Tissue Based Proteomics and Biomarker Discovery

Multivariate Data Mining Strategies for Mass Spectral Imaging

Raf VAN DE PLAS

Promotor:
Prof. dr. ir. B. De Moor

Co-Promotor:
Prof. dr. E. Waelkens

Dissertation presented
in partial fulfillment of the
requirements for the degree
of Doctor in Engineering

May 2010

Cover image:
Network representation of co-localization relationships between various ions extracted from a sagittal section of mouse brain. Some spatial distributions corresponding to ion node clusters are depicted as well, overlaid on a microscopic image of the mouse brain tissue section. Details on the mass spectral imaging measurements that led to this network are discussed in the chapter on 'co-localization networks'. Image colors were post-processed by JASC Paint Shop Pro 9.

Tissue Based Proteomics and Biomarker Discovery

Multivariate Data Mining Strategies for Mass Spectral Imaging

Raf VAN DE PLAS

Jury:

Prof. em. dr. ir. E. Aernoudt, chair
Prof. dr. ir. B. De Moor, promotor
Prof. dr. E. Waelkens, co-promotor
Prof. dr. ir. J.A.K. Suykens
Prof. dr. ir. J. Vanderleyden
Prof. dr. ing. R.M.A. Heeren (FOM AMOLF, Amsterdam)
Prof. dr. L.A. McDonnell (LUMC, Leiden)

Dissertation presented
in partial fulfillment of the
requirements for the degree
of Doctor in Engineering

May 2010

© 2010 Katholieke Universiteit Leuven
Groep Wetenschap & Technologie
Arenberg Doctoraatsschool
W. de Croylaan 6
B–3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of this publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

D/2010/0000/000

ISBN 000-00-0000-000-0

Voorwoord

To be completed during the jury's reading phase...

Raf Van de Plas
11 mei 2010

Korte inhoud

Recent is het alsmaar duidelijker geworden dat de moleculaire mechanismen die ten grondslag liggen aan vele van 's werelds meest invloedrijke ziekten, een veel grotere complexiteit vertonen dan oorspronkelijk gedacht. Aan diagnostische zijde kent de zoektocht naar ziekte-rapporterende biomerkers in makkelijk toegankelijke vloeistoffen, zoals bloed, slechts gemengd succes. De biomerkers in dergelijke vloeistoffen zijn vaak in zodanig lage concentraties aanwezig, dat ze enkel met immunotechnieken betrouwbaar kunnen worden gedetecteerd. Hoewel ze diagnostische waarde hebben, rapporteren ze meestal secundaire ziekte-effecten en verschaffen daardoor weinig inzicht in de daadwerkelijke oorzaak van een ziektebeeld. Voor studies die trachten het onderliggende patho-mechanisme te doorgronden, kan de direkte analyse van aangetast weefsel, in plaats van indirekte vloeistoffen, echter uitkomst bieden. Weefselanalyse laat toe om over een groot aantal verschillende moleculen en hun metabole wijzigingen informatie te verwerven en doet dit dicht bij de bron zodat de observaties niet beperkt blijven tot secundaire ziekte-effecten. Experimenten rechtstreeks op weefsel laten tevens toe om potentiële biomerkers te bestuderen in onverdunde vorm, bij hoogste concentratie en verbonden aan specifieke weefselregio's of celtypes.

Massaspectrale beeldvorming (MSI) kent een groeiend belang in het onderzoek naar de ruimtelijke spreiding van biomoleculen in weefsel. Het is een relatief nieuwe technologie die duizenden individuele massaspectrale metingen doorheen het ganse weefsel *in situ* verzamelt. MSI betekent een fusie tussen de ruimtelijke informatie van een beeldvormingsmethode en de moleculaire specificiteit en sensitiviteit van massaspectrometrie. De technologie biedt de mogelijkheid om allerlei biomoleculen, gaande van proteïnen en peptiden tot metabolieten en lipiden, te detecteren in weefsel zonder dat deze moleculen kunstmatig moeten worden gemarkeerd. Het feit dat er geen voorafgaande hypothese rond verdachte moleculen noodzakelijk is en de mogelijkheid wordt geboden om in één enkel experiment honderden moleculen tegelijkertijd te volgen, betekent dat MSI een ideaal instrument is voor exploratief onderzoek. De technologie vertoont hiermee sterk potentieel voor zowel voor exploratieve en confirmatieve biomedische weefselstudies.

De voornaamste toepassing van MSI is het genereren van ionbeelden. Een ionbeeld toont de ruimtelijke distributie van één specifiek ion, massa of molecule. Door hun label-vrije karakter en vaak ongeëvenaarde specificiteit, beschrijven ionbeelden een specifieke niche binnen het veld van de moleculaire beeldvorming. Het volle potentieel van de MSI-technologie overstijgt echter het maken van ionbeelden. De metingen van een MSI experiment omvatten typisch een grote hoeveelheid biochemische informatie, die in parallel verzameld wordt en zowel het spatiale als het moleculaire massadomein beschrijft. Hoewel de data zeer veel ruimte kan innemen, zorgt deze brede biochemische informatiebron ervoor dat MSI uitermate geschikt is voor het bestuderen van intermoleculaire eigenschappen zoals co-lokalisatie en gecorreleerde expressie. De toepassingen voor dit soort informatie gaan van inzichten in metabole processen tot de exploratie van co-lokaliserende biomerkers. Naast de univariate ionbeelden, ontbreekt het momenteel echter aan multivariate, computationele methoden die in staat zijn om, met het oog op een specifieke biomedische studie, uit deze zee van data, relevante informatie te vissen.

Deze thesis tracht deze nood te lenigen met verschillende bijdragen langsheen het ganse massaspectrale beeldvormingspad. Het betreft bewerkingsmethoden (o.a. normalisatie- en compressiemethoden), die specifieke ruistypes trachten te verwijderen, maar ook ongesuperviseerde methoden (zoals niet-negatieve matrixfactorisatie en co-lokalisatienetwerken), die gebruikt worden om de data te ontrafelen in onderliggende patronen. Dit wordt verder aangevuld met *spatial querying*, een bijdrage aan de categorie van de gesuperviseerde methoden die zich richten op specifieke biologische vragen. Verder beschrijft de thesis een pragmatische, wavelet-gebaseerde compressiemethode die toelaat problemen inzake datagrootte te vermijden. Het doctoraat wordt afgerond met een biomedische gevalstudie van diabetes mellitus type 2. Daarin wordt massaspectrale beeldvorming gecombineerd met de in deze thesis ontwikkelde computationele technieken, om zo, enkel op basis van weefsel, inzicht te verschaffen in de metabole processen van diabetes en aan te geven waar deze processen beginnen af te wijken van normaal, gezond metabolisme.

Abstract

In recent years the molecular mechanisms underlying many of the worlds most pressing diseases have been shown to exhibit a complexity that exceeds earlier expectations. On the diagnostic side, the search for disease-reporting biomarkers in easily accessible fluids such as blood has met with variable success. Most often, biomarkers are present in extremely low concentrations (only detectable by immune-based methods) and they often report secondary disease effects, thereby shedding only limited light on the true cause of the pathology. A direct analysis of the affected tissue, however, would prove very helpful, since it would report directly on the complex metabolic alterations of multiple molecules (close to the source, so including, but not restricted to, secondary effects). It would also present potential biomarkers undiluted and at their highest concentration, and potentially even tied to a specific tissue region or cell type.

Mass spectral imaging (MSI) has been gaining considerable momentum as a means of examining the spatial distribution of biomolecules throughout tissue. It is a relatively new technology that acquires thousands of individual mass spectra in situ from across a tissue section. MSI fuses the localization information of an imaging method with the molecular specificity and sensitivity of mass spectrometry. The ability to detect biomolecules such as proteins, peptides, metabolites, and lipids throughout tissue without the need for labeling, and its capability of measuring hundreds of molecules concurrently in a single experiment, have made MSI particularly valuable in exploratory settings where there is no prior hypothesis of relevant target molecules. As a result, MSI is rapidly becoming a potent exploratory and confirmatory instrument for tissue-based biomedical studies.

The most common use of MSI is to produce ion images. An ion image shows the spatial distribution throughout the tissue of one particular ion, mass, or molecule. Ion images fill a distinct niche within the molecular imaging field due to their label-free nature and often unrivaled specificity. However, considering ion images as the sole product of a MSI experiment would deny the full potential of the technology. The measurements of a MSI experiment typically comprise a vast amount of concurrently collected biochemical information, spanning both the spatial as well as the molecular mass domain. Although the data sets can become very large as a result, it makes MSI uniquely suited for studying inter-molecule relationships such as co-localization

and correlated expression. The applications for this type of information range from insights into metabolic pathways to the exploration of co-localizing biomarkers. Besides univariate ion images, there is currently a lack of multivariate computational methods capable of filtering from this sea of data, the nuggets of information relevant to the biomedical study at hand.

This thesis attempts to answer this challenge with contributions along the entire mass spectral imaging workflow. This includes preprocessing methods (such as normalization and compression methods) to cancel certain noise types, unsupervised methods (such as non-negative matrix factorization and co-localization networks) to unravel the underlying patterns present in the data, and supervised methods (such as spatial querying) to answer specific biological questions. The thesis further adds a pragmatic wavelet compression strategy to counter the difficulties introduced by the data size. The doctorate is concluded by a biomedical case study of diabetes mellitus type 2 in which a combination of mass spectral imaging and the computational techniques developed in this work is used to find out, only by studying the tissue, where the metabolic processing of the diabetes models starts to deviate from normal metabolism.

Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry.... if mathematical analysis should ever hold a prominent place in chemistry – an aberration which is happily almost impossible – it would occasion a rapid and widespread degeneration of that science.

Auguste Comte, Cours de philosophie positive, 1830

As cited by Livius: eventus docebit
("The outcome will tell")

The author, 2010

Contents

Contents ix

List of Acronyms xiii

List of Figures xv

List of Tables xix

1 Introduction 1
- 1.1 What is Mass Spectral Imaging? . . . 3
- 1.2 MSI data and ion images . . . 6
- 1.3 Structure of the thesis . . . 9

2 Ionization Efficiency Correction 13
- 2.1 Why normalization? . . . 14
- 2.2 What is the nature of ion intensity noise? . . . 15
- 2.3 The common approach: Total Ion Current (TIC) . . . 17
- 2.4 What is the problem with the common approach? . . . 19
- 2.5 A novel approach: Ionization Efficiency Correction . . . 21
- 2.6 Case study 1: synthetic mass spectrometry data set . . . 24
  - 2.6.1 Creation of the synthetic data set . . . 24
  - 2.6.2 Comparison of normalization performance . . . 28
  - 2.6.3 Comparison of normalization performance with additive noise 29
- 2.7 Case study 2: mass spectral imaging . . . 33
- 2.8 Conclusion . . . 36

3 Unmixing MSI Data 39
- 3.1 Principal Component Analysis . . . 42
  - 3.1.1 Standard Linear Principal Component Analysis . . . 43
  - 3.1.2 PCA Applied to MSI Data . . . 45
  - 3.1.3 Case Study: Rat Spinal Cord Nerve Tissue . . . 46
  - 3.1.4 PCA Summary . . . 51

3.2 Peak Intensity Weighted Principal Component Analysis 52
3.2.1 Standard PCA – Implicit Variable Weights 52
3.2.2 Peak Intensity Weighted PCA based Decomposition 53
3.2.3 PIW-PCA Summary 56
3.3 Independent Component Analysis 56
3.3.1 ICA versus PCA 56
3.3.2 ICA Applied to MSI Data 57
3.3.3 ICA Summary 58
3.4 Non-negative Matrix Factorization 59
3.4.1 What is the problem with PCA? Why enforce non-negativity? 59
3.4.2 Non-negative Matrix Factorization – The Concept 60
3.4.3 Case study: MSI Experiment on a Synthetic Tissue Immitation – The ESAT Logo Gold Standard 61
3.4.4 Case study: MSI Experiment on Mouse Brain 65
3.5 Conclusions 67

4 Spatial Querying 75
4.1 Exploring MSI Data 76
4.1.1 Unsupervised Data Mining 76
4.1.2 Supervised Mass Query 77
4.1.3 Supervised Spatial Query 78
4.2 Methods 79
4.2.1 Univariate Correlation-based Spatial Querying 79
4.2.2 Multivariate Least Squares-based Spatial Querying 80
4.3 Case Study 83
4.3.1 Materials & Methods 83
4.3.2 Results 84
4.3.3 Discussion 89
4.4 Conclusions 93

5 Enabling the Exploration of Large MSI Experiments Through Wavelets 95
5.1 Scaling issues 97
5.2 Methods 98
5.2.1 Wavelet Analysis 98
5.2.2 Discrete Wavelet Transform 99
5.2.3 DWT-based Decomposition of MSI Data 100
5.3 Case Study on a MALDI-TOF imaging experiment: Classical PCA versus DWT+PCA 101
5.3.1 Materials & Methods 102
5.3.2 Results & Discussion 103
5.4 Case Study on a MALDI-FTICR imaging experiment: Enabling Clustering 106

5.4.1 Materials & Methods 107
5.4.2 Results & Discussion 109
5.5 Conclusion 114

6 Co-localization Networks 115
6.1 Why examine ion co-localization in tissue? 116
6.2 Co-localization Network 117
6.3 Case study: An Example on Mouse Brain 118
6.3.1 Co-localization Measure 119
6.3.2 From Data Tensor to Network 120
6.3.3 Interpretation 125
6.4 Conclusions 128

7 Hormone Processing Pathway Analysis using Mass Spectral Imaging and Data-mining of Co-localization 131
7.1 Aim of the study 132
7.2 Introduction 133
7.3 Results 136
7.3.1 In situ unlabeled imaging of insulin in pancreatic islets using MSI: insulin-reporting ion image comparable to established immunostaining 136
7.3.2 Relative quantification in MSI data: insulin-reporting ion concentrations across the various mouse models 138
7.3.3 Other non-insulin ions that co-localize with insulin in the pancreatic islets: Extracting a multi-molecular signature of islet content 141
7.3.4 Comparison between islet signatures extracted from the different mouse models: Differential comparison of islet-content 143
7.3.5 Translation of the differential comparison of islet content to a metabolic representation: Mapping differential findings to hormone processing pathways 147
7.4 Discussion 150
7.5 Methods 154
7.5.1 Mouse models 154
7.5.2 Tissue preparation 154
7.5.3 Mass spectral imaging experiment 155
7.5.4 MSI data and ion images 156
7.5.5 Mapping 156

8 Conclusions and Perspectives 167
8.1 Conclusions 167
8.2 Future Perspectives 169

A Definitions **171**
A.1 Units 171
A.2 Definitions 171

B Supplementary Figures to Chapter 7 **173**

Bibliography **207**

List of Publications **232**

Curriculum Vitae **242**

List of Acronyms

CIC	Common Ion Current
CHCA	alpha-cyano-4-hydroxycinnamic acid (MALDI matrix molecule)
DIC	Differential Ion Current
DWT	Discrete Wavelet Transform
FTICR	Fourier Transform Ion Cyclotron Resonance
H and E	Hematoxylin and Eosin (staining)
ICA	Independent Component Analysis
IEC	Ionization Efficiency Correction
MALDI	Matrix Assisted Laser Desorption Ionization
MS	Mass Spectrometry
MSI	Mass Spectral Imaging
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
SIMS	Secondary Ion Mass Spectrometry
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
TIC	Total Ion Current
TOF	Time-of-flight

List of Figures

1.1 Four phases of a mass spectrometer. 5
1.2 Schematic overview of a MSI experiment. 5
1.3 Examples of ion images from a MSI experiment on mouse brain. 7
1.4 Examples of ion images from a differential MSI experiment on healthy versus diseased mouse spinal cord. 8
1.5 Structure of the thesis. 10
1.6 Screenshots from the "Unravel" software suite. 11

2.1 Example of normalization. 15
2.2 Rank-1 approximation of a matrix. 22
2.3 Schematic comparison of TIC-based normalization and IEC normalization. 23
2.4 Five content patterns. 25
2.5 Common and differential percentages of the total ion currents. 26
2.6 The noisy data set of 25 spectra. 26
2.7 Heat map of the spectra with ion intensity noise. 27
2.8 Heat map of the spectra without noise. 27
2.9 Heat map of the TIC normalized spectra. 28
2.10 Heat map of the IEC normalized spectra. 29
2.11 A zoomed-in look at the normalization of the mass spectrum from sample 6. 30
2.12 Heat map of the spectra with ion intensity noise in the presence of Gaussian additive noise. 31
2.13 Heat map of the TIC normalized spectra in the presence of Gaussian additive noise. 31
2.14 Heat map of the IEC normalized spectra in the presence of Gaussian additive noise. 32
2.15 Heat map of the un-normalized spectra from the MSI experiment. 34
2.16 Heat map of the TIC normalized spectra from the MSI experiment. 35
2.17 Heat map of the IEC normalized spectra from the MSI experiment. 36
2.18 Comparison of normalization results for three separate ion images. 37

2.19 Comparison of normalization results for three separate ion images, overlayed on a microscopy image of the tissue section to aid biological interpretation. 38

3.1 Unmixing of MSI data. 42
3.2 Graphical representation of matrix decomposition and reconstruction using principal component analysis. 44
3.3 Microscopic image of a transversal section of rat spinal cord. 47
3.4 Example ion images from the rat spinal cord MSI experiment detailed in section 3.1.3. 48
3.5 Percentage of variance explained by each principal component (PC). . . . 49
3.6 The first four loading vectors. 50
3.7 The first four score vectors. 51
3.8 Examples of ion count (or gray level) transformations. 54
3.9 Overview of the contrast stretching transformation performed on the spinal cord data set. 55
3.10 (top) Unweighted first PC. (bottom) Peak intensity weighted first PC. . . . 57
3.11 (top) Unweighted second PC. (bottom) Peak intensity weighted second PC. 58
3.12 Schematic comparison of PCA versus NMF based unmixing. 62
3.13 Construction of the ESAT logo data set. 64
3.14 PCA decomposition of the ESAT logo data set. 66
3.15 NMF decomposition of the ESAT logo data set. 67
3.16 PCA decomposition of the ESAT logo data set (surface plots). 69
3.17 NMF decomposition of the ESAT logo data set (surface plots). 71
3.18 Non-negative Matrix Factorization applied to a MSI experiment of a sagittal section of mouse brain. 73

4.1 A taxonomy of algorithms for the computational analysis of MSI data. . . 77
4.2 Overview of the spatial querying procedure. 81
4.3 Microscopic image of the sagittal section of mouse brain that is imaged in section 4.3. 83
4.4 Examining the upper *hippocampus* area in the sagittal mouse brain section using a binary query image. 85
4.5 Examining the upper *hippocampus* area in the sagittal mouse brain section using a gray level query image. 86
4.6 Examining the *corpus callosum* region in the sagittal mouse brain section using a registered microscopic image. 88
4.7 Downsampling the mask image from microscopic to MSI resolution. . . . 89
4.8 Results of the gray level query. 89

5.1 Overview of the DWT-enabled analysis of MSI data. 97

5.2 A multiple-level DWT performed on a mass spectrum from the case study of section 5.3. 100
5.3 Overview of the DWT-based PCA decomposition of MSI data. 102
5.4 Picture and ion image of the sagittal mouse brain section imaged in section 5.3. 103
5.5 Non-DWT-based principal components. 104
5.6 DWT-based principal components. 105
5.7 Clustering via the lower-dimensional wavelet coefficient space. 108
5.8 The MALDI-FTICR imaging data set is acquired from a section of rat testes. 109
5.9 A summary of the clustering results. 111
5.10 Clustering results. 113

6.1 Co-localization networks: going from a MSI data tensor to a network representation. 118
6.2 NMF decomposition of the peak picked MSI data set. 122
6.3 Initial co-localization network generated from the 92 most influential ion species in the mouse brain data set (92 nodes, 1198 edges). 124
6.4 Co-localization network of the mouse brain (92 nodes, 1198 edges, circular layout). 125
6.5 Co-localization network of the mouse brain (92 nodes, 1198 edges, organic layout). 126
6.6 Co-localization networks of the mouse brain at different amplitude thresholds. 127
6.7 Exploration of the network/spatial area duality. 129

7.1 The metabolic screening methodology discussed in chapter 7 fuses three distinct information sources. 132
7.2 Overview of the mass spectral imaging experiments and their computational follow-up. 135
7.3 Imaging of insulin in pancreatic tissue: the traditional labeled approach versus the unlabeled MSI approach. 158
7.4 Comparison of surface-normalized insulin concentration between the four mouse models. 159
7.5 Surface plot visualization of insulin content in the four mouse models. . . 161
7.6 Extraction of a set of data trends from the MSI measurements via nonnegative matrix factorization. 162
7.7 Comparison of islet specific signatures between WT and Ob/Ob, detail from m/z 2800 to 4000. 163
7.8 Differential fold changes between WT and Ob/Ob. 164
7.9 Hormone processing and cleavage products in Ob/Ob versus WT. 165

B.1 Tissue masks defined for the different pancreas sections. 174

B.2 The effect of preprocessing. 176
B.3 Further support for observed pattern in insulin concentration across the four mouse models. 177
B.4 NMF decomposition of WT. 178
B.5 NMF decomposition of DIA. 179
B.6 NMF decomposition of Ob/Ob. 180
B.7 NMF decomposition of SA/HF. 181
B.8 Islet-specific mean mass spectrum versus islet-specific NMF signature from the Ob/Ob experiment. 182
B.9 Comparison of islet specific signatures. Complete mass range. 184
B.10 Comparison of islet specific signatures. m/z 2100 to 4000. 186
B.11 Comparison of islet specific signatures. m/z 3800 to 5700. 188
B.12 Comparison of islet specific signatures. m/z 5500 to 7400. 190
B.13 Comparison of islet specific signatures. m/z 7200 to 12000. 192
B.14 Fold changes for the 419 peak-picked ion peaks in the signatures. 194
B.15 Fold changes for the 419 peak-picked ion peaks in the signatures, zoomed to -10 to + 10 range. 196
B.16 Differential expression comparisons between the islet signatures. 198
B.17 Differential expression comparisons between the islet signatures, zoomed versions. 199
B.18 Color-coded differential expression comparisons between the islet signatures, zoomed versions. 201
B.19 Color-coded mappings from peaks to molecules of interest. 204
B.20 Hormone processing and cleavage products in SA/HF versus WT. 205

List of Tables

5.1 Case study resource specifics, based on using a pixel×pixel covariance matrix for PCA. 106
5.2 Case study resource specifics, based on using a features×features covariance matrix for PCA. 107

CHAPTER 1

Introduction

You will have to brace yourselves for this – not because it is difficult to understand, but because it is absolutely ridiculous.

Richard Feynman

Anything will lase if you hit it hard enough.

Arthur Schawlow

The research summarized in this doctoral thesis finds itself at the crossroads between various different disciplines. From the different chapters, it will be apparent that the rationale and implementation of the various computational contributions draw upon fields ranging from chemistry and physics, over mathematics and computer science, to biology and medicine. The central goal, however, has always been clearly defined: develop tools to advance biomedical research and to acquire insight into the mechanisms of disease.

In recent years the molecular mechanisms underlying many of the worlds most pressing diseases have been shown to exhibit a complexity that exceeds earlier expectations. Groundbreaking developments such as the sequencing of the human genome in 2001 [1, 2] have enabled enormous leaps in our understanding of the molecular machinery of various diseases, but the post-genomic era quickly pointed out that studying the

genome alone will not be sufficient to truly solve some of the most important puzzles in pathology [3]. The need to examine proteins, peptides, and metabolites in biomedicine is further illustrated by the growing number of diseases that turn out to be caused (at least in part) by post-translational modifications [4] or other molecular processes that leave no trace in the genome. Inspired by these discoveries, this doctoral thesis focuses on advancing the study of proteins, and this specifically for measurements straight from organic tissue.

On the diagnostic side, the search for disease-reporting biomarkers in easily accessible fluids such as blood has met with variable success [5, 6, 7]. Most often, biomarkers are present in extremely low concentrations (only detectable by immune-based methods) and they often report secondary disease effects, thereby shedding only limited light on the true cause of the pathology. A direct analysis of the affected tissue, however, would prove very helpful, since it would report directly on the complex metabolic alterations of multiple molecules and it would present potential biomarkers undiluted and at their highest concentration, and potentially even tied to a specific tissue region or cell type. Aided by the abundance of tissue from biobanking efforts and the steadily growing range of animal disease models, there is an increasing need for *in situ* tissue analysis and molecular imaging methods to help unravel the spatial and biochemical complexities of disease phenotypes.

In the area of tissue based protein analysis and localized biomarker discovery, a relatively new technology called mass spectral imaging[1] (MSI) has received increased attention as a prime technique for exploratory pathological research [8, 9, 10]. The mass spectral imaging technology lays at the core of this doctoral thesis, the objective of which is to develop the computational analysis of MSI and thus, to unlock the true potential of this new molecular imaging modality.

Most areas of science employ a specialized vocabulary to convey their concepts and developments in a rigorous and succinct manner. The field of mass spectrometry is no exception, and it makes the research often difficult to understand for the uninitiated. Furthermore, mass spectrometry and its related disciplines seem to have a tendency towards using colorful acronyms, and a wide variety of these have become commonplace over the years (see Busch *et al.* for a summary introduction [11]). Most concepts are defined in the text as they first appear, but some of the fundamentals are listed for reference in appendix A. Acronyms employed throughout the thesis have been summarized for easy reference in the acronyms section at the front of the booklet.

[1] The names *mass spectral imaging* and *imaging mass spectrometry* are synonymous. For the remainder of this thesis, we wil use *mass spectral imaging*, as its acronym 'MSI' avoids confusion with the acronym for *ion mobility spectroscopy*.

1.1. What is Mass Spectral Imaging?

The field of proteomics is defined as the quantitative and/or qualitative analysis of the protein component within a organism, a tissue type or cell, and this under a set of well-defined environmental conditions (e.g. age, external/internal regulating factors, disease state,...). A proteomics study typically tries to acquire a snapshot of the current proteomic content in the cell. Assisted by highly selective separation techniques such as liquid chromatography and gel electrophoresis [12], mass spectrometry [13, 14] - the major driving force behind high-throughput proteomics - has been applied in several successful proteomics studies [15] and generated better understanding of various pathologies such as Alzheimer's disease [16, 17].

In fact, mass spectrometry (MS) has become the central technology for studies involving the large-scale characterization and identification of many different types of biomolecules in medicine and biology, and thus fulfills also a pivotal role in adjacent fields such as peptidomics (study of peptides), metabolomics (study of metabolites), lipidomics (study of lipids), and glycomics (study of sugars). An intermezzo on standard mass spectrometry is provided on the following page. In the wake of mass spectrometry, mass spectral imaging (MSI) has been gaining considerable momentum as a means of examining the spatial distribution of biomolecules in organic tissue. MSI is a new form of molecular imaging in tissue where tissue slices are analyzed directly in the mass spectrometer. Within the general framework of MSI, there is a variant based on matrix-assisted laser desorption ionization (MALDI) that is commonly applied for studying biomacromolecules such as polypeptides. Given the biomedical goals mentioned earlier, it will be this MALDI variant of MSI that this dissertation will focus on, and for the remainder of the text MSI will refer to MALDI-based MSI.

Since its initial application [8, 18], development of the MSI methodology has mainly focused on the physics of instrument development [19, 20, 21, 22, 23, 24, 25], research into novel applications of MSI [23, 26, 27, 28, 29, 30], and the chemistry of tissue preprocessing to enable reliable and reproducible ionization [9, 31, 32, 33, 34, 35]. MSI has shown significant growth in the past few years [10, 22, 36, 37, 38, 39] and has been applied successfully in pioneering studies on animal tissue [10, 18, 40, 41]. Now that wet-lab protocols for various tissue types are becoming increasingly available and stable, the major challenge in performing a MSI experiment is shifting more to the interpretation of the often massive amounts of acquired data.

An overview of the typical workflow of a MALDI-based MSI experiment is shown in Fig. 1.2. A very thin tissue section (e.g. 10 µm thick) is cut using a standard microtome. The tissue section is then mounted on a ITO-coated conductive glass slide. Subsequently a homogeneous layer of matrix crystals is applied on top of the tissue using a robotic spotter. The layer of matrix crystals is crucial to getting good signal from the tissue. To ensure reproducibility, the matrix is usually applied to the

Intermezzo: Mass spectrometry

Mass spectrometry is an analytical technique that delivers a high-grade measurement of the molecular masses present in a sample. It is most commonly used to characterize the molecular content of a sample, and if sufficient information is available it can go as far as to identify the molecules in question. The typical output of a measurement is a mass spectrum. This is a summary of all the molecular masses that were measured along a preset mass range, and usually takes the form of a plot showing ion abundance (or ion intensity) versus mass-to-charge ratio (m/z). An example of a mass spectrum is shown in Fig. 2.1. A mass spectrometer consists of four essential parts, as depicted in Fig. 1.1:

1. **ion source** – desorbes the analyte molecules from their medium, and ionizes the sample molecules so that they carry one or more electrical charges.
2. **mass analyzer** – separates ions according to their mass-to-charge ratio.
3. **ion detector** – detects the ions that make it through the mass analyzer and turns their abundance into an electrical signal.
4. **data acquisition** – turns the analog signal from the detector into an ion count number (and potentially performs some initial noise filtering).

For each phase, there are several different implementations possible, often using different physical principles. These implementations are typically optimized towards specific applications, conditions, or molecular species. Suffice it to say that the mass spectrometry name encapsulates a wide array of possible configurations.

Since the first studies by J.J. Thomson in 1912, MS has undergone several important evolutions that have carried the technique from anorganic fields such as material science to organic studies in medicine and biology. The importance of mass spectrometry for the study of biomolecules was further confirmed in 2002 when two of the inventors of the soft ionization techniques that enabled protein analysis, were awarded the Nobel Prize in Chemistry (John B. Fenn and Koichi Tanaka). These soft ionization techniques are Electrospray Ionisation (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI). The second ionization method, MALDI, is used to produce the MSI data described in this thesis. MALDI ionization involves a bombardment of the sample molecules with photons from a laser, to generate ions. The ionization technique requires that the sample is embedded into a chemical matrix solution, which upon crystallization enables maximal transfer of the laser energy into excitation energy for the sample.



Figure 1.1. Four phases of a mass spectrometer. (1) Ion source: desorption and ionization of the analytes from the sample. (2) Mass analyzer: filtering ions on the basis of their mass-to-charge ratio. (3) Ion detector: detection of ions that make it through the mass analyzer. (4) Data acquisition (and preprocessing): translation of ion-induced currents in the detector to ion counts in the mass spectrum.

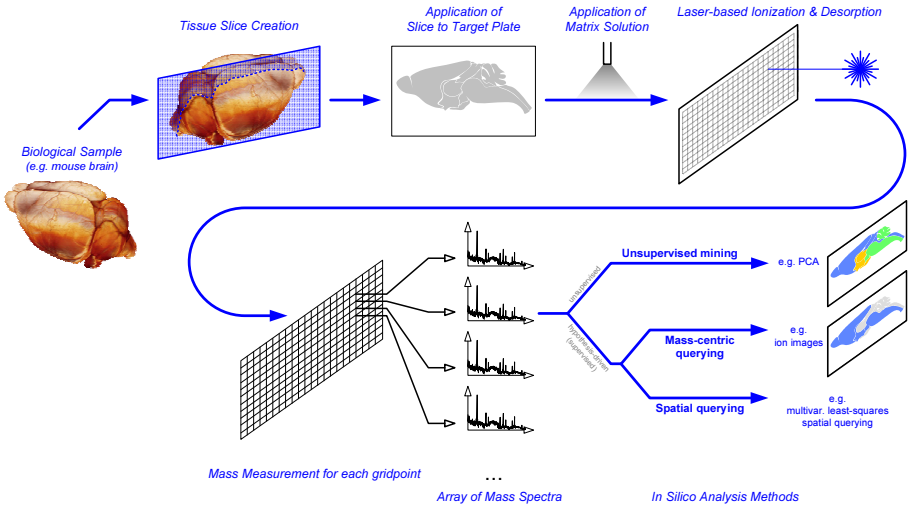


Figure 1.2. Schematic overview of a MSI experiment. The wet-lab side of the procedure consists of cutting an organic tissue section, mounting it on a MALDI target plate, applying an appropriate chemical matrix solution, and performing a MALDI mass spectral measurement at each grid point of a virtual array that has been superimposed on the tissue section. The result is an array of spots or 'pixels' covering the tissue section, with a mass spectrum linked to each individual pixel.

tissue using robotic means (typically Shimadzu's CHIP-1000® chemical inkjet printer, Labcyte's Portrait 630® acoustic droplet ejector, or Bruker Daltonic's ImagePrep® station). The slide carrying the matrix-coated tissue is then inserted into the MALDI mass spectrometer using a dedicated slide holder. A grid of measurement locations is defined on the tissue and a mass spectrum is measured at each grid location. The result is a large collection of individual mass spectral measurements, organized into a raster with known x and y dimensions.

Subsequently, the data generated by a MSI analysis are generally superimposed (*in silico*) on top of a microscopic image of the analyzed tissue section (see Fig. 1.3 and 1.4). By doing this, it becomes possible to acquire spatial density maps for var-

ious masses or ions, and these are essentially reporting the distribution of various biomolecules in the tissue section. In short, mass spectral imaging adds a spatial dimension to mass spectrometry-based biochemical analysis, and delivers insight into the localized distribution of biomolecules (e.g. proteins, peptides and metabolites) throughout an organic tissue section. MSI offers the important advantage that in a single analysis, a broad range of molecules can be registered, and this without the need for labeling. These properties make MSI a potent exploratory and confirmatory tool for tissue-based biomedical studies. A summary of the most apparent pros and cons of MSI, as compared to other imaging techniques, is presented in the following list.

Pros

- **Requires no chemical labeling of sample**.
- **Requires no prior hypothesis** – Excellent exploratory merit.
- **Very high specificity** – Better than most immunobodies.
- **Massively multiplexed** – Hundreds of ions detected in single experiment.
- **Insight into inter-molecule relationships** – Inherent multivariate nature.

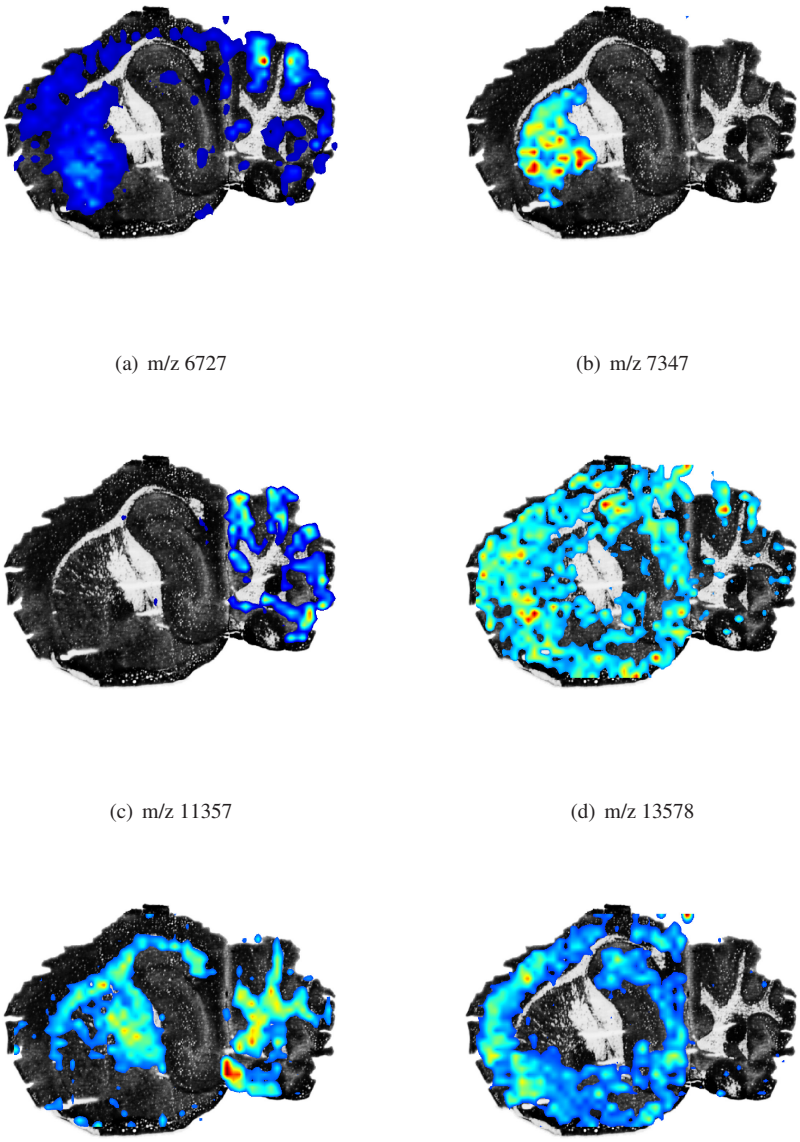
Cons

- **Relatively low spatial resolution** – For example compared to microscopy.
- **Sensitivity is tissue-dependent**.

1.2. MSI data and ion images

The data resulting from an MSI experiment consists of a grid of measurement locations or 'pixels' covering the tissue section, with an individual mass spectrum connected to each pixel. The data structure can be considered a three-mode array with two spatial modes (x and y) and one mass-to-charge mode (m/z). The size of an MSI data set is primarily influenced by two parameters: the number of measurement locations or pixels, and the number of scanned m/z bins. The pixel number increases as the tissue surface area that needs to be covered grows larger or the spatial resolution becomes finer. The number of m/z bins is proportional to the extent of the mass range and the granularity of the ion detector (= m/z step size).

Currently, MSI is most commonly used to produce ion images, showing the spatial presence of a single ion, mass, or molecule in the tissue (see Fig. 1.3 and 1.4 for



(a) m/z 6727

(b) m/z 7347

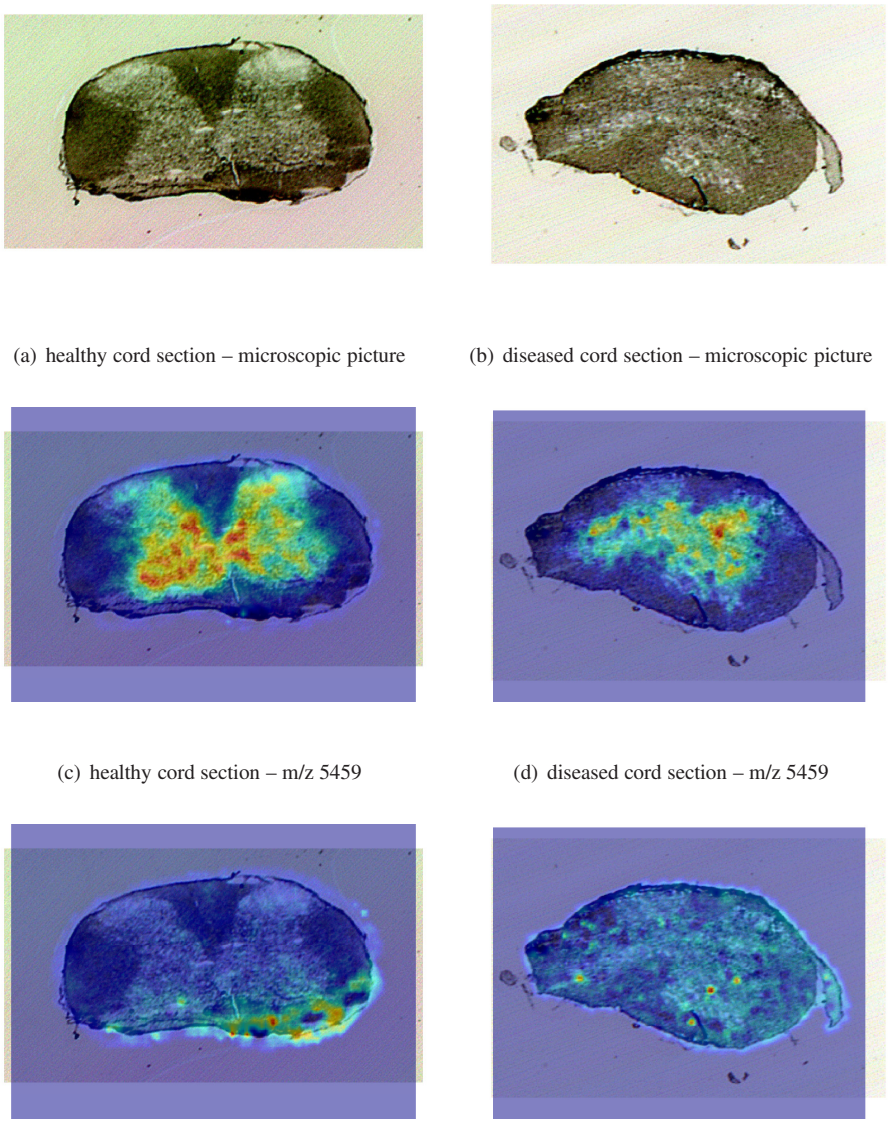
(c) m/z 11357

(d) m/z 13578

(e) m/z 17393

(f) m/z 22204

Figure 1.3. Examples of ion images from a MSI experiment on mouse brain (opaque overlay on a gray level microscopic image of the tissue section).



(a) healthy cord section – microscopic picture

(b) diseased cord section – microscopic picture

(c) healthy cord section – m/z 5459

(d) diseased cord section – m/z 5459

(e) healthy cord section – m/z 7944

(f) diseased cord section – m/z 7944

Figure 1.4. Examples of ion images from a differential MSI experiment on healthy versus diseased mouse spinal cord (transparent overlay on a microscopic image of the tissue section).

examples). From a mathematical standpoint, an ion image can be seen as a cross-section of the data tensor at a particular mass (or m/z). However, it should be noted that a univariate approach such as ion images, is under-exploiting the large amount of data that is acquired during a MSI experiment. As the biological questions become more challenging, there is an increased need to put that collected data to use. Indeed, since a few years now, there is a growing demand for multivariate data processing methods, capable of extracting nuggets of information from a sea of collected MSI data. The development of such methods is essentially the goal of this doctoral thesis.

1.3. Structure of the thesis

Figure 1.5 explains the general structure of this doctoral dissertation. It integrates the different achievements described in chapters 2 to 7 into the computational analysis workflow for MSI.

It should be noted that prior to all MSI data processing developments, a custom designed software tool had to be developed (the "Unravel" software suite, 20000+ lines of Java code) in order to make the computational analysis pipeline independent of the particular mass spectrometer platform the data was acquired on. Figure 1.6 shows two Unravel screenshots. Simultaneously a reverse engineering effort was necessary to extract MSI measurements from the proprietary vendor formats.

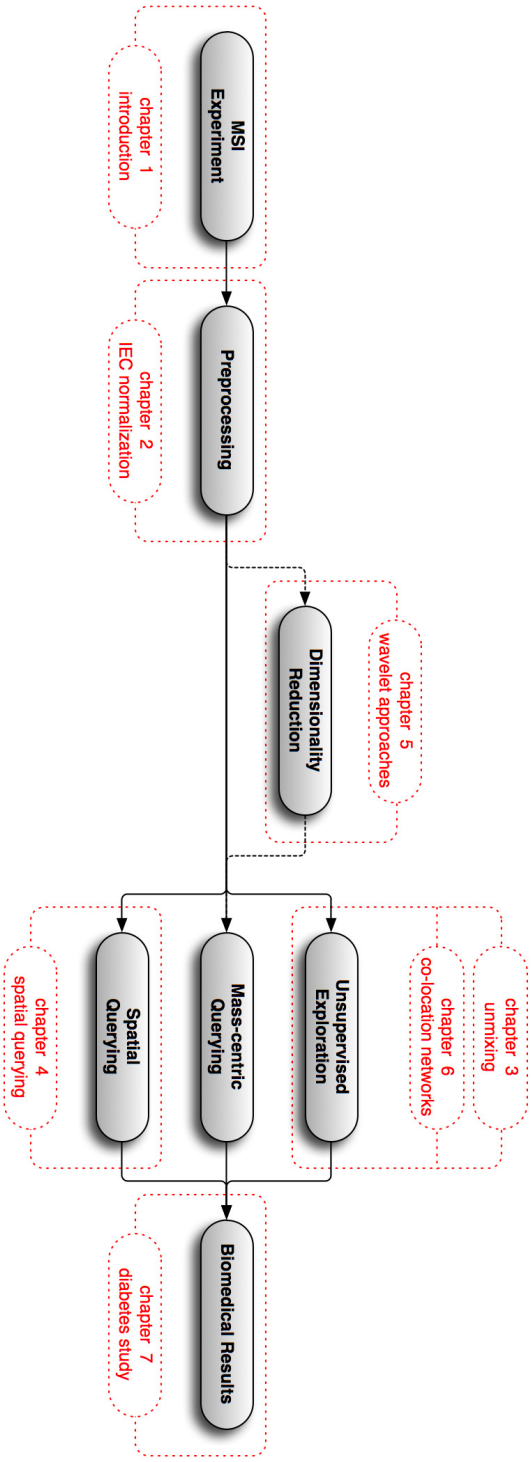
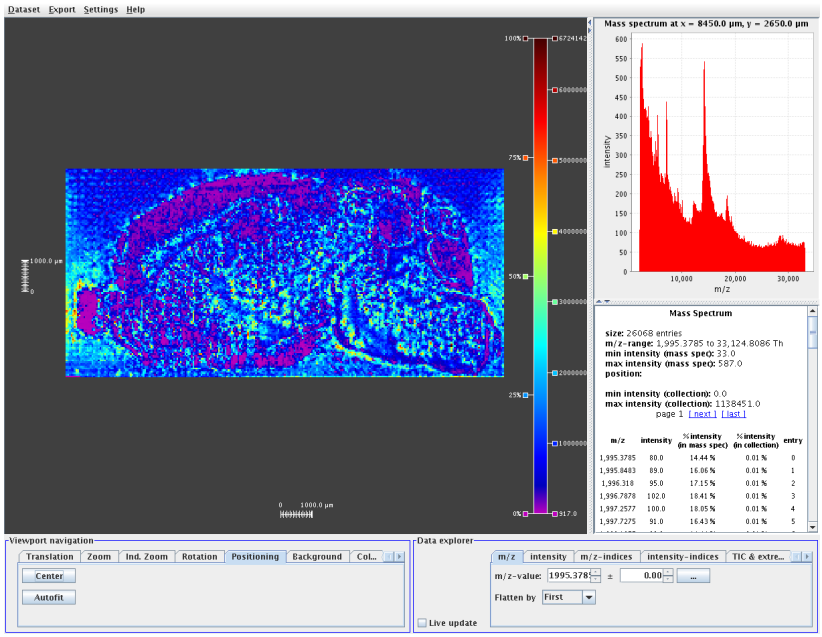
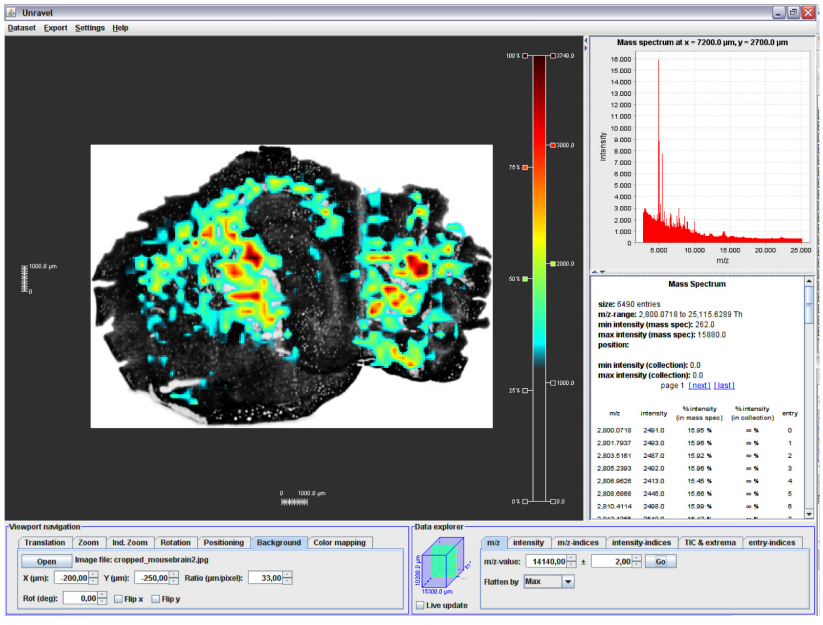


Figure 1.5. Structure of the thesis.



(a)



(b)

Figure 1.6. Screenshots from the “Unravel” software suite.

CHAPTER 2

Ionization Efficiency Correction

In physics, you don't have to go around making trouble for yourself - nature does it for you.

Frank Wilczek

This chapter is a derivative of patent application #GB1005959.0: Van de Plas R., De Moor B., and Waelkens E., "Ionization Efficiency Correction, a novel normalization method for mass spectrometry and mass spectral imaging".

In all fields of mass spectrometry, including mass spectral imaging, it has been shown that proper preprocessing of the acquired data is crucial to obtaining a good interpretation of the measurements [42, 43, 44, 45]. A preprocessing method will try to remove undesirable variation or noise from the mass spectral measurements prior to direct human interpretation or higher-level statistical analysis. Most preprocessing methods will attempt to counteract only a specific noise type, and as a result the preprocessing phase of a study can entail various steps. Typical examples include:

- **baseline correction**: quantifying and removing the chemical/instrumental noise background;
- **calibration**: adjusting the m/z values according to a set of known calibrants;
- **alignment:** projecting several spectra onto a common m/z scale;

- **normalization**: projecting peak heights from several spectra onto a common intensity scale;
- **smoothing/denoising**: removing ion detector and data acquisition induced jitter;
- **peak detection**: converting a mass spectral profile to a discrete set of peaks;

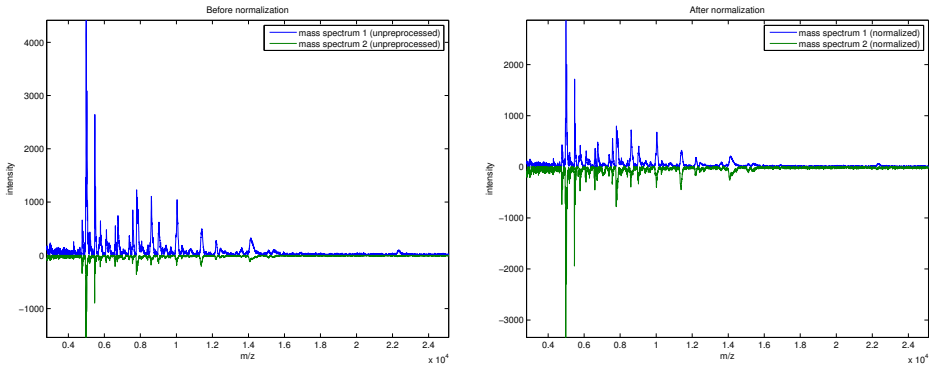
Overall the goal of the preprocessing phase is to filter undesirable influences from the raw measurements, and to provide a cleaned-up data set for further downstream analysis.

The central topic of this chapter is the normalization of ion intensities across different mass spectra. The ability to compare peak heights from one mass spectrum to another, is of prime importance both for standard mass spectrometry as well as mass spectral imaging. This chapter starts by explaining the rationale behind peak height normalization, and by discussing some the causes of ion intensity noise. Subsequently, there is a closer look at one of the most commonly used normalization methods in mass spectrometry today, and a problem inherent to this approach is highlighted. Finally, a new normalization method is introduced that overcomes this *caveat*. The performance of this new procedure, called Ionization Efficiency Correction (IEC), is demonstrated on two case studies. The first case follows a common experimental design in the field of biomarker discovery, in which mass spectrometry is used to compare the content of different samples. The data set is synthetically generated to provide a gold standard against which the algorithm's performance can be weighed. The second case study is a mass spectral imaging experiment on a sagittal section of mouse brain, and highlights the value of the new approach from an imaging standpoint.

2.1. Why normalization?

One of the primary preprocessing steps in mass spectrometry is normalization. The goal of a normalization procedure is to normalize the ion intensities and make peak heights comparable from one mass spectrum to another. Many applications of mass spectrometry require information not only on the presence or absence of an ion species, but they also require some indication of quantity regarding those ions. As there is a relationship between the concentration of an analyte and its ion count as reported in a mass spectrum, peak heights can serve as indicators of quantity. However, peak heights can only be used reliably when influences other than abundance can be minimized.

The need for reliable peak heights, comparable across different mass spectra, spans a very wide range of biochemical applications. In qualitative analyses aiming to understand the content of a sample, peak height is sometimes used to establish an indication



(a) Un-normalized (b) Normalized

Figure 2.1. Example of normalization. (a) Before ion intensity normalization is applied, the peak heights differ substantially between both spectra. (b) After normalization, the spectra are put on the same intensity scale, enabling direct comparison of peak height.

of confidence, by enabling the calculation of a signal-to-noise ratio (SNR) for each ion species under consideration. Qualitative analyses are typically found in areas such as protein identification [46, 47, 48, 49, 50] and the search for post-translational modifications [4, 51]. In quantitative applications, peak height as an indicator of abundance lies central to the analysis. Quantitative analyses span a multitude of approaches ranging from isotope labeling to label-free methods, and from absolute quantification to relative profiling. An example of absolute quantification is the use of mass spectrometry as a pharmacokinetic assay, tying an absolute peak height to a certain concentration of the target analyte per unit volume or mass of sample [52, 53, 54, 55, 56]. Most quantitative applications of mass spectrometry however are of the biomarker profiling type. These do not ascribe a meaning to absolute peak heights, but look rather for relative peak height changes that can be tied to a particular disease state or condition through classification [53, 57, 58]. One particular area in which peak heights need to be directly compared from one spectrum to the next is mass spectral imaging. An ion image produced from a MSI experiment is simply a false color representation of peak height across an organic tissue section. The need for comparable peak intensities from one pixel to another is therefore readily apparent.

2.2. What is the nature of ion intensity noise?

A mass spectrometer establishes a physical relationship between a particular molecular species and the peak height recorded at a certain mass-over-charge value. In

general, quantity is one of the most important factors in this relationship, meaning that a larger amount of molecules usually results in a larger ion count at the corresponding mass-over-charge bin. However, the link is rarely if ever as simple and as linear as that. In fact, peak height can be perturbed by wet lab factors such as differences in sample preparation and sample content [34]. It can also be influenced by instrument factors such as ionization efficiency [59] and ion detector saturation. In the case of mass spectral imaging, additional factors are introduced such as the topology and texture of the tissue, and the variation of matrix coating across the section. The best strategy is of course to minimize these noise factors on the wet lab side by taking care to keep all experimental parameters constant from one measurement to the next. A good example of such efforts includes the matrix deposition in MSI experiments, often performed by robotic spotters in an effort to put down a matrix coating as homogeneous as possible [34]. In practice however, some of the relevant parameters cannot be controlled to the extent necessary to do away with ion intensity noise. Compensation for this unavoidable noise type will therefore fall to computational methods on the *in silico* side.

Some of the ion intensity noise factors are molecule-specific, and their influence is therefore local to a particular m/z area or bin (e.g. an ion overshadowed in the detector by a more abundant co-ionizing ion of nearby mass, or ions that, due to conformational reasons, are not very inclined to ionize). Other noise factors have a global effect across the entire mass range (e.g. variation in the concentration of matrix crystals). The molecule-specific factors usually pose few problems for inter-spectrum comparisons as long as the same ion species is being compared across all spectra. More precisely, the goal of a differential analysis between spectra is not to compare abundance from one ion species to another species located elsewhere on the m/z scale, but rather to compare abundance of the same ion species from one sample to the next. This means that as long as the molecule-specific factors are kept the same across the different mass spectra, by keeping the experimental parameters as constant as possible, these effects usually need not be explicitly removed. The effect of the global noise factors however is usually much more extensive and can rarely be left unadjusted. This is why most normalization procedures in mass spectrometry target mass range-wide intensity noise.

An explicit enumeration of the various physical noise effects in the ion source, the mass analyzer, and the ion detector would be a difficult endeavor as some of these processes are not yet fully understood, while others can be described with only the most elaborate of mathematical models [60]. Instead of attempting to model each effect explicitly, normalization methods in mass spectrometry take an empirical stance, modeling global ion intensity noise simply as a straightforward linear scaling factor across the entire mass spectral profile. Although there is something to be said for more elaborate nonlinear noise models along the m/z axis, there are often insufficient clues as to the real ion intensities to fit a model more complex than a global scaling.

The assumption of a global scaling due to noise, and counteracting it with a reverse scaling factor, has empirically been shown to give good results. One of global scaling's strong points is that in most cases it captures the bulk of the ion intensity noise, while at the same time avoiding overfitting. The problem of overfitting a noise model in mass spectrometry is not trivial. Usually there is very little information available external to the measurement (unless well-characterized calibrants are spiked into the sample, which is often unpractical). Additionally, most general mass spectrometry studies have an insufficient number of replicate measurements to reliably generalize from. Some more advanced models have been formulated [61, 62], but in general a global scaling factor remains the standard model for ion intensity noise.

Given the MALDI-based nature of the imaging experiments described in this thesis, it serves to mention that any mass spectrometry experiment using this type of ionization is particularly prone to ion intensity noise, making a normalization preprocessing step practically a prerequisite. The reason for the intensity noise lies in the use of matrix molecules to enable ionization. In MALDI-based measurements, analytes need to be embedded into matrix crystals to keep them intact during the laser-induced desorption and ionization phase. It is clear that in such a setup the number of analyte ions that are formed in the ion source, will not only be dependent on the amount of analyte present, but also on the amount of matrix crystals that are present. However, growing crystals on an analyte in a sample well and later repeating that process on another sample in another well, while trying to obtain the same concentration of crystals, is not an easy task. In the MALDI mass spectrometry field substantial research has gone into improving the matrix molecules [35, 59, 63], achieving more reproducible matrix deposition [34], and understanding the topic of matrix *hotspots* [34]. Hotspots are certain areas in the sample showing better crystallization and therefore higher ion intensity and signal-to-noise ratio. Although significant progress has been made, the reproducibility of matrix conditions remains a point of attention in MALDI-based research, be it standard as well as imaging-oriented. MSI adds to the matrix difficulties with additional effects from the tissue layer, from which the analyte needs to be desorbed and on which the matrix needs to crystallize. This means that in MSI experiments it is not uncommon to see ion intensity not only influenced by matrix conditions, but also by the particular cell type from which the measurement is taken. Both the medium as well as the quality of the matrix embedding cause amplification or attenuation of ion formation in the source.

2.3. The common approach: Total Ion Current (TIC)

The common normalization methods in mass spectrometry operate on the global scaling assumption mentioned above. These algorithms consider ion intensity noise to be an undesirable scaling factor, which is different for every mass spectrum. The rem-

edy seems evident: rescale the noisy spectrum with a scaling factor multiplicatively inverse to the noise scaling factor. The problem thus presents itself as a two-step procedure:

1. Find an estimate of the noise scaling factor.

2. Scale back the mass spectrum with the inverse scale factor.

The problem is that the noise scaling factor is unknown, and the only information available is the noisy spectrum. Unless external information is provided regarding the true ion intensities of the spectrum (e.g. an externally calibrated peak intensity), the algorithm has little clues on which to base its estimate in step 1.

The problem definition above describes the situation where the goal is to remove the ion intensity noise from the mass spectrum altogether. However, most experimental setups that include a cross-comparison between mass spectra only seek relative peak height changes. As mentioned in the introduction, normalization aims to project peak heights from several spectra onto a common intensity scale to allow relative comparison. Whether or not that common intensity scale is the true intensity scale in the absence of noise, is irrelevant to most studies. The only requirement for relative comparisons of peak height is to establish a common ground between the spectra to scale towards. This common ground can be any measure that connects the ion intensities of the different spectra together. Various schemes have been suggested, particularly in the context of mass spectral search algorithms. One naive example is base peak normalization [64], where spectra are scaled relative to each other such that their highest peak is equally high in all spectra. Another example is a method based on total ion current, which has been shown to produce much better results [64, 65]. The total ion current (TIC) of a mass spectrum is the sum of the separate ion currents carried by the different ions contributing to the spectrum [66]. In mathematical terms, the TIC can be considered the sum of all ion counts collected in a mass spectrum, or the integral over the mass spectral profile. Scaling mass spectra such that they have the same TIC has become an *ad hoc* norm for normalization in many areas of mass spectrometry. The rationale behind scaling towards a common TIC value makes physical sense. Distinct experiments, but executed with identical experimental parameters (e.g. laser intensity, sample amounts,...), should arguably yield similar amounts of ions.

Most TIC-based normalization algorithms aimed at enabling relative comparisons, consist of these two steps:

1. For each mass spectrum, calculate its TIC

2. Scale back the mass spectrum with the inverse of its TIC.

Step 2 will scale all spectra to a TIC of one. Some algorithms however will scale towards a common TIC value (e.g. the median TIC of all spectra) instead. One reason for this is interpretation, in the sense that the normalized peak heights will be on a scale not too far removed from the original ion count values that were collected. Another reason is numerical precision and memory requirements. Because of memory and computation considerations, some implementations are better served with an integer based ion count than real values between zero and one. Theoretically all these TIC-based approaches are equivalent as they all retain the relative differences regardless of the absolute intensity scale of the normalization result.

Figure 2.1 gives an example of how a TIC-based normalization works for the comparison of two real mass spectra. The spectra come from a tissue profiling experiment, and the normalization algorithm that is applied, is a standard TIC-based implementation called *msnorm*, provided with the Bioinformatics Toolbox of MATLAB (The MathWorks, Natick MA, USA). The algorithm took 100 percent of the spectras' TIC into account to calculate the resulting scaling factor. Although both spectra only contain positive values, one spectrum is shown as negative to enable easy comparison. A schematic overview of the two-step procedure is also shown in Fig. 2.3.

2.4. What is the problem with the common approach?

Ionization efficiency is defined as the ratio of the number of ions formed to the number of electrons or photons used in the ionization process [66]. If the laser intensity of a MALDI mass spectrometer is kept constant from one measurement to the next, ionization efficiency equates to the yield of ions formed in a mass spectral measurement. Scaling spectra to have the same ionization efficiency could therefore be considered scaling towards the same ion yield across all mass spectra. When all other parameters are kept the same, such an operation would indeed counteract the amplifying and attenuating effects presented in section 2.2. The key point however is that the statement "when all other parameters are kept the same" includes the sample content.

At first glance it seems that if ionization efficiency at constant ionization energy is equal to ion yield, and if the total ion current is proportional to the ion yield hitting the detector, TIC can be a good measure for ionization efficiency. This reasoning forms the rationale behind most TIC-based normalization methods, and as a result these methods will try to equalize the total ion current across all spectra. The problem with TIC-based normalization methods is that this reasoning does not take into account differences in sample content, and their scaling is therefore done on the basis of a measure that is only partially proportional to ionization efficiency. The reason that sample content plays a role is that differences in molecular content will produce different peak patterns in the spectra. The ion counts tied into the peaks that are differential between the spectra are added to the TIC, but in reality these ion counts are not

the result of a change in ionization efficiency, which is what the TIC is expected to report. The potential harm this wrong assumption holds, will depend on the particulars of the measurements (e.g. How different are the peak patterns from sample to sample?, How much of the TIC is differential?,...). Although the repercussions might be negligible in some cases and the use of TIC normalization is certainly better than no normalization at all, the influence of differential peaks on TIC-based normalization is almost always present. Most studies compare samples that differ in molecular content. Often finding the differences is the reason for performing the study in the first place, and the alternative would invalidate any effort to look for biomarkers. In those cases the particulars of the measurements will decide whether TIC-based normalization will underperform.

As an example, lets consider two samples. Both contain the same amount of analyte A. Only the second sample additionally contains analyte B in an amount similar to A. If both samples are coated with matrix and measured using the same laser power and under the same experimental conditions, you will normally get one peak of analyte A in the spectrum[1] of sample one and two peaks (A and B) in the spectrum of sample two. Even if the A peak is somewhat diminished in height due to the presence of B, it is clear that a TIC-based inverse scaling of sample two would count both A and B ion counts and could severely bias the height of peak A in sample two downward. By including the ion counts from differential peak B, the TIC of sample two is estimated two times higher than the TIC of sample one. The result is that sample two is scaled down approximately twice too strongly. The result of a TIC-based normalization would be that peak A in sample two is only half as high as peak A in sample one, although they should represent the same quantity of analyte A. A graphic example of how TIC can steer normalization wrong is also shown in the schematic of Fig. 2.3.

Some TIC-based methods can be made to compensate for the bias somewhat, by removing ion counts from peaks whose height falls within a certain user-defined quantile range. However, the same problem remains as there are no real rules of thumb for setting the value of the quantile parameter and the same effect would still be happening, only focused on smaller peaks.

To summarize, ionization efficiency is a good basis for normalization between spectra. However, we posit that TIC is not a good measure for ionization efficiency. The reason is that ionization efficiency should be compared and rectified on the basis of the parts that are common between two spectra. Not on the basis of the parts that are differential, and the TIC cannot tell these two apart. In short, the more similar the content between samples, the better the TIC-based scaling. The more dissimilar the samples, the greater the bias. In studies that have a substantial amount of differential peaks from one sample to another, the bias of TIC-based methods can become troublesome. Particularly imaging mass spectrometry is vulnerable as these data sets

[1] Let us only consider molecular ions for this example and no fragmentation ions.

typically contain spectra from a wide range of different cell types and anatomical regions.

2.5. A novel approach: Ionization Efficiency Correction

Given the problem with TIC-based methods highlighted in the previous section, we formulate a new normalization approach, named ionization efficiency correction (IEC). The method follows a rationale similar to the one of section 2.3, namely that ionization efficiency provides clues towards projecting peak height intensities from different spectra onto a common scale. The difference with TIC-based methods lays in the fact that only ion counts from analytes common to all spectra are used to calculate the normalization. The participation of differential peaks in the scaling factor calculation is minimized. As we will demonstrate, TIC-based methods cannot tell the difference between common and differential content, while IEC can.

For reasons of clarity, let us define two additional concepts. The common ion current (CIC) of a mass spectrum is the sum of all ion counts belonging to the part of the mass spectrum that is common in relative profile to other mass spectra in a data set. The differential ion current (DIC) of a mass spectrum is the sum of all ion counts belonging to the part of the mass spectrum that is not common in relative profile to other mass spectra in a data set.

$$\mathrm{TIC} = \mathrm{CIC} + \mathrm{DIC} \tag{2.1}$$

Ionization efficiency correction is a three-step normalization process:

1. Separate the part common to all spectra from the parts that are differential.
2. For each mass spectrum, calculate its CIC.
3. For each mass spectrum, scale back the mass spectrum with the inverse of its CIC.

Step two and three are similar to the operations applied in a TIC-based algorithm. The difference is that the traditional TIC is replaced by the CIC value, which is a better estimate of ionization efficiency. The crux of the method lies in obtaining the estimate for the CIC, which is the responsibility of step one. Given a data set of multiple mass spectra, the task of step one is to identify which parts of the relative profiles of all these spectra are commonly found across the entire data set. If such a common relative expression profile can be found for all spectra with an individual scaling factor for each mass spectrum, the CIC of a spectrum is simply the area-under-the-curve of the scaled common profile.

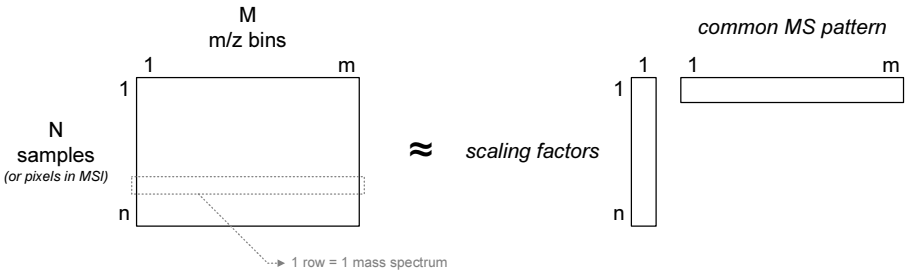


Figure 2.2. Rank-1 approximation of a matrix. A rank-1 approximation attempts to decompose the matrix into a single product of vectors, while potentially optimizing additional constraints. In the case of IEC, the matrix consists of mass spectra and the decomposition produces a common mass spectral profile in one vector (optimized to avoid differential peaks), and a set of scaling factors in the other vector.

The task of extracting from a collection of mass spectra a single pseudo-mass spectrum that only contains common ion peaks and relative peak heights can be approached in a number of different ways. Considered from a linear algebra perspective, the problem of step one can be translated to the mathematical domain in terms of the search for a rank-1 approximation of the two-mode array or matrix containing all the mass spectra[2]. A rank-1 approximation is a concept often used in the context of matrix decomposition methods. The goal of a rank-1 approximation is to approximate a matrix with the product of two vectors, as depicted in Fig. 2.2. In addition to looking for a rank-1 approximation, the search should be optimized towards avoiding the inclusion of differential peaks. Given a data set of N mass spectra that each contain M m/z bins, the task of step one will be to look for a rank-1 approximation of the $N \times M$ data matrix, while penalizing differential peaks in the profile vector. This approximation entails a $1 \times M$ vector containing the common mass spectral profile and a $N \times 1$ vector containing scaling factors that will be used to calculate the CICs. Written as an optimization problem, this becomes:

$$\begin{array}{rll} \text{minimize} & \|D - sp^T\|^2 & \\ \text{subject to} & p \text{ containing no differential peaks} & \\ \text{where} & D \in \mathbb{R}^{N \times M} & \text{(mass spectral data set)} \\ & s \in \mathbb{R}^N & \text{(scaling factors)} \\ & p \in \mathbb{R}^M & \text{(mass spectral profile)} \end{array} \tag{2.2}$$

Within the matrix decomposition field there are several different methods that can perform a rank-1 approximation, but most are tweaked towards optimizing different

[2]Notice that the mathematical definition of a matrix applies here. This concept has no relation to the chemical matrix in which analytes are embedded for MALDI measurements.

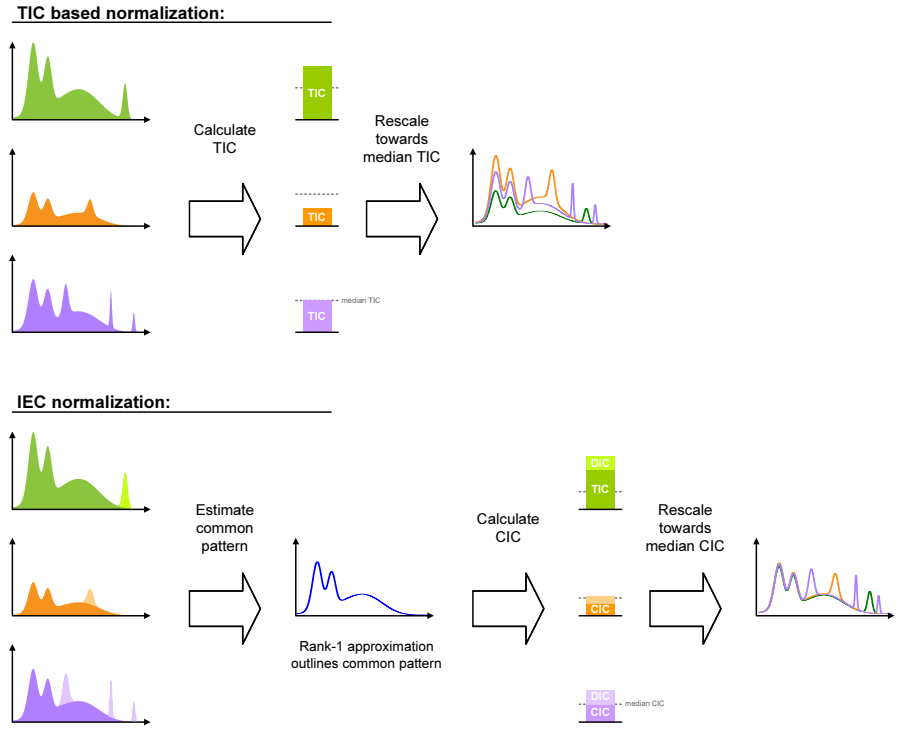


Figure 2.3. Schematic comparison of TIC-based normalization and IEC normalization. Notice the extra step in IEC, which employs a matrix decomposition method to discern common peak patterns (leading to CIC) from differential peak patterns (leading to DIC) in the mass spectra.

characteristics of the decomposition. Examples include principal component analysis (PCA) [67, 68], independent component analysis (ICA) [69], and singular value decomposition (SVD) [70, 71]. Because of the need to minimize differential peaks, none of these algorithms provides out-of-the-box the rank-1 approximation we need. Empirically however, we have attained good results with a modification of the non-negative matrix factorization (NMF) algorithm [72]. In our implementation the basic NMF algorithm is run several times on the data set in rank-1 mode, but each iteration the differential residuals are deducted from the data set. This approach converges towards a rank-1approximation with little or no remnants of non-common features in the profile. We use this decomposition algorithm as the driving force behind step one, but it is clear that this phase of the algorithm is an inviting area for more advanced developments in the future. Conceptually, the IEC method can be considered a normalization framework in which a particular decomposition engine can be dropped to estimate the CIC.

Once a pseudo-spectrum of the common mass peaks has been established, accompa-

nied by a scaling factor for each individual mass spectrum, we have the material necessary to tell the common ion counts and the differential ion counts apart. Step two then estimates the CIC of a mass spectrum as the area-under-the-curve of the common profile (determined in step one), scaled by that mass spectrum's individual scaling factor (also determined in step one). Step three scales back the entire mass spectrum, not just the common parts, with the inverse of the CIC. A schematic overview of the TIC-based methodology and the difference with the newer IEC approach is shown in Fig. 2.3. In analogy to TIC-based methods, IEC could be labeled a CIC-based method.

2.6. Case study 1: synthetic mass spectrometry data set

The objective of the first case study is to give a concrete demonstration of the problems inherent to TIC-based normalization methods, and to quantify the improvements introduced by IEC normalization. A thorough comparison of normalization methods requires the availability of a gold standard against which the methods' performance can be measured. As real-world biological case studies can rarely provide sufficient characterization of the ion intensity noise on the measurements, this first case study will center on a synthetic mass spectral data set.

2.6.1. Creation of the synthetic data set

The data set will mimic a typical experimental setup aimed at biomarker discovery. The data describes 25 individual mass spectral measurements, that are engineered to have both common as well as differential ion peaks. To ensure the authenticity of the study, the spectra are generated from a base pattern, which is a real mass spectrum from a profiling study on mouse brain. The peaks from the base pattern will be present in all 25 spectra at various peak heights, and will fulfill the role of common pattern across the data set. Adding differential peaks to the base pattern generates four additional classes of content. The added peaks are the mimicked by adding Gaussian distributions of various height and variance to the base pattern. To test the robustness of the algorithms, the shape of the additional peaks is varied. Pattern one adds only a few slim peaks. Pattern two contains different shapes through a fusion of peaks. Pattern three adds primarily wide peaks of low height, while pattern four contains a mixture. All five patterns span a m/z range from 2800 to 25000, as depicted in Figure 2.4.

By adding peaks to the base pattern, all patterns describe an area-under-the-curve or TIC that consists of both a common part and a pattern-specific differential part. The common part is equal to the base pattern and will give rise to the common ion current or CIC of the spectra generated from these spectra. The remainder of the TIC will

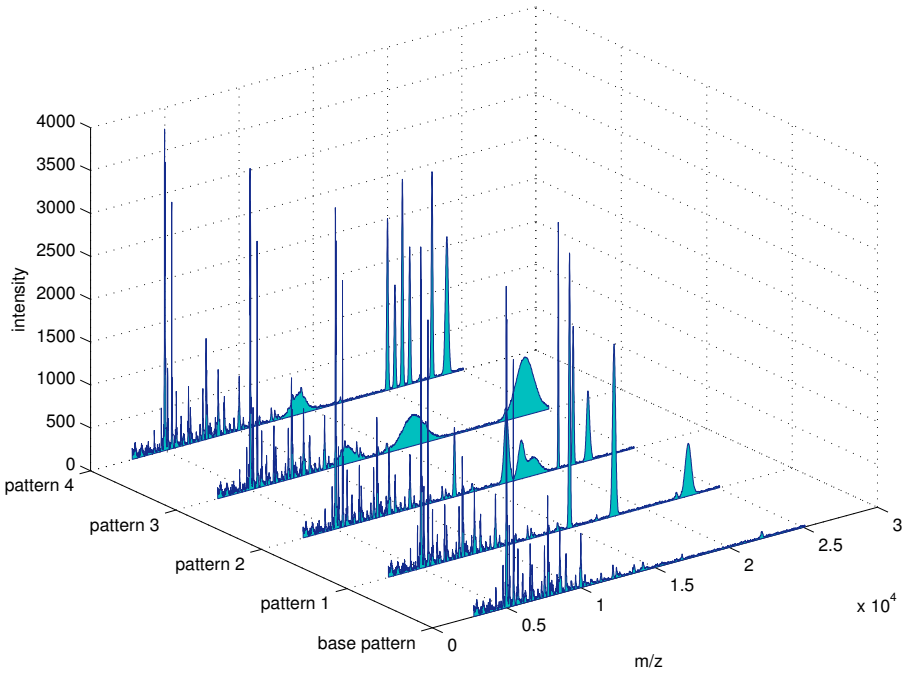


Figure 2.4. Five content patterns. These five patterns form the basis for the 25 individual mass spectra that make up the synthetic data set. Notice the common peaks, equal to the base pattern, and the pattern-specific differential peaks added to the second half of the mass range.

give rise to the differential ion current or DIC of the derived spectra. Figure 2.5 gives per pattern an overview of the proportion CIC versus DIC.

