML format to imzML [73].

2.6.2 Processing requirements

Data acquired from an imaging experiment represents a collection of spectra, the size of which depends upon the scanned area, spatial resolution, and mass-spectra range and resolution. Each spectrum consists of a vector of intensities generally encoded in 32-bit integer numbers, and the number of data points in each spectrum depends upon the resolution of the spectrometer. Furthermore, the raw data

size is proportional to the number of pixels in the image multiplied by the number of points in each spectrum. As an example of TOF-acquired data, an MS image at 100 pixels × 100 pixels with 50,000 points per spectrum consumes ~1.86 GB (100 × 100 pixels × 50,000 points × 4 bytes) when it is fully loaded in computer memory. Because pixels are arranged in a bi-dimensional space, an increase in image size is expressed as a scan-area expansion leading to a dramatic amount of memory usage. For instance, doubling the image size in X and Y dimensions as in the previous example produces ~7.45 GB of data. This memory requirement indicates the physical limitations of computer memory, making it difficult to efficiently handle such volumes of data. In addition to memory requirements, processing time and CPU use must also be taken into consideration. The large amounts of data produced by MSI experiments require heavy processing resources, especially when complex multivariate statistical algorithms are used.

Most mathematical and statistical software packages use a load-and-process approach where all data is first loaded into random access memory (RAM) and then processed with the desired algorithms. Although this approach makes it easy to handle the data, it requires large quantities of RAM. Moreover, most of these packages are based on interpreted languages that work in a mono-task approach. Due to these processing requirements, an interesting approach may be to adopt parallel-processing approaches to benefit from modern multicore-processor systems. However, many statistical libraries have not been designed to support multithreading; therefore, parallelization is often not an out-of-the-box solution. A possible strategy to overcome this could be splitting the data into fragments and processing each one in a different instance of the mono-task program. This approach was demonstrated as being effective when a processing platform with many processors is available [74]. Nevertheless, some packages are available for parallel processing, including those of R and MATLAB. These packages make the multithreaded implementation of algorithms more straightforward. Processing infrastructures that use graphics processing units have been tested with multithreaded algorithms, resulting in reduced computation times [75]. However, this reduction occurs at the expense of flexibility and simplicity.

2.6.3 Data-reduction strategies

Due to the high computational resources required for processing large MS-image datasets, data-reduction techniques play an important role. The goal of data re-

duction is to extract relevant information from the dataset, while minimizing both the memory footprint and information loss. A common approach involves peak picking, which stores only spectral peaks in a reduced memory space. Following peak picking, a feature-selection routine selects the most informative peaks, which helps to reduce the data requiring further processing. However, peak picking is the final step in the pre-processing chain; therefore, pre-processing actions performed prior to this one will not benefit from this first-stage data reduction. Another strategy involves peak binning after peak picking [31]. Once the mass resolution of every peak is determined, all peaks under the mass tolerance are grouped into the same bin. This represents an important reduction in the number of variables.

Each spectrum of an MS image contains peaks and a large collection of zeros and noise; therefore, MSI data can be considered as very sparse. Leendert A. Klerk's research group took advantage of this to develop a method to handle MSI data more efficiently using Harwell–Boeing-formatted matrices [76], where the data matrix was stored in a minimal-memory layout that discarded empty values. However, storing sparse matrices does not degrade the information, because data is retained, and null data points are prevented from being stored. In this scenario, the memory footprint is reduced further by associating each peak to a mass bin in the TOF domain. This method is useful for reducing memory requirements and computation time, although most of the algorithms found in commonly used libraries are not designed to handle sparse matrices. Consequently, the main drawback of these data structures is that alternative algorithm implementations must be written.

Methods of data compression based on raw-data transformations make the algorithms currently used for MSI experiments more efficient (in terms of computational resources). Using these methods, the spectral data is transformed by reducing dimensionality, but keeping the fundamental information. Processing is then executed in this transformed space. The results must be transformed back to the original space if understandable information is to be obtained. Using this workflow, Van de Plas et al. [55] demonstrated that performing a DWT for each spectrum effectively reduced the computational requirements when only larger wavelet coefficients were retained. Furthermore, when a PCA was computed in this reduced DWT-transformed space instead of using spectral data directly, the results were much more accurate than those acquired in a native data space due to the inherent dimensionality reduction achieved by DWT.

Other strategies rely on dimensionality reduction algorithms, such as PCA [77]. These algorithms transform the original variables (i.e., ions) into new ones using linear combinations of the original ions to maximize the variance using as few variables as possible. Here, data reduction is accomplished by removing less-significant variables from the dataset. An associated problem is that the new variables are not directly associated with a particular molecule, making biomarker discovery difficult.

Due to the complexity of processing data for a full MS image, manual segmentation is often chosen. In this case, some ROIs are drawn following the manually discovered patterns in an image. Spectra are then extracted from these images and used as input for data processing. Despite the simplicity and reduced processing time associated with this workflow, results are only obtained for some parts of the image, rendering the rest of the dataset meaningless. However, this approach may be useful for rapidly profiling well-known regions [28, 41].

2.7 MSI software packages

Recently, various software tools were developed to explore the data produced by MSI instruments and obtain biological tissue information. MALDI-MS equipment usually comes with dedicated software used to control acquisition and perform common imaging tasks. However, in some cases, the software supplied by the manufacturer might not fulfill all of the processing requirements. In such cases, functionality can be increased by including extra software packages in the processing chain. These packages can be obtained from the instrument manufacturer or third-party providers. In this section, we classified the available software packages according to their licensing agreements and features: commercial, freeware, and open source. Commercial tools are private software packages generally developed by companies and can only be used if a license is purchased. Freeware tools are also private, but are under a licensing agreement that allows their use free of charge in some situations. In contrast, an open-source tool provides access to the source code and is very often free of charge.

Below, we discuss the more common software tools that can work with MSI datasets. Each software-licensing group is introduced, describing its weakness and strengths. We hope that this discussion helps decide whether a given tool will be useful for a specific application. The main differences between these tools are

summarized in the appendix tables 2.5, 2.6 and 2.7, including input/output data format, build-in processing, and supported platforms.

2.7.1 Commercial software tools

Usually, commercial software tools come with the MSI instrument and provide the necessary functions to control acquisition and visualization of raw data. Despite its cost, commercial software tools are often the most user-friendly solution, enabling anyone without in-depth knowledge of MSI data to visualize the results.

FlexImaging

FlexImaging is the software portion of Bruker's imaging platform. FlexImaging provides a graphical front-end for user-friendly control of data acquisition and visualization. Images of various ions can be represented within a defined tolerance and combined with an optical image of the sample.

FlexImaging delegates the processing of the raw spectra to FlexAnalysis. By doing so, Bruker takes advantage of a long list of well-known algorithms implemented in FlexAnalysis, including baseline correction, normalization, and calibration. For statistical analysis, FlexImaging is designed to easily interface with Bruker ClinProTools, which can perform multivariate calculations, such as PCA, SVM, or hierarchical clustering, as well as univariate statistical tests. The results generated with ClinProTools can be plotted using FlexImaging. These results can then be mapped over the image to view the spatial distribution of the processed data. FlexImaging performs all computations using its own proprietary data format and allows the export of data using open formats, including ASCII, Analyze7.5, and imzML (since v4.1).

SCiLS Lab

SCiLS lab is designed for use with the Bruker platform and is part of Bruker's MALDI Molecular Imager solution. SCiLS imports data from FlexImaging in Bruker's native imaging format and can export results to an Excel spreadsheet and also back to FlexImaging. Because SCiLS is able to exchange data only with Bruker's platforms, its use is limited and should be considered as an extension of the Bruker imaging platform.

SCiLS Lab is a full-featured integrated solution for straightforwardly visualizing and statistically analyzing MALDI MSI data in order to make it more readily interpretable. It is able to perform common pre-processing steps, as well as univariate and multivariate statistical analyses. Additionally, it can spatially cluster biologically different tissue regions using supervised or unsupervised approaches. For supervised clustering, the algorithm learns patterns from user-defined tissue regions to obtain a segmentation map.

MALDIVision

MALDIVision is a platform-independent tool that is particularly strong in data visualization. In order to be compatible with most MSI instruments in the market, it uses standard file formats (Analyze7.5 and imzML) to import data. Multiple images can be overlaid and mapped to different colors to enable comparison of spatial distributions of selected ions or combined with optical images to perform histological validation. The images can also be displayed at an intensity normalized to that of a standard compound assumed to be homogenously distributed in tissue.

This software can calculate such typical statistical parameters as mean, median, or standard deviation from user-defined areas, and can also perform more advanced tasks, including the production of histograms and cumulative-probability graphs, to visualize ion-intensity distribution. Many features provided by MALDIVision are also available in some freeware and open-source tools, making MALDIVision an effective solution for simple MS-data exploration.

TissueView

TissueView is an MSI tool from Sciex (Framingham, MA, USA). The program can handle imaging data directly from instruments made by the same manufacturer or from Analyze7.5 files. The tool focuses on image visualization and can represent a single mass-ion bin by mapping the intensity onto a color scale. Additionally, up to three ions can be co-localized with each intensity being coded in a RGB-color channel and can also import optical images that can be overlaid with MS images. For data-processing purposes, TissueView can calculate the average spectrum and provide the ion distribution in a particular tissue region. However, for more advanced data analysis, spectra can be interfaced with Sciex MakerView to perform statistics with tools, such as PCA. The software can also

load the data in Sciex Data Explorer to identify the proteins using a Mascot server (http://www.matrixscience.com/server.html).

ImageQuest

ImageQuest is the MSI-visualization tool used by Thermo Fisher Scientific (Waltham, MA, USA) MALDI instruments and reads data in the raw data format used by the manufacturer. This program provides various image-reconstruction alternatives for representing spatially mapped ion intensity. To aid navigation, the optical image used during acquisition is also displayed, but not overlaid, with the MS image.

To rapidly identify which raster positions contain relevant information, ImageQuest introduces a plot window named "chromatogram", where the overall intensity of each pixel is represented versus each scan. Clicking on a chromatogram peak prompts ImageQuest to show where each scan is located on the 2D image, as well as its spectrum. The visualization can also be animated to find unknown peaks. In this mode, ImageQuest will scroll automatically through the defined mass range, enabling the user to observe how the image evolves for each selected mass.

High-definition imaging (HDI)

HDI is the integrated MSI software solution by Waters Corporation (Milford, MA, USA). It is designed to interface with Waters mass spectrometry instruments in a unified way, from data acquisition to processing and visualization, and is capable of exporting to standard formats, such as imzML and ASCII. Various images from different ions can be represented simultaneously, and images can be reconstructed from a given mass range, a peak selected from an automatically generated list, or a combination of three overlaid images mapped onto an RGB-color space.

In addition to image visualization, HDI also focuses on discovering meaningful information behind the data. In this regard, the typical pre-processing algorithms are included, as well as a set of statistical tools, such as PCA, PLS-DA, S-plots, and hierarchical clustering.

Quantinetix

Quantinetix is suited for specific molecule quantification in MALDI MSI experiments. It was designed to support a wide range of formats, enabling its integration into almost any instrument workflow. The data formats supported range from standard imzML and Anaylze7.5 to native proprietary formats, such as those used by Bruker, Sciex, Thermo Fisher Scientific, and Waters Corporation.

To accurately quantify a compound, three normalization techniques enable users to choose which one best fits their needs. These algorithms include on-tissue dilution, isotopic labeling, and Ion suppression. MS images can be overlaid with optical images and are generated from single-ion intensity distributions or with multiple ions assigned to various colors. In addition to the image representation, the tool also provides plot windows showing information about the quantification and normalization algorithms.

2.7.2 Freeware software tools

Freeware software tools have been widely used as a zero-cost solution for data visualization, providing a frontend for exchanging MSI data through various collaborators. However, freeware tools are limited in MS-data processing, and, consequently in some situations, they are not a viable alternatives to commercial tools.

msiQuant

msiQuant [28, 78] is a tool for assisting the labeled normalization and quantitation of drugs and neuropeptides directly in tissue sections. It includes a data-processing chain carefully designed to minimize peak alterations. The baseline correction is implemented with its novel sorted mass spectrum transform algorithm, and, depending on the features of the sample, the normalization approach can be chosen from four algorithms. The software is designed to work within user-defined ROIs and discard meaningless data, thereby saving computer resources.

In terms of data-importing facilities, it can load images from the original Bruker file format and imzML. This software should be considered as an alternative to Quantinetix and is a low-cost solution; however, it also introduces some novel algorithms than can improve quantification in some cases.

BioMap

BioMap provides a visualization platform that supports image modalities, such as optical, positron emission tomography, computed tomography, near-infrared fluorescence, and MSI. This makes it possible to combine images generated from several experimental techniques; however, because BioMap is not a MALDI-MS-specific software package, it lacks typical MS-processing algorithms. It can be extended by adding modules written in interactive data language (IDL) and capable of analyzing specific data. Because BioMap is a general imaging solution that does not specifically target MSI, the file format used for data storage is Analyze7.5.

Despite the frequent use of BioMap in MSI analysis, it also results in frequent memory errors during the processing of large MS datasets. This suggests that all processing is performed in RAM, making this a sub-optimal solution for use with current high-resolution MS images. Moreover, its execution in an IDL environment could complicate the installation procedure for the average user.

Datacube Explorer

Datacube Explorer [79] is an MSI-visualization tool that also includes the capability to perform clustering on images using Kohonen map algorithms. Despite the possibility of the Kohonen map-segmentation feature not being useful for many users, this package is compatible for the future integration of other algorithms. It is also capable of reconstructing 2D images by selecting a particular ion, although 3D reconstruction is also possible when a dataset contains a collection of consecutive tissue slices.

Datacube Explorer supports standard open formats (ImzMl and Analyze7.5), but also includes its own format optimized for better handling of large datasets. Datacube Explorer is an optimal low-cost alternative for simple exploration of MS data.

Mirion

Mirion [80] is an image-exploration tool that supports importing data from proprietary formats (XCalibur; Thermo Fisher Scientific) and the imzML format. Images can be generated from manual ion selection or automatically using an embedded algorithm based on the selection of the most repetitive peak. This automatic feature uses a mass histogram generated from the full dataset to select

the more dominant peaks. Mirion also enables ion images to be combined with optical images, forming a multilayer image that can be represented using different color channels.

Currently, Mirion is limited to run under 32 bits thus its memory is limited at 2 GB. Such limitations prevent exploration of many MS images acquired by modern instruments with high mass and/or spatial resolution. In such situations, Mirion offers the possibility of loading only a part of the data in order to explore it.

OpenMSI

OpenMSI [81] is a web platform that provides an application program interface plus an interface to retrieve and explore MSI data. Their website can be used to upload MSI data and explore it anywhere using a computer connected to the internet and without the requirement of specific software tools. Uploading MSI data to a web server overcomes storage problems derived from performing many acquisitions. Moreover, sharing data using a web browser through their OpenMSI interface drastically simplifies manual MS-data exploration in a large research groups. However, OpenMSI currently lacks the processing tools required for MSI-data analysis.

OpenMSI provides a file format based on the HDF5 format that is highly optimized for efficient storage and access of MSI information. Data is stored in chunks to improve input/output performance, with these chunks compressed into the lossless GZIP format to reduce network bandwidth required for file transfer and provide for efficient storage. Furthermore, to enable rapid access of individual spectra, data is replicated to overcome linearized binary format limitations. Despite data replication, the final stored data size is still compressed to reduce raw data size.

2.7.3 Open-source software tools

Open-source software tools are a great option for low-budget MSI-data analysis. Due the fact that anyone can read and modify the code of open-source software, advanced users can adapt them to their specific requirements, enabling anyone with a programing background to expand an open-source software tool to satisfy their processing pipeline. Most open-source tools used for MSI analysis execute under platforms, such as MATLAB or R, making some knowledge of these environments necessary. Because many users are unfamiliar with programing, some freeware

tools may be a better choice in situations where raw data needs to be visualized. However, the experienced user will likely discover the most flexible and powerful solutions using open-source tools.

MSiReader

MSiReader [82] is a tool developed in MATLAB and provides a full-featured graphical user interface for the loading and visualization of MSI data from various file formats, including mzXML, imzml, Analyze7.5, and ASCII. Data can be represented in a user-friendly way by selecting a representative ion and a mass tolerance. It also has processing capabilities supporting baseline correction, normalization to a specific peak or by TIC, peak picking, and background subtraction. Features can also be automatically extracted by selecting the most abundant peaks in a selected ROI. Despite its processing tools, MSiReader also enables individual spectra to be exported and custom processing algorithms to be integrated into the MATLAB environment. Despite MSiReader being open source, it has been implemented into the MATLAB platform, which is neither open nor free.

OmniSpect

OmniSpect [83] performs computationally intensive functions on a remote server, with the functions divided into data-converting tools and multivariate-analysis algorithms for MSI datasets. Similar to MSiReader, OmniSpect makes intensive use of the MATLAB environment to perform calculations; therefore, its code is open-source, but the runtime execution requires a proprietary backend. This tool can import data from most common imaging formats, including NetCDF, mzXML, imzML, and Analyze7.5, and convert it into a MATLAB representation, after which the user can select up to three ions to represent each image. Moreover, OmniSpect can perform multivariate analysis using the NMFA algorithm to detect similarities in spatial-ion distributions. OmniSpect provides a web interface to represent information and facilitate control, with such remote-processing features very useful when various users require analysis of MSI data. However, enabling a server to utilize MATLAB in a web interface may be more complicated than running a simple standalone program on a personal computer. Given this potential limitation, each user must consider whether remote processing will be beneficial in each situation.

Cardinal

Cardinal [84] is a software package for the R environment that enables data import using two standard formats: imzML and Analyze7.5. This toolset is built based on the R language and is distributed using the Bioconductor website (https://www.bioconductor.org/). It does not provide a unified graphical user interface to manipulate visualizations; however, many functions are available to enable direct execution in an R session or use in an R script file. Such functions include MSI-data-loading routines, pre-processing tools, segmentation and classification algorithms, and image visualization. Despite lack of a user-friendly command interface, the package provides adequate documentation. Moreover, the availability to create script files and mix Cardinal code with other R packages provides a powerful platform for MSI-data processing. Comparing Cardinal with MATLAB-based solutions, the open-source nature of the program and the language used to create it makes Cardinal the most cost-efficient solution. Furthermore, integration in a growing R environment with plenty of free packages containing multiple algorithm implementations makes Cardinal a suitable choice for advanced users.

2.8 Final conclusion

In the previous decade, MSI became a key technique used for molecular analysis of biological tissues due to its ability to locate ions in space (drugs, metabolites peptides, or proteins). Greater sophistication in sample preparation and improvements in MALDI-MS instruments have resulted in acquisition of high-quality MS images with resolutions ranging from 2 to 200 μm, making this technique useful for clinical diagnosis. The processing of MSI data remains challenging, with researchers confronted with alterations in the distribution of peak intensities caused by possible inhomogeneity in the organic matrix distribution between pixels (in the case of MALDI applications) or tissue inhomogeneity, effects of ion suppression, or reductions in ionization efficiency throughout extended imaging experiments. In such complex scenarios, the use of bioinformatics strategies for MSI analysis are mandatory, with the main conclusions from this review as follows:

- *The relevance of pre-processing steps.* Pre-processing stages compensate for variations and noise in raw data. Currently, there are a wide variety of available pre-processing algorithms whose suitability depends upon the purpose of

each MSI experiment. Peak alignment across all pixels is a crucial step necessary for well-resolved spectral images. When internal calibration signals are not available, peak-alignment strategies are based on cross-correlation of image pixels. If internal-calibration peaks are available (as in the case of matrix-free LDI), they can be used both for peak alignment and mass calibration. An accurate mass-calibration operation is essential for compound identification. Another critical pre-processing step is intensity spectra normalization. The TIC algorithm is effective at compensating for matrix inhomogeneity, but can lead to distorted images when there are different biological areas present in tissue. If an internal-calibration signal is available, it can be used for intensity calibration. Statistical algorithms can provide accurate results, but they are more difficult to implement and have additional computational requirements.

- *Fusion of images in supervised-classification algorithms.* We reported numerous studies using classical microscopy images as referenced for training multivariate models for the segmentation and classification of molecular images. The simultaneous interpretation of the two kinds of images, with the former providing high spatial resolution and pathological interpretation and the latter molecular information, could lead to a new generation of "fused imaging" strategies or techniques.

- *Molecular images for clinical diagnosis.* The ability to rapidly acquire and characterize MS images of tissues (i.e., <1 h) could enable a myriad of new applications for clinical diagnosis. The automation of matrix-deposition techniques, together with the increase in the frequent use of UV-pulsed lasers in MALDI-MS instruments, opens up many possibilities, including tumor recognition in clinical practices. However, different tissue samples have been compared with limited success, and real-time multivariate algorithms for tissue-image segmentation and classification need to be developed in the future.

- *Computational strategies.* Researchers developing bioinformatics tools for MSI analysis need to design and implement smart and powerful computational strategies due to the high-dimensionality of MSI-datasets, especially when images are taken at high spatial resolution. Currently, almost any high-level programing platforms and languages, such as R, MATLAB, and

Phyton, include libraries supporting parallel programming. Such resources would be beneficial at reducing the computation time necessary for MSI analyses. Moreover, data processing should be optimized to enable a smaller memory footprint. Compiled programming languages, including C, C++, C#, or Java, enable memory to be controlled more carefully as compared with interpreted platforms, such as R, Phyton, or MATLAB. However, interpreted languages usually provide larger algorithm libraries and higher abstraction layers. Data-reduction algorithms (i.e., binning, peak picking, sparse- matrix storing, etc.) are also desirable based on their reductions of computational load.

- *Software tools.* Many software packages suitable for MSI processing and visualization are currently available. In general, proprietary software tools are user-friendly and provide adequate features; however, open-source tools enable scalability and flexibility, with some providing unique processing methods. The primary bottleneck is the lack of compatibility between different software packages, which complicates the exchange of data. Nevertheless, several LDI-instrument software tools currently include imzML-exporting features as an attempt to overcome this lack of compatibility.

References

[1] Franz Hillenkamp and Jasna Peter-Katalinic. *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*. 2014.

[2] Tyler Greer, Robert Sturm, and Lingjun Li. "Mass spectrometry imaging for drugs and metabolites". In: *J. Proteomics* 74.12 (Nov. 2011), pp. 2617–2631.

[3] Brendan Prideaux and Markus Stoeckli. "Mass spectrometry imaging for drug distribution studies". In: *J. Proteomics* 75.16 (2012), pp. 4999–5013.

[4] Ruben D. Addie et al. "Current State and Future Challenges of Mass Spectrometry Imaging for Clinical Research". In: *Anal. Chem.* 87.13 (July 2015), pp. 6426–6433.

[5] Michaela Aichler and Axel Walch. "MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice". In: *Lab. Investig.* 95.4 (Apr. 2015), pp. 422–431.

[6] Jörg Kriegsmann, Mark Kriegsmann, and Rita Casadonte. "MALDI TOF imaging mass spectrometry in clinical pathology: A valuable tool for cancer diagnostics (Review)". In: *Int. J. Oncol.* 46.3 (Dec. 2015), pp. 893–906.

[7] Anna Nilsson et al. "Mass Spectrometry Imaging in Drug Development". In: *Anal. Chem.* 87.3 (2015), pp. 1437–1455.

[8] Jeremy L. Norris and Richard M. Caprioli. "Analysis of Tissue Specimens by Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry in Biological and Clinical Research". In: *Chem. Rev.* 113.4 (Apr. 2013), pp. 2309–2342.

[9] Emrys A Jones et al. "Imaging mass spectrometry statistical analysis." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 4962–89.

[10] Theodore Alexandrov. "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges." In: *BMC Bioinformatics* 13 Suppl 1.Suppl 16 (2012), S11.

[11] Herbert Thiele et al. "2D and 3D MALDI-imaging: Conceptual strategies for visualization and data mining." In: *Biochim. Biophys. Acta* 1844.1 Pt A (Jan. 2014), pp. 117–37.

[12] Richard J.A. Goodwin. "Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 4893–4911.

[13] Kamila Chughtai and Ron M a Heeren. "Mass spectrometric imaging for biomedical tissue analysis". In: *Chem. Rev.* 110.5 (2010), pp. 3237–3277.

[14] Achim Buck et al. "High-resolution MALDI-FT-ICR MS imaging for the analysis of metabolites from formalin-fixed, paraffin-embedded clinical tissue samples". In: *J. Pathol.* 237.1 (2015), pp. 123–132.

[15] Nidia Lauzon et al. "Development of Laser Desorption Imaging Mass Spectrometry Methods to Investigate the Molecular Composition of Latent Fingermarks". In: *J. Am. Soc. Mass Spectrom.* 26.6 (2015), pp. 878–886.

[16] Erin Gemperline, Stephanie Rawson, and Lingjun Li. "Optimization and Comparison of Multiple MALDI Matrix Application Methods for Small Molecule Mass Spectrometric Imaging". In: *Anal. Biochem.* 86.9 (2014), p. 1003010035.

