

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Computational tools for analysis of mass spectrometry imaging data

Permalink

<https://escholarship.org/uc/item/4xz0n52n>

Author

Bruand, Jocelyne

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational Tools for Analysis of Mass Spectrometry Imaging Data

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jocelyne Bruand

Committee in charge:

Professor Vineet Bafna, Chair
Professor Eduardo Macagno, Co-Chair
Professor Pieter C. Dorrestein
Professor Terry Gaasterland
Professor Pavel A. Pevzner
Professor Glenn Tesler

2012

Copyright
Jocelyne Bruand, 2012
All rights reserved.

The dissertation of Jocelyne Bruand is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2012

DEDICATION

To my family, for their love, inspiration and support.

EPIGRAPH

In theory, there is no difference between theory and practice.

But in practice, there is.

— Yogi Berra

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Acknowledgements	xii
Vita	xiv
Abstract of the Dissertation	xv
Chapter 1	Introduction	1
	1.1 What is Mass Spectrometry Imaging?	2
	1.2 Potentials of Mass Spectrometry Imaging	2
	1.3 Shortcomings of Mass Spectrometry Imaging	4
Chapter 2	Automated querying and identification of novel peptides using MALDI mass spectrometric imaging	6
	2.1 Introduction	7
	2.2 Methods	11
	2.2.1 MALDI Imaging Data Acquisition	11
	2.2.2 MALDI Imaging Data Normalization	13
	2.2.3 Query Definition	14
	2.2.4 Query Shift	14
	2.2.5 Testing for Localization of Protein Expression	15
	2.2.6 Simulations	16
	2.2.7 MS/MS Sample Preparation and Data Acquisition	17
	2.2.8 MS/MS Identification	18
	2.2.9 In situ hybridization	19
	2.3 Results and Discussion	19
	2.3.1 An overview of the pipeline	19
	2.3.2 Data Normalization	21
	2.3.3 Defining Regions of Interest	23
	2.3.4 Simulations	24
	2.3.5 Over-expressed molecules in the leech	25

	2.3.6 Peptide identification	31
	2.4 Conclusions	33
	2.5 Acknowledgements	35
Chapter 3	AMASS: Algorithm for MSI Analysis by Semi-supervised Segmentation	36
	3.1 Introduction	37
	3.2 Results	41
	3.2.1 AMASS: Algorithm for MSI Analysis by Semi-supervised Segmentation	41
	3.2.2 Initial Segmentation	42
	3.2.3 Querying	43
	3.2.4 Spot Partitioning	47
	3.3 Discussion	54
	3.4 Methods	55
	3.4.1 Data Acquisition	55
	3.4.2 Query	57
	3.4.3 Spot Partitioning	58
	3.4.4 Query consistency	59
	3.5 Acknowledgement	59
Chapter 4	Comparative Analysis of Mass Spectrometry Imaging Data	61
	4.1 Introduction	61
	4.2 Methods	62
	4.2.1 Data Acquisition and Preprocessing	62
	4.2.2 Defining the Query	64
	4.2.3 Obtaining the Molecular Signatures	64
	4.2.4 Filtering the datasets	64
	4.2.5 Clustering the datasets	65
	4.3 Results	69
	4.3.1 Surfactins and <i>Bacillus subtilis</i>	69
	4.3.2 SapB and <i>Streptomyces coelicolor</i>	72
	4.4 Discussion	73
	4.5 Acknowledgements	74
Chapter 5	On-Tissue Peptide Identification using Spectral Libraries	75
	5.1 Introduction	75
	5.2 Methods	76
	5.2.1 Data Acquisition	76
	5.2.2 Spectral Library Creation	76
	5.2.3 Spectral Library Search	78
	5.3 Results	78
	5.3.1 AMASS Run	78

5.3.2	Peptide Identifications	80
5.3.3	Spectral Library Matches with Good Peptide Identification	80
5.3.4	Other Spectral Library Matches	82
5.4	Discussion	83
Chapter 6	Conclusions	85
Appendix A	Supplemental: Automated querying and identification of novel peptides using MALDI mass spectrometric imaging	88
Appendix B	Supplemental: AMASS – Algorithm for MSI Analysis by Semi-supervised Segmentation	105
Bibliography	120

LIST OF FIGURES

Figure 1.1:	Mass Spectrometry Imaging (MSI) overview	3
Figure 2.1:	Overall process for detecting and identifying masses specifically expressed within an ROI or specific morphological feature	12
Figure 2.2:	Results of the <i>rho</i> -statistical test for the CNS query in the leech	22
Figure 2.3:	Mask and top results for LEECHE12A with other ROIs	26
Figure 2.4:	Annotated MS/MS spectrum for HmIF4, a novel peptide in the family of glial intermediate filament	29
Figure 2.5:	Annotated MS/MS spectrum for novel peptide with parent mass $\simeq 3666$ Da, targeted for being specifically expressed in the lateral/dorsal region	30
Figure 3.1:	Main workflow overview	41
Figure 3.2:	List of queries and their associated results	44
Figure 3.3:	Log-odds score matrix and hierarchical clustering	48
Figure 3.4:	Binary spot signatures and leech segmentation maps	50
Figure 3.5:	Results for the rat brain slice dataset	52
Figure 4.1:	Overview of the comparative analysis workflow	63
Figure 4.2:	Clustering results with query $m/z = 1045$ and $m/z = 1075$ (surfactins)	66
Figure 4.3:	<i>Bacillus subtilis</i> cluster for query $m/z = 1075$	67
Figure 4.4:	<i>Streptomyces coelicolor</i> cluster for query $m/z = 2026$	70
Figure 5.1:	Database model for spectral library	77
Figure 5.2:	AMASS results for MSI data	78
Figure 5.3:	Spectral library hit for peptide from myelin basic protein	79
Figure 5.4:	Spectral library hit for peptide from beta-globin	81
Figure 5.5:	Other Spectral Library Matches	82
Figure A.1:	Effects of normalization on data	89
Figure A.2:	Mask and query for the CNS in the LEECHE12A	91
Figure A.3:	MALDI images for CNS localization at different scores	91
Figure A.4:	Simulation results decreasing ROI signal over the entire region	93
Figure A.5:	Simulation results when degrading the signal in the ROI.	94
Figure A.6:	ClustalW alignment of the HmIF4 protein sequence with those of three other known intermediate filaments in <i>Hirudo medicinalis</i>	95
Figure A.7:	Annotated spectrum for a peptide from the histone H2B	96
Figure A.8:	Annotated spectrum for uncharacterized peptide	97
Figure A.9:	Distribution of the MS1 raw data peaks for 2 experiments	97
Figure B.1:	Querying with small random seeds	106

Figure B.2:	Segmentation results with and without smoothing after 10 iterations for random initial random segmentation in leech	109
Figure B.3:	Top 20 score peaks at least 10 Daltons apart for segments in leech at successive iterations.	110
Figure B.4:	Top 20 score peaks at least 10 Daltons apart for segments in rat at successive iterations.	111
Figure B.5:	Query consistency scores	118
Figure B.6:	Basic anatomy for the a) the leech embryo and b) the rat brain slice.	119

LIST OF TABLES

Table A.1:	Over-expressed masses for the CNS ROI in LEECHE12A	90
Table A.2:	Over-expressed masses in the leech CNS across different samples . .	92
Table A.3:	Identification of several other peptides	98
Table B.1:	Molecular signatures for anterior and posterior ganglia	107
Table B.2:	Molecules expressed in different regions of the rat brain cortex. . . .	112
Table B.3:	Molecules expressed in the amygdala and piriform cortex of the rat brain.	113
Table B.4:	Molecules expressed in different regions of the rat brain hippocampus.	114
Table B.5:	Molecules expressed in different regions of the rat brain thalamus and epithalamus.	115
Table B.6:	Molecules expressed in different regions of the rat brain hypothalamus.	116
Table B.7:	Molecules expressed in other various regions of the rat brain.	117

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Vineet Bafna, for his scientific guidance, his patience and academic support during my time at UCSD. This work could not have been completed without him. I would also like to thank my co-advisor, Eduardo Macagno, for sharing his vast knowledge in biology and taking the time to guide me throughout my Ph.D. In a similar fashion, I would like to thank the rest of my committee members, Terry Gaasterland for sharing her ever-improving assembly of the leech genome, Pieter Dorrestein for his bright novel ideas and data gold mine, Pavel Pevzner for his guidance and feedback, and Glenn Tesler for his mathematical genius and meticulousness.

I am appreciative of all my lab mates for their camaraderie, support and feedback throughout these years. In particular, I would like to thank all my officemates, past and present, especially Mark Chaisson, Shaojie Zhang, Fjola Bjornsdottir, Kyowon Jeong and Stefano Bonissone, for their company and laughter. Also, Sangtae Kim, Natalie Castellana, Ming Wang deserve special thanks for letting me pick their brains. Finally, I am grateful for the many climbing hours with my lab mates Stefano Bonissone, June Snedecor and Anand Patel.

I am truly indebted to all collaborators and co-authors, in particular the Salzet and Macagno lab, for sharing their experience and experiments.

Most of all, I would like to thank my family and friends for their company and for “supporting” me throughout this process, both in the English and French meaning of the term. I am particularly grateful to my sister, Corinne Bruand, for always making me laugh and my parents, Bernard and Yoke-Ping Bruand, for their motivation and support. Annie Pham, Sidric Torres, Alice Lo, Du Tran and Amy Che also deserve special thanks for their many years of moral support.

Finally, I would like to acknowledge the funding sources for this work, NSF and NIH.

Chapter 2, in full, was published as “Automated querying and identification of novel peptides using MALDI mass spectrometric imaging”. Bruand J, Sistla S, Mériaux C, Dorrestein PC, Gaasterland T, Ghassemian M, Wisztorski M, Fournier I, Salzet M, Macagno E, Bafna V. *J Proteome Res* 10(4): 1915-28 2011. The dissertation author was

the primary author of this paper.

Chapter 3, in full, was published as “AMASS: algorithm for MSI analysis by semi-supervised segmentation”. Bruand J, Alexandrov T, Sistla S, Wisztorski M, Meriaux C, Becker M, Salzet M, Fournier I, Macagno E, Bafna V. *J Proteome Res* 10(10): 4734-43. 2011. The dissertation author was the primary author of this paper.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Bruand J, Dorrestein PC, Bafna V. The dissertation author was the primary author of this material.

VITA

- 2005 Bachelor of Science in Information and Computer Science,
University of California, Irvine
- 2012 Doctor of Philosophy in Bioinformatics and Systems Biology,
University of California, San Diego

PUBLICATIONS

- Bruand J, Alexandrov T, Sistla S, Wisztorski M, Meriaux C, Becker M, Salzet M, Fournier I, Macagno E, Bafna V. 2011. AMASS: algorithm for MSI analysis by semi-supervised segmentation. *J Proteome Res* 10(10):4734-43.
- Deblasio D, Bruand J, Zhang S. 2011. A Memory Efficient Method for Structure-Based RNA Multiple Alignment. *IEEE/ACM Trans Comput Biol Bioinform.*
- Mériaux C, Arafah K, Tasiemski A, Wisztorski M, Bruand J, Boidin-Wichlacz C, Desmons A, Debois D, Lapr votte O, Brunelle A, Gaasterland T, Macagno E, Fournier I, Salzet M. 2011. Multiple changes in peptide and lipid expression associated with regeneration in the nervous system of the medicinal leech. *PLoS One* 6(4):e18359.
- Bruand J, Sistla S, M riaux C, Dorrestein PC, Gaasterland T, Ghassemian M, Wisztorski M, Fournier I, Salzet M, Macagno E, Bafna V. 2011. Automated querying and identification of novel peptides using MALDI mass spectrometric imaging. *J Proteome Res* 10(4):1915-28.
- DeBlasio D, Bruand J, Zhang S. 2009. PMFastR: A New Approach to Multiple RNA Structure Alignment.” *WABI* . p 49-61.
- Swamidass SJ, Chen JH, Bruand J, Phung P, Ralaivola L, Baldi P. 2005. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *ISMB (Supplement of Bioinformatics)* 21 Suppl 1:i359-368.
- Chen JH, Swamidass SJ, Dou Y, Bruand J, Baldi P. 2005. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* 21(22):4133-4139.

ABSTRACT OF THE DISSERTATION

Computational Tools for Analysis of Mass Spectrometry Imaging Data

by

Jocelyne Bruand

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2012

Professor Vineet Bafna, Chair
Professor Eduardo Macagno, Co-Chair

Imaging to assess the presence and localization of specific molecules in tissues and cells is central to the study of biological systems. However, most imaging technologies focus on specific molecules of interest. An exciting recent advance is the development of Mass Spectrometry Imaging (MSI), which allows for the generation of topographic 2D maps for various endogenous and some exogenous molecules (e.g., drugs and their metabolites) without prior specification. Advances in MSI have transformative potential, allowing us to answer questions about the localization of proteins, peptides, lipids, metabolites and other molecules. To help MSI realize its potential, we describe several algorithms for the analysis of MSI data from different angles.

In a first problem, we start with the premise that we are given a pre-defined region of interest (ROI) based on the morphology of the tissue or organism. We aim to find and identify molecules that are specifically expressed in the ROI. We solve this problem by using a statistics for localization specificity and a novel pipeline for identification.

Next, we extend the approach above to segment the MSI dataset into consistent regions of interest, and for each segment, we identify a molecular signature: a collection of peaks that are preferentially expressed in that segment. Our implementation, called AMASS (Algorithm for MSI Analysis by Semi-supervised Segmentation), relies on the discriminating power of a molecular signal instead of its intensity as a key feature, uses an internal consistency measure for validation, and allows significant user interaction and supervision as options.

A third problem examines the comparative analysis of many MSI datasets. We describe a new method which, given a set of pertinent query molecules, finds, in each dataset, all molecules that have a similar spatial distribution and clusters the datasets based on the resulting molecular signatures. The approach has the potential to identify unknown relationships between multiple data acquisitions.

Finally, we briefly touch on the peptide identification from on-tissue MS/MS data using a spectral library specific to MALDI imaging peptide identification. Our preliminary results highlight the potential of this approach.

Chapter 1

Introduction

Imaging to assess the presence and localization of specific molecules in tissues and cells is central to the study of biological systems. Historically, successful approaches usually involved labeling one/few proteins at a time either by attaching a fluorescent domain genetically or by treating a biological sample with labeled antibodies, and then recording two-dimensional (2D) micrographs of the sample, possibly also reconstructing them into a three-dimensional (3D) object or movie. Such imaging techniques are low-to-medium throughput approaches and give the biologist insight into just a small number of biological samples, limited to known proteins for which antibodies or tagged forms are available. By contrast, there is an increasing number of imaging technologies (transcriptomic or proteomic) that allow for the sampling and exploration of the entire complement of active molecules in the cell.

An exciting and innovative recent advance in mass spectrometry is the development of *Mass Spectrometry Imaging* (MSI). By applying mass spectrometry directly on tissue, MSI allows for the generation of topographic 2D and 3D maps for various endogenous and some exogenous molecules (e.g., drugs and their metabolites).

In this introduction, we will give a quick overview of mass spectrometry imaging, its potentials and its shortcomings.

1.1 What is Mass Spectrometry Imaging?

As mentioned above, mass spectrometry imaging is an imaging technique which allows us to visualize the spatial distribution of all detected molecules in a label-free manner. In Figure 1.1, we give an overview of the mass spectrometry imaging technique. Basically, we acquire a spectrum at each location of a pre-defined raster. By taking the intensity at each spectrum/location for specific m/z value, we create an intensity image which reflects the localization of the molecule (or set of molecules) having a mass corresponding to that m/z value. Thus, an MSI dataset can be visualized as a set of intensity image, one for each m/z value. By aligning the spectra as rows in a matrix, we can also represent an MSI dataset as a spot-by- m/z matrix containing the intensity value for the corresponding m/z value and the spot location. Three main techniques have been developed to acquire such datasets: Secondary Ion Mass Spectrometry (SIMS), Matrix-Assisted Laser Desorption/Ionization (MALDI) and Desorption Electrospray Ionization (DESI). SIMS typically allows for the detection of molecules up to m/z 1000-1500 with a spatial resolution of $400nm$ to $1 - 2\mu m$ [1]. Thus, it renders high spatial resolution images of small molecules, such as lipids or drugs. On the other hand, MALDI imaging can detect molecules for m/z values exceeding 100000, with some trade-off in spatial resolution ($25\mu m$) [2]. Thus, MALDI imaging has been the technique of choice for the study of peptides and proteins. Finally, DESI imaging has the capability of acquiring MSI data from ambient samples, eliminating the labor intensive sample preparation required in MALDI imaging. While the m/z detection range is higher than SIMS (up to 66000) [3], it has mostly been used for the study of lipids. The spatial resolution is lower than that of SIMS or MALDI imaging ($\sim 100\mu m$) [4], though the development of nano-DESI should allow for resolutions as low as $12\mu m$ [5].

1.2 Potentials of Mass Spectrometry Imaging

Mass spectrometry imaging has shown great success in a variety of areas. In the study of diseases, MSI has been applied to study of various types of cancer [6, 7, 8] and different neurodegenerative diseases [9, 10, 11] to define molecular pattern differences between healthy and unhealthy tissues, characterizing tissue types molecular profiles

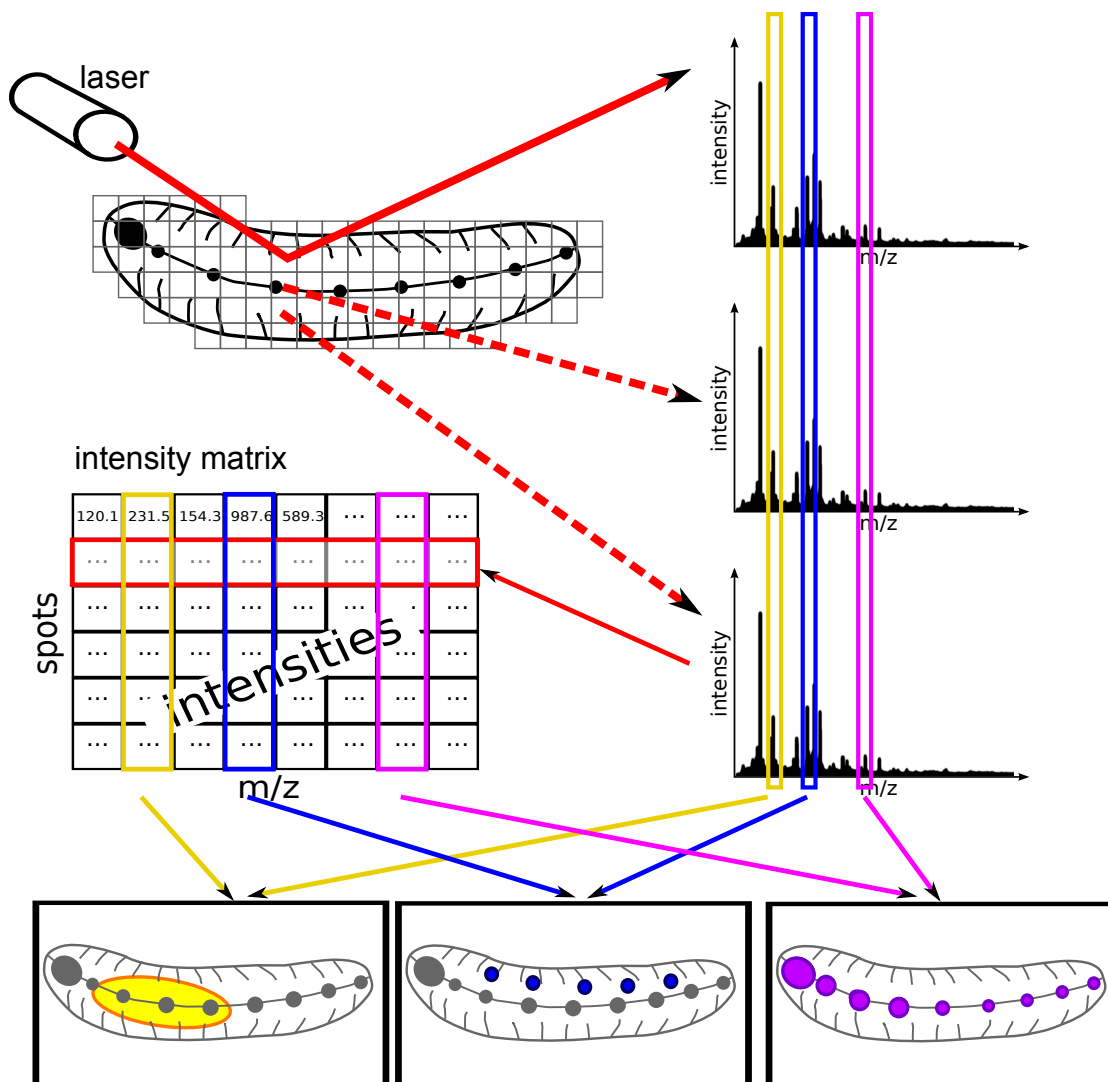


Figure 1.1: Mass Spectrometry Imaging (MSI) overview. Given an object of study (here, a leech) a spectrum is acquired for each point of pre-defined raster. Here, we show a laser hitting the sample across the raster, as would be the case in MALDI imaging. However, other MSI techniques may employ other acquisition methods (e.g., capillaries for DESI imaging). By taking the intensity at each spectrum/location for specific m/z value, we create an intensity image which reflects the localization of the molecule (or set of molecules) having a mass corresponding to that m/z value. Thus, an MSI dataset can be seen either as a set of intensity images, or as a spot-by- m/z matrix containing the intensities.

and identify biomarkers. Mass spectrometry imaging has also seen uses in drug development due to its ability to detect and localize both the parent drug and its metabolites throughout tissues or whole organisms [12].

MSI can also be used to further our understanding of biological systems. Studies have been done on various tissue types such as brain [13, 14], prostate [15, 16], ocular lens [17, 18], and whole organisms [11, 19, 20]. It allows us to create molecular profiles for different tissue types and monitor molecular changes across different conditions. Another interesting approach has been to use MSI to study the interaction between bacterial colonies Yang2009, Liu2010, Watrous2010.

Many new developments in MSI open doors to further biological investigations. Three-dimensional MSI allows for the visualization of molecules in the entire sample [21], instead of just a slice. The combination of quantitation and MSI enables us to know the concentration of a drug, peptide or protein across different tissue types [22, 23].

1.3 Shortcomings of Mass Spectrometry Imaging

There are various shortcomings to the mass spectrometry imaging technique. Currently, MALDI imaging requires very precise and labor-intensive sample preparation. Lack of homogeneity in matrix deposition and crystal formation creates biases in the spatial distribution of the molecules, though the development of robotic matrix deposition has helped reduce this effect [24, 25]. Second, MALDI imaging spatial resolution ($\sim 25\mu m$) is still inferior to optical resolution ($< 1\mu m$) and does not allow for cell-level imaging. In order to increase spatial resolution, it is necessary to form smaller matrix crystal and improve instrument acquisition speed [?].

Another major limitation is the ion suppression effect [26]. One molecule can completely suppress the signal of other equally abundant molecules. Moreover, intensities may not directly reflect molecules abundance as there are disparities in ionization efficiency.

Finally, one of the major current challenges is the identification of the compounds [27, 2, 28]. Two main approaches are used. In a top-down approach, proteins

of interest are extracted from tissue and identified. This process is labor-intensive and loses the spatial localization of the protein. In a bottom-up approach, trypsin is spotted on-tissue and MS/MS spectra are acquired directly on tissue. In this case, the sample is very complex and abundant molecules dominate the spectra, making identification challenging.

Chapter 2

Automated querying and identification of novel peptides using MALDI mass spectrometric imaging

MSI is a molecular imaging technique which allows the generation of topographic 2D maps for various endogenous and some exogenous molecules, without prior specification of the molecule.

In this chapter, we start with the premise that a *region of interest* (ROI) is given to us based on pre-selected morphological criteria. Given an ROI, we develop a pipeline, first, to determine m/z values with distinct expression signatures, localized to the ROI and, second, to identify the peptides corresponding to these m/z values.

To identify spatially differentiated m/z values, we implement a statistic that allows us to estimate, for each spectral peak, the probability that it is over-, or under-expressed within the ROI versus outside. To identify peptides corresponding to these masses, we apply LC-MS/MS to fragment endogenous (non-protease digested) peptides. A novel pipeline based on constructing sequence tags *de novo* from both original and de-charged spectra, and subsequent database search is used to identify peptides. As the MSI signal and the identified peptide are only related by a single mass value, we isolate the corresponding transcript, and perform a second validation via *in situ* hybridization of the transcript.

We tested our approach on a number of ROIs in the medicinal leech, *Hirudo*

medicinalis, including the central nervous system (CNS). The Hirudo CNS is capable of regenerating itself after injury, thus forming an important model system for neuropeptide identification. The pipeline helps identify a novel gene, HmIF4, a member of the intermediate filament family involved in neural development, and second novel, uncharacterized peptide. A third peptide, derived from the histone H2B, is also identified, in agreement with the previously suggested role of histone H2B in axon targeting.

2.1 Introduction

The use of multiple imaging techniques to assess the presence and location of specific proteins in tissues and cells is central to the study of biological systems. The prevailing approach is to label one or several proteins at a time either by attaching a fluorescent domain genetically or by treating a biological sample with labeled antibodies, and then to record two-dimensional (2D) micrographs of the sample, possibly reconstructing them into a three-dimensional (3D) object or movie. Such imaging techniques are low-to-medium throughput approaches and give the biologist insight into just a small number of biological samples, limited to known proteins for which antibodies or tagged forms exist.

In contrast to the low throughput of imaging technologies, some available genomic, transcriptomic, and proteomic (particularly via mass spectrometry) technologies allow for the sampling and exploration of the entire complement of active molecules in the cell.

An exciting and innovative recent advance in mass spectrometry is the development of *Mass Spectrometry Imaging* (MSI). MSI is a molecular imaging technique which allows the generation of topographic 2D maps for various endogenous and some exogenous molecules (e.g., drugs and their metabolites) involving the application of mass spectrometry directly on tissue.

In the Matrix-Assisted Laser Desorption/Ionization (MALDI) MSI workflow, thin tissue sections (10-15 μ m) from organs or even whole body animals are mounted onto a conductive glass slide allowing microscopic observation of the tissue prior to MS analysis. Important preparative steps include appropriate tissue treatments [29] and

ionic matrix deposition which must be optimized to reach highest analytical performance [30]. By incorporating a target scanning capability within the mass spectrometer itself, it is then possible to obtain mass spectra at a series of specified locations on the target. A major advantage of direct MALDI MSI analysis is to avoid time-consuming extraction, purification or separation steps, which have the potential for producing artefacts. After its introduction by Spengler et al. in 1994 [31], direct MALDI analysis of tissue sections was developed by various groups [32, 33, 34, 35, 36]. The studies performed by these groups demonstrated that acquisition of tissue expression profiles while maintaining cellular and molecular integrity was feasible. With automation and new analysis software, it also became possible to produce multiplex imaging maps of selected bio-molecules within tissue sections [37, 38, 39].

While the true abundance of a molecule cannot be measured using this approach, the intensity of its corresponding spectral peak (or, its expression level) often correlates with its abundance, albeit in a complicated manner (e.g., compounds with poorer ionization efficiency display lower intensity peaks than would be expected purely on their abundance). Molecules that are preferentially expressed in a region of the sample will show higher intensity in the image corresponding to a specific m/z value when represented with the intensity encoded by a color map. It is also important to note that when looking at these m/z images, it is possible to be looking at the combined intensity of several compounds with similar m/z values. Most bioinformatics approaches have focused on using MALDI MSI as a tool for the discovery of signature markers of particular physiological stages. One approach is to distinguish regions of the tissue presenting very different mass spectral signatures. This has been addressed by a number of researchers, who use unsupervised clustering methods to characterize a region of interest (ROI) [40, 41].

While unsupervised clustering is essential to the analysis of large datasets without user input, it ignores prior knowledge about tissue morphology. In many cases, a more targeted, or supervised, approach is desirable, allowing the user to pull out the molecular signature for a specific area of interest. In this chapter, we start with the premise that a *region of interest* (ROI) is given to us based on pre-selected morphological criteria. As an example of an ROI, consider the central nervous system (CNS) of

the medicinal leech, *Hirudo medicinalis*, one of the best-studied representatives of the phylum Annelida (segmented worms). Given a particular ROI, we ask (a) which masses have a distinct expression signature, localized to the ROI; and, (b) to which peptides do these masses correspond? The answer to these questions, in the context of, for example, several embryonic stages, can help us identify key peptides and proteins in leech neuronal development. As the leech CNS has a demonstrated capacity to repair itself after injury and to restore function [42, 43, 44], the discovery of peptides involved in neuronal development and regeneration could have therapeutic implications.

To answer our first question, i.e., what are the masses that are specifically expressed in a region of interest, we developed a statistical method that operates on MALDI MSI data. While some recently published methods seek to differentiate molecules between two regions (eg. cancerous vs. non-cancerous) [45, 46, 47, 7], we provide a publicly available tool which allows for the analysis of non-contiguous regions, using various methods. We also validate our method using simulations. An interactive tool allows the user to define the ROI on a histological image of a leech embryo. We defined several different ROI corresponding to CNS, the lateral/ventral regions, and the nephridia. We implemented a statistic that allows us to estimate, for each spectral peak, the probability that it is more highly or less highly expressed within the ROI versus outside. The m/z values that we find include some whose expression is very low relative to other peaks but strongly localized to the region of interest. The method was validated, using both simulated perturbations of the original intensities, as well as visual inspection of MALDI images restricted to the peaks of interest. All selected peak images displayed localization in the area of interest and the signal became visually weaker and less localized to the ROI as the score decreases. The statistic was used to identify peak masses specifically, expressed in the different ROIs.

The second question we pursued is the identification of the peptides associated with those mass values. In fact, identification of the species showing interesting spatial distributions remains one of the most challenging problems in MSI. Many recent studies have focused on this problem and some of these approaches obtained sequence information by performing MS/MS directly from the tissue. In the case of identifying larger peptides or proteins, these approaches have favored a bottom-up method comprising of

in situ tissue digestion by applying a proteolytic enzyme with a spotter or sprayer, followed by MS/MS on tissue [48, 49, 50]. While these methods are powerful in that they give us a broader overview of the proteome while combining imaging and identification in one step, they have several limitations. First, identification from on-tissue MS/MS has been restricted to high-abundance molecules and remains a challenge for low abundance molecules [48, 51]. Moreover, digestion greatly increases the complexity of the spectrum, especially for lower masses, although one proposed solution is to couple an ion mobility mass spectrometer to the MALDI-TOF instrument thus using drift time as an additional separating dimension [52]. Finally, enzymatic product diffusion [51], variation in peptide intensities [53], and the fact that many parent masses will have similar distributions, all increase uncertainties in the correlation between parent image and trypsin product images. It is worth noting that Chen et al. [54] opted for another bottom-up approach from the sample in the identification of neuropeptides in the lobster. While they successfully sequenced many neuropeptides from extracts, MALDI imaging was used independently only as a second step to visualize the localization of these identified neuropeptides, using mass value to correlate the images to the peptides. Thus, the approach does not necessarily identify specific molecules of interest.

In contrast, we focus on LC-MS/MS identification of endogenously processed peptides (2000-5000 Da). By not using a protease digestion step, we maintain the link between the observed mass and the identified peptide. The identification is challenging, as the fragmentation patterns of high-charge, non-tryptic peptides are poorly understood [55]. Currently, while top-down mass spectrometry allows for the identification of spectra of larger proteins, it requires either a) labor-intensive sample purification to isolate the protein of interest or b) a highly abundant protein in order to obtain spectra with good isotope resolution which is necessary for identification. In order to use complex sample and identify less abundant peptides, we limit the identification here to intermediate sized peptides despite a larger range of interesting m/z values. We developed a custom peptide identification pipeline based on constructing sequence tags *de novo* from both original and de-charged spectra, and performing a database search including modifications (see Figure 2.1). As the MSI signal and the identified peptide are only related by a single mass value, we isolate the corresponding transcript, and per-

form a second validation via *in situ* hybridization of the transcript. Using this method, we successfully identified a number of peptides (see Table A.3). One of these peptides belongs to a novel gene that we call HmIF4; it is a member of the family of intermediate filaments (IF), with strong sequence similarity to gliarin, macrolin and filarin, three previously characterized IFs in *Hirudo medicinalis*, which were known to be expressed in the CNS. Whole mount *in situ* hybridization (see Methods) with a probe to the corresponding RNA matched well with MALDI imaging data, supporting the identification. A second identified peptide corresponds to a segment of histone H2B, and showed consistent localization via *in situ* hybridization as well. A third identified peptide is completely novel, and not currently represented in the leech genomic databases (NCBI nr, Helobdella proteins and *Hirudo* EST, see Methods).

2.2 Methods

Figure 2.1 provides an overview of the method, which has two subprocesses: MSI based peptide/protein localization and MS/MS based peptide/protein identification.

2.2.1 MALDI Imaging Data Acquisition

In brief, the MSI data used to test the computational methods reported here were acquired from two 12-day old leech embryo specimens, herein referred to as LEECHE12A and LEECHE12B to reflect their embryonic age (12 days at 24°C). The specimens were opened along the dorsal midline, pinned flat and the yolk sack and endoderm removed to expose the central nervous system. Next they were exposed briefly (1-2 min) to methanol in order to lightly fix and permeabilize the tissues, then placed on glass slides coated with indium tin oxide (ITO) and immediately dried. Methanol was selected because it provided a quick, one-step fixation (non-cross-linking) and permeabilization that works well with leech embryos. We also found that it aids efficient peptide extraction following matrix application, though this was not assayed against other possible lipid solvents, as this was not the principal goal of the work reported here. The embryos were mounted so the internal surface of the body wall faced the laser beam. After recording optical images of the mounted embryos, they were coated with several

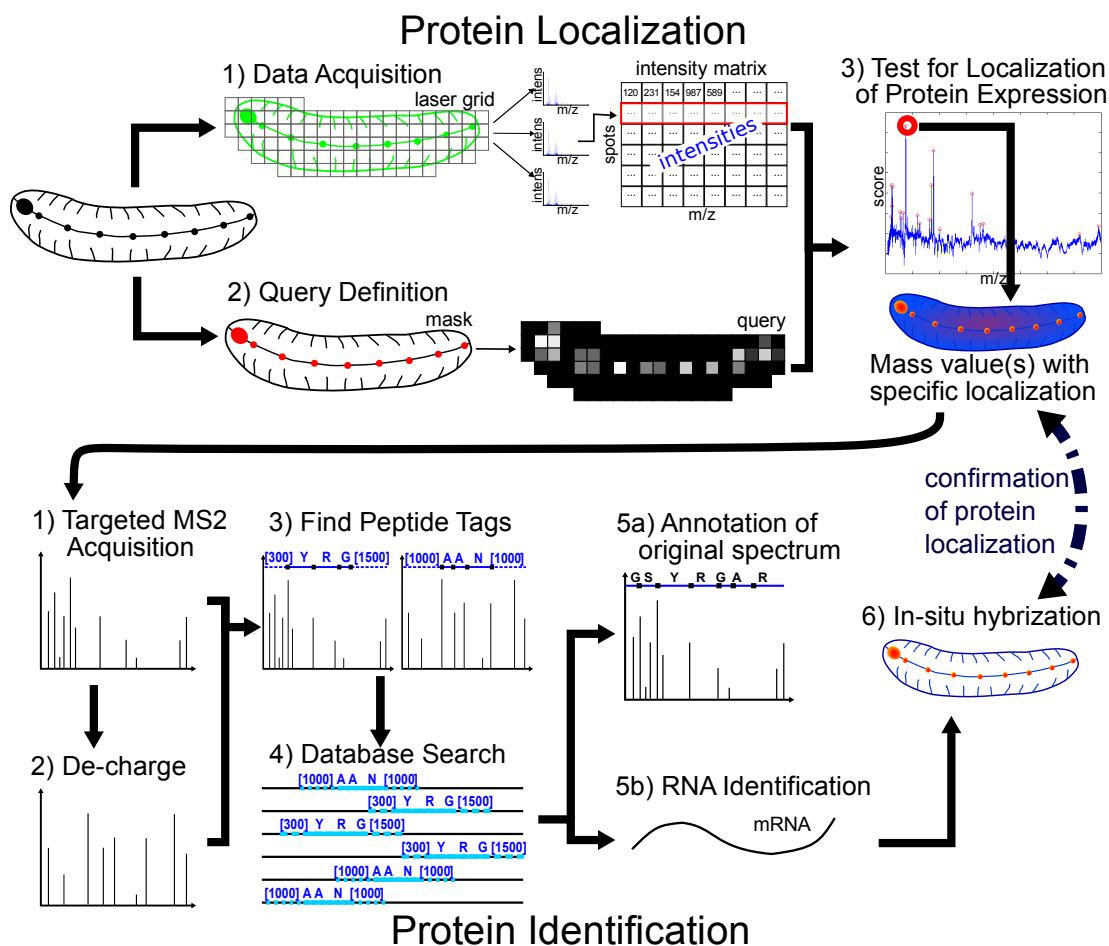


Figure 2.1: Overall process for detecting and identifying masses specifically expressed within an ROI or specific morphological feature. The process consists of two major parts: protein localization (top) and protein identification (bottom). Protein localization: We developed a pipeline to detect proteins that are preferably expressed in a given ROI using MALDI imaging data. This pipeline consists of 3 parts: 1) data acquisition and processing, 2) query definition and 3) analysis. Protein identification: We developed a pipeline to identify peptides specifically expressed in our ROI. This pipeline consists of 6 steps: 1) targeted MS2 acquisition, 2) de-charging of acquired spectra, 3) finding peptide tags, 4) database search, 5) annotation of spectrum and RNA identification, and 6) in-situ hybridization.

layers of special solid ionic matrices (CHCA/Aniline), using a manual pneumatic TLC sprayer (VWR, Strasbourg, France). Such matrices have proved to be quite efficient for peptide/protein analysis directly from tissue sections (increased signal intensity, increased number of detected peptides/proteins, higher stability under vacuum conditions, lower ablation rate) [29]. MALDI Imaging was performed on a MALDI-TOF/TOF instrument (Ultraflex II, Bruker Daltonics, Germany) at the University of Lille. While only MS1 spectra were acquired in the mass spectrometric imaging stage, it is worth noting that a TOF/TOF acquisition could be useful to help correlate the sequenced peptides with the original imaging data by using the TOF/TOF partial fragmentation. However, because many of the interesting molecules are of lower abundance (see Figure A.9), it is likely that even partial fragmentation may be hard to obtain straight on tissue without protein concentration. Spectra were acquired over 38837 m/z values from 12115 locations in a rectangular raster of points $60\mu\text{m}$ apart on LeechE12a and 37199 m/z values from 22230 locations at raster in a rectangular raster of points $35\mu\text{m}$ apart on LEECHE12B. Data was acquired on a wide range of m/z values to ensure that our software could detect spatially localized molecules on a large scale of values with different noise levels. Because the data was acquired on a wide m/z range, the m/z resolution did not allow us to detect isotopic patterns on the imaging data. However, peaks were well matched across spectra and across samples. The complete data-set is a collection of spectra, each associated with a ‘spot’ on the leech surface. Conceptually, the data can be represented as a collection of triplets $\langle m, s, I_{m,s} \rangle$ describing the spectral intensity $I_{m,s}$ at each spot s , and m/z value m .