The patterns represent five different content classes from which five 'samples' per class are added to the data set. Each pattern gives rise to five separate mass spectra, individually scaled with a noise scaling factor to mimic ion intensity noise. The noise scaling factors are randomly generated, and include both amplified as well as attenuated cases. By performing this scaling, ion intensity noise is added to the data set, and the ion intensity scales of the different mass spectra are dispersed. The goal of the normalization algorithms will now be to reverse the situation to a level that allows for solid relative comparison. Unlike in real-world case studies, here the noise scaling factors are known to us and can serve as a gold standard for normalization performance. Figure 2.6 shows the noisy spectra that were generated and their respective scaling factors.

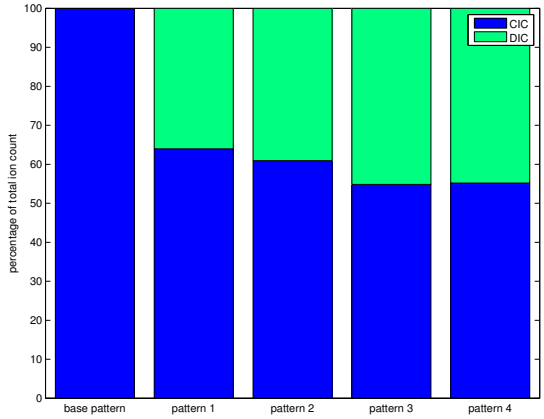


Figure 2.5. Common and differential percentages of the total ion currents. The base pattern consists completely of peaks common to all spectra. The additional patterns 1 through 4 contain respectively 64, 61, 55, and 55 percent common ion counts.

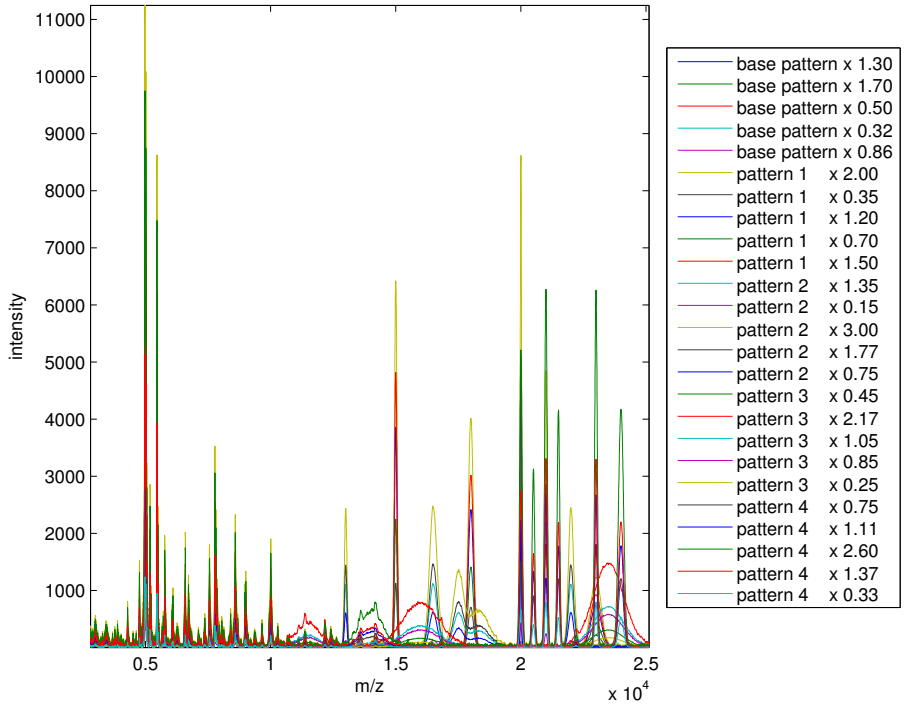


Figure 2.6. The noisy data set of 25 spectra. The legend indicates the pattern from which the spectrum stems and with what scaling factor the spectrum was perturbed.

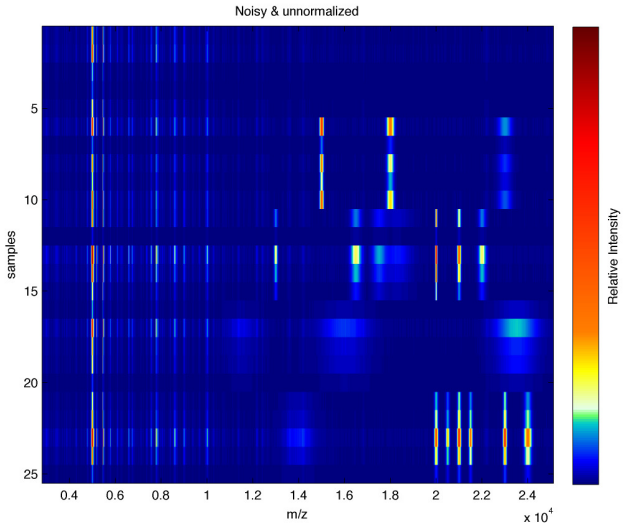


Figure 2.7. Heat map of the spectra with ion intensity noise. The attenuations and amplifications of the rows are clearly visible, and make direct comparison of peak height from one sample to another impossible. The multiplicative noise gives a false sense of variation in quantity. In reality, every peak stems from the same 'amount' of ions in its pattern.

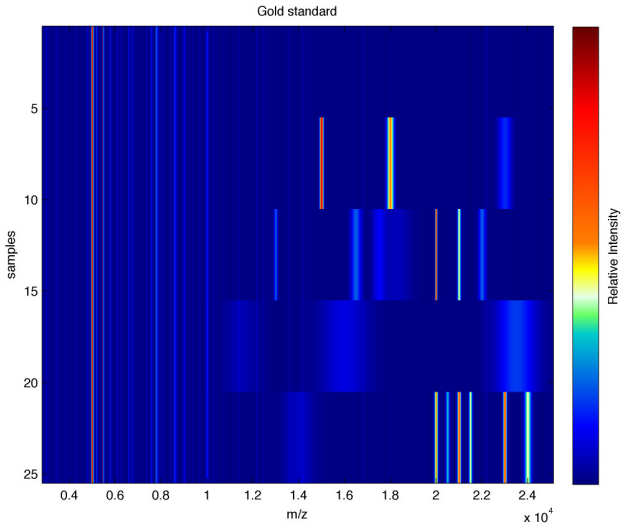


Figure 2.8. Heat map of the spectra without noise. This is the result of reverse scaling using the known noise factors, and can serve as a gold standard for the normalization process.

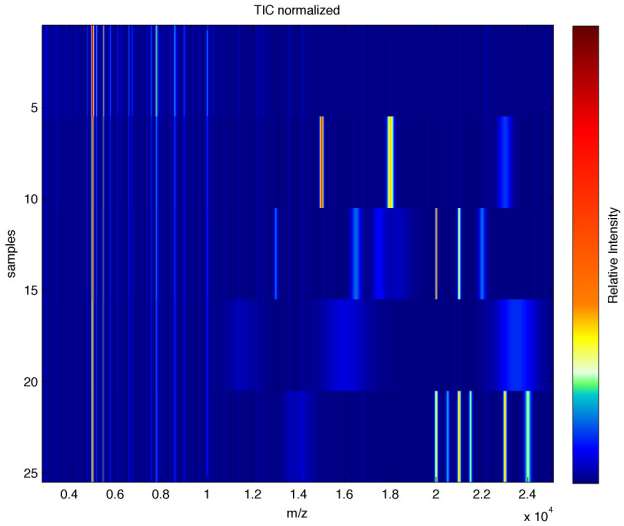


Figure 2.9. Heat map of the TIC normalized spectra. Notice that samples 1 through 5 are correctly normalized relative to each other. These stem from the base pattern and their TIC is equal to their CIC, explaining why TIC based normalization performs well in these cases. However, samples 6 through 25 show a clear spectrum-wide peak height difference when compared to the first five samples. This striping effect is artificial peak height variation introduced by the TIC-based algorithm, and could be mistaken for genuine biological variation.

2.6.2. Comparison of normalization performance

First, a gold standard for normalization is established as the inverse of the noise scaling factors that were used to create the data set. Then, both TIC-based and IEC normalization are applied to the noisy data set. The results are summarized in Fig. 2.7, 2.8, 2.9, and 2.10. Figure 2.6 shows a heat map representation of the noisy spectra. Notice the ion intensity noise-induced striping of the spectra. Good normalization will need to remove that striping effect maximally. Figure 2.8 shows the result after reverse scaling with the real noise factors. It is clear that the striping is gone and the noise is removed. Figure 2.9 shows the results after using the classical TIC-based method. It shows good performance within content classes, but its normalization across different contents is incomplete. There is still definite striping between the spectra that only contain the base pattern and the ones that contain differential peaks as well. The heat map shows that the presence of non-common peaks increases the total ion count of the spectrum, and as a result overestimates the spectrum's ion yield from the sample. The ramification of this is an underestimation of the necessary noise-canceling factor. As a result, peak heights are too low compared to spectra that contain less differential peaks. Figure 2.10 shows the results from the IEC algorithm. It shows no striping and is visually indiscernible from the gold standard pattern. Note that in most of

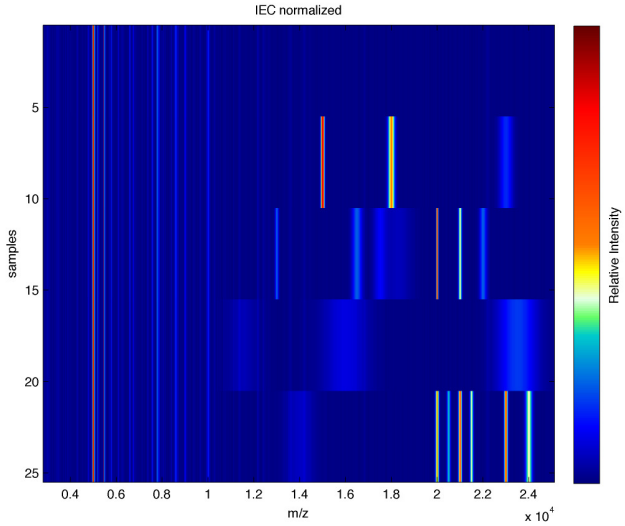


Figure 2.10. Heat map of the IEC normalized spectra. Notice the absence of any striping effect, even when comparing spectra with many differential peaks to spectra with little or none. The relative scaling by IEC restores the pre-noise data set well, and the heat map is visually indiscernible from the gold standard heat map. Note that normalization does not necessarily restore the absolute peak height. It is only concerned with relative peak height changes.

these methods the absolute peak heights are never restored exactly. They only make spectra comparable at a relative level. The heat map illustrates that by using a rank-1 approximation of the spectra, IEC is able to avoid bias from differential peaks.

Figure 2.11 provides a closer look at some of the normalization results, by focusing on a zoomed section of the mass spectrum of sample 6. It shows the gold standard and both reconstructions for the sixth mass spectrum (with their intensities scaled between zero and one to enable direct comparison). The IEC traces the profile of the gold standard closely, while the TIC approach clearly underestimates the true peak heights. The excellent matchup between IEC and the gold standard keeps the gold standard, indicated in blue, largely hidden behind the red line of the IEC result. Only at the very tops of the peaks do the blue tips of the gold standard improve over the IEC result.

2.6.3. Comparison of normalization performance with additive noise

To demonstrate the robustness of the approaches, the experiment is repeated with additive noise. First, the same multiplicative ion intensity noise from the previous run is added. Then, Gaussian additive noise is put on top of all m/z bins with a standard deviation equal to one percent of the noisy data set's intensity range (s.d. of approx.

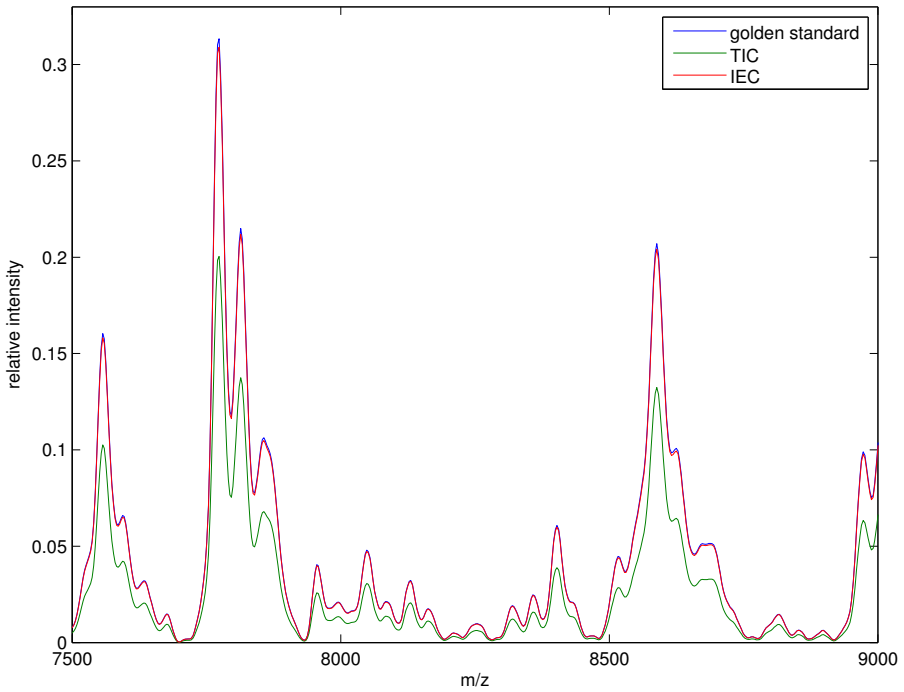


Figure 2.11. A zoomed-in look at the normalization of the mass spectrum from sample 6. The TIC-based normalization (green) clearly underestimates, while the IEC (red) is almost perfectly matched with the gold standard (blue). Only at the very tips does the gold standard become really visible. (The gold standard in scaling factors is the inverse of the noise scaling factors employed to generate the ion intensity noise in this data set.)

120 ion counts). The results are shown in Fig. 2.12 (noisy version), Fig. 2.13 (TIC normalized), and Fig. 2.14. Again the heat maps show that IEC outperforms TIC, even with a significant amount of additional noise added on top of the normalization problem.

Whether the difference between IEC and TIC-based methods is significant, depends on the study at hand. The answer is tied to elements such as the available instrument, the sensitivity required to prove or disprove a hypothesis, and most importantly whether or not the samples in question are heterogeneous in content. The less heterogeneous, the more overlap between the spectra and the better TIC will perform. In general however, IEC will outperform TIC-based normalization in most cases because it takes the 'common versus differential' issue out of the hands of the experimental design, and provides *in silico* means of compensating for whatever form the measurements may take. This is an important asset as more and more studies are collecting ever larger amounts of measurements, cross-comparing spectra from a very wide va-

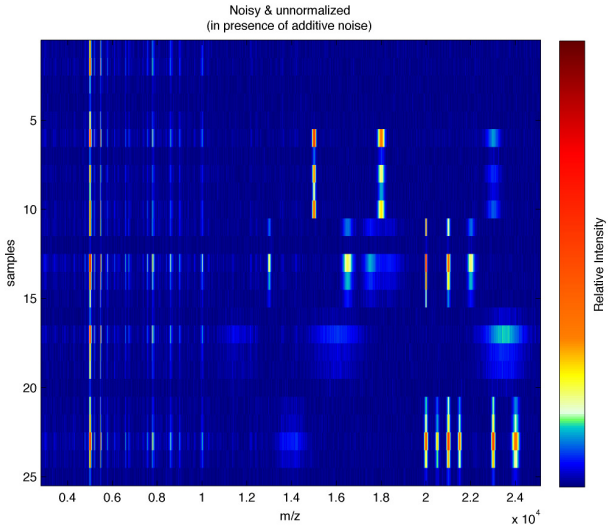


Figure 2.12. Heat map of the spectra with ion intensity noise in the presence of Gaussian additive noise.

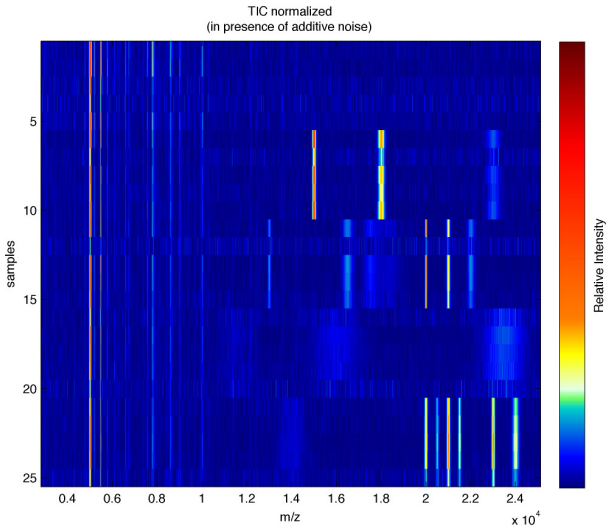


Figure 2.13. Heat map of the TIC normalized spectra in the presence of Gaussian additive noise. Notice the bad scaling of sample 12. This is the result of low ion intensity values being swamped by the additive noise. The TIC-based method is not able to discern peaks produced by the additive noise from real common low abundance peaks. The additive noise peaks are in the same relative intensity range as the real ion peaks, and contribute almost as much to the TIC as real peaks. As a result the amount of TIC is overestimated (a large part of it being noise), which leads to an underscaling of the spectrum.

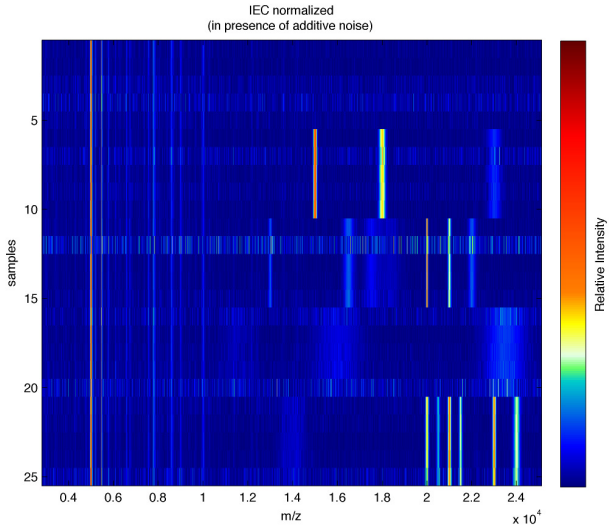


Figure 2.14. Heat map of the IEC normalized spectra in the presence of Gaussian additive noise. IEC does a better job than the TIC-based method of bringing the spectrum of sample 12 up to comparable peak heights. Although noise peaks are scaled up as well, it is preferable to have at least the real and common peaks at their correct height. Noise peaks that are scaled upward can always be removed from the analysis later, by removing peaks that only appear in a single sample.

riety of biological origins.

2.7. Case study 2: mass spectral imaging

Section 2.2 introduced mass spectral imaging as an area of mass spectrometry that is particularly sensitive to ion intensity noise. In the case of MALDI based MSI, one reason for this sensitivity is the matrix crystallization required by the ionization method. Also the fact that analytes are measured in situ without first separating the molecules from the surrounding tissue often plays an important role. The influence of the multiplicative ion intensity noise becomes readily visible in MSI. The most common use of MSI technology is in fact to produce ion images that show peak height across an entire organic tissue section, making the comparability of peak height from one pixel to the next crucial. Additionally, as computational MSI analysis develops further, the influence of reliable peak heights will become even more important than is currently the case with ion images. For this reason, this second case study centers on MSI and takes a closer look at IEC performance on a MSI experiment on mouse brain. A benefit of a MSI case study is that it enables intuitive assessment of the normalization results through pictures. Unlike the synthetic case study, real-world experiments are a noisy business and do not provide a gold standard against which to grade the performance of the algorithms. However, a MSI experiment is one of the few experiment types that provides the opportunity to see in the spatial domain whether the normalization results make biological sense. For example, if an anatomical region is more homogeneously filled or outlined, there is a high probability that the algorithm succeeded in extracting more useful information from the measurements, and thus exhibits better performance. This case study therefore takes a closer look at ion images to assess TIC versus IEC performance.

The MSI experiment is performed on a sagittal section of mouse brain, using a MALDI mass spectrometer. Further details on the materials and methods of the experiment are provided in section 5.3.1. The data set consists of 1734 mass spectra collected from the mouse brain section in a rectangular grid of 51×34 pixels. Each mass spectrum captures 6490 m/z bins spanning a range from m/z 2800 to 25000. As the rectangular grid has to circumscribe the entire tissue section, certain mass spectra stem from outside the tissue area. These are removed from the case study to avoid introducing non-tissue derived variation into the analysis. After retaining only the on-tissue spectra, the data set consists of 1381 mass spectra that make up a data matrix of 1381×6490 ion count values. The baseline is removed from the analysis to avoid it being an influential factor in the assessment of normalization performance.

Similar in approach to the synthetic case study, we first collect a heat map representation of the 1381 spectra in their un-normalized form in Figure 2.15. The figure shows

only the highest peaks clearly in some of the rows. The other peaks largely disappear at the lower ion count values. Then, TIC-based normalization is applied to the data set, resulting in the spectra of Figure 2.16. Compared to Fig. 2.15, the heat map clearly demonstrates that normalization is a worthwhile endeavor in MSI. TIC-based normalization succeeds in pulling several new peaks from the measurements. A good sign for the reliability of these peaks is that they show up consistently across different spectra, appearing as vertical lines in the heat map. Unlike in the synthetic data set where horizontal striping in the heat map was used as a clue to point out incomplete normalization, the striping effects in these heat maps have a different cause. They have little to do with normalization, but are the result of pushing measurements that are acquired from a rectangular grid into a list format. The 'breaks' in the vertical lines therefore usually occur at intervals equal to the width of the measurement grid (e.g. roughly every 51 or 34 spectra). Finally, the ionization efficiency correction algorithm is applied. The results, shown in Figure 2.17, clearly show that IEC is capable of extracting even more consistent peaks from the spectra than the TIC-based method could. The lower mass range shows much richer variation, and IEC seems particularly successful in pulling low intensity peaks over the noise threshold. This property of IEC is very interesting as sensitivity is a big topic of concern in MSI. Overall, the observations from the heat maps confirm the conclusions from the synthetic case study: any form of normalization is better than none, but the new IEC method does remarkably better than the standard TIC-based approach.

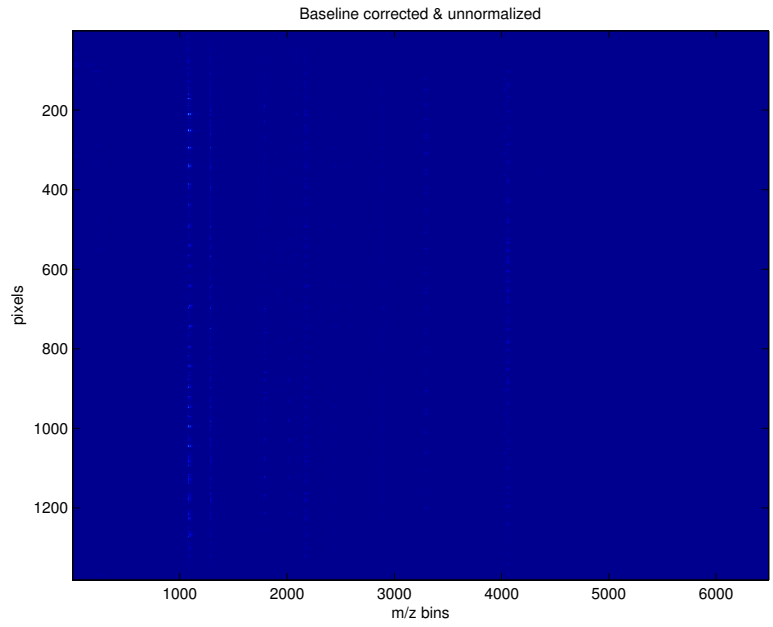


Figure 2.15. Heat map of the un-normalized spectra from the MSI experiment.

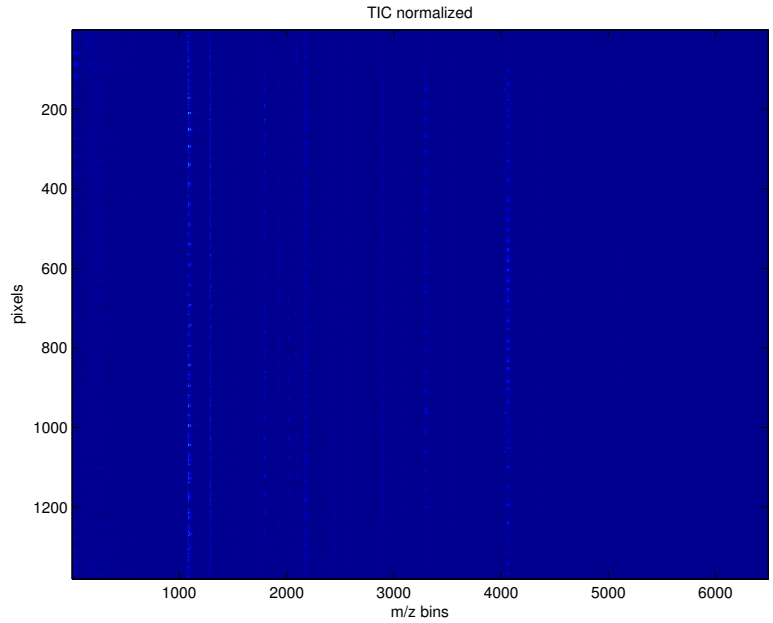


Figure 2.16. Heat map of the TIC normalized spectra from the MSI experiment. Notice the increased amount of common peaks pulled from the data.

However, to truly assess the value of IEC for MSI experiments, we need to take a look at the ion images produced from these data sets. Based on the heat maps from Fig. 2.15, 2.16, and 2.17, three ion peaks are selected for comparison in the spatial domain. These ions are m/z 4977, 12181, and 18416, and ion images stemming from the un-normalized, the TIC normalized, and the IEC normalized data set are extracted for each of them. The results are shown in grid form in Fig. 2.18. To ease biological interpretation they are shown again in Fig. 2.19, transparently overlaid on a microscopic image of the mouse brain tissue. Again, the need for ion intensity normalization in MSI is clearly demonstrated. The un-normalized ion images only make anatomical sense for the tallest of peaks. Ion m/z 12181 shows little or no structure in the raw form, while as soon as some form of normalization is applied it clearly shows increased presence in the cerebellar nucleus on the right hand side, and a marked absence from the central hippocampal area. Similar observations can be made for ion m/z 4977 and 18416. The differences between the TIC-based method and IEC are more nuanced. However, they do become abundantly clear when the anatomical background is taken into account. Figure 2.19 highlights some of the differences with arrows. The general observation seems to be that after performing IEC, the anatomical regions are more homogeneously filled and their outlines more clearly traced. For m/z 4977, this means that its presence is more widely confirmed throughout the upper and lower hippocampal area. For m/z 12181, its absence from the hippocampus is

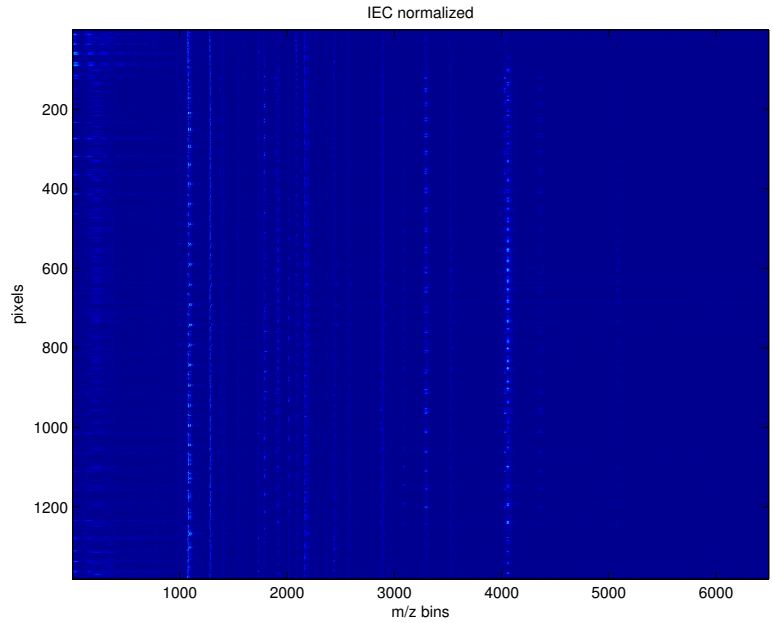


Figure 2.17. Heat map of the IEC normalized spectra from the MSI experiment. IEC pulls more consistent peaks from the data than the TIC-based method could.

more strongly emphasized (also versus the un-normalized ion image), and its presence in the cerebellar nucleus is more evenly spread. The same is also true for m/z 18416, which shows up in the central white ventricle area and in the elongated corpus callosum that touches it at the top.

2.8. Conclusion

This chapter introduced a novel normalization method for use in standard mass spectrometry as well as mass spectral imaging. The ionization efficiency correction method comes closer to the goal of using ionization efficiency for normalization purposes than the current industry standard based on total ion current. The reason for its improved performance lies in its ability to discern in the mass spectra common peak patterns from differential peak patterns, and to adjust its scaling factors accordingly. IEC does this by fusing the chemistry and physics considerations towards normalization with the mathematical concepts of matrix decomposition. It is in this unique fusion of approaches that the novelty of the method lays. IEC provides both a general framework for normalization and a concrete implementation of that framework using non-negative matrix factorization. Further development will particularly focus

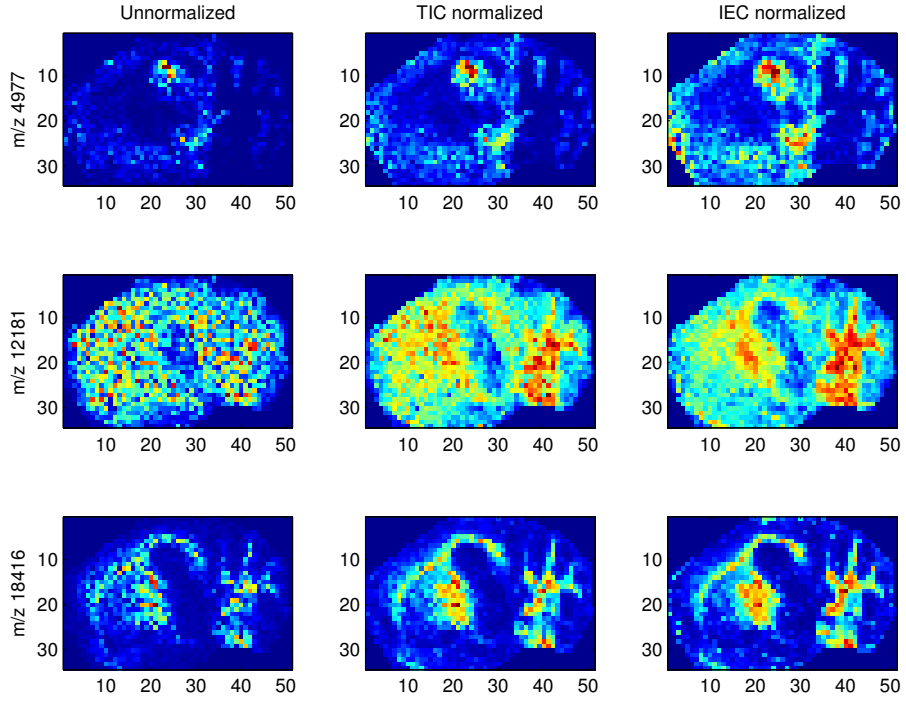


Figure 2.18. Comparison of normalization results for three separate ion images. The ion images of three different ions, m/z 4977, 12181, and 18416 are shown in three situations: unnormalized, TIC normalized, and IEC normalized. Notice that IEC succeeds in extracting more biologically relevant structure from the data set than the TIC-based method. A version of these images overlaid on a microscopic image of the tissue is available in Fig. 2.19.

on improving the rank-1 approximation engine of the IEC framework, and to reduce its computational requirements.

The results from both case studies confirm the thinking behind the approach, and demonstrate the applicability of IEC for various forms of mass spectrometry. However, as the algorithm is not particularly customized towards an ionization method or spectrometer configuration, it can equally well be applied in related fields such as secondary ion mass spectrometry or electrospray ionization measurements. In fact, its approach is presumed to be valid in any area or technology that has issues with multiplicative noise that for practical reasons cannot be transformed through for example log-transformation. This includes hyphenated techniques such as chromatography, as well as various forms of spectroscopy outside mass spectrometry.

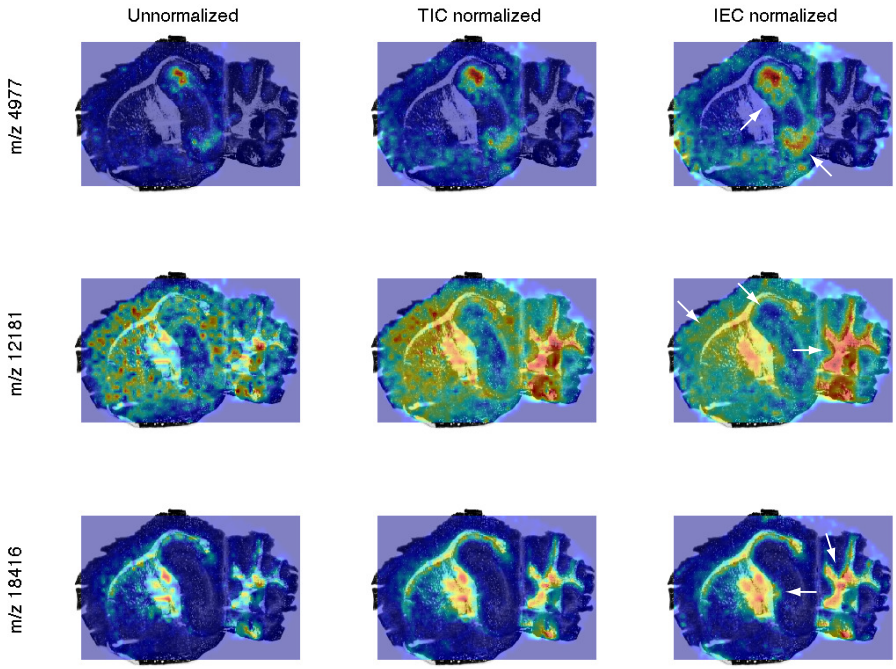


Figure 2.19. Comparison of normalization results for three separate ion images, overlayed on a microscopy image of the tissue section to aid biological interpretation. The ion images of three different ions, m/z 4977, 12181, and 18416, are shown for three sitations: un-normalized, TIC normalized, and IEC normalized. Particular areas where IEC outperforms the TIC-based method are highlighted with an arrow.

CHAPTER 3

Unmixing MSI Data

Essentially, all models are wrong, but some are useful.

George E. P. Box

This chapter is based on the following publications: • Van de Plas R. et al., "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis", Proc. of the Pacific Symposium on Biocomputing 12 (PSB), 2007, pp. 458-469; • Van de Plas R. et al., "Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA", Proc. of the Third IEEE/NIH BISTI Life Science Systems and Applications Workshop, 2007.

Organic tissue is a complex environment that consists of many different cell types and cell states. As a result, tissue is often composed of a wide range of molecular content signatures. In addition to the molecular complexity, there is also abundant variability in the spatial configuration of the different cell types or cell states. Some types form tightly grouped globules with crisp borders, while others scatter themselves between neighboring cell types in a gradient-like fashion or exhibit some other exotic dispersion pattern. Both the molecular as well as the spatial intricacies cause a big diversity in the spatial distribution of different molecular species throughout the tissue.

In addition to the biological effects that cause a particular molecule to show up at a certain location, there are also wet lab and instrumental reasons that influence spatial

distribution. When the tissue is cut for example, certain molecules might get smeared. After wet lab steps such as fixation or washing for example, treatment molecules might stay behind on the tissue. The mass spectral ionization process might produce additional fragment ions that co-localize with their mother molecules. Although these non-biological influences are kept to a minimum, they are never completely avoided and further perturb the spatio-biochemical complexity of the tissue.

In a MSI experiment, the molecular content of a tissue slice will be sampled at various locations across the section. Rarely will the MALDI laser sample only a single cell type or cell state under its laser diameter, and consequently the acquired mass spectra will usually report on a mixture of cell type contents. In fact, one might say that the mass spectra collected in a MSI experiment are the end result of a complex mixing process between various biological as well as non-biological effects that put molecules at certain locations in the tissue.

If we are to make sense of this complex mixture of ions and locations, we need methods that can unravel the MSI measurements and group together ions that share the same biological source or non-biological cause. Such methods should at least provide a more digestible set of patterns for further analysis, and are a prerequisite for truly leveraging the multivariate information encoded in a MSI experiment. This chapter aims to answer this challenge by providing several methods aimed at reversing the mixing process discussed above. If the spectra in a MSI experiment are a mixture of several spatio-biochemical effects, the goal is to unmix the spectra into a set of their underlying components.

Overall, the (unsupervised) unmixing of MSI spectra into a set of underlying components can be compared to what in signal processing is called *blind source separation*. Blind source separation is the separation of a set of signals or patterns from mixed versions of those patterns, and this with minimal external information [73]. The problem of source separation can be tackled through various approaches, but the general idea is that different source signals do not correlate with each other. Whether this assumption holds is dependent on the particular data type under consideration. In the MSI case however, the general premise of blind source separation translates into grouping together ions that show correlated spatial expression (co-localization) as well as correlated mass spectral expression (correlated or anti-correlated peaks across the different spectra). Figure 3.1 gives a general overview of the blind source separation goal in a MSI context. There are several different uses for the reduction of a MSI data tensor into a reduced set of patterns or trends:

- **Aid interpretation (trend detection)** – By disassembling the data tensor into a set of underlying patterns, the human interpretation of a MSI experiment becomes more straightforward. Ions that co-localize in particular areas of the tissue are grouped together into a component or trend, making it easier for a

human expert to learn which biomarkers are tied to a certain anatomical region or cell type.

- **Filtering** – The decomposition of the data tensor into a linear mixture of components provides the researcher with the opportunity to remove components that are irrelevant to the study at hand. If one is interested in a particular area of the tissue, or a particular subset of the measured ion peaks, it is possible to only select the components relevant to that area or those ions, and to remove the other components from further analysis. This operation essentially filters irrelevant variation from the data set, providing easier interpretation and more robust downstream analysis.

- **Dimensionality reduction** – By replacing a MSI data tensor with a lower rank approximation model obtained through blind source separation, a form of dimensionality reduction is achieved. This operation essentially transforms the data from its high-dimensional input space to a much lower-dimensional space defined by the components. This usually has favorable effects on the robustness of any follow-up with machine learning methods such as clustering and classification. A more thorough treatment of dimensionality reduction in MSI is provided in chapter 5.

- **Denoising** – Unraveling the data tensor into a sum of underlying rank-1 components allows a critical view into the contribution of each component to the overall data variance. This enables removal of components whose influence on the overall data set is minimal, or that exhibit a noisy spatial or mass spectral pattern. If the components that are modeling noise in the data, are removed before re-assembly into a data tensor, the resulting data tensor is essentially denoised. This is particularly useful for methods that employ absolute values in the original input space, such as ion images.

Depending on which aspects one wants to optimize during the unmixing process, different algorithms are available to implement the source separation:

- ***Principal Component Analysis (PCA)*** – Components are uncorrelated – section 3.1.

- ***Peak Intensity Weighted Principal Component Analysis (PIW-PCA)*** – Components are uncorrelated and less sensitive to outlier peak heights – section 3.2.

- ***Independent Component Analysis (ICA)*** – Components are statistically independent – section 3.3

- ***Non-negative Matrix Factorization (NMF)*** – Components are non-negative – section 3.4

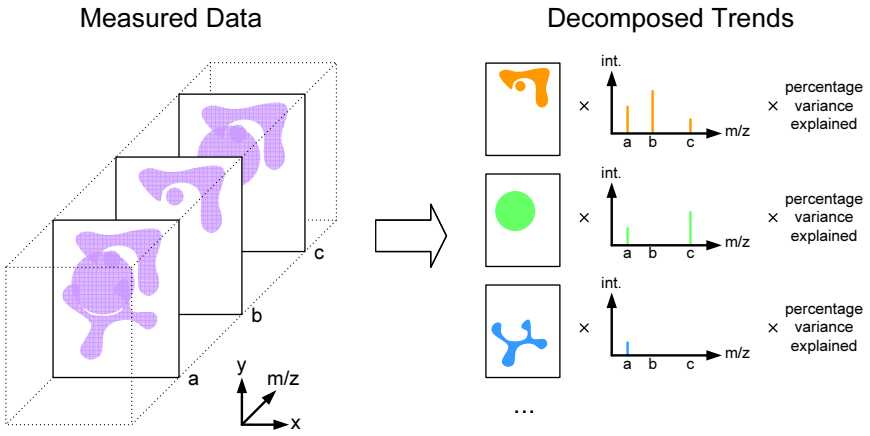


Figure 3.1. Unmixing of MSI data. The goal of the unmixing process is to decompose the original data set into a sum of components. Each component is characterized by a spatial distribution, highlighting a particular anatomical region or cell type in the tissue. Every component also has a mass spectral signature tied to it. The mass spectral signature shows which ions are specific to the area selected in the spatial distribution, and what their relative ion count expression is.

The most common member of this family is principal component analysis (PCA) which attempts to maximize the covariance captured by its components [67, 68]. Its application to MSI data is discussed in section 3.1 and in [74, 75]. Section 3.2 discusses a mass spectrometry-specific adaptation of the PCA algorithm that takes into account peak height ranges rather than absolute peak height intensities [76]. Section 3.3 briefly touches upon the use of independent component analysis (ICA) [69, 77] as a means of unmixing MSI data. The final section highlights non-negative matrix factorization (NMF) [72] as an unmixing algorithm particularly affine to MS and MSI data through its enforced non-negativity. The focus of this chapter will be primarily on aiding interpretation through trend detection. An example of the use of unmixing for filtering can be found in chapter 7. Examples of its use for dimensionality reduction are available in [78, 79, 80].

3.1. Principal Component Analysis

In this section we investigate the use of principal component analysis (PCA) as a guide for the prospective exploration of data coming out of an MSI experiment. The goal is to use the PCA results as a first stepping stone towards more elaborate multivariate analysis of MSI data. In section 3.1.1 we first discuss the general idea behind the PCA technique, followed by a treatise on the specific use of PCA in an MSI context

in section 3.1.2. In the case study discussed in section 3.1.3, we apply the technique to real data from a MSI experiment where a rat nerve tissue section was imaged.

3.1.1. Standard Linear Principal Component Analysis

Principal component analysis is a latent variable[1] data analysis technique, widely employed in many areas for uses such as dimensionality reduction [67, 68]. It was mentioned in a MALDI-MSI context by McCombie *et al.* [78] and Broersen *et al.* [74] for the purpose of dimensionality reduction and denoising. In this study the method is used for trend detection in both the mass and the image domain.

Before formulating a definition of PCA, it is necessary to explain some aspects of the concept of the rank of a matrix. The rank of a matrix M is the maximum number of linearly independent rows (or columns) of M. This means that the rank of M is the smallest number of outer products of vectors that can be used to reproduce the matrix M exactly. Another definition is that the rank of M is equal to the number of nonzero eigenvalues of $M^T M$. A matrix of rank 1 can therefore be completely represented as the outer product of two vectors, while a matrix of rank 5 requires the sum of (at least) five such outer products for it to be completely reconstructed.

PCA is a decomposition of a matrix X, of size $N \times K$ and of rank R, into matrices of rank 1, designated F_a:

$$X = \sum_{a=1}^{A} F_a. \tag{3.1}$$

Given the definition of rank, it is evident that the smallest value of A for which this equation still holds is equal to the rank R of the matrix X. The matrices F_a have the same size as X ($N \times K$), but as they are rank 1 they can be replaced by the outer product of two vectors s_a ($N \times 1$) and l_a ($K \times 1$) in equation 3.1:

$$X = \sum_{a=1}^{A} F_a = \sum_{a=1}^{A} s_a l_a^T = SL^T. \tag{3.2}$$

The vectors[2] s_a are generally called the score vectors of the decomposition, while vectors l_a are named loading vectors. Each pair of vectors s_a and l_a can be designated a principal component of matrix X and has a particular coefficient connected to it (stemming from the eigenvalue of the underlying $X^T X$), which indicates the relative amount of variance of X explained by its particular principal component. By further

[1] A latent variable is a variable which we do not observe directly, but its existence can be inferred from the observed variables.

[2] In this chapter we follow the convention of representing a vector as a column vector unless explicitly transposed.

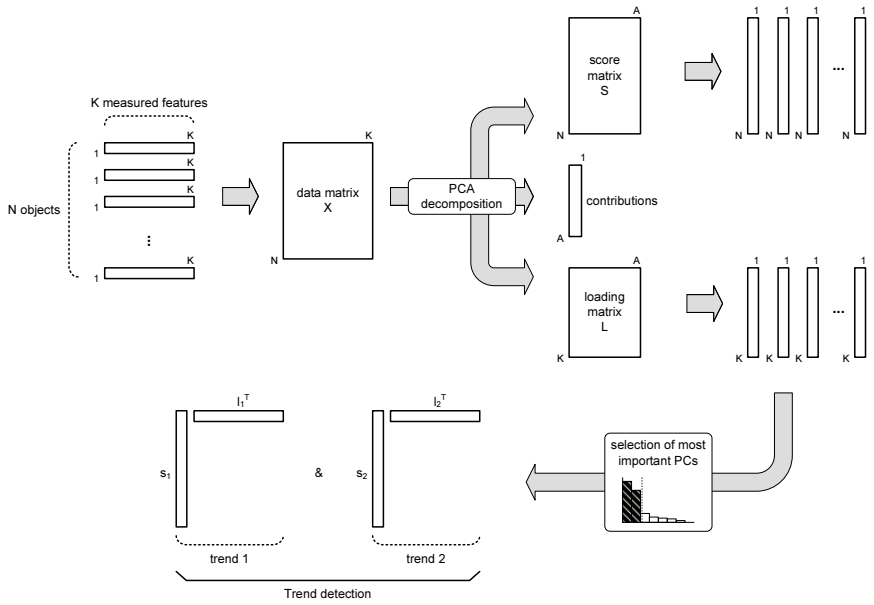


Figure 3.2. Graphical representation of matrix decomposition and reconstruction using principal component analysis. The upper half shows the decomposition of data matrix X into A principal components, or the outer product of score matrix S (grouping score vectors s_1 to s_A) and loading matrix L (grouping loading vectors l_1 to l_A). The lower half depicts selecting the principal components with the largest contribution of variance (or information) in order to study the uncorrelated trends found in X.

utilizing matrix notation, PCA can be written more consicely as the decomposition of the data matrix X into the single product of a matrix S of size $N \times A$, holding all the score vectors, and a matrix L of size $K \times A$, holding all the loading vectors. In order to complete the decomposition with a minimum value for A, the matrices F_a and their composing vectors s_a and l_a are necessarily maximally uncorrelated. A schematic overview of these steps is available in Fig. 3.2. Using the coefficients (actually eigenvalues) connected to them, one can order the various principal components according to their contribution in terms of variance of X explained and information contained. A number of uses follow from this notion such as dimensionality reduction and denoising [67], but we focus specifically on the use of trend detection, in which the principal components with the largest contribution characterize the major uncorrelated trends underlying the data.

One additional remark about PCA is that it has a unique solution up to a rotation. This means that sometimes more interesting and interpretable versions of the components can be obtained by applying a factor rotation criterion such as the varimax rotation. This however changes little for further computational analysis, but it is usually pre-

ferred if human interpretation is the end goal.

3.1.2. PCA Applied to MSI Data

In applying PCA to MSI data, our primary goal is to identify the major uncorrelated trends that can be found in the spatial domain (x and y) as well as in the mass domain (m/z). These trends, which tell us which pixels or m/z bins behave similarly or dissimilarly, can be used as a guide in exploring these often complex and very large data sets, and to avoid the proverbial 'drowning in information' that can be experienced when classical ion images are employed without a prior hypothesis of target mass.

As mentioned in introductory chapter 1, the data measured during an MSI experiment can be stored as an array of order 3, or tensor, $\overline{D}$ with two spatial dimensions (x and y) and one m/z dimension (cfr. the abscissa in a mass spectrum). Each scalar value d_{ijk} in the tensor represents the absolute intensity of a particular mass peak at a certain x-position i, a certain y-position j, and measured at a certain m/z bin k (with $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $k = 1,\ldots,K$).

One way of applying PCA to an MSI tensor $\overline{D}$ is to refold the tensor into an array of order 2, or matrix, D to fit the expression shown in equation 3.2. This refolding process is done by reordering all discrete spatial positions in the x- and y-dimensions, or 'pixels' if you will, into one long vector holding $I.J$ elements. The result is a matrix D of size $(I.J) \times K$, holding all information contained within the original tensor $\overline{D}$. Applying PCA in the manner discussed in section 3.1.1 delivers a score matrix S of size $(I.J) \times A$, a loading matrix L of size $K \times A$, and a vector of eigenvalues λ indicating each principal component's variance contribution.

Based on the amount of variance explained, we can now identify and take a closer look at the most important principal components. A single principal component is characterized by one score vector and one loading vector. The score vector is of size $(I.J) \times 1$ and does not easily allow for direct exploration. However, a reordering operation that reverses the effect of the operation performed to go from $\overline{D}$ to D allows us to refold this vector of size $(I.J) \times 1$ to the image space defined by the two spatial dimensions x and y, resulting in an image matrix of size $I \times J$. In their image form the score vectors deliver a more human-interpretable view on the underlying spatial correlations. This type of images gives us an idea of which pixels, or laser spots, have a similar mass spectral footprint when all m/z bins are taken into account (note the difference with univariate ion images). The corresponding loading vector of size $K \times 1$ does not require a refolding operation as it can be expressed directly in the m/z-domain. A visualisation of the loading vector gives us an indication of which m/z bins behave similarly within the context of one principal component.

It is necessary to mention here that the above linkup of score vectors with the image domain and loading vectors with the mass domain is based on the assumption

that PCA is performed on a data matrix where the rows represent pixels and the columns represent m/z bins. This assumption would be in line with the convention of an objects $\times$ features data matrix used in most PCA literature. However, when there are more features available than objects, which is usually the case for MSI (e.g. 7451 versus 1302 in the spinal cord data set), it is more computationally efficient to use the transpose of D instead [67]. The results of this more 'economic' PCA are identical to the ones from the procedure described earlier, with the only difference being that the loading vectors are now linked to the image domain and the score vectors to the mass domain. This economized PCA was used in the case study of section 3.1.3.

3.1.3. Case Study: Rat Spinal Cord Nerve Tissue

In this section we demonstrate the use of PCA in a MSI context by applying it to the MSI measurement of rat nerve tissue. For reference, Fig. 3.3 shows a microscopic image of a nerve tissue section taken from the same animal as the one used in this case study. Figure 3.3's tissue slice has undergone histological staining to bring out the gray/white tissue differentiation which is not visually apparent in untreated samples, but which does show up in the PCA analysis performed in this section.

Materials and methods

The tissue section (15 µm thick) was taken from a transversal section of the spinal cord of a standard control rat. The recorded mass range extended from m/z 5000 to 12000 and alpha-cyano-4-hydroxy cinnamic acid (7 mg/ml, in acetonitrile 50%, 0.05% TFA) was used as a chemical matrix. A MALDI mass spectral measurement was performed on each grid point of a virtual raster of size 31×42 that was superimposed on the tissue section with an interspot distance of 100 µm in both the x and y-directions. The mass spectrometer that was used is the ABI 4800 MALDI TOF/TOF Analyzer from Applied Biosystems Inc in linear mode. The data collection in the mass spectrometer was guided by the *4000 Series Imaging* module, available at `http://www.maldi-msi.org`. Processing was done using in-house developed software.

Preprocessing

As Fig. 3.4(a) to 3.4(f) show, the MSI raster was slightly off center with regards to the tissue section, resulting in a tissue-free area in the bottom right corner (shown in purple). To avoid these empty measurements consuming variance and influencing the PCA-results, we disregarded them when their total ion current fell below a 10% threshold.

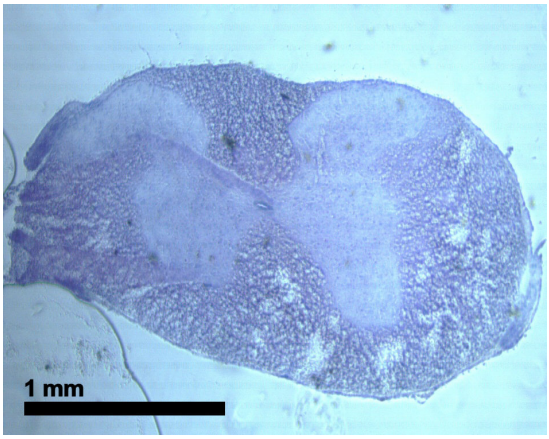
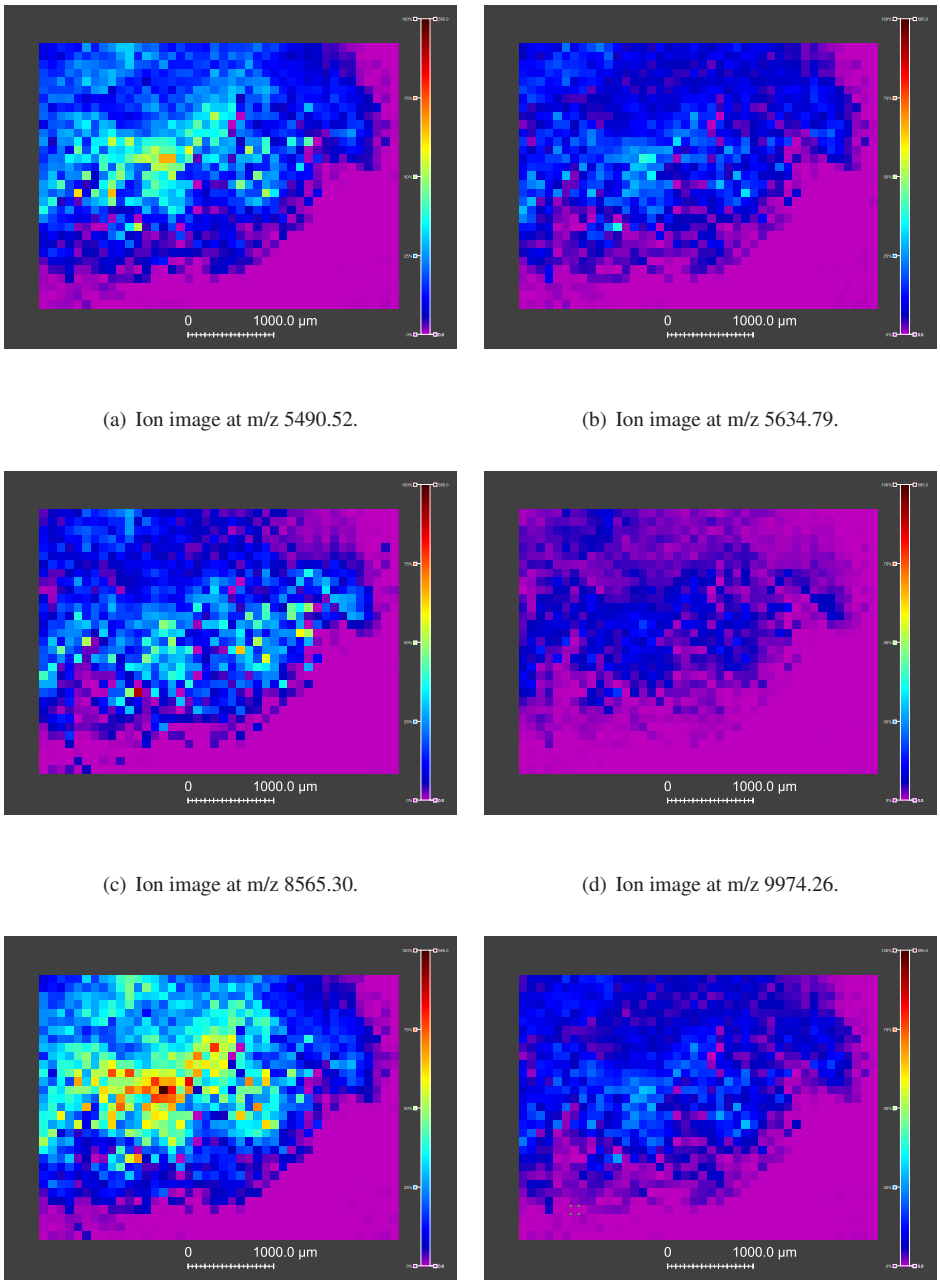


Figure 3.3. Microscopic image of a transversal section of rat spinal cord, histologically stained to show the butterfly-shaped central area known as the *substantia grisea* (grey matter), surrounded by white matter nerve tissue.

Analysis Results

We applied PCA via singular value decomposition of the covariance matrix of the data matrix, using the economized version of PCA mentioned in section 3.1.2. This means that the loading vectors are represented in the image domain, while the score vectors are shown in the mass domain. When interpreting these visualisations, the relative differences in value are important, not the absolute value or sign (e.g. see first score vector in Fig. 3.7). In the image domain (Fig. 3.6) low valued areas in blue are discriminated from high valued pixels in red, indicating zones within which the mass spectral footprint (and the underlying chemical composition) is similar. In the mass domain (Fig.3.7) m/z bins carrying similar values correlate strongly in behavior across the tissue (the peaks vary together), and can be discriminated from bins with dissimilar values.

The bar plot in Fig. 3.5 shows us the relative amount of information contained in each principal component (PC) (above a 0.1% cut-off). It is apparent that the first PC is very prominent with more than 90% variance explained. This means that the spatial and mass-related correlations connected to the first PC can be considered as the primary structure found in the chemical composition of the tissue slice. Secondary and tertiary uncorrelated trends are also apparent as the second and third PC still hold a non-negligible amount of information. However, from the fourth PC onwards the contributions become less influential, tending towards noise in the data. Therefore, we will focus on the first three loading and score vectors shown in Fig. 3.6 and 3.7. The strong reduction in complexity indicates a large amount of correlation in the spatial domain (indicating region formation) and the mass domain (indicating ions, or m/z



(a) Ion image at m/z 5490.52.

(b) Ion image at m/z 5634.79.

(c) Ion image at m/z 8565.30.

(d) Ion image at m/z 9974.26.

(e) Ion image at m/z 5484.29.

(f) Ion image at m/z 6275.55.

Figure 3.4. Example ion images from the rat spinal cord MSI experiment detailed in section 3.1.3.

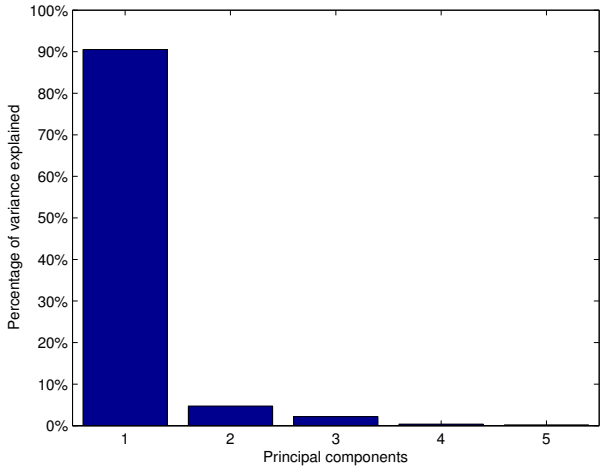


Figure 3.5. Percentage of variance explained by each principal component (PC). This graph only shows PCs with a contribution larger than 0.1%. The total number of PCs is actually given by $\min(N, K)$, but most have a negligible or zero contribution.

bins, behaving similarly; e.g. by coming from the same parent molecule).

The primary trend, characterized by the first loading vector in Fig. 3.6 and the first score vector in Fig. 3.7, shows that a butterfly-shaped region in the center of the tissue has a dissimilar chemical composition from the areas surrounding it (blue vs. red). This area correlates strongly in location and shape with the anatomical region called the *substantia grisea* (grey matter), surrounded by white matter nerve tissue (also visible in Fig. 3.3). The first score vector shows that all m/z bins have negative values with differing relative amounts, indicating that the spatial discrimination between the grey and white matter areas is mainly explained through quantitative differences in the chemistry, rather than qualitative ions showing up or disappearing. Also notice the two characteristic peaks at m/z 5484 and 8564, that show up consistently across the nerve tissue but whose relative quantity can be employed as a mass marker for grey matter.

The second trend differentiates the blue region of tissue at the top of the raster from the red/yellow area at the center. When studying the second score vector it becomes clear that the differences between these areas are mainly caused by the dense peak area between m/z 5000 and m/z 8000 showing up more prominently while the peaks at m/z 5484 and 8564 lose intensity. One has to bear in mind that this secondary trend only accounts for some 4% of the chemical variation across the slice.

With a contribution coefficient of 1 to 2%, the third loading vector differentiates between a ventral and dorsal area in the tissue. This third trend correlates strongly with

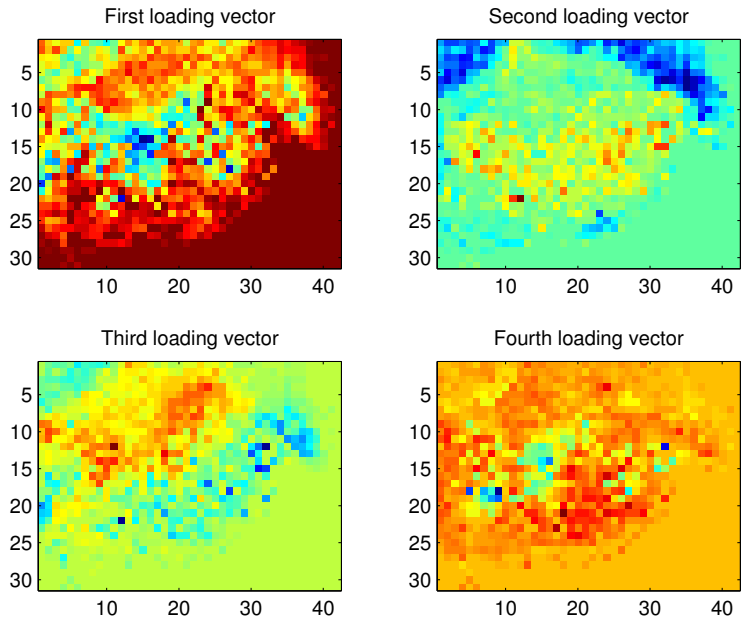


Figure 3.6. The first four loading vectors. Folded back into the image domain, they show correlations at the pixel level.

the biochemical differences in ventral/dorsal cellular composition of the spinal cord. In the mass domain of the third score vector we see that this ventral/dorsal difference is mainly due to the 5484-peak showing up and the 8564-peak diminishing in one area while the reverse happens in the other area. In the primary and secondary trends, the differences were mainly quantitative in nature. However, in this third trend we see an example of a more qualitative difference with the presence of a particular ion characterizing a particular area in the tissue. The fourth loading and score vectors are shown for completeness, but it is evident that spatially the correlations are less localized and structured, tending towards spread-out noise. In the mass domain we see differentiation between m/z bins which are close together. This is rather unlikely to be structured given that there are isotopic and other ties between m/z values this close together, which further seems to indicate that from this trend onwards we are dealing with modeled noise.

In summary, the PCA-results tell us that in this particular MSI data set the chemical composition is dominated by the difference between grey matter nerve tissue and white matter, and two quantitative ion markers for these areas are observed at m/z 5484 and 8564. In addition to that, a ventral/dorsal difference was measured which

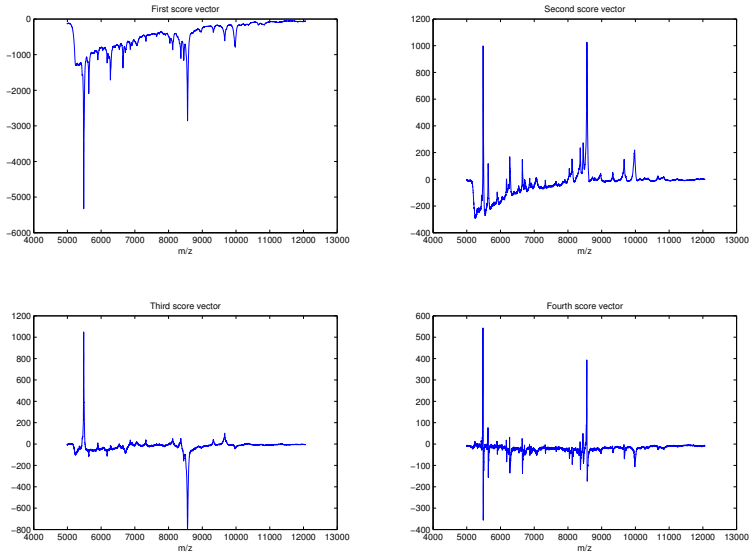


Figure 3.7. The first four score vectors. Represented in the mass domain, they show correlations between m/z bins.

can be related to known ventral/dorsal differences in the spinal cord.

3.1.4. PCA Summary

This section described a procedure for using PCA in an MSI context as an intrument for guiding prospective analysis of the chemical tissue composition. In the spatial domain it can show the human observer which regions have a particular mass spectral footprint, and it can differentiate these from other areas in the tissue slice without the need to perform invasive chemistry on the sample as is the case with e.g. histological staining. In the mass domain, specific molecular masses responsible for these differences (in m/z form) are identified, and lend themselves for further downstream analysis using, for example, ion images. The case study on rat nerve tissue demonstrated these uses by delineating grey matter from white matter and by identifying two mass markers that can be used to differentiate between these zones. It also illustrates how, in addition to the visual aspect of differentiating zones in the tissue, MSI as a technology permits a direct measurement of the chemical reality responsible for these differing areas, in the form of molecular masses.

3.2. Peak Intensity Weighted Principal Component Analysis

This section elaborates on the principal component analysis (PCA)-based unsupervised decomposition of an MSI-measured organic tissue section into its underlying biochemical trends. We introduce a method to control the weight that particular peak intensity ranges are allowed to exert on the final decomposition model. The extension provides a way for peak intensity-based scaling to be incorporated directly into the decomposition process, for the purpose of denoising or contrast enhancement. Different peak height ranges can be given different weights and influence in the final trends enabling, for example, the attenuation of noise signals characterized by low ion counts. The method makes use of peak height transformations that are conceptually equivalent to what is known in digital image processing as gray level transformations, but rather than aiming to enhance contrast for human interpretation they are used to influence the unsupervised decomposition process. As an example, we apply a combined denoising/contrast stretching measure to the same MSI measurement of rat spinal cord used in section 3.1.

3.2.1. Standard PCA – Implicit Variable Weights

PCA is usually implemented as a singular value decomposition (SVD) of a square matrix, which represents the relationships between the variables in the data matrix D (M observations $\times$ N variables). Commonly this is the covariance matrix, but in some cases the correlation matrix is used as well. The difference lies in the relative influence that every variable is allowed to exert on the final principal component model. In the covariance matrix C each element c_{pq} represents the covariance between variables p and q (see 3.3). Calculating the covariance involves subtraction of the mean per variable (mean-centering), but absolute deviations from the mean are left intact. As a result variables with large observations give large entries in the covariance matrix, and thus, exert more influence on the components.

$$c_{pq} = \frac{1}{M-1} \sum_{m=1}^{M} (d_{mp} - \overline{d}_p)(d_{mq} - \overline{d}_q) \tag{3.3}$$

In the correlation matrix R, every element r_{pq} is further standardized by scaling all entries to a range between -1 and 1, dividing the covariance by the standard deviations of the variables p and q. The scaling has the effect of equalizing the weight that each variable contributes to the decomposition. Variations over small absolute values will become equally influential as large absolute variations. This makes sense when the data consists of heterogenous variables whose values do not have a direct physical relationship to each other. However, in the case of mass spectrometry the variables do have a physical relationship and are directly comparable in that they are

all ion counts measured by the ion detector of the same mass spectrometer, only with a different m/z parameter for the mass analyzer. Therefore, it makes physical sense to assign large peaks more importance than smaller peaks, making the covariance matrix the preferred basis for decomposition (and the extension of section 3.2.2).

Case Study Results

Applying standard PCA based on the covariance matrix to the data set of section 3.1.3 results in a compact set of principal components (PCs). A full treatment of the results is given in section 3.1.3, but for comparison with the peak intensity weighted PCA discussed in section 3.2.2, we show two of the most interesting trends in Fig. 3.10 and 3.11. Fig. 3.10 (top) shows the first PC responsible for 90% of the variance in the data, which delineates the butterfly-shaped region of gray matter in the center from the chemically different surrounding white matter (see also Fig. 3.3). Mass spectrally, the differences can be primarily contributed to the relative presence of m/z 5484 and 8564 that function as a marker for these regions. The second PC in Fig. 3.11 (top) differentiates the blue region at the top from the central area. The mass signature tell us that the difference between these areas is primarily the result of a consistently higher baseline showing up in the blue area, indicating that ionization took place more easily there.

3.2.2. Peak Intensity Weighted PCA based Decomposition

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

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

Ion Count Transformations

An ion count histogram describes the distribution of peak heights throughout a collection of mass spectra. The type of scaling we are looking for, also uses it as the

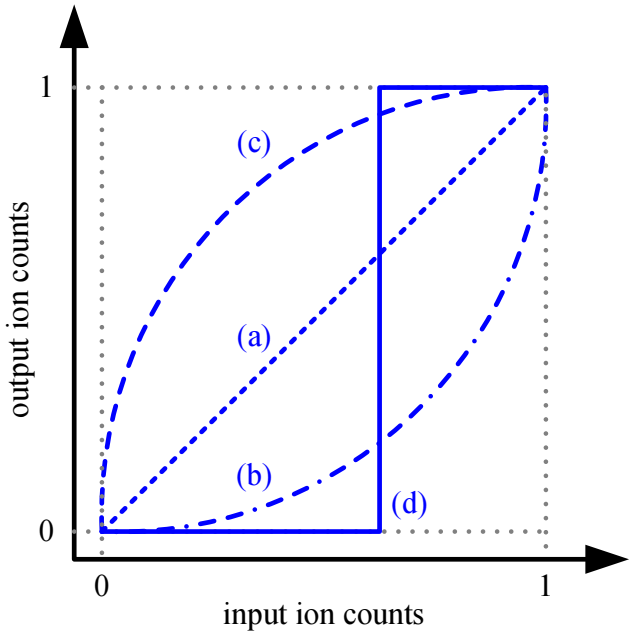


Figure 3.8. Examples of ion count (or gray level) transformations. (a) Identity transformation. (b) *n*th Power transformation. (c) *n*th Root transformation. (d) Threshold transformation.

input for a transformation function that translates the original ion count d_{ijk} in tensor $\overline{D}$ to a modified ion count d'_{ijk}. Examples of such histograms can be found in Fig. 3.9. Output ion count d'_{ijk} is related to input ion count d_{ijk} by a transformation T such that $d'_{ijk} = T(d_{ijk})$. T can take many different forms, depending on the goal in mind.

An interesting observation is that this type of histogram manipulation is conceptually equivalent to an image enhancement technique called gray level transformations [81], often employed in digital image processing. In a similar way as with the ion counts, pixel intensities (or gray levels) are mapped to different intensity values with the intent of increasing contrast. In digital image processing this contrast enhancement is focused on improving human interpretability towards a certain purpose. In an MSI setting the contrast modification serves to make certain types of responses more 'visible' to the PCA decomposition. A link between digital image processing and MSI is also mentioned in [82].

Figure 3.8 shows a number of example transformation curves. The most basic example of the family of linear transformations is the identity transformation shown as curve (a). It is a trivial transformation in that it leaves all ion counts the same. Another interesting family of curves is known as power-law transformations, which typically take the form of $d'_{ijk} = \alpha(d_{ijk})^{\beta}$ where α and β are constants. These include *n*th power

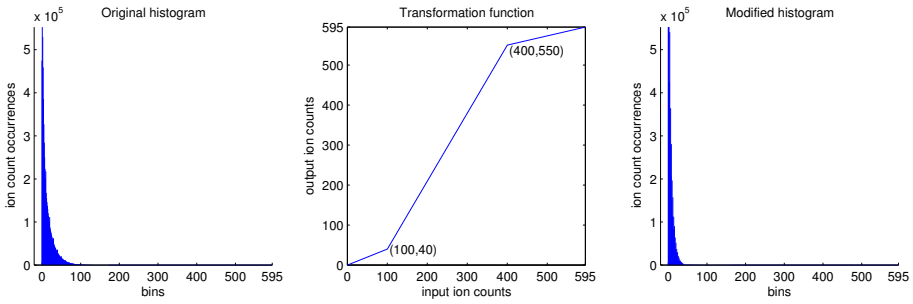


Figure 3.9. Overview of the contrast stretching transformation performed on the spinal cord data set. (left) The original histogram of all ion counts in the data set. Given the predominating presence of low ion counts, this graph does not show the distribution of measurements at higher ion counts. However, the entire histogram is populated up to the maximum response in this experiment at 595 ion counts. (middle) The gray level transformation that was applied. It pushes lower ion counts (< 100) together in narrow range close to zero, reducing their impact on the decomposition. Higher ion counts (> 400) are compressed into a narrow range close to the maximum response, equalizing their importance in the decomposition. (right) The modified histogram showing the effect (only visible in the low ion count range) of the contrast stretching procedure.

transformations such as curve (b), which map a wide range of lower ion counts to a narrower range, while higher ion count levels are allowed to expand into a wider range of values. The inverse happens when nth root transformations (see curve (c)), which are also part of the power-law family. With this type of curve, high ion count levels are compressed together into a narrower band of the available range, while differences between lower ion count levels spread out over a larger area of the histogram and become more prominently visible. Another type of transformation is the logarithmic transform, which demonstrates behavior similar to the nth root curves. T can also be a piecewise linear function, such as the thresholding curve (d). Another good example of this type is shown in Fig. 3.9 (middle), where it is applied to the case study of section 3.1.3. This type of mapping is known as a *contrast stretching* transformation. The lower ion counts are compressed into a narrow ion count range close to zero, while higher ion counts that cross a certain level are compressed together at the high end of the range, equalizing their influence. The ion counts in between are spread out over a wider range, thinning that area of the histogram. In conclusion one can state that the transformation function T can take any form as long as it is monotonically increasing (in order to avoid artificially created intensity artefacts).

Case Study Results

As an example, we apply a contrast stretching transformation to the data set of section 3.1.3 (Fig. 3.9). It acts as a soft thresholding filter, diminishing the influence

of lower (< 100) ion counts, while ion counts above a certain threshold (> 400) get grouped together and are given a more equal weight. The chosen parameters are the result of an empirical assessment of the input ion count histogram. The automatic assignment of such parameters is an avenue of further research. An evaluation of the influence of the imposed weights for each component is given in the captions of Fig. 3.10 and 3.11. Overall, the results demonstrate a particular use of the weighing scheme as a means of denoising mass spectral data, without losing too much of the original information. The parameterized simplification demonstrated by the mass spectral signatures is an interesting example of this.