[17] Suming Chen et al. "2,3,4,5-Tetrakis(3,4-dihydroxylphenyl)thiophene: A New Matrix for the Selective Analysis of Low Molecular Weight Amines and Direct Determination of Creatinine in Urine by MALDI-TOF MS". In: *Anal. Chem.* 84.23 (Dec. 2012), pp. 10291–10297.

[18] Nina Bergman, Denys Shevchenko, and Jonas Bergquist. "Approaches for the analysis of low molecular weight compounds with laser desorption/ionization techniques and mass spectrometry". In: *Anal. Bioanal. Chem.* 406.1 (2014), pp. 49–61.

[19] Yuliya E Silina and Dietrich a Volmer. "Nanostructured solid substrates for efficient laser desorption/ionization mass spectrometry (LDI-MS) of low molecular weight compounds". In: *Analyst* 138.23 (2013), p. 7053.

[20] Cheng-Kang Chiang, Wen-Tsen Chen, and Huan-Tsung Chang. "Nanoparticle-based mass spectrometry for the analysis of biomolecules." In: *Chem. Soc. Rev.* 40.3 (Feb. 2011), pp. 1269–1281.

[21] Trent R Northen et al. "Clathrate nanostructures for mass spectrometry." In: *Nature* 449.7165 (Oct. 2007), pp. 1033–6.

[22] Martin Dufresne et al. "Silver-Assisted Laser Desorption Ionization For High Spatial Resolution Imaging Mass Spectrometry of Olefins from Thin Tissue Sections". In: *Anal. Chem.* 85.6 (Mar. 2013), pp. 3318–3324.

[23] Liam A McDonnell and Ron M A Heeren. "Imaging mass spectrometry." In: *Mass Spectrom. Rev.* 26.4 (2007), pp. 606–43.

[24] Matthias Rainer, Muhammad Nasimullah Qureshi, and Günther Karl Bonn. "Matrix-free and material-enhanced laser desorption/ionization mass spectrometry for the analysis of low molecular weight compounds". In: *Anal. Bioanal. Chem.* 400.8 (2011), pp. 2281–2288.

[25] Raul Calavia et al. "Nanostructure Initiator Mass Spectrometry for tissue imaging in metabolomics: future prospects and perspectives." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5061–8.

[26] Dirk Hölscher et al. "High resolution mass spectrometry imaging reveals the occurrence of phenylphenalenone-type compounds in red paracytic stomata and red epidermis tissue of Musa acuminata ssp. zebrina cv. 'Rowe Red'". In: *Phytochemistry* 116 (Aug. 2015), pp. 239–245.

[27] Marcel van Herk. "A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels". In: *Pattern Recognit. Lett.* 13.7 (July 1992), pp. 517–521.

[28] Patrik Källback et al. "Novel mass spectrometry imaging software assisting labeled normalization and quantitation of drugs and neuropeptides directly in tissue sections." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 4941–51.

[29] Dante Mantini et al. "LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise." In: *BMC Bioinformatics* 8.1 (Jan. 2007), p. 101.

[30] Glen a. Satten et al. "Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens". In: *Bioinformatics* 20.17 (2004), pp. 3128–3136.

[31] Jeremy L. JL Norris et al. "Processing MALDI mass spectra to improve mass spectral direct tissue analysis". In: *Int. J. Mass Spectrom.* 260.2-3 (Feb. 2007), pp. 212–221.

[32] Abraham. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." In: *Anal. Chem.* 36.8 (July 1964), pp. 1627–1639.

[33] Michael Hanselmann et al. "Toward digital staining using imaging mass spectrometry and random forests." In: *J. Proteome Res.* 8.7 (July 2009), pp. 3558–67.

[34] Maureen B Tracy et al. "Precision enhancement of MALDI-TOF MS using high resolution peak detection and label-free alignment." In: *Proteomics* 8.8 (Apr. 2008), pp. 1530–8.

[35] Taryn M Guinan et al. "Silver Coating for High-Mass-Accuracy Imaging Mass Spectrometry of Fingerprints on Nanostructured Silicon". In: *Anal. Chem.* 87.22 (Nov. 2015), pp. 11195–11202.

[36] Q Peter He et al. "Self-calibrated warping for mass spectra alignment." In: *Cancer Inform.* 10 (Jan. 2011), pp. 65–82.

[37] Gregor McCombie et al. "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis." In: *Anal. Chem.* 77.19 (Oct. 2005), pp. 6118–6124.

[38] Judith M. Fonville et al. "Robust Data Processing and Normalization Strategy for MALDI Mass Spectrometric Imaging". In: *Anal. Chem.* 84.3 (Feb. 2012), pp. 1310–9.

[39] Sören-Oliver Deininger et al. "Normalization in MALDI-TOF imaging datasets of proteins: practical considerations." In: *Anal. Bioanal. Chem.* 401.1 (July 2011), pp. 167–81.

[40] Kirill A Veselkov et al. "Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer". In: *Proc. Natl. Acad. Sci.* 111.3 (Jan. 2014), pp. 1216–1221.

[41] Gregory Hamm et al. "Quantitative mass spectrometry imaging of propranolol and olanzapine using tissue extinction calculation as normalization factor." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 4952–61.

[42] Theodore Alexandrov et al. "Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering." In: *J. Proteome Res.* 9.12 (Dec. 2010), pp. 6535–46.

[43] Loic Denis, Dirk A Lorenz, and Dennis Trede. "Greedy Solution of Ill-Posed Problems: Error Bounds and Exact Inversion". In: *Inverse Probl.* 25.11 (2009), p. 115017.

[44] Do Yup Lee et al. "Resolving brain regions using nanostructure initiator mass spectrometry imaging of phospholipids." In: *Integr. Biol. (Camb).* 4.6 (June 2012), pp. 693–9.

[45] M Reid Groseclose et al. "High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry". In: *Proteomics* 8.18 (Sept. 2008), pp. 3715–3724.

[46] Lipo Wang. *Support Vector Machines: Theory and Applications*. Springer Science & Business Media, 2005, p. 431.

[47] Joseph F. Hair Jr et al. *Multivariate Data Analysis*. Pearson Education Limited, 2013, p. 752.

[48] Sandhya Samarasinghe. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*. CRC Press, 2006, p. 570.

[49] Liam A McDonnell et al. "Mass spectrometry image correlation: quantifying colocalization." In: *J. Proteome Res.* 7.8 (Aug. 2008), pp. 3619–27.

[50] Sandra Rauser et al. "Classification of HER2 Receptor Status in Breast Cancer Tissues by MALDI Imaging Mass Spectrometry". In: *J. Proteome Res.* 9.4 (Apr. 2010), pp. 1854–1863.

[51] T Kohonen et al. "Self organization of a massive document collection." In: *IEEE Trans. Neural Netw.* 11.3 (May 2000), pp. 574–585.

[52] Raf Van de Plas et al. "Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping". In: *Nat. Methods* 12.4 (Feb. 2015), pp. 366–372.

[53] Yi-Tzu Cho et al. "Combining MALDI-TOF and molecular imaging with principal component analysis for biomarker discovery and clinical diagnosis of cancer". In: *Genomic Med. Biomarkers, Heal. Sci.* 4.1-2 (Mar. 2012), pp. 3–6.

[54] Mário A T M.A.T. Figueiredo and Robert D R.D. Nowak. "An EM algorithm for wavelet-based image restoration." In: *IEEE Trans. Image Process.* 12.8 (Aug. 2003), pp. 906–16.

[55] Raf Van de Plas, Bart De Moor, and Etienne Waelkens. "Discrete wavelet transform-based multivariate exploration of tissue via imaging mass spectrometry". In: *Proc. 2008 ACM Symp. Appl. Comput. - SAC '08*. New York, New York, USA: ACM Press, Mar. 2008, p. 1307.

[56] Michael Hanselmann et al. "Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis". In: *Anal. Chem.* 80.24 (Dec. 2008), pp. 9649–9658.

[57] Rasmus Bro. "PARAFAC. Tutorial and applications". In: *Chemom. Intell. Lab. Syst.* 38.2 (Oct. 1997), pp. 149–171.

[58] Pietro Franceschi and Ron Wehrens. "Self-organizing maps: A versatile tool for the automatic analysis of untargeted imaging datasets". In: *Proteomics* 14.7-8 (Apr. 2014), pp. 853–861.

[59] David Bonnel et al. "Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: Application to prostate cancer". In: *Anal. Bioanal. Chem.* 401 (2011), pp. 149–165.

[60] Judith M. Fonville et al. "Hyperspectral Visualization of Mass Spectrometry Imaging Data". In: *Anal. Chem.* 85.3 (Feb. 2013), pp. 1415–1423.

[61] Sören-Oliver Deininger et al. "MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers." In: *J. Proteome Res.* 7.12 (Dec. 2008), pp. 5230–6.

[62] Emrys A. Jones et al. "Multiple Statistical Analysis Techniques Corroborate Intratumor Heterogeneity in Imaging Mass Spectrometry Datasets of Myxofibrosarcoma". In: *PLoS One* 6.9 (Sept. 2011). Ed. by Maria Gasset, e24913.

[63] Theodore Alexandrov and Jan Hendrik Kobarg. "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering." In: *Bioinformatics* 27.13 (July 2011), pp. i230–8.

[64] T Alexandrov et al. "Super-resolution segmentation of imaging mass spectrometry data: Solving the issue of low lateral resolution." In: *J. Proteomics* 75.1 (Dec. 2011), pp. 237–245.

[65] Jocelyne Bruand et al. "AMASS: algorithm for MSI analysis by semi-supervised segmentation." In: *J. Proteome Res.* 10.10 (Oct. 2011), pp. 4734–43.

[66] Theodore Alexandrov et al. "Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity". In: *Anal. Chem.* 85 (2013), pp. 11189–11195.

[67] Andrew D. Palmer and Theodore Alexandrov. "Serial 3D imaging mass spectrometry at its tipping point". In: *Anal. Chem.* 87.8 (Apr. 2015), pp. 4055–4062.

[68] Xingchuang Xiong et al. "Data processing for 3D mass spectrometry imaging". In: *J. Am. Soc. Mass Spectrom.* 23.February (2012), pp. 1147–1156.

[69] Dennis Trede et al. "Exploring Three-Dimensional Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry Data: Three-Dimensional Spatial Segmentation of Mouse Kidney". In: *Anal. Chem.* 84.14 (July 2012), pp. 6079–6087.

[70] Janina Oetjen et al. "MRI-compatible pipeline for three-dimensional MALDI imaging mass spectrometry using PAXgene fixation." In: *J. Proteomics* 90 (Oct. 2013), pp. 52–60.

[71] Thorsten Schramm et al. "imzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5106–10.

[72] Lennart Martens et al. "mzML–a community standard for mass spectrometry data." In: *Mol. Cell. Proteomics* 10.1 (Jan. 2011), R110.000133.

[73] Alan M. Race, Iain B. Styles, and Josephine Bunch. "Inclusive sharing of mass spectrometry imaging data requires a converter for all". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5111–5112.

[74] Donald F. Smith et al. "Distributed computing strategies for processing of FT-ICR MS imaging datasets for continuous mode data visualization". In: *Anal. Bioanal. Chem.* 407.8 (Mar. 2015), pp. 2321–2327.

[75] Emrys A Jones et al. "High speed data processing for imaging MS-based molecular histology using graphical processing units." In: *J. Am. Soc. Mass Spectrom.* 23.4 (Apr. 2012), pp. 745–52.

[76] Leendert a. Klerk et al. "Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets". In: *Int. J. Mass Spectrom.* 260.2-3 (Feb. 2007), pp. 222–236.

[77] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemom. Intell. Lab. Syst.* 2.1-3 (Aug. 1987), pp. 37–52.

[78] Patrik Källback et al. "msIQuant - Quantitation Software for Mass Spectrometry Imaging Enabling Fast Access, Visualization, and Analysis of Large Data Sets". In: *Anal. Chem.* 88.8 (Apr. 2016), pp. 4346–4353.

[79] Ivo Klinkert et al. "Methods for full resolution data exploration and visualization for large 2D and 3D mass spectrometry imaging datasets". In: *Int. J. Mass Spectrom.* 362 (Apr. 2014), pp. 40–47.

[80] C Paschke et al. "Mirion - A Software Package for Automatic Processing of Mass Spectrometric Images". In: *J. Am. Soc. Mass Spectrom.* 24.8 (Aug. 2013), pp. 1296–1306.

[81] Oliver Rübel et al. "OpenMSI: a high-performance web-based platform for mass spectrometry imaging." In: *Anal. Chem.* 85.21 (Nov. 2013), pp. 10354–61.

[82] Guillaume Robichaud et al. "MSiReader: An Open-Source Interface to View and Analyze High Resolving Power MS Imaging Files on Matlab Platform". In: *J. Am. Soc. Mass Spectrom.* 24.5 (May 2013), pp. 718–721.

[83] R. Mitchell Parry et al. "OmniSpect: An open MATLAB-based tool for visualization and analysis of matrix-assisted laser desorption/ionization and desorption electrospray ionization mass spectrometry images". In: *J. Am. Soc. Mass Spectrom.* 24 (2013), pp. 646–649.

[84] Kyle D Bemis et al. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". In: *Bioinformatics* 31.14 (July 2015), pp. 2418–2420.

2.9 Appendix

Ref.	*Baseline Correction*	*Background noise removal*	*Noise reduction*	*Normalization*	*Alignment*	*Peak Picking*	*Binning*	*Matrix peak removal*
(Norris et al., 2007)	DataExplorer, FlexAnalysis, BioMap			Cube root, Logarithmic, TIC, Noise scaled (wavelet based)	By a subset of peaks common in 90% of pixels, Iterative cross-correlation			
(Källback et al., 2012)	SMQ1, SMST, SMA, SMM		Savitsky–Golay	TIC, RMS, Median, Labeled peak	Labeled peaks	1st derivative interpolated centroid		
(Mantini et al., 2007)	Kaiser filter	Kaiser filter	Kaiser filter					
(Fonville et al., 2012)				Reference molecule, TIC "matrix peaks", TIC "all data", TIC "informative peaks", PQN, Histogram matching, Median intensity of the informative peaks			Fixed mass binning to 0.2 Da bins prior to peak picking.	Matrix is removed using information extracted from out of tissue acquired regions.
(Mccombie et al., 2005)		Multivariate PCA and DA identification matrix peaks			Linear shift based on Cross correlations to a reference			PCA based
(Savitzky & Golay, 1964)			Savitzky-Golay					
(Jones et al., 2012a)				Isotopically labeled standard, TIC, Mean intensity				
(Deininger et al., 2011)				TIC, RMS, Median noise, SQRT-TIC, Log-TIC				
(Hamm et al., 2012)				TEC factor				
(Satten et al., 2004)	Custom method		Noise estimation threshold					
(Tracy et al., 2008)					Label-Free	Gaussian fitting	Binning after peack picking comparing all reported masses.	
(He et al., 2011)					Self-Calibrated Warping			
(Alexandrov et al., 2010)						OMP		

Table 2.1: Pre-processing methods summary.

Author (Ref)	Study	Kind of samples	Multivariate algorithm used	Main Results
(McCombie et al., 2005)	Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis	Brain from mouse with Alzheimer	PCA, hierarchical clustering (HC), k-means and ISODATA	Differentiation of tissue regions according to pixel spectral similarities
(Schwamborn et al., 2007)	Identifying prostate carcinoma by MALDI-Imaging	Human prostatic samples with or without cancer	SVM and GA	Differentiation of prostatic tissues with and without cancer
(McDonnell et al., 2008)	Mass Spectrometry Image Correlation: Quantifying Colocalization	Rat brain	Pearson coefficient correlations	Determination of the similarity of the distributions of specific molecules within and between MS Images
(Hanselmann et al., 2009)	Toward Digital Staining using Imaging Mass Spectrometry and Random Forests research articles	Animal models of human breast cancer	Random Forest classifier	Differentiation of different regions in a tumor sample and comparison between different tumor samples
(Rauser et al., 2010)	Classification of HER2 Receptor Status in Breast Cancer Tissues by MALDI Imaging Mass Spectrometry	Human breast cancer tissues	Supervised clustering of tissue regions using ANN i SVM	Definition of HER+ and HER- tissues
(Veselkov et al., 2014)	Chemo-informatics strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer	Human colorectal cancer tissues	Images co-registration, PLS-DA for tissue region differentiation	In house H&E and MSLDI Determination of region-specific lipid signatures in colorectal cancer tissues
(Van de Plas et al., 2015)	Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping	Mouse brain	Fusion image by multivariate regression	'fusing' two distinct technologies: mass spectrometry and microscopy images

Table 2.2: Overview of the literature about **supervised** multivariate analysis applied to MSI.

Author (Ref)	Study	Kind of samples	Multivariate algorithm used	Main Results
(Hanselmann et al., 2008)	Concise representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis	Highly and weakly metastatic tumor	pLSA, PCA, ICA and non-negative PARAFAC	Image reconstruction by pLSA
(Van de Plas et al., 2008)	Discrete wavelet transform-based multivariate exploration of tissue via imaging mass spectrometry	Mouse brain in the context of a resource-hungry study	DWT and PCA	Analysis of a sagittal section of mouse brain
(Bonnel et al., 2011)	Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: Application to prostate cancer	Prostate cancer samples	PCA-SDA and HC	identification and localization of specific markers in histological samples
(Cho et al., 2012)	Combining MALDI-TOF and molecular imaging with principal component analysis for biomarker discovery and clinical diagnosis of cancer	Rat renal samples	PCA	differentiation the different biologically regions in a tissue comparing lipid, peptide and protein profiles
(Lee et al., 2012)	Resolving brain regions using nanostructure initiator mass spectrometry imaging of phospholipids	Mouse brain	NMFA	Resolving of neuronal and glial reach brain regions
(Franceschi & Wehrens, 2014)	Self-organizing maps : A versatile tool for the automatic analysis of untargeted imaging datasets	Apple slices	SOM	spatial distribution of ions associated with the regions

Table 2.3: Overview of the literature about **unsupervised** multivariate analysis applied to MSI.

Author (Ref)	Study	Kind of samples	Multivariate algorithm used	Main Results
(Deininger et al., 2008)	MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers	Gastric cancer and non-neoplastic mucosa tissues	HA coupled to PCA	Determination of regions within and between tissues
(Alexandrov et al., 2010)	Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering	Rat brain and neuroendocrine tumor samples	image clustering process, using (HDDC) method	New procedure for spatial segmentation of MALDI-imaging data sets, that clusters all spectra into different groups based on their similarity.
(Alexandrov & Kobarg, 2011)	Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering	Rat brain and neuroendocrine tumor samples	Novel strategies for spatial segmentation that incorporates spatial relations between pixels into cluster regions.	Tumor heterogeneity exploration
(Bruand et al., 2011)	AMASS: algorithm for MSI analysis by semi-supervised segmentation	Rat brain	AMASS method: Automatic segmentation of maps according to patterns of co-expression of individual molecules	Discovery of novel molecular signatures
(Jones et al., 2011)	Multiple Statistical Analysis Techniques Corroborate Intratumor Heterogeneity in Imaging Mass Spectrometry Datasets of Myxofibrosarcoma	Myxofibrosarcoma	PCA, ICA, non-negative matrix factorization, pLSA, k-means clustering and hierarchical clustering).	Automatically localization of clusters in datasets
(Alexandrov et al., 2013)	Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity	Rat brain coronal section	Clustering method according to the spectral similarity of the pixels	Segmented rat brain with 10 regions and the spectrum with the ions associated to every segment
(Fonville et al., 2013)	Hyperspectral Visualization of Mass Spectrometry Imaging Data	Rat brain	The visualization strategy was applied to results of PCA, SOM and t-distributed stochastic neighbor embedding.	The image color-coding is based on spectral characteristics, such that pixels with similar molecular profiles are displayed with similar colors

Table 2.4: Overview of the literature about **unsupervised with further expert evaluation** multivariate analysis applied to MSI.

Software tool	MS Image visualization options	Main preprocessing methods	Main data Analysis methods	Operating systems or platform	Supported input data formats	Supported output data formats
flexImaging (*Bruker*) www.bruker.com	- Multiple *m/z* ion visualization using user-selectable colors - Optical image overlay	- Baseline correction - Noise reduction - Alignment - Normalization - Peak Picking (using FlexAnalysis)	- Univariate statistics - Genetic Algorithm - SVM -Supervised Neural Network -QuickClassifier - K-Nearest Neighbor - PCA (using ClinProTools)	Windows	Bruker instruments acquisition	ASCII imzML (since version 4.1) Analyze7.5
SCiLS Lab (*SCiLS*) scils.de	- Multiple *m/z* ion visualization using user-selectable colors - Optical image overlay - 3D volume reconstruction	- Baseline correction - Noise reduction - Alignment - Normalization - Peak Picking	- spatial segmentation with edge-preserving spatial de-noising - pLSA - pixel classification	Windows	FlexImaging	FlexImaging
MALDIVision *(PREMIER Biosoft) www.premierbiosoft.com*	- Multiple *m/z* ion from various images visualization using user-selectable colors - Optical image overlay	- Normalization to reference compound	- Histogram - Cumulative Probability Graph	Windows	imzML Analyze7.5	
TissueView *(Sciex)*	- Single *m/z* ion representation - Three *m/z* ions visualization in RGB - Optical image overlay	- Alignment - Normalization - Peak Picking (using MarkerView)	- PCA - PCA-DA (discriminant analisis) - PCVG (Principal Component Variable Grouping) (using MarkerView)	Windows	Analyze7.5	
ImageQuest *(Thermo Scientific) www.thermofisher.com*	- Single *m/z* ion representation - Side by side presentation of optical image - *m/z* ion scroll animation	- Normalization		Windows	Thermo Scientific instruments acquisition	
High Definition Imaging *(Waters)* www.waters.com	- Multiple *m/z* ion visualization using user-selectable colors - Three *m/z* ions visualization in RGB	-Peak picking	- PCA - PLS-DA - S-plots - hierarchical clustering.	Windows	Waters instruments acquisition	ASCII imzML
Quantinetix *(ImaBiotech) www.imabiotech.com*	- Multiple *m/z* ion visualization using selectable colors - Optical image overlay	- Normalization to reference compound	- Quantification	Windows	FlexImaging Thermo Waters imzML Analyze7.5	

Table 2.5: Summary of **commercial** software tools for MS imaging.

Software tool	MS Image visualization options	Main preprocessing methods	Main data Analysis methods	Operating systems or platform	Supported input data formats	Supported output data formats
msiQuant *(Uppsala University)* *www.maldi-msi.org*	- Single *m/z* ion representation - Optical image overlay	- Normalization	- Quantification	Windows	FlexImaging imzML	
BioMap *(Novartis)* *www.maldi-msi.org*	- Multiple m/z ions using various images - Side by side presentation of optical image			Windows, OSX, Linux	Analyze7.5	
Datacube Explorer *(AMOLF)* *amolf.nl/download/datacubeexplorer*	- Single *m/z* ion representation - 3D volume reconstruction		- Kohonen map clustering	Windows	imzML Analyze7.5	
Mirion *(Justus Liebig University)* (Paschke et al., 2013)	- Three *m/z* ions visualization in RGB - Optical image overlay			Windows	imzML	
OpenMSI *(Lawrence Berkeley National Laboratory)* *openmsi.nersc.gov*	- Three *m/z* ions visualization in RGB			any web browser	Author must be contacted for data uploading.	OpenMSI data file based on HDF5

Table 2.6: Summary of **freeware** software tools for MS imaging.

Software tool	MS Image visualization options	Main preprocessing methods	Main data Analysis methods	Operating systems or platform	Supported input data formats	Supported output data formats
MSiReader *(NC State University)* *http://www4.ncsu.edu/~dcmuddim/msireader.html*	- Single *m/z* ion representation - Three *m/z* ions visualization in RGB - Optical image overlay	- Baseline correction - Noise reduction - Alignment - Normalization - Peak Picking		Matlab (Windows, OSX, Linux)	ASCII mzXML imzml Analyze 7.5	
OmniSpect *(Emory University)* *cs.appstate.edu/omnispect*	- Single *m/z* ion representation - Three *m/z* ions visualization in RGB		- NMF Algorithm	Matlab (Windows, OSX, Linux)	mzXML imzML Analyze 7.5 NetCDF	
Cardinal (*Purdue University)* cardinalmsi.org	- Command line: interface, images are generated using R instructions	- Baseline correction - Alignment - Normalization - Peak Picking	- PCA - PLS-DA - Classification based on regularized nearest shrunken centroids	R (Windows, OSX, Linux)	imzML Analyze 7.5	

Table 2.7: Summary of **open-source** software tools for MS imaging.

Chapter 3

Assessing the potentiality of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications

3.1 Abstract

Mass spectrometry imaging (MSI) is a molecular imaging technique that maps the distribution of molecules in biological tissues with high spatial resolution. The most widely used MSI modality is matrix-assisted laser desorption/ionization (MALDI), but some organic matrices used in classical MALDI may impact the quality of the molecular images due to limited lateral resolution and strong background noise in the low mass range, hindering its use in metabolomics. Here we present a matrix-free LDI technique based on the deposition by sputtering of gold nanolayers on tissue sections. This gold coating method is quick, fully automated and repetitive and allows growing highly controlled nanolayers, necessary for high quality and high resolution MS image acquisition. The performance of the developed method has been tested on the acquisition of MS images of brain. The obtained spectra showed a high number of MS peaks on the low mass region (m/z below 1000 Da) with few background peaks, demonstrating the viability of the sputtered gold nanolayers of promoting the desorption/ionization of a wide range of metabolites. These results, together with the reliable MS spectrum calibration using gold peaks, make the developed method a valuable alternative for MSI applications.