2.2.2 MALDI Imaging Data Normalization

Each spectrum was normalized to correct for systematic biases, including an m/z dependent bias, and a region specific bias. The spatial bias is clearly seen in Figure A.1, with an order of magnitude difference in total intensity across different regions. A median baseline correction (flexAnalysis) was employed to correct for the m/z bias. Note that baseline correction causes some intensity values to become negative. To correct for spatial bias, we performed normalization after baseline correction. The average intensity

for positive intensities after baseline correction at each spot was computed as

$$A(s) = \frac{\sum_m I_{m,s}}{\#\{m | I_{m,s} > 0\}}.$$

The data was normalized by recomputing the intensities as

$$I_{m,s} \leftarrow \frac{I_{m,s}}{A(s)} \sum_{s,m} A(s).$$

This data was written into a custom compressed lossless format in order to facilitate data analysis.

2.2.3 Query Definition

We defined an ROI by manually creating a *mask*, which is an image that can be superposed onto the histological image. Formally, a mask *mask* \mathcal{M} maps each high-resolution pixel to a binary value $\mathcal{M}_p \in \{0, 1\}$, indicating whether or not the pixel is part of the ROI. These masks are easily created by adding layers onto the image using any standard editing tool with layer capabilities. These layers can then be exported as separate images. We developed a plug-in that extends the open-source GNU Image Manipulation Program (GIMP, <http://www.gimp.org/>) to facilitate the exportation process, allowing the user to export any combination of layers into new images.

Because the MALDI spots are at much lower resolution than the mask, they can be partially inside and partially outside the ROI. Thus, we define our *query* to designate how much of each MALDI spot belongs to the ROI. Thus, the query \mathcal{Q} maps each low-resolution laser spots s onto a real value $\mathcal{Q}_s \in [0; 1]$. For each low-resolution laser spot s , we can define the collection of pixels $p \in s$ on the light-transmitted image which belong to the laser spot. We assign the following value to each spot:

$$\mathcal{Q}_s = \frac{\sum_{p \in s} \mathcal{M}_p}{\#\{p \in s\}}.$$

Figure A.2 shows the user-defined mask for the leech CNS, and the resulting grey-image query \mathcal{Q}_s for all spots s .

2.2.4 Query Shift

Mapping of the MALDI spots to the histological image is done here by manually defining *teaching points*, which are a set of spots with coordinates on both images, in

flexImaging software (Bruker Daltonics) prior to acquisition. By using these matching sets of coordinates, it is possible to calculate the relative scale ratios and establish a correspondence between the coordinates of both images. However, because of the lack of precision in the definition of the teaching points, the mapping and/or scaling of MALDI images to the histological image may be slightly off. Because the ROI is defined on the optical image, which shows the morphological features, it is important to minimize mapping errors between the two types of images. In order to correct for imprecision in defining the teaching points, we introduced the possibility of manually setting shifts by translation and/or scaling of the mapping from the MALDI image to the optical image. Both shifts are done independently on the x-axis and y-axis, as teaching points can have lack of precision on either axis.

2.2.5 Testing for Localization of Protein Expression

The input to this process is the set of normalized intensities $I_{m,s}$ and one or two *query(ies)*. If only one query is specified, the given ROI is compared to the rest of the spots, as done in Figure 2.2. If two queries are specified, then the intensities from the first ROI are compared to the intensities from the second ROI, ignoring the rest of the data, as done in Figure 2.3. Depending on the statistical test, it may be necessary to define two sets of spots from the query: those in the ROI and those outside the ROI. For a query \mathcal{Q} , we can define two thresholds t_1 and t_2 , such that $t_1 \geq t_2$. Then, the set of spots such that $\mathcal{Q}_s \geq t_1$ are within the ROI and the set of spots such that $\mathcal{Q}_s \leq t_2$ and $\mathcal{Q}_s > t_1$ are outside the ROI. Note that there is no overlap between the two sets but that there may be spots not belonging to either set. If the input is two queries, one threshold t_1 is given for each query and spots belonging to both query are arbitrarily assigned to the first ROI. While the user has the option to define those query thresholds, all values are defaulted to 0.5.

A set of intensities for all spots exhibits the same pattern as the given ROI if the intensities are distributed such that there is a separation between those within the ROI and those outside the ROI. While our software allows the user to choose between several statistics, we use the ρ statistic, which is the Mann-Whitney U statistic, which ranges between 0 and $n_{\text{ROI}}n_{\text{bg}}$, normalized by its maximum possible value. In the case of

ties, we assign the average rank to all corresponding values. For each m/z , we calculate the Mann-Whitney U statistic for the average intensity over a range of $\pm 2Da$ as $U = R_{\text{ROI}} - \frac{n_{\text{ROI}}(n_{\text{ROI}}+1)}{2}$ where n_{ROI} is the number of spots in the ROI and R_{ROI} is the sum of the ranks of the intensities in the ROI. The ρ statistic is calculate as $\rho = \frac{U}{n_{\text{ROI}}n_{\text{bg}}}$, where n_{bg} is the number of spots outside the ROI. High-scoring peaks for $m/z < 2200$ were discarded because many spectra did not have any peaks in that region causing a bias in the localization.

2.2.6 Simulations

In order to assess the performance of our method, we generate some simulated data. The first simulated data aims to see how our method performs when the ROI signal decreases in terms of area, that is we want to see how the statistic behaves as less spots in the ROI show higher expression. Let n_1 and n_2 be the number of spots inside and outside the ROI respectively, and let $I_{\text{ROI}}(r_1, \dots, r_{n_1})$ and $I_{\text{bg}}(b_1, \dots, b_{n_2})$ be the corresponding sets of intensities. We sort the ROI spots by location and the background spots by intensities. In order to generate a random background intensity, we randomly select a background spot b_i such that $1 \leq i \leq (n_2 - 1)$ and we generate a random intensity I_{rand} sampled uniformly in $[I_{\text{bg}}(b_i), I_{\text{bg}}(b_{i+1})]$. To generate the simulated data, we incrementally set k ROI spots r_1, \dots, r_k to a random background intensity. Thus, the ROI spots now have intensities I_{sim} assigned to them such that $I_{\text{sim}}(r_1, \dots, r_k)$ are random background intensities, and $I_{\text{sim}}(r_i) = I_{\text{ROI}}(r_i)$ for $k + 1 \leq i \leq n_1$. In the case of re-balancing the total intensities, we distribute the subtracted intensity by setting

$$I_{\text{sim}}(r_i) = I_{\text{ROI}}(r_i) + \frac{\sum(I_{\text{ROI}}(r_1, \dots, r_k)) - \sum(I_{\text{sim}}(r_1, \dots, r_k))}{n_1 - k}$$

for $k + 1 \leq i \leq n_1$.

Our second simulation aims to see how our method performs when the ROI signal decreases over the entire region. In this case, we simply decrease the intensities of each ROI spots by a certain percentage until the average intensity inside the ROI is the same as the non-ROI (or background) intensities. This means that for each spot r_i within the ROI, we assign the intensity $I_{\text{sim}}(r_i) = x * I_{\text{ROI}}(r_i)$ where $x \in [\text{mean}(I_{\text{bg}})/\text{mean}(I_{\text{ROI}}), 1]$.

2.2.7 MS/MS Sample Preparation and Data Acquisition

Identification of the molecules with particular m/z values selected from the MALDI-TOF imaging required the acquisition of high-resolution MS/MS spectra. We therefore tested several extraction procedures for obtaining intact proteins/peptides without enzymatic digestion that yielded good MS and MS/MS results with either MALDI or ESI methods. These included extraction with 1N acetic acid, 1N HCL, TCA, basic extraction with ammonia, 50:50:1 Methanol:water:FA, and PBS. For the purposes of the work described here, the most consistent results were obtained with a simple PBS extraction (a comparative study of these methods will be published elsewhere). Peptides were extracted from leech embryos of embryonic ages between 6 and 12 days old. Forty embryos with or without yolk were snap frozen in liquid nitrogen and then pulverized in a Dounce homogenizer. The homogenized tissue was stored at -80°C after thorough lyophilization. The homogenized tissue was then dissolved in $200\mu\text{L}$ of ice cold 100mM PBS (pH 7.4) containing protease inhibitor cocktail and 0.1M PMSF and vortexed. Tissue was dissolved by stirring at 4°C for another 4 hours. Samples were sonicated for 5 mins with 10 second pulses at constant voltage. The extract was then centrifugated at 12000 rpm for 30 mins at 4°C . The supernatant was separated from the pellet and was reextracted using PBS. Supernatants from all the extractions were pooled and samples were desalted and then dried using a Speedvac before proceeding to mass spectral identification.

The samples were analyzed by liquid chromatography (LC) coupled with tandem mass spectrometry with electrospray ionization. All nanospray ionization experiments were performed by using a QSTAR-Elite hybrid mass spectrometer (AB/MDS Sciex) interfaced to a nanoscale reversed-phase high-pressure liquid chromatograph (Tempo) using a $10\text{cm}-180\text{ ID}$ glass capillary column packed with $5\text{-}\mu\text{m}$ C18 ZorbaxTM beads (Agilent). The buffer compositions were as follows: buffer A was composed of 98% H₂O, 2% ACN, 0.2% formic acid, and 0.005% TFA; buffer B was composed of 100% ACN, 0.2% formic acid, and 0.005% TFA. Peptides were eluted from the C-18 column into the mass spectrometer using a linear gradient of 5-80% buffer B over 140 min at $400\mu\text{L}/\text{min}$. Time-of-flight MS were acquired at m/z 400 to 2000 Da for 0.5 s with 12 time bins to sum. MS/MS data were acquired from m/z 50 to 2000 Da by using “enhance

all” and 24 time bins to sum, dynamic background subtract, automatic collision energy, and automatic MS/MS accumulation with the fragment intensity multiplier set to 6 and maximum accumulation set to 2s before returning to the survey scan. LC-MS/MS data were acquired in a data-dependent fashion by selecting the 5 most intense peaks with charge state of 2 to 5 that exceeds 20 counts, with exclusion of former target ions set to “360 seconds” and the mass tolerance for exclusion set to 100 ppm. The data dependent acquisition was also operated with inclusion and exclusion lists to include an ion selection list for MS/MS analysis and exclusion of ions already analyzed.

2.2.8 MS/MS Identification

The *Hirudo* genome has not been sequenced, so we create a custom database, LeechProtsDB, comprised of *Hirudo* EST sequences [56] (<http://genomes.sdsc.edu/leechmaster/database/>), the predicted *Helobdella robusta* proteins (JGI, v1.0, <http://genome.jgi-psf.org/Helro1/Helro1.download.ftp.html>) and all *Hirudo* protein sequences from the NCBI nr database (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). We cluster and de-charge the raw MS/MS spectra and predict peptide tags on both the original and the de-charged spectra using PepNovo[57] version 20090907. Peptide tags were predicted using with no enzyme, fragment tolerance of 0.5 Da, parent mass tolerance of 2.5 Da, considering the post-translational modifications M+16, Q+1 and N+1. A *peptide tag* is a string of residues with flanking masses on the C-terminus and the N-terminus (Figure 2.1, Database Search). The fragment-ions comprising the top scoring tags were manually investigated for charge confirmation. Tags that passed this validation were searched against our custom database. A database peptide was considered a candidate if it matched the tag perfectly, and the residues on either end had masses that matched the tag masses. All candidate peptides were scored according to the fraction of explained intensity, which is the proportion of the total spectrum intensity which can be explained by annotated peaks.

2.2.9 In situ hybridization

A complementary coding strand probe was obtained by *in vitro* transcription of the PCR product using T3 polymerase (Invitrogen), and subsequently hydrolyzed into shorter fragments. *In situ* hybridization was then performed with a digoxigenin-labeled RNA probe as described previously [58]. In brief, embryos of various ages were fixed in paraformaldehyde and further by Pronase E digestion. Day 6-8 embryos were digested for 20-25 minutes, older embryos for longer times. Digested embryos were then hybridized overnight at 59°C in 50% formamide with approximately 1 *ng/mL* digoxigenin labeled RNA. Washed embryos were treated with RNAase A (Sigma) to degrade unhybridized probe. Hybridized probe was visualized immunohistologically with an alkaline phosphatase (AP)- conjugated anti-digoxigenin (Roche) reacted for periods ranging from 15 hours to 3 days, using NBT and X-phosphate color reagents (Roche). Intact embryos were cleared in 80% glycerol, mounted under a coverslip and photographed.

2.3 Results and Discussion

2.3.1 An overview of the pipeline

In Figure 2.1, we show our customized pipeline for detecting and identifying masses specifically expressed in a given morphological feature. It consists of two major subprocesses: protein localization (top) and peptide identification (bottom). The first subprocess allows us to detect peptides or proteins that are preferentially expressed in a given ROI using MALDI imaging data and consists of 3 parts: 1) data acquisition and processing, 2) query definition and 3) analysis. First, we acquire spectra across a raster of locations across the entire tissue or specimen, obtaining a list of spectra associated with specific spatial coordinates. A histological image of the specimen taken prior to matrix deposition serves to localize raster points, or MALDI spots, on the specimen. We use the histological image to manually define a *mask* of the ROI, which is stored as a transparent layer with black pixels only within the ROI. Because MALDI spots are generally at lower resolution than the mask, the mask must be converted into a *query* as described in Methods. The query defines which MALDI spots are in the ROI, or

more precisely, how much of each MALDI spot belongs to the ROI. Queries can also be searched against each other in order to find molecules that are differentially expressed from one ROI to another. An example of a mask and query for the central nervous system of a leech embryo specimen (LEECH12A) is shown in Figure A.2. For each m/z , we apply a statistical test to decide if it is preferentially expressed in the ROI.

The critical performance issue is to rank results correctly. We use the ρ statistic, which is the Mann-Whitney U statistic normalized by its maximum possible value. Conceptually, it represents the probability that, given two random spots, one in ROI and one outside the ROI (or in the second ROI), the intensity of the spot inside the ROI is higher than that of the spot outside the ROI. Thus, a ρ statistic value of 0.5 indicates that the expression is not specific to the ROI. As the statistic approaches 1, it becomes increasingly likely that the intensity of a random ROI spot is greater than that of a random non-ROI spot. This means that the expression becomes more localized to the ROI. Conversely, as the statistics approaches 0, it becomes increasingly likely that the expression is inversely localized to the ROI. There is no hard line between “good” and “bad” localizations, but rather gradual decrease in specificity of the localization to the ROI. Therefore, we leave it up to the user to decide a probability threshold based on the desired quality of results. For our purpose, we often used a threshold of 0.65. The performance of the statistic for simulated, and actual leech data is discussed in detail in the following sections.

The second stage of the pipeline is aimed at identifying the peptides specifically expressed in our ROI. MS/MS spectra are acquired by specifically targeting the list of masses detected by the first stage (Figure 2.1). This is done in a data-dependent or semi-data-dependent manner (see Methods). To maintain the connection with the m/z values, the MS/MS spectra are acquired using a non-proteolytically digested sample. We use a high-accuracy QTOF instrument (sub-3 ppm), with multiple collision energies to provide a high-quality fragmentation. Multiply charged fragment ions are de-charged using isotopic peaks. Next, we generate peptide sequence *tags* on both the original and the de-charged spectra. We define a tag as a short string of amino-acids flanked by mass values (see Figure 2.1 : Database Search). We search all tags in a modification-tolerant manner against a custom protein database and annotate and score the resulting

candidate peptides from the search. At this stage, we have a top-scoring candidate peptide identified through MS/MS, with a parent mass value that shows preferential expression in the ROI. As a test of this identification, we synthesize a probe from the corresponding mRNA with embryonic cDNA as a template, which is then used in *in situ* hybridization assays to verify that mRNA and peptide co-localize to our ROI. We would expect peptide co-localization to the mRNA, but in some instances the mRNA could have a wider distribution, suggesting post-translational regulation. Conversely, it is also possible that the peptide is transported to a sub-cellular location, that is different from the region of synthesis. Hence, co-localization is supportive of identification, but lack of it cannot be taken as proof of mis-identification.

2.3.2 Data Normalization

Similar to other studies [59], we observe significant and systematic bias in the distribution of the intensities, both on the m/z axis and spatially. Reasons for the spatial bias include differing tissue composition or thickness and heterogeneity of ionic matrix crystallization [29]. In order to eliminate this bias, we must first do a baseline correction on all spectra (m/z -dependent bias), then we must normalize the intensities across all spectra (see Methods). In Figure A.1a, we see that there is large variation in the average spot intensities. In panel *b*, we see the distributions of the average spot intensities across the leech surface before (top) and after (bottom) normalization. After normalization, all spot spectra have the same average peak intensity. Not correcting the bias can lead to erroneous conclusions on the localization of expression for many molecular species. As an example, the leech brain appears to have overall significantly lower total intensity compared to other regions. In panel *c*, correcting for this bias reveals the species $m/z = 10357.1$ (right) as being higher in the brain (top vs. bottom panels). On the other hand, the species at $m/z = 8563.04$ (left) appears to have significantly localized expression initially. However, the significance is diminished after correction.

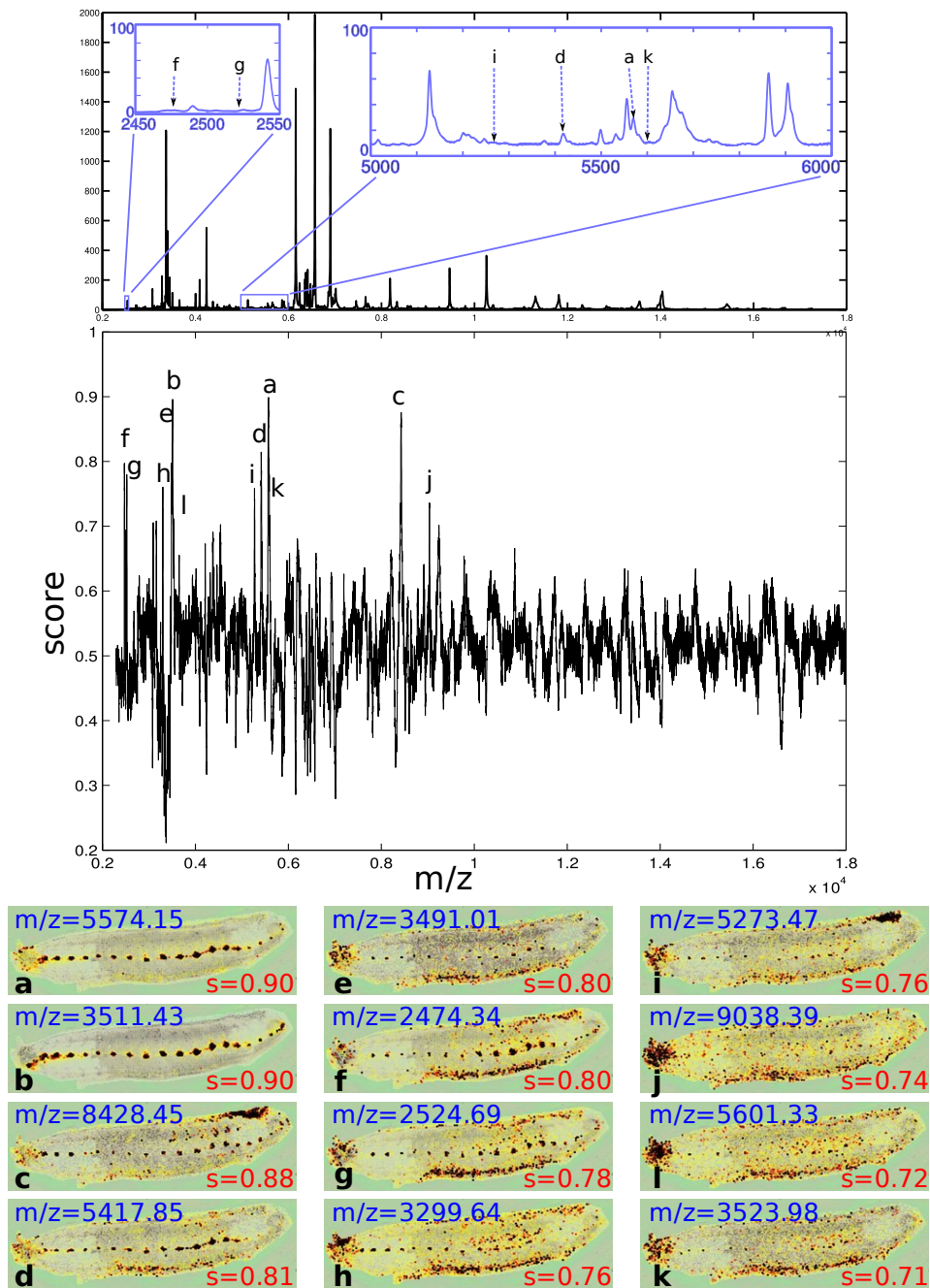


Figure 2.2: Results of the ρ -statistical test for the CNS query in the leech. We identified 43 m/z values that were significantly present in the CNS of LEECHE12A (score ≥ 0.65), which are listed in Table A.1. Visual inspection clearly demonstrates the power of the method as illustrated by the images for the top 12 most significant m/z values. Score decreases along with quality of CNS localization.

2.3.3 Defining Regions of Interest

As described in Methods, MSI-Query was applied to two samples denoted as LEECHE12A and LEECHE12B. A total of 12115 and 22230 MALDI MS spectra were acquired on LEECHE12A and LEECHE12B respectively. We created masks for three distinct ROIs (*CNS*, *lateral-ventral*, and *nephridia*) onto the histological images and converted them into queries (see Methods). These masks correspond to the following embryonic tissues.

CNS: The leech segmental ganglia are very similar to each other and comprise of about 400 neurons, $\simeq 180 - 190$ pairs and $\simeq 30$ unpaired [60], many of which are very well characterized developmentally, anatomically, physiologically and neurochemically [61, 62]. Furthermore, unlike the mammalian CNS, the leech CNS has a demonstrated capacity to repair itself after injury and to restore function [42, 43, 44]. We defined the central nervous system (CNS) as the segmental ganglia, the head ganglion and the tail ganglion in each specimen.

Lateral-Ventral: As in other animals, mechanosensory neurons in the leech innervate the skin in a regular pattern of domains also known as tiles. One example of these types of cells are those that respond to light touch on the skin surface (TV, TL and TD cells). These cells have an interesting difference in how they set up their sensory arbors, subdividing each segment: the TV cells innervate left and right ventral tiles, the TL cells innervate left and right lateral tiles, and the TD cells innervate left and right dorsal tiles. A very interesting problem is to identify what peptides/proteins (or other molecules) might mark the boundaries or areas of each tile and signal to the sensory cells where to locate their arbors. We defined the lateral domains as those extending from head to tail between two lines drawn along the ventral-most and dorsal-most boundaries of the laterally positioned nephridia (see Figure 2.3 top left). The ventral domains were then defined as lying between the lateral domains and the ventral mid-line. The area of the CNS was subtracted from the ventral domains in order to examine ventral domain information without the CNS contribution and in order to reduce noise.

Nephridia: The nephridia are the segmentally-iterated excretory organs of the leech and as such serve the purposes of ridding the animal of waste products and maintaining water and electrolyte balance. Both the structure and transport mechanisms of the leech nephridia have been studied in some detail [63, 64]. Because the nephridia connect to the outside of the animal (through the nephridiopores), they can serve as pathways for bacterial or other microbial invasion, and some preliminary data suggests that cells in the nephridia may be expressing and releasing antibacterial peptides. We created a mask of the nephridia as shown in Figure 2.3 (top right).

2.3.4 Simulations

Currently, there are no standard datasets to assess performance, or standards to generate simulated data. To assess performance, we chose a mass value ($m/z = 5574.15$) that was significantly localized in the leech CNS. Our first simulation tests the performance of our method when the ROI signal decreases over the entire region. In this case, we simply decrease the intensities in the ROI spots by a certain percentage until the average intensity inside the ROI is the same as the non-ROI (or background) intensities. In Figure A.4, we see that the score decreases slowly at first, and starts dropping drastically once the average ROI intensity is less than about twice the average background intensity. Visually, we can see that the signal also starts dropping more rapidly around the same point. In the original image, the average ROI intensity is about 6.8 times the non-ROI average intensity, and the signal is very clear. In the second image, the ratio of average ROI intensity to average background intensity are approximately 2.74. While the signal is visually not as pronounced as in the original image, especially in the posterior ganglia, we can still see clear CNS expression and the score is still high ($s = 0.75$). However, in the third and fourth images (intensity ratios 1.87 and 1.58, scores 0.66 and 0.62), we observe a lower signal. We can also see a decrease of signal in the anterior ganglia between the two images. Finally, the last two images (ratios 1.29 and 1.0, scores 0.57 and 0.5) show almost no CNS localization. Again, similar intensities in the ROI and in the background lead to a score close to 0.5 as expected.

Our second simulation tests our method when degrading the signal in the ROI. In this case, we set a proportion of the ROI spots to have random non-ROI (or background)

intensities (see Methods). It is then possible to balance the total ROI intensities by distributing the subtracted intensity to the remaining spots; that way, the total intensities in ROI and outside ROI remain the same throughout the simulation. In Figure A.5, we show the results for two simulated runs for the two cases described above: with and without balancing the ROI intensities. In both cases, the score linearly decreases as more ROI spots are set to background intensity. In the balancing case, the intensities of the remaining spots increase to compensate for the other spots; consequently, the score remains higher in the balanced case than in the unbalanced case, as expected. Note that because the background intensities are set in a random manner, the results differ slightly for each run. This explains why the ending scores are different in the two runs. Visually, we can see that around $s = 0.65$, which indicates a 65% chance that a random ROI spot has higher intensity than a random non-ROI spot, the signal is still CNS-specific, but is targeted to a sub-region. When all spots are set to background intensities, the signal is lost and the probability score decreases to 0.5 as expected. The results show that the ρ statistic provides a direct interpretation of the strength of the ROI signal.

2.3.5 Over-expressed molecules in the leech

CNS: Any MALDI spot partially hitting the CNS is considered in the ROI. Even so, only 2.66% and 3.28% of the spots in LEECHE12A and LEECHE12B respectively were in our ROIs, and each ganglion had at most 2-3 laser spots 50% or more coverage by the ROI (see Figure A.2). Despite this challenge, we find that our method performs extremely well. We identified 43 m/z values that were significantly present in the CNS of LEECHE12A (score ≥ 0.65), which are listed in Table A.1. Visual inspection clearly demonstrates the power of the method as illustrated by the images for the top 12 most significant m/z values shown in Figure 2.2. We can see from these images that the ions corresponding to these m/z values are clearly more highly represented in the CNS and that specific expression can be detected even in the presence of other signal. For example, mass 8428.45 (case *c*) is detected as having significantly higher expression in the CNS even though it is expressed in another area in a dorsal posterior region. Likewise, we can see signal in other areas for cases *e*, *f*, *g*, *h* ($m/z = 2474.11, 2524.06, 9240.11, 3299.11$). Panels *i-k* are marked by lower (albeit significant) scores which

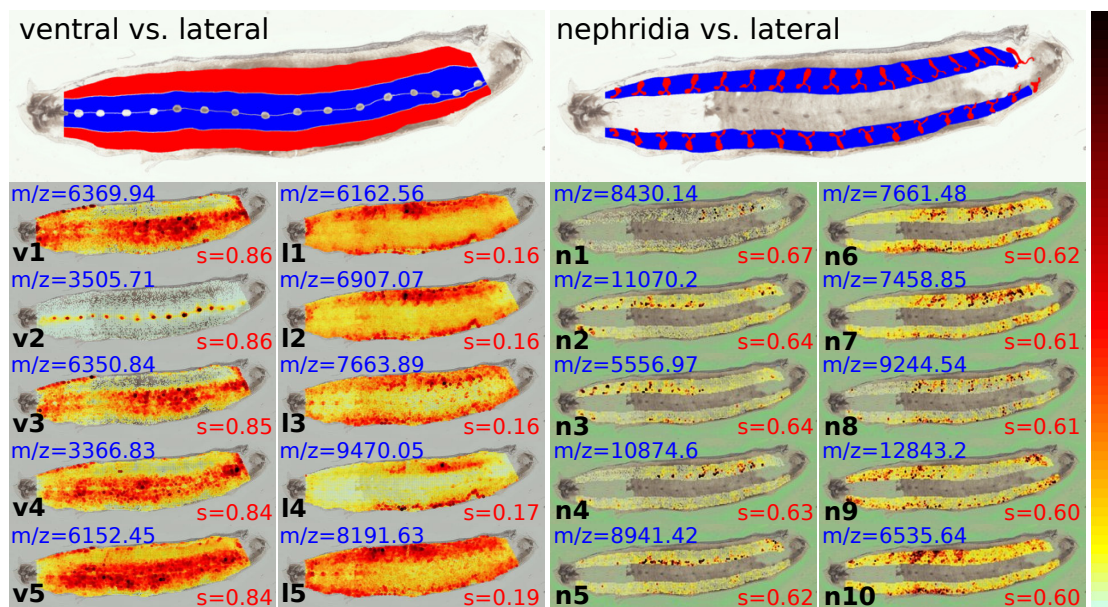


Figure 2.3: Mask and top results for LEECHE12A with other ROIs: (left) ventral vs. lateral query and (right) nephridia vs. lateral. For ventral vs. lateral query, top 5 high scoring images (v1-v5) and top 5 low scoring images (l1-l5) are shown. A high score indicates strong expression in the ventral region against the lateral region, while a low score indicates the inverse. In all images, there is a clear division between the two sections (ventral and lateral). For nephridia vs. lateral, top 10 results are shown. Scores are lower than for the other queries. The expression pattern is noisier and non-homogeneous.

corresponds to the decrease in CNS specificity. In those cases, not only is the noise higher in the rest of the leech, but the uniformity of the expression inside the CNS decreases. Specifically, we can see that fewer spots within the ganglia display strong expression while expression in the head ganglion increases compared to the rest of the ganglia. However, even in those cases, the CNS intensities are uniformly higher than non-CNS intensities.

The correlation between decreasing score and decreasing quality of CNS localization is also evident in the lower ranked masses. In Figure A.3, we show the expression pattern for 3 representative m/z values. The first image is taken for $m/z = 2797.28$ which was assigned a score of $s = 0.62$, just below our cut-off of 0.65. It still shows regional distribution specific to the CNS, but signal in other regions significantly impairs the ROI signal compared to the top-scoring images in Figure 2.2. In the second panel, the intensities outside the nervous system almost perfectly balance out the intensities within the CNS, and thus we get a score of 0.51, which represents no CNS localization. Finally, at the other end of the range, we can detect ions which have specific expression to outside the ROI. In the third image, at score $s = 0.23$, the molecule is highly expressed in the ventral region of the leech but shows distinct under-expression in the ganglia and the brain, displaying the inversed CNS expression pattern expected by such a low score.

Lateral-Ventral: In order to detect some of the potential signaling peptides that might be involved in the development of mechano-sensory arbors, we looked for m/z values expressed differentially between the ventral and the lateral regions by using our algorithm. Given that there are more spots in these ROIs than in the CNS ROI, and that they are more evenly distributed, we expected the algorithm to perform well for these masks. Indeed, we had many high-scored results for both ventral and lateral regions. Figure 2.3 shows the images for the top 5 highest scores ($v1-v5$) and for the top 5 lowest scores ($l1-l5$) when running the algorithm for the ventral region against the lateral region. A high score indicates strong expression in the ventral region against the lateral region, while a low score indicates the inverse. In all images, there is a clear division between the two sections (ventral and lateral) on the left side of the leech. The right side of the leech seems to have more noise, but the demarcation between the two sections is maintained

throughout the results. Interestingly, we pick out a nervous system signal at $m/z = 3505$. However, when looking at the image we can see that there is also a clear separation in signal between the ventral and lateral regions. This signal could be from a molecule that is expressed in both the ventral region and the CNS yielding a higher intensity in the CNS, possibly due to higher abundance there. Alternatively, the different signal intensities might reflect a mixture of two molecules with similar m/z values, one molecule producing a high intensity signal in the CNS and the other a lower intensity signal in the ventral region. When looking at the localization of the lower signal molecule, we can see the same intensity separation between the ventral and lateral regions.

Nephridia: When querying the nephridia against the rest of the leech, the top results were clearly in the nephridia alone; however, the fact that many masses are expressed more strongly in the lateral section than the ventral section caused some of the lower scoring results to be noisy. As a consequence, we queried the nephridia against the lateral section of the leech. The top 10 results are shown in Figure 2.3. It is important to note that the scores for nephridia are much lower than for the previous queries. In fact, only the top score $s = 0.67$ is above our threshold of $s = 0.65$. When looking at the images, we can see that even though there is nephridia localization, the expression pattern is noisy and non-homogeneous. For example, $m/z = 5557$ and $m/z = 10875.6$ are expressed more in the anterior and posterior sections of the leech, respectively. Interestingly, several masses show a co-localization with the nervous system ($m/z = 8429, 7663, 6535$) (data not shown), reflecting perhaps the accumulation of strongly expressed and secreted neuropeptides in the CNS and the secretory organs [65]. Besides detecting interesting masses for the ROI, we can also see regional differences in the expression of the molecules. Corresponding molecules may represent segmental functional differentiation.