3.2.3. PIW-PCA Summary

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

3.3. Independent Component Analysis

For the sake of completeness, we mention here independent component analysis (ICA) and its potential application to MSI data. ICA is another blind source separation method that has been gaining considerable popularity in recent years. The approach ICA [69] uses is comparable to that of PCA, but there are some remarkable differences.

3.3.1. ICA versus PCA

First of all, where PCA attempts to decompose the MSI data set into a set of uncorrelated components or trends, ICA strives towards statistically independent components. It does this by not only considering the second moment about the mean (variance),

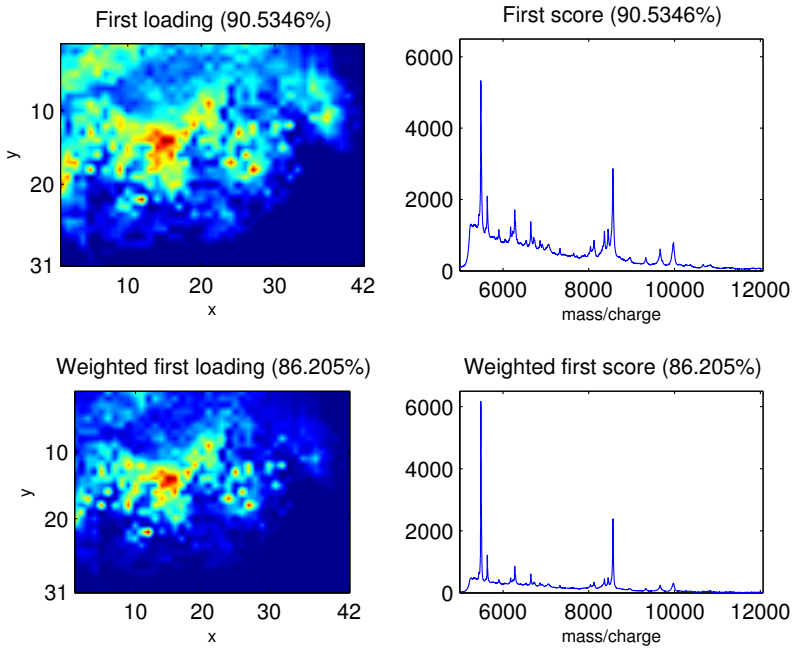


Figure 3.10. (top) Unweighted first PC. (bottom) Peak intensity weighted first PC. By weighing down the smaller peaks we get a cleaner delineation of the *Substantia grisea* and a considerable simplification of the mass spectral signature. Particularly the top area, specifically selected by PC2 in Fig. 3.11, is prevented from causing a *bleed-through* via its many low ion count values.

which is the sole criterion for PCA, but it additionally considers the third and fourth standardized moment about the mean as well (resp. skewness and kurtosis). By using these extra criteria in differentiating its components, ICA is able to describe and handle a much wider array of statistical distributions. PCA on the other hand, assumes the data to have a Gaussian distribution. There are many variants of the ICA approach (e.g. JADE, infomax, FastICA,...) but most are implemented as the iterative solving of an optimization problem. Calculation-wise such an approach cannot compete with a direct calculation as in PCA.

3.3.2. ICA Applied to MSI Data

The use of ICA in a MSI context has been modest up to this point. The author did a prospective study using the FastICA algorithm in 2007, and since that time the topic has been touched upon again by Hanselmann *et al.* [79]. The data is not shown here, but a general observation was that the ICA algorithm delivered more components and a more complex unmixing model. As a result, ICA is typically less well-suited than

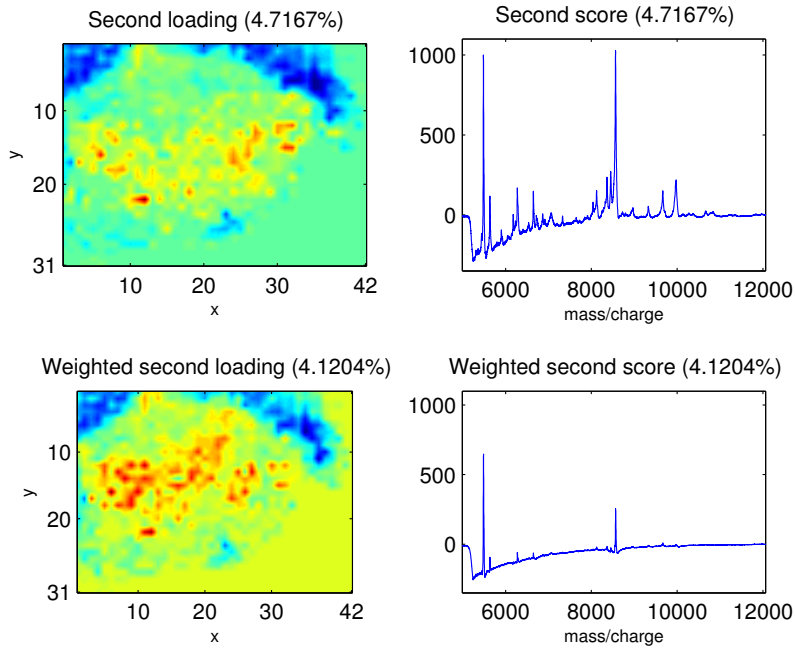


Figure 3.11. (top) Unweighted second PC. (bottom) Peak intensity weighted second PC. Most apparent is the strong simplification of the mass spectral signature, which in its reduced form, according to the spatial images, still delineates the same zone quite accurately with just two of its original peaks. The smaller peaks that were attenuated in the weighted version turn out to influence the central area of the spinal cord.

PCA for visual interpretation, and it can probably best serve in a preprocessing role before higher-level analysis. However, its heavier calculation requirements make it a difficult unmixing choice to defend, given the large data sizes in MSI.

3.3.3. ICA Summary

Due to its extra constraints, ICA delivers a more stringent decomposition than PCA. This manifests itself in MSI data sets by having additional trends branch off from what PCA still considers the same component. Generally this causes ICA to deliver more components than PCA to describe the same amount of data, but these components do have a stronger statistical foundation regarding their independence. Overall, one might consider ICA a more advanced unmixing tool than PCA, but due to its added computational cost and lesser affinity for visual interpretation, ICA will deliver most use in MSI experiments where the data starts to deviate significantly from PCA's assumptions.

3.4. Non-negative Matrix Factorization

In this section, we introduce non-negative matrix factorization (NMF) as a novel matrix decomposition method to perform the blind source separation task central to this chapter. Although NMF's statistical properties are less strong than those of methods such as PCA or ICA, its advantages have been empirically demonstrated. The method has been gaining considerable momentum in the signal processing community [72]. When applied to MSI data, NMF enables primarily easier and more straightforward interpretation.

Section 3.4.1 first discusses the difficulties tied to the interpretation of PCA results, and motivates the enforcement of a non-negativity constraint during the unmixing process. Then, section 3.4.2 gives a conceptual overview of the NMF algorithm, and the particular implementation employed here. Next, the differences between the traditional PCA method and the novel NMF method, are demonstrated on a synthetic data set that mimics the mixing process present in organic tissue. This synthetic data set has been artificially created in the wet lab, and thus provides a gold standard against which the performance of the algorithms can be weighed. Finally, section 3.4.4 gives a demonstration of the NMF algorithm on a real tissue experiment on mouse brain.

3.4.1. What is the problem with PCA? Why enforce non-negativity?

PCA and related factorization methods have proven their worth as computational preprocessing steps for more elaborate *in silico* analyses such as clustering or classification. However, their applicability for the purpose of enhancing human interpretation of MSI data, leaves something to be desired. One of the prime problems standing in the way of interpretation by mass spectrometry experts is the fact that the spectra are decomposed into signals that hardly resemble the mass spectral profiles the field is familiar with (although the decomposition may be mathematically correct). To be more precise, mass spectrometry is a field of ion *counts*, meaning that all collected data in standard mass spectrometry as well as mass spectral imaging experiments will inherently be either zero or positive. However, the non-negative nature of mass spectrometry data is usually not taken into account by the blind source separation algorithms, and as a result their output consists of signatures that contain both positive as well as negative values. Essentially, the mass spectra are not unmixed into a set of sub-spectra, each component grouping a subset of peaks together, as would be preferable. Instead the spectra are decomposed into complex signatures containing both peak additions as well as peak subtractions, which make little physical sense. Examples of such signatures are shown in Fig. 3.7. Although it is possible to interpret components such as these (as we did in section 3.1.3), it is a laborious process and requires intimate knowledge of the employed algorithm and its optimization goals. In

PCA for example, it is not uncommon to encounter a component that describes negative ion counts at a particular location. This means that at that location, ion counts for that particular peak need to be removed from the previous components to approximate the real data. The fact that these traditional decomposition models allow for deductions and cancellations of ion counts, causes the components to take a rather holistic turn and detracts from the intended visual separation of MSI data. It is clear that this cross-talk between components can lead to a very intricate mixing model of the original spectra. Although the inter-component relationships might indeed provide the most compact mathematical model for the data given the goals of the algorithm, they work counter-intuitive for human eyes as they prevent focus on a single component. Given these observations, two requests for a new decomposition method emerge. The first requirement is the ability to consider components completely separate from other components, and as such to effectively isolate a spatio-biochemical effect from other effects in the tissue. The second requirement is for the method to produce only non-negative source components to describe its approximation. This would force the spectral signatures of the components to describe a measured mass spectrum as a mixture of mass spectra-like patterns. Having the spectra decomposed into a set of pseudo-spectra further facilitates direct interpretation of the tissue effects in terms of ion presence and peak height.

3.4.2. Non-negative Matrix Factorization – The Concept

In response to the challenge proposed in the previous section, we adapted the algorithm of NMF to a MSI context. The NMF method steps away from the holistic component models established in approaches such as PCA and ICA, and instead provides a basis consisting of localized features along the mass domain as well as the spatial domain. Similar to the linear mixing models of PCA, PIW-PCA, and ICA, this method also decomposes a matrix X, of size $N \times K$ and with a certain rank, into a matrices of rank 1, designated F_a:

$$X = \sum_{a=1}^{A} F_a. \tag{3.4}$$

The matrices F_a have the same size as X ($N \times K$), but as they are rank 1 they can be replaced by the outer product of two vectors w_a ($N \times 1$) and h_a ($K \times 1$) in equation 3.4:

$$X = \sum_{a=1}^{A} F_a = \sum_{a=1}^{A} w_a h_a^T = WH^T. \tag{3.5}$$

The difference between NMF and other factorization methods comes down to the constraints that are put on the approximating matrices W and H, and the fact that one has to provide a rank r for the approximation beforehand. Where PCA requires

the columns of W to be orthonormal and the columns of H to be orthogonal to each other, NMF requires the values of W and H only to be non-negative. The rank-r approximation of matrix X thus becomes:

$$X \approx \sum_{a=1}^{r} w_a h_a^T = W' H'^T. \tag{3.6}$$

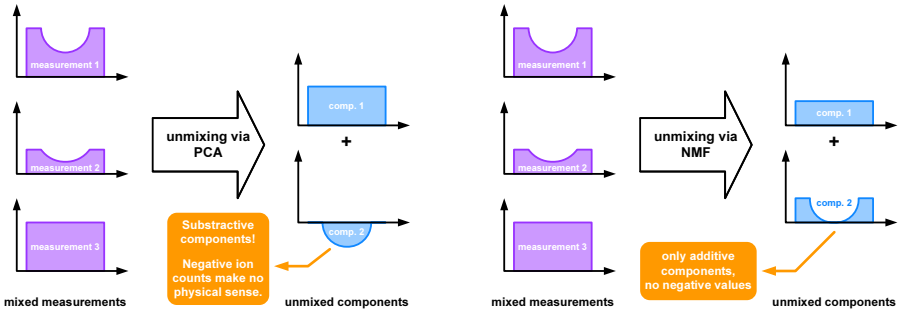
where W' is non-negative and of size $N \times r$ and H' is non-negative and of size $K \times r$. By only allowing zero or positive values in the component matrices W' and H', only additive combinations of components are allowed to approximate real measurements. Contrary to PCA, no cancellations or subtractions of ion counts are possible. The result of this additive-only model is that NMF tends to generate parts-based representations of the real measurements. In MSI terms this means that mass spectra are only approximated by summation of different pseudo-spectra carrying different peak patterns. The solution to finding the rank-r unmixing model that approximates MSI data matrix D the best, is to formulate the constraints into an optimization problem such as

$$\begin{aligned} &\text{minimize} && \|D - W'H'^T\|^2 && \\ &\text{where} && D \in \mathbb{R}_+^{N \times K} && \text{(mass spectral data set)} \\ & && W' \in \mathbb{R}_+^{N \times r} && \text{(spatial distributions of components)} \\ & && H' \in \mathbb{R}_+^{K \times r} && \text{(mass spectral signatures of components)} \end{aligned} \tag{3.7}$$

Figure 3.12 gives a simple example of the difference between PCA and NMF based decomposition. As motivated in Lee *et al.* [72], a parts-based representation lays closer to the human understanding of a situation and will usually provide easier insight into the content of the data.

3.4.3. Case study: MSI Experiment on a Synthetic Tissue Immitation – The ESAT Logo Gold Standard

In this section we will demonstrate the difference between classical PCA based decomposition of a MSI experiment versus NMF based decomposition. To truly assess the PCA versus NMF results, we need to be able to discern tissue induced variation from algorithm induced variation. In other words, a gold standard is required. However, when real organic tissue is imaged using MSI, we never know the complete ground truth regarding its molecular content. This means, that to acquire a gold standard of spatio-biochemical effects, we need to construct a synthetic data set ourselves.



(a) PCA based unmixing

(b) NMF based unmixing

Figure 3.12. Schematic comparison of PCA versus NMF based unmixing. Ion counts are strictly non-negative values. As a result, PCA's subtractive components (representing negative ion counts) make little physical sense. NMF, with purely additive components, allows for more straightforward part-based insight into the structure of the data.

To assemble a synthetic data set at the *in silico* level is a definite possibility, but it is very difficult to create an artificial set of mass spectra that exhibits all the quirks and characteristics inherent to mass spectrometry. The reason for this difficulty is that a mass spectrometer is a very complex instrument, and to model its behavior to a sufficient extent would require a great deal of additional knowledge and verification. Instead we opt to construct a synthetic data set at the wet lab level by putting known molecules in certain configurations on an ITO-coated glass slide, and using a real mass spectrometer instead. The idea is to mimic organic tissue with a simple spatial configuration of a limited, but known, set of molecules. The next step is then to treat the synthetic configuration as if it were true tissue, and measure the pattern as if it were a normal MSI experiment. As we know which molecular species was put where on the glass, we have essentially a map of the different effects present in the data. Having the ground truth available, enables immediate evaluation of whether an unmixing algorithm gets it wrong or right.

Materials and methods

In order to construct a synthetic data set as close to real tissue as possible, we need the capability to deposit chemical substances at scales common to real tissue experiments (e.g. 300 µm). Luckily, the effort to make MALDI matrix deposition and crystallization as homogeneous and repeatable as possible, has triggered substantial interest from industry into building chemical spotters. Notable examples include Shimadzu's chemical inkjet printer, the CHIP-1000, and Labcyte's Portrait 630 reagent

multi-spotter. Although these robotic instruments are primarily developed with the deposition of matrix molecules in mind, they can serve equally well as methods for depositing analyte molecules in a user-specified pattern. In fact, the data set discussed in this section is manufactured using the Portrait 630 acoustic droplet ejection system (Labcyte Inc., Sunnyvale, CA, USA).

As indicated in Fig. 3.13, the data set consists of three synthetic peptides deposited as three distinct spatial distributions throughout a grid of 50 by 47 locations on the glass slide (280 µm interspot distance). The peptides have a mass of approximately 1559 Da (peptide A), 1305 Da (peptide B), and 1220 Da (peptide C). Their mass spectral profile from a pure sample is shown in Fig. 3.13. The spatial patterns inspired by and acquired from the logo of the institution under the auspices of which this doctoral thesis takes place. After using the acoustic droplet ejection robot to put down the three analyte molecules one after the other, the entire 50×47 grid (280 µm interspot distance) is coated with an alpha-CHCA matrix solution using the same device. Finally, the synthetic mixture of patterns is measured just like a normal MSI experiment using an ABI 4800 MALDI-TOF/TOF Analyzer (Applied Biosystems Inc., USA). The result of this procedure is a MSI data set of 2350 pixels by 18319 m/z bins.

Results

The goal of this case study is to have both the PCA and the NMF algorithm perform their unmixing of the synthetic data set, and then to assess which one comes closest to learning the original combination of molecular signatures. As we put down only three distinct molecular signatures[3], this comparison will only look at the rank-3 approximations of the data for both the PCA and NMF case.

The results of the PCA decomposition of the ESAT logo data set are shown in Fig. 3.14. PCA does a good job of separating the text pattern from both wave patterns, using principal component 1. The wave patterns however, demonstrate a less intuitive effect of PCA. They are differentiated from each other using a combination of both principal component 2 and 3. First component 2 establishes a basis for both wave patterns by providing a 'common' mass spectral signature containing elements from both the upper and lower wave patterns. The differentiation of the lower wave from the upper wave is then accomplished by principal component 3, by removing the peaks of peptide A from the upper wave, and concurrently boosting the peaks from peptide B in that same upper wave. A more telling surface plot representation of the principal components is available in Fig. 3.16. The problems of PCA regarding interpretation

[3]The matter of deciding how many components to consider in a decomposition is a nontrivial problem in the field of factorization, and lies outside the scope of this study. Mostly this parameter is set on the basis of rules of thumb or empirical considerations, but in this synthetic data set we are in the rare situation that we know exactly how many molecular species are involved.

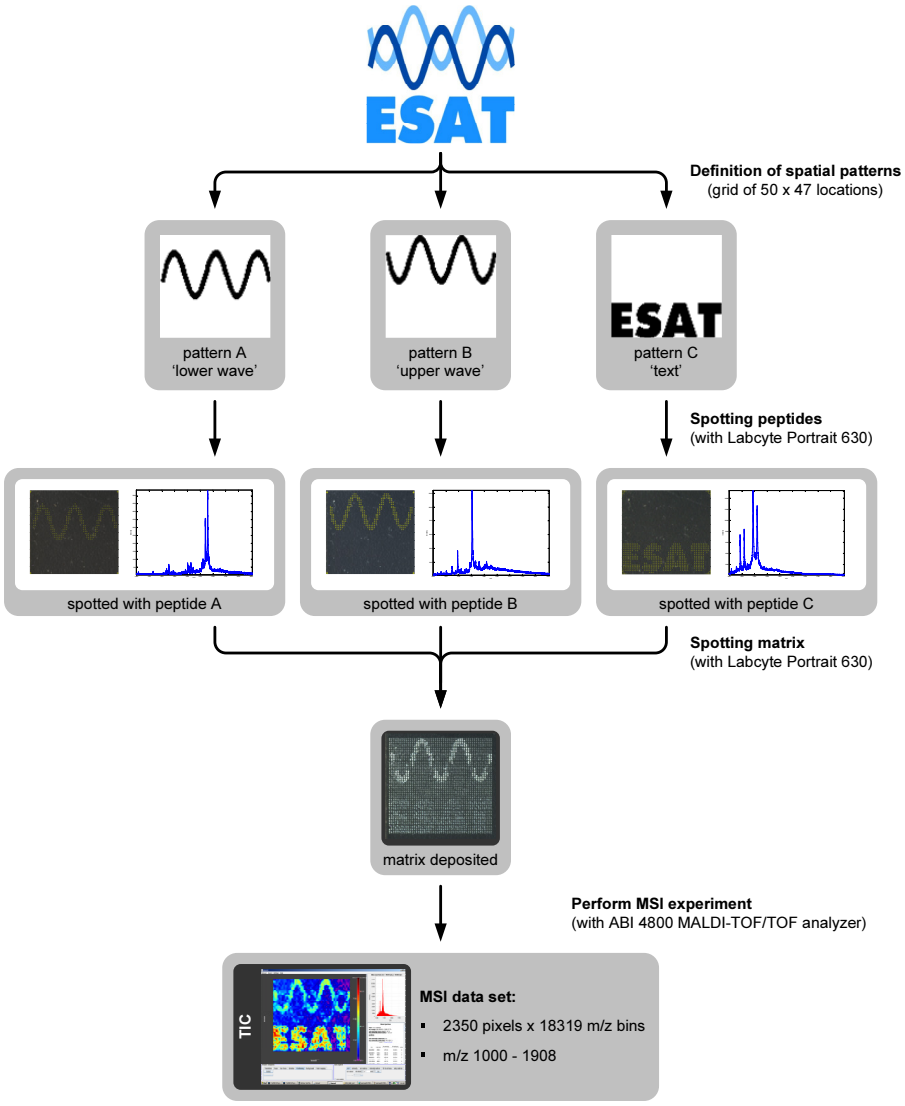


Figure 3.13. Construction of the ESAT logo data set. (a) Definition of three distinct spatial distributions (*lower wave*, *upper wave*, and *text*). (b) Spotting peptide A to pattern A, peptide B to pattern B, and peptide C to pattern C using the Portrait 630 acoustic ejection platform. (c) Spotting MALDI matrix solution (alpha-CHCA) on top of the earlier positions. (d) MSI measurement of the entire 50×47 grid, from m/z 1000 to 1908. The result is a MSI data set of 2350 pixels by 18319 m/z bins.

of the data are well demonstrated by this simple example. By allowing for subtractive components, PCA has a tendency to create 'ghost' components that never existed in the real world. A good example of such a ghost component is principal component 2. Besides the very small overlaps between the two wave patterns, there was never a mixture of peptide A and B. Nevertheless the PCA decomposition gives the impression that such a signature is really there. It is only when the subtractive third component is taken into account that there is any differentiation between both wave patterns. Although such deviations from the expected source mixture poses no problem for a computational follow-up using clustering or classification, it might definitely bias the human interpretation of the tissue content.

The results of the NMF decomposition are shown in Fig. 3.15. The unmixing process of the NMF method succeeds much better in unraveling the expected three source patterns. Just like PCA, NMF successfully separates the text pattern from both wave components, using NMF component 3. The real difference lies in how the waves are handled. As subtractive components are not allowed in a NMF decomposition, the algorithm is left little choice but to separate each wave into a separate component. The points of overlap between the upper and lower wave patterns are represented in the expected manner, as a linear mix of two finite amounts of components 1 and 2. In fact, the mass spectral patterns that NMF learns from the data set are visually indistinguishable from the pure mass spectral patterns of the peptides shown in Fig. 3.13. This further demonstrates how enforcing non-negativity in NMF causes the decomposition to be much closer to the natural mixing process that is taking place. The reason for this seems to be that mass spectrometry and MSI is the end result of additive mixing by nature itself. Figure 3.17 gives further insight into the spatial distribution of the NMF components.

A nice side effect of employing NMF decomposition is that the mass spectral signatures not only become more readily interpretable by humans, but that they can also be directly fed into algorithms traditionally developed for the processing of real mass spectra. Overall, this synthetic data set demonstrates that NMF unmixing does not suffer from the 'ghost' components typically found in PCA unmixing. Interpretation of the various forms of inter-component cross-talk that show up in PCA will become increasingly difficult as data sets grow and the number of decomposition components increases. For this reason, it seems that NMF is the preferred method for uses where the unmixing process will be followed by human expertise and interpretation.

3.4.4. Case study: MSI Experiment on Mouse Brain

To finish this section on MSI, we demonstrate its use on a MSI measurement of real tissue. The tissue is a sagittal section of mouse brain, and the acquired data set describes

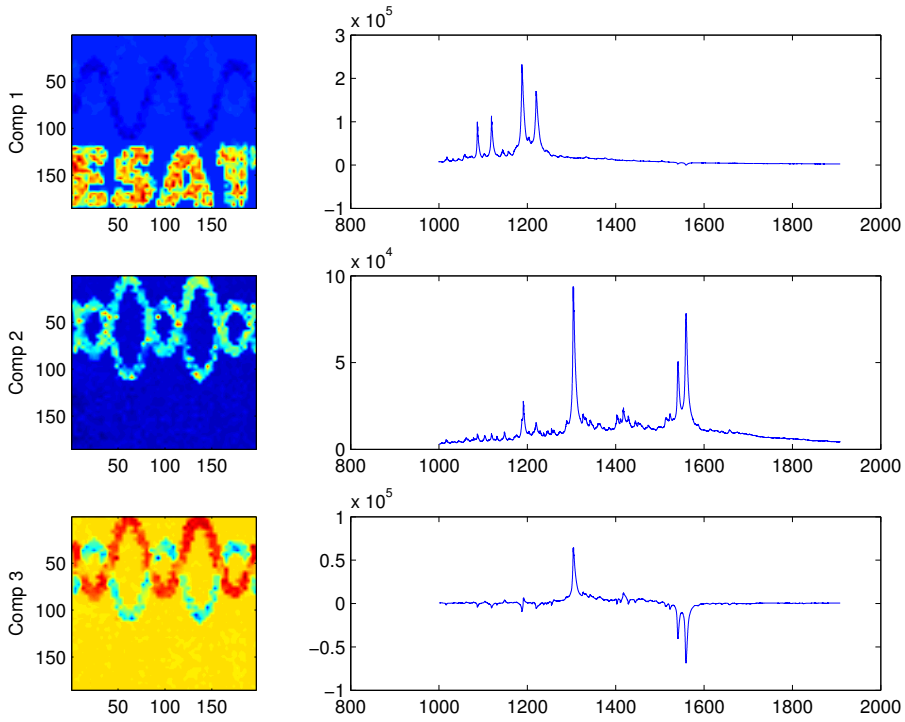


Figure 3.14. PCA decomposition of the ESAT logo data set. Notice the counter-intuitive way of differentiating between the upper and lower wave pattern. Instead of separating the two patterns into a single component each, they are differentiated by combination of an additive component (PC2) and a subtractive component (PC3). PC2 suggests a combined molecular signature of peptide A and B to be present in the 'tissue', while none such existed at creation (barring the small overlap between both wave patterns). Although mathematically correct, this decomposition has potential to mislead human interpretation.

1734 pixels (51 $\times$ 34) and 6490 m/z bins (m/z 2800 to 25000). A detailed record of the materials and methods of this MSI experiment is available in section 4.3.1.

Results

The results of a rank-6 NMF decomposition is shown in Fig. 3.18. Although similar areas are highlighted as would be the case with PCA, the NMF components present in a cleaner and more straightforward fashion. Particularly notice the mass spectrum-like signatures accompanying each NMF component. Using this decomposition, every mass spectrum in the data set is written as a linear combination of these six signatures.

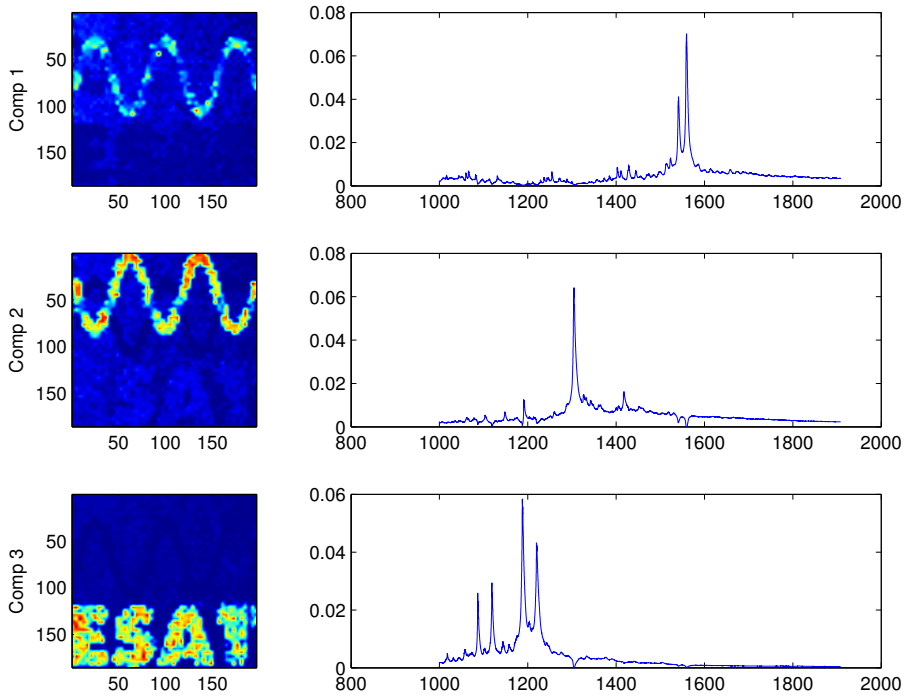
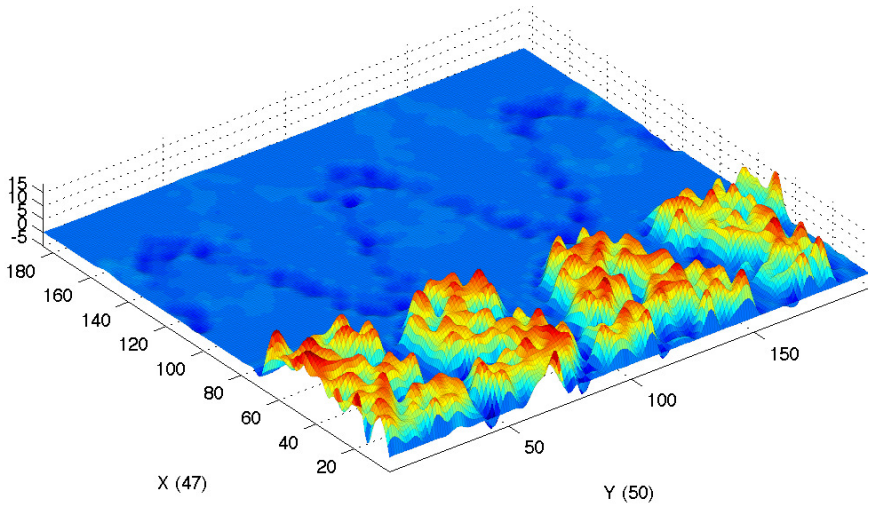


Figure 3.15. NMF decomposition of the ESAT logo data set. Notice the clear separation between the upper and lower wave pattern, each encoded by their own NMF component. This decomposition lays much closer to the real mixture process that was employed to create the data set, and avoids any 'ghost' components that potentially creep into the analysis when subtractive components are allowed.

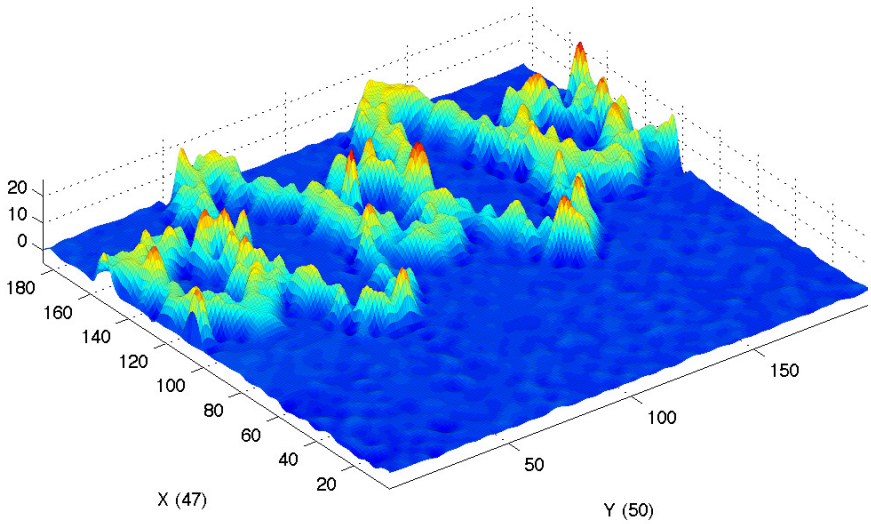
3.5. Conclusions

In concluding this chapter, it seems we can safely consider unmixing of MSI data an important tool in the multivariate analysis of imaging mass spectrometry. The basic unmixing framework discussed in this chapter, proves useful across the board with applications ranging from dimensionality reduction, denoising, and the filtering of specific signatures from the data, to aiding human interpretation of the massive amounts of data that are collected. The unmixing of MSI data basically takes the form of a matrix factorization or decomposition problem in linear algebra, and as such can be resolved by a wide variety of factorization algorithms. One can consider the process of unmixing a chassis in which a particular algorithmic engine can be dropped, customized towards the particular goals of the study at hand.

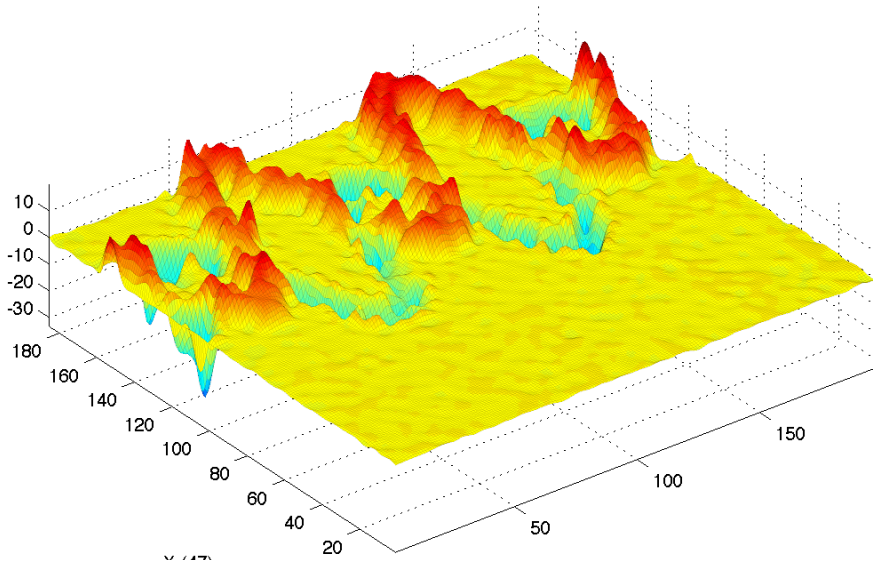
The chapter further examined and compared possible 'unmixing engines' such as PCA, PIW-PCA, and NMF. The general conclusion of these specific implementations



(a) PCA component 1 – Isolates the text pattern and peptide C for further analysis.



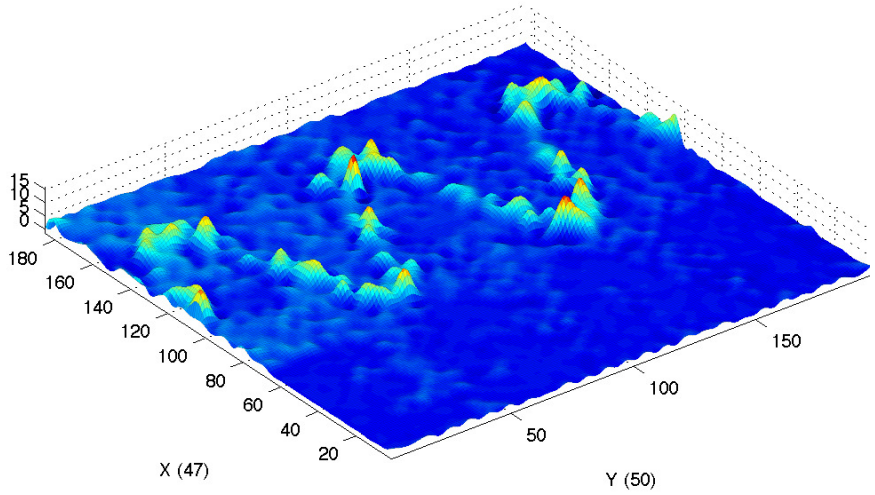
(b) PCA component 2 – Describes a 'ghost' pattern combining peptides A and B. In real life, no combination of A and B were both present at all these locations. This pattern only approximates the real data when combined with principal component 3, and demonstrates the holistic nature of PCA decompositions, and the potential for misguiding human interpretation.



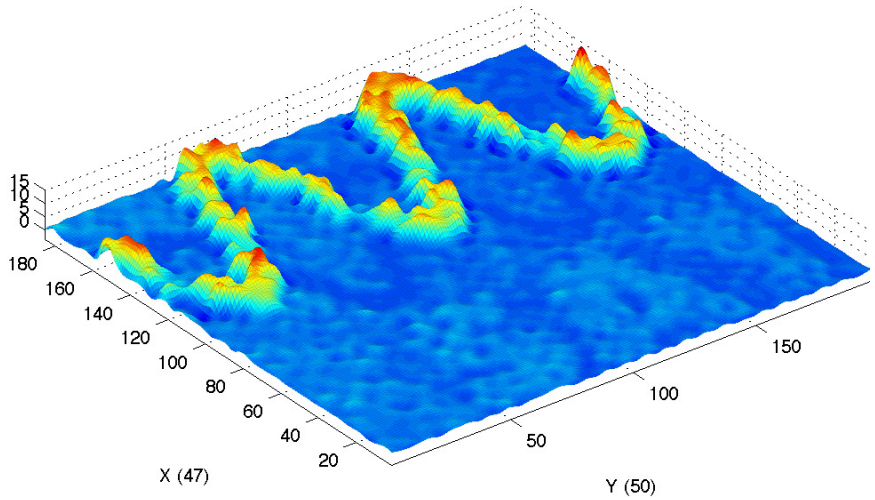
(c) PCA component 3 – Differentiates the upper from the lower wave pattern. This is a subtractive component that suggest the presence of *negative ion counts* in its blue areas. These ion count cancellations make little physical sense and only gain relevance to the real measurements when combined with principal component 2.

Figure 3.16. PCA decomposition of the ESAT logo data set (surface plots).

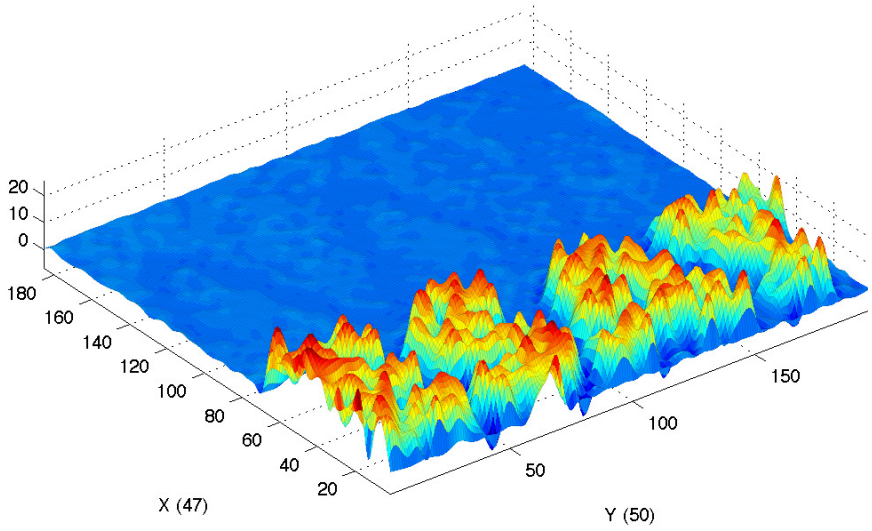
of the unmixing process, is that most probably there is not one true best unmixing algorithm for mass spectrometry and MSI. The data suggests that different unmixing goals will need different aspects to come forward in the unmixing process and as such, different algorithms will need to be developed or repurposed. PCA and PIW-PCA for example are well-established methods that have a solid statistical grounding. They are particularly useful for preprocessing roles such as dimensionality reduction and denoising prior to classification or clustering. Their calculation is rather fast for the massive data sets that can be encountered in MSI, which is the result of them having a single deterministic solution. They are, however, much less suited to produce images and signatures for human interpretation, as was shown in section 3.4.3. Their mixing model fails to capture the non-negative aspect of MS data and as a result can produce mathematically correct, but interpretation-wise misleading ghost components. On the other hand there are also statistically less stringent methods such as NMF and latent semantic indexing (LSA) [79], that are better at providing insight into MSI data. These methods usually take more calculation time as they are often implemented as iterative algorithms, and do not necessarily have a unique solution. They enforce additional constraints such as non-negativity to copy more closely the real mixing effects taking



(a) NMF component 1 – Isolates the lower wave pattern and peptide A for further analysis.



(b) NMF component 2 – Isolates the upper wave pattern and peptide B for further analysis.



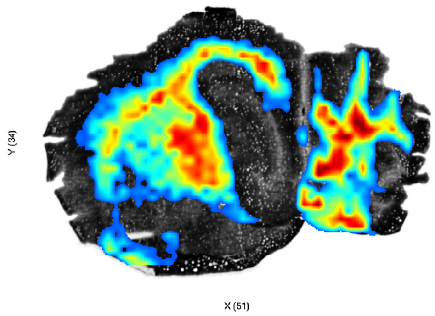
(c) NMF component 3 – Isolates the text pattern and peptide C for further analysis.

Figure 3.17. NMF decomposition of the ESAT logo data set (surface plots).

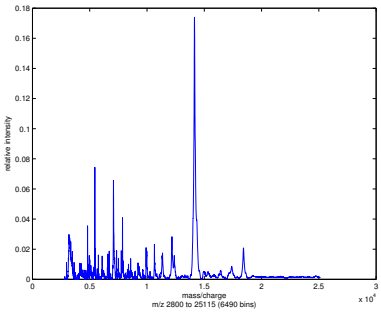
place in the tissue. These methods will produce primarily insights aimed at human consumption.

Acknowledgements

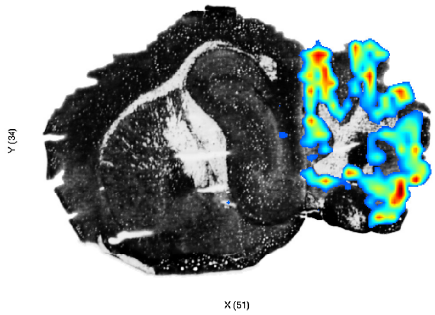
We kindly acknowledge the support from Siobhan Pickett, Jovica Pavlović, and Elaine Heron from Labcyte Inc.



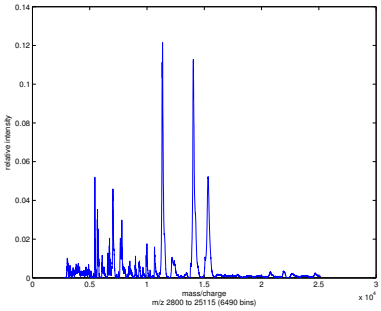
(a) Component 1 – spatial distribution



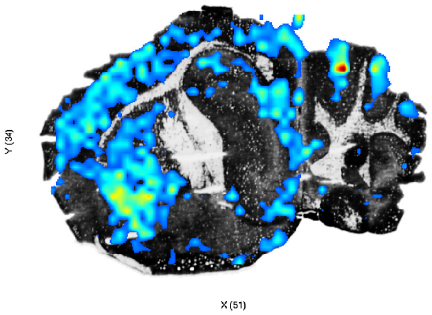
(b) Component 1 – mass spectral signature



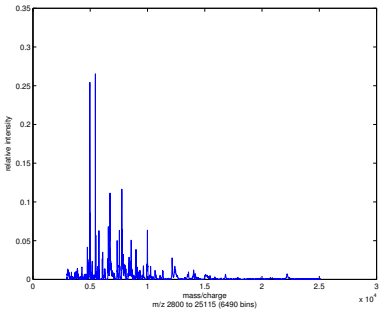
(c) Component 2 – spatial distribution



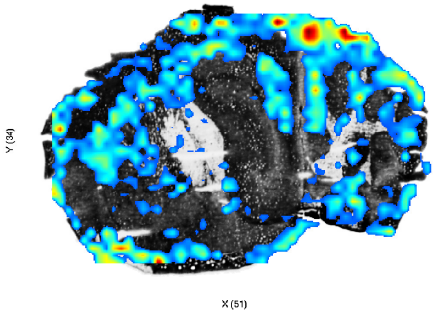
(d) Component 2 – mass spectral signature



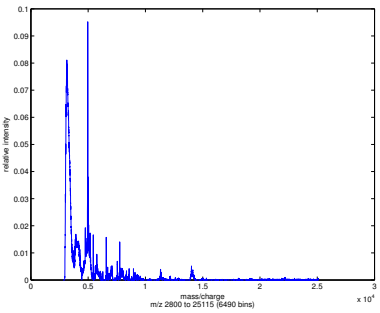
(e) Component 3 – spatial distribution



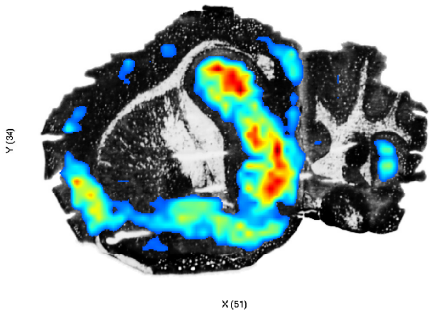
(f) Component 3 – mass spectral signature



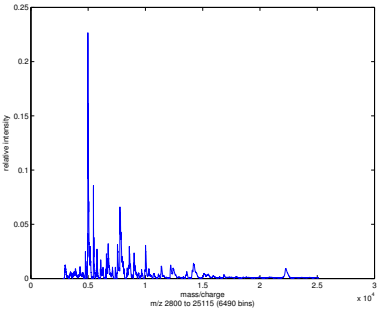
(g) Component 4 – spatial distribution



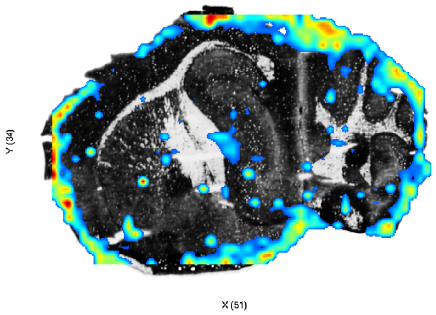
(h) Component 4 – mass spectral signature



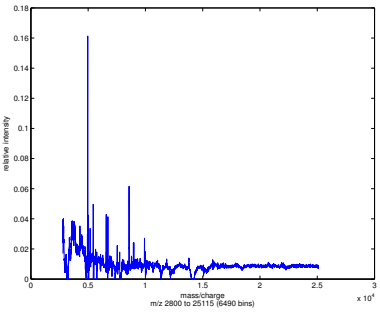
(i) Component 5 – spatial distribution



(j) Component 5 – mass spectral signature



(k) Component 6 – spatial distribution



(l) Component 6 – mass spectral signature

Figure 3.18. Non-negative Matrix Factorization applied to a MSI experiment of a sagittal section of mouse brain.

f

CHAPTER 4

Spatial Querying

The beginning of knowledge
is the discovery of something we do not understand.

Frank Herbert

This chapter is based on the following publication: • Van de Plas R. et al., "Spatial Querying of Imaging Mass Spectrometry Data: A Nonnegative Least Squares Approach", In The NIPS workshop on Machine Learning in Computational Biology, 2007.

In this chapter we develop an efficient computational method that enables the researcher to interrogate MSI data from a spatial standpoint, rather than with a traditional mass-centric approach such as ion images. The objective of such spatial querying is to retrieve the ions (or masses) that are specific to a certain localized area of interest (e.g. an anatomical region) or that exhibit a certain relative expression pattern across the tissue. Such questions arise for example in pathomechanisms that show location-specific behavior (e.g. Parkinson's and Huntington's disease), the search for anatomical region-specific biomarkers, the study of localized biochemical phenomena, and the addition of spatial information into biological models.

At its core spatial querying can be seen as a type of search engine to plough through the sea of data that make up a MSI experiment. A user can present the spatial querying algorithm with a spatial expression pattern that he is interested in, and the algorithm

will return a list of masses that best fit that expression pattern. The query pattern can be defined on the MSI grid of measurement locations directly, but it can also be defined on a registered microscopic image of the tissue section by a pathologist (on an imaging modality they are familiar with).

This chapter focuses on the development of a multivariate approach to spatial querying in which the search for biochemical distributions that resemble the query image is formulated as a convex least squares problem. Interpretability in the form of sparse coefficients is obtained by exploiting the positivity of the involved quantities, rather than through the use of dedicated priors. As a model organism case study we apply the method to the MSI measurement of a sagittal section of mouse brain and demonstrate how both synthetic tissue area specifications as well as microscopic pictures can be used for spatial queries.

4.1. Exploring MSI Data

A single tissue section scanned by MSI typically delivers a massive amount of measurements. The size and dimensionality of the data set is proportional to the spatial resolution that is used, the extent of the mass spectral range that is scanned, and the number of bins covering that range. This high dimensionality prohibits direct interpretation by a human expert. As a result, a number of *in silico* analysis methods have been formulated that either mine the MSI data for underlying trends according to a certain criterion (unsupervised mining), or examine the data in response to a scientific question or hypothesis (supervised querying). The method put forward in this paper attempts to fill a void in the supervised category by enabling to search MSI data for ions that exhibit a particular spatial pattern of relative abundance.

To frame the work discussed in this chapter, Fig. 4.1 gives a rough overview of the types of computational algorithms currently available for the analysis of MSI data. Sections 4.1.1 and 4.1.2 give a short introduction to unsupervised and mass-centric supervised approaches commonly used in MSI studies. Section 4.1.3 introduces the problem of spatial querying and explores the potential for algorithms that can render such a service.

4.1.1. Unsupervised Data Mining

Unsupervised multivariate data mining approaches [75, 76, 79, 83] are an appropriate tool for cases where (i) no hypothesis is available regarding which biomolecules, ions, or masses are relevant for the pathway or pathomechanism under study, or (ii) in prospective studies where one wants to avoid biasing the analysis towards such a hypothesis. These methods decompose a MSI data tensor into a reduced set of major

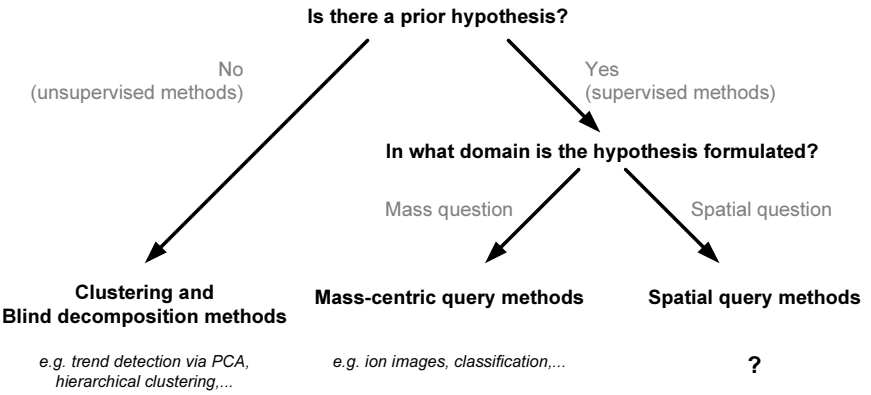


Figure 4.1. A taxonomy of algorithms for the computational analysis of MSI data. Recent studies have formulated several methods for the unguided exploration of MSI data, ranging from blind decomposition algorithms (see chapter 3) to clustering approaches. Also the ability to interrogate a data set with specific masses in mind and asking a mass-centric question to receive a spatial answer, has been catered for by e.g. classical ion images. The reverse situation however, to ask a biological question about a particular area and receive an mass answer, seems to be markedly neglected. The algorithm put forward in this chapter aims to fill this niche by providing the opportunity to query a MSI experiment with a specific spatial expression pattern of interest, and receive a hit list of masses that best fit that expression profile.

biochemical trends found within the tissue by grouping together masses that show, for example, correlated behavior. The resulting trends are each characterized by a spatial distribution (showing which tissue areas differ significantly from other areas) and a mass spectral signature (identifying the masses responsible for these differences). An example of such a decomposition can be found in [75]. As these unsupervised methods require no user-specified query or information and perform their decomposition according to a predefined similarity measure, they have particular merit in exploratory or lead discovery settings.

4.1.2. Supervised Mass Query

In cases where a scientific hypothesis requires validation, unsupervised methods are much less suited. They do not take hypothesis specifics into account while performing their decomposition, which means that one has to examine all resulting trends for information on the particular masses or spatial areas one is interested in. Let us refer to such an examination as to a *query*. An unsupervised approach runs the risk that interesting masses or areas become overshadowed by dominating effects since the employed distance measures might not be appropriate for the question or hypothesis at hand. If however the biological question or query can be formulated purely in the mass domain as a certain mass or ion of interest, a straightforward method exists in

ion images. An ion image is a cross-section of a MSI data tensor and shows the spatial distribution throughout the tissue of a particular mass (or m/z). It is a common approach for interrogating MSI data, albeit with the *caveat* of being a univariate method that ignores the between-mass relationships. Examples of ion images are shown in Fig. 1.3 and Fig. 1.4.

4.1.3. Supervised Spatial Query

In addition to the queries mentioned in the previous section, there exists a type of queries that present themselves in the spatial domain rather than the mass domain. They arise from scientific questions that focus on a particular zone in the tissue and they try to obtain insight into the chemical signatures and relationships specific to that zone. One example is the search for ions that show up only or particularly abundant in a specific anatomical region or tissue type in the section (e.g. increased presence of a marker in a patch of tumor tissue). Another example is the search for masses that show a relative difference in expression level between two particular areas (e.g. upregulation in the gray matter zone of a spinal cord section versus its surrounding white matter areas or vice versa). Although these types of questions arise quite frequently in biology and medicine, a method for handling such spatial queries to MSI experiments is currently lacking. This absence is in stark contrast to the potential uses for such a method, particularly for examining co-localization-type chemical relationships and for elucidating pathomechanisms that show spatially specific behavior. These include pathologies such as Parkinson's disease (where dopamine producing brain nuclei such as the *amygdala* and the *putamen* are affected [84]), Huntington's disease (with cell death in the *striatum* [85]), amyotrophic lateral sclerosis (with a striking degeneration of motor neuron regions in the spinal cord [86, 87]), and spinocerebellar ataxia affecting the Purkinje cell structure with an anatomical shrinking of the *cerebellum* (demonstrating variable degrees of neurodegeneration in the *cerebellum*, brainstem, and spinocerebellar tract [88]). Additionally, it has been shown that the affected regions can behave differently depending on the stage of the disease. An example in Parkinson's disease is the upregulation of the dopaminergic function in the *pallidal*, *amygdala*, and *cingulate* regions in the early stage and the reduction later on. The loss of *putamen* dopaminergic function has been shown to correlate with rigidity and bradykinesia [89]. For all these diseases, mouse models are currently available and in addition, mouse models demonstrating Tau accumulation at the spinal motor neuron regions have been described [90]. Concluding, it is apparent that in all these examples imaging mass spectrometry can be applied to study the alterations in the affected regions at various stages of the disease, with particular advantages if the analyses can be further focused on the known areas involved.

With these applications in mind, this chapter attempts to address an apparent lack of robust supervised spatial querying methods for MSI data, by formulating a multi-

variate approach based on a least squares optimization problem. It delivers an interpretable and ordered view into the masses that are active in the region of interest.

4.2. Methods

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

Definition 4.2.1 (Spatial Query to a MSI Experiment)**.** A *spatial query* posed to a MSI data array aims to retrieve molecular masses (or ions) which *'exhibit'* a spatial pattern of interest through their ion counts. A spatial pattern can typically be represented as a query image. The understanding is to retrieve a small number of such masses which are most likely contributing to the pattern, helping the researcher to investigate a location-specific effect or to design a specialized experiment.

Subsection 4.2.1 first explains a naive univariate approach to the problem, which establishes a reference for the task and exposes a number of important *caveats*. Subsection 4.2.2 presents the multivariate method that we propose in response, and explains a number of data constraints that allow us to efficiently solve such a problem in a reasonable timeframe. Both methods are demonstrated in the case study of section 4.3 using a MSI measurement of mouse brain.

4.2.1. Univariate Correlation-based Spatial Querying

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

$$\left\{\phi^m \in \mathbb{R}_+^K\right\}_{m=1}^M. \tag{4.1}$$

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

Arguably the most simple approach to answer a spatial query is to return the masses corresponding to the top correlated features. This amounts to a univariate method were we scan each mass and rank according to a measure of correlation. The Pearson correlation coefficient provides a most convenient measure $\rho \in [-1,1]$ defined for any $m = 1,\ldots,M$ as

$$\rho^m = \frac{q^T\phi^m}{\|q\|_2\|\phi^m\|_2}. \tag{4.2}$$

Further, it can be argued to use a rank correlation coefficient such as Kendall's τ to rank the different features. This makes sense when the underlying distributions can be expected to deviate substantially from normality. Kendall's tau taking a value between -1 and 1 is defined for any $m = 1, \ldots, M$ as

$$\tau^m = \frac{2}{(1-K)K} \sum_{1 \le k < l \le K} \operatorname{sign}\left((q_k - q_l)(\phi_k^m - \phi_l^m)\right). \tag{4.3}$$

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

While the profile of those coefficients over different masses (see Fig. 4.4(b)) gives immediately feedback on the query, it remains up to the user how many top ranked features should be studied. This amounts to thresholding the correlations at a user-defined value. It can be hoped for to find such a threshold based on a statistical test of correlation, see e.g. [91]. However, since the test is applied uniformly over all M features, one needs to apply a statistical correction for multiple testing such as Bonferroni's. Because of the large M, this results in inferior or even non-informative results. A second disadvantage is that this approach does not take the correlation between different features into account. This can be seen as follows: if all features reflect one spatial pattern (e.g. the contour of the tissue), then the correlations of all features with a related query image will simultaneously increase. Therefore, the use of such a data-independent threshold method (e.g. based on a significance test) would not work reliably.

4.2.2. Multivariate Least Squares-based Spatial Querying

This subsection describes a multivariate approach based on a least squares argument. It looks for the most optimal (and smallest) combination of ion images that when multiplied by their mass contribution coefficients add up to the target image specified in the query. It is this optimal profile of mass contribution coefficients p that essentially answers our spatial question (see Fig. 4.2). Ions corresponding to high coefficients will exhibit an expression pattern that more closely resembles the requested pattern q than the other ions available in the data set. The following linear model is adopted for any $k = 1, \ldots, K$

$$q_k = \sum_{m=1}^{M} \phi_k^m p_m + \varepsilon_k, \tag{4.4}$$

where the coefficients $p = (p_1, \ldots, p_M)^T$ are restricted to positivity, encoding the assumption that the image query q is a weighted average of the features, up to the residuals $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)^T \in \mathbb{R}^K$. This means that the query image is assumed to be a sum of positive contributions from a finite set of molecular masses. An assumption we make is that the noise in the data does not exactly mimic the spatial pattern of the

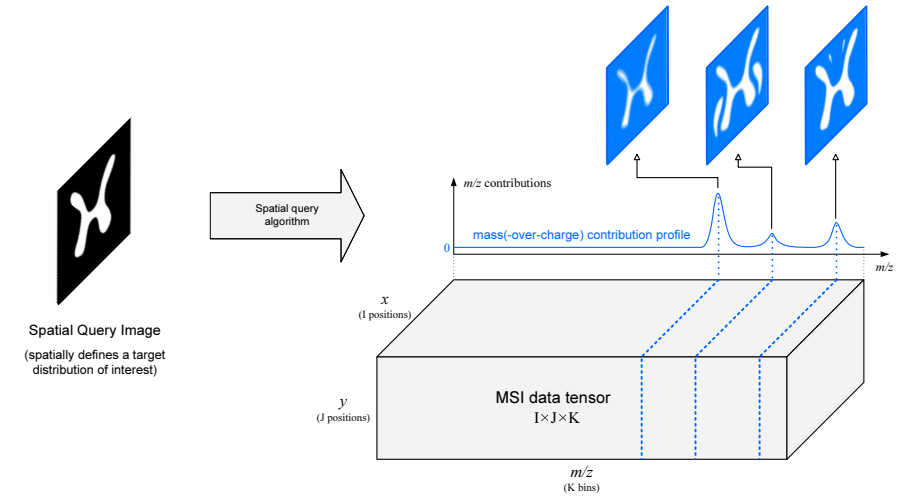


Figure 4.2. Overview of the spatial querying procedure. The query is formulated in the spatial domain as a region of interest in the tissue (drawn on the MSI measurement grid). The method returns a solution vector in the mass domain, indicated as the mass(-over-charge) contribution profile. The masses that have a nonzero contribution describe a distribution in the tissue that (to some extent) mimics the shape and relative gray level topology specified in the spatial query.

query image. A classical approach to approximate linear coefficients based on a set of measurements is to minimize the squared norm of the residuals, or

$$p^* = \underset{p \geq 0_M}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{m=1}^{M} \phi_k^m p_m - q_k \Big)^2. \tag{4.5}$$

Remark that the choice of a Poisson regression model could be more actual as the involved data are count data. We, however, stick to a least squares approach for computational reasons. Let 0_K denote a vector of size K with all-zeros. Then one can reformulate the problem (4.5) as a standard convex quadratic problem as follows

$$\min_p \left[\frac{1}{2} p^T H p - f^T p \right] \quad \text{s.t.} \quad p \geq 0_K \tag{4.6}$$

where the positive semi-definite matrix $H \in \mathbb{R}^{M \times M}$ is defined as $H_{ij} = \phi^{iT} \phi^j$ for all $i, j = 1, \ldots, M$, and $f \in \mathbb{R}^M$ is defined as $f_m = q^T \phi^m$ for all $m = 1, \ldots, M$. This estimator has the peculiar property that many mass coefficients are set to zero as a consequence of the positivity constraints. This sparseness of p improves the interpretability considerably when compared to the univariate method discussed in section 4.2.1. The positivity constraints also allow these convex programs to be solved efficiently given the large amount of data (e.g. 30 minutes for the case study queries using the MOSEK Toolbox[1]). It must be emphasized that sparsity is not originating from the use of an

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

L1 regularization scheme as in LASSO, see e.g. [92], or equivalently from assuming a Laplace prior on the parameters. It however derives naturally from a direct property of the involved data.

Moreover, if one knew *a priori* where the positive non-zero coefficients would be located, duality theory [92] states that the optimal non-zero coefficients can be computed by solving an ordinary least squares problem of correspondingly reduced size. This implies that the solution is fairly reliable in case of small number of non-zero coefficients. On the other hand, since usually $K < M$ the effect of ill-conditioning would deteriorate the solution if many coefficients are nonzero. A classical method to cope with those numerical issues is to impose an extra regularization term on the Hessian matrix $H' = H + \gamma I_M$ for an appropriate choice of $\gamma > 0$. The resulting quadratic program (4.6) solves then the regularized least squares problem

$$p^*_\gamma = \underset{p \geq 0_M}{\arg\min} \left[\frac{1}{K} \sum_{k=1}^{K} \left(\sum_{m=1}^{M} \phi_k^m p_m - q_k \right)^2 + \gamma \sum_{m=1}^{M} p_m^2 \right]. \tag{4.7}$$

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

Now, we illustrate how this formulation can be extended for pixels where we do not have a specific query requirement, that is, the query model (4.4) is extended with the occurrence of *don't care* pixels. A naive solution is to omit such pixels in the least squares formulation (4.5) so that we effectively answer the query based on far less information. Since the inference problem (4.4) is in general already ill-conditioned as $M > K$, we prefer the use of a proper weighting scheme to decrease the influence of pixels which are of minor importance to the query at hand. The interpretation is that we do not mind the values of down-weighted pixels, unless it could help the algorithm to make a choice between a set of equally good hypotheses. This is encoded by a positive vector $w = (w_1, \ldots, w_M)^T \in \mathbb{R}^M_+$ as follows

$$p^*_w = \underset{p \geq 0_M}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} w_k \left(\sum_{m=1}^{M} \phi_k^m p_m - q_k \right)^2. \tag{4.8}$$

These pixel weights can be interpreted as a *mask* image, focusing attention of the search strategy to the relevant spatial neighborhood. In the experiments of section 4.3, we use the weights $w_k = 1$ for relevant pixels, and only one percent influence $w_k = 0.01$ for *don't care*s.

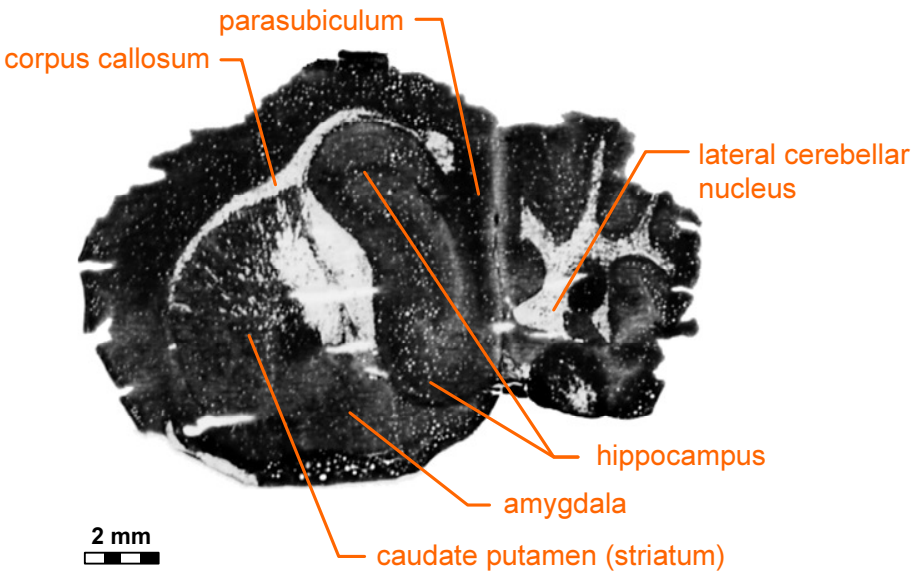


Figure 4.3. Microscopic image of the sagittal section of mouse brain that is imaged in section 4.3. Several anatomical regions are labeled for reference.

4.3. Case Study

In this section, we apply the spatial query methods developed in section 4.2 to the MSI measurement of a mouse brain tissue section. Figure 4.3 shows a microscopic picture of the tissue slice, with a number of anatomical regions specified for reference.

4.3.1. Materials & Methods

The example is a MSI measurement of a tissue slice (10 µm thick) that was taken from an off-center sagittal section of the brain of a BL57/6 mouse. The recorded mass range extended from m/z 2800 to 25000 divided into 6490 m/z bins. Sinapinic acid (in 75% acetonitrile and 0.2% TFA) was applied as a chemical matrix solution, using an ImagePrep station from Bruker Daltonics. A MALDI mass spectral measurement was performed on each grid point of a virtual raster of size 51×34 that was superimposed on the tissue section with an interspot distance of 300 µm in both the x and y-directions. The mass spectrometer was the Autoflex III MALDI TOF/TOF from Bruker Daltonics in linear mode, and the data collection was guided by the FlexImaging module. The spatial query method is applied on the raw spectra: no extra preprocessing (such as baseline substraction or smoothing) was performed since we would like to assess the usefulness of the method for explorational purposes and without

dependency on preprocessing model assumptions or parameters. The described data analysis was done using in-house developed software, MATLAB from MathWorks Inc., and the MOSEK Optimization Toolbox from MOSEK ApS.

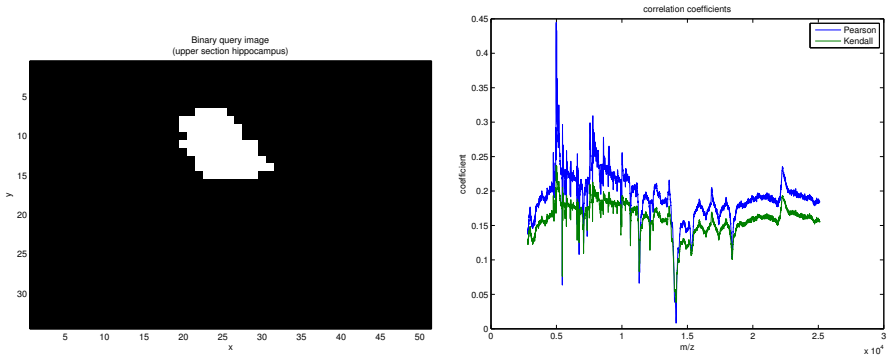
4.3.2. Results

We perform spatial queries on two specific anatomical regions present in the sagittal mouse brain section. The goal is to assess which molecular masses show a spatial distribution that resembles as a whole or in part the area of interest specified by the query. The first region of interest is the upper half of the *hippocampus*, visible in Fig. 4.3 as the central bulbous region. The second area under investigation is the elongated *corpus callosum*, shown in Fig. 4.3 as the narrow pale area.

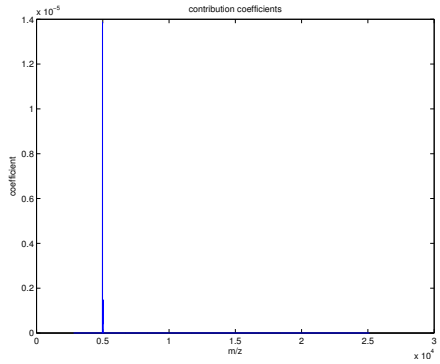
Upper Hippocampus

The upper *hippocampus* area is examined using two synthetic query images, drawn as overlays on the MSI measurement grid. The first query, shown in Fig. 4.4(a), is binary encoded. The pixels that are considered part of the area of interest are given a value of 1. The search for masses whose spatial presence corresponds to the drawn area, is performed using both the univariate and the multivariate methods described in section 4.2. The univariate correlation-based method is run twice, once with the Pearson and once with the Kendall correlation coefficient, in order to obtain a reference result with which to compare the multivariate performance. For each m/z bin, the correlation between its ion image and the query image is calculated. The resulting profile of correlation coefficients spanning the mass range is shown in Fig. 4.4(b). The binary query is then processed using the multivariate least squares method, with no *don't care* pixels specified and with the regularization parameter γ set to 5×10^8. The resulting profile of contributing masses is given in Fig. 4.4(c) (although not clearly visible due to strong compression along the m/z axis). It shows two mass windows that know an increased presence in the upper *hippocampus*: one from m/z 4975 to 4984 (highest contribution at m/z 4980), and the other from m/z 5014 to 5016 (highest at m/z 5014).

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



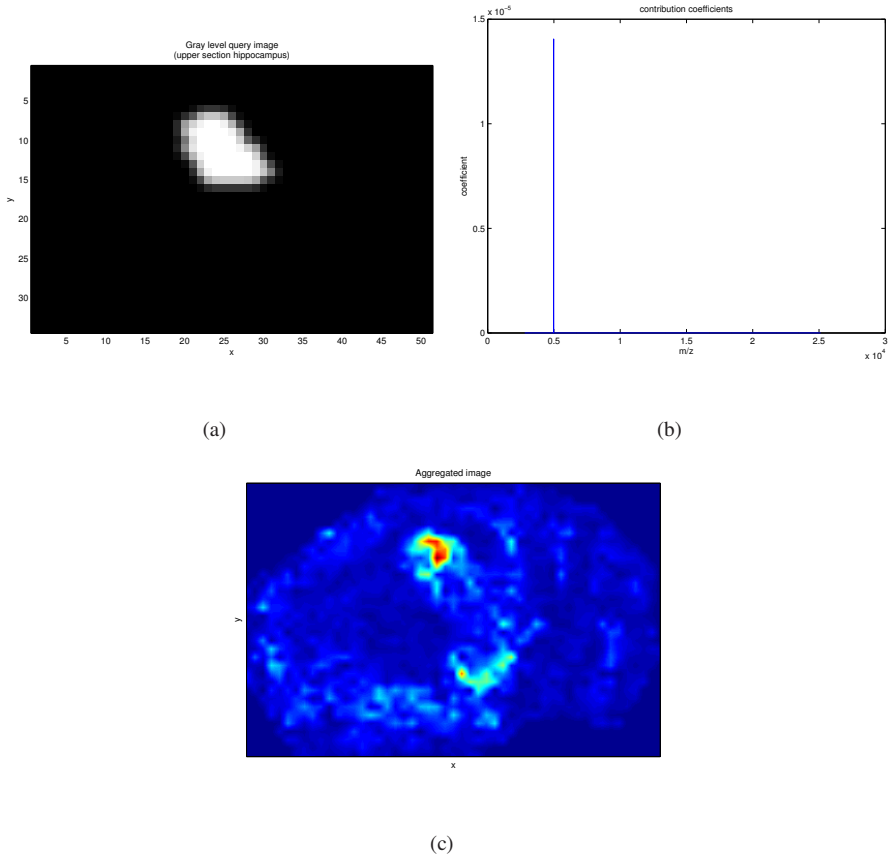
(a) (b)



(c)

Figure 4.4. (a) Binary query image examining the upper *hippocampus* area in the sagittal mouse brain section. (b) Univariate method – Correlation profile of the masses (or ions) that are spatially specific to the upper *hippocampus* area, using both the Pearson and the Kendall correlation coefficients. (c) Multivariate method – Contribution profile of the masses (or ions) whose presence corresponds spatially to the binary specification of the upper *hippocampus* area. Due to the extent of the mass range the graph seems to show only one specific mass contribution, but this peak is actually a narrow representation of two neighboring peak families at m/z 4980 and 5016.

version of the upper *hippocampus* query to the mouse brain measurements (see Fig. 4.5(a)). The difference with the binary query is the gradual decrease of molecular presence at the edges of the area. This effect is closer to real biochemical distributions than the crisp area edges specified in Fig. 4.4(a). The corresponding profile of involved masses is presented in Fig. 4.5(b). The retrieved masses are identical to the binary query results, but the individual coefficients differ somewhat (primarily a



(a) (b)

(c)

Figure 4.5. (a) Gray level query image examining the upper *hippocampus* area in the sagittal mouse brain section. (b) Contribution profile of the masses (or ions) whose presence corresponds spatially to the gray level specification of the upper *hippocampus* area. Due to the extent of the mass range the graph seems to show only one specific mass contribution, but this peak is actually a narrow representation of two neighboring peak families at m/z 4980 and 5016. (c) Aggregated image calculated from the mass mixture found to be specific to the upper *hippocampus* area. The image indicates that the lower *hippocampus* area has an identical mass spectral signature to the query region.

decrease of contribution from the window at m/z 5016.)

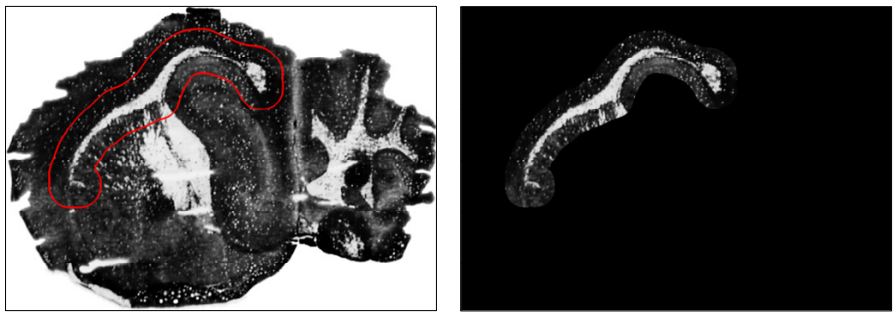
An interesting extension is the construction of an *aggregated* image from these masses by multiplying their ion images with their contribution coefficients and summing the products. The result is a type of mass-constrained weighted average image. Masses whose ion image mimics the spatial query more closely will have a higher contribution coefficient and thus, a higher influence on the aggregated image. However, in practice an ion image will never be an exact copy of the spatial query, which means

that the aggregated image will pull in other areas as well. The aggregated image is a means of examining which other areas in the tissue show a chemical composition that is identical to the one found in the region of interest originally queried for. In essence, one is asking a spatial question, retrieving the answers in the mass domain, and then using this mixture of masses to pull in other spatial areas that exhibit a similar chemical mixture. Figure 4.5(c) shows the aggregated image that was calculated from the results of the gray level *hippocampus* query (Fig. 4.5(b)).