3.2 Introduction

Classic histopathological analysis, in which the visual inspection of stained tissue sections is used for the identification of specific morphological regions and minute structures in the tissue, is one of the essential tools in medical diagnosis. Although often successful, in ambiguous cases the pathologist is not always capable of determining the correct diagnosis using histopathological methods alone. Complementary techniques that elucidate the chemical composition of those tissues aid the pathologist in these cases. In the recent years, mass spectrometry imaging (MSI) has emerged as a useful tool for the untargeted, and spatially correlated molecular analysis of clinical tissues, providing chemical information directly from the tissue [1, 2].

The most widely used ionization technique in MSI is matrix-assisted laser desorption/ionization (MALDI) [3], where an organic matrix is applied on a sample surface to promote the desorption/ionization of the analytes (e.g. proteins, lipids and metabolites). Classical MALDI-matrix application techniques may in-

troduce artefacts like compound diffusion, deteriorating the lateral resolution of the image [4], and/or inhomogeneities during deposition over the tissue and/or co-crystallization leading to increased differences in pixel-to-pixel ion intensities [5]. It is also known that some highly volatile organic matrices like 2,6-dihydroxyacetophenone (DHA) and dithranol evaporate during their time in the high vacuum ion source of the mass spectrometer, resulting in measurement artefacts over long data acquisition times [6]. To overcome some of these problems, matrix sublimation has been introduced for matrix deposition allowing higher spatial resolution analyses, as it is a solvent-free matrix deposition method and therefore results in smaller sized matrix crystals and less lateral diffusion of analytes [7]. Nevertheless, one of the main drawbacks of MALDI is that the organic matrices introduce a considerable number of MS signals in the low m/z range of the spectrum ($<$ 1000 Da). These signals interfere severely with the MS peaks of endogenous low weight compounds, complicating the application of MSI to metabolomics studies [5].

Matrix-free LDI-MS techniques have emerged as valuable alternatives for the analysis of low molecular weight metabolites. Commonly used matrix-free techniques are surface-assisted laser desorption/ionization (SALDI), in which ionization is supported by the surface of the target plate [5, 8, 9, 10, 11, 12], and nanostructure-initiator mass spectrometry (NIMS) [13], which uses molecules of an initiator compound trapped in nanostructured surfaces promoting the ionization of the metabolites. Moreover, metal (Au, Ag, Pt, etc.) and metal oxide (WO_3, TiO_2, Fe_3O_4, ZnO, etc.) nanoparticles and nanolayers, frequently called nanoparticle-assisted LDI (nano-PALDI) have also been used for the LDI-MS analysis of biomolecules [5]. In this context, gold nanoparticles are likely the ideal substrate because they present high stability, can be easily functionalized [14, 15], are able to absorb the UV light emitted by the laser and effectively transfer this absorbed energy to the metabolites promoting its absorption and providing a source of ionization. Several studies have used gold nanoparticles for the analysis of biofluids by LDI-MS, and for MSI applications achieving an effective ionization of low mass range metabolites with very low background signal [16, 17, 18, 19, 20, 21, 22]. In these studies, gold nanoparticles were deposited on the tissues by mixing them with organic matrices or solvents. This "wet" deposition of gold does not prevent the potential lateral diffusion of the metabolites and the inhomogeneous distribution of the gold nanoparticles. To overcome this, sputter deposition, which is a solvent-free and reproducible deposition technique, would

allow the deposition of high purity, homogeneous metal or metal oxide nanolayers onto biological tissues whilst avoiding molecular delocalization associated with solvent-based application methods. In a previous publication by Dufresne et al. sputter deposition of silver was used prior to the MSI of olefins from tissue sections [23]. This study demonstrated the viability of sputter depositions for MSI metabolomics applications, with high spatial resolution (down to 5 μm). More recently, a sodium deposition followed by a sputtered gold layer has been introduced as a powerful method for the analysis of triacylglycerols [24]. Furthermore, the characteristic gold and silver peaks and clusters can be used for internal mass calibration along the different m/z regions of the obtained spectrum [23, 24, 25].

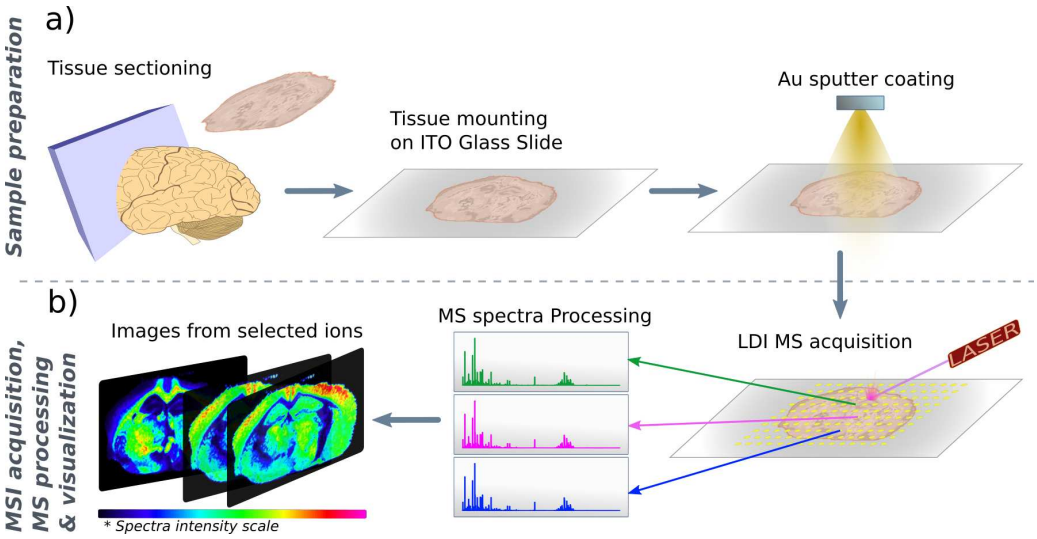


Figure 3.1: Experimental workflow of the developed gold nanolayer-assisted LDI-MSI method. **A)** Sample preparation, including sectioning of 10μm-thick sections, tissue mounting on indium-tin oxide-coated (ITO) glass slides and the tissue coating with gold nanoparticles by sputtering. **B)** Summary of LDI-MSI acquisition, spectral pre-processing, image reconstruction and visualization.

Hence, in this study we present the application of gold nanolayers deposited by sputtering directly onto the tissue section to obtain metabolomic MS images of animal tissues by LDI-MS. In comparison with silver, gold has only one stable isotope, thus reducing the number of peaks and facilitating the detection of trace compounds. Gold ionizes polar and heavier metabolites more effectively [26], and provides highly stable nanolayers. The experimental workflow used in this study is summarized in Fig. 3.1 and includes the cryosectioning of the tissues into thin sections (10 μm), the mounting of these tissue sections on conductive indium-tin oxide-coated (ITO) glass slides and the coating of the sections with gold nanoparti-

cles using sputter deposition. After the spatially correlated LDI-MS acquisition of a spectrum at each pixel, spectra were processed and the images of the molecular distributions reconstructed and visualized. In this study, we report the optimization of the sputter deposition conditions, based on the ionization efficiency of the gold nanolayers by using mice liver sections and the optical and morphological characterization of the deposited gold nanolayers. Finally, the viability of the optimized gold nanolayers was checked by the acquisition of metabolomics MSI data from mouse brain tissue.

3.3 Materials and Methods

3.3.1 Materials

Indium tin oxide (ITO)-coated glass slides were obtained from Bruker Daltonics (Bremen, Germany). The gold-target (purity grade $>$ 99.995%) used for sputtering was obtained from Kurt J. Lesker Company (Hastings, England). The reagents and solvents for staining were hematoxylin and HPLC grade xylene supplied by Sigma-Aldrich (Steinheim, Germany) and ethanol (96% purity, supplied by Scharlau, Sentmenat, Spain).

3.3.2 Sample preparation

The liver tissues used for gold-sputtering optimization and the brain tissues used for the example of MSI metabolites assignation were obtained from C57BL/6 mice of 6 months old. These tissues were provided by Professor Martins-Green's research group at the Cell Biology Department of the University of California Riverside. The tissues were snap frozen at -80°C after collection and stored and shipped at this temperature until analysis. Animal experimental protocols were approved by the University of California, Riverside, Institutional Animal Care and Use Committee (IACUC).

The brain tissues used for the high-lateral resolution MSI analysis were obtained from three month-old, male, C57BL/6J mice. These tissues were obtained from Leiden University Medical Center where all the high lateral resolution experiment was carried out. The brains were excised, flash-frozen on dry-ice and stored at -80°C until analysis. All experiments were approved by the Animal Experiment Ethics Committee of Leiden University Medical Center.

For MSI acquisition, the tissues were sectioned at -20°C into 10 µm thick sections using a Leica CM-1950 cryostat (Leica Biosystems Nussloch GmbH) located at the Centre for Omics Sciences (COS) of the University Rovira i Virgili and mounted on ITO coated slides by directly placing the glass slide at ambient temperature onto the section. To remove residual humidity, samples where dried in a vacuum desiccator for 15 minutes after tissue mounting.

3.3.3 Gold sputter coating

Gold nanolayers were deposited onto the 10 µm tissue sections using a sputtering system ATC Orion 8-HV (AJA International, N. Scituate, MA, USA). An argon atmosphere with a pressure of 30 mTorr was used to create the plasma in the gun. The working distance of the plate was set to 35 mm. The deposition times were determined from the deposition rate, which is directly proportional to the layer thickness. Since deposition times used in this study were very short, the substrate temperature remained cold during the deposition, thereby avoiding degradation of the tissue metabolites. The final optimized sputtering conditions for MSI were at ambient temperature, using RF mode at 60 W for 35 s.

3.3.4 Sample characterization

Reflectance measurements of the gold-coated tissues were carried out with a Lambda-950 spectrophotometer, equipped with deuterium and tungsten lamps (Perkin-Elmer, Waltham, MA, USA) scanning in the 250 to 800 nm wavelength range.

Morphology of the gold layer was characterized by transmission electronic microscopy (TEM) using a JEOL 1011 microscope (Jeol, Peabody, MA, USA). A TEM grid was used to deposit a gold layer using the optimized conditions described above.

3.3.5 LDI-MS acquisition

MSI data used for the Au-layer optimization and characterization were acquired using a MALDI TOF/TOF UltrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics, also at the COS facilities. Acquisitions were carried out using the medium and large laser spot size settings, operated at 2 kHz at an attenuated power of 60 %, collecting a total of 500 shots per pixel.

High spatial resolution MSI data were recorded using a MALDI TOF/TOF rapifleX with SmartBeam 3D II Nd:YAG/355 nm laser from Bruker Daltonics, located at Leiden University Medical Center (LUMC). The laser was operated at 10 kHz collecting 200 shots per pixel .

Raster sizes from 10 to 1000 μm were used during the optimization. The TOF mass spectrometer was operated in positive ion, reflectron mode, with a digitization rate of 1.25 GHz, m/z range 70 to 1200 Da, with a manually optimized extraction delay. The spectrometer was calibrated prior to MSI data acquisition using $[Au]^+$ peaks as reference masses. Following the LDI-MSI experiment, the sections were stained with hematoxylin.

3.3.6 Spectra pre-processing and image visualization

MSI data was acquired using the Flex-software suite (v3.0 Bruker Daltonics). Each MSI dataset was exported to the XMASS data format using instrument manufacturer software packages (FlexImaging and Compass export) and a custom script. The data stored in XMASS was converted to a custom format based on segmented matrices storage highly optimized for processing large MSI datasets in R language [27]. Mass spectra were aligned using a novel unlabeled method developed to handle our custom data format efficiently and which is included in our rMSIproc package (http://github.com/prafols/rMSIproc). After alignment, the whole dataset shared the same mass axis and, therefore mass calibration was applied to the whole dataset by only calibrating the mean spectrum. Following this method, masses were calibrated using gold peaks as reference: 196.9666, 393.9331, 590.8997, 787.8662 and 984.8328 Da. Moreover, m/z 96.9223 and 112.8962 associated with $[KNaCl]^+$ and $[K_2Cl]^+$ were also used as mass reference peaks to better calibrate the low-mass range ions [26]. In order to show the actual performance of the gold layer, no normalization was performed. MSI datasets were explored manually to select a set of peaks localized on different morphological structures. This exploration stage was accomplished using our dedicated graphical user interface included in the rMSI R package, specially developed to rapidly explore MSI data [27]. MSI image reconstruction and visualization was also performed with the same in-house software package.

3.3.7 Metabolite identifiaction

MS peaks were obtained using an in-house peak picking algorithm included in the rMSIproc R package with S/N>5. The obtained list of MS peaks was matched with HMDB [28] data base within a tolerance of 20 ppm and the possible ion adducts: H, Na, K and NH_4. In order to obtain a list of possible metabolites, the obtained search results were filtered using the biological information of molecules provided by the HMDB. We have also used the information provided by the HMDB to highlight the putative identified metabolites that have previously been reported in brain tissues.

3.4 Gold nanolayer optimization and characterization

3.4.1 Sputter coating optimzation for LDI

The sputtered deposition of gold nanoparticles has been optimized to achieve the highest LDI-MS signal intensities at the lowest laser fluencies. The gold nanolayer deposited on the tissue must provide enough gold nanoparticles to promote the desorption/ionization of the metabolites, but also thin enough to enable the laser to reach the tissue-gold interface. Moreover, the deposited gold nanolayer must ensure the correct identification of the gold MS peaks to enable in-situ mass calibration using these peaks.

To optimize the layer thickness we have used liver sections from C57/BL mice. We have selected a liver tissue from a healthy mouse for the optimization steps because it usually presents high biological homogeneity at the spatial resolution used for MSI analysis, facilitating the comparison of the performance of the different gold layers in a real sample. To further ensure the comparability of the tests, the various gold layers were deposited over consecutive liver sections. Moreover, each acquisition was performed on identical regions of tissue, which were selected by optical inspection of the liver sections. Then, a wide random walk of 1000 μm per pixel was used in order to obtain an averaged spectrum of each laser shot.

As a starting point we tested three different Au nanolayers designed to cover a broad range of Au thicknesses. Once an approximate optimal layer was obtained, the next step was to fine-tune the sputter coating time to fine-tune the Au thickness. Two modes can be selected for gold sputtering: direct current (DC), which is the fastest method, and radio frequency (RF), which provides higher control over

the deposited gold layers. Since the desired Au nanolayer must be a thin layer according to previous studies [23, 24], we performed all Au depositions operating the sputtering system in RF mode to better control the tissue coating process. The first three deposition times tested as first thickness exploration were 25, 100 and 300 s at 60 W and ambient temperature.

The laser attenuation reported in FlexImaging was adjusted for each one of the three sputtered layers. Note: this laser attenuation setting is part of the user interface, designed to give the average user fine control over the laser powers commonly used in MALDI experiments (0% corresponds to the laser power offset, and 100% to the laser power offset + laser power range, both of which can be found in the instrument specific settings). The laser fluence varies approximately linearly with the attenuation throughout this range. Here we report the laser attenuation value to compare the laser fluence used for each sputtered layer. This laser attenuation parameter was adjusted in order to achieve the highest MS peak intensities in the m/z 700 – 900 range for each gold layer. Based on their molecular masses, the metabolites that can be found in this mass range are likely to correspond to phosphatidylcholines and triacylglycerides, compounds of high biological relevance, but easily fragmented. Therefore, during the laser fluence adjustment we monitored the intensity of m/z 184, which corresponds to the head group fragment of the phosphatidylcholines. We selected the best performing laser power fluence for each Au layer to obtain a good tradeoff between peak intensity and molecular fragmentation. The optimal laser powers were found to be 60, 70 and 75% for the 25, 100 and 300 s sputter coating times respectively. The thicker layers required higher laser power, suggesting that the ionization efficiency is lower for thicker layers and were more prone to suffer fragmentation due to higher laser power.

Fig. 3.2 shows the average spectra of the liver sections obtained for each gold layer. In agreement with the results described above, the 25 s gold layer provided the highest number of MS peaks with higher intensity in all the areas of the selected mass range, including the 700-900 Da range (see Fig. 3.2 B, D and F). These results confirm the better performance of the thinner gold nanolayer.

We designed a second Au nanolayer optimization set up considering as starting point the 25 s Au layer, considered optimal in the previous experiment. We applied various sputtered Au layers using deposition times ranging from 15 to 45 s in steps of 5 s onto consecutive sections of liver tissues. The laser power was kept at 60 % for all the layers since all of them must be compared in the same conditions.

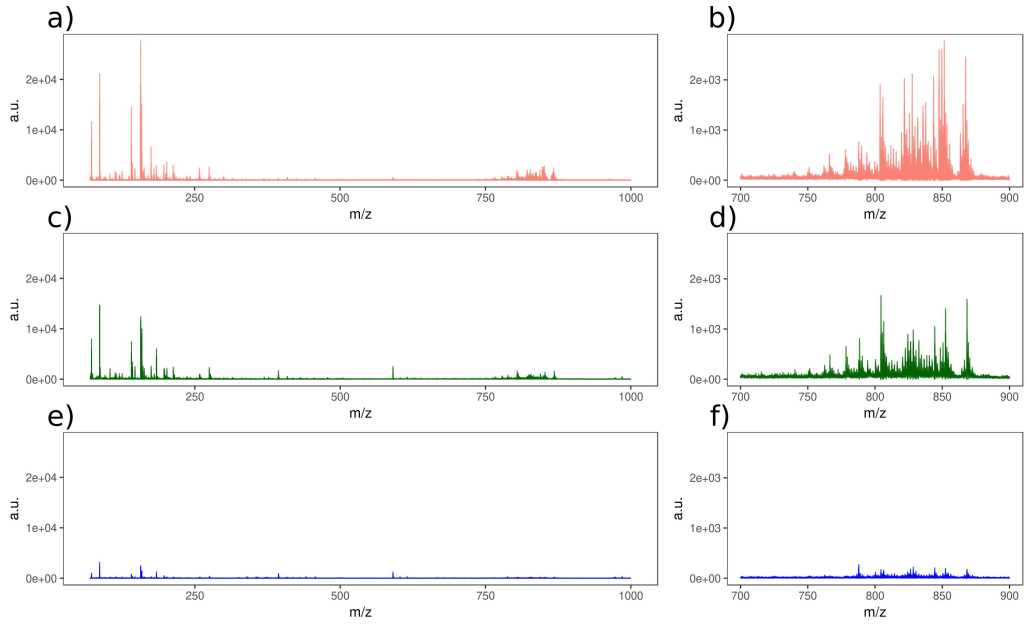


Figure 3.2: Average spectra of mouse liver sections obtained with each of the three tested gold layers. **A)** 25 s Au coating time at a laser power of 60%, **C)** 100 s Au coating time at laser power of 70 %, **E)** 300 s Au coating time at laser power of 75 %, **B, D** and **F** figures plot the m/z spectrum between 700 and 900 Da to illustrate the performance of the tested gold layers in a specific area of the spectrum.

Moreover, we also acquired a tissue section with no Au coating as reference of ionization without Au (0 s). In each case we acquired a complete MSI dataset containing approximately 300 pixels with a pixel size of 100x100 μm.

After LDI-MS acquisition, the MS data obtained with each of the tested gold layers were compared to determine the optimal gold coating time. As specified in the materials and methods section, spectra were acquired in reflectron positive mode and processed using in-house developed R packages rMSI and rMSIproc. We retained the first 250 most intense pixels of each tissue section for the data analysis. This discards regions of the tissue with holes or bad MS performance and provides the same number of sampling points for each sputtered layer. In order to provide an objective comparison criterion between different sputtering conditions, we have calculated two parameters from each gold layer: the total ion count (TIC) defined as the summing up intensities of all MS peaks; and the fragmentation ratio calculated by dividing the intensity of the head group fragment of the phosphatidylcholines (m/z 184.07) by the sum of intensities of the MS peaks found in the 500 to 1000 m/z range. For the estimation of these three parameters, we have only considered the peaks of the analyzed liver section with a signal to

noise ratio (S/N) over 5, excluding the MS peaks of the gold clusters (m/z 196.97, 393.93, 590.90, 787.87 and 984.83). Fig. 3.3 shows the results of comparison of the gold deposition times tested. As can be seen in Fig. 3.3A, the highest TIC was obtained with the 35 s gold layer. Moreover, the lowest fragmentation ratio value was also obtained for the 35 s. Fig. 3.3C confirms that the 35 s coating time provides the optimal Au layer since higher peaks were detected with same laser conditions in the 500 to 1000 m/z range.

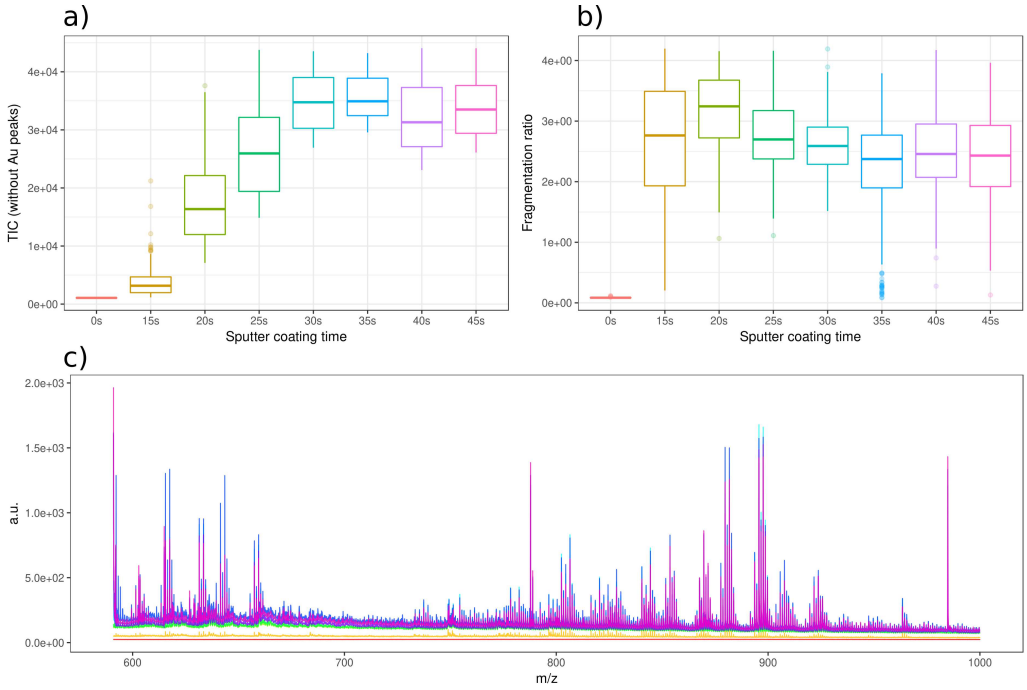


Figure 3.3: Comparison of various Au coating times MSI performance using 250 pixels of each MS image. **A)** TIC vs. Au coating time at a laser power of 60%. Au cluster peaks were removed to avoid biasing the experiment since Au MS intensity increases with the sputter coating time. **B)** Fragmentation ratio of each Au layer was calculated dividing the intensity of to the head group fragment of the phosphatidylcholines (m/z 184) peak by the sum of all peak intensities in the 500 to 1000 m/z. **C)** Plot of average spectra from all Au layers with the same coloring as boxplots A and B.

Acquisitions in negative ionization mode could enhance the MS signal of some metabolites and, therefore, we have also explored the performance of the sputtered gold layers in this ionization mode. As an example, Fig. 3.7 in the appendix shows the average MS spectrum obtained with a 35 s gold layer of a consecutive section of the liver used for the tests in positive mode, with the same laser conditions. A total of 298 MS peaks, 33 of them in m/z 700 – 900 range, were detected in negative mode with an S/N>5 and TIC 2.63x10^5. Gold peaks were clearly

identified in the negative spectrum also enabling the internal calibration process also for this ionization mode. These results demonstrate the suitability of gold sputtered layers to acquire MS images in negative ionization mode, opening a wide range of possibilities for specific applications like the analysis of low-weight acids [29] or fatty acids [30]. Nevertheless, in this study we have focused on positive ionization mode as the deposition was optimized for use in this mode.

The results above demonstrate that the best LDI-MS performance was obtained with the 35 s gold layer. This short gold deposition deposited enough gold particles to ensure the desorption/ionization of the metabolites, but also allowed the laser to easily reach the tissue surface. Moreover, the Au cluster peaks were detected with enough intensity to provide for a reliable mass calibration. Longer gold coating times may prevent the proper desorption/ionization of the underlying metabolites because of the dissipation of more laser energy into the thicker gold layer before reaching the tissue surface.

