DEVELOPMENT AND APPLICATION OF CHEMOMETRIC METHODS FOR MODELLING METABOLIC SPECTRAL PROFILES

Judith Marlou Fonville

A thesis submitted for the degree of Doctor of Philosophy of Imperial College London, 2011
Abstract

The interpretation of metabolic information is crucial to understanding the functioning of a biological system. Latent information about the metabolic state of a sample can be acquired using analytical chemistry methods, which generate spectroscopic profiles. Thus, nuclear magnetic resonance spectroscopy and mass spectrometry techniques can be employed to generate vast amounts of highly complex data on the metabolic content of biofluids and tissue, and this thesis discusses ways to process, analyse and interpret these data successfully.

The evaluation of $J$-resolved spectroscopy in magnetic resonance profiling and the statistical techniques required to extract maximum information from the projections of these spectra are studied. In particular, data processing is evaluated, and correlation and regression methods are investigated with respect to enhanced model interpretation and biomarker identification. Additionally, it is shown that non-linearities in metabonomic data can be effectively modelled with kernel-based orthogonal partial least squares, for which an automated optimisation of the kernel parameter with nested cross-validation is implemented. The interpretation of orthogonal variation and predictive ability enabled by this approach are demonstrated in regression and classification models for applications in toxicology and parasitology. Finally, the vast amount of data generated with mass spectrometry imaging is investigated in terms of data processing, and the benefits of applying multivariate techniques to these data are illustrated, especially in terms of interpretation and visualisation using colour-coding of images. The advantages of methods such as principal component analysis, self-organising maps and manifold learning over univariate analysis are highlighted.

This body of work therefore demonstrates new means of increasing the amount of biochemical information that can be obtained from a given set of samples in biological applications using spectral profiling. Various analytical and statistical methods are investigated and illustrated with applications drawn from diverse biomedical areas.
Declaration of originality

I would hereby like to mention that all the material presented in this thesis is the result of my own work and ideas, unless referenced specifically. Certain sections have been adapted from papers which I have written or been involved with during my research, and these papers are included as appendices B–F. The glucuronide nuclear magnetic resonance data were acquired by Dr. C. Johnson, the $J$-resolved nuclear magnetic resonance data were acquired in collaboration with Dr. A. Maher and T. Wright and the in vivo spectroscopic liver data were acquired by members of the Metabolic and Molecular Imaging group at the Hammersmith Hospital Campus, Imperial College, UK. The galactosamine and mercuric chloride toxicity data were courtesy of the COMET projects, and the *T. brucei brucei* data were kindly provided by Dr. Y. Wang. The mass spectrometry imaging and histology data were acquired by C. Carter, University of Birmingham, UK.
Acknowledgements

The work presented in this thesis would not have been possible without the support of my two supervisors Prof. Elaine Holmes and Prof. Jeremy Nicholson, and the guidance provided by Prof. John Lindon. In addition to their scientific input and generosity, they have taught me a lot about academic and independent research, publishing and strategy.

Additionally, I would like to explicitly mention Dr. Mattias Rantalainen, Dr. Muireann Coen and Dr. Anthony Maher, as their enthusiasm and energy is infectious, and the time they have taken to teach and supervise me is greatly appreciated. I would like to thank Dr. Tim Ebbels, Dr. Liz Want and Dr. Hector Keun for their additional support. I am also indebted to Dr. Jake Pearce, Dr. Olivier Cloarec, Dr. Selena Richards and Dr. Olaf Beckonert, as their knowledge and helpfulness has made life a lot easier than it could have been.

I would like to acknowledge the funding from the Analytical Chemistry Trust Fund of the Royal Society of Chemistry. Additionally, the willingness of Claire Carter and Dr. Josephine Bunch from the University of Birmingham to share their data and knowledge has helped me greatly in the conception of my last research chapter.

Finally I would like to thank my fellow PhD students and the other members of the Department of Biomolecular Medicine for their (im)moral support and friendship throughout my PhD, in particular Silke, Alison, Caroline J., Claire, Volker, Steve, Caroline S., Nkazi and Andreas.

Most of all, I would like to thank my friends and family, and especially my parents, brothers and my boyfriend Joe for being there for me and making sure I take a break!
Contents

List of Tables ......................................................... 9
List of Figures ....................................................... 12
Abbreviations and notations ......................................... 13

1 Introduction ....................................................... 14
1.1 Systems biology .................................................. 14
1.2 Metabonomics .................................................... 15
1.3 Metabonomic data analysis ...................................... 17
1.4 Development and application of chemometric methods for modelling metabolic spectral profiles .................................................. 20

2 Theory and methods ............................................... 24
2.1 NMR: nuclear magnetic resonance ................................ 24
2.1.1 Spins and magnetic field ...................................... 24
2.1.2 Detection of the NMR spectrum ................................ 25
2.1.3 Chemical shift ................................................... 26
2.1.4 Relaxation ....................................................... 27
2.1.5 J-coupling ....................................................... 28
2.1.6 ‘1D’ and Carr–Purcell–Meiboom–Gill spectra ............... 29
2.1.7 Two-dimensional NMR methods ............................ 30
2.2 MS: mass spectrometry .......................................... 31
2.2.1 Basics of mass spectrometry .................................. 31
2.2.2 MSI: mass spectrometry imaging ............................ 32
2.3 Multivariate modelling ........................................... 35
2.3.1 Data processing ............................................... 35
2.3.2 Univariate versus multivariate approaches .................. 36
2.3.3 PCA: principal component analysis ........................ 37
2.3.4 PLS: partial least squares .................................... 39
### 2.3.5 OPLS: orthogonal partial least squares ................................. 40
### 2.3.6 Model validation ..................................................................... 43
### 2.3.7 STOCSY: statistical total correlation spectroscopy .................. 45
### 2.3.8 SOM: self-organising map ..................................................... 46

#### 3 The use of 1D projections of J-resolved NMR spectra in metabonomics 50

3.1 Aims and objectives ................................................................. 50
3.2 Introduction .............................................................................. 51
3.3 Materials and methods ............................................................. 53
  3.3.1 Sample preparation ............................................................... 53
  3.3.2 \(^1\)H NMR spectroscopy ...................................................... 54
  3.3.3 Data processing ................................................................. 54
  3.3.4 Signal-to-noise ratio and line width analyses ......................... 55
  3.3.5 Statistical and chemometric analyses .................................... 55
  3.3.6 STORSY: statistical total regression spectroscopy .................. 56
3.4 Results and discussion ............................................................... 57
  3.4.1 Projection of JRES spectra ................................................. 57
  3.4.2 Symmetrisation of JRES spectra .......................................... 59
  3.4.3 Peak alignment of full-resolution JRES projections ................. 60
  3.4.4 Necessity of alignment of JRES projections for statistical analysis .. 61
  3.4.5 Improved STOCSY results by use of aligned JRES projections ... 65
  3.4.6 STORSY analysis of JRES projections ..................................... 66
  3.4.7 Over-representation of metabolite peaks in NMR spectroscopy .... 69
  3.4.8 Complementarity of 1D and CPMG spectra and JRES projections . 73
  3.4.9 An example of information recovery using JRES spectroscopy .... 74
3.5 Conclusions ............................................................................... 76

#### 4 Simulated annealing optimised K-OPLS in metabonomics 77

4.1 Aims and objectives ................................................................. 77
4.2 Introduction .............................................................................. 78
4.3 Methods .................................................................................... 79
  4.3.1 Kernel-OPLS ..................................................................... 79
  4.3.2 Kernel function optimisation ................................................. 79
  4.3.3 Simulated annealing ............................................................. 80
  4.3.4 Nested cross-validation ....................................................... 83
  4.3.5 Number of Y-orthogonal components .................................... 84
  4.3.6 Software implementation ..................................................... 87
  4.3.7 Visualisation of K-OPLS variable importance .......................... 87
4.4 Data sets ................................................................. 88
  4.4.1 Galactosamine toxicity ........................................... 88
  4.4.2 Mercuric chloride toxicity ...................................... 88
  4.4.3 *Trypanosoma brucei brucei* parasitic infection .......... 88
  4.4.4 Data processing .................................................. 89
4.5 Results and discussion .................................................. 89
  4.5.1 Prediction performance and optimisation of the kernel parameter .... 89
  4.5.2 Prediction of non-responders in the galactosamine toxicity study .... 92
  4.5.3 Time-course modelling of mercuric chloride toxicity ........... 95
  4.5.4 Classification of *T. brucei brucei* parasitic infection ....... 98
  4.5.5 Prediction of *T. brucei brucei* parasitic infection time-course .... 99
  4.5.6 Visualisation of variable importance in K-OPLS models ....... 100
4.6 Conclusions ............................................................ 106

5 Multivariate analysis of mass spectrometry imaging data ............... 107
  5.1 Aims and objectives .................................................. 107
  5.2 Introduction ......................................................... 108
  5.3 Materials and methods ............................................... 109
    5.3.1 Materials .......................................................... 109
    5.3.2 Tissue preparation for MALDI MSI of phospholipids ........... 109
    5.3.3 Mass spectrometry imaging ...................................... 110
    5.3.4 Histology ........................................................ 110
  5.4 Processing of MSI data ............................................... 111
    5.4.1 Binning, alignment and baseline correction .................... 111
    5.4.2 Unfolding of the data: from 3D to 2D .......................... 112
    5.4.3 Peak selection .................................................. 112
    5.4.4 Pixel selection .................................................. 116
    5.4.5 Normalisation ................................................... 117
    5.4.6 Data transformation ............................................. 121
    5.4.7 Smoothing ....................................................... 121
    5.4.8 Centring and scaling ............................................ 123
    5.4.9 Overview and generalisation of this approach ................. 124
  5.5 Multivariate analysis of MSI data .................................. 126
    5.5.1 Correlation analyses of MSI data ............................... 126
    5.5.2 Principal component analysis of MSI data ...................... 129
    5.5.3 Analysis of MSI data with self-organising maps ............... 135
    5.5.4 Evaluation of MSI data with manifold learning ............... 144
    5.5.5 Comparison of histology and multivariate modelling of MSI data .... 149
5.5.6 MSI versus magnetic resonance spectroscopy imaging ........................................ 150
5.6 Conclusions ............................................................................................................. 152

6 Discussion and future work ....................................................................................... 153
6.1 Discussion ................................................................................................................. 153
6.1.1 Analytical methodologies ................................................................................... 153
6.1.2 Chemometric methodologies ............................................................................. 154
6.1.3 Potential application areas .................................................................................. 155
6.1.4 Biological interpretation ....................................................................................... 156
6.1.5 Extensions to real-time use .................................................................................. 157
6.2 Future work ............................................................................................................... 159
6.3 Conclusions ................................................................................................................. 163

Bibliography ..................................................................................................................... 164

Appendix A: Analysis of MSI data from fresh tissue ..................................................... 181
Appendix B: Paper I ......................................................................................................... 185
Appendix C: Paper II ........................................................................................................ 197
Appendix D: Paper III ...................................................................................................... 212
Appendix E: Paper IV ...................................................................................................... 222
Appendix F: Paper V ........................................................................................................ 234
List of Tables

2.1 Biomolecular NMR nuclei ............................................. 24
3.1 Signal-to-noise ratios of JRES projections .......................... 59
3.2 Linewidths of JRES projections ...................................... 59
3.3 Principal component analysis and peak alignment of JRES projections ........... 63
3.4 OPLS-DA results on aligned and unaligned JRES projections ............... 64
3.5 STORSY on the glucuronide data set .................................. 67
4.1 Variables in the simulated annealing algorithm .......................... 83
4.2 Details of the evaluated data sets ................................... 90
5.1 MSI data set sizes during processing .................................. 123
5.2 List of m/z values selected in figure 5.16 ............................ 134
### List of Figures

1.1 Central dogma of molecular biology and metabolic pathways .................. 15  
1.2 Factors influencing metabolism .................................................. 16  
1.3 The necessity of multivariate statistics ....................................... 18  
1.4 The contribution of this work to the metabonomic workflow ................. 22  
2.1 Basics of NMR ................................................................. 26  
2.2 Chemical shift ................................................................. 27  
2.3 Complex 1D $^1$H NMR biofluid spectrum .................................... 28  
2.4 $J$-coupling splitting patterns ............................................... 29  
2.5 Spectral editing with CPMG .................................................. 30  
2.6 Example mass spectrum ....................................................... 32  
2.7 Mass spectrometry imaging experiment and data analysis .................... 33  
2.8 Scheme of chemometric methods ............................................. 37  
2.9 Explanation of PCA .......................................................... 38  
2.10 Example of PCA of spectral data ............................................ 39  
2.11 Results from an exemplar PLS model ...................................... 42  
2.12 Multicollinearity in spectra .................................................. 45  
2.13 Peak assignment with STOCSY ............................................ 46  
2.14 Exemplar interpretation of SOM ............................................ 47  
3.1 Peak overlap and over-representation ........................................ 51  
3.2 Effect of bucketing ........................................................... 52  
3.3 Exemplar 2D JRES spectrum ............................................... 58  
3.4 Alignment of raw NMR spectral data ....................................... 60  
3.5 PCA before and after peak alignment of JRES projections .................... 62  
3.6 Effect of alignment on STOCSY results .................................... 65  
3.7 Chemical structure of 1–β–O–acyl glucuronide ................................ 66  
3.8 STORSY results of lactate .................................................... 68  
3.9 Correlation between spectra of different experiments and biofluids ........ 70
3.10 Over-representation of glucose ........................................... 72
3.11 Biomarker discovery with STOCSY of JRES projections ............. 74
3.12 Comparison of 1D, CPMG and JRES for biomarker identification .... 75

4.1 Optimisation with simulated annealing .................................. 81
4.2 Nested cross-validation ..................................................... 84
4.3 Convergence plots ......................................................... 85
4.4 Example of SA-K-OPLS using the glucuronide data set ................. 86
4.5 Prediction performances .................................................... 91
4.6 SA-K-OPLS score plot of the galactosamine toxicity study ........... 92
4.7 Galactosamine classification receiver operating characteristic curve .. 93
4.8 KPLS and OPLS score plots of the galactosamine toxicity study ...... 94
4.9 External validation results .................................................. 95
4.10 SA-K-OPLS score plot of the mercuric chloride toxicity study ......... 96
4.11 Clinical chemistry of mercuric chloride toxicity study ................ 96
4.12 KPLS and OPLS score plots of the mercuric chloride toxicity study .. 97
4.13 TBB classification receiver operating characteristic curve ........... 99
4.14 Score plots displaying SA-K-OPLS models of T. brucei brucei infection 99
4.15 KPLS and OPLS score plots of the TBB data ........................ 100
4.16 PCA modelling of the kernel matrix in the TBB classification ........ 102
4.17 Variables important for the classification of TBB infection ........ 104

5.1 Histology image of the sample ........................................... 110
5.2 Peak selection process ..................................................... 113
5.3 Peak selection approach A, based on matrix peaks ..................... 114
5.4 Peak selection approach B, based on image structure ................ 115
5.5 Pixel selection based on TIC ............................................. 117
5.6 Normalisation of MSI data ............................................... 119
5.7 Differential co-crystallisation of analytes with the MALDI matrix ...... 120
5.8 Smoothing of the images .................................................. 122
5.9 Flowchart of MSI data processing ....................................... 124
5.10 Correlation enhancement for improved contrast between pixels ........ 127
5.11 Correlation enhancement histograms ................................... 128
5.12 Principal component analysis of MSI data ................................ 130
5.13 Loadings of the PCA model ............................................. 131
5.14 RGB encoding of PCA scores .......................................... 131
5.15 Images of m/z = 810.6 and 760.4 .................................... 132
5.16 Interpretation of the PCA score space .................................. 133
# Abbreviations and notations

These tables list the common abbreviations and notations used throughout this thesis.

## Abbreviations:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>αCHCA</td>
<td>Alpha-cyano-4-hydroxycinnamic acid</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the receiver operating characteristic curve</td>
</tr>
<tr>
<td>BMI</td>
<td>Body-mass-index</td>
</tr>
<tr>
<td>CPB</td>
<td>Carr–Purcell–Meiboom–Gill</td>
</tr>
<tr>
<td>CV</td>
<td>Cross-validation</td>
</tr>
<tr>
<td>FID</td>
<td>Free induction decay</td>
</tr>
<tr>
<td>GaIN</td>
<td>Galactosamine</td>
</tr>
<tr>
<td>ISOMAP</td>
<td>Isometric feature mapping</td>
</tr>
<tr>
<td>JRES</td>
<td>J-resolved NMR pulse-sequence</td>
</tr>
<tr>
<td>K-OPLS</td>
<td>Kernel-based OPLS</td>
</tr>
<tr>
<td>KPLS</td>
<td>Kernel-based PLS</td>
</tr>
<tr>
<td>MALDI</td>
<td>Matrix-assisted laser desorption/ionisation</td>
</tr>
<tr>
<td>MCCV</td>
<td>Monte Carlo cross-validation</td>
</tr>
<tr>
<td>MCCI</td>
<td>Magnetic resonance spectroscopy imaging</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MSI</td>
<td>Mass spectrometry imaging</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass-to-charge ratio</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>OC</td>
<td>Orthogonal component</td>
</tr>
<tr>
<td>OPLS</td>
<td>Orthogonal partial least squares</td>
</tr>
<tr>
<td>OPLS-DA</td>
<td>OPLS discriminant analysis</td>
</tr>
<tr>
<td>auc</td>
<td>Area under the receiver operating characteristic curve</td>
</tr>
<tr>
<td>bmi</td>
<td>Body-mass-index</td>
</tr>
<tr>
<td>cpmg</td>
<td>Carr–Purcell–Meiboom–Gill</td>
</tr>
<tr>
<td>cv</td>
<td>Cross-validation</td>
</tr>
<tr>
<td>fid</td>
<td>Free induction decay</td>
</tr>
<tr>
<td>galN</td>
<td>Galactosamine</td>
</tr>
<tr>
<td>isomap</td>
<td>Isometric feature mapping</td>
</tr>
<tr>
<td>jres</td>
<td>J-resolved NMR pulse-sequence</td>
</tr>
<tr>
<td>k-opls</td>
<td>Kernel-based OPLS</td>
</tr>
<tr>
<td>kpls</td>
<td>Kernel-based PLS</td>
</tr>
<tr>
<td>maldi</td>
<td>Matrix-assisted laser desorption/ionisation</td>
</tr>
<tr>
<td>mccv</td>
<td>Monte Carlo cross-validation</td>
</tr>
<tr>
<td>mrsi</td>
<td>Magnetic resonance spectroscopy imaging</td>
</tr>
<tr>
<td>m/2</td>
<td>Mass-to-charge ratio</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass-to-charge ratio</td>
</tr>
<tr>
<td>nmr</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>oc</td>
<td>Orthogonal component</td>
</tr>
<tr>
<td>opls</td>
<td>Orthogonal partial least squares</td>
</tr>
<tr>
<td>opls-da</td>
<td>OPLS discriminant analysis</td>
</tr>
</tbody>
</table>

## Notations:

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Number of components</td>
</tr>
<tr>
<td>δ or ppm</td>
<td>Chemical shift</td>
</tr>
<tr>
<td>E</td>
<td>X-residuals [n × k]</td>
</tr>
<tr>
<td>F</td>
<td>Y-residuals [n × m]</td>
</tr>
<tr>
<td>k</td>
<td>Number of descriptor variables</td>
</tr>
<tr>
<td>m</td>
<td>Number of response variables</td>
</tr>
<tr>
<td>n</td>
<td>Number of observations</td>
</tr>
<tr>
<td>P</td>
<td>Model loadings [k × a]</td>
</tr>
<tr>
<td>q²</td>
<td>Goodness of prediction</td>
</tr>
<tr>
<td>r</td>
<td>Pearson’s correlation coefficient</td>
</tr>
<tr>
<td>T</td>
<td>Model scores [n × a]</td>
</tr>
<tr>
<td>X</td>
<td>Descriptor data [n × k]</td>
</tr>
<tr>
<td>Y</td>
<td>Response data [n × m]</td>
</tr>
</tbody>
</table>

---

13
Chapter 1

Introduction

1.1 Systems biology

Understanding the human body, health and disease, has classically been performed with the traditional molecular biology ‘one at a time’ reductionist approach. In contrast, the modern top-down -omics approaches representing the interdisciplinary field of systems biology are capable of giving a holistic overview of the interaction of any multi-cellular superorganism with the environment, through integration of information acquired at different levels of biomolecular organisation. This is achieved by use of high-throughput techniques that allow the quantification of the molecular constituents of cells, biofluids, organs or even whole organisms. Examples include genomics, transcriptomics, proteomics and metabolomics. Genomics, where various generations of DNA sequencing techniques enable the sequencing of the genetic make-up, provides a large amount of information, however genomics does not create a complete picture of biology. For example, epigenetic structural changes eventually influence which regions of the DNA are actually being transcribed and not all genes and their functions are known. Transcriptomics attempts to evaluate genetic information in action, as it quantifies messenger RNA transcripts, which is related to the DNA that is actually being expressed at a time point. The next downstream step in the traditional molecular biology paradigm, shown in figure 1.1 A, is proteomics, which quantifies protein expression, usually with gel-electrophoresis separation followed by mass spectrometry identification. This is a very interesting branch of systems biology, although inconsistent method performance and a limited number of comprehensive libraries for protein identification have hampered the implementation of reliable large-scale proteomic approaches. A further way to directly investigate the overall physiological status of an organism is by evaluating the concentrations of the metabolites present, as these are indicative of biochemical reactions in cells and can also act as response mechanisms and crucial effectors of biological systems, see figure 1.1 B.
1.2 Metabonomics

Metabolites are the low molecular weight substrates, intermediates and products of metabolism, the chemical reactions occurring within a living organism that are necessary for the maintenance of life. Examples of metabolites include amino acids, nucleotides, antioxidants, organic acids, polyols, vitamins and alcohols. A large amount of information on metabolic pathways and function is known and has been captured in databases such as KEGG, MetaCyc and ChEBI, see figure 1.1 B.

The study of the complete metabolic content of an organism and how this is changed by some stimulus or external perturbation is known as metabonomics, which is defined as ‘the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification’. The related field of metabolomics has been defined in plant biology and microbiology as ‘the systematic study of the metabolic complement of the cell’. The analytical and measurement procedures for both are the same, and thus the two terms are often used interchangeably. The evaluation of a large set of metabolic descriptors can also be referred to more generally as metabolic profiling.

Metabolic profiling is a fast and cost-effective method to obtain a biological snapshot of an organism. Metabolic profiles are of great aid in the evaluation of both genetic and environmental factors. For example, this approach has been successfully applied in the diagnosis of inborn errors.
of metabolism and temporary or chronic exposure to stress factors also causes changes in concentrations of certain metabolites. Other examples of metabolic profiling studies include the modelling of drug use by the observation of exogenous metabolites, that are not normally present in healthy organisms, and toxicity, by studying altered levels of endogenous metabolites. In general, any change in metabolic concentration is typically caused by some biological underpinning, which may or may not be easy to uncover, depending on the severity and complexity of the challenge and the system.

Metabolic profiling is a versatile tool for investigating biological functions at the cell and systems levels, and thus inter alia for improving our understanding of biology, for studying disease and other pathological processes and for innovating pharmacological treatments. Examples of metabonomic applications include the study of cancer, parasitic infection, drug toxicity and personalised medicine. Metabolism is also influenced by such factors as genetics, gut microbial metabolism, and the consumption of food and drugs, as illustrated in figure 1.2. The interplay of the different exogenous factors and upstream effects, e.g. transcriptomic and proteomic changes, is reflected in the measured metabolite levels downstream. This is illustrated by a metabolome-wide association study, where the metabolite profiles of four population groups were assessed: whereas genomics would find that, for example, the ancestry of a person is the (main) determinant for genomic status, metabonomics is influenced by diet, lifestyle, environment and self-medication. The study showed that metabolite profiles were closely related to risk factors of interest, thus emphasising the value of metabonomics in systems biology.

Figure 1.2: Some of the various endogenous and exogenous factors influencing metabolism and the observed metabolic profile.
Metabolic profiling can be performed on readily available biofluids, such as urine and plasma\textsuperscript{39, 40} or other biofluids such as cerebrospinal fluid,\textsuperscript{41, 42} sperm\textsuperscript{43, 44} and saliva.\textsuperscript{45} More invasive methods can be used to obtain tissue samples, such as biopsies or surgical specimens, appropriate measurements on which can provide compartmentalisation information. Different information is obtained from the various biofluids: urine, for example, gives a snapshot profile of the excretion products,\textsuperscript{46} whereas plasma is typically less variable since it is under homoeostatic control and is thus expresses subtle concentration variations.

Experimental designs for metabonomic studies with animals are often based on different classes, e.g. healthy versus diseased. Various exposure levels can be studied, such as different doses of toxin, and temporal studies can be performed, where data are acquired for a series of time points to give a more dynamic model of the system under study. In human populations, observational studies are typically used, where samples are acquired from a pre-selected cohort. For such studies, a large number of participants is often necessary, because there are many factors besides the study objective that can influence the metabolic profile, as is illustrated in figure 1.2.

The metabolic content of the collected biofluid or tissue sample is then measured with a high throughput analytical technique such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) spectroscopy.\textsuperscript{47} Spectral analysis allows the untargeted (partial) resolution of the components in the complex mixture of metabolites. The advantage of NMR and MS over other analytical techniques is that both methods are capable of providing structural and compositional information, which is of paramount importance for unambiguous metabolite identification. NMR, discussed in §2.1, is a non-destructive method requiring minimal sample preparation. Although the sensitivity of NMR is lower than MS, a great number of diverse metabolite signals can be observed, especially at high magnetic field strengths.\textsuperscript{48, 49} The study of biofluids with MS, discussed in §2.2, is commonly preceded by a chromatography-based separation step, providing a second measurement dimension to reduce signal overlap and aid peak identification.\textsuperscript{50} The sensitivity of MS is higher than NMR and a smaller sample volume is required, but the analytical reproducibility of MS spectra is lower than for NMR spectra.\textsuperscript{47} A recently introduced approach to performing mass spectrometry, which is typically performed on liquids, is based on imaging of tissue samples, where mass spectra are acquired at different positions of the sample.\textsuperscript{51} Mass spectrometry imaging (MSI) generates a large amount of high-dimensional data and has already been used in protein and lipid profiling studies. It shows great promise for metabolic profiling, as spectral information as well as spatial information is obtained, allowing the localisation of different compounds across tissue sections to be studied.

1.3 Metabonomic data analysis

The spectral data that are acquired with NMR and MS can be very complex, as illustrated in figure 1.3 A. In this example, the different dose levels and time points before (0 hr) and after (> 0 hr) administration of mercuric chloride demonstrate that the variation related to time and dose level
needs to be accounted for before the interpretation of the main response to the toxic challenge, including inter-individual differences. An example of mass spectrometry imaging data is shown in figure 1.3 B, where the mass spectra of a few selected pixels are shown. Analysis of a full mass spectrometry imaging data set needs to address both the spatial and spectroscopic aspects of the data.

While visual analysis of spectral data is biased towards existing biological knowledge and peaks with high intensity, pattern recognition provides a more thorough and reliable analysis. Thus, computational multivariate data analysis can be performed to discriminate and evaluate different

Figure 1.3: (A) The raw $^1$H NMR spectral data from the mercuric chloride toxicity study, that will be discussed more thoroughly in chapter 4, are superimposed. It is clear that a lot of inherent variation needs to be accounted for, related to dose level as well as time-course. These plots demonstrate that a visual analysis would be cumbersome and unlikely to reveal all underlying information. (B) Mass spectrometry imaging data, where mass spectra are acquired at 10000’s of spatial positions, create a large amount of data (only 15 spectra are shown here for clarity). The analysis of these data, which will be discussed in chapter 5, must take into account both the spatial and spectral nature of the data.
classes of samples, or to perform regression, e.g. modelling of time-courses, based on multivariate spectral inputs. Modelling strategies should account for various effects influencing the data, including:

- the complexity of the spectra, e.g. the large number of measured metabolites;
- the experimental design of the study, for example the different classes, dose levels and time points; and the interactions of these different factors;
- the possibility of confounding factors known as orthogonal variation, such as batch effects, analytical instrument drift and gender and age differences.

The complexity of metabolic networks, as shown in figure 1.1 B, indicates that the response to a perturbation of homoeostasis is unlikely to incur a change in only one variable; in contrast, a collection of up- and downregulated metabolites is expected. Additionally, a metabolite is typically represented by a number of variables. Thus, there is a need to use methods of multivariate nature which are able to deal with complex, collinear spectral data. Moreover, the selected algorithm should be relatively insensitive to noise and able to cope with more variables (spectral data points) than observations (samples). The analysis of spectral information from the metabolic species present in a number of samples can be successfully performed using multivariate statistics, also known as pattern recognition and chemometrics; these terms are used interchangeably in this thesis.

Two widely used multivariate approaches are principal component analysis (PCA) and partial least squares (PLS, also known as projections to latent structures); these methods can be used to give an overview of the data and perform predictive modelling, respectively. Both methods project the data onto a low-dimensional space based on the variance structure. In addition to these linear modelling approaches, non-linear and stochastic models have been found useful for metabolic analysis. For example, a non-linear descendant of PLS known as K-OPLS allows for predictive non-linear modelling and shows great promise for metabolic studies. Additionally, data analysis with self-organising maps, also known as Kohonen maps, can aid the visualisation of complex non-linear data and will be used to study MSI data.

A plethora of other statistical methods, including both uni- and multivariate approaches, can be used in metabonomic studies but are not discussed in this thesis. These include ANOVA-based and multi-way methods that allow the inclusion of a third data set axis, e.g. different bodily compartments. Other types of non-linear and stochastic approaches used in metabolic profiling include neural networks, genetic algorithms and random forests. In addition, various other projection-based techniques exist, such as independent component analysis, ridge regression, principal components regression etc. that can be used for data modelling, but which are not commonly used in metabonomics.
A main goal of metabonomics is biomarker identification, and to this end, the variables responsible for e.g. clustering of different classes need to be available and transparent. After a selection of ‘interesting’ peaks is made, a first step towards compound identification is with statistical, correlation-based approaches, as discussed in §2.3.7. Further validation consists of multi-dimensional NMR experiments that allow the study of structural connectivity of a compound, or MS approaches that accurately determine the mass and fragments of a pre-selected peak. Finally, the pure compound can be added (‘spiked’) into a sample of interest to verify that the suggested peaks indeed show the corresponding increased intensity.

Some researchers choose to identify and quantify the metabolites prior to statistical analysis. This approach, however, does not allow for biomarker discovery, as unknowns are not tabulated. Moreover, with such an approach substantial effort might be focussed on the identification of peaks that are of minor interest, as many of the peaks might not directly be involved in the biological process, and therefore might not be of great importance in the subsequent model. In this thesis, the spectroscopic profiling data are used rather than the quantified metabolites.

1.4 Development and application of chemometric methods for modelling metabolic spectral profiles

The work presented in this thesis aims to extract a maximum amount of information from a set of biological samples, using spectroscopic techniques. In particular, the focus is on combining analytical spectroscopy and statistical multivariate modelling of the resulting data sets. Three results chapters are presented to show how a combination of appropriate techniques can allow metabolic profiling to be performed successfully on different experimental and analytical systems. A key factor that should be considered is that different data sets, acquired with different analytical methods or experimental designs, should be addressed with processing and modelling algorithms best suited to the particular study.

The first example, as seen in chapter 3, discusses the necessity of acquisition of a simplified version of the complex spectroscopic fingerprints typically generated in NMR-based metabolic profiling studies. NMR can be used to observe a wide range of relatively high-concentration hydrogen-containing molecules, but this very strength of NMR is simultaneously a weakness: the large number of chemical species that are NMR-visible, often with multiple peaks for each compound, causes the spectrum to appear relatively crowded, especially for low-field NMR spectrometers. Thus, the information latent in the samples can be obscured, as smaller peaks can go unnoticed if they overlap with other peaks. This crowding can be overcome elegantly through the usage of multi-dimensional spectroscopy, whereby the signals are dispersed on a second dimension. One particularly fast 2D NMR experiment, J-resolved spectroscopy (JRES), has been suggested to overcome this overlap problem and the associated over-representation of certain abundant molecules. The full-resolution
projections of these JRES spectra, which effectively show proton-decoupled $^1$H NMR spectra, were evaluated in this thesis in terms of the various processing steps, such as projection method and peak alignment, and the quantitative properties were assessed. Most importantly, the use of full-resolution JRES projections in biomarker discovery with multivariate modelling and metabolite annotation based on statistical approaches was demonstrated. Therefore, it is suggested that future studies involving complex mixtures should consider the use of JRES NMR spectra in their protocols, especially if complex multiplet structures, such as those arising from carbohydrate compounds, are likely to occur and dominate the spectra, obscuring smaller signals.

The second example of development and application of chemometric methods to model metabonomics, discussed in chapter 4, shows that certain data sets can display non-linear characteristics, and multivariate modelling provides better predictive ability and more robust and meaningful models if this non-linearity is taken into account. Kernel-based orthogonal partial least squares (K-OPLS) enables the non-linear modelling of data sets and separates the orthogonal variation from the predictive variation. Good predictive ability is possible using a simulated annealing approach to find the optimal kernel parameter value, which is essential for kernel-based modelling. The simulated annealing-based K-OPLS (SA-K-OPLS) was implemented in the existing software using a nested cross-validation to report reliable predictive performance. The use of SA-K-OPLS in metabonomics studies is demonstrated with regression and classification problems in toxicology and parasitology based on $^1$H NMR urinary and plasma data sets. Superior prediction compared to the linear equivalent OPLS, and meaningful interpretation of the orthogonal components are reported. Methods for obtaining information on the variables involved in the non-linear SA-K-OPLS modelling, which is normally not available due to the absence of usable loadings in kernel-based models, are introduced. The results from this study show the potential of non-linear modelling in time-course studies and in personalised medicine settings. The freely available SA-K-OPLS software should find use in various high-dimensional non-linear prediction problems, including the modelling of metabolic spectral data.

The final example, presented in chapter 5, entails the analysis of mass spectrometry imaging (MSI) data. The processing and analysis of these highly complex spectral imaging data is still in its infancy. An intuitive and robust approach to processing MSI data is presented. The suggested methods use the presence of matrix peaks in addition to informative signals to perform the necessary but non-trivial peak and pixel selection steps. After a clean and reasonably-sized data set was obtained, multivariate modelling was performed. The use of correlation analysis, principal component analysis and self-organising maps in the visualisation and interpretation of processed MSI data is illustrated, with a strong emphasis on both the multivariate as well as the imaging nature of the data. Additionally, the visualisation with manifold learning techniques of MSI data is demonstrated: the clustering of image pixels in multi-dimensional space corresponds to their spatial distribution and known anatomy. A comparison between the two analytical spectroscopic...
techniques MSI and in vivo magnetic resonance spectroscopic imaging is made. Thus, it is shown how a large amount of information can be extracted from the MSI hyper-spectral data, consisting of spatial information and spectral signatures, with the presented set of processing, visualisation and modelling tools.

A typical ‘metabonomic workflow’ is schematically shown in figure 1.4, and the developed methods are highlighted; these include the different types of analytical spectroscopy followed by the

---

Figure 1.4: The metabonomic workflow consisting of the experimental design, sample collection, analytical spectroscopy, data processing and analysis steps. The focus of this thesis is on the generation of spectroscopic metabolic profiles with analytical techniques followed by data processing and unbiased analysis of the large amount of spectroscopic data with statistical and pattern recognition techniques. Chapter 3 focuses on J-resolved \(^1\)H NMR spectra and the subsequent data processing and analysis steps; Chapter 4 details the optimisation of the kernel parameter in K-OPLS with simulated annealing and the use of this non-linear SA-K-OPLS method in metabolic profiling; Chapter 5 concerns the de novo development of a data processing approach for mass spectrometry imaging data followed by the introduction of several methods for data modelling and visualisation.
appropriate data processing, modelling and visualisation steps. The application of the described approaches to biological samples should help the experienced researcher to extract a maximum amount of information from the study efficiently and successfully. This would aid biological interpretation, for example through improved biomarker identification based on J-resolved spectroscopy, the generation of transparent, high-quality predictive models with SA-K-OPLS and simultaneous interpretation of spatial and spectral profiles of MSi spectra. Thus, a greater amount of biological information can be obtained from the samples, and predictive models and visualisation tools to evaluate interventions and interpret metabolic variables are developed.
Chapter 2

Theory and methods

This chapter will describe the analytical spectroscopic methods discussed in this thesis and the multivariate modelling techniques employed. Certain sections were adapted from an article that was published recently, which is included as appendix C.56

2.1 NMR: nuclear magnetic resonance

2.1.1 Spins and magnetic field

Nuclear magnetic resonance (NMR) is based on a fundamental property of atoms: the nuclear spin (I). This quantum mechanical property is related to the total angular momentum of the nucleus, and the discrete values that can be assumed are determined by the number of protons and neutrons constituting the nucleus, the spin number is therefore different per isotope. NMR spectra are easiest to interpret when they are generated by isotopes that have only two allowed values for the spin quantum number (spin $I = \frac{1}{2}$ nuclei, with $2I + 1 = 2$ eigen states and hence two spin projection quantum numbers $m_s = +\frac{1}{2}$ and $m_s = -\frac{1}{2}$). Table 2.1 lists nuclei commonly measured in biomolecular NMR, only spin systems with $I = \frac{1}{2}$ will be discussed here. The presence of protons in a large number of biological molecules, combined with high natural abundance and large magnetogyric ratio $\gamma$ (see equation 2.1 and table 2.1) makes $^1$H the most relevant nucleus for generating metabolic fingerprints.39,40

<table>
<thead>
<tr>
<th>Isotope</th>
<th>Ground state spin</th>
<th>Natural abundance</th>
<th>Magnetogyric ratio $\gamma$ (rad/s-T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^1$H</td>
<td>$\frac{1}{2}$</td>
<td>-100%</td>
<td>$267.522 \times 10^6$</td>
</tr>
<tr>
<td>$^{13}$C</td>
<td>$\frac{1}{2}$</td>
<td>1.1%</td>
<td>$67.283 \times 10^6$</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>$\frac{1}{2}$</td>
<td>0.37%</td>
<td>$-27.126 \times 10^6$</td>
</tr>
<tr>
<td>$^{19}$F</td>
<td>$\frac{1}{2}$</td>
<td>-100%</td>
<td>$251.815 \times 10^6$</td>
</tr>
<tr>
<td>$^{31}$P</td>
<td>$\frac{1}{2}$</td>
<td>-100%</td>
<td>$108.394 \times 10^6$</td>
</tr>
</tbody>
</table>

Table 2.1: Nuclei that are commonly measured in biomolecular NMR.69
When a nucleus is placed in an external magnetic field $B_0$, there will be an energetic advantage for spins to align with the magnetic field. The energy difference between the states $m_s = +\frac{1}{2}$ and $m_s = -\frac{1}{2}$ in a magnetic field is described by equation 2.1, where $\gamma$ is the magnetogyric ratio (see table 2.1) and $\hbar$ is the reduced Planck constant ($1.0546 \times 10^{-34}$ J·s).

$$E = -\gamma m_s \hbar B_0$$

(2.1)

The result is a net overall magnetisation of the sample caused by a slight excess population of spins in the lower energy state, as described by the Boltzmann distribution, given in equation 2.2.

$$\frac{N^-}{N^+} = e^{-\Delta E/kT}.$$  

(2.2)

In equation 2.2, $\Delta E$ is the energy difference between the spin states, $T$ is the temperature, $k$ is Boltzmann's constant ($1.38066$ J/K), and $N^+$ and $N^-$ are the number of spins in the lower and higher energy levels, respectively.

### 2.1.2 Detection of the NMR spectrum

For each nucleus with a non-zero spin quantum number, there is an associated magnetic moment. This magnetic moment is parallel to the spin angular momentum (if the magnetogyric ratio $\gamma$ is positive, or anti-parallel otherwise) and precesses around the axis of the direction of the $B_0$ field (e.g. the $z$-axis), which is indicated in figure 2.1 A. The angular frequency of the precession is called the resonance or Larmor frequency, and is related to the strength of the external magnetic field and the magnetogyric ratio $\gamma$ as shown in equation 2.3. For commercial NMR spectrometers with a $B_0$ of $14.1$ T, this Larmor frequency $\omega_0$ corresponds to $600$ MHz for $^1$H detection.

$$\omega_0 = -\gamma B_0$$

(2.3)

If the magnetisation of the ensemble of spins is calculated, a net magnetisation vector aligned with the $B_0$ field results ($M_0$ in figure 2.1 B). It is possible to introduce transitions of the spins between the 2 states (for $I = \frac{1}{2}$) by applying an oscillating voltage that exactly matches the energy difference in equation 2.1: the Larmor frequency. The irradiation at $\omega_0$ is applied in the form of a short radio frequency pulse to rotate the net magnetisation for each type of nucleus (e.g. different protons in a molecule), present when a sample is at thermal equilibrium in an external magnetic field, see figure 2.1 B. Commonly used pulses rotate the magnetisation vector by either $90^\circ$, to achieve maximum orthogonal magnetisation (i.e. no magnetisation component remains in the direction of the $B_0$ field, and the two spin energy level populations are equalised), or $180^\circ$, to invert the current magnetisation (and thus to invert the spin energy level populations).
Figure 2.1: (A) Each nuclear spin has a magnetic moment that will precess around the applied magnetic $B_0$ field at the Larmor frequency. (B) Rather than evaluating all the precessing individual spins, a vector representation is used, where the net nuclear magnetic moment for each type of nucleus will be aligned with the $B_0$ field at thermal equilibrium. The applied radio frequency ($rf$) pulse rotates the direction of the magnetic moment. (C) Once the $rf$ pulse is stopped, the nuclear spins will again precess around the $B_0$ field at the Larmor frequency, see equation 2.3. This magnetisation precession can now be detected as it induces a current in the receiver coil, which is measured as the free induction decay (FID). (D) Through $T_1$ relaxation processes, the net magnetic moment will over time become realigned with the $B_0$ field, and thus if this field was along the $z$-axis, magnetisation along the $z$-axis, $M_z$, gradually returns to thermal equilibrium. (E) The $T_2$ relaxation mechanisms (e.g. inhomogeneous magnetic fields, different local magnetic fields) cause the gradual loss of synchronisation of the precessing spins. This causes the decay of the net transverse relaxation ($M_{xy}$ if $B_0$ was along the $z$-axis).