Comparative analysis and reproducibility: To address the reproducibility of CNS specific mass values, we repeated the experiment in another leech. LEECHE12B, like LEECHE12A, is a 12 day embryo and expected to show similar distribution of peptides. It is worth noting that while the two samples were prepared using the same methods on comparable samples, spectra were acquired on a slightly greater m/z range and

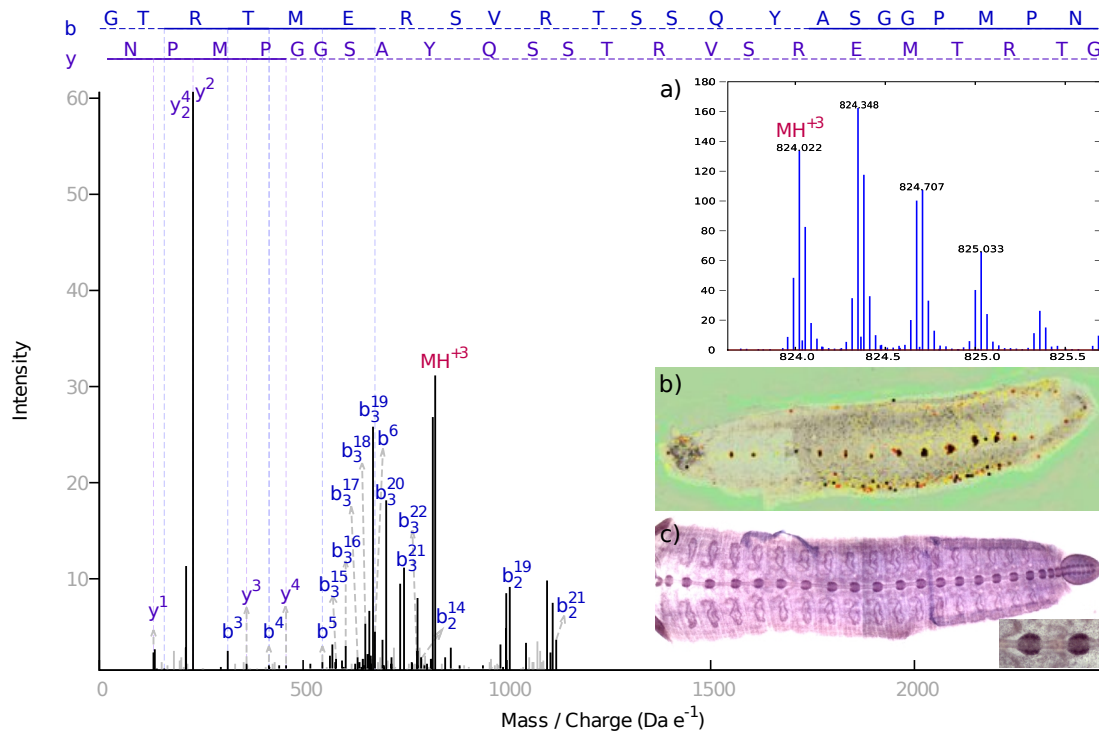


Figure 2.4: Annotated MS/MS spectrum for HmIF4, a novel peptide in the family of glial intermediate filament. The annotation explains 78.41% of the total peak intensity, with strong b and a fragment-ion series. Corresponding MALDI and *in situ* hybridization images both show CNS localization. MS1 (top right) has good isotope resolution. Charges for all fragments were manually verified by examining isotope patterns. Text annotation is available by request.

on a much tighter raster for LEECHE12B. In Figure A.2, we display the top scoring masses for the two samples. All high-scoring masses in LEECHE12B were found in LEECHE12A, but the inverse is not true. This is attributed to a lower overall intensity of the data in LEECHE12B, as can be seen by the high-scoring images from LEECHE12B. This difference in data quality is attributed to the difference in data acquisition parameters as mentioned above. Notwithstanding the low overall quality of LEECHE12B, the scores between the samples are comparable and reflect the true quality of the localization.

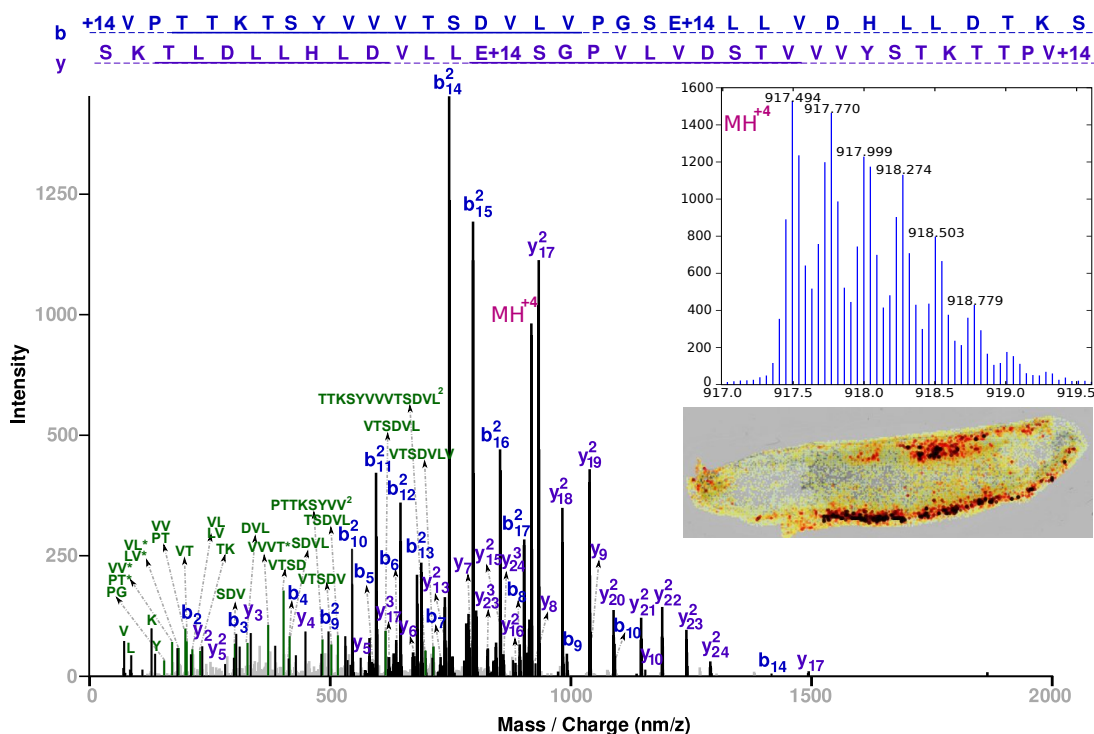


Figure 2.5: Annotated MS/MS spectrum for novel peptide with parent mass $\simeq 3666$ Da, targeted for being specifically expressed in the lateral/dorsal region. The entire annotation explains 78.92% of the total intensity, including a very strong 7-residue tag, [1089.62]VTSDVLV[1863.22], with a complete doubly-charged b ion series b_8 - b_{15} and a complete y ion series y_8 - y_{15} . A degenerate database search that matched Leucines with Isoleucine, and Aspartate against deaminated Asparagine did not find a match in our database. This suggests that this is a novel peptide, not represented in the LeechProtsDB database.

2.3.6 Peptide identification

Initial experiments were performed on extracts from 39 whole embryos. The samples were subjected to nano-LC-ESI-QTOF MS. We selected peaks corresponding to interesting imaging masses from the MS1 data and acquired MS-MS spectra in a data-dependent manner or semi-data-dependent manner (see Methods). We increased acquisition time in order to acquire high-quality spectra of lower abundance peptides. In Figure A.9, we can see that two of the identified spectra had lower MS1 intensity counts than many of the other molecules present in the sample. No trypsin digestion was performed so we could use the peptides' parent masses to link the MS/MS data to the MALDI imaging data which captures endogenously processed peptides. Identification of intermediate sized endogenous peptides is difficult due to a limited understanding of fragmentation chemistry and high charge on fragment ions. To overcome these issues, we developed a novel peptide identification pipeline, based on construction of sequence tags from both original and deconvoluted spectra, and a database search including modifications. We identified a number of peptides (see Table A.3). A few are described below.

We first identified a candidate sequence for parent mass $\simeq 2474$ Da that was specifically expressed in the CNS. In Figure 2.4, we present the annotated spectrum for the peptide that was derived from an EST transcript in our LeechProtsDB database, as well as the corresponding MALDI image showing CNS specific expression. The annotation explains 78.41% of the total peak intensity, with strong b and a fragment-ion series. As expected, we observe a very strong peak for fragments at the N-terminus of the prolines. It is also worth noting that a large part of the *b* ion series, namely *b*₁₅-*b*₂₂, is built on validated charge 3 fragment ions, which makes identification of this peptide difficult using standard tools.

A BLAST search of the EST sequence against the NCBI nr database established that the peptide came from a previously unreported protein. The strongest similarities (but not identity) are all to intermediate filament proteins, and specifically to gliarin, macrolin and filarin, the three known intermediate filaments in *Hirudo medicinalis* [66], which have been described as important in neuronal development in leech. Thus, we believe to have found a novel member of the family of intermediate filaments, which

we call HmIF4. We aligned the four protein sequences using ClustalW 2.0.12 [67], as shown in Figure A.6. The EST open reading frame aligned particularly well in the conserved rod domain, and has more variability outside of that domain. The peptide we identified is located in a variable region in the 5' end of the rod domain where the sequences are quite dissimilar, thus confirming the discovery of a novel protein. Finally, we examined the distribution of the HmIF4 transcript using *in situ* hybridization (see Methods). In Figure 2.4c, we can clearly see that the mRNA is indeed preferentially expressed in the CNS. The concordance of the spatial distributions at the peptide and corresponding mRNA strongly support the specificity of expression, and also suggests that differential gene expression, not protein targeting, is the reason for the spatial distribution.

A second peptide, with parent mass $\simeq 3666$ Da, was targeted for being specifically expressed in the lateral/dorsal region. The identified sequence, is noted as

[201.13]TLTVV[277.15]VVTSDVLV[258.18]VLDTTD[976.55]

and explains 78.92% of the total intensity. The annotated spectrum is shown in Figure 2.5. Moreover, most of the rest of the intensity can be explained by internal ions from breaks at the dominant peaks. Out of this long annotation, we have a very strong 7-residue tag, [1089.62]VTSDVLV[1863.22], with a complete doubly-charged *b* ion series *b*10-*b*17 and a complete *y* ion series *y*17-*y*24. There is also a partial *a*-ion ladder supporting the direction of this tag. On its own, the tag explains 67.3% of the spectral intensity. A BLAST search of the protein sequence did not get any good hits suggesting that this is a novel peptide. Another spectrum of the same peptide, unmodified and 2 amino acids longer confirms the identification. The parent mass of this peptide, $\simeq 3841$ Da, also shows dorsal localization (see Figure A.8).

A third molecule (parent mass $\simeq 2500$ Da) was shown to have a CNS specific expression. A database search identified the peptide LPGELAKHAVSEGTKAVK-TYTSSK, which is part of the histone H2B in a related species, *Helobdella robusta* (Figure A.7). Histones, which are part of the DNA packaging complex, are highly conserved. Indeed, a BLAST search [68] of the translated EST sequence against the NCBI nr database, returned complete perfect matches to 179 sequences in many different species. Therefore, it is very likely that the peptide is conserved between the

sequences in *Helobdella robusta* and *Hirudo medicinalis*. Regarding CNS localization, it is worth noting that Shimma et al. [50] have previously identified a histone H2B expressed in the mouse brain using MALDI imaging. *In situ* hybridization of the mRNA again shows a preferential location in the CNS, but with a relatively weaker signal (see Figure A.7).

2.4 Conclusions

Recent years have seen a tremendous improvement in instrumentation for mass spectrometric imaging employing MALDI, DESI and SIMS techniques [69]. MALDI MSI is particularly useful for the study of the tissue distribution of biologically interesting molecules because it affords both access to large range of intact molecules and a relatively higher spatial resolution when compared to other ion sources. MALDI MSI is therefore the approach of choice when studying tissue distributions of larger molecules, such as peptides or proteins.

While MALDI imaging offers great advantages for detecting and mapping unknown molecules in their native, processed state, it does have some important physical limitations. For example, despite recent and potential further improvements, MSI cannot achieve either the level of detectability, single or a few molecules, or the spatial resolution of conventional light microscopic cell imaging techniques. The pixel resolution obtained with MSI, as reported in most published studies, is $\sim 50\text{-}300\mu\text{m}$, though a possible lower limit of $\sim 5\mu\text{m}$ for more abundant species has been reported [69]. In comparison, resolution of 200nm is achievable using laser scanning confocal microscopy of cells immunostained with fluorescent tags [70].

The advantages of MSI, then, are twofold: first, a pixel is not simply a pixel, but a complex array of mass values that can be resolved to high-accuracy. Second, MSI allows for an unsupervised (label-free) interrogation of the sample, allowing for the discovery of previously unknown species that are active in specific spatio-temporal contexts. The approach we report, referred as MSI-query, addresses and exploits these two facets, developing a novel analysis methodology.

As noted above, a hurdle in the assessment and analysis of the large amounts of

data inherent in MSI is to determine which of the many masses represented in the spectra are worth pursuing, given that the identification of the corresponding protein requires a great investment in time and effort. In our approach, we have started with the premise that the topographic distribution of a particular m/z value can be a first order filter for selecting those molecules of particular interest. Thus, we first developed a statistical technique to identify mass values that are specifically expressed in a morphological region specified by the user. Using this constraint, we then obtained a collection of mass values (presumably endogenously processed peptides or proteins) that are specifically expressed in the CNS, nephridia, and ventral/dorsal segments of the medicinal leech embryo, the model system we used to test our technique. In order to obtain amino acid sequence information for less abundant species, we decided to use a procedure for identifying the peptides/proteins corresponding to these interesting masses from secondary fragmentation (MS/MS) data that required decoupling the imaging and MS/MS [71] performed with LC-MS separation of non-digested proteome extracts. While other approaches obtain sequence information by performing MS/MS directly from the tissue while maintaining spatial information, our method allows for the concentration of peptides, helping identify peptides and proteins with intermediate abundance [55].

Our approach does have some shortcomings that should be noted. The identification of endogenous peptides based on fragmentation of intermediate sized, highly charged precursors is challenging given available tools. We developed a customized pipeline for identification. As a second issue, the link between MS/MS and MSI parent masses is tenuous due to the lower accuracy of mass resolution in MSI. However, we test our results by using a second independent validation through *in situ* hybridization of the identified mRNA. While co-localization of the ISH and MSI signals can provide only a measure of consistency, it may also lead us to interesting differences between mRNA and protein localization that can be further explored.

Initial tests of our methods on the leech embryo MSI data have thus far resulted in the identification of a few novel proteins, including a member of the intermediate filament (IF) family and a completely novel peptide sequence. The discovery of a new IF expressed by neurons in the leech CNS is also of significant biological interest. IFs form a diverse family of proteins important for cytoskeletal architecture. Invertebrate IF

proteins are relatively less analyzed and might be evolutionarily and functionally distinct from their vertebrate counterparts. The three known IFs in leech have distinct patterns of expression. The expression of macrolin is limited to macroglia, gliarin is expressed in both glial, and macroglial cells, and filarin is selectively expressed in neurons [72, 66]. While their function is poorly understood, the neuronal IFs are suggested to be developmentally regulated, and may be involved in stabilizing the neural cytoskeleton. Our discovery of a novel IF protein adds to the diversity of invertebrate neuronal IFs.

Our results also include the detection of CNS expression of a fragment of histone H2B in early leech development. Interestingly, although histones are mainly known for their essential roles in chromosome packaging, histone H2B has been reported previously to be localized to the mouse brain [50]. Moreover, recent studies in *Drosophila* have suggested that the specific targeting of some axons (R1-R6) in the optic ganglia is mediated by the selective deubiquitination of the fly ortholog of histone H2B [73, 74]. Further, the deubiquitination is mediated by the SAGA complex, which has analogs from yeast to human [73, 74]. The discovery of these and other peptides using MSI shows the power of mass spectrometric imaging in a label-free identification of spatially differentiated proteins.

2.5 Acknowledgements

This chapter, in full, was published as “Automated querying and identification of novel peptides using MALDI mass spectrometric imaging”. Bruand J, Sistla S, Mériaux C, Dorrestein PC, Gaasterland T, Ghassemian M, Wisztorski M, Fournier I, Salzet M, Macagno E, Bafna V. *J Proteome Res* 10(4):1915-28 2011. The dissertation author was the primary author of this paper.

Chapter 3

AMASS: Algorithm for MSI Analysis by Semi-supervised Segmentation

Mass Spectrometric Imaging (MSI) is a molecular imaging technique that allows the generation of 2D ion density maps for a large complement of the active molecules present in cells and sectioned tissues. Automatic segmentation of such maps according to patterns of co-expression of individual molecules can be used for discovery of novel molecular signatures (molecules that are specifically expressed in particular spatial regions). However, current segmentation techniques are biased towards the discovery of higher abundance molecules and large segments; they allow limited opportunity for user interaction and validation is usually performed by similarity to known anatomical features.

We describe here a novel method, AMASS (Algorithm for MSI Analysis by Semi-supervised Segmentation). AMASS relies on the discriminating power of a molecular signal instead of its intensity as a key feature, uses an internal consistency measure for validation, and allows significant user interaction and supervision as options. An automated segmentation of entire leech embryo data images resulted in segmentation domains congruent with many known organs, including heart, CNS ganglia, nephridia, nephridiopores, and lateral and ventral regions, each with a distinct molecular signature. Likewise, segmentation of a rat brain MSI slice dataset yielded known brain features, and provided interesting examples of co-expression between distinct brain regions. AMASS represents a new approach for the discovery of peptide masses with

distinct spatial features of expression.

Software source code and installation and usage guide are available at <http://bix.ucsd.edu/AMASS/>.

3.1 Introduction

The use of multiple imaging techniques to assess the presence and location of specific proteins in tissues and cells is central to the study of biological systems. Historically, successful approaches usually involved labeling one/few proteins at a time either by attaching a fluorescent domain genetically or by treating a biological sample with labeled antibodies, and then recording two-dimensional (2D) micrographs of the sample, possibly also reconstructing them into a three-dimensional (3D) object or movie. Such imaging techniques are low-to-medium throughput approaches and give the biologist insight into just a small number of biological samples, limited to known proteins for which antibodies or tagged forms are available. By contrast, there is an increasing number of imaging technologies (transcriptomic or proteomic) that allow for the sampling and exploration of the entire complement of active molecules in the cell.

Mass Spectrometric Imaging (MSI) is a molecular imaging technique which allows the generation of 2D ion density maps for a large complement of the molecules present in the tissue under study [37]. In the Matrix-Assisted Laser Desorption / Ionization (MALDI) MSI workflow, thin tissue sections (10-15 μm) from organs, or even whole dissected specimens, are mounted onto a transparent glass slide, allowing microscopic observation of the material prior to MS analysis. After deposition of the MALDI matrix, automated direct MALDI analysis of tissue sections provides information on masses of the desorbed molecules in a 2D raster defined by the selected positions of the laser beam [38]. The studies performed by various groups [32, 33, 34, 35, 36] have demonstrated that acquisition of tissue expression profiles while maintaining cellular and molecular integrity is feasible. With automation and new analysis software, it has also become possible to produce multiplex imaging maps of selected bio-molecules within tissue sections [37, 38, 39]. Molecules that are preferentially expressed in a region of the sample will show higher intensities in that region when looking at the image

corresponding to the specific m/z value associated with the molecule. Discovery of these molecules often involved observing the images for each mass value sequentially in a movie, to short-list ones with interesting patterns.

Most bioinformatics approaches have focused on making the discovery process easier by allowing computational queries of MSI data-sets. In previous work [75], we started with a supervised approach in which we assumed that the *region of interest* (ROI) is specified based on pre-selected morphological criteria. As an example of an ROI, consider the central nervous system (CNS) of the medicinal leech, *Hirudo medicinalis*, one of the best-studied representatives of the phylum Annelida (segmented worms). Given a particular ROI, we asked if (a) there were specific molecular signatures or collections of peptide mass values that are specific to the ROI; and, (b) which peptides correspond to these masses. We identified molecular signatures for many ROIs, including 43 m/z values in the CNS, and identified 35 peptides, one of which was a novel member of the intermediate filament family (which we named HmIF4), which appears to be involved in neural development.

By contrast, unsupervised approaches (no pre-specified ROI) seek to computationally segment (or partition) MALDI spots into regions, each characterized by a specific *molecular signature* or profile. In most cases, the idea is to treat each MALDI spot as a vector of expressed masses, and to apply unsupervised clustering techniques for segmentation. Principal Component Analysis (PCA) and hierarchical clustering (HC) are classic non-parametric clustering techniques, and have been used successfully for MSI [40, 41, 76, 16]. Alexandrov et al. argue that these methods do not take advantage of the spatial clustering of MSI spots and develop a technique based on edge detection and smoothing [77]. While these clustering-based methods show promising results, they need to be optimized both in memory and runtime to be able to process the full MSI datasets which are typically large. For example a dataset acquired on 20000 MALDI spots with 40000 m/z values for each spectrum yields a dataset of 800 million values (3.2GB). Typically, MSI datasets are reduced for processing by decreasing mass resolution [40, 78], by applying a discrete wavelet transform [79] to each spectrum, or by explicit peak selection on each spectrum [78, 77]. Normally, the peak-picking is performed at a pre-processing stage in a spectrum-wide manner based only on the intensity

and the shape of a potential peak. If a region of interest is characterized by a single (or a few) peaks that are not among the most intense peaks in a region, these peaks may be omitted during peak-picking, making the region indistinguishable from others. The standard clustering approaches also does not rely on any a priori knowledge about tissue morphology. Finally, unsupervised clustering-based segmentation methods are useful but limited in providing a user an opportunity to go deeper into the data analysis. Most significantly, we find in our investigations that segments overlap because they share peaks so it is important to allow the user to make a reasoned choice.

In this paper, we address these issues explicitly. We start with the difficult question of what constitutes a ‘good’ segmentation. Prevailing methods implicitly equate good segmentation to ones that match known morphological features of tissues observed through optical methods [40, 76, 77]. While this validation is natural and provides direct visual feedback – indeed we use it as one technique in this report (see Figures 3.4 and 3.5) – it has problems. Often, molecules are expressed in multiple, morphologically distinct, regions. Segmenting images so as to conform to known morphology will inhibit the discovery of novel molecular signatures. Second, MSI resolution (20-70 μm) is still inferior to optical resolution ($< 1\mu\text{m}$). The potential of MSI is not as a replacement of optical methods, but to help identify the molecular basis of morphological differentiation. Therefore, we judge image-segmentation quality with alternative criteria based on molecular signatures.

A key finding of our previous work [75] was that, given a region of interest (ROI) defined by an image-segment (or collection of MSI spots) I , we usually obtain a strong molecular signature for I , a collection of mass values that are preferentially expressed in spots in I . Then, the spectrum of each spot s can be compared to the molecular signature associated with I . We use this idea to judge the quality of segmentation. Informally, a segmentation is *consistent* if each segment I has a unique molecular signature that is shared with all spots in I and not with other spots. This consistency measure is independent of morphology, and allows us to discover signatures that cross known morphological boundaries.

Using the molecular signature defined by I as a ‘query’, we can recruit other spots to the segment, refining the segmentation. Our method is reminiscent of iterative

unsupervised clustering methods, like a K -means clustering. It starts by choosing an initial segmentation, each with a molecular signature (or ‘center’). Subsequent iterations repeat two steps: (a) each spot is assigned to the nearest of the K signatures (based on a query) and, (b) K new signatures are described from the recruited spots. Earlier methods consider each MALDI spot as a vector of intensities over mass-bins, causing the clustering is dominated by high intensity peaks. This has been typically circumvented by using scaling techniques, such as autoscaling, which have their own problems. We propose a different representation of each spot. Starting with a current image-segmentation \mathcal{S} , each spot is represented as an $|\mathcal{S}|$ -dimensional vector of query-scores to each of the segments in $|\mathcal{S}|$, where $|\mathcal{S}|$ is the number of clusters in the segmentation. Thus two spots are similar if they have similar scores against all clusters. To start the algorithm we need an initial segmentation. In our case, the initial segments can be chosen at random, or by partial user-input (semi-supervised). The initial segments are chosen to be small groups (only a few) of contiguous spots, but otherwise no spatial correlation is assumed.

In summary, three ideas describe AMASS (Algorithm for MSI Analysis by Semi-supervised Segmentation). (a) Rank based statistics are a useful discriminator for any current cluster, and this allows us to query. (b) Query-result consistency is a valid score for the validity of a cluster. (c) The scores of a spot against existing clusters can be used to compare and re-partition spots. In addition, we make available a computational tool implementing the algorithm which allows many other controls for user intervention.

We applied AMASS on multiple datasets, including a leech embryo dataset obtained from a 12-day (E12) specimen that was dissected and prepared flat before mounting on the MALDI target, and a dataset of a rat brain coronal section of 4.16 mm from Bregma with known anatomical structures. We show in the detailed results below that, in each case, a completely automated run provided fine-grained, biologically meaningful segmentations and their molecular signatures. The leech dataset was segmented into regions corresponding to head, tail and segmental ganglia of the central nervous system, nephridia, heart, and lateral and ventral regions. The rat brain dataset was segmented into many domains corresponding to well-defined anatomical regions, with some signatures corresponding to co-expression of molecules in distinct morphological regions.

3.2 Results

3.2.1 AMASS: Algorithm for MSI Analysis by Semi-supervised Segmentation

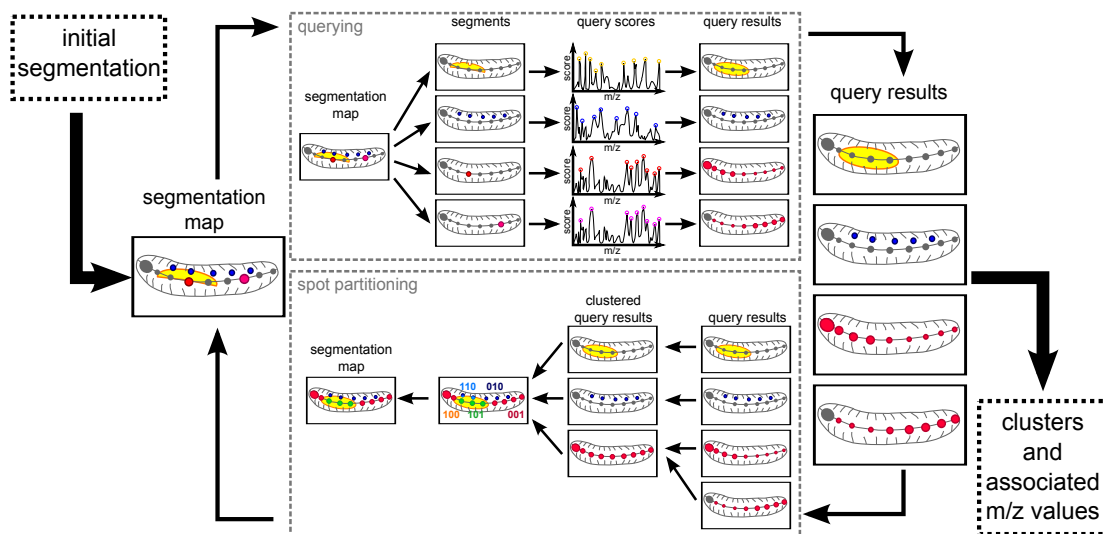


Figure 3.1: Main workflow overview. First, we need an initial image-segmentation, which can either be defined randomly or by the user. In the querying component, each of the segment from the image-segmentation is used as a query and top-scoring m/z peaks are retained. A log-odds score is calculated for each spot and each query; this score represents the likelihood of a spot belonging to that query. The resulting set of scores per query forms a set of query-results. These are used as input to the spot partitioning component. In this component, the highly similar query-results are clustered together. We then obtain binary signatures for each of the spots and retain the dominating ones as cluster centroids. Clustering the all spots to the closest centroid results in a new image-segmentation. The whole process can be run iteratively until the quality of the segmentation is satisfactory.

The input to AMASS is a set of MSI spots S . Each spot in S is defined by a spectrum: a collection of m/z values and associated intensities. Define an *image-segment* I simply as a collection of spots. An *image-segmentation* $\mathcal{S} (= \cup I)$ of an MSI data-set is an incomplete partitioning of the spots into image-segments. By incomplete, we mean that each spot is assigned to at most one image-segment, but could be assigned to none. The output of AMASS is a segmentation $\mathcal{S} = \cup I$ into consistent segments such that most spots are assigned. AMASS works with an iterative refinement of segments.

procedure $\text{AMASS}(S, \text{spectra}) \rightarrow \mathcal{S}, \mathbf{A}$, molecular signatures

1. Select an initial image-segmentation \mathcal{S} , chosen either by the user, or via random spot selection.
2. Repeat until $(|S| < \epsilon)$
 - (a) Calculate $\mathbf{A} = \text{Query}(\mathcal{S})$.
 (* $\mathbf{A}[I, s]$ denotes the score for spot s against each segment $I \in \mathcal{S}$ *)
 - (b) For all *consistent* segments $I \in \mathcal{S}$
 (* see Methods 3.4.4, equation 3.10 for definition of consistency *)
 - Output I ; Set $S = S \setminus I$.
 (* A consistent spot is fixed and output *)
 - (c) Set $\mathcal{S} \leftarrow \text{Spot-partition}(\mathcal{S}, \mathbf{A})$.
 (* Recompute non-consistent segments based on scores in \mathbf{A} *)

In practice, we iterate for a small number of rounds before terminating. The three main steps are a choice of Initial segmentation, the Query procedure, and the Spot-partition, and these are described below, along with results.

3.2.2 Initial Segmentation

The initial segmentation can be done in either a guided mode or in a blind mode. In a guided mode, the user provides the initial clustering. Typically, it is a list of regions of interest (ROIs) for which he/she would like to get additional information with spots outside the ROIs unassigned. Examples of guided initial segments are shown in Figure 3.2. The semi-supervised component of the algorithm will then return additional information about the segments or ROIs, specifically which areas have similar molecular signatures as well as the actual molecular signatures. In a blind approach, the algorithm automatically generates a large set of small random seed clusters. Subsequently, it merge and expand the appropriate seeds. An example of such random segmentation is shown in Figure B.2a.

3.2.3 Querying

The goal of querying is to compute $\mathbf{A}[I, s]$, a log-odds measure of similarity between the spectrum at a spot s and the molecular signature of spots in segment I . Denote the MALDI spectrum (m/z values and intensities) associated with spot s by a vector of intensities \mathbf{v}_s ; $\mathbf{v}_s[m]$ is the intensity at m/z value m (Methods 3.4.2). We use the following steps.

1. For each m/z value m and segment I , compute weight \mathbf{w}_I , with $\mathbf{w}_I[m]$ describing the ‘importance’ of m in discriminating I from $S \setminus I$ (Methods 3.4.2, equation 3.6).
2. Compute a weighted-intensity $\mathcal{Z}(I, s) = \mathbf{w}_I \cdot \mathbf{v}_s$.
3. Optionally, smooth the weighted intensities image.
4. Compute $\Pr(s \in I)$, and $\Pr(s \notin I)$ using the distribution of $\mathcal{Z}(I, s)$ over spots in I and $S \setminus I$, respectively (see Methods 3.4.2, equations 3.2 and 3.3).
5. Set $\mathbf{A}[I, s] = \log \left(\frac{\Pr(s \in I)}{\Pr(s \notin I)} \right)$

To showcase AMASS’s ability to work with user defined queries (initial segments), we prepared queries informed by our knowledge of morphology. However, the queries were *not* precisely defined, as seen in Figure 3.2a. For example, ventral and lateral regions were defined by simple lines (for anterior, central and posterior) across the corresponding sections, while three of the ganglia were queried independently.

For each query I , we show three consecutive images in Figure 3.2a. The first two panels correspond to weighted-intensities $\mathcal{Z}(I, s)$ (before and after smoothing) and the third panel corresponding to the log-odds score $\mathbf{A}[I, s]$ computed as above. In every case, the scored images all highlight exactly the areas we would expect to see, illustrating the power of querying. Queries that are fairly complete, such as the skin, essentially recapture the region of the original query. Partial queries, such as the three single ganglia, each recover the entire central nervous system.

Figure 3.2b shows the advantages and costs of smoothing. The granularity inherent in MALDI imaging data is reduced by smoothing allowing for evenness in spot to spot weighted-intensities. Larger regions, such as the ventral central region, benefit by coalescing disjointed spots. This allows us to define unified regions in different section of the leech. However, very small and finely defined regions, such as the nephridiopores,

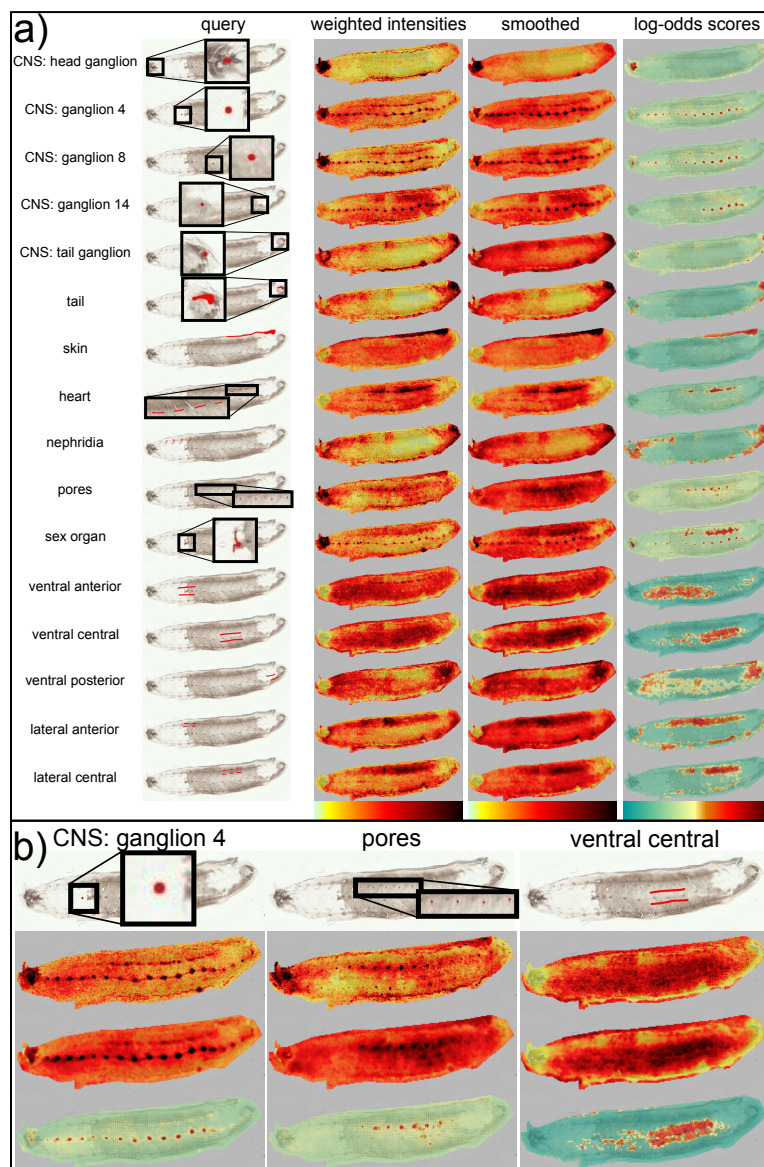


Figure 3.2: (a) List of queries and their associated results. Shown are on each row the original query, the corresponding weighted intensities image, the smoothed weighted intensities image and the log-odds scores image. Querying with specific image-segments results in the recruitment of other spots with similar molecular signatures. For example, querying with one ganglion or a few pores recruits the whole CNS or the rest of the pores respectively. (b) Detailed images for 3 different queries. We can see that while smoothing helps in cleaning noise on larger queries such as the ventral query, it can also cause the loss of some MALDI spots in the case of smaller regions, such as the pores.

lose in accuracy and localization. While we can see higher intensities in almost all the pores throughout the leech in the non-smoothed image, only the highest intensity pores are detectable in the smoothed image. We can also notice some diffusion of the signal in the CNS after smoothing. Spatial smoothing is an important part of some MALDI imaging analysis tools [77]. While AMASS provides smoothing as an option to reduce granularity, it is not used in our final segmentation results.

In the log-odds score images, we contrast the negative scores and the positive scores by showing them in green and red respectively (see scale in Figure 3.2). Thus, dark red spots in these images represents spots with molecular signatures very similar to that of the original query, and thus are recruited by the query, while dark green spots represent spots that are very unrelated. Partially related spots typically obtain scores closer to 0 (shown in pale yellow to light orange). For example, querying with the ventral posterior region (Figure 3.2b), expectedly results in partial recruitment across the entire ventral region including the ventral anterior region, with the highest scores in the ventral posterior regions. Thus, while the automated segmentation chooses a score threshold based on the distribution of the scores in the original query (see Methods 3.4.3), we make this an adjustable parameter.

Random Queries: While the algorithm is designed to let the user guide the study by choosing initial segments, choosing random spots as initial queries also results in a remarkably high quality segmentation. In Figure B.1, we show several examples of random seed-segments and the corresponding query-results, which are very similar to user-defined queries from the same morphological region (such as the CNS). In addition, query-results gain specificity in the next iteration as the new queries are based on molecular signatures found from each current iteration. Regions that are only defined by a few spots, such as the pores, are less likely to show on every random run; however, in general, several runs of the algorithm on a random seeding eventually find that region (data not shown).

Molecular signatures: In Figure 3.2a, one can observe that while a query consisting only of ganglion 4 recruits the entire CNS, more or less evenly, the query-results associated with ganglion 14 show stronger association in the more posterior ganglia. Thus,

there are some differences in the molecular signatures associated with these queries in different regions from the same gross morphological feature (i.e., CNS). This specific case can be attributed to the rostrocaudal gradient in leech embryo development [80]. The head of the embryo is ~ 3 days older than the tail, and thus the ganglia may show different protein expression depending on their relative “age”. It is also possible that the differences reflect the innervation of different organs along the rostrocaudal axis.