3.4.2 Au nanolayer characterization

The morphology of the RF-deposited gold layer was characterized by Transmission Electronic Microscopy (TEM). Fig. 3.4A shows the TEM image of the optimized gold nanolayer deposited over a TEM grid at a magnification of 400,000. TEM images could only be taken by coating a TEM grid and could present a different morphology compared to the gold layer sputtered over a biological tissue. Nevertheless, TEM images could be used as reference. As can be seen in Fig. 3.4A, the gold nanolayer (represented by the dark grey and black areas) is discontinuous. This gold nanolayer forms irregular nanoislands, surrounded by free spaces with a dimension between 5 and 10 nm. A pixel integration over the TEM images showed that the gold particles covered approximately the 65% of the sputtered surface. The sputtered layer over a biological tissue might adapt to the roughness and morphology of the different tissue surfaces, and might be more discontinuous. Although there was a discontinuity of the deposited gold at the nanoscale level, the layer is homogeneous at the LDI-MSI acquisition scale (μm-scale) and therefore, it does not affect the reproducibility of the MSI analysis.

As commented above, one of the most important features of the surfaces developed for LDI-MS applications is the ability to absorb the maximum energy at the wavelength of the instrument laser beam (355 nm for this study). To characterize the performance of the optimized gold nanolayer, we have measured the absorp-

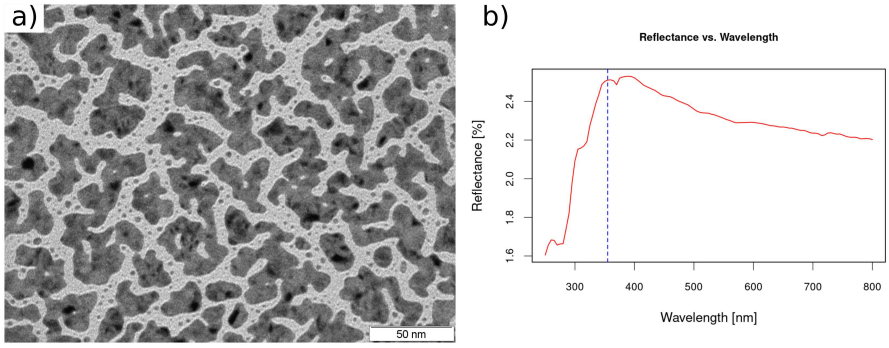


Figure 3.4: A) TEM image at magnification of 400,000 of the optimized gold nanolayer, sputtered in RF mode at 60 W and ambient temperature for 35 s. The gold nanolayer is represented by the dark grey and black areas. **B)** Reflectance spectrum of the sample system formed by a ITO-coated glass slide, a 10 µm mice brain section and the optimized gold layer. The vertical dashed blue line corresponds to the Nd:YAG laser wavelength (355 nm) used for the LDI-MS acquisitions.

tion spectrum of an optimized gold coated, 10 µm tick mouse brain tissue section mounted on a ITO-covered glass slide. This absorption spectrum was measured using a Vis-UV spectrometer with a light incidence angle of 30° in order to mimic as much as possible the acquisition conditions of the laser configuration in the UltrafleXtreme MALDI-TOF instrument [31]. Under the acquisition conditions, the light reflection of the sample system was ca. ~2.5 % at 355 nm, which indicates that the tissue-Au-layer system absorbs most of the laser energy achieving high optical efficiency. Fig. 3.4B shows the obtained reflectance spectrum.

3.5 Results: MSI of animal tissues with gold-sputtered layers

The viability of the optimized gold nanolayers for metabolomics MSI applications by LDI-MS was checked by acquiring MSI data of different animal organ tissue sections. The deposition of the optimized gold layer in RF mode, under highly controlled conditions, allowed the acquisition of MSI data using a lower attenuation laser power (60%). C57/BL mouse brain was also used to test the spatial resolution of the method. Fig. 3.8 in the appendix shows results of LDI-MSI analyses acquired at a spatial resolution of 10 and 20 µm. As an example, one ion was manually selected to show the highly detailed morphological structure in the corpus callosum. In these images we were able to reproducibly reveal small brain tissue structures, demonstrating the capabilities of the sputtered gold layer for high

spatial resolution LDI-MSI.

In this study, C57/BL mouse brain tissue was acquired with raster sizes of 80 μm, a resolution previously reported to be sufficient to reveal the tissue structures in these organs [23]. Moreover, we verified that with this pixel size the laser shots did not overlap and thus detection sensitivity was not compromised. Fig. 3.5 shows the MSI visualizations of three selected ions (m/z 845.46, m/z 849.64 and m/z 213.04, Fig. 3.5A, B and C, respectively) obtained from a sagittal section of a mouse brain. These figures represent the relative abundance of the selected ions in the color scale showed in each figure, where red represents the areas with maximum ion signals and dark blue the minimum ion signals. As can be seen, the selected ions present different region selectivity in the mouse brain. Fig. 3.5D plots the combined image of these three ions using the RGB color scheme (m/z 845.45 in red, m/z 849.64 in green and m/z 213.04 in blue). In this figure, different brain regions can be clearly distinguished and labeled. The reliability of the developed MSI method was confirmed by comparing the brain morphology obtained with the MSI images with the same brain slice stained with Hematoxilyn (Fig. 3.5E), stained shortly after the MSI acquisition (note the gold layer is porous, which allows the hematoxylin to stain the underlying tissue). In contrast to matrix assisted LDI, the tissue staining can be done without performing any washing step. The ions at m/z 845.46 and 849.64 were putatively identified as the potassium adducts of two lipids commonly found in brain tissues (see Table 1 for further details).

Fig. 3.6 shows the average MS spectrum from the gold coated mouse brain tissue section. The gold peaks used for the calibration of this spectrum are also indicated. As can be seen, the gold nanolayers are able to promote the desorption/ionization of different metabolites throughout a wide mass range. A total of 356 peaks were detected with a S/N>5, with a TIC of $2.63\text{x}10^5$.

The detected peaks were putatively assigned on the basis of mass accuracy, by matching the experimental mass with the Human Metabolome Data Base (HMDB) [28] database. Thirty endogenous metabolites have been putatively assigned with an identity in the brain section, listed in Table 1, with a mass error below 15 ppm for most metabolites. The list of putative assignments includes amino acids, carbohydrates and other small metabolites and several kinds of lipids, such as fatty acyls, glycerolipids, glycerophospholipids, sphingolipids and sterol lipids. In bold we have marked the metabolites previously reported in brain by the HMDB to give more confidence to the assignments. To check the accuracy of the identifications

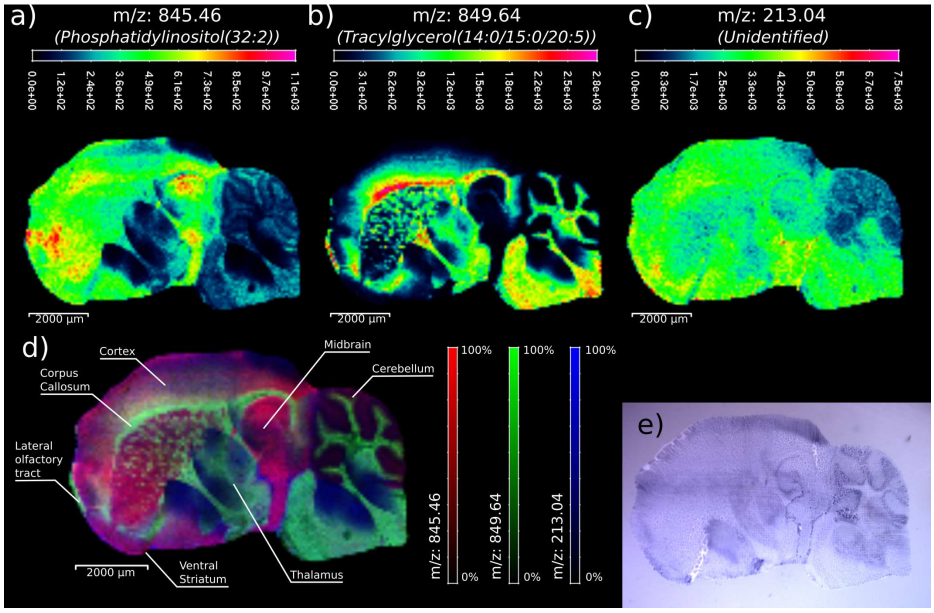


Figure 3.5: Sagittal section of a mouse brain acquired with the optimized sputtered gold layer at a raster size of 80 µm. Figures **A, B** and **C** plots the relative abundance of three ions found to reproduce the brain morphology (845.46 Da, 849.64 and 213.04, respectively). **D)** shows the combined RGB color encoded representation of the three ions that plots different brain areas of the sagittal section. Some of the identified brain regions are labeled. E Optical image of a consecutive brain section slice stained with a Hematoxilyn.

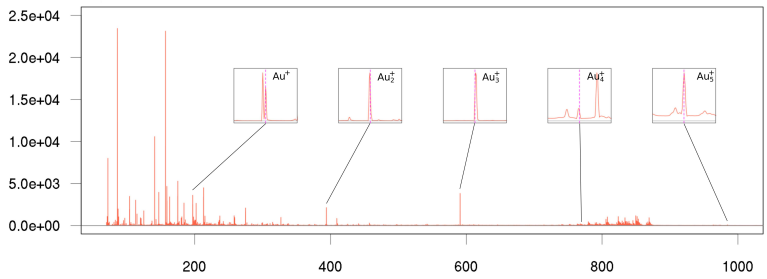


Figure 3.6: Average MS spectrum of a mice brain section. The MS peaks of gold used for the spectra mass calibration are also indicated.

we studied the spatial distribution of the ions identified as cholesterol. Cholesterol was detected as sodium and potassium adducts (m/z 409.33 and m/z 425.31, respectively). As seen in Fig. 3.9 at the appendix, the spatial distribution of cholesterol ions is similar, thus corroborating that both ions come from the same metabolite. These coherent distributions of cholesterol reinforce the suitability of the gold-induced ionization for MSI.

Name	Ion formula	m/z $_{exp}$[a]	m/z $_{calc}$[b]	Δm/z (ppm)
Citrulline	$[C_6H_{13}N_3O_3+Na]^+$	198.0864	198.0849	-7.8
DAG (35:0)	$[C_{38}H_{74}O_5+H+NH_4]^+$	314.7974	314.7971	0.8
Monoacylglycerol (18:2)	$[C_{21}H_{38}O_4+K]^+$	393.2330	393.2402	18.3
Cholesterol	**$[C_{27}H_{46}O+Na]^+$**	**409.3409**	**409.3441**	**7.8**
Cholesterol	**$[C_{27}H_{46}O+K]^+$**	**425.3091**	**425.3180**	**21.0**
Palmitoyl glucuronide	$[C_{22}H_{42}O_7+Na]^+$	441.2787	441.2823	8.2
Palmitoyl glucuronide	$[C_{22}H_{42}O_7+K]^+$	457.2581	457.2562	-4.0
dimethylphosphatidylethanolamine	$[C_{41}H_{78}NO_8P+Na]^+$	766.5271	766.5357	11.2
Phosphatidylcholine(34:4)	**$[C_{42}H_{76}NO_8P+K]^+$**	**792.4857**	**792.4940**	**10.5**
Phosphatidylserine(34:3)	**$[C_{40}H_{72}NO_{10}P+K]^+$**	**796.4646**	**796.4525**	**-15.2**
Phophatidylethanolamine(40:10)	**$[C_{45}H_{70}NO_8P+Na]^+$**	**806.4772**	**806.4731**	**-5.0**
Phosphatidylserine(36:5)	$[C_{42}H_{72}NO_{10}P+K]^+$	820.4617	820.4525	-11.2
Phosphatidylcholine(38:3)	$[C_{46}H_{88}NO_7P+Na]^+$	820.6139	820.6191	6.4
Phosphatidylserine(36:4)	**$[C_{42}H_{74}NO_{10}P+K]^+$**	**822.4692**	**822.4682**	**-1.2**
Phosphatidylcholine(38:1)	$[C_{46}H_{90}NO_7P+Na]^+$	822.6361	822.6347	-1.8
Phosphatidylserine(38:7)	$[C_{44}H_{72}NO_{10}P+Na]^+$	828.4711	828.4786	9.1
Phophatidylethanolamine(42:11)	**$[C_{47}H_{72}NO_8P+Na]^+$**	**832.4840**	**832.4888**	**5.8**
3,4-dihydroxy-5-all-trans-decaprenylbenzoate	$[C_{57}H_{85}O_4+H]^+$	834.6539	834.6526	-1.5
Phosphatidylglycerol(38:4)	$[C_{44}H_{79}O_{10}P+K]^+$	837.5056	837.5042	-1.7
Phosphatidylcholine(38:2)	$[C_{46}H_{90}NO_7P+K]^+$	838.6038	838.6086	5.8
Phosphatidylserine(38:7)	$[C_{44}H_{72}NO_{10}P+K]^+$	844.4639	844.4525	-13.5
Phosphatidylinositol(32:2)	**$[C_{41}H_{75}O_{13}P+K]^+$**	**845.4620**	**845.4577**	**-5.1**
Glucosylceramide	**$[C_{48}H_{91}NO_8+K]^+$**	**848.6360**	**848.6376**	**1.9**
Tracylglycerol(49:5)	$[C_{52}H_{90}O_6+K]^+$	849.6369	849.6369	0.0
DMPE(40:10)	$[C_{47}H_{74}NO_8P+K]^+$	850.4673	850.4784	13.0
Phosphatidylcholine(40:2)	$[C_{48}H_{94}NO_7P+Na]^+$	850.6576	850.6660	9.9
Phosphatidylserine(40:9)	$[C_{46}H_{72}NO_{10}P+Na]^+$	852.4730	852.4786	6.5
Phosphatidylcholine(40:3)	$[C_{48}H_{92}NO_7P+K]^+$	864.6167	864.6243	8.7
Phosphatidylcholine(40:1)	$[C_{48}H_{94}NO_7P+K]^+$	866.6282	866.6399	13.5
Phosphatidylserine(40:9)	$[C_{46}H_{72}NO_{10}P+K]^+$	868.4617	868.4525	-10.6

Table 3.1: Putative identification of metabolites in the brain tissue section including the chemical name, ion formula, the experimental m/z obtained in our experiment (a), the m/z calculated from the database (b), and the mass error of the identification in ppm.

The optimized gold-induced ionization system presents several advantages regarding other MSI sample preparation techniques. On the one side, sputtering allows the deposition of high purity gold nanoparticles (>99.995%) avoiding contamination of the sample and, therefore, the presence of interfering peaks in the MS spectra. Furthermore, since gold only presents one stable isotope there is no additional dilution of the ion current (as occurs with silver assisted LDI), less mass spectral congestion and the mass spectral peaks can be assigned identities more easily. Furthermore during the MSI data acquisition the mass calibration of

the instrument can drift; the presence of gold cluster ions in each pixel facilitates the alignment and calibration between pixels, thus potentiating a more reliable identification of the metabolites. In this study we have achieved a mass accuracy below 15 ppm for most compounds using gold clusters for internal calibration. Furthermore, compared with the wet deposition of gold layers or organic matrices, the sputtering deposition process used here is known as a fast and highly reproducible technique. The total time needed for the gold layer deposition over a tissue section is around 5 min including sample mounting, pumping the vacuum chamber and deposition. A recent study suggested gold as possible universal material for LDI-MS analysis and imaging [26] because of the high detectability and high mass determination accuracy achieved with this material. The detection of MS peaks in a wide m/z range obtained in brain also confirms the potentialities of the application of this new MSI methodology in clinical diagnostics. The rapid and reproducible dry deposition of gold optimized here would promote the use of gold for MSI applications, without the metabolite delocalization inherent to wet deposition methodologies.

3.6 Conclusions

In this study we present the development of an alternative method for the acquisition of MSI data based on the sputter deposition of gold nanolayers over thin tissue slices. Gold is a highly stable material and neither degradation nor oxidation occurs after sample preparation or during the LDI-MS acquisition. The presented sample preparation method is fast, fully automated and reproducible. Furthermore, this dry deposition method avoids the delocalization of metabolites in biological tissues improving the spatial resolution (down to 10 μm, and only limited by the laser configuration) of the obtained MS images.

The capacity of gold to acquire a wide range of metabolites has been demonstrated through the acquirement of MS images of brain tissue. The mass spectra obtained from the analyzed tissues are very rich in the m/z range under 1000 Da. Background MS peaks from gold nanoparticles are just five single signals homogenously distributed across the spectra. These signals have a minimal interference on metabolites detection and can also be used for a reliable spectrum alignment and mass calibration between pixels. Moreover, we have been able to putatively identify thirty endogen metabolites in brain demonstrating the reliability of the

acquired spectra. Therefore, the gold-assisted sputtering MSI method presented here could open up new possibilities for a reliable use of MSI in clinical diagnostics.

References

[1] Jeremy L Norris and Richard M Caprioli. "Imaging mass spectrometry: A new tool for pathology in a molecular age". In: *PROTEOMICS - Clin. Appl.* 7.11-12 (Dec. 2013), pp. 733–738.

[2] Tiegang Li et al. "In situ biomarker discovery and label-free molecular histopathological diagnosis of lung cancer by ambient mass spectrometry imaging". In: *Sci. Rep.* 5.1 (Nov. 2015), p. 14089.

[3] Michael. Karas and Franz. Hillenkamp. "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons". In: *Anal. Chem.* 60.20 (Oct. 1988), pp. 2299–2301.

[4] Satu M Puolitaival et al. "Solvent-free matrix dry-coating for MALDI imaging of phospholipids." In: *J. Am. Soc. Mass Spectrom.* 19.6 (June 2008), pp. 882–6.

[5] Cheng-Kang Chiang, Wen-Tsen Chen, and Huan-Tsung Chang. "Nanoparticle-based mass spectrometry for the analysis of biomolecules." In: *Chem. Soc. Rev.* 40.3 (Feb. 2011), pp. 1269–1281.

[6] Nina Ogrinc Potočnik et al. "Use of advantageous, volatile matrices enabled by next-generation high-speed matrix-assisted laser desorption/ionization time-of-flight imaging employing a scanning laser beam". In: *Rapid Commun. Mass Spectrom.* 29.23 (Dec. 2015), pp. 2195–2203.

[7] Joseph A Hankin, Robert M Barkley, and Robert C Murphy. "Sublimation as a method of matrix application for mass spectrometric imaging". In: *J. Am. Soc. Mass Spectrom.* 18.9 (Sept. 2007), pp. 1646–1652.

[8] K. P. Law and James R. Larkin. "Recent advances in SALDI-MS techniques and their chemical and bioanalytical applications". In: *Anal. Bioanal. Chem.* 399.8 (Mar. 2011), pp. 2597–2622.

[9] Dominic S. Peterson. "Matrix-free methods for laser desorption/ionization mass spectrometry". In: *Mass Spectrom. Rev.* 26.1 (Jan. 2007), pp. 19–34.

[10] Angelina Yimei Lim, Jan Ma, and Yin Chiang Freddy Boey. "Development of Nanomaterials for SALDI-MS Analysis in Forensics". In: *Adv. Mater.* 24.30 (Aug. 2012), pp. 4211–4216.

[11] Liang Qiao, BaoHong Liu, and Hubert H Girault. "Nanomaterial-assisted laser desorption ionization for mass spectrometry-based biomedical analysis". In: *Nanomedicine* 5.10 (Dec. 2010), pp. 1641–1652.

[12] P. A. Kuzema. "Small-molecule analysis by surface-assisted laser desorption/ionization mass spectrometry". In: *J. Anal. Chem.* 66.13 (Dec. 2011), pp. 1227–1242.

[13] Trent R Northen et al. "Clathrate nanostructures for mass spectrometry." In: *Nature* 449.7165 (Oct. 2007), pp. 1033–6.

[14] Michael E. Kurczy et al. "Comprehensive bioimaging with fluorinated nanoparticles using breathable liquids". In: *Nat. Commun.* 6.May 2014 (2015), p. 5998.

[15] Rosa Pilolli, Francesco Palmisano, and Nicola Cioffi. "Gold nanomaterials as a new tool for bioanalytical applications of laser desorption ionization mass spectrometry". In: *Anal. Bioanal. Chem.* 402.2 (Jan. 2012), pp. 601–623.

[16] Jicheng Duan et al. "CHCA-modified Au nanoparticles for laser desorption ionization mass spectrometric analysis of peptides." In: *J. Am. Soc. Mass Spectrom.* 20.8 (Aug. 2009), pp. 1530–9.

[17] Yu-Fen Huang and Huan-Tsung Chang. "Analysis of adenosine triphosphate and glutathione through gold nanoparticles assisted laser desorption/ionization mass spectrometry." In: *Anal. Chem.* 79.13 (July 2007), pp. 4852–9.

[18] John A McLean, Katherine A Stumpo, and David H Russell. "Size-selected (2-10 nm) gold nanoparticles for matrix assisted laser desorption ionization of peptides." In: *J. Am. Chem. Soc.* 127.15 (Apr. 2005), pp. 5304–5.

[19] Jeongjin Son, Gwangbin Lee, and Sangwon Cha. "Direct analysis of triacylglycerols from crude lipid mixtures by gold nanoparticle-assisted laser desorption/ionization mass spectrometry." In: *J. Am. Soc. Mass Spectrom.* 25.5 (May 2014), pp. 891–4.

[20] Chih-Lin Su and Wei-Lung Tseng. "Gold nanoparticles as assisted matrix for determining neutral small carbohydrates through laser desorption/ionization time-of-flight mass spectrometry." In: *Anal. Chem.* 79.4 (Feb. 2007), pp. 1626–33.

[21] Hsin-Pin Wu et al. "Gold nanoparticles as assisted matrices for the detection of biomolecules in a high-salt solution through laser desorption/ionization mass spectrometry." In: *J. Am. Soc. Mass Spectrom.* 20.5 (May 2009), pp. 875–82.

[22] Naoko Goto-Inoue et al. "The detection of glycosphingolipids in brain tissue sections by imaging mass spectrometry using gold nanoparticles." In: *J. Am. Soc. Mass Spectrom.* 21.11 (Nov. 2010), pp. 1940–3.

[23] Martin Dufresne et al. "Silver-Assisted Laser Desorption Ionization For High Spatial Resolution Imaging Mass Spectrometry of Olefins from Thin Tissue Sections". In: *Anal. Chem.* 85.6 (Mar. 2013), pp. 3318–3324.

[24] Martin Dufresne, Jean-François Masson, and Pierre Chaurand. "Sodium-Doped Gold-Assisted Laser Desorption Ionization for Enhanced Imaging Mass Spectrometry of Triacylglycerols from Thin Tissue Sections". In: *Anal. Chem.* 88.11 (June 2016), pp. 6018–6025.

[25] Taryn M Guinan et al. "Silver Coating for High-Mass-Accuracy Imaging Mass Spectrometry of Fingerprints on Nanostructured Silicon". In: *Anal. Chem.* 87.22 (Nov. 2015), pp. 11195–11202.

[26] Justyna Sekuła et al. "Gold nanoparticle-enhanced target (AuNPET) as universal solution for laser desorption/ionization mass spectrometry analysis and imaging of low molecular weight compounds". In: *Anal. Chim. Acta* 875 (May 2015), pp. 61–72.

[27] Pere Ràfols et al. "rMSI: an R package for MS imaging data handling and visualization". In: *Bioinformatics* 33.15 (Mar. 2017).

[28] David S Wishart et al. "HMDB 3.0–The Human Metabolome Database in 2013." In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D801–7.

[29] Rohit Shroff, Alexander Muck, and Aleš Svatoš. "Analysis of low molecular weight acids by negative mode matrix-assisted laser desorption/ionization time-of-flight mass spectrometry". In: *Rapid Commun. Mass Spectrom.* 21.20 (Oct. 2007), pp. 3295–3300.

[30] Takahiro Hayasaka et al. "Imaging mass spectrometry with silver nanoparticles reveals the distribution of fatty acids in mouse retinal sections." In: *J. Am. Soc. Mass Spectrom.* 21.8 (Aug. 2010), pp. 1446–54.

[31] Rüdiger Frey and Armin Holle. "Effect of different laser beam angles on performance of MALDI TOF with gridless ion optics". In: *Bruker Daltonik Tech. Notes* 15 (Apr. 2015).

3.7 Appendix

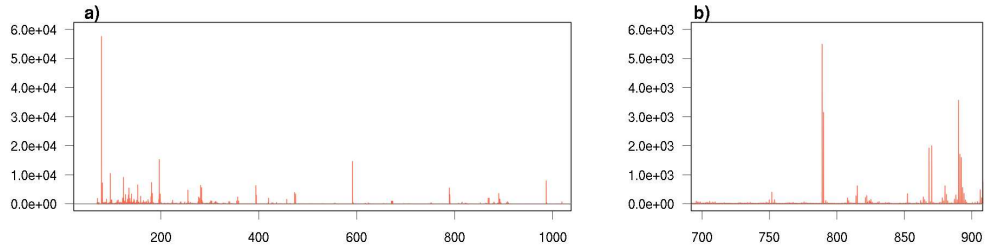


Figure 3.7: Average spectrum of a mouse liver section acquired in reflectron negative mode using the 35 s sputter coated gold layer. **A)** The full MS spectrum until m/z 1000 and **B)** Zoom of the MS spectrum between m/z 700 and 900.