When the radio frequency pulse is stopped, the nuclear spin polarisations will once again precess around the $B_0$ field with a periodicity governed by the Larmor frequency, see equation 2.3 and figure 2.1 C. This precession of the spins around the magnetic field induces a weak radio frequency signal in the transverse plane, which is detected and amplified to give the ‘free induction decay’ (FID), which will be the sum of all the individual contributions of the precessing nuclei in the sample. This time-domain signal is Fourier transformed to create a spectrum representing the resonance frequencies of the nuclei. NMR can be considered a quantitative technique, as the induced signal and thus the intensity of the peak in the Fourier transformed spectrum is proportional to the number of nuclei at a given resonance frequency.

2.1.3 Chemical shift

The exact magnetic field experienced by a given nucleus is determined by a combination of both $B_0$ and the local electronic environment, because the external magnetic field induces a magnetic field in the electron cloud. For example, the lactic acid (lactate, $\text{CH}_3\text{CH(OH)}\text{COO}^-$) protons in the $\text{CH}_3$ moiety will experience an increased electron density which effectively decreases the
Figure 2.2: The different nuclei in the lactic acid molecule will resonate at different frequencies, displayed on the ‘ppm’ or chemical shift scale. Note how the chemical shifts are reported relative to the reference molecule defined to resonate at 0.00 ppm. The intensities of the peaks are related to the number of protons giving rise to the signal (indicated in red in the chemical structures). The J-coupling that causes the appearance of the quartet (4 peaks) and the doublet (two peaks) will be described in §2.1.5.

experienced magnetic field, as the nucleus is relatively shielded from the magnetic field compared to the proton close to the electronegative carboxylic acid group. Thus, the CH$_3$ proton nuclei precess with a lower resonance frequency than the CH proton, resulting in a lower chemical shift (sometimes referred to as ‘upfield’, see figure 2.2 for nomenclature).

The chemical shift, defined in equation 2.4, is a field-independent difference in Larmor frequency for two nuclei ($\nu$ and $\nu_{\text{ref}}$), often the nucleus of interest and a pre-defined reference nucleus.

$$\delta = 10^6 \times \frac{\nu - \nu_{\text{ref}}}{\nu_{\text{ref}}}$$  \hspace{1cm} (2.4)

This chemical shift is a unitless entity denoted as ‘$\delta$’ or ‘ppm’, the latter denotes parts per million, as the difference is typically in the order of Hz whereas the Larmor frequency is measured in MHz (hence the multiplication with $10^6$ in equation 2.4). This definition results in consistent chemical shift values, irrespective of the machine field strength: thus at both 80 MHz and 600 MHz $^1$H observation frequency, the chemical shift of succinate will be 2.41 ppm from TSP (3-(trimethylsilyl)-[2,2,3,3-$^2$H$_4$]-propionic acid sodium salt, a commonly used reference molecule, for which the chemical shift will be defined as $\delta$ 0.00).

### 2.1.4 Relaxation

After the radio frequency pulse has been applied, the net magnetisation, that is precessing around the $B_0$ field, will slowly return to the thermal equilibrium magnetisation (i.e. aligned with the externally applied $B_0$ field) through a process known as relaxation, see figure 2.1 D and E. $T_1$ or spin–lattice relaxation is the build-up of the longitudinal magnetisation, that is the Boltzmann distribution, through the dissipation of gained excess energy from the radio frequency pulse (compared to equilibrium) with thermal motion. $T_1$ relaxation therefore is related to the restoration
of the net magnetisation in the direction of the $B_0$ field, see $M_z$ in figure 2.1 D. In addition to this relaxation mechanism, $T_2$ or transverse relaxation is caused by the disappearance of the coherence between the large number of precessing nuclei, caused by the small microscopic magnetic field differences experienced by each spin. This is sometimes referred to as ‘dephasing’ or chemical exchange, see figure 2.1 E, and results in a loss of transverse magnetisation $M_{xy}$.

2.1.5 $J$-coupling

An expansion of a typical NMR spectrum of a biofluid is presented in figure 2.3. In this figure, it is clear that the complexity in composition of biological fluids is mirrored in the NMR spectrum, with a multitude of (overlapping) signals. Some of the observed complexity in the spectra is the result of a phenomenon known as ‘$J$-coupling’. In addition to the local electronic environment, the Larmor frequency is also dependent on the presence of other magnetic nuclei in the same molecule, e.g. whether a neighbouring nucleus has a $m_s = +\frac{1}{2}$ or $m_s = -\frac{1}{2}$. For example, the lactate CH$_3$ group has 3 ‘chemically equivalent’ protons, which means all protons are of the same isotopic species and are exchangeable through a molecular symmetry operation.$^{69,70}$ This CH$_3$ group will split the neighbouring CH proton into 4 peaks, as there are 4 combinatorial possibilities of 3 spins with value $m_s = \pm\frac{1}{2}$, see figure 2.4. The energy difference between spins in a multiplet is known as $J$-coupling (typically a few Hz for $^1$H–$^1$H coupling). The splitting pattern thus gives information about the number and chemical equivalence of the neighbouring nuclei: for a CH$_x$ group, the multiplicity will be $x + 1$ (e.g. a quartet for a CH$_3$ group peak splitting) and the relative intensities of the multiplet can be deduced from Pascal’s triangle, see figure 2.4.

![Figure 2.3: The complexity of a biofluid spectrum is clear from this ‘1D’ spectrum of a urine sample from a healthy rat acquired at 800 MHz as part of the galactosamine toxicity study described in chapter 3. The expansion illustrates that there is a large degree of overlap between the different peaks constituting the spectrum.](image-url)
Figure 2.4: Depending on the number of chemically equivalent neighbouring nuclei \( x \), a signal is split into \( x + 1 \) peaks, based on the combinatorial combinations of \( m_s = \pm \frac{1}{2} \) (for \( I = \frac{3}{2} \)); the relative intensities are shown on the right and can be deduced from Pascal’s triangle (the values in brackets are normalised to a peak area of 1).

The coupling constant and splitting information is of great importance for structure elucidation and peak identification. However, the drawback is that the presence of \( J \)-coupling also increases the complexity of the acquired NMR spectra, and results in an increased overlap between peaks, especially in complex mixtures. Methods to obtain proton-decoupled proton spectra, i.e. spectra where no \( J \)-coupling is present, have been suggested early in the history of NMR\(^7\) and will be thoroughly discussed in the context of metabonomics in chapter 3.

2.1.6 ‘1D’ and Carr–Purcell–Meiboom–Gill spectra

The NMR spectrum presented in figure 2.3 was acquired from a urine sample. To achieve such a spectrum, the solvent peak (water in biological samples) has to be suppressed in order to avoid an overflow of the analogue-to-digital signal converter, which would hinder the digitisation and detection of the much smaller metabolite signals. In order to achieve quantitative, interpretable spectra, the sample is commonly ‘irradiated’ at the water peak frequency, which effectively dephases the water resonance. The standard ‘1D’ experiment typically performed in metabonomics consists of the following pulse sequence:

\[
(d_1 - 90\mu s - 90\mu s - t_m - 90\mu s - \text{acquire FID})
\]

with water peak suppression during \( d_1 \) (the relaxation delay) and \( t_m \) (the ‘mixing’ time). In this sequence, the three repeated 90° pulses give a ‘standard’ pulse-and-acquire spectrum, while enhancing the suppression of the residual water resonances from sample regions outside the receiver coil, that have a small contribution to the spectral appearance.

The pulse sequence known as the Carr–Purcell–Meiboom–Gill sequence (CPMG) may be used to suppress the resonances from larger molecules due to their shorter \( T_2 \) relaxation time constants:

\[
(d_1 - 90\mu s - [\tau - 180\gamma - \tau]_n - \text{acquire FID})
\]

The spin–spin relaxation delay \( 2\tau \) is the key feature of the sequence, attenuating signals from slow tumbling molecules with small \( T_2 \) values. The CPMG pulse sequence is particularly useful
2.1.7 Two-dimensional NMR methods

One approach to overcome the overlap of signals in one-dimensional NMR spectra of complex mixtures is to disperse the signals using a second frequency axis to create two-dimensional (2D) NMR spectra. This second axis is generated by incrementing a time-delay ($t_1$) in the pulse sequence and collecting an FID during time $t_2$ at each delay value. Thus, the NMR signal is recorded as a function of two time variables, $t_1$ and $t_2$. A general 2D sequence contains a preparation period, often to generate transverse magnetisation, an evolution period to allow the spins to interact, a mixing period to manipulate the spins and create observable signals, which are recorded during the detection period. This 2D array of data, created by increments in the evolution time $t_1$, is then Fourier transformed twice to create a spectrum where the two axes correspond to the frequencies $F_1$ and $F_2$. The arraying concept is easily extended to three or more dimensions.

The nature of the preparation and mixing pulse sequences determine the information that is measured in the spectrum, and with different approaches, information on various structural properties can be obtained. Multidimensional encoding of NMR spectra can be based on correlation techniques, which allows the investigation of spin connectivity via chemical bonds (through $J$-coupling in COSY, TOCSY, etc.) or the distance between nuclei, based on the nuclear Overhauser effect (through dipolar coupling in NOESY, ROESY, etc.). This information aids spectral interpretation in two ways: 1) determining which peaks arise from the same molecule, 2) reducing the overlap between the peaks that was visible in the 1D spectrum.

Another form of 2D spectral encoding is based on $J$-resolved techniques ($J$RES, see chapter 3), where the encoding in the second dimension is not based on chemical shifts, but the $J$-coupling, for example $^1$H–$^1$H coupling (homonuclear $J$RES) or $^1$H–$^{13}$C coupling (heteronuclear $J$RES). The
encoding of a second axis can also be based on diffusion coefficients, an experiment known as diffusion edited spectroscopy.

Usually, 2D pulse sequences are repeated many times to build up sensitivity, and with numerous increments in \( t_1 \). Thus, the experiment time of 2D NMR pulse sequences is often much greater than for 1D NMR experiments and in general, any form of 2D spectroscopy suffers from lower sensitivity than a 1D experiment.

2.2  MS: mass spectrometry

2.2.1 Basics of mass spectrometry

The molecular composition of samples can also be studied with mass spectrometry (MS),\(^72\),\(^73\) which has a high sensitivity and the ability to measure a diverse array of metabolites, and is thus considered one of the main techniques for metabolic profiling complementary to NMR, discussed in §2.1. Molecules with different chemical compositional formulae including different naturally abundant isotopes will have different molecular weights. Therefore, molecular information can be extracted by measuring the molecular weight of the species, and this is what MS achieves. MS uses the property that a charged particle placed in an electric or magnetic field experiences a Lorentz force \( F \). This force is dependent on the ion charge \((q)\), the electric field \((E)\), the ion velocity \((v)\) and the magnetic field \((B)\), according to equation 2.5 (where \( \times \) represents the vector cross product).\(^74\),\(^75\)

\[
F = q(E + v \times B)
\]  
(2.5)

Because the exerted force equals the product of the acceleration \( a \) and the mass \( m \) (Newton’s second law, see equation 2.6), these equations can be combined into a differential equation from which the ratio \( \frac{m}{q} \) can be determined, see equation 2.7. The charge \( q \) is often divided by the elementary charge \( e \) (\(1.602 \times 10^{-19} \text{ C}\)) to obtain the charge number \( z \). This ratio of mass to charge number is often referred to as mass-to-charge ratio, and represented as \( m/z \) in this thesis.

\[
F = ma
\]  
(2.6)

\[
\frac{m}{q}a = E + v \times B
\]  
(2.7)

The first step in MS is the ionisation of molecules (to create the charge \( q \)), which can be performed in various ways. The focus in this thesis will be on a method known as matrix-assisted laser desorption/ionisation (MALDI). This is an ionisation method that allows molecular ions to be formed, in contrast to other methods that can create a large number of charged fragments from each molecular ion. It is therefore known as a ‘soft’ ionisation technique and is very suitable for large biomolecules, as those are very fragile and would be comprehensively fragmented if subjected to other ionisation approaches. A matrix is applied to the sample of interest and the matrix molecules will co-crystallise with the analyte upon vaporisation of the solvents, and when a laser
beam is targeted to the sample plate, the matrix molecules will absorb the energy of the laser irradiation to become desorbed and ionised. Gas-phase chemistry causes the transfer of ionisation to the matrix-trapped analyte molecule [M] to create ionised analyte molecules such as [M+H]^+, [M+Na]^+ and [M+K]^+; double or multiple charged ions can also be formed.74,75

The ions can then be measured using time-of-flight (TOF) MS, where the ions experience an electric field. The potential energy $E_p$ based on the electric potential $V$ is converted to a kinetic energy that will be equivalent for equally charged molecules, see equation 2.8.

$$qV = E_p = E_k = \frac{1}{2}mv^2$$

The velocity $v$ of the molecule will depend on the mass $m$ and charge $q$. Thus, by measuring the time-lapse to reach the detector one can infer the mass of the molecule: the heavier particles will travel more slowly, hence reaching the detector at later times.74,75 The detector monitors the charge induced or the current produced when an ion passes by or hits a surface, and this signal is amplified and recorded for the different flight times to produce a mass spectrum. An example mass spectrum (acquired as part of the MALDI imaging of the sample discussed in chapter 5) is shown in figure 2.6.

### 2.2.2 MSI: mass spectrometry imaging

Rather than acquiring a mass spectrum from a biofluid sample, e.g. after separation by chromatography, it is possible to directly analyse a tissue or other sample of interest using a MALDI imaging approach. Whereas in conventional MALDI, the sample (e.g. a protein purified from a gel) is ‘spotted’ on a plate, and then the matrix is applied to that in order to get a mass spectrum, one can simply use the tissue as a ‘plate’. For example, a cryo-sectioned tissue slice can be mounted on a slide, see figure 2.7 A1 and A2, and after the matrix is applied to the sample, figure 2.7 A3, a mass spectrum can be generated without additional preparation, as shown in 2.7 A4.51

Matrices used for MALDI on biological substrates need to efficiently absorb the laser irradiation, have good vaporisation characteristics and be able to act as a proton source to encourage ionisation; examples of crystallising matrix molecules are 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid).
Figure 2.7: (A) A sample is cryo-sectioned (step $A_1$) and mounted on a plate (step $A_2$), to which a matrix is applied (step $A_3$) in preparation for mass spectrometry imaging (step $A_4$). In mass spectrometry imaging, a mass spectrum is acquired for each position along a grid. (B) The MSI data can be presented in a data tensor of size $x \times y \times n$ (step $B_1$) and analysed with multi-way methods such as PARAFAC. Alternatively, the full data set can be unfolded into a $x \times y \times n$ matrix (step $B_2$) and analysed with two-way chemometric techniques. Common univariate data analysis involves the evaluation of each of the $n$ individual $m/z$ values, represented as an image (step $B_3$), or each of the $x \times y$ pixels, which can be represented by a characteristic mass spectrum (step $B_4$).

acid) and alpha-cyano-4-hydroxycinnamic acid ($\alpha$CHCA). The matrix molecules are dissolved in a mixture of organic and aqueous solvent to facilitate both hydrophobic and hydrophilic analyte molecules to dissolve and subsequently co-crystallise with the matrix molecules.

The acquisition of a spectrum using MALDI-imaging is analogous to the approach frequently taken in metabolomics, where instead of purifying, identifying and quantifying each individual constituent, a characteristic fingerprint of the sample composition is generated with a powerful analytical approach. As a result, the information density of spectra generated in this manner is very high.
Once a spectrum has been acquired in one location, the set-up is repositioned in order to acquire a mass spectrum from a different part of the sample; using this mass spectrometry imaging (MSI) approach, a ‘grid’ of mass spectra across the sample is obtained, see figure 2.7.\textsuperscript{51,76,77} Thus, localised information is obtained, reporting the chemical make-up of the different regions with mass spectrometry-based molecular fingerprints. It is not hard to imagine that these distribution maps of biomolecules, e.g. shown in figure 2.7 B\textsubscript{3}, should consistently reflect properties of different tissue regions. This can, for example, aid characterisation of a tissue into different structural units, and additionally make molecular masses available that typify these segments.\textsuperscript{78} A plethora of data analysis objectives and methods are available to evaluate the images and the spectroscopic data, and this will be the main topic of chapter 5: how to extract the maximum amount of useful information from an MSI data set. In general, these approaches can be divided into univariate methods:

- Studying single $m/z$ images, as shown in figure 2.7 B\textsubscript{3}.
- Evaluation of selected mass spectra for different regions in a ‘profiling’ manner, shown in figure 2.7 B\textsubscript{4}.

and multivariate methods:

- Analysis of the complete 3-way data tensor with multi-way analysis methods such as PARAFAC and TUCKER, see figure 2.7 B\textsubscript{1}.\textsuperscript{4,79}
- Unfolding the data into a two-way matrix, where the pixels are displayed in the different rows and the $m/z$ variables in the different columns, shown in figure 2.7 B\textsubscript{2}. Then multivariate methods such as principal component analysis and self-organising maps can be performed.

There are a large number of other ionisation, mass analysing and detection methods, which were not discussed here. It should be noted though, that some of these can also be used in an imaging set-up, for example secondary ion mass spectrometry (SIMS). However, MALDI is the method of choice for biological matrices as it is a soft ionisation technique and therefore the compounds are not fully fragmented, and TOF is the most used complementary mass analyser companion. MALDI and TOF are compatible through the pulse-process nature of the laser-based ionisation, moreover, TOF is able to analyse a large mass range. Imaging data acquired with the MALDI-TOF MSI experiment are discussed in chapter 5.
2.3 Multivariate modelling

Multivariate statistical modelling is essential for a thorough and systematic analysis of spectral data: it evaluates the patterns of variance within the spectra reflecting biochemical properties. Metabonomic spectral data have some characteristics that prohibit the application of many conventional univariate and regression methods. This section will discuss chemometric methods that can successfully model and visualise the measured spectral data. \(^65,66\)

Notation

The obtained experimental spectroscopic data are conventionally arranged in matrices with rows representing different samples or measurements and columns corresponding to variables. Matrices will be represented in bold upper-case, vectors in bold lower case and scalars in plain lower case. Let \(X\) be the matrix of descriptor variables of dimensionality \([n \times k]\) for \(n\) observations and \(k\) variables and \(Y\) be the matrix of response variables of dimensionality \([n \times m]\) for \(m\) variables in \(Y\).

In the case of classification, \(Y\) is a ‘dummy’ matrix consisting of ones and zeros, forming a design matrix indicating the class of each observation. The matrix transpose is indicated as \(T\).

2.3.1 Data processing

Normalisation

When performing any data analysis, it is important to consider the need and cause for normalisation. Spectral data such as those generated by NMR and MS have to be normalised (a row operation) because the specific settings of acquisition as well as dilution due to sample preparation or inter-sample variability, e.g. urine concentration, can influence peak intensity levels. There are several methods available for normalisation of metabonomic data:\(^80,81\)

- Samples can be normalised to total area, i.e. unit integral. For this method of normalisation, it is necessary to remove large, non-quantitative peaks such as water or other solvent peaks, and urea in urinary NMR spectra acquired with pre-saturation.
- Normalisation to the absolute concentration of a reference compound, e.g. TSP or creatinine.
- Probabilistic quotient normalisation (PQN, also known as median fold change normalisation)\(^82\) is a robust method to calculate the most probable dilution factor with respect to a chosen reference spectrum \(x_{ref}\) (which is e.g. the median of all samples), see equation 2.9, where \(x\) is the original spectrum and \(x_n\) is the normalised spectrum.

\[
x_n = \frac{x}{\text{median}(x \times \frac{1}{x_{ref}})} \tag{2.9}
\]
Centring and scaling

Generally, the most interesting aspect from the data is its variation, rather than the magnitude of the actual values: information about change and continuity is often of great biological interest. Hence, data are normally mean-centred prior to further analysis, which is achieved by subtracting the mean of each variable as a column operation, to remove the effect from the average.

Equalisation of the importance of all variables, by reducing the effect of high valued variables and emphasising small variables, is achieved through scaling each variable. A common scaling method divides each column by its standard deviation, in order to achieve unit variance. Other methods include pareto scaling (division by the square root of the standard deviation), range and log scaling, and a combination of different approaches can be used as well. In general, data are mean-centred before any scaling is applied. Although scaling to unit variance changes the relative influence of the variables on the model, so that the study of low-concentration metabolites is possible, it simultaneously increases the influence of spectral noise on the model.

2.3.2 Univariate versus multivariate approaches

Initially, analysis of metabolic spectral profiles was performed visually, but it was soon recognised that pattern recognition methods had great potential for metabolomic studies. High-dimensional metabonomic data are commonly analysed by multivariate (latent-variable based) methods; where a latent variable is a weighted linear combination of the original spectral variables. These latent variables can delineate groupings, which are often not visible with any classical univariate analytical approach, where one variable is considered at a time. Only multivariate methods can really take into account the relationships between variables: multi-variable patterns can be significant, even if the individual variables are not.

Two commonly used multivariate methods are principal component analysis (PCA) and partial least squares regression (PLS). With these projection methods, the enormous spectroscopic data set can be represented by a few interpretable latent variables, as the dimensionality of the data is greatly reduced. These projection approaches are advantageous for a reliable and flexible analysis of NMR and MS data as they can handle the multicollinearity and information redundancy commonly observed in spectral data sets.

Multicollinearity is a prominent characteristic of metabolic data sets, as a chemical species may contain multiple signals (ions, multiplets) in several spectral regions, and multiple variables constitute one peak. The data are high-dimensional, with the number of variables exceeding the number of samples, which is problematic for classical linear regression methods. Additionally, the different variables are often correlated; all of these issues are appropriately addressed through the use of latent variable methods.
Unsupervised and supervised methods

A distinction between methods to model spectral data is whether they are unsupervised or supervised methods, see figure 2.8 for some examples used in this thesis. Unsupervised methods are based solely on the measured data, and no information about different classes or other response variables of interest are provided to the model. Examples of this approach are principal component analysis (PCA) and self-organising maps (SOMs). The other branch consists of supervised models, which model the relationship between the (spectral) descriptor data, \( X \), and the response data, \( Y \). Examples of these are partial least squares (PLS) and its derivative methods such as orthogonal partial least squares (OPLS) and the kernel-based version of OPLS, K-OPLS (see chapter 4).

2.3.3 PCA: principal component analysis

Principal component analysis (PCA) describes the main directions of variation in a high-dimensional data set through multivariate projection. PCA models are commonly used for global exploration of the data and identification of outliers. The data are summarised in a new coordinate system that explains the variation from the original data set using a number of orthogonal latent variables, which are the axes, or principal components (PCs) in the new coordinate system and are linear combinations of the original variables. The first principal component is in the direction that explains the largest variance in the data, subsequent components are orthogonal to the previous ones and explain the largest remaining variance.
The method is based on a singular value matrix decomposition of the $X$-data matrix, see equation 2.10.

$$X = TP^T + E$$  \hspace{1cm} (2.10)

The $X$ data can be projected on the new axes (principal components): the new coordinates of the objects are called ‘scores’ ($T$, $[n \times a]$, where $a$ is the number of principal components calculated). The weights of the original variables that create the principal components are represented by the loadings ($P$, $[k \times a]$), see figure 2.9. The remaining variance is contained in the residual matrix ($E$, $[n \times k]$). There is a direct relation between the scores and loadings: scores having a higher coordinate on a PC have higher values for the loading on that principal component, so in figure 2.9:

(A) 3D plot of three collinear variables: age, body-mass-index ($BMI$) and waist-to-hip ratio ($WHR$). PCA created two principal components, indicated with red lines, and the newly created coordinate plane spanned by them is indicated. (B) The score plot shows the position of the samples in the new PCA coordinate system. (C) The contribution of the original variables to the new principal component axes are shown in the loadings plot. In the first component, higher BMI, WHR and age are related, but the second component models variation of subjects with lower age and high BMI and increased WHR. The colour-coding of the points is related to internal fat levels (high levels are red) to aid identification of corresponding points in A and B. Data were acquired by members of the Metabolic and Molecular Imaging group at the Hammersmith Hospital Campus, Imperial College London, UK.
2.9 B the green points on the right hand side have a higher age, body-mass-index and waist-to-hip ratio, as can be seen in the loadings plot in figure 2.9 C.

As additional components are included in the PCA model, the variation that is modelled typically contains more noise, whereas the first components represent the global data structure. The number of necessary components is often a lot smaller than the number of original variables, and thus PCA is suitable for data compression. The percentage of total variance that is explained by each component in PCA can be used to assess the importance of the individual PCs and to determine the number of components that should be used in the model.

Projections created with PCA allow an unbiased overview of the data set, and can directly summarise and visualise the prominent metabolic changes, see for example the chemical reactions monitored in figure 2.10.

![PCA scores](image)

![PCA loadings](image)

Figure 2.10: To demonstrate the use of PCA on spectral data, this model is calculated from mean-centred NMR data of a sample that changes in composition over time. Plotted are the scores of PC 1 versus PC 2 (top), the order in which spectra were acquired is indicated with a colour (blue for the first experiment), demonstrating the effect of time. The corresponding loadings of PC 1 (bottom), demonstrate how the global shape of the NMR spectrum changes over time. These data were also used to evaluate ‘STORSY’ in chapter 3.

### 2.3.4 PLS: partial least squares

Partial least squares (PLS, also known as ‘projections to latent structures’) models the variation in X, whilst being directed by a response data set Y: PLS derives the components from the descriptor data set X which best describe the specified Y structure. This is done by maximising the covariance, expressing the common structure between the X and Y matrices. It can therefore be used for prediction purposes, and is a ‘supervised’ technique. This is useful for regression cases such as
prediction of time after a certain event (e.g. infection, exposure to a toxin) or discriminant analysis of groups, e.g. ‘control’ versus ‘diseased’.

**PLS** decomposes the $X$ and $Y$ matrices according to equations 2.11 and 2.12, where $T [n \times a]$ and $U [n \times a]$ are the score matrices and $P [k \times a]$ and $C [m \times a]$ are the loading matrices for $X$ and $Y$, respectively. The unmodelled residuals are contained in the $E [n \times k]$ and $F [n \times m]$ matrices.\(^{65,66,87}\)

$$X = TP^T + E \quad (2.11)$$

$$Y = UC^T + F \quad (2.12)$$

An alternative way to formulate the relation between $X$ and $Y$ is by means of the **PLS** regression coefficients $b_{PLS}$ in equation 2.13, where $G$ represents the model residuals. This equation can be used to predict $Y$ from $X$ data, $\hat{Y}$.

$$Y = Xb_{PLS} + G \quad (2.13)$$

**PLS** is sometimes subdivided into regression analysis (**PLS-Re**) and discriminant analysis (**PLS-Da**). In classification or discriminant analysis, samples are allocated to discrete classes, which are represented by so-called ‘dummy’ variables (encoded with 0 and 1).

**PLS**, and supervised approaches in general, are often applied to tackle more subtle problems in metabolomics, in which the large data sets require an approach that permits relationships buried in a background of other large and multiplexed effects to be uncovered. Several processes and their metabolic contributions control the composition of the same biosample, as was seen in figure 1.2, and the main source of variation may not address the biological question: not all influences are relevant to the variable of interest and these should be disregarded or down-weighted during analysis. For example, human urine can display large differences between individuals,\(^{46}\) and different metabolic phenotypes have been suggested,\(^{89}\) which can complicate studies on effects such as physiological challenges, the effect of probiotics and prebiotics, and nutritional status on the metabolome. A virtue of latent variable projection methods such as **PLS** compared to other predictive modelling approaches, is the transparency of the models with respect to the scores, that would allow the identification of e.g. possible sub-groupings, and loadings, which allow metabolic interpretation. Latent variables may be visually presented in pseudo-spectral form, to aid interpretation by evaluation through a spectroscopic appearance.\(^{90}\)

### 2.3.5 OPLS: orthogonal partial least squares

**Orthogonal signal correction**

Orthogonal signal correction (**OSC**) is a filtration procedure that was introduced for **PLS** in order to remove the information orthogonal to $Y$ from the $X$ matrix.\(^{91}\) This procedure can be seen as a filter aimed at cleaning $X$ of the information unrelated to $Y$, such as caused by run order, temperature drift or batch effects.\(^{92}\)
Orthogonal partial least squares (OPLS) is a supervised latent variable approach, similar to PLS, but with an integrated OSC filter. Originally, OSC was applied as a separate step in a data processing sequence, and, as such, was troublesome regarding validation and detection of overfitting. Therefore, OSC and PLS were integrated within a single-procedure decomposition, OPLS, which performs a regression with orthogonal filtration of $X$ on a vector or matrix $Y$. This resolved the validation difficulties of OSC-PLS and was a powerful addition to the repertoire of PLS methods, because the variance in a data set can be split by OPLS into predictive and orthogonal variance.

The OPLS algorithm decomposes the descriptor data matrix $X$ into a set of predictive (response-related) latent variables and a set of response-orthogonal latent variables. Let $T_p [n \times a_p]$ and $T_o [n \times a_o]$ represent the predictive and response-orthogonal score matrices respectively, with $a_p$ predictive components and $a_o$ orthogonal components (OC). $P_p [k \times a_p]$ and $P_o [k \times a_o]$ are the corresponding loading matrices. The OPLS model is outlined in equation 2.14, where $E [n \times k]$ is the residual matrix.

$$X = T_p P_p^T + T_o P_o^T + E$$  \hspace{1cm} (2.14)

The orthogonal components should be evaluated to suggest or determine the basis of this ‘structured noise’. The variation described in the orthogonal components can, if not filtered, decrease the interpretability of the resulting model, which happens in classical PLS models. OPLS allows an enhanced focus on variation of interest, whilst minimising other variation, which can be of e.g. biological or analytical source. Importantly, these orthogonal components add an extra dimension in which the data can be interpreted: both predictive and orthogonal components display highly relevant information from the data set. For example, the predictive and orthogonal components can be interpreted in relation to the experimental design, sample run order and other external variables.

In the discriminant analysis (DA) version of OPLS, OPLS-DA, the maximum separation between the class samples encoded in $Y$ (with $n$ dummy variables for $n$ classes) is achieved using the input data $X$, by decomposing the covariance matrix ($Y^T X$) into $n - 1$ predictive components and a sufficient number of orthogonal components, which can be determined by for example cross-validation (see §2.3.6). Subsequently, the loadings identify the variables associated with a specific class. The predictive ability of OPLS models is identical to PLS models for a single response variable, i.e. a $y$-vector.

An example of modelling and interpretation of spectral data with OPLS is shown in figure 2.11. This figure shows the results of an OPLS model that was calculated using mean-centred NMR data of a sample that changes in composition over time (the same data set as in figure 2.10).

In figure 2.11 A, a 3D plot of raw data from three chemical shift values is shown, representing three typical spectral regions: noise (x-axis, $\delta$ 0.8), a peak that decreases over time (from the starting compound, on the y-axis, $\delta$ 1.54) and a peak that increases over time (from a product
compound, on the z-axis, δ 6.13). In figure 2.11 B, the ‘inner relation’ plot of U versus T shows the degree to which the PLS components are of predictive value. To avoid over-optimism in the presented results, cross-validated values were plotted. A slight non-linear effect can be observed at the tails of this plot. A score plot of the predictive and orthogonal component, see figure 2.11 C, demonstrates the clear relation of the predictive component with time (again encoded with colour). Additionally, the orthogonal axis indicates the presence of orthogonal variation, which is possibly explained by the formation of intermediate compounds, that appear and disappear over time and are therefore of no value for the linear prediction, yet form ‘structured’ orthogonal

Figure 2.11: Demonstration of OPLS on a mean-centred NMR data set of a sample that changes in composition over time. Using 100 MCCV rounds (see §2.3.6, 75% of the data in the training set) and 1 OC, q² = 0.995. The order in which spectra were acquired is indicated with a colour (blue for the first experiment). (A) 3D plot of raw data of three typical spectral regions: noise (x-axis, δ 0.8), a peak that decreases over time (from the starting compound, on the y-axis, δ 1.54) and a peak that increases over time (from a product compound, on the z-axis, δ 6.13). (B) The ‘inner relation’ plot based on cross-validated values. (C) A score plot of the predictive and orthogonal component. (D) The weights of the predictive component of the OPLS model demonstrate how the global shape of the NMR spectrum changes over time, an expansion is shown for the three spectral values plotted in A.
variation. Finally, metabolic interpretation is possible using figure 2.11 D, where the weights of the variables in the predictive component of the OPLS model demonstrate how the global shape of the NMR spectrum changes over time. An expansion of the three peaks shown in figure 2.11 A shows that the noise region has no weight (zoomed in 100× at δ 0.8), the original compound decreases, and was of high concentration, hence a high weight value (negative weight, δ 1.54) and a low-concentration product is beginning to form (positive weight, zoomed in 100× at δ 6.13).

2.3.6 Model validation

One of the main advantages of PLS and OPLS is the predictive ability of the obtained models. However, without caution, any supervised modelling is prone to over-fitting. This section will discuss several validation approaches, and these are equally applicable to PLS, OPLS and K-OPLS (see chapter 4) models.

Ideally, the availability of a second, independent data set would allow for validation of the calculated PLS models. Alternatively, there are a number of methods to perform cross-validation (CV) based on the data that are used for modelling. These include methods that leave either one or a fraction of the samples out (such as n-fold cross-validation), but preference might be given to the use of resampling techniques, such as Monte Carlo cross-validation (MCCV) and bootstrapping. Monte Carlo techniques create a cross-validation training data set by randomly selecting e.g. 75% of the possible samples. This training data set is used to build a predictive model, and the model is validated using the test set, which consists of the samples that were not selected for the training set; this procedure is repeated a number of times.

Cross-validation gives predictions for the left-out samples, the test set, and one can advantageously use these test set predictions to create cross-validated score and inner relation plots, in order to get more realistic information on the predictivity and quality of the model. Moreover, cross-validation permits the calculation of model statistics such as $r^2$ and $q^2$. $r^2$ describes the proportion of variance modelled from the $X$ and $Y$ matrices and is sometimes referred to as the ‘goodness of fit’ of the model, i.e. how well the model actually explains the data, see equations 2.15 and 2.16 (assuming mean-centred data).

$$r^2_x = 1 - \frac{\sum E^2}{\sum X^2}$$  \hspace{1cm} (2.15)

$$r^2_y = 1 - \frac{\sum G^2}{\sum Y^2}$$  \hspace{1cm} (2.16)

The prediction performance of regression models will be evaluated with $q^2$, commonly referred to as the ‘goodness of prediction’, which is inversely proportional to the generalisation error, see equation 2.17.

$$q^2_y = 1 - \frac{\sum (\hat{Y}_{cv} - Y)^2}{\sum Y^2}$$  \hspace{1cm} (2.17)

Here, $\hat{Y}_{cv}$ is the vector (or matrix) of predicted values for $Y$ using cross-validation and the summation is over all elements of $Y$. 

43
In discriminant analysis, class membership is determined by the decision function \( f_{\text{max}}(\hat{Y}_{\text{cv},i}) \), returning the class label for which \( \hat{Y}_{\text{cv},i} \) has the highest value; \( \hat{Y}_{\text{cv},i} \) is a vector of the predicted values for each class for observation \( i \). The prediction performance of discriminant analysis models can be evaluated with the ‘mean sensitivity’, which is calculated across all \( k \) classes via equation 2.18, where \( tp_i \) is the number of true positives (i.e. number of detected cases) for class \( i \) and the denominator is the total number of cases for class \( i \), equivalent to the sum of the number of false negatives (\( fn_i \)) and true positives. A mean sensitivity equal to 1 implies perfect discrimination between the classes.

\[
\text{mean sensitivity} = \frac{1}{k} \sum_{i=1}^{k} \frac{tp_i}{tp_i + fn_i} \tag{2.18}
\]

In the special case of two classes, an alternative performance metric for discriminant analysis models uses a standard trapezoid approximation of the area under the receiver operating characteristic curve (AUC). This is calculated with equation 2.19, where the classification threshold is set at the predicted value (these numbers were sorted) for each of the \( n \) observations subsequently.\(^{100}\)

\[
\text{AUC} = \sum_{i=1}^{n-1} \frac{\Delta(1 - \text{specificity}) \times \Delta \text{sensitivity}}{2} \tag{2.19}
\]

Here, \( \Delta \) indicates the difference between two subsequent sensitivity (or specificity) values. The sensitivity is defined above, and the specificity is the ratio of true negatives to the total number of controls (true negatives + false positives). The AUC is interpreted as the probability that the model will rank a true positive observation higher than a true negative observation and is related to the non-parametric Wilcoxon rank sum test as well as the Mann-Whitney U test. An AUC value of 1 corresponds to perfect classification, a value of 0.5 is equivalent to random performance. The AUC metric has favourable properties in classification scenarios as it is independent of the class frequencies and does not depend on a cut-off point between the classes, which is used for dichotomisation.

Interestingly, the number of components in PLS and OPLS models is often determined based on predictive measurements such as the \( q^2_y \) value, but as the \( q^2_y \) is later used to report the predictive ability, some degree of ‘over-fitting’ is likely. Nested cross-validation can be used to give an unbiased optimisation of the number of orthogonal components whilst still reporting valid prediction results,\(^{101}\) and was employed in this thesis to optimise a parameter necessary in kernel-based OPLS, see §4.3.4. Another sophisticated validation approach is to use computer-intensive model resampling under the null hypothesis: a population of models is generated by random permutations of the \( Y \) data (i.e. the correspondence between \( X \) and \( Y \) is no longer present).\(^{102,103}\) By evaluating the prediction performance of these ‘random’ models in comparison to the original model, the value of the predictive measures can be interpreted. These approaches enable the validation of the obtained prediction, resulting in increased confidence in the results obtained with supervised modelling.
Whichever method of validation is chosen, it should be stressed that proper experimental design, data acquisition and cautious modelling are necessary to prevent over-interpretation of resultant models: a predictive model that cannot be generalised or has not been validated is of no value. 

2.3.7 STOCSY: statistical total correlation spectroscopy

A widely used univariate method, primarily employed to aid metabolite assignments, is statistical total correlation spectroscopy (STOCSY). Although this method is not multivariate, it is often used in conjunction with multivariate models as it constitutes a powerful tool for biomarker identification and structural elucidation. STOCSY is based on the multicollinearity of spectral peaks arising from the same compound across a series of samples: the relative peak intensities of a given molecule remain constant, as is illustrated in figure 2.12. Hence, STOCSY can be used to evaluate whether signals are from the same molecule.

![Figure 2.12: The peak heights of 2-oxoglutarate in the galactosamine toxicity study discussed in chapter 3 are directly related for the different samples. STOCSY uses this property: the constant relative peak heights of a compound result in a high correlation between peaks from the same molecule across a series of samples.](image)

The Pearson correlation $r(\mathbf{X}_\delta, \mathbf{X}_{\delta \text{ref}})$ between the signal intensities $\mathbf{X}_\delta$ at any resonance frequency $\delta$ and the signal intensities $\mathbf{X}_{\delta \text{ref}}$ at the selected frequency of interest $\delta_{\text{ref}}$ is calculated based on the n available spectra, following equation 2.20. This equation shows that the correlation is equivalent to the covariance of the intensities $\mathbf{X}_\delta$ and $\mathbf{X}_{\delta \text{ref}}$ divided by the standard deviations $\sigma$ of the peak intensities.

$$r(\mathbf{X}_\delta, \mathbf{X}_{\delta \text{ref}}) = \frac{\sum_{i=1}^{n}(X_{i,\delta} - \text{mean}(\mathbf{X}_\delta))(X_{i,\delta \text{ref}} - \text{mean}(\mathbf{X}_{\delta \text{ref}}))}{(n-1)\sigma_{\mathbf{X}_\delta} \sigma_{\mathbf{X}_{\delta \text{ref}}}} \tag{2.20}$$

The set of correlations calculated for all values of $\delta$ is often visualised by producing a pseudospectral appearance, on which the correlation coefficient $r$ is projected as a colour, see figure 2.13. In this figure, the correlations of the spectral intensities with the driver peak at $\delta 2.45$ confirms the identity of this peak as 2-oxoglutarate. These peak-like displays are generated from, for example, sample or mean spectra, or by using the covariance between the X and Y data to generate a spectral representation. Alternatively, STOCSY can be displayed in a 2D fashion, analogous to a 2D TOCSY (NMR) spectrum. STOCSY has proven useful for biomarker identification, especially
in metabonomic studies, because the need for time-consuming two-dimensional NMR experiments is reduced. Moreover, for NMR data, STOCSY may provide a larger and more complete number of correlations than obtained by pulse sequence approaches, which are limited by bond connectivities and intra-molecular distances. Exemplar studies using STOCSY included evaluation of analgesic usage in a large epidemiological study and the study of drug metabolism kinetics.

The correlation-based approach of STOCSY not only highlights signals from the same molecule, but also other highly-correlated and anti-correlated signals, which can for example be pathway-based, e.g. connectivities in relation to metabolic effects of a renal toxin. An example of inter-molecular correlation is seen in figure 2.13, where STOCSY driven from a peak of 2-oxoglutarate also highlights the peaks from citrate (δ 2.69 and δ 2.55), both compounds are intermediates in the citric acid cycle. An approach to differentiate between intra- and inter-molecular correlations based on the slope of the regression coefficient, a method named statistical total regression spectroscopy (STORSY), has been investigated and is described in §3.3.6 of this thesis.

Figure 2.13: Statistical total correlation spectroscopy (STOCSY) was performed on the urinary 
JRES data set described in chapter 3. The correlations with the resonance at δ 2.45 (•) are presented as a colour projected onto an expansion of one of the spectra, and the red peak at δ 3.02 (•) would confirm the assignment of 2-oxoglutarate (see also figure 2.12). The molecular structure of 2-oxoglutarate is shown. Slightly lower inter-molecular correlations are seen with the two citrate peaks at δ 2.69 and δ 2.55 (•).