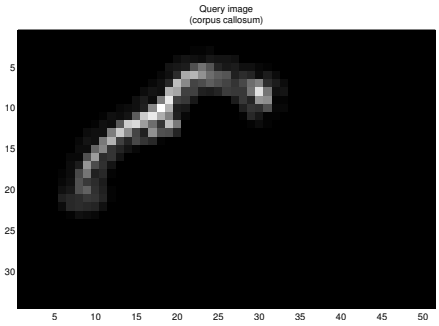
Corpus Callosum

The spatial query used for the examination of the *corpus callosum* is not synthesized by specifying an area of interest on top of the MSI measurement grid. Instead the query image is the result of delineating a region on a registered microscopic image of the tissue section, and by using the visual intensity levels to specify the texture of the query area. The goal is to demonstrate that a spatial query does not have to be an abstract selection of pixels at the usually crude spatial resolution of the MSI experiment, but can also be constructed by an expert who has experience with the tissue type under study, by selecting regions in a microscopic image of the section. It is expected that this will improve accuracy and ease of use as the higher spatial resolution of the microscopic picture allows for clearer tissue navigation. Additionally, it allows experts to pose questions to MSI data working from pictures that lie much closer to the traditional imaging modalities they usually have experience with (e.g. fluorescence miscrocopy). Figure 4.6(a) shows the registered microscopic image with the borders of the selected *corpus callosum* region indicated in red. The area is selected sufficiently wide in order to incorporate its border gradient from pale to dark into the query texture later on. The result of the selection procedure is shown Fig. 4.6(b). This image is downsampled to the MSI grid size using a standard nearest-neighbor algorithm, and then used as the (gray level) spatial query for investigating the *corpus callosum* (see Fig. 4.6(c)).

In this example, we also make use of the *don't care* pixel extension developed in section 4.2.2. As earlier, we want to find the masses that show a spatial presence resembling the one specified with gray levels inside the query area. However, in the earlier *hippocampus* queries the area outside the specified region was set to zero, forcing to reconstruct also the black plateau of the query. For the *corpus callosum* query we want to make the mass search solely dependent on the selected area, mostly ignoring the ion images' content outside that area. This is done using a *mask* image that gets translated into a pixel weight vector for the optimization problem. The mask image for the spatial query of Fig. 4.6(c) is constructed from the same user-specified area as earlier, and is binary encoded using a value of 1 for all pixels that fall outside the *corpus callosum* region and that are to be ignored. This mask, shown in Fig.



(a) (b)

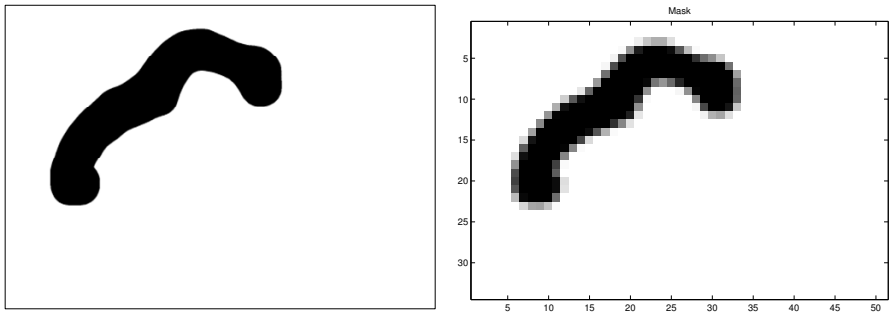


(c)

Figure 4.6. (a) Microscopic image of the brain tissue section, registered to the MSI measurement grid. The borders of the selected *corpus callosum* region are indicated in red. (b) Selected area from the registered microscopic image at its original resolution. (c) Downsampled version of the selected region, which is used as the query image for examining *corpus callosum*-specific masses.

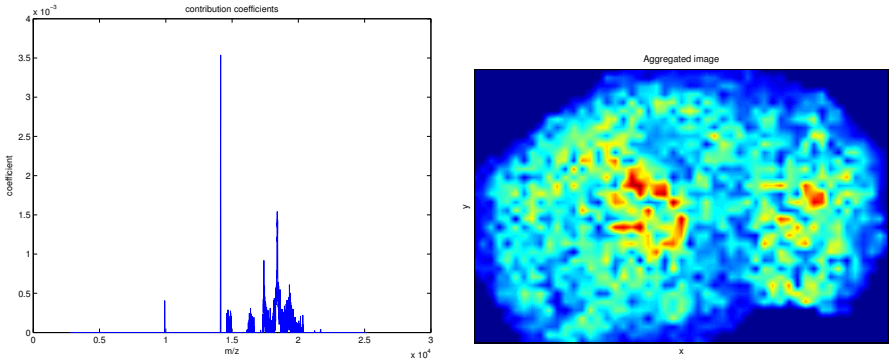
4.7(a), is downsampled to the mask image at MSI resolution of Fig. 4.7(b) and fed to the multivariate query method.

The gray level query of Fig. 4.6(c), combined with the mask of Fig. 4.7(b), is processed by the multivariate spatial query method using again a regularization parameter $\gamma = 5 \times 10^8$, and produces the contribution profile shown in Fig. 4.8(a). The profile shows that several mass windows along the range demonstrate *corpus callosum*-presence. Particularly the masses at m/z 9918, 14140, 17400, and 18410 have ion images that follow the visual appearance that was extracted from the picture. An aggregated image is calculated as well and shown in Fig. 4.8(b).



(a) (b)

Figure 4.7. (a) Mask image at original microscopic resolution. The white area represents the *don't care* pixels, having a value of 1. (b) Downsampled version of the mask image at MSI resolution. It is translated to a vector of weights for the optimization problem.



(a) (b)

Figure 4.8. (a) Contribution profile of the masses (or ions) whose presence corresponds spatially to the visual picture gray levels of the *corpus callosum* area. (b) Aggregated image calculated from the mass mixture found to be specific to the *corpus callosum* area.

4.3.3. Discussion

Univariate Spatial Querying

Although based on a simple design, the correlation profile of Fig. 4.4(b) provides fast and qualitative answers to the query. It indicates for example the high relevance of bins occurring around m/z 4980. However, the disadvantages of the univariate method become apparent as well. There are no zero contribution masses, defying

the purpose somewhat as in practice the entire mass range is selected as part of the location-specific mass mixture. Thresholding the coefficients would help, but the profile is noisy to such an extent that developing a dynamic thresholding strategy is not a trivial problem. The disadvantages show up both using the Pearson and the Kendall correlation coefficients. An additional concern is the effect of the multiple testing problem inherent to the approach and the need to employ a suitable correction for the results affected by it.

Multivariate Spatial Querying Using Binary Image

Using the same binary query, the multivariate method delivered a much more informative contribution profile (Fig. 4.4(c)). The solution is much sparser, with only two mass windows having nonzero contribution coefficients. This sparseness avoids the need for a thresholding strategy as with the univariate approach. The nonzero coefficients, indicating which masses (or ion images) approximate the spatial query best, also turn out to be grouped into windows of related neighboring m/z bins. This concurs with the notion of adjacent m/z bins often receiving ion counts from the same or related mother molecules due to effects such as their isotopic distribution or the binning strategy of the mass spectrometer. Additionally, the size of the individual contribution coefficients implies an ordering on the selected masses according to their overall resemblance to the query image. These observations illustrate the added value of the proposed multivariate method over the univariate approach.

Multivariate Spatial Querying Using Gray Level Image

We performed the same analysis for the gray level query as for the binary query and the contributing masses were found to be virtually identical (see Fig. 4.4(c) versus Fig. 4.5(b)). As expected, there are small differences in their individual contribution, indicating that certain masses from the set approximate the gradient edges of the gray level query better than others. In the example, m/z 5014 and 5016 show a diminished contribution when compared to the binary query results.

The aggregated image of Fig. 4.5(c) pulls in other spatial areas that show a chemical profile similar to the upper *hippocampus* area that was queried for. The tissue-wide spatial presence of the found mass mixture is shown from yellow to red. The primary added region is the lower area of the hippocampus. This is no surprise as we set out using a spatial query that selected only part of the *hippocampus*, which as a known anatomical region in the brain has a more or less homogeneous composition throughout. None of the selected ion images described only the upper *hippocampus* area. Given the positivity-enforced sparsity, this means that throughout the entire data set no ion image could be found that shows increased ion presence only in the upper

hippocampus and not in the lower *hippocampus*. This example indicates the capability of the aggregated image to complete anatomical regions if only part is selected purely on the basis of relative ion presence. It can equally well be used in a discovery setting to identify unknown regions that show a chemical signature similar to an area known to be active in a certain pathway or mechanism. It provides an interesting added source of information, and has particular value in exploratory settings or for experiment design.

Multivariate Spatial Querying Using a Visual Picture Area

The contribution profile of Fig. 4.8(a) demonstrates the capability for a visual picture to be used as the source for spatial querying. This type of application is very interesting for the interrogation of MSI data by non-MSI experts such as pathologists. It provides an interface closer to traditional imaging technologies such as microscopy. Also, by using the visual texture from the microscopic image as the gray level query target, we are in a sense combining information from a separate data source, a microscopic image of the tissue section, with the MSI measurements performed on that same tissue slice.

The aggregated image, shown in Fig. 4.8(b), pulls in a number of different anatomical regions that exhibit a similar mass spectral signature to what was found in the *corpus callosum*. These regions (indicated in yellow and red) include the *lateral cerebellar nucleus* and its extensions (which are also clearly visible on the picture of Fig. 4.3) and the *caudate putamen* or *striatum* region. These inclusions indicate a mass spectral signature common to all three regions.

A primary aspect of the method is that it takes a *global* view towards spatial querying. This means that it incorporates all measurement locations into the model, rather than only selecting the spectra that fall within the region of interest (as would be the case by simply cropping to that region). The reasons for this are multiple. One reason is the importance of the *gradient* of the query, capturing the border of the region of interest. Specifically, in the *don't care*-extended model it is left to the user to specify how 'confident' one is in the exact shape of the query. In many cases one wants to leave open the possibility to find masses of regions which only roughly correspond to the spatial query at hand. A second reason is the increased numerical stability of the optimization problem provided by the additional measurements. This way one avoids ill-conditioning of the procedure by keeping the number of pixels versus the number of m/z bins as large as possible.

Towards Customization?

Besides the spatial connections implicitly encoded into the query image, the model described in eq. 4.4 considers all pixels and m/z bins as separate entities and does not incorporate additional prior information along the spatial or the mass spectral domain. Examples of such additional prior information include neighborhood relations between pixels and chemical relationships between m/z bins due to post-translational modifications of a similar protein. Although the results of the basic model are quite satisfactory, the extension towards using such prior information is an interesting area of further investigation. An extension of the model is one possibility but such information could equally well be incorporated into a more application-specific form of regularization while keeping the current model.

In this chapter we use a common L2-regularization scheme to demonstrate its use for steering the optimalization effort to a more robust and interpretable solution. However, the regularization term can be further customized for the spatial querying problem in particular and the fact that we are dealing with mass spectral data with a certain chemical structure underlying it. The weight γ given to the regularization term is dependent on the ion count values found in the original MSI data. As no normalization of the ion counts was performed, the value of γ was set to an empirically determined value for the case study, but an automatic estimation of γ is a possibility as well. Such a procedure needs to take into account the type of regularization term that is used and the types of additional prior information and constraints incorporated in it. Concluding, the parameter γ provides the researcher with a means of exploring the robustness of a given solution and a means of investigating the strength of the link between the inclusion of certain ions in the solution profile and the prior chemical or spatial information explicitly embedded in the regularization term.

Regarding the robustness and noise sensitivity of the method one can state that the results of this approach seem to be largely orthogonal to spatially uncorrelated artefacts present in the spectra. Spatially correlated noise could potentially influence the results in the not very likely event that it happens to mimic to a significant degree the search pattern provided in the query image. It is apparent that prior knowledge of the exact noise-model of the measured spectra could be incorporated in the computational model. Specifically, if the noise can be described as a spatial process with certain parameters, this could be used to adapt the least squares object function.

At this point, we deem it not yet relevant to elaborate on these thoughts for a number of reasons. At first, it remains largely unknown how the *complete* spatial structure due to the chemical relationships between all m/z bins would look like. Similarly, it is an open question which spatial process would capture the actual noise process well. It is clear that such descriptions will probably be highly instrument-dependent. Secondly, any deviation from the prior spatial/noise model would deteriorate results.

For example, if two ion bins are in reality not chemically related and this is contrary to what is conjectured in the prior spatial model, the data will have to provide much evidence to counter that assumption and decouple their mass contribution coefficients in the resulting profile. In such cases where there is a mismatch between the prior model and reality the computational method would consequently yield many spurious nonzero coefficients. As such one can argue that the schemes discussed in Subsection 2.2. are more robust. Thirdly, since the method is conceived as an efficient computational method to assist in exploring the raw data, it makes sense to let the data speak for themselves, rather than pushing the outcome towards a certain direction *a priori*. These findings however do not contradict the fact that future work on exactly this topic is to be expected to refine the methodology.

4.4. Conclusions

The technology of imaging mass spectrometry adds spatial information to the traditional mass spectral biochemical analysis. With the method of spatial querying presented in this chapter, we provide the researcher with a means of interrogating MSI experiments from the spatial viewpoint, rather than the traditional mass-centric approach. The objective is to learn which among the thousands of measured molecular masses or ions show increased presence in a specific area of the tissue (e.g. at the center of a tumor) or exhibit a relative up- or downregulation when comparing various tissue areas (e.g. white matter versus gray matter brain stem tissue, motoric versus pain-related nerve tissue areas in the spinal cord,...). This type of information has particular value e.g. in the search for anatomical region-specific biomarkers, for the study of local biochemical phenomena, and for the hypothesis-driven study of pathomechanisms. Therefore, a multivariate approach is developed based on a least squares argument. It can be solved efficiently as a convex programming problem and has numerical properties that greatly enhance its applicability to large MSI data sets. The methods can handle purely synthetic queries in a binary or more nuanced gray-level format as well as queries based on a microscopic picture of the measured tissue section. Although the examples presented in this paper were primarily focused on neurobiological tissue, these methods have no inherent link to a certain type of sample and can be put to use in any setting where MSI has value.

CHAPTER 5

Enabling the Exploration of Large MSI Experiments Through Wavelets

Problems worthy of attack
prove their worth
by hitting back.

Piet Hein

This chapter is based on the following publication: • Van de Plas R. et al., "Discrete Wavelet Transform-based Multivariate Exploration of Tissue via Imaging Mass Spectrometry", Proc. of the 23rd Annual ACM Symposium on Applied Computing, 2008.

Mass spectral imaging is a high-throughput technique that mines massive amounts of measurements from a single tissue section. As various parameters such as the covered tissue surface area, the spatial resolution, and the extent of the mass range grow, MSI data sets rapidly become very large, making analysis from a computational and memory standpoint increasingly difficult. The overall goal of this chapter is to provide a strategy for those situations where the multivariate analysis of a MSI data set becomes either (i) infeasible due to hard constraints on resources such as memory, (ii) impractical due to exorbitant calculation times or scalability of the algorithm implementation, or (iii) unreliable due to the 'curse of dimensionality'. As the volume and dimensionality of MSI data keeps growing with new instrumental developments and wet lab improvements, almost all of the more advanced data mining methods currently

gaining ground in MSI, run the risk of falling victim to a bottleneck in available computational resources. Some examples of methods at risk include clustering methods aimed at digitally 'staining' the tissue [78, 80, 93], trend detection methods such as principal component analysis [75, 78], and supervised methods aimed at answering specific biological questions (e.g. spatial querying [94]).

In answer to this challenge, this chapter introduces the discrete wavelet transform (DWT) as a means of reducing the dimensionality of the data, while retaining a maximum amount of biochemical information. More precisely, performing the DWT allows for a more compact description of each mass spectrum, expressed as wavelet coefficients. The proposed strategy is to transform the MSI data from its original input space to the (much) lower-dimensional wavelet coefficient space, and to perform the multivariate analyses in the wavelet space. Figure 5.1 gives a schematic overview of the strategy. Although the transform is lossy, our examples will show that most of the biochemically relevant information is retained. As a result, performing a multivariate analysis in the wavelet space usually gives results that are very similar if not identical to the results obtained through direct analysis of the raw data. The major advantage of going to wavelet space however, is that much less memory and calculation time will be required to obtain identical results. The DWT of MSI data can therefore be seen as a compression scheme that removes as much 'irrelevant' information from the data as possible prior to analysis, thereby enabling the processing of otherwise 'problematic' data sets.

Besides the resource rationale, another motivation for going to the wavelet space is that the transformation generally denoises the data in the process. This way, going to the wavelet space can also be seen as a traditional preprocessing step, aimed at removing noise from the measurements and helping the multivariate analysis to obtain more robust results. A third motivation is the dimensionality reduction that DWT entails. High-dimensional data such as MSI data, contains so many variables that to learn a concept or trend with confidence from the data requires an enormous amount of observations. This 'curse of dimensionality' [95, 96] can be avoided by reducing the number of dimensions prior to analysis, for example with DWT. A final motivation is the scalability of the analysis algorithm that needs to operate on the data. These algorithms are not necessarily designed to operate on very large data volumes, and might become practically unusable as a result. These final three arguments demonstrate that even in cases where resources are not a problem, there is usually some benefit to be gained from performing DWT on the data prior to the analysis.

The efficacy of performing analyses in the DWT-reduced space is illustrated using two MSI case studies. The first one assesses the unsupervised trend detection via principal component analysis (PCA) on a MALDI-TOF based MSI measurement of a sagittal section of mouse brain. The second case study demonstrates the true potential of the technique by performing clustering on a MALDI-FTICR based imaging data set on

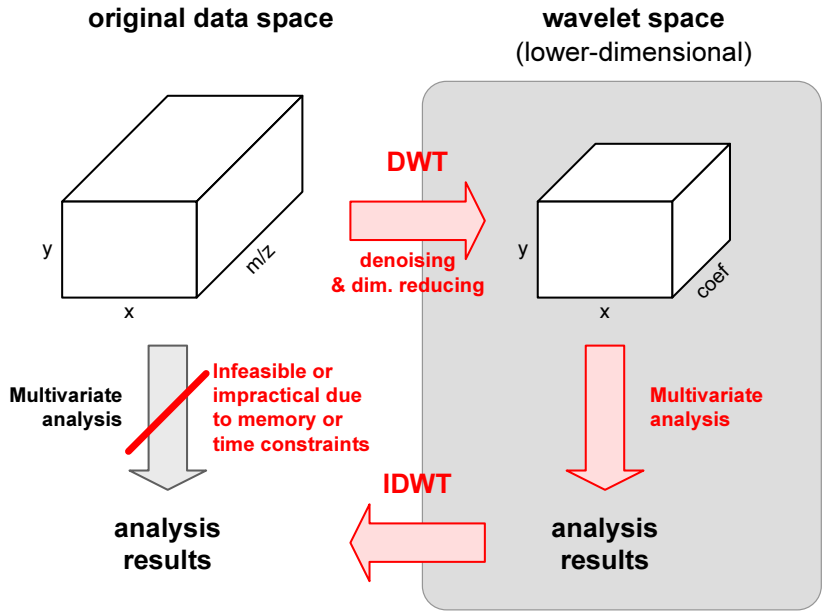


Figure 5.1. Overview of the DWT-enabled analysis of MSI data. Transforming the MSI data to the lower-dimensional wavelet coefficient space side-steps some of the difficulties that can arise with direct analysis of the raw data. Performing DWT can help tackle practical issues (e.g. hard constraints on memory and computational resources) as well as analytical issues (e.g. high-dimensionality of the data mining problem, and the scalability of the algorithm implementation), and can help reduce the presence of noise.

tissue from rat testes. This second data set particularly highlights the potential of wavelet-transformed MSI, as it enables the analysis of a very large data set that would otherwise be very difficult, if not impossible, to process at all.

5.1. Scaling issues

The size of a MSI data set is primarily influenced by two parameters: the number of measurement locations or pixels and the number of scanned m/z bins. The pixel number increases as the tissue surface area that needs to be covered grows larger or the spatial resolution becomes higher. The number of m/z bins is proportional to the extent of the mass range and the granularity of the mass resolution. As this technology develops, both parameters get pushed upward and data sets appear that stretch the computational and memory resources generally available, or challenge the scalability of the algorithms that operate on them.

As an example, let's consider a MSI experiment with a 130×180 measurement grid

(23400 pixels) and a mass range extending from 3000 to 28000 Dalton at 1 m/z-bin per Dalton (25000 bins). From a data matrix of this size, PCA would require a covariance matrix of at least size 23400×23400. At a precision of 8 byte, this means more than 4.3 GB required memory, which is beyond the addressing capability of 32-bit processors. Additionally, calculation times rise as the number of features grows, making processing potentially impractical. A third issue is the complexity of the analysis algorithms and the scalability of their implementations that comes into play. Due to its tensor properties and the 'curse of dimensionality', MSI has, more so than standard mass spectrometry, a need for strong dimensionality reduction methods with minimal loss of biochemical information.

5.2. Methods

In response to the issues raised in section 5.1, this chapter introduces the discrete wavelet transform (DWT) as a method for transforming high-dimensional MSI data into a lesser-dimensional space with most relevant information left intact. This section first explains the basics behind wavelet analysis, and then examines the particular discrete wavelet variant employed in this study. Finally, the details of executing analyses in the reduced space are discussed more closely using basic trend detection via PCA as an example analysis.

5.2.1. Wavelet Analysis

Wavelet analysis [97] is a widely used signal processing method that employs a probing function, known as a wavelet, to analyze a given signal. Similar to the use of sinusoid probing functions in Fourier analysis, wavelets of different scale are compared to the signal under study. However, unlike the sinusoids in Fourier analysis, which have infinite length, wavelets are waveforms of limited duration that have an average value of zero. This means they can be shifted along the time axis of the signal to measure the correlation with the wavelet at that particular time step. This adds a step dimension to the analysis, which is of particular interest for studying nonstationary signals that change significantly over time through aspects such as drifting baselines and peaks. Essentially, wavelet analysis decomposes a signal into shifted and scaled versions of a user-specified mother wavelet.

This chapter applies wavelet analysis to mass spectra, which are inherently nonstationary in that their primary information comes from larger or smaller peaks showing up at particular positions along its m/z axis[1]. A number of studies have employed wavelet

[1] Shifting along the time steps of a signal is common nomenclature in wavelet texts. However, a mass spectrum is a one dimensional signal that has no time dimension but a mass-to-charge dimension instead,

analysis to standard mass spectrometry for the purposes of denoising and detection of localized phenomena such as peaks [98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112], but its use in mass spectral imaging remains under-exploited.

5.2.2. Discrete Wavelet Transform

A discrete wavelet transform (DWT) [113] is a form of wavelet analysis where the scales and positions are chosen to be powers of two, which allows for an efficient implementation. In applying DWT, a signal s of length N is separated into an approximation signal a carrying the high-scale (low-frequency) parts of the signal, and a detail signal d carrying the low-scale (high-frequency) parts. Both a and d have the same length N as the original vector s, but via downsampling DWT only requires to keep $N/2$ wavelet coefficients, c_a and c_d, for each to be reconstructed. The DWT coefficients c_a and c_d can be considered a more compact description of the signals a and d.

The inverse operation to reconstruct a signal from its coefficients is known as the inverse discrete wavelet transform (IDWT). If the IDWT is provided the approximation coefficient vector c_a and the detail coefficient vector c_d (both of length $N/2$), it will reconstruct the original signal s (of length N). However, if the compact description of the details is left out by setting the elements of c_d to zero, only the approximation coefficients will be used and IDWT will reconstruct the approximation signal a.

DWT can be repeated on successive approximation signals to perform a multiple-level decomposition. On every level of the decomposition a detail signal is split off and the number of retained coefficients is halved (approximately, due to padding and convolution aspects). Theoretically, the decomposition can be repeated until only one coefficient is retained, but in practice the number of levels is user-specified according to the goal of the application. Fig. 5.2 shows the wavelet decomposition tree from a multiple-level DWT of a single mass spectrum from the case study of section 5.3. A wavelet from the Daubechies family was chosen ("db5", see Fig. 5.2) and the decomposition depth was three levels.

One aspect of the DWT-IDWT-operation that needs to be watched, is that it is a shift-variant [114] process. This means that as detail coefficients are left out, the reconstructed signal will shift slightly along the m/z-axis. In most cases however this will present no problem as the overall goal is to group pixels along the spatial domain. The mass aspects of the results need not be explicitly reconstructed from the m/z coefficients, but can be extracted from the original input space, thus avoiding the shift-variant problem. If reconstruction via IDWT is absolutely necessary, the amount to

which shows the amount of ions counted at a particular m/z window. From this point onward, it is implied that the shifting of wavelets takes place along the m/z steps in a mass spectral signal.

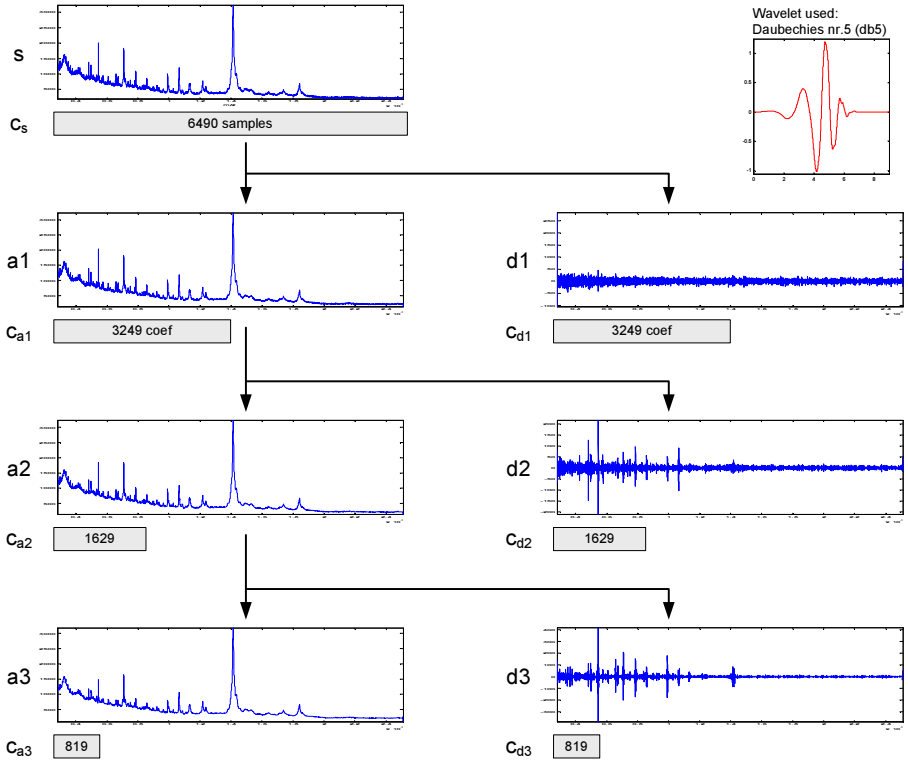


Figure 5.2. A multiple-level DWT performed on a mass spectrum from the case study of section 5.3. The chosen wavelet is db5 (top-right) and the decomposition tree has a depth of three levels. The profile of the original mass spectrum s is well preserved in approximations $a1$ to $a3$, while the required coefficients describing the waveform have been reduced from 6490 to 819. There is some loss of peak height due to narrow tops being filtered out as detail signals (see $d2$ and $d3$), but the general pattern is well preserved and becomes less noisy.

which this aliasing behavior is acceptable is dependent on the application and the mass precision required, but evidently it should be kept to a minimum (see section 5.2.3) by using the number of DWT levels sparingly.

5.2.3. DWT-based Decomposition of MSI Data

The first step is to transform the data from the high-dimensional m/z-space to a lower-dimensional space expressed in wavelet coefficients, while keeping a maximum of the information intact. To this end, a multiple-level DWT with a user-specified level-depth and wavelet, is performed on each individual mass spectrum in the data set. The ion counts describing the profile of each mass spectrum are replaced by a more compact description using only the approximation coefficients. The original data tensor (with

an x, y, and m/z-mode) is replaced by a smaller denoised three-way array (with an x, y, and coef-mode).

Here we will use standard linear PCA [67] as an example of an analysis that could benefit from being performed in wavelet space. Of course other multivariate methods such as clustering can be used in a similar fashion. PCA is performed on the reduced tensor, making the calculation less memory and computationally intensive. Following the setup of [75], the array is decomposed into a number of principal components, characterized by a spatial distribution, an amount of variance explained, but also by a signature in the coefficient domain rather than the m/z-domain. The components' score vectors are expressed in terms of correlated wavelet coefficients instead of correlated m/z bins or masses.

If the objective of the analysis is to get insight into the spatial distributions apparent in the data, these components are ready for interpretation in their current form. If an interpretation of the masses responsible is required, a third step can be performed where the score vectors are translated back into m/z-terms using the IDWT. Fig. 5.3 shows an overview of the procedure, and related approaches formulated in other fields are available in [115, 116].

One thing to keep in mind is that this operation performs a form of lossy compression and denoising by removing the detail signals from the analysis. The influence of the lost information is generally negligible compared to other noise sources, but it can be controlled through the choice of wavelet and the specified level-depth. This, in addition to the shift-variant behavior mentioned in section 5.2.2, implies that for best results the decomposition depth should be restricted to the minimum depth necessary to make the analysis practical or feasible within a certain set of resources.

5.3. Case Study on a MALDI-TOF imaging experiment: Classical PCA versus DWT+PCA

To establish an empirical test case, this section applies both the DWT-based method as well as direct PCA to the MSI measurement of a mouse brain section. The main goal is an assessment on a real example of how well the information in the data set is retained, given a serious reduction in dimensionality. Additionally, the example also demonstrates whether factor analysis methods such as PCA, performed in the coefficients-space, yield results that approximate those obtained directly from the original m/z-space. The details are provided in section 5.3.1 and section 5.3.2 elaborates on the findings.

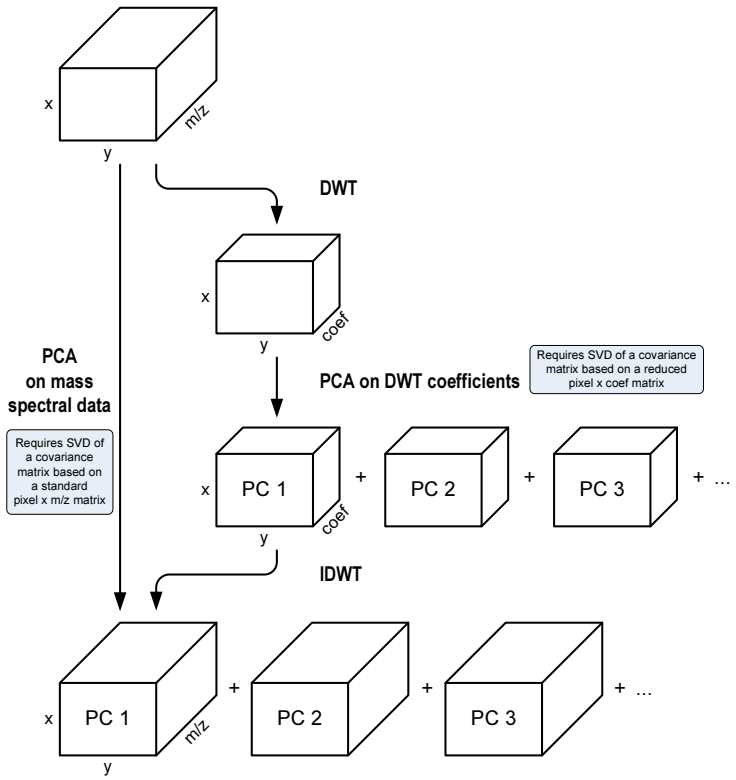


Figure 5.3. Overview of the DWT-based PCA decomposition of MSI data. The procedure involves three steps: DWT is used to reduce the feature space, PCA is performed in the reduced space. If required, IDWT can be performed on the feature signatures of the principal components to facilitate mass spectral interpretation of them.

5.3.1. Materials & Methods

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

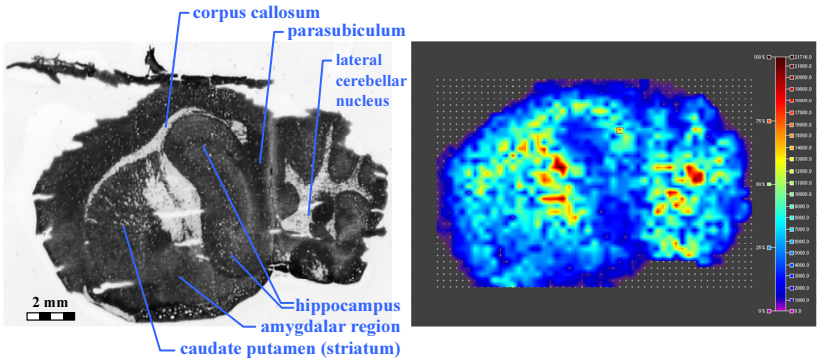


Figure 5.4. (left) Picture of the sagittal mouse brain section imaged in section 5.3. (right) Ion image showing the presence of m/z 14148 primarily in the corpus callosum and the lateral cerebellar nucleus.

results, we disregarded them when their total ion current fell below a 10% threshold. This essentially selects only the on-tissue measurements.

5.3.2. Results & Discussion

Fig. 5.5 shows the first three principal components (PCs) derived from direct application of PCA. The first PC shows a pattern found throughout the tissue section, corresponding to a general signature for brain nerve tissue. Basic brain cell chemistry is present everywhere in the tissue, and as expected it explains 91.8% of all variance in the data. It is dominated by masses at 4980, 5461, and 14150 Da[2]. There is a slightly dimished peak intensity in the central *caudate putamen* region, the top-to-left *corpus callosum*, and the lobe-like extensions around the *lateral cerebellar nucleus* (for guidance see Fig. 5.4), which is probably the result of a dimished ionization efficiency in these areas due to particular chemical composition. PC 2 accounts for 3.1% of the information in the data set and differentiates the *hippocampus* and the bottom *amygdalar* region from the rest of the tissue. The difference lies primarily in the significantly higher ion counts for the 4980 Da mass with a corresponding dimished appearance of m/z 14150. The third PC explains 2.0% of the data and delineates a number of known areas in the brain, primarily due to a higher expression around 14140 Da, 15320 Da, and a peak around 18420 Da. PC 3 describes from right to left the *lateral cerebellar nucleus* and its extensions (which are also clearly visible on the picture of Fig. 5.4), the *parasubiculum* which is not really apparent on the visual picture, the *corpus callosum*, and the *caudate putamen*. Also notice the significant absence of the *hippocampus* in PC3. These results illustrate MSI's use for chemically

[2]MALDI-based ionization typically results in ions with a charge z equal to one, which is why we equate and use interchangeably the concepts of mass-to-charge and molecular mass in this discussion.

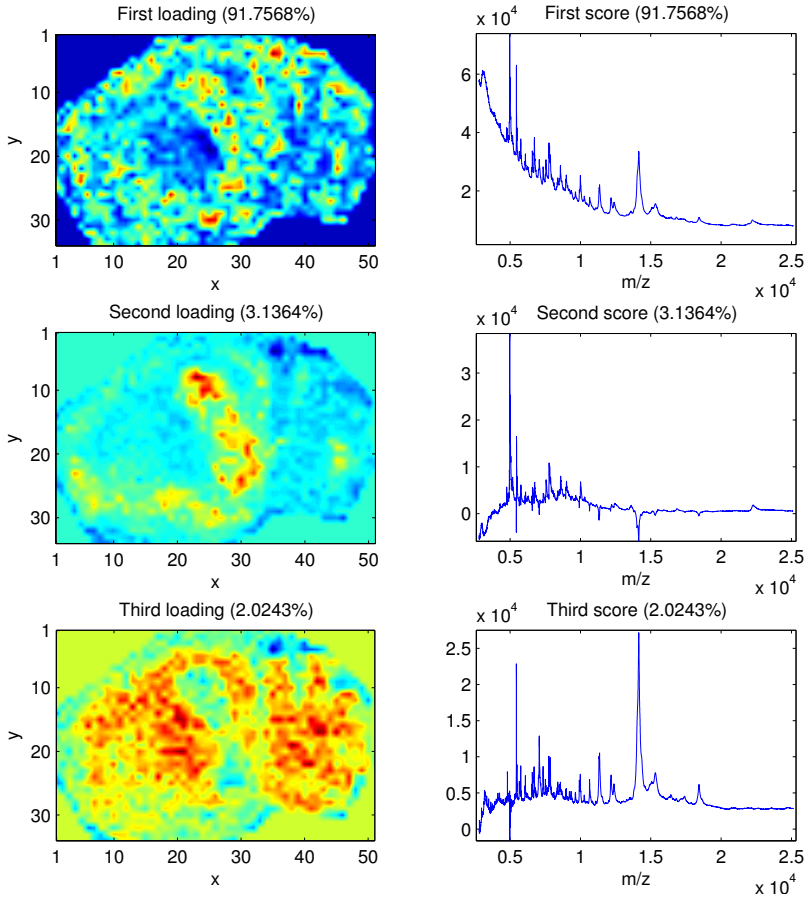


Figure 5.5. Non-DWT-based principal components. These are the three most important PCs found via direct application of PCA. They chemically delineate a number of known anatomical zones in the mouse brain (see section 5.3.2).

delineating homogeneous regions in tissue, be it visually apparent ones (e.g. *corpus callosum*) or others (e.g. *parasubiculum*).

For the trend detection in reduced space, we applied, similar to Fig. 5.2, a DWT using wavelet db5 and a depth of three levels. The features were reduced from 6490 to 819, compressing the data tensor by a factor of approximately eight to size $51 \times 34 \times 819$. PCA was applied and the three primary principal components are shown in Fig. 5.6. The components turn out to be almost identical. There are small differences in the percentages of variance explained, a small reduction in the loading vector of PC 2 (manifested as decreased blue), and dimished peak intensities consistent with the removal of detail signals. This seems to indicate that although the chosen parameters

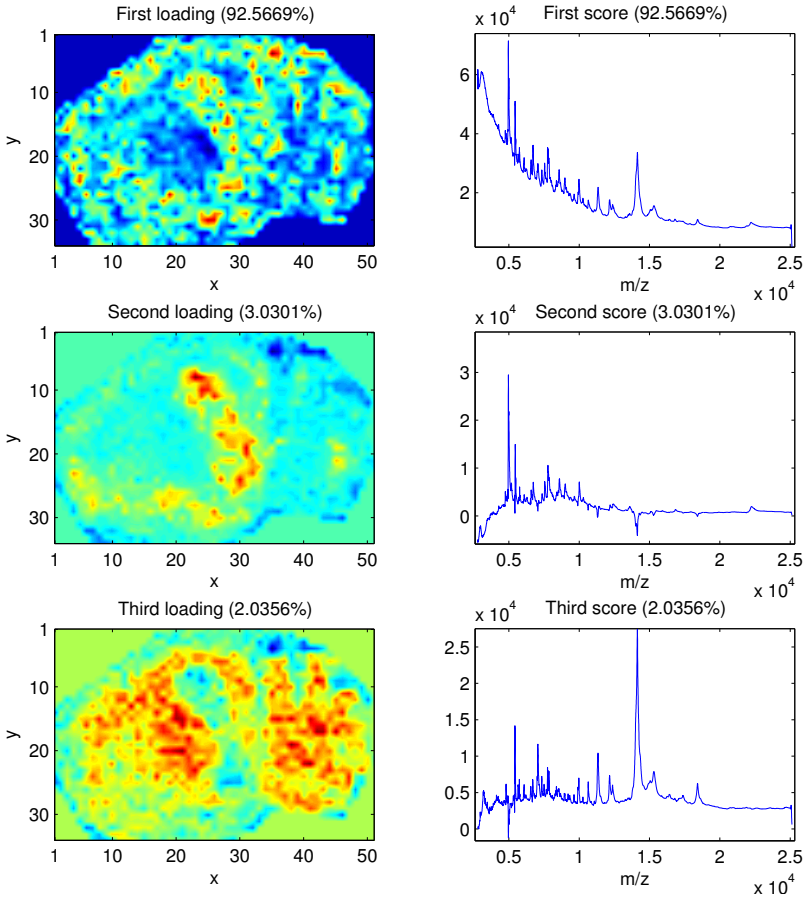


Figure 5.6. DWT-based principal components. These are the three primary PCs found via the method of section 5.2.3, using the DWT for dimensionality reduction. Although PCA was applied in the lower-dimensional coefficients-space, comparison with Fig. 5.5 indicates that the decomposition results are almost identical.

(db5 and three levels deep) produced a substantial reduction in required resources, they are still far from the critical values at which they would start to push the interpretation astray through loss of information.

For the PCA shown above, we chose to employ the pixel $\times$ pixel covariance matrix as the basis (see the "economized" approach in [75]). As a result, the dimensionality reduction manifests itself as a calculation time reduction, because there are less features to be used in the calculation of each covariance entry. To verify this, calculation times for the various steps were recorded on an Intel Celeron processor at 1400 MHz with 2 GB of RAM, shown in Table 5.1. As expected, the combined calculation

time of the DWT and the reduced PCA (13.2 s + 75.6 s) is less than performing PCA directly (158.3 s). **In other words, performing the analysis in wavelet coefficient space produced a 44% reduction in required calculation time.** An interesting observation is that the IDWT-operation on the coefficient signatures takes rather long, potentially losing the time advantage that was gained earlier. However, this is of no consequence as IDWT will not need to be performed in most MSI analysis situations. The application of DWT and IDWT does not influence the findings along the spatial domain, only those along the mass domain. This means that learning a relationship between pixels in wavelet space is directly transferable to the original space, without the need to do any back-transformation. In most cases, knowing which pixels belong together is sufficient to learn their mass characteristics directly from the original mass spectra, and thus there will be no need to complete the round trip to wavelet space by employing the IDWT. IDWT can thus be dropped from the comparison. It would be sufficient to retrieve only the spatial signatures after DWT and PCA, and to use them to obtain the score vectors directly from the original data tensor (which additionally would avoid the shift-variant behavior as well). Another other option for applying PCA is to base it on the features $\times$ features covariance matrix. In that case the dimensionality reduction would primarily manifest itself as a memory advantage (Table 5.2), needing only a 819 $\times$ 819 matrix rather than a 6490 $\times$ 6490 matrix. This result highlights the true memory advantage that DWT can provide for very large MSI data sets. **In this example we obtain interpretation-wise identical PCA results by only using 1.5 percent of the memory that would normally be required (5.3 MB versus 337 MB): a 63-fold reduction in memory requirements.** In short, there are various configurations possible which deliver different types of advantages. Which is the most advantegeous depends on the size and type of resources available and on the particular *x*, *y*, and m/z-sizes of the case under study.

Table 5.1. Case study resource specifics, based on using a pixel$\times$pixel covariance matrix for PCA.

	Non-DWT PCA	DWT PCA	DWT PCA IDWT
size cov. matrix	pixels$\times$pixels	pixels$\times$pixels	pixels$\times$pixels
required memory	1734$\times$1734 (24MB)	1734$\times$1734 (24MB)	1734$\times$1734 (24MB)
contrib./cov. entry	# m/z bins	# coefficients	# coefficients
required calc. time	158.3 s	13.2+75.6 s = 88.8 s	13.2+75.6+250.0 s = 338.8 s
	(PCA)	(DWT) + (PCA)	(DWT) + (PCA) + (IDWT)

5.4. Case Study on a MALDI-FTICR imaging experiment: Enabling Clustering

The intent of this second case study is to demonstrate the wavelet strategy on a data set that, due to its size, would otherwise be very difficult to analyze with multivariate

Table 5.2. Case study resource specifics, based on using a features×features covariance matrix for PCA.

	Non-DWT PCA	DWT PCA
size cov. matrix	m/z bins×m/z bins	coef×coef
required memory	6490×6490 (337MB)	819×819 (5.3MB)
contrib./cov. entry	# pixels	# pixels

methods. After having compared the performance of the DWT approach versus the classical approach in the previous section (using a data set that still allowed for classical analysis), this case study moves squarely into an area that is beyond most classical methods.

Most of the chapter has focused on PCA as an example of a typical multivariate analysis one might want to perform on a MSI data set. Here we diversify this notion of a multivariate analysis and show that this wavelet strategy is not tied to one particular analysis type. To this end, we perform an operation increasingly common to MSI, which is the clustering of pixels on the basis of their chemical content. Uses for this have been reported for example in relation to segmentation of the tissue and the assessment of correlated expression [78, 80, 117]. In this example, we choose to use the basic and established *k-means* clustering method [118], to make clear that the results are due to the wavelet strategy and not tied to the intricacies of a customized clustering algorithm. The overall goal of the analysis is to have the algorithm derive a set of zones from the tissue that exhibit common chemical content. The basic setup is shown in Fig. 5.7.

5.4.1. Materials & Methods

The case study describes a mass spectral imaging experiment performed on a section of rat testis (*epidymis*), with the particular aim of studying the maturation of spermatocytes through the distribution of phospholipids. The tissue section was coated with an alpha-CHCA matrix using the ImagePrep station from Bruker Daltonics. The measurements were done in positive mode using a 9.4 Tesla ApexUltra FTICR-MS from Bruker, with a DualSource and Smartbeam laser (laser spot diameter of 45 µm). The spatial resolution is 50 µm inter-spot distance, and the MSI grid consists of 86 measurement locations along the x axis and 66 positions along the y axis. The experiment thus describes 5676 individual pixels and mass spectra. A microscopic image of the tissue section together with the MSI measured area (in green) are shown in Fig. 5.8. The scanned mass range runs from m/z 385 to 3000. The employed FT-ICR mass analyzer provides some of the highest resolving power currently available in mass spectrometry (a mass resolution larger than 100000). As a result, the scanned

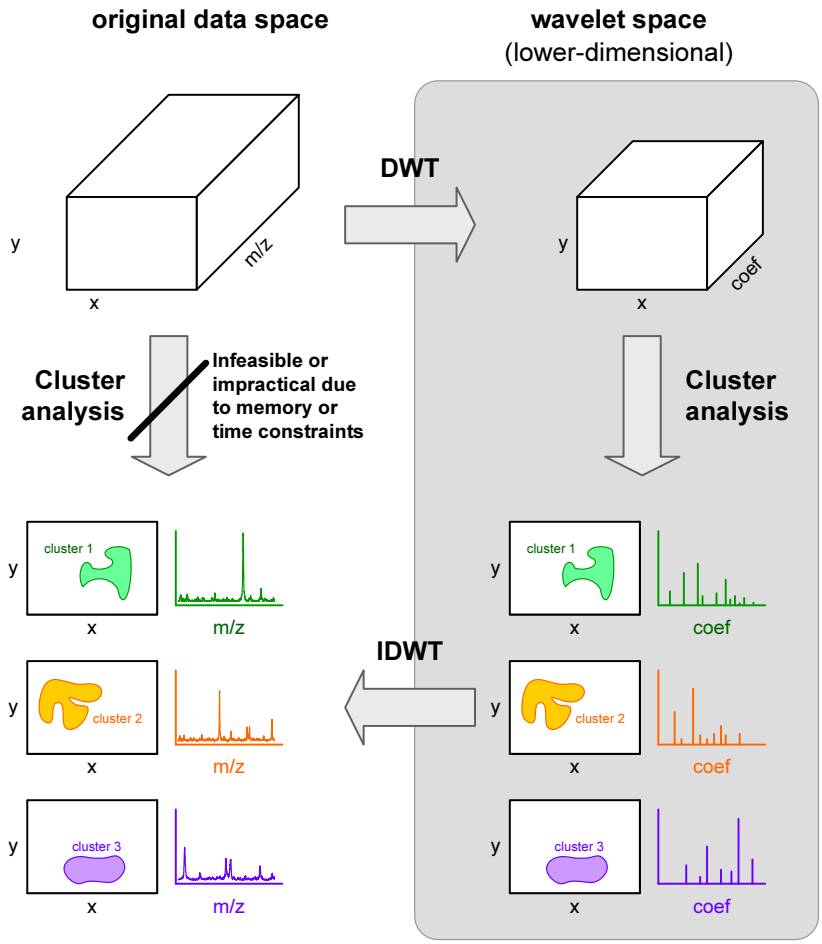


Figure 5.7. Clustering via the lower-dimensional wavelet coefficient space. Notice that the cluster partitions obtained along the spatial domain in wavelet space need not be explicitly transformed back to the original input space. They are immediately available in their desired form because the DWT only transforms along the mass domain, not the spatial domain.

mass range is sampled at a rate of 457035 individual mass bins per spectrum. The FT-ICR MSI data set therefore consists of 5676 $\times$ 457035 recorded values, or a total of 2,594,130,660 individual numbers. At a precision of 8 byte (double precision), the raw data requires 20,753,045,280 bytes or 19.3 GB of storage or memory. Data analysis and preprocessing (baseline removal) was done using in-house developed software and MATLAB (The MathWorks, Natick MA, USA).

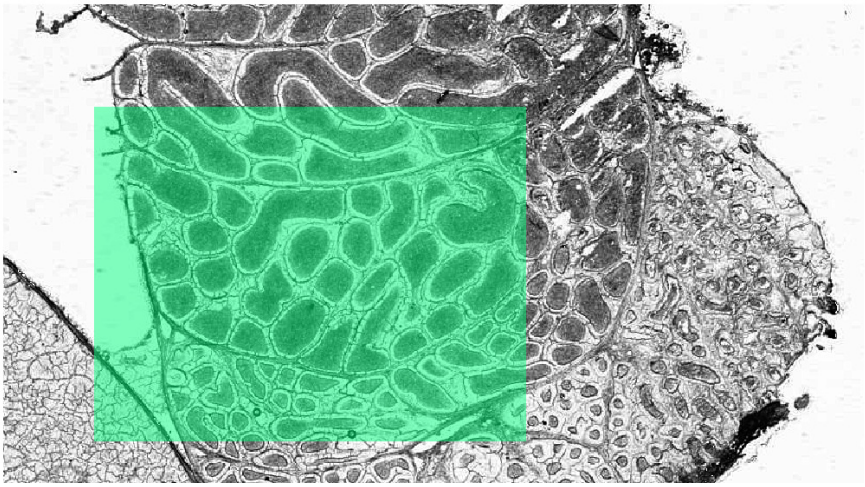


Figure 5.8. The MALDI-FTICR imaging data set is acquired from a section of rat testes. The MSI measurement area is indicated above in green on a microscopic image of the tissue slice. The grid consists of 86 by 66 measurement location, acquired at 50 µm intervals. Due to the high mass resolution that can be attained with a FT-ICR mass analyzer, each spectrum describes 457035 individual m/z bins in a mass range from m/z 300 to 3000. In total, the experiment collects 5676 $\times$ 457035 values, which at *double* precision (8 bytes per value) amounts to 19.3 GB of raw data.

5.4.2. Results & Discussion

The application of a clustering method directly on the (raw or preprocessed) data, has little use in this case. Although the k-means algorithm does not preclude its application to large data sets per se, the difficulties with doing so often manifest themselves at the level of the distance measure that is being used. Being presented with so many variables, many distance measures become either unworkable because their calculation requires too much time, or they get misguided by the multitude of variables along which almost any partition might be motivated. Besides the difficulty of finding a suitable distance measure that still gives reliable results in very high-dimensional spaces, a second problem is that most clustering techniques require the distance to be calculated for each possible pair of spectra in the data set. This usually amounts to $n(n-1)/2$ calculations of the distance measure, if the data set consists of n spectra and the distance measure is symmetric. Although this n is manageable for most standard mass spectrometry studies, it is clear that in a MSI context n can run into the thousands, thus increasing the amount of required calculation by several orders of magnitude. It is often the combination of the amount of spectra and the calculation time of the distance measure that make a full clustering effort infeasible on large data sets. In our FT-ICR example however, the reason for not being able to perform the clustering on the original data is more conceptual. Because the spectra have lots of

mass bins that have a value close to zero (more than 95% of the m/z range), most distance measures see little or no difference between any of the spectra, and conclude that all spectra report the same content. Although this could be remedied by more advanced or customized distance measures, it still illustrates that many of the traditional approaches are ill-equiped to handle very high-dimensional data spaces without modification. Short of an effort to construct customized clustering algorithms for MSI data, our best option is to make a traditional analysis feasible by performing some form of dimensionality reduction.

For this reason, we employ the same DWT strategy as earlier, transforming the data to wavelet coefficient space and performing the clustering analysis in this reduced space. The DWT transformation used Daubechies' *db3* wavelet as a basis, and decomposed the spectra seven levels deep. This entails roughly a 2^7-fold dimensionality reduction. To be precise, the data went from 457035 variables per pixel to only 3575 variables per pixel. It is on this reduced data set of 5676×3575 that we perform k-means clustering. The k-means algorithm was set to look for six clusters in the data, while employing the *cosine* distance measure and repeating the process through ten replicate runs with random start initialization. Although the formulation of a good distance measure for mass spectrometry data is a research topic in its own right, reasonably good results have been obtained with the classical measure called the *cosine* distance. Actually, it is a measure of similarity between two vectors by calculating the cosine of the angle between them. The cosine similarity is then transformed into a distance measure by taking one minus the cosine. The reason for its good performance in mass spectrometry is that it primarily compares relative profiles and is largely independent of the absolute amplitudes of the variables (or peaks). Thus it is able to avoid being influenced disproportionately by multiplicative ionization efficiency noise (see also chapter 2). Here, the cosine distance is applied to wavelet coefficients rather than ion counts, but because the coefficients mirror many of the amplitude characteristics of the ion peaks, the same reasoning behind its use holds. A summary of the results of the clustering analysis are shown in Fig. 5.9, which depicts every cluster of pixels with a different color. Pixels that have the same color, also share similar molecular content according to the clustering algorithm. The individual clusters, together with their average mass spectrum over the clustered tissue area, are shown separately in Fig. 5.10.

Unlike the case study of section 5.3, here there is no reference against which to weigh the performance of the algorithm. However, the clustering results cause pixels to group together nicely in the spatial domain into areas that match well with the structures seen in the microscopic image. As there was never any spatial information incorporated into the analysis in the first place, this observation gives evidence to the validity of the results. By overlaying the results on a microscopic picture of the tissue, it becomes apparent that the clusters clearly correlate with visible structures in the section. The 'tubules' shown in the picture are actually one long, tightly coiled tubule cut

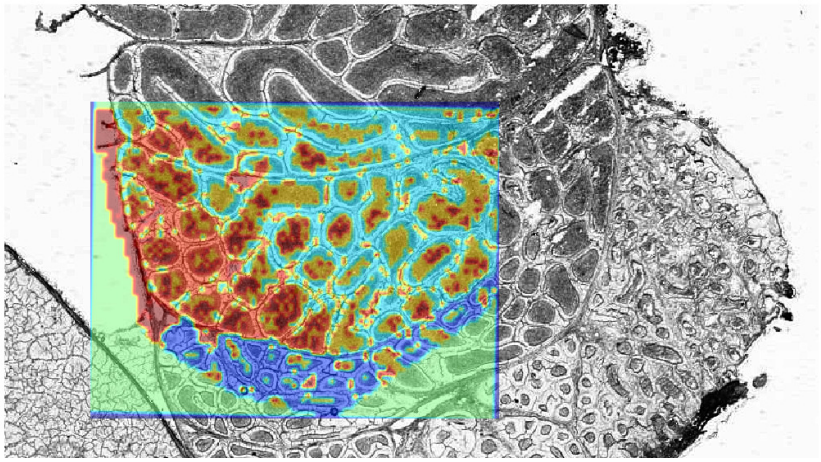
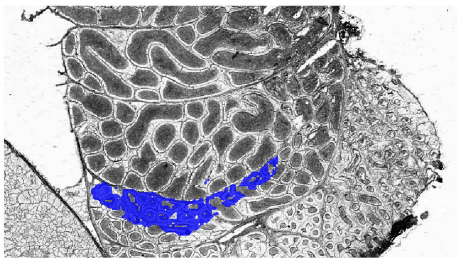


Figure 5.9. A summary of the clustering results. Each cluster is indicated as a different color, grouping together pixels that according to the analysis exhibit a similar chemical content. This image gives an idea of the variation in molecular content across the tissue section. The clusters are also shown separately in Fig. 5.10, together with the average mass spectrum for each cluster.

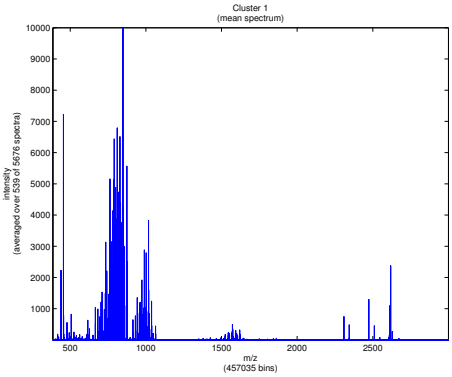
at several places. Cluster 4, which is indicated in yellow, seems to specifically select the inside of this tubule, while cluster 3 (in green) selects areas where no molecular content was found at all[3] (notice the low ion counts for cluster 3 versus the others). Cluster 1 firmly selects the bottom area of the tubular walls, while 2 and 5 delineate primarily wall structures in respectively the right and left hand side of the tissue. Presumably these differences are related to the differences observed between the 'head' and the 'tail' of this tubule structure. Cluster 6 rounds up the list by selecting ions that show up internally to the tubule (but only the bottom left hand side), and also seem to accumulate outside the *epidymis* structure.

The main message to take away from this case study is that by using the wavelet strategy discussed in section 5.2.3, it becomes possible to analyze data sets that would otherwise be impractical or infeasible to process. This ability will presumably only rise in importance as data sets keep growing in response to more advanced demands from the biology, and pushed by new instrumental developments.

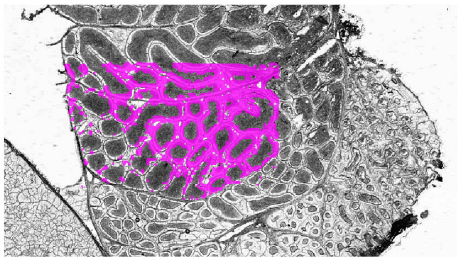
[3]The reason cluster 3 indicates no findings also in areas where there is clearly tissue in the picture, is that the acquisition software from Bruker was set to collect measurements in a user-specified polygon shape rather than a rectangular grid, and thus these locations were never measured. This is also evident from the straight lines that border this cluster.



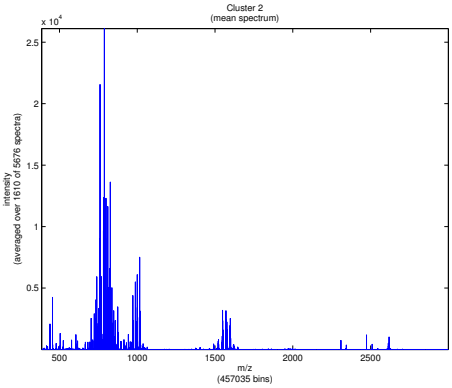
(a) Cluster 1



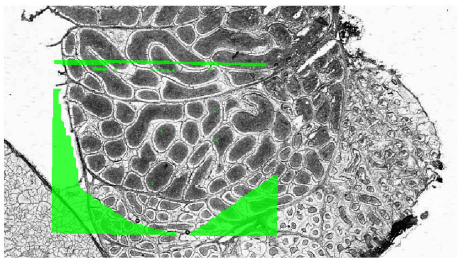
(b) Cluster 1 – average mass spectrum



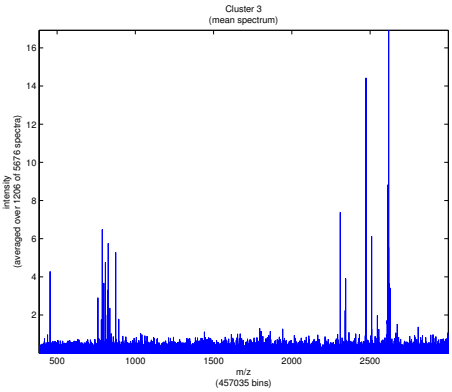
(c) Cluster 2



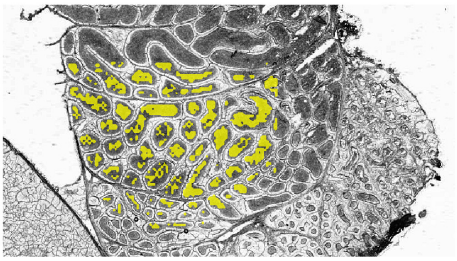
(d) Cluster 2 – average mass spectrum



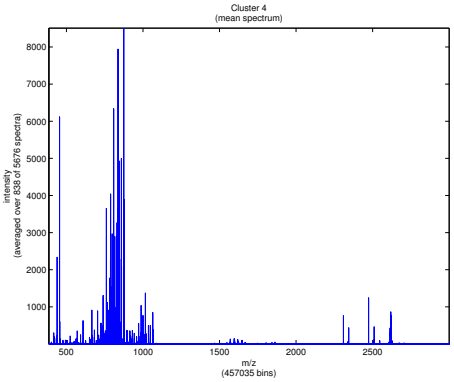
(e) Cluster 3



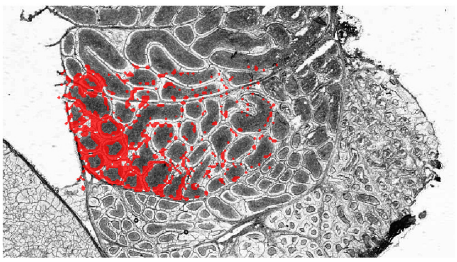
(f) Cluster 3 – average mass spectrum



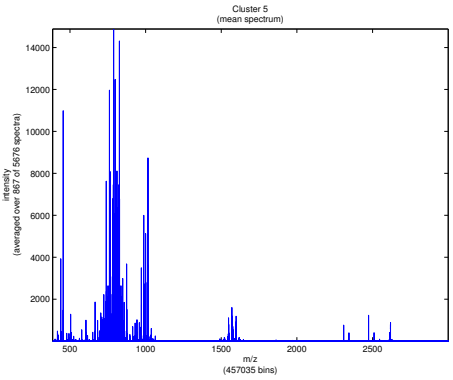
(g) Cluster 4



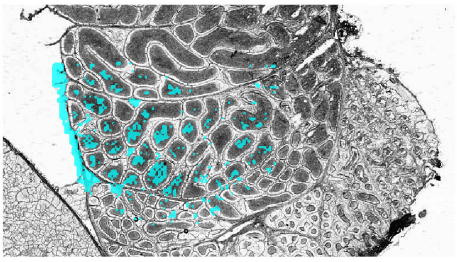
(h) Cluster 4 – average mass spectrum



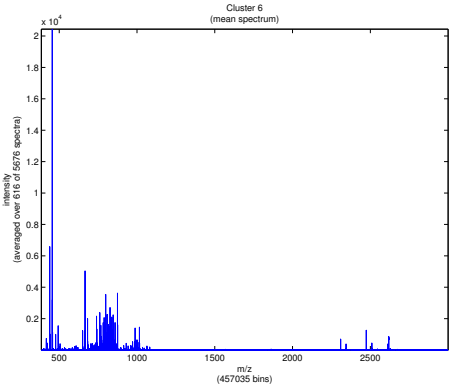
(i) Cluster 5



(j) Cluster 5 – average mass spectrum



(k) Cluster 6



(l) Cluster 6 – average mass spectrum

Figure 5.10. Clustering results.

5.5. Conclusion

The excellent compression characteristics of the DWT provide us with a means of dimensionality reduction that retains the mass spectral information to a high degree while offering significant reductions in resource requirements. In this chapter we have shown specifically that resource-hungry operations, such as biochemical trend detection via PCA or clustering, can be performed directly on the compacter description of the data with meaningful biochemical results that deviate little from non-DWT-reduced results. The approach has calculation time reducing, noise removing, and memory requirements dimishing aspects that can be useful in any MSI context. Additionally, it holds promise as a tool to make multivariate exploration feasible on very large MSI data sets, that might otherwise surpass available resources.

Acknowledgments

We kindly acknowledge Chris Hendrix, Dagmar Niemeyer, Jens Fuchser, Jens Decker, Michael Becker, and Sören-Oliver Deininger from Bruker Daltonics in Bremen, Germany. We also thank Charles Pineau from the University of Rennes for the section of rat testis.

CHAPTER 6

Co-localization Networks

That's the nature of research –
you don't know what in hell you're doing.

Harold "Doc" Edgerton

This chapter is based on the following paper: • Van de Plas R. et al., "Co-localization Networks: Assessing Inter-molecule Relationships in Tissue", manuscript in preparation.

In introductory chapter 1, it was shown that a mass spectral imaging experiment is essentially a fusion of biochemical expression and spatial information for a particular tissue section. Due to its favorable characteristics of providing label-free imaging for hundreds of molecular species at a time, it has primarily gained attention as a molecular imaging modality. It is therefore not surprising that the majority of the studies employing MSI have done so from an imaging perspective, and have put the focus on the ion images it can produce. This is however only half of the story, and in this chapter we intend to approach MSI from a slightly different perspective. More specifically, this chapter introduces a novel network visualization that depicts the inter-molecule relationships that can be found in a MSI experiment, such as co-localization and correlated expression. The network representation, dubbed a co-localization network, provides putative insight into potential biological or mass spectrometry induced relationships between ions.

Traditional imaging modalities such as microscopy often measure only a few variables per pixel (e.g. red, green, and blue). This means that there is usually little point in analyzing microscopy data along its variable mode, and as a result almost all analysis occurs along its spatial mode in the form of pictures. In MSI the situation differs significantly. MSI routinely procures thousands of variables per pixel, and it can therefore act as a much richer source of biochemical information than any of the traditional modalities. The *in silico* analysis of MSI experiments is thus presented with the unique opportunity to mine this additional data source. In other words, the wealth of information along the biochemical domain causes an inherent duality to be encoded into MSI measurements that allows for both the examination of spatial relationships as well as biochemical relationships. Still, however, almost all MSI studies have focused on approaching MSI data along its spatial mode, and have left the relationships between molecules largely unexplored. We posit that to unlock the full potential of MSI, there will need to be methods that enable analysis from both the spatial mode as well as from the biochemical (molecular mass) mode.

This push, away from univariate approaches and towards multivariate approaches that exploit and explore inter-molecule relationships, has been the driving force behind the majority of developments documented in this thesis. In chapter 3, the focus is on unraveling MSI data into its underlying components by mining correlations along the spatial domain as well as along the mass domain. Chapter 4 provides the means to explicitly look for correlations between masses along the mass domain and a hypothetical expression pattern. Chapter 7 employs the concepts of co-localization and correlated expression to filter out mass spectral signatures tied to a specific anatomical region. Finally, this chapter takes some of the concepts of chapter 3 to step away from the picture-centric view on MSI data, and instead provides an ion-centric view into the tissue by providing a network representation of 'ion-to-ion' relationships. The focus will be on the concept of co-localization, but correlated expression will become part of the analysis as well.

6.1. Why examine ion co-localization in tissue?

The main inter-molecule relationship that will be examined in this chapter is the degree of co-localization between molecules. Co-localization is a concept that is well-studied in areas such as cytology and fluorescence microscopy, and it refers to the degree of overlap between the spatial distributions of two substances. Two molecules are said to co-localize strongly if they consistently show up in the same cell type or tissue area. Co-localization can also be considered as a measure for the spatial correlation between two molecules. Within a MSI context, McDonnell *et al.* touched upon the concept of co-localization with the aim of assessing experiment-to-experiment variability [117]. In this chapter however, the goal is to use co-localization as a guide

along which further exploratory insight into the content of a MSI data set can be developed.

The reason that co-localization is of interest for the interpretation of MSI measurements is two-fold:

1. **biological reasons for co-localization** – Molecules that consistently show up in each others vicinity, often share a biological reason for doing so. Examples are proteins that make up part of a protein complex, or metabolic processing partners such as a protease and its metabolite product. By mapping the co-localization, the aim is to learn something about the possible biological relationships that are to be found in the MSI experiment.

2. **instrumental reasons for co-localization** – During the process of mass spectral measurement, various effects can cause the creation of related ions. Examples include in-source decay, or intentionally induced fragmentation of a mother ion into fragment ions. As these fragment ions stem from the same parent molecule, they will also share spatial expression. Again, co-localization can help in finding which ions out of the thousands that are measured, potentially share a fragmentation relationship. This type of information can be extremely valuable for learning the underlying processes that give rise to the mass spectra we detect.

6.2. Co-localization Network

The concrete goal of this study is to devise a way of going from a MSI data tensor to a more biologically interpretable representation of the mass data. The intent is to use such a representation as a way to guide further analysis according to potential biological or instrumental hypotheses. Given the familiarity of biologists with network representations in areas such as regulatory relationships and protein-protein interactions [119], we propose to use a network visualization to depict the co-localization associations that can be extracted from the MSI experiment. For the remainder of this thesis, this representation will be referred to as a *co-localization network*.

A co-localization network consists of nodes and edges. The nodes represent ions (or masses), found as peaks in the data. The edges represent the co-localization relationship between two nodes. If there is an edge between two nodes, it means that these two ions are considered 'co-localized' according to the co-localization measure that is employed. Once a network is built, it becomes clear which subsets of ions are tightly interconnected, implicitly highlighting which ions show up consistently together in the same areas. If the network is laid out to keep tightly interconnected *cliques*[1] of

[1] A clique is a concept from graph theory. It is a subset of a graph's vertices or nodes such that every two vertices in the subset are connected by an edge

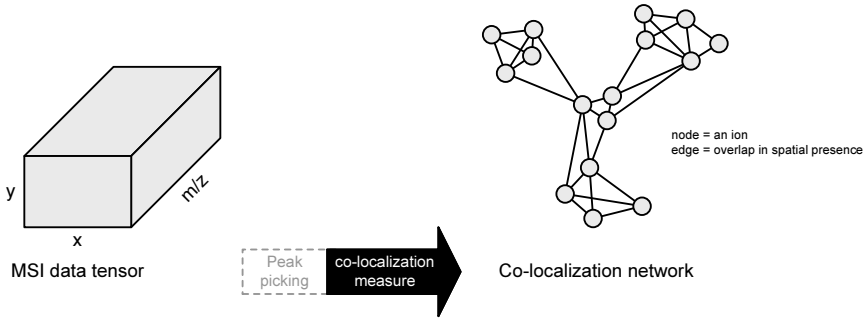


Figure 6.1. Co-localization networks: going from a MSI data tensor to a network representation. The spatial correlations present in the data are mined using a user-specified co-localization measure. These co-localization relationships are then turned into edges on a network of ions found in the data.

ions together, it will also become clear which ions share a 'looser' relationship. This could mean for example an ion A showing up in the same place as ion B, but additionally also showing up in another anatomical area from which B is absent. Overall, the network will try to connect all ions together that share space somewhere in the tissue, and thus it creates a clear representation of which spatial distribution patterns are to be found in the tissue, and which ions are markers for these patterns. By not tying these notions of spatial presence to physical areas in a measurement grid, but rather having a more abstract representation, it also becomes easier to compare spatial distributions and co-localization between different MSI experiments.

Figure 6.1 gives an overview of the basic setup in going towards co-localization networks, and section 6.3 gives an example of its capabilities on a real MSI experiment on mouse brain.

6.3. Case study: An Example on Mouse Brain

With this case study we aim to demonstrate how the availability of a co-localization network for a MSI experiment can help guide the exploration of the data set. First, section 6.3.1 and 6.3.2 will detail the construction of a co-localization network by examining respectively the co-localization measure and the process of going from co-localization distances to a network representation. Then, section 6.3.3 will employ the constructed network to explore some of the ions that seem to share a spatial relationship. Take note that this is but one example of how a co-localization network can be constructed. At various stages during the process, methods can be switched out for different implementations, dependent on the goals of the study at hand.

The case study employs the same MALDI-TOF MSI data set that was used in chapter 4. The tissue under study is a sagittal section of mouse brain, from which a MSI data tensor of size $51 \times 34 \times 6490$ is acquired. Further details on the materials and methods of the experiment can be found in section 4.3.1.

6.3.1. Co-localization Measure

The co-localization measure is an aspect of the network construction process that nicely illustrates the duality that was hinted at in the introductory paragraph. When MSI is approached from the spatial domain (the default approach), pixels are considered observations and the expressions along the various m/z bins are considered variables measured at those observations. When clustering pixels together into tissue areas that exhibit homogeneous molecular content (e.g. in [78, 80]), a distance measure is needed to quantify the correlation between two pixels along the mass domain. If the approach is now switched around to the other mode, ions are considered observations, while their expressions along different tissue locations (pixels) are considered the variables connected to those observations. If we now want to cluster together ions into groups of co-localizing molecules, also a distance measure is needed. This distance measure needs to quantify the correlation between two ions along the spatial domain. It is this category of distance measures that the co-localization measure belongs to.

The co-localization measure that is tied to a network, can essentially be any distance measure that quantifies the amount of co-localization between two ions or molecules. Possible candidates run the gamut from very simple measures (e.g. counting the spatial overlap percentage) to more advanced approaches using correlation coefficients. A co-localization measure can even be customized to highlight particular aspects that might be meaningful to the study at hand. One suggestion is to use an established correlation coefficient such as Kendall's τ or Pearson's ρ to compare two vectors of pixels to each other (also mentioned in [117]). In section 4.2.1 of chapter 4 however, we already mentioned some of the potential problems with such a univariate approach (e.g. a multiple testing problem, data set-specific thresholding,...). In our choice of co-localization measure for the case study we therefore leverage some of the developments done in the unmixing chapter (ch. 3).

For the purpose of this demonstration, we will use co-membership to a component in a rank-12 non-negative matrix factorization of the data as a measure for co-localization. The advantage of piggy-backing on an established matrix decomposition method such as NMF, is that the co-localization measure gains in robustness and can profit from the denoising effect inherent to an unmixing method. Another advantage is that the unmixing of the data set is influenced by co-localization, but also by correlated ion expression. By using a decomposition as the basis for a network-generating measure,

the network inherits both from co-localization as well as from the correlated peak expression, which further strengthens the network edge confidence. In short the measure employed in this case study, can be defined as follows: ion peaks that show up together in the signature of a NMF component, are linked. Some ions will appear in multiple components because their spatial presence extends beyond the boundaries of a single component, and will act as hubs in the network.

6.3.2. From Data Tensor to Network

In the previous section it was already implied that the unsupervised exploration of the co-localization aspect of MSI data can essentially be seen as clustering along the mass domain, while using a distance measure that quantifies spatial correlations between ions. By extension, one can consider a co-localization network to be a user-friendly visualization of such a clustering effort. Going from a MSI data tensor to a network is therefore a two-step procedure:

1. Constructing a distance matrix of ions versus ions, which quantifies the spatial (distribution) similarity between the different ions.
2. Construct a network visualization of the co-localization values that exceed a user-specified threshold.

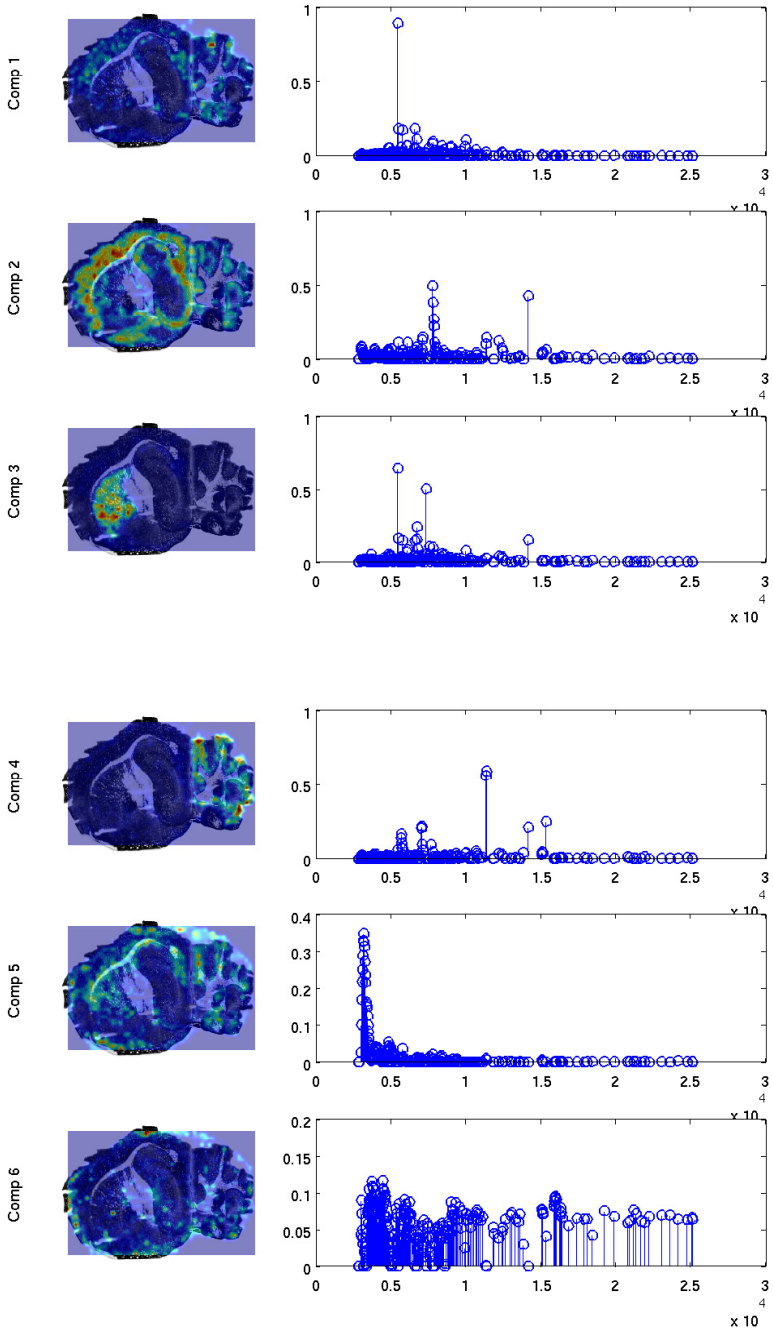
Step 1

In our example, acquiring the relationships between ions comes down to performing a non-negative matrix factorization of the MSI data set and checking component membership. Although NMF is usually performed on the entire mass spectrum rather than a stick-representation of it, there is a need to avoid visually overloading the resulting network. Consequently, the unmixing effort is preceded by a feature selection procedure using a standard peak picking algorithm (MATLAB, Matick, MA, USA). The result is a selection of 252 ion peaks, and the NMF is performed on the reduced data set of size 1734 pixels by 252 peaks. Figure 6.2 shows the spatial and mass spectral signatures of this NMF decomposition.