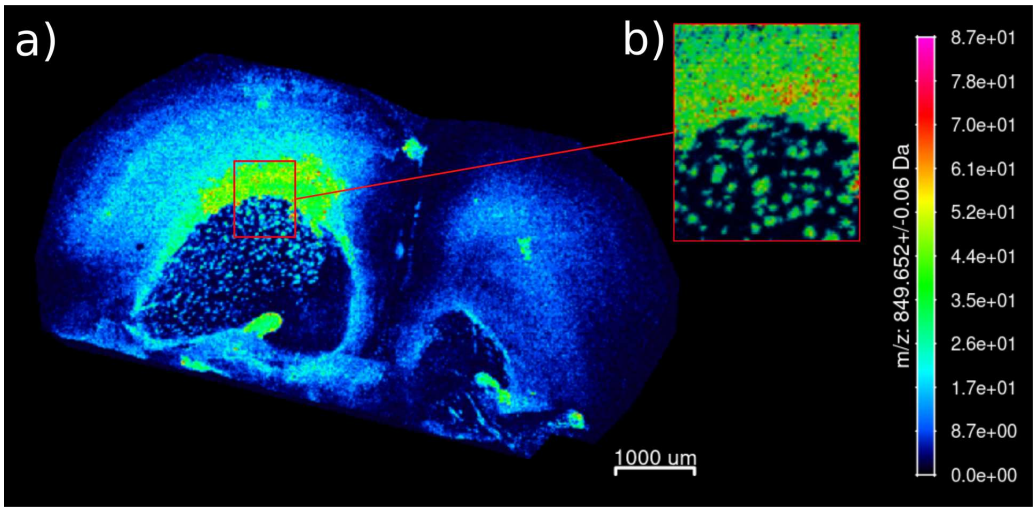


Figure 3.8: Mouse brain tissue section acquired high spatial resolution using a Bruker MALDI-TOF/TOF rapifleX instrument. **A)** Complete brain coronal section acquired at 20 µm. **B)** A small part of corpus callosum and striatum acquired at 10 µm using the same tissue section.

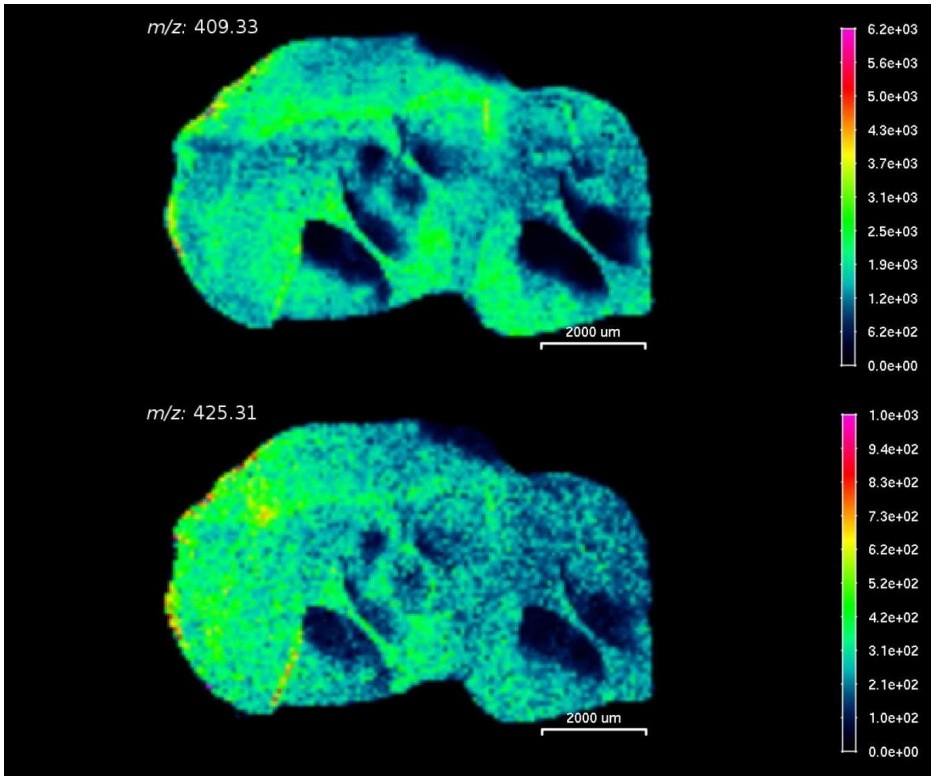


Figure 3.9: MS images of ions m/z 409.33 and 425.31 of mouse brain tissue section. These ions presents two highly correlated images that have been putatively assigned as cholesterol ($[C_{27}H_{46}O+Na]^+$ and $[C_{27}H_{46}O+K]^+$ respectively).

Chapter 4

rMSI: an R package for MS imaging data han-dling and visualization

4.1 Abstract

R platform provides some packages that are useful to process mass spectrometry imaging (MSI) data; however, none of them provide an easy to use graphical user interface (GUI). Here, we intro-duce rMSI, an R package for MSI data analysis focused on providing an efficient way to manage MSI data together with a GUI integrated in R environment. MS data is loaded in rMSI custom format optimized to minimize the memory footprint yet maintaining a fast spectra access. The rMSI GUI is designed for simple and effective data exploration and visualization. Moreover, rMSI is designed to be integrated in the R environment through a library of functions that can be used to share MS data across others R packages. The release of rMSI for R environment establishes a novel and flexible platform for MSI data analysis, completely free and open-source.

4.2 Introduction

Mass spectrometry imaging (MSI) is an emerging technique capable of mapping the spatial distribution of molecules in biological tissues with high spatial resolution. The MSI instrument scans the sample in a defined raster acquiring a MS spectrum for each pixel [1]. The MSI experiment produces large datasets that require specific software tools to be processed. Proprietary software packages, generally associated to each specific mass spectrometer, are available to analyze MSI data. However, they are either expensive or exclusive to each vendor and their closed-source model makes impossible to modify the code to explore all MSI possibilities. A few MSI software tools have been released under an open source license allowing a wide availability and easy code modification. MSiReader [2] is an open-source toolbox for Matlab platform that provides a full-featured graphical user interface (GUI) for MSI data exploration. MSiReader is also freely available as a standalone program which does not require a Matlab license. However, the main drawback of MSiReader is the lack of a memory optimized data handling model. Recently, SpectralAnalysis [3] has been released as another Matlab tool that provides an efficient data handling model for MSI data. However, both of these Matlab tools require a commercial software license to develop and modify its source code. On the other hand, R platform is becoming a largely used solution for the development of bioinformatics tools in a completely open-source model. To date, Cardinal [4] is the only available MSI specific tool developed for the R

environment. Cardinal provides many pre-processing algorithms and image segmentation routines. Cardinal is able to handle large datasets by pre-processing MS data directly from disk, however no GUI for fast and easy exploration of the data is provided. Here, we present rMSI, an R package focused on integrating MSI exploration in an R-based environment through a GUI that allows a rapid, responsive and easy visualization and comparison of MS images. The integration of rMSI in R establishes a flexible and reliable platform for MSI data analysis.

4.3 The graphical user interface (GUI)

A user-friendly graphical interface is included in rMSI to facilitate fast MSI data exploration. A screenshot of rMSI main GUI is displayed in Figure 4.1. The GUI is divided in four areas: "spectra list" (Fig. 4.1A), "MS image" (Fig. 4.1B), "intensity scales" (Fig. 4.1C), and "spectra visualization" (Fig. 4.1D). MS images are displayed together with their color codification represented in "intensity scales". The spectra visualization window shows the average spectrum of the whole MS image by default, but the spectra from different pixels can be also overlaid. Ion images can be reconstructed by selecting a m/z range in the spectra view area or entering m/z and tolerance values using the keyboard. Up to three ions can be plotted simultaneously, encoding each ion image in a color channel of an RGB color system. The GUI also allows drawing a rectangular region of interest (ROI) over the MS image (Fig. 4.1E) to perform actions to a selected set of pixels. Moreover, rMSI provides a special mode for comparing two MS images simultaneously, for example in a disease versus control study. This feature displays two MS images laid out side by side with a common spectra view (Fig. 4.2 in the appendix). In this mode, selected ions are automatically rendered in both MS images areas at once.

4.4 MS data handling strategy

MS images use a large amount of memory since they contain a large collection of spectra. The memory needed could range from hundreds of megabytes to several tens of gigabytes, depending on the m/z resolution and the total number of acquired pixels. Such amount of spectra may be difficult to handle in the computer's memory (RAM), especially if more than one image has to be loaded at once, e.g.

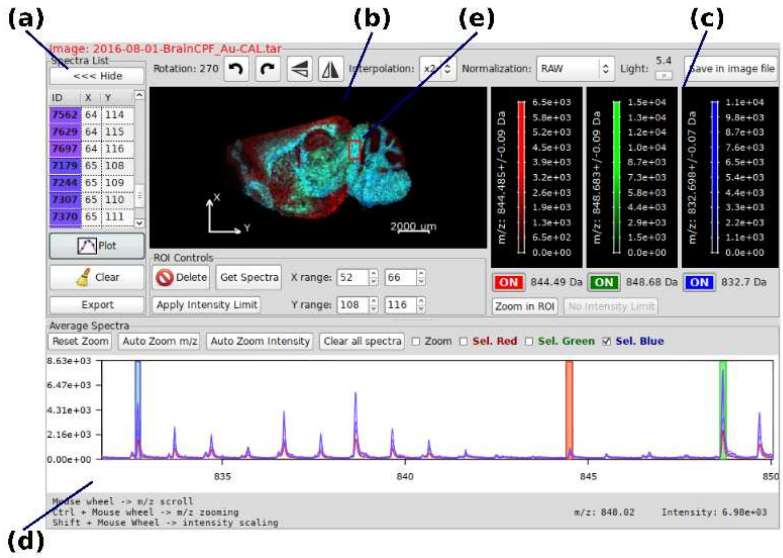


Figure 4.1: Screenshot of rMSI's main GUI displaying a MS image from a mouse brain sagittal section. Spectra were acquired using a Bruker MALDI-TOF/TOF UltrafleXtreme in a mass range of m/z 80 to m/z 1000 and a spatial resolution of 80 µm. **A)** Spectra list area displays a collection of selected pixels. **B)** MS Image area displays the ion image in a single channel or triple channel. **C)** Intensity scale area displays the color intensity mapping for each selected ion. **D)** Spectra viewer allows selecting an ion on the spectrum. **E)** Region of Interest (ROI) user selected ROI is displayed as a red rectangle.

to compare tissues from a case-control study. To overcome this, rMSI provides a data format designed to combine RAM memory with the available hard disk drive (HDD) free space. First, the mass spectrum from each pixel is uniquely identified with an ID number allowing a fast and controlled data access. Then, MS data is split into different blocks that are stored uncompressed in the HDD applying neither m/z binning nor any data reduction strategy. Each data block is stored in an R matrix where each mass spectrum is located in a row. These matrices are stored in HDD files sorting data by columns; hence intensities of a set of neighbor m/z channels can be obtained in a single disk reading operation. This design allows a fast ion image reconstruction. To improve the spectra loading time, each data matrix is limited to 50 MB in size. This allows loading a whole matrix to RAM for fast row access during spectral processing. The rMSI GUI takes advantage of this data model by only loading the part of the image that is being represented. This low memory footprint design allows exploring various high resolution (spatial and spectral) MS images simultaneously in a standard laptop computer.

Data access performance has been tested for various datasets on a laptop with an Intel Core2Duo 2 GHz processor, 4 GB of RAM and a 5400 rpm disk. Despite

of the bottleneck that represents accessing data from the HDD, rMSI provided ion image reconstruction times ranging from 2 to 8 seconds depending on the data size (see appendix Table 4.1 and Fig. 4.3). The best performance was obtained for a 3.7 GB dataset and the worst performance was obtained for a 31.5 GB dataset containing almost 106 m/z channels synthetically created to simulate a large image acquired with a MS high resolving power spectrometer.

In addition to its own format, rMSI allows importing data from the open standard imzML [5] and Bruker's XMASS. Once the data is loaded in rMSI, the generated files are organized to be reusable in future R sessions allowing an immediate loading time for next R sessions.

4.5 Conclusion

The developed R package presented here fulfills the requirement of a user-friendly interface integrated in the popular and open-source R environment. The developed data format allows a fast and snappy MSI data exploration of high-resolution images in standard computers. The usage of rMSI integrated in R environment provides a flexible and powerful platform for MSI data handling and analysis.

References

[1] Jeremy L Norris and Richard M Caprioli. "Imaging mass spectrometry: A new tool for pathology in a molecular age". In: *PROTEOMICS - Clin. Appl.* 7.11-12 (Dec. 2013), pp. 733–738.

[2] Guillaume Robichaud et al. "MSiReader: An Open-Source Interface to View and Analyze High Resolving Power MS Imaging Files on Matlab Platform". In: *J. Am. Soc. Mass Spectrom.* 24.5 (May 2013), pp. 718–721.

[3] Alan M. Race et al. "SpectralAnalysis: Software for the Masses". In: *Anal. Chem.* 88.19 (Oct. 2016), pp. 9451–9458.

[4] Kyle D Bemis et al. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". In: *Bioinformatics* 31.14 (July 2015), pp. 2418–2420.

[5] Thorsten Schramm et al. "imzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5106–10.

4.6 Appendix

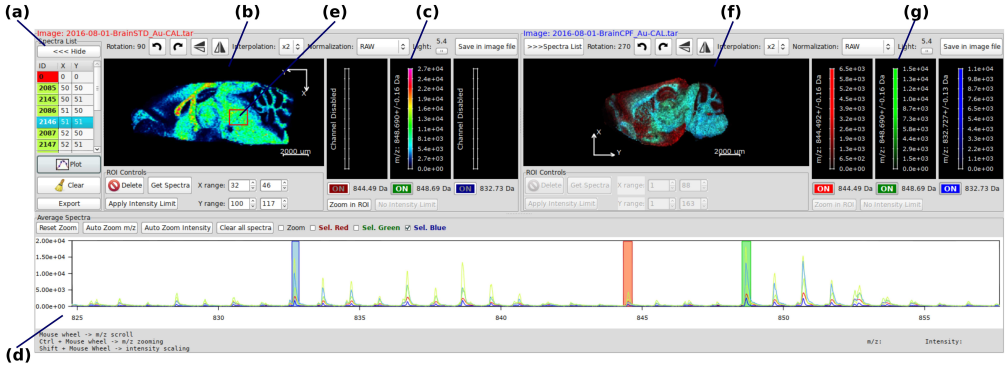


Figure 4.2: Screenshot of rMSI's main GUI used to explore two MS images from a mouse brain sagittal section in dual mode. Here, two MS images are laid out side by side to compare case vs. control tissues. Spectra were acquired using a Bruker MALDI-TOF/TOF UltrafleXtreme in a mass range of m/z 80 to m/z 1000 and a spatial resolution of 80 µm. **A)** Spectra list area displays a collection of manually selected pixels with their ID's. **B)** MS image 1 area displays the ion intensity distribution of the first MS image in a single channel. **C)** Image 1 intensity scale window displays the color intensity mapping for the selected ion. A rainbow color scale is used here because only one ion is selected. **D)** Spectra viewer allows selecting an ion on the spectrum to render its image in **(B)** and **(F)** areas. **E)** Region of interest (ROI) user selected ROI is displayed as a red rectangle. **F)** MS image 2 area displays three selected ion intensities distributions of the second MS image in triple channel mode. Here an RGB color encoding is used to represent the three ions in a single image. **G)** Image 2 intensity scale window displays the color intensity mapping for the selected ions. Here, each ion intensity is represented by a red, green or blue color.

	Dataset 1	*Dataset 2*	*Dataset 3*	*Dataset 4*	*Dataset 5*	*Dataset 6*
Tissue sample:	Mouse brain sagittal section	Mouse Liver	Mouse brain sagittal section	Mouse brain sagittal section*	Mouse Liver*	Mouse brain sagittal section*
Spatial resolution:	80 µm	40 µm	40 µm	40 µm	40 µm	80 µm
Number of pixels:	10020 pixels	14318 pixels	27955 pixels	27955 pixels	14318 pixels	8530 pixels
TOF detector sample rate:	2,5 GS/s	2,5 GS/s	1,25 GS/s	2,5 GS/s	1,25 GS/s	2,5 GS/s
Number of m/z channels:	92330 *m/z* bins	92330 *m/z* bins	52565 *m/z* bins	157695 *m/z* bins	369320 *m/z* bins	923300 *m/z* bins
Data size uncompressed:	3,7 GB	5,3 GB	5,9 GB	17,6 GB	21,2 GB	31,5 GB
Data size compressed:	1,4 GB	2,0 GB	2,0 GB	5,9 GB	7,9 GB	12,5 GB
rMSI unpacking time:	3 minutes	4 minutes	5 minutes	11 minutes	15 minutes	25 minutes
Ion image reconstruction time (0,05 Da tolerance):	2,3 seconds	3,1 seconds	3,5 seconds	5,2 seconds	6,3 seconds	8,2 seconds

Table 4.1: Results of rMSI performance tests using an outdated laptop computer featuring an Intel Core2Duo 2 GHz processor, 4 GB of RAM and standard 5400 rpm hard disk drive. Six MSI datasets with different number of pixels and m/z channels were tested to measure the rMSI data format's scalability. All data was acquired using a Bruker MALDI-TOF/TOF UltrafleXtreme spectrometer with various raster size settings and sampling rates. *Datasets 5, 4 and 6 are synthetic datasets created to simulate rMSI behavior for high mass resolution datasets (FT-ICR). Synthetic datasets (*) were created by extending the m/z axes of real datasets.

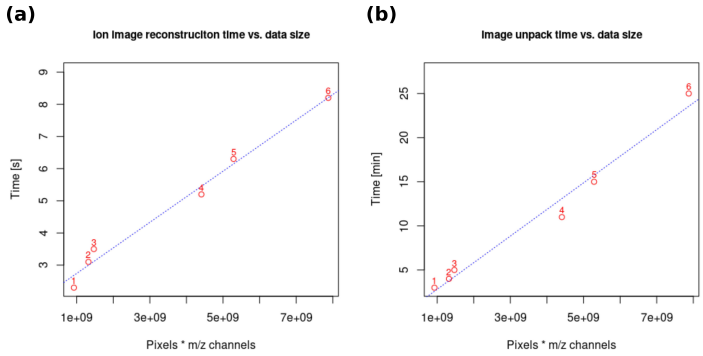


Figure 4.3: Plot of spectra access time using data from table 4.1. Here, data size is calculated as the multiplication of number of pixels by the number of mass channels (all spectra intensities are encoded in 32 bits integers). The corresponding dataset is displayed for each data point as an overlaid number in red color. **A)** Ion image reconstruction time vs. data size displays the relationship between the data size and the time used to display an ion image. **B)** Image unpacking time vs. data size displays the relationship between the data size and the time used to unpack data from rMSI's compressed format. As can be seen, the data access time scales pretty linearly independently if the data size increments are due to number of pixels or number of m/z channels.

Chapter 5

Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer

5.1 Abstract

Mass spectrometry imaging (MSI) is a technique that can map analyte spatial distribution directly onto a tissue section. This enables the spatial correlation of molecular entities with a tissue morphology to be investigated. Analyte annotation in MSI is intrinsically linked to the mass accuracy of the data. Mass accuracy and analyte identification are determined by such factors as the experimental set up and the data processing workflow. We present an MSI data processing workflow that uses a label-free approach to compensate for mass shifts. The algorithms developed were designed to perform efficiently even for large datasets generated from an FTICR mass spectrometer. We assessed the overall mass accuracy in the range m/z 400 to 1200 using silver and gold sputtered nanolayers. With our novel processing workflow we were able to obtain mass errors as low as 5 ppm using a TOF instrument.

5.2 Introduction

Classical histology visually inspects stained tissue sections and identifies specific regions, and is an essential tool in medical diagnosis. However, histology is not capable of determining the chemical composition of tissue regions and must be complemented with other techniques if molecules are to be characterized. This requires additional sample preparation steps and analysis. In recent years, mass spectrometry imaging (MSI) has emerged as a straightforward alternative to this end [1, 2]. MSI, also called molecular histology, consists of acquiring molecular images from biological tissue sections and extracts the chemical information directly from the tissue.

The most widely used ionization technique in MSI applications is Matrix Assisted Laser Desorption Ionization (MALDI) [3, 1]. It applies an organic matrix to thin tissue sections to promote the desorption/ionization of proteins, lipids and other metabolites. Nevertheless, one of the main drawbacks of MALDI is that the organic matrices introduce a considerable number of MS signals in the low m/z range of the spectrum (below 1000 Da). These signals severely interfere with the MS peaks arising from endogenous low molecular weight compounds, which makes their application to metabolomic studies a challenge [4].

Matrix-free LDI-MS techniques have emerged in recent years as valuable alternatives for the analysis of metabolites. Metal sputter coating has been introduced

as an efficient deposition method suitable for MSI using gold and silver nanolayers [5, 6]. Sputtering is a dry deposition technique, which deposits highly pure and homogeneous metal or metal oxide nanolayers on biological tissues. Furthermore, the characteristic peaks from metal clusters can be used for internal mass calibration throughout the m/z regions of the spectrum obtained [5, 7].

MS spectral peaks obtained on tissue sections display some degree of mass misalignment between pixels because of several experimental factors. In the case of time of flight (TOF) mass analyzers, tissue-surface irregularities (or sample topography) in conjunction with spec-trometer drift due to small dilatations/contractions of the flight tube, or slight variations in the high-voltage power supply commonly trigger mass drifts [8, 9]. Due to this mass shift, a wider mass tolerance may be necessary if ion images are to be properly displayed. Therefore, in peak crowded regions of the spectrum different compounds may be mixed when images are reconstructed. This experimental variability hampers molecular identification because various peaks originating from different compounds may be detected within the same m/z window. To improve ion image reconstruction, several alignment and recalibration strategies have been proposed to improve feature selection and molecule annotation. Classical methods for improving measurement mass accuracy (MMA) are related to mass calibration. Two main strategies are generally used: external calibration and internal calibration. External calibration methods consist of determining a calibration function using standard compounds placed side by side with the tissue section. Then, the same calibration function is applied to the whole dataset. This methodology is useful for calibrating the instrument before starting the MSI measurement, but it cannot compensate for mass shifts introduced during acquisition. The internal calibration approach uses known molecules present in every pixel of the tissue section as mass references. These reference molecules can be known endogenous tissue compounds, standard compounds sprayed with the organic matrix, or peaks of the matrix itself. Although there is no justification in the papers, a mass accuracy better than 20 ppm was reported using silver cluster peaks as references for internal mass calibration [5]. More recently, a 10 ppm mass accuracy was achieved using gold cluster peaks as calibration references [6]. Nevertheless, the mass accuracy has not been validated throughout the analyzed mass range. MS peaks associated with cationic gold nanoparticles has also been used to calculate the mass accuracy [10], although the error has not been quantified. The main drawback of the internal calibration method is that not all the reference

molecules can be properly detected in all the pixels of the MS image. This means that the MMA cannot be ensured for every peak list in a dataset. An alternative method is to align all MS spectra in the same mass axis so an overall recalibration function can be applied to the whole dataset. This methodology can be referred to as label-free alignment because no reference compound is required for all the data to be aligned. This alignment strategy calibrates a complete dataset as accurately as a single spectrum calibrated using internal reference peaks. Furthermore, with a proper alignment, molecules can be putatively identified directly from the peak lists of the whole dataset. It has been demonstrated that it is possible to successfully align spectra acquired with Fourier Transform Ion Cyclotron Resonance (FTICR) spectrometers using ion abundance [11] or ambient peaks [12]. FTICR detectors are very robust to mass drifts because the mass shift is only related to the charge in the ICR cell. Therefore, if ion intensities are used to predict the mass error of each spectrum there is a considerable improvement in MMA. In order to compensate for mass shifts in data generated with a TOF instrument other algorithms must be used because, in this case, the mass shift is not related to peak intensity. Moreover, spectra generated by TOF instruments display far more variability than spectra recorded with FTICR. Tracy et al. [9] introduced a method to align TOF data in the time domain instead of the m/z domain. This method states that corrections in the time domain provide better accuracy than in the m/z domain since most TOF variations are linear in the time domain. More recently, a novel approach that models the mass shifts using sorting algorithms has been introduced [13]. These alignment studies have tested for MMA validation using some compounds found to be present throughout the tissue section. Nevertheless, the mass error across the full m/z range has not been reported since not enough reference compounds have been confidently identified.