2.3.8 SOM: self-organising map

Self-organising maps (SOMs), initially developed by Kohonen, are a class of unsupervised artificial neural networks that map the high-dimensional input data into a low-dimensional space, in a topology-preserving manner. After training the map, the weight vectors associated with each unit (or neuron) comprising the map should approximate the patterns of the objects that were used to train the map, for example the molecular spectral profiles. The units are typically hexagons connected in a rectangular map shape. Units become more dissimilar to each other as the distance in the map between two units increases, leading to a ‘self-organised’ appearance that allows an instant overview of the data, an example is shown in figure 2.14.
Figure 2.14: Results of a self-organising map based on in vivo $^1$H hepatic spectral NMR data acquired by members of the Metabolic and Molecular Imaging group at the Hammersmith Hospital Campus, Imperial College London, UK. (A) Raw $^1$H spectral data of the 338 volunteers, the two visible peaks are from water (4.7 ppm) and lipids (1.3 ppm). The spectra were coloured with increasing body-mass-index (BMI); generally higher BMI values correspond to high lipid:water ratios. (B) The weight vectors corresponding to each unit of the self-organising map resemble the spectral data that were used to build the map. Higher fat levels are observed in the lower part of the map. (C) The distance map compares the weight vectors of neighbouring units, with similar units in black, and increasing levels of blue for units differing from their surrounding units. The bottom left units are highly dissimilar from the rest of the map and indicate a different class of spectra, as seen in B these have higher lipid and lower water peaks. (D) After training, the data are projected onto the map and the best matching unit is found for each spectrum, based on comparison of the measured spectral profile and the weight vectors shown in B. One unit will then be representing a number of objects, each represented as a pie slice in this hit map. The size of the pie corresponds to the number of objects that it was the best matching unit for. The colouring of each pie slice is based on the BMI value also used in A (higher BMI values are red), and show a correspondence to the organisation of the map based on spectral data: higher BMI values tend to map on units with higher internal hepatic fat levels.
To create a self-organising map, a map of pre-defined size is initialised with weight vectors for every unit, corresponding to intensities for each variable that has to be modelled (comparable to a loading in PCA). These weight vectors can be chosen randomly, but are commonly based on the first two principal components of the data set to decrease computational time and give a deterministic SOM and therefore reproducible result. Subsequently, the map is ‘trained’ to start resembling the topology of the data, allowing for non-linearities to be mapped efficiently. This is done by presenting an item from the training data set (e.g. a spectrum) to the map, and finding the ‘best matching unit’ by evaluation of the, typically Euclidean, distance between the training item \(\mathbf{x}\) and each neuron on the map. This competitive learning step identifies the weight vector of the map with closest resemblance to the training item, and this weight vector \(\mathbf{w}_j(t)\) is then updated, \(\mathbf{w}_j(t + 1)\), to more closely resemble the data, by ‘learning’ from the training item \(\mathbf{x}\). This is done in a topology-preserving manner, whereby neighbouring units also get updated to a degree dependent on their distance on the map to the best matching unit, see equation 2.21.

\[
\mathbf{w}_j(t + 1) = \mathbf{w}_j(t) + \eta(t) N(t, r) [\mathbf{x} - \mathbf{w}_j(t)]
\]  

(2.21)

The learning is governed by the learning rate \(\eta(t)\), which decreases linearly as a function of the time \(t\) during the overall training time \(T\), see equation 2.22; \(\eta(t_0)\) is the initial training rate.

\[
\eta(t) = \eta(t_0) (1 - \frac{t}{T})
\]  

(2.22)

The neighbourhood function \(N(t, r)\), see equation 2.23, determines the degree to which units are updated, and is dependent on the training time \(t\) as well as the distance \(r\) on the map grid to the best matching unit. This is similar to a smoothing kernel, and in this work a Gaussian shape with a neighbourhood radius of \(\sigma(t)\), defining the region of influence of the training sample at time \(t\), was used.

\[
N(t, r) = \exp\left(\frac{r^2}{2\sigma(t)^2}\right)
\]  

(2.23)

The gradually decreasing neighbourhood radius enables a global mapping of the SOM based on the samples, followed by a fine-tuning of the appearance of the map. The updating of units is repeated by subsequently presenting all items in the training data set, and can be performed efficiently in a batch algorithm. This process of presenting the training data set to update the SOM is repeated until convergence. The result of the learning process is that the weight vectors start to represent the data, see figure 2.14 A and B. Similarities, measured as distances, between neighbouring units can be visualised using a distance map (based on U-matrix methods) and can be used to identify clusters and outliers, shown in figure 2.14 C.

Data, either training or newly acquired (test) data, can be presented to the converged SOM and best matching units for each of the items can be calculated and visualised, see figure 2.14 D. The mapping of data to the SOM highlights their visualisation capabilities: SOMs enable the detection of clusters and classes, show relative (dis)similarities of samples and present the responsible weight.
vectors, see figure 2.14. SOMs allow non-linear mapping of data in low dimensions and are relatively insensitive to outliers, as they will occupy their own region in the map.

The use of SOM in metabonomic studies was nicely illustrated with a study of $^1$H NMR plasma lipoprotein data, where the unsupervised SOM approach was able to characterise the lipoprotein subclass profiles in a clinically relevant way. SOM maps have also been used for variable importance and selection.
Chapter 3

The use of 1D projections of $J$-resolved NMR spectra in metabonomics

3.1 Aims and objectives

Spectroscopic profiling of biological samples is an integral part of metabolically-driven top-down systems biology and can be used for identifying biomarkers of toxicity and disease. However, optimal biomarker information recovery and resonance assignment still pose significant challenges in NMR-based complex mixture analysis. This chapter describes a method for reducing peak overlap, which is achieved when projecting two-dimensional $J$-resolved (JRES) NMR spectra, and is based on the published work included in appendix B. This is done by means of:

1. Evaluating different processing steps to obtain high quality full-resolution JRES spectral projections;

2. Demonstrating the necessity of peak alignment for modelling of full-resolution JRES projections;

3. Investigation of the application of statistical total correlation spectroscopy to JRES data;


5. Comparing the recoverable information content in full-resolution $^1$H JRES projections with conventional one-dimensional spectroscopy.
3.2 Introduction

$^1$H NMR based metabolic profiling suffers from the considerable peak overlap in biofluid spectra, especially when chemical shift variation occurs or the data are binned into discrete chemical shift regions for intensity calculations (bucketing), and this often inhibits complete interpretation of the spectra and accompanying pattern recognition models.$^{90,116,117}$

One specific problem for NMR spectroscopy is that many metabolites give rise to a considerable number of spin-coupled multiplets over a range of chemical shifts, see also figure 2.3. This has two consequences: impaired peak dispersion due to spectral overlap and over-representation of the metabolite in any statistical classification exercise, as illustrated in figure 3.1. One method for reducing peak overlap is to disperse signals into a second dimension. $J$-resolved ($J$RES) NMR spectroscopy,$^{71}$ is one of the simplest two-dimensional NMR experiments with which 2D spectra can be rapidly obtained. This can be more efficient than physical separation of the compounds (for example using a chromatography step before NMR analysis)$^{118}$ or the use of lower sensitivity 2D NMR experiments (such as homo- and heteronuclear correlation spectroscopy).$^{119}$

The JRES experiment is formed of an array of spin–echo pulse elements, in which an incremented delay period is used to define a second frequency dimension. After suitable data processing,$^{40}$ the chemical shift and $J$-coupling information are resolved on two orthogonal axes, thereby increasing signal dispersion. The 2D JRES spectrum can be projected onto the chemical shift axis,$^{120,121}$ effectively yielding a $^1$H broadband decoupled proton spectrum consisting exclusively of ‘singlets’, as will be discussed in figure 3.3, which can then be subjected to pattern recognition analysis. Potentially, this approach can reduce the overlap and over-representation problems encountered in biological samples.

![Figure 3.1: PEAK OVERLAP AND OVER-REPRESENTATION](image)

Figure 3.1: The presence of multiplets in NMR spectra has a major effect on the complexity as well as the information content of the spectra. Although the multiplicity of peaks contains vital structural information, it also increases the level of peak overlap and gives a metabolite peak a wider base in a spectroscopic data table, causing highly-split multiplets to be ‘over-represented’.
The value of JRES NMR spectra in metabolic profiling studies of biofluids was demonstrated two decades ago\textsuperscript{122–125} and JRES spectra have since then been routinely acquired\textsuperscript{40, 126–128} including, inter alia, studies of human urine,\textsuperscript{122} plasma\textsuperscript{48, 123} and cerebrospinal fluid.\textsuperscript{124} JRES NMR spectroscopy has also been utilised in studies of the components of plants,\textsuperscript{129} fish\textsuperscript{130} and beer,\textsuperscript{131} in kinetic drug metabolism studies,\textsuperscript{88} and combined with magic-angle spinning NMR methods on tissues\textsuperscript{132} and even in vivo.\textsuperscript{133} The most readily apparent advantages of JRES NMR spectroscopy stem from improved dispersion, through reduced peak congestion. The refocusing character of the pulse sequence also results in $T_2$-editing, which attenuates broad macromolecular signals as well as resonances from motionally constrained compounds and species with chemical exchange processes at intermediate rates on the NMR time-scale. The spin–spin coupling information is retained in the 2D spectrum and can aid peak identification. JRES has proven to be particularly useful in metabolic profiling studies\textsuperscript{122, 128, 134} because many of the nuclear spin systems of such small molecules exhibit first-order NMR spectra; hence, artefactual peaks associated with second-order effects are rare. However, there are limitations caused by the non-standard line-shape and $T_2$-editing in JRES NMR spectra and the effects on metabolite quantitation will be discussed in §3.4.6.\textsuperscript{135} Other studies have investigated the effect of JRES NMR data acquisition parameters applied to metabonomics.\textsuperscript{40, 136} Improved spectral clustering of data in PCA models built from JRES data versus conventional one-dimensional (1D and Carr–Purcell–Meiboom–Gill (CPMG)) spectra has been reported.\textsuperscript{127}

To date, most studies evaluating JRES spectra have used binned (or bucketed) data based on a low-resolution representation of the peak intensities rather than the full-resolution spectral data.\textsuperscript{127–130, 132} The limitations of binning spectral data on data interpretation and information recovery are well-documented\textsuperscript{80, 137} and are illustrated in figure 3.2: an NMR peak can be divided across different bins, or a bin may contain a number of peaks. Therefore, statistical effects become

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure32.png}
\caption{Full datapoint resolution spectra (red, 44001 points: $\delta$ 0.1–4.5, 6.0–10) will allow for better biomarker recovery than the binned equivalent, shown as grey bars. For clarity, only a part of the full spectral region is shown.}
\end{figure}
less pronounced and small peaks may become completely obscured. The main reason for binning data is to minimise the effects of chemical shift variability, which is necessary when statistical techniques are applied. The development of alignment algorithms for NMR data have lessened the need for this compromise, because they resolve all peak positional variation whilst retaining the high resolution of the data.\textsuperscript{116, 117, 138, 139}

Here, the JRES NMR experiment will be comprehensively re-evaluated with respect to its use in mathematical modelling of biological spectra. Full-resolution one-dimensional projections are used, and it is anticipated that the increased level of apparent resolution will aid biomarker identification, especially for complicated, highly overlapped spectral regions. The processing steps required for routine and robust use of JRES NMR spectra in metabolomics will be considered, including peak alignment (following a recommendation from Dr. M. Coen, Imperial College London, UK). The qualitative and quantitative aspects in conjunction with correlation spectroscopy approaches will be assessed, and a method based on regression for peak assignment based on JRES projections will be evaluated. Finally, the degree of information recovery compared to conventionally acquired one-dimensional \textsuperscript{1}H NMR spectra (1D and CPMG) will be considered. As an example where NMR peak overlap and over-representation are acute, spectra acquired from plasma and urine samples from a galactosamine toxicity study in the rat were used.\textsuperscript{140} These contain exogenous and endogenous metabolites that are spectrally high-represented, caused by the presence of many multiplets, thus obscuring a considerable portion of the spectral range.

3.3 Materials and methods

3.3.1 Sample preparation

Samples were collected and analysed previously as part of the COMET-2 project\textsuperscript{141, 142} and were stored at -80 °C. Rats were administered 10 mL/kg vehicle (0.9% saline, \(n = 8\)) or galN (galactosamine hydrochloride dissolved in 0.9% saline to give a free base concentration of 41.5 mg/mL, \(n = 8\)).\textsuperscript{140} Urine specimens were collected for 24 hr after administration of galactosamine or dosing vehicle, and plasma was obtained upon sacrifice, 24 hr after administration, as described previously.\textsuperscript{140} Urine samples were thawed, vortexed and allowed to stand for 10 minutes prior to mixing aliquots (400 \(\mu\)L) with phosphate buffer (200 \(\mu\)L, 0.2 M containing 10% deuterium oxide (D\(_2\)O), 3 mM TSP (3-(trimethylsilyl)-[2,2,3,3-\textsuperscript{2}H\(_4\)]-propionic acid sodium salt) and 3 mM sodium azide) and centrifuged at 13000 rpm for 10 minutes. Samples (550 \(\mu\)L) were placed in 5 mm outer diameter NMR tubes. Stored plasma samples, previously prepared and in 5 mm outer diameter NMR tubes\textsuperscript{140} were thawed and used for analysis.
3.3.2 $^1$H NMR spectroscopy

Proton NMR spectra were acquired on a Bruker Avance-800 spectrometer operating at an 800.32 MHz $^1$H frequency and a temperature of 27 °C. D$_2$O provided a field frequency lock and TSP provided a chemical shift reference for the urine samples. Urine and plasma $^1$H NMR spectra were acquired with water peak pre-saturation using the pulse sequence ($d_1$–90$\_x$–4 $\mu$s–90$\_x$–$t_m$–90$\_x$–acquire FID). For each sample, 64 transients were collected into 64k data points using a spectral width of 16025 Hz, with a relaxation delay ($d_1$) of 2 s, an acquisition time of 2.04 s and a mixing time ($t_m$) of 100 ms. The water resonance was selectively irradiated during $d_1$ and $t_m$. CPMG $^1$H NMR spectra of urine and plasma, acquired with the pulse sequence ($d_1$–90$\_x$–$\tau$–180$\_y$–$\tau$–acquire FID), with a spin–spin relaxation delay 2n$\tau$ of 102.4 ms ($\tau$ = 0.4 ms, n = 128) were collected with 64 transients and a $d_1$ of 2 s into 64k data points with a spectral width of 16025 Hz and an acquisition time of 2.04 s. The water peak was irradiated during $d_1$. The 2D JRES urine and plasma $^1$H NMR spectra were acquired with the pulse sequence, ($d_1$–90$\_y$–$\tau$–180–$\tau$–acquire FID, with suppression of the water resonance during $d_1$) into 64k data points in F2 with a $d_1$ of 2 s and 8 transients using 32 increments of $\tau$; the spectral widths in F2 and F1 were 16025 and 50 Hz, respectively.

Initially, the magnet could not be shimmed adequately for the urinary data, which could be reduced by equilibrating the sample in the NMR machine for about 30 minutes before acquisition. Some samples showed bad shimming and peak splitting, especially evident in the JRES projections, which was resolved by shimming in the ‘GS mode’ of the JRES experiment. The replicated spectra were plotted and the highest quality spectra were selected. Two of the urine samples were reprepared (with an extra centrifuge step and careful pipetting, resulting in a dilution). The repreparation, however, did not reduce the need for the sample to settle. The main result of repreparation is dilution of the sample, which can easily be compensated for by using correct normalisation methods. All samples have gone through a number (3–5) of freeze-thaw cycles.

3.3.3 Data processing

A line broadening function of 1 Hz and one level of zero filling were applied to all 1D and CPMG spectra prior to Fourier transformation (FT). Automated phase and baseline correction and referencing to TSP (δ 0.000) or α-glucose (δ 5.233) resonances for urine and plasma, respectively, were performed using in-house software (NMRproc v0.3, Drs. T.M.D. Ebbels and H.C. Keun, private communication). JRES data were zero filled to 128k data points in F2, and up to 256 increments in F1. Apodisation of the JRES data was carried out using an unshifted sine-bell function in both dimensions prior to FT. As is necessary, the absolute value spectra were calculated. Data were tilted, symmetrised with respect to the horizontal through the centre of F1, and skyline or sum projected as indicated. The JRES projections were referenced and baseline corrected as above.

All spectra were imported to MATLAB (R2008a, The Mathworks, MA, USA) and data were interpolated to form a vector running from δ -1.0 to δ 10.0 in steps of δ 0.00025. This maintains the
resolution of the spectral data while allowing for increased accuracy in chemical shift referencing. Following removal of the TSP, water and urea regions, the data between $\delta$ 0.1–4.5 and $\delta$ 6.0–10.0 were used for further analysis (resulting in 44001 data points per spectrum) and subjected to probabilistic quotient normalisation, see §2.3.1.\textsuperscript{82}

There were six data sets in total: urine and plasma data for each of the three types of NMR experiment (1D, CPMG and JRES). To investigate the effect of peak alignment, recursive segment-wise peak alignment\textsuperscript{116} was performed, where the reference spectrum for each set of 16 samples was chosen to be the one with maximum correlation with the other spectra for that particular experiment and biofluid type.

### 3.3.4 Signal-to-noise ratio and line width analyses

The signal-to-noise ratio ($S/N$) for different processing options of the JRES projections was calculated as the mean ratio across the 16 samples of a given peak maximum with respect to the standard deviation of noise ($\delta$ 0.2–0.5 for JRES (1201 data points), $\delta$ 9.8–10.0 for CPMG (800 data points)). The $S/N$ was calculated for the alanine methyl signal at $\delta$ 1.47 and the lactate methyl signal at $\delta$ 1.33 in both urine and plasma and the TSP singlet at $\delta$ 0.00 and the succinate singlet at $\delta$ 2.41 in urine; for the CPMG and untilted JRES spectra, the maximum value of the lactate and alanine doublet signals was used. It should be noted that the term noise is not absolutely correct in JRES spectra since it is not truly random having no negative components in the absolute value mode, prior to any baseline correction.

The line widths of the JRES projections were approximated as the median of the full width at half height. The line widths were calculated for the lactate methyl singlet and the alanine methyl singlet in plasma and urine in addition to the TSP and succinate singlets in urine; for the untilted data, the higher frequency component of the lactate and alanine doublets was used.

### 3.3.5 Statistical and chemometric analyses

Various multivariate modelling methods and univariate correlation approaches were employed:

- PCA\textsuperscript{86} models (see §2.3.3) were generated using mean-centred spectral data to evaluate the data before and after alignment. The direction of differentiation between the scores of the two groups was used to evaluate the PC at which separation occurred.

- To evaluate if alignment removed orthogonal variation, OPLS\textsuperscript{93} models (see §2.3.5) were constructed with in-house software on mean-centred data scaled to unit variance, with a dummy vector representing the two classes (encoded as $0 =$ control, $1 =$ dosed).

- Pearson’s correlation coefficient ($r$) was calculated between spectra to estimate the similarity between different samples and experiments. The correlation $r(s_1, s_2)$ is calculated based on all data points in two spectra $s_1$ and $s_2$. 

55
• Statistical total correlation spectroscopy (STOCSY)\(^6\) displays the Pearson correlation between the signal intensities \(X_\delta\) at each resonance frequency \(\delta\) and a reference variable of interest, see §2.3.7. The used reference values were: the alanine CH\(_3\) resonance (for doublets the high frequency peak was used) and the creatinine CH\(_3\) resonance at \(\delta\) 3.05 (for figure 3.12), or a dummy vector representing the class (in figures 3.6 and 3.11).

### 3.3.6 STORSY: statistical total regression spectroscopy

Peak intensities from the same molecule should, in principle, approximate the proportional ratio of the contributing protons. Whether this proportionality can be used for improved correlation analysis is investigated by using a regression of two signal intensities from the same molecule across a series of samples, which is a method coined statistical total regression spectroscopy (STORSY). It is hypothesised that calculating a regression slope of two resonances, similar to the correlation in STOCSY, would aid peak identification and could differentiate significant correlations into inter- and intra-molecular correlations.\(^{143}\) For example, if STORSY is performed for the CH-proton of lactate, one would expect the methyl group to have a relative regression slope of 3, whereas other pathway or related high correlations may have regression values high or low enough to indicate these cannot arise from the same molecule; e.g. a peak ratio exceeding 10 is unlikely to arise from signals within the same metabolite.

The most straightforward way in which STORSY could be implemented in standard metabonomic research is by performing this calculation on \(J\)RES projections, since that eliminates the need to take multiplicity patterns into account, such as present in typical one-dimensional proton spectra. Theoretically, the sum projection retains the quantitative information whereas the skyline projection does not. The projections are performed along \(F_1\), which is the \(J\)-coupling direction (the vertical direction in figure 3.3). In the skyline projection, the maximum signal along \(F_1\) is retained and will be a fraction of the original intensity, dependent on the splitting pattern, hence peak integrals of the skyline projection will not correspond to the original proton ratios.\(^{121}\) With a CH\(_3\) integral set to 3, the ratio of the lactate peaks for the sum projection, which sums all signals along \(F_1\), should remain 1:3, whilst the ratio would be 0.75:3 for the skyline projection (the relative intensities for the CH quartet at \(\delta\) 4.11 are 0.125:0.375:0.375:0.125, and for the CH\(_3\) doublet at \(\delta\) 1.33 these are 1.5:1.5, with the retained skyline intensity displayed bold, see also figure 2.4).

Inclusion of an intercept in the regression equation can be used to allow for baseline offset and overlapping peaks, although both are diminished in \(J\)RES spectra. A verification of the proportions is to perform a regression from peak B to peak A: the slope should be the inverse of the slope from peak A to peak B. Before interpretation of the STORSY values, it should be verified that the results are not affected by outliers in the peak intensity data, as these can have a large leverage effect on the regression slope value.
**STORSY analysis of glucuronidation time-series**

To evaluate STORSY, data from a previous study were used\(^8\) where the data were reprocessed with and without symmetrisation as indicated and zero filled to 128\(^k\) in \(F_2\) and 256 in \(F_1\). Data were sum projected, baseline corrected (zero-order polynomial correction using Topspin, Bruker Biospin) and referenced to TSP (by \(\text{NMRproc v0.3}\), Drs. T.M.D. Ebbels and H.C. Keun) and interpolated from \(\delta\) -0.1 to \(\delta\) 9.0 in steps of \(\delta\) 0.00025 (36401 data points). As all data were acquired from one sample, normalisation was performed using TSP area (\(\delta\) -0.02 to \(\delta\) 0.02) and peak positional variability is minimal, alleviating the need for alignment.

**STORSY for galactosamine toxicity study**

Correlation and regression analyses were performed on the peak maxima (peak height) or peak integrals (peak area) of the CH and CH\(_3\) peaks of lactate. Pearson correlations were calculated between the CH and CH\(_3\) data, and a standard linear regression for the CH versus CH\(_3\) data was calculated with and without intercept, i.e. \(y = bx\) and \(y = bx + a\).

### 3.4 Results and discussion

#### 3.4.1 Projection of JRES spectra

JRES spectra may be projected in the chemical shift dimension after tilting the 2D spectrum, to effectively produce a homonuclear decoupled 1D proton spectrum. The effect of this ‘collapse’ of multiplets is demonstrated in figure 3.3: the multiple, highly overlapped glucose resonances can now be unambiguously assigned in the JRES spectrum from their chemical shift and coupling pattern information. Two types of projection are routinely employed: the sum projection, summing all signals along \(F_1\), and the skyline projection, retaining the highest signal along \(F_1\).\(^{120,121}\) Generally, the skyline projection results in signals with a better signal-to-noise ratio (S/N).\(^{136}\) This is illustrated in figure 3.3, where a representative ‘control’ 2D spectrum from the galactosamine toxicity study and associated projections are shown, scaled to equivalent noise levels.

In table 3.1, the mean S/N is displayed for the various processing methods as applied to both urine and plasma spectral peaks, demonstrating that for both symmetrised and unsymmetrised projections, the S/N is better for skyline than sum projections. Because most increments of the JRES spectrum do not contain signal, the commonly observed S/N increase with the square root of summed increments is not achieved.\(^{136}\) Naturally, this S/N advantage of skyline projection is more pronounced for singlets than multiplets. For example, in figure 3.3 the highly split H5 \(\beta\)-glucose resonance at \(\delta\) 3.464 (ddd) shows similar sum and skyline projection peak intensities after scaling to constant noise levels. Moreover, it is expected that the level of noise is distributed normally, and so taking the maximum signal along \(F_1\), as in a skyline projection, could effectively decrease the noise level, if the noise maxima are similar in value;\(^{136}\) this results in a slightly increased height of
Figure 3.3: Example of a symmetrised 2D JRES $^1$H NMR spectrum acquired at 800 MHz from plasma of a control rat in the galactosamine study. Sum (blue) and skyline (red) projections without symmetrisation are shown, scaled to identical noise levels (measured at $\delta$ 0.2–0.5). For comparison, the conventional 1D $^1$H NMR spectrum is shown (black). The use of increased dispersion of the JRES experiment and the incorporation of multiplicity information in peak identification is demonstrated with assignment of all individual glucose resonances of both the $\alpha$- and the $\beta$-anomers (structures are shown on the bottom). $^{48}$

the baseline, but with a lower standard deviation of the noise level, hence resulting in an apparent increase in $S/N$.

It has been reported that the skyline projection gives the best spectral quality based on binned data combined with a sine-bell-exponential apodisation function in the direct dimension. $^{136}$ Most applications of JRES projections in metabonomics have used the skyline projection, with only a few exceptions. $^{126, 142}$ Although the $S/N$ is lower for all of the JRES projections compared to the CPMG results with the current acquisition parameters (see table 3.1), this can be remedied by adjustment of the spectral acquisition parameters such as an increased number of scans or increments.
Table 3.1: The mean signal-to-noise ratios ($\times 10^3$, arbitrary units) of various spectral data types.

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Symmetrisation</th>
<th>Projection</th>
<th>Urinary Alanine</th>
<th>Plasma Alanine</th>
<th>Urinary Lactate</th>
<th>Plasma Lactate</th>
<th>Urinary Succinate</th>
<th>TSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>skyline</td>
<td>0.277</td>
<td>0.413</td>
<td>0.733</td>
<td>2.026</td>
<td>4.958</td>
<td>18.456</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>sum</td>
<td>0.074</td>
<td>0.136</td>
<td>0.215</td>
<td>0.630</td>
<td>0.711</td>
<td>2.662</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>skyline</td>
<td>0.262</td>
<td>0.485</td>
<td>0.702</td>
<td>2.316</td>
<td>5.860</td>
<td>21.814</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>sum</td>
<td>0.077</td>
<td>0.141</td>
<td>0.218</td>
<td>0.639</td>
<td>0.728</td>
<td>2.794</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>skyline</td>
<td>0.198</td>
<td>0.358</td>
<td>0.636</td>
<td>1.752</td>
<td>4.324</td>
<td>16.057</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>sum</td>
<td>0.039</td>
<td>0.075</td>
<td>0.127</td>
<td>0.378</td>
<td>0.887</td>
<td>3.334</td>
</tr>
<tr>
<td>CPMG</td>
<td></td>
<td></td>
<td>0.616</td>
<td>1.271</td>
<td>1.347</td>
<td>4.921</td>
<td>5.381</td>
<td>19.974</td>
</tr>
</tbody>
</table>

a Results for CPMG and untilted JRES NMR spectra are based on values from the higher peak in each listed doublet. The standard 1D NMR spectrum values are not given, because the peak height and noise measurements for plasma are confounded, due to the presence of macromolecular signals.

b Plasma lactate peaks in the CPMG spectra were overlapped with a broad peak around $\delta$ 1.28.

It is shown (vide infra) that the disadvantage of the lower overall S/N of JRES spectra is outweighed by a considerably improved apparent resolution due to reduced peak overlap in the projections.

### 3.4.2 Symmetrisation of JRES spectra

The full widths at half height of specified peaks are given in table 3.2 and these show little difference between symmetrised skyline and sum JRES projection peaks. Symmetrisation of the JRES data results in narrower peaks of the subsequent projections, a phenomenon which is directly related to the symmetrisation procedure. Without symmetrisation, sum projections resulted in broader peaks than skyline projections, see table 3.2, and this is in agreement with the superior resolution of skyline projections found previously. Symmetrisation can either create or eliminate spectral artefacts; however neither effect was noticed here. The S/N values for symmetrised and unsymmetrised spectra were comparable, see table 3.1.

Table 3.2: The median line width (Hz) measured as the full width at half height of various processed projections of JRES NMR spectra.

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Symmetrisation</th>
<th>Projection</th>
<th>Plasma Alanine</th>
<th>Plasma Lactate</th>
<th>Urinary Succinate</th>
<th>TSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>skyline</td>
<td>3.0</td>
<td>3.0</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>sum</td>
<td>3.6</td>
<td>3.6</td>
<td>3.4</td>
<td>3.2</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>skyline</td>
<td>2.0</td>
<td>1.7</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>sum</td>
<td>1.8</td>
<td>1.8</td>
<td>1.3</td>
<td>1.2</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>skyline</td>
<td>2.0</td>
<td>1.8</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>sum</td>
<td>1.9</td>
<td>1.8</td>
<td>1.4</td>
<td>1.2</td>
</tr>
</tbody>
</table>
3.4.3 Peak alignment of full-resolution JRES projections

Urinary $^1$H NMR spectra show more peak positional variation than plasma spectra, which is caused by, for example, greater variability in pH, metal and ion concentrations; metabolite–protein interactions and chemical exchange processes can also interfere. JRES data are more sensitive to misalignment effects compared to 1D and CPMG spectra because the reduction in ‘chemical shift bandwidth’ for each resonance (multiplets are collapsed into singlets and are even more narrow if symmetrised, see figure 3.1) causes small variations in chemical shift to manifest as marked misalignments in JRES projections. Figure 3.4 shows JRES projections of $^1$H NMR spectra of urine from the dosed and control groups of the galactosamine toxicity study before and after recursive segment-wise peak alignment,\(^{116}\) with expansions demonstrating the substantial improvement by alignment.

![Graph showing unaligned and aligned JRES projections of urine samples](image)

Figure 3.4: Symmetrised skyline projections of 800 MHz $^1$H JRES spectra of urine samples from the galactosamine toxicity study before (top) and after (bottom) alignment.\(^ {116}\) Spectra from dosed animals are coloured red and those from control rat spectra are shown in blue; expansions are shown for δ 2.67–2.74 and δ 3.42–3.48. Asterisks (*) denote that the creatinine δ 4.06 and H2 β-galactosamine δ 3.14 resonances are not aligned perfectly with the used algorithm and settings.
Previous analyses of JRES data of metabolite mixtures did not require peak alignment, because binned integral data were used. However, it should be appreciated that peak alignment is preferable to binning, as multiple signals may become concatenated, or individual signals split in the latter procedure (see also figure 3.2). Moreover, the use of bins precludes direct interpretation of multivariate models. Previously, some major peak shifts had to be accounted for by combining multiple bins, indicating that even binning itself may not compensate for all peak shifts (e.g. citrate in urine). Of course, it would be possible in theory to stabilise the pH of all samples, but this is prohibitive for large sample sets such as from biobank or metabolome-wide association studies. Moreover, stabilisation of urine samples to neutral pH is only temporary since precipitation within the sample tends to cause a shift in the pH, and samples stabilised to acidic pH will result in chemical shifts that are different from those in most commonly used databases.

Therefore, it appears appropriate to perform peak alignment, which proved to be uncomplicated for the singlets in JRES projections (in figure 3.4 most peaks are well aligned, and possibly other algorithms will give equally good results). A similar situation occurs in the alignment of $^{13}$C NMR spectra, where peak shift correction is more straightforward than for overlapping $^1$H NMR spectra, because $^{13}$C shifts are less affected by pH changes and there are fewer or no complicated multiplet resonances. Improvements in applicable alignment algorithms can be tailor-made for the specific data structure found in JRES spectra (e.g. the absence of multiplets) to increase effectiveness and further decrease the computational time.

3.4.4 Necessity of alignment of JRES projections for statistical analysis

Using the unsupervised statistical method principal component analysis (PCA), it was clear that discrimination of galactosamine treated and control groups for urinary JRES projections in the first principal component (PC) is conditional on proper peak alignment, see figure 3.5 and table 3.2. Alignment significantly improved the PCA model, with a much higher proportion of the variance explained in PC 1 (76.4% versus 33.2% for the symmetrised skyline projection) and the separation of classes on the basis of toxicity being promoted to higher principal components (compare figure 3.5 A and B).

It is not only important to align the data to create improved models, but additionally the interpretation of the data is greatly increased upon alignment. Before alignment, the resultant dispersive line shape of the loadings due to peak position variability makes it ambiguous as to whether a certain peak is actually contributing positively or negatively to a PCA loading. Peak positional variation was indicated for virtually all resonances, e.g. taurine, dimethylamine, succinate, citrate, 2-oxoglutarate and creatinine (see figure 3.5 C). The galactosamine and $N$-acetylglucosamine resonances are unaligned to such a high degree that these variables do not contribute significantly to the variance modelled in the first principal component, much in contrast to the aligned model.
Figure 3.5: (A, B) Scores of the PCA model calculated for the mean-centred symmetrised skyline projections of jres urine spectra before and after alignment,\textsuperscript{116} for galactosamine-dosed animals (red squares) and vehicle-dosed controls (blue circles). PC: principal component and associated percentage of explained variance. (C, D) The corresponding PCA loadings displayed for the region $\delta$ 2.0–4.3 clearly demonstrate the disappearance of dispersive shapes upon peak alignment for, e.g., taurine (1), dimethylamine (2), succinate (3), citrate (4), 2-oxoglutarate (5) and creatinine (6), while resonances of galactosamine (7) and N-acetylglucosamine (8, a galactosamine metabolite) become a dominant source of variance in the first PC.

(figure 3.5 D). For plasma data, only small improvements in loading interpretation are observed after alignment, and a good separation between the two classes is present in both aligned and unaligned data, this is due to much lower positional variation in the original data.

The parameters for each PCA model of the differently processed urinary data sets are given in table 3.3. Symmetrised skyline projections separated in PC 3, whereas the unsymmetrised skyline projections separated in PC 2. This effect of symmetrisation is most likely related to peak width, as established in table 3.2, which reduces the number of variables representing a proton resonance. This results in an even smaller span of the data peak and thus if these data are not well-aligned,
Table 3.3: Principal component analysis results before and after spectral peak alignment, models were based on the mean-centred NMR urine data. Principal component models were evaluated in terms of variance explained (Var. exp.), which is the fraction of variance in the total data set contained by the indicated principal component (PC), and the PC at which separation between the galactosamine dosed and control classes was observed.

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Symmetrisation</th>
<th>Projection</th>
<th>Var. exp. PC 1</th>
<th>Var. exp. PC 2</th>
<th>PC of separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNALIGNED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>skyline</td>
<td>0.37</td>
<td>0.21</td>
<td>2</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>sum</td>
<td>0.40</td>
<td>0.24</td>
<td>1</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>skyline</td>
<td>0.33</td>
<td>0.17</td>
<td>3</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>sum</td>
<td>0.31</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>skyline</td>
<td>0.29</td>
<td>0.21</td>
<td>2</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>sum</td>
<td>0.26</td>
<td>0.26</td>
<td>1 and 2</td>
</tr>
<tr>
<td>1D</td>
<td></td>
<td></td>
<td>0.61</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td>CPMG</td>
<td></td>
<td></td>
<td>0.56</td>
<td>0.15</td>
<td>1</td>
</tr>
</tbody>
</table>

| ALIGNED |                |            |                |                |                  |
| yes  | no             | skyline    | 0.80           | 0.08           | 1                |
| yes  | no             | sum        | 0.79           | 0.07           | 1                |
| yes  | yes            | skyline    | 0.76           | 0.07           | 1                |
| yes  | yes            | sum        | 0.82           | 0.07           | 1                |
| no   | no             | skyline    | 0.76           | 0.07           | 1                |
| no   | no             | sum        | 0.79           | 0.06           | 1                |
| 1D   |                |            | 0.90           | 0.03           | 1                |
| CPMG |                |            | 0.88           | 0.04           | 1                |

the classes are not separated in the PCA model. The models based on sum projections separate in PC 1, possibly because highly split signals, for example the galactosamine multiplet peaks, manifest higher intensity and hence greater importance in the sum projection than the corresponding peak in the skyline projection. The PCA models of 1D and CPMG data without alignment did split in PC 1, which could be explained by the ‘over-representation’ of relevant peaks. The resonances of interest are highly split multiplets and therefore dominate the variance in the PCA model, even when not perfectly aligned.

When the JRES projections were obtained without tilting these data, the multiplicity of the peaks remains, but the separation is not in PC 1 as might be expected from the larger variable span of multiplets, which should be less susceptible to peak shift influences. Hence, it should be appreciated that added to the multiplicity collapse, the reduced density of signals due to T2-editing, the sine-bell processing and the projection step could all cause the JRES projection data to be inherently more variable than the 1D and CPMG spectra.
Orthogonal partial least squares discriminant analysis (OPLS-DA) models were made with 1 predictive component for the different processed data sets. Data were mean-centred and scaled to unit variance. The mean sensitivity of the prediction for the two classes is displayed; the values marked with an asterisk are the maximum obtained (evaluated up to 5 orthogonal components, \( oc \)), and for these instances, no perfect separation was achieved. The \( q^2 \) values are given for the model with no orthogonal components. The values were calculated with 100 Monte Carlo class-balanced cross-validation rounds, where 75% of the data were used to calculate the model.

### UNALIGNED

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Symmetrisation</th>
<th>Projection</th>
<th>Sensitivity 0 ( oc )</th>
<th>Sensitivity 1 ( oc )</th>
<th>( q^2 ) 0 ( oc )</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>skyline</td>
<td>0.96</td>
<td>1</td>
<td>0.7352</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>sum</td>
<td>0.9925</td>
<td>1</td>
<td>0.7495</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>skyline</td>
<td>0.9575</td>
<td>0.9925*</td>
<td>0.7057</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>sum</td>
<td>0.9950</td>
<td>1</td>
<td>0.7424</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>skyline</td>
<td>0.9500</td>
<td>0.9875*</td>
<td>0.7002</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>sum</td>
<td>0.9600</td>
<td>1</td>
<td>0.7137</td>
</tr>
<tr>
<td>1D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8830</td>
</tr>
<tr>
<td>CPMG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8984</td>
</tr>
</tbody>
</table>

### ALIGNED

<table>
<thead>
<tr>
<th>Tilt</th>
<th>Symmetrisation</th>
<th>Projection</th>
<th>Sensitivity 0 ( oc )</th>
<th>Sensitivity 1 ( oc )</th>
<th>( q^2 ) 0 ( oc )</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>skyline</td>
<td>1</td>
<td>1</td>
<td>0.7621</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>sum</td>
<td>1</td>
<td>1</td>
<td>0.8093</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>skyline</td>
<td>1</td>
<td>1</td>
<td>0.7945</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>sum</td>
<td>1</td>
<td>1</td>
<td>0.7818</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>skyline</td>
<td>1</td>
<td>1</td>
<td>0.7895</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>sum</td>
<td>1</td>
<td>1</td>
<td>0.7443</td>
</tr>
<tr>
<td>1D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9194</td>
</tr>
<tr>
<td>CPMG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9355</td>
</tr>
</tbody>
</table>

Orthogonal partial least squares discriminant analysis (OPLS-DA, see §2.3.5) of the urinary JRES projection data before and after alignment confirmed the findings from PCA: before alignment orthogonal components are necessary to achieve optimal models in terms of the mean sensitivity of class prediction (see equation 2.18), whereas after alignment no orthogonal components are necessary, see table 3.4. Evaluation of the loadings confirmed the dispersive components are modelled as orthogonal variation. From this table, it is also clear that the goodness of prediction for the first component, \( q^2_{0} \), is higher for the aligned models. Interestingly, OPLS-DA models of the plasma data actually show better prediction performance for the unaligned spectra than for the aligned spectra, which is possibly caused by the creation of minor artefacts with the current alignment protocol, and the high quality of the original data.

In conclusion, the data type (multiplets or singlets) and processing (skyline or sum projection, usage of tilt and symmetrisation) will greatly influence the resultant pattern recognition model. However, the variation induced by a combined effect of peak shifts and peak width differences,
Alignment can cause some minor artefacts, see e.g. figure 3.4 and for plasma data, alignment is less necessary: OPLS models showed good separation on unaligned data and PCA model scores and loadings were not largely improved with alignment either. Therefore, it is not clear if the advantages of alignment outweigh the disadvantages for the analysis of plasma data.

### 3.4.5 Improved STOCSY results by use of aligned JRES projections

Information retrieval from correlation analyses such as statistical total correlation spectroscopy (STOCSY)\(^\text{67}\) is also improved after peak alignment. STOCSY plots corresponding to the data shown in figure 3.4 are shown in figure 3.6, in which peak height represents the covariance of the variable with the class dummy vector and the colour represents the correlation with this vector. The severity

![STOCSY plots](image)

Figure 3.6: STOCSY plots based on urine 800 MHz \(^1\)H JRES symmetrised skyline projections obtained before and after peak alignment, as shown in figure 3.4. The correlation (colour, \(r^2\)) and covariance (peak height) of the JRES projections with respect to a class dummy vector are displayed; peaks pointing upward covary positively with the toxic insult. Upon alignment, correlations are more pronounced and peaks lose the dispersive shape that is characteristic of unaligned data, improving interpretability. Strong J-coupling spectral artefacts are seen in the original spectra between the signals of galactosamine resonances at \(\delta \approx 3.8\) and are also visible here. Legend: DMA, dimethylamine; 2-OG, 2-oxoglutarate; and \(\alpha/\beta\)G, \(\alpha/\beta\)-galactosamine.
of the toxic insult, and inter-individual differences in response\textsuperscript{31,140} cause a broad concentration range for certain metabolites in the different samples, increasing the reliability of the obtained correlation coefficients. Spectral alignment removes the dispersive shapes in the STOCSY plot and reveals correlations that were previously unobserved, for example those of the galactosamine resonances.