For this proof-of-concept study, we opted for a simple binary measure of co-localization. Any ions that both show up in the same NMF component and exceed a certain threshold, are marked as connected.

Step 2

The binary distance measure of step 1, has implicitly constructed a binary distance matrix to manufacture the new network from. The distance matrix is first filtered to


Comp 1
Comp 2
Comp 3
Comp 4
Comp 5
Comp 6
x 10^4

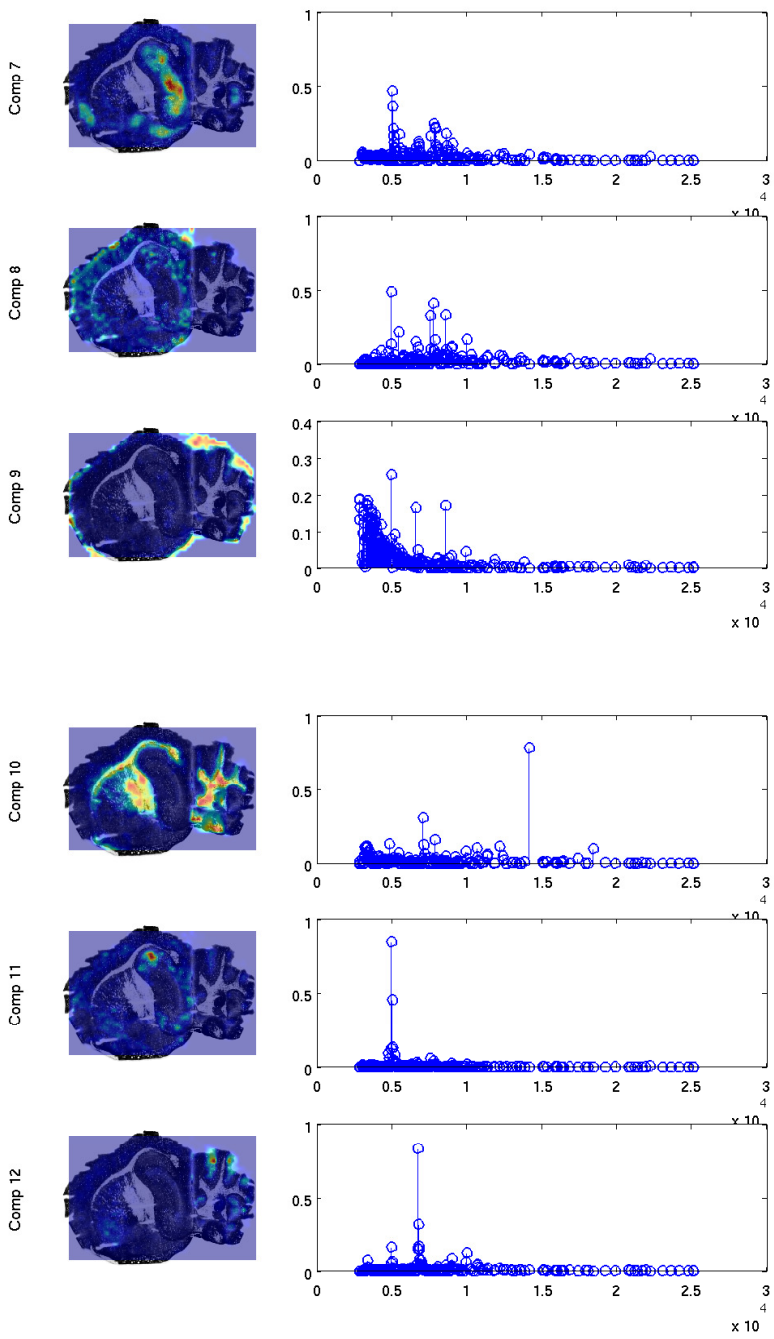


Figure 6.2. NMF decomposition of the peak picked MSI data set. All ion peaks that appear together in one of these twelve components are linked in the co-localization network.

only retain edges whose participating ion peaks exceed 10% of a component's total ion intensity range. This threshold avoids some of the low amplitude noise visible in Fig. 6.2, and ensures that the network will contain only the most robust edges. After filtering, the binary distance matrix is turned into a long list of edges. The list of edges is imported into Cytoscape, a specialized network software suite that is well-known in the systems biology field (see http://www.cytoscape.org/).

The initial network generated from the list of 1198 co-localization edges is shown in Fig. 6.3. The nodes of the network represent the participating ions, and are shown with their m/z value in the image. The network shows only 92 of the 252 ion peaks that were considered. These 92 ion species are the ones whose expression levels exceeded the user-specified threshold. The others consist mainly of low amplitude noisy peaks that were found by the peak picking algorithm, but that show an insufficient signal-to-noise ratio to participate in this analysis. Notice that the network is not laid out yet, and that in its current form it can contribute little further insight into the MSI data set.

Figure 6.3 demonstrates that generating a network structure without optimizing the layout towards depicting the co-localization aspect, provides only a marginal improvement in understanding. Optimizing the layout of a network structure is a nontrivial problem in computer science, but in recent years some progress has been made towards automating the process. First, an attempt is made to layout the co-localization network in a circular configuration. Figure 6.4 depicts the result of this layout. Although the circular layout is often used in genomic studies, we find that here it is not very intuitive. The main reason is that the circular layout implies an inherent ordering of the ions on the circle, which our data set does not provide.

The goal of the network layout is to visually convey which ions are tightly co-localized, which ones share only a loose relationship, and which ones have little to do with each other. With this goal in mind, the *organic* layout algorithm is applied to the mouse brain network. This layout algorithm is a type of spring-embedded algorithm that optimizes towards showing clustered relationships. This means that tightly interconnected cliques are kept close together, while the overall shape of the network is driven by the looser edges connecting to some members of these cliques. The resulting network is shown in Fig. 6.5. It is clear that this network is much more telling than the ones generated earlier. The network shows which molecular masses apparently co-localize in the mouse brain (e.g. m/z 7013, 5678, and 15333 at the bottom left). It also shows which ions form hubs in the network by connecting to a very wide variety of spatial patterns in the tissue (e.g. m/z 6581). To further highlight the structure embedded in the network, Fig. 6.5 colors the edges according to the NMF component they came from.

To keep the network visualization useful, it is important that the graph representation remains visually digestible. This means that in addition to presenting analytical information, a network visualization will often need to be pruned to show only the most

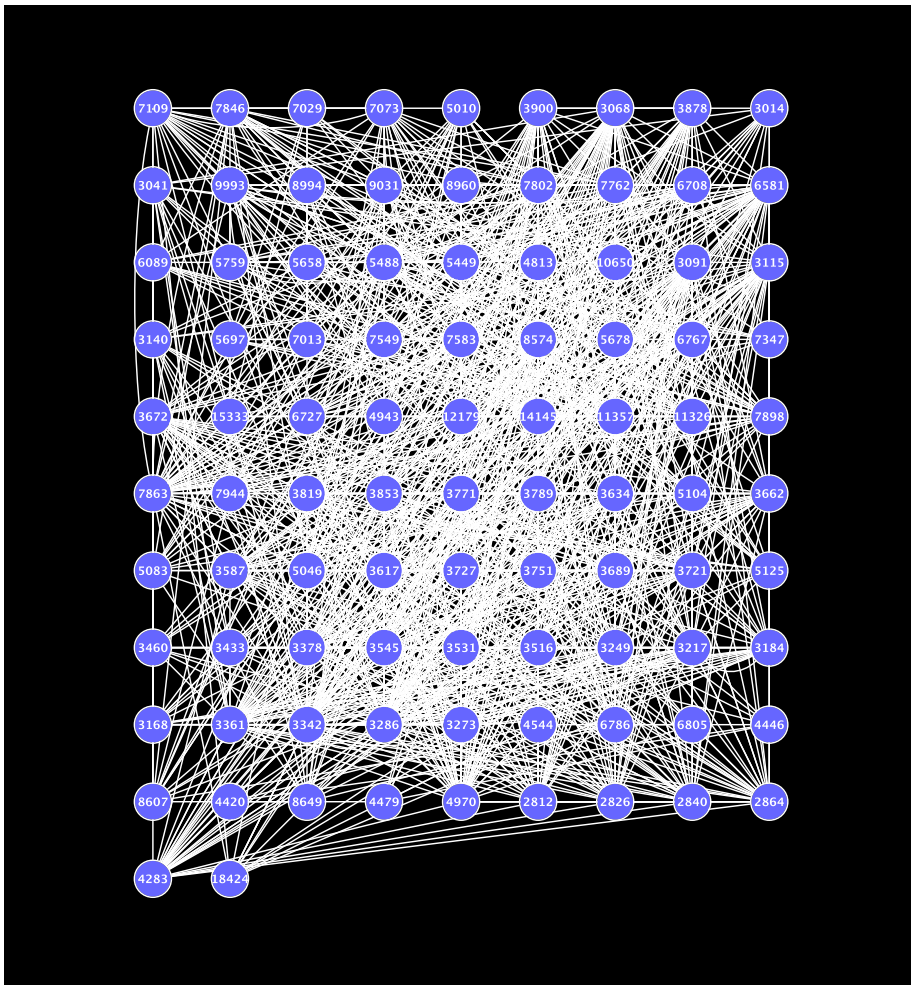


Figure 6.3. Initial co-localization network generated from the 92 most influential ion species in the mouse brain data set (92 nodes, 1198 edges). The nodes represent ion species and show their m/z value in the image. The edges are generated from an edge list constructed using the co-localization measure. Notice that the network is not laid out yet, and that in its current form it can contribute little to further insight into the MSI data set.

prominent relationships and to avoid overloading the graph. In our example the only pruning parameter is the peak amplitude threshold used in the co-localization measure. In the network of Fig. 6.5 this parameter was set to 10% of a component's total ion intensity range. As this threshold is shifted downward, more ion species will participate in the network and the graph will become more crowded. An example of this effect is shown in Fig. 6.6(b). In network representations there is a trade-off between including information and overcrowding the network to the point of rendering the vi-

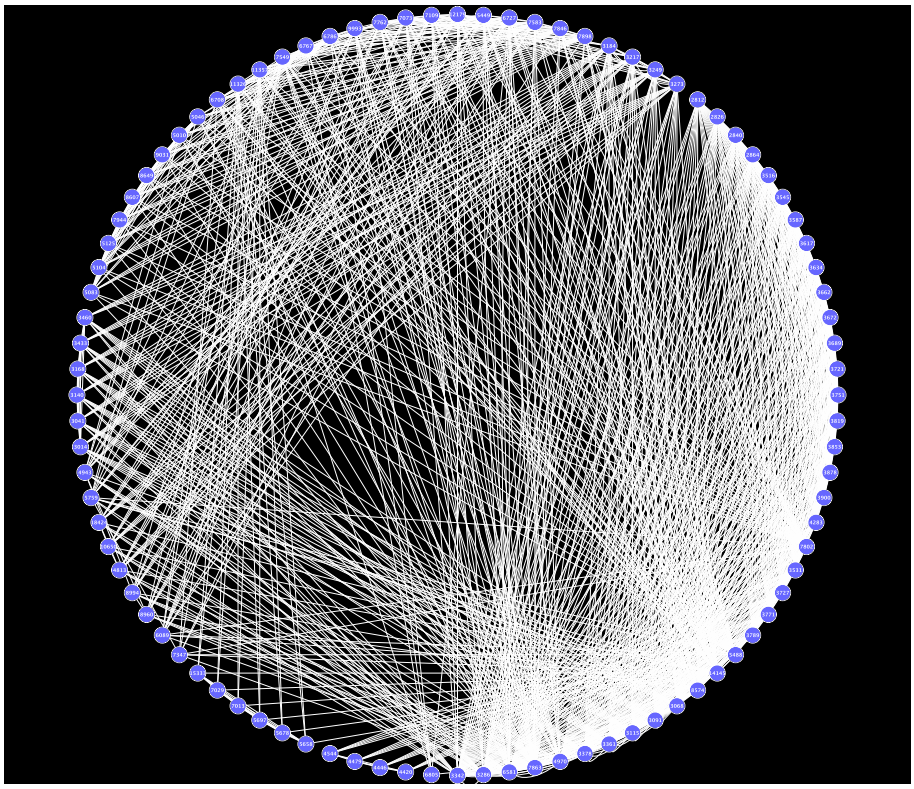


Figure 6.4. Co-localization network of the mouse brain (92 nodes, 1198 edges, circular layout). The circular layout implies an inherent ordering of the ion species, which is not provided by our data set.

sualization uninformative. Overall, the threshold parameter works as a pruning device on top of the unsupervised analysis, and only influences the visualization aspects, not the analytical results. As the particular characteristics of a MSI data set determine the optimal trade-off between insight and broadness of scope, the threshold will need to be re-assessed on an experiment-by-experiment basis. The analytical part of the analysis, however, remains identical.

6.3.3. Interpretation

The previous two sections demonstrated how to go from a MSI data tensor to a network representation of the co-localization aspects encoded in that data set. Basically, the process can be wrapped into an unsupervised analysis building block, requiring the specification of only two parameters (pruning threshold and rank of the decomposition).

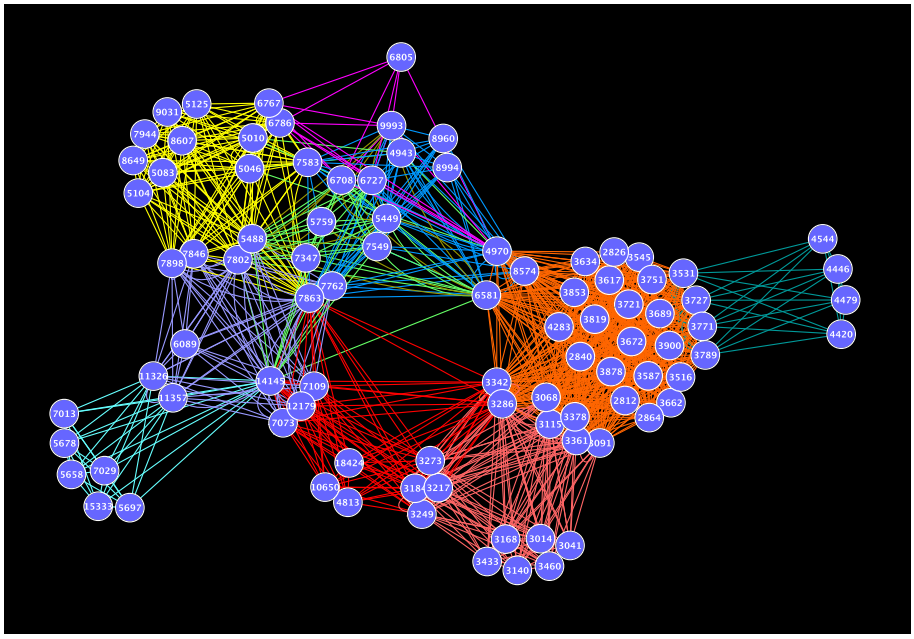
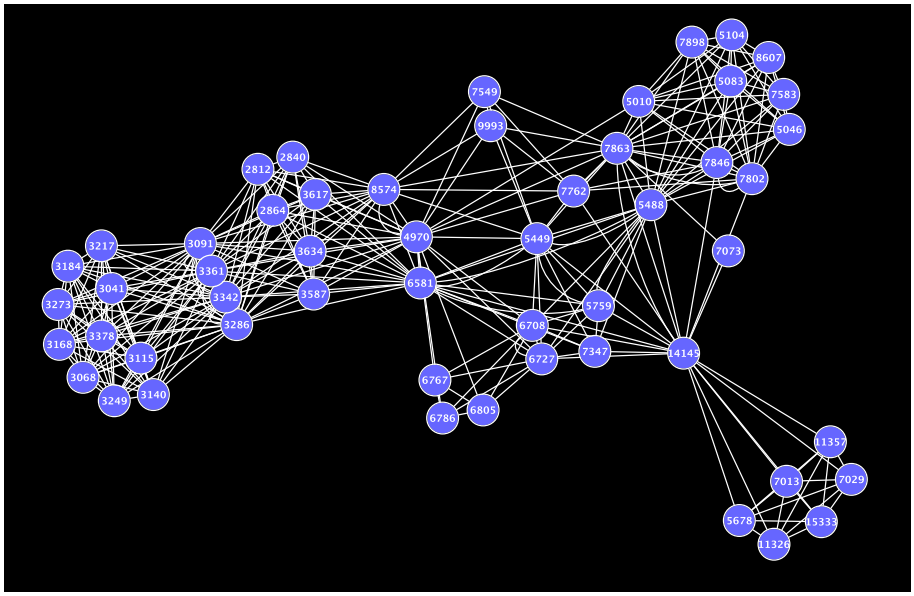


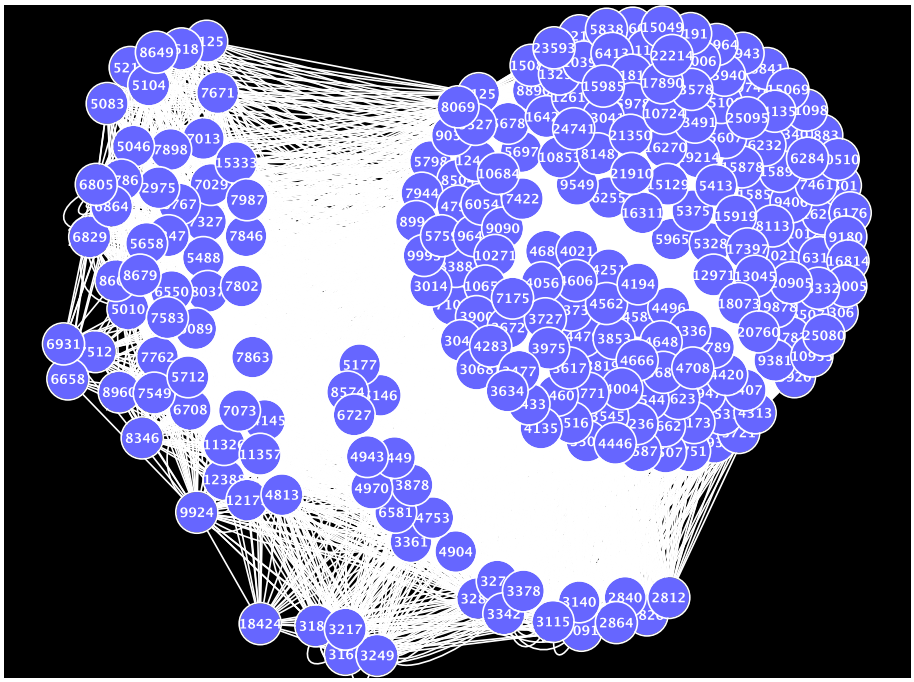
Figure 6.5. Co-localization network of the mouse brain (92 nodes, 1198 edges, organic layout). The network shows cliques of ions that co-localize very specifically (e.g. m/z 7013, 5678, and 15333 at the bottom left). It also shows which ions form hubs in the network by connecting to a very wide variety of spatial patterns in the tissue (e.g. m/z 6581).

Once a network is available, it can be interrogated from different viewpoints. It can be used in a steered fashion, to answer specific chemical questions. To find out, for example, which ions show up in the same places as a particular molecule of interest. It can also be used in a purely unsupervised manner, to discover which are the most important spatial expression patterns in the data, and by what ions they are defined.

A clique of nodes in a co-localization network highlights a subset of ions in the data set that co-localize strongly. However, a group of nodes also implicitly defines a particular area in the tissue. This link between the network and anatomical tissue regions is explored further in Fig. 6.7, which zooms in on a section of the co-localization network of Fig. 6.5. When traced to the spatial domain, the clique of ions located on the lower left (m/z 7013, 5678, 5658, 7029, 5697, and 15333) selects specifically and only the lobes of the *cerebellum*. All the way at the top, the clique of m/z 5125, 9031, 7944, 8607, 8649, 5083, and 5104 is shown to localize to the *hippocampus*. The cliques do not overlap in surface area, and therefore there are no direct links between these areas of the network. However, there are indirect links via other clusters of nodes. For example, if we move along the edges from the lower left to related ions such as m/z 11326 and 11357, it becomes apparent that these ions have a spatial distribution simi-



(a) threshold = 0.15



(b) threshold = 0.05

Figure 6.6. Co-localization networks of the mouse brain at different amplitude thresholds.

lar to that of the lobe-related clique, but at the same time their spatial presence differs sufficiently to warrant a visual separation in the network. As we progress upward to m/z 6089 and further to m/z 7898 and 7846, we are actually moving from spatial expression pattern to spatial expression pattern with incremental differences.

A co-localization network can be considered to report on the spatial relationships between molecules in tissue, but abstracts the matter away from the particular measurement grid employed by the experiment. It thereby opens up opportunities to examine co-localization within a single experiment as well as across different experiments.

6.4. Conclusions

Conceptually, the construction of a co-localization network from a MSI data set is a multivariate exploratory tool. Once a network representation is available, it can be used as a map to the 'chemical territory' sampled by the MSI experiment. In a relatively rapid and clear fashion, the network provides insight into which ions out of the hundreds or thousands measured, exhibit a potential biological or instrumental relationship (reported by their co-localization). Just like the developments of chapters 3 and 4, the co-localization measure is a tool to plough through the multitude of data that a MSI experiment commonly produces, but it does so from an angle that, until recently, has been largely unexploited.

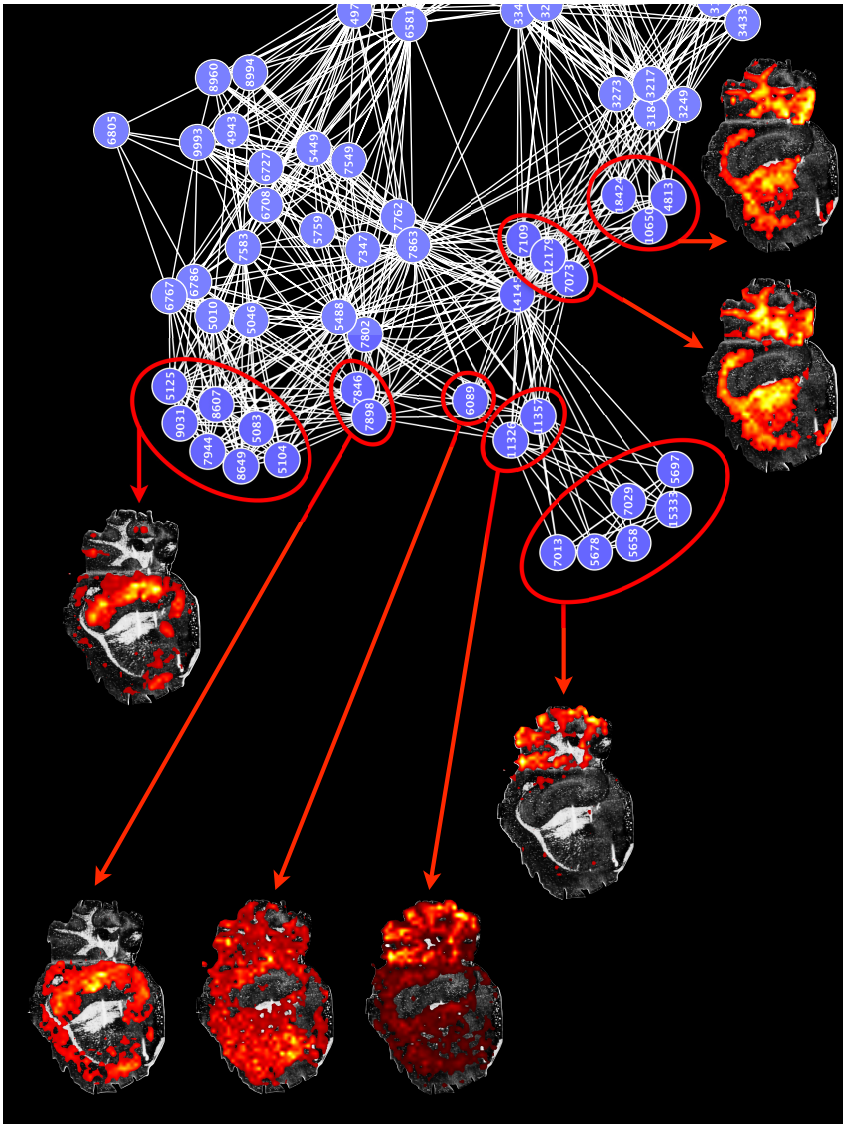


Figure 6.7. Exploration of the network/spatial area duality. Clusters of tightly interconnected nodes signify strong co-localization between those ions, but additionally they also translate to a specific area in the tissue. Tracing the link from one area to another along the edges of the network, identifies which ions share surface area with the regions or ions of interest.

CHAPTER 7

Hormone Processing Pathway Analysis using Mass Spectral Imaging and Data-mining of Co-localization

To do successful research, you don't need to know everything,
you just need to know of one thing that isn't known.

Arthur Schawlow

This chapter is an extended version of the following publication: • Van de Plas R. et al., "Hormone Processing Pathway Analysis using Mass Spectral Imaging and Data-mining of Co-localization", manuscript submitted for review.

Supplementary figures are available in Appendix B.

The previous chapters have mainly focused on specific computational developments for the *in silico* analysis of MSI experiments. This chapter however, takes a more holistic perspective, and describes a concrete biomedical study from A to Z. This case study is a testament to the importance of computational analysis for MSI, as it demonstrates how the combination of mass spectral imaging with appropriate *in silico*

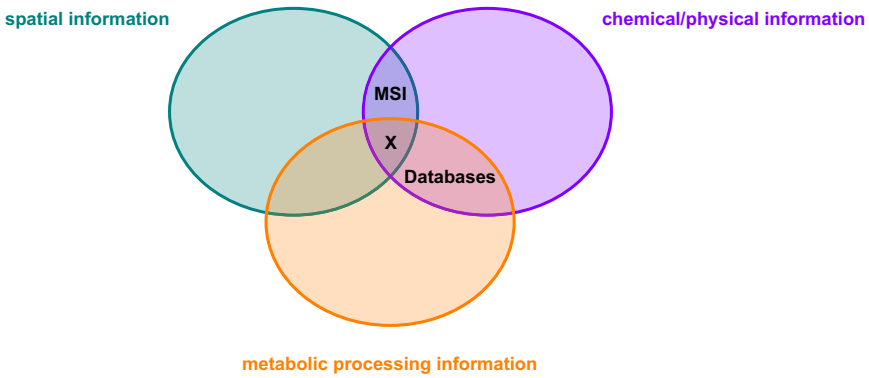


Figure 7.1. The metabolic screening methodology discussed in chapter 7 fuses three distinct information sources. Spatial information and biochemical characterization are fused empirically through MSI experiments, and mined computationally to enable filtering on the basis of molecular co-localization. The link with metabolic pathway information is procured *in silico* from curated information in literature and databases (e.g. UniProt).

methods has the potential to go significantly beyond what is currently possible with MSI alone. The overall goal of the study is to develop a metabolic pathway screening method that works directly from tissue, and is able to provide rapid prospective information on where the metabolic processing of diseased tissue starts deviating from normal metabolic function.

The study combines spatial information and biochemical characterization using MSI experiments, performs several computational preprocessing, filtering, and differential analysis steps, and finally ties the results into a metabolic pathway representation to allow for a more direct biological interpretation. The chapter essentially fuses information from three distinct data sources, as depicted in Fig. 7.1.

7.1. Aim of the study

Mass spectral imaging (MSI) allows high throughput screening of the localization of hundreds of biomolecules in tissue. However, little attention has been paid to the development of approaches that examine co-localization and correlated expression between these molecules. This chapter demonstrates that a combination of mass spectral imaging technology with *in silico* data-mining of the co-localization and correlated expression aspects of the measurements, unlocks a largely untapped intrinsic potential of MSI. It generates a powerful analytical technology for screening major biochemical processes directly in tissue. The potential of this approach is demonstrated by mapping the differential analysis of biomolecules in normal versus diseased tissue to hormone processing pathways. This strategy highlights metabolic pathways rather

than individual biomarkers perturbed by the disease, and enables deeper biological interpretation and hypothesis building. As an example, several diabetes model systems are examined for their well-studied hormone processing system and compared at the metabolic level. However, this approach can be applied to any field where sufficient literature and databases of interactions and potential processing are available.

7.2. Introduction

In recent years the molecular mechanisms underlying many of the world?s most pressing diseases have been shown to exhibit a complexity that exceeds most of the early expectations [120, 121, 122, 123, 124, 125]. Although such pathomechanisms make up only part of the interaction map sought after in the growing field of systems biology, it is clear that even in this subarea the sheer amount of variables involved presents a considerable challenge [126, 127, 128, 129]. While it is true that biology and medicine are becoming increasingly data-rich sciences, an accurate understanding of disease-causing interactions will require a fusion of multiple data sources, an examination at various biological scopes and scales, and the development of methods acquiring more information per experiment. Inspired by these insights, this chapter presents a protocol that employs three distinct data sources to provide insight into one of the most prevalent diseases in the world today: diabetes [130, 131, 132, 133]. The information sources are spatial/co-localization information, biophysical/biochemical characterization, and metabolic processing information. The first two are empirically acquired via a small number of experiments using a novel imaging methodology based on mass spectrometry, and the third is extracted from curated databases and literature.

Mass spectrometry (MS) has become the central technology for studies involving the large-scale characterization and identification of biomolecules in medicine and biology [15, 50, 128, 134, 135]. In its wake mass spectral imaging (MSI) has been gaining considerable momentum as a means of examining the spatial distribution of biomolecules in organic tissue [22, 23, 25, 28, 30, 38, 136, 137]. A MSI variant based on matrix-assisted laser desorption ionization (MALDI), is commonly applied for studying the localization in tissue of biomacromolecules such as polypeptides. This study centers on MSI experiments employing this variant, and for the remainder of the text MSI will refer to MALDI-based MSI. Although its spatial resolution (e.g. 50x50 μm^2/pixel) is currently relatively low compared to traditional molecular imaging techniques such as microscopy (e.g. 1x1 μm^2/pixel), and the identification of biomacromolecules straight from tissue remains challenging, MSI has several prominent advantages. It can concurrently detect a substantial part of the molecular complement present in tissue, and thus deliver a broad view on differentially expressed molecules with only a small number of experiments. The MSI technique essentially fuses spatial information with the molecular specificity of MS, and as such provides a

molecular imaging technology that avoids the labeling steps required by other imaging modalities such as fluorescence microscopy and immunolabeling. As labeling is avoided and MS can detect the presence of a wide range of chemical species in a single measurement, a MSI experiment can map the localization of hundreds of biomolecules within a single tissue section. As MSI does not require a prior hypothesis regarding a target molecule, it is particularly suited for exploratory studies.

Since its initial application [8, 10, 18], development of the MSI methodology has mainly focused on the physics of instrument development and the chemistry of tissue preprocessing to enable reliable ionization [20, 34, 35, 138, 139, 140, 141, 142]. Now that wet-lab protocols for various tissue types are becoming increasingly available and stable, the major challenge in performing a MSI experiment is shifting more to the interpretation of the often massive amounts of acquired data. Currently MSI based imaging is most commonly used to produce ion images, showing the spatial presence of a single ion, mass, or molecule in the tissue. However, there is a growing need for data processing methods capable of extracting nuggets of information from a sea of collected MSI data. Some initial work in this area has appeared recently [75, 79, 93, 94], but many important questions remain largely untouched at this time.

This study describes an experimental setup to screen organic tissue for rapid insight into the biochemical processing of spatially related molecules. The screening method relies on MSI to simultaneously collect both the biochemical characterization of multiple molecules and their spatial distribution throughout the tissue. The computational part of the screening method combines MSI with multivariate factor analysis to enable filtering on the basis of co-localization. It also provides an automated differential comparison, and finally it maps the differential findings to a curated set of metabolic pathways. It is demonstrated that having both the biochemical characterization and spatial aspects available has great potential for the elucidation of biochemical pathways and protein processing, as the products of these operations tend to manifest themselves in a co-localized fashion.

As a proof of concept, the platform is used to explore pancreatic islet-specific hormone processing in a wild type mouse versus three well-established diabetes mouse models: a chemically induced diabetes mouse model (DIA), and two genetically induced diabetes mouse models, Ob/Ob and SA/HF. Initially a uni-molecular stance is taken, tracking the spatial distribution throughout the pancreas of a single target molecule, common in diabetes studies, insulin. We demonstrate how MSI enables unlabeled *in situ* tracking of insulin in tissue, and also how a basic quantitative density estimate can be obtained from this type of measurement. Then a full multi-molecular approach is applied, performing a broad spectrum differential analysis of islet-specific content between the different mouse models, and revealing perturbations in various metabolic pathways linked to the islets of Langerhans.

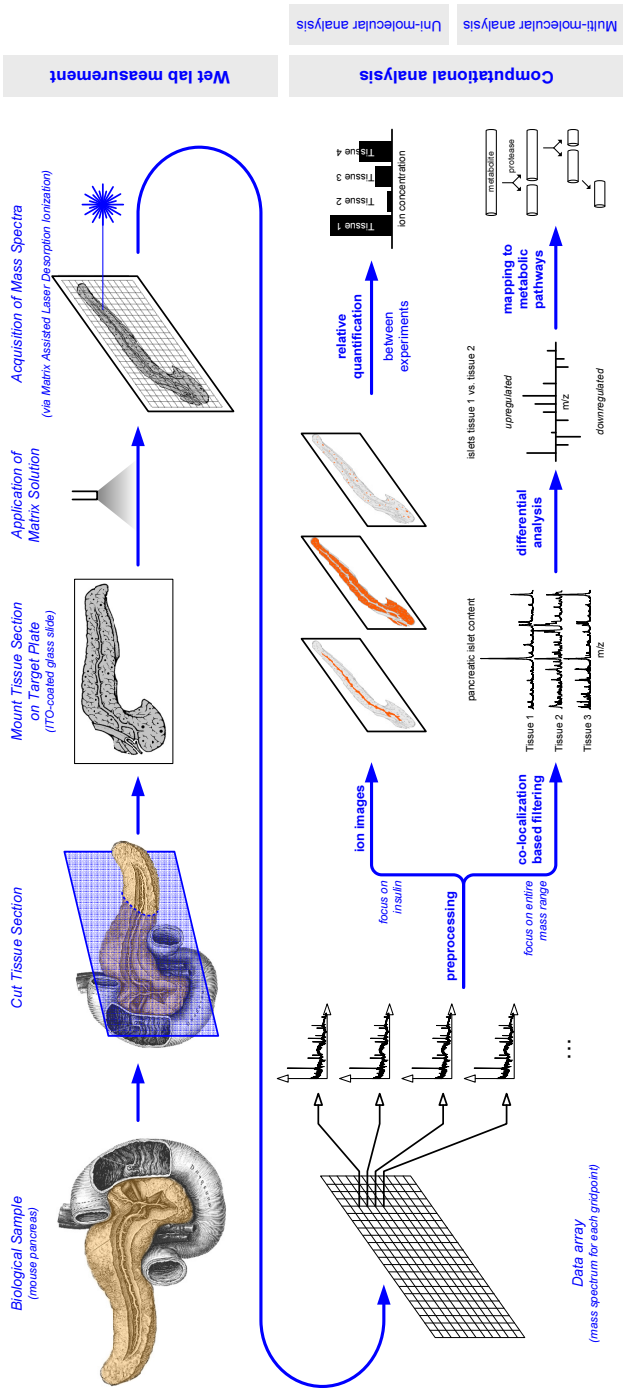


Figure 7.2. Overview of the mass spectral imaging experiments and their computational follow-up. (a) A thin tissue section (10 µm) is cut from the pancreas using a microtome and mounted on an electrically conductive glass slide. (b) After washing with ethanol, a layer of chemical matrix solution (CHCA) is deposited on the tissue section to enable MALDI-based ionization of the molecules embedded in the tissue. The deposition is done via a robotic acoustic droplet ejection system to ensure homogenous matrix coating. (c) The tissue carrying glass slide is introduced into the mass spectrometer and a grid of measurement locations is defined across the tissue. A mass spectrum is collected at each grid location and stored into a data array for preprocessing and further computational analysis. The computational analysis entails both a uni-molecular path focused on insulin, as well as a multi-molecular path aimed at screening for metabolic pathway perturbations.

7.3. Results

We describe a workflow containing both wet-lab and computational components to go from tissue to hormone processing pathways. Additionally, we detail a comparison of these pathways between multiple disease models. Five consecutive issues are addressed in the following result sections:

1. When imaging the presence of insulin in pancreatic tissue, can a MSI-produced ion image match the performance of an established immunostaining method? And if so, how can this be done without the need for labeling or indirect tracking via a related molecule such as the C-peptide?

2. Can an estimate of normalized insulin quantity or density be extracted from the MSI data for cross-comparison between the various mouse models?

3. How can we surpass ion images and extend the analysis to reveal other ions that co-localize with insulin in the pancreatic islets? How can we extract without physical excision a multi-molecule signature of tissue content specific to the islets of Langerhans?

4. How can we make a direct comparison between the islet signatures extracted from the different mouse models?

5. How can the differential comparison of islet content be translated to a metabolic representation? How can MSI be used to explore the intracellular processing of individual biomolecules such as hormones across different disease models? Or how can the results be lifted from a mass context to a metabolic pathway context using only a small number of experiments.

7.3.1. In situ unlabeled imaging of insulin in pancreatic islets using MSI: insulin-reporting ion image comparable to established immunostaining

This section and the next focus on the single molecular species of insulin. In this section we describe how the MSI methodology can be used for imaging the presence of insulin *in situ* without the need for labeling. Also the performance of this unlabeled approach is compared to an established labeled method employing immunostaining.

Pancreas tissue is harvested from the four mouse models (one wild type and three diabetes models: DIA, Ob/Ob, and SA/HF), flash-frozen in liquid nitrogen, and cut into sections using a standard microtome. One section was stained with standard hematoxylin and eosin staining (H&E). Another section undergoes an immunostaining for insulin (red) and cell cores (blue). A third section is left unstained and undergoes MSI

measurement. The MSI experiments, detailed in Fig. 7.2 and in the methods section, are identical across all four mouse models. By collecting mass spectral measurements at regular intervals (50 μm) directly from the tissue, the MSI procedure acquires *in situ* the presence of insulin and other molecules throughout the pancreas. The measurements are collected into four data sets, one for each mouse model, and preprocessed as specified in the methods section. The effect of the preprocessing is highlighted by supplementary Fig. B.1 and B.2. All findings in this study are based on these data sets, either directly (e.g. this section and the next) or via trends extracted from the measurements through data-mining (e.g. the final three result sections). Knowing the native molecular mass of insulin in mouse, we can extract from our data sets the location in the tissue of ions corresponding to this mass. As there are two forms of insulin in mouse, Ins1 (5810 Da) and Ins2 (5803 Da), two ions are of interest: m/z (5810+1)/1 and m/z (5803+1)/1. The complexity of tissue-derived samples causes the collected mass spectra to exhibit a cruder mass resolution than is normally the case for standard proteomic analyses using the same instrument. Therefore it is difficult to make a clear distinction between both insulin forms and their various naturally occurring isotopes. As a detailed examination of the differences between the two insulin forms lays outside the scope the study, an encapsulating stance is taken where any ions falling within a mass-over-charge (m/z) window from 5800 to 5820 are considered to report insulin presence. Focusing only on this window, we extract from all four data sets an ion image showing the spatial distribution of insulin reporting ions. Confirmation for this approach is obtained through standard proteomic analysis detailed in the methods section, which identified the selected ions as a mixture of both forms of insulin. The ion image is a false color image showing the spatial presence of insulin in the tissue. To aid interpretation it is overlaid on a higher spatial resolution gray level image of the tissue section, which is taken before the matrix coating is applied. The images acquired from all three modalities across all four models are collected in Fig. 7.3, allowing for a direct comparison between the classical insulin staining and the novel insulin ion image approach. In the ion images only pixels showing a substantial insulin presence are colored. A rigorous comparison between the different modalities using numerical performance measures is difficult given that different tissue sections have to be used. However, by selecting neighboring sections there is a sufficient overlap in shape and physiological layout between the sections to allow for an intuitive assessment. Two aspects are examined: specificity and sensitivity. Specificity is meant in the sense of correctly detecting insulin and not falsely reporting another molecule. If the insulin staining is taken as a gold standard for detecting insulin presence and delineating beta cell containing islets of Langerhans in tissue, the unlabeled MSI approach does very well, evidenced by the strong commonality in detected islet layout. Although the ion images demonstrate a lower spatial resolution than the microscopy-based modalities, the images exhibit almost identical constellations of islets in both the WT and Ob/Ob cases. The SA/HF and DIA cases reveal very little insulin presence in either modal-

ity, making evaluation difficult. Nevertheless even in these cases, the larger beta cell clusters are detected at identical locations. These results support the idea that the unlabeled MSI protocol can indeed match the specificity of labeled approaches when it comes to detecting insulin presence in tissue (see Fig. 7.3). Sensitivity is concerned with whether insulin is reported if it is present. Using the same gold standard, it is apparent that the MSI approach performs adequately for our purposes but also that the insulin stainings perform better in delineating smaller beta cell clusters in the tissue and show higher sensitivity. This is not surprising given the courser spatial resolution and the fact that during the measurement process the insulin ions undergo severe competition for detection with hundreds of other ion species present in the sample. Also traditional techniques to counter this effect such as fractionation through liquid chromatography are usually incompatible with the *in situ* sampling goal of MSI. The sensitivity of a MSI experiment is dependent on both the instrumental setup as well as the particular tissue under study, and therefore should be considered on a case-by-case basis.

The results of Fig. 7.3 illustrate that MSI makes a viable technique for imaging the presence of native form insulin in tissue. It should be noted that the technique samples directly from flash-frozen tissue *in situ* and has no need for chemical or genetic labeling of the target molecule. As such it stays plausibly closer to the biological truth than techniques that require invasive chemical processing of the sample. If its spatial resolution (currently limited to cellular scales) and its somewhat lower sensitivity (dependent on the complexity of the particular tissue) is not prohibitive for the study at hand, MSI based insulin imaging can be a serious alternative to traditional labeled approaches when it comes to tracking a single species like insulin. The strong correspondence that is demonstrated between the ion image of a molecule and a well-established assay for that molecule provides credence for tracking other chemical species as well.

7.3.2. Relative quantification in MSI data: insulin-reporting ion concentrations across the various mouse models

After establishing MSI as a suitable means of tracking insulin presence in tissue, we describe a basic quantification follow-up to illustrate the tissue-wide and cross-experiment quantitative capabilities available via MSI. MSI also provides the opportunity to obtain, through its peak heights, some basic relative quantification information across multiple tissue samples. Such relative quantification requires ion count values to be comparable across tissue locations as well as across different experiments, making normalization a required preprocessing step. As specified in the methods section, measurements outside the tissue are removed (as indicated in Fig. B.1), and the data sets undergo simultaneous normalization across all four experiments and have their

baseline removed. The effects of preprocessing are illustrated in Fig. B.2. The relative peak height of a particular ion can now be considered an indicator for the relative quantity of its corresponding species. This approach is subject to the same comments and critiques as traditional non-labeled quantification via mass spectrometry, but has the added benefit of typically having a hundred-fold more spectra to average out any large deviations. The reliability of these peak height-based quantity indications is dependent on the quality of the data and will not outperform established quantitative techniques. However, the fact that these indications of quantity are readily available in any MSI experiment without modification and for all ions present within the measured mass range, makes them particularly suited for finding interesting leads which can be further confirmed using more rigorous techniques. In this insulin-centric section we use this approach to compare insulin concentrations between the wild type and the three disease models.

The ion images of Fig. 7.3 examined the spatial presence and relative concentration of insulin-reporting ions within the context of a single tissue class. Here these aspects are compared across all four tissue classes. First, the relative concentration of insulin is compared without taking the spatial distribution into account. Then, the spatial aspect is brought in to produce a 3D surface plot comparison illustrating insulin findings across all four tissue sections. In order to obtain a robust estimate of the amount of insulin from the tissue sections, only ion counts reporting insulin are selected and only measurement locations that exhibit a substantial signal to noise are taken into account. Along the mass domain this translates into the same m/z window constraint used for the insulin ion images. Along the spatial domain only pixels showing an insulin-reporting ion count exceeding 10000 are selected (1.36 % of the full range of insulin-reporting ion counts across all tissue classes; max = 735330). Given these constraints, the WT, Ob/Ob, DIA, and SA/HF classes return resp. 1.3345e8, 3.5151e7, 4.6227e4, and 6.1064e6 insulin-reporting ion counts. From the images shown in Fig. 7.3 however it is clear that all tissue classes do not show the same amount of islets surface area. This could be due to the particular cutting angle of the section, or due to the particular model producing typically larger islets (e.g. Ob/Ob). Either way, sections that contain larger and more islet areas than others would provide a larger amount of detected insulin, regardless of whether there is a true difference in insulin concentration in the islet cells. To account for this effect, the number of insulin-reporting ion counts is divided by the number of insulin-containing pixels (according to the threshold defined above). Essentially this surface-normalization provides a measure of concentration in insulin-reporting counts per surface area. The WT mouse model delivers 8.2840e4 ion counts/pixel or 33.1 ion counts/μm^2. The Ob/Ob, DIA, and SA/HF models contain resp. 9.3, 4.6, and 14.6 ion counts/μm^2. As the MALDI laser does not probe the entire pixel area in these experiments, and the numbers are tied to the specific laser intensity settings and environmental conditions for this set of experiments, these numbers are expected to be underestimations of

the true density. However as all experimental conditions were kept constant as much as possible between the experiments, they do lend themselves for cross-comparison between the various mouse models. Figure 7.4 summarizes the quantitative results in a bar graph and reveals a clear pattern. The WT model contains the largest concentration of insulin, which surpasses the concentrations in any of the three disease models several fold. The mouse suffering from chemically induced diabetes, DIA, exhibits the smallest insulin concentration of all examined models both in islet surface area and ion counts per pixel. Both genetically induced diabetes models, Ob/Ob and SA/HF, exhibit severely reduced insulin content when compared to WT but the reduction is not as dramatic as with the streptozotocine-treated DIA. It is also interesting to note that SA/HF exhibits a higher insulin concentration per pixel than Ob/Ob while Ob/Ob demonstrates substantially larger islet surface area. For further confirmation, two additional graphs are provided in Fig. B.3, where the maximum insulin ion counts and a box plot of the selected ion counts are compared. The same general pattern is observed.

The color representations of the ion images shown in Fig. 7.3 are scaled to optimally represent the ion count range detected within a single experiment. If all ion images are put on the same ion count range and color scheme, the large differences in insulin concentration apparent from Fig. 7.4 make the comparative visualization of the lower concentration levels in the disease models difficult. Instead Fig. 7.5 provides a combined visualization of spatial distribution and insulin concentration by showing a comparatively scaled surface plot representation of insulin ion counts overlaid on gray level pictures of the tissue sections. The surface plots reflect the summarizing numbers from Fig. 7.4 and provide easier visual interpretation for the relative quantification results. The selection of insulin-reporting pixels using the noise threshold mentioned above provides us with a rough estimate of the percentage of pancreas surface area that is populated by beta cell islets. For the WT model, insulin is present in detectable amounts in 30 % of the tissue area that was imaged. As expected, the introduction of streptozotocine in the DIA model reduces the presence of insulin to an almost undetectable level of 0.06 % of the surface area. The Ob/Ob model exhibits remarkably large beta cell islets of low insulin concentration when compared to WT, covering 16.8 % of the tissue surface. In the SA/HF section, insulin shows a surface coverage of 2.1 %. These percentages are of course linked to the specific tissue sections examined here and serve purely a descriptive and not an inferential function. Although the imaging experiments were not initially designed for relative quantification and further confirmation via replicates is advised, the fact that all relative quantitative information gleaned from the experiments is in line with earlier studies and literature gives further support to the validity of the data that has been collected.

These first two result sections focused on the qualitative and quantitative uni-molecular capabilities of MSI, demonstrated on insulin. The goal of the following three result sections is to unlock the multi-molecular potential of MSI to perform a differential

comparison of the content of the islets of Langerhans between the different diabetes mouse models. Unlike the first two result sections where only insulin is studied and the focus lay on imaging and spatial distribution through combined wet-lab and computational means, the following, purely computational, result sections enable an islet comparison that encompasses multiple chemical species, and that due to its multivariate nature enables differential analysis at the metabolic pathway level.

7.3.3. Other non-insulin ions that co-localize with insulin in the pancreatic islets: Extracting a multi-molecular signature of islet content

The different variants of diabetic stress studied here cause different hormone-processing events to take place in the tissue. The study of such metabolic processing pathways requires information on more molecules than insulin alone. Due to the wide mass range employed in the MSI experiments such data is certainly available, but a robust data mining effort is required to extract any nuggets of information from this sea of cross-correlating data.

This section aims to answer the question: How do we go from a uni-molecular to a multi-molecular analysis, and how do we focus on a subset of chemical species specific to the islets of Langerhans? The previous sections showed that insulin-reporting ions were a good qualitative marker for islet location. Here we use an algebraic matrix decomposition method on the four data sets to find ion peaks co-localizing with the insulin peak. This effectively filters islet-specific content from the data sets. Ion counts that are acquired in non-islet areas or that appear not solely in islet areas are discarded from further analysis. The matrix decomposition method is essentially used as a filtering mechanism to obtain from the four pancreas sections a mass spectral signature specific for their islets of Langerhans. The result can be considered an *in silico* equivalent to an islet-targeting enrichment of the sample in the wet lab, but without the need for chemical treatment or physical excision.

A MSI experiment provides a spatial distribution for every ion species that falls within its mass range and that exhibits a reasonable signal-to-noise ratio. As their number can run into the hundreds to thousands for a simple experiment, it stands to reason that a manual examination to find ions that exhibit for example co-localization becomes increasingly difficult and prone to human error. Instead such correlation in spatial expression should be mined computationally to keep the analysis practically feasible and devoid of human bias. If ions that consistently show up together in the same tissue area can be grouped together, we have a specific content signature for that area. In current ion image oriented approaches to MSI such intermolecular information is largely left untouched.

The central aim of this case study is to unravel changes induced by diabetic stress on the endocrine peptide complement of the pancreas. This causes us to be particu-

larly interested in the islets of Langerhans, who are made up of at least five known cell types, alpha, beta, delta, epsilon, and PP cells, which in turn are responsible for some of the most important hormones related to digestion and glucose homeostasis. We therefore need a data-mining algorithm that is capable of extracting from each mouse model a mass spectral profile of the chemical content of specifically the islets of Langerhans. Such an islet-specific trend should essentially select only ions found specifically at the islets, weeding out ions that do not appear at the islets or ions that do appear there but also in many other non-islet areas. Additionally the trend should describe a common relative expression pattern for the selected ions as encountered across the islet-derived mass spectra. The objective can be compared to retrieving an average relative mass spectrum from the pixels covering the islet areas of the tissue, but with all non-islet specific peaks removed to a maximal extent.

The sought-after trend detection shares many characteristics with what in data mining is considered the problem of matrix decomposition. Matrix decomposition algorithms seek to disassemble a body of data into a set of underlying patterns or components. Some of these algorithms have been applied to MSI data for interpretation [74, 75, 83] and de-noising [78, 79]. In this study we introduce non-negative matrix factorization (NMF) [72] as a suitable matrix decomposition method for filtering MSI data. An overview of its application is shown in Fig. 7.6. NMF tries to decompose the data into a preset number of components while enforcing non-negativity as a hard constraint on these components. This non-negativity constraint is the reason for the excellent fit between NMF and mass spectral data. As mass spectral data by definition consists of only nonnegative values (ion counts) the components extracted by NMF allow for a more straightforward mass spectral interpretation than was ever possible with algorithms that do not enforce a non-negativity constraint such as PCA. PCA applied to MSI data typically produces mathematically correct, but interpretation-wise less desirable signatures in the mass spectral domain because they exhibit a mixture of positive as well as negative ion counts. On the spatial side, NMF delivers (like PCA) tissue zones of common ion content. On the mass spectral side, NMF provides (unlike PCA) interpretable pseudo-mass spectra to accompany these tissue zones. This second characteristic is crucial if we need to arrive at mass spectral-like signatures for our islet content comparison. As a result we employ NMF to perform the filtering role discussed above.

Each of the four MSI data sets is decomposed into six NMF components. Supplemental Fig. B.4 to B.7 show the results for respectively WT, DIA, Ob/Ob, and SA/HF. Various patterns embedded in the measurements emerge, but our goal is to select the pattern corresponding to the pancreatic islets. As we could only stain tissue sections neighboring to the MSI-measured tissue, the fluorescence images of Fig. 7.3 provide only a guiding idea of where islets are located. The previous result sections however showed that for all four models the spatial distribution of the insulin ion corresponds well with the locations of the islets in the tissue. As such the search for a pattern of

islet-specific ions can be considered the search for a pattern of ions co-localizing with the insulin peak. As the NMF algorithm is not steered, we have no inherent guarantee that the six most dominant trends extracted from the data will include an insulin ion-containing islet-specific pattern. However in the WT, Ob/Ob, and SA/HF models the ion count of the insulin peak is sufficiently large to guarantee an insulin-component to be present within the dominating trends. As indicated in Fig. B.4, B.6, and B.7, we select the component that exhibits the dominating insulin peak as a representative mass spectral signature for the islets of Langerhans in those models. As is apparent from Fig. 7.4, the amount of insulin in the DIA model is reduced to almost undetectable levels and as such the NMF decomposition of the DIA experiment does not present an insulin-dominated islet trend for further analysis. The following two result sections will therefore focus on the differential comparison between the WT, Ob/Ob, and SA/HF models.

7.3.4. Comparison between islet signatures extracted from the different mouse models: Differential comparison of islet-content

This section builds upon the islet signatures developed in the previous phase and aims to extract which masses, ions, or molecules - specific to the islets of Langerhans - show differential expression between the different mouse models. As an example of the screening capabilities provided by the method developed here, we focus on the m/z 2800 to 4000 region in the Ob/Ob versus WT case. This mass range shows an interesting set of differential ion peaks that bear strong witness to a particular metabolic pathway deviation in the Ob/Ob mouse model.

The NMF decomposition performed in the preceding analysis step provides us with a mass spectral signature of islet content for the WT, Ob/Ob, and SA/HF models. Each signature shows a relative profile of ion counts as it was commonly found across islet-derived mass spectra. On the one hand an islet signature is the result of an averaging across all islet-derived mass spectra in a model. On the other hand the signature is filtered to only contain ion peaks specifically showing up at the islets, reducing ion peaks that are not exclusive to the islet areas of the tissue. The initial goal is to perform a qualitative comparison at the mass spectral level, singling out differential peaks that either show up or disappear from model to model. A secondary, more quantitative aim is to assess relative peak height changes between the signatures. In order to compare peak height, the signatures need to be scaled relative to each other. As the NMF employed a linear sum of signatures to approximate the original data, all ion peaks within a signature should be scaled by a single common scaling factor. The problem of scaling two islet signatures relative to each other is thus reduced to finding an appropriate scaling factor a single common peak. The obvious choice for a common peak between the signatures is the insulin ion peak. The insulin peak has the

best signal-to-noise ratio of all detected ions and is present in all three islets signatures, making it a prime candidate for internal standard to which other peak heights can be linked. Simply assigning the insulin peak to 100 % and scaling the other ion peaks accordingly is not a valid option. As shown in Fig. 7.4, there is a substantial difference in insulin concentration between the different mouse models. However, the surface-normalized insulin ion count shown in Fig. 7.4 gives us a guiding estimate of the relative insulin concentration. Each islet signature is scaled according to its corresponding surface-normalized insulin ion count and plotted relative to each other in Fig. B.9 (Fig. B.9(a): WT versus Ob/Ob; Fig. B.9(b): WT versus SA/HF; Fig. B.9(c): Ob/Ob versus SA/HF). To aid comparison, each plot shows one signature as a negative version. Due to the size of the insulin peak in these plots it is hard to discern other ion peaks that exhibit much smaller peak heights. Nevertheless each signature contains additional peaks that become visible when zoomed in (Fig. B.10 to B.13, resp. covering m/z 2100 to 4000, m/z 3800 to 5700, m/z 5500 to 7400, and m/z 7200 to 12000).

As an example of how these figures can lead to new insights and richer hypothesis building, we focus on the m/z 2800 to 4000 range in the Ob/Ob vs. WT case, shown in Fig. 7.7. Both WT and Ob/Ob report the presence of an ion peak (m/z 2904) corresponding to the Ins2 signal peptide (2886 Da). As expected from the one-to-one metabolic relationship with mature insulin in the beta cells, the relative peak height mirrors in both cases the relative height of the mature insulin peaks. A similar story is valid for the ion peak found at m/z 3130. This peak corresponds for example to the C-peptide of either Ins1 (3121 Da) or Ins2 (3133 Da), or possibly to the Chromogranin C precursor signal peptide (3118 Da). The fact that the relative peak height mirrors relative expression of other representatives from the Ins2 hormone processing pathway suggests that the Ins2 C-peptide is most probably responsible for the 3130-peak. Both ion peaks that were discussed have matching m/z positions and peak heights in both signatures, providing evidence for the reliability and reproducibility of the method from one experiment to the next. However, the real goal is to find qualitatively different ions between the two tissue classes. An example is found at m/z 3287 and m/z 3352 which correspond resp. to glucagon-like peptide 1 versions 7-36 and 7-37. Both are largely absent from the WT islets, while the Ob/Ob islets show a marked presence indeed. These observations seem to indicate that the Ob/Ob tissue deviates in its processing of the glucagon precursor. One possible explanation for these observations is an increase in the amount of glucagon precursors in Ob/Ob compared to WT. Support for this hypothesis is found in an observed increase in the presence of glucagon precursor signal peptides in Ob/Ob at ion m/z 2301 (not shown in Fig. 7.7, but in Fig. B.10). However, further evidence suggests that upregulation of glucagon precursors is only part of the story. Although more glucagon precursors seem to be made available, there is no detectable change in the expression of the precursor?s primary product, the hormone glucagon. In fact the amount of standard glucagon seems to remain constant

(see ion m/z 3493). The upregulation of the precursor therefore seems to be aimed at increasing the availability of a different cleavage product from the precursor. This observation suggests particularly the major proglucagon fragment (MPGF, 10202 Da) as a possible suspect for upregulation. The MPGF represents largely the second half of the glucagon precursor polypeptide sequence, and can be cleaved to various glucagon-like peptides. A possible trace of the suggested MPGF increase can be found at m/z 10270. Our mass resolution however does not allow us to rule out possible contributions to this ion from the IGF-2 E peptide, urging us to consider the MPGF increase as unconfirmed. Whether the observed ion at m/z 10270 corresponds to MPGF is actually not crucial to the story as we see clear upregulation elsewhere in the mass range in several of MPGF's glucagon-like peptide offspring. Two particular observations supporting the hypothesis that Ob/Ob exhibits perturbed glucagon-like peptide presence, are the strongly upregulated versions of glucagon-like peptide 1 at m/z 3287 and m/z 3352 that started our discussion, indicated as GLP-1 (7-36) and GLP-1 (7-37). Additional support is found at ion m/z 4171, which corresponds to the standard GLP-1 (4170 Da) and exhibits an increase in abundance, albeit smaller than with the GLP-1 variants. Although one might be tempted to conclude that upregulation of glucagon-like peptides can be considered a putative marker for the Ob/Ob mouse model, the data suggests that even further refinement is warranted. Of the four known cleavage products of the MPGF, the standard GLP-1, the GLP-1 (7-36) variant, and the GLP-1 (7-37) variant suggest upregulation. However, the fourth cleavage product from the MPGF known as glucagon-like peptide 2 (GLP-2, 3769 Da), does not seem to share this trend. Instead a marked downregulation is observed at ion m/z 3770, debunking a MPGF-wide upregulation hypothesis for Ob/Ob across all glucagon-like peptides. Overall the measurements seem to suggest a very specific upregulation of GLP-1 and its variants in the Ob/Ob mouse model (while GLP-2 is downregulated). With this glucagon-centric example we demonstrate that the availability of observations across different members of a metabolic pathway allows for a deepening of hypotheses that would not have been possible if only a few molecular species were observed. If only glucagon would have been measured, there would have been very little evidence to suggest an investigation into the rich variation in glucagon-like peptide abundance that became apparent in the example. The assignment of ions to metabolic pathway members is not fully confirmed via traditional proteomic identification (see later). However this does not pose a real problem as the procedure is meant as a screening method to acquire fast putative insight into which metabolic pathways deviate from normal metabolic processing and are worthy of further exploration. Also because the method allows for observation of multiple players in a metabolic network, the putative conclusions show a remarkable robustness to missassignment on a small number of these ions. Besides the glucagon-related species, Fig. 7.7 also shows some ions from other islet-related pathways such as the Ins1 and Ins2 B-chain, and GRPP (m/z 3440), IAPP (m/z 3932), and preptin (m/z 3949).

The example above demonstrates direct interpretation of a signature comparison. However, our aim for this methodology is to provide a screening method, and premature narrowing of the research focus to specific pathways might cause some interesting observations to be missed. Thus, to avoid human bias at this stage of the analysis, an automated manner of highlighting interesting differential ions is required. The analysis thus far has considered the islet signatures as mass spectral profiles, providing an ion intensity value for every m/z bin measured during the experiments. In this stage we narrow the focus to particular features of the islets signatures by using a peak picking method to select m/z positions of interest along the m/z axis. Following the setup suggested by [111], a standard peak picking algorithm, called mspeaks (Bioinformatics toolbox, MATLAB 2009a, The MathWorks, Natick, MA, USA) is used on the mean of the three scaled islet signatures (default parameters). A total of 419 peaks are found, each reporting the abundance of a different ion species in the islets. Each peak is characterized by a m/z value and three peak heights extracted from the scaled islet signatures of the WT, Ob/Ob, and SA/HF models. The list of peaks is available in the supplementary information as Table 1. The comparison of islet content is now translated into comparing the abundance of these 419 ion species in three pair-wise setups: WT versus Ob/Ob, WT versus SA/HF, and Ob/Ob versus SA/HF. The differences in peak height, expressed as ion counts, can be derived from Table 1 and are readily apparent in Fig. 7.7 and B.9 to B.13. However, the relationship established in the mass spectrometer between acquired ion counts and true abundance is complex and molecule-specific. This makes differences expressed as ion counts only useful for comparing the same ion species from signature to signature, but says little about expression changes from one molecular species to another. Thus, there is a need to transform the peak height changes into more generally comparable relative fold changes that are not tied to the absolute peak heights, which are largely ion-specific. We suggest a representation using the ratio of peak height in one islet signature over the peak height in the other signature. Given the absolute values involved, the range of ratios can become very wide. To counter this effect and aid interpretation they are put on a logarithmic scale of base two, effectively comparing peak heights as log ratios. Such a log ratio indicator serves the qualitative goal of strongly highlighting ion species that have a peak appear or disappear. However it also serves the more nuanced quantitative aim of providing an indication of fold change from one tissue class to another. The log ratios for the selected ion peaks are shown along the m/z axis in supplemental Fig. B.14 for the three pair-wise comparisons. Given the broad range of calculated log ratios, supplemental Fig. B.15 provides a zoomed-in version to aid further interpretation. The log ratios are also available as separate columns in Table 1.

Although automated peak picking followed by a log-ratio-versus-m/z representation successfully highlights the differences in islet content, its interpretation is hampered by a MS-specific issue. A ratio-based approach takes absolute peak height out of the equation. However a high absolute peak height exhibits a better signal-to-noise ratio

(SNR) and is therefore more reliable than a low absolute peak height. Also, in a purely ratio-based comparison small changes between short peaks can produce equally large fold changes as larger peaks, although there is less basis (read ion counts) to support those small differences. It is clear that a comparison only in relative terms could potentially be misguiding, and that there is a need to rank according to SNR or absolute peak height as well. We suggest a scatter plot that shows relative abundance changes as log ratios along the vertical axis and ranks the ion species according to absolute peak height along the horizontal axis. An example is shown in Fig. 7.8, which summarizes the differential comparison between the Ob/Ob islets and the WT islets. Each point in the plot is one of the 419 ion species that are being compared. To keep the plot interpretable only a few ions have been labeled with their corresponding (rounded) m/z value. The position of an ion species along the vertical axis indicates the expression change between the Ob/Ob signature and the WT signature. The position along the horizontal axis gives an indication of the signal-to-noise ratio linked to that expression change. More precisely the horizontal position indicates the absolute height of the higher of the two peaks that make up the comparison. Ions where both the Ob/Ob and WT peaks are small have less basis for a reliable fold change calculation and end up on the left hand side of the plot. The farther to the right, the higher the peaks involved and the more reliable the fold change calculation. A reliability threshold of 20 normalized ion counts is indicated as a dashed line. The ion expression change along the vertical axis is indicated as a fold change of Ob/Ob over WT, expressed as a log ratio of the Ob/Ob peak height over the WT peak height. Positive values indicate ions that are upregulated in Ob/Ob compared to WT. Negative values indicate downregulated peaks. As the base of the logarithmic scale is two, an ion located at a value of 3 (e.g. m/z 3396) exhibits a 23 or 8-fold upregulation in Ob/Ob, while a value of -1.5 (e.g. m/z 3769) exhibits a 21.5 or 2.8-fold downregulation. Due to the sensitivity of the technique and the quality of our data, fold changes smaller than two are considered rather unreliable and difficult to differentiate from no change. Therefore, our analysis will consider ions that fall within this range as 'unchanged' in expression. Dotted lines indicate the +1 and -1 log ratio values, indicating resp. 2-fold up- and downregulation. Similar plots for the SA/HF versus WT and Ob/Ob versus SA/HF setups are available in supplemental Fig. B.16. Figure B.17 gives a zoomed-in version of these plots for easier interpretation (Fig. 7.8 was already a zoomed-in version of the Ob/Ob vs. WT comparison).

7.3.5. Translation of the differential comparison of islet content to a metabolic representation: Mapping differential findings to hormone processing pathways

The final question to be answered is: What can the differential findings tell us about the hormone processing pathways that appear to be activated or deactivated in these

diabetes models? This last part of the study aims to find the overlap between the chemical species tracked in our tissue experiments and the body of metabolic knowledge on that tissue available in literature and databases.

After preprocessing, multivariate filtering, and differential analysis, the measured masses can now be cross-referenced to a database of potentially relevant metabolites. The list of metabolites is extracted from prior information on cellular processing of precursor peptides/proteins, and can be narrowed to the tissue, anatomical region, or cell type being explored in the MSI experiments. This information is readily available for diabetes, explaining why this disease was selected as a model system for this study. By looking for the overlap between the theoretical masses of the metabolically related relevant chemical species and the masses of the detected ions in the experiments, the differential findings from earlier can be placed within a more systems biology context. In this manuscript, the expression of ions in the diseased mice relative to a healthy wild type mouse are overlaid on a schematic of hormone processing and cleavage products in the islets of Langerhans, as extracted from literature and databases. The goal is to present the differences observed at the mass spectral level in a visualization of hormone processing known to take place in the islets of Langerhans. Putting the results in a cleavage pathway form goes beyond an exhaustive list of biomarkers, and highlights where a metabolic pathway starts deviating from normal in these diabetes models. Having such metabolic information available straight from tissue provides a fast way of obtaining deeper insight into the underlying pathomechanism and can help steer hypothesis building.

The islets are a complex mixture of at least five different cell types, designated alpha, beta, delta, epsilon, and PP cells. Each cell type is responsible for a different set of hormones related to digestion and glucose homeostasis. As the spatial resolution of the MSI applied does not allow for a differentiation of the different cell types individually, the extracted islet-specific signatures report on the content of all these cell types. Using these cell types as a guide, eleven potentially relevant metabolic pathways are extracted from the UniProt database (depicted in Fig. 7.9, only cleavage products are shown). For the alpha cells, we track the hormonal processing of the glucagon precursor and its thirteen possible cleavage products (including pro-peptides as well as true hormone products). For the beta cells five different metabolic pathways are tracked: the processing of both the insulin-1 and insulin-2 precursors (seven molecular species each), the cleavage of the islet amyloid polypeptide (IAPP) precursor (six species), the processing of the insulin-like growth factor II (IGF2) precursor (nine species), and finally products from the chromogranin-A precursor. This last processing pathway however consists primarily of very large molecules, of which pancreastatin is the only one that falls within the scanned mass range. From the delta cells we trace the presence of the seven species tied to the somastatin precursor, and from the epsilon cells the ghrelin precursor and its six possible products are examined. The PP cells' primary product is of course the pancreatic prohormone (pancreatic polypep-

tide; PP) and its five representative species are included into the analysis. Besides the nine metabolic pathways linked to the islet cell types, we also included two additional pathways known from the literature to have ties to the islets and their products: one is the processing of the proSAAS precursor (twelve members), and the other captures products from cholecystokinin (CCK) processing (eight members). Whether a particular cleavage product is supposed to come to expression within the pancreas or not is currently not taken into account as such information is usually not absolute and we want to avoid prematurely narrowing the screening results. Only cleavage products whose theoretical mass lies outside of the measured mass range are pruned. Besides the visual representation of the pathways shown in Fig. 7.9 and B.20, you can find a table of the molecules of interest in the supplementary information as Table 2, and a second version of the table ordered according to theoretical mass in Table 3. Essentially the method requires a library of metabolically related molecules (relevant to the tissue under study) to be extracted from literature or curated databases. In our case the extraction was done manually from the 'cleavage product' indication provided in UniProt, but the automated extraction of such information via text mining and multiple database interrogation is an obvious research target for the future.

The correlation between the library of metabolic pathway members and the experimentally tracked ions is accomplished through straightforward matching of the theoretical masses in the library (increased by one[1]) to the m/z values of the differentially expressed ions (as shown in Fig. 7.8, B.16, and B.17). As the mass resolution from the experiments does not reach unit precision and some noise in the measurements cannot be avoided, this matchup is performed within a preset accuracy of plus or minus 20 Daltons. This should allow some level of noise along the m/z axis while not going as far as to add or substract an amino acid. Using this 20 Dalton window, there are 60 unique assignments (one ion – one name), 23 double assignments (one ion – two names), 6 triple assignments (one ion – three names), and zero assignments of more than three possible metabolic nodes to a single ion. In summary 89 of the 419 peaks are assigned a metabolic node, and 49 of the 57 metabolic nodes (theoretically detectable within our mass range) are assigned an ion peak with expression information. The final step is to lift the differential findings from their mass spectral context to a biological context through an appropriate visualization. We suggest a color-coded mapping of the relative expression of the ions depicted in Fig. 7.8, B.16, and B.17 to a graphical representation of the metabolic pathways. Figure 7.9 demonstrates our suggestion for the Ob/Ob versus WT case, while supplemental Fig. B.20 does the same for SA/HF versus WT. The coloring of the nodes in the metabolic representation follows the same thresholds on fold change and peak height as set earlier. Upregulation beyond two-fold is indicated in green. Downregulation beyond two-fold is indicated

[1] When a molecule of mass M is ionized in positive mode it gains a proton, accounting for a mass increase of one Dalton. MALDI ionization produces primarily single charged ions, meaning that the charge z is normally one. The mass-over-charge ratio reporting molecule M in a mass spectrum is therefore (M+H)/z, which can be reduced to (M+1)/1 or simply M+1.

in red. Expression changes within two-fold are considered indiscernible and colored in yellow. The reliability of the fold-change calculation is taken into account as a threshold on the largest participating peak of twenty normalized ion counts. If both peaks fall below that threshold the coloring is adapted to a lighter version to indicate the reduced support for this expression observation. Supplemental Fig. B.18 explicitly shows the coloring scheme in the context of the differential plots that were used earlier. The area in which an assigned point falls, determines the coloring of its node in Fig. 7.9 and B.20.

7.4. Discussion

It is clear that characterization of dysregulated metabolic pathways in various diseases is critical for the development, the selection and evaluation of therapy as well as for the screening for sensitive and early responding biomarkers. Recent advances in metabolomic approaches are typically based on MS or NMR spectroscopy, which both provide a large amount of data and share a number of major challenges, particularly regarding the interpretation of the data obtained. However, in all these approaches, the spatial localization, a crucial property of biomolecules, is missing. Using MSI, this property can be recombined with the biochemical data resulting in a better defined differential analysis of molecules grouped together in similar anatomical regions.

This chapter presents a combined wet lab/computational protocol that merges biophysical characterization of multiple chemical species in parallel with a spatial context of known anatomical regions in tissue, filters out signatures of biochemical content for regions of interest, performs a differential analysis between disease and normal tissue, and maps the findings to a curated structure of metabolic interactions to aid interpretation. The overall aim is to demonstrate a technical platform that leverages the aspects of spatial co-localization and correlated expression between biomolecules to provide a more holistic view into disease while keeping the number of required experiments low.

In this study, the platform is used to explore pancreatic islet-specific hormone processing in a wild type mouse versus a chemically induced diabetes mouse model, DIA, and two genetically induced diabetes mouse models, Ob/Ob and SA/HF. The multi-molecular nature of MSI experiments and the presented computational analyses provide differential insight into the presence of insulin and other islet-produced hormones such as IAPP, glucagon-like proteins, and ghrelin. Moreover, by comparing the data with a list of known islet-related metabolites, a number of putative dysregulations in the metabolic processing of the islets becomes apparent. Although such observations require further confirmation through more traditional lower-throughput techniques, they have inherent value as a means of obtaining a wide range of metabolic information straight from tissue and as a means of steering further research by filtering

a list of interesting targets and hypotheses from the massive amount of data available. Essentially, the method lifts the differential findings collected in the mass spectral domain to a hormone processing pathway framework. It should be noted that this study uses the well-known diabetes mouse model systems primarily for delivering a proof-of-principle of how *in situ* imaging of multiple hormones and their cleavage products in pancreatic islets can provide an uncommon insight into hormone processing via the spatial behavior of several of the molecular players that are involved.

Imaging of insulin in tissue is a nontrivial topic in recent diabetes research, and various methods have been formulated [143, 144]. This chapter describes how MSI can be used to image insulin throughout pancreas tissue, and compares the insulin-reporting ion images to a traditional immunostaining. The ion image method provides localization up to the cellular level, but not to sub-cellular scales as with microscopy. However, the MSI based approach has several particular advantages. Since it is a completely unlabeled method, there is no need for an elaborate labeling scheme, neither by chemical means (e.g. immunolabeling), nor by genetic approaches (e.g. transgenic mice expressing a GFP attached to insulin). This avoids chemical perturbation of the tissue sample and enables experimentation on established disease models without modification. In fact, unlike most insulin-oriented methods, the MSI approach does not track co-localizing insulin-related fragments such as the C-peptide, but instead tracks true native insulin. Secondly, the method acquires its information *in situ* from the tissue without the need for resource-intensive physical excision (e.g. via laser capture microdissection) or chemical separation of the islets of Langerhans. We show that with proper computational methods the complex data content can be unraveled adequately to enable interpretation. The final benefit is that MSI provides these insulin-specific advantages while simultaneously measuring the presence of hundreds of other chemical species in the tissue as well. In summary, insulin imaging via MSI has a number of characteristics that avoid chemical or physical perturbation, and bring its measurements arguably closer to the true biological situation. Also, its multi-molecular acquisition and unlabeled nature make it particularly suitable for exploratory high-throughput screenings.