Herein we present a novel and complete MSI pre-processing workflow which can align and recalibrate large MSI datasets automatically. Unlike most previous alignment methods, the alignment algorithm developed is designed to work directly in the MS spectra domain rather than use peak-lists, which means that the peak shape is taken into account for the alignment. This is an advantage over other strategies because the resulting spectra can be accessed afterwards and used for interactive ion image reconstruction. The algorithm presented uses a spectral cross-correlation approach to obtain a mass axis warping method that can compensate for non-linear mass shifts. All the processing is designed to make intense use

of the Fast Fourier Transform (FFT) which reduces the use of computer resources even for large datasets. The workflow presented here has been tested on TOF and FTICR MS data. We also introduce an experimental methodology based on sputtered metal-nanolayer deposition that is used to assess mass error throughout the spectrum. We used an Ag and Au bilayer deposited directly onto the tissue section to promote LDI ionization and provide enough reference peaks to study the MMA. The peaks of the Ag, Au and AgAu clusters are easily identified due to the theoretical isotopic pattern of Ag. These peaks are detected throughout the mass range from m/z 400 to 1200. The peaks are used to provide a reliable methodology for studying the mass accuracy for a given instrumental platform and processing workflow. Using a reflector TOF instrument and our pre-processing workflow we demonstrate that we can obtain a peak matrix of the whole dataset with mass errors as low as 5 ppm in a mass range from m/z 600 to 1200 when Ag peaks are used as references for calibration. The pre-processing methodology presented here considerably improves the MMA of the complete peak matrix, which contains all the relevant peak information from an MSI experiment using an Ag sputtered layer or any other set of reference peaks properly distributed across the m/z range.

5.3 Materials and methods

5.3.1 Materials

Indium tin oxide (ITO) coated glass slides were obtained from Bruker Daltonics (Bremen, Germany). The gold and silver targets used for sputtering coating were obtained from Kurt J. Lesker Company (Hastings, England) with a purity grade higher than 99.995%.

5.3.2 Sample preparation

Liver was obtained from C57BL/6 mice, snap frozen at -80°C after collection and stored and shipped at this temperature until analysis. Dr M. Teresa Colomina, professor of Psychobiology at the Research Center for Behavioral Assessment (CRAMC) of the Universitat Rovira i Virgili provided the animal tissues. The tissues were sectioned at -20°C in slices 10 μm thick using a Leica CM-1950 cryostat (Leica Biosystems Nussloch GmbH) located at the Centre for Omic Sciences (COS) of the Universitat Rovira i Virgili and mounted on ITO coated slides by

directly placing the glass slide at ambient temperature onto the section. To remove residual humidity, samples were dried in a desiccator under vacuum for 15 minutes after tissue mounting.

5.3.3 Sputter coating

Silver and gold nanolayers were deposited over the 10 μm tissue sections using an ATC Orion 8-HV (AJA International, N. Scituate, MA, USA) sputtering system. An argon atmosphere with a pressure of 30 mTor was used to create the plasma in the gun. The working distance of the plate was set to 35 mm. The silver layer was deposited in DC mode at 100 W for 10 s. The gold layer was deposited in RF mode at 60 W for 35 s. These deposition modes were selected so that the liver tissue could be coated with both metals faster and without pumping the sputter chamber vacuum various times to replace the target. Since the deposition times used in this study were very short, the substrate temperature did not increase during the deposition.

5.3.4 LDI-MS acquisition

MS tissues images were acquired using a MALDI TOF/TOF UltrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics, also at the COS facilities. Acquisitions were carried out using a large laser spot, operated at 2 kHz at an attenuated power of 50 %, collecting a total of 500 shots per pixel with a raster size of 100 μm. MS spectra were acquired in positive reflection mode, at 2.5 GHz in a mass range between m/z 400 to 1200, with a manually optimized extraction delay. The spectrometer was calibrated prior to tissue image acquisitions using Ag^+ peaks as reference masses.

5.4 Processing workflow

MS images were acquired using FlexImaging 3.0 software from Bruker. Each image was exported to XMASS data format using instrument manufacturer software packages (FlexImaging and Compass export). The raw data was loaded using the rMSI package written in-house [14]. This package provides a data storage format based on segmented matrices and optimized for processing large MSI datasets in R language. Then a pre-processing workflow consisting of smoothing, alignment, recalibration and peak detection was applied.

5.4.1 Smoothing

The first step in the processing chain is a smoothing stage using the well-known Savitzky-Golay algorithm [15] which has the important property of retaining the exact position of mass peaks. The smoothing stage improves the performance of the following processing methods because it reduces noise. Therefore, the alignment routine becomes more robust since the random noise reduction provides a better correlation between spectra. Peak detection is also improved because it becomes less sensitive to noise peaks.

5.4.2 Label-free alignment

The spectra of an MS image present some degree of mass miss-alignment between pixels for several experimental reasons. The goal of a label-free alignment algorithm is to project all the spectra onto the same mass axis without using any reference peaks. Here, we describe an alignment method based on spectral correlations that can compensate for the experimental mass drifts.

The first step of the algorithm is to select a single spectrum as an internal reference. The simplest approach is to use the average spectrum of all pixels as a reference. However, for long acquisitions the average spectrum may exhibit very wide peaks or even double peaks as a consequence of instrumental mass drift. To avoid this, we use the spectrum of a single selected pixel in the MS image as reference. The algorithm calculates the correlations of each spectrum to the average spectrum. Then, the pixel with the highest correlation is selected as the reference. Since this reference spectrum is automatically chosen and no standard compounds are required, this methodology is considered label-free.

The alignment algorithm presented here is based on the cross-correlation theorem through the Fast Fourier transform (FFT). However, the method presented not only compensates for an offset in the m/z axis; it also performs non-linear corrections. This is done by calculating two correlation coefficients for each spectrum. The alignment algorithm is summarized in Fig 5.1.

Each spectrum in the dataset is split into two parts, the bottom part that includes the lower m/z channels and the top part for the higher m/z channels. Each part is calculated by windowing the spectrum using a Hanning function which allows a smooth transition in the central part (Fig. 5.1A). Using a Hanning instead of rectangular functions provides more accurate correlations because the spectrum

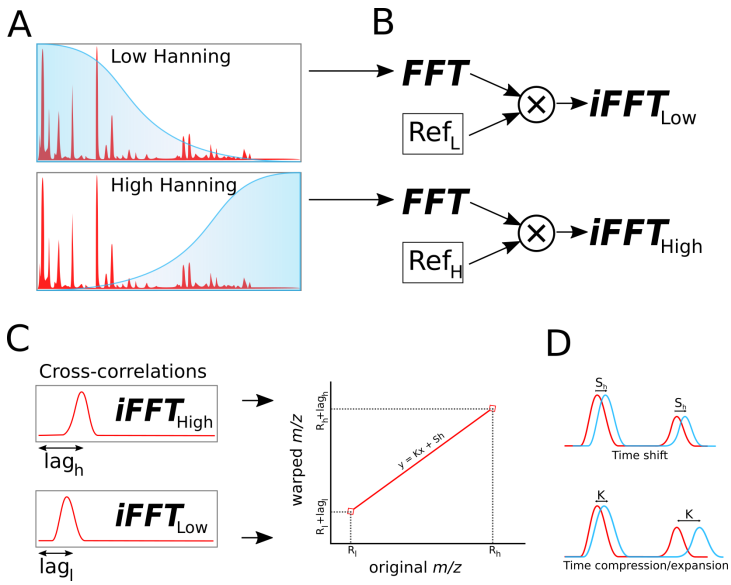


Figure 5.1: Spectra alignment algorithm flow chart. **A)** A Hanning window is applied to each part of the spectrum to emphasize only the lowest or highest region in the cross-correlations calculation. **B)** Each part of the windowed spectrum is Fourier transformed and multiplied by the FFT of the windowed spectrum selected to be used as the internal reference. **C)** The time shifts (lag_l and lag_h) for each part of each spectrum are obtained by locating the peak in the cross-correlation function. The lags are mapped to the S_h offset and k coefficient in order to be applied to the spectrum. **D)** Each original spectrum is time shifted and time warped according to its S_h and k values.

is not abruptly cut at the middle. After applying the Hanning window, the FFT of each part of the spectrum is calculated. To improve the FFT performance, zero padding is used to obtain a power of two lengths of all vectors. These steps are done in first place using the selected reference spectrum. Then, the resulting frequency transformations are conjugated and kept in memory of the computer. Once the reference transformations have been calculated, the same procedure is applied for each spectrum except for the complex conjugate operation (Fig. 5.1B). Then, each part of the spectrum is multiplied by the complex conjugate of reference in the frequency domain. The result of this operation is a function that corresponds to the Fourier transform of the cross-correlation of each spectrum with the reference. See equation 5.1 where r corresponds to each one of the half parts of the reference spectrum and x_k corresponds to the half-spectrum of a given pixel k.

$$\mathcal{F}\{\mathrm{r} \star \mathrm{x_k}\} = (\mathcal{F}\{\mathrm{r}\}^*) \cdot \mathcal{F}\{\mathrm{x_k}\} \tag{5.1}$$

The time shift that maximizes the correlation between each part of the spectrum (bottom and top) is obtained by locating the maxima at the inverse Fourier

transform of each cross-correlation value $F\{\mathrm{r} \star \mathrm{x_k}\}$ (see Fig. 5.1C). Two time shifts are obtained for each pixel, one for the bottom part of the spectrum (lag_l) and the other for the top part of the spectrum (lag_h). However, these time shifts cannot be applied directly to the spectrum because this would involve breaking the m/z axis at its center. Instead, these two time shifts are transformed to two equivalent parameters: time shift offset S_h and time scaling coefficient K. The time shift represented by S_h consists of a shift that will be applied to the whole spectrum. The shift parameter aims to compensate for measurement drifts because the flight distance varies with tissue roughness. The time scaling, K will take values smaller than one if the spectrum must be contracted and values higher than one if the spectrum must be expanded. Thus, the peak widths will narrow when the spectrum is compressed and the peaks will widen when it is expanded. The K parameter is designed to compensate for TOF drifts introduced by flying tube thermal contraction/expansion and accelerating voltage variation.

To obtain the scaling constant K and the time shift offset S_h a linear transformation is applied. The aligned spectrum must display the times shifts (lag_l) and (lag_h) at the bottom and top, respectively. Therefore, the K and S_h values are calculated using equations 5.2 and 5.3 for each pixel in the dataset. This process will obtain the line equation that maps the original mass axis to the warped mass axis with proper shifts applied (Fig. 5.1C). Here, two new variables are introduced: R_l and R_h which represent the m/z positions where the (lag_l) and (lag_h) will be applied to the mass axis. Generally, R_l and R_h are set at the minimum and maximum m/z values of the spectrum, respectively. However, these parameters can be tuned to compensate for the effect of some really intense peaks that may influence the cross-correlation procedure excessively.

$$K = \frac{R_h + lag_h - R_l - lag_l}{R_h - R_l} \tag{5.2}$$

$$S_h = \frac{R_h \cdot lag_l - R_l \cdot lag_h}{R_h - R_l} \tag{5.3}$$

Then the time shift S_h and time scaling K are applied to all spectra by means of the FFT properties. First, the spectra are scaled using FFT downsampling or oversampling methods depending on whether the scaling parameter is greater or less than one. Downsampling is achieved by removing samples from the center section of the frequency domain. On the other hand, the oversampling adds zeros

to the center section of the frequency domain. The algorithm calculates how many sample points must be added or removed to properly approximate the K parameter. Thus, applying signal interpolation before the whole alignment methodology is useful if this step requires extra accuracy. After the scaling stage, the time shift S_h is applied in the frequency domain by means of the Fourier transform time shift property:

$$\mathcal{F}\left(x_k\left(t - S_h\right)\right) = X_k\left(\omega\right) \cdot e^{-j\omega S_h} \tag{5.4}$$

Here, the frequency domain spectrum is multiplied by the complex exponential time shift coefficient to obtain the shifted Fourier transform of the spectrum (Fig. 5.1D). Finally, the inverse FFT transform is applied to retrieve the aligned spectrum.

5.4.3 Mass recalibration

After the label-free alignment stage, all the spectra share a common m/z axis. Thus, the whole dataset can be recalibrated using the same reference. Here, the average spectrum is recalculated using the aligned spectra to avoid double peaks and wider peak issues. Then, the average spectrum is used to calculate the mass calibration function using reference masses. Each reference peak position is located in the average spectrum and recorded together with its theoretical mass. A loess smoother is used to predict the calibrated m/z axis. In this study, we used the Ag peaks from the AgAu layer as internal reference [5, 6, 10]. However, standard compound or matrix peaks can be used for this purpose as well. Refer to Table 5.1 in the appendix for the complete list of peaks used for calibration.

5.4.4 Peak detection

After the alignment and recalibration stage, peaks can be detected with reduced mass error. However, the number of data points used to represent a peak is not accurate enough if the peak is represented directly by a data point. Previous work has been done to improve peak detection mass accuracy using algorithms that model each peak in the spectrum as a mathematical function. Algorithms such as OMP [16] or centroid Gaussian fitting [9] can detect peak mass positions with high accuracy although processing time increases. On the other hand, it has also been demonstrated that good mass accuracy can be achieved using peak

centroid cubic interpolation [17] which requires far less computation time. Here we present a fast and simple peak picking algorithm based on FFT interpolation. In the first stage, all the peaks in a spectrum are detected according to the classical mathematical definition: in a peak, the first curve derivative is zero and the second derivative is negative. However, this produces a long list of peak candidates, most of which actually come from noise. In order to decide which peak candidate is an actual peak we use a noise estimation model based on FFT filtering. Here, the whole spectrum is converted to the frequency domain and low-pass filtered using a decaying exponential function. The inverse Fourier transform of this signal provides a computationally fast estimation of the noise floor and the baseline. Then, the intensity of each peak candidate is compared to the corresponding value of the noise estimation. The peak candidates that are less intense than the desired signal to noise threshold are discarded. The next step consists of calculating each peak mass accurately. A modified Hanning window with three center samples set to one are applied for each detected peak. Then each peak is interpolated to higher resolution using FFT. The position of the maximum of the interpolated peak in the window is used to calculate the peak mass. The Hanning window assures no ringing effect in the peak interpolation stage which allows better mass prediction. This method enables fast peak detection with improved mass accuracy which is suitable for large datasets.

When all spectra have been converted to a list of peaks the next step is to merge everything into a single matrix to represent the whole dataset. Each matrix row corresponds to each pixel in the MS image and the columns contain the peak intensity. A vector with the same length as the number of columns is recorded to keep the m/z value of each peak. This matrix makes it possible to make a statistical analysis of the data as long as columns are treated as variables and rows as observations. The process of converting all the peaks list to a single matrix is known as peak binning. Here two filters are defined: binning tolerance and peak filtering. The first parameter is used to merge peaks from different pixels into a single peak matrix column when the mass difference is below the defined tolerance in ppm. The tolerance filter is related to the spectrometer resolving power and must be small enough to prevent peaks of different masses from merging and large enough to prevent peaks from the same molecule splitting into multiple peak matrix columns. The peaks that are not detected in any spectrum are written as zeros in the corresponding peak matrix cell. When all the peaks across the

dataset have been placed in the right location of the peak matrix, the m/z vector is calculated as the centroids of peak masses. Then, the second filter is applied to remove peak matrix columns that contain too few peaks. At this point the peaks in each column are counted and if there are not more than the threshold specified in % of total number of pixels the whole column is removed from the peak matrix. This filter ensures that less frequent peaks from noise will not bias any further statistical analysis.

5.5 Results and discussion

In order to validate the method described for full MSI dataset pre-processing, we developed a novel test methodology based on sputtered metal layer deposition. Instead of using standard compounds to test the performance of the alignment and calibration algorithms we deposited a bilayer of silver and gold nanoparticles on a section of mouse liver tissue. This bimetal layer provides many peaks distributed across the mass spectrum that can be easily identified and used to validate the mass calibration methodology. Since the metal bilayer contains silver and gold, the mass spectrum displays peaks from gold, silver and clusters of both metals. Table 5.1 in the appendix shows the complete list of all peaks detected from the sputtered layer. Here, the tissue section is only used to provide an efficient ionization surface because the metal layer directly on the ITO glass slide does not ionize properly.

After MS image acquisition we performed the processing steps described above. To test the accuracy of the alignment algorithm, a peak picking was conducted on the whole data set before and after the alignment stage. Here, the peak list of each spectrum is stored before and after alignment without a binning step. Thus, the detected peak mass of the same molecule in each spectrum shows a slight variation. This mass variation depends on factors like mass miss-alignment and peak resolution. Therefore, the aim of the alignment algorithm is to reduce the detected mass difference of the same molecule across the whole data set which allows the binning step to be executed at lower tolerance. The reduction in mass variation is also shown for larger acquisition images (more than 10000 pixels) in the appendix (Fig. 5.4). Here, we observed that the alignment algorithm can compensate for the TOF drift. An R script was created to calculate the mass drift of each peak list for each theoretical silver and gold peak. This information was used to construct plots of the detected mass for a given compound before

and after alignment (Fig. 5.2). As can be seen, after the alignment stage the detected mass displays far less variation (Fig. 5.2A). Histograms of detected mass distribution were also calculated (Fig. 5.2B). Refer also to Fig. 5.5 in the appendix which contains all the histograms of AgAu reference peaks. Here, we observe that the detected masses present a narrower distribution after the alignment stage (Fig. 5.2C). These plots are also useful for selecting an appropriate value for peak binning tolerance.

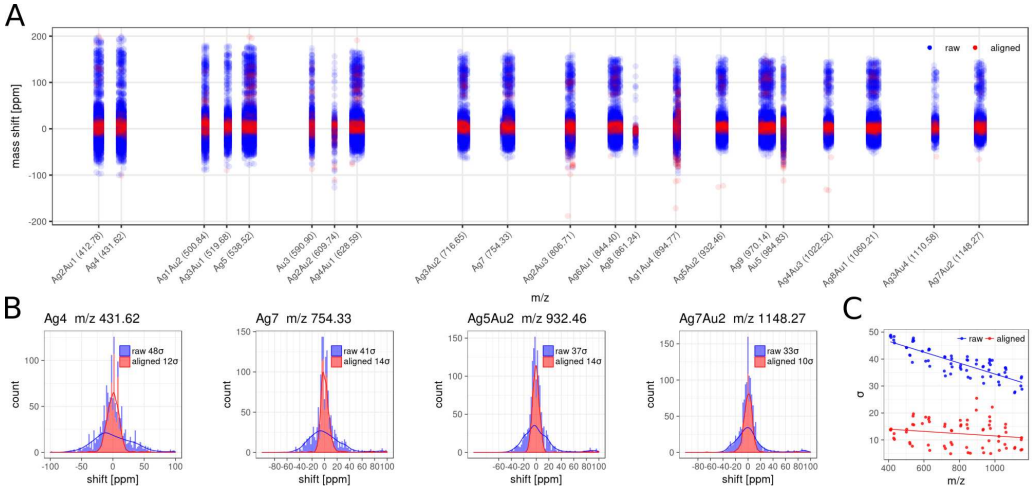


Figure 5.2: Label-free alignment evaluation using Ag, Au and AgAu peaks. **A)** Comparison of mass drift variation across the whole mass range before and after spectra alignment. **B)** Histograms of the detected peak mass distribution for several reference peaks: Ag_4, Ag_7, Ag_5Au_2 and Ag_7Au_2 (more histogram plots are provided in the appendix). **C)** comparison of standard deviation before and after the alignment routine.

After the alignment stage, the spectra were recalibrated using the average spectrum and the most intense peak of each silver cluster (see Table 5.1 in the appendix for further reference on masses used for calibration). Silver peaks were selected for mass calibration since they were more intense and distributed throughout the mass range. Moreover, silver peaks will also be present if an experiment is performed with just the silver layer. Then, the peak matrix of the whole dataset was obtained using the peak-picking and peak-binning procedure described above. The peak masses from the sputtered layer were identified in the final peak matrix using the theoretical silver/gold spectrum. Moreover, the isotopic distribution of the silver cluster was taken into account to provide a more robust identification. For each metal peak identified, the difference with the theoretical mass was calculated and is presented in Fig. 5.3 and Table 5.1 in the appendix. Using the mass errors obtained for each metal peak an error function can be predicted for the complete

mass range by applying loess smoothing to calculate errors throughout the m/z range. Fig. 5.3 shows the mass error obtained using Ag peaks as the calibration reference. Here we obtained errors as low as 5 ppm in the m/z range 400 to 1200 using TOF data in reflectron mode. Peaks with an error above 10 ppm are labeled in Fig. 5.3. These peaks present a higher mass error because they overlap with some endogenous compounds from the liver tissue, which distorts the peak shape. Fig. 5.6 is provided in the appendix to show the spectra of all the selected peaks in which this overlapping phenomenon can be seen.

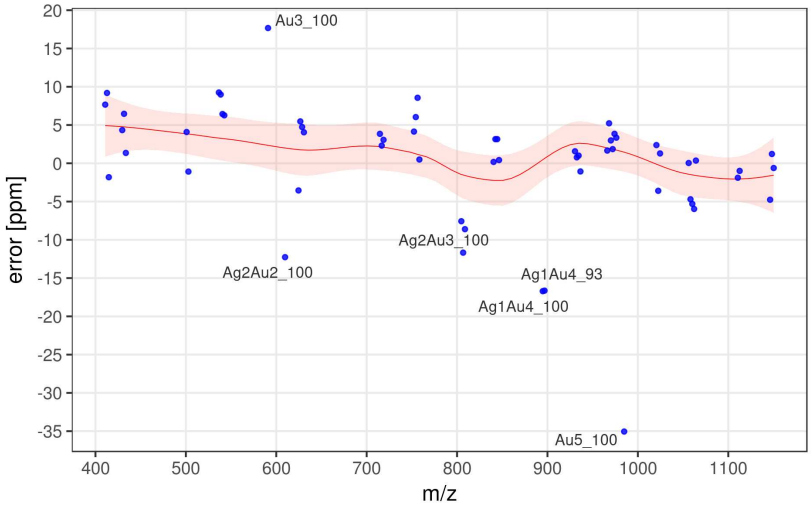


Figure 5.3: Mass measurement accuracy (MMA) prediction using Ag, Au and AgAu clusters as reference peaks. Peak masses were obtained after all processing steps (smoothing, alignment, recalibration, peak-picking and peak-binning). Then, peak masses corresponding to the AgAu sputtered layer were identified in the final binned peak matrix.

The pre-processing methodology discussed can take an MSI dataset and generate a mass-aligned and mass-calibrated version together with a peak matrix that summarizes all the peak information. A novel approach based on an AgAu sputtered layer was taken to provide a reliable validation of the MMA and demonstrate that mass errors as low as 5 ppm can be achieved for the complete peak matrix. This makes it possible to putatively identify molecules directly from the peak matrix instead of selecting specific spectra from the dataset. Unlike other TOF alignment strategies [9, 13], our alignment algorithm gives a new MS image that contains complete spectra instead of a peak list. This allows manual exploration and ion image reconstruction of the aligned dataset using the same MSI visualization software as can be used for raw data visualization. All of the processing

is done automatically and all the user needs to do is select the reference peaks to be used for mass calibration just after the alignment stage. This increases the throughput of MSI data since statistical analysis can be performed and molecules putatively identified directly from the peak matrix generated, which is an accurate low size representation of the original dataset.

5.6 Conclusion

We have presented a complete MS image pre-processing pipeline that can provide a data-reduced representation of MS data with high accuracy. We demonstrated that the label-free alignment algorithm developed can reduce the mass miss-alignment in the spectra by calculating cross-correlations to an internal reference. This method has shown to provide accurate results when data is highly correlated which is the case of most MS imaging applications, since all the spectra within a tissue section share a lot of common features. We also increased the correlations between spectra by using a homogeneously distributed sputtered nano-layer of metal which is also used as an ionization matrix. The resulting aligned spectra share a common mass axis which allows a full MS image mass recalibration simply by calibrating the average spectrum. The peaks of the metal nanolayer were successfully used as internal mass references for calibration.

A sputtered silver-gold nano-layer was used as a novel methodology to validate the complete workflow. The peak matrix mass errors were verified using theoretical metal peak masses of the Ag and Au clusters. Furthermore, the MS data processing pipeline described uses an accurate data reduction strategy in which the peak matrix can represent the whole MS image with an MMA as low as 5 ppm in the m/z 400 to 1200 range. This enables further statistical analysis of MS peaks to be performed more accurately and efficiently.

References

[1] Jeremy L Norris and Richard M Caprioli. "Imaging mass spectrometry: A new tool for pathology in a molecular age". In: *PROTEOMICS - Clin. Appl.* 7.11-12 (Dec. 2013), pp. 733–738.

[2] Tiegang Li et al. "In situ biomarker discovery and label-free molecular histopathological diagnosis of lung cancer by ambient mass spectrometry imaging". In: *Sci. Rep.* 5.1 (Nov. 2015), p. 14089.

[3] Anna Nilsson et al. "Mass Spectrometry Imaging in Drug Development". In: *Anal. Chem.* 87.3 (2015), pp. 1437–1455.

[4] Cheng-Kang Chiang, Wen-Tsen Chen, and Huan-Tsung Chang. "Nanoparticle-based mass spectrometry for the analysis of biomolecules." In: *Chem. Soc. Rev.* 40.3 (Feb. 2011), pp. 1269–1281.