### 3.4.6 STORSY analysis of JRES projections

**STORSY analysis of glucuronidation time-series**

The proposed STORSY method (see §3.3.6) was evaluated in a straightforward data set: the regression slopes were calculated for the different molecular species present in a sample displaying acyl glucuronide intra-molecular transacetylation reactions\textsuperscript{88}. Many carboxylate drugs are metabolised to ester glucuronides by conjugation with glucuronic acid. These $1\beta-O$-acyl glucuronides (e.g. see figure 3.7) can be reactive and can isomerise to the $2\beta-3\alpha$ and $4\beta-O$-acyl forms with mutarotation to the corresponding $\alpha$ isomers; the $1\alpha$ isomer is formed by a back reaction from the $2\alpha$ isomer. All isomers can also hydrolyse to free aglycone and glucuronic acid. Reactions of various aglycones have been studied including the glucuronides of $(R)\alpha$- and $(S)\alpha$-methyl phenylacetate acid\textsuperscript{88,146,147}. In this data set, the internal acyl migration reactions of $1\beta-O$-acyl glucuronide were monitored by NMR acquisitions at regular intervals, resulting in spectra with different concentrations of the isomers as a function of time. A scheme illustrating the rearrangement and various NMR assignments can be found in Johnson et al.\textsuperscript{88}

NMR spectra are acquired as the spontaneous reactions proceed, to monitor the changing proportions of the substances present in the sample. The results of STORSY on this data set are presented in table 3.5, where it is observed that the regression slopes between the H1’ of the $1\beta$-isomer (see figure 3.7) and several intra-molecular protons are not in correspondence with

![Figure 3.7](https://via.placeholder.com/150)

Figure 3.7: Structure of the $1\beta-O$-acyl glucuronide of (S)$\alpha$-methyl phenylacetic acid, the H1’ atom that is used to calculate regressions and correlations is coloured pink.
Table 3.5: The regression coefficients (b) and correlation coefficients (Corr.) with the H1’ resonance (*) of 1-β-O-acyl glucuronide, shown in figure 3.7.

<table>
<thead>
<tr>
<th>Atom</th>
<th>$^1$H Proton ratio (ppm)</th>
<th>Proton ratio</th>
<th>Corr.</th>
<th>Slope $y = bx$</th>
<th>Slope $y = bx + a$</th>
<th>Corr.</th>
<th>Slope $y = bx$</th>
<th>Slope $y = bx + a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1’*</td>
<td>5.576</td>
<td>1</td>
<td>1.0000</td>
<td>1.00</td>
<td>1.00</td>
<td>1.0000</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H5</td>
<td>1.540</td>
<td>3</td>
<td>0.9997</td>
<td>2.27</td>
<td>2.53</td>
<td>0.9998</td>
<td>2.30</td>
<td>2.40</td>
</tr>
<tr>
<td>H2’</td>
<td>3.494</td>
<td>1</td>
<td>0.9995</td>
<td>0.71</td>
<td>0.72</td>
<td>0.9997</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>H4’</td>
<td>3.503</td>
<td>1</td>
<td>0.9997</td>
<td>0.59</td>
<td>0.59</td>
<td>0.9997</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>H3’</td>
<td>3.588</td>
<td>1</td>
<td>0.9999</td>
<td>0.75</td>
<td>0.74</td>
<td>1.0000</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>H5’</td>
<td>3.811</td>
<td>1</td>
<td>0.9999</td>
<td>0.87</td>
<td>0.86</td>
<td>1.0000</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>H4</td>
<td>4.001</td>
<td>1</td>
<td>0.9995</td>
<td>0.29</td>
<td>0.29</td>
<td>0.9995</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>H3</td>
<td>7.383</td>
<td>2</td>
<td>0.9976</td>
<td>0.37</td>
<td>0.40</td>
<td>0.9980</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>H1</td>
<td>7.416</td>
<td>1</td>
<td>0.9991</td>
<td>0.33</td>
<td>0.38</td>
<td>0.9980</td>
<td>0.36</td>
<td>0.42</td>
</tr>
<tr>
<td>H2</td>
<td>7.450</td>
<td>2</td>
<td>0.9985</td>
<td>0.41</td>
<td>0.47</td>
<td>0.9984</td>
<td>0.41</td>
<td>0.47</td>
</tr>
</tbody>
</table>

When using orthogonal least squares regression (orthogonal linear regression function, [http://www.mathworks.com/matlabcentral/fileexchange/16800](http://www.mathworks.com/matlabcentral/fileexchange/16800)), which takes into account the equivalent measurement error for the ‘x’ and ‘y’ variables, the values remain almost unchanged. As an example, the orthogonal regression with intercept from H1’ to H3, the lowest correlation, and therefore most likely to improve with this approach, gives 0.399 for the symmetrised and 0.430 for the unsymmetrised case, whereas the theoretical value of the ‘proton ratio’ is 2 (see table 3.5). This suggests that due to T2-editing and various other processing steps in JRES spectra, the usual stoichiometric area ratios are not observed, which was confirmed by evaluation of peak integrals.

**STORSY in the galactosamine toxicity study**

A demonstration of the regression between JRES projection signal intensities of the same molecule across a series of samples is shown in figure 3.8. STORSY was used to calculate the peak ratios between the CH and CH$_3$ peaks from lactate in the galactosamine toxicity study, and a useful degree of variation between the samples is observed. Unfortunately, the computed ratios of either the peak height (used for correlation analyses such as STOCSY) or the peak integral (traditionally used for quantitation) of the JRES sum projections are incorrect. Despite the high correlations, due to the linear relationship between peak signals being maintained, the value of the regression slope for the lactate signals is at least 4.6 for both peak height and peak integral, as opposed to the theoretical value of 3. Similar discrepancies were observed for other molecules investigated.
Figure 3.8: Regression slope with (blue) and without (pink) intercept, and the correlation coefficient ($r$) for the peak height (A) and peak area (B) data as calculated for the lactate peak in JRES spectra of plasma, processed without symmetrisation and using a sum projection. Although the regression is highly significant, there is a large discrepancy between the expected slope of 3 and the calculated slopes ranging between 4.6 and 5. Pearson’s correlation values ($r$) for the skyline projections without symmetrisation were 0.977 for peak height and 0.973 for peak area.

It should be noted that both the peak area and the peak height did not agree with the molecular stoichiometry, indicating that the quantitation is also absent for traditional approaches which would employ peak area.

**Feasibility of STORSY**

In conclusion, the reproducibility of JRES spectra is high, as is demonstrated by consistently high correlation values, but the sum projections of these JRES data are non-quantitative. This phenomenon is explained by a combination of factors, including differential relaxation effects: the $T_2$ and $T_1$ relaxation time constants are not the same for different protons in a given molecule. Peak quantitation is further compromised by processing using absolute value mode, which is inherently non-quantitative, and the use of sine-bell apodisation to create better-resolved peaks with smaller dispersive tails, which can differentially affect the peak intensity from peaks of different linewidth. The lack of quantitation of JRES data has been discussed previously: some signal intensities were found to be decreased whilst others increased, and this was explained by the different $T_2$ relaxation time constants. Composite window functions have been proposed for optimal data processing that claim to provide accurate quantitation; in general the quantitation is heavily dependent on the processing parameters.

Although the applied processing procedure is non-quantitative, a high correlation was displayed between intra-molecular protons of sum projections, validating the use of STOCSY for JRES sum projections. These correlations were also maintained in skyline projections, because a given molecule has consistent peak multiplicities and relative intensities for different samples. Semi-quantitative (relative) changes in metabolite concentrations have also been achieved by others using binned JRES projection spectra, and recently a line shape fitting strategy to aid quantitative
analysis of metabolites was proposed.  

The severe disturbances in the measured peak ratios indicate that there are some major limitations that need to be overcome in order to acquire spectra with a quality suitable for STORSY analysis. The requirements for an NMR experiment that can provide high quantitation accuracy include minimal relaxation-editing, and the removal of peak overlap and ideally multiplicity.

Alternative methods of obtaining JRES spectra include the usage of Zangger–Sterk pulse sequence elements that encode the second dimension spatially rather than with time, resulting in quantitative spectra with absorption mode line shapes. The quantitative JRES spectroscopy method suggested by Pell et al. has been investigated and it was found that the observed S/N is prohibitively low for use with currently available technology for metabolic profiling studies in biofluids. In addition, the involved pulse power optimisation is not appropriate for high-throughput applications.

3.4.7 Over-representation of metabolite peaks in NMR spectroscopy

In many NMR-based metabonomic studies, the spectra show prominent peaks from drugs, toxins and their metabolites. These can have undue weight in multivariate statistical models and the identification and means of down-weighting such peaks have been demonstrated previously using STOCSY type approaches. In addition, many of these substances give rise to multiple, often highly J-coupled, signals. This situation is also encountered here, where galactosamine and its metabolites, as well as endogenous compounds such as glucose, have many resonances that are in turn extensively coupled, and thus can be considered to be ‘over-represented’ in the NMR spectrum with respect to the number of variables that correspond to a metabolite. This has a major influence on variance-modelling approaches such as PCA, which could be dominated by these resonances, and makes the identification of other, possibly more discriminatory, markers more difficult. JRES projections could mitigate this problem, as the number of data points representing a given peak is markedly reduced upon the collapse of multiplets.

The spectral information content in the various types of 1D 1H NMR representations is evaluated by estimating the overall similarities between the different spectral data types through computing Pearson’s correlation coefficient between every pair of spectra. The results are plotted in figure 3.9, where each ‘pixel’ represents the correlation between two individual spectra; black lines separate different blocks representing six discrete data sets, related to the type of pulse sequence (1D, CPMG and JRES) and treatment class (galactosamine-treated and control). These figures allow for a direct comparison between the complete spectral information (not focused on peaks but also including baseline regions, although higher intensity peaks will have a greater leverage effect than smaller baseline peaks) of the different NMR experiments for the two biofluid matrices.
Figure 3.9: Heat maps showing Pearson’s correlation coefficient (r) between the 1D, CPMG and skyline projected JRES NMR spectra for urine and plasma, before and after recursive segment-wise peak alignment. The solid black lines separate the galactosamine dosed (‘D’) and control (‘C’) groups and the different experiment types (‘1D’, ‘CPMG’, ‘JRES’).

It is apparent from figure 3.9 that spectral alignment of urine NMR data greatly improved the overall clustering with experiment and class subtypes. Dose-related clustering patterns are highly disturbed for 1D and CPMG data and almost absent from JRES data before alignment, confirming the necessity for peak alignment; similar results were obtained for sum projected spectra. The correlation patterns for plasma demonstrate clear differences between the different experiments: 1D and CPMG plasma data are more similar to each other than to the JRES projections. As expected, the JRES projections are slightly more correlated to the CPMG data than the conventional 1D spectra, which can be explained by the known $T_2$ relaxation editing in both pulse sequences.

The aligned urinary NMR data in figure 3.9 show that the 1D experiments from the control group have a higher correlation to the CPMG spectra of the control group than to the dosed 1D data. This is not unexpected, as for urine the difference between a CPMG and 1D spectrum is small due to the limited number of molecules edited out by the CPMG, since urine is predominantly comprised of low molecular weight compounds. Moreover, urine from the dosed group contains
high levels of galactosamine and its metabolites (with high multiplicity and therefore containing many variables), which are absent from the control spectra, in addition to the toxicity-induced alteration of endogenous components. Ultimately, the differences in biochemical content of urine for the dosed and control samples in this study are much larger than the spectral differences between these pulse sequences. If the differences between the groups were more subtle, this would not necessarily have been the case.

The correlations between the dosed and control groups for urinary spectra are higher for the JRES projections than for the 1D and CPMG spectra. This can be explained by the smaller number of variables that are changed following the toxic insult and this in turn is due to the collapse of the multiplets, leaving the JRES projections from the two classes more alike. This indicates a higher representation of xenobiotic-related peaks in 1D and CPMG urinary spectra of galactosamine-treated animals. Similarly, the JRES spectra of control urine have a lower correlation coefficient with dosed 1D and CPMG spectra (figure 3.9, dotted blocks), compared to dosed JRES spectra with either control or dosed 1D and CPMG spectra (figure 3.9, dashed block). Naturally, the difference is smaller when the JRES spectra were also from dosed samples, or when the 1D and CPMG are control spectra as well. Thus the toxin-related metabolites can be regarded as ‘over-represented’ in the 1D and CPMG data and the effective ¹H broadband decoupling as achieved through projection of JRES data reduces this effect. This over-representation of multiplets in 1D and CPMG spectra is comparable to the over-representation of protein envelope signals in PCA models based on plasma NMR data scaled to unit variance.⁴⁴

An example of over-representation of glucose is seen in figure 3.10, where the variables representing glucose are coloured for a 1D spectrum and a JRES projection. The ‘over-representation’ is not completely remedied by the multiplet collapse. For example, glucose has more resonances than most other metabolites even when collapsed into JRES projections (12 chemically shifted protons for the α and β anomers). From a biological point of view, there is no reason why metabolites with more (non-equivalent) protons should be allowed to dominate any pattern recognition model: the over-representation is caused by molecular structure and not by biological importance. Peak or curve fitting (e.g. possible with commercial software), factor analysis, curve resolution¹⁵⁴ and STOCSY-E¹⁵³ might be able to compress signals in a JRES projection further into one variable per metabolite, i.e. deconvoluting the constituting metabolites into the pure spectral profiles, to obtain relative concentrations across a sample series, which can then be subjected to unbiased multivariate analysis.

When projecting the JRES data without tilting, the multiplicity is retained. It was found that the untilted JRES projections were still dissimilar from the 1D and the CPMG data in the correlation maps. This indicates that it is not only the multiplicity that caused these data to be inherently different, but also the nature of the JRES experiment with the different degree of T₂-editing and data processing applied such as usage of absolute value mode and sine-bell apodisation.
Figure 3.10: An illustration of the fraction of variables that is representing the highly-split multiplets of the glucose resonances in a 1D spectrum and a \textit{jres} projection. The over-representation is more pronounced for the glucose peaks in a 1D spectrum, than in a \textit{jres} projection.

The patterns that were observed in figure 3.9 were not found when the data were scaled to unit variance prior to analysis, or if the more robust Spearman rank correlation was used. This is because there are a few large peaks that are the main cause of the correlation coefficients, as they exert a ‘leverage effect’ (where a few high values profoundly influence the regression and correlation coefficients); the same effect will also occur with methods such as PCA and PLS, and can be partially remedied by scaling of the variables. Therefore, this figure is not only a clear example of the over-representation of signals, but in addition it emphasises the need to reduce the influence of high-intensity peaks, for example through the use of robust methods or by employing appropriate scaling.

\textbf{Normalisation of \textit{jres} spectra}

The previously reported bias in total area normalisation associated with toxic insults\textsuperscript{80} should theoretically decrease when using \textit{jres} projections, due to the reduction in presence of high-concentration exogenous and endogenous signals contributing to the total spectral area.

Probabilistic quotient normalisation is based on the assumption that less than 50\% of the data points change across different spectra\textsuperscript{82} but this requirement may not be fulfilled in \textit{jres} spectra. In the \textit{jres} urine projections, for example, the majority of variables correspond to noise regions that may be less representative of the sample concentration (in \textit{jres} projections, \textasciitilde60\% of the spectrum is ‘noise’ as opposed to \textasciitilde10\% for 1D; to calculate the percentage of data points comprising noise in the spectra, 10 times the standard deviation of the noise over the region $\delta$ 0.2–
0.5 in JRES spectra and δ 9.8–10.0 in 1D spectra was used as a cut-off value). These considerations demonstrate that the choice of ‘best’ normalisation method is complicated; however because the probabilistic quotient normalisation is the more robust option, this was performed.

### 3.4.8 Complementarity of 1D and CPMG spectra and JRES projections

It might be thought that the information content of a CPMG NMR spectrum from a biofluid sample may essentially be obtained from acquisition of a JRES spectrum combined with a standard pulse-and-acquire ‘1D’ experiment. The main virtue of CPMG, T2-editing, is also obtained in the JRES experiment, with the added advantage that JRES spectra provide projections that are easier to interpret in heavily overlapped regions, especially when combined with the scalar coupling information available in the 2D data. A potential disadvantage of the JRES experiment is more pronounced T2-editing than with the CPMG experiment since conventional sine-bell apodisation minimises the contribution of rapidly relaxing protons even further. As a result, increased information extraction is obtained in highly overlapped spectral regions, for example around δ 3.0–4.0 (see the assignments in figure 3.3) at the expense of information density for rapidly relaxing species.

Because it is unnecessary to acquire a large number of increments (as zero filling and linear prediction\textsuperscript{155} can be of help), it is possible to acquire good-quality JRES spectra of typical biofluid samples very quickly, in a period of less than 5 minutes.\textsuperscript{156} There are multiple approaches for fast JRES data acquisition: one is based on a reduction of the number of increments and transients in the traditional JRES experiment,\textsuperscript{156} whereas other methods encode the second dimension spatially rather than through time, reducing data acquisition times, albeit at the cost of sensitivity.\textsuperscript{150–152, 157}

The sensitivity of JRES data is only slightly lower than the sensitivity of 1D spectra acquired over the same amount of time,\textsuperscript{158} and can be increased further through the use of cryoprobes. Therefore time is not prohibitive to supplement or substitute acquisition of CPMG spectra with JRES spectra, even though more scans could be required. Moreover, it has been shown that the CPMG experiment can be reliably approximated from mathematical manipulation of 1D data.\textsuperscript{159}

Given the above, JRES NMR spectroscopy is found to be advantageous and complementary to the more usual one-dimensional NMR methods for routine use in metabonomic research, as previously suggested.\textsuperscript{123, 128} An additional benefit is that JRES spectra, in contrast to 1D and CPMG spectra, have the same appearance at different observation frequencies when spectra are displayed on the ppm scale, and thus data can be more easily compared across different spectrometers. This allows for rapid comparative analysis of spectral data sets from multiple field instruments and will simplify generation of universal spectral databases.\textsuperscript{134} Thus, JRES NMR spectra can be extremely useful as part of the metabonomic NMR toolbox.\textsuperscript{115, 134}
3.4.9 An example of information recovery using JRES spectroscopy

As illustrated in figure 3.11 with data from the galactosamine toxicity study, the reduced overlap of biofluid spectral peaks in JRES spectra not only aids model interpretation, but can also resolve previously unidentified correlations; that is, the reduction in peak congestion resolves overlapped multiplets, thereby allowing previously concealed peaks to be recovered. The JRES projections of the CH$_3$ peak of alanine (in plasma, δ 1.47) and creatinine (in urine, δ 3.05) show improved intra-molecular correlations; the correlation with the alanine CH peak at δ 3.77 and the creatinine CH$_2$ peak at δ 4.06, respectively, are more pronounced, see figure 3.11. The plasma region was previously crowded with amino acid and glucose signals, whilst the creatinine signals in urine were overlapped with galactosamine resonances. The individual resonances are completely resolved upon JRES projection (despite the imperfect alignment of peaks in this region for urine JRES projections) and hence a more reliable STOCSY plot is obtained, demonstrating improved intra-molecular correlations. This high correlation was not observed in the untilted JRES projections, which demonstrates that the obtained correlation is indeed uncovered by the reduced overlap due to effective broadband decoupling. Notably, this is achieved despite a lower S/N for the presented JRES spectra (see table 3.1). These findings are particularly relevant to the study of endogenous changes when the conventional 1D spectra are heavily obscured by peaks from the administered toxin and/or its metabolites.

Figure 3.11: STOCSY plots represent the correlation ($r^2$) and covariance, using colour and height, respectively. (A, B) The STOCSY driven from the alanine CH$_3$ peak at δ 1.47 shows improved correlation with the alanine α–CH signal for the JRES spectra of plasma, because the overlap with glucose signals is reduced. (C, D) STOCSY driven from the creatinine CH$_3$ δ 3.05 peak shows good correlation to its CH$_2$ signal in urine, demonstrating that this intra-molecular correlation is only present in JRES projections, where the overlap with galactosamine resonances is removed.
Similarly, in figure 3.12 the improved identification of biomarker metabolites when using JRES projections compared to 1D and CPMG spectra in the galactosamine toxicity study is demonstrated through a STOCSY plot related to class (note that some correlations are not optimal because the creatinine δ 4.06 and H2 β-galactosamine δ 3.14 resonances are not aligned perfectly in the JRES projection). The correlation patterns identify the same biomarkers for response to the galactosamine toxic insult in the 1D, CPMG and JRES spectra, but the reduced overlap in the JRES projections (and the availability of splitting patterns in the 2D spectrum) significantly aids in a complete and unambiguous assignment of all the individual metabolic resonances, contrary to previous results where the region δ 3.4–4.2 had to be simply labelled ‘galactosamine’. It should also be noted that the 2-oxoglutarate peaks display a much higher correlation in the JRES projection and therefore this metabolite is unveiled as a potential biomarker.

Figure 3.12: Correlations of the spectral data points (colour, $r^2$) with a dummy variable representing the control or galactosamine-treated class, projected onto the covariance spectrum with the dummy variable for the different aligned spectral NMR types of urine: (A) 1D, (B) CPMG, (C) JRES symmetrised skyline projection. The interpretation of highly overlapped signals, for example in the region of δ 3.0–4.0, is greatly improved for JRES, resulting in improved biomarker identification. Legend: DMA, dimethylamine; 2-OG, 2-oxoglutarate; α/βG, α/β-galactosamine.
3.5 Conclusions

The most important advantage of JRES NMR spectroscopy over 1D and CPMG NMR spectroscopy is the reduction in peak overlap, which aids biomarker identification and data interpretation in ‘congested’ spectral areas. The usefulness of full-resolution JRES projections was evaluated, and ways to realise the potential of the technique for metabonomic studies have been specified, particularly with respect to processing and spectral alignment of the data. The chosen galactosamine toxicity data set is particularly suitable, as many previously identified differentiating biomarkers of galactosamine toxicity represent complex multiplets and thus a considerable effect resulting from the collapse of signals in the JRES spectral projections was observed. Due to the inherently non-quantitative nature of the standard JRES spectrum, there is no need to use the stoichiometry-conserving sum projection, and thus the skyline projection is often preferred, as this gives an increased S/N and is still highly reproducible. Correlations of interest retrieved with STOCSY are conserved or improved compared to conventional one-dimensional experiments for both projection methods, whilst regression slopes obtained using STORSY are unreliable in the current experimental system. The effective peak span widths decrease upon projection compared to highly split multiplets, and thus pattern recognition and correlation methods require alignment of the full-resolution JRES spectral projections. The subsequently reduced peak overlap in projected JRES spectra can improve the quality and interpretation of multivariate models and statistical correlation analyses, and will enhance biomarker identification and reduce the differential degree of over-representation of molecules. Hence, judicious acquisition of JRES data and application of alignment can improve the interpretation of pattern recognition models and increase the information content extractable from NMR-based metabonomic studies, resulting in enhanced biomarker identification.
Chapter 4

Simulated annealing optimised K-OPLS in metabonomics

4.1 Aims and objectives

In the previous chapter, the choice of NMR experiment, peak alignment and subsequent use of linear multivariate projection methods, such as principal component analysis and orthogonal partial least squares (OPLS), which find use in modelling of spectroscopic data, were discussed. However, when the relationship between the descriptor variables and the response is non-linear, conventional linear prediction models will perform sub-optimally. In this chapter, the focus is shifted to evaluate the mathematical non-linear predictive modelling of spectroscopic data with kernel-based orthogonal partial least squares (K-OPLS), and its characteristics are illustrated with three separate metabonomic data sets: a study on the liver toxin galactosamine, a study of the nephrotoxin mercuric chloride and a study of Trypanosoma brucei brucei parasite infection. This work, for which the accepted publication is included in appendix D, was done by means of the following steps:

1. Demonstrating that the optimisation of the parameter $\sigma$ of the Gaussian kernel transformation performed in K-OPLS can be done effectively and in a user-friendly manner by means of simulated annealing;

2. Implementation of the simulated annealing optimisation and calculation of prediction performance metrics (AUC and $q^2_y$) using a nested cross-validation approach;

3. Comparing the predictive ability of non-linear K-OPLS and its linear equivalent OPLS;

4. Evaluating model interpretation enabled by the separate modelling of predictive and orthogonal variation and accompanying score plots;

5. Development of methods to approximate and visualise which variables play a main role in the non-linear model based on ‘pseudo-samples’.
4.2 Introduction

Multivariate projection methods such as partial least squares (PLS)\(^7\) and the related orthogonal
PLS (OPLS, see §2.3.5)\(^3\) are frequently applied for modelling of spectroscopic biological data, as
they provide predictive and interpretable models.\(^23,55,64\) The OPLS algorithm enables separate
modelling of \(Y\)-predictive (response-related) and systematic \(Y\)-orthogonal (response-orthogonal)
variation in data, ‘structured noise’.\(^92–95,160\) Thus, OPLS is beneficial in terms of model interpre-
tation compared to PLS, and has been successfully employed in metabonomics.\(^56\) The concept
of \(Y\)-orthogonal variation can be understood as systematic effects that are needed to characterise
the system but are unrelated to the question at hand, i.e. the model predictions. For instance,
when aiming to classify a group of responders versus non-responders to a particular treatment,
the structured noise could be composed of inter-sample differences that are needed to describe
variability of the system but are not useful for separating responders from non-responders.

Most of the commonly used multivariate prediction models in metabonomics assume a linear
relationship between the \(X\) (descriptor) and \(Y\) (response) variables. However, many biological
systems display non-linear characteristics in response to a perturbation. Under such conditions,
non-linear methods are expected to provide improved models, which is particularly important in
predictive applications such as disease diagnostics,\(^161\) assessment of toxicity\(^142\) and characterisation
of variable response of individuals to drugs in personalised healthcare.\(^33,34\)

A particular class of non-linear models are kernel-based models,\(^162\) with an early chemometric
application in the form of radial basis functions–PLS.\(^163,164\) Other examples of kernel-based mod-
els include support vector machines (SVM),\(^165–167\) kernel-based partial least squares (KPLS\(^168,169\))
and kernel-based least squares regression.\(^170\) Kernel-OPLS (K-OPLS) is the non-linear extension
of the OPLS model, a commonly used multivariate model in metabonomic studies.\(^55,56\) In con-
trast to separate (linear) orthogonal signal correction (OSC) followed by KPLS modelling,\(^171\) or
kernel-OSC followed by KPLS modelling,\(^172\) K-OPLS provides an integrated orthogonal signal cor-
rection property that allows for separate modelling of predictive and \(Y\)-orthogonal variation in the
feature space, removing drawbacks associated with multi-step solutions, such as separate (K)OSC
and (K)PLS steps.\(^93\) Although the K-OPLS method does not necessarily provide improved pre-
diction performance compared to other kernel-based models,\(^162\) K-OPLS facilitates an improved
model interpretation compared to alternative models, which can aid quality control and further
understanding of the model and data.

Kernel-based models require an optimisation of the kernel-function parameter, which may be
challenging for the non-specialist, as the kernel parameter is often a continuous parameter with
an undefined upper limit that may have multiple local optima. The optimisation step is essential
to produce a model with a good predictive performance.\(^171,173,174\) Here, an automated procedure
is implemented for optimisation of the kernel parameter based on simulated annealing (SA), a
stochastic optimisation method,\(^175,176\) which greatly simplifies the use of K-OPLS, as suggested by
Bylesjö et al.\textsuperscript{57} This optimisation has been incorporated into the freely available K-OPLS software package for both R and MATLAB (http://sourceforge.net/projects/kopls/).

The main objective in this study is to evaluate if non-linear prediction models provide an advantage compared to linear alternatives in two common application areas of metabonomics: toxicology and disease diagnostics. Using the non-linear SA-K-OPLS method, the possibility for improved prediction performance in comparison to the linear OPLS model is demonstrated for three separate spectral NMR metabonomic data sets. In particular, the focus is on problems where prediction is of paramount importance, and it is also shown how structured Y-orthogonal variation can be interpreted to gain further insight into the data. To further increase model transparency, the influence of variation of each variable in the K-OPLS model is approximated.

4.3 Methods

4.3.1 Kernel-OPLS

The K-OPLS algorithm is a kernel reformulation of the OPLS algorithm, described in §2.3.5, utilising the computationally efficient ‘kernel-trick’, allowing the elements of the kernel Gram matrix $\mathbf{K}$ [$n \times n$] to be treated as dot-products in the higher-dimensional feature space $\mathcal{F}$ to which the data are (non-linearly) mapped. K-OPLS\textsuperscript{57,58} has a set of predictive latent variables represented by the predictive scores matrix $\mathbf{T}_p$ [$n \times a_p$], where $a_p$ is the number of predictive components, and a set of Y-orthogonal latent variables represented by the scores matrix $\mathbf{T}_o$ [$n \times a_o$], where $a_o$ is the number of extracted orthogonal components, in analogy to the OPLS model.

The kernel-trick is common among all kernel-based pattern recognition methods, and is computationally efficient since no explicit transformation of the data is needed. Estimating a kernel-based model is a two-step procedure. First, $\mathbf{K}$ is calculated using a suitable kernel function, see equation 4.1, where $x_i$ and $x_j$ represent rows $i$ and $j$ in $\mathbf{X}$. This circumvents the explicit mapping of the descriptor data to the (higher dimensional) feature space, and facilitates modelling of non-linear variation.

$$K_{i,j} = k(x_i, x_j)$$ (4.1)

Then, K-OPLS is used to establish a regression or discriminant analysis model of the, potentially non-linear, relationship between the descriptor data, now represented by the kernel-matrix, $\mathbf{K}$, and the response variable(s), $Y$.

4.3.2 Kernel function optimisation

There are many kernel functions available to construct a kernel matrix $\mathbf{K}$; here the commonly used Gaussian kernel function is employed, see equation 4.2.

$$k(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)$$ (4.2)
This kernel function has a tuning parameter related to the size of the Gaussian kernel ($\sigma$). The value of the kernel parameter greatly influences the predictive performance of the resulting model (see the black line in figure 4.1, showing the predictive ability as a function of the kernel parameter), emphasising the importance of optimisation of this parameter.

To find a good parameter value for the kernel function, grid search or gradient minimisation could be applied,\cite{166, 173} but these approaches may be computationally and labour intensive or inefficient, can get trapped in local minima and are also dependent on the selected grid range and resolution. An exhaustive search over all possible values is not computationally tractable due to the continuous nature of the parameter. There is a multitude of global optimisation methods available, such as simplex-based methods\cite{177} and genetic algorithms,\cite{178, 179} previously reported in the context of SVM applications. In this work, it is demonstrated how simulated annealing (SA) can be used to automatically optimise the kernel parameter in the K-OPLS model in order to achieve good prediction performance in a user-friendly manner, but other approaches could also have been applicable to this optimisation problem. The goal is not to compare different optimisation methods, but to provide and demonstrate an automated and easy-to-use procedure.

### 4.3.3 Simulated annealing

In analogy with thermodynamic principles,\cite{181} where an infinitely slowly cooled ‘sample’ will occupy the lowest energy state, simulated annealing enables stochastic sampling of the kernel parameter space as a function of the decreasing ‘temperature’.\cite{175} The temperature influences the probability of temporarily moving to sub-optimal parameter values in order to overcome local minima, and with decreasing temperature the search range of the kernel parameter is slowly narrowed down (see the bars in figure 4.1).

The pseudo-code for the SA algorithm is provided in table 4.1 and algorithm 1 (also see Corana et al.\cite{176} for further details). In line 1 of the SA algorithm, the kernel parameter $x$ is initialised. Subsequently, and until convergence, the prediction performance is evaluated and a new kernel parameter is proposed in the area surrounding the previous value, based on the step size $v$. If the prediction performance, determined by cross-validation (see §4.3.4), of the evaluated value is improved, the proposed parameter value will always be accepted, otherwise, the proposed parameter value is accepted with a probability related to the current temperature $T$, according to the Metropolis criterion (line 14).\cite{181} High temperatures have a higher probability of accepting kernel parameter states that result in sub-optimal prediction performance, enabling adequate sampling of the parameter space and avoiding local minima. The SA algorithm is designed to maintain an acceptance rate of around 50% of the proposed kernel parameter values, and the step size is sequentially updated to accommodate this (line 25 in algorithm 1). As the temperature is gradually decreased (line 29), the chance of accepting high-energy values, i.e. kernel parameter values that result in worse predictive ability, decreases until a defined convergence criterion (line 7) is fulfilled.
The optimisation of a Gaussian kernel parameter $\sigma$ with simulated annealing is performed by minimising $1 - q^2_\sigma$ (left hand axis). The predictive ability for the sampled $\sigma$ values is plotted in black, here for an exemplar run of the mercuric chloride high dose regression data set, see §4.4.2, and the presence of local minima is clear. Subsequent ‘optimal’ points (i.e. a value that is a new minimum when it was evaluated) are connected with the red line. A number of values for the kernel parameter $\sigma$ are sampled, the spread of which becomes smaller as the temperature (right hand axis) decreases. The bars indicate the region in which kernel parameters were evaluated for a given temperature $T$. The temperature cooling rate $rT$ was 0.7 for demonstration purposes, for all SA-K-OPLS results $rT = 0.1$ was used.

Here, results are presented from SA-K-OPLS using a Gaussian kernel function, however, the simulated annealing approach is widely applicable, and can be used to optimise other kernel functions with a continuous parameter. The kernel parameter $\sigma$ in the Gaussian kernel function defined in equation 4.2 is optimised for values $> 0$. There are a number of parameter settings, that were implemented as defaults in the available code, that could influence the simulated annealing procedure. These values were optimised heuristically, based on literature recommendations and time and convergence considerations. Although it may seem counter-intuitive to set more than one parameter in order to optimise the one kernel parameter setting, the default simulated annealing settings appear robust and typically provide good convergence, whereas the predictive performance as a function of the kernel parameter is a complex, non-trivial optimisation problem as is illustrated in figure 4.1.
Algorithm 1 The algorithm used for optimisation of the kernel parameter based on simulated annealing, see also Corana et al.\textsuperscript{176}

1: $x(0) \sim U(0, 1)$ \{Initialise kernel parameter $x$\}
2: $v = x(0)$ \{Initialise step size $v$\}
3: $R(0) = f(x(0))$ \{Calculate performance metric $R$\}
4: $R_{\text{opt}} = R(0)$ \{Initialise performance metric $R_{\text{opt}}$\}
5: $h = 0$
6: $i = 0$
7: while $(RT(h) - R_{\text{opt}} > \epsilon$ \& \max\{RT$_{(i \in \mathbb{N}_+ \land h \in h)} - RT(h)\} > \epsilon)$ do
8: \hspace{0.5em} for $j = 1$ to $N_T$ do
9: \hspace{1em} $N_{\text{accepted}} = 0$
10: \hspace{1em} for $k = 1$ to $N_S$ do
11: \hspace{1.5em} $x_{\text{prop}} = x_i + U(-1, 1) \times v$ \{Propose new kernel parameter\}
12: \hspace{1.5em} $R_{\text{prop}} = f(x_{\text{prop}})$ \{Evaluate performance at proposal\}
13: \hspace{1.5em} $\alpha \sim U(0, 1)$
14: \hspace{1.5em} if $\alpha < \min\{\exp(R(i) - R_{\text{prop}})/T, 1\}$ then
15: \hspace{2em} $x_{i+1} = x_{\text{prop}}$ \{Accept and save current state\}
16: \hspace{2em} $R_{i+1} = R_{\text{prop}}$ \{Save current performance\}
17: \hspace{2em} if $R_{i+1} < R_{\text{opt}}$ then
18: \hspace{2.5em} $x_{\text{opt}} = x_{i+1}$ \{Save if best state so far\}
19: \hspace{2.5em} $R_{\text{opt}} = R_{i+1}$ \{Save the optimal performance\}
20: \hspace{1em} end if
21: \hspace{1em} $N_{\text{accepted}} = N_{\text{accepted}} + 1$
22: \hspace{1em} $i = i + 1$
23: end if
24: end for
25: Update $v$: 
\[
    v = \begin{cases} 
        v \left(1 + \frac{N_{\text{accepted}}}{N_S} - 0.6 \right) & \text{if } N_{\text{accepted}} > 0.6N_S \\
        v \left(1 - \frac{N_{\text{accepted}}}{N_S} + 0.2 \right) & \text{if } N_{\text{accepted}} < 0.4N_S \\
        v & \text{otherwise}
    \end{cases}
\]
26: end for
27: $RT(h) = R_{i+1}$ \{Save last accepted value before temperature update\}
28: $h = h + 1$
29: $T = r_T \times T$ \{Update temperature $T$\}
30: $x_{(i)} = x_{\text{opt}}$
31: $R_{(i)} = R_{\text{opt}}$
32: end while
4.3.4 Nested cross-validation

A nested cross-validation procedure\textsuperscript{182,183} was used to evaluate prediction performance during optimisation of the kernel parameter, and the optimisation was performed separately for models with different numbers of Y-orthogonal components, see §4.3.5. In each ‘outer cross-validation loop’, the data were split into a training set and a test set, and the kernel parameter was optimised in the ‘inner cross-validation loop’ (using only the training data from the outer cross-validation loop, see figure 4.2 A). The prediction performance was then evaluated using the test set from the outer cross-validation loop. Thus, the outer loop test set is never used for model optimisation, but only for evaluation of overall prediction performance. This nested cross-validation procedure (see also §2.3.6) is repeated for n outer cross-validation rounds, and figure 4.2 B provides a schematic overview of the procedure.

The predictive ability of regression models was evaluated using the ‘goodness of prediction’ measure $q_y^2$ (see equation 2.17), and discriminant analysis models were characterised by the mean sensitivity (see equation 2.18) and the area under the receiver operating characteristic curve ($AUC$, see equation 2.19). For the presented results, Monte Carlo cross-validation (MCCV, see §2.3.6) was used for regression problems and the class-balanced MCCV approach (MCCVB) for discriminant analysis, with 10 and 100 repetitions in the inner and outer cross-validation loop, respectively, and 75% of the data were used in the training set for both the inner and the outer cross-validation loop. Identical training and test sets were used to generate SA-K-OPLS and OPLS models in order to achieve a fair comparison. The score plots and reported kernel parameter values ($\sigma$, see table 4.2) were generated based on a kernel-optimisation and K-OPLS model using the full data set, as
**Figure 4.2:** (A) The nested cross-validation procedure minimises the risk of model over-fitting and overestimating the prediction performance while optimising the kernel parameter. In this procedure, $n$ optimal kernel parameters are determined, where $n$ is the number of repetitions of the 'outer' cross-validation (CVo) round. The kernel parameter for each CVo training set is optimised using simulated annealing in the 'inner' cross-validation loop CVi, while the CVo test set is used to evaluate the prediction performance. (B) This schematic demonstrates that the simulated annealing step is performed in each inner cross-validation loop (step 1) to find the kernel parameter (step 2) which is used to predict the CVo test set (step 3). The full procedure is repeated $n$ times (step 4) to generate the cross-validated estimate of performance.

This model would be the one that could be used for prediction of additional data (an example of external validation is shown in figure 4.9). The convergence plots shown in figure 4.3 demonstrate that the prediction performances stabilise within the 100 outer cross-validation rounds to provide reliable estimates of $q^2$ and AUC.

### 4.3.5 Number of Y-orthogonal components

The number of Y-orthogonal components for the OPLS and SA-K-OPLS models presented was chosen by evaluating results for models calculated with 0 to 10 orthogonal components (OC); all models required the use of 1 predictive component (because Y was a single vector or a two-class dummy matrix). The kernel parameter is optimised separately for each number of calculated components in the model (e.g. with 2 OC, the kernel parameter will be slightly different than with 1 OC). The final number of orthogonal components was determined as follows: additional components were included if $q^2_y$ increased by more than 1% per component for regression analysis, and in discriminant analysis the number of components corresponding to the maximum AUC was chosen.

Table 4.2 lists the final number of components used for each data set. The optimal number of components is determined separately for SA-K-OPLS and OPLS models, as the models are different in nature.
Figure 4.3: The prediction results of the 100 outer cross-validation rounds are shown for both the OPLS and SAK-OPLS models in the top panels for the five data sets presented in this work. The cumulative mean (solid line) and median (thin line) for one to 100 cross-validation rounds are shown in the middle panels. This demonstrates the convergence of the $q^2_Y$ predictive values for regression studies: ‘HgCl$_2$ high dose’ (B), ‘HgCl$_2$ both doses’ (C) and ‘TBB infection time’ (E); and stabilisation of the AUC measure for the discriminant analysis studies: ‘Galactosamine’ (A) and ‘TBB classification’ (D). The prediction performance appears to be converged for all data sets, as the means appear consistent after 100 cross-validation rounds. The cumulative standard deviation (std, solid line) and median absolute deviation (mad, thin line) are shown in the bottom panels.
A toy example

The glucuronide data set (see §3.4.6 and figures 2.10 and 2.11) will be used to demonstrate the use of SA-K-OPLS in non-linear spectral modelling. This data set consists of NMR spectra acquired at regular intervals as the transacetylation and isomerisation reactions of 1–β–O–acyl glucuronide proceed. The time after solvation of the glucuronide is predicted based on the binned NMR spectral data (δ 0.01 resolution) of the evolving mixture using 100 outer and 10 inner MCCV rounds (75% of the data was used in the MCCV). The results are shown in figure 4.4 and it was found that the predictive performance improved significantly compared to the linear model: the p-value of a t-test based on the 100 cross-validation $q^2_y$ values in SA-K-OPLS versus the OPLS $q^2_y$ results (both with 4 orthogonal components) is $\ll 10^{-15}$. The prediction is improved, because compounds that are neither starting nor final components will show a non-linear behaviour as a function of time, as their concentration initially increases and then decreases.

Figure 4.4: SA-K-OPLS predictions of time (colour-coded) in the glucuronide data set were improved compared to the already high-quality linear OPLS model.
4.3.6 Software implementation

The updated K-OPLS software (version 1.1.1)\(^5\) is freely available at [http://sourceforge.net/projects/kopls/](http://sourceforge.net/projects/kopls/). Implementations are provided for MATLAB and R, and are compatible with Windows, Mac OS X and Linux operating systems and include functionality for nested cross-validation and kernel parameter optimisation using simulated annealing and grid search. All results presented here were obtained using MATLAB (R2008a, The Mathworks, MA, USA).

4.3.7 Visualisation of K-OPLS variable importance

A method recently presented to infer variable importance in SVM\(^1\) was altered to visualise the important variables for K-OPLS discriminant analysis models. Three different approaches were undertaken, all based on the evaluation of variations in an individual variable represented by a ‘pseudo-matrix’. There are \(m\) pseudo-matrices, one for each variable, and each pseudo-matrix consists of 20 pseudo-samples (so that the matrix dimensions are [20 × \(m\)]). The intensity of all variables in the pseudo-matrix was set to zero, except the variable under investigation, of which the value was incremented from the minimum to the maximum value of that variable in the original data matrix (here, this data matrix had been scaled to unit variance). Thus, a pseudo-matrix \(X_{ps}\) consists of zeros [20 × \(m\)] with the column of the selected variable represented by 20 values of increasing intensity.