As AMASS is a query/molecular-signature based segmentation, we can easily extract the molecular signatures associated with the query. As a test, we chose anterior ganglia 2-4, and posterior ganglia 13-15, and extracted differentiating score peaks from the querying module. In Table B.1, we show the score peaks with weight greater than 0.7. Note that these are the weight associated to the m/z value, and not the rank statistics. While many of the m/z values show expression throughout the entire CNS, such as $m/z \simeq 2524$ and $m/z \simeq 5418$, some m/z values show a bias in intensities between the anterior and posterior regions. For example, at $m/z \simeq 3299$ and $m/z \simeq 5273$, high intensities values are present in ganglia 1-10 and 1-12 respectively but not in the rest of the CNS. On the other hand, at $m/z \simeq 4377$, high intensities are prevalent in posterior ganglia (8 – 14), but not anterior ganglia. These molecules will be prime candidates for targeted identification of peptides involved in specific stages of the leech neuronal development. In previous work [75], we identified one of the molecule in the table ($mz \simeq 2474$) which shows expression in both the anterior and posterior ganglia as a peptide from a novel gene, HmIF4, in the family of neurofilaments. Similar targeted identification can be done to target peptides for m/z values specific to anterior or posterior ganglia.

AMASS iteratively improves segmentation in a way that will create distinct molecular signatures for each segment. To test the signature strength of specific molecules, we observed the top 20 score peaks at least 10 Daltons apart for segments at successive iterations in leech and rat respectively (Figures B.3 and B.4). In both cases, we can see that the peaks are overall conserved throughout the iterations. However, there are some changes from one iteration to the next. For peaks $m/z \simeq 3508$, 5417, 5570, we find that the weights increase with number of iterations while in peaks at $m/z \simeq 3295$ and 4007, the weight is high for the ganglia 5-6 initial segment, much lower in iterations

1 and 2, and not even in the top peaks for the ganglion 14 initial segment. These changes happen as the entire CNS is recruited to a segment starting with a single ganglion, and can be explained by observing the intensity images in Table B.1. Peaks $m/z \simeq 3508, 5417, 5570$ show high intensity throughout the CNS; peaks at $m/z \simeq 3653, 4377, 8564$ show up in posterior ganglia, but not in the anterior ones. While $m/z \simeq 8526$ shows up as expressed in both, its intensities are high in ganglia 2-4 but lower in ganglia 5-6. Thus the contribution of individual peaks to the molecular signature, rises and falls with its expression in the segment, and allows for a fine grained exploration.

Molecular signatures for different regions of the rat brain also show interesting patterns (Tables B.2, B.2, B.3, B.4, B.5, B.6, B.7) as well as the corresponding m/z images in Tables B.3 and B.4). We observe that several of the m/z values specifically expressed in the piriform cortex also show expression in the CA1-CA3 cell bodies, the CA3 cell bodies and the dentate gyrus ($m/z \sim 3454, 6223, 6272, 6646$ in Table B.3). Reciprocally, when querying the CA3 cell bodies, we find many of the same m/z values that also show expression in the piriform cortex ($m/z \sim 6226, 6275, 6648$). However, the two queries do not share all peaks. There are many peaks from the piriform cortex query which do not show expression in the CA3 cell bodies, and there is peak which shows very strong signal in the CA cell bodies ($m/z \sim 8447$) but no signal in the piriform cortex. The molecular relation between the two areas may be due to both containing apical dendrites of pyramidal neurons which are located in these regions. These shared peaks, illustrate the need for a tool that allows exploration of different segments, instead of a ‘black box’ approach to segmentation.

3.2.4 Spot Partitioning

Hierarchical clustering of query-results: The result of the querying component is a matrix $\mathbf{A}[I, s]$ which contains the log-odds score of each spot s against each segment-query I . Each row of the matrix represents the result of querying a segment I , while each column is a vector of scores against each segment for a spot s . In Figure 3.3, we show the resulting matrix from querying the previous initial segments on the leech dataset, with scores encoded in a green-red color map. Spots are sorted by (x,y) coordinates; thus they are ordered from the top-left spot to the bottom-right spot, scanning vertically

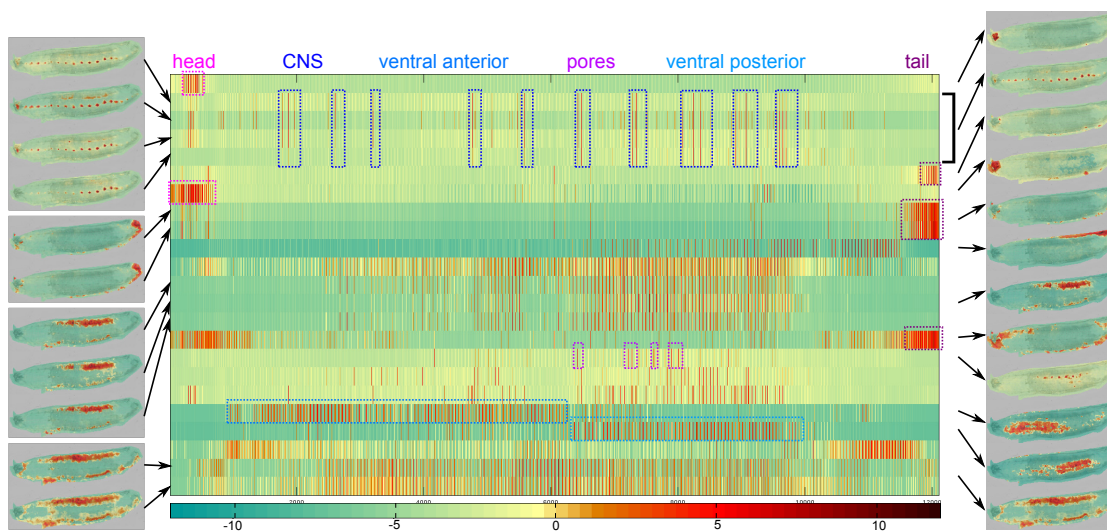


Figure 3.3: Log-odds score matrix and hierarchical clustering. Each row of the matrix represents a query-result, with some of the corresponding log-odds images shown on the left-hand side. Spots are sorted by (x,y) coordinates; thus they are ordered from the top-left spot to the bottom-right spot, scanning vertically from left to right. When looking at the columns of the matrix, we can see high-scoring columns throughout several rows corresponding to specific morphological features, such as the ganglia in rows 2-5. Certain rows of the matrix also show very high similarity. These rows are clustered together and the result clustered query-result image is shown on the right-hand side. Rows (or query-results) that do not show high similarity to other rows end up in singleton clusters.

from left to right. When looking at the columns of the matrix, we can find columns with high scores throughout the same query-results, corresponding to certain morphological features. For example, the first four score images in the left column show higher log-odds scores in the CNS. The corresponding rows (2-5) show several bundles of vertical red lines (highlighted in the figure) which are consistent throughout the 4 rows and represent some of the ganglia. Some of the anterior ganglia do not show as strong scores in row 5, consistent with the image.

When looking at either the log-odds score images or the corresponding rows, we can see that the query-results from different segments are often very similar. This is expected as disjoint segments from the same morphological feature will have similar molecular signatures and thus MALDI spots will have similar scores against these segments. To merge these query-results, we perform hierarchical clustering on the matrix rows, or query-result vectors $\mathbf{A}[I, *]$ (Methods 3.4.3, equation 3.7), using the Tanimoto coefficient as a distance measure [81]. Here, we cluster to a Tanimoto coefficient of 0.65, but empirically AMASS is robust to a large range of thresholds. The left-side of Figure 3.3 shows images for query-results, while the right-hand side shows the clustered results (with mean scores). Regions that covered the same morphological features, such as CNS or lateral, ended up as one cluster, while regions that are only partially similar, such as full-lateral vs. posterior-lateral, remain separate. Some rows, such as the pores or the ventral regions, do not cluster with any other query-results and are shown as clusters of size 1 on the right-hand side.

Binary Signatures and Spot-partitioning: While the query-results clustering is robust, we expectedly find overlapping regions in the clustered query-results (Figure 3.3). For example, while some query-results cover the entire lateral region (last image on right-hand side), others cover only the posterior lateral region (7th image on right-hand side). This means that most spots in the posterior lateral region have high scores against two clustered query-results. Recomputing the segmentation involves clustering the spots that have similar pattern of scores across the current set of segments I . We do the following (also see Methods 3.4.3):

1. Set $\mathbf{B} = \text{Binarize}(\mathbf{A})$. Each distinct binary column \mathbf{b}_s is a binary signature. (See

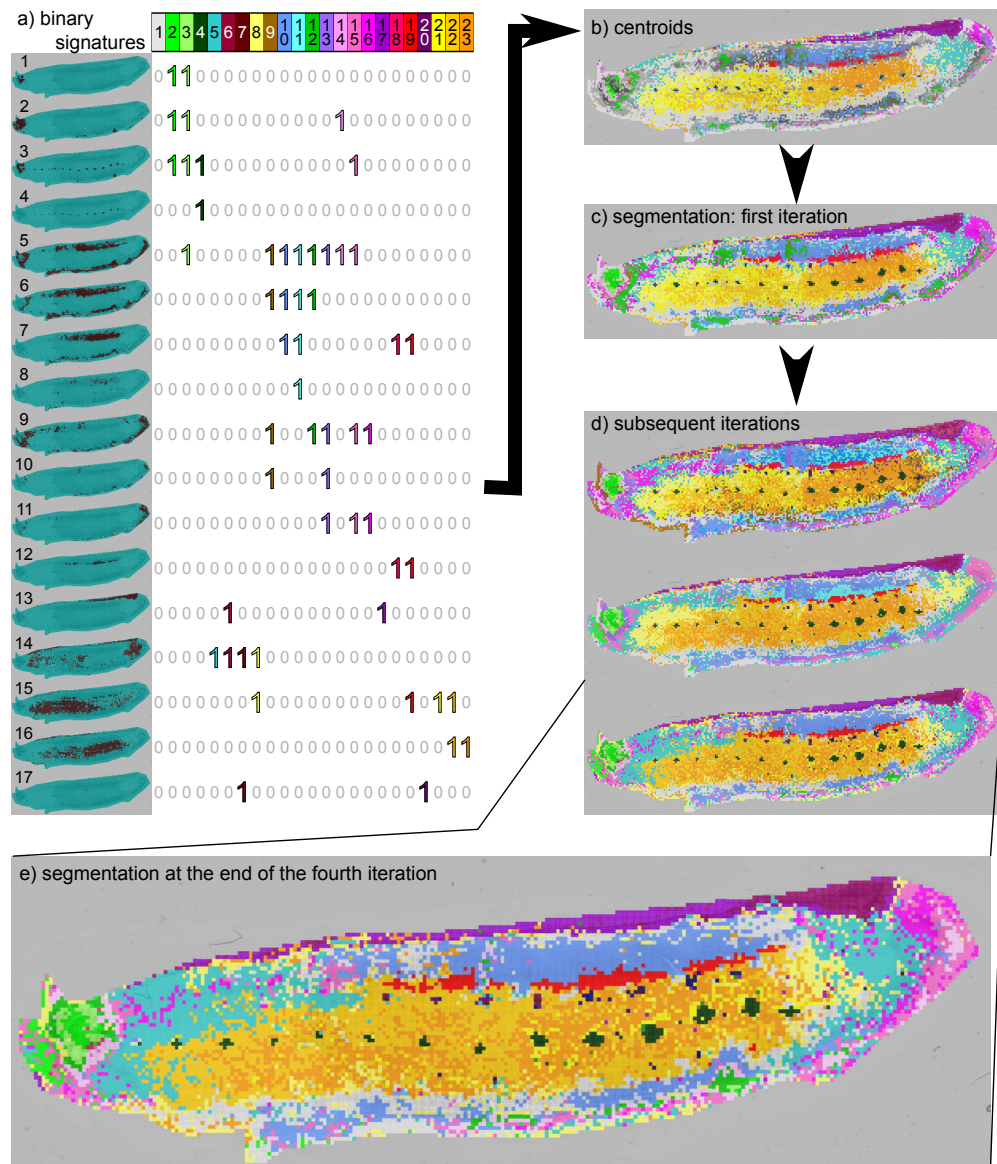


Figure 3.4: Binary spot signatures and leech segmentation maps. a) The dominating binary signatures. Each row represents a clustered query-result and each column represents a selected binary signature. Regions of interest may show some overlap. For example, the centroid for the heart (columns 18 and 19) also shows expression in the lateral region (row 7) thus resulting in binary signatures containing 1's in both rows 7 and 12. b) Spots corresponding to each of these binary signatures. These are used as centroid for clustering. These centers already reveal the major segments. c) New image-segmentation resulting from the reassignment of spots to the closest centroids. d) Refinement of the segmentation over subsequent iterations. e) The segmentation at the end of 4 iterations (run without any user intervention).

Methods 3.4.3, equation 3.8.)

2. Choose a subset $\mathcal{H} \subseteq \mathcal{G}$ of *dominating* signatures: signatures that are common to many spots.
3. For each dominating signature $\mathbf{b} \in \mathcal{H}$, calculate the center $\mathbf{c}_{\mathbf{b}}$ as the mean of spots that binarize to \mathbf{b} . (See Methods 3.4.3, equation 3.9.)
4. Reassign each spot s to the center $\arg \min_{\mathbf{b}} \|\mathbf{a}_s - \mathbf{c}_{\mathbf{b}}\|_2$.

While the user has some ability to choose which binary signatures are to be maintained in the interactive mode, the algorithm can automatically determine which binary signatures are ‘dominating’ (see Methods). Figure 3.4a describes the dominating binary signatures from the first iteration. For example, column 4 of the matrix (dark green) describes the binary signature for spots in the CNS ganglia (1 in rows 3, 4, and 0 elsewhere). Also, the last 3 columns (yellow, gold, orange) describe the ventral region (rows 15, 16). However, the figure reveals the complexity of segmentation. These 3 binary signatures specify molecules in the anterior ventral region only, in both the anterior and posterior ventral region, and the posterior ventral region only, respectively. The ‘correct’ segmentation could be obtained by any combination of these 3 binary signatures. Moreover, if we look at the anterior ventral region (row 15), we see representation from multiple signatures, including those from the heart (column 19), and an undefined region (column 8), illustrating spatial distribution of molecules that would not be apparent in a final segmentation. It is worth noting that there are few spots in the heart only, thus resulting in binary signatures that cover both the heart and the lateral region (columns 18 and 19). Similarly, there are two query-results covering the head (row 1 and 2). Thus, in the resulting segmentation, the spots are divided between those in the “inner” part of the head, present in both query-results (columns 2 and 3) and those present in the “outer” part of the head, i.e., only in row 2 (column 14).

These dominating binary signatures are used to compute new centers. Figure 3.4b illustrates the spots that matched exactly to a center signature. We can see that these centers already reveal the major segments. The reassignment of spots to the center creates a new segmentation (Figure 3.4c). In the next iteration, we use each of the segments of this new segmentation as queries in the semi-supervised component, thus re-iterating through the process described above. Subsequent iterations result in a

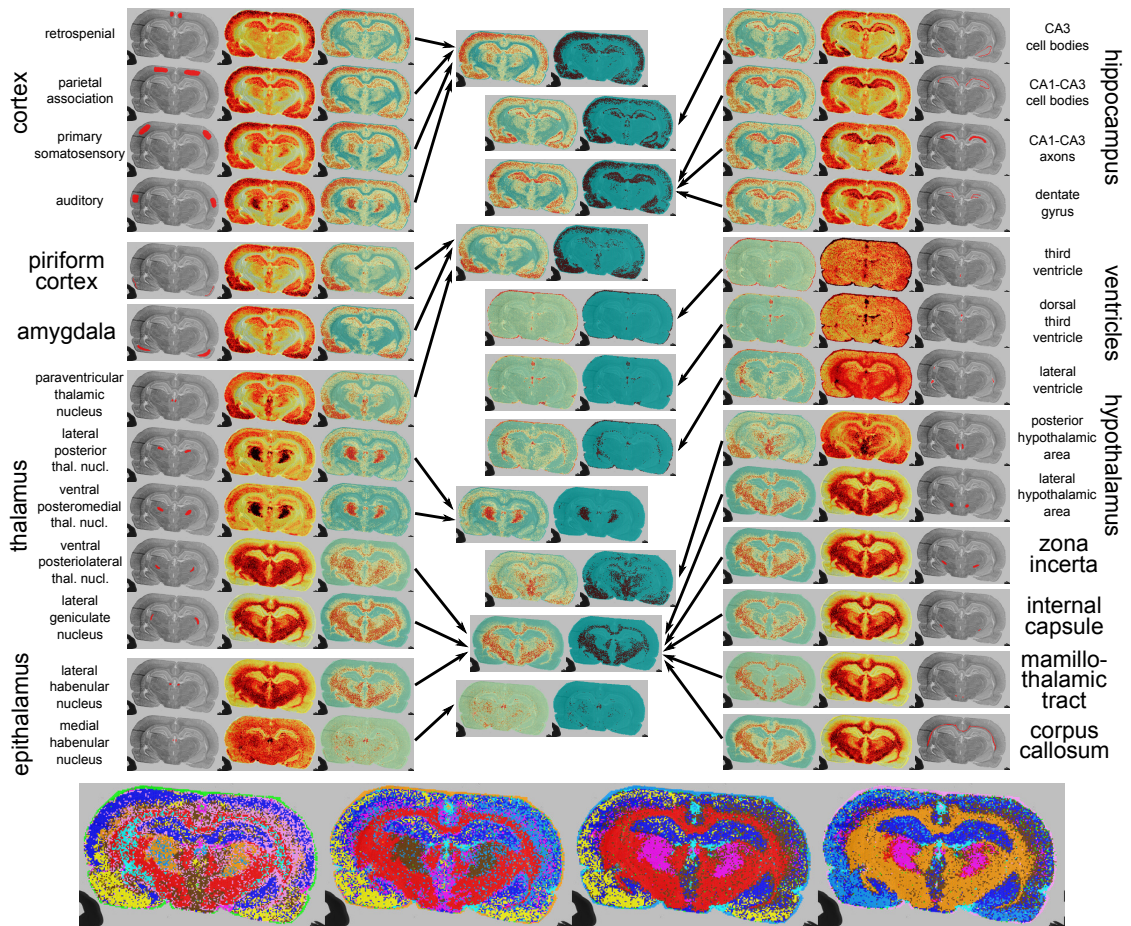


Figure 3.5: Results for the rat brain slice dataset. The results show clear demarcations of the morphology. At the top, on the left-hand and right-hand sides, weighted intensities and log-odd scores images are shown for each of the original queries. In the center, we show the log-odds and binary images resulting from clustering of the query-results. At the bottom, we show the image-segmentation resulting from subsequent iterations.

refinement of the segmentation (Figure 3.4d). The segmentation at the end of 4 iterations (run without any user intervention) is highlighted (Figure 3.4e), and reveals the power of AMASS. Unlike other clustering methods, the final segmentation clearly reveals small and large morphological regions including ganglia, pores, brain, lateral, and ventral regions along with their molecular signatures.

Similar results were obtained for the rat brain segmentation (Figure 3.5), with clear demarcations of the morphology. Basic anatomy is provided for reference in Figure B.6b. Specific initial queries behave as expected with a few surprises. In Figure 3.5,

we have separated the queries based on the brain substructure to which they belong (cortex, thalamus, hippocampus, etc). The triplet of images associated with each query is composed of the corresponding original query, the weighted intensity image and the log-odds score image (outward towards the middle). When looking at the cortex queries, we can see that the different upper cortex queries (retrosplenial, parietal, primary somatosensory) all result in the larger upper cortex region. However, the region demarcated as the auditory cortex interestingly recruits a portion of the thalamus. The piriform cortex and amygdala, which are related to the neocortex, show some signal in the cortex with the majority of the signal in their respective regions. Interestingly, the paraventricular thalamic nucleus also shares a similar molecular signature to that of the amygdala and the piriform cortex. Other parts of the thalamus seem to split between two different regions; the lateral posterior thalamic nucleus and the ventral posteromedial thalamic nucleus recruit one shared region, while the ventral posterolateral thalamic nucleus and the lateral geniculate nucleus recruit white matter. The internal capsule, mamillothalamic tract and corpus callosum, which are part of the white matter of brain, which consists mostly of myelated axons, also recruit all white matter regions of the brain. This suggests that there is a distinct molecular signature for white matter, possibly due to myelin. As expected, all ventricles share the same molecular signature, which in this case should correspond to that of the matrix, explaining the signal at the edge of the sample. It is worth noting that some regions, such as the medial habenular nucleus and posterior hypothalamic area, have very particular molecular signatures, resulting in the recruitment of very specific regions. The middle panels describe the result of hierarchical clustering after the first iteration. The two images in each cluster represent the resulting average log-odd scores and the binary image after votes. The clustering step behaves as expected, with the different cortex query-results ending up in one cluster and all white matter query-results ending up in another. The bottom panels show the image-segmentation results after subsequent iterations. The rat brain is segmented in the different anatomical regions.

Finally, in Figure B.2, we show results for a completely random run on the leech embryo dataset. The algorithm automatically generated an initial random segmentation (shown in panel a), composed of 100 seed-segments each consisting of a 1 to 3 adjacent

MALDI spots. We ran 10 automatic iterations of the algorithm, once using 3x3 median smoothing, and once without any smoothing (panels b and c respectively). Segments resulting from the random segmentation also show the major morphological features of the leech and does not differ much from the guided approach. A few things to note are that the distinction between the anterior ventral and posterior ventral regions of the leech is not as well defined as in the guided approach, although it is still present. Also, in this specific run, a part of the top body margin clustered with the ventral region of the leech. Moreover, the nephridia, which have a weak signal, are not shown in this specific run, while the anterior ones are maintained in the guided analysis. This is due to the fact that small regions of interest do not show on every random run as there is a chance that no seed-segment is generated in the region. However, we do see the nephridiopores in the non-smoothed version of this specific run, signaling that a random segment must have been generated in one the nephridiopore. Finally, it is worth noting that the smooth segmentation provides much cleaner and more unified segments, but at the cost of some of the smaller segments, such as the pores or some of the anterior ganglia, which are completely lost by the final segmentation.

3.3 Discussion

MALDI imaging is rapidly becoming a technique of choice for surveying and discovering proteins and peptides that have spatially distinct signatures of expression. The large multi-dimensional nature of the datasets (expression of $\sim 10^3$ molecular species in $\sim 10^4$ spots) makes the mining for knowledge difficult. Unsupervised approaches seek to segment the tissue section into regions, each with a distinct molecular signature. However, classical segmentation techniques are often based on clustering molecules that have similar expression patterns. The quality of segmentation is often judged by its congruence with known morphology.

Here, we argue that these approaches do not work as well if there are small segments with low to medium abundance mass values. Instead, we propose a semi-supervised approach that ranks mass values by their spatial discrimination. Our results lead to consistent discovery of very fine segments (organs with 2 – 3 spots at $50\mu m$

resolution). Also, our query based techniques often reveal novel relationships, such as co-expression of molecules in the auditory cortex and portions of the thalamus in rat brain.

The next step in the process, is the actual identification of peptides corresponding to the molecular signatures. This remains a challenge even with progress in *in situ* trypsinization and other MS/MS fragmentation techniques. Further refinement of the discover of molecular signatures, and the identification of peptides will contribute to a novel tool for exploring the role of molecules in specific cellular phenotypes.

3.4 Methods

We first acquire MS Imaging data on the animal/section. We convert the data into our own lossless format and normalize it. As shown on Figure 3.1, the algorithm consists of two main components: a semi-supervised component and a partitioning component. The semi-supervised component performs a query for each of the original segments. It returns the molecular signature specific to the segment, as well as all areas sharing similar molecular signatures. The partitioning component assigns each spot to 0 or 1 cluster creating a (potentially partial) segmentation map. After selection of initial clusters, the algorithm iteratively runs these approaches fixing high-accuracy clusters along the way. While the algorithm can be run in a completely automatic mode, the main goal is to provide the user with easy control at each step of the way. Thus, it is possible for the user to choose which clusters to fix, keep or discard at each iterations. This allows the user to fine tune the results without “tweaking” parameters. The final output is a segmentation map with associated areas and molecular signatures for each cluster.

3.4.1 Data Acquisition

Leech embryo: For the leech embryo analysis, we selected a specimen at stage E12 (12 days of development at room temperature), when the segmented nervous system and other organs like the nephridia, have clearly defined boundaries and are in a sufficiently advanced degree of molecular differentiation that specific signatures can be expected.

The embryo was opened along the dorsal mid-line and the yolk removed, then pinned flat, exposed for 1 – 2 min to methanol to harden the tissues, finally placed on metal-coated (ITO) glass slides with the internal surface exposed and immediately dried. After recording transmitted light images to document the gross morphology of the specimen, it was coated with several layers of special solid ionic matrices (CHCA/Aniline), using a manual pneumatic TLC sprayer (VWR, Strasbourg, France). Such matrices that have proven to be very efficient for peptide/protein analysis directly from tissue sections. MALDI direct analyses of tissues and MALDI Imaging were performed on a MALDI-TOF/TOF instrument (Ultraflex II, Bruker Daltonics, Germany) over 38837 m/z values from 12115 locations, generally sampling the embryo completely in a rectangular raster of points $60\mu\text{m}$ apart. We refer to previous work [75] for a more detailed description of the sample preparation. The complete data-set is a collection of spectra, each associated with a ‘spot’ on the leech surface. Conceptually, the data can be represented as a collection of triples $\langle m, s, I_{m,s} \rangle$ describing the spectral intensity $I_{m,s}$ at each spot s , and m/z value m . The spectral intensity depends upon the abundance of the molecular species among other factors. While the intensities of different molecules cannot be compared directly, the relative intensity of the same molecule (mass value m) at different spots is a measure of the relative abundance of the molecule.

Rat brain slice: Cryosections of $10\mu\text{m}$ thickness were cut on a cryostat (CM 1900 UV, Leica Microsystems GmbH, Wetzlar, Germany) and transferred to a precooled, conductive indium-tin-oxide (ITO) coated glass slide (Bruker Daltonik GmbH, Bremen, Germany). The sections were washed twice for 1 min in 70% ethanol, and once for 1 min in 96% ethanol and then dried in a vacuum desiccator. The matrix (Sinapinic acid at 10 mg/mL in 60% acetonitrile and 40% water with 0.2% trifluoroacetic acid) was applied using the ImagePrep device (Bruker Daltonik GmbH) following a standard protocol. Mass spectra were acquired on a MALDI-TOF instrument (Autoflex III; Bruker Daltonik GmbH) equipped with a 200 Hz smartbeam II laser. MALDI measurements were performed in linear positive mode at a mass range of 2.5 kDa to 25 kDa. The lateral resolution for the MALDI image was set to $80\mu\text{m}$. A total of 200 laser shots were summed up per position.

3.4.2 Query

We compute $\mathbf{A} : S \times 2^S \rightarrow \mathfrak{R}$, where

$$\mathbf{A}[I, s] = \log \frac{\Pr(s \in I)}{\Pr(s \notin I)}. \quad (3.1)$$

The probability estimates are computed empirically. Consider a score function $\mathcal{Z} : S \times 2^S \rightarrow \mathfrak{R}$ where $\mathcal{Z}(I, s)$ denotes the ‘score’ of spot s against the segment I . The only requirement on \mathcal{Z} (see next subsection) is that the scores in I are higher than $S \setminus I$, and well separated. We estimate $\Pr(s \in I)$ by empirically computing the probability that a randomly chosen spot in I would score lower than $\mathcal{Z}(I, s)$

$$\Pr(s \in I) \simeq \Pr(\mathcal{Z}(I, t) \leq \mathcal{Z}(I, s) | t \in I). \quad (3.2)$$

Likewise,

$$\Pr(s \notin I) \simeq \Pr(\mathcal{Z}(I, t) \geq \mathcal{Z}(I, s) | t \in S \setminus I). \quad (3.3)$$

Weighted intensity scores: The spectrum acquired on spot s is a collection of m/z values and intensities. We do a simplified peak selection, choosing the top 5 scoring m/z values (averaged over a 1 Da window) in a scrolling window of 50 Daltons. The selected peaks are represented by a vector \mathbf{v}_s , where $\mathbf{v}_s[m]$ is the intensity at m/z value m . Second, we compute a vector of weights \mathbf{w}_I , where $\mathbf{w}_I[m]$ describes the ‘importance’ of a peak at m in separating spots in I from $S \setminus I$. Intuitively a spot s belongs to I if \mathbf{w}_I and \mathbf{v}_s are correlated. Therefore, we choose the weighted-intensity score function

$$\mathcal{Z}(I, s) = \mathbf{w}_I \cdot \mathbf{v}_s. \quad (3.4)$$

In earlier work [75], we computed the Wilcoxon-Mann-Whitney ρ -statistic. $\rho_I[m]$ is a measure of how well the peak at m separates spots in I from those in $S \setminus I$. Formally, for randomly chosen spots $s \in I, t \in S \setminus I$

$$\rho_I(m) \simeq \Pr(s[m] \geq t[m]). \quad (3.5)$$

While we could use $\rho_I[m]$ directly as the weighting function, we choose

$$\mathbf{w}_I[m] = \begin{cases} 2 \|\mathbf{w}_I\|^{-1} (\rho_I[m] - 0.5))^p & \text{for } \rho_I[m] \geq 0.5 \\ 0 & \text{for } \rho_I[m] < 0.5 \end{cases}. \quad (3.6)$$

Here, $\|\mathbf{w}_I\|^{-1}$ is a normalizing constant. For $p > 1$, $w_I[m]$ increases sub-linearly, staying close to 0 for intermediate values of $\rho_I[m]$, and then increasing sharply to 1, thus allowing the strongly discriminative m/z values to be sharply upweighted versus multiple low-discriminating mass-values. As p increases, so does the weight of the top m/z discriminative values, causing the query-result to be more specific to the original query.

Smoothing: Optionally, image smoothing may be applied on the weighted intensity images in order to suppress the pixel-to-pixel variability. As shown in Alexandrov et al. [77], the advanced image smoothing methods applied to mass intensity images significantly improve the segmentation results. In contrast to Alexandrov et al. [77], we use simple median smoothing (3x3 window).

3.4.3 Spot Partitioning

Hierarchical clustering Since highly related queries return very similar results, we cluster the rows of matrix A . We use hierarchical clustering with the Tanimoto coefficient as a distance function between segments I_1, I_2 , computing distance to the average log-odds image in the case of clustered-segment.

We denote the clustered-segments as I' , and let $I \rightarrow I'$ if and only segment I is clustered into I' . We compute cluster-scores for spots as

$$A'[I', s] = \text{mean}_{\{I \rightarrow I'\}}(A[I, s]). \quad (3.7)$$

Spot-vector binarization: We select a threshold score t_I for each I based on the distribution of scores in I and $S \setminus I$. Intuitively spot s belongs to I if $A[I, s] \geq t_I$. Next we merge the segments in C by taking a majority vote. Denote matrix B as a binary matrix with rows corresponding to segment-clusters.

$$B[I', s] = \begin{cases} 1 & \text{if } \#\{I \rightarrow I' : A[I, s] \geq t_I\} / \#\{I \rightarrow I'\} \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} . \quad (3.8)$$

Dominating signatures as centers: The columns of B corresponding to spot s describe a ‘binary-signature’ for spot s . In the ideal case, strong segment-clusters should have a unique signature and all spots contained in the cluster have the corresponding signature. For each cluster, denote the most frequent signatures as dominating, if it has sufficient frequency.

Spot partitioning: In the final step of the iteration, we use the dominating signatures to determine cluster centroids and partition spots by assigning the remaining spots to these clusters. Let \mathbf{a}'_s denote the cluster-scores ($\mathbf{A}'[*,s]$) for spot s . For each dominating signature \mathbf{b} , we define the associated set of spots $\mathcal{S}_{\mathbf{b}} = \{s : \mathbf{b}_s = \mathbf{b}\}$. We then define a centroid $\mathbf{c}_{\mathbf{b}}$ for each dominating signatures as the mean of the cluster-scores of spots

$$\mathbf{c}_{\mathbf{b}} = \text{mean}_{\{s \in \mathcal{S}_{\mathbf{b}}\}}(\mathbf{a}'_s). \quad (3.9)$$

We also add a zero centroid to the set, which is either the centroid of all spots not belonging to any query-result if such spots exists, or a zero vector if there are no such spots. Each spot s is reassigned to the closest centroid $\arg \min_b \|\mathbf{a}'_s - \mathbf{c}_{\mathbf{b}}\|_2$. Overall, spot partitioning corresponds to a single pass of K -means clustering, and provides the segmentation for the next iteration.

3.4.4 Query consistency

We measure segmentation based on consistency of query (see Figure B.5). For segment I , denote \mathcal{S}_I as the set of all spots such that $\mathbf{A}[I,s] \geq t_I$. In the ideal scenario, querying a segment will return all the spots in that cluster and only those spots ($I = \mathcal{S}_I$). We use the Jaccard similarity coefficient to measure consistency of I as

$$\text{consistency}(I) = J(I, \mathcal{S}_I) = \frac{|I \cap \mathcal{S}_I|}{|I \cup \mathcal{S}_I|}. \quad (3.10)$$

3.5 Acknowledgement

This chapter, in full, was published as ‘‘AMASS: algorithm for MSI analysis by semi-supervised segmentation’’. Bruand J, Alexandrov T, Sistla S, Wisztorski M,

Meriaux C, Becker M, Salzet M, Fournier I, Macagno E, Bafna V. *J Proteome Res* 10(10):4734-43. 2011. The dissertation author was the primary author of this paper.

Chapter 4

Comparative Analysis of Mass Spectrometry Imaging Data

4.1 Introduction

Due to the nature of biological data, noise is inherent and signal in a single dataset may not always reflect a true phenomenon. Confidence in results is typically boosted by replicating an experiment and using statistics to measure the likelihood of a signal reflecting a true event. Moreover, large scale studies require many experiments across several organisms and/or conditions.

However, comparative analysis is rarely done in mass spectrometry imaging. The time and cost associated with sample preparation and running the instruments have prohibited the generation of many replicates in the past. However, as the technology improves, not only does the data quality increase, but the time and cost of data generation decreases. Thus, it is essential to apply automated comparative analysis to the field of mass spectrometry imaging.

In this chapter, we introduce a new method for large scale comparative analysis of MSI datasets. Given a set of pertinent query molecules, our tool finds, in each dataset, all molecules that have a similar spatial distribution and clusters the datasets based on the resulting molecular signatures. By comparing the molecular signatures in the clusters, we can confirm the existence of signal across replicates and identify signal changes for

different conditions. This approach has the potential to identify unknown relationships between multiple data acquisitions.

We apply our method to a large number of bacterial interaction MSI datasets which have been acquired over previous years. Querying with known natural products results in clusters of datasets from organisms are known to produce these molecules. In-depth analysis of the corresponding molecular signatures shows many known products, as well as some yet uncharacterized molecules.

4.2 Methods

Figure 4.1 gives an overview of the workflow. First, we collect an assortment of MSI datasets and preprocess each in a lossless manner. Each dataset can be viewed as a series of m/z images, one for each possible m/z values. To compare the datasets, we select a query m/z value, corresponding to a known molecule. It is also possible to give a set of query m/z values as input. For each dataset a query region is defined from the corresponding m/z image. The query region is used to define a molecular signature of m/z values with similar spatial distribution than our query m/z . The molecular signatures are clustered using hierarchical clustering. A filtering step allows to reduce the number of datasets in the final clustering.

4.2.1 Data Acquisition and Preprocessing

We collect a medley of existing MSI datasets which were previously acquired for different purposes. Current data was acquired on a Bruker Microflex (Bremen, Germany) but additional data from a Bruker Autoflex (Bremen, Germany) is available and soon to be added to our collection. Each dataset is converted into a in-house lossless format for efficient data processing. When an experiment consists of data acquisition over several separate regions, the data is split and each region is considered an independent dataset. In Figure 4.1, we show each dataset as a series of m/z images, for all possible m/z values. Currently, our collection consists of 898 datasets from 345 experiments.