[5] Martin Dufresne et al. "Silver-Assisted Laser Desorption Ionization For High Spatial Resolution Imaging Mass Spectrometry of Olefins from Thin Tissue Sections". In: *Anal. Chem.* 85.6 (Mar. 2013), pp. 3318–3324.

[6] Martin Dufresne, Jean-François Masson, and Pierre Chaurand. "Sodium-Doped Gold-Assisted Laser Desorption Ionization for Enhanced Imaging Mass Spectrometry of Triacylglycerols from Thin Tissue Sections". In: *Anal. Chem.* 88.11 (June 2016), pp. 6018–6025.

[7] Taryn M Guinan et al. "Silver Coating for High-Mass-Accuracy Imaging Mass Spectrometry of Fingerprints on Nanostructured Silicon". In: *Anal. Chem.* 87.22 (Nov. 2015), pp. 11195–11202.

[8] Jeremy L. JL Norris et al. "Processing MALDI mass spectra to improve mass spectral direct tissue analysis". In: *Int. J. Mass Spectrom.* 260.2-3 (Feb. 2007), pp. 212–221.

[9] Maureen B Tracy et al. "Precision enhancement of MALDI-TOF MS using high resolution peak detection and label-free alignment." In: *Proteomics* 8.8 (Apr. 2008), pp. 1530–8.

[10] Justyna Sekuła et al. "Gold nanoparticle-enhanced target (AuNPET) as universal solution for laser desorption/ionization mass spectrometry analysis and imaging of low molecular weight compounds". In: *Anal. Chim. Acta* 875 (May 2015), pp. 61–72.

[11] Donald F. Smith et al. "Advanced Mass Calibration and Visualization for FT-ICR Mass Spectrometry Imaging". In: *J. Am. Soc. Mass Spectrom.* 23.11 (Nov. 2012), pp. 1865–1872.

[12] Jeremy A. Barry, Guillaume Robichaud, and David C. Muddiman. "Mass recalibration of FT-ICR mass spectrometry imaging data using the average frequency shift of ambient ions". In: *J. Am. Soc. Mass Spectrom.* 24.7 (July 2013), pp. 1137–1145.

[13] Purva Kulkarni et al. "Correcting mass shifts: A lock mass-free recalibration procedure for mass spectrometry imaging data". In: *Anal. Bioanal. Chem.* 407.25 (Oct. 2015), pp. 7603–7613.

[14] Pere Ràfols et al. "rMSI: an R package for MS imaging data handling and visualization". In: *Bioinformatics* 33.15 (Mar. 2017).

[15] Abraham. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." In: *Anal. Chem.* 36.8 (July 1964), pp. 1627–1639.

[16] Theodore Alexandrov et al. "Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering." In: *J. Proteome Res.* 9.12 (Dec. 2010), pp. 6535–46.

[17] Patrik Källback et al. "Novel mass spectrometry imaging software assisting labeled normalization and quantitation of drugs and neuropeptides directly in tissue sections." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 4941–51.

5.7 Appendix

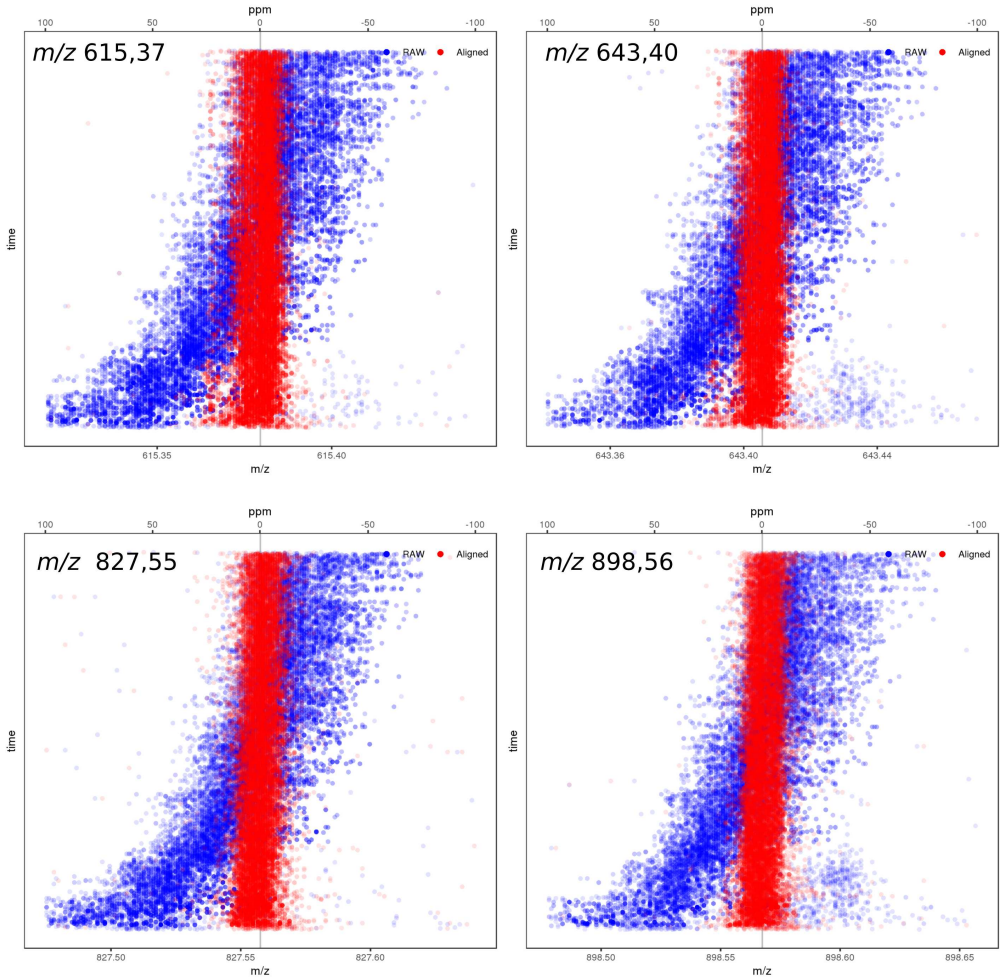


Figure 5.4: Mass shift observed at four endogenous peaks (m/z 615.37, m/z 643.40, m/z 827.55 and m/z 898.56) of a liver section before (blue) and after (red) the alignment routine. This data set contains more than 10000 pixels that are plotted here sorted according the MSI acquisition order (time). It can be seen how the mass drift introduced by a large acquisition time is compensated by the alignment algorithm.

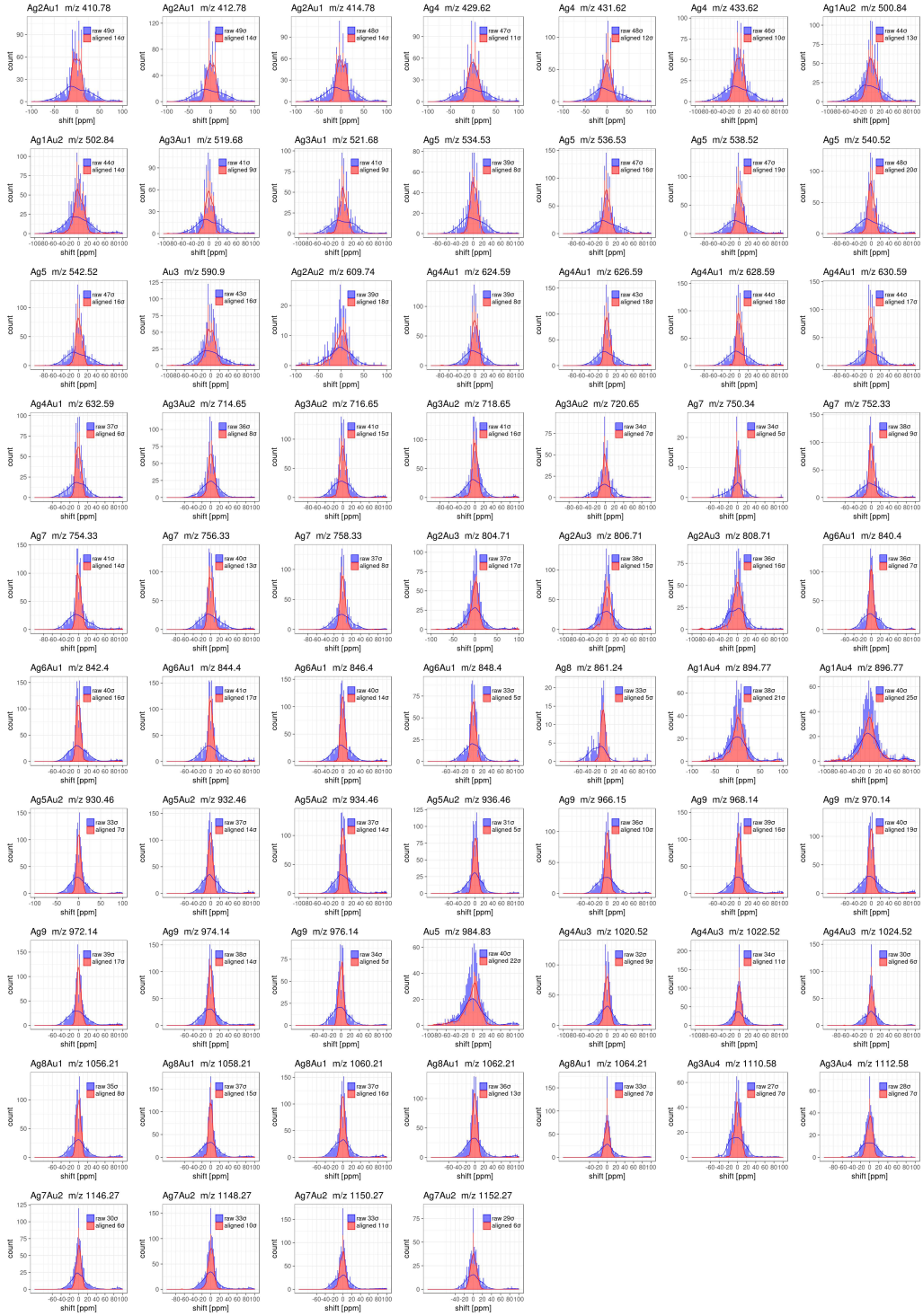


Figure 5.5: Mass shift histograms before (blue) and after (red) the alignment routine at Ag, Au and AgAu reference peaks.

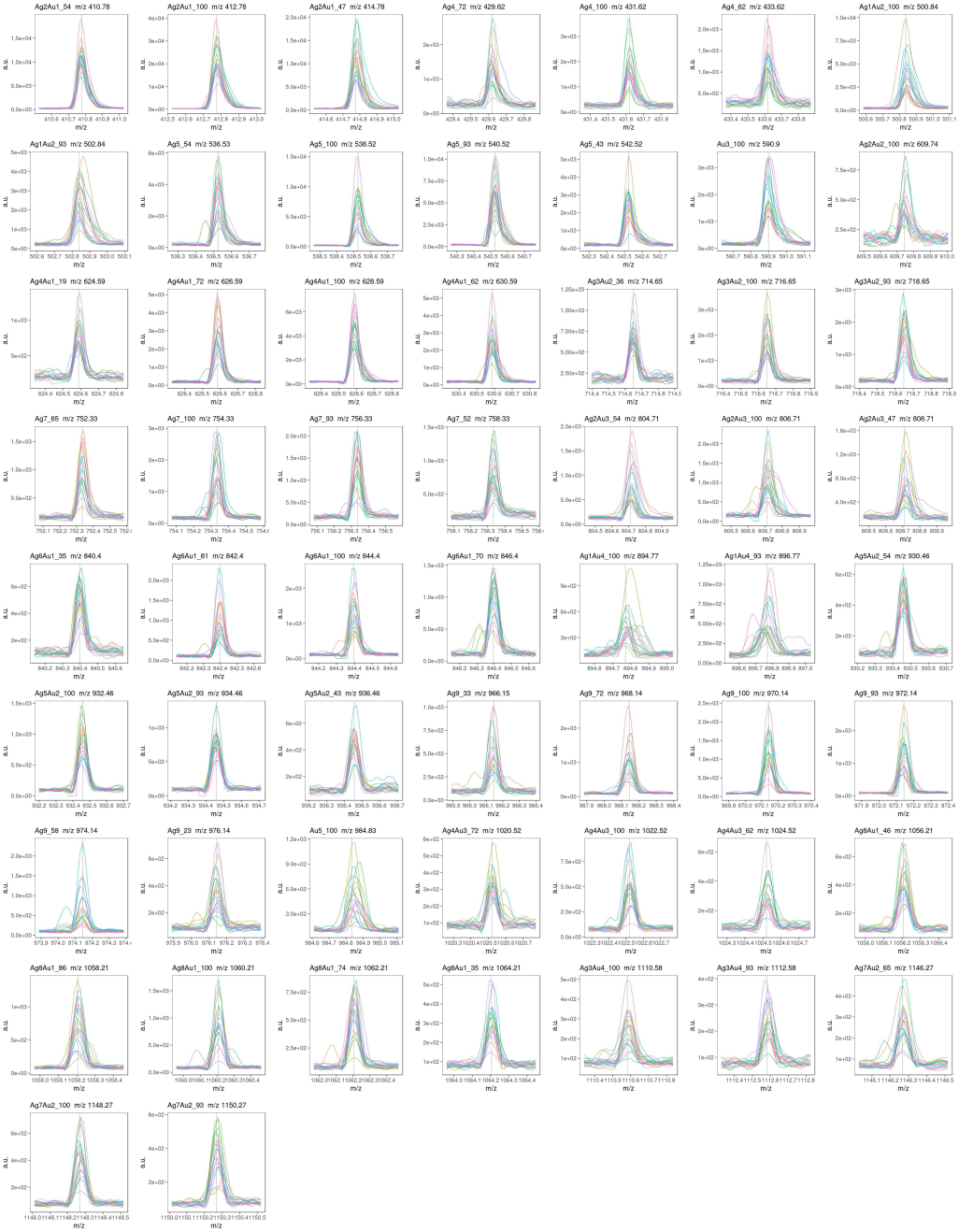


Figure 5.6: Spectra plots focused on reference peaks of 20 randomly selected calibrated pixels. Peaks Au_3, Ag_2Au_2, Ag_2Au_3, Ag_1Au_4 and Au_5 present a larger mass error because the reference peaks are overlapped with some endogenous compounds.

Metal peak ion formula	Relative isotope abundance	m/z_{calc} [a]	m/z_{exp} [b]	Error [ppm]
$[Ag_2Au_1]^+$	54%	410.7768	410.7736	7.7
$[Ag_2Au_1]^+$	100%	412.7764	412.7726	9.2
$[Ag_2Au_1]^+$	47%	414.7761	414.7768	-1.8
$[Ag_4]^+$	72%	429.6200	429.6182	4.3
$[Ag_4]^+$*	100%	431.6197	431.6169	6.5
$[Ag_4]^+$	62%	433.6194	433.6188	1.4
$[Ag_1Au_2]^+$	100%	500.8382	500.8362	4.1
$[Ag_1Au_2]^+$	93%	502.8379	502.8384	-1.1
$[Ag_5]^+$	54%	536.5251	536.5202	9.3
$[Ag_5]^+$*	100%	538.5248	538.5200	9.0
$[Ag_5]^+$	93%	540.5245	540.5210	6.4
$[Ag_5]^+$	43%	542.5241	542.5207	6.3
$[Au_3]^+$	100%	590.8997	590.8893	17.7
$[Ag_2Au_2]^+$	100%	609.7430	609.7504	-12.3
$[Ag_4Au_1]^+$	19%	624.5869	624.5892	-3.5
$[Ag_4Au_1]^+$	72%	626.5866	626.5832	5.5
$[Ag_4Au_1]^+$	100%	628.5863	628.5833	4.7
$[Ag_4Au_1]^+$	62%	630.5859	630.5834	4.0
$[Ag_3Au_2]^+$	36%	714.6484	714.6457	3.8
$[Ag_3Au_2]^+$	100%	716.6481	716.6464	2.3
$[Ag_3Au_2]^+$	93%	718.6477	718.6455	3.1
$[Ag_7]^+$	65%	752.3350	752.3319	4.1
$[Ag_7]^+$*	100%	754.3346	754.3301	6.0
$[Ag_7]^+$	93%	756.3343	756.3278	8.6
$[Ag_7]^+$	52%	758.3340	758.3336	0.5
$[Ag_2Au_3]^+$	54%	804.7099	804.7160	-7.6
$[Ag_2Au_3]^+$	100%	806.7095	806.7190	-11.7
$[Ag_2Au_3]^+$	47%	808.7092	808.7161	-8.6
$[Ag_6Au_1]^+$	35%	840.3968	840.3966	0.2
$[Ag_6Au_1]^+$	81%	842.3965	842.3938	3.2
$[Ag_6Au_1]^+$	100%	844.3961	844.3934	3.2
$[Ag_6Au_1]^+$	70%	846.3958	846.3954	0.4
$[Ag_1Au_4]^+$	100%	894.7713	894.7863	-16.7
$[Ag_1Au_4]^+$	93%	896.7710	896.7859	-16.6
$[Ag_5Au_2]^+$	54%	930.4583	930.4568	1.6
$[Ag_5Au_2]^+$	100%	932.4579	932.4572	0.8
$[Ag_5Au_2]^+$	93%	934.4576	934.4566	1.0
$[Ag_5Au_2]^+$	43%	936.4572	936.4582	-1.1
$[Ag_9]^+$	33%	966.1452	966.1436	1.7
$[Ag_9]^+$	72%	968.1448	968.1398	5.2
$[Ag_9]^+$*	100%	970.1445	970.1416	3.0
$[Ag_9]^+$	93%	972.1442	972.1423	1.9
$[Ag_9]^+$	58%	974.1438	974.1401	3.9
$[Ag_9]^+$	23%	976.1435	976.1402	3.3
$[Au_5]^+$	100%	984.8329	984.8674	-35.1
$[Ag_4Au_3]^+$	72%	1020.5197	1020.5173	2.4
$[Ag_4Au_3]^+$	100%	1022.5194	1022.5231	-3.6
$[Ag_4Au_3]^+$	62%	1024.5190	1024.5177	1.3
$[Ag_8Au_1]^+$	46%	1056.2066	1056.2066	0.0
$[Ag_8Au_1]^+$	86%	1058.2063	1058.2113	-4.7
$[Ag_8Au_1]^+$	100%	1060.2060	1060.2116	-5.3
$[Ag_8Au_1]^+$	74%	1062.2056	1062.2120	-6.0
$[Ag_8Au_1]^+$	35%	1064.2053	1064.2049	0.4
$[Ag_3Au_4]^+$	100%	1110.5812	1110.5833	-1.9
$[Ag_3Au_4]^+$	93%	1112.5808	1112.5819	-1.0
$[Ag_7Au_2]^+$	65%	1146.2681	1146.2736	-4.8
$[Ag_7Au_2]^+$	100%	1148.2678	1148.2664	1.2
$[Ag_7Au_2]^+$	93%	1150.2674	1150.2681	-0.6

Table 5.1: List of AgAu sputtered layer detected peaks used for MMA validation. [a] peak theoretical mass, [b] peak detected mass, * Peaks used as references for m/z calibration.

Chapter 6

rMSIproc: an R package that efficiently implements a complete pre-processing workflow for mass spectrometry imaging

6.1 Abstract

Mass spectrometry imaging (MSI) is a molecular histology technique that can reveal biochemical information directly from a tissue section. However, it generates a large quantity of complex spectral data. Translating the MSI raw data into relevant biochemical information is still a challenging task due to factors such as experimental variation and the huge size of MSI data. This requires implementing computationally efficient routines to process the raw MSI data. We present rMSIproc, an open-source R package that implements a full data pre-processing workflow for MSI experiments performed using TOF or FT-ICR spectrometers. The package provides an original strategy for spectral mass alignment and mass recalibration, with enhanced peak matrix mass measurement accuracy. rMSIproc is designed to work with files larger than the computer memory capacity and the algorithms are implemented using a multi-threading strategy.

6.2 Introduction

Mass spectrometry imaging (MSI) is an emerging technique capable of mapping the spatial distributions of molecular ions in biological tissues [1]. The size and complexity of MSI data requires specialized software to extract relevant information. Several software packages have been released to address these demands, which have been recently reviewed describing the most challenging processes in MSI data analysis [2]. More recently, a new release of MSiReader has been published [3]. MSiReader is a Matlab written package that is gaining popularity, as it features intuitive graphical tools, and is freely available online. MSiReader is predominantly used for the visualization of MSI data; however it only includes a few algorithms for MS spectral processing. The major limitation of MSIreader is that the full dataset is loaded into computer memory, which impedes processing a dataset larger than the available memory. SpectralAnalysis [4] is another software package written in Matlab that overcomes the memory limitations and provides common pre-processing algorithms: smoothing of mass spectra, baseline correction, intensity normalization, and peak detection. In comparison to these Matlab-based software solutions, the R platform is a truly open alternative that allows a straightforward modification and combination of different tools. Indeed, the last version of the R-based Cardinal package [5] has improved the data model in order to handle larger-than-memory datasets through the 'matter' package [6].

Nevertheless, Cardinal does not provide a graphical user interface (GUI) to explore the MSI data. None of these software tools have exploited the full potential of multicore processors by writing multithread-ready code.

Here we present rMSIproc, an open-source MSI data pre-processing package developed in R, to complement the previously released rMSI package [7]. The rMSI package was designed to allow an efficient access to large MSI datasets combined with a data visualization GUI. rMSIproc takes advantage of the rMSI data handling strategy and adds a full data pre-processing workflow designed to extract relevant mass-to-charge (m/z) features from large datasets. The entire data processing is implemented using a multi-thread approach that takes advantage of modern multicore processors.

6.3 rMSIproc features

The main goal of the rMSIproc package is to produce a peak matrix that is a reduced and robust representation of the complete MSI dataset, being small enough to fit within the computer's memory. Therefore, this format enables all the available R statistical analysis packages to be used for MSI data analysis. rMSIproc can also record and store the mass spectra pre-processing results in an rMSI formatted file. This allows the rMSI visualization tools to use the pre-processed data to enhance ion image reconstruction. The pre-processing package includes algorithms to perform the following: Savitzky-Golay smoothing [8], spectral alignment, m/z recalibration, intensity normalization, peak detection and peak binning. A diagram of the pre-processing workflow is provided in the appendix Fig. 6.1. The mass spectra alignment tool uses a novel algorithm that can compensate for instrument-induced mass shifts in a fully automated way without using any known molecule as internal reference. The peak detection method is optimized for a fast mass peak centroid prediction. Both the alignment- and the peak detection algorithms rely on the fast Fourier transform (FFT) to calculate interpolations and cross-correlations efficiently. After the peak detection, a binning tolerance is used to merge all peaks from all spectra in a binned peak matrix. The peak matrix stores each spectrum in a row and retains the different m/z species in the columns. The peak matrix follows the standard R language conventions, so it can be directly used for statistical data analysis, as each row is understood to be an observation and each column as a variable. All the pre-processing parameters are integrated

in a GUI for easy operation (Fig. 6.2 in the appendix). However, the user can still integrate rMSIproc in any R script by calling the required functions of rMSIproc as a standard R package.

6.4 Implementation details

rMSIproc uses rMSI to efficiently handle MSI data. Therefore, the same data formats as rMSI are supported. This includes the open-standard format imzML [9] in both 'continuous' and 'processed' modes. Data is loaded using rMSI functions and then pre-processed by accessing rMSI objects directly from inside the rMSIproc methods. The internals of rMSIproc are mainly implemented in C++ to provide efficient memory management and highly optimized multi-threading execution. However, all the user-relevant methods are exposed as R functions following the classical structure of an R package.

6.5 Results

Several MSI datasets up to 200 gigabytes in size have been successfully processed using rMSIproc, including mass spectra smoothing and alignment, m/z calibration, normalization of mass spectra intensities, peak detection and binning. In all cases, we obtained a balanced CPU load distributed across all processing cores on a machine with four available CPUs. The memory consumption was managed by exclusively loading the data chunk being processed at each given time. The performance of rMSIproc is reported in the appendix Table 6.1. rMSIproc can also merge and process various datasets simultaneously, producing a single m/z matrix from all the aligned mass spectra. The alignment algorithm has proven to compensate for mass shifts in both TOF and FTICR datasets. After the alignment, all mass spectra share a common mass axis and can be re-calibrated together. Our alignment routine can properly resolve isobaric m/z species in ultra-high mass resolution MALDI-FTICR datasets that are otherwise impossible to detect accurately (an example is provided in the appendix Fig. 6.3).

6.6 Conclusions

rMSIproc is a valuable tool for pre-processing MSI files containing both high mass and high spatial resolution MSI datasets in R environment. It features two novel algorithms for mass spectral alignment and fast peak-detection. The combination of rMSI and rMSIproc provides a full MSI data visualization and pre-processing platform that uses modern computer architectures in a novel and open-source manner.

References

[1] Jeremy L Norris and Richard M Caprioli. "Imaging mass spectrometry: A new tool for pathology in a molecular age". In: *PROTEOMICS - Clin. Appl.* 7.11-12 (Dec. 2013), pp. 733–738.