- **PCA of the kernel-transformed data**

The kernel matrix of the spectral data with the optimum parameter \(\sigma\) in the K-OPLS model is centred (\(K^c\ [n \times n]\)) after subtracting variability described by the response-orthogonal components, and a PCA model is calculated. The score plots of this PCA model are subjectively evaluated to find a combination of two PCs for which the classes show the best separation. Then, each pseudo-matrix is subjected to the same kernel-transformation, including the orthogonal signal subtraction, as the original spectral data. Each kernel pseudo-matrix is subsequently centred (\(K^c_{ps}\ [20 \times n]\)) and projected into the previously constructed PCA model, which results in 20 score values for each of the pseudo-matrices. In the score plot of the two selected PCs that showed a separation of the classes, the pseudo-matrix scores are connected to form the trajectory of the variable in the non-linear score space. Trajectories in the direction of discrimination between the classes are expected to have higher importance in the model.\(^1\)

- **PLS of the kernel-transformed data**

A PLS model with one predictive component is calculated based on the kernel transformed spectral data \(K^c\). Each kernel pseudo-matrix \(K^c_{ps}\) is projected into the PLS model to give 20 predictive score values for each of the pseudo-matrices; variables that cause a large change in the predictive component are more important.
• **K-OPLS prediction of pseudo-samples**

The scores of $X_{ps}$ are predicted by a previously-built K-OPLS model. The maximum difference between the predictive scores for the 20 pseudo-samples is calculated, and this range of the predictive scores for each variable is used as a measure of variable importance.

### 4.4 Data sets

#### 4.4.1 Galactosamine toxicity

The galactosamine toxicity data set comprises $^1$H NMR urinary spectra from rats treated with galactosamine hydrochloride ($n = 40$, 500 mg/kg). Galactosamine is a model hepatotoxin that has been shown to induce a highly variable pathological response in a given cohort of rats: it is common that a proportion of animals show no adverse effects (non-responders) and responders display varied severities of hepatic necrosis. Independent histopathological and clinical chemistry analyses were used to determine the degree of response to galactosamine (galN). Urine samples were obtained at 8 and 16 hours collection intervals representing -48 to -40 (8 hr), -40 to -24 (16 hr), -24 to -16 (8 hr) and -16 to 0 (16 hr) hours before dosing (samples courtesy of the COMET–2 project), as described previously in Coen et al. For this data set, the goal was the classification of non-responders ($n = 10$) versus responders ($n = 30$) from the pre-dose urinary $^1$H NMR spectra (samples acquired at 40, 24, 16 and 0 hours before dosing), and this was performed using discriminant analysis with SA-K-OPLS.

#### 4.4.2 Mercuric chloride toxicity

Rats were administered mercuric chloride, and full details are given in Lindon et al. Mercuric chloride and its possible metabolic derivatives are NMR-invisible, and thus all observed spectral changes can be attributed to endogenous metabolites. Doses were given at two levels ($n = 10$ ‘low’ dose, 0.5 mg/kg, and $n = 10$ ‘high’ dose, 2.0 mg/kg) and urinary samples were obtained at time points before (0 hr) and after dosing (8, 24, 48, 72, 96, 120, 144 and 168 hr post-dose) and $^1$H NMR spectra were acquired. Half of the animals at each dose level were sacrificed at 24 hours after dosing, the other animals were sacrificed at 168 hours. The time-course (9 time points) is modelled with SA-K-OPLS regression.

#### 4.4.3 Trypanosoma brucei brucei parasitic infection

To evaluate the effect of *T. brucei brucei* parasitic infection in mice, plasma samples were collected from an infected group ($n = 12$), and a control group ($n = 12$) of animals. Animals in both groups provided samples (108 in total) at a series of time points pre- and post-infection (-2, 1, 7, 14, 21, 28 and 33 days after infection), as described by Wang et al. The presence of infection was
established using discriminant analysis of pre-infection and control animals versus infected animals. This was followed by a regression analysis, on the infected animals only (64 samples), to predict the time after infection.

4.4.4 Data processing

Proton (1H) NMR spectra were acquired as previously detailed. As is standard, spectra were imported into MATLAB (R2008a, The Mathworks, MA, USA) with δ 0.0005 resolution. Spectral regions, excluding the water resonance for all spectra and the urea peak for urine samples, were uniformly binned with δ 0.01 resolution, where a summed spectral intensity is calculated for each bin. For plasma NMR spectra the spectral regions δ 0.00–4.60 and δ 5.10–10.00 were used and for urine NMR spectra, this was δ 0.60–4.70, δ 5.05–5.50 and δ 6.25–10.00. Data quality control was performed by identifying and removing outliers from an initial investigation with principal component analysis. Samples were only omitted if they occupied an extreme value in score space and have an identified technical or biological reason, for example poor water peak suppression or lower than acceptable S/N. The resulting spectra were subsequently normalised with probabilistic quotient normalisation, and variables were mean-centred and scaled to unit variance. The data set sizes and SA-K-OPLS model details can be found in table 4.2.

4.5 Results and discussion

4.5.1 Prediction performance and optimisation of the kernel parameter

A summary of prediction performances, AUC for discriminant analysis and q^2_y for regression models, and optimised Gaussian kernel parameter (σ) values for each data set can be found in table 4.2. Detailed results for each data set are presented in the following sections. For the current data set type, scaling and size, the kernel parameters had comparable orders of magnitude, but this cannot be assumed in general (it is possibly related to the similar NMR spectroscopy acquisition procedures and data processing), and therefore these values should not be used as a future guideline. The scaling of the data will have a large influence on both the chosen kernel parameter as well as the q^2 metric; therefore, a kernel parameter found optimal for mean-centred data is most likely not appropriate for data scaled to unit variance, and the predictive ability of models built on differently scaled data sets may also differ.

It should be noted that simulated annealing is a stochastic optimisation method, and therefore the end result of every optimisation will vary slightly. To ensure that the prediction performance estimates were reliable for comparison, and not unduly effected by random variability relating to the cross-validation procedure, the convergence of the prediction performance metrics during cross-validation was monitored, see figure 4.3.
Table 4.2: Details of the evaluated data sets: the prediction of responders in galactosamine toxicity (GalN), the prediction of time after intoxication with mercuric chloride (HgCl$_2$) for a ‘high dose’ level and two dose levels and the classification of infection with *T. brucei brucei* parasite (*TBB*) and the time after infection with *TBB* were modelled (using discriminant analysis (DA) or regression (RE) models). The Gaussian kernel parameter ($\sigma$) of the full model is shown. The prediction performance is given in terms of AUC for discriminant analysis and $q^{2}_y$ for regression models. The number of orthogonal components (#OC) is indicated for both OPLS and SA-K-OPLS models.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Biofluid type</th>
<th>Model type</th>
<th>Size [n x k]</th>
<th>Response variable</th>
<th>$\sigma$</th>
<th>prediction performance</th>
<th>#OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GalN -responders</td>
<td>urine</td>
<td>DA</td>
<td>159 x 830</td>
<td>2 classes</td>
<td>20.06</td>
<td>OPLS</td>
<td>0.834 5</td>
</tr>
<tr>
<td>HgCl$_2$ -high dose</td>
<td>urine</td>
<td>RE</td>
<td>60 x 830</td>
<td>9 time points</td>
<td>21.62</td>
<td>OPLS</td>
<td>0.615 4</td>
</tr>
<tr>
<td>HgCl$_2$ -two doses</td>
<td>urine</td>
<td>RE</td>
<td>123 x 830</td>
<td>9 time points</td>
<td>24.28</td>
<td>OPLS</td>
<td>0.584 4</td>
</tr>
<tr>
<td><em>TBB</em> -classification</td>
<td>plasma</td>
<td>DA</td>
<td>108 x 950</td>
<td>2 classes</td>
<td>20.29</td>
<td>OPLS</td>
<td>0.956 2</td>
</tr>
<tr>
<td><em>TBB</em> -time</td>
<td>plasma</td>
<td>RE</td>
<td>64 x 950</td>
<td>6 time points</td>
<td>22.66</td>
<td>OPLS</td>
<td>0.640 3</td>
</tr>
</tbody>
</table>

The prediction results are shown in figure 4.5, where the prediction improvement of SA-K-OPLS over its linear equivalent OPLS is visible for all models. For certain metabonomic applications, predictive performance is of paramount importance, such as in toxicology and disease diagnostics. Although linear models often provide a basic level of prediction performance for multivariate data sets, in the presence of substantial non-linear variability in data, non-linear models are expected to provide improved predictions compared to linear models. This is explained by the potential ability of a more flexible non-linear model to incorporate additional predictive information if the predictor and response variables are related non-linearly. The aggregation of information across many weakly associated descriptor variables, linearly and non-linearly related to $Y$, also contributes to an improved, and potentially also more robust, prediction model. For the data sets analysed here, it was observed that predictions by SA-K-OPLS were better than, or equivalent to, the linear OPLS model. However, the additional assumption of linearity in the OPLS model may result in better prediction performance under certain specific conditions, for example if the problem is truly linear and the available training data are very limited in the number of observations.
Figure 4.5: The differences of the prediction performance results for SA-K-OPLS and OPLS models were evaluated using 100 outer cross-validation rounds (mean and the standard error of the mean, representing the variability in the 100 cross-validation results, are indicated as bars with error bars). Regression analysis results from the mercuric chloride (HgCl$_2$) data sets and the *T. brucei brucei* (TBB time) time after infection data were evaluated with $q^2_y$. Discriminant analysis results are presented with both the mean sensitivity and the area under the receiver operating characteristic curve (AUC), for the separation of infected versus uninfected samples from the *T. brucei brucei* infection study (TBB class.) and for classification of non-responders versus responders in the galactosamine toxicity data set (GalN resp.).
4.5.2 Prediction of non-responders in the galactosamine toxicity study

Galactosamine has been widely used as a toxin to model hepatitis, although numerous studies have identified significant variability in response to this hepatotoxin.\textsuperscript{140} In drug discovery, the ability to prospectively identify and account for non-responders would greatly reduce the number of patients needed in clinical trials.\textsuperscript{186} Attempts have been made to predict non-responders from their urinary metabolic profiles prior to,\textsuperscript{33,34,187} or soon after dosing\textsuperscript{188} in pharmacometabonomics. Non-responders (n = 40) and responders (n = 119) to galactosamine, as determined by histopathological and clinical chemistry markers of necrosis, were classified from the baseline (i.e. pre-dose) urinary $^1$H NMR spectra using SA-K-OPLS discriminant analysis.

The discrimination between the responder (circles) and non-responder (squares) classes is clear in figure 4.6, and an interesting, initially unexpected source of variation is visible on the orthogonal component, which could be associated with the collection period (urine was collected from -48 to -40 hr and -24 to -16 hr, whereas two other fractions represented collections from -40 to -24 hr.

![Figure 4.6: A score plot of the predictive and first orthogonal component of the galactosamine toxicity data set shows a clear discrimination in the predictive component (x-axis) between responders and non-responders (circles and squares, respectively). The first orthogonal component (y-axis) shows a trend related to diurnal variation, as inferred from the urine collection time-period (either -48 to -40 hr and -24 to -16 hr: 8 hours (day) collection; and -40 to -24 hr and -16 to 0 hr: 16 hours (night) collection).](image-url)
and -16 to 0 hr). It is visible that the 16 hours (night) collections have lower values on the first orthogonal component (\(oc_1\)), and the 8 hours (day) collections appear to have a higher value on \(oc_1\). The orthogonal variance is therefore likely to reflect diurnal variation, which has been shown to exert a subtle impact on metabolite profiles of urine.\(^{94,189}\)

An increased prediction performance is observed with respect to the linear OPLS model: the mean sensitivity increased from 0.75 to 0.79, and the AUC from 0.84 to 0.88, using 5 and 3 orthogonal components for OPLS and SA-K-OPLS, respectively, see figure 4.5. The ROC curve of SA-K-OPLS is constantly higher than the OPLS ROC curve, see figure 4.7, indicating the overall high performance of the non-linear approach for different specificity and sensitivity choices.

\[\text{Figure 4.7: The receiver operating characteristic curves for the galactosamine toxicity study.}\]

The difference between K-OPLS and KPLS is analogous to the difference between OPLS and PLS: K-OPLS provides a separate set of predictive and \(Y\)-orthogonal model components, a property that caused OPLS to gain great attention for analysis of metabonomic data.\(^{56}\) For comparison, the equivalent KPLS model was made using the same kernel parameter \(\sigma = 20.06\) and 4 predictive components (equal to the 1 predictive and 3 orthogonal components in SA-K-OPLS); the KPLS code was provided by Dr. R. Rosipal (Medical University of Vienna, AUT). The score plots from this model are shown in figure 4.8, and do not allow this clear observation and separation of the diurnal effects as orthogonal variation; moreover the separation between the two classes is not solely on the first component, as was the case for SA-K-OPLS. This demonstrates one of the unique features of K-OPLS compared to other kernel-based methods: the ability to capture and interpret non-linear structured variation in a data set that is unrelated to the response variable (\(Y\)). \(Y\)-orthogonal variation may be interpreted as multivariate latent variables that describe unspecified covariate variation, for example unknown batch effects, animal variability or drift in analytical instruments over time. Interpretation of orthogonal variation thus provides additional means for quality control and understanding of data and model. For this data set, similar diurnal orthogonal variation is observed in the OPLS model, which had a lower predictive ability, see figure 4.8.
External validation

Prediction performances reported here were calculated by nested cross-validation, which reduces the risk of over-fitting when optimising the kernel parameter. If obtainable, an external test set should ideally be used to evaluate prediction performance, however the nested cross-validation is expected to give an unbiased estimate of the predictive ability if no external test data are available. An example of the prediction performance of SA-K-OPLS on an external data set was generated by splitting the galactosamine toxicity data set into a modelling and an (external) validation set. Thus, the nested cross-validation was used to create a model based on 120 observations, and the 39 removed samples from the validation set were predicted using this model. Also in this instance, the prediction improved for SA-K-OPLS compared to OPLS, see figure 4.9. Of course, it would be preferred to validate the model with an external test set that was acquired on a different occasion, to fully test the robustness of the method, but these data are not always available.
Figure 4.9: The classification results of non-responders versus responders in the galactosamine toxicity data set were evaluated using a model solely built on 75% of the data (\(n = 120\)), the ‘Training data’. The results are presented for both the mean sensitivity and the AUC. The OPLS model was based on 5 OC, which gave the maximum AUC for the training data (calculated with nested cross-validation), and for SA-K-OPLS 2 OC were used. The mean and standard error of the mean of the 100 outer cross-validation rounds are shown, representing the variability in prediction performance of the training data set. Results for the ‘External validation data’, the remaining 25% (\(n = 39\)), indicate large improvements for both mean sensitivity and AUC for the SA-K-OPLS model compared to the linear modelling equivalent, OPLS.

4.5.3 Time-course modelling of mercuric chloride toxicity

Mercuric chloride is a well-studied nephrotoxin targeting the proximal tubules of the kidney.\(^{190, 191}\) A single dose induces a reversible lesion demonstrating a clear dynamic profile of metabolic consequences progressing to maximum insult and then recovery. Urinary \(^1\)H NMR data were used to predict the time after dosage with mercuric chloride (high dose level). This is of relevance for toxicology, because a good delineation of the time-course describing the onset, evolution and regression of a toxic lesion is of value for comparison with the gold standard histopathology approach, to identify fast and slow responding animals, and for selecting the most appropriate sampling times for other assays such as transcriptomics.\(^{192}\) The prediction of time after dosing is improved using the non-linear kernel method compared to linear modelling: \(q^2_y\) is 0.67 compared to 0.61 (4 orthogonal components for both OPLS and SA-K-OPLS), see figure 4.5. Interpretation of the orthogonal component reveals inter-animal variation in response: figure 4.10 A shows that animal 28 exhibits a different response compared to the other animals, and animal 29 also diverges from the other animals at later time points.
Figure 4.10: (A) A score plot of the prediction of time after dosage with a high dose of mercuric chloride based on metabolic changes was used to delineate the onset and regression of a toxic lesion. The predictive score component ($x$-axis) describes variability related to the $Y$ variable (time), while the orthogonal score component ($y$-axis) captures some systematic metabolic changes occurring at predominantly 24–48 hours after dosing. Interestingly, animal 28 (squares) and later animal 29 (diamonds) diverge from the behaviour of the other animals over time. (B) A score plot of the prediction of time after dosage with ‘low’ and ‘high’ doses of mercuric chloride. The orthogonal score component ($y$-axis) captures systematic metabolic differences between the low and the high dose group. Animal 28 and later animal 29, which diverge from the behaviour of the other high dose samples over time, appear to recover in a path similar to low dose samples.

Figure 4.11: A principal component analysis of serum clinical chemistry data acquired at 24 and 168 hours post-dose (mean-centred and scaled to unit variance) of the animals in the mercuric chloride toxicity study demonstrates that the high dose animals cluster separately from the low dose animals, except the data acquired from animal 28, and the later time point of animal 29; these findings are in agreement with the results in the orthogonal component of the SA-K-OPLS models.
A regression model was also constructed where data from the low and high dose levels of mercuric chloride were combined. Thus, this model is focussed on the common time-related metabolic response across toxin dose-levels, allowing prediction of the time from dosing, in analogy with clinical situations where the dose level and exposure time may be unknown. Also for this model, the prediction performance increased for SA-K-OPLS compared to OPLS (see figure 4.5, $q^2_y$ is 0.66 compared to 0.58). Moreover, the aberrant observations for the two animals 28 and 29 in the high dose group discussed above now cluster with the low dose animals in the score plot, see figure 4.10 B, indicating that they show a less severe response to the toxin.

Figure 4.12: Score plots for KPLS and OPLS models for the two mercuric chloride data sets, the colour-coding is identical to figure 4.10. (A) The score plots of the KPLS model for the high dose mercuric chloride toxicity study, using the same kernel parameter and number of components as the SA-K-OPLS model. (B) Score plots of the KPLS model on the mercuric chloride ‘both’ doses data set, using the kernel parameter and number of components that were used for the SA-K-OPLS model. (C) A score plot showing the predictive and first orthogonal component of the OPLS model for the high dose mercuric chloride data set. (D) A score plot showing the predictive and first orthogonal component of the OPLS model based on the combined low and high dose mercuric chloride data set.
The trends found for animal 28 and 29 of the high dose group were confirmed independently in the serum clinical chemistry data, where at 24 and 168 hours after dosing animal 28 showed lower levels of typical markers such as creatinine, gamma-glutamyl transpeptidase, urea nitrogen and phosphate than the rest of the high dose group, and similar effects were observed for animal 29 at 168 hours after dosing. The disparate metabolic behaviour of the identified ‘outlier’ animals is also clear from a PCA model created on the clinical chemistry data measured in serum, see figure 4.11. It should be noted that the spectra were based on urinary samples, so both biofluids demonstrate the altered metabolism of the animals. Thus, these animals were correctly identified as having a different response, based on the orthogonal variation in the SA-K-OPLS score plot, without prior knowledge or consultation of the metadata.

This example confirms that SA-K-OPLS not only improves the predictive ability for these two data sets compared to OPLS, but moreover, unlike other non-linear kernel-based methods, provides a visual interpretation of the variation in the data set, in terms of predictive variation as well as structured Y-orthogonal variation. The orthogonal variation information is not observed for the corresponding KPLS models, see figure 4.12 A and B where the score plots are shown for the high dose (5 components) and both doses (4 components) data sets (using \( \sigma = 21.62 \) and 24.28, respectively). Interestingly, a clear interpretation of the orthogonal variation was not possible in the equivalent OPLS models, see figure 4.12 C and D.

4.5.4 Classification of *T. brucei brucei* parasitic infection

In a third exemplar data set, the \(^{1}\text{H} \text{NMR}\) data of plasma samples from a *T. brucei brucei* parasitic infection (African sleeping sickness) in a murine model were evaluated. Spectra were acquired from plasma samples of mice before and at various stages during the infection and similarly from control animals. Data from infected animals formed the ‘infected group’; the ‘uninfected group’ consisted of the combined pre-infection and control animal data. The discrimination between the infected and uninfected groups resembles the situation in a clinical setting, where it is crucial to establish the occurrence of infection, irrespective of the time elapsed after initial infection. Improved classification accuracy between infected and uninfected individuals with SA-K-OPLS was observed compared to linear modelling, see figure 4.5 and table 4.2, with an increased mean sensitivity (SA-K-OPLS: 0.92 with 1 OC, OPLS: 0.88 with 2 OC), as well as a larger AUC (0.97 compared to 0.96). The ROC curves for these two models intersect, see figure 4.13, and thus there is a small region of specificity and sensitivity where OPLS would perform better than SA-K-OPLS.

In figure 4.14 A, it is apparent that the early time point of infection (day 1) is more difficult to discriminate from the uninfected spectra than the later time points (this was also observed for the OPLS and KPLS models, see figure 4.15 A and B), which can be expected from the development of a parasitic infection.
4.5.5 Prediction of \textit{T. brucei brucei} parasitic infection time-course

The status of the \textit{T. brucei brucei} infection in terms of the time after contraction is highly relevant as the nature of therapeutic intervention and the prognosis for recovery are greatly dependent on the event of blood-brain barrier crossing of the parasite that occurs in this infection. As with the mercuric chloride toxicity data set, a metabolic trajectory reflecting the progression of the infection can be characterised using spectral profiling.\textsuperscript{185} To stage the infection time point, the time after infection was modelled non-linearly as a regression, and prediction results are displayed in figure 4.14 B; the corresponding OPLS and KPLS models are shown in figure 4.15 C and D. It was found that the prediction was improved for the SA-K-OPLS model compared to the results from the linear OPLS model: $q^2_y$ increased from 0.64 in OPLS to 0.68 in SA-K-OPLS, both with 3 orthogonal components, see also figure 4.5 and table 4.2.
4.5.6 Visualisation of variable importance in K-OPLS models

In contrast to linear methods, there are no model loadings accessible for interpretation of variables in the case of kernel-based methods, due to the usage of the kernel trick. The non-linear mapping to the high-dimensional feature space removes the direct relation between model scores and loadings. This may be a disadvantage of kernel-based non-linear modelling but the importance of this potential disadvantage will depend on the application. For example it could be a problem in studies where biomarker identification was the main goal, but in studies where maximum predictive performance is imperative, for example in toxicology and disease diagnostics using multivariate biomarker panels, the improved predictive ability is a major advantage.
Evaluation of kernel-transformed pseudo-samples in a PCA biplot

As the visualisation and interpretation of important variables could greatly aid biological understanding, an attempt is made to approximate the variable importance for the $TBB$ classification model. The sensitivity of a model to variations in a selected variable is analysed by means of a ‘pseudo-matrix’ $X_{ps}$, representing the effect of the variation of one variable, see §4.3.7. This method was recently introduced to achieve a similar purpose in SVM modelling, and is based on the concept of biplots, where loadings and scores of a PCA model are simultaneously displayed; it is proposed that the change of a score value as a function of variation of a variable is a measure of the variable importance. This property has been proven for linear models, and it is postulated that the same approach also gives a measure of variable importance for non-linear models. The visualisation of this PCA-based variable importance metric is discussed here for K-OPLS models, and two similar metrics of variable importance are introduced.

The scores of a PCA model on $X$, shown in figure 4.16 A, do not discriminate between the infected and uninfected groups on PC 1 and 2. When a PCA model is calculated based on the mean-centred kernel-transformed descriptor data, the two classes still show a large degree of overlap, see figure 4.16 B. However, the scores of a PCA model from the mean-centred kernel-transformed spectral data matrix after orthogonal signal correction ($K_c$) separate the infected and uninfected classes more effectively, as is indicated with the dotted line in figure 4.16 C. Interestingly, the majority of the samples that are misclassified are from time point 1, which was also observed for the OPLS, KPLS and K-OPLS models.

The importance of the different variables in this PCA model of the kernel-transformed data is subsequently evaluated with a kernel-transformed pseudo-matrix $K_{ps}$, consisting of pseudo-samples, to represent the variation of one variable as described in §4.3.7. The scores of $K_{ps}$ in the PCA model are calculated, and connected to form a trace. This process is repeated for each variable (which is represented by its own pseudo-matrix), and the resulting score trajectories are interpreted in a similar manner as loadings in a linear model. The score trajectories are shown in figure 4.16 D, where each line corresponds to a series of connected PCA scores for one variable. The colouring is related to the variable importance, which was determined by the range of the projection of the score trajectory on the (solid) class-separating line that was determined in figure 4.16 C: the maximum distance between projections of the twenty pseudo-samples in each pseudo-matrix on the discriminant line determines the variable importance. Thus, variables with a larger component on the solid line (which is mostly in the direction of PC 2), are shown in red and are considered most responsible for the visible score patterns and class separation in figure 4.16 C.
Figure 4.16: (A) Scores from a PCA model of the descriptor data scaled to unit variance X (for colour-coding of symbols, see figure 4.14). (B) Scores from a PCA model of the mean-centred kernel matrix before orthogonal signal correction. (C) Scores from a PCA model of mean-centred kernel matrix after orthogonal signal correction $K^c$. A separating plane between the infected and uninfected samples is indicated with the dotted line (calculated with linear discriminant analysis, classify in the MATLAB statistics toolbox). The vector orthogonal to this discriminant axis is shown as a solid line. (D) The pseudo-sample trajectories for each variable are colour-coded according to the variation in the direction of separation (black line): the most ‘important’ variables are represented as red lines and are visualised in a spectral form in figure 4.17 A.

Evaluation of three pseudo-sample based variable importance metrics

The PCA biplot method was extended with two other approaches to form a set of three related variable importance metrics, all based on this concept of ‘pseudo-matrix’ scores, that were used to evaluate the variables in the TBB classification model:

- Two PCs showing discrimination between the groups in a PCA model of $K^c$ are found. The scores of $K^c_{ps}$ in this PCA model are evaluated to find the variables showing the largest variance in the direction of group separation, see figure 4.16 and 4.17 A. This is based on the property of PCA that loadings and scores can be superimposed in a biplot.\textsuperscript{184,193,194} The kernel matrix that was used is based on the kernel parameter and the number of orthogonal components of the original SA-K-OPLS model.
To make a predictive model, rather than an unsupervised model, PLS is used to predict the two classes based on $K^c$. Subsequently, the score values of $K^c_{ps}$ are predicted by this PLS model. The scores on the predictive component are used to evaluate the ‘sensitivity’ of this model to variations in the respective variable, see figure 4.17 B. As the kernel matrix was already corrected for orthogonal variation and it is a two-class problem, only 1 predictive component and no orthogonal components were used.

The score values of the pseudo-matrix $X_{ps}$ are predicted from the K-OPLS model, established in §4.5.4, to give an indication of the effect of changes in a variable to the score value on the predictive component, see figure 4.17 C. This variable importance metric makes use of the fact that K-OPLS provides score values, unlike SVM, for which the non-linear biplot model was necessary.\(^\text{184}\) However, there is no theoretical ground to assume that variables found important by this metric are related to the model predictions in a similar way as loadings are for linear models, since the effect of the kernel transformation is not well-defined.

For the PCA-approach, a linear discriminant analysis is performed to find the direction of separation between the two classes. The range of the projections of the pseudo-samples for each variable on the discriminating line (the solid line in figure 4.16 D) is used to evaluate variable importance. The discriminating plane of the PCA-based approach is not always in the first PCs: for the galactosamine toxicity model, the best separating plane was a linear combination of PC 20 and 21. For the galactosamine model, a clear effect of diurnal variation on the PCA model was noticed, which confirms that the kernel-transformed matrix after orthogonal signal correction should be used. For the PLS and K-OPLS models, the range of the scores on the predictive component was used to evaluate variable importance.

A spectral representation of the different metrics is shown in figure 4.17, where the mean spectrum is colour-coded with variable importance. It should be noted that these approximations are not true loadings, and are unable to take interactions between variables into account. The resulting values can be viewed as an indication of the ‘sensitivity’ of the model scores to changes in a variable. The absolute values of the loadings ($P = X^T_p/T_pT_p^T$) of the OPLS model are included in figure 4.17 D for comparison. Note that the sign of the loadings from the OPLS model would give an indication of the up- or down-regulation of a metabolite, whereas the kernel method allows the modelling of non-linear patterns that have no clear directionality (see e.g. the parabolic shape of some of the score trajectories in figure 4.16 D), and thus only provides an unsigned importance metric.

From figure 4.17, it is clear that the three related pseudo-sample approaches give very similar results in terms of variable importance. This could be the result of the similar nature of the three methods, which all project the pseudo-samples in a non-linear model, using the same kernel-transformation. There are only minimal differences between the variable importances of the linear and non-linear approaches, including the slightly increased importance of variables in the OPLS
Figure 4.17: A colour-coding representing the importance of the variables is projected on the mean spectrum of TBB study samples; the colour-coding for all plots ranges from blue (unimportant) to red (most important). (A) The variable importance is based on the range of projections of the *PCA* score trajectories of the pseudo-samples on the discriminating axis shown in figure 4.16 D. (B) A colour-coding representing the change in predictive score value of the pseudo-data based on a *PLS* model of the kernel-transformed data. (C) A colour-coding representing the change in the score on the predictive component when predicting the pseudo-samples with the previously generated *K-OPLS* model, see §4.5.4. (D) The loadings of the equivalent *OPLS* model (with two orthogonal components, see §4.5.4) from the original, i.e. not kernel-transformed, spectral data.
model in the range $\delta$ 6.8–8.8 compared to the non-linear models, and vice versa around $\delta$ 1.90, $\delta$ 2.43 and $\delta$ 2.91. Although the pseudo-sample approach needs to be further validated, a comparison of the linear and non-linear models confirms that signals important in linear models typically remain useful in non-linear modelling with K-OPLS.

The approach of pseudo-samples to determine important variables in non-linear models is not restricted to K-OPLS discriminant analysis and SVM: once it is verified to be an accurate measure of variable importance, the method could also find use in other non-linear projection models such as KPLS, and it could be extended to aid the interpretation of regression models. The pseudo-samples were created in the range of the original variable, which is expected to have a large effect on the projection trajectories if the data would not have been scaled to unit variance. In this situation, it could be worthwhile to normalise the trajectories in order to correct for the overall size of the trajectory.

Summarising, three different approaches to indicate variable importance for this non-linear model based on pseudo-matrices are suggested, and they show large correspondence with each other and the loadings of OPLS, indicating that this could be a viable way to increase the transparency of SA-K-OPLS models.
4.6 Conclusions

Currently, predictive modelling in metabonomic studies is almost exclusively performed using linear methods. Here, it is demonstrated that a non-linear kernel-based model can provide valuable prediction performance improvements compared to the linear model in three typical metabonomic data sets. The kernel-based OPLS model (K-OPLS) was used in this study, which is analogous in its model structure to the commonly used linear OPLS model. It was found that the non-linear (K-OPLS) prediction performance is better than or equal to the linear OPLS model. K-OPLS provides added value compared to alternative kernel-based models by enabling separate modelling of predictive and Y-orthogonal model components, facilitating improved model interpretation and providing a means for quality control. The introduced variable importance approximations could potentially provide additional biological interpretation of non-linear models.

A complicating factor for non-expert users in the application of kernel-based models can be the optimisation of the kernel function parameter. It was demonstrated how simulated annealing can be applied successfully to automate optimisation of the kernel parameter, providing a user-friendly procedure based on nested cross-validation. Thus, it was shown that SA-K-OPLS provides a flexible framework for non-linear classification and regression modelling and it is anticipated that the method can also be employed successfully in studies with data collected by other analytical methods and for modelling of other -omics data types, as well as high-dimensional and potentially non-linear data in general.

These results should facilitate increased utilisation of non-linear prediction models in metabolic profiling studies, especially in application areas such as disease diagnostics and toxicology, where prediction performance is of utmost importance. The SA-K-OPLS algorithm applied here is available in the K-OPLS software packages for R and MATLAB, including cross-validation procedures, kernel-optimisation and calculation of prediction performance metrics.
Chapter 5

Multivariate analysis of mass spectrometry imaging data

5.1 Aims and objectives

The previous chapters described methods for processing and modelling of NMR spectral data that summarise global profiles of tissues or biofluids. An exciting new development with great potential for metabonomic studies, is the acquisition of localised molecular spectra with mass spectrometry imaging (MSI). An MSI data set consists of a series of mass spectra obtained from various sequential locations forming a grid. Thus, the data can be interpreted as a full mass spectrum at a given spatial point (pixel), or as a map of a given substance level over a two-dimensional set of pixels. In terms of modelling, the previous chapters already explored the modelling of spectral data, and the MSI data additionally required to take the image information into account. Therefore, sophisticated approaches are needed and the following steps are discussed in this chapter to address different aspects of MSI data, in order to extract the maximum amount of useful information from the acquired hyper-spectral data:

1. Develop and evaluate various approaches to processing MSI data;

2. Implement and discuss a correlation-based measure of spectral similarity of different pixels;

3. Perform principal component analysis and evaluate the distribution of pixels in the score plots and the mass spectrometry image;

4. Compare the distribution of pixels on self-organising maps with the localisation of these pixels in the image;

5. Visualise the global and local structure of the data in a few dimensions using non-linear dimensionality reduction based on manifold learning techniques.
5.2 Introduction

Mass spectrometry imaging (MSI) provides localised spectroscopic information about the contents of a sample, e.g. a tissue slice, across a grid, see §2.2.2.\textsuperscript{51,76–78,195,196} It reports a spatial biochemical topology with high molecular specificity, providing an \textit{ex vivo} view of tissue. MSI can be exploited to find tissue biomarkers or characteristic metabolic profiles of sample regions. For example, classification based on MSI spectral data is suggested as a method to obtain a digital staining.\textsuperscript{197}

MSI has been used to study proteins,\textsuperscript{77,198–204} lipids,\textsuperscript{205} metabolites\textsuperscript{206,207} and drugs,\textsuperscript{208,209} and has been performed on a large number of sample types, including plants,\textsuperscript{207} bacterial colonies,\textsuperscript{210} drug-treated tissue,\textsuperscript{209,211} rat brain, and human surgical specimens,\textsuperscript{77,212} including human cancerous tissue in breast\textsuperscript{204,213} and ovarian cancer.\textsuperscript{201} Even whole animals have been imaged as part of a drug study.\textsuperscript{208}

MSI data combine spatial resolution and spectroscopy; characteristic of this type of data is a high dimensionality (see table 5.1) and complexity, especially for biological samples. MALDI is a soft ionisation technique, and therefore there is little fragmentation, however a large number of adducts can be formed. Variable tissue height and differential ionisation patterns further complicate the obtained spectra. Data are typically evaluated with standard MSI analysis tools and software, such as BIOMAP,\textsuperscript{214,215} ClinProTools\textsuperscript{216} and others.\textsuperscript{217,218}

Analysis of single images (i.e. one \textit{m/z} value) is most commonly employed,\textsuperscript{77,208–210,213} and univariate analysis of selected peaks can be performed.\textsuperscript{212} The complexity of human biology and disease, however, is unlikely to be captured with this approach. Single or manually obtained biomarkers are probably not robust, as a plethora of different mechanisms will lead to subtle variations in observed molecular concentrations. Some work has addressed the multivariate nature of the mass spectrometry data in matrix-assisted laser desorption/ionisation (MALDI) MSI data sets, for example, principal component analysis (PCA) has been performed on selected MSI pixels\textsuperscript{201,219} or the complete sample.\textsuperscript{205,220–223} PCA can also be used to compare anatomical regions of different animals\textsuperscript{224} and to align images of multiple MSI data sets from one sample in a mosaic.\textsuperscript{225} PCA has further been employed to compress the data prior to clustering\textsuperscript{203,226} or to perform discriminant analysis of regions of interest.\textsuperscript{219,226} Classification of MSI data can also be performed with k-nearest neighbour methods,\textsuperscript{202} partial least squares,\textsuperscript{227} support vector machines,\textsuperscript{166,202,204,212,220} neural networks,\textsuperscript{204} random forests\textsuperscript{197} and hierarchical clustering.\textsuperscript{198,203,204}

Multivariate techniques have also been suggested and employed for other types of MSI data, for example generated by secondary ion mass spectrometry (SIMS).\textsuperscript{218,228,229} PCA followed by a VARIMAX rotation to yield physically realistic spectral loadings and increase image contrast was useful in an analysis of TOF-SIMS imaging data, but at the cost of chemical specificity.\textsuperscript{229,230} An alternative was suggested through the use of probabilistic latent semantic analysis,\textsuperscript{231} which also produces non-negative components.
Often, MSI is applied to analyse high molecular weight species, but the approach can be extended to metabolite data,\textsuperscript{206,207} and it is expected to be advantageous compared to ‘classical’ metabolonomic mass spectrometry through the combination of relatively high sensitivity with localised metabolic information. One problem is that the applied matrix in MALDI imaging causes peaks that interfere with the spectral profile. This is one of the challenges that is addressed and discussed in this thesis. Moreover, there has been little research that has taken the multivariate nature as well as the image nature of the MSI data into account. Therefore, the focus is on visualisation rather than classification, and the full data set is used rather than a small selection of sample regions and single m/z values. Intuitive and straightforward approaches for analysis, that retain the link between data modelling and biological interpretation by cross-referencing between the spectral data and the image, are explored in this work. The methods are demonstrated with a formalin fixed sagittal rat brain section, with spatial resolution 100 $\mu$m\texttimes{} 100 $\mu$m and from which mass spectra were acquired in the range m/z = 50–1000.

5.3 Materials and methods

The sample preparation and histology and MSI data acquisition were all performed by Claire Carter, University of Birmingham.

5.3.1 Materials

Alpha-cyano-4-hydroxy cinnamic acid ($\alpha$CHCA) and trifluoracetic acid (TFA) were purchased from Sigma-Aldrich (Dorset, UK). 10% neutral buffered formalin (NBF) was supplied by the Histology Department of the Medical School at the Royal Hallamshire Hospital (Sheffield, UK). Methanol was purchased from Fisher Scientific (Loughborough, UK). Whole control rat brains were supplied by Covance (Harrogate, UK). A TM Sprayer (automated matrix deposition system) was supplied by Leap Technologies (North Carolina, USA).

5.3.2 Tissue preparation for MALDI MSI of phospholipids

A whole rat brain was removed from the -80 °C freezer and allowed to thaw at room temperature. Once thawed, the brain was placed in 10% NBF and fixed for 48–72 hr. Fixed brains were bisected and half of each was mounted onto a cryostat chuck, using water-ice slush. 12 $\mu$m serial sections of each brain were thaw mounted onto stainless steel MALDI target plates for mass spectrometry and glass slides for histology. Tissue samples were coated in 5 mg/mL $\alpha$CHCA matrix material, prepared in 80% methanol (0.1% TFA) using the Leap TM sprayer. Plates were sprayed at 150 °C, 10 psi, a flow rate of 0.25 mL/min with a stage velocity of 500 mm/min. Each target insert was sprayed with 8 cycles.
5.3.3 Mass spectrometry imaging

All experiments were carried out on a QqTOF (Qstar Elite) mass spectrometer (Applied Biosystems, Foster City, USA), operated in positive ion reflectron mode. Instrument parameters had been previously optimised for the analysis of glycerophosphatidylcholines from rat brain. The Nd:YAG (355 nm) laser was operated at 20% available power (2.1 µJ) with a repetition rate of 500 Hz. The target plate stepping distance was set to 100 µm for both the x and y dimension using the imaging acquisition software. Data were acquired using the ‘dynamic pixel’ setting, which involves the laser being moved within the 100 µm² pixel area and resultant data being summed to give an accumulated mass spectrum over one second for each pixel area. The resulting data were converted from .wiff data format to ‘analyst’ files with the wiff-to-analyse script (supplied by MDS Analytical Technologies, Ontario, CAN), and the data were imported into BIOMAP image software, available from www.maldi-msi.org.

Mass peaks were identified based on published literature values for MALDI MSI and were confirmed by fragmentation studies observing the head group. Molecules were named according to the lipid maps nomenclature: PC is glycerophosphatidylcholine, SM is sphingomyelin, and the number of carbon atoms in the lipid chain is indicated followed by a colon and the number of double bonds; examples are [PC 34:1] and [SM 20:0].

5.3.4 Histology

For histological evaluation of tissue architecture, the adjacent serial section was stained with haematoxylin and eosin. A histology image of the stained sagittal rat brain section with several anatomical regions labelled is shown in figure 5.1.

![Histology Image](image.png)

Figure 5.1: A stained histology image from the rat brain, anatomical regions are indicated. CB: cerebellum; CC: corpus callosum; CTX: cerebral cortex; HP: hippocampus; HY: hypothalamus; M: medulla; MD: mid brain; OC: optic chiasm; P: pons; PG: pituitary gland; S: septum; TH: thalamus. Note that this image is rotated through 180° with respect to the other figures in this chapter.
5.4 Processing of MSI data

Here, the processing of the acquired MSI data from the formalin-fixed sample of rat brain tissue is described. The data were stored in a BIOMAP-generated file, which is a software program frequently used to analyse the MSI data.\textsuperscript{214,215} Although BIOMAP allows for various forms of image analysis and manipulation (region of interest analysis, segmentation, clipping, histogram display etc.), it is based on an image approach, where one \textit{m/z} value is evaluated at a time. This implies that > 1000 images have to be analysed if one wants to thoroughly evaluate the data set, which is not only very time- and labour-intensive, but also highly likely to miss relations between different \textit{m/z} values, because the multivariate aspect of the data is ignored. To process the data and perform multivariate analysis, the data were imported into MATLAB using the \textit{readanalyze} function (available from \url{http://www.mathworks.com/matlabcentral/fileexchange/3760}).

The requirement of processing MSI data has been emphasised previously,\textsuperscript{232,233} and the aim of this work is to discuss the necessary data processing steps. The processed data can then be analysed by various multivariate approaches, described in §5.5.

5.4.1 Binning, alignment and baseline correction

MSI data have an extremely high information density and the analysis of these imaging data is in its infancy. Binning, taking the sum of all signals over a range, may not be ideal for certain types of data (see chapter 3). However, the data were already binned to $\Delta \textit{m/z} = 0.2$ width when imported into the BIOMAP software. The use of unbinned data is not further pursued here for two reasons: the difficulty in obtaining the full-resolution spectra, and the increased data set size and concomitant increase in memory and computational burden. This is at the cost of identification of single molecules, as it is not unlikely that different molecular species are together contributing to the observed intensity of a bin, i.e. different signals can overlap within the bin width of $\Delta \textit{m/z} = 0.2$. The effects of binning can be re-evaluated at a later stage combined with the possible development of a data alignment algorithm, if necessary.