The second result uses the normalized peak height of insulin to estimate insulin concentrations across the different disease models. Although it is only a rough estimate, the observed pattern matches expectations regarding the various diabetes models, and demonstrates that with proper preprocessing a quantitative unlabeled estimate of quantity can be obtained.

The third result formulates a concrete data-mining approach where a matrix decomposition method is used as a computational means of filtering the massive amount MSI data for patterns of interest. Although MSI has seen the application of matrix decomposition methods before, they have been used primarily as a means of preprocessing (e.g. dimensionality reduction prior to clustering [78]), for denoising [74, 83], or to

help interpretation (e.g. [75, 76, 79]). However, in this study the matrix decomposition is used in a new filtering role, allowing for undesirable patterns to be removed from further analysis and focusing on particular mass spectral profiles of interest (e.g. in this case peak patterns specific to the islets of Langerhans). In addition to defining a new matrix decomposition role in MSI, a decomposition method new to the field of MSI needed to be introduced. As the subsequent differential analysis requires signatures that resemble true mass spectral profiles containing only positive values, a non-negativity constraint is introduced. This constraint rules out many methods popular in MSI such as PCA. Instead this chapter successfully introduces an established method called NMF, to perform the filtering role. When trying to reveal underlying structure in MSI data or performing the filtering sought after in this study, the main advantage of matrix decomposition methods over other unsupervised methods such as clustering, is that they not only consider correlated peak expression but simultaneously consider spatial co-localization as well. This aspect turned out to be crucial to obtaining the islet-specific content signatures we were looking for. The filtering step responds to the need for more advanced data mining of MSI data and in this particular implementation functions as a type of *in silico* enrichment of pancreatic islets, avoiding labor-intensive and invasive procedures such as physical excision from the tissue.

The differential comparison of islet content from one model to the next makes up the fourth result of this chapter. The extracted signatures are scaled and plotted opposite to each other to enable direct interpretation. Although the scaling of islet signatures relative to each other is vulnerable to the validity of using (normalized) ion counts as a means of estimating molecule quantity, and might impact expression estimations, the scaling will have little influence on the search for appearing or disappearing peaks. Nevertheless, the current scaling is supported by a panel of common ions, apparently uninvolved in the disease mechanism and showing matching peak height across the models. The perturbation of the processing of the glucagon precursor in the Ob/Ob model is used as an example of the multi-molecular insights that can be gleaned from this broad range screening technique. In addition to the manual example, an automated differential analysis method is introduced that captures both expression changes as well as a measure of reliability for these change calculations. The large cloud of differentially expressed ions in these plots serves to illustrate the vast amount of molecules being tracked simultaneously by this approach. It also helps the researcher to recognize the more promising ions that deserve further analysis. By including a mass spectrometry specific hint of reliability, the plotting method presented there goes beyond standard biomarker representations that usually take the form of Fig. B.14 and B.15. Also, this new differential representation has no particular ties to the filtering method preceding it here, and should be equally useful in experimental biomarker setups where mass spectra are compared directly to each other.

The color-coded cleavage product visualization (Fig. 7.9 and B.20) in the final result

summarizes within a biological context the detected molecular content of the islets of Langerhans. It depicts the observed cross-section between molecules theoretically present at the islets, and molecules measured via mass spectrometry at the islets in these specific mouse models, while simultaneously highlighting putative expression changes. The color coding is obviously tied to the particular up and downregulation thresholds that are employed. As these thresholds are pushed downward and the sensitivity of the underlying mass spectrometry instruments increases, more subtle up and downregulation will be revealed for products that are currently labeled in yellow. By making the link between information sources, a multi-molecular imaging method such as MSI can be used to explore the intracellular processing of individual biomolecules such as hormones across different disease models, thereby generating data not easily obtained via traditional approaches such as immunostaining.

The reasons for making the link to metabolic pathways at the ion level rather than performing standard proteomic identification on these ions first, are multiple. First and foremost, this is meant to be a high-throughput screening method to quickly highlight differentially behaving metabolic pathways for further investigation. Mass spectral identification is currently still a non-trivial step and an active area of research, requiring expert knowledge, sufficient sample material, and a reasonable amount of time and work. These properties are hardly compatible with a high-throughput screening method that emphasizes wide-range coverage rather than a specific focus on particular molecules. Secondly, the screening method is aimed at the early lead discovery phase where the avoidance of false positives comes second to providing a broad insight. It is clear that using only the theoretical and observed mass to link a metabolic pathway member to an ion in the tissue has the potential for wrong assignment. However the chances of wrong assignments across several different members of a processing pathway are arguably smaller, and even then a false positive pathway will be found out through further investigation. The possibility for wrong assignment is also diminished by use of the spatial information uniquely provided by MSI. Having the biological scope spatially narrowed to a particular cell type or anatomical region, and further aided by filtering steps as described in the third result, means a smaller list of potential candidates. Thirdly, as the field of mass spectral identification develops instrumentally and computationally, identification direct from tissue is bound to become more reliable, and as the screening method is successfully demonstrated at the ion level, there is little doubt that in the future the presented strategy can be further strengthened by tying directly into identified molecules.

It should be clear that this platform does not have to be restricted to the pancreas. If sufficient prior information on metabolic pathways is available in the area of study, the MSI-based metabolic screening protocol can be applied in any biomedical or plant-based area of research that can supply organic tissue samples of sufficient quality. Besides the biological observations related to diabetes mellitus that are obtained through this new approach, it is clear that the computational methods developed in

this manuscript can be applied equally well in other disease cases to help elucidate the underlying pathomechanism. The particular advantage of this approach lies in the new type of insight it can provide with regard to disease mechanisms by providing putative areas where diseased metabolic processing starts deviating from normal processing. This type of information was previously unavailable or very difficult to gather practically using traditional techniques, especially for a large enough complement of molecules to be able to consider the experimental setup a screening setup. The method represents a novel combination of information sources, data mining of previously untouched aspects in biochemical tissue content such as co-localization and correlated expression, the incorporation of prior information from databases into the MSI analysis, and mapping to a more telling visualization of the results to aid biological interpretation. It is demonstrated that having both the biochemical characterization and spatial aspects available has great potential for the elucidation of biochemical pathways and protein processing, as the products of these operations tend to manifest themselves in a co-localized fashion.

7.5. Methods

7.5.1. Mouse models

The pancreas of one healthy and three disease mouse models are compared. The healthy wild type mouse, labeled WT, is a standard BL57/6 mouse. The chemically induced diabetes model, labeled DIA, is a BL57/6 mouse where the islet function has been destroyed by the introduction of streptozotocine. The Ob/Ob and SA/HF mice are two well-studied genetic models for type 2 diabetes mellitus.

7.5.2. Tissue preparation

The pancreas is harvested from each mouse and immediately flash-frozen using liquid nitrogen. The frozen pancreas is then cut into sections using a standard microtome. Three neighboring sections are captured for further analysis. The first section (5-7 µm thick) is captured on a standard glass microscopy slide and stained with hematoxylin and eosin (H&E). The second section (5-7 µm thick) is captured on a standard microscopy slide as well and undergoes insulin (red) and cell core (blue) staining. The third section (10 µm thick) is captured on electrically conductive ITO-coated glass slides (Bruker Daltonics, Germany) to enable MSI measurement. This third section is left unstained and is only washed with ethanol 30 %. All experimental conditions were kept constant as much as possible between the experiments to allow cross-comparison between the various mouse models.

7.5.3. Mass spectral imaging experiment

A thin tissue section (10 µm) is cut from the flash-frozen pancreas using a standard microtome. The section is mounted on an electrically conductive ITO-coated glass slide (Bruker Daltonics, Germany) suited for measurement in a MALDI based mass spectrometer. The glass slide with the tissue is washed with a gradient concentration of ethanol. Subsequently, a layer of chemical matrix solution (alpha-cyano-4-hydroxycinnamic acid) is deposited on the tissue section to enable MALDI-based ionization of the molecules embedded in the tissue. The deposition is done via a robotic acoustic droplet ejection system, the Portrait 630 (Labcyte Inc., USA), to ensure homogenous matrix coating and to increase reproducibility across measurement locations and tissue sections. The tissue carrying glass slide is introduced into the mass spectrometer, an Applied Biosystems 4800 MALDI-TOF/TOF Analyzer (Applied Biosystems Inc., USA). Using the MSI4000 software, freely available at www.maldi-msi.org, a grid of measurement locations across the tissue is defined. A mass spectrum is collected at each grid location and stored into a data array structure for preprocessing and further computational analysis. Each mass spectrum is the accumulated result of 400 laser shots at its tissue location. The spectra cover the molecular mass range from 2000 Da to 12000 Da using 12380 m/z-bins. The spatial resolution of all four experiments is 50 µm distance between neighboring measurement locations. As a result the WT experiment captured 9880 mass spectra on a grid of 104 $\times$ 95 pixels. The DIA experiment describes 8455 mass spectra on a grid of 89 $\times$ 95 pixels. The Ob/Ob and SA/HF experiments collect respectively 10414 tied to 127 $\times$ 82 pixels and 10769 mass spectra from 121 $\times$ 89 pixels. As the measurement grids have to encapsulate the outline of the tissue sections, a reasonable amount of measurement locations falls outside the tissue. To avoid the introduction of variance from non-tissue sources and to reduce the complexity of the data sets, tissue masks are defined manually for each tissue section (shown in Fig. B.1). Only mass spectra that fall within the green tissue area are used for the study, while mass spectra from the blue area are discarded. This reduces the data sets to 6434 spectra for WT, 9002 for Ob/Ob, 6441 for DIA, and 8695 spectra for SA/HF. To enable comparison between measurements at different locations in the tissue or across different tissue sections, two additional preprocessing steps are applied to the data sets. First the peak height in the spectra is normalized across all four data sets by scaling the spectra towards a common total ion count (TIC) value, which is taken to be the median of all TICs found in the raw spectra of all four experiments. Then the baseline is removed from each spectrum using MATLAB's standard baseline correction function using a window width of 150 Da and a step size of 150 Da. The effect and necessity of the preprocessing steps is illustrated in Fig. B.2.

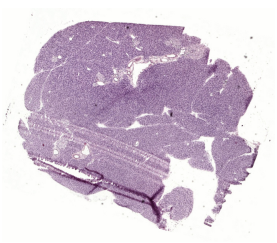
7.5.4. MSI data and ion images

The data from a MSI experiment consists of a grid of measurement locations spanning the tissue section. Every grid element has a complete mass spectrum connected to it. It is usually represented as a three-mode data array or tensor with two spatial modes (x and y) and one mass-over-charge mode (m/z). For the four MSI experiments on pancreas tissue, the acquisition parameters were kept identical, meaning that the mass range and the number of m/z bins is the same for all four data tensors. However, the tissue sections differ in size, shape, and orientation, causing the x and y mode to differ from model to model. The content of the wild type tissue was captured in a data tensor of size $104 \times 95 \times 12380$ or 122 million values. The DIA tensor is of size $89 \times 95 \times 12380$ (105 million values). The Ob/Ob and SA/HF tensors are respectively of size $127 \times 82 \times 12380$ (129 million values) and size $121 \times 89 \times 12380$ (133 million values).

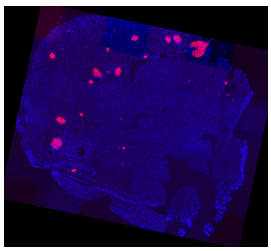
An ion image showing the spatial abundance of a particular m/z value, can be considered a cross-section of the MSI data tensor at that particular m/z bin. Cutting the tensor at a specific value along the m/z axis produces a matrix array that holds the corresponding ion counts and can be shown as a false color image overlaid on the tissue.

7.5.5. Mapping

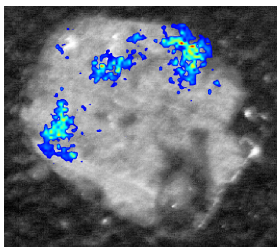
The color code of Fig. B.18 was transferred directly to the metabolic pathway representations of Fig. 7.9 and B.20. Due to the peak picker algorithm's insistence on reporting the ion amount through peak height at the peak centroid and not through integration over a small m/z window, two molecules in the Ob/Ob study (glucagon and pancreastatin) and three molecules in the SA/HF study (preptin, GLP-1, and somatostatin precursor signal peptide) were compared using a peak shoulder rather than the true peak maximum. This was manually adjusted.



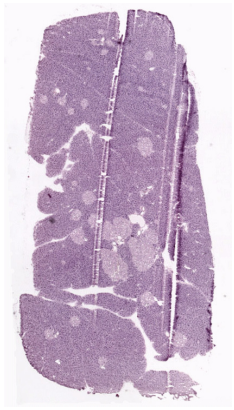
(a) WT – H&E staining



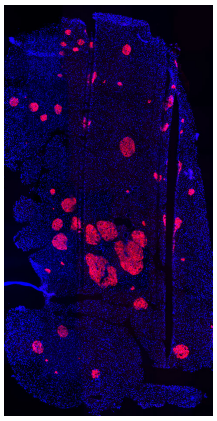
(b) WT – insulin staining



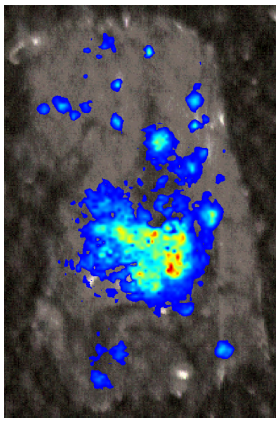
(c) WT – ion image



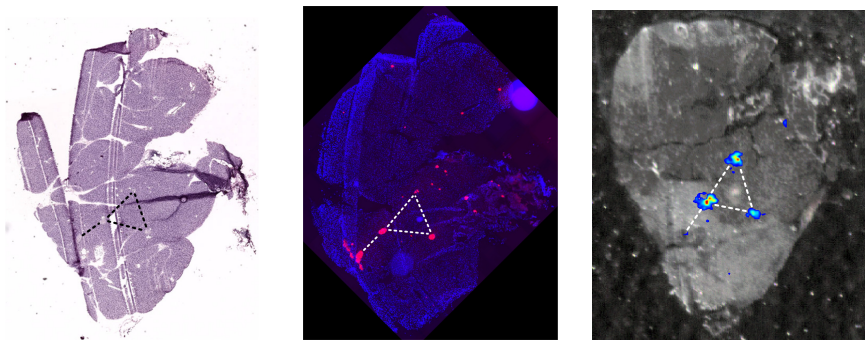
(d) Ob/Ob – H&E staining



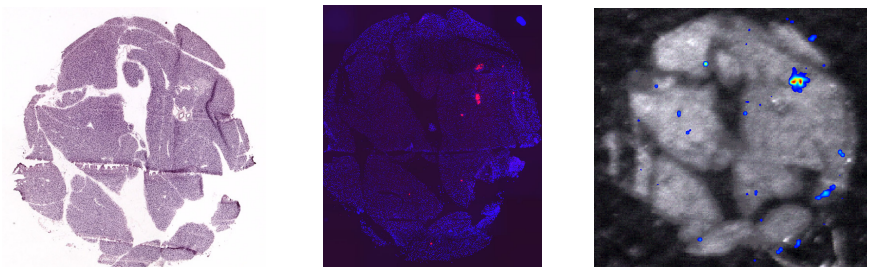
(e) Ob/Ob – insulin staining



(f) Ob/Ob – ion image



(g) SA/HF – H&E staining (h) SA/HF – insulin staining (i) SA/HF – ion image



(j) DIA – H&E staining (k) DIA – insulin staining (l) DIA – ion image

Figure 7.3. Imaging of insulin in pancreatic tissue: the traditional labeled approach versus the unlabeled MSI approach. Each row shows a comparison within a mouse model class: (1st row) WT ; (2nd row) Ob/Ob; (3rd row) SA/HF; (4th row) DIA. (left) H&E stained microscopic image, introduces general shape of tissue sections. Notice slight texture changes corresponding to beta cell islets. (middle) Fluorescence microscopy image, retrieved via an established staining for insulin (red is insulin, blue is cell core). (right) Ion image, reports spatial distribution of ions corresponding to insulin, and is overlaid on higher-resolution gray level picture of the tissue. Ion image sensitivity does not match that of immunostainings, but is sufficient to locate most islets without the need for labeling. As the different imaging methods are not mutually compatible within a single tissue section, every method is applied on directly neighboring sections from the same pancreas. The general shape and physiology is therefore largely maintained from one column to the next, but minor warping effects of the tissue can be observed. The structure of the islets of Langerhans is well maintained across all imaging modalities. (As the H&E and insulin stained tissue sections of the SA/HF model were damaged in the cutting process, a corresponding constellation of islets is indicated on all three images to aid interpretation.)

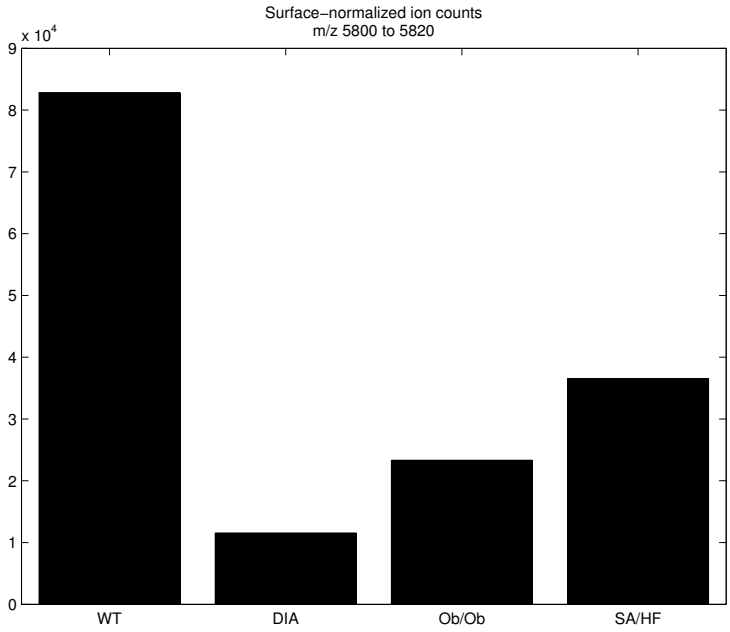
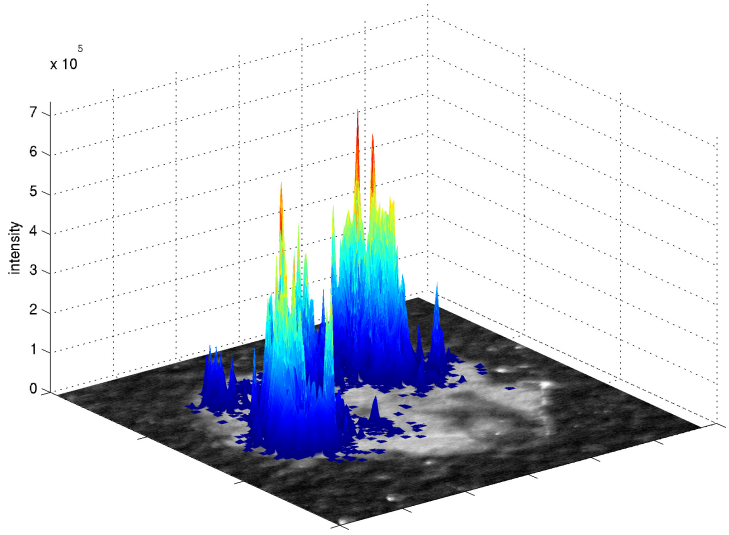
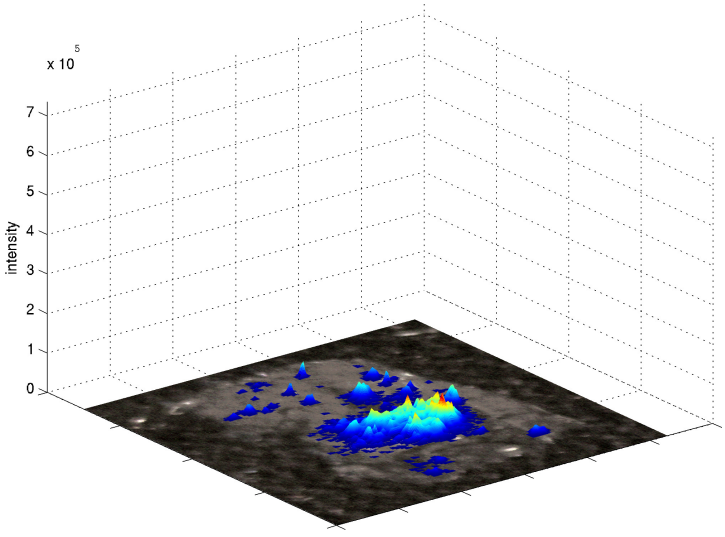


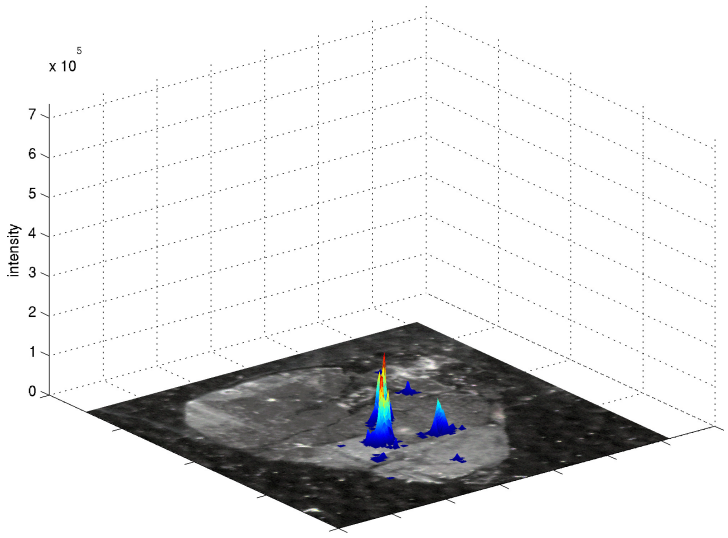
Figure 7.4. Comparison of surface-normalized insulin concentration between the four mouse models. As all four tissue sections have a different shape and surface area and the size and amount of insulin containing beta cell islets varies substantially between the different mouse models, a proper normalization for surface area is required. In all four experiments a hard threshold of at least 10000 preprocessed ion counts within the m/z 5800 to 5820 range was used to identify the insulin containing pixels in the image grid. The threshold corresponds to 1.36% of the highest detected insulin ion count in the experiments and selects 1611 insulin-containing pixels of a total of 5273 pixels in WT, 1507 of 8968 in Ob/Ob, 167 of 7957 in SA/HF, and only 4 of 6440 in DIA. For each model, the total sum of insulin ion counts (collected from the insulin-containing pixels) is divided by the number of insulin-containing pixels. The plot shows average insulin ion counts per pixel where insulin was detected.



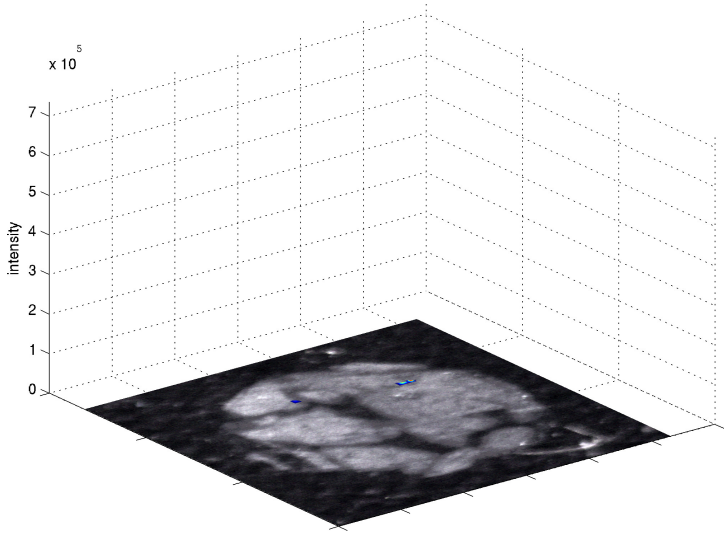
(a) Insulin ions in WT



(b) Insulin ions in Ob/Ob



(c) Insulin ions in SA/HF



(d) Insulin ions in DIA

Figure 7.5. Surface plot visualization of insulin content in the four mouse models. Due to the strong differences in insulin concentration between the four models, a common color range cannot suitably highlight the relative differences between the four experiments. Instead a surface plot representation is used where the intensity is plotted along a common height scale. In addition to making the relative differences in concentration more interpretable, this representation also provides insight into islet size and surface area spread.

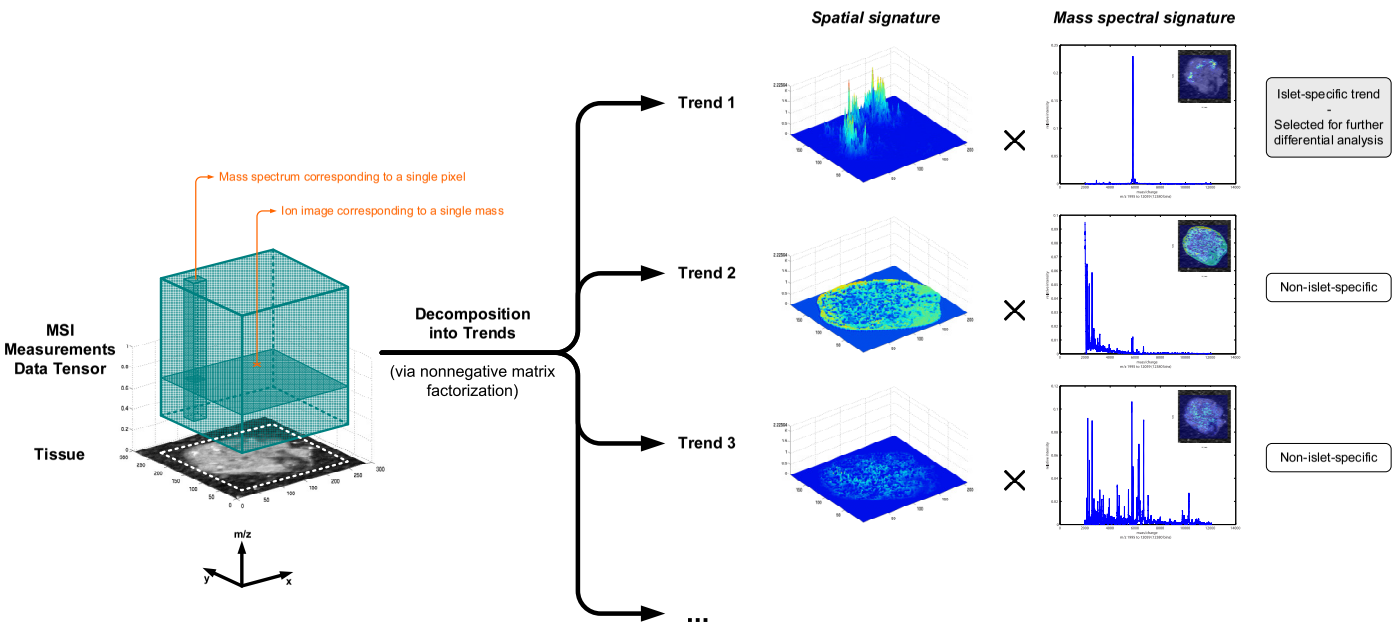


Figure 7.6. Extraction of a set of data trends from the MSI measurements via nonnegative matrix factorization. Each trend describes a set of co-locating ions and defines an area in the tissue where these ions show up together. A spatial signature and a mass spectral signature characterize each trend. The mass spectral signature can be considered a pseudo-mass spectrum showing only the ion peaks connected to this trend and defining a relative expression pattern for these peaks. The spatial signature describes a distribution in the tissue where the trend ion panel shows up in the given relative expression, and can be visualized as an image overlaid on the tissue. We perform decomposition into six trends on each of the examined mouse models. The trend describing the content of the islets of Langerhans is selected for further differential analysis between the different mouse models.

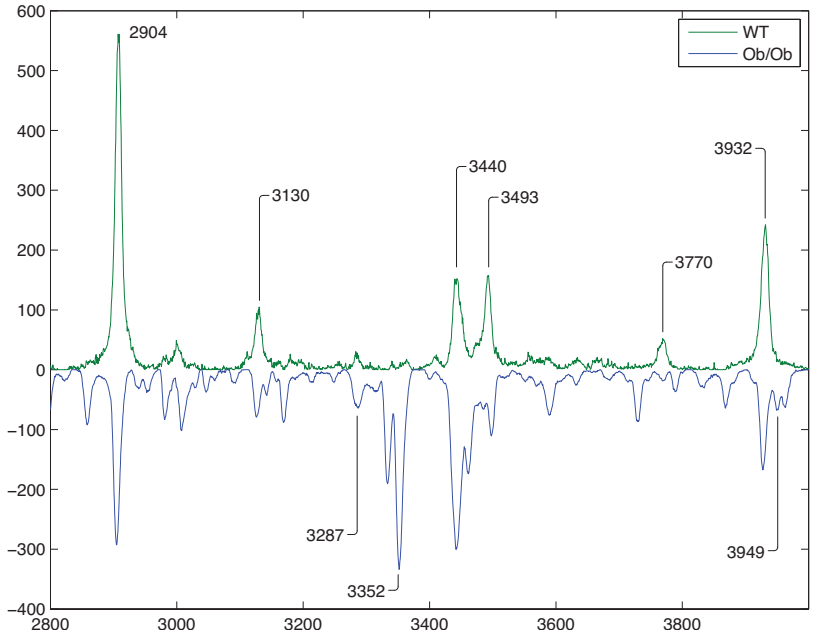


Figure 7.7. Comparison of islet specific signatures between WT and Ob/Ob, detail from m/z 2800 to 4000. Full signatures are available in supplementary Fig. B.9.

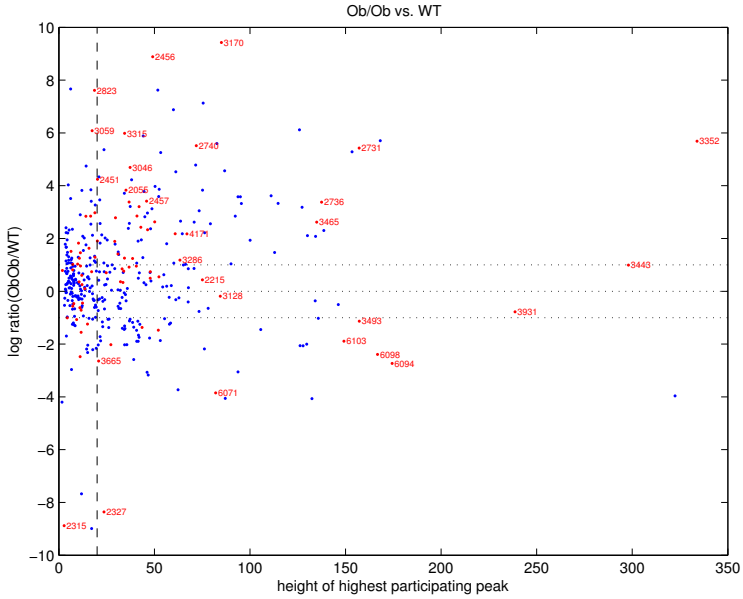


Figure 7.8. Differential fold changes between WT and Ob/Ob. Each dot is an ion (peak) found in the islet signatures. A selection is labeled with their corresponding m/z value. Expression change is plotted along the y-axis. The position along the x-axis gives an indication of the signal-to-noise ratio linked to that expression change. The ion expression change is indicated as a fold change of Ob/Ob over WT, expressed as a log ratio of the Ob/Ob peak height over the WT peak height. Positive values indicate ions that are upregulated in Ob/Ob compared to WT. Negative values indicate downregulated peaks. As the base of the logarithmic scale is two, an ion located at a value of 3.4 (e.g. m/z 2736) exhibits a 23.4 or 10-fold upregulation in Ob/Ob, while a value of -1.9 (e.g. m/z 6103) exhibits a 21.9 or 3.7-fold downregulation. Fold changes smaller than two are considered *no reliable change*, and these ions are labeled as *unchanged* in expression. Dotted lines indicate the +1 and -1 log ratio values, indicating resp. two-fold up and downregulation. Along the x-axis the ions are ordered according to the height of their highest participating peak in the comparison. Ions where both the Ob/Ob and WT peaks are small have less basis for a reliable fold change calculation and end up on the left hand side of the plot. The farther to the right, the higher the peaks involved and the more reliable the fold change calculation. A reliability threshold of 20 normalized ion counts is indicated as a dashed line.

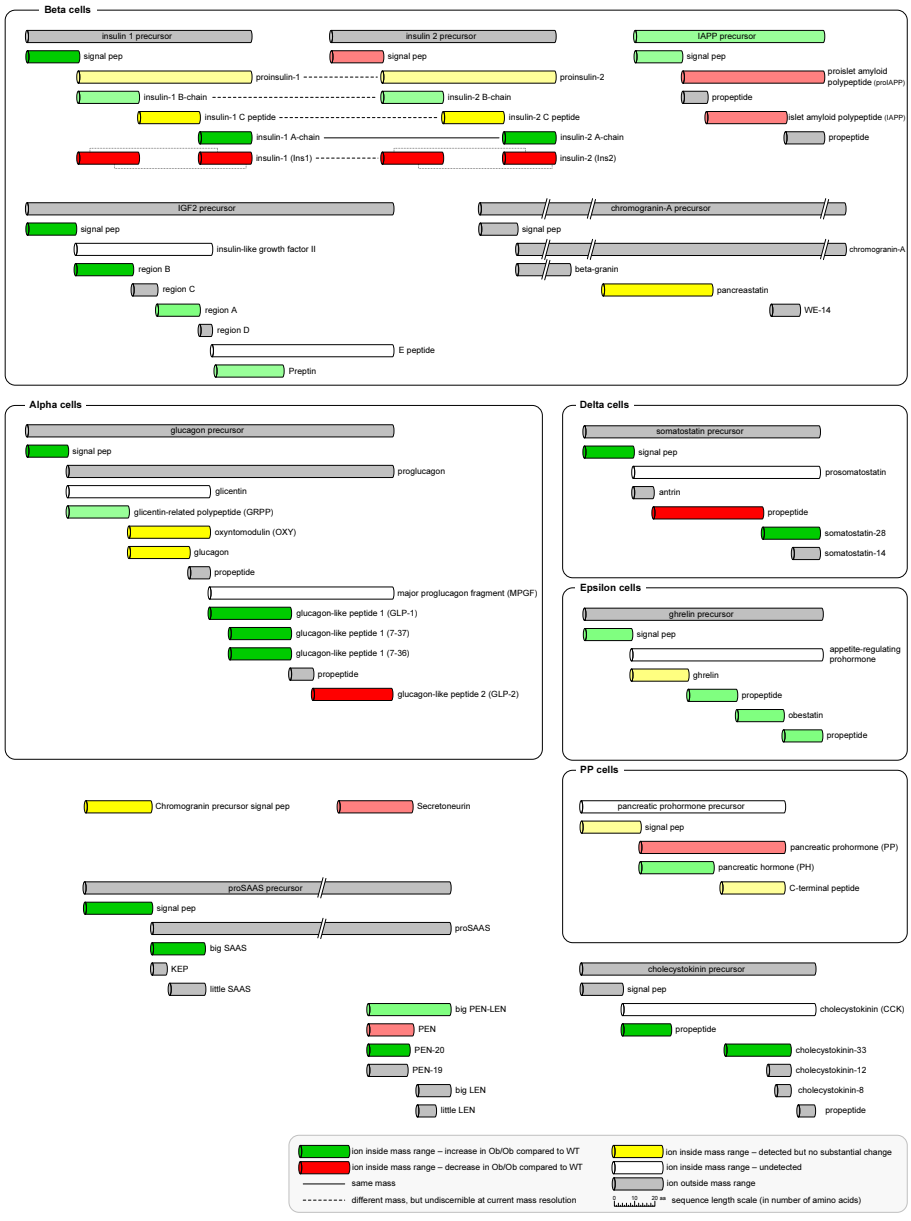


Figure 7.9. Hormone processing and cleavage products in Ob/Ob versus WT.

CHAPTER 8

Conclusions and Perspectives

From the moment I picked up your book until I laid it down, I was convulsed with laughter. Some day I intend reading it.

Groucho Marx

8.1. Conclusions

Mass spectral imaging is a novel molecular imaging technology that employs the sensitivity and specificity of standard mass spectrometry to measure the spatial distribution of molecules throughout organic tissue sections. Its primary advantages are that it requires no prior labeling of the molecules of interest, and that it is capable of concurrently tracking hundreds of molecules within the same tissue experiment. Both aspects make MSI particularly well suited for exploratory analysis of tissue, and MSI is gaining considerable momentum in the tissue biomarker community as a result. Although until recently most MSI development has been dominated by physics and chemistry advances, the massive amounts of data generated in a MSI experiment have made it clear that the computational analysis of MSI will become one of the most pressing challenges in the field. The data from an MSI experiment is typically huge in size, extremely complex, and contains much more information than is currently exploited. In this doctoral thesis, various strategies are explored and developed for the analysis of MSI experiments from a computational standpoint.

Several preprocessing methods were explored and developed to enhance peak normalization, noise filtering, and reliable data compression. A new preprocessing method

for enhanced normalization (IEC, Ionization Efficiency Correction) was developed, facilitating the direct comparison of mass spectra from different tissue regions or from different samples. The importance of this development is that it makes spectra directly comparable in both standard mass spectrometry as well as MSI experiments. The method thereby boosts the reliability and information that can be gained from ion images, and improves the results of any high-level analyses further downstream. As MSI moves towards more differential analyses across different tissue samples, reliable and robust normalization of the spectra will become a key issue.

Several unsupervised methods (such as PCA, PIW-PCA, and NMF) were applied to disassemble MSI data into underlying patterns, useful for interpretation, filtering, denoising, and dimensionality reduction. It is demonstrated that different algorithms will be suitable for different applications, and the superiority of NMF for human interpretation is highlighted.

A new supervised method (spatial querying) was developed, aimed at answering a particular biological question. This method enables MSI data to be queried from the spatial domain rather than the traditional mass-centric approach. Such an approach can be useful for diseases demonstrating well-defined anatomical regions (e.g. Parkinson's disease, Huntington's disease, Amyotrophic lateral sclerosis,...).

A pre-analysis dimensionality reduction method was developed to enable analysis of MSI experiments that would otherwise exceed available computational and memory resources. Using wavelet transformation, the MSI data is compressed into a lower-dimensional space. This method also provides a way for efficient data storage and has denoising characteristics.

The co-localisation network analysis is another newly developed high-level unsupervised method. This method provides a visualization of inter-molecule relationships in a MSI experiment, such as co-localization and correlated expression. The method proves very informative when looking for co-localizing partners, and this without having to explicitly traverse the thousands of ion images present in the data set.

The doctorate is concluded by a biomedical case study of diabetes mellitus type 2. Here the combination of mass spectral imaging and the computational techniques developed in this work, is applied to explore metabolic processing of various cellular regulators present in pancreatic islets. This example provides evidence for the intrinsic analytical potential of MSI as a powerful exploratory technology to screen for chemical (metabolic) alterations in tissue. It also highlights the absolute requirement of specifically designed mathematical approaches to unlock this intrinsic potential.

8.2. Future Perspectives

The push towards more computational and statistical analysis of MSI data, signals two things about mass spectral imaging as a technology. First of all, MSI is slowly but surely becoming a trustworthy source of information in biomedical research. A move towards data mining in a new technology typically happens when the technology has reached a certain maturity and the biggest physical or chemical issues are under control. By definition, data mining and statistics can only come into play once the technology becomes repeatable and trusted. The second aspect, brought on by increased computational analysis, is that new avenues of research open up. By adding computer science to the physics and chemistry mix that is mass spectral imaging, opportunities for different insights into the data open up, and it becomes possible to extract information that otherwise would remain buried under the noise or high-dimensionality of the data.

With regards to scientific research, mass spectral imaging ties together three distinct disciplines: biological/biomedical research, instrumental/analytical research, and machine learning/statistical research. As the successful analysis of a MSI experiment draws on ideas and techniques from all three disciplines, it stands to reason that MSI will benefit from new developments in any and all of the three areas.

Biomedical perspectives for MSI focus on it becoming part of the clinical path. As the mass spectrometry technology itself is still under fast development, various instrumental perspectives are already offered currently. These include a projected increase in MSI's spatial resolution, adding new dimensions to the data (e.g. ion mobility separation, fragmentation analysis,...), and overall improvements in accuracy, sensitivity, and mass resolution. Although the computational methods described here are fully operational, it should be clear that they are but a starting point for further development. Examples for future development include methods for differential analysis, kernel-based clustering and classification methods, the increased incorporation of non-linearity, the deepening of mass spectrometry statistics, and a further fine tuning of compression techniques to finally enable differential analysis across multiple large data sets with minimal loss of biochemical information.

APPENDIX A

Definitions

A.1. Units

dalton (Da) The unified atomic mass unit (u), also called the dalton (Da), is defined as one twelfth of the mass of a carbon-12 atom [66, 145].

$$1 \text{ u} = 1 \text{ Da} = 1.6605402 \times 10^{-27} \text{ kg}$$

charge unit (e) The charge unit is the charge of a single electron, and is an absolute value and a physical constant [66].

$$e = 1.602176487 \times 10^{-19} \text{ C}$$

m/z unit (Th) Mass to charge ratio of an ion is an abstract, unitless number, usually written as $m/z = 80$. Sometimes an explicit unit for m/z is used, called the thomson (Th).

A.2. Definitions

biomarker Any indicator of a particular biological state of an organism. In medicine, a biomarker usually refers to a metabolite or protein in an easily accessible fluid such as blood or urine, whose concentration reflects the presence and severity of the disease.

ion An atom or molecule that has a net electric charge.

charge (z) The number of charges an ion carries, or the ratio of the charge of a particle to the elementary charge [66].

m/z The m/z ratio of an ion is the dimensionless quantity formed by dividing the mass number of the ion (in Da) by its charge number (in e) [66].

desorption The formation of ions in the gas phase from a material deposited on a solid surface.

ionization A process that yields one or more ions from an atom or a neutral molecule [145].

assay An assay is an analysis or procedure to determine the presence or absence of one or more molecular species in an organism or an organic sample. A quantitative assay also measures the quantity of its target analyte in the sample.

total ion current The sum of the separate ion currents carried by the different ions contributing to the spectrum [66]. From a mathematical standpoint, the sum of all ion counts in a mass spectrum irrespective of ion species, or the integral over the mass spectral profile.

ionization efficiency The ratio of the number of ions formed to the number of electrons or photons used in an ionization process [66].

nonstationarity A time series is nonstationary if its statistical properties vary over time. Although there is no time-dimension, a mass spectrum can also be considered as nonstationary, as its statistical properties change depending on the particular m/z bin one examines.

APPENDIX B

Supplementary Figures to Chapter 7

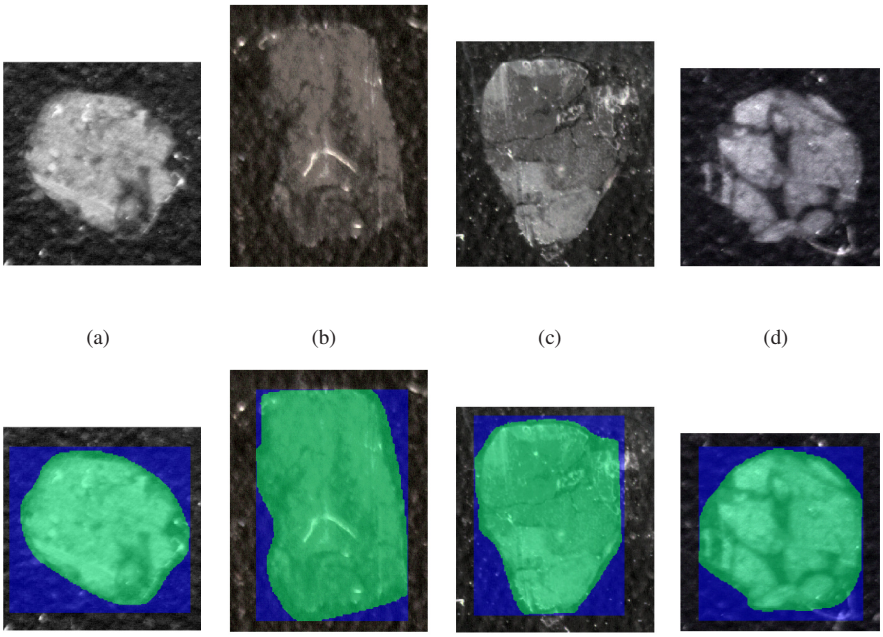
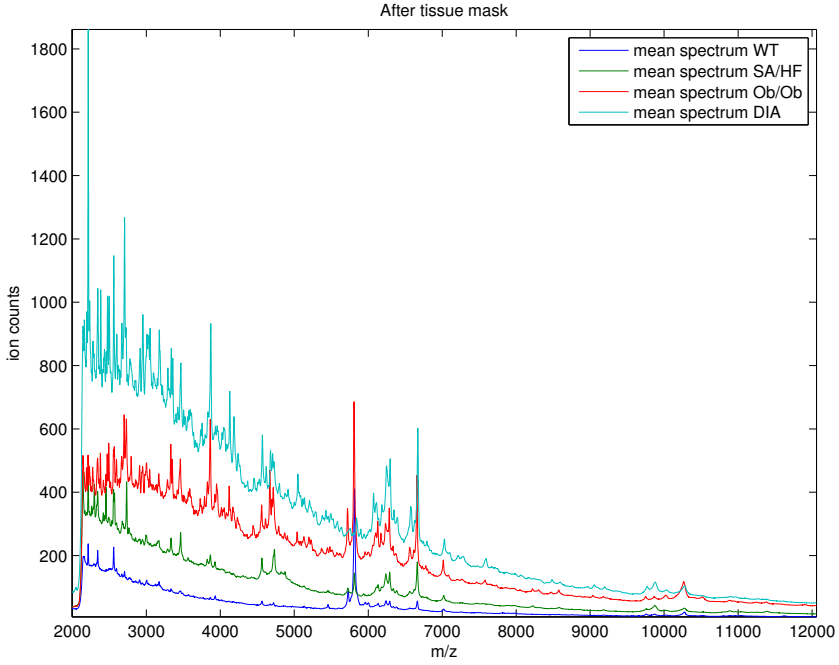
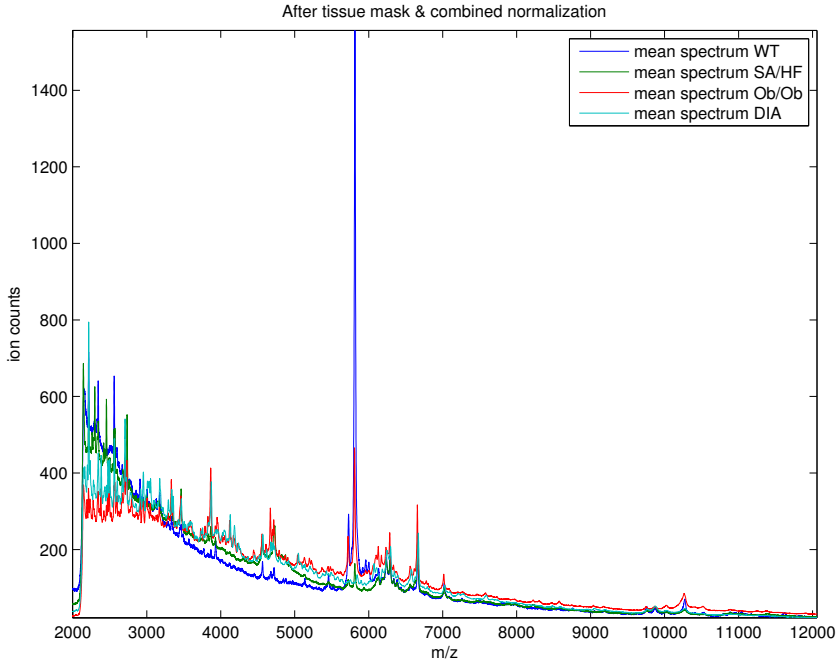


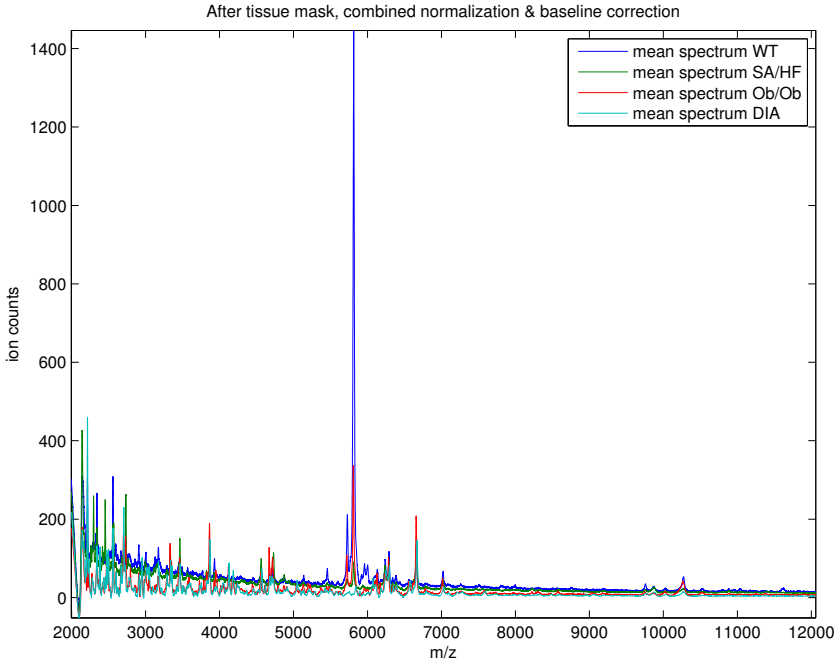
Figure B.1. Tissue masks defined for the different pancreas sections. Only mass spectra that fall within the green tissue area are used in the study. Mass spectra in the blue area are removed from the study to avoid the introduction of variance from non-tissue sources and to reduce the complexity of the combined data set. (a-d) Microscopic gray level images of the unstained tissue sections. (e-h) Tissue masks overlaid on top of the microscopic images. (a,e) WT ; (b,f) Ob/Ob; (c,g) SA/HF; (d,h) DIA.


After tissue mask
mean spectrum WT
mean spectrum SA/HF
mean spectrum Ob/Ob
mean spectrum DIA
ion counts
m/z
1800
1600
1400
1200
1000
800
600
400
200
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

(a)

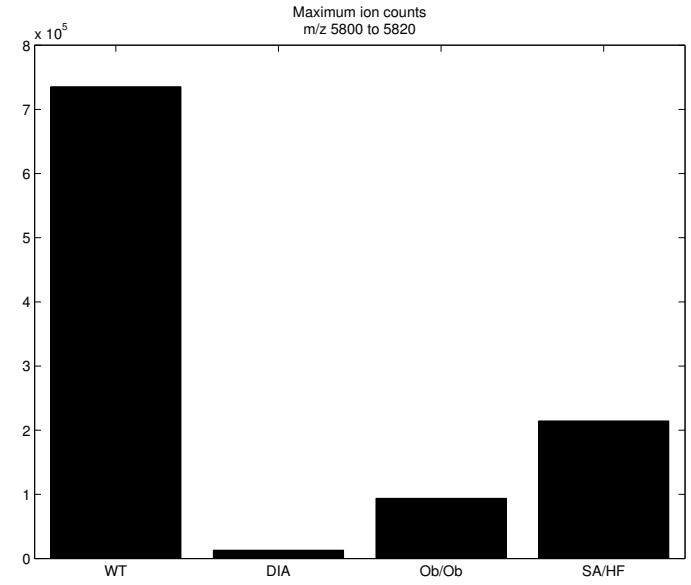

After tissue mask & combined normalization
mean spectrum WT
mean spectrum SA/HF
mean spectrum Ob/Ob
mean spectrum DIA
ion counts
m/z
1400
1200
1000
800
600
400
200
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

(b)

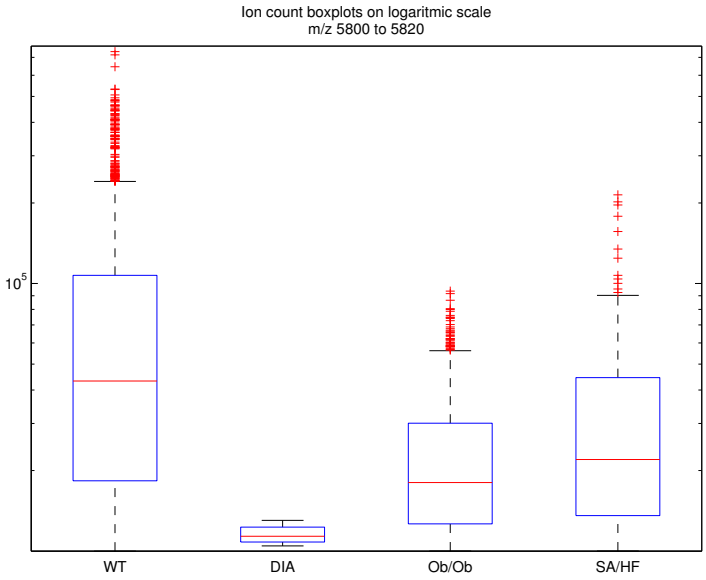


(c)

Figure B.2. The effect of preprocessing. (a) A comparison of the mean mass spectrum across the tissue surface for each tissue class, with the tissue masks from Fig. S1 applied but without normalization or baseline correction. Notice the substantial differences in intensity between the different experiments, preventing direct comparison of peak heights between them. (b) Mean mass spectrum per experiment after combined normalization across all tissue classes. Notice the improved matching in peak heights of common peaks. (c) Mean mass spectrum per experiment after combined normalization and subsequent baseline removal. After preprocessing, the absolute peak heights no longer have a link to the physical meaning of ion counts. Instead the peaks have been scaled to a common intensity range that allows cross-comparison between the four experiments.



(a)



(b)

Figure B.3. Further support for observed pattern in insulin concentration across the four mouse models. (a) Comparison of maximum insulin ion counts acquired in the four pancreas sections. (b) Boxplot comparison of insulin ion counts collected from insulin-containing pixels. Note that the y-axis is a logarithmic scale. Notice that both plots follow the same pattern established in Fig. 7.4.

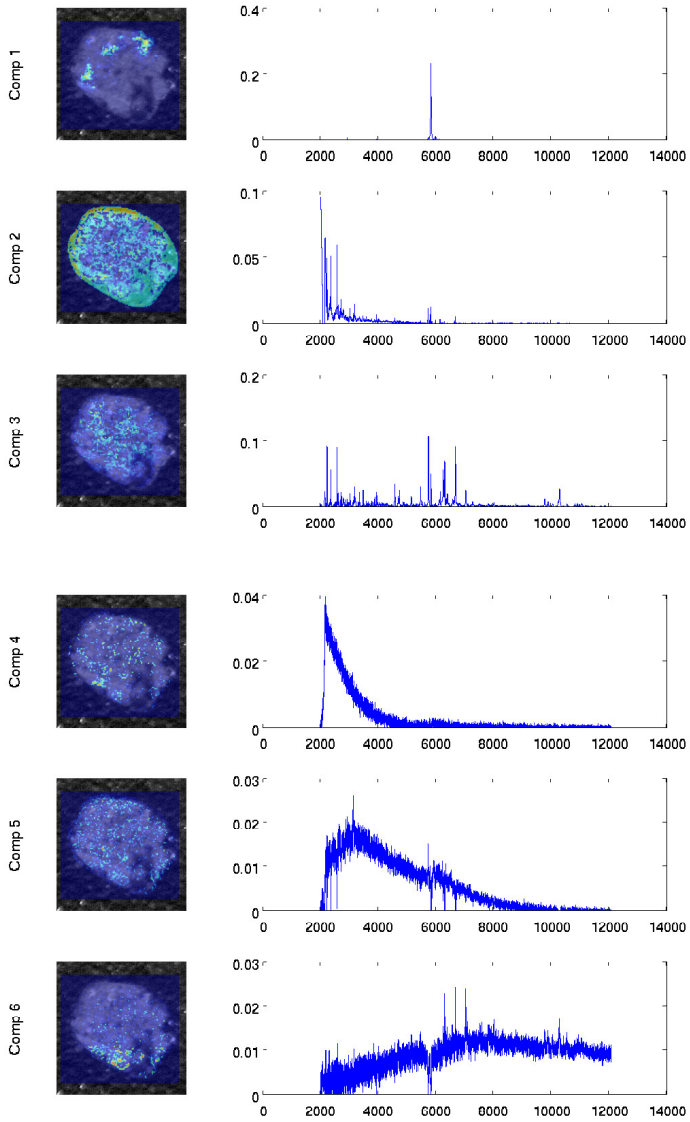


Figure B.4. NMF decomposition of WT. Component 1 is taken as the islet signature.

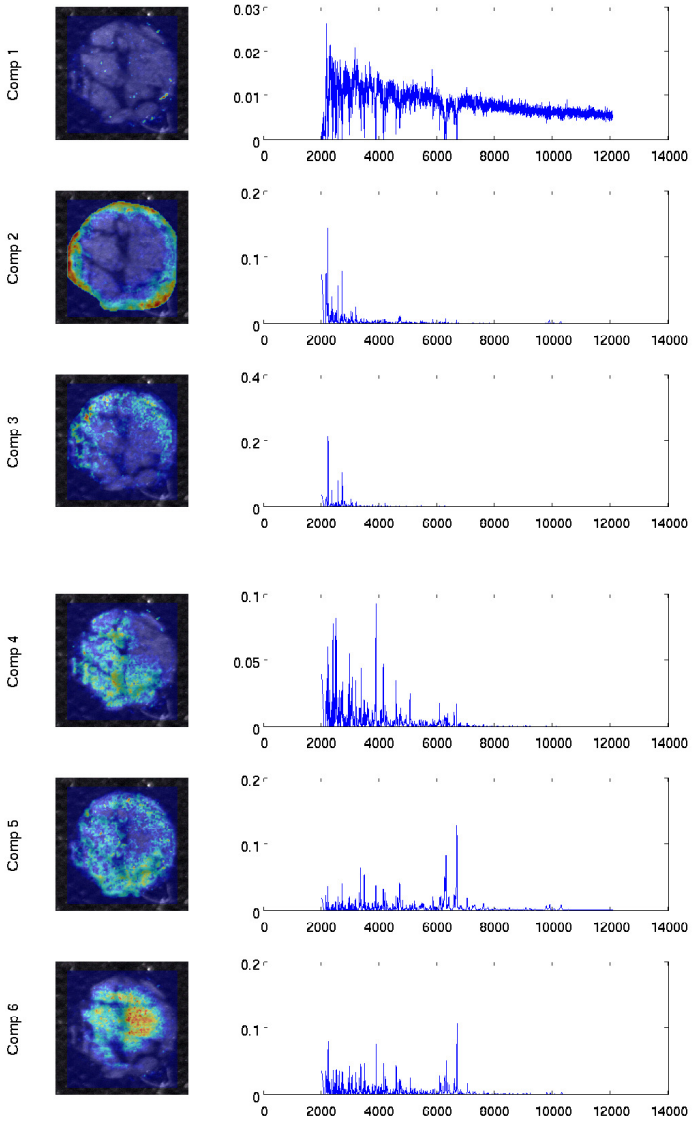


Figure B.5. NMF decomposition of DIA. The DIA section contains insufficient insulin to outline an islet specific signature.

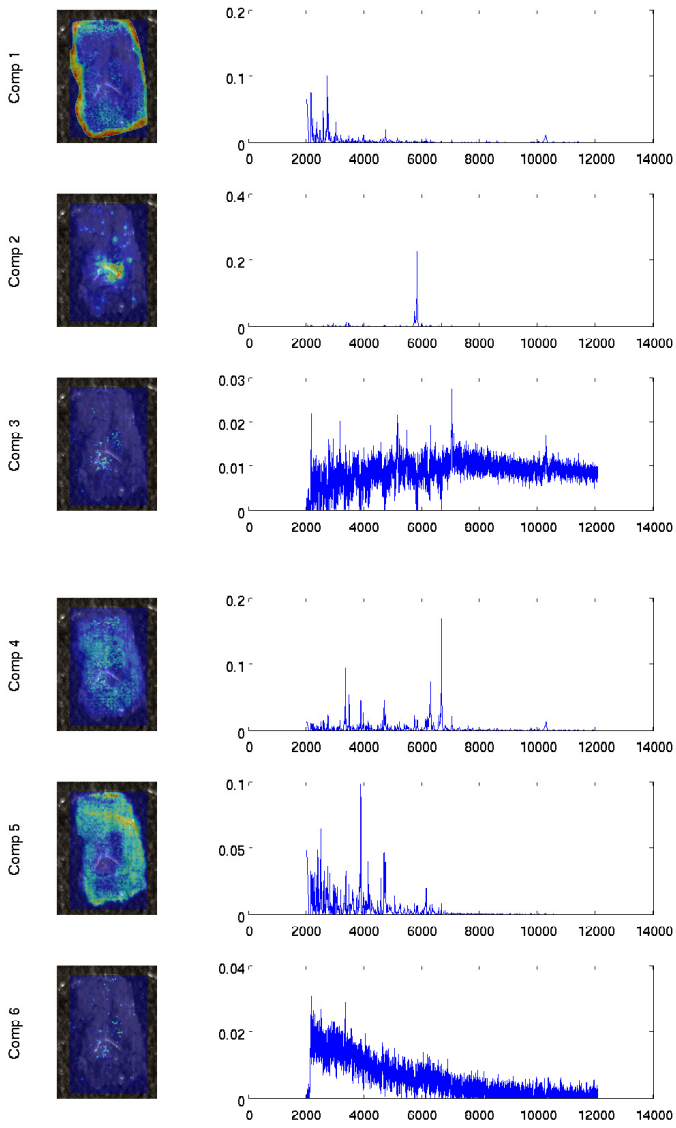


Figure B.6. NMF decomposition of Ob/Ob. Component 2 is taken as the islet signature.

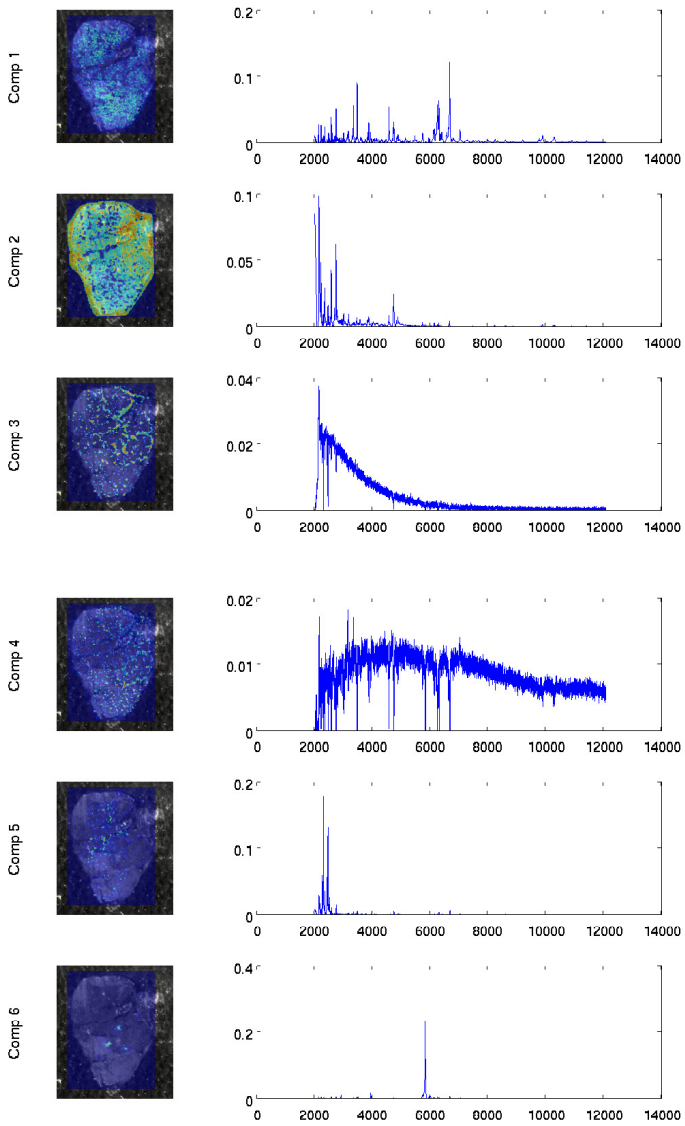
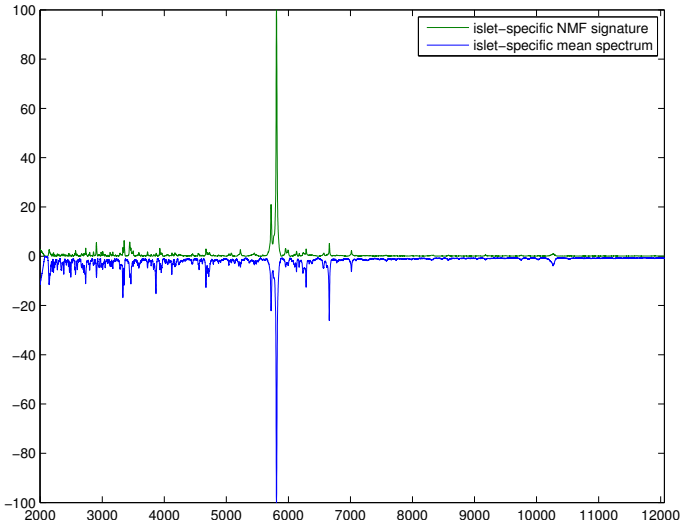
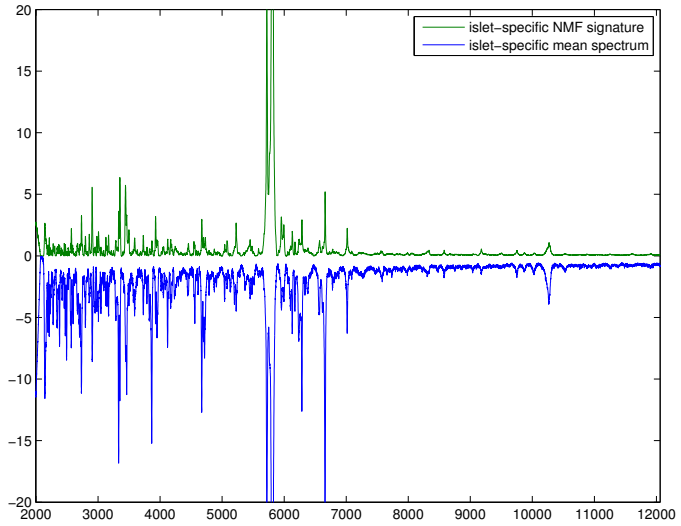


Figure B.7. NMF decomposition of SA/HF. Component 6 is taken as the islet signature.

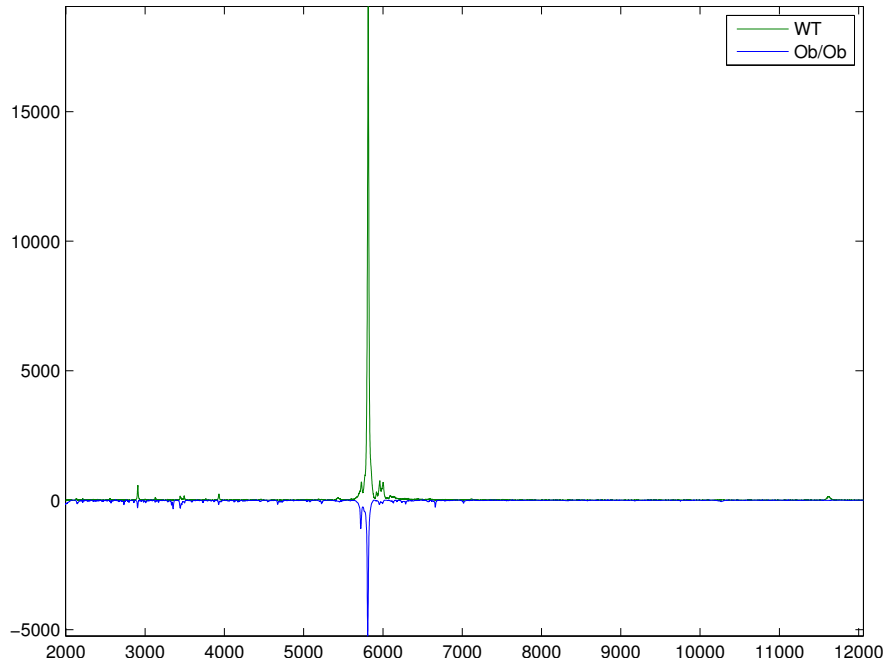


(a) unzoomed

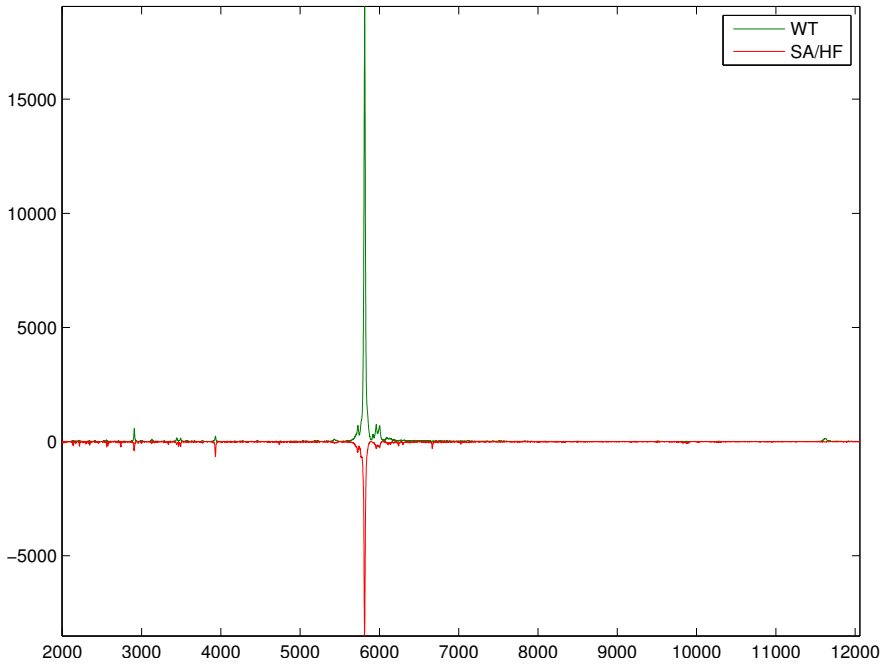


(b) zoomed

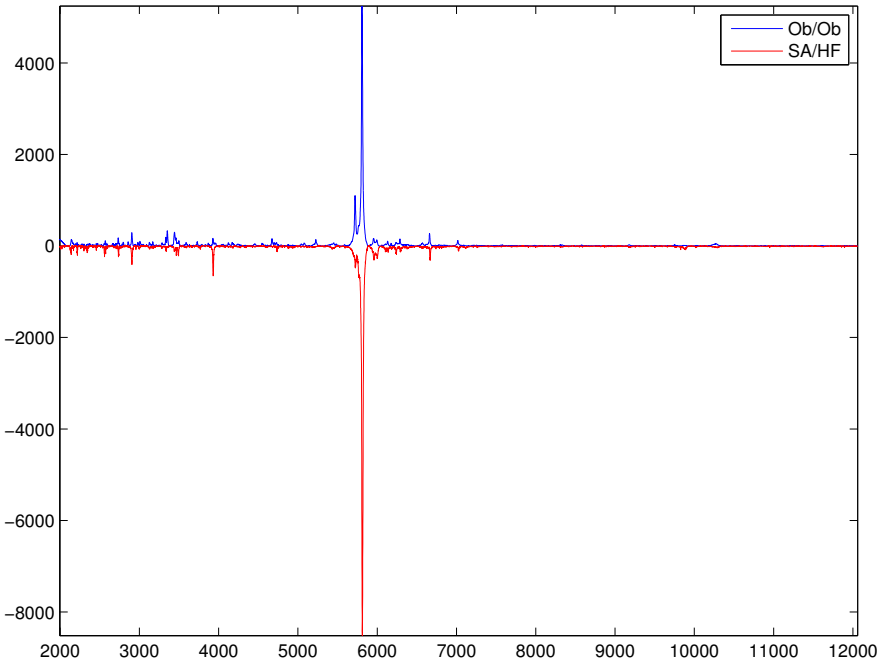
Figure B.8. Islet-specific mean mass spectrum versus islet-specific NMF signature from the Ob/Ob experiment. In addition to ions exclusive to the islets, the mean spectrum contains also ions that appear both in the islet areas and elsewhere. The NMF signature attempts to retain only the islet-exclusive ion peaks, making for a more suitable signature of islet content. (Above) Comparison plot, largest peak scaled to hundred. (Below) zoomed version.


WT
Ob/Ob
15000
10000
5000
0
−5000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

(a)

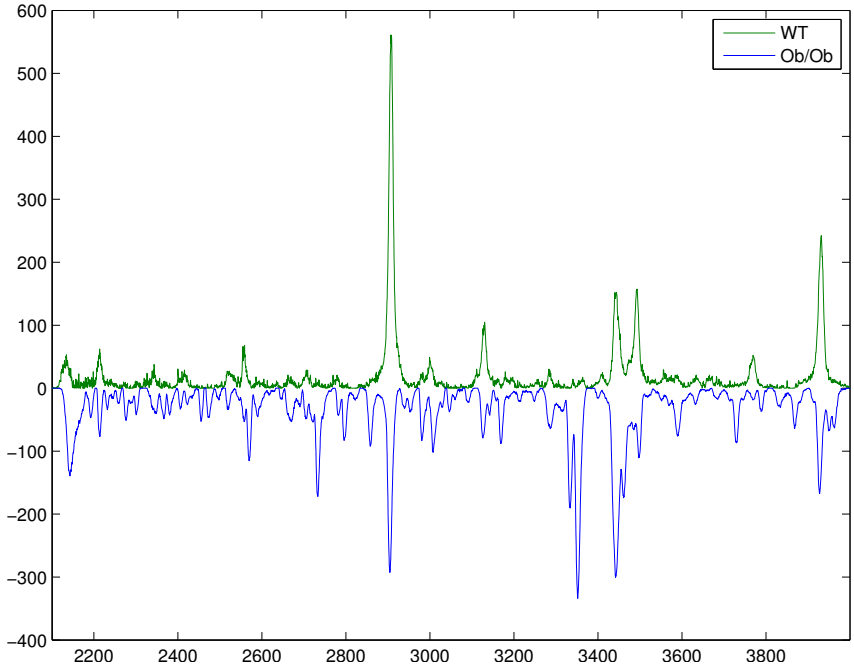

WT
SA/HF
15000
10000
5000
0
−5000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

(b)

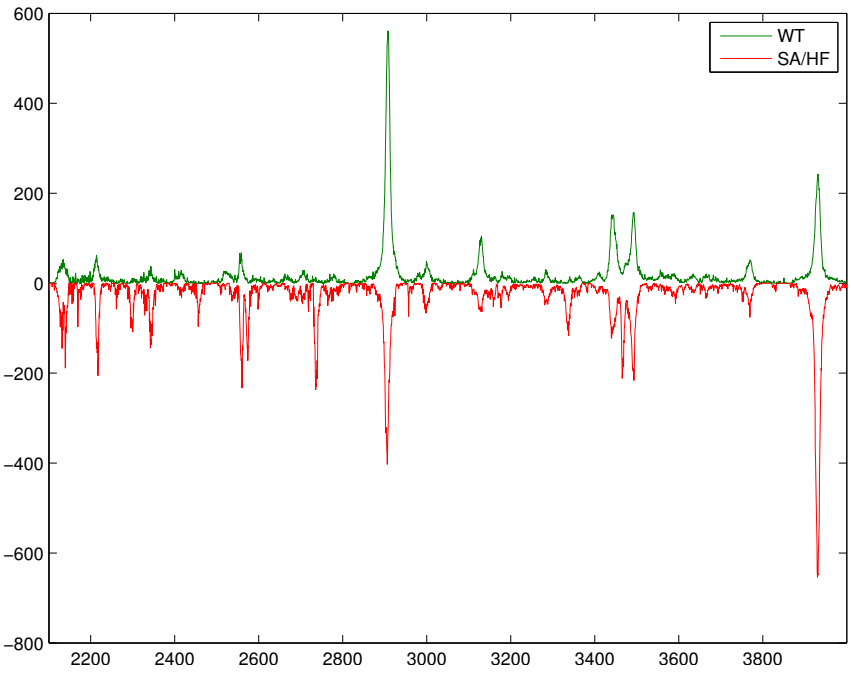


(c)

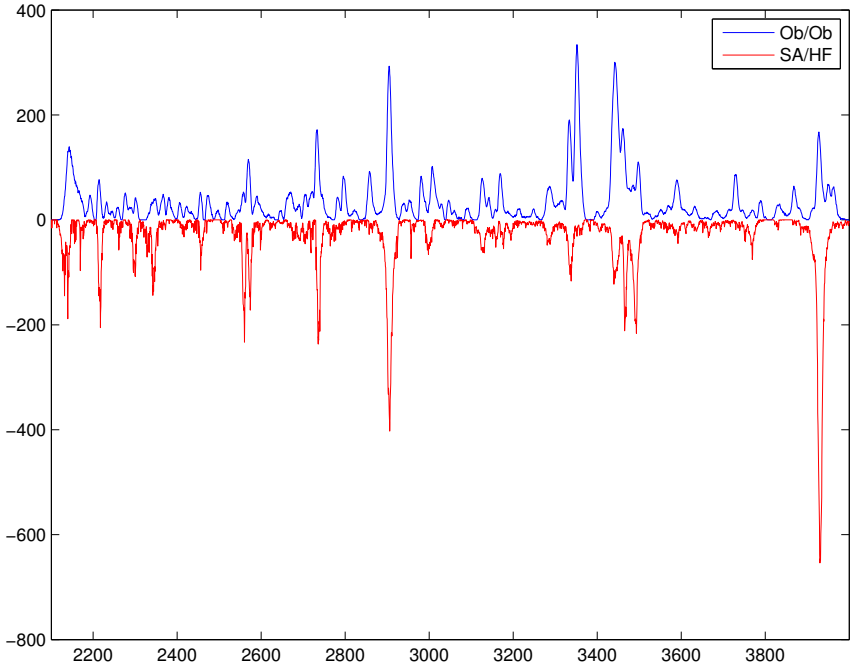
Figure B.9. Comparison of islet specific signatures. Complete mass range.