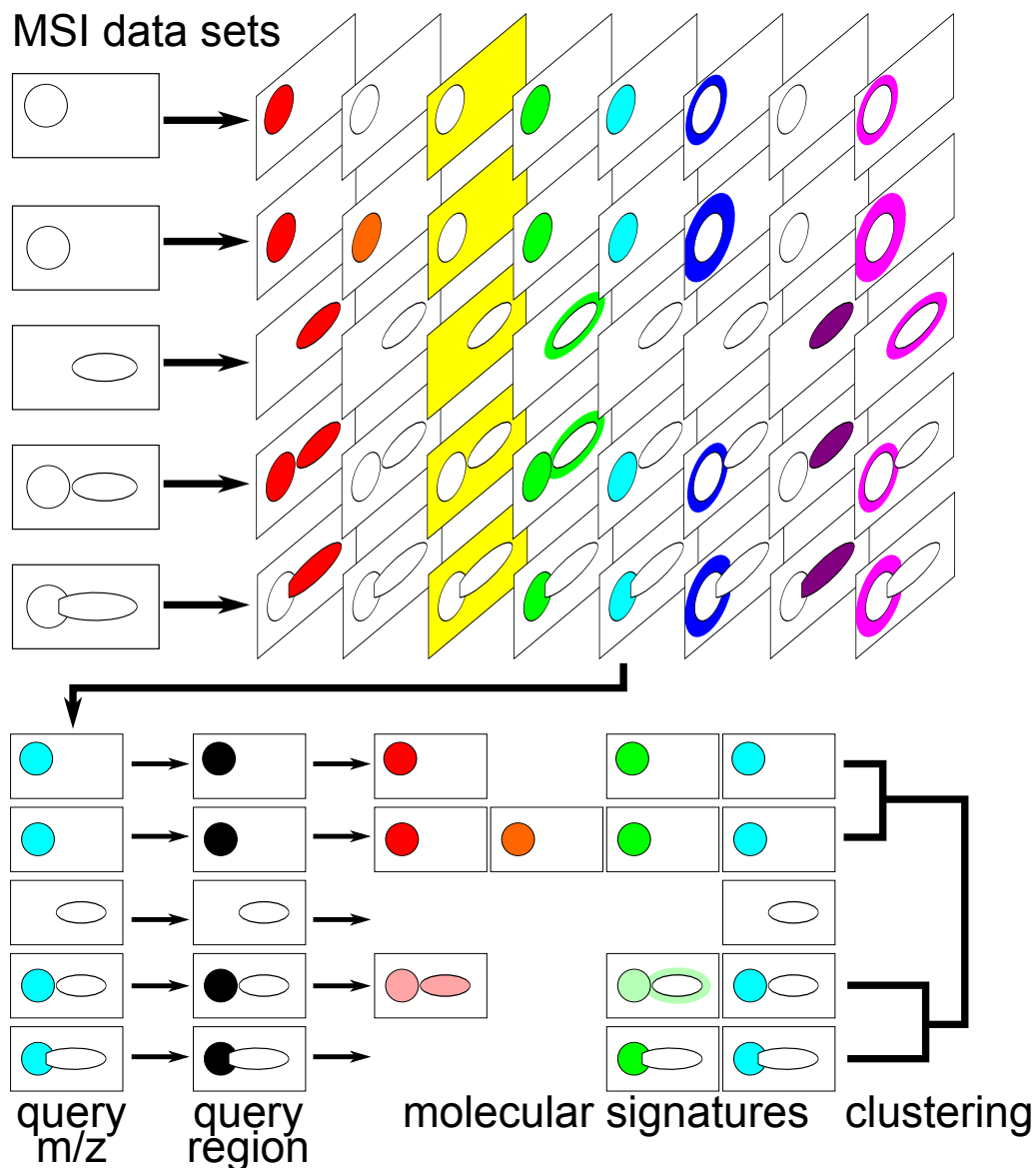


Figure 4.1: Overview of the comparative analysis workflow. First, we collect an assortment of MSI datasets and preprocess each in a lossless manner. Each dataset can be viewed as a series of m/z images, one for each possible m/z values. To compare the datasets, we select a query m/z value. Here, we select the cyan m/z value. For each dataset a query region is defined from the corresponding m/z image. The query region is used to define a molecular signature of m/z values with similar spatial distribution than our query m/z . The molecular signatures are clustered using hierarchical clustering.

4.2.2 Defining the Query

To compare the datasets, we define a query for each dataset based on the intensity image for one or several query m/z value(s). When several query m/z values are given, the query intensity image is generated by averaging intensities at the given m/z values. From this intensity image, we define a query region in following manner. Intensities are clustered using 1D k-means clustering with 2 clusters (high-intensity vs. low-intensity). We choose as original centroids the maximum and median intensity values for the high-intensity and low-intensity clusters respectively. We define the query region as the set of spots having intensity at least 3 times closer to the final high-intensity centroid than the final low-intensity centroid.

4.2.3 Obtaining the Molecular Signatures

Molecular signatures were acquired using the AMASS library [82]. We calculate the previously described weight for the average intensities in each 1 Da bins. We also obtain the weighted intensities image and the log-odds images as previously described. Since the molecular signatures are binned in the same manner for all dataset, we have a score for each dataset and each m/z bin. This can be represented as a matrix where each row has the molecular signature for a dataset and each column is a m/z value. A score of 0 is assigned if there is no intensity image (data not acquired) for a dataset at an m/z value. If the full m/z range is used, molecular signatures will be defined over the range of m/z values from the minimum m/z value of all datasets to the maximum m/z value of all datasets. It is also possible to define the minimum and maximum m/z values, in which case all data beyond these boundaries is ignored when calculating the molecular signatures and weighted intensities images.

4.2.4 Filtering the datasets

Since we have a medley of datasets, the query molecule will not be expressed in all datasets, and thus the query intensity image is unlikely to show an “interesting” distribution for all datasets. For each dataset, we aim to distinguish whether it is of interest by employing several filters. This allows us to reduce the number of datasets in

the final clustering.

Consistency filtering: We previously describe a consistency score for clusters [82]. The premise is the image resulting from the molecular signature should be similar to the original query. If they differ, then the molecular signature does not reflect the query, and thus should be discarded. This typically happens when the query m/z image captures noise. While some m/z values may show somewhat similar distribution, they all differ from the original query by a certain extent and greatly from each other. We offer two types of consistency filter: a query consistency filter and a weighted intensity consistency filter. The query consistency score is calculated exactly as described previously [82]. The weighted intensity consistency score is the cosine distance between the weighted intensity image and the original query m/z image.

Number of peaks filtering: We aim here to filter two different cases. In the first case, the query m/z image shows a unique spatial distribution which is not reflected by any other m/z values. This is typical if the original m/z image captures noise. Thus we filter out any datasets which have less than 10 m/z bins with score $s \geq 0.5$. In a second case, the query m/z value shows a distribution which is reflected by most other m/z values. This is typical when the query m/z value is a matrix element for that dataset. We filter out any datasets which have more than 1000 m/z bins with a score $s \geq 0.5$.

4.2.5 Clustering the datasets

We create a dendrogram of the datasets by clustering the molecular signatures using hierarchical clustering with the correlation distance and UPGMA linkage. Other distances and linkage options are made available to the user, but we find this setting to perform better (data not shown). All clustering is done via the SciPy python package [83].

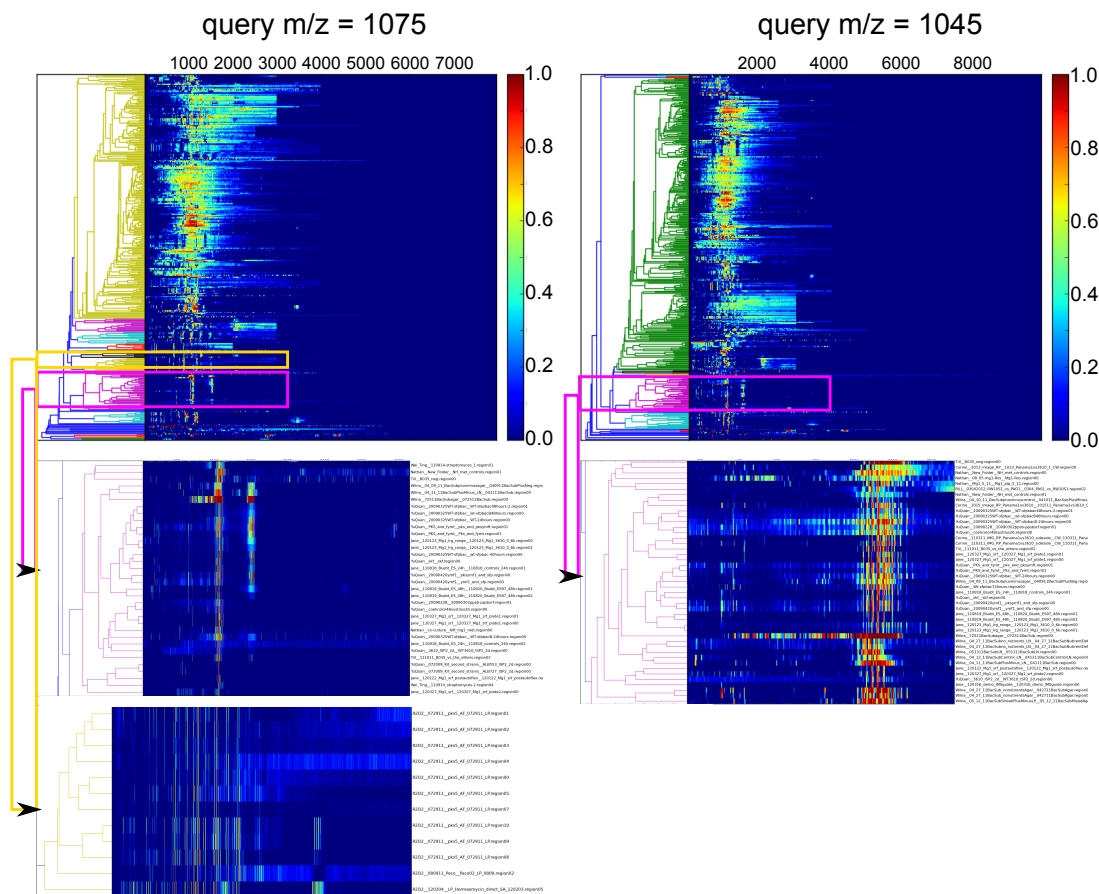


Figure 4.2: Clustering results with query $m/z = 1045$ and $m/z = 1075$ (surfactins). The top two panels show the full dendrogram with molecular signatures when query with $m/z = 1045$ and $m/z = 1075$, which correspond to two surfactins. The trees are very similar as expected since the two molecules have similar spatial distribution. When looking at the magenta sub-cluster, which consists mostly of *Bacillus subtilis* datasets, we find that the subtrees between the two queries are almost identical. The yellow subtree in the $m/z = 1075$ query consists of pks knockout *Bacillus subtilis* datasets.

Figure 4.3: *Bacillus subtilis* cluster for query $m/z = 1075$. We use the intensity images at $m/z = 1075$ (surfactin) to create our query region (right hand size). The top matrix shows the molecular signatures for each dataset, as well as the resulting clustering (left of matrix). Several specific m/z values (columns) were selected from the matrix for more thorough examination, and corresponding m/z images were pulled for each dataset (bottom). Numbered datasets indicate separate replicate experiments.

Str - *Streptomyces*.

Bsubt - *Bacillus subtilis*.

ES - environmental strain.

WT - wild type.

pks - polyketide synthase knockout.

bacB - bacB knockout.

sfp - sfp knockout.

ppsb - plipastatin knockout.

skf - skf knockout.

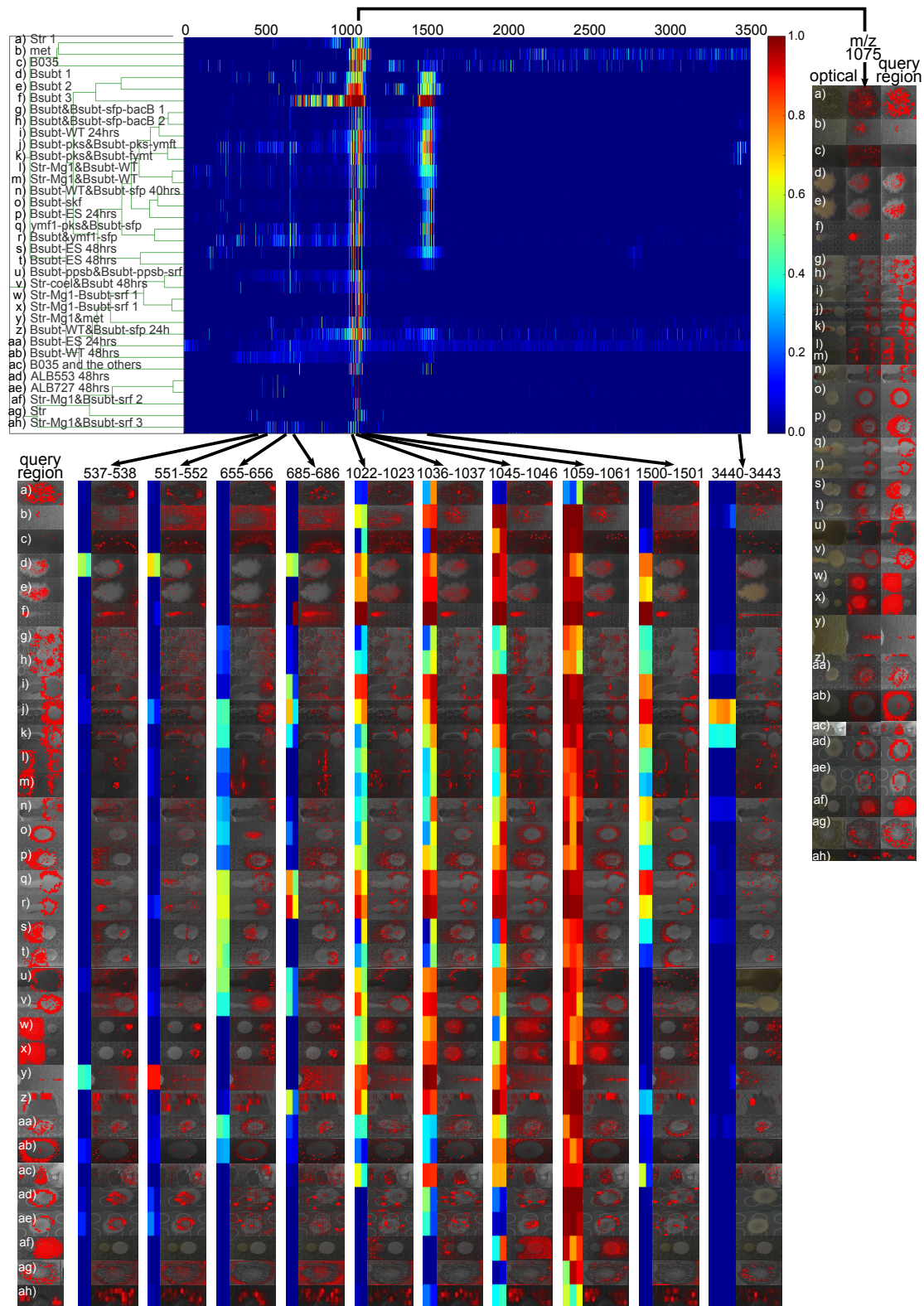
srf - surfactin knockout.

ymft - ymfT knockout.

ymf1 - YMF1 strain.

Mg1 - Mg1 strain.

ALB553 and ALB727 - marine environmental isolates.



4.3 Results

While we are still incrementing our collection of datasets, we show here some preliminary results for two clusters: a cluster consisting of *Bacillus subtilis* datasets from querying surfactin and a cluster consisting of *Streptomyces coelicolor* datasets from querying sapB.

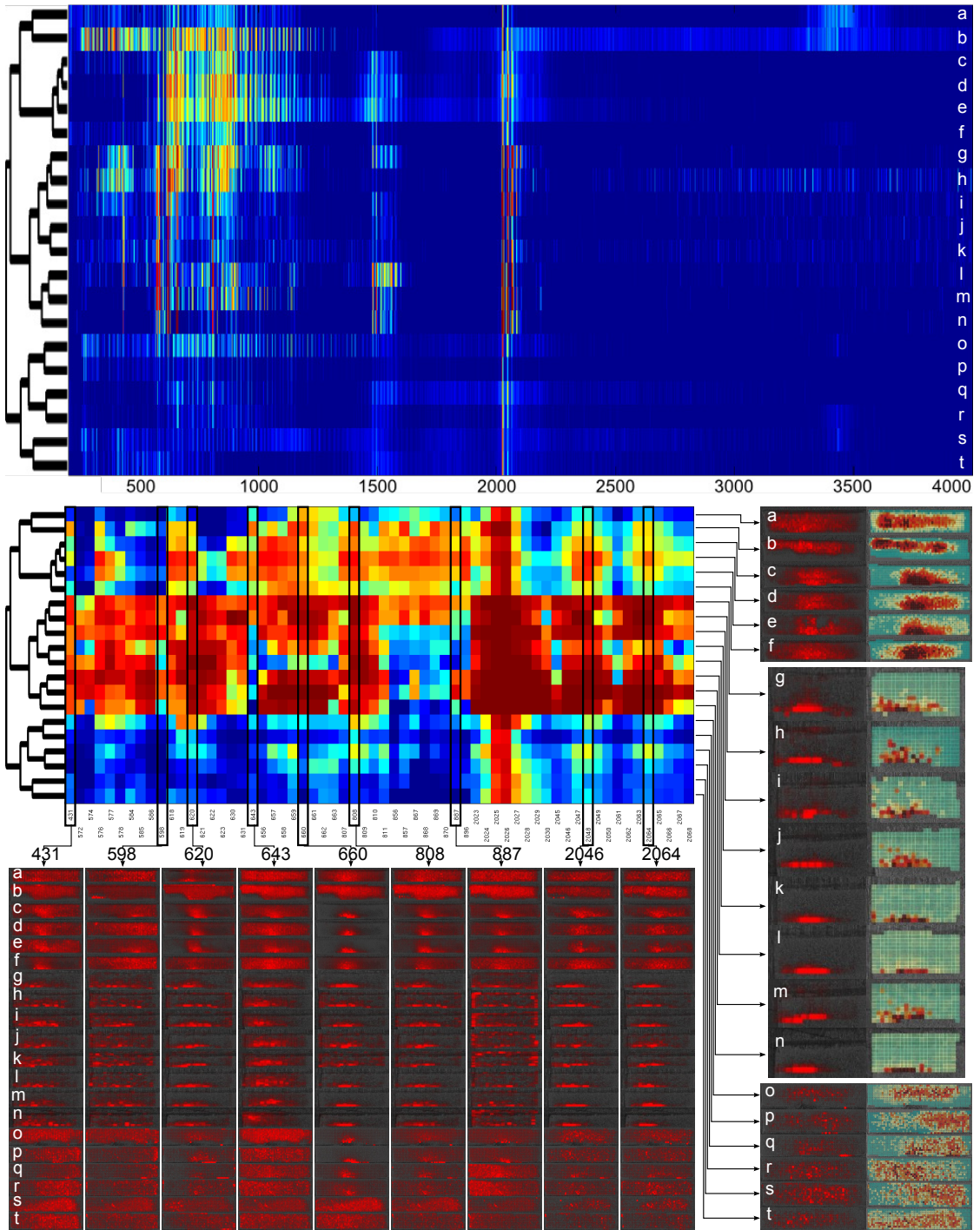
4.3.1 Surfactins and *Bacillus subtilis*

We query our datasets with $m/z = 1045$ and $m/z = 1075$, which are mass values of different forms of surfactin in *Bacillus subtilis*. Figure 4.2 shows the full dendrograms (with filtering) resulting from the clustering of the molecular signatures. The trees are very similar as expected since the two molecules have similar spatial distribution. When looking at the magenta sub-cluster, which consists mostly of *Bacillus subtilis* datasets, we find that the subtrees between the two queries are almost identical. The yellow sub-tree in the $m/z = 1075$ query consists of pks knockout *Bacillus subtilis* datasets.

In Figure 4.3, we focus on the magenta sub-tree from the $m/z = 1075$ query. Some of these datasets have been published elsewhere [84, 85].

In most cases, replicate datasets have similar molecular signatures and form sub-clusters. For example, datasets for *Bacillus subtilis* colony by itself form a sub-cluster on rows *d*, *e*, *f*, and datasets for wild-type *Bacillus subtilis* interacting with sfp knockout *Bacillus subtilis* forms a sub-cluster on rows *g*, *h*. These replicates originate from different acquisitions showing consistency of results across multiple experiments. However, some replicates do not cluster together. It is more often the case for replicates originating from different acquisitions, most likely due to experimental and data quality differences. For example, if we look at the replicates for interaction between *Streptomyces* Mg1 strain and *Bacillus subtilis* srf knockout, rows *w* and *x*, which originate from the same experiment, form a cluster, but rows *af* and *ah*, which originate from different experiments are separated from that cluster and from each other. When looking at the molecular signatures, we can see that the signal is much stronger in rows *w* and *x*.

Figure 4.4: *Streptomyces coelicolor* cluster for query $m/z = 2026$. We use intensity images at $m/z = 2026$ (sapB) as query. The top matrix shows the molecular signature for each dataset for all m/z values. The datasets share many of the m/z values with high scores. Dendrogram on the left shows the clustering. datasets originating for the 3 mass spectrometry imaging experiment form 3 sub-clusters. The second matrix is a reduced matrix. We only show columns from the original matrix which have at least 5 rows of score $s \geq 0.7$. For each dataset, images on the right-hand side correspond to the intensity distribution at $m/z = 2026$ (left column) and log-odds image resulting from the query (right column). At the bottom, we show the m/z images for all datasets for specific m/z values selected from the reduced matrix. As expected, higher scores are given when the intensity distribution for that m/z value matches that of $m/z = 2026$.



In the molecular signature matrix, we can see a set of m/z values in the 1020-1140 range which have high scores for almost all datasets. We pull some of these high scoring m/z columns ($m/z = 1022-1023, 1036-1037, 1045-1046, 1059-1061$) and generate the corresponding m/z intensity image for each dataset. These high-scoring m/z values correspond to the mass of different forms of surfactin. This is to be expected since we queried with a surfactin, thus other forms of surfactin should show similar spatial distribution. The intensity images for these m/z values indeed show similar spatial distribution to that of the query intensity image ($m/z = 1075$, right).

Another set of m/z values in the range of 1440-1560 show high scores for the sub-cluster in rows $d-t$. These values correspond to the masses of several forms of plipastatin, which is also known to be expressed by *Bacillus subtilis*.

Additionally, a few rows show high 3440-3443, which corresponds to subtilisin. It is likely that we only see a clear expression in few datasets because it is a lower abundance molecule and because detection tends to be more challenging in higher m/z range. MS1 spectra show much lower signal for subtilisin [86].

Finally, we find a few still uncharacterized m/z values which show similar spatial distribution for several datasets. For example, $m/z = 655-656$ has mid-range scores for rows $j-v$ and the corresponding intensity images are similar to the query intensity image. Many other m/z values show an interesting distribution in this cluster and warrant further investigation.

4.3.2 SapB and *Streptomyces coelicolor*

We query our datasets with $m/z = 2026$, which is the mass of sapB in *Streptomyces coelicolor* [84, 87]. Figure 4.4 shows the resulting molecular signature and clustering. The clustering consists of three *Streptomyces coelicolor* datasets which were acquired for 3D reconstruction [88].

On the top matrix, we show the full molecular signatures for these datasets. For better visualization, we reduced the columns of the molecular signature and form a reduced matrix (middle). We only retain m/z columns with at least 5 rows of score $s \geq 0.7$. Because of the strictness of these thresholds, the reduced matrix does not reflect all interesting values, but it makes visualization much more manageable.

The strongest signal is in the m/z range of 2020-2050 which correspond to sapB, the query molecule. While there are many other strong signals in rows $a-n$, the signal is much fainter in rows $o-t$, which correspond to datasets from a different experiment. When looking at the the query intensity images and resulting log-odds images (right-hand side), we can see that is less uniform and overall lower in these last rows. Most likely, the datasets from experiment rows $o-t$ are of lower quality than those of rows $a-f$ and $g-n$.

As in the previous case, we see signal for m/z values corresponding to various molecules known to be expressed by *Streptomyces coelicolor*. These include ferri-coelichelin at $m/z = 619$ [89], γ -actinorhodin in the mz range 630-660, and CDA in the mz range 1490-1600 (top matrix only). Some other m/z values ($m/z = 431, 808, 867, 887$) are yet to be characterized. If we look at the individual images for $mz = 808$, we can see clear signal in the colony in rows $b-n$, as reflected by the scores in that column, indicating that there is indeed a signal there.

4.4 Discussion

We describe here a computational tool for large-scale comparative analysis of MSI data. We use our tool to compare ~ 900 datasets of bacterial interactions and pull out interesting clusters with pertinent molecular signatures. We find strong signals for various expected natural products, as well as for a few uncharacterized molecules.

Many simple extensions can be added to this method to make it more powerful. For example, a region of interest can be defined on a single dataset and the resulting molecular signature can used as the set of query m/z values to search the rest of the data. This would prevent the bias introduced by selecting the query m/z values.

Taking this idea further, we could segment one single dataset into different regions, each with its own molecular signature, use the different molecular signatures as queries for the other datasets and build the final clustering based on the concatenated molecular signatures. This would allow for the search and clustering of several component simultaneously.

Finally, while the tool has an interactive python interface, we would like to make

available a web interface which would allow users across the world to browse all the current data in an interactive manner.

4.5 Acknowledgements

This chapter, in part, is currently being prepared for submission for publication of the material. Bruand J, Dorrestein PC, Bafna V. The dissertation author was the primary author of this material.

Chapter 5

On-Tissue Peptide Identification using Spectral Libraries

5.1 Introduction

Developments in instrumentation and experimental techniques have allowed for the acquisition of MS/MS spectra of on-tissue trypsinized peptides. However, these spectra are generally of lower quality than those acquired from protein extract, due to several factors. First, the MSI samples are acquired at a single spot instead of a whole organelle, resulting in less molecules present in the sample. Second, typical protein extraction process involves steps such as centrifugation to increase protein concentration which are eliminated in the case of direct on-tissue acquisition. Finally, the lack of a nano-LC column before analysis by the mass spectrometer results in poor protein separation and mixture peptides.

Most identification approaches have focused on using database search tools. However, for complex spectra or spectra of low quality, spectral library search tools have shown to greatly increase the number of identification [90, 91]. Moreover, existing tools allow for the search of mixture peptides against spectral libraries [92], which would be essential in the case of spectra acquired directly on-tissue. We describe here the preliminary work on the development of spectral libraries for MALDI imaging peptide identification and the corresponding search tools.

5.2 Methods

5.2.1 Data Acquisition

We obtained tissue sections from a 4-month old frozen rat half-brain. Tissue were placed on glass slide, washed and subjected to on-tissue trypsin digestion using a piezoelectric microspotter.

For protein extract spectra acquisition, off-line Liquid Extraction Surface Analysis (LESA) was performed. Samples were subjected to nano-LC separation and directly spotted on MALDI target. We acquire MS/MS spectra on a MALDI-orbitrap platform. Currently, we have two datasets containing 580 and 980 MS/MS spectra. We also acquired a set of 50415 MS/MS spectra from an ESI-orbitrap platform with a similar protocol.

For MALDI imaging acquisition, we obtained MS and MS/MS spectra directly from tissue from a MALDI-orbitrap instrument. At each MALDI spot, the top 5 most intense peaks were selected for MS/MS. No exclusion window was specified. We acquired 2118 MS/MS spectra from tissue.

5.2.2 Spectral Library Creation

We perform noise filtering, baseline filtering and peak detection on the resulting spectra using the OpenMS suite.

MS/MS data is stored and organized in a MySQL database. This database allows for the creation of custom spectral libraries by organism, tissue type, mass spectrometer, enzyme and other parameters. The database schema is shown in Figure 5.1. All acquired MS/MS spectra from protein extract samples, both raw and processed, are referenced in the database with their experimental information.

Peptide identification is done using Mascot [93] and MS-GFDB [94]. Identifications are stored in the database if the top hit had a Mascot E-value ≤ 0.1 or MS-GF spectral probability $\leq 10 \times 10^{-6}$. The reason for such permissive thresholds is that the database is filtered in an on-line manner.



Figure 5.1: Database model for spectral library. All spectra were inserted in the database, allowing user to search against both identified and unidentified spectra.



Figure 5.2: AMASS results for MSI data. a) optical image with MALDI spots. b) (left) segments resulting from an AMASS run with random seeds after 3 iterations. (right) log-odds images for the AMASS segments. We can see the different regions of the cerebellum.

5.2.3 Spectral Library Search

MS/MS spectra obtained directly from tissue are pre-processed using the same method as those acquired from protein extract. They are searched against a subset of database which is filtered in an on-line manner. The database can be filtered by organism, instrument type, and/or identification score threshold.

Similarity is measured using the cosine distance of the squared intensities. Top-similarity spectra are reported, as well as corresponding identifications. Due to the nature of the database, we can pull other interesting spectra corresponding to each hit for comparison purposes. These could be spectra having the same peptide identification, modified or unmodified, potentially from different organisms or acquired by different instruments.

5.3 Results

5.3.1 AMASS Run

As a first pass, we run AMASS on the imaging data. We show the results in Figure 5.2. The cerebellum is divided into 3 segments (first row) corresponding to morphological regions: the *arbor vitae* (left) which is white matter, and the granular layer (middle) and molecular layer (right) which consists of gray matter. We also find 2 clusters for matrix signal (second row).

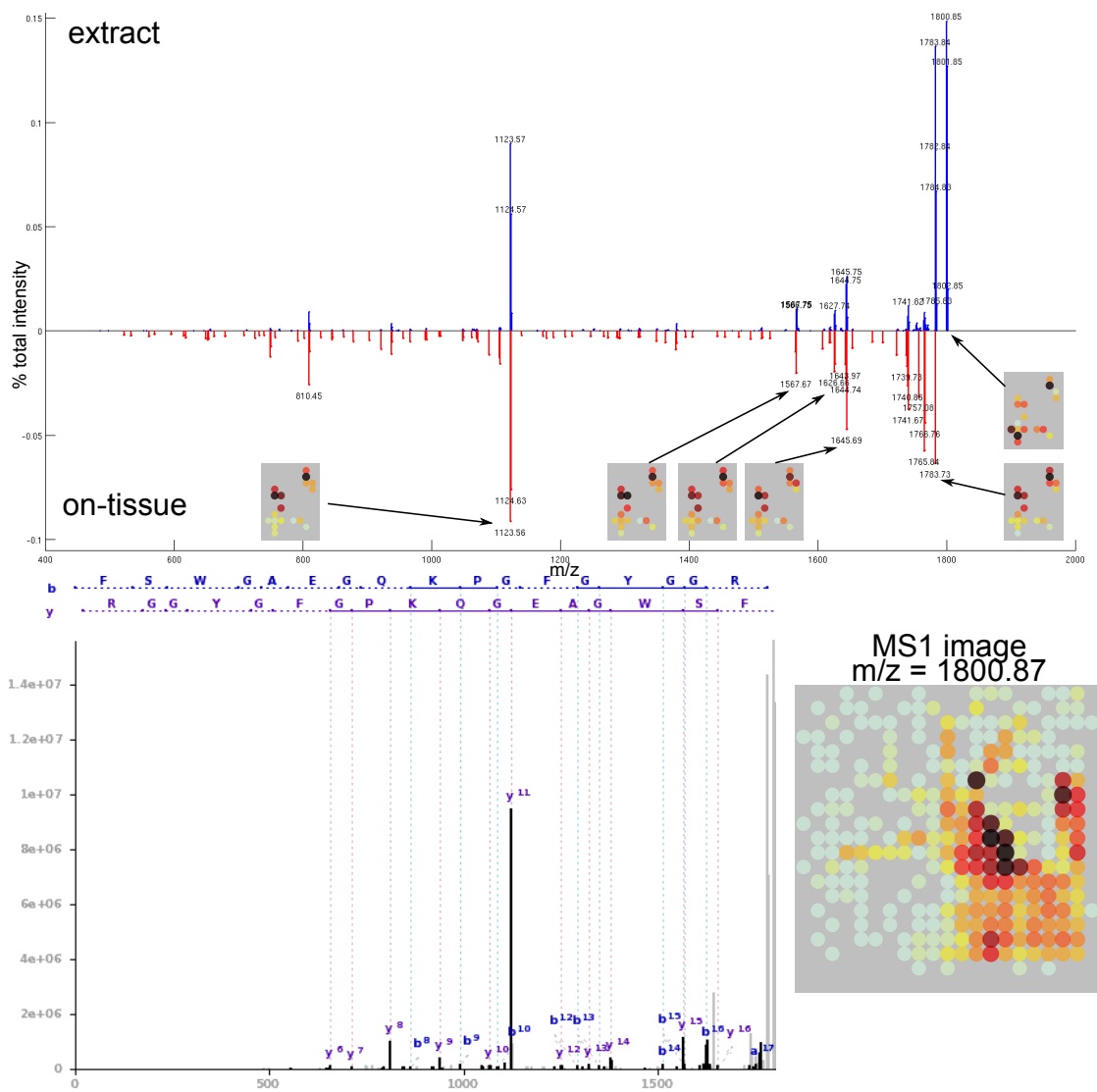


Figure 5.3: Spectral library hit for peptide from myelin basic protein. On the top we show the match for the MSI spectrum (on-tissue) and the database spectrum (extract). The peaks of the two spectra match well. The identification for the extract spectrum (bottom left) has a mascot e-value of 0.02 and MS-GF spectral probability of 7.85×10^{-15} . The MSI data shows localization in the *arbor vitae* (bottom right) which is white matter and thus rich in myelin, congruent with the identification.

5.3.2 Peptide Identifications

From the first MALDI dataset, we obtain 112 unique peptide identifications with MS-GF spectral probability $\leq 10 \times 10^{-10}$ or mascot score > 30 . From the second MALDI dataset, we obtain 132 unique peptide identifications with the same thresholds. While we have many identifications from the ESI-orbitrap dataset, we had no MSI data to compare it with. Searching the MALDI imaging MS/MS spectra against both the MALDI and ESI datasets in the database only yielded to MALDI vs. MALDI hits, emphasizing the need for a MALDI-specific spectral library.

5.3.3 Spectral Library Matches with Good Peptide Identification

Due to the small size of our database, we only identified 3 proteins with spectral library matches to peptides with “good” identifications (MS-GF spectral probability $\leq 10 \times 10^{-10}$).

A first hit is to a peptide from myelin basic protein MBP. In Figure 5.3, we show the match between the on-tissue and the extract MS/MS spectra. The peaks between the two spectra match well. However, while the parent ion has the highest intensity in the extract spectrum, it is completely fragmented in the tissue spectrum. We also generated the MS/MS fragment localization images for those spots at which the precursor ion $m/z = 1800 \pm 1$ was selected for MS/MS. The localization is almost identical for all fragment ions, except for the parent ion which is missing in the imaging spectrum. The extract spectrum was identified to be a fragment of myelin basic protein with mascot e-value of 0.02 and MS-GF spectral probability of 7.85×10^{-15} . The MSI shows localization in the *arbor vitae* which is white matter and thus rich in myelin, congruent with the identification.

Another hit is to a peptide from beta-globin, a protein from hemoglobin, which is known to be expressed in the rat brain [95, 96]. We show the match and the identification in Figure 5.4. Again, the peaks match well. The identification has a MS-GF spectral probability of 3.57×10^{-10} . The MSI data shows localization in the gray matter of the cerebellum, probably due to the fact that the gray matter is more cellular than the white matter and thus has more blood vessels.

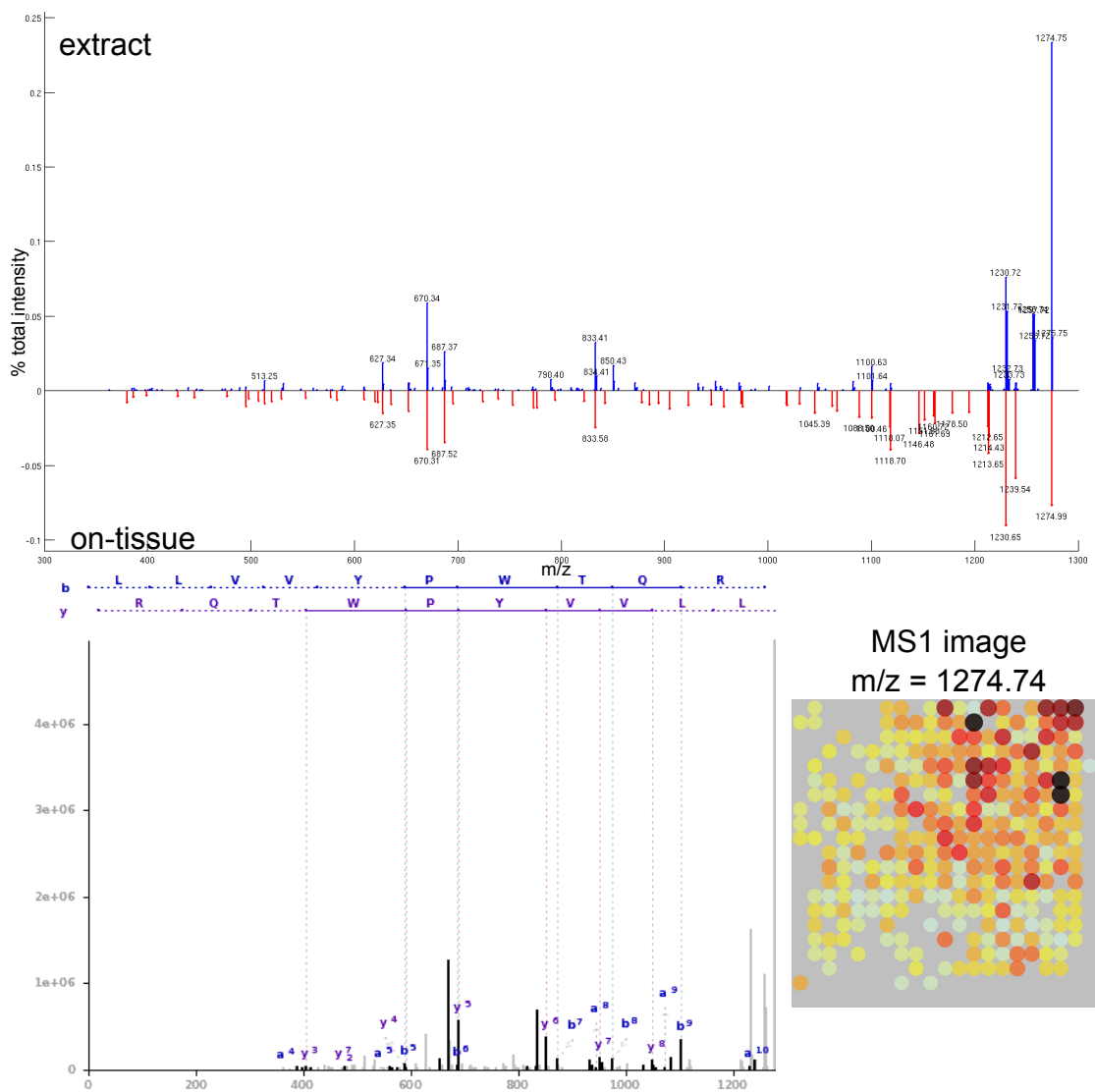


Figure 5.4: Spectral library hit for peptide from beta-globin. On the top, we show the match for the on-tissue and extract spectra. The peaks match well. The identification has a MS-GF spectral probability of 3.57×10^{-10} . The MSI data shows localization in the gray matter of the cerebellum.