Calibration of MSI data can be necessary when large \textit{m/z} ranges are acquired\textsuperscript{212} or the spectral resolution of the \textit{m/z} bins is very high, but was not necessary for the current data set. From evaluation of the plotted spectral data, it was decided that baseline correction and alignment of the binned data are not necessary: a flat baseline was observed and it appeared that peak misalignment was not present, as isotope patterns were consistent and did not shift across the different pixels.
5.4.2 Unfolding of the data: from 3D to 2D

The MSI data can be thought of as an array, see figure 2.7 B1, either of various mass spectra at the different spatial positions (figure 2.7 B4), or as images for various m/z values (figure 2.7 B3). However, image analysis of single m/z values or the analysis of individual mass spectral profiles are univariate approaches, and in this chapter a different route is taken: the application of various multivariate approaches allows a thorough, unbiased overview of the data, taking the multivariate and collinear nature of the data into account (see also §2.3.2).

The two directional dimensions of the MSI data were ‘collapsed’ to form a large matrix where each row is the mass spectrum at a given x, y position and each column one m/z variable, see figure 2.7 B2. The nature of the MSI data is very similar in the x and y directions of the original data tensor, as both are identifiers of location, and therefore the unfolding step seems appropriate. Unfolding increases the speed of computations (e.g. compare parallel factor analysis (PARAFAC) and PCA) and reduces the complexity of the resulting models. Therefore, data analysis with multi-way methods, that would analyse the data tensor shown in figure 2.7 B1, is not further explored and the unfolded matrix [20535 × 4751] is used. The information of the x, y location of each spectrum is retained and used to reconstruct images.

5.4.3 Peak selection

Peak selection has been shown to be integral to obtain useful multivariate models of MSI data and here it is confirmed that peak selection is useful: it decreases the number of non-informative variables, and it reduces the data size and calculation times. Peak selection is not trivial because spectral information and quality varies between the pixels, so an easy definition that can be used to reject noisy variables is not directly available. A number of approaches have been evaluated, and here, two methods are applied consecutively, and both fulfil two important criteria. Firstly, a thorough selection of peaks is achieved, where only a small fraction of the original number of variables is retained. Secondly, the approaches were pragmatic and relatively intuitive to understand and calculate.

Approach A for peak selection: correlation with matrix peaks

Preliminary analysis of the data set demonstrated that some peaks are more prominent in the region surrounding the tissue sample than on the sample itself, and these therefore probably arise from the applied matrix solution used in MALDI. These peaks, although possibly informative, do not directly convey information on endogenous metabolites. Hence, as a first approach, it was decided to remove these peaks. One such peak at m/z = 172.0 is from the αCHCA matrix solution ([M - H2O + H]+, m/z = 172.04, C10H6NO2). The correlation of each m/z variable with this selected matrix peak m/z = 172.0 across all pixels was evaluated, because it is anticipated that peaks with a negative correlation with the matrix peaks are more prominent on the tissue than in
Figure 5.2: (A) The 10 peaks that were selected to represent the main matrix peaks (those with maximum covariance with \( m/z = 172.0 \)) are shown in red. The y-axis is the mean intensity, calculated across all 20535 pixels for the 4751 variables. (B) The selection of those peaks of which the summed correlation with the 10 selected matrix peaks is negative is shown in orange (1224 variables, approach A). (C) Peaks from the selection in B that fulfilled the second criterion, based on the variance explained in a PCA model, are shown in green (564 variables, approach B).

The region outside the tissue, and correspond to interesting variables. Conversely, peaks correlating positively with a matrix peak are more likely to be related to the matrix and other experimental settings, than endogenous metabolic variation. To correct for any chance correlations, 10 prominent matrix peaks were used, selected by calculating the covariance of each \( m/z \) value with the peak at \( m/z = 172.0 \). These 10 peaks displaying the highest covariance with the selected matrix peak are indicated in figure 5.2 A, where covariance was used rather than correlation in order to select high-intensity peaks only and to avoid selection of isotopes instead of different matrix molecules.

The correlations for each variable in the data set (4751 in total) with each of the 10 matrix peaks were calculated and summed; only those peaks with a negative sum of the 10 correlations were retained, shown in orange in figures 5.2 and 5.3. It is clear from examples in the different correlation regions, shown in figure 5.3, that positive correlations indeed correspond to variables...
Figure 5.3: The correlations of each variable with 10 selected matrix peaks, see figure 5.2 A, were summed, and the summed correlations were sorted. Note that the x-axis is arranged according to decreasing correlation with the matrix peaks, which is plotted on the y-axis, and does not to relate to the individual m/z values. Only peaks with a negative summed correlation are retained (coloured orange). Images of 3 selected variables demonstrate that positive correlations correspond to higher intensity outside the sample (m/z = 650.2), low correlations often correspond to non-informative peaks (m/z = 207.6), and large negative correlations show clear relevance to the biological tissue (m/z = 761.6).

with a higher signal intensity outside the tissue region than within the tissue, and are therefore unlikely to be biologically relevant (e.g. m/z = 650.2). Peaks with a low correlation, e.g. m/z = 207.6, are mostly representing noisy variables, and peaks with a negative overall correlation, e.g. m/z = 761.6, display a clear structure and distribution in the sample.

**Approach B for peak selection: variance explained in PCA on the image**

It is clear that the peak selection made with approach A, which was based on correlation with the matrix, could be improved, since noisy variables were still included (e.g. see m/z = 207.6 in figure 5.3). Although many multivariate approaches are able to cope with noisy variables, the model strength decreases with a large number of non-informative variables. For univariate methods, the effects of noise can be even more problematic. Therefore, it was proposed to identify the variables
in the selection of approach A that lack any relation to anatomy, and are likely to be noise. This was achieved using a PCA-based decomposition of each m/z image: the intensity values of the pixels for the selected m/z are represented in a matrix, where rows correspond to different y-locations and columns to the x-position in the sample. The variance explained in the first principal component of the PCA model (with only mean-centring) was used as an indicator of image-related intensity distribution of the variable. If the intensity differences are randomly distributed, the variance explained in the PCA model will be low, e.g. for m/z = 989.6, see figure 5.4. On the other hand, if there is any structure in the image, more variance is modelled with PCA, as is shown for m/z = 873.6 in figure 5.4.

To select a pragmatic and user-friendly cut-off, an h-index was used as an appropriate heuristic, and calculated as the sum of all explained variances divided by the number of original variables (1224). Variables that have explained variances in PC 1 higher than this h-index of 24.3% were retained and coloured green in figure 5.4. The expansions in figure 5.2 show that the variables selected with approach A but not retained in approach B are mostly corresponding to variables that have a lower mean intensity. However, this PCA-based approach avoids removing low-intensity, informative ions; or retaining artefactual high-intensity, noisy ions. A similar approach would be the investigation of structure in the image using an entropy-based criterion.

Figure 5.4: The variance explained for the first principal component of a PCA-based decomposition of the image for each variable (mean-centred) is plotted. Note that the x-axis corresponds to variables that were sorted with decreasing variance explained, and does not correspond to the m/z values. Variables with lower levels of explained variance contain less biological and anatomical relevance (compare m/z = 873.6 and m/z = 989.6). An h-index of 24.3% was used as a cut-off for variable selection: only variables with a higher percentage of variance explained were selected and coloured green.

The m/z images and values of the deleted variables verified that not many informative peaks were removed with approach B: if any signal was found at all, the related main isotopes were still selected, so information regarding the parent ion was not lost. Thus, a potential loss of information is not problematic, since there is a lot of redundancy in the data. Even if 10% of the potentially informative peaks were removed, it is unlikely that this would result in the disappearance of characteristic molecular fingerprints. Moreover, there are clear computational and interpretative
advantages for smaller and cleaner data.

It should be noted that a very similar variable selection resulted if the images were rotated through 90° (swapping the columns and rows for the PCA, i.e. a matrix transpose). If for any future sample, anatomical structures are expected to be directional, e.g. more horizontally or vertically oriented, this information should be used in the decision of transposing the data matrix before PCA-based variable selection. Columns with only 0 were removed prior to PCA calculations. Mean-centring of the data prior to PCA-decomposition is necessary as it removes the overall mean, which would otherwise be the main contributor to the variance in the first principal component. Unit variance scaling was not performed, as this would give equal weight to all columns, which could negatively affect the selection of small anatomical features and emphasise noise in the data. A high level of explained variance would also be obtained for m/z images that are high in the surrounding and low in the sample region. Therefore, the variable selection with approach A has to precede peak selection with approach B.

5.4.4 Pixel selection

After peak selection with approaches A and B, it was decided to only discard those pixels that clearly do not arise from sample regions, although certain studies only retain a few selected spectra. A fast method to select the sample pixels was found based on the total ion count (TIC) ratios of the informative peaks versus all original variables. The TIC is the sum of the intensity values for a pixel, and was calculated for all original 4751 variables, and the data set after variable selection with approaches A and B (564 variables). As can be seen from figure 5.5 A, the ratio of both TICs shows two classes of pixels (the colour-coding refers to the threshold that was used to select the pixels shown in figure 5.5 C). A plot of $\log_{10}\left(\frac{\text{TIC}_{564}}{\text{TIC}_{4751}}\right)$, as shown in figure 5.5 B, demonstrates that a higher TIC from selected variables compared to the total TIC corresponds to sample pixels. This can be explained by the relative increase of informative peaks in the sample region, compared to mostly peaks from matrix variables in the surrounding region.

Application of the manually chosen threshold $\log_{10}\left(\frac{\text{TIC}_{564}}{\text{TIC}_{4751}}\right) = -0.5$ results in the selection of the pixels coloured green in figure 5.5 A, and figure 5.5 C is obtained. Some discontinuities in the selected pixels are clear, but these are not problematic for data analysis (see also §5.4.7). The data set sizes of the data after unfolding, peak and pixel selection steps can be evaluated in table 5.1; the consecutive processing steps achieve a reduction of data set size, which decreases computational cost and increases model quality.
Figure 5.5: (A) Pixels from the sample were selected based on the ratio of the total ion count (TIC) from the 564 selected variables to the TIC from all 4751 variables. (B) An image coloured with \( \log_{10}(\frac{\text{TIC}_{564}}{\text{TIC}_{4751}}) \) illustrates that a relatively high value of TIC\(_{564}\) versus TIC\(_{4751}\) corresponds to pixels from the sample. (C) A threshold of \( \log_{10}(\frac{\text{TIC}_{564}}{\text{TIC}_{4751}}) = -0.5 \) results in the selection of the green-coloured pixels (the points coloured green in A).

5.4.5 Normalisation

Normalisation is the procedure whereby a correction is made for an experimental bias that would result in higher (or lower) signal intensities in some pixels than in others, for example as a result of uneven MALDI matrix coating or altered ionisation efficiency. Thus, the correction should address analytical and technical variation, rather than biological variation, and is widely known to be necessary for analysis of MSI data, especially to correct for differences in matrix solution coverage. MSI spectra are frequently normalised to total ion count (TIC), which is the sum of all signal intensities. Normalisation can also be performed with respect to a matrix peak. Other commonly used methods of normalisation in mass spectrometry and their influence on the identification of biomarkers in a non-image setting have been described, and in §2.3.1 various normalisation approaches for NMR data are discussed. Here, four normalisation methods are evaluated:

- A reference molecule of which the concentration is constant throughout the sample could be used to normalise the pixels. One variable that showed a fair degree of homogeneity,
and no distribution clearly related to anatomical regions (which would be of biological origin rather than analytical, and it is only the latter that needs to be corrected), is \( m/z = 749.6 \), see figure 5.6 A. Note that figure 5.6 A displays the peak intensities after pixel selection, however the complete image of this variable showed that the concentration is much lower in the surrounding region than in the sample region, implying this is a genuine biological signal. Usage of such a reference molecule is based on the assumption that there exists a molecule that would have an identical concentration in each pixel. This assumption seems unrealistic: even cell membrane components, such as certain phospholipids, will not be uniformly abundant as membrane composition, cell sizes and packing density can vary.

- **Normalisation to the TIC of ‘matrix-positively correlated peaks’** is a correction for uneven matrix coating, which is performed by the dividing by the sum of all variables that were deleted with approach A for peak selection, see §5.4.3. This normalisation factor is displayed in figure 5.6 B, where it is observed that edges have overall higher TIC values. Normalisation to matrix peaks is based on the assumption that more analyte is measured if there is more matrix signal present, as it is the matrix that co-crystallises with the analytes. However, it commonly occurs that the matrix crystallises without analyte, or that different analytes are distributed unevenly within the crystals,\textsuperscript{76,236} which causes this normalisation method to be inappropriate.

- **Normalisation to the TIC of all (original) data**, the division by the sum of all 4751 variables in this study, is another normalisation option, see figure 5.6 C. Although this is a commonly used normalisation method, no rationale for using this normalisation method was found in the literature. The problems observed for normalisation to matrix peaks are persistent in this approach: high matrix concentration without co-crystallised analyte, crystal inhomogeneity and differential ionisation efficiency will negatively affect the accuracy of this normalisation.

- **Normalisation to the TIC of ‘informative peaks’**, the division by the sum of the 564 peaks after variable selection (see figure 5.2 C), is displayed in figure 5.6 D. Here, the rationale is that biological constituents might have varying concentrations in different regions, but these will largely average out: the concentration of molecule A might be higher in tissue part 1 than 2, but that of molecule B might show a reverse pattern.

In figure 5.6, the factor by which each pixel would be divided is shown for the different normalisation methods, and it is clear that the different approaches will give drastically different results. Interestingly, a ‘halo’ effect is seen in figure 5.6 B and C, where a higher TIC is observed around the edge of the sample.
Figure 5.6: Four methods for normalisation of MSI data are evaluated and the factor by which each individual pixel would be divided is shown on a colour scale. (A) The intensity values of a ‘uniformly distributed’ reference peak (e.g. \( m/z = 749.6 \)). (B) The TIC from the matrix-related peaks (i.e. all 3527 peaks that were deleted in §5.4.3). (C) The TIC from all data (4751 variables), this is the most commonly performed normalisation. (D) The TIC from the 564 selected, ‘informative’ variables; the normalisation approach proposed here.

The differential co-crystallisation efficacy of the matrix is confirmed in figure 5.7 A, which invalidates the normalisation to the TIC of matrix-positively correlated peaks and the normalisation to the TIC of all (original) data. In this figure, a direct correlation between the TIC from the 564 ‘informative peaks’ and the TIC from all 4751 variables, including the matrix peaks, is observed for the majority of points, but a cloud of points above the diagonal indicates the discrepancy between the two measures. A similar TIC from informative peaks, but different TIC from all peaks was observed for the two indicated pixels, shown in the image of figure 5.7 B. To evaluate the nature of the TIC differences, the ratio between the peak intensities of the two pixels \( \frac{I_{\text{blue}}}{I_{\text{green}}} \) was calculated for the ‘informative’ variables (564 selected peaks), ‘all’ variables (4751 original peaks), and the ‘matrix’ variables (3527 peaks that correlated with matrix signal). Histograms of these peak intensity ratios are visible in figure 5.7 C, where only the lower 90% quantile is shown for
Figure 5.7: (A) The TIC from the ‘informative peaks’ (the sum of 564 variables) versus the TIC from all peaks (4751 variables) is plotted. A correlation between the two values is found for a large proportion of the sample pixels. However, some pixels show a relatively large TIC from all peaks compared to the TIC from informative peaks. Two points with a similar TIC from informative peaks but different TIC from all peaks were selected. (B) The pixels that were selected in A are shown in the image, the blue circle demonstrates a higher TIC from all data than the pixel highlighted with the green circle. (C) Histograms of the peak intensity ratios $I_{\text{blue}}/I_{\text{green}}$, for only the informative peaks (left), all data (middle), or the matrix peaks (the m/z values that were not-orange in figure 5.3, right). Only the 90% lower quantile is plotted for clarity and the median of the ratios is indicated in red. The peak intensity ratios were calculated based on the variables that were non-zero in both pixels: 556 (from 564) for the informative peaks, 2325 (from 4751) for all data, and 1329 (from 3527) for the matrix peaks. 

clarity, and the median is indicated in red. The histograms illustrate that the ‘informative’ peaks vary little between the two pixels, and this was confirmed by inspection of the spectral data. Thus, normalisation to the TIC of 4751 variables or the TIC of matrix peaks could possibly create a non-existing difference between the two pixels.

The normalisation based on informative peaks seems most robust to differential matrix co-crystallisation effects, matrix inhomogeneity and inconsistent ionisation efficiency. However, if two pixels had similar compositional profiles but with a two-fold concentration difference (i.e. the measured content is two-fold diluted between the two pixels), the result is that these profiles would be identical after normalisation to the TIC of informative peaks. It is deemed unlikely that
this phenomenon would be the result of a biological effect, if it would occur it is most likely of
analytical nature, for example as a result of matrix crystallisation without analyte, and therefore
correction would be beneficial, rather than problematic. Nevertheless, it should be noted that
different biological complexity, e.g. across different tissues, would negatively skew normalisation to
TIC of informative peaks, as well as normalisation to the TIC of all peaks.

Probabilistic quotient normalisation (PQN), discussed in §2.3.1, is frequently used to normalise
NMR data. The use of PQN in MSI data normalisation was deemed inappropriate due to the potential
lack of ‘common variables’ in mass spectra from different anatomical regions. This is prohibitive
for the application of PQN, which is based on the assumption that less than 50% of the variables
change across the different spectra. If a large number of variables are not shared between two
tissue regions, this would affect PQN a lot more than TIC; TIC normalisation, on the other hand,
would suffer from the presence of some high intensity peaks in one region that are not present
in other regions. Thus, normalisation based on TIC is more robust than PQN in terms of variable
molecular composition, but potentially less accurate, as occurrence of high-intensity peaks could
skew the normalisation.

5.4.6 Data transformation

The correlation between the mean and standard deviation of the peak intensity indicated that for
this MSI data set, the variance increases with peak intensity: the Spearman correlation between
the mean and standard deviation of the 564 variables is 0.9726. To stabilise the variance, the
data were transformed as follows: $x_{\text{new}} = \log_{10}(1 + x)$. The transformation should be performed
prior to smoothing (discussed in §5.4.7), since variance-stabilised data will improve smoothing
results. Logarithmic scaling increases the importance of lower intensity but structurally informative
variables in subsequent modelling steps. Other transformations of MALDI MSI data have been
suggested, including power, root or threshold transformation; however, choosing from the large
number of possible transformation functions and settings might be troublesome for an inexperienced
user.

5.4.7 Smoothing

A smoothing step was performed, based on the convolution of each m/z image with a smoothing
window, see figure 5.8, to reduce the effects of noise on the image. This procedure is performed
for each variable: a separately smoothed image results for each m/z. Smoothing of SIMS MSI data
with a Markov random field has been reported in a study of tumour xenografts.
Here, a Gaussian function was used to calculate the smoothing window, see equation 5.1, where 
\( x \) and \( y \) are the coordinates of the pixel in the window, and \( x_c \) and \( y_c \) are the coordinates of the window centre, which is located at the pixel that is being smoothed, see figure 5.8 B.

\[ C(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-x_c)^2 + (y-y_c)^2}{2\sigma^2}\right) \text{ if } |x-x_c| < \frac{l}{2} \text{ and } |y-y_c| < \frac{l}{2} \]  
(5.1)

The two parameters \( \sigma \) and \( l \) are the width of the Gaussian and the length of the convolution window, respectively, and \( C \) is the resulting convolution window, see figure 5.8 A. In this figure, four window functions are shown with \( \sigma = 1 \) or 2, and \( l = 7 \) or 15. The edges of the larger convolution windows have negligible contributions to the smoothed result due to the fast decay of the Gaussian with the chosen \( \sigma \) values.

The current implementation of the convolution results in the ‘removal’ of edges: only those pixels completely surrounded by selected pixels (see figure 5.5) in the applied convolution window are retained, see figure 5.8 C. This can be exploited to remove the edge regions which may suffer from artefacts, as indicated by high tic levels seen in figure 5.6 B and C, and resolves the discontinuities in the pixel selection (see §5.4.4). Comparison of the top and bottom rows in figure 5.8 C shows that the effect of \( l \) on the image is small, so the main decision factor for the width \( l \) is edge removal. The width was chosen as \( l = 7 \) here, which decreases the sample width by removing approximately 4 pixels around the edges (see table 5.1 for the final data set size). In cases where smoothing is not necessary, a convolution window with the identity transformation was used to remove edge pixels.

Figure 5.8: (A) Gaussian window functions with \( \sigma = 1 \) or 2, and \( l = 7 \) or 15. (B) An exemplar m/z image before smoothing. Each pixel is replaced with the smoothed value based on the convolution window, centred on the pixel (indicated with a circle); windows of width \( l = 7 \) and 15 are drawn. (C) The result of convolution of the image in B with each of the window functions shown in A. The settings \( \sigma = 1; \ l = 7 \) were used for this data set (top left figure).
The choice of $\sigma$ is based on visual inspection of the smoothed results for a variety of $m/z$ images, and was chosen to be $\sigma = 1$. This is a balance between higher $\sigma$ levels removing too much of the local variation, possibly ‘over-smoothing’ and blurring the image (compare the left and right images of figure 5.8 C), and lower values of $\sigma$ performing minimal smoothing, which results in highly pixelated images. It should be appreciated that the matrix crystal size and sample preparation can influence the necessity and optimal settings for smoothing. The effective pixel size or spatial resolution is reduced by application of smoothing, but is likely still comparable to or better than the ‘experimental spatial resolution’, which includes losses of spatial integrity as a result of sample preparation.$^{51}$

Smoothing enhances the spectral quality of pixels, for example, noise spikes are down-weighted and poor pixel quality is improved. Moreover, smoothing enhances the data set with local information; modelling on smoothed data therefore effectively incorporates a local component. It would be interesting to evaluate the results of a ‘smart smoothing’ procedure, where a quality criterion for each pixel is used in the smoothing weighting to obtain a high-quality data set. This criterion could for example be based on information from the TIC ratio discussed in figure 5.5, where lesser quality pixels would be down-weighted in order to avoid artefactual signals. It should be noted that edge-preserving smoothing methods could be used to improve the image sharpness,$^{237}$ but evaluation of different smoothing methods is outside the scope of this work, and sufficient smoothing results were obtained using the straightforward Gaussian window.

Table 5.1: The data set sizes after the various processing steps, including peak selection, pixel selection and smoothing.

<table>
<thead>
<tr>
<th>Description</th>
<th>Data set size</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data</td>
<td>$4751 \times 111 \times 185$</td>
</tr>
<tr>
<td>unfolded data</td>
<td>$20535 \times 4751$</td>
</tr>
<tr>
<td>peak selection A</td>
<td>$20535 \times 1224$</td>
</tr>
<tr>
<td>peak selection B</td>
<td>$20535 \times 564$</td>
</tr>
<tr>
<td>pixel selection</td>
<td>$13324 \times 564$</td>
</tr>
<tr>
<td>smoothing</td>
<td>$11439 \times 564$</td>
</tr>
</tbody>
</table>

5.4.8 Centring and scaling

The log transformation that was performed in §5.4.6 stabilises the variance, and hence the need for data scaling is greatly reduced. Mean-centring of the data was performed, and subtracts the average intensity for each $m/z$ variable.
5.4.9 Overview and generalisation of this approach

The steps described in the previous sections are summarised in figure 5.9, where the following methods are shown:

1. **Peak selection** is based on anti-correlation of biologically relevant peaks with the peaks from the applied matrix; this selects only those variables that are more intense in the sample region than outside the tissue section. Subsequently, peaks that do not represent (anatomical) structure are removed, based on the variance explained in a PCA model of each m/z image.

2. **Pixel selection** uses the ratio of TIC from all peaks (including the matrix peaks) to the TIC from the selected peaks.

3. **Normalisation** divides signal intensities by the TIC of the selected variables for each pixel.

![Diagram showing the 6 main steps executed in the processing of the rat brain MSI data set](image)

Figure 5.9: A flowchart showing the 6 main steps executed in the processing of the rat brain MSI data set, including peak and pixel selection, normalisation, data transformation, smoothing and centring. Note that all steps should be executed consecutively, changing the order for any of these steps can have a large effect on the resulting data set.
4. **Transformation** is performed by a \((1 + \log_{10}(x))\) operation on all data.

5. **Smoothing** using a Gaussian window function of width \(l = 7\), and a smoothing parameter \(\sigma = 1\) results in noise reduction and removes sample edges.

6. **Centring and scaling** was performed by mean-centring the variables, no scaling was performed.

The heuristic approach that was taken to process the MSI data may appear ad hoc. However, a systematic evaluation of all possible processing steps is difficult, firstly because there are many combinatorial options for all the processing steps (so an experimental design could prove useful), but most importantly there is no objective evaluation criterion: there is no clear metric that could be used to assess which of the different processing options is better. Therefore the pragmatic processing choices were based on rational arguments, such as pareto-efficiency for the \(h\)-index as a cut-off, and visual inspection. The order of the different processing steps is important: it is necessary to identify informative peaks in order to perform the pixel selection and normalisation steps, log-transformation has to be performed prior to smoothing, etc. One could argue that peak selection is not absolutely necessary, as some modelling methods are relatively insensitive to the presence of noisy variables. However, the memory and time requirements of modelling the MSI data are considerable, and a reduction in data set size makes the calculations significantly faster and the models easier to interpret.

The processing of MSI data of a different nature can use the flowchart shown in figure 5.9 as a guideline, but may require fine-tuning and tailoring of e.g. the pixel and peak selection steps. The data of a sample that was ‘fresh’, rather than formalin-fixed, could without further amendments be analysed with the presented processing and multivariate analysis tools, some resultant image visualisations are shown in appendix A.
5.5 Multivariate analysis of MSI data

The processing of the MSI data, as described above, resulted in a data set of 11,439 pixels × 564 m/z variables. Various methods can be used to evaluate these processed MSI data; here the use of correlation analyses and multivariate modelling with PCA are discussed, and connections between the image and the latent space of PCA are investigated. Subsequently, non-linear modelling with self-organising maps is introduced for the analysis of MSI data, where the distribution of pixels over these maps is visualised. Finally, data modelling with two examples of manifold learning techniques is reported and a critical comparison between MSI data and in vivo magnetic resonance spectroscopy data is made.

5.5.1 Correlation analyses of MSI data

Modified correlation coefficient for higher contrast in pixels similarity

To demonstrate similarity in spectral profile of a pixel with the remainder of the sample, correlations (r) are calculated between the pixel of interest and all other pixels. The result of this pixel correlation is shown in the top of figure 5.10 for a pixel in the hippocampus: grey and white matter areas show up as highly correlated and anti-correlated, respectively. Here, it is proposed to employ enhanced correlations to get a clearer contrast of the different levels of correlation between pixels, an approach previously used in the evaluation of pharmaceutical tablets. This is done by scaling the correlations between 0 to 1, representing the lowest and highest correlation, respectively, according to equation 5.2.

\[ r_{\text{scaled}} = \frac{r - \min(r)}{\max(r) - \min(r)} \]  

(5.2)

Subsequently, an ‘enhanced correlation’ \( r_{\text{enh}} \) is obtained via equation 5.3, where \( \alpha \) is a parameter governing the discrimination power: a value close to 0 implies high discrimination power, a value of \( \alpha \) close to 1 will give a correlation profile similar to the original correlation coefficients.

\[ r_{\text{enh}} = \frac{r_{\text{scaled}} - \min(r_{\text{scaled}})}{\max(r_{\text{scaled}}) + \alpha - r_{\text{scaled}}} \]  

(5.3)

This approach can be viewed as a non-linear transformation of the colour scale of the images, and will therefore not uncover new information, but enhance the contrast in the correlation figures.

The correlation enhancement was implemented for both Pearson and Spearman correlations, and the result is visible in figure 5.10. The top row shows the Pearson (left) and Spearman (right) correlations with the selected pixel, and a large number of correlated pixels are observed. The smaller plots indicate the evolution of the observed correlation profiles with increasing levels of \( \alpha \), where \( \alpha = 1 \) is almost equivalent to the normal correlation profiles, highlighting all grey matter, and lower levels of \( \alpha \) accentuate the hippocampus. This is exactly the use of correlation enhancement: small differences in correlation patterns are emphasised, to allow more differentiation between otherwise highly correlated regions. Notably, the Pearson and Spearman correlation patterns are
Figure 5.10: The Pearson (left) and Spearman (right) correlations of image pixels with a selected pixel (centre of white circle in the hippocampus). The smaller maps show images of the enhanced correlations for different levels of $\alpha$, see equations 5.2 and 5.3, for both the Pearson and Spearman correlation data.
very similar, but Spearman correlations need a higher $\alpha$ value to achieve the same pattern, e.g. compare $\alpha = 0.2$ for Pearson with $\alpha = 0.5$ for the Spearman correlation.

Figure 5.11 A shows the frequency histogram of the Pearson correlations from the pixel selected in figure 5.10 before correlation enhancement. The correlation was enhanced with $\alpha = 0.01$ to create the histogram shown in figure 5.11 B, and an expansion is shown in figure 5.11 C. It is clear that the number of highly positively correlated peaks is greatly reduced after correlation enhancement, in correspondence with the analogous image for $\alpha = 0.01$ in figure 5.10.

**STOCSY and the quality of MSI data**

Correlation analysis of MSI data has previously shown the ability to relate different $m/z$ values based on co-localisation, and it was found that results of correlation analysis are greatly dependent on data processing steps, such as smoothing, baseline subtraction and normalisation. The use of statistical total correlation spectroscopy (STOCSY, see §2.3.7) to relate different molecular compounds was briefly evaluated for this MSI data set, as similar spatial distributions of isotopes and fragments could help identification of unknowns. However, there are a number of effects that hinder the successful application of STOCSY. Firstly, MALDI is a soft ionisation approach and therefore there will not be many fragments, whereas e.g. in GC-MS there will be many fragments and in NMR there are often a number of resonances from one molecule. Additionally, adducts are formed in the ionisation step, for example the cationisation of the analyte with alkali metal ions such as potassium and sodium, dimerisation and other interactions with the MALDI matrix solution. The formation of adducts is known to vary in different parts of the tissue, and this will negatively affect correlation approaches: if more sodium than potassium adduct is formed in one part of the tissue, and more potassium than sodium adduct in a second part, these peaks will show a low or even negative correlation. Studies have found that the distributions of metal adduct forms are different: the hippocampus had nearly twice as much potassium adduct of $m/z = 734$ as the protonated form, while the corpus callosum had nearly equal amounts. It could be an option to...
identify molecules by performing STOCSY only on the pixels of a selected anatomical region (with comparable adduct formation), rather than on the complete image.

Matrix choice and application method are crucial determinants of the information obtained from the sample, in terms of measured (desorbed) metabolites as well as in terms of quantitation and reproducibility.\textsuperscript{206,213} Spectral quality is additionally influenced by suppression effects, which is that certain desorbed species can preferentially become ionised, and therefore appear to be at higher abundance in the spectrum; this process is at the cost of species that have lower ionisation efficiency and therefore appear to have lower relative abundances than the true sample composition. Suppression effects can differ between tissue types and analyte classes,\textsuperscript{226} which can cause serious artefacts when comparing regions and tissues in a quantitative manner. Thus, MALDI imaging data should be treated as semi-quantitative, or qualitative, especially because the variable pixel quality can skew the distributions of peak intensity.

5.5.2 Principal component analysis of MSI data

Although it has been stated that MSI data acquired with SIMS could be more suitable for multivariate analysis than MALDI data due to the presence of correlating fragments in SIMS,\textsuperscript{51} the following work aims to demonstrate that multivariate methods are very useful in the analysis of MALDI imaging data.

Multivariate versus univariate analysis

Principal component analysis (PCA) of the MSI data, processed as described in §5.4, with 564 variables and 11439 pixels and mean-centred, is able to identify a number of features, see figure 5.12. In this figure, the scores of the PCA model on a given PC axis are used to colour-code the image, so a red pixel corresponds to a high value on that PC. These images show that PC 1 models the overall difference between the grey and white matter, and subsequent PCs show substructures, such as the hippocampus and the cerebellum, that are blue in PC 3.

The loadings of the PCA model can be visualised by ‘back-projection’,\textsuperscript{90} as is shown in figure 5.13. In this figure, the colour is the absolute value of the loading, and thus red peaks are most important for this component. The sign of the loading is combined with the unscaled data to present the height of the peaks. The example shown in figure 5.13 is for PC 1, and can similarly be calculated for all subsequent components. The peak at $m/z = 835.6$ was assigned as a sphingomyelin ([SM 24:1 + Na]), the peak at $m/z = 788.6$ as a phosphatidylcholine ([PC 36:1 + H]) and $m/z = 810.6$ as [PC 36:1 + Na]; these compounds are higher in the areas coloured red for PC 1 in figure 5.12, such as the corpus callosum, medulla and the white matter in the cerebellum. In contrast, $m/z = 782.4$, which could be from [PC 34:1 + Na], $m/z = 756.4$ and $m/z = 697.4$ are higher in the areas coloured blue for PC 1 in figure 5.12, including the cerebral cortex, hippocampus and the cerebellar grey matter.
Figure 5.12: Images colour-coded with the scores of the pixels on the different principal components (PC 1 to 6, the respective variance explained is indicated). The PCA model was built on the mean-centred data set after processing (11439 pixels × 564 variables), and the scores on the PCs are ‘refolded’ to the image to aid visual interpretation.
Figure 5.13: The absolute values of the loading for the first PC are used as a colour-coding; peaks coloured red can be considered important variables. The sign of the loading is combined with the data without mean-centring to form the peak height and direction and create a mass spectral appearance. Peaks pointing up have a higher intensity in the pixels with high scores on PC 1, whereas negative bars correspond to pixels with low scores on PC 1.

A different way of representing the scores is based on red–green–blue (RGB) encoding, where each colour corresponds to the scores on one of the principal components. An example is shown in figure 5.14 A, where PC 1 (red), PC 2 (green) and PC 3 (blue) can be summarised in one image, (compare with figure 5.12). This differentiates e.g. the corpus callosum and the white matter and cortex of the cerebellum. The same can be done for any combination of principal components, e.g. the RGB image based on PCs 4, 5 and 6 is shown in figure 5.14 B, where the hippocampus, thalamus and septum are visible. Comparison of these two figures with the histology image (figure 5.1) shows an advantage of the modelled MSI data in terms of differentiating anatomical parts.

Figure 5.14: The scores on three different principal components can be summarised in one image using red, green and blue encoding for the different PCs. (A) The combined image of PC 1 (red), PC 2 (green) and PC 3 (blue). (B) The RGB encoded image of PC 4 (red), PC 5 (green) and PC 6 (blue).
Multivariate analysis\textsuperscript{241} of the high-dimensional MSI data is preferred to univariate analysis. As an example, two $m/z$ images are shown in figure 5.15: $m/z = 810.6$ [PC 36:1 + Na], which had the highest loading on PC 1, and the variable with the highest average intensity, $m/z = 760.4$ [PC 34:1 + H]. As expected, there is some similarity between these images and PC 1 and PC 2 (inverted), respectively, but the clarity of the images differ (compare with figure 5.12). Moreover, the choice of these two values is biased: 2 from the 564 variables were selected manually, whereas PCA is an unbiased method to create an overview of the data. To find specific structures, it would be necessary to evaluate each individual $m/z$ variable. Thus, manual variable selection is likely to overlook interesting substructures or over-interpret artefacts.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{images.png}
\caption{Images of the variable with the highest loading on PC 1 ($m/z = 810.6$, left) and the highest average value in the image ($m/z = 760.4$, right).}
\end{figure}

\textbf{Score space versus image space}

Evaluation of PCA results is not limited to mapping back the scores of a principal component in the image: the score plot itself can be used to evaluate clustering of certain image pixels.\textsuperscript{242, 243} The identified outliers, classes or trends can then be mapped back onto the image, similar to a mask, where only the selected pixels are presented. Examples of this are given in figure 5.16 A–C, where two clusters of pixels in the score plot of PC 1 versus PC 5 are coloured, and these pixels correspond to the outer right side of the image: the top (pink) and bottom (turquoise) regions.

This separation in the score plot could be the result of the different composition of these two regions in the cerebral cortex, but because PCA models the variation in the spectral data, it is not possible to differentiate between biologically interesting variation and analytical artefacts. The only thing that is certain is that the measured signal differs between these two regions. Thus, the fact that separate clusters are detected can indicate both analytical variation as well as biological variation and can therefore be used in data quality control. One example of an analytical artefact would be the observation of a raster-effect, which could happen because the MSI data are currently acquired in a grid: artefact modelling would allow the diagnosis of sample degradation over time.
Ideally, the spectra should be measured in a random order, but with the current technology this would result in a great increase in measurement time or incomplete coverage of the image.

A score plot of PC 8 versus PC 9, see figure 5.16 D–F, has a distinct separate cluster (yellow) corresponding to the anatomical structure of the hippocampus; the cluster coloured blue corresponds to pixels in the region around the pons and pituitary gland.

Figure 5.16: (A) The scores of PC 1 versus PC 5, two clusters in the score space are selected. (B, C) The position of the selected pixels from the score space in the image is visualised, where a grey scale and colour scale, both based on PC 1, are used to indicate the unselected and selected pixels, respectively. (D) The scores of PC 8 versus PC 9, the yellow and blue clusters are selected. (E, F) The positions of the selected pixels from the score space are visualised, where the grey scale and colour scale are based on the scores on PC 8. (G) The loadings of the two clusters in A are shown as turquoise and pink bars. The loadings are calculated as the weighted linear combination of both PC axes, PC 1 and 5, based on the median position of each cluster, indicated with crosses in A. An exemplar expansion of these composite loadings is shown. (H) The loadings can also be visualised with a biplot, where the scores (black) and loadings (green, cyan and blue) of PC 8 and 9 are superimposed. Variables increased in the hippocampus (yellow cluster in D) are shown in blue, negative loadings are shown in cyan, and the corresponding m/z values are listed in table 5.2.
Interpretation of the variables important for the clusters in the PCA score plots is possible using the median of the selected pixel clusters in the score space region (indicated with crosses). The linear combination of both principal component loadings was calculated, and visualised in the barplot of figure 5.16 G for the clusters indicated in figure 5.16 A. An alternative option is to create a biplot, see figure 5.16 H, which superimposes the scores (black) and loadings (green, cyan and blue).\textsuperscript{194} The loadings were multiplied by 0.1 to match the range of the scores, and loadings that are in the same region as the hippocampal cluster (yellow in D) are coloured blue. Therefore, the yellow pixels are characterised by higher levels of the blue and lower levels of the cyan loadings; the m/z values are listed in table 5.2. The example shown in figure 5.16 D and E is an excellent illustration of the use of multivariate approaches in the analysis of MSI data: the score and image space visualisations identify patterns in the data of biological relevance with accompanying characteristic m/z profiles.

Table 5.2: The selected variables indicated as blue and cyan loadings in figure 5.16 H, representative for the yellow cluster of the hippocampus in figure 5.16 D, are given in terms of m/z value; tentative assignments of some of the isotopes are shown where known.

<table>
<thead>
<tr>
<th>Blue m/z</th>
<th>Assignment</th>
<th>Cyan m/z</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>264.2</td>
<td></td>
<td>752.4</td>
<td>SM 16:0+Na</td>
</tr>
<tr>
<td>363.2</td>
<td></td>
<td>769.4</td>
<td>SM 18:0+K</td>
</tr>
<tr>
<td>459.2</td>
<td></td>
<td>784.4</td>
<td></td>
</tr>
<tr>
<td>496.2</td>
<td></td>
<td>784.6</td>
<td>PC 34:0+Na</td>
</tr>
<tr>
<td>518.2</td>
<td></td>
<td>785.4</td>
<td></td>
</tr>
<tr>
<td>599.4</td>
<td></td>
<td>797.4</td>
<td></td>
</tr>
<tr>
<td>694.4</td>
<td></td>
<td>828.4</td>
<td></td>
</tr>
<tr>
<td>734.6</td>
<td>PC 32:0+H</td>
<td>828.6</td>
<td>PC 38:6+Na</td>
</tr>
<tr>
<td>735.6</td>
<td></td>
<td>829.4</td>
<td></td>
</tr>
<tr>
<td>745.4</td>
<td></td>
<td>856.4</td>
<td></td>
</tr>
<tr>
<td>753.4</td>
<td>SM 18:0+Na</td>
<td>856.6</td>
<td>PC 40:6+Na</td>
</tr>
<tr>
<td>754.4</td>
<td>PC 32:1+Na</td>
<td></td>
<td></td>
</tr>
<tr>
<td>761.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>776.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>783.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>804.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>804.6</td>
<td>PC 36:4+Na</td>
<td></td>
<td></td>
</tr>
<tr>
<td>805.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>805.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>809.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>820.4</td>
<td>PC 36:4+K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>832.4</td>
<td>PC 38:4+Na</td>
<td></td>
<td></td>
</tr>
<tr>
<td>833.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>833.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Finally, one can also evaluate the residuals of the PCA model \(E\), see equation 2.10), by calculating the sum of squares of the distances from the data to the model. An example of the residuals from the mean-centred PCA model after calculating nine components is shown in figure 5.17. It is clear that after nine components, unmodelled structural variation is still present, as some structures are highlighted red when the residues are plotted. The choice for nine components is arbitrary, used for illustration purposes; the variance explained for each subsequent component is smaller than 0.4%. The sum of squares for the residuals of each variable was also plotted, and some of the highest residuals are annotated in figure 5.17. It is clear that investigation of these residuals is necessary when an exhaustive PCA analysis is performed, as some features might still be latent in the data and can be discovered with this approach.

5.5.3 Analysis of MSI data with self-organising maps

This section introduces the use of self-organising maps (SOMs) in MSI data analysis. The non-linear approach as enabled by SOM has not been previously used for the analysis of MSI data. Classification with support vector machines (SVM) allows for modelling non-linearities and has been applied to MSI data,\(^{166,202,204,212,220}\) but does not have the localised modelling, visualisation and interpretation possibilities that SOM has. The SOM is a type of artificial neural network that maps high-dimensional data to a two (or three) dimensional space, as described in §2.3.8. A SOM map is self-organised, which means that neighbouring units are more similar to each other than further removed units. Data analysis with SOM is versatile and effective for three main reasons: 1) it allows for non-linear mapping of the data, unlike e.g. PCA which is based on a linear decomposition of the variance, 2) SOMs are unsupervised and therefore not biased by prior assumptions and 3) the resulting maps present an overview of the data with clear visualisations.