WT
Ob/Ob
600
500
400
300
200
100
0
−100
−200
−300
−400
2200
2400
2600
2800
3000
3200
3400
3600
3800

(a)

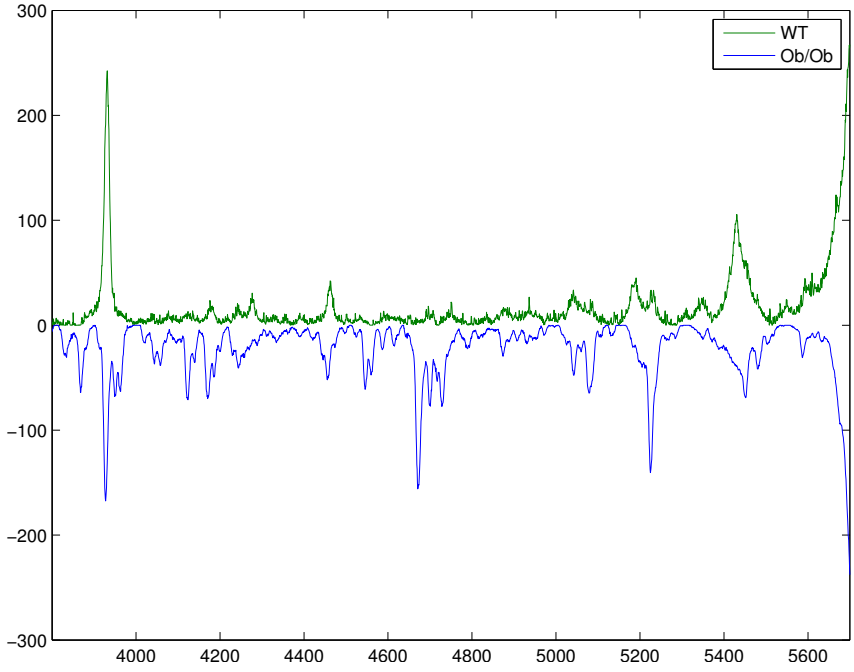

WT
SA/HF
600
400
200
0
−200
−400
−600
−800
2200
2400
2600
2800
3000
3200
3400
3600
3800

(b)

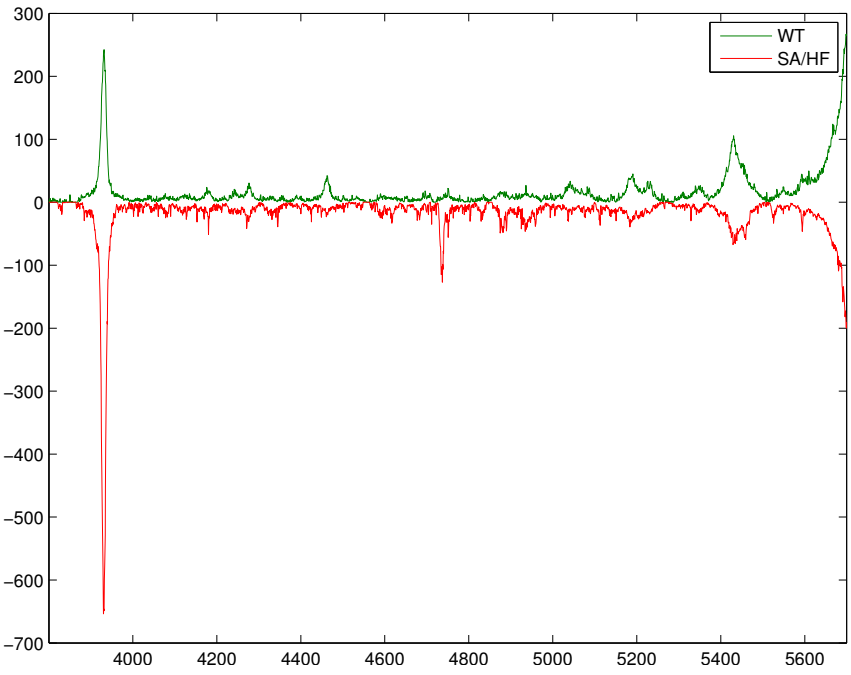


(c)

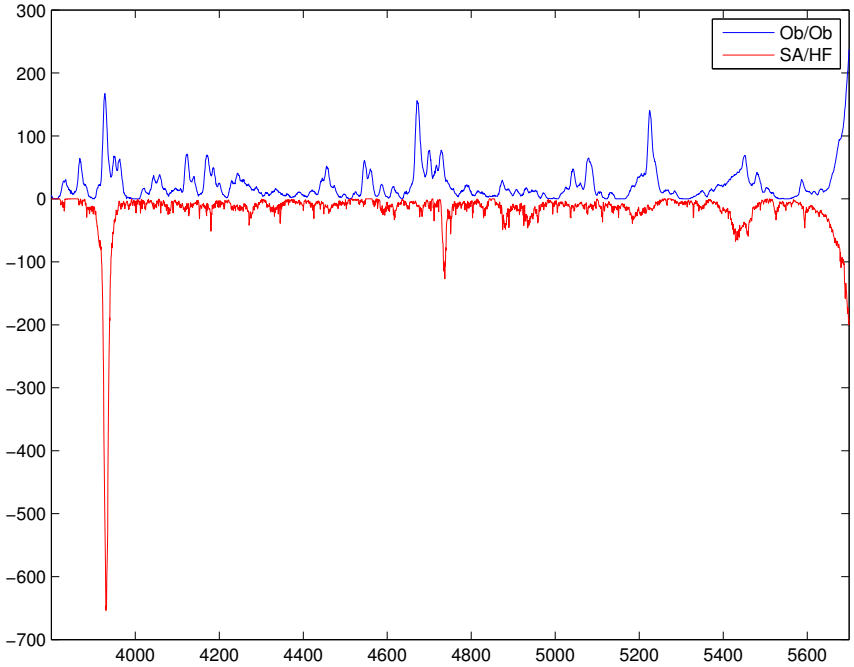
Figure B.10. Comparison of islet specific signatures. m/z 2100 to 4000.


WT
Ob/Ob
300
200
100
0
−100
−200
−300
4000
4200
4400
4600
4800
5000
5200
5400
5600

(a)

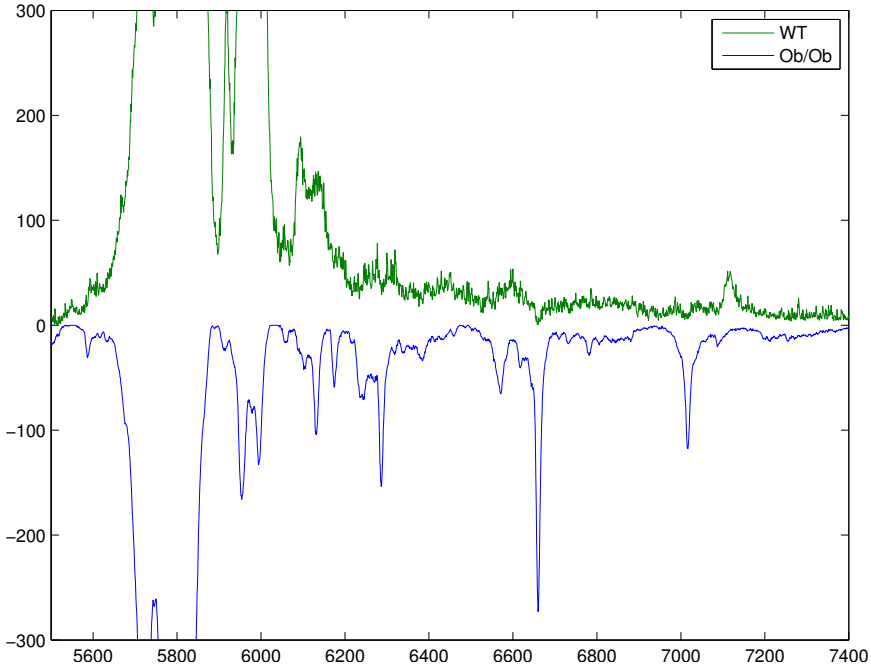

WT
SA/HF
300
200
100
0
−100
−200
−300
−400
−500
−600
−700
4000
4200
4400
4600
4800
5000
5200
5400
5600

(b)

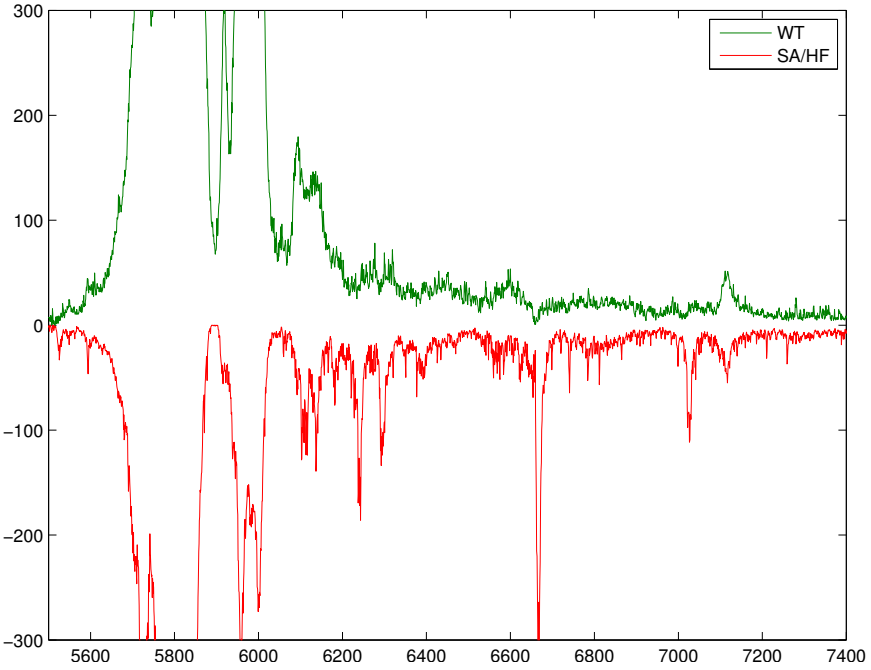


(c)

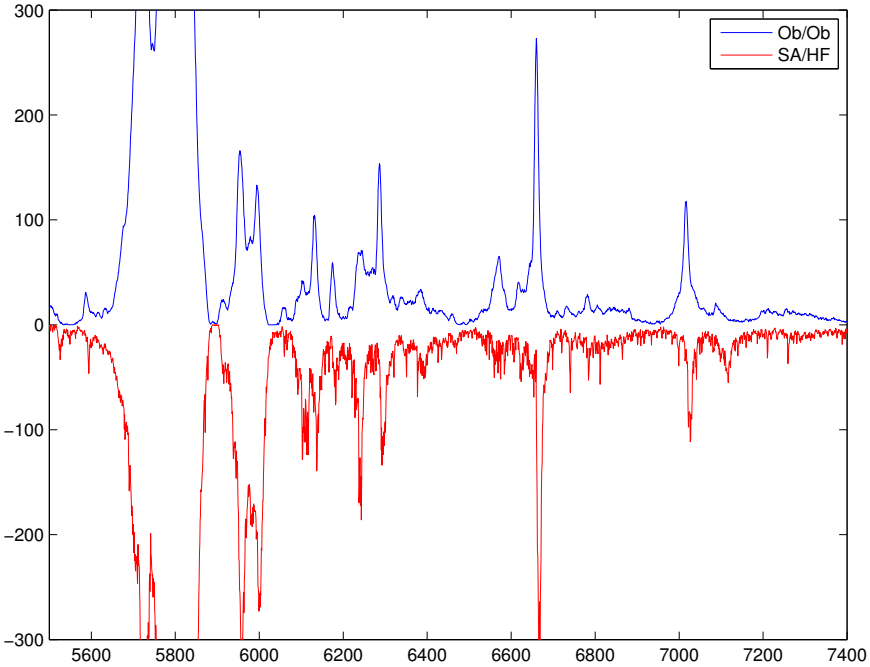
Figure B.11. Comparison of islet specific signatures. m/z 3800 to 5700.


WT
Ob/Ob
300
200
100
0
−100
−200
−300
5600
5800
6000
6200
6400
6600
6800
7000
7200
7400

(a)

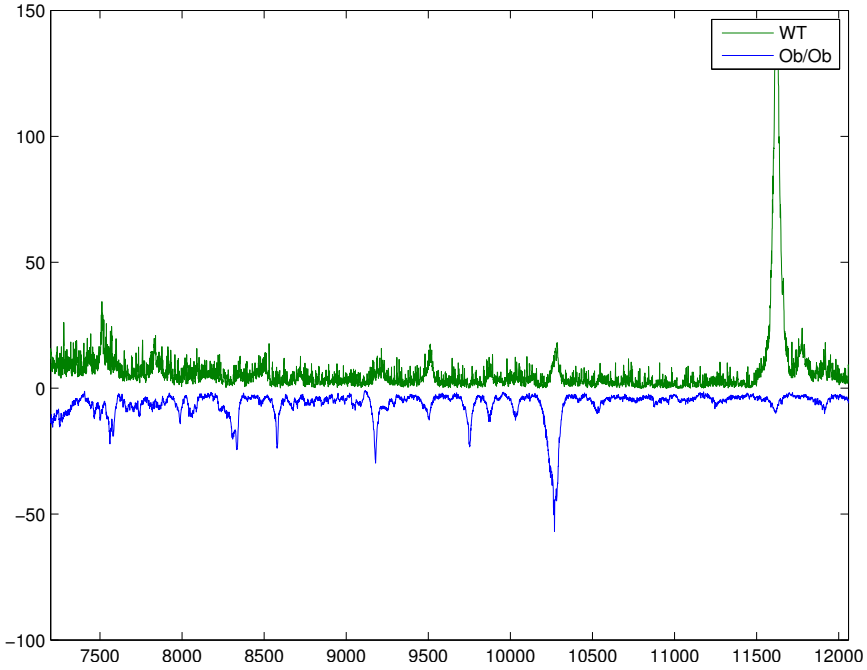

WT
SA/HF
300
200
100
0
−100
−200
−300
5600
5800
6000
6200
6400
6600
6800
7000
7200
7400

(b)

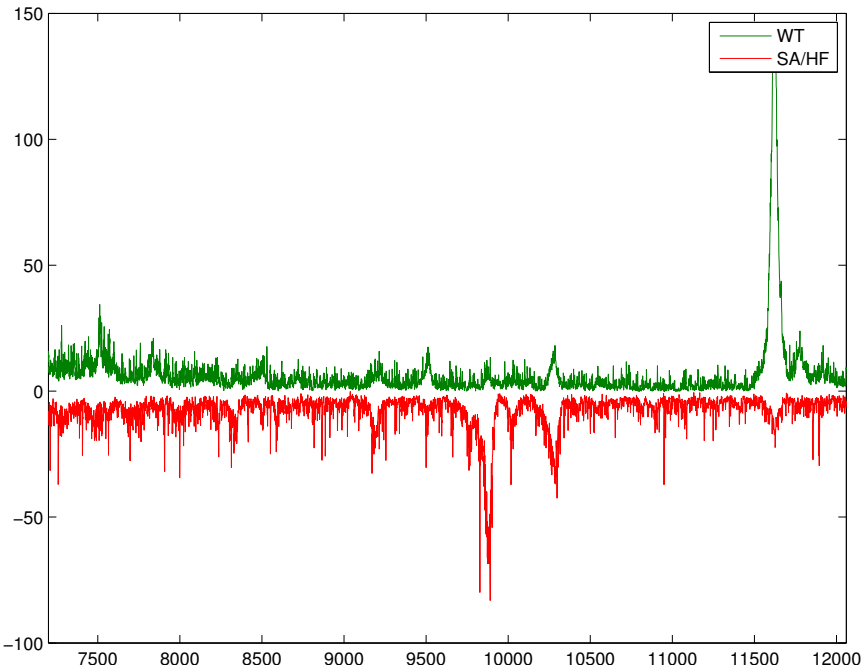


(c)

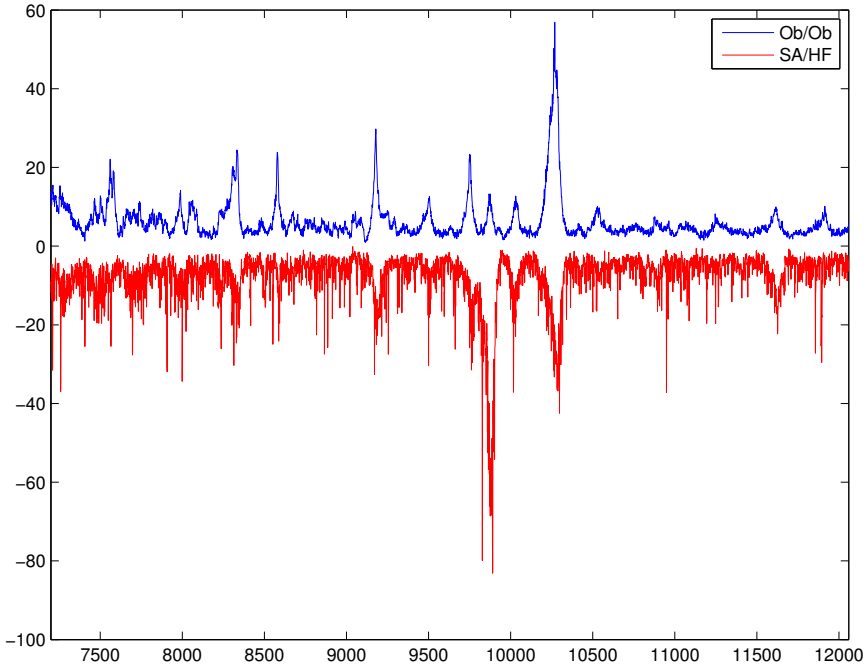
Figure B.12. Comparison of islet specific signatures. m/z 5500 to 7400.


WT
Ob/Ob
150
100
50
0
−50
−100
7500
8000
8500
9000
9500
10000
10500
11000
11500
12000

(a)

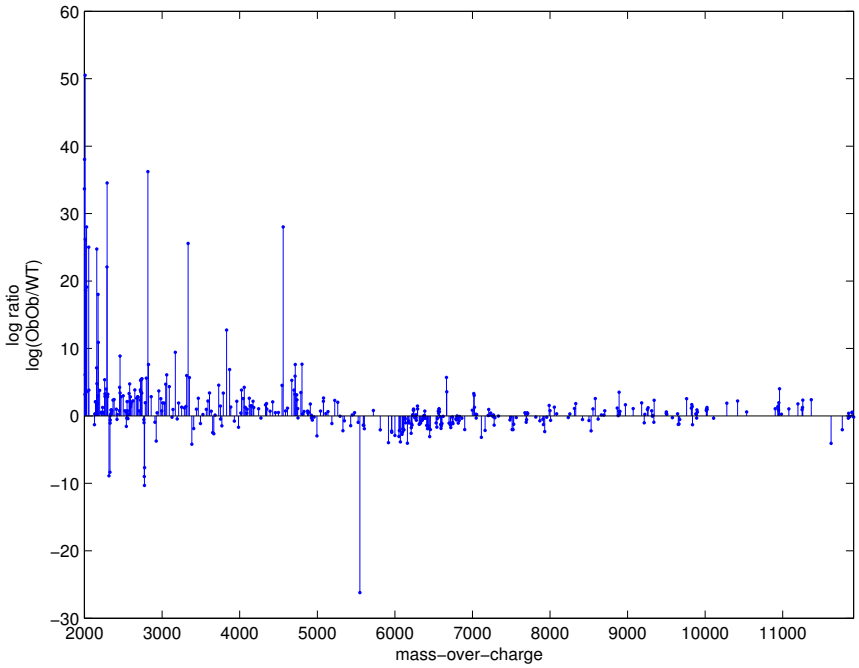

WT
SA/HF
150
100
50
0
−50
−100
7500
8000
8500
9000
9500
10000
10500
11000
11500
12000

(b)

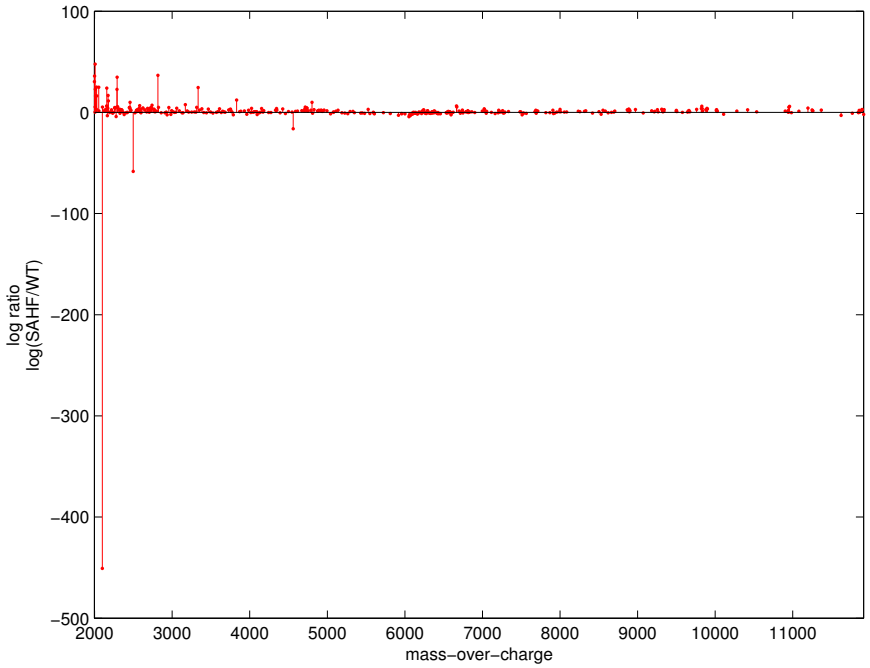


(c)

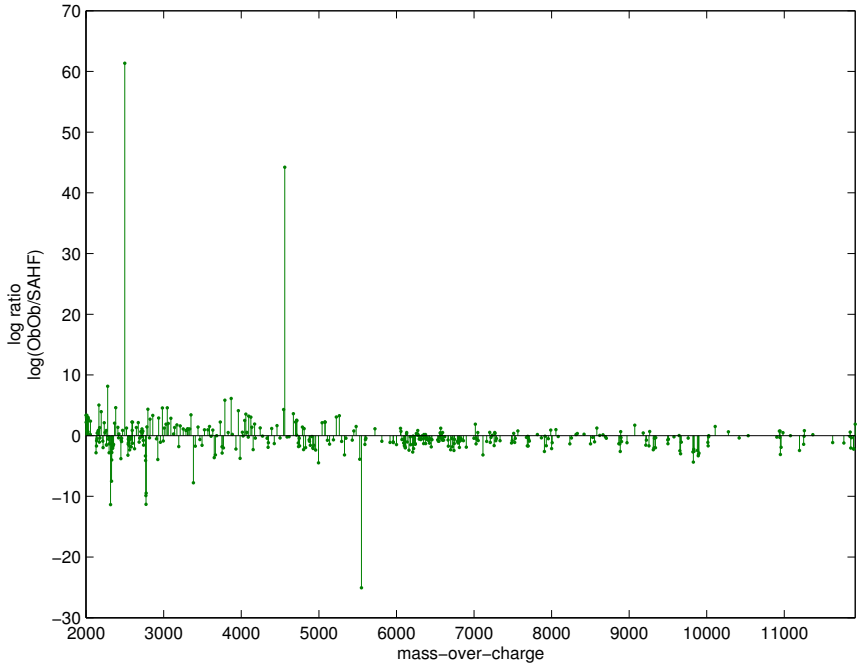
Figure B.13. Comparison of islet specific signatures. m/z 7200 to 12000.


log ratio
log(ObOb/WT)
mass−over−charge
60
50
40
30
20
10
0
−10
−20
−30
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000

(a)

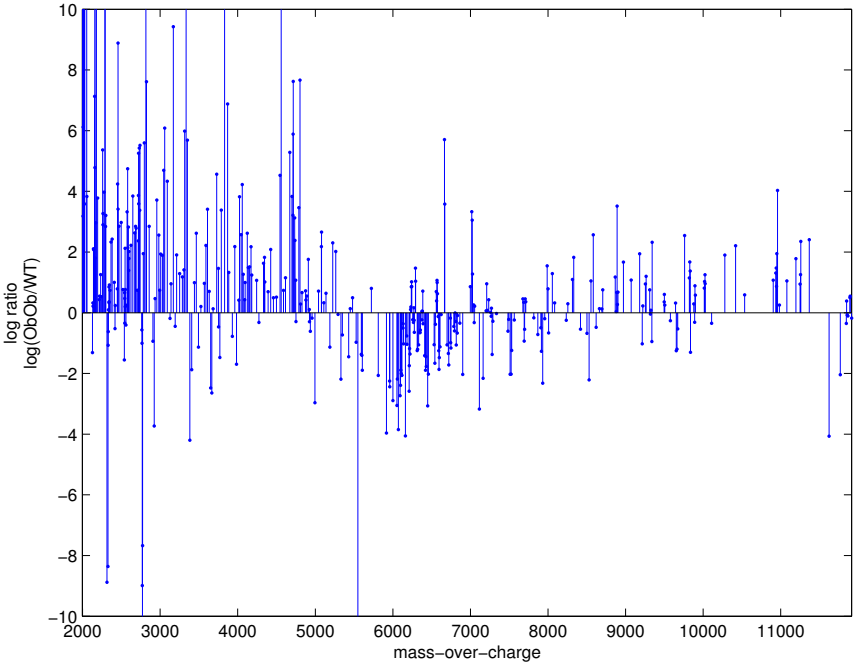

log ratio
log(SAHF/WT)
mass−over−charge
100
0
−100
−200
−300
−400
−500
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000

(b)

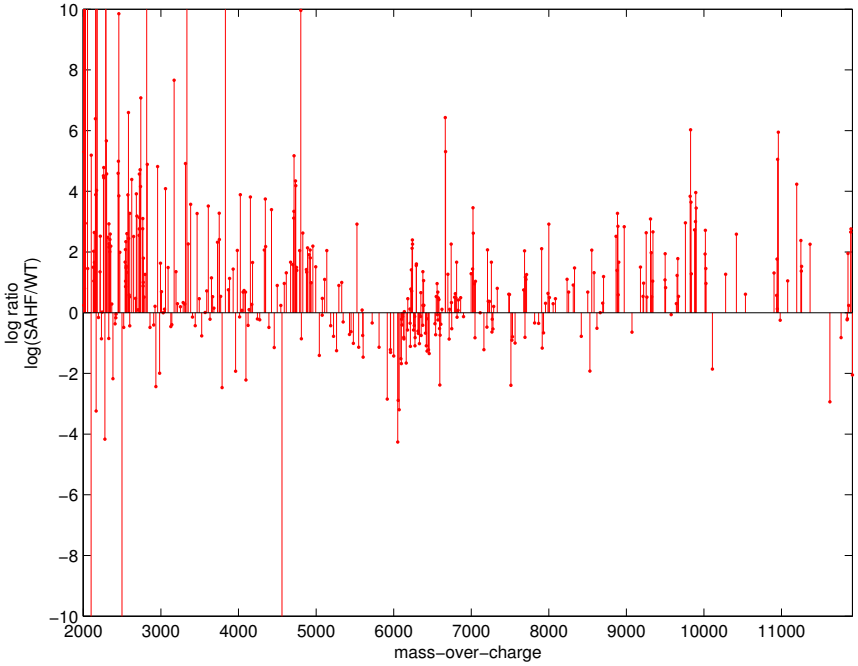


(c)

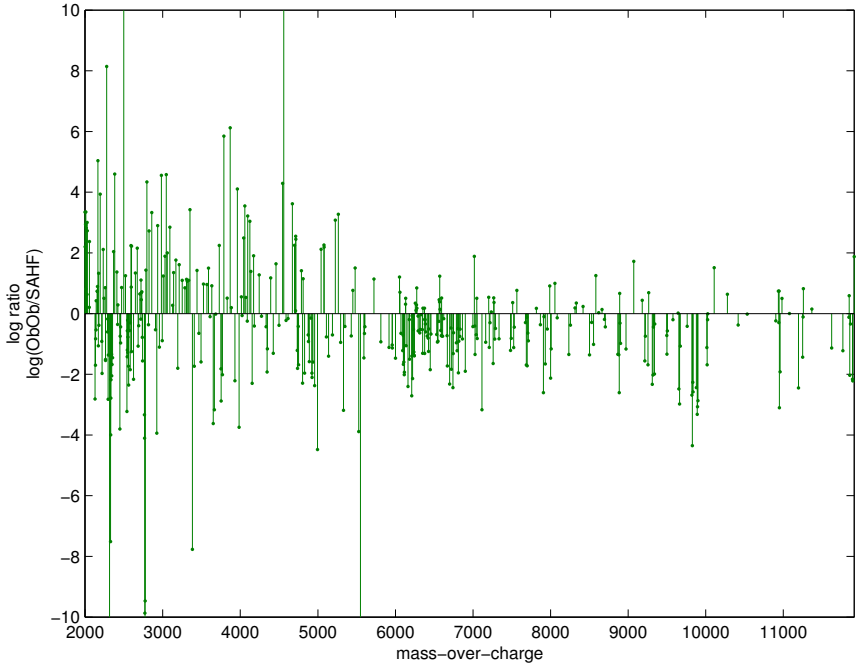
Figure B.14. Fold changes for the 419 peak-picked ion peaks in the signatures. (a) Ob/Ob versus WT. (b) SA/HF versus WT. (c) Ob/Ob versus SA/HF.


log ratio
log(ObOb/WT)
mass−over−charge
10
8
6
4
2
0
−2
−4
−6
−8
−10
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000

(a)

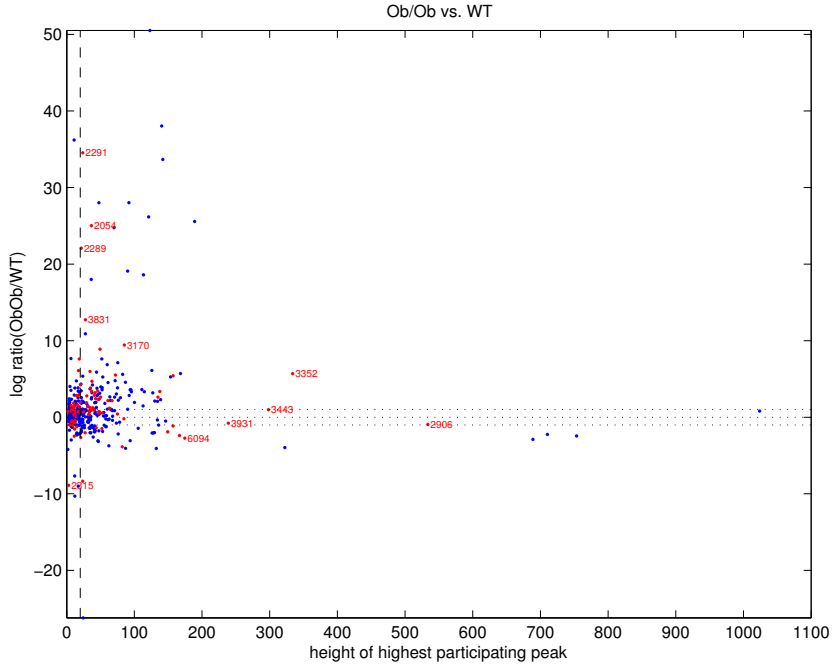

log ratio
log(SAHF/WT)
mass−over−charge
10
8
6
4
2
0
−2
−4
−6
−8
−10
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000

(b)

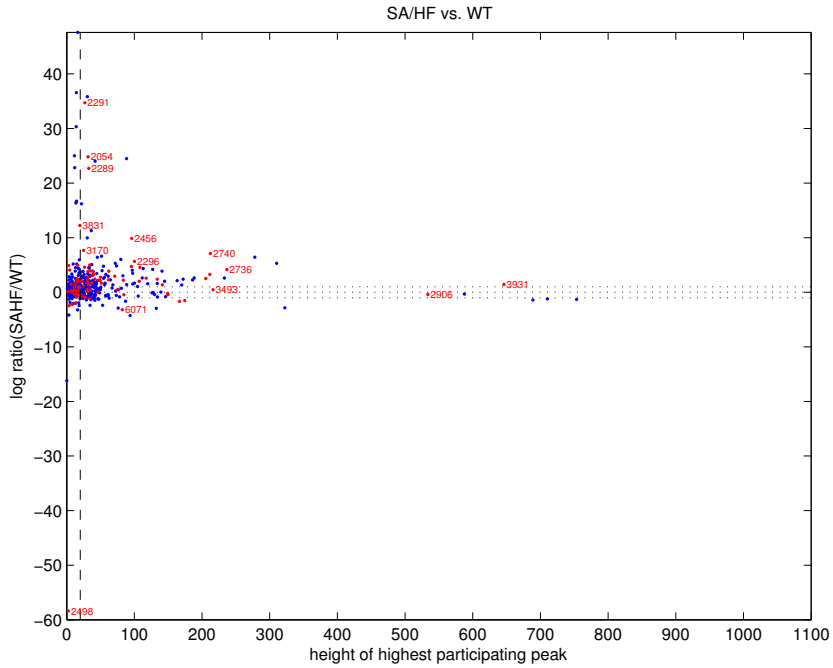


(c)

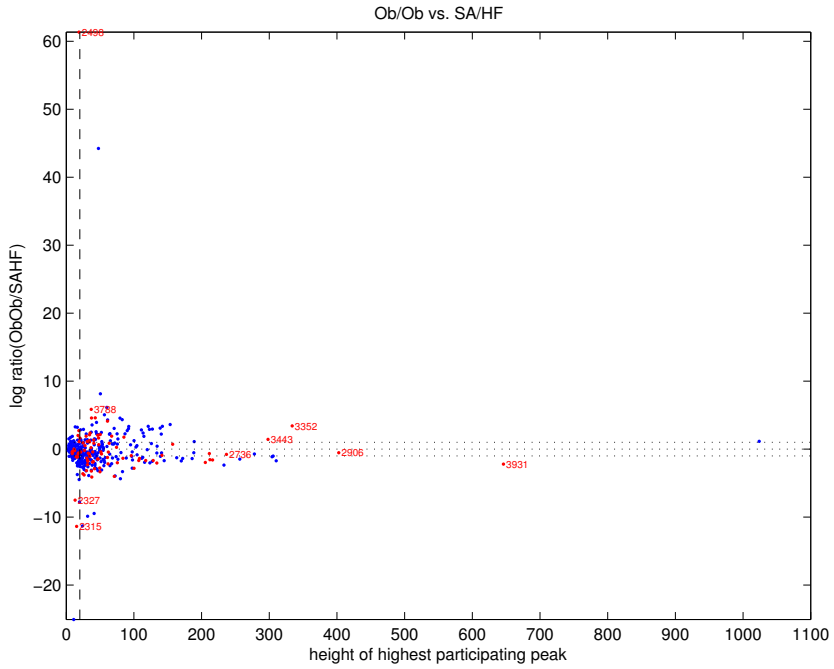
Figure B.15. Fold changes for the 419 peak-picked ion peaks in the signatures, zoomed to -10 to + 10 range. (a) Ob/Ob versus WT. (b) SA/HF versus WT. (c) Ob/Ob versus SA/HF.


Ob/Ob vs. WT
log ratio(ObOb/WT)
height of highest participating peak
2291
2054
2289
3831
3170
3352
3443
3931
2906
6094
2415

(a)

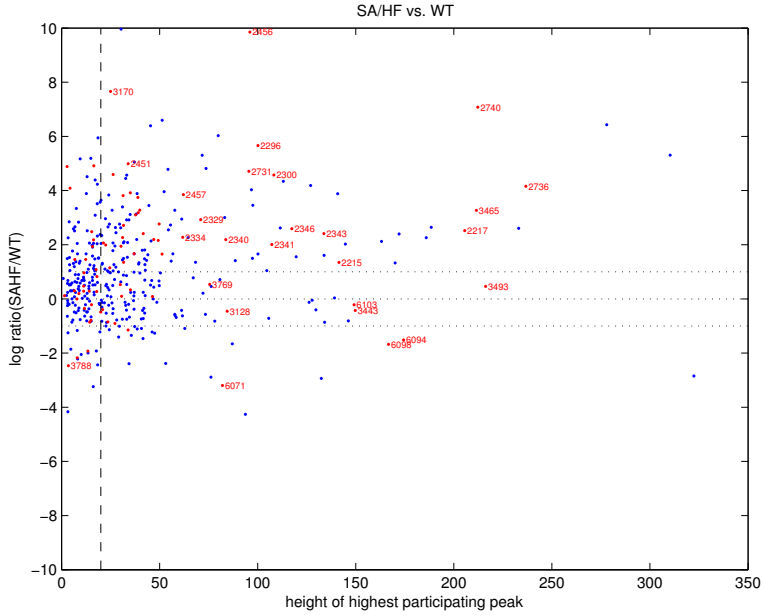

SA/HF vs. WT
log ratio(SAHF/WT)
height of highest participating peak
2291
3054
2289
3931
3170
2456
2296
2740
2736
3493
2906
3931
6071
2498

(b)

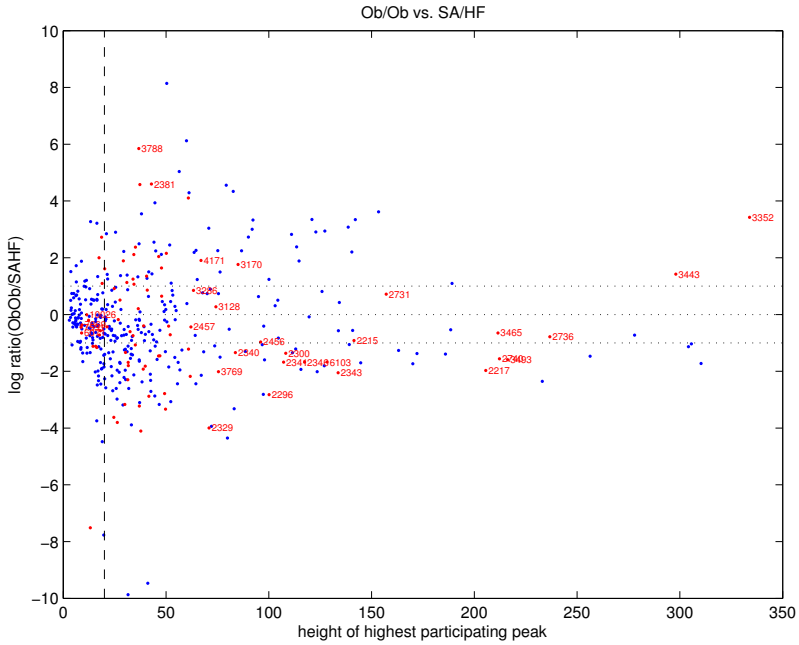


(c)

Figure B.16. Differential expression comparisons between the islet signatures. All peak-picked ions from the signatures are plotted according to their log ratio versus their highest participating peak (see Fig. 7.8). (a) Ob/Ob versus WT. (b) SA/HF versus WT. (c) Ob/Ob versus SA/HF. Notice how both diabetes models differ more from the WT model (larger data cloud) than from each other (smaller data cloud).

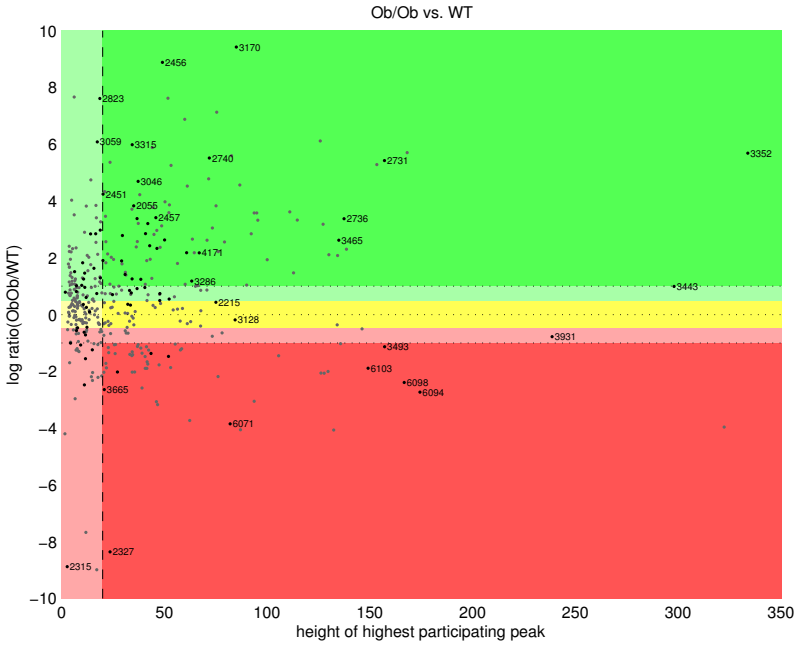


(a)

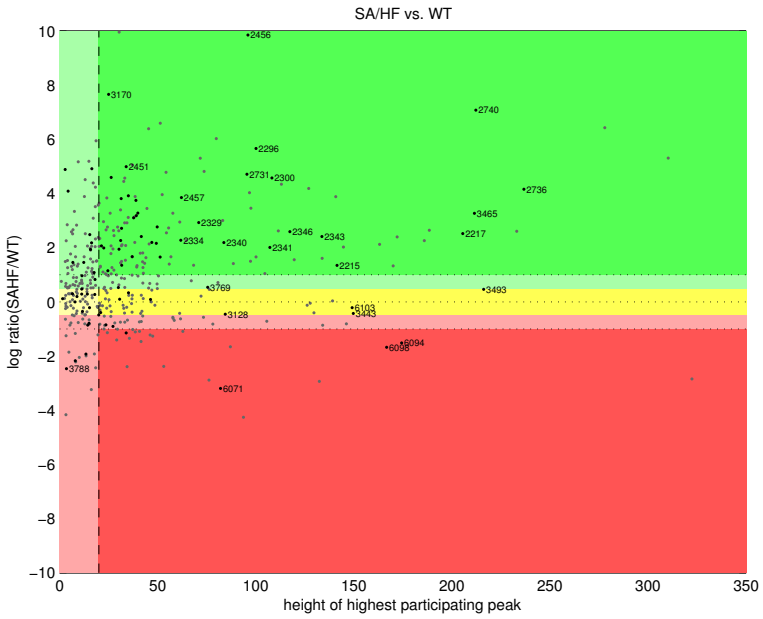


(b)

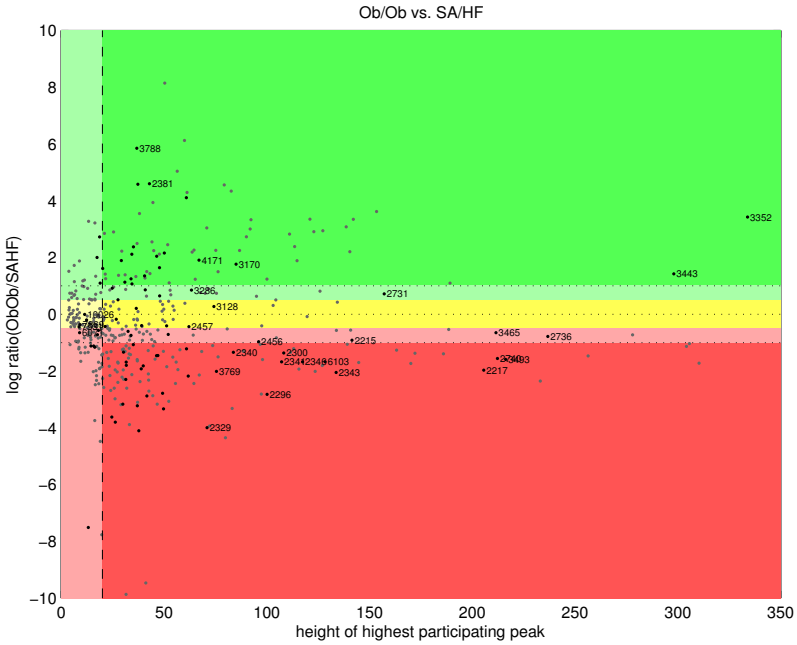
Figure B.17. Differential expression comparisons between the islet signatures, zoomed versions. (a) SA/HF versus WT. (b) Ob/Ob versus SA/HF. Ob/Ob versus WT is shown in Fig. 7.8.


Ob/Ob vs. WT
log ratio(ObOb/WT)
height of highest participating peak
10
8
6
4
2
0
−2
−4
−6
−8
−10
0
50
100
150
200
250
300
350
3170
2456
2823
3059
3315
2740
2731
3352
3046
2451
2055
2457
2736
3465
4171
3286
3443
2215
3128
3931
3493
6103
6098
6094
3665
6071
2327
2315

(a)

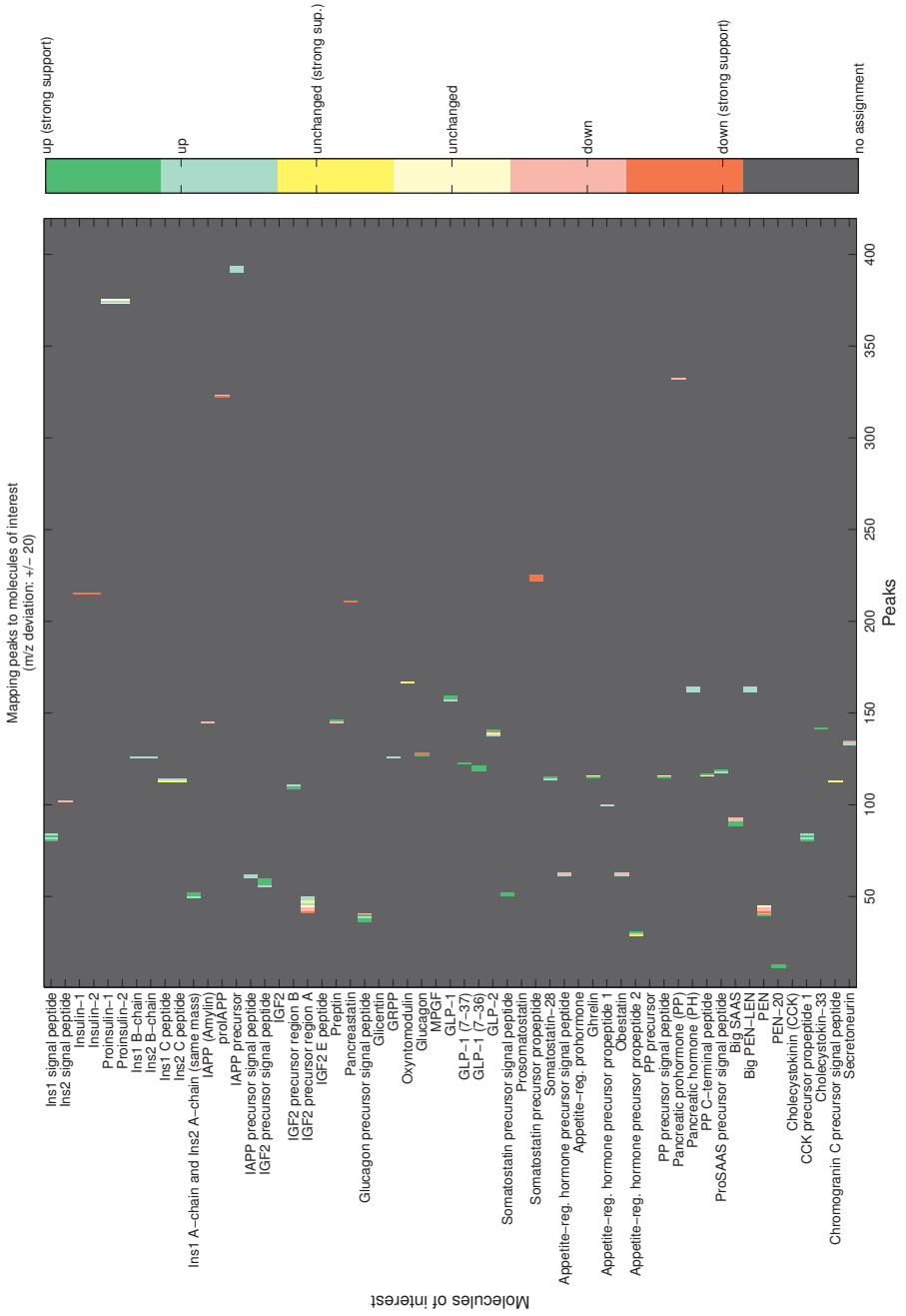

SA/HF vs. WT
log ratio(SAHF/WT)
height of highest participating peak
10
8
6
4
2
0
−2
−4
−6
−8
−10
0
50
100
150
200
250
300
350
2456
3170
2740
2296
2451
2731
2300
2736
2457
3465
2329
2346
2343
2217
2334
2340
2341
2215
3769
3493
6103
3128
3443
6094
6098
3788
6071

(b)

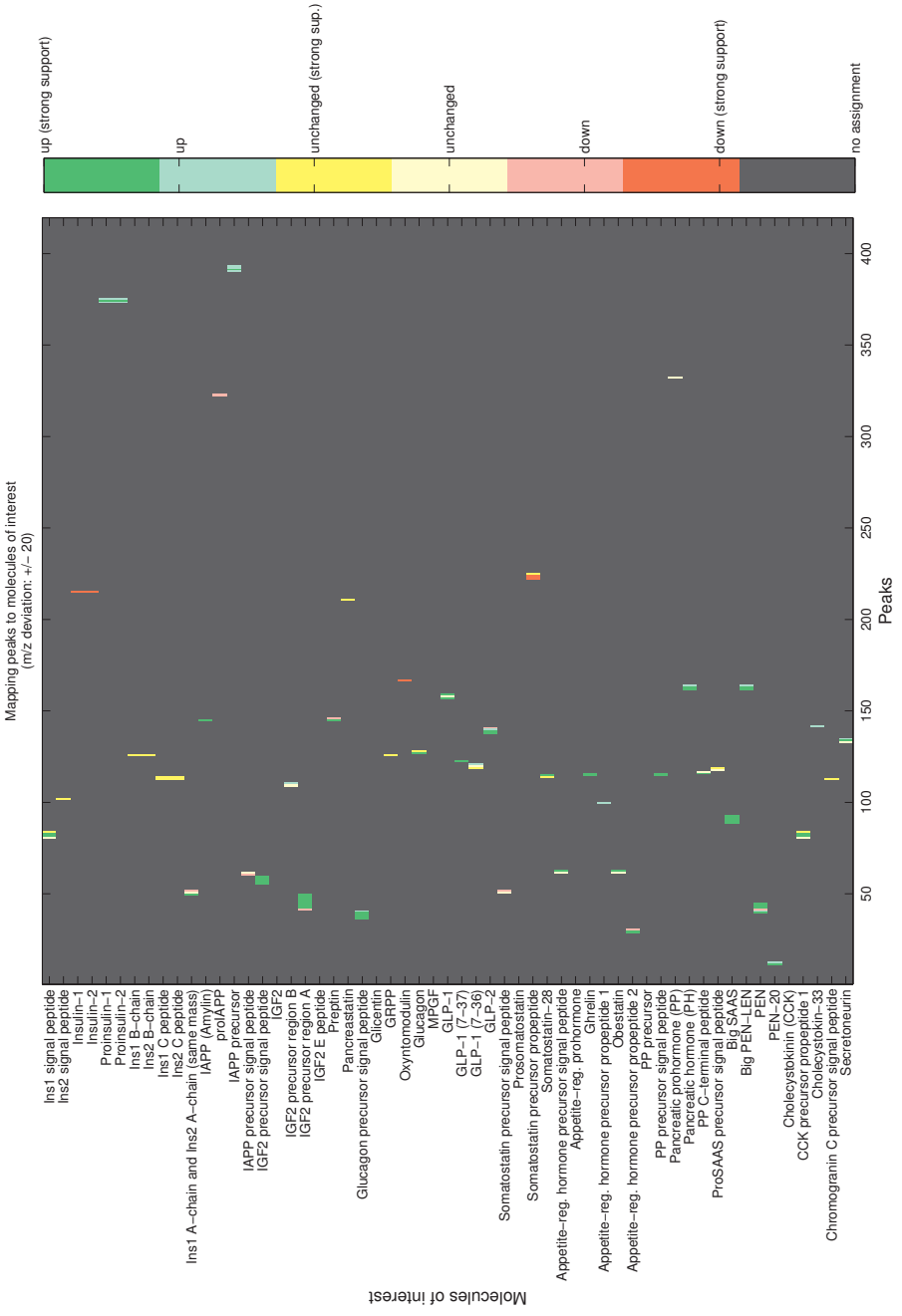


(c)

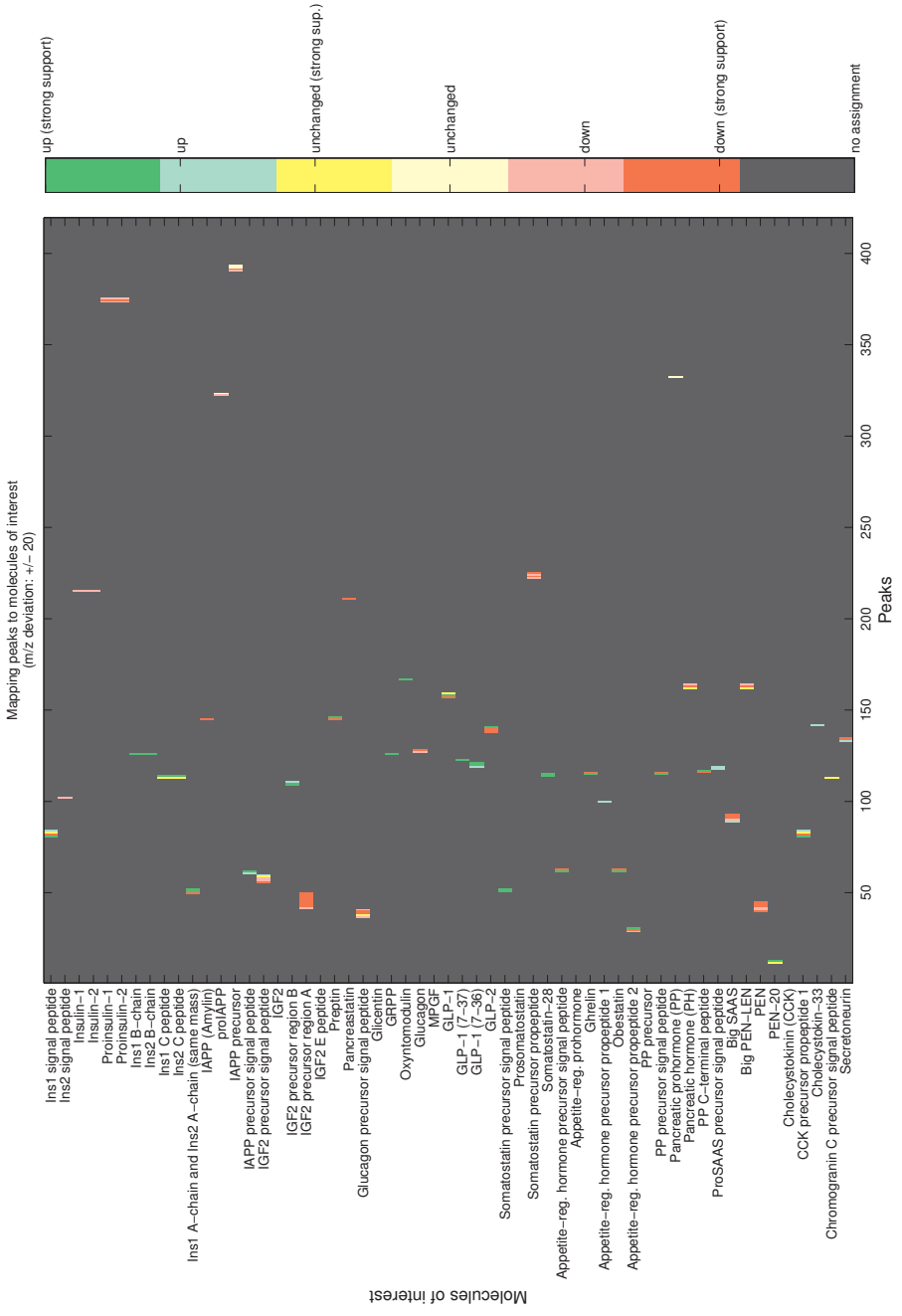
Figure B.18. Color-coded differential expression comparisons between the islet signatures, zoomed versions. Ions in the green area are upregulated. Ions in the red area are downregulated. Ions in the yellow area show a fold change of less than 1.4 (0.5 on the log scale) and are considered unchanged. Taking the sensitivity of the mass spectral measurements into account, ions showing up or downregulation of less than two-fold (1 on the log scale, indicated as dotted lines) are given a lighter color to differentiate them from the more reliable fold changes that exceed two-fold. Ions left of the reliability threshold present a fold change calculation based on two small peak heights, and are considered less reliable than those right of the threshold and are therefore indicated in a lighter color as well. (a) Ob/Ob versus WT. (b) SA/HF versus WT. (c) Ob/Ob versus SA/HF.


up (strong support)
up
unchanged (strong sup.)
unchanged
down
down (strong support)
no assignment
Mapping peaks to molecules of interest
(m/z deviation: +/− 20)
Peaks
50
100
150
200
250
300
350
400
Molecules of interest
Ins1 signal peptide
Ins2 signal peptide
Insulin−1
Insulin−2
Proinsulin−1
Proinsulin−2
Ins1 B−chain
Ins2 B−chain
Ins1 C peptide
Ins2 C peptide
Ins1 A−chain and Ins2 A−chain (same mass)
IAPP (Amylin)
proIAPP
IAPP precursor
IAPP precursor signal peptide
IGF2 precursor signal peptide
IGF2
IGF2 precursor region B
IGF2 precursor region A
IGF2 E peptide
Preptin
Pancreastatin
Glucagon precursor signal peptide
Glicentin
GRPP
Oxyntomodulin
Glucagon
MPGF
GLP−1
GLP−1 (7−37)
GLP−1 (7−36)
GLP−2
Somatostatin precursor signal peptide
Prosomatostatin
Somatostatin precursor propeptide
Somatostatin−28
Appetite−reg. hormone precursor signal peptide
Appetite−reg. prohormone
Ghrelin
Appetite−reg. hormone precursor propeptide 1
Obestatin
Appetite−reg. hormone precursor propeptide 2
PP precursor
PP precursor signal peptide
Pancreatic prohormone (PP)
Pancreatic hormone (PH)
PP C−terminal peptide
ProSAAS precursor signal peptide
Big SAAS
Big PEN−LEN
PEN
PEN−20
Cholecystokinin (CCK)
CCK precursor propeptide 1
Cholecystokin−33
Chromogranin C precursor signal peptide
Secretoneurin

(a)


up (strong support)
up
unchanged (strong sup.)
unchanged
down
down (strong support)
no assignment
Mapping peaks to molecules of interest
(m/z deviation: +/− 20)
Peaks
50
100
150
200
250
300
350
400
Molecules of interest
Ins1 signal peptide
Ins2 signal peptide
Insulin−1
Insulin−2
Proinsulin−1
Proinsulin−2
Ins1 B−chain
Ins2 B−chain
Ins1 C peptide
Ins2 C peptide
Ins1 A−chain and Ins2 A−chain (same mass)
IAPP (Amylin)
proIAPP
IAPP precursor
IAPP precursor signal peptide
IGF2 precursor signal peptide
IGF2
IGF2 precursor region B
IGF2 precursor region A
IGF2 E peptide
Preptin
Pancreastatin
Glucagon precursor signal peptide
Glicentin
GRPP
Oxyntomodulin
Glucagon
MPGF
GLP−1
GLP−1 (7−37)
GLP−1 (7−36)
GLP−2
Somatostatin precursor signal peptide
Prosomatostatin
Somatostatin precursor propeptide
Somatostatin−28
Appetite−reg. hormone precursor signal peptide
Appetite−reg. prohormone
Ghrelin
Appetite−reg. hormone precursor propeptide 1
Obestatin
Appetite−reg. hormone precursor propeptide 2
PP precursor
PP precursor signal peptide
Pancreatic prohormone (PP)
Pancreatic hormone (PH)
PP C−terminal peptide
ProSAAS precursor signal peptide
Big SAAS
Big PEN−LEN
PEN
PEN−20
Cholecystokinin (CCK)
CCK precursor propeptide 1
Cholecystokin−33
Chromogranin C precursor signal peptide
Secretoneurin

(b)



(c)

Figure B.19. Color-coded mappings from peaks to molecules of interest. (a) Ob/Ob versus WT. (b) SA/HF versus WT. (c) Ob/Ob versus SA/HF.

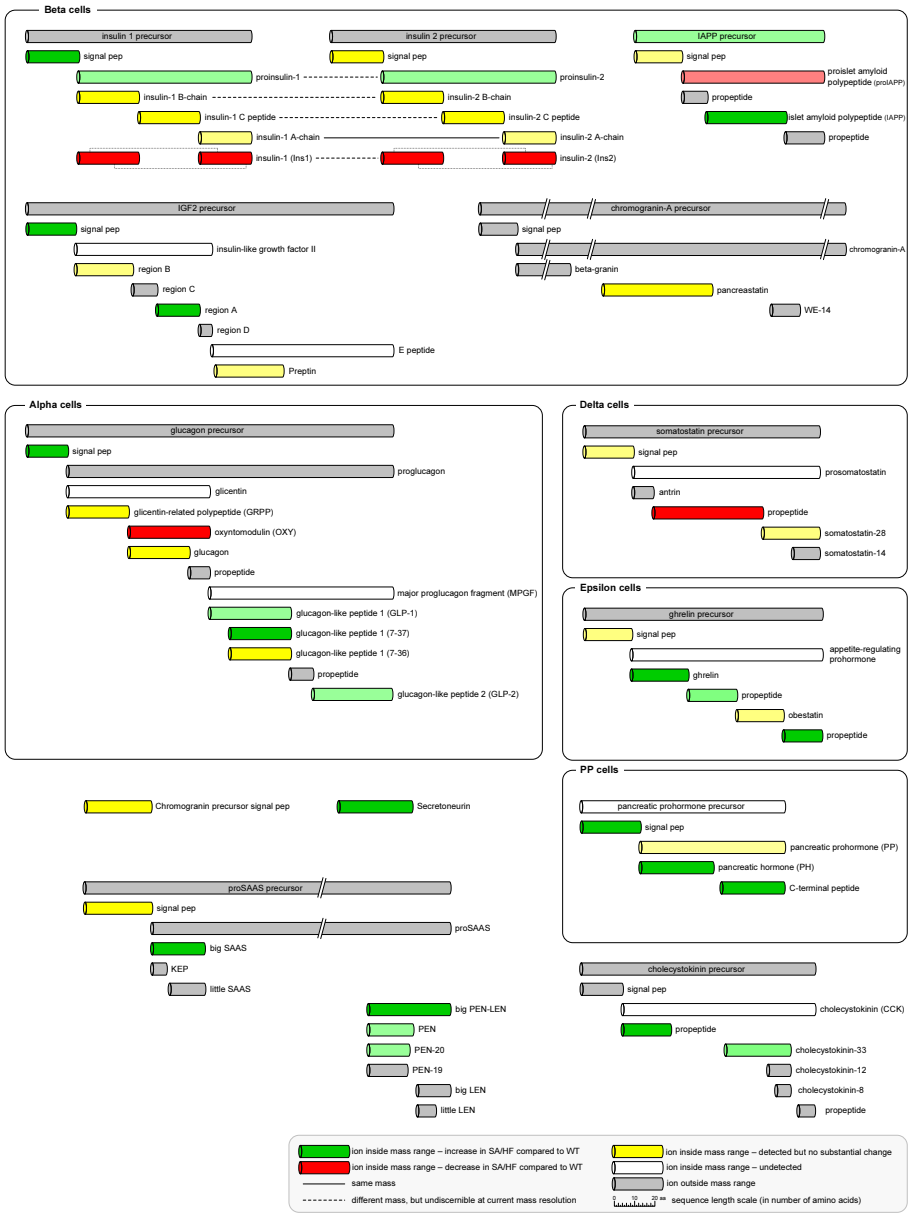


Figure B.20. Hormone processing and cleavage products in SA/HF versus WT.

Bibliography

[1] J. C. Venter *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–51, Feb 2001.

[2] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb 2001.

[3] S. Hanash, "Disease proteomics," *Nature*, vol. 422, no. 6928, pp. 226–32, Mar 2003.

[4] N. L. Young, M. D. Plazas-Mayorca, and B. A. Garcia, "Systems-wide proteomic characterization of combinatorial post-translational modification patterns," *Expert Rev Proteomics*, vol. 7, no. 1, pp. 79–92, Feb 2010.

[5] A. W. Bell, E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, J. J. M. Bergeron, and HUPO Test Sample Working Group, "A hupo test sample study reveals common problems in mass spectrometry-based proteomics," *Nat Methods*, vol. 6, no. 6, pp. 423–30, Jun 2009.

[6] I. Cadron, T. Van Gorp, F. Amant, I. Vergote, P. Moerman, E. Waelkens, A. Daemen, R. Van de Plas, B. De Moor, and R. Zeillinger, "The use of laser microdissection and seldi-tof ms in ovarian cancer tissue to identify protein profiles," *Anticancer Res*, vol. 29, no. 4, pp. 1039–45, Apr 2009.

[7] H. J. Issaq and J. Blonder, "Electrophoresis and liquid chromatography/tandem mass spectrometry in disease biomarker discovery," *J Chromatogr B Analyt Technol Biomed Life Sci*, vol. 877, no. 13, pp. 1222–8, May 2009.

[8] 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, no. 23, pp. 4751–60, Dec 1997.

[9] J. Franck, K. Arafah, M. Elayed, D. Bonnel, D. Vergara, A. Jacquet, D. Vinatier, M. Wisztorski, R. Day, I. Fournier, and M. Salzet, "Maldi imaging mass spectrometry: state of the art technology in clinical proteomics," *Mol Cell Proteomics*, vol. 8, no. 9, pp. 2023–33, Sep 2009.

[10] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, "Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues," *Nat Med*, vol. 7, no. 4, pp. 493–6, Apr 2001.

[11] K. Busch, "Sams – speaking with acronyms in mass spectrometry," *Spectroscopy*, vol. 17, no. 2, pp. 54–63, 2002.

[12] P. L. Ferguson and R. D. Smith, "Proteome analysis by mass spectrometry," *Annu Rev Biophys Biomol Struct*, vol. 32, pp. 399–424, 2003.

[13] J. R. Yates, 3rd, "Mass spectral analysis in proteomics," *Annu Rev Biophys Biomol Struct*, vol. 33, pp. 297–316, 2004.

[14] J. Godovac-Zimmermann and L. R. Brown, "Perspectives for mass spectrometry and functional proteomics," *Mass Spectrom Rev*, vol. 20, no. 1, pp. 1–57, 2001.

[15] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, Mar 2003.

[16] M. S. Lim and K. S. J. Elenitoba-Johnson, "Proteomics in pathology research," *Lab Invest*, vol. 84, no. 10, pp. 1227–44, Oct 2004.

[17] K. Kanninen, G. Goldsteins, S. Auriola, I. Alafuzoff, and J. Koistinaho, "Glycosylation changes in alzheimer's disease as revealed by a proteomic approach," *Neurosci Lett*, vol. 367, no. 2, pp. 235–40, Sep 2004.

[18] 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, no. 1, pp. 67–71, Jan 1999.

[19] S. L. Luxembourg, T. H. Mize, L. A. McDonnell, and R. M. A. Heeren, "High-spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface," *Anal Chem*, vol. 76, no. 18, pp. 5339–44, Sep 2004.

[20] A. F. M. Altelaar, I. Klinkert, K. Jalink, R. P. J. de Lange, R. A. H. Adan, R. M. A. Heeren, and S. R. Piersma, "Gold-enhanced biomolecular surface imaging of cells and tissue by sims and maldi mass spectrometry," *Anal Chem*, vol. 78, no. 3, pp. 734–42, Feb 2006.

[21] L. A. McDonnell, R. M. A. Heeren, R. P. J. de Lange, and I. W. Fletcher, "Higher sensitivity secondary ion mass spectrometry of biological molecules for high resolution, chemically specific imaging," *J Am Soc Mass Spectrom*, vol. 17, no. 9, pp. 1195–202, Sep 2006.

[22] L. A. McDonnell and R. M. A. Heeren, "Imaging mass spectrometry," *Mass Spectrom Rev*, vol. 26, no. 4, pp. 606–43, 2007.

[23] R. M. A. Heeren, D. F. Smith, J. Stauber, B. Kükrer-Kaletas, and L. MacAleese, "Imaging mass spectrometry: hype or hope?" *J Am Soc Mass Spectrom*, vol. 20, no. 6, pp. 1006–14, Jun 2009.

[24] M. Froesch, S. L. Luxembourg, D. Verheijde, and R. M. A. Heeren, "Imaging mass spectrometry using a delay-line detector," *Eur J Mass Spectrom (Chichester, Eng)*, vol. 16, no. 1, pp. 35–45, 2010.

[25] E. R. Amstalden van Hove, D. F. Smith, and R. M. A. Heeren, "A concise review of mass spectrometry imaging," *J Chromatogr A*, Feb 2010.

[26] P. Fragu, J. Clerc, C. Briançon, C. Fourré, J. Jeusset, and S. Halpern, "Recent developments in medical applications of sims microscopy," *Micron*, vol. 25, no. 4, pp. 361–70, 1994.

[27] M. L. Pacholski and N. Winograd, "Imaging with mass spectrometry," *Chem Rev*, vol. 99, no. 10, pp. 2977–3006, Oct 1999.

[28] P. Chaurand, S. A. Schwartz, M. L. Reyzer, and R. M. Caprioli, "Imaging mass spectrometry: principles and potentials," *Toxicol Pathol*, vol. 33, no. 1, pp. 92–101, 2005.

[29] J. B. Weitzman, "Imaging with isotopes: high resolution and quantitation," *J Biol*, vol. 5, no. 6, p. 17, 2006.

[30] M. Wisztorski, D. Croix, E. Macagno, I. Fournier, and M. Salzet, "Molecular maldi imaging: an emerging technology for neuroscience studies," *Dev Neurobiol*, vol. 68, no. 6, pp. 845–58, May 2008.

[31] 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, no. 3, pp. 809–19, Feb 2006.

[32] R. Lemaire, S. A. Menguellet, J. Stauber, V. Marchaudon, J.-P. Lucot, P. Collinet, M.-O. Farine, D. Vinatier, R. Day, P. Ducoroy, M. Salzet, and I. Fournier, "Specific maldi imaging and profiling for biomarker hunting and validation: fragment of the 11s proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker," *J Proteome Res*, vol. 6, no. 11, pp. 4127–34, Nov 2007.

[33] J. Stauber, R. Lemaire, J. Franck, D. Bonnel, D. Croix, R. Day, M. Wisztorski, I. Fournier, and M. Salzet, "Maldi imaging of formalin-fixed paraffin-embedded tissues: application to model animals of parkinson disease for biomarker hunting," *J Proteome Res*, vol. 7, no. 3, pp. 969–78, Mar 2008.

[34] 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, no. 19, pp. 8193–202, Oct 2009.

[35] C. Meriaux, J. Franck, M. Wisztorski, M. Salzet, and I. Fournier, "Liquid ionic matrixes for maldi mass spectrometry imaging of lipids," *J Proteomics*, Feb 2010.

[36] R. M. Caprioli, "Perspectives on imaging mass spectrometry in biology and medicine," *Proteomics*, vol. 8, no. 18, pp. 3679–80, Sep 2008.

[37] D. S. Cornett, M. L. Reyzer, P. Chaurand, and R. M. Caprioli, "Maldi imaging mass spectrometry: molecular snapshots of biochemical systems," *Nat Methods*, vol. 4, no. 10, pp. 828–33, Oct 2007.

[38] I. Fournier, M. Wisztorski, and M. Salzet, "Tissue imaging using maldi-ms: a new frontier of histopathology proteomics," *Expert Rev Proteomics*, vol. 5, no. 3, pp. 413–24, Jun 2008.

[39] B. Långström, P. E. Andrén, O. Lindhe, M. Svedberg, and H. Hall, "In vitro imaging techniques in neurodegenerative diseases," *Mol Imaging Biol*, vol. 9, no. 4, pp. 161–75, 2007.

[40] P. Chaurand, S. A. Schwartz, and R. M. Caprioli, "Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections," *Curr Opin Chem Biol*, vol. 6, no. 5, pp. 676–81, Oct 2002.

[41] P. Chaurand, S. A. Schwartz, D. Billheimer, B. J. Xu, A. Crecelius, and R. M. Caprioli, "Integrating histology and imaging mass spectrometry," *Anal Chem*, vol. 76, no. 4, pp. 1145–55, Feb 2004.

[42] M. Hilario, A. Kalousis, C. Pellegrini, and M. Müller, "Processing and classification of protein mass spectra," *Mass Spectrom Rev*, vol. 25, no. 3, pp. 409–49, 2006.

[43] R. Hussong and A. Hildebrandt, "Signal processing in proteomics," *Methods Mol Biol*, vol. 604, pp. 145–61, 2010.

[44] L. Nie, G. Wu, and W. Zhang, "Statistical application and challenges in global gel-free proteomic analysis by mass spectrometry," *Crit Rev Biotechnol*, vol. 28, no. 4, pp. 297–307, 2008.

[45] 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, no. 2-3, pp. 212–221, Feb 2007.

[46] M. Kinter and N. E. Sherman, *Protein sequencing and identification using tandem mass spectrometry*. New York: John Wiley, 2000.

[47] B. Lu, A. Motoyama, C. Ruse, J. Venable, and J. R. Yates, 3rd, "Improving protein identification sensitivity by combining ms and ms/ms information for shotgun proteomics using ltq-orbitrap high mass accuracy data," *Anal Chem*, vol. 80, no. 6, pp. 2018–25, Mar 2008.

[48] L. Martens and R. Apweiler, "Algorithms and databases," *Methods Mol Biol*, vol. 564, pp. 245–59, 2009.