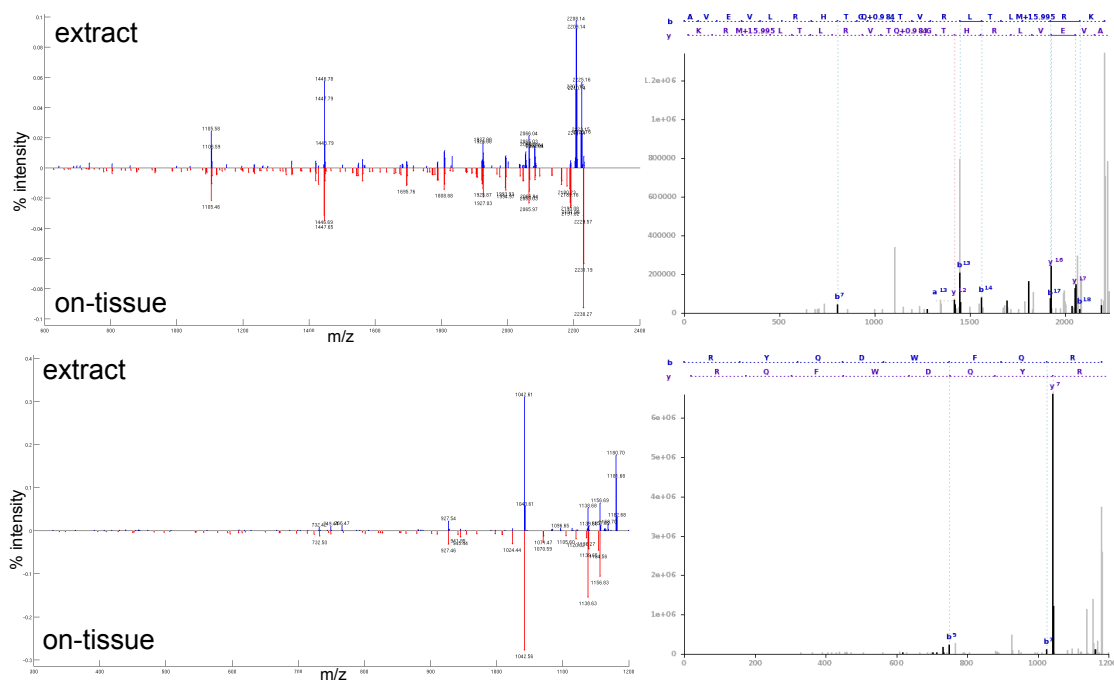


Figure 5.5: Other Spectral Library Matches. (top) Spectral library match to a spectrum identified as a peptide from a multiple PDZ domain protein. The identification got a MS-GF spectral probability of 4.35×10^{-7} . (bottom) Spectral library match to a spectrum identified as a peptide from the integrator complex subunit 3. The identification got a MS-GF spectral probability of 1.64×10^{-8} .

Finally, we found hits to two peptides from trypsin with MSI localization outside the sample. Because most of the signal should be matrix in that area, it makes sense for trypsin to be one of the highest peaks.

5.3.4 Other Spectral Library Matches

We had other hits to spectra with poor or no identifications. We show two such examples in Figure 5.5. One spectrum is identified as a peptide from a multiple PDZ domain protein with a MS-GF spectral probability of 4.35×10^{-7} (top). The other spectrum is identified as a peptide from the integrator complex subunit 3 with a MS-GF spectral probability of 1.64×10^{-8} . These results are preliminary and a more complete spectral library would yield in a better and more hits.

5.4 Discussion

We present here preliminary results for using spectral libraries as a mean to identify peptides from on-tissue MS/MS. Despite the fact that our current database is small, we still obtained some interesting hits. Several improvements on this database would greatly increase the number of hits and their quality. First, the current extract spectra were acquired via LESA. A whole rat brain sample would contain more material and yield to higher quality spectra, increasing the number of identifications in the database. Second, the number of spectra in the database is small, and needs to be increased. However, we are confident that the development of a complete high-quality MALDI-orbitrap spectral library would greatly increase the number of identifications from on-tissue MS/MS spectra, as most on-tissue MS/MS spectra do not have a high-quality enough for database search.

The development of an imaging spectral library would open many other possibilities. One interesting use of spectral libraries is the identification of mixture peptides. This would be particularly relevant to spectra acquired directly from tissue as there is no separation from an LC column. Wang et al. show that on average they can identify 15% more mixture spectra using a spectral library instead of a protein database [92, 97]. It is also possible to create other molecules to this database, as not all MSI focuses on peptides and proteins.

It is worth noting that there are other strategies for MSI identification. Recently, the development of nano-DESI [5] allows which uses an electrospray ionization source, and thus the acquisition of multiple-charge spectra. While DESI has low spatial resolution ($150\mu m$), it should theoretically be possible to acquire images in as high resolution as $12\mu m$ using nano-DESI. However, the lack of material in the sample and the peptide mixture problems will still be present, making spectral libraries a great asset. One great benefit is that there currently exists spectral libraries from ESI-orbitrap sources, thus reducing the need to acquire new data to build the library. However, the spectral search tools would still need to be modified to compare tissue versus extract spectra, and it will be crucial to consider mixture spectra.

Another approach is to use micro-dissection to identify spectra from a specific location on the tissue slice or organism. Peptides are extracted from a $300\mu m$ region

of the tissue and sent for identification through a high resolution ESI-orbitrap platform [98]. Because the peptide are extracted and a nano-LC column is used, the number of identification greatly increases. However, the process is labor-intensive and only peptides a few pre-selected areas are identified.

Finally, we should note that the information in the images could help to validate on-tissue identifications. Peptides from the same protein are likely to show the same distribution. While this is not enough to give an identification on its own, it should be incorporated in a search tool for MSI peptide identification.

Chapter 6

Conclusions

We first started to explore mass spectrometry imaging data in a supervised manner, asking if we could find and identify peptides or proteins specifically expressed in a pre-defined region of interest. In Chapter 2, we acquired a list of many mass values present in different regions of the leech by means of a statistical analysis. Using a middle-down approach and a novel identification pipeline, we identified several interesting peptides, including one encoded by a novel gene, HmIF4, a member of the intermediate filament family involved in neural development. We also performed a second validation via *in situ* hybridization of the corresponding mRNA transcripts.

We extended this approach in Chapter 3, in which we described a novel method, AMASS (Algorithm for MSI Analysis by Semi-supervised Segmentation), that automatically segments the MSI dataset into consistent regions of interest, and determines for each segment a molecular signature, or collection of peaks that are preferentially expressed in the segment. AMASS relies on the discriminating power of a molecular signal instead of its intensity as a key feature, uses an internal consistency measure for validation, and allows significant user interaction and supervision as options. We show that automated segmentation of a whole leech embryo dataset and or a rat brain dataset yield to segments congruent to known morphological features.

In Chapter 4, we take a step back and consider the case of many mass spectrometry imaging datasets, across different organisms, conditions, time scales, etc. We describe a method for automated large scale comparative analysis of these datasets. Given a set of pertinent query molecules, we find in each dataset all molecules that have a

similar spatial distribution to these molecules. We then cluster the datasets based on the resulting molecular signatures. By comparing the molecular signatures in specific clusters, we can confirm the existence of signal across replicates and identify signal changes for different conditions. This approach has the potential to identify unknown relationships between multiple data acquisitions.

Finally, in Chapter 5, we briefly touched the problem of peptide identification in mass spectrometry imaging. In Chapter 2, we employed a middle-down approach to identify several peptides with interesting distribution. Here, we take a bottom-up approach and show how using spectral libraries can greatly improve peptide identification from on-tissue MS/MS data. We presented some preliminary results to highlight the potential of this approach and discussed its other benefits.

The use of these different approaches constitute a preliminary set of tools for the analysis of mass spectrometry imaging data to enable us to further our understanding of biological systems and discover new biomarkers to known diseases. However, the presented algorithms are just the tip of the iceberg and much work still has to be done to mine MSI data to its fullest.

Mass spectrometry imaging could make great use of a centralized repository for the data. Currently, there is no repository for MSI data and it is kept on the hard drives of the various labs that generated it. A centralized repository would allow for the MSI field to become a more collaborative effort. It would allow for labs to compare their data to other experiments without needing to repeat them. Integration of a repository with web-based analysis tool kit would allow users to analyze and possibly share their data simultaneously. Integrating a comparative analysis tool would also allow automated comparison to all shared datasets, allowing the community to explore all available data. There are several barriers to establishing such a repository. First, the datasets are typically very large, and thus great bandwidth and great storage space is required. Second, a universal data format must be adopted. Rompp et al. have developed the imzML format [99], but tools to convert from vendor format into this format are still lacking, though will hopefully soon be incorporated into vendor software.

Another interesting challenge is the co-registration of MSI datasets to optical images and to each other. Co-registration to the optical image is currently done manually

by the definition of several teaching points on the optical image and in laser coordinates. Human error causes small shifts in registration. It becomes particularly challenging in the case of very high definition optical images, where a most minor error will greatly offset the MSI images from the optical image. Because the MSI data follows the boundaries of the optical image, co-registration can and should be done automatically. This can be taken a step further, in which MSI datasets can be co-registered to optical images from other related datasets and to other MSI datasets. This would greatly increase the power of comparative analysis, as each region of interest could be mapped to that of other datasets based on not only the mass spectrometry data, but also the optical data. Similarity and differences in molecular signatures could then be spotted as previously described.

Finally, new approaches to mass spectrometry imaging, such as three-dimensional and quantitative mass spectrometry imaging, have allowed for a surge in new types of datasets with additional information. New methods need to be developed to automatically visualize and analyze this data.

This dissertation provides a set of computational tools to analyze mass spectrometry imaging data, but it is only a first pass to a much greater set of methods yet to be developed.

Appendix A

Supplemental: Automated querying and identification of novel peptides using MALDI mass spectrometric imaging

Note: The quality of some figures was reduced from the original publication.
Original supplemental material is available free of charge at <http://pubs.acs.org>.

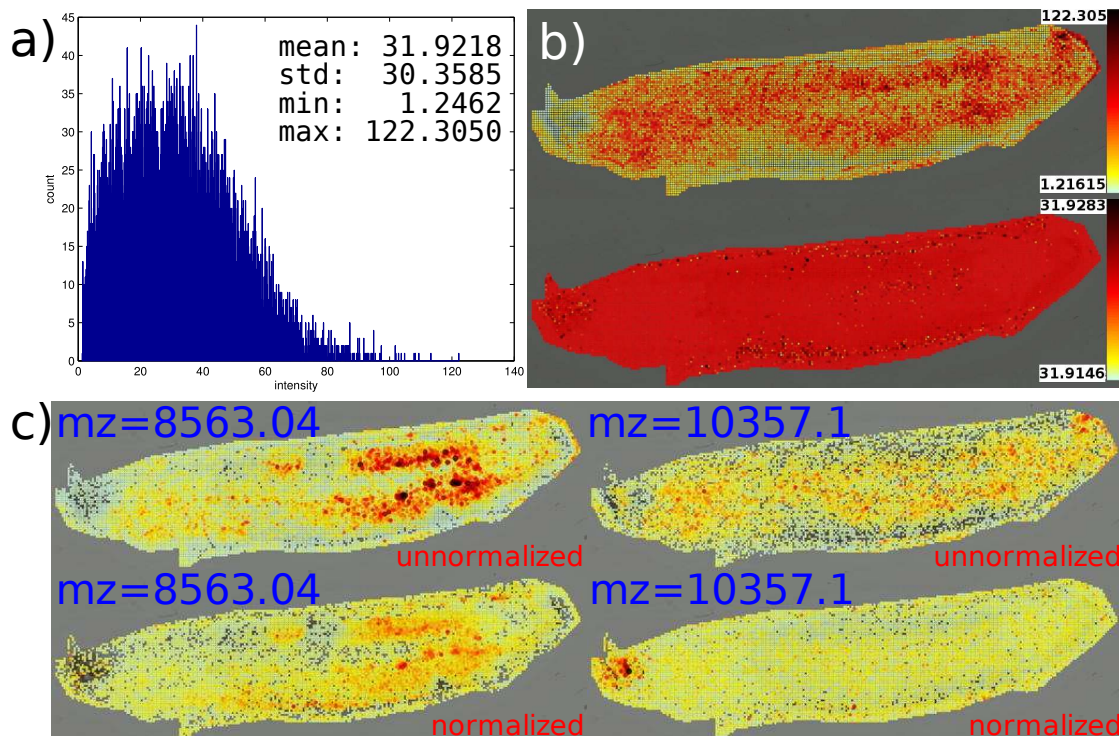


Figure A.1: Effects of normalization on data. a) Average peak intensity for spots before and after normalization. There is a spatial bias for the posterior region in the original data. Spatial bias is removed by normalization. b) Distribution of spot average peak intensities. The average intensity for each laser spots differ greatly. Lack of homogeneity in crystallization can cause this spatial bias. c) Resulting changes in spatial distributions. In both unnormalized images, we see the same bias as in a). At $m/z = 8563.04$, the intensities are more evenly distributed throughout the nervous system are distributed after normalization. At $m/z = 10357.1$, normalizing the data allowed us to see a signal in the head of the leech.

Table A.1: Over-expressed masses at threshold 0.65 for the CNS ROI in LEECHE12A.

<i>m/z</i> -range	peak	score	<i>m/z</i> -range	peak	score
2470.68-2481.44	2474.34	0.80	6186.44-6186.8	6186.44	0.65
2505.32-2508.31	2506.7	0.69	6187.53-6200.57	6196.95	0.68
2521.69-2527.93	2524.69	0.78	6212.91-6212.91	6212.91	0.65
3090.6-3096.75	3093.93	0.71	6216.54-6220.89	6217.99	0.66
3157.97-3165.99	3160.3	0.71	6223.8-6223.8	6223.8	0.65
3295.94-3303.86	3299.64	0.76	6593.09-6593.46	6593.09	0.65
3485.31-3496.45	3491.01	0.80	6594.58-6595.33	6595.33	0.66
3496.45-3531.63	3511.43	0.90	8205.81-8205.81	8205.81	0.65
3536.01-3542.31	3540.11	0.71	8207.89-8216.66	8213.32	0.66
3651.01-3651.57	3651.29	0.65	8217.49-8218.33	8218.33	0.65
3652.4-3652.4	3652.4	0.65	8220-8220	8220	0.65
4211.19-4214.18	4211.79	0.67	8223.76-8223.76	8223.76	0.65
4371.25-4383.43	4378.86	0.69	8399.29-8400.14	8399.72	0.66
4523.78-4529.97	4526.88	0.69	8400.98-8447.91	8428.45	0.88
4539.58-4547.64	4542.06	0.70	8448.76-8449.61	8449.18	0.66
5268.46-5278.82	5273.47	0.76	9032.26-9041.02	9038.39	0.74
5279.83-5280.16	5279.83	0.66	9224.17-9249.86	9240.11	0.70
5410.4-5429.38	5417.85	0.81	9250.74-9257.83	9253.4	0.67
5430.06-5432.43	5431.41	0.68	9779.68-9779.68	9779.68	0.65
5434.81-5436.5	5435.48	0.68	9780.59-9781.05	9780.59	0.65
5560.06-5606.15	5574.15	0.90	10868.4-10869.8	10869.4	0.67
6024.92-6024.92	6024.92	0.65			
6025.63-6026.7	6026.7	0.66			

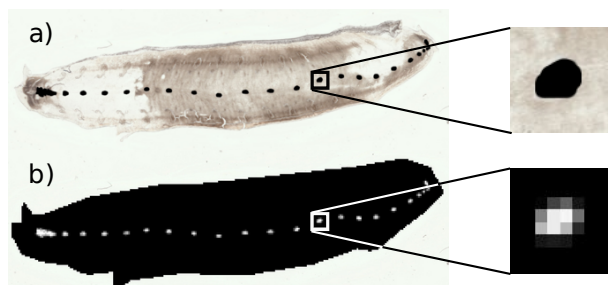


Figure A.2: Mask (top) and query (bottom) for the CNS in the LEECHE12A. We can see in the query that a MALDI spot covers about a 10x10 pixels square on the histological image; thus, the query is a set of gray squares instead of the binary black and transparent pixels of the mask.

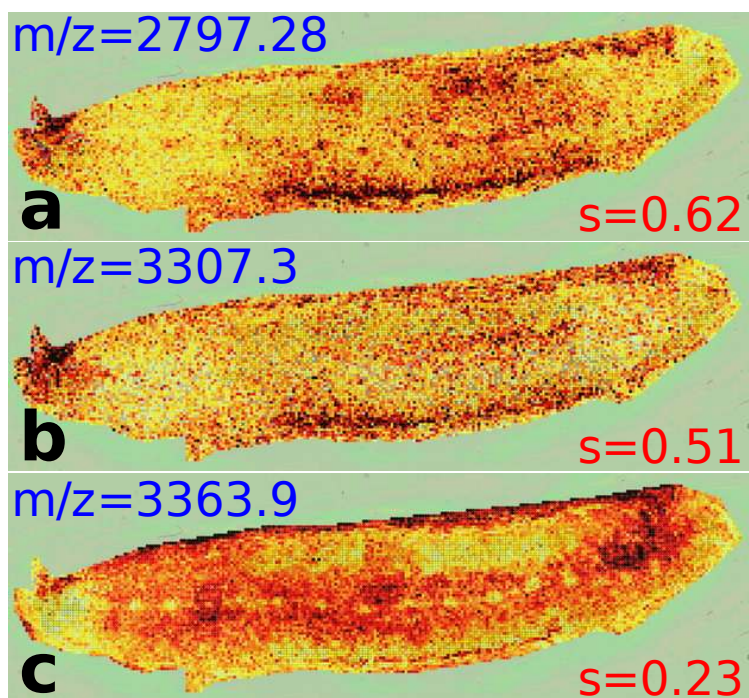
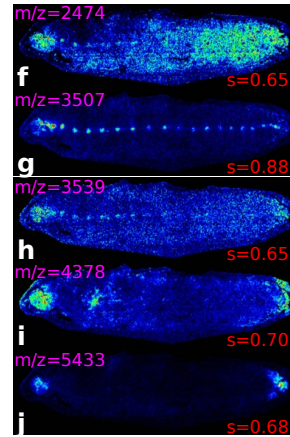
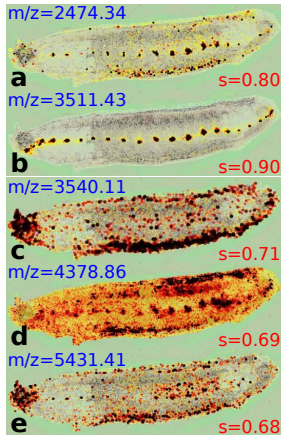


Figure A.3: MALDI images for CNS localization at different scores. a) At score $s = 0.62$, just below our cut-off of 0.65, we still see localization to CNS, but noise significantly impairs the signal compared to the top-ranked images. b) At score $s = 0.51$ we expect no localization. Indeed, the intensities outside the nervous system almost perfectly balance out the intensities within the CNS. c) At score $s = 0.23$, we are at the other end of the spectrum. We detect a molecule which have inversed expression to the CNS; the molecule is highly expressed in the ventral region of the leech but shows distinct under-expression in the ganglia and the brain.

Table A.2: Over-expressed masses in the leech CNS across different samples. We displayed all m/z values in LEECHE12B with score $s \geq 0.65$, and all m/z values with score $s \geq 0.70$ in LEECHE12A. We only showed m/z values with score $s \geq 0.65$ in LEECHE12A if it corresponded to a hit in LEECHE12B as the full table is quite extensive (Table A.1). Results for $m/z < 2200$ were discarded because spectra were still in the noise area.

LEECH12A		LEECH12B	
m/z peak	score	m/z peak	score
2474.34	0.797	2474.34	0.653
2479.6	0.727		
2524.69	0.78		
3093.93	0.705		
3160.3	0.708		
3299.64	0.76		
3491.01	0.797		
3511.43	0.896	3507.89	0.882
3523.98	0.713		
3540.11	0.71	3539.84	0.654
4378.86	0.691	4377.95	0.697
4542.06	0.702		
5273.47	0.758		
5417.85	0.814		
5426.66	0.708		
5431.41	0.675	5433.79	0.681
5574.15	0.899		
5586.87	0.805		
5601.33	0.717		
8428.45	0.899		
9038.39	0.736		
9240.11	0.701		



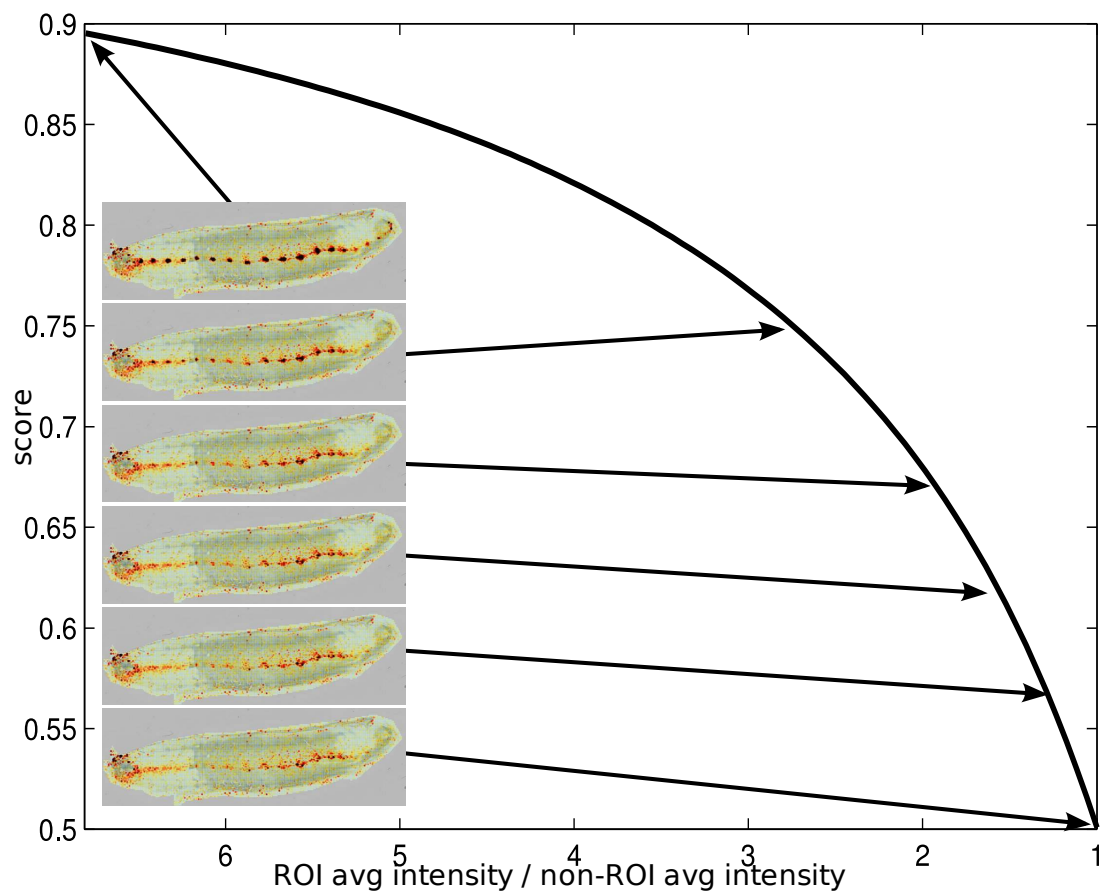


Figure A.4: Simulation results decreasing ROI signal over the entire region. Intensities are decreased in the ROI spots by a certain percentage until the average intensity inside the ROI is the same as the average intensity outside the ROI. Score and signal decrease in a similar fashion. Similar intensities in the ROI and in the background lead to a score close to 0.5 as expected.

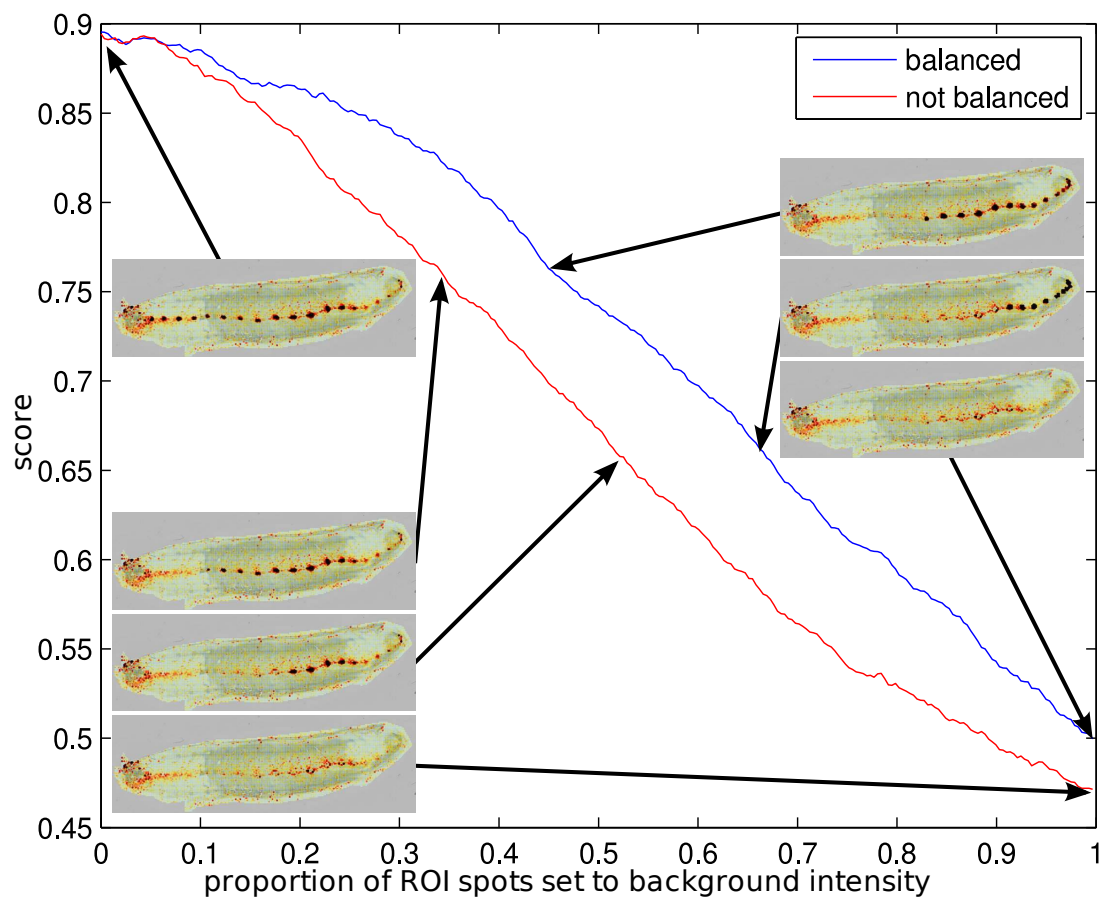


Figure A.5: Simulation results when degrading the signal in the ROI. We set a proportion of the ROI spots to have random non-ROI (or background) intensities (see Methods). It is then possible to balance the total ROI intensities by distributing the subtracted intensity to the remaining spots; that way, the total intensities in ROI and outside ROI remain the same throughout the simulation. Results are shown for two simulated runs: with and without balancing the ROI intensities. In both cases, the score linearly decreases as more ROI spots are set to background intensity. In the balancing case, the intensities of the remaining spots increase to compensate for the other spots; consequently, the score remains higher in the balanced case than in the unbalanced case, as expected. When all spots are set to background intensities, the signal is lost and the probability score decreases to 0.5 as expected.

```

CLUSTAL 2.0.12 multiple sequence alignment

leech_EST  --TFTEMERTRVNTTYTSSGVDGVDDGASTFRES--SYGAGGPITG 46
gliarin    --MAEEVITKTRRVYKTEVSGEGSSLLATTYR-----PSVT 37
filarin    -----MESGNFAEYERTIRSN-----I0 19
macrolin   ARQIEPLGMDTRIDKSSVNSGQKVVSTKSMWDLDTGSTYTKATVV 50
          * . . . . .

leech_EST  GRTLIVSRIGTGSPPMGSGIGLGGTRMERSVRTSSQVASGGPPNYSV 96
gliarin    PRNVIHRSDIAPLGSMSHSS----TIRREKTIQYGNAYALSP--SSYAP 81
filarin    PRNLIQRATPGAFNVSRS-----VTRSVGVNYGAGG----- 54
macrolin   PHOLIIORTLTGGLSSGGGSL-----RSTADFRFSVMVPGVH----- 88
          : : * . . . . .
          >rod domain

leech_EST  ITATGVSGIKESRDQEKMDQDLNERFANYIDKVRNLEAQRKLAEDLSR 146
gliarin    LASSGVSVKNSREREKMDQDLNERFASYIEKVRFLAQNKRITDELDK 131
filarin    VAGGAATSYTDQRNKEKREMDLNERFAGYIEKVRFLAQNKKLADELDA 164
macrolin   LATKEVDSARTTRRREKMDQDLNRLTRYIETVRFLAQNKQDLNETKT 138
          : : . . * : : : * : : * : : * : : * : : * : : :
          >rod domain

leech_EST  LKEKWKDTPVQKAMFQVDLDCRHLQDEAEKEKARLETRLASLEEEED 196
gliarin    LKSRWKDTPQIKAMFQVLDIARRLLDDEKEKARLETKIASLEEINEE 181
filarin    LKSRWKDTPVQKAMFQVLDLQVQKRLDDCEKETORLQIQVASHKEKDD 154
macrolin   LKAKWKETSQVRAMFADLEEARIKDDELEKTDKAKLEIRISSVIEALDV 188
          ** : : * : : * : : * : : * : : * : : * : : * : : * : :
          * : : * : : * : : * : : * : : * : : * : : * : :

leech_EST  LRQELAAQQLSENQAFISKNNQLLDVSEIQLGRKRRIEQLNEKERD 246
gliarin    LAVKLEALQTNIEQRKIDRONQQLSDYEGEISLLRRRVEGLEADKDD 231
filarin    LRRKLEAANAADVSRDKLEKIQQIAEIQSEVHLRLRSLDLDGKRYN 204
macrolin   EKRRNATSEKTIIEYREKIQENQRQLVDLQANNLLQRLELLEGGDRDR 238
          : : * : : * : : * : : * : : * : : * : : * : : * : :

leech_EST  KKNIAQLKELAKARQDLNETLEHIAENRCOTLOEETDFLKSINHEQEM 296
gliarin    RKTITATLNAALNARANIDDETLRHIDAEINRRQTLLEELFLKSVHEQEL 281
filarin    KAILSKLQENLRRTDFDQAQVEHDAEAKRLALEELAFIKLQVHEQEL 254
macrolin   KKLVGELKAVTRYRTDLSQTLVYDAADRQSLLEELDFIKQVHEQEM 288
          : : . * : : * : : * : : * : : * : : * : : * : : * : :
          : : * : : * : : * : : * : : * : : * : : * : : * : :

leech_EST  KELAALAYRDTAPE-RDYWKHMAQALREIQEYDDKFDOSIRTEETHYT 345
gliarin    KELAALAYRDTTENDRFPWKEMGNALREIQEYDEKLDLNRTEIESSY 331
filarin    RELAAKAYFDSTASNREYKSEMSHKLQEHYGEKIDELQNEHSLNYS 304
macrolin   KELNILIKDYSIVNRQYKTEMERALKIQDLYDELSDMRDETFYQ 338
          : : * : : * : : * : : * : : * : : * : : * : : * : :
          : : * : : * : : * : : * : : * : : * : : * : : * : :

```

Figure A.6: ClustalW alignment of the HmIF4 protein sequence with those of three other known intermediate filaments in *Hirudo medicinalis*. The EST open reading frame aligned particularly well in the conserved rod domain, and has more variability outside of that domain. The peptide we identified is located in a variable region in the 5' end of the rod domain where the sequences are quite dissimilar, thus confirming the discovery of a novel protein.

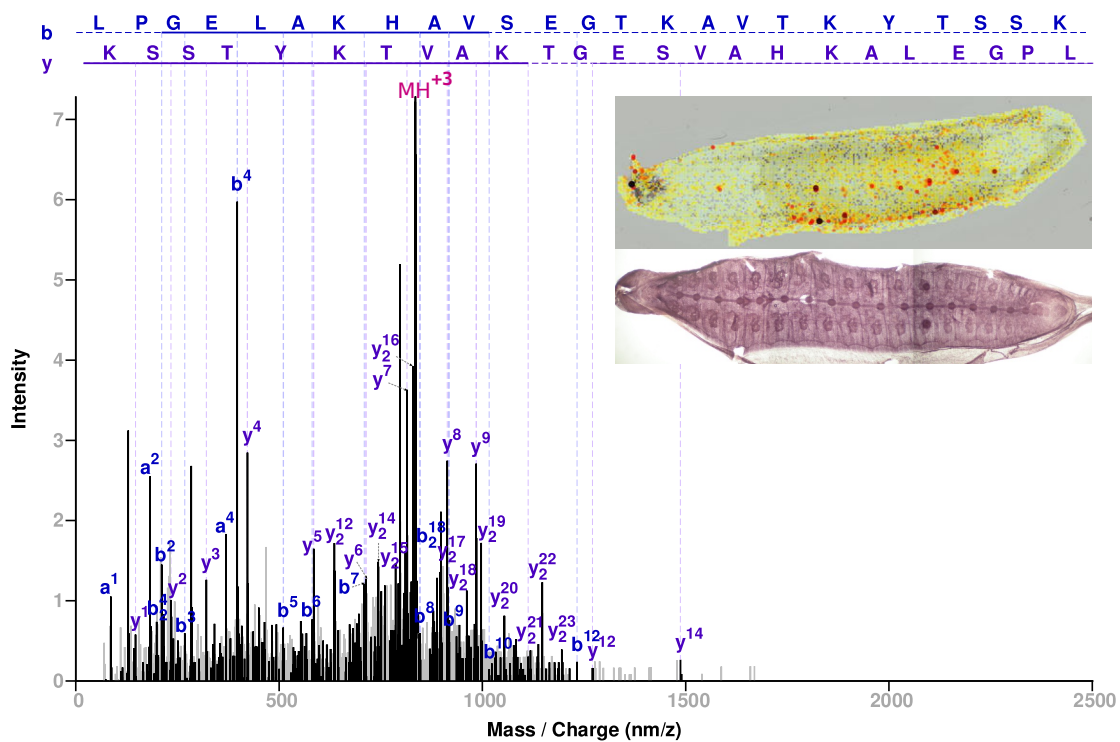


Figure A.7: Annotated spectrum for a peptide from the histone H2B. Parent mass \simeq 2500 Da was shown to have a CNS specific expression. *In situ* hybridization of the mRNA shows a preferential location in the CNS, but with a relatively weaker signal.