All models were made with the SOM toolbox for MATLAB,\(^{109}\) and in-house written scripts were developed in conjunction for visualisation improvement. One disadvantage of SOM is that the choice of the size, shape and dimensionality of the map is not straightforward. In this study, the

---

Figure 5.17: The sum of squared residuals of each pixel in the PCA model is plotted in the image. Sums of squared residuals for each variable are calculated and plotted on the right hand side.
default settings of the SOM package and the batch implementation were used. The reduced number of variables (564 instead of the original 4751) aids the calculation of the SOM, firstly in terms of speed and secondly by improving the quality of the map.

Pixel distribution across a small SOM

The visualisation capabilities of SOMs are demonstrated in figure 5.18 for a map of ten units (or neurons). Each pixel is mapped onto the most similar neuron, and a small number of neurons generates a clustering-like result, where many pixels are mapped on each neuron. The size of the hexagon in figure 5.18 A is indicative of the number of pixels mapped onto a unit (the hit map, see §2.3.8). The pixels that were mapped onto each unit in figure 5.18 A can be coloured correspondingly in the image, as shown in figure 5.18 B. A clear spatial clustering pattern is the result, where different anatomical structures are mapped onto different units.

The similarity between neighbouring units is shown in figure 5.18 C: a unit similar to its neighbour will have a short distance and a black colour. From figure 5.18 C it is clear that there are two clusters: the top units (1, 6, 2 and 7) versus the bottom units (4, 9, 5 and 10), which are separated by units 3 and 8. The distances shown in figure 5.18 C can also be visualised in the image, see figure 5.18 D. In this figure, lighter blue corresponds to pixels that were mapped onto units that are dissimilar from other units (e.g. units 3 and 8). This aids in the identification of pixels in connective tissue, such as between the pink (unit 10) and light blue (unit 6) coloured pixels of the cerebral cortex and corpus callosum, respectively, shown in figure 5.18 B.

A gradient colour-coding of the units can also be employed, see figure 5.18 E: units more to the right have a green component in them, and from top to bottom, a larger red component is added (hence appearing from dark red for unit 1 to bright red for unit 5). The combination of green and red results via additive colour mixing in the yellow colour for unit 10 etc. The pixels can be coloured correspondingly in the image, as shown in figure 5.18 F. This gradient colour-coding does justice to the self-organising character of the SOM, as it retains the relation to the organisation created by mapping. For example, the colour-coding by SOM of the image pixels changes from red via yellow to green between the cerebral cortex region and the corpus callosum. The compositional difference between the white and grey matter in the cerebellum is also more evident in figure 5.18 F than in figure 5.18 B.

The blue encoding shown in figure 5.18 D representing the dissimilarity of units can be added to the mapping shown in figure 5.18 F, the result of which is shown in figure 5.18 G. For example, the tissue between the grey and white matter of the cerebellum has a large blue component and is coloured purple in figure 5.18 G, confirming that the two neighbouring tissues are very different. This method adds distance information to the self-organisation colouring, but the current encoding does not improve the differentiation of two neighbouring neurons that are highly dissimilar from each other, as they could both have a large blue component. One solution would be to use
Figure 5.18: Visualisation of a self-organising map consisting of 10 units based on the mean-centred MSI data (564 variables, 11439 pixels), see figure 2.14 for an explanation of SOM visualisation. (A) A colour-coding for each unit is provided, and the hexagon size of each unit corresponds to the number of pixels mapped onto each unit. (B) The pixels that were mapped in A can be shown in the original image with the same colour-coding. (C) A distance map shows how similar each unit is to its neighbours, blue units have weight vectors dissimilar from their surroundings. (D) The pixels can be colour-coded in the image based on the distance map shown in C. (E) An alternative encoding of the units uses red and green values for y and x map-coordinates, respectively. (F) The image pixels are coloured with the encoding shown in E, to retain the localised aspect of the SOM. (G) The blue component shown in D is overlaid with the map displayed in F.

A distance-based gradient for the colour-coding of the axes shown in E, where rather than the current linear scheme for red and green components, the distance information is included to map the organisation as well as the dissimilarity between neighbouring neurons.

The clusters identified in figure 5.18 can also be displayed individually, as is shown in figure 5.19. Here, a PCA-based colour-coding in grey is used to visualise anatomical structures, and the pixels belonging to the displayed unit are shown in colour (using the same colour-scheme as in figure 5.18 A). This figure illustrates that white matter pixels are mapped in the upper region of the SOM, and grey matter pixels in the bottom part. A similar visualisation could use digital image data of histological staining (if available), rather than PCA-based colour gradients.
Figure 5.19: The clusters established in figure 5.18 are shown individually as coloured onto a grey map, all colour gradients are based on the scores on PC 1.
The clustering pattern and distance map of a SOM based on the MSI data scaled to unit variance are shown in figure 5.20 A and B: the results are quite similar to the mean-centred figures, indicating that the effect of scaling after log-transformation is small for this analysis. The distribution between clusters 5 and 10, and 1 and 6 is shifted, but since these neighbouring units are similar to each other, the effect on interpretation is minor.

Figure 5.20: Visualisation of the distribution of pixels across the 10 units of a self-organising map (left) and the corresponding distance map (right). (A, B) A SOM based on the data set described in figure 5.18 after scaling to unit variance. (C, D) A SOM calculated from the scores of a PCA model with 3 components based on the mean-centred data. (E, F) Results of a SOM based on a single variable (m/z = 756.4).
SOMs were also calculated based on ‘compressed’ data, namely the scores from PCA models with 3, 7 and 12 components (bending points when the explained variance is plotted), with 91.05%, 97.39% and 99.10% variance explained, respectively. The results based on 7 and 12 components were very similar to the results of the original data, as can be expected from the high percentage of variance retained after this compression. The results of a SOM based on 3 components is shown in figure 5.20 C and D. The orientation of the SOM map is reversed along the y-axis with respect to figure 5.18, interchanging e.g. unit 5 (dark green) and 10 (pink). Irrespective of colour, some clustering structures are retained, e.g. in the cerebral cortex and corpus callosum (the bottom right), whereas the pons and medulla regions (top left) are different.

The advantage of multivariate analysis over univariate analysis is demonstrated with a SOM based on m/z = 756.4, which is the variable with the maximum sum of the absolute intensities of the SOM weights. The results are summarised in figure 5.20 E and F, where only two main structures can be discerned (unit 5 in green versus unit 7 in blue), and most of the fine-structure that was found in the other maps has disappeared. Notably, this map is based on the most highly weighted variable, so the SOM results of other variables could be even less clear; moreover, the individual analysis of each of the 564 variables with SOM maps would be very cumbersome.

**SOM weight vectors**

The interpretation of the variables responsible for the observed distribution in the SOM map is crucial for biological investigations. This is enabled by plotting the weight vectors of each unit, as shown in figure 5.21 A. Alternatively, the original data can be used, where the median of the intensities of all pixels mapped onto a unit is calculated for each variable, shown in figure 5.21 B. Small differences are observed when comparing the median intensities and the weight vectors, e.g. for m/z = 782.4, 788.4 and 788.6. Finally, it is possible to investigate whether the SOM weight values of an m/z value correspond to a certain unit: the correlation of the (ten) weight values for a variable with each individual unit is determined, by calculating the correlation with a ‘dummy variable’. This dummy variable is a vector which is set to 0 for all units, except the unit of interest, which is set to 1 (similar to the dummy matrix in PLS-DA). Only those variables which correlate with p < 0.05 are presented in figure 5.21 C. The correlation threshold for this p-value was empirically determined by calculating the correlations of 100000 normally distributed random draws representing ‘fake variables’ with the dummy variable, see also §2.3.6 on validation methods. These results were not corrected for multiple testing, and hence no firm conclusions should be drawn from figure 5.21 C, but the approach is useful for obtaining a quick overview of the potentially interesting variables in the different regions of the map.
Figure 5.21: (A) The contribution of each variable to the weight vectors of the SOM is shown, colour-coded per unit. (B) The median intensity of each variable across all pixels that were mapped onto a unit can be used to indicate variable importance, again colour-coded per unit. (C) Bars are shown for those variables of the weight vector that correlate ($p < 0.05$) with a specific unit. For clarity, only the region $m/z = 780–790$ is shown.

**RGB colour-coding of larger SOMs**

The map size in the previous examples was manually set to be 10 units large, to allow an easy interpretation of the individual clusters. However, SOMs can also be used for overview purposes, rather than the semi-segmentation that was discussed above. As an example, a SOM was made using the setting `mapsize=small` in the toolbox developed by Vesanto et al.,\(^{109}\) where a heuristic rule determines the optimum map size, creating a map of 133 units [$19 \times 7$], see figure 5.22. The RGB encoding discussed for the map in figure 5.18 is now preferable to the standard colour map coding, compare figure 5.22 E and F. Previously established anatomical patterns such as the corpus callosum are not observed clearly with the default colour map for a large number of units, as the 133 unique colours are not individually distinguishable, and additionally do not convey information on the location of the unit, losing the self-organised information (see figure 5.22 E).
In contrast, when the RGB encoding is used, based on the y and x location of the units and the
distance map, as discussed previously, figure 5.22 F is obtained. This image with RGB encoding is
more homogeneous and interpretable than the previous colouring, which is the result of the RGB
encoding conserving local organisation structure of the SOM.

Finally, it is possible to increase the dimensionality of the SOM map. Although this is not
widely used, the three-dimensional maps allow for interpretable visualisation when combined with
the RGB image encoding. For example, a map of size $[6 \times 9 \times 3]$ was calculated and the RGB
encoding (6 red, 9 green and 3 blue levels) is visualised by showing each individual colour in figure
5.23 A, and the overlay is shown in figure 5.23 B. Again, RGB encoding is highly preferred over

Figure 5.22: Visualisation of a self-organising map $[19 \times 7]$ based on the mean-centred MSI data (564 variables).
(A) A unique colour for each unit is assigned using a standard colour map; the size of the hexagons corresponds
to the number of pixels mapped onto each unit. (B) The distance map shows how dissimilar each unit is
from its neighbours, a separate cluster in the top right hand corner is indicated by the blue colouring. (C)
Visualisation using a gradient colouring, with red and green values for y and x map-coordinates, respectively.
(D) The gradient encoding from C is combined with the blue distance encoding shown in B. (E) The pixels
can be mapped into the original image with the colour-coding used in A. (F) The pixels were coloured based
on the colour-coding shown in D, retaining the relation to the organisation of the map (red and green) and the
distances (blue).
encoding each neuron with the standard colour map, as is clear from a comparison of figure 5.23 C and D: the standard encoding (27 unique colours) of a \([3 \times 3 \times 3]\) SOM map in figure 5.23 C does not remotely give the clear anatomical information and sense of (dis)similarity of structure as the same map with RGB encoding shown in figure 5.23 D, e.g. in terms of the similarity between the substructures in the hippocampus and the cerebral cortex. Thus, this novel RGB encoding of MSI pixels in SOMs has clear advantages for interpretation, and as shown in figure 5.21, the weight vectors can be used to interpret the responsible m/z values. The inclusion of the distance map, as was possible with the blue-level in 2D SOM maps, should be feasible using opacity or other visual tools,\(^\text{221}\) but was not further explored in this work since the current SOM software does not return distance maps for three-dimensional map grids.

Figure 5.23: Visualisation of a three-dimensional SOM. (A) The individual colour components for a 6 (red) \(\times\) 9 (green) \(\times\) 3 (blue) map. (B) RGB coloured result for a SOM map of size \([6 \times 9 \times 3]\). (C) The result of a \([3 \times 3 \times 3]\) 3D SOM map where each unit is assigned an individual standard colour, similar to e.g. figure 5.22 A. (D) RGB coloured result of the \([3 \times 3 \times 3]\) SOM map shown in C.
5.5.4 Evaluation of MSI data with manifold learning

Manifold learning approaches attempt to embed high-dimensional data in a new low-dimensional space.244-249 From this definition, PCA and SOM can be seen as special subclasses of the versatile manifold learning approach. Manifold learning methods are typically well-suited for data visualisation and investigation of non-linear structure and classes in the data, but are difficult to interpret in terms of responsible variables, or explained variance of the data.

**ISOMAP: isometric feature mapping**

One commonly used manifold learning approach is isometric feature mapping or ISOMAP.245 It solves the dimensionality reduction problem using geodesic metrics, curve-based distances, to visualise the underlying non-linear global geometry. ISOMAP firstly defines which points are neighbours, either within a fixed radius, or with a nearest-neighbour approach. This neighbourhood information is combined in a graph, where only neighbouring points are connected and the edge length depends on the distance between the two neighbours. Subsequently the shortest path for any two points through this graph is calculated. Finally, a multidimensional scaling is performed on the matrix of graph distances, to embed the data in a Euclidean space that preserves the manifold’s geometry. ISOMAP will find the globally optimal solution of the convex, well-determined embedding problem, and shows a low-dimensional Euclidean representation of a high-dimensional data set based on geodesic distances.245

The result of ISOMAP analysis of the MSI data set (suggested by Dr. L. Pizarro, Imperial College London, UK) is shown in figure 5.24 A (based on k = 12 nearest neighbours, the default setting, to calculate the graph). Anatomical regions of interest such as the cerebellum, hippocampus and optical chiasm can clearly be discerned in this 2D visualisation with RGB encoding for the three calculated ISOMAP axes. The number of components calculated for the ISOMAP models was based on a maximum likelihood estimation of the dimensionality as implemented in the MATLAB toolbox for dimensionality reduction, which was used to calculate all models (obtained from [http://homepage.tudelft.nl/19j49/home.html](http://homepage.tudelft.nl/19j49/home.html)).244 The result of ISOMAP on the unsmoothed MSI data is shown in figure 5.24 B, which gives a more noisy, but sharper, better resolved picture. The black dots are the data points that were not embedded in the manifold learning: if the neighbourhood graph is not fully connected, only the largest connected component of the neighbourhood graph is embedded in the last step of ISOMAP.

The location of pixels on the ISOMAP axes can also be plotted (similar to a score plot in PCA), and an example is shown in figure 5.24 C for the first two calculated ISOMAP components of the smoothed data. The pixels selected in figure 5.24 C can be evaluated in image-form, as is shown in figure 5.24 D. The pixel clusters correspond to high degree with anatomical structures such as the cerebral and cerebellar cortex, corpus callosum and hippocampus. A great similarity between pixels of anatomical units is observed, even though ISOMAP is unsupervised and no information
Figure 5.24: (A) The results of ISOMAP are visualised with RGB encoding of the location of pixels on the three ISOMAP axes. (B) Results of ISOMAP on the data without smoothing, 26 dimensions were calculated but only the first three dimensions are used in this visualisation. Not all pixels could be embedded for this data set, and the non-embedded data appear as black pixels on the image. (C) A plot showing the position of each pixel on the first two ISOMAP axes for the smoothed data set. (D) Images showing the pixels corresponding to the selections coloured in C. (E) The results of ISOMAP where images are used as variables and the m/z values are the observations, shown as points. (F) Images of the highlighted m/z from E show related distributions if they had comparable ISOMAP coordinates, i.e. similar images are typically mapped close together.
about the origin of the pixels is included (the mapping is only based on the mass spectra). This clustering is unlikely to be an analytical artefact: although a bias could arise because neighbouring pixels are acquired close in time, this would only be observed in one direction, so a line-shaped pattern would be found rather than the clear biological structures observed here.

Finally, instead of using the mass spectra as variables, one can consider the images to be the variables, and the distribution of \( m/z \) values in a low dimensional embedding can be evaluated. The result of this approach is shown in figure 5.24 E, where a set of 9 \( m/z \) values was randomly selected. The distribution profiles of these \( m/z \) are shown in figure 5.24 F, and distributions are more similar for variables with comparable ISOMAP coordinates. The swapping of variables and observations can also be done in PCA and other approaches; in general the transposed data matrix will give additional information for most data analysis approaches.\(^{221}\)

**t-SNE: t-distributed stochastic neighbour embedding**

The recently introduced technique known as t-distributed stochastic neighbour embedding (t-SNE) is another manifold learning method.\(^ {247, 248}\) This dimensionality reduction algorithm (available in the MATLAB toolbox for dimensionality reduction\(^ {244}\)) is particularly valuable for the visualisation of high-dimensional data sets, as it creates a map that reveals structure at many different scales. This is for example useful for analysis of high-dimensional data that lie on several different, but related, low-dimensional manifolds. t-SNE parametrises the non-linear mapping between the original data and the latent space with a feed-forward neural network and is unsupervised.\(^ {247, 248}\)

The results of the t-SNE visualisations on the MSI data are shown in figure 5.25, for both two and three calculated axes and on both smoothed and unsmoothed data. In these figures, different regions corresponding to anatomical features are well-defined; especially for the unsmoothed data. Interestingly, the hippocampus appears as two separate clusters in the smoothed data, but as one cluster for the unsmoothed data. It is also worth noting that the hippocampus is more dissimilar from the cerebral cortex in the t-SNE image than for the ISOMAP results in figure 5.24 A and B. This is likely to be the result of the preservation of both global and local structure with t-SNE, resulting in tighter and more dispersed sets of clusters.\(^ {248}\) This results in a contrasting colouring for the different anatomical regions such as the pons, pituitary gland, hippocampus, and thalamus, in addition to the global differences between the white and grey matter of the cerebellum and the difference between the corpus callosum and the cerebral cortex that were also modelled with other multivariate approaches.

The t-SNE results can also be plotted as positions on the manifold axes, which is done in figure 5.26 A. In this figure, it is clear that t-SNE mapping results in well-defined clusters. Some of these clusters have been selected and the corresponding pixels in the image are coloured, see figure 5.25 B. The anatomical relevance is even more pronounced than for the ISOMAP results, and different regions of the cerebellum are easily distinguished, which was not obtained with e.g. PCA. This is
Figure 5.25: The results of t-sne with two (top) or three (bottom) calculated axes. An encoding with red and green (and blue for the third axes) corresponding to the position of the pixels in the t-sne map was used. (A,C) t-sne results for the mean-centred data set, with all previously discussed processing steps. (B,D) t-sne results for the data set without smoothing but with mean-centring.

A most promising result, and it is intriguing that t-sne, an unsupervised approach, is able to find this level of anatomical detail.

When using the images as variables in the t-sne model, figure 5.26 C is obtained; a group of 7 variables was arbitrarily selected and their distribution profiles are shown in figure 5.26 D. The clustering of similar spatial distributions for these examples is less defined than with isomap as was shown in figure 5.24 F. The calculation of isomap and t-sne models for the ‘normal’ matrix [11439 × 564] is quite computer-intensive (~5–8 hr), but when using the 564 images as observations [564 × 11439], the calculation is performed within a minute.
Figure 5.26: The results of t-SNE for the smoothed data set, based on two t-SNE components, using either the mass spectra as variables (top) or the images as variables (bottom). (A) The scatter plot of t-SNE results based on the spectral data shows clear clustering patterns. Several clusters are selected and their points coloured. (B) The pixels contained in the selections in A correspond to defined anatomical features, showing that clustering in t-SNE space corresponds to meaningful results in the image space. (C) Results of t-SNE using the images as variables. The m/z are now the observations, and a cluster of 7 m/z values is indicated in red. (D) Distribution maps of the m/z selected in C.
5.5.5 Comparison of histology and multivariate modelling of MSI data

For completeness, figure 5.27 displays the histological image shown in figure 5.1 and the RGB encoded images for the PCA, SOM and t-SNE multivariate models, which are rotated to allow easy comparison. This figure clearly demonstrates the interpretational advantages of the ‘digital staining’ allowed by multivariate modelling of MSI data compared to traditional histology images. The anatomical regions can be distinguished to various degrees with the different modelling methods, where PCA models major features, whereas SOM and t-SNE additionally model local and non-linear variations. It is clear that the choice of modelling method will influence the results and visualisation, and it should be remembered that the MSI data have additional molecular information which is not available in histological data, and less accessible when using the manifold learning methods.

Figure 5.27: Anatomical differentiation as allowed by histological and digital image staining. (A) The histology image. (B) The RGB-encoded PCA model of the MSI data. (C) The colouring as achieved with a 3D SOM map. (D) The results of t-SNE modelling.
5.5.6 MSI versus magnetic resonance spectroscopy imaging

This chapter has focussed on the analysis of mass spectrometry data, whereas the other chapters were based on magnetic resonance data. There is an equivalent of MS imaging in the magnetic resonance world, known as magnetic resonance spectroscopy imaging (MRSI) or chemical shift imaging. Commonly, only a few pre-selected volume elements (voxels) in different anatomical regions are acquired in MRSI. If, however, a full grid is acquired, the similarity between MRSI and MSI data is clear: they both entail the acquisition of analytical, localised data, creating a hyperspectral imaging data set. However, the typical S/N for a given voxel size is a lot larger for the MSI data than for the MRSI data.

The lower S/N of MRSI data is caused by a combination of factors, including: magnetic field inhomogeneity, limited field strength and movement of the sample. To obtain a good S/N, a typical voxel size in MRSI is 2 cm × 2 cm × 2 cm while for MSI the used pixel size was 100 µm × 100 µm. An additional complicating factor for high-field MRSI machines is that tuning and matching of the probe is sometimes done manually, which can be very difficult in long-bore machines. Consequently, consistent spectral quality is rarely obtained, and only a limited number of signals are observed. The number of observable peaks and the S/N of MSI data greatly exceeds the current MRSI experimental results.

As an example of the S/N of MRSI spectra, figure 5.28 shows the in vivo 1H MRSI spectrum of a rat brain, acquired at 11 T (470 MHz 1H), which is a very high field for in vivo studies. However, even at this field strength, only a limited number of high-concentration metabolites are visible, some of which would not qualify above the S/N threshold in an in vitro NMR experiment, e.g. see figure 2.3, and the S/N is clearly a lot lower than for MSI spectra, as shown in figure 2.6.

Another main difference between the two methods is that MRSI data can be acquired in vivo, whereas MSI data are acquired ex vivo. Naturally, the information content is of most biological relevance if obtained in vivo, rather than ex vivo: MSI can only reliably analyse stable metabolites.

Figure 5.28: Exemplar in vivo 1H magnetic resonance spectrum acquired from a rat brain at 11 T (spectrum acquired in collaboration with the AMRIS facility at the University of Florida, FL, USA). Tentative assignments of the peaks are shown. Legend: PCr, phosphocreatine; Cr, creatine; Glx, glutamine and glutamate; Tau, taurine; Cho, choline; NAA, N-acetylaspartate.
The acquisition of MRSI data is limited in time, as studies on living objects can typically not last over 0.5–1 hr. MSI acquisitions typically take hours and because the sample is in vacuum, a change in sample composition and metabolic degradation can occur during the acquisition time; moreover, lipids are known to ‘migrate’ from the sample. Thus, the main, obvious advantage of MRSI is that the technique is non-invasive and operates in vivo, while MSI is based on surgical specimens such as tissue slices and is destructive for the sample. This destructiveness limits the ability to perform repeat experiments for MSI, the closest possibility is to use neighbouring cryo-sectioned slices. Replicate experiments in MRSI can also be troublesome, as different positioning, experimental set-up (e.g. shimming) and metabolic status of the subject will influence the spectra.

Both imaging approaches have been mostly applied to problems of relatively limited difficulty, such as classification of different cancer types, or diagnosis of healthy and diseased patients. There is a lot of scope for improvement, on the instrumental and data acquisition side, as well as in terms of appropriate and sophisticated data analysis that takes the multivariate and image nature of these data sets into account.
5.6 Conclusions

This work suggests some solutions for a number of initial bottlenecks in the analysis of MALDI MSI data, such as an intuitive method for peak and pixel selection. The normalisation method choice is non-trivial, especially as no background was found in the literature, even though there seems to be a consensus to use the overall total ion count, and an alternative, robust normalisation is suggested to mitigate artefacts from differential co-crystallisation of the analyte with the matrix. The proposed workflow, with simple and pragmatic processing steps, is not theoretically limited to any type of tissue and can be tailored for different types of MSI data.

Quantitative approaches cannot easily be used for MSI data analysis due to the nature of the data: not all molecules are ionised and desorbed to the same extent during the MALDI process, due to dependencies on the matrix composition, differences in co-crystallisation and different proton affinities, causing intense signals to suppress others. Univariate analysis of images or selected spectral profiles is likely to miss interesting aspects of the data or to over-interpret artefacts, and instead a non-targeted approach was taken to visualise the data set effectively and to obtain characteristic profiles.

The suggested methods, such as the modified correlation coefficient for enhanced contrast, the score-space and image-space visualisation of PCA results, and the RGB encoded visualisation of pixel distributions in SOMs were all newly introduced to the analysis of MSI data. Manifold learning analyses of MSI data showed remarkable clustering in relation to anatomy, and resulted in high-quality visualisation of the data, that was clearly different in character from the PCA and SOM-based visualisations.

These various results show that rather than analysis of single m/z values, a multivariate approach should be preferred, as it is less biased, more robust and incredibly versatile. The main virtue of the presented approaches is that the link between mass spectra and images is always retained, greatly aiding the interpretation of the resulting models. The various RGB-encoded images show a very sharp anatomical relevance, more pronounced than individual m/z images, and are an exceptional improvement compared to the traditional histology image.
Chapter 6

Discussion and future work

6.1 Discussion

The aim of this project was to develop and implement chemometric approaches for metabonomics to achieve improved data interpretation, visualisation and predictive modelling. This is based on an appropriate choice of spectroscopic method, data processing and modelling approaches and would give researchers the ability to extract more information from their samples. One example is the acquisition of \( J \)-resolved NMR spectral profiles to reduce the frequently occurring peak overlap, which can result in more interpretable data and models, especially if full-resolution projections are used. In addition, optimal predictive ability for non-linear data and improved model transparency are enabled with a user-friendly implementation of SA-K-OPLS, which is demonstrated for metabonomic NMR data sets. Another example of increased information extraction is the analysis of tissue with MSI instead of using tissue extracts or magic-angle spinning based NMR techniques, to provide localised molecular information. If the tissue is inhomogeneous, for instance a small or heterogeneous tumour sample, the investigation will greatly benefit from an imaging approach, as this provides separate biochemical profiles for the different locations, rather than an average or ‘diluted’ spectral profile. Here, several aspects of the presented work are discussed in the context of current analytical and chemometric methodologies, biological interpretation, and applications and implications in other research.

6.1.1 Analytical methodologies

Current data acquisition and analysis tools have been extended by the approaches discussed in this thesis, and figure 1.4 illustrates the various spectroscopic methods, data processing and modelling steps that were addressed.

For example, although JRES spectroscopy has been used in the past,\textsuperscript{122,128} the evaluation of projections was based on binned data, whereas in this work full-resolution spectra were evaluated.
A comparison of full-resolution JRES projections and 1D spectra based on multivariate models and correlation-based techniques showed improved interpretation through reduced peak overlap, and demonstrated that peak alignment was a necessary processing step. Knowledge about these evaluated processing and modelling aspects of JRES data should improve the quality of future JRES data and models. This work has highlighted the advantages of additional dimensions in NMR spectra, and it is expected that the development of both conventional and fast methods in multidimensional NMR spectroscopy will profit from the observations stated here; examples include the need for aligning data, and the ability to appreciate and exploit the dispersion across the extra dimension(s).

Typically, the S/N of JRES projections is lower than for 1D NMR experiments, and T2-editing also reduces the intensity of fast-relaxing metabolites in JRES spectra. Strong coupling artefacts can occur, although these are not common for structurally-simple metabolites. Thus, the severity of peak overlap will be a determining factor in deciding on the usefulness of JRES in the analysis of a sample set. For example, certain correlations retrieved with STOCSY will be improved by reduced peak overlap, but a decrease in correlation as a result of lower signal intensity and S/N can also result. Generally, JRES can be considered as complementary to the conventional 1D and CPMG techniques.

Spectroscopic developments, such as the move from in vitro mass spectrometry to imaging, will further our understanding of the biological specimen under study: the localised information obtained with the imaging approach would not have been available with traditional in vitro approaches. Moreover, it is also expected that in vivo magnetic resonance approaches will benefit from the chemometric and image analysis tools presented in this thesis.

### 6.1.2 Chemometric methodologies

Chemometrics will have to cope with increasingly large and complex data sets in the near future. It is especially in these cases that non-linear modelling such as with SA-K-OPLS can be essential rather than simply beneficial. Time-course studies become more feasible with technical advances and reduced experimental costs (also for other -omics), and complex experimental designs are being used. A hallmark example of increased complexity is the shift from ‘simple’ toxicity classification studies to complicated, multifaceted studies of personalised medicine and disease prevention.

The modelling of MSI data is currently often limited to PCA or analysis of selected pixels, and this thesis presents a new set of modelling tools and develops and carefully evaluates a rationalised data processing scheme. The application of multivariate methods to the information-dense MSI data provides visualisations where pixels cluster in relation to anatomical features. This is a significant improvement compared to manual evaluation of the individual m/z images, where prior knowledge is necessary. The non-linear visualisation of the MSI data using SOMs and the developed colour-coding that retains the localised information characteristic of the SOM, presented in this thesis, have
not been reported before. Other methods for MSI data segmentation can be used, but the great virtue of the presented methods is their ability to account for both the multivariate and the image nature of the data. The expected rapid expansion of the field of MSI should increase the appreciation of the need for multivariate analysis of these data. More generally, there is a need for methods to study large, multi-way data sets. For example, over the past years many researchers have gained access to various -omics platforms, and thus there is a strong drive to integrate these data and perform ‘true systems biology’: combining e.g. transcriptomic, metabolomic and proteomic data. The use of chemometric methods by wet-lab research groups and clinicians is greatly dependent on the ease of use and interpretability of the modelling methods, especially in terms of visualisation, which was provided with the suggested MSI data analysis approaches that retain the link with the image during multivariate modelling.

Finally, the new implementation and interpretation of SA-K-OPLS has been described. It was shown how SA-K-OPLS can give insight into the data with score plots, and is able to differentiate predictive and orthogonal variation, which is not possible with other kernel-based methods such as KPLS and SVM, although they may give comparable prediction performance. The performance of SA-K-OPLS is always at least equivalent to the linear model; that is, the predictive ability of SA-K-OPLS is unlikely to be outperformed by OPLS, and improved predictive performance was demonstrated for various data sets. The interpretation of relevant variables in a non-linear model based on pseudo-samples, which was recently introduced in the analysis of SVM models and was adapted for the K-OPLS models, has helped to open up the ‘black box’ that normally surrounds kernel-based models. Although the developments presented for K-OPLS are more incremental than the contributions to the relatively untapped field of MSI data analysis, they are extremely useful: the automated optimisation and nested cross-validation make SA-K-OPLS more user-friendly and provide reliable predictive ability estimates, whilst the interpretation of orthogonal variation increases model transparency. Moreover, clear examples of non-linear modelling in metabonomic studies have been provided, including applications to pharmacometabonomics (prediction of drug responsiveness prior to dosing), toxicological time-course studies, and parasitic infection predictions.

### 6.1.3 Potential application areas

The methods developed here are not limited to the data for which they were presented. SA-K-OPLS was demonstrated for metabolic spectral data of biofluids, and is anticipated to be equally useful for analysis of other types of NMR data and MS spectra, and other high-dimensional data sets such as from genetics, transcriptomics and proteomics. It should be noted that the number of variables that were used in the SA-K-OPLS modelling was quite low (~900 bins), but bigger data sets (e.g. with higher resolution) could be analysed at the cost of increased computational time. Using data sets with a higher resolution could result in models with improved predictive ability,
and would also increase the ease of interpretation of the variable importance approximations, as peak overlap within bins would be reduced. All kernel-based methods need optimisation of the kernel parameter, and thus simulated annealing could also be applied to these problems, where it is strongly suggested that a nested cross-validation is used to prevent over-fitting.

Acquisition of MSI data can be targeted to study all sorts of molecular classes in addition to lipids, such as proteins,\textsuperscript{77,201} metabolites\textsuperscript{206,207} and drugs.\textsuperscript{208,209} The data processing can provisionally be used for other data sets, provided that the acquisition is done in a similar manner, e.g. with the characteristic that there is more matrix signal from pixels outside the sample than from within the sample. As an example, MSI data from a fresh tissue sample (which is different from the formalin-fixed tissue that was presented in chapter 5) were processed and modelled with the tools described in this thesis, and the results are shown in appendix A. Any type of correctly processed MSI data can be subjected to the multivariate modelling techniques, and it is anticipated that these methods will be of great use in the analysis of MSI data from different classes of molecules.

The hyper-spectral data analysis demonstrated with PCA, SOM and manifold learning could also be applied to data obtained with other analytical techniques such as magnetic resonance spectroscopy imaging. Interestingly, manifold learning is well-known in the study of images, but it has been illustrated here that it can also be very informative in the study of spectral data, especially for large data sets. The application of manifold learning approaches to other metabolic profiling data could be useful. As manifold learning techniques are especially beneficial for the analysis of data sets with a large number of samples, they could be employed to evaluate the metabolic profiles resulting from the administration of different toxins (such as the COMET consortia data). Another virtue of manifold learning is the simultaneous non-linear modelling of both local and global variation, which could for example be used in the evaluation of individual differences, in the context of personalised medicine.

### 6.1.4 Biological interpretation

Knowledge generation is facilitated by high data quality and model transparency. Improved detection and identification of biomarkers as allowed with the reduced overlap in \textit{JRES} projections will enhance our understanding of biochemistry e.g. in toxicology and diabetes models, where carbohydrate signals dominate the spectra. Improvements from \textit{JRES} arise firstly because the weight of dominating molecules in modelling is reduced, and secondly because signals previously overlapped or covered can be unveiled, for example the amino acid resonances in the range $\delta \text{3.7-4.0}$.\textsuperscript{156}

The ability to investigate orthogonal variation with score plots, and to approximate the importance of the variables in the \textit{SA-K-OPLS} models, aids in the interpretation of these powerful predictive models. This was illustrated in the pharmacometabonomic study on galactosamine responsiveness, where diurnal variation was initially unknown and later discovered as the source of orthogonal variation. Knowledge of this confounding effect can aid future experimental design and
quality control. For example, one could model the samples from the two collection times separately, or future studies can collect samples at only one time during the day.

The improved interpretation of MSI data with the suggested processing and modelling approaches consists of various contributions, starting with an increased understanding of the contribution from the MALDI matrix solution to the acquired spectra. As an example, ~20 main matrix peaks would typically be identified by researchers and removed prior to analysis, but in this study 3527 peaks were found to be correlated with the matrix distribution patterns. Another example is that the different normalisation methods highlight TIC effects at the edge of the biological sample; this information should be used during quality control, as this widely observed halo effect is detrimental for image interpretation because it compromises the reliability of peak intensity. Then, the advantages of multivariate modelling of MSI data compared to univariate approaches have been emphasised, and PCA and SOM provide a basis for biochemical interpretation based on loadings and variable weights. Ionisation patterns and adduct formation can further contribute to anatomical differentiation, and can provide clues about matrix effects and quantitation as well as potential information on salt concentrations, provided the underlying physical phenomena are understood.

Translational research, i.e. studies providing results and models which could eventually be used in the clinic by non-specialists, requires biological models to use fewer parameters and reduce the number of user intervention steps. An example of increased user-friendliness is predictive modelling with SA-K-OPLS, where the automated optimisation of the kernel parameter and the nested cross-validation allow the usage of this method by a wider target audience. This thesis illustrated the use of SA-K-OPLS in a clinical setting through a murine model of a T. Brucei Brucel parasitic infection, where the diagnosis and time of blood–brain barrier crossing must be accurately predicted for optimal treatment outcome.

The processing of MSI data could potentially be developed as a push-button technique, and unsupervised multivariate models can easily be made with minimal user input. This removes the need for either manual investigation of every single m/z image or prior knowledge of anatomy and biochemical composition in the analysis of tissue based on MSI data. A clear illustration is the (unsupervised) t-SNE manifold learning method, that could identify many anatomical (sub)structures, which is not easily achieved with traditional simple staining methods.

6.1.5 Extensions to real-time use

Instantaneous data modelling is not directly feasible with the current implementation for some of the methods, as several hours can be required to calculate the SA-K-OPLS and manifold learning models. SA-K-OPLS is a computationally intensive method, specifically because of its use of nested cross-validation and optimisation steps. Although the calculation time is longer than conventional linear models, the advantage of non-linear modelling is that it gives improved predictive ability. Similarly, the MSI data sets are very large, which causes a computational and memory burden, but
the MSI data are much more informative than unlocalised MS data. With a parallel implementation of algorithms, and with the progression of electronics to faster machines with bigger memory, the modelling can be speeded up significantly and thus these computational considerations are not an insurmountable problem.

Importantly, the modelling step is still many orders of magnitude faster than the data acquisition. In fact, real-time data acquisition might pose a problem, since, for example, the acquisition of MSI data can take hours. There are several tricks to decrease the MSI experimental time, which include a reduction of the dimensions of the sampled region and the use of a smaller spatial and spectral resolution; more sophisticated solutions include the use of faster lasers and better analysers.252 The acquisition of JRES NMR data takes about twenty minutes per spectrum, which can be reduced by limiting the number of increments and scans.156 Different pulse sequences have been suggested to collect data even faster, and thus this method might be feasible for use in a real-time setting.150–152, 157

The slowest step in data analysis is often the identification of peaks, both in NMR and MS experiments. This is part of the scientific process, and is used to obtain biological information (in addition to the predictions and visualisation that can be achieved real-time, as discussed above). Tentative peak assignments and hypotheses about ongoing biological processes need to be validated and investigated in further, targeted studies, but are not necessarily involved in the decision-making steps that need to be real-time. Once a developed chemometric model is well-validated and biochemically understood, it can be used in a real-time situation, provided that the new analytical data can be acquired quickly and reliably, as prediction of new data points using an existing model is fast.
6.2 Future work

One of the main features of this work is the use of multivariate projection methods to overcome the collinearity and noise present in spectral data sets. However, one avenue that is at the moment under-explored, is the increase in model accuracy if a data table of deconvoluted molecular concentrations is used instead of the spectra; such a table would consist of an entry for each metabolite and the measured concentration in each of the samples. This would completely overcome the over-representation problem that is partially addressed by using JRES spectral projections, and can additionally be used to remove confounding exogenous signals. There are a number of approaches that could be investigated to achieve this, for example the use of multivariate curve resolution combined with a reference library, although the full deconvolution of the spectra of unknown peaks would be difficult. An alternative approach is iterative spectral editing based on correlations.

The peak alignment of JRES projections proved essential for the urinary data sets, but less pronounced improvements were observed for plasma, as peak shifts are relatively small in that biofluid. One could argue that, because the used alignment procedure is not fully artefact-free (see figure 3.4), alignment of plasma data is optional. However, a preferable option is to develop a more sophisticated approach to aligning the spectral data. Namely, although the exact chemical shift may vary, the same compound in different samples should exhibit identical $J$-coupling patterns and relatively stable $J$-coupling constants, and this information is available in the 2D spectra. Incorporation of this information in the alignment algorithm is expected to greatly reduce or remove the occurrence of artefacts (based on a remark from F. Geier, Imperial College London, UK).

There is a wealth of multi-dimensional NMR techniques available, with JRES being the main 2D technique that has commonly been applied in metabolite profiling settings. Other methods have been used, but mostly on a pre-selected sample with the goal of peak identification. It is often not feasible to perform 2D NMR experiments on all samples, because it takes longer to acquire 2D data than 1D data, as a second axis needs to be encoded. The $J$-coupling as a second axis requires relatively few increments compared to encoding the chemical shifts of a second nucleus such as $^1$H, $^{13}$C or $^{15}$N, and thus JRES is a relatively fast multi-dimensional technique. However, the advent of fast multi-dimensional NMR methods, which are based on the encoding of one axis with location in the sample rather than with a time increment, would decrease the experimental time. In the following years, pulse sequences and hyper-polarisation methods should be developed further to achieve a good s/N and reproducibility for these fast methods, which would be complemented by the trend of increasing magnetic field strength for NMR machines. One point that needs to be addressed is the optimisation of pulse lengths and other experimental parameters: at the moment the time necessary to set up a 1D experiment is very short, as identical settings can be used for a series of similar samples. It would be practical if fast 2D methods also require minimal experimental set-up, as otherwise they would not be as time-effective as intended. 2D NMR
spectra can be analysed with chemometric methods\textsuperscript{119,254–257} and it seems worthwhile to focus pattern recognition and other modelling efforts on multi-dimensional NMR spectra acquired with fast methods. It is also worth investigating the use of peak detection in the 2D spectra followed by the calculation of peak intensity. When these values are used as variables, the information on both axes is retained, which was not the case in the current JRES analyses, where projections were taken and coupling information was discarded during modelling. Finally, the inclusion of multi-dimensional spectral data could improve the performance of (semi-)automated peak assignment algorithms.

To fully automate SA-K-OPLS, the number of orthogonal components, and if necessary the number of predictive components, should be optimised within the simulated annealing process. To optimise the kernel parameter in SA-K-OPLS faster, it could be useful to tune the parameters of the simulated annealing approach and to evaluate if the shape of predictive ability as a function of $\sigma$ is informative (see figure 4.1). Alternatively, other optimisation approaches that provide a good solution in a shorter amount of time can be investigated. One method would be to use a logarithmic series of kernel parameters\textsuperscript{173,258} as it is a more reasonable option than trial and error\textsuperscript{172} or the linear grid-search that is commonly employed\textsuperscript{162,259}. Although this deterministic search would be faster than simulated annealing, it might also be less precise and prone to miss the global optimum. Further research should evaluate whether the decreased modelling time is outweighed by possibly poorer prediction performance. Whichever method is used, it should always be implemented with a nested cross-validation in order to provide reliable predictive performance metrics.

The importance of variables in the non-linear models, which is available as a loading for linear models, was approximated by the use of pseudo-samples\textsuperscript{184,193}. This approach needs to be further validated and contrasted with other methods, such as those that could accommodate interactions between variables. Examples of alternative methods include the calculation of the correlation between the input data and the kernel matrix\textsuperscript{260} and feature selection, for example based on bootstrapping and resampling approaches\textsuperscript{261,262}. An important factor to consider in this respect is calculation time: if a bootstrapped model would need full calculation of the kernel parameter with simulated annealing and nested cross-validation, then this approach would be more computationally expensive than the conceptually simple pseudo-data approach that was used here. Extension of the variable importance approximations to multi-class and regression problems appears straightforward, and would be useful to investigate.