[49] J. Stauber, L. MacAleese, J. Franck, E. Claude, M. Snel, B. K. Kaletas, I. M. V. D. Wiel, M. Wisztorski, I. Fournier, and R. M. A. Heeren, "On-tissue protein identification and imaging by maldi-ion mobility mass spectrometry," *J Am Soc Mass Spectrom*, vol. 21, no. 3, pp. 338–47, Mar 2010.

[50] J. R. Yates, C. I. Ruse, and A. Nakorchevsky, "Proteomics by mass spectrometry: approaches, advances, and applications," *Annu Rev Biomed Eng*, vol. 11, pp. 49–79, 2009.

[51] A. R. Farley and A. J. Link, "Identification and quantification of protein post-translational modifications," *Methods Enzymol*, vol. 463, pp. 725–63, 2009.

[52] M. W. Duncan, P. J. Gale, and A. L. Yergey, *The principles of quantitative mass spectrometry*, 1st ed. Denver, Colo.: Rockpool Productions, 2006.

[53] M. W. Duncan, H. Roder, and S. W. Hunsucker, "Quantitative matrix-assisted laser desorption/ionization mass spectrometry," *Brief Funct Genomic Proteomic*, vol. 7, no. 5, pp. 355–70, Sep 2008.

[54] H. Humbert, M. D. Cabiac, J. Barradas, and C. Gerbeau, "Evaluation of pharmacokinetic studies: is it useful to take into account concentrations below the limit of quantification?" *Pharm Res*, vol. 13, no. 6, pp. 839–45, Jun 1996.

[55] G. Liebisch, M. Binder, R. Schifferer, T. Langmann, B. Schulz, and G. Schmitz, "High throughput quantification of cholesterol and cholesteryl ester by electrospray ionization tandem mass spectrometry (esi-ms/ms)," *Biochim Biophys Acta*, vol. 1761, no. 1, pp. 121–8, Jan 2006.

[56] D. Mims and D. Hercules, "Quantification of bile acids directly from plasma by maldi-tof-ms," *Anal Bioanal Chem*, vol. 378, no. 5, pp. 1322–6, Mar 2004.

[57] N. G. Ahn, J. B. Shabb, W. M. Old, and K. A. Resing, "Achieving in-depth proteomics profiling by mass spectrometry," *ACS Chem Biol*, vol. 2, no. 1, pp. 39–52, Jan 2007.

[58] P. C. Carvalho, J. Hewel, V. C. Barbosa, and J. R. Yates, 3rd, "Identifying differences in protein expression levels by spectral counting and feature selection," *Genet Mol Res*, vol. 7, no. 2, pp. 342–56, 2008.

[59] F. Hillenkamp, M. Karas, R. C. Beavis, and B. T. Chait, "Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers," *Anal Chem*, vol. 63, no. 24, pp. 1193A–1203A, Dec 1991.

[60] R. Knochenmuss and R. Zenobi, "Maldi ionization: the role of in-plume processes," *Chem Rev*, vol. 103, no. 2, pp. 441–52, Feb 2003.

[61] Y. V. Karpievitch, T. Taverner, J. N. Adkins, S. J. Callister, G. A. Anderson, R. D. Smith, and A. R. Dabney, "Normalization of peak intensities in bottom-up ms-based proteomics using singular value decomposition," *Bioinformatics*, vol. 25, no. 19, pp. 2573–80, Oct 2009.

[62] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *J Comput Biol*, vol. 7, no. 6, pp. 819–37, 2000.

[63] M. Mank, B. Stahl, and G. Boehm, "2,5-dihydroxybenzoic acid butylamine and other ionic liquid matrixes for enhanced maldi-ms analysis of biomolecules," *Anal Chem*, vol. 76, no. 10, pp. 2938–50, May 2004.

[64] G. Rasmussen and T. Isenhour, "The evaluation of mass spectral search algorithms," *J. Chem. Inf. Comput. Sci*, vol. 19, no. 3, pp. 179–186, 1979.

[65] Z. B. Alfassi, "On the normalization of a mass spectrum for comparison of two spectra," *J Am Soc Mass Spectrom*, vol. 15, no. 3, pp. 385–7, Mar 2004.

[66] A. D. McNaught and A. Wilkinson, *Compendium of chemical terminology: IUPAC recommendations*, 2nd ed. Oxford: Blackwell Science, 1997. [Online]. Available: http://goldbook.iupac.org/index.html

[67] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer, 2002.

[68] M. Ringnér, "What is principal component analysis?" *Nat Biotechnol*, vol. 26, no. 3, pp. 303–4, Mar 2008.

[69] J. V. Stone, *Independent component analysis: a tutorial introduction*. Cambridge, Mass.: MIT Press, 2004.

[70] B. De Moor and P. Van Dooren, "Generalizations of the qr and the singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 4, pp. 993–1014, Oct 1992.

[71] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial and Applied Mathematics, Series B*, vol. 2, no. 2, pp. 205–224, 1965.

[72] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[73] L. De Lathauwer and P. Comon, "Blind source separation and multichannel deconvolution," *Signal Processing*, vol. 73, pp. 1 – 2, 1999.

[74] A. Broersen, "Feature visualization in large scale imaging mass spectrometry data," Ph.D. dissertation, Technische Universiteit Eindhoven, Mar 2009.

[75] 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. 458–69, 2007.

[76] R. Van de Plas, B. De Moor, and E. Waelkens, "Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted pca," in *Proceedings of the Third IEEE/NIH BISTI Life Science Systems and Applications Workshop (LISSA)*. IEEE, Nov 2007, pp. 209 –212.

[77] L. De Lathauwer, B. De Moor, and J. Vandewalle, "An introduction to independent component analysis," *Journal of chemometrics*, vol. 14, no. 3, pp. 123–149, 2000.

[78] 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, no. 19, pp. 6118–24, Oct 2005.

[79] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Concise representation of mass spectrometry images by probabilistic latent semantic analysis," *Anal Chem*, vol. 80, no. 24, pp. 9649–58, Dec 2008.

[80] S.-O. Deininger, M. P. Ebert, A. Fütterer, M. Gerhard, and C. Röcken, "Maldi imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers," *J Proteome Res*, vol. 7, no. 12, pp. 5230–6, Dec 2008.

[81] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 2002.

[82] Y. C. Ling, M. T. Bernius, and G. H. Morrison, "Simips: secondary ion mass image processing system," *J Chem Inf Comput Sci*, vol. 27, no. 2, pp. 86–94, May 1987.

[83] L. Klerk, A. Broersen, I. Fletcher, R. van Liere, and R. Heeren, "Extended data analysis strategies for high resolution imaging ms: New methods to deal with extremely large image hyperspectral datasets," *International journal of mass spectrometry*, vol. 260, no. 2-3, pp. 222–236, 2007.

[84] N. Hattori and Y. Mizuno, "Pathogenetic mechanisms of parkin in parkinson's disease," *Lancet*, vol. 364, no. 9435, pp. 722–4, 2004.

[85] F. O. Walker, "Huntington's disease," *Lancet*, vol. 369, no. 9557, pp. 218–28, Jan 2007.

[86] V. Silani, L. Cova, M. Corbo, A. Ciammola, and E. Polli, "Stem-cell therapy for amyotrophic lateral sclerosis," *Lancet*, vol. 364, no. 9429, pp. 200–2, 2004.

[87] A. M. Clement, M. D. Nguyen, E. A. Roberts, M. L. Garcia, S. Boillée, M. Rule, A. P. McMahon, W. Doucette, D. Siwek, R. J. Ferrante, R. H. Brown, Jr, J.-P. Julien, L. S. B. Goldstein, and D. W. Cleveland, "Wild-type nonneuronal cells extend survival of sod1 mutant motor neurons in als mice," *Science*, vol. 302, no. 5642, pp. 113–7, Oct 2003.

[88] W. M. C. van Roon-Mom, S. J. Reid, R. L. M. Faull, and R. G. Snell, "Tata-binding protein in neurodegenerative disease," *Neuroscience*, vol. 133, no. 4, pp. 863–72, 2005.

[89] D. J. Brooks and P. Piccini, "Imaging in parkinson's disease: the role of monoamines in behavior," *Biol Psychiatry*, vol. 59, no. 10, pp. 908–18, May 2006.

[90] J. Götz, N. Deters, A. Doldissen, L. Bokhari, Y. Ke, A. Wiesner, N. Schonrock, and L. M. Ittner, "A decade of tau transgenic animal models and beyond," *Brain Pathol*, vol. 17, no. 1, pp. 91–103, Jan 2007.

[91] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*, 4th ed. New York: Marcel Dekker, 2003, vol. v. 168.

[92] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.

[93] M. Hanselmann, U. Köthe, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Toward digital staining using imaging mass spectrometry and random forests," *J Proteome Res*, vol. 8, no. 7, pp. 3558–67, Jul 2009.

[94] R. Van de Plas, K. Pelckmans, B. De Moor, and E. Waelkens, "Spatial querying of imaging mass spectrometry data: A nonnegative least squares approach," *The NIPS workshop on Machine Learning in Computational Biology (NIPS MLCB), Whistler (B.C.), Canada*, Dec 2007.

[95] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*. River Edge, NJ: World Scientific, 2002.

[96] D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, 2000.

[97] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, 1992, vol. 61.

[98] E. Lange, C. Gröpl, K. Reinert, O. Kohlbacher, and A. Hildebrandt, "High-accuracy peak picking of proteomics data using wavelet techniques," *Pac Symp Biocomput*, pp. 243–54, 2006.

[99] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–65, Sep 2006.

[100] W. Zhao and C. E. Davis, "Swarm intelligence based wavelet coefficient feature selection for mass spectral classification: an application to proteomics data," *Anal Chim Acta*, vol. 651, no. 1, pp. 15–23, Sep 2009.

[101] Y. Liu, "Feature extraction and dimensionality reduction for mass spectrometry data," *Comput Biol Med*, vol. 39, no. 9, pp. 818–23, Sep 2009.

[102] H.-M. Hsueh, H.-C. Kuo, and C.-A. Tsai, "Multispectra cwt-based algorithm (mcwt) in mass spectra for peak extraction," *J Biopharm Stat*, vol. 18, no. 5, pp. 869–82, 2008.

[103] P. Chen, Y. Lu, and P. B. Harrington, "Application of linear and nonlinear discrete wavelet transforms to maldi-ms measurements of bacteria for classification," *Anal Chem*, vol. 80, no. 19, pp. 7218–25, Oct 2008.

[104] D. Kwon, M. Vannucci, J. J. Song, J. Jeong, and R. M. Pfeiffer, "A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise," *Proteomics*, vol. 8, no. 15, pp. 3019–29, Aug 2008.

[105] T. W. Randolph, "Scale-based normalization of spectral data," *Cancer Biomark*, vol. 2, no. 3-4, pp. 135–44, 2006.

[106] X. Li, J. Li, and X. Yao, "A wavelet-based data pre-processing analysis approach in mass spectrometry," *Comput Biol Med*, vol. 37, no. 4, pp. 509–16, Apr 2007.

[107] T. W. Randolph and Y. Yasui, "Multiscale processing of mass spectrometry data," *Biometrics*, vol. 62, no. 2, pp. 589–97, Jun 2006.

[108] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung, and H. M. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Proteomics*, vol. 5, no. 16, pp. 4107–17, Nov 2005.

[109] J. S. Yu, S. Ongarello, R. Fiedler, X. W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics*, vol. 21, no. 10, pp. 2200–9, May 2005.

[110] H. Montazery-Kordy, M. H. Miran-Baygi, and M. H. Moradi, "A data-mining approach to biomarker identification from protein profiles using discrete stationary wavelet transform," *J Zhejiang Univ Sci B*, vol. 9, no. 11, pp. 863–70, Nov 2008.

[111] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, vol. 21, no. 9, pp. 1764–75, May 2005.

[112] S. Nikolov, H. Hutter, and M. Grasserbauer, "De-noising of sims images via wavelet shrinkage," *Chemometrics and Intelligent Laboratory Systems*, vol. 34, no. 2, pp. 263–273, 1996.

[113] S. Mallat, "Multiresolution approximation and wavelets," *Trans. Amer. Math. Soc*, vol. 315, pp. 69–88, 1989.

[114] A. Bradley, "Shift-invariance in the discrete wavelet transform," *Proceedings of VIIth Digital Image Computing: Techniques and Applications. Sydney, December, 10th-12th*, 2003.

[115] B. Bakshi, "Multiscale pca with application to multivariate statistical process monitoring," *AIChE Journal*, vol. 44, no. 7, pp. 1596–1610, 1998.

[116] Y. Qu, B.-L. Adam, M. Thornquist, J. D. Potter, M. L. Thompson, Y. Yasui, J. Davis, P. F. Schellhammer, L. Cazares, M. Clements, G. L. Wright, Jr, and Z. Feng, "Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data," *Biometrics*, vol. 59, no. 1, pp. 143–51, Mar 2003.

[117] L. A. McDonnell, A. van Remoortere, R. J. M. van Zeijl, and A. M. Deelder, "Mass spectrometry image correlation: quantifying colocalization," *J Proteome Res*, vol. 7, no. 8, pp. 3619–27, Aug 2008.

[118] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[119] B. H. Junker and F. Schreiber, *Analysis of biological networks*. New York: John Wiley & Sons, 2008.

[120] R. Kaddurah-Daouk and K. R. R. Krishnan, "Metabolomics: a global biochemical approach to the study of central nervous system diseases," *Neuropsychopharmacology*, vol. 34, no. 1, pp. 173–86, Jan 2009.

[121] C. Abate-Shen and M. M. Shen, "Diagnostics: The prostate-cancer metabolome," *Nature*, vol. 457, no. 7231, pp. 799–800, Feb 2009.

[122] D. H. Geschwind and G. Konopka, "Neuroscience in the era of functional genomics and systems biology," *Nature*, vol. 461, no. 7266, pp. 908–15, Oct 2009.

[123] T. J. R. Harris and F. McCormick, "The molecular pathology of cancer," *Nat Rev Clin Oncol*, Mar 2010.

[124] D. Thomas, "Gene-environment-wide association studies: emerging approaches," *Nat Rev Genet*, vol. 11, no. 4, pp. 259–272, Mar 2010.

[125] M. C. Hall, "Proteomics modifies our understanding of cell cycle complexity," *Sci Signal*, vol. 3, no. 106, p. pe4, 2010.

[126] D. B. Kell, "Metabolomics and systems biology: making sense of the soup," *Curr Opin Microbiol*, vol. 7, no. 3, pp. 296–307, Jun 2004.

[127] X. Feng, X. Liu, Q. Luo, and B.-F. Liu, "Mass spectrometry in systems biology: an overview," *Mass Spectrom Rev*, vol. 27, no. 6, pp. 635–60, 2008.

[128] J. K. Nicholson and J. C. Lindon, "Systems biology: Metabonomics," *Nature*, vol. 455, no. 7216, pp. 1054–6, Oct 2008.

[129] M. Gstaiger and R. Aebersold, "Applying mass spectrometry-based proteomics to genetics, genomics and network biology," *Nat Rev Genet*, vol. 10, no. 9, pp. 617–27, Sep 2009.

[130] D. Daneman, "State of the world's children with diabetes," *Pediatr Diabetes*, vol. 10, no. 2, pp. 120–6, Apr 2009.

[131] T. Mandrup-Poulsen, "Diabetes," *BMJ*, vol. 316, no. 7139, pp. 1221–5, Apr 1998.

[132] M. Reimann, E. Bonifacio, M. Solimena, P. E. H. Schwarz, B. Ludwig, M. Hanefeld, and S. R. Bornstein, "An update on preventive and regenerative therapies in diabetes mellitus," *Pharmacol Ther*, vol. 121, no. 3, pp. 317–31, Mar 2009.

[133] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–53, May 2004.

[134] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *Science*, vol. 312, no. 5771, pp. 212–7, Apr 2006.

[135] W. J. Griffiths and Y. Wang, "Mass spectrometry: from proteomics to metabolomics and lipidomics," *Chem Soc Rev*, vol. 38, no. 7, pp. 1882–96, Jul 2009.

[136] R. J. A. Goodwin, S. R. Pennington, and A. R. Pitt, "Protein and peptides in pictures: imaging with maldi mass spectrometry," *Proteomics*, vol. 8, no. 18, pp. 3785–800, Sep 2008.

[137] L. MacAleese, J. Stauber, and R. M. A. Heeren, "Perspectives for imaging mass spectrometry in the proteomics landscape," *Proteomics*, vol. 9, no. 4, pp. 819–34, Feb 2009.

[138] B. K. Kaletaş, I. M. van der Wiel, J. Stauber, Lennard J. Dekker, C. Güzel, J. M. Kros, T. M. Luider, and R. M. A. Heeren, "Sample preparation issues for tissue imaging by imaging ms," *Proteomics*, vol. 9, no. 10, pp. 2622–33, May 2009.

[139] E. H. Seeley, S. R. Oppenheimer, D. Mi, P. Chaurand, and R. M. Caprioli, "Enhancement of protein sensitivity for maldi imaging mass spectrometry after chemical treatment of tissue sections," *J Am Soc Mass Spectrom*, vol. 19, no. 8, pp. 1069–77, Aug 2008.

[140] S. M. Puolitaival, K. E. Burnum, D. S. Cornett, and R. M. Caprioli, "Solvent-free matrix dry-coating for maldi imaging of phospholipids," *J Am Soc Mass Spectrom*, vol. 19, no. 6, pp. 882–6, Jun 2008.

[141] A. F. M. Altelaar, S. L. Luxembourg, L. A. McDonnell, S. R. Piersma, and R. M. A. Heeren, "Imaging mass spectrometry at cellular length scales," *Nat Protoc*, vol. 2, no. 5, pp. 1185–96, 2007.

[142] M. Wisztorski, R. Lemaire, J. Stauber, S. A. Menguelet, D. Croix, O. J. Mathé, R. Day, M. Salzet, and I. Fournier, "New developments in maldi imaging for pathology proteomic studies," *Curr Pharm Des*, vol. 13, no. 32, pp. 3317–24, 2007.

[143] S. Speier, D. Nyqvist, M. Köhler, A. Caicedo, I. B. Leibiger, and P.-O. Berggren, "Noninvasive high-resolution in vivo imaging of cell biology in the anterior chamber of the mouse eye," *Nat Protoc*, vol. 3, no. 8, pp. 1278–86, 2008.

[144] S. Speier, D. Nyqvist, O. Cabrera, J. Yu, R. D. Molano, A. Pileggi, T. Moede, M. Köhler, J. Wilbertz, B. Leibiger, C. Ricordi, I. B. Leibiger, A. Caicedo, and P.-O. Berggren, "Noninvasive in vivo imaging of pancreatic islet cell biology," *Nat Med*, vol. 14, no. 5, pp. 574–8, May 2008.

[145] E. d. Hoffmann and V. Stroobant, *Mass spectrometry: principles and applications*, 2nd ed. Chichester: Wiley, 2001.

[146] Z. Takáts, J. M. Wiseman, and R. G. Cooks, "Ambient mass spectrometry using desorption electrospray ionization (desi): instrumentation, mechanisms and applications in forensics, chemistry, and biology," *J Mass Spectrom*, vol. 40, no. 10, pp. 1261–75, Oct 2005.

[147] R. G. Cooks, Z. Ouyang, Z. Takats, and J. M. Wiseman, "Detection technologies. ambient mass spectrometry," *Science*, vol. 311, no. 5767, pp. 1566–70, Mar 2006.

[148] J. M. Wiseman, D. R. Ifa, A. Venter, and R. G. Cooks, "Ambient molecular imaging by desorption electrospray ionization mass spectrometry," *Nat Protoc*, vol. 3, no. 3, pp. 517–24, 2008.

[149] S. P. Pasilis, V. Kertesz, G. J. Van Berkel, M. Schulz, and S. Schorcht, "Hptlc/desi-ms imaging of tryptic protein digests separated in two dimensions," *J Mass Spectrom*, vol. 43, no. 12, pp. 1627–35, Dec 2008.

[150] M. C. Wood, D. K. Busby, and P. B. Farnsworth, "Microscopic imaging of glass surfaces under the effects of desorption electrospray ionization," *Anal Chem*, vol. 81, no. 15, pp. 6407–15, Aug 2009.

[151] R. Wolstenholme, R. Bradshaw, M. R. Clench, and S. Francese, "Study of latent fingermarks by matrix-assisted laser desorption/ionisation mass spectrometry imaging of endogenous lipids," *Rapid Commun Mass Spectrom*, vol. 23, no. 19, pp. 3031–9, Oct 2009.

[152] E. Esquenazi, Y.-L. Yang, J. Watrous, W. H. Gerwick, and P. C. Dorrestein, "Imaging mass spectrometry of natural products," *Nat Prod Rep*, vol. 26, no. 12, pp. 1521–34, Dec 2009.

[153] C. Wu, D. R. Ifa, N. E. Manicke, and R. G. Cooks, "Molecular imaging of adrenal gland by desorption electrospray ionization mass spectrometry," *Analyst*, vol. 135, no. 1, pp. 28–32, Jan 2010.

[154] J. Watrous, N. Hendricks, M. Meehan, and P. C. Dorrestein, "Capturing bacterial metabolic exchange using thin film desorption electrospray ionization-imaging mass spectrometry," *Anal Chem*, vol. 82, no. 5, pp. 1598–600, Mar 2010.

[155] D. R. Ifa, C. Wu, Z. Ouyang, and R. G. Cooks, "Desorption electrospray ionization and other ambient ionization methods: current progress and preview," *Analyst*, vol. 135, no. 4, pp. 669–81, Apr 2010.

[156] Y. Li, B. Shrestha, and A. Vertes, "Atmospheric pressure molecular imaging by infrared maldi mass spectrometry," *Anal Chem*, vol. 79, no. 2, pp. 523–32, Jan 2007.

[157] P. Nemes and A. Vertes, "Laser ablation electrospray ionization for atmospheric pressure, in vivo, and imaging mass spectrometry," *Anal Chem*, vol. 79, no. 21, pp. 8098–106, Nov 2007.

[158] Y. Li, B. Shrestha, and A. Vertes, "Atmospheric pressure infrared maldi imaging mass spectrometry for plant metabolomics," *Anal Chem*, vol. 80, no. 2, pp. 407–20, Jan 2008.

[159] P. Nemes, A. A. Barton, Y. Li, and A. Vertes, "Ambient molecular imaging and depth profiling of live tissue by infrared laser ablation electrospray ionization mass spectrometry," *Anal Chem*, vol. 80, no. 12, pp. 4575–82, Jun 2008.

[160] P. Nemes, A. A. Barton, and A. Vertes, "Three-dimensional imaging of metabolites in tissues under ambient conditions by laser ablation electrospray ionization mass spectrometry," *Anal Chem*, vol. 81, no. 16, pp. 6668–75, Aug 2009.

[161] P. Nemes, A. S. Woods, and A. Vertes, "Simultaneous imaging of small metabolites and lipids in rat brain tissues at atmospheric pressure by laser ablation electrospray ionization mass spectrometry," *Anal Chem*, vol. 82, no. 3, pp. 982–8, Feb 2010.

[162] A. Kindness, C. N. Sekaran, and J. Feldmann, "Two-dimensional mapping of copper and zinc in liver sections by laser ablation-inductively coupled plasma mass spectrometry," *Clin Chem*, vol. 49, no. 11, pp. 1916–23, Nov 2003.

[163] R. W. Hutchinson, A. G. Cox, C. W. McLeod, P. S. Marshall, A. Harper, E. L. Dawson, and D. R. Howlett, "Imaging and spatial distribution of beta-amyloid peptide and metal ions in alzheimer's plaques by laser ablation-inductively coupled plasma-mass spectrometry," *Anal Biochem*, vol. 346, no. 2, pp. 225–33, Nov 2005.

[164] T. Hirata and Z. Miyazaki, "High-speed camera imaging for laser ablation process: for further reliable elemental analysis using inductively coupled plasma-mass spectrometry," *Anal Chem*, vol. 79, no. 1, pp. 147–52, Jan 2007.

[165] J. S. Becker, A. Matusch, C. Depboylu, J. Dobrowolska, and M. V. Zoriy, "Quantitative imaging of selenium, copper, and zinc in thin sections of biological tissues (slugs-genus arion) measured by laser ablation inductively coupled plasma mass spectrometry," *Anal Chem*, vol. 79, no. 16, pp. 6074–80, Aug 2007.

[166] J. S. Becker, J. S. Becker, M. V. Zoriy, J. Dobrowolska, and A. Matucsh, "Imaging mass spectrometry in biological tissues by laser ablation inductively coupled plasma mass spectrometry," *Eur J Mass Spectrom (Chichester, Eng)*, vol. 13, no. 1, pp. 1–6, 2007.

[167] J. Seuma, J. Bunch, A. Cox, C. McLeod, J. Bell, and C. Murray, "Combination of immunohistochemistry and laser ablation icp mass spectrometry for imaging of cancer biomarkers," *Proteomics*, vol. 8, no. 18, pp. 3775–84, Sep 2008.

[168] B. Wu, M. Zoriy, Y. Chen, and J. S. Becker, "Imaging of nutrient elements in the leaves of elsholtzia splendens by laser ablation inductively coupled plasma mass spectrometry (la-icp-ms)," *Talanta*, vol. 78, no. 1, pp. 132–7, Apr 2009.

[169] M. V. Zoriy, D. Mayer, and J. S. Becker, "Metal imaging on surface of micro- and nanoelectronic devices by laser ablation inductively coupled plasma mass spectrometry and possibility to measure at nanometer range," *J Am Soc Mass Spectrom*, vol. 20, no. 5, pp. 883–90, May 2009.

[170] J. S. Becker, M. Zoriy, A. Matusch, B. Wu, D. Salber, C. Palm, and J. S. Becker, "Bioimaging of metals by laser ablation inductively coupled plasma mass spectrometry (la-icp-ms)," *Mass Spectrom Rev*, vol. 29, no. 1, pp. 156–75, 2010.

[171] M. C. Santos, M. Wagner, B. Wu, J. Scheider, J. Oehlmann, S. Cadore, and J. S. Becker, "Biomonitoring of metal contamination in a marine prosobranch snail (nassarius reticulatus) by imaging laser ablation inductively coupled plasma mass spectrometry (la-icp-ms)," *Talanta*, vol. 80, no. 2, pp. 428–33, Dec 2009.

[172] A. Matusch, C. Depboylu, C. Palm, B. Wu, G. U. Höglinger, M. K.-H. Schäfer, and J. S. Becker, "Cerebral bioimaging of cu, fe, zn, and mn in the mptp mouse model of parkinson's disease using laser ablation inductively coupled plasma mass spectrometry (la-icp-ms)," *J Am Soc Mass Spectrom*, vol. 21, no. 1, pp. 161–71, Jan 2010.

[173] N. Kamaly, J. A. Pugh, T. L. Kalber, J. Bunch, A. D. Miller, C. W. McLeod, and J. D. Bell, "Imaging of gadolinium spatial distribution in tumor tissue by

laser ablation inductively coupled plasma mass spectrometry," *Mol Imaging Biol*, Nov 2009.

[174] D. S. Gholap, A. Izmer, B. De Samber, J. T. van Elteren, V. S. Selih, R. Evens, K. De Schamphelaere, C. Janssen, L. Balcaen, I. Lindemann, L. Vincze, and F. Vanhaecke, "Comparison of laser ablation-inductively coupled plasma-mass spectrometry and micro-x-ray fluorescence spectrometry for elemental imaging in daphnia magna," *Anal Chim Acta*, vol. 664, no. 1, pp. 19–26, Apr 2010.

[175] A. Végvári, T. E. Fehniger, L. Gustavsson, A. Nilsson, P. E. Andrén, K. Kenne, J. Nilsson, T. Laurell, and G. Marko-Varga, "Essential tactics of tissue preparation and matrix nano-spotting for successful compound imaging mass spectrometry," *J Proteomics*, vol. 73, no. 6, pp. 1270–8, Apr 2010.

[176] K. Sköld, M. Svensson, A. Nilsson, X. Zhang, K. Nydahl, R. M. Caprioli, P. Svenningsson, and P. E. Andrén, "Decreased striatal levels of pep-19 following mptp lesion in the mouse," *J Proteome Res*, vol. 5, no. 2, pp. 262–9, Feb 2006.

[177] R. Castaing and G. Slodzian, "Optique corpusculaire—premiers essais de microanalyse par emission ionique secondaire," *Microscopie*, vol. 1, pp. 395–399, 1962.

[178] T. L. Colliver, C. L. Brummel, M. L. Pacholski, F. D. Swanek, A. G. Ewing, and N. Winograd, "Atomic and molecular imaging at the single-cell level with tof-sims," *Anal Chem*, vol. 69, no. 13, pp. 2225–31, Jul 1997.

[179] 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 Applied Computing (ACM SAC)*, Association for Computing Machinery (ACM). New York, NY, USA: ACM, Mar 2008, pp. 1307–1308, conference Chair-Wainwright, Roger L. and Conference Chair-Haddad, Hisham M.

[180] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, and J. A. K. Suykens, "A tutorial on support vector machine-based methods for classification problems in chemometrics," *Analytica Chimica Acta*, vol. 665, no. 2, pp. 129–145, Apr 2010.

[181] U. Sauer and N. Zamboni, "From biomarkers to integrated network responses," *Nat Biotechnol*, vol. 26, no. 10, pp. 1090–2, Oct 2008.

[182] J.-D. J. Han, "Understanding biological functions through molecular networks," *Cell Res*, vol. 18, no. 2, pp. 224–37, Feb 2008.

[183] R. Lemaire, A. Desmons, J. C. Tabet, R. Day, M. Salzet, and I. Fournier, "Direct analysis and maldi imaging of formalin-fixed, paraffin-embedded tissue sections," *J Proteome Res*, vol. 6, no. 4, pp. 1295–305, Apr 2007.

[184] O. Jardin-Mathé, D. Bonnel, J. Franck, M. Wisztorski, E. Macagno, I. Fournier, and M. Salzet, "Mitics (maldi imaging team imaging computing system): a new open source mass spectrometry imaging software," *J Proteomics*, vol. 71, no. 3, pp. 332–45, Aug 2008.

[185] P. J. Todd, T. G. Schaaff, P. Chaurand, and R. M. Caprioli, "Organic ion imaging of biological tissue with secondary ion mass spectrometry and matrix-assisted laser desorption/ionization," *J Mass Spectrom*, vol. 36, no. 4, pp. 355–69, Apr 2001.

[186] P. Chaurand and R. M. Caprioli, "Direct profiling and imaging of peptides and proteins from mammalian cells and tissue sections by mass spectrometry," *Electrophoresis*, vol. 23, no. 18, pp. 3125–35, Sep 2002.

[187] P. Chaurand, S. Fouchécourt, B. B. DaGue, B. J. Xu, M. L. Reyzer, M.-C. Orgebin-Crist, and R. M. Caprioli, "Profiling and imaging proteins in the mouse epididymis by imaging mass spectrometry," *Proteomics*, vol. 3, no. 11, pp. 2221–39, Nov 2003.

[188] P. Chaurand, S. A. Schwartz, and R. M. Caprioli, "Assessing protein patterns in disease using imaging mass spectrometry," *J Proteome Res*, vol. 3, no. 2, pp. 245–52, 2004.

[189] P. Chaurand, M. E. Sanders, R. A. Jensen, and R. M. Caprioli, "Proteomics in diagnostic pathology: profiling and imaging proteins directly in tissue sections," *Am J Pathol*, vol. 165, no. 4, pp. 1057–68, Oct 2004.

[190] A. C. Crecelius, D. S. Cornett, R. M. Caprioli, B. Williams, B. M. Dawant, and B. Bodenheimer, "Three-dimensional visualization of protein expression in mouse brain structures using imaging mass spectrometry," *J Am Soc Mass Spectrom*, vol. 16, no. 7, pp. 1093–9, Jul 2005.

[191] H. Meistermann, J. L. Norris, H.-R. Aerni, D. S. Cornett, A. Friedlein, A. R. Erskine, A. Augustin, M. C. De Vera Mudry, S. Ruepp, L. Suter, H. Langen, R. M. Caprioli, and A. Ducret, "Biomarker discovery by imaging mass spectrometry: transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat," *Mol Cell Proteomics*, vol. 5, no. 10, pp. 1876–86, Oct 2006.

[192] P. Chaurand, D. S. Cornett, and R. M. Caprioli, "Molecular imaging of thin mammalian tissue sections by mass spectrometry," *Curr Opin Biotechnol*, vol. 17, no. 4, pp. 431–6, Aug 2006.

[193] S. Khatib-Shahidi, M. Andersson, J. L. Herman, T. A. Gillespie, and R. M. Caprioli, "Direct molecular analysis of whole-body animal tissue sections by imaging maldi mass spectrometry," *Anal Chem*, vol. 78, no. 18, pp. 6448–56, Sep 2006.

[194] M. L. Reyzer and R. M. Caprioli, "Maldi-ms-based imaging of small molecules and proteins in tissues," *Curr Opin Chem Biol*, vol. 11, no. 1, pp. 29–35, Feb 2007.

[195] 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, no. 2, pp. 254–62, Feb 2007.

[196] P. Chaurand, K. E. Schriver, and R. M. Caprioli, "Instrument design and characterization for high resolution maldi-ms imaging of tissue sections," *J Mass Spectrom*, vol. 42, no. 4, pp. 476–89, Apr 2007.

[197] J. A. McLean, W. B. Ridenour, and R. M. Caprioli, "Profiling and imaging of tissues by imaging ion mobility-mass spectrometry," *J Mass Spectrom*, vol. 42, no. 8, pp. 1099–105, Aug 2007.

[198] 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, no. 2, pp. 411–23, Feb 2008.

[199] T. K. Sinha, S. Khatib-Shahidi, T. E. Yankeelov, K. Mapara, M. Ehtesham, D. S. Cornett, B. M. Dawant, R. M. Caprioli, and J. C. Gore, "Integrating spatially resolved three-dimensional maldi ims with in vivo magnetic resonance imaging," *Nat Methods*, vol. 5, no. 1, pp. 57–9, Jan 2008.

[200] M. Andersson, M. R. Groseclose, A. Y. Deutch, and R. M. Caprioli, "Imaging mass spectrometry of proteins and peptides: 3d volume reconstruction," *Nat Methods*, vol. 5, no. 1, pp. 101–8, Jan 2008.

[201] W. M. Hardesty and R. M. Caprioli, "In situ molecular imaging of proteins in tissues using mass spectrometry," *Anal Bioanal Chem*, vol. 391, no. 3, pp. 899–903, Jun 2008.

[202] K. E. Burnum, S. Tranguch, D. Mi, T. Daikoku, S. K. Dey, and R. M. Caprioli, "Imaging mass spectrometry reveals unique protein profiles during embryo implantation," *Endocrinology*, vol. 149, no. 7, pp. 3274–8, Jul 2008.

[203] D. S. Cornett, S. L. Frappier, and R. M. Caprioli, "Maldi-fticr imaging mass spectrometry of drugs and metabolites in tissue," *Anal Chem*, vol. 80, no. 14, pp. 5648–53, Jul 2008.

[204] P. Chaurand, J. C. Latham, K. B. Lane, J. A. Mobley, V. V. Polosukhin, P. S. Wirth, L. B. Nanney, and R. M. Caprioli, "Imaging mass spectrometry of intact proteins from alcohol-preserved tissue specimens: bypassing formalin fixation," *J Proteome Res*, vol. 7, no. 8, pp. 3543–55, Aug 2008.

[205] M. R. Groseclose, P. P. Massion, P. Chaurand, and R. M. Caprioli, "High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using maldi imaging mass spectrometry," *Proteomics*, vol. 8, no. 18, pp. 3715–24, Sep 2008.

[206] 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, no. 47, pp. 18 126–31, Nov 2008.

[207] L. A. McDonnell, A. van Remoortere, R. J. M. van Zeijl, H. Dalebout, M. R. Bladergroen, and A. M. Deelder, "Automated imaging ms: Toward high throughput imaging mass spectrometry," *J Proteomics*, Nov 2009.

[208] L. A. McDonnell, T. H. Mize, S. L. Luxembourg, S. Koster, G. B. Eijkel, E. Verpoorte, N. F. de Rooij, and R. M. A. Heeren, "Using matrix peaks to map topography: increased mass resolution and enhanced sensitivity in chemical imaging," *Anal Chem*, vol. 75, no. 17, pp. 4373–81, Sep 2003.

[209] R. M. A. Heeren, A. J. Kleinnijenhuis, L. A. McDonnell, and T. H. Mize, "A mini-review of mass spectrometry using high-performance fticr-ms methods," *Anal Bioanal Chem*, vol. 378, no. 4, pp. 1048–58, Feb 2004.

[210] A. F. M. Altelaar, J. van Minnen, C. R. Jiménez, R. M. A. Heeren, and S. R. Piersma, "Direct molecular imaging of lymnaea stagnalis nervous tissue at subcellular spatial resolution by mass spectrometry," *Anal Chem*, vol. 77, no. 3, pp. 735–41, Feb 2005.

[211] L. A. McDonnell, S. R. Piersma, A. F. MaartenAltelaar, T. H. Mize, S. L. Luxembourg, P. D. E. M. Verhaert, J. van Minnen, and R. M. A. Heeren, "Subcellular imaging mass spectrometry of brain tissue," *J Mass Spectrom*, vol. 40, no. 2, pp. 160–8, Feb 2005.

[212] R. M. A. Heeren, "Proteome imaging: a closer look at life's organization," *Proteomics*, vol. 5, no. 17, pp. 4316–26, Nov 2005.

[213] S. L. Luxembourg, A. R. Vaezaddeh, E. R. Amstalden, C. G. Zimmermann-Ivol, D. F. Hochstrasser, and R. M. A. Heeren, "The molecular scanner in microscope mode," *Rapid Commun Mass Spectrom*, vol. 20, no. 22, pp. 3435–42, 2006.

[214] I. Klinkert, L. A. McDonnell, S. L. Luxembourg, A. F. M. Altelaar, E. R. Amstalden, S. R. Piersma, and R. M. A. Heeren, "Tools and strategies for visualization of large image data sets in high-resolution imaging mass spectrometry," *Rev Sci Instrum*, vol. 78, no. 5, p. 053716, May 2007.

[215] A. Broersen, R. van Liere, A. F. M. Altelaar, R. M. A. Heeren, and L. A. McDonnell, "Automated, feature-based image alignment for high-resolution imaging mass spectrometry of large biological samples," *J Am Soc Mass Spectrom*, vol. 19, no. 6, pp. 823–32, Jun 2008.

[216] A. F. M. Altelaar and R. M. A. Heeren, "Tissue analysis with high-resolution imaging mass spectrometry," *Methods Mol Biol*, vol. 492, pp. 295–308, 2009.

[217] L. A. Klerk, N. P. Lockyer, A. Kharchenko, L. MacAleese, P. Y. W. Dankers, J. C. Vickerman, and R. M. A. Heeren, "C60+ secondary ion microscopy using a delay line detector," *Anal Chem*, vol. 82, no. 3, pp. 801–7, Feb 2010.

[218] J. Krause, M. Stoeckli, and U. P. Schlunegger, "Studies on the selection of new matrices for ultraviolet matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," *Rapid Commun Mass Spectrom*, vol. 10, no. 15, pp. 1927–33, 1996.

[219] M. Faupel, D. Bonenfant, P. Schindler, E. Bertrand, D. Mueller, M. Stoeckli, F. Bitsch, T. Rohner, D. Staab, and J. Van Oostrum, "Biophotonics applied to proteomics," *Subcell Biochem*, vol. 43, pp. 323–38, 2007.

[220] E. G. Solon, A. Schweitzer, M. Stoeckli, and B. Prideaux, "Autoradiography, maldi-ms, and sims-ms imaging in pharmaceutical discovery and development," *AAPS J*, vol. 12, no. 1, pp. 11–26, Mar 2010.

[221] P. Chaurand, M. Stoeckli, and R. M. Caprioli, "Direct profiling of proteins in biological tissue sections by maldi mass spectrometry," *Anal Chem*, vol. 71, no. 23, pp. 5263–70, Dec 1999.

[222] J. M. Koomen, M. Stoeckli, and R. M. Caprioli, "Mapping of surrogate markers of cellular components and structures using laser desorption/ionization mass spectrometry," *J Mass Spectrom*, vol. 35, no. 2, pp. 258–64, Feb 2000.

[223] M. Stoeckli, D. Staab, M. Staufenbiel, K.-H. Wiederhold, and L. Signor, "Molecular imaging of amyloid beta peptides in mouse brain sections using mass spectrometry," *Anal Biochem*, vol. 311, no. 1, pp. 33–9, Dec 2002.

[224] M. Rudin, P. Allegrini, N. Beckmann, H. U. Gremlich, R. Kneuer, D. Laurent, M. Rausch, and M. Stoeckli, "Noninvasive imaging in drug discovery and development," *Ernst Schering Res Found Workshop*, no. 48, pp. 47–75, 2004.

[225] T. C. Rohner, D. Staab, and M. Stoeckli, "Maldi mass spectrometric imaging of biological tissue sections," *Mech Ageing Dev*, vol. 126, no. 1, pp. 177–85, Jan 2005.

[226] M. Rudin, M. Rausch, and M. Stoeckli, "Molecular imaging in drug discovery and development: potential and limitations of nonnuclear methods," *Mol Imaging Biol*, vol. 7, no. 1, pp. 5–13, 2005.

[227] M. Stoeckli, R. Knochenmuss, G. McCombie, D. Mueller, T. Rohner, D. Staab, and K.-H. Wiederhold, "Maldi ms imaging of amyloid," *Methods Enzymol*, vol. 412, pp. 94–106, 2006.

[228] M. Stoeckli, D. Staab, A. Schweitzer, J. Gardiner, and D. Seebach, "Imaging of a beta-peptide distribution in whole-body mice sections by maldi mass spectrometry," *J Am Soc Mass Spectrom*, vol. 18, no. 11, pp. 1921–4, Nov 2007.

[229] G. Hopfgartner, E. Varesio, and M. Stoeckli, "Matrix-assisted laser desorption/ionization mass spectrometric imaging of complete rat sections using a triple quadrupole linear ion trap," *Rapid Commun Mass Spectrom*, vol. 23, no. 6, pp. 733–6, Mar 2009.

[230] K. Lemaire, M. A. Ravier, A. Schraenen, J. W. M. Creemers, R. Van de Plas, M. Granvik, L. Van Lommel, E. Waelkens, F. Chimienti, G. A. Rutter, P. Gilon, P. A. in't Veld, and F. C. Schuit, "Insulin crystallization depends on zinc transporter znt8 expression, but is not required for normal glucose homeostasis in mice," *Proc Natl Acad Sci U S A*, vol. 106, no. 35, pp. 14 872–7, Sep 2009.

[231] Y. Yasui, M. Pepe, M. L. Thompson, B.-L. Adam, G. L. Wright, Jr, Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, no. 3, pp. 449–63, Jul 2003.

[232] H. Nygren and P. Malmberg, "High resolution imaging by organic secondary ion mass spectrometry," *Trends Biotechnol*, vol. 25, no. 11, pp. 499–504, Nov 2007.

[233] H. Hutter, C. Brunner, S. Nikolov, C. Mittermayer, and M. Grasserbauer, "Imaging surface spectroscopy for two- and three-dimensional characterization of materials," *Anal Bioanal Chem*, vol. 355, no. 5-6, pp. 585–90, Jun 1996.

[234] H. Shin, M. P. Sampat, S. F. Bish, J. M. Koomen, and M. K. Markey, "Statistical characterization of chemical noise in maldi tof ms by wavelet analysis of multiple noise realizations," *AMIA Annu Symp Proc*, p. 1092, 2006.

[235] K. Noy and D. Fasulo, "Improved model-based, platform-independent feature extraction for mass spectrometry," *Bioinformatics*, vol. 23, no. 19, pp. 2528–35, Oct 2007.

[236] F.-M. Schleif, T. Villmann, M. Kostrzewa, B. Hammer, and A. Gammerman, "Cancer informatics by prototype networks in mass spectrometry," *Artif Intell Med*, vol. 45, no. 2-3, pp. 215–28, 2009.

[237] R. Tautenhahn, C. Böttcher, and S. Neumann, "Highly sensitive feature detection for high resolution lc/ms," *BMC Bioinformatics*, vol. 9, p. 504, 2008.

[238] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for maldi mass spectrometry data analysis," *BMC Bioinformatics*, vol. 10, p. 4, 2009.

[239] T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. E. M. Tollenaar, P. Maass, and H. Thiele, "Biomarker discovery in maldi-tof serum protein profiles using discrete wavelet transformation," *Bioinformatics*, vol. 25, no. 5, pp. 643–9, Mar 2009.

[240] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–91, Aug 2003.

[241] S. Zhang, T. J. DeGraba, H. Wang, G. T. Hoehn, D. A. Gonzales, A. F. Suffredini, W.-K. Ching, M. K. Ng, X. Zhou, and S. T. C. Wong, "A novel peak detection approach with chemical noise removal using short-time fft for protof ms data," *Proteomics*, vol. 9, no. 15, pp. 3833–42, Aug 2009.

[242] L. Minerva, S. Clerens, G. Baggerman, and L. Arckens, "Direct profiling and identification of peptide expression differences in the pancreas of control and ob/ob mice by imaging mass spectrometry," *Proteomics*, vol. 8, no. 18, pp. 3763–74, Sep 2008.

[243] Y. Sugiura, Y. Konishi, N. Zaima, S. Kajihara, H. Nakanishi, R. Taguchi, and M. Setou, "Visualization of the cell-selective distribution of pufa-containing phosphatidylcholines in mouse brain by imaging mass spectrometry," *J Lipid Res*, vol. 50, no. 9, pp. 1776–88, Sep 2009.

[244] J. S. Becker and N. Jakubowski, "The synergy of elemental and biomolecular mass spectrometry: new analytical strategies in life sciences," *Chem Soc Rev*, vol. 38, no. 7, pp. 1969–83, Jul 2009.

[245] R. Matthiesen and A. S. Carvalho, "Methods and algorithms for relative quantitative proteomics by mass spectrometry," *Methods Mol Biol*, vol. 593, pp. 187–204, 2010.

[246] S. D. Patterson, "Data analysis–the achilles heel of proteomics," *Nat Biotechnol*, vol. 21, no. 3, pp. 221–2, Mar 2003.

[247] E. O. Voit and J. Almeida, "Decoupling dynamical systems for pathway identification from metabolic profiles," *Bioinformatics*, vol. 20, no. 11, pp. 1670–81, Jul 2004.

[248] G. A. Satten, S. Datta, H. Moura, A. R. Woolfitt, M. d. G. Carvalho, G. M. Carlone, B. K. De, A. Pavlopoulos, and J. R. Barr, "Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens," *Bioinformatics*, vol. 20, no. 17, pp. 3128–36, Nov 2004.

[249] T. Fushiki, H. Fujisawa, and S. Eguchi, "Identification of biomarkers from mass spectrometry data using a "common" peak approach," *BMC Bioinformatics*, vol. 7, p. 358, 2006.

[250] W. Urfer, M. Grzegorczyk, and K. Jung, "Statistics for proteomics: a review of tools for analyzing experimental data," *Proteomics*, vol. 6 Suppl 2, pp. 48–55, Sep 2006.

[251] B. J. Tyler, G. Rayal, and D. G. Castner, "Multivariate analysis strategies for processing tof-sims images of biomaterials," *Biomaterials*, vol. 28, no. 15, pp. 2412–23, May 2007.

[252] D. Fenyö and R. C. Beavis, "Informatics development: challenges and solutions for maldi mass spectrometry," *Mass Spectrom Rev*, vol. 27, no. 1, pp. 1–19, 2008.

[253] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, and M. Oresic, "Normalization method for metabolomics data using optimal selection of multiple internal standards," *BMC Bioinformatics*, vol. 8, p. 93, 2007.

[254] K. R. Lee, X. Lin, D. C. Park, and S. Eslava, "Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method," *Proteomics*, vol. 3, no. 9, pp. 1680–6, Sep 2003.

[255] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, vol. 20, no. 5, pp. 777–85, Mar 2004.

[256] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, no. 9, pp. 1667–72, Sep 2003.

[257] D. A. Barkauskas, S. R. Kronewitter, C. B. Lebrilla, and D. M. Rocke, "Analysis of maldi ft-icr mass spectrometry data: a time series approach," *Anal Chim Acta*, vol. 648, no. 2, pp. 207–14, Aug 2009.

[258] D. A. Barkauskas and D. M. Rocke, "A general-purpose baseline estimation algorithm for spectroscopic data," *Anal Chim Acta*, vol. 657, no. 2, pp. 191–7, Jan 2010.

[259] M. A. Grayson, *Measuring mass: from positive rays to proteins*. Philadelphia: Chemical Heritage Press, 2002.

[260] R. W. Kiser, *Introduction to mass spectrometry and its applications*, ser. Prentice-Hall international series in chemistry. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

[261] S. M. Carlson, A. Najmi, and H. J. Cohen, "Biomarker clustering to address correlations in proteomic data," *Proteomics*, vol. 7, no. 7, pp. 1037–46, Apr 2007.

[262] T. Skov, D. Ballabio, and R. Bro, "Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks," *Anal Chim Acta*, vol. 615, no. 1, pp. 18–29, May 2008.

[263] S. Chen, M. Li, D. Hong, D. Billheimer, H. Li, B. J. Xu, and Y. Shyr, "A novel comprehensive wave-form ms data processing method," *Bioinformatics*, vol. 25, no. 6, pp. 808–14, Mar 2009.

[264] J. E. Eckel-Passow, A. L. Oberg, T. M. Therneau, and H. R. Bergen, 3rd, "An insight into high-resolution mass-spectrometry data," *Biostatistics*, vol. 10, no. 3, pp. 481–500, Jul 2009.

[265] I. M. Taban, A. F. M. Altelaar, Y. E. M. van der Burgt, L. A. McDonnell, R. M. A. Heeren, J. Fuchser, and G. Baykut, "Imaging of peptides in the rat brain using maldi-fticr mass spectrometry," *J Am Soc Mass Spectrom*, vol. 18, no. 1, pp. 145–51, Jan 2007.

[266] L. F. Waanders, K. Chwalek, M. Monetti, C. Kumar, E. Lammert, and M. Mann, "Quantitative proteomic analysis of single pancreatic islets," *Proc Natl Acad Sci U S A*, vol. 106, no. 45, pp. 18 902–7, Nov 2009.

[267] M. S. Burns, "Applications of secondary ion mass spectrometry (sims) in biological research: a review," *J Microsc*, vol. 127, no. Pt 3, pp. 237–58, Sep 1982.

[268] O. Fiehn, "Metabolomics–the link between genotypes and phenotypes," *Plant Mol Biol*, vol. 48, no. 1-2, pp. 155–71, Jan 2002.

[269] J. Hirsch, K. C. Hansen, A. L. Burlingame, and M. A. Matthay, "Proteomics: current techniques and potential applications to lung disease," *Am J Physiol Lung Cell Mol Physiol*, vol. 287, no. 1, pp. L1–23, Jul 2004.

[270] R. L. Caldwell and R. M. Caprioli, "Tissue profiling by mass spectrometry: a review of methodology and applications," *Mol Cell Proteomics*, vol. 4, no. 4, pp. 394–401, Apr 2005.

[271] P. Chaurand, J. L. Norris, D. S. Cornett, J. A. Mobley, and R. M. Caprioli, "New developments in profiling and imaging of proteins from tissue sections by maldi mass spectrometry," *J Proteome Res*, vol. 5, no. 11, pp. 2889–900, Nov 2006.

[272] V. Espina, M. Heiby, M. Pierobon, and L. A. Liotta, "Laser capture microdissection technology," *Expert Rev Mol Diagn*, vol. 7, no. 5, pp. 647–57, Sep 2007.

[273] S. G. Boxer, M. L. Kraft, and P. K. Weber, "Advances in imaging secondary ion mass spectrometry for biological samples," *Annu Rev Biophys*, vol. 38, pp. 53–74, 2009.

[274] R. Kruse and J. V. Sweedler, "Spatial profiling invertebrate ganglia using maldi ms," *J Am Soc Mass Spectrom*, vol. 14, no. 7, pp. 752–9, Jul 2003.

[275] S. S. Rubakhin, W. T. Greenough, and J. V. Sweedler, "Spatial profiling with maldi ms: distribution of neuropeptides within single neurons," *Anal Chem*, vol. 75, no. 20, pp. 5374–80, Oct 2003.

[276] D. B. Thibault, C. J. Gillam, A. C. Grey, J. Han, and K. L. Schey, "Maldi tissue profiling of integral membrane proteins from ocular tissues," *J Am Soc Mass Spectrom*, vol. 19, no. 6, pp. 814–22, Jun 2008.

[277] 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, no. 7, pp. 3278–83, Jul 2009.

[278] A. R. Kherlopian, T. Song, Q. Duan, M. A. Neimark, M. J. Po, J. K. Gohagan, and A. F. Laine, "A review of imaging techniques for systems biology," *BMC Syst Biol*, vol. 2, p. 74, 2008.

[279] J. A. McLean, "The mass-mobility correlation redux: the conformational landscape of anhydrous biomolecules," *J Am Soc Mass Spectrom*, vol. 20, no. 10, pp. 1775–81, Oct 2009.

[280] S. M. Lin, L. Zhu, A. Q. Winter, M. Sasinowski, and W. A. Kibbe, "What is mzxml good for?" *Expert Rev Proteomics*, vol. 2, no. 6, pp. 839–45, Dec 2005.

[281] R. Stanislaus, L. H. Jiang, M. Swartz, J. Arthur, and J. S. Almeida, "An xml standard for the dissemination of annotated 2d gel electrophoresis data complemented with mass spectrometry results," *BMC Bioinformatics*, vol. 5, p. 9, Jan 2004.

[282] E. Deutsch, "mzml: a single, unifying data format for mass spectrometer output," *Proteomics*, vol. 8, no. 14, pp. 2776–7, Jul 2008.

[283] M. A. Brown and R. C. Semelka, "Mr imaging abbreviations, definitions, and descriptions: a review," *Radiology*, vol. 213, no. 3, pp. 647–62, Dec 1999.

[284] E. W. Deutsch, "Mass spectrometer output file format mzml," *Methods Mol Biol*, vol. 604, pp. 319–31, 2010.

[285] P. Price, "Standard definitions of terms relating to mass spectrometry: a report from the committee on measurements and standards of the american society for mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 2, no. 4, pp. 336–348, 1991.

[286] A. K. Smilde, R. Bro, and P. Geladi, *Multi-way analysis with applications in the chemical sciences*. Chichester, West Sussex, England: J. Wiley, 2004.

[287] R. M. Smith, *Understanding mass spectra: a basic approach*, 2nd ed. Hoboken, N.J.: Wiley Interscience, 2004.

[288] F. W. McLafferty and F. Turecek, *Interpretation of mass spectra*, 4th ed. Mill Valley, Calif.: University Science Books, 1993.

List of Publications

Articles in International Journals

Van de Plas R., Vander Mierde D., Lemaire K., De Moor B., In't Veld P., Schuit F., Waelkens E. "Hormone Processing Pathway Analysis using Mass Spectral Imaging and Data-mining of Co-localization". *Nature Medicine*. Submitted for review.

Luts J., Ojeda F., Van de Plas R., De Moor B., Van Huffel S., Suykens J.A.K. "A tutorial on support vector machine-based methods for classification problems in chemometrics". *Analytica Chimica Acta*, vol. 665, no. 2, Apr. 2010, pp. 129-145.

Lemaire K., Ravier M.A., Schraenen A., Creemers J.W.M., Van de Plas R., Granvik M., Van Lommel L., Waelkens E., Chimienti F., Rutter G.A., Gilon P., in't Veld P.A., Schuit F.C. "Insulin crystallisation depends on zinc transporter ZnT8 expression, but is not required for normal glucose homeostasis in mice". *Proceedings of the National Academy of Sciences*, vol. 106, no. 35, Sep. 2009, pp. 14872-14877.

Cadron I., Van Gorp T., Amant F., Vergote I., Moerman P., Waelkens E., Daemen A., Van de Plas R., De Moor B., Zeillinger R. "The use of laser microdissection and SELDI-TOF MS in ovarian cancer tissue to identify protein profiles". *Anticancer Research*, vol. 29, no. 4, Apr. 2009, pp. 1039-1046.

Kyama C.M., Mihalyi A., Gevaert O., Waelkens E., Simsa P., Van de Plas R., Meuleman C., De Moor B., D'Hooghe T.M. "Evaluation of endometrial biomarkers for semi-invasive diagnosis of endometriosis". *Fertility and Sterility*. Submitted for review.

Articles in the Proceedings of International Conferences

Van de Plas R., De Moor B., Waelkens E. "Discrete Wavelet Transform-based Multivariate Exploration of Tissue via Imaging Mass Spectrometry". In *Proceedings of the*

23rd Annual ACM Symposium on Applied Computing (ACM SAC), Fortaleza, Brazil, Mar. 2008.

Van de Plas R., Pelckmans K., De Moor B., Waelkens E. "Spatial Querying of Imaging Mass Spectrometry Data: A Nonnegative Least Squares Approach". In *The NIPS workshop on Machine Learning in Computational Biology (NIPS MLCB)*, Whistler (B.C.), Canada, Dec. 2007.

Kyama CM., Mihalyi A., Gevaert O., Simsa P., Waelkens E., Van de Plas R., Mwenda JM., Meuleman C., De Moor B., D'Hooghe TM. "Endometrial biomarkers for semi-invasive diagnosis of endometriosis". In *Proceedings of the 23rd Annual Meeting of ESHRE*, Lyon, France, Human Reproduction, 2007.

Van de Plas R., De Moor B., Waelkens E. "Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA". In *Proceedings of the Third IEEE/NIH BISTI Life Science Systems and Applications Workshop 2007 (LISSA 2007)*, Bethesda, Maryland USA, Nov. 2007.

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

Patent Application

Van de Plas R., De Moor B., Waelkens E. "Ionization Efficiency Correction, a novel normalization method for mass spectrometry and mass spectral imaging". Patent application #GB1005959.0, under examination.

Articles in Preparation

Van de Plas R., Konings P., De Moor B., Waelkens E. "Spatial Autocorrelation in Mass Spectral Imaging and Profiling: A Word of Caution Regarding Inferential Statistics". Manuscript in preparation.

Van de Plas R., De Moor B., Waelkens E. "Unmixing of Mass Spectral Imaging Data". Manuscript in preparation.

Van de Plas R., Pelckmans K., De Moor B., Waelkens E. "Spatial Querying of Imaging Mass Spectrometry Data for the Biochemical Characterization of Anatomical Regions in Tissue". Manuscript in preparation.

Van de Plas R., De Moor B., Waelkens E. "Discrete Wavelet Transform-based Multivariate Exploration of Tissue via Imaging Mass Spectrometry". Manuscript in preparation.

Van de Plas R., De Moor B., Waelkens E. "Enhanced Peak Detection for Mass Spectral Imaging". Manuscript in preparation.

Van de Plas R., De Moor B., Waelkens E. "Co-localization Networks: Assessing Inter-molecule Relationships in Tissue". Manuscript in preparation.

Oral Talks at National and International Conferences or Workshops

24 May 2010 - *The 58th Conference on Mass Spectrometry and Allied Topics (ASMS 2010)* - Salt Lake City, Utah, USA. "Going Beyond Images: Exploration of Hormone Processing Pathways in Diabetic Mouse Models via Mass Spectral Imaging and Datamining". Raf Van de Plas, Dirk Vander Mierde, Katleen Lemaire, Bart De Moor, Peter In't Veld, Frans Schuit, and Etienne Waelkens.

7 Apr 2010 - *Interdisciplinary Workshop on Sparsity and Modern Mathematical Methods for High Dimensional Data* - Brussels, Belgium. "Wavelet approaches to enable the exploration of organic tissue via multivariate analysis of mass spectral imaging data". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

14-15 Dec 2009 - *Benelux Bioinformatics Conference (BBC 2009)* - Liège, Belgium. "Anatomically Guided Differential Analysis of Neurodegenerative Alterations in Tissue via Mass Spectral Imaging and Spatial Querying". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Bart De Moor, and Etienne Waelkens.

18 Sep 2009 - *Flanders Proteomics Workshop on Data Mining and Bioinformatics for Proteomics (KVCV-P Comp Proteo 2009)* - Antwerp, Belgium. "Computational Analysis of Mass Spectral Imaging Data: Recent Advances". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

11 Sep 2009 - *Core Facility Seminars: ProMeta, K.U.Leuven's Interfaculty Centre for Proteomics and Metabolomics* - Leuven, Belgium. "Recent Advances in the Computational Analysis of Mass Spectral Imaging Data". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

3 Sep 2009 - *The 18th International Mass Spectrometry Conference (IMSC 2009)* - Bremen, Germany. "Spectroscopy Europe: Video discussion on Spatial Querying of Mass Spectral Imaging Data Allows for Differential Analysis of Neurodegenerative

Alterations in Tissue". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Justin Vijay Louis, Bart De Moor, and Etienne Waelkens.
http://www.spectroscopyeurope.com/blogs/imsc-2009/1606-discussion-with-raf-van-de-plas-katholieke-universiteit-leuven

3 Jun 2009 - *The 57th Conference on Mass Spectrometry and Allied Topics (ASMS 2009)* - Philadelphia, Pennsylvania, USA. "Spatial Querying of Mass Spectral Imaging Data Allows for Differential Analysis of Neurodegenerative Alterations in Tissue". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Justin Vijay Louis, Bart De Moor, and Etienne Waelkens.

28 Apr 2009 - *UHasselt Master in Biostatistics Course "Proteome Analysis"* - Leuven, Belgium. "High-throughput Proteomics and Metabolomics via Mass Spectrometry". Raf Van de Plas.

26 Mar 2009 - *NVMS/BSMS 2009 Annual Meeting* - Rolduc, Kerkrade, The Netherlands. "Spatial Querying: Differential Analysis of Neurodegenerative Alterations in Tissue via a Novel Spatial Querying Approach". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Bart De Moor, and Etienne Waelkens.

19 Mar 2009 - *AMOLF Mass Spectral Imaging Course* - Amsterdam, The Netherlands. "Co-localization Networks: Exploring Ion Co-localization in Tissue". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

29 Jan 2009 - *GOA Meeting* - Leuven, Belgium. "Spatial Querying: Spatial Driven Exploration of Imaging Mass Spectrometry Data". Raf Van de Plas, Justin Vijay Louis, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

12 Jan 2009 - *ESAT/BioStat Brainstorm* - Leuven, Belgium. "Spatial Querying: Spatial Driven Exploration of Imaging Mass Spectrometry Data". Raf Van de Plas, Justin Vijay Louis, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

23 May 2008 - *FNRS Meeting on Glycosylation and other Post-Translational Modifications* - Louvain-la-neuve, Belgium. "Biochemical Exploration of Tissue via Imaging Mass Spectrometry (with a post-translational twist)". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

15 Feb 2008 - *LKI Oncoforum V 2008* - Leuven, Belgium. "Exploring the Biochemical Composition of Tissue via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

8 Dec 2007 - *NIPS 2007 Workshop on Machine Learning in Computational Biology* - Whistler, British Columbia, Canada. "Spatial Querying of Imaging Mass Spectrometry Data: A Nonnegative Least Squares Approach". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

29 May 2007 - *The 2nd SymBioSys Workshop* - Leuven, Belgium. "Tutorial * Proteomic Analysis from Gel to Mass Spectrometry: Possibilities, Limitations, and Recent Developments * Focus on Mass Spectrometry". Raf Van de Plas.

2 May 2007 - *K.U.Leuven Master of Bioinformatics Course "Case Studies in Bioinformatics System Integration"* - Leuven, Belgium. "High-throughput Proteomics and Metabolomics via Mass Spectrometry". Raf Van de Plas.

16 Feb 2007 - *Annual Meeting of the Belgian Society for Mass Spectrometry (BSMS 2007)* - Leuven, Belgium. "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis". Raf Van de Plas, Fabian Ojeda, Maarten Dewil, Ludo Van Den Bosch, Bart De Moor, and Etienne Waelkens.

19 Oct 2006 - *Sista Seminar* - Leuven, Belgium. "Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry guided by Multivariate Data Analysis". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

29 Sep 2006 - *FWO Visitation Commission* - Leuven, Belgium. "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry guided by Principal Component Analysis". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

16 Jun 2006 - *Knowledge for Growth 2006* - Ghent, Belgium. "MALDI-based Imaging Mass Spectrometry applied to Rat Nerve Tissue". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

17 May 2006 - *K.U.Leuven Master of Bioinformatics Course "Case Studies in Bioinformatics System Integration"* - Leuven, Belgium. "ProMeta: High-throughput Proteomics and Metabolomics via Mass Spectrometry". Raf Van de Plas.

30 Jan 2006 - *IBM Integrator Demo* - Leuven, Belgium. "Introduction to ProMeta, The K.U.Leuven Interfacultary Centre for Proteomics and Metabolomics". Raf Van de Plas.

13 May 2005 - *ESAT-SCD and IBM Biomedical Meeting* - Leuven, Belgium. "Imaging Mass Spectrometry Applied to Amyotrophic Lateral Sclerosis". Raf Van de Plas, Maarten Dewil, Ludo Van Den Bosch, Bart De Moor, and Etienne Waelkens.

Abstracts at National and International Conferences or Workshops

24 May 2010 - *The 58th Conference on Mass Spectrometry and Allied Topics (ASMS 2010)* - Salt Lake City, Utah, USA. "Improved Wavelet Analysis of Mass Spectral Imaging Data for Feature Selection and Data Compression through Incorporation of Spatial Information". Nico Verbeeck, Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

29 Sep-2 Oct 2009 - *The 45th Annual Meeting of the European Association for the Study of Diabetes (EASD 2009)* - Vienna, Austria. "No insulin crystallisation in beta cells in the absence of ZnT8". K. Lemaire, M.A. Ravier, A. Schraenen, L. J.W.M. Creemers, R. Van de Plas, M. Granvik, E. Waelkens, F. Chimienti, G. Rutter, P. Gilon, P. in't Veld, and F. Schuit.

30 Aug-4 Sep 2009 - *The 18th International Mass Spectrometry Conference (IMSC 2009)* - Bremen, Germany. "Spatial Querying of Mass Spectral Imaging Data Allows for Differential Analysis of Neurodegenerative Alterations in Tissue". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Justin Vijay Louis, Bart De Moor, and Etienne Waelkens.

27 Jun-2 Jul 2009 - *The 17th Annual International Conference on Intelligent Systems for Molecular Biology and 8th European Conference on Computational Biology (ISMB/ECCB 2009)* - Stockholm, Sweden. "Differential Analysis of Neurodegenerative Alterations in Tissue via a Novel Spatial Querying Approach". Raf Van de Plas, Kristiaan Pelckmans, Thomas Philips, Bart De Moor, and Etienne Waelkens.

29 May 2009 - *Annual Meeting IUAP P6/28* - Gent, Belgium. "Tissue Proteomics: Spatial Driven Exploration of Imaging Spectrometry Data". Raf Van de Plas, Justin Vijay Louis, Veerle Janssens, Jozef Goris, Bart De Moor, and Etienne Waelkens.

18-19 Dec 2008 - *KVCV-P Flanders Proteomics Meeting* - Antwerpen, Belgium. "Spatial Querying of Imaging Mass Spectrometry Data for the Biochemical Characterization of Anatomical Regions in Tissue". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

15-16 Dec 2008 - *Benelux Bioinformatics Conference (BBC 2008)* - Maastricht, The Netherlands. "Spatial Querying of Imaging Mass Spectrometry Data". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

19-23 Jul 2008 - *The 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)* - Toronto, Canada. "Spatial Querying of Imaging

Mass Spectrometry Data". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

1-5 Jun 2008 - *The 56th Conference on Mass Spectrometry and Allied Topics (ASMS 2008)* - Denver, Colorado, USA. "Enhanced Interpretation of Imaging Mass Spectrometry Data Using Non-negative Matrix Factorization". Raf Van de Plas, Bart Vanluyten, Bart De Moor, and Etienne Waelkens.

1-5 Jun 2008 - *The 56th Conference on Mass Spectrometry and Allied Topics (ASMS 2008)* - Denver, Colorado, USA. "Enabling Multivariate Exploration of Large Imaging Mass Spectrometry Data Sets Using Discrete Wavelet Transform". Etienne Waelkens, Raf Van de Plas, and Bart De Moor.

1-5 Jun 2008 - *The 56th Conference on Mass Spectrometry and Allied Topics (ASMS 2008)* - Denver, Colorado, USA. "Peak Intensity Weighted PCA for the Multivariate Exploration of Tissue via Imaging Mass Spectrometry". Bart De Moor, Raf Van de Plas, and Etienne Waelkens.

28 May 2008 - *SymBioSys/BioMAGNet Event* - Leuven, Belgium. "Enabling Multivariate Exploration of Large Imaging Mass Spectrometry Data Sets Using Discrete Wavelet Transform". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

23 May 2008 - *FNRS Meeting on Glycosylation and other Post-Translational Modifications* - Louvain-la-neuve, Belgium. "Biochemical Exploration of Tissue via Imaging Mass Spectrometry (with a post-translational twist)". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

16-20 Mar 2008 - *The 23rd Annual ACM Symposium on Applied Computing (SAC 2008)* - Fortaleza, Brazil. "Discrete Wavelet Transform-based Multivariate Exploration of Tissue via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

15 Feb 2008 - *LKI Oncoforum V 2008* - Leuven, Belgium. "Exploring the Biochemical Composition of Tissue via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

15 Feb 2008 - *LKI Oncoforum V 2008* - Leuven, Belgium. "Proteomic analysis via Mass Spectrometry: In silico processing and model-building". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

15 Feb 2008 - *Annual Meeting of the Belgian Society for Mass Spectrometry (BSMS 2008)* - Beerse, Belgium. "Making Trend Detection in Large Imaging Mass Spectrometry Data Sets Feasible: A Wavelet Approach". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

7-8 Dec 2007 - *NIPS 2007 Workshop on Machine Learning in Computational Biology* - Whistler, Britisch Columbia, Canada. "Spatial Querying of Imaging Mass Spectrometry Data: A Nonnegative Least Squares Approach". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

12-13 Nov 2007 - *Benelux Bioinformatics Conference (BBC 2007)* - Leuven, Belgium. "Spatial Querying of Imaging Mass Spectrometry Data for the Biochemical Characterization of Anatomical Regions in Tissue". Raf Van de Plas, Kristiaan Pelckmans, Bart De Moor, and Etienne Waelkens.

8-9 Nov 2007 - *The 3rd IEEE/NIH BISTI 2007 Life Science Systems and Applications Workshop (LISSA 2007)* - Bethesda, Maryland, USA. "Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

28 Oct-1 Nov 2007 - *The 15th International Meeting of the European Society of Gyneacological Oncology (ESGO 2007)* - Berlin, Germany. "Serum biomarker discovery workflow with MALDI-TOF MS to distinguish between benign and malignant adnexal masses". Toon Van Gorp, Raf Van de Plas, Olivier Gevaert, Isabel Cadron, Etienne Waelkens, Frédéric Amant, Bart De Moor, Dirk Timmerman, and Ignace Vergote.

17-20 Sep 2007 - *Annual Joint Meeting of the French Societies for Mass Spectrometry, Electrophoresis and Proteomics (SMAP 2007)* - Pau, France. "Independent Component Analysis as a Method for the Detection of Biochemical Trends in Tissue via Imaging Mass Spectrometry". Jacques Pruvost, Raf Van de Plas, Helen Montgomery, Bart De Moor, and Etienne Waelkens.

21-25 Jul 2007 - *The 15th Annual International Conference on Intelligent Systems for Molecular Biology and 6th European Conference on Computational Biology (ISMB/ECCB 2007)* - Vienna, Austria. "Independent Component Analysis as a Guide for Prospective Biochemical Tissue Exploration via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

1-4 Jul 2007 - *The 23rd Annual Meeting of the European Society of Human Reproduction and Embryology (ESHRE 2007)* - Lyon, France. "Evaluation of endometrial biomarkers for semi-invasive diagnosis of endometriosis". Cleophas M. Kyama, Attila Mihalyi, Olivier Gevaert, Peter Simsa, Etienne Waelkens, Raf Van de Plas, Jason M. Mwenda, Christel Meuleman, Bart De Moor, and Thomas M. D'Hooghe.

3-7 Jun 2007 - *The 55th Conference on Mass Spectrometry and Allied Topics (ASMS 2007)* - Indianapolis, Indiana, USA. "Independent Component Analysis as a Method

for the Detection of Biochemical Trends in Tissue via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

18-22 Jan 2007 - *The 18th Sanibel Conference on Mass Spectrometry: Imaging Mass Spectrometry* - Sanibel Island, Florida, USA. "Adaptation of Principal Component Analysis towards Inherent Incorporation of Spatial Information for Trend Detection in Imaging Mass Spectrometry Data". Raf Van de Plas, Tijl De Bie, Bart De Moor, and Etienne Waelkens.

18-22 Jan 2007 - *The 18th Sanibel Conference on Mass Spectrometry: Imaging Mass Spectrometry* - Sanibel Island, Florida, USA. "Independent Component Analysis as a Guide for Prospective Biochemical Tissue Exploration via Imaging Mass Spectrometry". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

7-8 Dec 2006 - *Royal Flemish Chemical Society Symposium on Proteome Analysis in Systems Biology* - Antwerp, Belgium. "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry and Principal Component Analysis". Raf Van de Plas, Fabian Ojeda, Etienne Waelkens, and Bart De Moor.

17-18 Oct 2006 - *Benelux Bioinformatics Conference (BBC 2006)* - Wageningen, The Netherlands. "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry and Principal Component Analysis". Raf Van de Plas, Fabian Ojeda, Etienne Waelkens, and Bart De Moor.

16 Jun 2006 - *Knowledge for Growth 2006* - Ghent, Belgium. "MALDI-based Imaging Mass Spectrometry applied to rat nerve tissue using the ABI 4800 MALDI TOF/TOF Analyzer". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

28 May-1 Jun 2006 - *The 54th Conference on Mass Spectrometry and Allied Topics (ASMS 2006)* - Seattle, Washington, USA. "MALDI-based Imaging Mass Spectrometry applied to rat nerve tissue using the ABI 4800 MALDI TOF/TOF Analyzer". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

2-5 Apr 2006 - *The 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006)* - Venice, Italy. "Detecting Chemically Distinct Anatomical Regions in Rat Nerve Tissue Using Imaging Mass Spectrometry". Raf Van de Plas, Etienne Waelkens, and Bart De Moor.

26 Jan 2006 - *Visit of European Commissioner for Science and Research Dr. Janez Potocnik* - Leuven, Belgium. "In Silico Analysis for Imaging Mass Spectrometry in Proteomics". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

28 Sep-1 Oct 2005 - *The 4th European Conference on Computational Biology (ECCB 2005)* - Madrid, Spain. "In Silico Analysis for Imaging Mass Spectrometry in Proteomics". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

20 Apr 2005 - *K.U.Leuven PhD Symposium* - Leuven, Belgium. "In Silico Analysis for Imaging Mass Spectrometry in Proteomics". Raf Van de Plas, Bart De Moor, and Etienne Waelkens.

Curriculum Vitae

Raf VAN DE PLAS was born in 1980, in Aarschot, Belgium. He received his Master in Engineering in Electronics and Computer Science from the Group T Leuven Engineering College, Belgium in 2002. The thesis entailed the study, design, and realization of a communication unit for large-scale audiovisual equipment, using GSM connectivity for control and diagnostics. The thesis received a Barco/VIK Award and a KPMG Management Award in 2002. He received his Advanced Master of Science in Artificial Intelligence from the Katholieke Universiteit Leuven (University of Leuven), Belgium in 2003 with a thesis on the design and implementation of algorithms for the automated reconstruction of shattered objects, with an eye towards craniofacial applications. He is a founding member of K.U.Leuven's Interfaculty Centre for Proteomics and Metabolomics, ProMeta, and a member of the IEEE, the Belgian Society for Mass Spectrometry, and the American Society for Mass Spectrometry. In Januari 2005, he obtained a scholarship from the Agency for Innovation by Science and Technology in Flanders (IWT) to pursue his Ph.D. degree at the SCD division of the Department of Electrical Engineering of the K.U.Leuven. In April 2010, he was awarded an Honorary Fulbright Grant to continue his research as a postdoc in the United States. His doctoral research focuses on mass spectral imaging of organic tissue, and the computational analysis of the massive data sets generated by this technology.