[2] Pere Ràfols et al. *Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications*. Nov. 2016.

[3] Mark T. Bokhart et al. "MSiReader v1.0: Evolving Open-Source Mass Spectrometry Imaging Software for Targeted and Untargeted Analyses". In: *J. Am. Soc. Mass Spectrom.* (Sept. 2017), pp. 1–9.

[4] Alan M. Race et al. "SpectralAnalysis: Software for the Masses". In: *Anal. Chem.* 88.19 (Oct. 2016), pp. 9451–9458.

[5] Kyle D Bemis et al. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". In: *Bioinformatics* 31.14 (July 2015), pp. 2418–2420.

[6] Kylie A. Bemis and Olga Vitek. "matter: an R package for rapid prototyping with larger-than-memory datasets on disk". In: *Bioinformatics* 33.19 (Oct. 2017), pp. 3142–3144.

[7] Pere Ràfols et al. "rMSI: an R package for MS imaging data handling and visualization". In: *Bioinformatics* 33.15 (Mar. 2017).

[8] Abraham. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." In: *Anal. Chem.* 36.8 (July 1964), pp. 1627–1639.

[9] Thorsten Schramm et al. "imzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5106–10.

6.7 Appendix

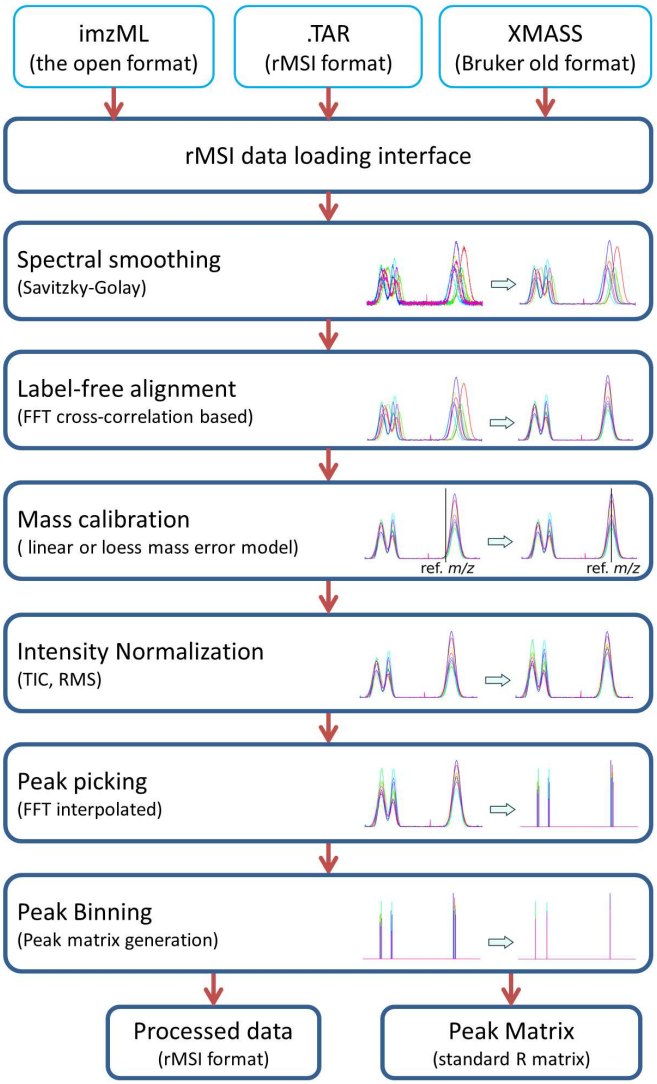


Figure 6.1: rMSIproc processing workflow schematic.

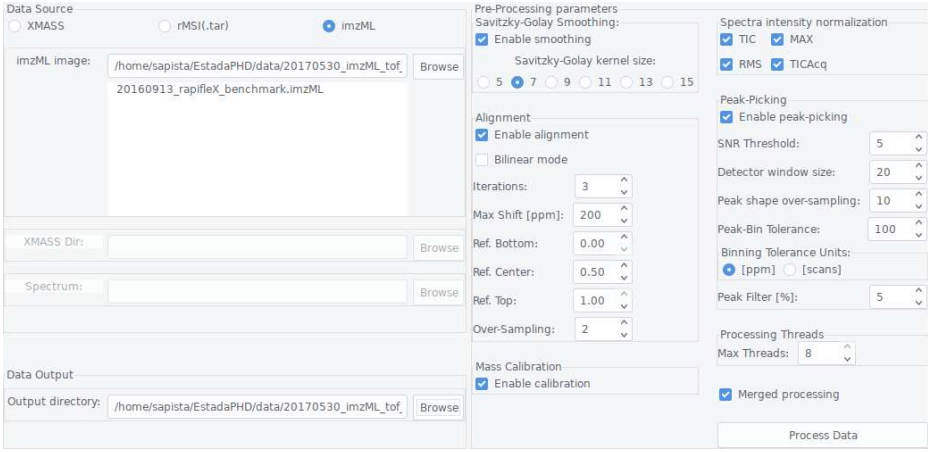


Figure 6.2: Screenshot of the rMSIproc's GUI used to easily configure the pre-processing settings. All processing parameters are available through the GUI for user-friendly interaction. The GUI is launched by issuing the command rMSIproc::ProcessWizard() on an R console.

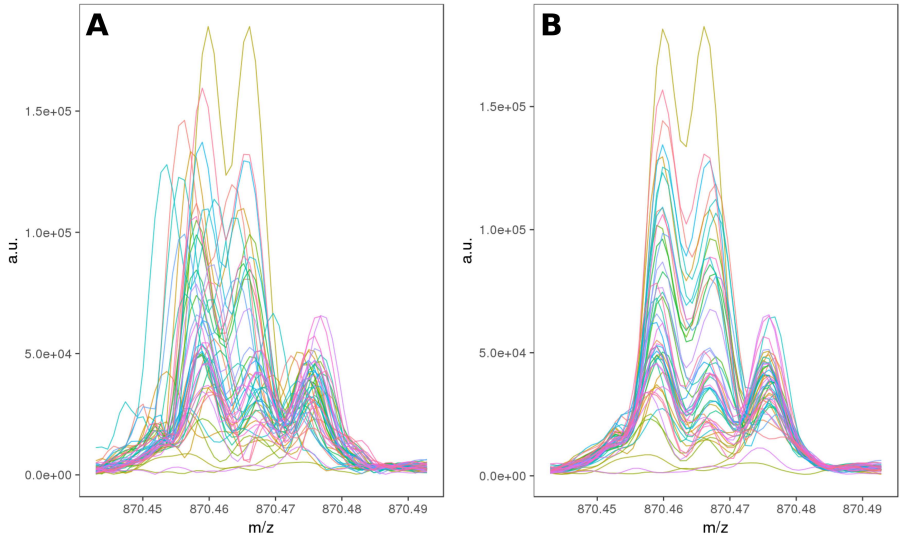


Figure 6.3: Example of the performance of the spectral alignment algorithm on a dataset acquired using an FTICR spectrometer. This plot was constructed by randomly selecting fifty pixels in the dataset. Then, a small mass range with three different ion species is represented. The spectrum corresponding to each pixel is plotted using a randomly chosen color. It can be easily observed the performance of the alignment algorithm by comparing the spectra of these fifty pixels before **(A)** and after **(B)** the alignment stage. After the alignment algorithm **(B)**, the three peaks can be properly resolved in this mass range. Without alignment **(A)** theses three peaks will appear mixed if a single peak matrix is constructed from the whole dataset.

Detector	Data size [GB]	Number of pixels	Points per spectrum	Number of threads	Processing time [minutes]
TOF	10,81	73620	19709	1	83
TOF	10,81	73620	19709	2	34
TOF	10,81	73620	19709	4	24
FTICR	32,11	3255	1324231	1	743
FTICR	32,11	3255	1324231	2	269
FTICR	32,11	3255	1324231	4	196

Table 6.1: Processing performance of rMSIproc using an AMD Opteron four-core computer at 3 GHz. We processed two MSI datasets of different sizes and recorded the time required in each case. The complete rMSIproc workflow includes mass spectra smoothing and alignment, m/z calibration, intensity normalization, peak picking and peak binning. The number of CPUs used was controlled by setting the maximum number of threads parameter of rMSIproc.

Chapter 7

Final discussion

The work carried out in this thesis can be clearly separated in two parts. Firstly, the experimental workflow set up for sample preparation and LDI-MS image acquisition. Secondly, the signal processing and bioinformatics part. This work is focused in developing a reliable and simple workflow to obtain high quality metabolomics images. We believe that these goals have been successfully accomplished since a gold sputtering deposition workflow has been optimized and is used routinely in our lab. The optimized gold deposition methodology allowed us to acquire the metabolomics imaging datasets which have been used to develop the pre-processing steps and the bioinformatics tools for MSI. At this point, we are able to accomplish the necessary processing steps to transform the complex MSI raw data into a simplified peak matrix. Therefore, the objective of developing a MSI processing platform capable of working with high-dimensionality data has been accomplished. This will facilitate the task of making sense from the MSI data in future applications.

7.1 Research on new methods for spatial metabolomics

The work conducted in this thesis is focused on the development of LDI technologies to obtain spatial metabolomics information. The other two MSI ionization techniques: SIMS and DESI, have not been explored here. The reasons why we focused in LDI are two-fold: Firstly, we had the opportunity to easily use a modern MALDI instrument (Bruker ultraFleXterme) installed in our lab plus the expertise of our lab technicians in classic MALDI experiments. Secondly, the MALDI spectrometer is nowadays the most used platform for MSI. Generally speaking, SIMS instruments are mostly limited in terms of mass accuracy at high spatial resolution and MS/MS capability. In case of DESI, it has great potential for measurements under ambient conditions, but has poor spatial resolution compared to MALDI and SIMS [1].

MALDI based MSI has been successfully applied in peptidomics proteomics. However, the adoption of MSI is still not a generalized technique for metabolomics studies due to the challenge of obtaining high quality low molecular weight spatially resolved spectral data. The main drawback of traditional MALDI imaging technique is related to the organic matrix MS peaks interfering with the low mass range. In this work, it has been demonstrated the reliability of a sputtered gold nano-layer to promote the ionization of low molecular weight compounds. Besides,

the use of sputtered metals like gold and silver has been reported by Dufresne et al. in MSI applications [2, 3]. This opens up the door to explore the benefits of depositing other metals using sputtering technique. In our opinion, it is preferred to test metals with only one stable isotope, because metals tend to form clusters during the LDI process. These metal clusters are detected as MS peaks that follow the isotopic pattern of the material. Therefore, these MS peaks might interfere severely with the peaks of the endogenous compounds.

There are other techniques with the potential of outperforming organic matrices for spatial metabolomics. Nanostructure-initiator mass spectrometry (NIMS) was introduced as a novel alternative for MSI, in where a silicon porous structure is used to trap the molecules of the initiator material to promote the ionization of metabolites [4]. However, it is necessary to cut the tissue in very thin sections (ca. 5 μm) to allow the laser to reach the active surface. This hampers the adoption of NIMS for MSI because it is difficult in practice to prepare these thin tissue sections and placing them onto the substrate. Moreover, it has been demonstrated that with this sample configuration the LASER system included in MALDI spectrometers like UltrafleXtream are not able to ionize the tissue surface compounds. More recently it has been demonstrated that it is possible to obtain metabolomics images by imprinting the tissue over a surface. For example, a gold substrate has been used to obtain low molecular weight images of a fingerprint [5]. This suggests that a nano-structured surface would be able to promote an efficient LDI process for imprinted tissues. Moreover, the surface could be functionalized to improve the detection of specific spices. For example, a thin metal layer could be deposited over the surface using sputtering in order to optimize the ionization of some compounds. The development of such nanostructured surfaces has established an emerging research field for spatial metabolomics.

On the other hand, spatial metabolomics datasets can also be acquired using organic matrices. However, to perform an accurate data analysis, the MS peaks belonging to the organic matrix must be discarded from the tissue endogenous MS peaks. This is a difficult task since the low mass region is very crowded with matrix signals and the detection of these interfering signals may need some advanced processing techniques [6, 7]. Even with the appropriate processing strategies to filter matrix signals, the interference of the strong matrix peaks below m/z 700 is so strong that hampers the detection of many metabolites [8]. Therefore, in our opinion, the final solution to spatial metabolomics will likely be provided by

advanced nano-material technologies in contraposition to organic matrices.

7.2 The challenges of MSI data processing

MSI instrumental platforms have evolved rapidly in last year allowing the acquisition of MSI datasets with higher resolutions, both lateral and mass, with less time. Nevertheless, the development of software tools for MSI has not been able to follow this trend. Many new features have been added to MSI software's and many new packages have been created, but the challenge of obtaining a reliable MSI data processing platform to extract biologically relevant information rapidly is still a work in progress. Currently, the most advanced proprietary software package for MSI is probably SCiLS (http://scils.de) which provides a powerful graphical user interface (GUI) with many data analysis and visualization tools. However its closed-source development model obstructs the full comprehension of how data is being processed and its adaptation to specific needs. In addition, the high license cost hinders its usage in the research area. In contraposition, open-source developed software provides a grade of flexibility that facilitates the implementation of custom MSI processing strategies. Moreover, the open-source software is usually free of cost so it can be easily adopted by every research group with independence of the budget. In the last years, the MSI community interest in a fully open solution has led to the open data format: imzML [9]. This open format has enabled the data exchange between different instrumental platforms and software tools which is crucial for the progress of MSI in particular, and science in general. Currently, the major MSI instrument manufacturers provide the option to export the data to imzML format. This empowers the open-source alternatives to fully develop their potential to take MSI technologies further.

Once MSI data has been converted to imzML format, the next step is to process the spectral information with the purpose of carrying out a statistical data analysis or just representing some ion images. The big size of MSI data demands specific software tools to manage the spectral information efficiently. The common approach in R programming consists in loading all the dataset in computer's RAM memory. However, this workflow is not viable for MSI. In this work, the package rMSI has been developed to overcome the "larger than memory" data limitation of the R platform. This allowed us to continue developing more advanced tools to process and visualize MSI data inside the R environment without having to

worry about memory constrains. In fact, rMSI was the first released package that overcomes such memory limitations for MSI in R. Recently, the well-known MSI focused R package Cardinal [10, 11] has followed our trend and evolved to provide support for larger than memory MSI datasets.

Besides the MSI data processing strategies and algorithms it is also important that an MSI software package provides a good usability. This is often understood as providing an easy user interaction through a polished GUI. Therefore, rMSI comes with a GUI designed to explore MSI data interactively. The developed GUI is able to construct ion images loading only the essential necessary part of MS data in computer's memory. Thanks to this low memory footprint, a dual view mode is possible and has been implemented. This visualization mode allows loading two MSI datasets side by side sharing the spectra viewer. This enables the fast manual comparison of two tissues in, for example, a healthy versus diseased tissue. The GUI contains many features also available in other MSI software packages but the novelty resides in the fact that the GUI is completely integrated in R. Thus, all the MSI data is shared with the running R session. This allows the execution of some R scripts over the currently loaded data and the immediate visualizations of the results through the GUI.

The first stage of MSI data processing is known as pre-processing and is based on reducing the experimental variability by processing each spectrum in the dataset. The common MS pre-processing workflow includes algorithms like spectral smoothing, baseline correction, intensity normalization and peak detection [12]. In our opinion, the peak detection step is a very simple and robust approach to reduce the data dimensionality without losing valuable information. We consider that the first stage of MSI data processing has been completely resolved in this work. The developed package rMSIproc provides all the necessary routines to convert a huge amount of raw MS data into a reduced peak matrix preserving the relevant information. The generated peak matrix is small enough to fit in the computer's memory and this enables the easy and efficient use of third parties developed algorithms for MSI applications. This will foster the research of MSI data analysis methods since the developed approach overcomes the memory usage concerns. Moreover, rMSIproc arranges the output data in a matrix-like object which is the common data structure for most statistical analysis methods.

We believe that this approach is the most efficient because it allows an accurate and reduced representation of the complete MSI dataset. Nevertheless, it is im-

portant to highlight the importance of the developed mass alignment algorithm. Without an alignment stage it is not possible to reduce the peak binning tolerance enough to properly separate MS peaks that are slightly overlapping in pixel to pixel spectrum. The benefits of our alignment strategy have been also observed in FTICR data, where some isobaric spices were properly resolved in the resulting peak matrix.

The current version of rMSIproc produces a peak matrix where all peaks with a signal to noise ratio (SNR) over a user defined threshold are retained. This means that all peaks corresponding to isotopes or adducts of the molecules are kept. This introduces redundant information in the peak matrix that may hamper the subsequent statistical analysis. The annotation of those peaks according to its possible common origin would be a valuable feature to add to rMSIproc. The MSI spatial information can be used to calculate image correlations between peaks that match the mass rules to be annotated as the same molecule [13]. This will provide a more robust annotating algorithm especially in the case of FTICR data where all isotopic distributions are resolved. The future perspective of rMSIproc development is to implement such annotation strategies but without modifying the peak matrix. The goal is to add a secondary data structure together with the peak matrix that provides all peak annotations. Besides, some R functions should be written allowing the creation of a new peak matrix using the original peak matrix and the desired annotations. The new peak matrix could be the input data for the subsequent statistical analysis algorithms.

Currently, the outputs of rMSIproc are the processed spectral data and the peak matrix, both in a custom data format designed to be efficient inside an R session. However, this custom data format hinders the integration of rMSIproc with other non-R based software tools. In order to facilitate the data exchange with third party tools it is necessary to implement imzML format exportation in future version of rMSIproc. The goal is to be able to write the MSI spectral data in a continuous imzML file and the peak matrix in the processed imzML format. This will enable other software tools to take advantage of the processing algorithms available in rMSIproc.

One of the biggest strengths of rMSIproc is related to its processing parallelization design. To the date and to our knowledge, rMSIproc is the only open-source MSI software package implemented using a multithreading approach. Moreover, the parallelization has been implemented using a structured C++ class style where

all multithreading synchronization mechanisms have been encapsulated inside an abstraction layer. This allows an easy addition of new algorithms to rMSIproc without having to concern anymore for the multithreaded implementations. This strategy enables rMSIproc to take advantage of modern multicore CPU's to drastically reduce the processing time required to generate the final peak matrix and the corrected MS spectra.

With rMSIproc we can process huge MSI datasets in a routinely basis. This allows us to obtain the peak matrix of an MSI experiment rapidly in order to perform the statistical analysis of the data. Therefore, the work to be developed in the future will be focused on the research of unsupervised segmentation algorithms for MSI. This should enable our research group to make untargeted metabolomics studies directly in the tissue sections and to differentiate tissue morphologies according to metabolomics criteria.

7.3 The future of histopathology and MSI

Spatial metabolomics is still an emerging field from both sides: the experimental workflows and the bioinformatics approaches. But the advances on this discipline will certainly integrate in histopathology to better understand chemical processes inside the tissues. MSI is a potentially extraordinary tool for pathological analysis and the investigation of disease mechanisms because, it provides the ability to image multiple molecules simultaneously with high sensitivity [1]. The MSI capability of determining the spatial localization of molecules has revolutionized our approach to diseases by allowing us to directly examine the pathological process. However, there is a tradeoff between spatial resolution and sensitivity making impossible for MSI to achieve a similar lateral resolution to the classical histology. This means that MSI will probably never be able to replace classic histology. In the future, it is more likely to have mass spectrometrists working together with histopathologists to provide an accurate diagnostic. The images generated using two different technologies like optical microscope and MSI should be interpreted together in order to obtain the fine-grained texture of microscopy combined with the rich chemical specificity of MSI. Image fusion is a concept that further explores the ability to combine images from different techniques through bioinformatics tools. It has been demonstrated that using image fusion approaches it is possible to predict ion distributions with a finer detail by combining MSI data with an

optical microscope image of the same tissue [14]. These strategies will improve the integration of MSI in pathological workflows since better spatially resolved chemical spices will simplify the tissue morphology interpretation.

MSI technologies are the link between spatial metabolomics and histopathology. MSI must provide confident molecular assignments in order to establish a solid relationship between both worlds. The efficacy of on-tissue molecular identifications is strongly related with the instrumental mass resolution and mass accuracy. Independently of the used spectrometer, it is possible to enhance the mass accuracy through the clever application of spectral pre-processing strategies. In this thesis, we developed a label-free alignment routine that allows minimizing the mass shift between pixel-to-pixel spectra. After the alignment stage all spectra shares a common mass axis so we can then re-calibrate the complete dataset. We have also successfully used the metal clusters peaks from a sputtered layer as internal mass references to increase the mass accuracy. Using a bilayer of silver and gold, we have proved that it is possible to achieve a mass accuracy down to 5 ppm with a TOF detector after applying the processing workflow developed. Nevertheless, this high mass accuracy is not viable in practice because the complexity of real tissue samples produce many overlapping peaks originated from different endogenous compounds. A spectrometer with higher mass resolution like FTICR must be used to address this problem. A higher mass resolution will allow selecting a smaller bin size for ion image reconstruction. Hence, molecular signatures could be assigned to spatial ion distributions more confidently [15]. In contrast, the extra time needed to acquire a high lateral resolution FTICR MSI dataset and the huge amount of data generated may hampers the practical application of ultra-high mass resolution to spatial metabolomics. In conclusion, the future of molecular histopathology is probably a multimodal approach where traditional histological stains, high lateral resolution TOF-MSI and high mass resolution FTICR-MSI will be combined to completely understand the diseases represented by the biochemical structures in the tissue.

References

[1] Andreas Römpp et al. "Histology by Mass Spectrometry: Label-Free Tissue Characterization Obtained from High-Accuracy Bioanalytical Imaging". In: *Angew. Chemie Int. Ed.* 49.22 (Apr. 2010), pp. 3834–3838.

[2] Martin Dufresne et al. "Silver-Assisted Laser Desorption Ionization For High Spatial Resolution Imaging Mass Spectrometry of Olefins from Thin Tissue Sections". In: *Anal. Chem.* 85.6 (Mar. 2013), pp. 3318–3324.

[3] Martin Dufresne, Jean-François Masson, and Pierre Chaurand. "Sodium-Doped Gold-Assisted Laser Desorption Ionization for Enhanced Imaging Mass Spectrometry of Triacylglycerols from Thin Tissue Sections". In: *Anal. Chem.* 88.11 (June 2016), pp. 6018–6025.

[4] Raul Calavia et al. "Nanostructure Initiator Mass Spectrometry for tissue imaging in metabolomics: future prospects and perspectives." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5061–8.

[5] Justyna Sekuła et al. "Gold nanoparticle-enhanced target (AuNPET) as universal solution for laser desorption/ionization mass spectrometry analysis and imaging of low molecular weight compounds". In: *Anal. Chim. Acta* 875 (May 2015), pp. 61–72.

[6] Gregor McCombie et al. "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis." In: *Anal. Chem.* 77.19 (Oct. 2005), pp. 6118–6124.

[7] Judith M. Fonville et al. "Robust Data Processing and Normalization Strategy for MALDI Mass Spectrometric Imaging". In: *Anal. Chem.* 84.3 (Feb. 2012), pp. 1310–9.

[8] Daisuke Miura, Yoshinori Fujimura, and Hiroyuki Wariishi. "In situ metabolomic mass spectrometry imaging: recent advances and difficulties." In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5052–60.

[9] Thorsten Schramm et al. "imzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data". In: *J. Proteomics* 75.16 (Aug. 2012), pp. 5106–10.

[10] Kyle D Bemis et al. "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments". In: *Bioinformatics* 31.14 (July 2015), pp. 2418–2420.

[11] Kylie A. Bemis and Olga Vitek. "matter: an R package for rapid prototyping with larger-than-memory datasets on disk". In: *Bioinformatics* 33.19 (Oct. 2017), pp. 3142–3144.

[12] Jeremy L. JL Norris et al. "Processing MALDI mass spectra to improve mass spectral direct tissue analysis". In: *Int. J. Mass Spectrom.* 260.2-3 (Feb. 2007), pp. 212–221.

[13] Andrew Palmer et al. "FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry". In: *Nat. Methods* 14.1 (Nov. 2016), pp. 57–60.

[14] Raf Van de Plas et al. "Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping". In: *Nat. Methods* 12.4 (Feb. 2015), pp. 366–372.

[15] Andreas Römpp and Bernhard Spengler. "Mass spectrometry imaging with high resolution in mass and space". In: *Histochem. Cell Biol.* 139.6 (June 2013), pp. 759–783.

List of Figures

List of Tables

Software source code

The source code of the software developed during this thesis can be found freely available and under the terms of general public license (GPL) agreement at:

https://github.com/prafols/rMSI

https://github.com/prafols/rMSIproc

Fundings

El treball desenvolupat en aquesta tesis ha estat possible gràcies a les fonts de finançament següents: Al Ministeri d'Economia i Competitivitat de l'Estat Espanyol amb els projectes TEC2012-31074 i TEC2015-69076-P així com l'ajuda que m'ha permès realitzar aquest doctorat: BES-2013-065572. Al suport de la Direcció General de Recerca de la Generalitat de Catalunya amb el projecte 2014 SGR 01267.