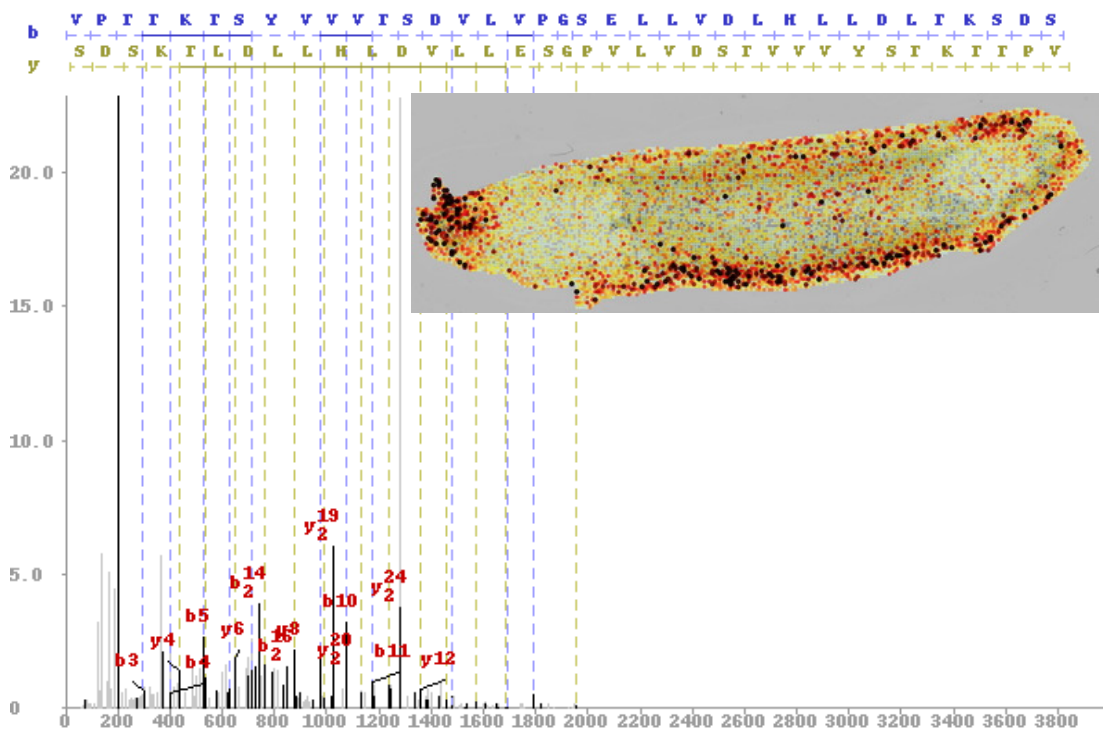


Figure A.8: Annotated spectrum for uncharacterized peptide. Parent mass $\simeq 3841$ Da was shown to have a dorsal specific expression similar to another fragment 2 amino acids shorter showing in Figure 2.5.

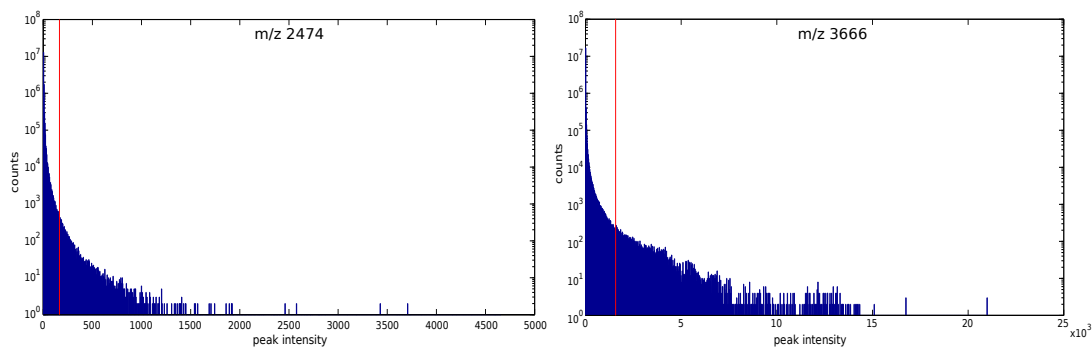


Figure A.9: Distribution of the MS1 raw data peaks for 2 experiments. Intensity of top peaks (162.2 and 1526.5) for annotated spectra are indicated by red line.

Table A.3: Identification of several other peptides. Peptides were identified using In-spect with an FDR cut-off of 0.01. Biological annotations were achieved by doing a Blast search of the protein sequence against NCBI nr and keeping the top hit if the E-value was less than or equal to $1e-5$.

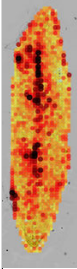
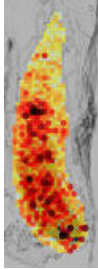
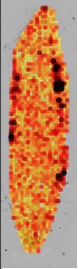
parent mass	annotation	signal	blast top hit ($E \leq 1e-5$)	image
1353.81	G.ASGGIGQPLSLLK.L	CNS	gi 75040807 sp Q5NVR2.1 MDHM_PONAB RecName: Full=Malate dehydrogenase, mitochondrial	N/A
1993.94	T.SYIEDFDVSTLPEHQLTG	CNS	N/A	
2107.00	L.SLWANNSEKINFQ-17LDGNSS.R	ventral	gi 22095553 sp Q9ROM0.2 CEL2_MOUSE RecName: Full=Cadherin EGF LAG seven-pass G-type receptor 2; AltName: Full=Flamingo homolog;	
1829.91	K.SDQVHGLFSVNVDRK.C	ventral	gi 117949389 sp Q6YHK3.2 CD109_HUMAN RecName: Full=CD109 antigen; AltName: Full=150 kDa TGF-beta-1-binding protein; AltName: Full=C3 and PZP-like alpha-2-macroglobulin domain-containing protein 7;	
1940.00	L.FFTQFTPLFFKPGLSY.V	lateral	N/A	
1918.04	N.ILSVGIGSAPYMSPQLTAL.A	dorsal	gi 189082905 sp A2AX52.2 C06A4_MOUSE RecName: Full=Collagen alpha-4(VI) chain;	

Table A.3: Identification of several other peptides. (continued)

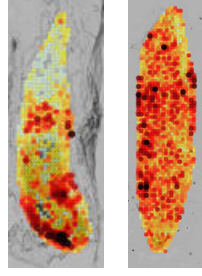
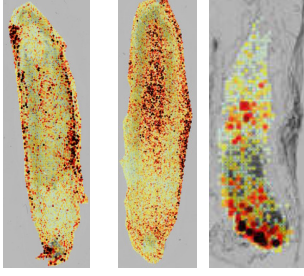
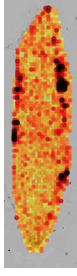
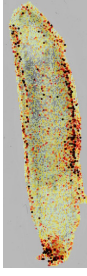
1633.84	L.FFSQFTPAFFKPGGL.T	dorsal	gi 117949389 sp Q6YHK3.2 CD109_HUMAN RecName: Full=CD109 antigen; AltName: Full=150 kDa TGF-beta-1-binding protein	
2598.29	R.HVADIRADDFSKEGQGVIGLQAG TN.Q	dorsal	gi 584954 sp Q08093.1 CNN2_MOUSE RecName: Full=Calponin-2; AltName: Full=Calponin H2, smooth muscle;	
1916.89	R.NVAEVPNVADENDFPSSL.I	dorsal	gi 52783213 sp Q9CY58.2 PAIRB_MOUSE RecName: Full=Plasminogen activator inhibitor 1 RNA-binding protein; AltName: Full=PAI1 RNA-binding protein 1;	
2977.39	N.GRKLLEDDEVPDLVENFDEASKT EVN.M	dorsal	gi 66774043 sp Q64152.3 BTF3_MOUSE RecName: Full=Transcription factor BTF3; AltName: Full=RNA polymerase B transcription factor 3	

Table A.3: Identification of several other peptides. (continued)

2456.10	D.ASNQDTTTCIYISSRYENYLN.A	dorsal	g1 88909565 sp Q64449.2 MRC2_MOUSE RecName: Full=C-type mannose receptor 2; AltName: Full=Lectin lambda; AltName: Full=Macrophage mannose receptor 2;	
1270.67	R.GEEVVLEVHLF.N	dorsal	g1 259016204 sp Q8IZJ3.2 CPMD8_HUMAN RecName: Full=C3 and PZP-like alpha-2-macroglobulin domain-containing protein 8	
2842.66	M.TKNHLNASMIKTKIPGLITLYL.L	dorsal	N/A	
2467.33	R.PNLAPVTRDVLVTFVDQSHVLF.F	dorsal	N/A	
1869.07	S.YVGQVLLKPVDEVLP.LS.S	dorsal	N/A	
2447.36	K.PNLVVPVTRDVLVTFVDQSHVLF.F	dorsal	N/A	

Table A.3: Identification of several other peptides. (continued)

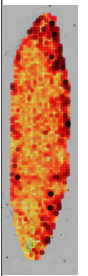
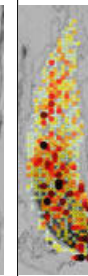
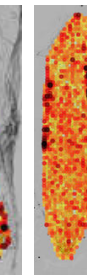
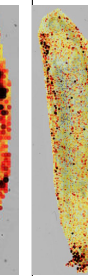
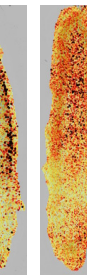

1887.98	*.AKFEDKESGKVVTAHLE.E	dorsal	N/A	
1154.63	T.KRKYYVAFDQ.K	dorsal	N/A	
1918.01	L.QLWFTCLTFHGHQ+1PI.N	dorsal	N/A	
2949.65	N.LRFVRRSSLRQHLYLRKFD W+16T.C	dorsal	N/A	
1735.87	L.FFSQFTPVFFKPGGL.S	not specific	gi 81879137 sp Q8R422.1 CD109_MOUSE RecName: Full=CD109 antigen; AltName: Full=GPI-anchored alpha-2 macroglobulin-related protein	
1661.87	E.FTPVFFKPGLSYVG.Q	not specific	gi 6225157 sp Q15417.1 CNM3_HUMAN RecName: Full=Calponin-3; AltName: Full=Calponin, acidic isoform	
1232.64	C.ASQAGMTAIGAVR.H	not specific	gi 6225157 sp Q15417.1 CNM3_HUMAN RecName: Full=Calponin-3; AltName: Full=Calponin, acidic isoform	

Table A.3: Identification of several other peptides. (continued)

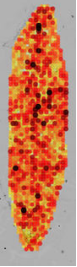
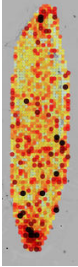
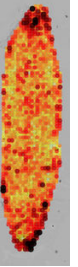
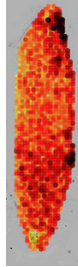
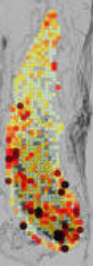
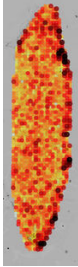
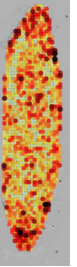
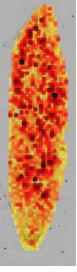
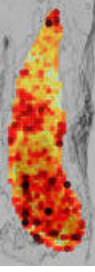
1967.99	L.FFSEFTPQFFKPGLLY.V	not specific	gi 259016204 sp Q8IZJ3.2 CPMD8_HUMAN RecName: Full=C3 and PZP-like alpha-2-macroglobulin domain-containing protein 8	
1957.96	L.VQNVPEFNEAVPSSRPTS.V	not specific		
1673.78	R.SFAPMSLKDASKDNY.F	not specific	N/A	
1814.90	R.CQSFLKSIL.SHCQAH.L.V	not specific	gi 55583881 sp Q68Y21.1 GRID2_DANRE RecName: Full=Glutamate receptor delta-2 subunit; Short=GluR delta-2 subunit;	
1318.67	R.GEEFVLEVHLEN	not specific	gi 117949389 sp Q6YHK3.2 CD109_HUMAN RecName: Full=CD109 antigen; AltName: Full=150 kDa TGF-beta-1-binding protein	
1842.90	L.VQNVPEFNEAAPSSRPTS.S			
1824.17	E.VVLKPKKFEVTVPLK.T			

Table A.3: Identification of several other peptides. (continued)

2073.02	G.YDERQQSSAVETKGFKL.R	not specific	<p>gi 215273864 sp Q08174.2 PCDH1_HUMAN RecName: Full=Protocadherin-1; AltName: Full=Cadherin-like protein 1; AltName: Full=Protocadherin-42; Short=PC42;</p> <p>CHECK</p>	
1284.58	N.QQ-17PVHMGMPYN.N	not specific	<p>gi 55584092 sp Q08473.3 SQD_DROME RecName: Full=RNA-binding protein squid; AltName: Full=Heterogeneous nuclear ribonucleoprotein 40;</p>	
1168.53	S.VTSSSSASIGEGS.R		<p>gi 205371790 sp Q9P2Q2.3 FRM4A_HUMAN RecName: Full=FERM domain-containing protein 4A</p>	N/A
1082.66	Q.LLGLGINIGVN.I		<p>gi 34921426 sp O96790.2 DPGN_DIPMA RecName: Full=Serine protease inhibitor dipetalogastin; (E = 5.7e-04)</p>	N/A
824.32	D.DGPFCAIN.G		<p>gi 1346041 sp P47931.1 FST_MOUSE RecName: Full=Follistatin; Short=FS; AltName: Full=Activin-binding protein;</p>	N/A

Appendix B

Supplemental: AMASS – Algorithm for MSI Analysis by Semi-supervised Segmentation

Note: The quality of some figures was reduced from the original publication.
Original supplemental material is available free of charge at <http://pubs.acs.org>.

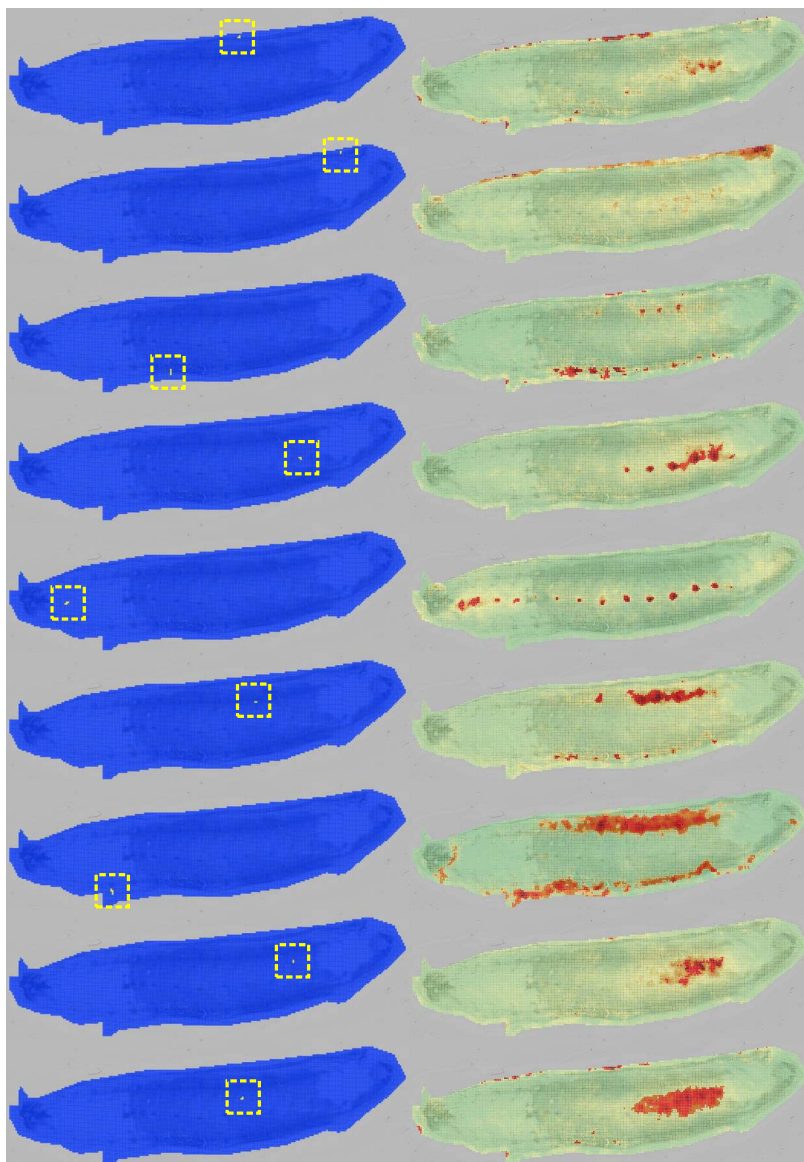
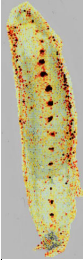
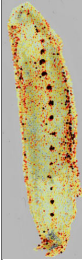
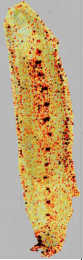
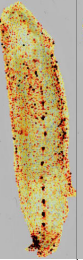
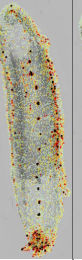
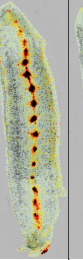
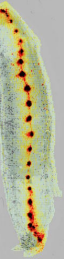

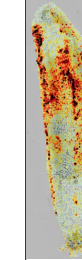
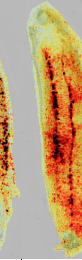
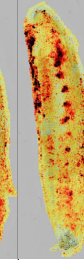
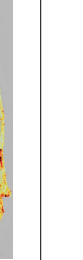


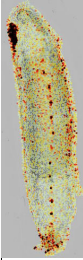
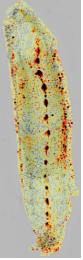
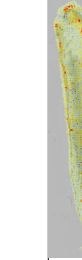
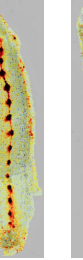
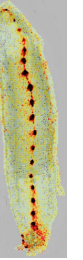

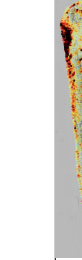
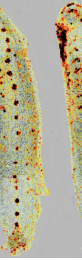
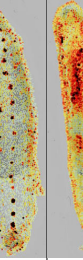
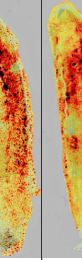
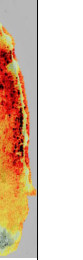





Figure B.1: Querying with small random seeds results in meaningful regions. Unlike the case of user-defined queries, many of the queries lead to similar results and some queries lead to lower quality results. Shown here are a few hand-selected random results. While the resulting log-odds images are in general not as specific as their user-defined counterpart, they still highlight different regions with specific molecular signatures. We can also expect the regions to gain specificity on the next iteration of the algorithm.

Table B.1: Molecular signatures for anterior and posterior ganglia. Specific molecules from the same morphological segment sometimes have slightly different signatures. We look at the molecular signatures of two queries: anterior and posterior ganglia. While many molecules are expressed throughout the CNS, some m/z values (3299, 4377, 5293) have differentiated signatures, consistent with the rostricaudal gradient of leech.

ganglia 2-4		ganglia 13-15		images
m/z	weight	m/z	weight	
		2473.88	0.82	
		2475.03	0.77	
		2476.86	0.76	
		2478.69	0.78	
2523.77	0.75	2523.77	0.90	
2525.62	0.75	2524.92	0.73	
3091.63	0.77	3095.98	0.73	
3299.11	0.80			
3300.69	0.78			
		3488.03	0.77	
		3490.74	0.84	
3494.28	0.73	3497.00	0.82	
3501.62	0.88	3502.71	0.96	
		3504.35	0.95	
		3505.98	0.95	
3510.61	0.98	3511.16	0.93	
3511.70	0.98			
3512.79	0.98			
3513.89	0.97			
3514.98	0.97			
3516.07	0.93			
3517.16	0.88			
3518.25	0.75			
		3652.40	0.74	
		3654.07	0.72	
4006.51	0.74			
4008.26	0.81			
		4374.30	0.80	
		4376.12	0.82	
		4377.95	0.82	
		4379.17	0.75	
ganglia 2-4		ganglia 13-15		images
m/z	weight	m/z	weight	
5273.47	0.76			
		5414.12	0.78	
		5416.16	0.77	
5418.19	0.82	5418.19	0.80	
5422.26	0.78	5419.21	0.71	
5425.31	0.70			
5564.53	0.90	5567.96	0.94	
		5570.37	0.96	
		5571.40	0.94	
		5572.43	0.92	
5573.80	0.98	5573.46	0.82	
5574.83	0.98	5574.49	0.78	
5575.86	0.98	5576.21	0.71	
5576.90	0.97			
5577.93	0.93			
5578.96	0.90			
5581.71	0.93			
5582.74	0.87			
5587.90	0.82			
8423.80	0.90	8428.03	0.89	
8427.60	0.90	8429.29	0.80	
8431.41	0.93			
8432.68	0.89			
8433.95	0.88			
8435.22	0.77	8561.77	0.74	
		9461.98	0.77	
		9463.33	0.74	

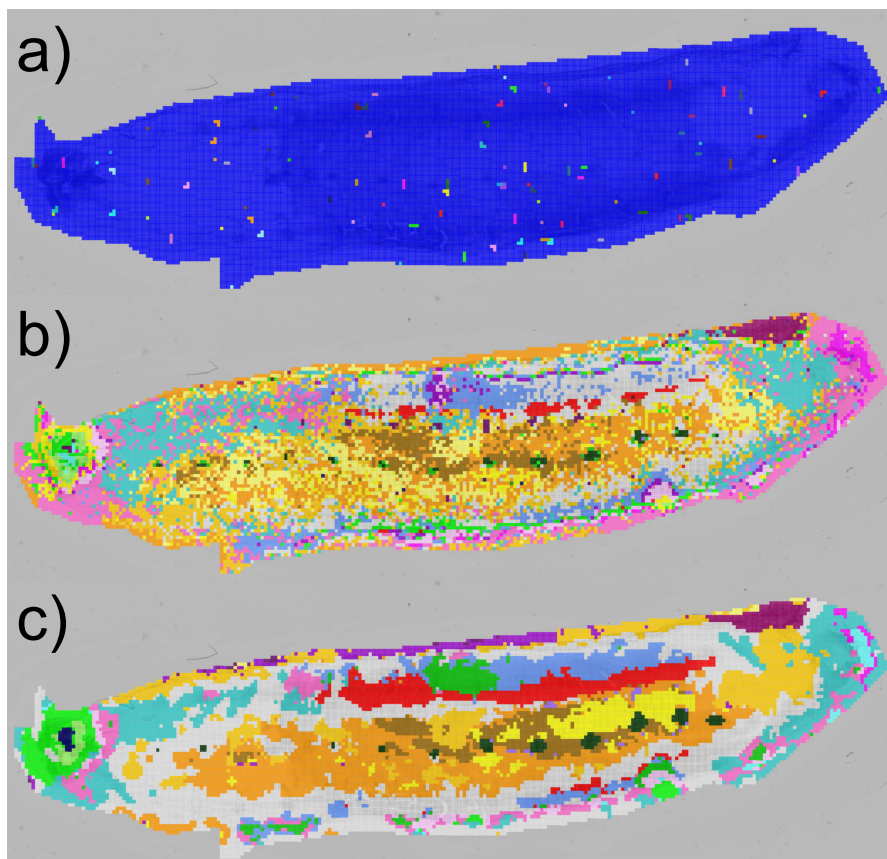


Figure B.2: Segmentation results with and without smoothing after 10 iterations for random initial random segmentation in leech. a) Initial random segmentation. b) Resulting segmentation map without smoothing. c) Resulting segmentation map with 3x3 median smoothing.

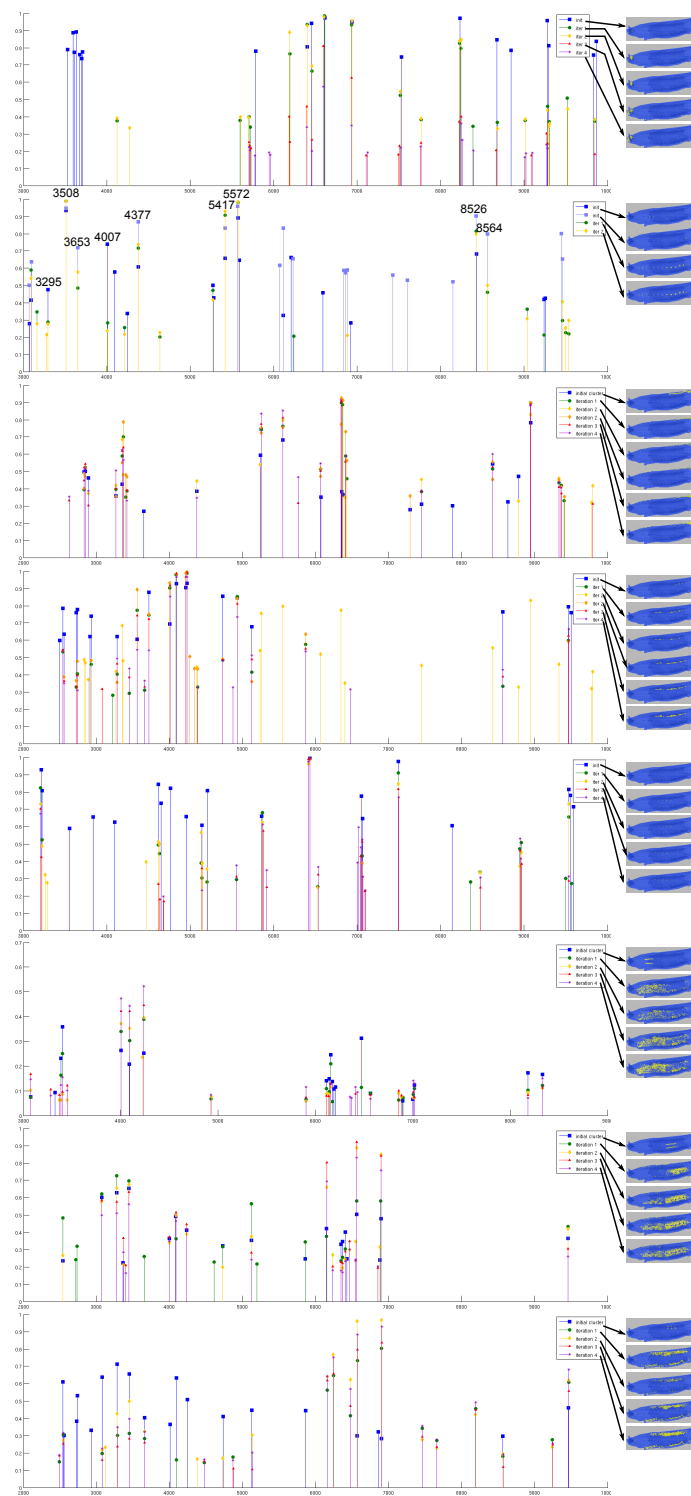


Figure B.3: Top 20 score peaks at least 10 Daltons apart for segments in leech at successive iterations.

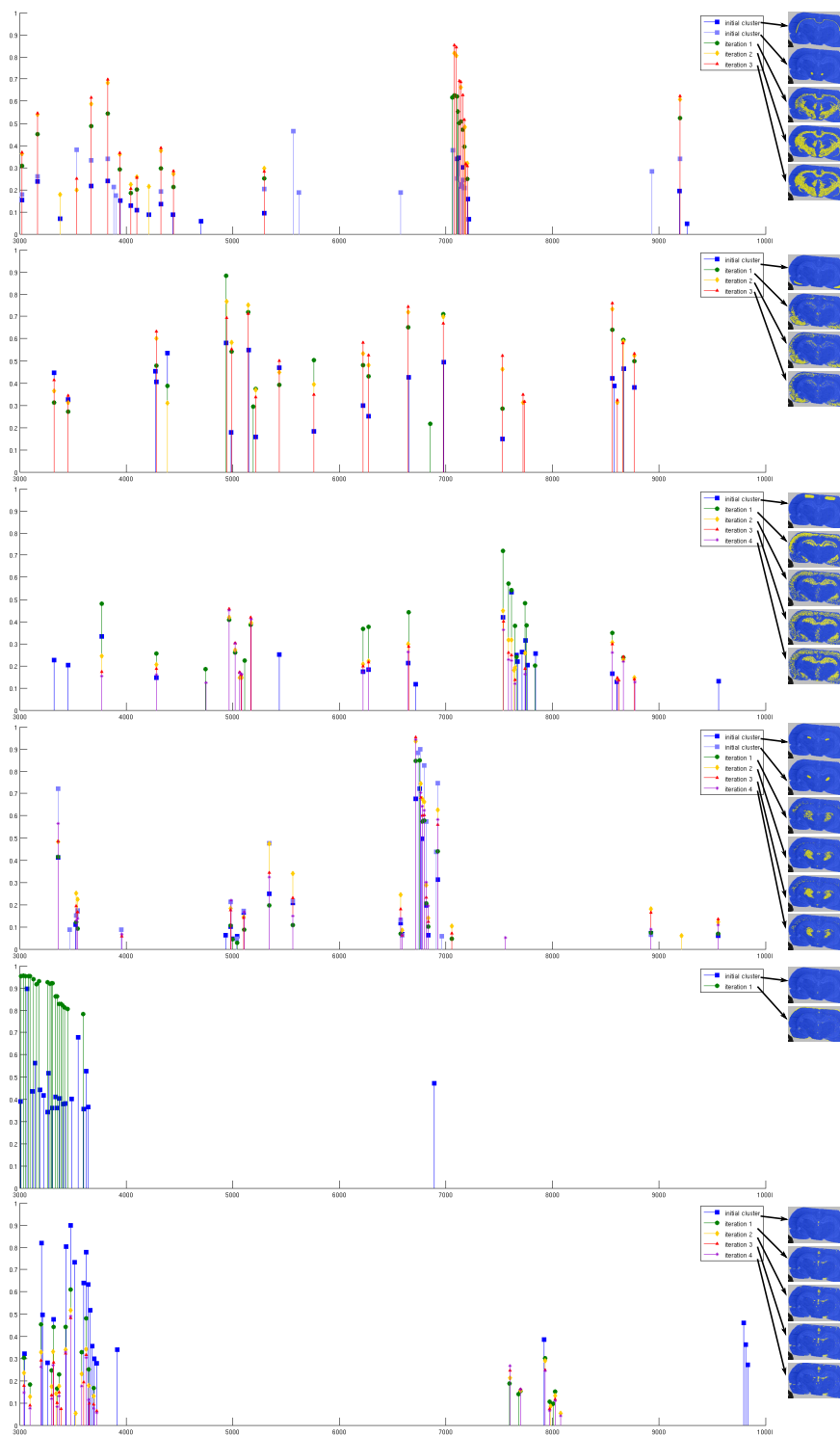


Figure B.4: Top 20 score peaks at least 10 Daltons apart for segments in rat at successive iterations.

Table B.2: Molecules expressed in different regions of the rat brain cortex.

retrosplenial cortex		parietal association cortex		primary somatosensory cortex		auditory cortex	
m/z	weight	m/z	weight	m/z	weight	m/z	weight
		3765.5	0.32				
		3767.5	0.33				
						6720.5	0.34
						6722.5	0.33
						6725.0	0.34
						6728.5	0.34
						6731.0	0.34
7534.5	0.33	7532.5	0.42	7529.0	0.33		
7537.5	0.34	7534.5	0.42	7532.5	0.33		
7541.0	0.34						
7543.0	0.33						
7546.5	0.31						
				7568.5	0.33		
				7571.5	0.36		
				7574.0	0.35		
				7580.0	0.31		
		7611.5	0.51				
		7614.0	0.53				
		7617.0	0.51				
		7619.5	0.46				
		7747.5	0.31				
7840.0	0.31						
7842.5	0.32						
7845.5	0.32						
9559.5	0.39						
9563.0	0.39						
9981.0	0.32						

Table B.3: Molecules expressed in the amygdala and piriform cortex of the rat brain.

piriform cortex		amygdala		image
m/z	weight	m/z	weight	
3321.0	0.40			
3322.5	0.47	3322.5	0.52	
3325.0	0.45	3324.5	0.45	
3326.5	0.35	3327.0	0.33	
3451.0	0.33			
3453.5	0.53			
3455.0	0.52	3455.0	0.33	
3457.0	0.50	3457.5	0.30	
3459.0	0.35			
4276.5	0.39	4270.5	0.47	
		4273.0	0.45	
		4275.0	0.37	
		4281.0	0.44	
		4283.0	0.40	
4285.5	0.36	4283.5	0.41	
		4383.5	0.42	
		4385.5	0.53	
		4388.0	0.50	
		4390.5	0.41	
4931.5	0.32	4931.0	0.44	
4933.5	0.37	4932.0	0.54	
4936.0	0.38	4936.0	0.59	
4938.0	0.34	4938.0	0.57	
		4940.5	0.52	
5146.0	0.38	5141.5	0.32	
		5143.5	0.42	
		5146.5	0.52	
		5148.0	0.55	
5151.0	0.33	5151.0	0.49	
5436.5	0.31	5434.5	0.36	
		5436.5	0.47	
		5439.0	0.46	
		5442.0	0.32	
6218.0	0.47			
6221.0	0.50			
6223.5	0.54			
6225.5	0.50			
6228.0	0.48			
6270.0	0.39			
6272.5	0.48			
6275.0	0.47			
6277.5	0.42			
6279.5	0.35			

piriform cortex		amygdala		image
m/z	weight	m/z	weight	
6643.0	0.60	6643.0	0.35	
6646.0	0.65	6646.0	0.41	
6647.5	0.62	6648.5	0.43	
6651.0	0.61	6650.5	0.39	
6653.5	0.61	6654.0	0.33	
6854.0	0.33			
6972.5	0.39	6972.5	0.34	
6975.5	0.48	6976.0	0.48	
6978.0	0.46	6978.5	0.49	
6981.0	0.33	6981.0	0.42	
8558.0	0.34			
8561.0	0.37	8560.5	0.40	
8564.0	0.37	8563.0	0.46	
8566.5	0.34	8567.0	0.47	
		8569.5	0.41	
8581.5	0.33	8582.0	0.39	
8663.0	0.30	8660.0	0.40	
		8663.0	0.46	
		8666.5	0.37	
		8669.0	0.35	
		8672.5	0.32	
		8672.5	0.41	
8767.0	0.31	8766.5	0.38	
8770.0	0.33	8769.5	0.38	
		8772.5	0.37	
8776.0	0.32	8776.0	0.38	
		8778.5	0.37	
		10942.0	0.33	
		10944.5	0.40	
		10948.0	0.38	
		10951.5	0.30	
		21860.5	0.46	
		21870.5	0.48	
		21875.5	0.47	
		21880.0	0.46	
		21885.0	0.44	
		21918.0	0.31	

Table B.4: Molecules expressed in different regions of the rat brain hippocampus.

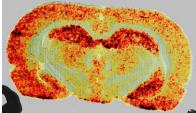
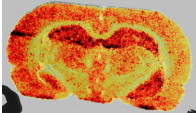
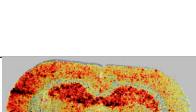
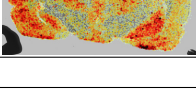
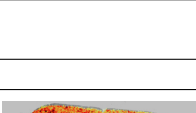
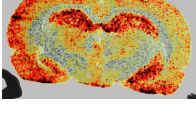
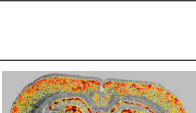
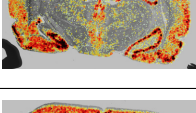
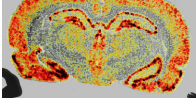
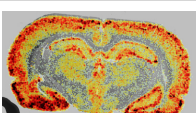
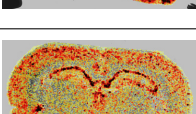
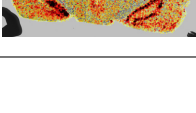
dentate gyrus		CA1-CA3 axons		CA1-CA3 cell bodies		CA3 cell bodies		images
m/z	weight	m/z	weight	m/z	weight	m/z	weight	
		4958.5	0.44					
		4961.5	0.47					
		4963.0	0.46					
		4966.0	0.44					
5000.0	0.32	5000.5	0.43	5000.5	0.23			
5002.5	0.32			5003.0	0.22			
		5021.0	0.38					
		5023.5	0.38					
5039.0	0.33	5037.5	0.41	5037.0	0.21			
5042.0	0.31	5039.5	0.41	5039.0	0.21			
		5057.5	0.33					
5080.5	0.32							
		5102.0	0.33					
		5104.5	0.34					
		5113.5	0.33					
		5165.0	0.43					
		5167.5	0.46					
		5169.5	0.47					
		5171.5	0.46					
		5174.0	0.44					
		5762.5	0.37					
		5765.0	0.34					
						6221.0	0.61	
						6223.5	0.62	
						6226.0	0.64	
						6228.5	0.60	
						6233.5	0.55	
						6272.5	0.58	
						6275.0	0.63	
						6277.0	0.59	
						6280.0	0.55	
						6282.5	0.53	
						6643.5	0.34	
						6645.5	0.36	
						6648.0	0.39	
						6650.5	0.37	
						6654.0	0.32	
						8444.0	0.32	
						8446.5	0.60	
						8449.5	0.54	
						8452.5	0.53	
						8456.0	0.36	

Table B.5: Molecules expressed in different regions of the rat brain thalamus and epithalamus.

paraventricular thal. nucleus		ventral posteromedial thal. nucleus		lateral habenular nucleus		medial habenular nucleus (cont'd)	
m/z	weight	m/z	weight	m/z	weight	m/z	weight
3953.5	0.50	3355.0	0.33	3530.5	0.34	7003.0	0.43
3956.0	0.61	3356.5	0.62	3544.0	0.38	7014.0	0.64
4273.0	0.32	3359.0	0.72	5562.5	0.43	7016.5	0.62
4383.5	0.61	3360.5	0.54	5564.5	0.47	7025.0	0.40
4385.5	0.50	5341.5	0.41	5567.0	0.59	8450.0	0.35
4388.0	0.56	5343.5	0.48	5569.5	0.33	9563.5	0.38
4390.0	0.40	5346.5	0.44	10600.50	0.41	9938.5	0.35
4392.0	0.38	6739.0	0.88	10604.00	0.40	10607.00	0.40
4934.0	0.32	6752.5	0.88	10607.50	0.54	10614.00	0.38
4936.5	0.34	6754.5	0.89	10611.00	0.36	11256.50	0.37
4938.5	0.32	6757.5	0.90	10614.50	0.41	11277.50	0.42
5765.0	0.39	6760.5	0.89	12121.50	0.31	11301.00	0.49
6221.0	0.33	6795.0	0.83			11322.00	0.54
6223.0	0.35	6802.5	0.59			11336.50	0.50
6225.5	0.30	6810.5	0.48			11340.00	0.52
6228.5	0.30	6813.5	0.57			11342.50	0.51
6275.0	0.34	6816.0	0.55			11356.50	0.43
6277.0	0.36	6911.0	0.44			11378.00	0.55
6279.5	0.32	6921.5	0.65			11388.50	0.51
9537.5	0.47	6924.5	0.75			11409.00	0.34
9560.0	0.46	6926.5	0.70			11481.50	0.30
9563.0	0.62	6929.5	0.53			13758.00	0.40
9569.5	0.47					13769.50	0.49
9573.0	0.46					13773.50	0.51
						13777.50	0.42
						13781.50	0.57
						13812.00	0.38
						13823.50	0.41
						13831.00	0.41
						13862.00	0.47
						13869.50	0.41
						13873.00	0.42
						13895.50	0.42
						13900.00	0.45
						13911.00	0.30
						13934.00	0.36
						13938.00	0.37
						13957.00	0.35
						13988.00	0.38
						14000.00	0.36
						14007.00	0.39
						14011.00	0.36
						14022.50	0.35
						15281.50	0.32

lateral post. thal. nucleus		ventral posteromedial thal. nucleus		medial habenular nucleus	
m/z	weight	m/z	weight	m/z	weight
3356.5	0.32	6739.0	0.59	5346.0	0.39
3359.0	0.41	6744.5	0.57	5567.5	0.33
6717.5	0.68	6754.5	0.61	6215.5	0.30
6752.5	0.68	6757.5	0.57	6223.5	0.32
6755.0	0.72	6760.0	0.59	6270.0	0.30
6757.0	0.71	6792.5	0.41	6272.0	0.33
6760.0	0.69	6795.0	0.43	6277.0	0.36
6773.0	0.50	6926.5	0.31	6545.5	0.33
6924.5	0.31			6595.5	0.35
				6741.5	0.71
				6747.0	0.74
				6754.5	0.73
				6760.0	0.78
				6762.5	0.75
				6794.5	0.53
				6798.0	0.54
				6889.5	0.46
				6891.5	0.42
				6924.0	0.59
				6926.5	0.44
				6929.5	0.54
				6951.5	0.30
				6986.0	0.40

Table B.6: Molecules expressed in different regions of the rat brain hypothalamus.

posterior hypothal. area		lateral hypothal. area		lateral hypothal. area (cont'd)	
m/z	weight	m/z	weight	m/z	weight
3066.0	0.40	3528.0	0.31	13965.5	0.31
3526.0	0.38	3530.0	0.36	14108.0	0.39
3528.0	0.45	3532.0	0.38	14111.5	0.40
3530.5	0.38	3534.5	0.36	14115.0	0.39
3542.0	0.39	3536.0	0.32	14119.0	0.39
3543.5	0.39	3670.5	0.33	14123.5	0.39
5564.5	0.43	3826.0	0.32	14165.5	0.32
5567.0	0.44	3828.0	0.34	14177.0	0.32
9557.0	0.30	5562.0	0.33	14181.5	0.32
10607.0	0.30	5565.0	0.45	14189.0	0.32
		5567.0	0.47	14192.5	0.32
		5569.5	0.37	14215.5	0.32
		7057.0	0.33	14231.5	0.30
		7060.0	0.37	14277.5	0.33
		7063.0	0.38	14282.0	0.34
		7065.5	0.36	14286.0	0.34
		7068.0	0.34	14294.0	0.34
		9195.0	0.30	14313.0	0.34
		9197.5	0.34	14328.5	0.34
		9207.0	0.30	14332.0	0.36
		10597.5	0.44	14340.5	0.34
		10600.5	0.51	18347.0	0.31
		10604.5	0.54	18382.5	0.32
		10607.0	0.59	18386.5	0.31
		10611.0	0.53	18396.0	0.32

Table B.7: Molecules expressed in other various regions of the rat brain.

zona incerta		internal capsule		internal capsule (cont'd)		internal capsule (cont'd)	
m/z	weight	m/z	weight	m/z	weight	m/z	weight
3668.5	0.32	3165.5	0.42	11960.5	0.37	14403.0	0.38
3670.5	0.32	3167.5	0.41	11964.5	0.35	14406.5	0.37
5564.5	0.35	3169.0	0.34	11968.0	0.30	14414.5	0.38
7046.0	0.34	3669.0	0.33	13850.0	0.32	14418.0	0.40
7051.5	0.34	3671.0	0.44	13907.0	0.41	14441.5	0.34
7055.0	0.33	3672.5	0.41	13961.5	0.46	14453.0	0.37
7057.0	0.36	3823.5	0.43	13980.5	0.50	14461.0	0.45
7060.0	0.34	3826.0	0.55	13991.5	0.48	14563.0	0.31
7137.0	0.30	3827.5	0.49	14007.5	0.49	18303.0	0.64
9195.0	0.33	3829.5	0.52	14050.0	0.53	18308.0	0.65
9198.0	0.31	3834.0	0.32	14061.0	0.53	18316.0	0.64
9201.5	0.31	4042.0	0.35	14072.5	0.55	18321.0	0.63
10597.0	0.30	4044.5	0.32	14088.5	0.53	18352.0	0.67
10600.5	0.38	4322.0	0.33	14095.5	0.52	18360.0	0.72
10604.0	0.39	4324.0	0.36	14100.0	0.52	18374.0	0.66
10607.0	0.40	4326.0	0.37	14177.0	0.61	18378.0	0.63
10611.0	0.34	4699.5	0.36	14184.5	0.58	18382.0	0.67
14088.5	0.35	5291.5	0.34	14189.0	0.57	18404.5	0.58
14099.5	0.35	7044.0	0.52	14200.0	0.59	18409.0	0.58
14104.0	0.35	7098.5	0.55	14208.0	0.57	18427.0	0.66
14107.5	0.35	7101.5	0.61	14227.5	0.57	18431.0	0.63
14119.5	0.35	7103.5	0.56	14239.0	0.61	18435.5	0.65
14142.0	0.31	7123.5	0.55	14247.0	0.56	18457.0	0.57
14154.0	0.31	7142.5	0.61	14258.5	0.62	18461.5	0.52
14161.5	0.32	7153.5	0.51	14266.5	0.57	18483.5	0.54
14165.5	0.31	7159.0	0.52	14285.5	0.59	18493.0	0.47
14177.5	0.30	7161.5	0.52	14290.0	0.52	18502.0	0.50
18342.5	0.36	7178.0	0.50	14305.5	0.58	18510.0	0.51
18360.5	0.35	7197.0	0.36	14312.5	0.52	18528.5	0.41
18373.5	0.36	7202.5	0.30	14321.0	0.52	18550.5	0.41
18382.5	0.33	9191.5	0.43	14348.0	0.52	18554.5	0.54
18387.0	0.34	9195.0	0.60	14351.5	0.53	18563.5	0.38
18404.0	0.32	9198.0	0.49	14356.0	0.54	18576.5	0.45
		9201.0	0.46	14363.5	0.66	18581.5	0.38
		9204.5	0.50	14391.0	0.55	18607.5	0.36

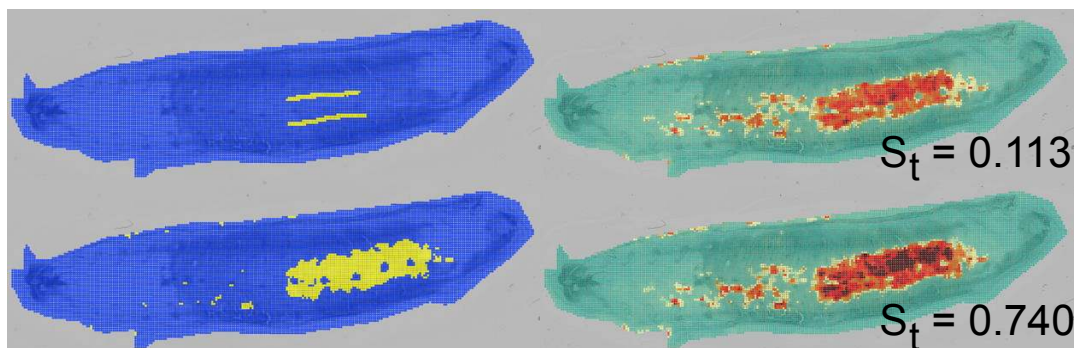


Figure B.5: Query consistency scores. In the top panel, the original query (left) recruits many other spots (right) sharing a similar molecular signature, thus resulting in a low consistency score (0.113). In the bottom panel, the query and query-result are very similar indicating that all spots within the query have similar molecular signatures and spots outside the query have different molecular signatures. This results in a much higher consistency score (0.74).

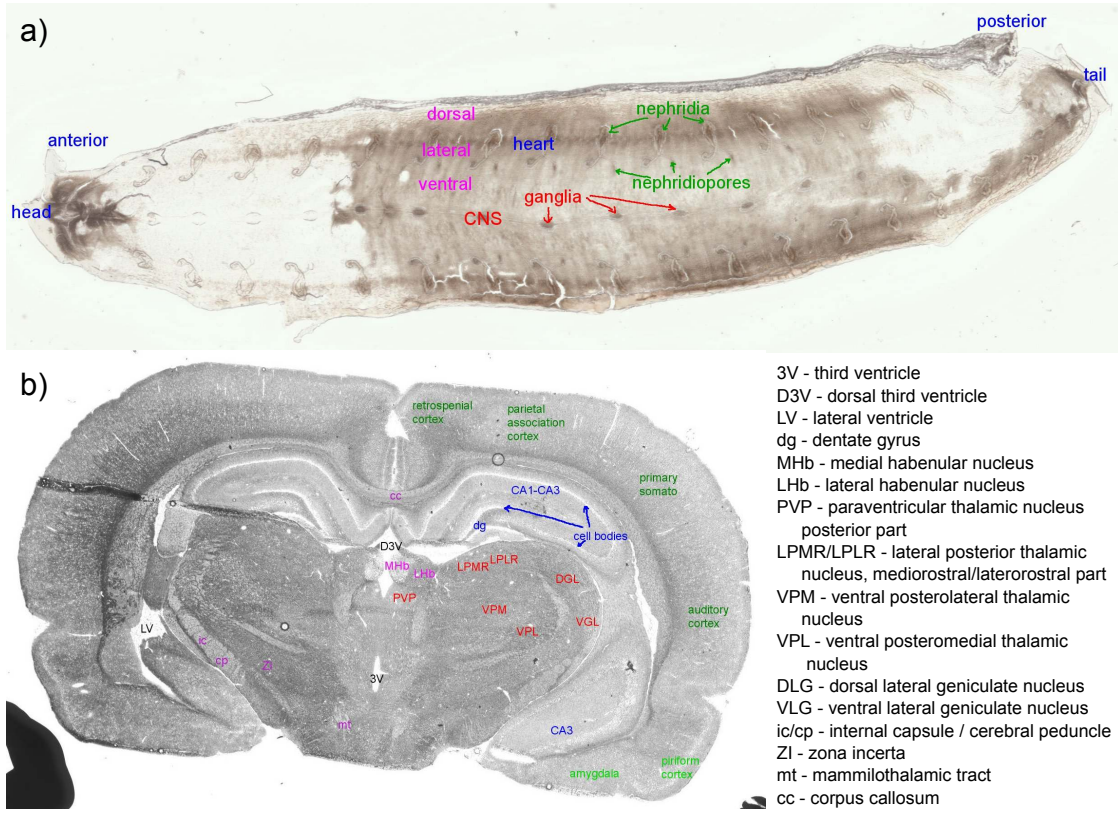


Figure B.6: Basic anatomy for the a) the leech embryo and b) the rat brain slice.

Bibliography

- [1] M. Brulet, A. Seyer, A. Edelman, A. Brunelle, J. Fritsch, M. Ollero, and O. Laprevote, “Lipid mapping of colonic mucosa by cluster TOF-SIMS imaging and multivariate analysis in cftr knockout mice,” *J. Lipid Res.*, vol. 51, pp. 3034–3045, Oct 2010.
- [2] P. Chaurand, “Imaging mass spectrometry of thin tissue sections: A decade of collective efforts,” *J Proteomics*, Apr 2012.
- [3] E. R. Amstalden van Hove, D. F. Smith, and R. M. Heeren, “A concise review of mass spectrometry imaging,” *J Chromatogr A*, vol. 1217, pp. 3946–3954, Jun 2010.
- [4] L. S. Eberlin, C. R. Ferreira, A. L. Dill, D. R. Ifa, and R. G. Cooks, “Desorption electrospray ionization mass spectrometry for lipid characterization and biological tissue imaging,” *Biochim. Biophys. Acta*, vol. 1811, pp. 946–960, Nov 2011.
- [5] J. Laskin, B. S. Heath, P. J. Roach, L. Cazares, and O. J. Semmes, “Tissue imaging using nanospray desorption electrospray ionization mass spectrometry,” *Anal. Chem.*, vol. 84, pp. 141–148, Jan 2012.
- [6] L. A. McDonnell, G. L. Corthals, S. M. Willems, A. van Remoortere, R. J. van Zeijl, and A. M. Deelder, “Peptide and protein imaging mass spectrometry in cancer research,” *J Proteomics*, vol. 73, pp. 1921–1944, Sep 2010.
- [7] M. El Ayed, D. Bonnel, R. Longuespee, C. Castelier, J. Franck, D. Vergara, A. Desmons, A. Tasiemski, A. Kenani, D. Vinatier, R. Day, I. Fournier, and M. Salzet, “MALDI imaging mass spectrometry in ovarian cancer for tracking, identifying, and validating biomarkers,” *Med. Sci. Monit.*, vol. 16, pp. R233–245, Aug 2010.
- [8] H. Bateson, S. Saleem, P. M. Loadman, and C. W. Sutton, “Use of matrix-assisted laser desorption/ionisation mass spectrometry in cancer research,” *J Pharmacol Toxicol Methods*, vol. 64, no. 3, pp. 197–206, 2011.

- [9] S. S. Rubakhin, N. G. Hatcher, E. B. Monroe, M. L. Heien, and J. V. Sweedler, "Mass spectrometric imaging of the nervous system," *Curr. Pharm. Des.*, vol. 13, no. 32, pp. 3325–3334, 2007.
- [10] B. Langstrom, P. E. Andren, O. Lindhe, M. Svedberg, and H. Hall, "In vitro imaging techniques in neurodegenerative diseases," *Mol Imaging Biol*, vol. 9, no. 4, pp. 161–175, 2007.
- [11] M. Wisztorski, D. Croix, E. Macagno, I. Fournier, and M. Salzet, "Molecular MALDI imaging: an emerging technology for neuroscience studies," *Dev Neurobiol*, vol. 68, pp. 845–858, May 2008.
- [12] S. Castellino, M. R. Groseclose, and D. Wagner, "MALDI imaging mass spectrometry: bridging biology and chemistry in drug development," *Bioanalysis*, vol. 3, pp. 2427–2441, Nov 2011.
- [13] F. Benabdellah, A. Seyer, L. Quinton, D. Touboul, A. Brunelle, and O. Laprevote, "Mass spectrometry imaging of rat brain sections: nanomolar sensitivity with MALDI versus nanometer resolution by TOF-SIMS," *Anal Bioanal Chem*, vol. 396, pp. 151–162, Jan 2010.
- [14] A. M. Delvolve, B. Colsch, and A. S. Woods, "Highlighting anatomical substructures in rat brain tissue using lipid imaging," *Anal Methods*, vol. 3, pp. 1729–1736, Aug 2011.
- [15] P. Chaurand, M. A. Rahman, T. Hunt, J. A. Mobley, G. Gu, J. C. Latham, R. M. Caprioli, and S. Kasper, "Monitoring mouse prostate development by profiling and imaging mass spectrometry," *Mol. Cell Proteomics*, vol. 7, pp. 411–423, Feb 2008.
- [16] D. Bonnel, R. Longuespee, J. Franck, M. Roudbaraki, P. Gosset, R. Day, M. Salzet, and I. Fournier, "Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: application to prostate cancer," *Anal Bioanal Chem*, vol. 401, pp. 149–165, Jul 2011.
- [17] A. C. Grey, P. Chaurand, R. M. Caprioli, and K. L. Schey, "MALDI imaging mass spectrometry of integral membrane proteins from ocular lens and retinal tissue," *J. Proteome Res.*, vol. 8, pp. 3278–3283, Jul 2009.
- [18] M. Ronci, S. Sharma, T. Chataway, K. P. Burdon, S. Martin, J. E. Craig, and N. H. Voelcker, "MALDI-MS-imaging of whole human lens capsule," *J. Proteome Res.*, vol. 10, pp. 3522–3529, Aug 2011.
- [19] M. L. Reyzer, P. Chaurand, P. M. Angel, and R. M. Caprioli, "Direct molecular analysis of whole-body animal tissue sections by MALDI imaging mass spectrometry," *Methods Mol. Biol.*, vol. 656, pp. 285–301, 2010.

- [20] P. Chaurand, D. S. Cornett, P. M. Angel, and R. M. Caprioli, "From whole-body sections down to cellular level, multiscale imaging of phospholipids by MALDI mass spectrometry," *Mol. Cell Proteomics*, vol. 10, p. O110.004259, Feb 2011.
- [21] H. Ye, T. Greer, and L. Li, "From pixel to voxel: a deeper view of biological tissue by 3D mass spectral imaging," *Bioanalysis*, vol. 3, pp. 313–332, Feb 2011.
- [22] S. L. Koeniger, N. Talaty, Y. Luo, D. Ready, M. Voorbach, T. Seifert, S. Cepa, J. A. Fagerland, J. Bouska, W. Buck, R. W. Johnson, and S. Spanton, "A quantitation method for mass spectrometry imaging," *Rapid Commun. Mass Spectrom.*, vol. 25, pp. 503–510, Feb 2011.
- [23] E. J. Clemis, D. S. Smith, A. G. Camenzind, R. M. Danell, C. E. Parker, and C. H. Borchers, "Quantitation of spatially-localized proteins in tissue samples using MALDI-MRM imaging," *Anal. Chem.*, vol. 84, pp. 3514–3522, Apr 2012.
- [24] L. Ferguson, R. Bradshaw, R. Wolstenholme, M. Clench, and S. Francese, "Two-step matrix application for the enhancement and imaging of latent fingerprints," *Anal. Chem.*, vol. 83, pp. 5585–5591, Jul 2011.
- [25] L. A. McDonnell, A. van Remoortere, R. J. van Zeijl, H. Dalebout, M. R. Bladergroen, and A. M. Deelder, "Automated imaging MS: Toward high throughput imaging mass spectrometry," *J Proteomics*, vol. 73, pp. 1279–1282, Apr 2010.
- [26] R. M. Heeren, D. F. Smith, J. Stauber, B. Kukrer-Kaletas, and L. MacAleese, "Imaging mass spectrometry: hype or hope?," *J. Am. Soc. Mass Spectrom.*, vol. 20, pp. 1006–1014, Jun 2009.
- [27] P. Chaurand, "Imaging Mass Spectrometry: Current Performance and Upcoming Challenges," *Spectroscopy Online*, Jul 2011.
- [28] J. H. Jungmann and R. M. Heeren, "Emerging technologies in mass spectrometry imaging," *J Proteomics*, Mar 2012.
- [29] R. Lemaire, J. C. Tabet, P. Ducoroy, J. B. Hendra, M. Salzet, and I. Fournier, "Solid ionic matrixes for direct tissue analysis and MALDI imaging," *Anal. Chem.*, vol. 78, pp. 809–819, Feb 2006.
- [30] J. Franck, K. Arafah, A. Barnes, M. Wisztorski, M. Salzet, and I. Fournier, "Improving tissue preparation for matrix-assisted laser desorption ionization mass spectrometry imaging. Part 1: using microspotting," *Anal. Chem.*, vol. 81, pp. 8193–8202, Oct 2009.
- [31] B. Spengler, M. Hubert, and R. Kaufmann, "MALDI ion imaging and biological ion imaging with new scanning UV-laser microprobe," in *Proceedings of the 42nd ASMS Conference on Mass Spectrometry and Allied Topics*, ASMS, 1994.

- [32] P. Chaurand, M. Stoeckli, and R. M. Caprioli, "Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry," *Anal. Chem.*, vol. 71, pp. 5263–5270, Dec 1999.
- [33] I. Fournier, R. Day, and M. Salzet, "Direct analysis of neuropeptides by in situ MALDI-TOF mass spectrometry in the rat brain," *Neuro Endocrinol. Lett.*, vol. 24, pp. 9–14, 2003.
- [34] K. Dreisewerd, R. Kingston, W. P. M. Geraerts, and K. W. Li, "Direct mass spectrometric peptide profiling and sequencing of nervous tissues to identify peptides involved in male copulatory behavior in *lymnaea stagnalis*," *International Journal of Mass Spectrometry and Ion Processes*, vol. 169-170, pp. 291 – 299, 1997. Matrix-Assisted Laser Desorption Ionization Mass Spectrometry.
- [35] C. R. Jiménez, K. W. Li, K. Dreisewerd, S. Spijker, R. Kingston, R. H. Bateman, A. L. Burlingame, A. B. Smit, J. van Minnen, and W. P. Geraerts, "Direct mass spectrometric peptide profiling and sequencing of single neurons reveals differential peptide patterns in a small neuronal network," *Biochemistry*, vol. 37, pp. 2070–2076, Feb 1998.
- [36] K. W. Li, R. M. Hoek, F. Smith, C. R. Jiménez, R. C. van der Schors, P. A. van Veelen, S. Chen, J. van der Greef, D. C. Parish, and P. R. Benjamin, "Direct peptide profiling by mass spectrometry of single identified neurons reveals complex neuropeptide-processing pattern," *J. Biol. Chem.*, vol. 269, pp. 30288–30292, Dec 1994.
- [37] R. M. Caprioli, T. B. Farmer, and J. Gile, "Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS," *Anal. Chem.*, vol. 69, pp. 4751–4760, Dec 1997.
- [38] M. Stoeckli, T. B. Farmer, and R. M. Caprioli, "Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument," *J. Am. Soc. Mass Spectrom.*, vol. 10, pp. 67–71, Jan 1999.
- [39] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, "Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues," *Nat. Med.*, vol. 7, pp. 493–496, Apr 2001.
- [40] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis," *Anal. Chem.*, vol. 77, pp. 6118–6124, Oct 2005.
- [41] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, and E. Waelkens, "Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis," *Pac Symp Biocomput*, pp. 458–469, 2007.

- [42] S. E. Blackshaw, L. P. Henderson, J. Malek, D. M. Porter, R. H. Gross, J. D. Angstadt, S. M. Levasseur, and R. A. Maue, "Single-cell analysis reveals cell-specific patterns of expression of a family of putative voltage-gated sodium channel genes in the leech," *J. Neurobiol.*, vol. 55, pp. 355–371, Jun 2003.
- [43] B. D. Burrell, C. L. Sahley, and K. J. Muller, "Progressive recovery of learning during regeneration of a single synapse in the medicinal leech," *J. Comp. Neurol.*, vol. 457, pp. 67–74, Feb 2003.
- [44] B. A. Skierczynski, R. J. Wilson, W. B. Kristan, and R. Skalak, "A model of the hydrostatic skeleton of the leech," *J. Theor. Biol.*, vol. 181, pp. 329–342, Aug 1996.
- [45] L. H. Cazares, D. Troyer, S. Mendrinós, R. A. Lance, J. O. Nyalwidhe, H. A. Beydoun, M. A. Clements, R. R. Drake, and O. J. Semmes, "Imaging mass spectrometry of a specific fragment of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase kinase 2 discriminates cancer from uninvolved prostate tissue," *Clin. Cancer Res.*, vol. 15, pp. 5541–5551, Sep 2009.
- [46] S. Rauser, C. Marquardt, B. Balluff, S. O. Deininger, C. Albers, E. Belau, R. Hartmer, D. Suckau, K. Specht, M. P. Ebert, M. Schmitt, M. Aubele, H. Hofler, and A. Walch, "Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry," *J. Proteome Res.*, vol. 9, pp. 1854–1863, Apr 2010.
- [47] K. Schwamborn and R. M. Caprioli, "Molecular imaging by mass spectrometry—looking beyond classical histology," *Nat. Rev. Cancer*, vol. 10, pp. 639–646, Sep 2010.
- [48] M. R. Groseclose, M. Andersson, W. M. Hardesty, and R. M. Caprioli, "Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry," *J Mass Spectrom*, vol. 42, pp. 254–262, Feb 2007.
- [49] E. H. Seeley and R. M. Caprioli, "Molecular imaging of proteins in tissues by mass spectrometry," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 18126–18131, Nov 2008.
- [50] S. Shimma, Y. Sugiura, T. Hayasaka, N. Zaima, M. Matsumoto, and M. Setou, "Mass imaging and identification of biomolecules with MALDI-QIT-TOF-based system," *Anal. Chem.*, vol. 80, pp. 878–885, Feb 2008.
- [51] R. J. Goodwin, S. R. Pennington, and A. R. Pitt, "Protein and peptides in pictures: imaging with MALDI mass spectrometry," *Proteomics*, vol. 8, pp. 3785–3800, Sep 2008.

- [52] J. Stauber, L. Macaleese, J. Franck, E. Claude, M. Snel, B. Kkrer Kaletas, I. M. Wiel, M. Wisztorski, I. Fournier, and R. M. Heeren, "On-Tissue Protein Identification and Imaging by MALDI-Ion Mobility Mass Spectrometry," *J. Am. Soc. Mass Spectrom.*, Sep 2009.
- [53] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold, "Computational prediction of proteotypic peptides for quantitative proteomics," *Nat. Biotechnol.*, vol. 25, pp. 125–131, Jan 2007.
- [54] R. Chen, X. Jiang, M. C. Prieto Conaway, I. Mohtashemi, L. Hui, R. Viner, and L. Li, "Mass Spectral Analysis of Neuropeptide Expression and Distribution in the Nervous System of the Lobster *Homarus americanus*," *J Proteome Res*, Jan 2010.
- [55] J. E. Lee, N. Atkins, N. G. Hatcher, L. Zamdborg, M. U. Gillette, J. V. Sweedler, and N. L. Kelleher, "Endogenous peptide discovery of the rat circadian clock: a focused study of the suprachiasmatic nucleus by ultrahigh performance tandem mass spectrometry," *Mol. Cell Proteomics*, vol. 9, pp. 285–297, Feb 2010.
- [56] E. R. Macagno, T. Gaasterland, L. Edsall, V. Bafna, M. B. Soares, T. Scheetz, T. Casavant, C. Da Silva, P. Wincker, A. Tasiemski, and M. Salzet, "Construction of a medicinal leech transcriptome database and its application to the identification of leech homologs of neural and innate immune genes," *BMC Genomics*, vol. 11, p. 407, 2010.
- [57] A. Frank, S. Tanner, V. Bafna, and P. Pevzner, "Peptide sequence tags for fast database search in mass-spectrometry," *J. Proteome Res.*, vol. 4, pp. 1287–1295, 2005.
- [58] D. Nardelli-Haeffliger and M. Shankland, "Lox10, a member of the NK-2 homeobox gene class, is expressed in a segmental pattern in the endoderm and in the cephalic nervous system of the leech *Helobdella*," *Development*, vol. 118, pp. 877–892, Jul 1993.
- [59] J. L. Norris, D. S. Cornett, J. A. Mobley, M. Andersson, E. H. Seeley, P. Chaurand, and R. M. Caprioli, "Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis," *Int J Mass Spectrom*, vol. 260, pp. 212–221, Feb 2007.
- [60] E. R. Macagno, "Number and distribution of neurons in leech segmental ganglia," *J. Comp. Neurol.*, vol. 190, pp. 283–302, Mar 1980.
- [61] C. Lefebvre and M. Salzet, "Annelid neuroimmune system," *Curr. Pharm. Des.*, vol. 9, pp. 149–158, 2003.
- [62] R. Sawyer, *Leech biology and behaviour*. Oxford University Press, 1986.

- [63] I. Zerbst-Boroffka and A. Wenning, "Mechanism of regulatory salt and water excretion in the leech, *Hirudo medicinalis*," *L. Zool. Beitr. N. F.*, vol. 30, p. 359377, 1986.
- [64] I. Zerbst-Boroffka, B. Bazin, and A. Wenning, "Chloride secretion drives urine formation in leech nephridia," *J. Exp. Biol.*, vol. 200, pp. 2217–2227, Aug 1997.
- [65] D. Schikorski, V. Cuvillier-Hot, M. Leippe, C. Boidin-Wichlacz, C. Slomianny, E. Macagno, M. Salzert, and A. Tasiemski, "Microbial challenge promotes the regenerative process of the injured central nervous system of the medicinal leech by inducing the synthesis of antimicrobial peptides in neurons and microglia," *J. Immunol.*, vol. 181, pp. 1083–1095, Jul 2008.
- [66] Y. Xu, B. Bolton, B. Zipser, J. Jellies, K. M. Johansen, and J. Johansen, "Gliarin and macrolin, two novel intermediate filament proteins specifically expressed in sets and subsets of glial cells in leech central nervous system," *J. Neurobiol.*, vol. 40, pp. 244–253, Aug 1999.
- [67] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, Nov 1994.
- [68] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, Sep 1997.
- [69] E. R. Amstalden van Hove, D. F. Smith, and R. M. Heeren, "A concise review of mass spectrometry imaging," *J Chromatogr A*, Feb 2010.
- [70] B. N. Giepmans, S. R. Adams, M. H. Ellisman, and R. Y. Tsien, "The fluorescent toolbox for assessing protein location and function," *Science*, vol. 312, pp. 217–224, Apr 2006.
- [71] R. Lemaire, A. Desmons, J. C. Tabet, R. Day, M. Salzert, and I. Fournier, "Direct analysis and MALDI imaging of formalin-fixed, paraffin-embedded tissue sections," *J. Proteome Res.*, vol. 6, pp. 1295–1305, Apr 2007.
- [72] K. M. Johansen and J. Johansen, "Filarin, a novel invertebrate intermediate filament protein present in axons and perikarya of developing and mature leech neurons," *J. Neurobiol.*, vol. 27, pp. 227–239, Jun 1995.
- [73] V. M. Weake, K. K. Lee, S. Guelman, C. H. Lin, C. Seidel, S. M. Abmayr, and J. L. Workman, "SAGA-mediated H2B deubiquitination controls the development of neuronal connectivity in the *Drosophila* visual system," *EMBO J.*, vol. 27, pp. 394–405, Jan 2008.

- [74] V. M. Weake and J. L. Workman, “Histone ubiquitination: triggering gene activity,” *Mol. Cell*, vol. 29, pp. 653–663, Mar 2008.
- [75] J. Bruand, S. Sistla, C. Mériaux, P. C. Dorrestein, T. Gaasterland, M. Ghassemian, M. Wisztorski, I. Fournier, M. Salzet, E. Macagno, and V. Bafna, “Automated Querying and Identification of Novel Peptides using MALDI Mass Spectrometric Imaging,” *J Proteome Res*, vol. 10, pp. 1915–1928, Apr 2011.
- [76] S. O. Deininger, M. P. Ebert, A. Futterer, M. Gerhard, and C. Rocken, “MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers,” *J. Proteome Res.*, vol. 7, pp. 5230–5236, Dec 2008.
- [77] T. Alexandrov, M. Becker, S. O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, and P. Maass, “Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering,” *J. Proteome Res.*, vol. 9, pp. 6535–6546, Dec 2010.
- [78] L. A. McDonnell, A. van Remoortere, N. de Velde, R. J. van Zeijl, and A. M. Deelder, “Imaging mass spectrometry data reduction: automated feature identification and extraction,” *J. Am. Soc. Mass Spectrom.*, vol. 21, pp. 1969–1978, Dec 2010.
- [79] R. Van de Plas, B. De Moor, and E. Waelkens, “Discrete Wavelet Transform-based Multivariate Exploration of Tissue via Imaging Mass Spectrometry,” in *Proceedings of the 23rd Annual ACM Symposium on Allied Computing (ACM SAC)*, (Fortaleza, Brazil), Mar 2008.
- [80] J. Fernandez and G. S. Stent, “Embryonic development of the hirudinid leech *Hirudo medicinalis*: structure, development and segmentation of the germinal plate,” *J Embryol Exp Morphol*, vol. 72, pp. 71–96, Dec 1982.
- [81] T. T. Tanimoto, “IBM Internal Report,” Nov 17 1957.
- [82] J. Bruand, T. Alexandrov, S. Sistla, M. Wisztorski, C. Meriaux, M. Becker, M. Salzet, I. Fournier, E. Macagno, and V. Bafna, “AMASS: algorithm for MSI analysis by semi-supervised segmentation,” *J. Proteome Res.*, vol. 10, pp. 4734–4743, Oct 2011.
- [83] E. Jones, T. Oliphant, P. Peterson, *et al.*, “SciPy: Open source scientific tools for Python,” 2001–.
- [84] Y. L. Yang, Y. Xu, P. Straight, and P. C. Dorrestein, “Translating metabolic exchange with imaging mass spectrometry,” *Nat. Chem. Biol.*, vol. 5, pp. 885–887, Dec 2009.

- [85] W. T. Liu, Y. L. Yang, Y. Xu, A. Lamsa, N. M. Haste, J. Y. Yang, J. Ng, D. Gonzalez, C. D. Ellermeier, P. D. Straight, P. A. Pevzner, J. Pogliano, V. Nizet, K. Pogliano, and P. C. Dorrestein, "Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of *Bacillus subtilis*," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 16286–16290, Sep 2010.
- [86] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein, "Mass spectral molecular networking of living microbial colonies," *Proc Natl Acad Sci U S A*, May 2012.
- [87] W. de Jong, E. Vijgenboom, L. Dijkhuizen, H. A. Wosten, and D. Claessen, "SapB and the rodlinins are required for development of *Streptomyces coelicolor* in high osmolarity media," *FEMS Microbiol. Lett.*, vol. 329, pp. 154–159, Apr 2012.
- [88] J. Watrous, V. V. Phelan, H. C., M. W., D. B. M., T. Alexandrov, and P. C. Dorrestein, "Microbial metabolic exchange in 3D," [submitted], 2012.
- [89] P. Patel, L. Song, and G. L. Challis, "Distinct extracytoplasmic siderophore binding proteins recognize ferrioxamines and ferricoelichelin in *Streptomyces coelicolor* A3(2)," *Biochemistry*, vol. 49, pp. 8033–8042, Sep 2010.
- [90] X. Zhang, Y. Li, W. Shao, and H. Lam, "Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis," *Proteomics*, vol. 11, pp. 1075–1085, Mar 2011.
- [91] C. Y. Yen, S. Houel, N. G. Ahn, and W. M. Old, "Spectrum-to-spectrum searching using a proteome-wide spectral library," *Mol. Cell Proteomics*, vol. 10, p. M111.007666, Jul 2011.
- [92] J. Wang, J. Perez-Santiago, J. E. Katz, P. Mallick, and N. Bandeira, "Peptide identification from mixture tandem mass spectra," *Mol. Cell Proteomics*, vol. 9, pp. 1476–1485, Jul 2010.
- [93] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–3567, Dec 1999.
- [94] S. Kim, N. Gupta, and P. A. Pevzner, "Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases," *J. Proteome Res.*, vol. 7, pp. 3354–3363, Aug 2008.
- [95] F. Richter, B. H. Meurers, C. Zhu, V. P. Medvedeva, and M. F. Chesselet, "Neurons express hemoglobin alpha- and beta-chains in rat and human brains," *J. Comp. Neurol.*, vol. 515, pp. 538–547, Aug 2009.

- [96] Y. He, Y. Hua, J. Y. Lee, W. Liu, R. F. Keep, M. M. Wang, and G. Xi, "Brain alpha- and beta-globin expression after intracerebral hemorrhage," *Transl Stroke Res*, vol. 1, pp. 48–56, Mar 2010.
- [97] J. Wang, P. E. Bourne, and N. Bandeira, "Peptide identification by database search of mixture tandem mass spectra," *Mol. Cell Proteomics*, vol. 10, p. M111.010017, Dec 2011.
- [98] J. Quanico, J. Franck, C. Daully, K. Strupat, J. Dupuy, R. Day, M. Salzet, I. Fournier, and M. Wisztorski, "From MALDI-MSI to proteome mapping: a new strategy for proteins identification from tissues," *[submitted]*, 2012.
- [99] A. Rompp, T. Schramm, A. Hester, I. Klinkert, J. P. Both, R. M. Heeren, M. Stockli, and B. Spengler, "imzML: Imaging Mass Spectrometry Markup Language: A common data format for mass spectrometry imaging," *Methods Mol. Biol.*, vol. 696, pp. 205–224, 2011.