The processing and analysis of MSI data is expected to be increasingly important in the future, as this method becomes more commonly used. The realisation that not all the information that is present in the data is extracted by the majority of researchers is an issue that future research should urgently address; for example through the development of accessible and understandable methods and by collating the different approaches in an easy-to-use toolbox. A difficulty that can arise is the lack of compatibility between certain data set sizes and current computing power:
faster processors with more memory and smarter algorithms to perform the processing steps and calculate the models are necessary.

There is a wide range of methods that have not yet been applied but are potentially very useful for analysis of MSI data, in terms of processing as well as data modelling. In terms of data processing, it was proposed to develop a ‘smart’ smoothing step (see §5.4.7), where the quality of each pixel is weighted in the smoothing procedure to obtain the highest-quality data set possible, which for example can deal with noisy pixels and pixels near the edge of the sample; edge-preserving smoothing is another interesting research avenue. A multivariate approach that could be added to the suggested set of modelling tools is OPLS-DA, which has already been applied to an FT-IR hyper-spectral image data set. Predictive methods such as PLS can be used in conjunction with histology and the orthogonal variation in OPLS, which is currently often ignored, could be used to understand sampling, experimental and data processing features and artefacts. Another option, although possibly highly computationally intensive, is the use of robust methods that are able to deal with inferior pixel quality and noisy spectral components. The assessment of various methods requires good evaluation criteria, and extra data sets to test the reproducibility, reliability and generalisation of the results. It is also necessary to obtain expert input on the resolved anatomical structures and interesting m/z values to evaluate the visualisation and modelling.

Manifold learning approaches such as ISOMAP and t-SNE showed promising results for the visualisation of MSI data, and these methods will need to be further explored and validated. At the moment, the main disadvantage of these non-linear methods is the lack of a variable importance metric: manifold learning techniques have no clear equivalent to the ‘loading’ that is calculated in PCA or to the ‘weights’ in SOM. This information is important for biological understanding, as it would provide a measure of the contribution of individual peaks to the separation of classes or visualisation. There are, however, some approximations that can be suggested, such as taking a representative profile of a cluster in the t-SNE plot, or performing PCA on a pre-selected set of pixels identified with the manifold learning method. Future work should be aimed at labelling those variables that formed the basis of the visualisation as important, in order to aid biological interpretation.

It was shown that manifold learning techniques could be used to study mass peaks, since related m/z values cluster in groups (see figure 5.24 E and F and figure 5.26 C and D). Other methods such as PCA and SOM can also be used in this manner, where the m/z values are used as observations and the images are used as variables, in order to find clustering patterns of related peaks. This approach can aid in the assignment of peaks, and the tentative identity can then be further validated with e.g. the use of tandem mass spectrometry.

The data presented were extracted from a BIOMAP-generated file, with a resolution of $\Delta m/z = 0.2$. Future research should evaluate whether the use of full-resolution data, which is likely to need peak alignment and additional processing steps, would instead provide an increased amount of
knowledge. Advancements in technology are likely to increase $m/z$ resolution and spatial resolution further and some imaging instruments incorporate an ion mobility separation step prior to the mass measurements, increasing dispersion and aiding peak identification. This further increases the dimensionality and size of the data, and thus also increases the need for appropriate multivariate modelling and efficient computational techniques, such as the storage of data as sparse matrices.\textsuperscript{229} As an example, manifold learning techniques have proven to be extremely powerful unsupervised visualisation tools, but since these methods take a few hours to calculate one model, algorithmic improvements might be required.

With the advent of high-throughput, automated imaging MS,\textsuperscript{265} and the possibility of three-dimensional imaging,\textsuperscript{250} more samples will be available. This will necessitate the co-registration of the different tissue slices and the development of methods that can analyse multiple images. Additionally, it will be necessary to address the experimental and analytical reproducibility, such as variable adduct formation. Other developments in MSI include imaging on (sub)cellular-resolution level;\textsuperscript{199} analysis of these data with extremely high resolution will greatly increase current biological knowledge, provided appropriate data analysis is performed.

Overall, the challenge is to clearly demonstrate the advantages, reliability and reproducibility of MSI compared to histology as the gold standard. This thesis shows how multivariate approaches can extract a wealth of information from the vast MSI data set. In fact, it would be very interesting to compare and integrate the results of MSI with histology or other imaging modalities, such as fluorescence methods or \textit{in vivo} magnetic resonance (see §5.5.6).\textsuperscript{266} A drawback of MSI is that obtaining a sample is an invasive procedure, and thus the combination of MSI with a non-invasive approach such as magnetic resonance imaging or MRSI could have a synergistic effect, where the simultaneous interpretation of both data sets could enhance the understanding of each individual method. Another exciting application is the ‘intelligent knife’, which analyses the mass spectra of the tissue-originated gaseous ions generated with surgical methods employing thermal ablation;\textsuperscript{267, 268} the integration of these surgical data from different tissues with the equivalent MSI spectra, analysed with tools based on the work presented here, has a great potential for healthcare.
6.3 Conclusions

The work in this thesis has shown how the development and application of analytical spectroscopic methods, data processing and chemometric modelling tools can allow a researcher to extract a maximum amount of information from a provided set of samples. Firstly, the use of J-resolved $^1$H NMR spectroscopy for improved biomarker identification was exemplified with a galactosamine toxicity study. The advantages of reducing peak overlap and over-representation of signals and the analysis of JRES spectra with multivariate projection techniques and correlation approaches were demonstrated. The use of JRES spectral data in metabolic profiling is advantageous, due to the ability to quickly generate 2D spectral data that increase peak dispersion and provide J-coupling information. Secondly, simulated annealing optimised kernel-based orthogonal partial least squares was used to perform predictive modelling in metabolic profiling data sets. The interpretation of orthogonal components and approximation of variable importance increases the transparency of this kernel-based modelling approach. The user-friendly implementation with nested cross-validation should allow this non-linear predictive modelling method to become attractive for researchers, not only in the field of metabonomics, but also in the wider research community. Finally, the processing and multivariate analysis of mass spectrometry imaging data demonstrated how this vast spectral data set can be analysed to provide a large amount of information. The visualisation and modelling of these hyper-spectral data were done by accounting for both the multivariate and the imaging nature of the method. The presented approaches will allow researchers to generate an unbiased and meaningful overview of these data, and to interpret the corresponding biological variables: both the anatomy of the image and the corresponding m/z peaks.

Overall, the metabonomic toolbox has been expanded with various spectroscopic, data processing and modelling methods, which aim to increase the information content and facilitate the interpretation of the data. Therefore, this work is expected to further biological understanding, drug development and healthcare.
Bibliography


[191] Kim KB, Um SY, Chung MW, Jung SC, Oh JS, Kim SH, Na HS, Lee BM, Choi KH: Toxicometabolomics approach to urinary biomarkers for mercuric chloride (HgCl\textsubscript{2})-induced nephrotoxicity using proton nuclear magnetic resonance (^1\textsubscript{H} NMR) in rats. *Toxicology and Applied Pharmacology* **249**: 114–126 [2010].


Appendix A

Analysis of MSI data from fresh tissue
Figure A.1: The results of multivariate analysis of the MSI data set [14531 × 643] from the fresh brain tissue sample, processed using the same approach and settings as the formalin-fixed sample discussed in chapter 5. A PCA model was calculated, and the first three components are mapped with RGB colouring (see also figure 5.14). Additionally, a 3D [3 × 3 × 3] SOM map was calculated and the x, y and z locations of the map encoded with RGB, compare with figure 5.23. The results of manifold learning with ISOMAP are also shown (the data set was too big to perform t-SNE); the first three manifold axes are encoded with RGB colouring, similar to figure 5.24 A.
Figure A.2: The results of multivariate analysis of the MSI data set [12641 × 643] from the fresh brain tissue sample, processed with a smoothing window of size 13 (to remove a large number of edge pixels and create a smaller data set). A PCA model was calculated, and the first three components are mapped with RGB colouring (see figure 5.14). Additionally, a 3D [3 × 3 × 3] SOM map was created and the x, y and z locations on the map are displayed with RGB encoding, compare with figure 5.23. The results of ISOMAP are also shown; the three calculated ISOMAP coordinates for each pixel are encoded with RGB colouring, similar to figure 5.24 A. Finally, the visualisation based on t-SNE embedding, using three RGB encoded axes, as was done in figure 5.25 C, is also displayed.
Figure A.3: The results of multivariate analysis of the MSI data set [12641 × 643] from the fresh brain tissue sample, processed without smoothing using the same pixels as in figure A.2. A PCA model was calculated, and the first three components are displayed with RGB colouring. Also shown are the results of a 3D [3 × 3 × 3] SOM map, where the x, y and z locations were mapped with RGB encoding. The results of ISOMAP are displayed, where the first three of the 53 calculated ISOMAP axes are encoded with RGB colouring. Finally, the results of t-SNE with three RGB colour-coded axes are shown.
Appendix B

Paper I
Evaluation of Full-Resolution J-Resolved $^1$H NMR Projections of Biofluids for Metabonomics Information Retrieval and Biomarker Identification

Judith M. Fonville, Anthony D. Maher,† Muireann Coen, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson*

Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, United Kingdom

Spectroscopic profiling of biological samples is an integral part of metabolically driven top-down systems biology and can be used for identifying biomarkers of toxicity and disease. However, optimal biomarker information recovery and resonance assignment still pose significant challenges in NMR-based complex mixture analysis. The reduced signal overlap as achieved when projecting two-dimensional (2D) J-resolved (JRES) NMR spectra can be exploited to mitigate this problem and, here, full-resolution $^1$H JRES projections have been evaluated as a tool for metabolic screening and biomarker identification. We show that the recoverable information content in JRES projections is intrinsically different from that in the conventional one-dimensional (1D) and Carr–Purcell–Meiboom–Gill (CPMG) spectra, because of the combined result of reduction of the over-representation of highly split multiplet peaks and relaxation editing. Principal component and correlation analyses of full-resolution JRES spectral data demonstrated that peak alignment is necessary. The application of a statistical total correlation spectroscopy (STOCSY) to JRES projections improved the identification of previously overlapped small molecule resonances in JRES $^1$H NMR spectra, compared to conventional 1D and CPMG spectra. These approaches are demonstrated using a galactosamine-induced hepatotoxicity study in rats and show that JRES projections have a useful and complementary role to standard one-dimensional experiments in complex mixture analysis for improved biomarker identification.

Metabolic profiling is a versatile tool for investigating biological functions at the cell and systems levels and, thus, inter alia for studying disease and other pathological processes.1–6 Biofluids, such as plasma and urine, are easily collected and contain a wealth of metabolic information, which can be captured using analytical techniques such as mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy.7 NMR is a nondestructive technique requiring minimal sample preparation, and, although the sensitivity is lower than mass spectrometry, a great number of diverse metabolite signals can be observed, especially at high magnetic field strengths.8,9 The natural abundance of hydrogen in biological materials and the inherently high NMR sensitivity of the proton makes $^1$H an obvious choice for generating metabolic fingerprints.10 Computational pattern recognition methods can then be applied to evaluate and discriminate different classes of samples, e.g., control versus disease, based on multivariate spectral inputs.11,12

Challenges associated with $^1$H NMR-based metabolic profiling include the considerable peak overlap in biofluid spectra and chemical shift variation for some peaks, and this often inhibits complete interpretation of the spectra and accompanying pattern recognition models13–15. One specific problem for NMR spectroscopy is that many metabolites give rise to a considerable number of spin-coupled multiplets over a range of chemical shifts. This spectral overlap has two consequences: impaired peak dispersion and over-representation of the metabolite in any statistical classification exercise.


Analytical Chemistry, Vol. 82, No. 5, March 1, 2010 1811
One method of reducing peak overlap is to disperse signals into a second dimension. J-resolved (JRES) NMR spectroscopy is one of the simplest two-dimensional (2D) NMR experiments where data can be rapidly obtained. This can be more efficient than physical separation of the compounds (for example, using a chromatography step before NMR analysis) or the use of lower-sensitivity 2D NMR experiments (such as homonuclear and heteronuclear correlation spectroscopy). The JRES experiment is formed of an array of spin–echo pulse elements, in which an incremented delay period is used to define a second frequency dimension. After suitable data processing, the chemical shift and coupling information are resolved on two orthogonal axes, thereby increasing signal dispersion, while the resonance intensities are edited based on the proton T$_2$ relaxation times. The 2D JRES spectrum can be projected onto the chemical shift dimension. After suitable data processing, the chemical shift and T$_2$-editing, which attenuates broad macromolecular signals. JRES has proven to be particularly useful in metabolic profiling studies, because many of the nuclear spin systems of such small molecules exhibit first-order NMR spectra; hence, artifactual peaks associated with second-order effects are rare. However, there are limitations caused by the nonstandard line-shape and T$_2$-editing in JRES NMR spectra, in terms of metabolite quantitation.

Improved spectral clustering of data in principal component analysis (PCA) models built from JRES data versus conventional one-dimensional (1D) spectra has been reported. However, to date, most studies to evaluate JRES spectra have used binned data based on a low-resolution representation of the peak intensities rather than full-resolution spectral data. The limitations of binning spectral data on data interpretation and information recovery are well-documented. The main reason for binning data is to minimize the effects of chemical shift variability; however, the development of alignment algorithms for NMR data has lessened the need for this compromise. Other studies have investigated the effect of JRES NMR data acquisition parameters applied to metabolomics.

Here, we comprehensively re-evaluate the JRES NMR experiment, together with the use of full-resolution one-dimensional projections, because it is anticipated that the increased level of apparent resolution will aid biomarker identification, especially for complicated, highly overlapped spectral regions. We consider the processing steps required for the routine and robust use of JRES NMR spectra in metabolomics, including peak alignment and the use of full-resolution data. We assess qualitative and quantitative aspects, in conjunction with correlation spectroscopy approaches, and consider the degree of information recovery, compared to conventionally acquired 1D and Carr–Purcell–Meiboom–Gill (CPMG) H NMR spectra. As an exemplar where NMR peak over-representation is acute, we have used spectra acquired from plasma and urine samples from a galactosamine toxicity study in rats. These contain compounds that are spectrally high-represented with respect to the presence of many multiplets arising from exogenous and endogenous metabolites, thus obscuring a considerable portion of the spectral range.

MATERIALS AND METHODS

Sample Preparation. Samples were collected and analyzed previously as part of the COMET-2 project and were stored at −80 °C. Rats were administered a 10 mL/kg vehicle (0.9% saline, 0.34)

n = 8) or galN (galactosamine hydrochloride dissolved in 0.9% saline to give a free base concentration of 41.5 mg/mL, n = 8).\(^{40}\)

Urine specimens were collected for 24 h after administration and plasma was obtained upon sacrifice, 24 h after administration, as described previously.\(^{40}\) Urine samples were thawed, vortexed, and allowed to stand for 10 min prior to mixing aliquots (400 µL) with phosphate buffer (200 µL, 0.2 M containing 10% deuterium oxide (D₂O), 3 mM TSP (3-(trimethylsilyl)-[2,2,3,3-²H₄]-propionic acid sodium salt) and 3 mM sodium azide) and centrifuged at 13 000 rpm for 10 min. Samples (550 µL) were placed in NMR tubes with an outer diameter of 5 mm. Stored plasma samples, previously prepared and placed in NMR tubes with an outer diameter of 5 mm,\(^{40}\) were thawed and used for analysis.

\[^{1}H\] NMR Spectroscopy. Proton NMR spectra were acquired on a Bruker Avance-800 spectrometer operating at a \[^{1}H\] frequency of 800.32 MHz and a temperature of 27 °C. D₂O provided a field frequency lock and TSP provided a chemical shift reference for the urine samples. Urine and plasma 1D \[^{1}H\] NMR spectra were acquired with water peak presaturation using the pulse sequence \((d_1 - 90^\circ - 4 \mu s - 90^\circ - t_m - 90^\circ - \text{acquire FID})\). For each sample, 64 transients were collected into 64k data points using a spectral width of 16 025 Hz, with a relaxation delay \((d_1)\) of 2 s, an acquisition time of 2.04 s, and a mixing time \((t_m)\) of 100 ms. The water resonance was selectively irradiated during \(d_1\) and \(t_m\). CPMG \[^{1}H\] NMR spectra of urine and plasma, acquired with the pulse sequence \((d_1 - 90^\circ - [\tau - 180^\circ - \tau]_n - \text{acquire FID})\), with a spin–spin relaxation delay of 2\(\tau_n = 102.4\) ms \((\tau = 0.4\) ms, \(n = 128)\) were collected with 64 transients and a \(d_1\) of 2 s into 64k data points with a spectral width of 16 025 Hz and an acquisition time of 2.04 s. The water peak was irradiated during delay \(d_1\). The 2D JRES urine and plasma \[^{1}H\] NMR spectra were acquired with the pulse sequence \((d_1 - 90^\circ - \tau - 180^\circ - \tau - \text{acquire FID})\), with suppression of the water resonance during \(d_1\) into 64k data points in F2 with a \(d_1\) of 2 s and 8 transients using 32 increments of \(\tau\); the spectral widths in F2 and F1 were 16 025 and 50 Hz, respectively.

Data Processing. A line broadening function of 1 Hz and one level of zero filling were applied to all 1D and CPMG spectra prior to Fourier transformation (FT). Automated phase and baseline correction and referencing to TSP \((\delta = 0.000)\) or α-glucose \((\delta = 5.233)\) resonances for urine and plasma, respectively, were performed using in-house software (NMRproc v0.3; Drs. T. M. D. Ebbels and H. C. Keun, private communication\(^{43}\)).

Data for each of the three NMR experiments (1D, CPMG, and JRES) was chosen to be the one with maximum correlation with the other spectra.

Signal-to-Noise Ratio and Line Width Analyses. The signal-to-noise ratios for different processing options of the JRES projections were calculated as the mean ratio across the 16 samples of a given peak maximum, with respect to the standard deviation of noise \((\delta = 0.2-0.5)\). The signal-to-noise ratio was calculated for the alanine methyl singlet at \(\delta 1.47\) and the lactate methyl singlet at \(\delta 1.33\), in both urine and plasma, and the TSP singlet at \(\delta 0.00\) and the succinate singlet at \(\delta 2.41\) in urine; for the CPMG spectra, the highest value of the lactate and alanine doublet signals was used. Note that the term noise is not absolutely correct in JRES spectra, because it is not truly random, having no negative components in the absolute value mode, prior to any baseline correction. The line widths of the JRES projections were approximated as the median of the full width at half height for the lactate methyl singlet and the alanine methyl singlet in plasma and urine, in addition to the TSP and succinate singlets in urine.

Statistical and Chemometric Analyses. The Pearson’s correlation coefficient \(r\) was calculated between spectra to estimate the (dis)similarity between different samples. Principal component analysis\(^{47}\) (PCA) models were generated using mean-centered spectral data before and after alignment. Axes of separation were determined based on the direction of differentiation between the scores of the two groups. Statistical TOtal Correlation Spectroscopy\(^{48}\) (STOCSY) plots display the Pearson correlation (color) and covariance (height) of all spectral data points with a variable of interest, typically either a peak of interest, e.g., the alanine \(\delta 3.05\) and \(\delta 4.5\) resonance (for doublets, the high-frequency peak was used), the creatinine \(\delta 3.05\), or a dummy vector representing the class (encoded as 0 = control, 1 = dosed). Orthogonal partial least squares\(^{49}\) models were constructed with in-house software on unit-variance scaled and mean-centered data, with a dummy vector as above to represent the two classes.
RESULTS AND DISCUSSION

Projection and Symmetrization of JRES Spectra. JRES spectra may be projected in the chemical shift dimension after tilting the 2D spectrum, to effectively produce a homonuclear decoupled 1D proton spectrum. The effect of this “collapse” of multiplets is demonstrated in Figure 1: for example, the multiple, highly overlapped glucose resonances can now be unambiguously assigned in the full JRES spectrum from their chemical shift and coupling pattern information. As is well-known, two types of projection can be routinely employed:19,20 the sum projection, summing all signals along F1, and the skyline projection, retaining the highest signal along F1. Generally, the skyline projection results in signals with a better signal-to-noise ratio (S/N).39 This is illustrated in Figure 1, where a representative “control” 2D spectrum and associated projections are shown, scaled to equivalent noise levels. In Table 1, the mean S/N is displayed for the various processing methods as applied to both urine and plasma spectral peaks, demonstrating that, in all cases, the S/N is better for skyline than sum projections. Because most increments of the JRES spectrum do not contain signal, the commonly observed S/N increase with the square root of summed increments is not achieved.39 Naturally, this S/N advantage of skyline projection is more pronounced for singlets than multiplets. For example, in Figure 1, the H5 β-glucose resonance at δ 3.464 (ddd) shows similar sum and skyline projection peak intensities after scaling to constant noise levels.

Figure 1. Example of a symmetrized 2D JRES 1H NMR spectrum acquired at 800 MHz from plasma of a control rat in the galactosamine study (see the Materials and Methods section for details). Sum (blue) and skyline (red) projections without symmetrization are shown, scaled to identical noise levels (measured at δ 0.2–0.5). For comparison, the conventional 1D 1H NMR spectrum is shown (black). The use of increased dispersion of the JRES experiment and the incorporation of multiplicity information in peak identification is demonstrated with assignment of glc (which denotes glucose). Some minor peaks remain unassigned.

Table 1. Mean Signal-to-Noise Ratios (S/N) and Full Width Half Height Line Widths of Various Spectral Data Types

<table>
<thead>
<tr>
<th>spectral data</th>
<th>urinary alanine</th>
<th>plasma alanine</th>
<th>urinary lactate</th>
<th>plasma lactate</th>
<th>urinary succinate</th>
<th>urinary TSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Signal-to-Noise Ratios (S/N)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPMG spectrum</td>
<td>0.616 × 10^3</td>
<td>1.27 × 10^3</td>
<td>1.35 × 10^3</td>
<td>4.92 × 10^3</td>
<td>5.38 × 10^3</td>
<td>20.0 × 10^3</td>
</tr>
<tr>
<td>JRES skyline projection</td>
<td>0.262 × 10^3</td>
<td>0.483 × 10^3</td>
<td>0.702 × 10^3</td>
<td>2.32 × 10^3</td>
<td>5.86 × 10^3</td>
<td>21.8 × 10^3</td>
</tr>
<tr>
<td>JRES sum projection</td>
<td>0.077 × 10^3</td>
<td>0.141 × 10^3</td>
<td>0.218 × 10^3</td>
<td>0.639 × 10^3</td>
<td>0.728 × 10^3</td>
<td>2.80 × 10^3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Full Width Half Height Line Widths (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRES skyline projection</td>
<td>1.4</td>
</tr>
<tr>
<td>JRES sum projection</td>
<td>1.4</td>
</tr>
</tbody>
</table>

* S/N results for CPMG NMR spectra are based on values from the highest peaks in the listed doublet. The standard 1D NMR spectrum values are not given, because the peak height and noise measurements for plasma are confounded, because of the presence of macromolecular signals.
JRES spectra is outweighed by a drastically improved apparent resolution due to reduced peak overlap in the projections. The fullwidths of specified peaks at half height are also given in Table 1, and these show little difference between symmetrized skyline and sum JRES projection peaks. In addition, symmetrization of the JRES data results in narrower peaks of subsequent projections, which is a phenomenon that is directly related to the symmetrization procedure. Without symmetrization, sum projections resulted in broader peaks than skyline projections and this is in agreement with the superior resolution of skyline projections found previously (data not shown). All further results presented are based on symmetrized spectra.

**Necessity of Spectral Alignment of Full-Resolution JRES Projections for Statistical Analysis.** Using the unsupervised statistical method of principal component analysis (PCA), it was clear (see Figure 2 and Table 2) that the discrimination of galactosamine-dosed animals (solid squares) and vehicle-dosed controls (solid circles). PC: principal component and associated percentage of explained variance. (C, D) The corresponding PCA loadings displayed for the region δ 2.0–4.5 clearly demonstrate the disappearance of dispersive shapes upon peak alignment for, e.g., taurine (1), dimethylamine (2), succinate (3), citrate (4), 2-oxoglutarate (5) and creatinine (6), while resonances of galactosamine (7) become a dominant source of variance in the first two PCs.

**Figure 2.** (A, B) Principal component analysis (PCA) scores for the skyline-projected JRES urine spectra before and after alignment, mean-centered; for galactosamine-dosed animals (solid squares) and vehicle-dosed controls (solid circles). PC: principal component and associated percentage of explained variance. (C, D) The corresponding PCA loadings displayed for the region δ 2.0–4.5 clearly demonstrate the disappearance of dispersive shapes upon peak alignment for, e.g., taurine (1), dimethylamine (2), succinate (3), citrate (4), 2-oxoglutarate (5) and creatinine (6), while resonances of galactosamine (7) become a dominant source of variance in the first two PCs.

| Table 2. Principal Component Analysis Results before and after Spectral Peak Alignment* |
|-----------------------------------------------|----------------|----------------|
| % Variation explained | PC1  | PC2  | PC showing group separation |
| unaligned spectra skyline JRES projection | 33  | 17  | 3  |
| sum JRES projection | 31  | 23  | 1  |
| aligned spectra skyline JRES projection | 76  | 7   | 1  |
| sum JRES projection | 82  | 7   | 1  |

* Principal component analysis (PCA) was based on mean-centered urine NMR data before and after recursive segment-wise peak alignment. Alignment significantly improved the PCA model, with a much higher proportion of the variance explained in PC1 (76.4% versus 33.2%) and biological separation on the basis of toxicity being
promoted to higher principal components (compare Figures 2A and 2B).

It is not only important to align the data to create improved models, but in addition, the interpretation of the data is greatly increased upon alignment. Before alignment, the resultant dispersive line shape of the loadings due to peak position variability makes it ambiguous, with regard to whether a certain peak is actually contributing positively or negatively to a PCA loading. Peak positional variation was indicated for virtually all resonances, e.g., taurine, dimethylamine, succinate, citrate, and 2-oxoglutarate (see Figure 2C). The galactosamine resonances are unaligned to such a high degree that these variables do not contribute significantly to the variance modeled in the first principal components (PCs), much in contrast to the aligned model (see Figure 2D). For plasma data (not shown), the difference in separation of the two classes upon alignment is minimal, but the improvement in loading interpretation remains, similar to that observed in Figure 2.

Urinary \(^1\)H NMR spectra show more peak positional variations than plasma spectra, and this is caused by, for example, greater variability in pH, metal, and ion concentrations. The parameters for each PCA urine model are given in Table 2, where the skyline projections of the JRES data did not separate in PC1 before alignment, but in PC3 (Figure 2). In contrast, the sum projection did separate in PC1, possibly because highly split multiplets, for example, from the galactosamine multiplet peaks, manifest higher intensity and hence greater importance in the sum projection than the corresponding peak in the skyline projection.

JRES data are more sensitive to misalignment effects, compared to 1D and CPMG spectra, because the reduction in "chemical shift bandwidth" for each resonance (multiplets are collapsed into singlets) causes small variations in chemical shift to manifest as marked misalignments in JRES projections. Therefore, the data type and the applied processing parameters (e.g., skyline or sum projection, symmetrization) will greatly influence the resultant pattern recognition model. However, this variation, induced by a combined effect of peak shifts and peak width differences, can be completely removed using an adequate alignment approach.

**Improvement of STOCSY Analyses by Use of Aligned JRES Projections.** Information retrieval from correlation analyses such as STOCSY (statistical total correlation spectroscopy)\(^{36}\) is also improved after peak alignment. Figures 3A and 3B shows JRES projections from \(^1\)H NMR spectra of urine from the dosed and control groups before and after alignment, with expansions demonstrating the substantial improvement by alignment. The corresponding STOCSY plots for these data are shown in Figures 3C and 3D, in which the peak height represents the covariance of the variable with the class dummy-vector and the color represents the correlation with this vector. The severity of the toxic insult, and interindividual differences\(^{40,50}\) in response, cause a broad concentration range for a given metabolite in the different samples, increasing the reliability of the obtained correlation coefficients. Spectral alignment removes the dispersive shapes in the STOCSY plot and clearly reveals correlations that were previously unobserved (for example, those between the galactosamine resonances).

Orthogonal partial least squares models of the urinary JRES projection data before and after alignment confirmed the findings from PCA: before alignment, one orthogonal component is necessary to achieve optimal models in terms of the goodness of prediction (\(Q^2\)) and the sensitivity and specificity of class prediction, whereas after alignment, no orthogonal components are necessary and, thus, standard partial least squares models would suffice (data not shown).

Previous analyses of JRES data of metabolite mixtures did not suffer from peak alignment artifacts because binned integral data were used.\(^{26–29,32}\) However, it should be appreciated that peak alignment is preferable over binning, because multiple signals may become concatenated or individual signals may split in the latter procedure. Moreover, the use of bins precludes the direct interpretation of multivariate models.\(^{14,15}\) Previously, some major peak shifts had to be taken into account by combining multiple bins, indicating that even binning itself may not compensate for major peak shifts. Alignment proved to be uncomplicated for the singlets in JRES projections (in Figure 3, most peaks are well-aligned) and possibly other algorithms will give equally good results. A similar situation occurs in the alignment of \(^{13}\)C NMR spectra,\(^{36}\) where peak shift correction is more straightforward than for overlapping \(^1\)H NMR spectra that exhibit complicated multiplet resonances.

**Investigation of the Inclusion of Stoichiometric Information into STOCSY Analyses.** Peak intensities in JRES sum projections from different protons in the same molecule should, in principle, be related to the number of protons contributing to each peak. However, the situation is complicated by the probability of the various NMR resonances having different \(T_2\) values. Whether this proportionality can be used for improved correlation analysis has been investigated using a regression of two signal intensities from the same molecule across a series of samples (Statistical Total Regression Spectroscopy, STORSY, full details are given in the Supporting Information). This could hypothetically help with the assignment of unidentified molecules and aid the differentiation between intramolecular and intermolecular correlations. Interpretation of STORSY data would only be trivial for spectra without multiplicity patterns, such as JRES projections. The computed intensity ratios were determined to not be consistent with the theoretical values (see Figure S-1 in the Supporting Information), because of the protons having different \(T_2\) values,\(^{21}\) together with the nature and processing of the JRES experiment. Despite this, a high correlation was displayed between intramolecular protons, validating the use of STOCSY for JRES sum projections. These correlations were also maintained in skyline projections where a given molecule has consistent peak multiplicities and relative intensities across different samples. Semiquantitative (relative) changes in metabolite concentrations have also been achieved by others using binned projections of JRES projection spectra,\(^{27}\) and, recently, a line-shape fitting strategy was proposed.

**Effects of Over-Representation of Metabolite Peaks in NMR Spectroscopy on Statistical Analysis.** In many NMR-based metabolomics studies, the spectra show prominent peaks from drugs, toxins, and their metabolites. These can have undue weight in multivariate statistical models and the identification and

---

Figure 3. (A, B) 800 MHz 1H JRES skyline projections of urine samples from the galactosamine toxicity study before and after alignment. Spectra from dosed animals are colored red and those from control rat spectra are colored blue; expansions are shown for δ 2.67–2.74 and δ 3.42–3.48. Asterisks (*) denote that the creatinine δ 4.06 and H2 β-galactosamine δ 3.14 resonances are not aligned perfectly. (C, D) STOCSY plots based on urine 800 MHz 1H JRES skyline projections obtained before and after alignment. The correlation (color, $r^2$) and covariance (peak height) of the JRES projections, with respect to a class dummy vector, are displayed; peaks pointing upward co-vary positively with the toxic insult (see the Materials and Methods section). Upon alignment, correlations are more pronounced and peaks lose the dispersive shape that is characteristic of unaligned data, greatly improving interpretability. Strong-coupling artifacts occur between the signals of galactosamine resonances at δ ~3.8. (Legend: DMA, dimethylamine; 2-OG, 2-oxoglutarate; and α/β-G, α/β-galactosamine.)

means of down-weighting of such peaks has been demonstrated previously using STOCSY-type approaches. In addition, many of these substances give multiple, and often highly $J$-coupled multiplets. This situation is also encountered for some endogenous species (such as glucose), and, here, it is shown that the use of JRES projections can mitigate this problem.
The spectral information content in the various types of 1D $^1$H NMR representations is elucidated as a prior step. To evaluate this, the overall similarities between the different spectral data types were estimated by computing the Pearson’s correlation coefficient between every pair of spectra. The results are plotted in Figure 4, where each “pixel” represents the correlation between two individual spectra; black lines separate different blocks representing six discrete datasets, related to the type of pulse sequence (1D, CPMG, and JRES) and analyzed separately by class (galactosamine-treated and control). These figures allow for a direct comparison between the complete spectral information (not focused on peaks but also including baseline regions) of the different NMR experiments for the two biofluid matrices.

It is apparent from Figure 4 that spectral alignment of urine data greatly improved the overall clustering with experiment and class subtypes. Dose-related clustering patterns are highly disturbed for 1D and CPMG data and are almost absent from JRES data before alignment, confirming the necessity for peak alignment. The correlation patterns for plasma demonstrate clear differences between the different experiments: 1D and CPMG plasma data are more similar to each other than to the JRES projections. As expected, the JRES projections are slightly more correlated to the CPMG data than the conventional 1D spectra, which can be explained by the known $T_2$ relaxation editing in both pulse sequences.

The correlation coefficient between the dosed and control groups for urinary spectra is higher for the JRES projections than for the 1D and CPMG spectra. This can be explained by the smaller number of variables that are changed following the toxic insult and this, in turn, is due to the collapse of the multiplets, leaving the JRES projections from the two classes more alike. This indicates a higher representation of xenobiotic-related peaks in 1D and CPMG urinary spectra of galactosamine-treated animals. Similarly, the control JRES spectra of urine have a lower correlation coefficient with dosed 1D and CPMG spectra (see Figure 4, dotted blocks), compared to dosed JRES spectra with either control or dosed 1D and CPMG spectra (Figure 4, dashed block). Thus, the toxin-related metabolites can be regarded as “over-represented” in the 1D and CPMG data, and the effective $^1$H broadband decoupling, as achieved through projection of JRES data, reduces this effect. This over-representation of multiplets in 1D and CPMG spectra is comparable to the over-representation of protein envelope signals in unit-variance scaled PCA.
models based on plasma NMR data.\textsuperscript{52} The previously reported bias in total area normalization associated with toxin insults,\textsuperscript{35} should theoretically decrease when using JRES projections, because of the reduction in the presence of high-concentration exogenous and endogenous signals that contribute to the total spectral area.

The “over-representation” is not completely remedied by the multiplet collapse. For example, glucose has inherently more resonances than other metabolites, even when collapsed into JRES projections. From a biological point of view, there is no reason why metabolites with more (nonequivalent) protons should be allowed to dominate any pattern recognition model, as is the case for variance-based techniques. Factor analysis, curve resolution,\textsuperscript{53} and STOCSY-E\textsuperscript{54} might be able to compress signals in a JRES projection further into one variable per metabolite, i.e., deconvoluting the constituting metabolites into the pure spectral profiles, to obtain relative concentrations across a sample series, which can then be subjected to multivariate analysis.

**Complementarity of 1D, CPMG, and JRES Projections.** It might be thought that the information content of a CPMG NMR spectrum from a biofluid sample may essentially be obtained from acquisition of a JRES spectrum combined with a standard pulse- and-acquire “1D” experiment. The main virtue of CPMG, $T_2$-editing, is also obtained in the JRES experiment, with the added advantage that JRES spectra provide projections that are easier to interpret in heavily overlapped regions, especially when combined with the scalar coupling information available in the 2D data. A potential disadvantage of the JRES experiment is more-pronounced $T_2$-editing than with the CPMG experiment, because conventional sine-bell apodization minimizes the contribution of rapidly relaxing protons even further. As a result, increased information extraction is obtained in highly overlapped spectral regions, for example, in the range of $\delta$ 3.0–4.0 (see assignments in Figure 1) at the expense of information density for rapidly relaxing species.

Because it is unnecessary to acquire a large number of increments (as zero filling and linear prediction\textsuperscript{24} can be helpful), it is possible to acquire good-quality JRES spectra of typical biofluids samples very quickly, in a period of $<5\text{ min.}$.\textsuperscript{55} There are multiple approaches for fast JRES data acquisition: one is based on reduction of the number of increments and transients in the traditional JRES experiment,\textsuperscript{55} whereas other methods encode the second dimension spatially rather than through time, reducing data acquisition times, although at the cost of sensitivity.\textsuperscript{56–59}

The sensitivity of JRES data is only slightly lower than 1D spectra acquired over the same amount of time\textsuperscript{60} and can be increased further through the use of cryoprobes. Therefore, time is not prohibitive to supplement or substitute acquisition of CPMG spectra with JRES spectra, even though more scans could be required. Given that we have shown that the CPMG experiment can be reliably approximated from mathematical manipulation of 1D data,\textsuperscript{61} its substitution may be more efficient in high-throughput studies.

Given the above, JRES NMR spectroscopy is observed to be advantageous and, thus, complementary to the more-usual 1D NMR methods for routine use in metabonomic research, as previously suggested.\textsuperscript{21,27} An additional benefit is that JRES spectra, in contrast to 1D and CPMG spectra, all have the same appearance at different observation frequencies when spectra are displayed on the ppm scale and, thus, data can be more easily compared across different spectrometers. This allows for rapid comparative analysis of spectral datasets from multiple field instruments and will simplify generation of universal spectral databases. Thus, JRES NMR spectra can be extremely useful as a part of the metabonomics NMR toolbox.

**An Example of Information Recovery Using JRES Spectroscopy.** As illustrated in Figure 5, the reduced overlap of biofluid spectral peaks in JRES spectra not only aids model interpretation, but also can resolve previously unidentified correlations; that is, the reduction in peak congestion resolves overlapped multiplets, thereby allowing previously concealed peaks to be recovered.\textsuperscript{27} The JRES projections of the CH$_3$ peak of alanine (in plasma, $\delta$ 1.47) and creatinine (in urine, $\delta$ 3.05) allow improved correlation with the other peaks in the same molecules, i.e., the alanine CH peak at $\delta$ 3.77 and the creatinine CH$_3$ peak at $\delta$ 4.06, respectively. The plasma region was previously crowded with amino acid and glucose signals, while the creatinine signals in urine were overlapped with galactosamine resonances. The individual resonances are completely resolved upon JRES projection (despite the imperfect alignment of peaks in this region for urine JRES projections) and, hence, a more-reliable STOCSY plot is obtained, demonstrating improved intramolecular correlations. Notably, this is achieved despite a lower S/N ratio for the presented JRES spectra (see Table 1). These findings are particularly relevant when the conventional 1D spectra are heavily obscured by peaks from the administered toxin and/or its metabolites.

Similarly, in Figure 6, we demonstrate the improved identification of biomarker metabolites when using JRES projections, compared to 1D and CPMG spectra through a correlation/covariance plot with class (note that some correlations are not optimal, because the creatinine $\delta$ 4.06 and H2 $\beta$-galactosamine $\delta$ 3.14 resonances are not aligned perfectly in the JRES projection). The correlation patterns identify the same biomarkers for response to the galactosamine toxic insult in the 1D, CPMG, and JRES spectra, but the reduced overlap in the projections (and availability of splitting patterns in the 2D spectrum) significantly aids in complete and unambiguous assignment of all individual metabolic resonances;\textsuperscript{40} we also note that the 2-oxoglutarate peaks display a much higher correlation, and, therefore, this metabolite is clearly unveiled as a biomarker in the JRES projection.

**CONCLUSION**

The most important advantage of J resolved (JRES) nuclear magnetic resonance (NMR) spectroscopy over conventional one-
NMR spectroscopy is the reduction in peak overlap, which aids biomarker identification and data interpretation in “congested” spectral areas. We have evaluated the usefulness of full-resolution JRES projections and demonstrated means to realize the potential of the technique for metabonomic studies, specifically in terms of processing and spectral alignment of the data. The skyline projection is often preferred over the sum projection, because this gives an increased signal-to-noise ratio and is still highly reproducible: correlations of interest retrieved with statistical total correlation spectroscopy (STOCSY) are conserved or improved. The effective peak span widths decrease for singlet projections versus highly split multiplets, and thus pattern recognition and correlation methods require alignment of the full-resolution JRES spectral projections. The subsequently reduced peak overlap in projected JRES spectra can improve the quality and interpretation of multivariate models and statistical correlation analyses, and it will enhance biomarker identification and reduce the differential degree of over-representation of molecules. Hence, judicious application of JRES data and alignment can improve the interpretation of pattern recognition models and increase the information content extractable from NMR-based metabonomic studies, resulting in enhanced biomarker identification.

Figure 5. STOCSY plots driven from (A, B, C) the CH₃ alanine peak at δ 1.47 for the various plasma datasets as shown and (D, E, F) the creatinine CH₃ peak at δ 3.05 in the urine datasets. The color and height represent the correlation ($r^2$) and covariance, respectively. ((A, D) 1D spectra; (B, E) CPMG spectra; and (C, F) JRES skyline projections.)

Figure 6. Correlations of the spectral data points (color, $r^2$) with a dummy variable representing the control or galactosamine-treated class, projected onto the covariance spectrum with the dummy variable for the different aligned spectral NMR types of urine: (A) 1D, (B) CPMG, (C) JRES skyline projection. The interpretation of highly overlapped signals, for example, in the region of 3–4 ppm, is greatly improved for JRES, resulting in improved biomarker identification.

ACKNOWLEDGMENT

This work was supported by the award of an RSC Ph.D. studentship to J.M.F. A.D.M. acknowledges funding from the EU Framework 6 MolPACE Project (LSHG-512066). The MRC Integrative Toxicology Training Partnership (ITTP) is acknowledged for financial support to M.C. The samples studied here were obtained as part of the COMET-2 project, and we thank the Steering Committee for allowing their provision. We also thank Dr. K. Veselkov for provision of the software for the recursive segment-wise peak alignment and Drs. T. Ebbels and O. Beckonert are acknowledged for valuable discussions. We acknowledge the support of the Imperial College MRC-HPA Centre for Environment and Health for this and related projects.

SUPPORTING INFORMATION AVAILABLE

Details and results for Statistical Total Regression Spectroscopy (STORSY) are given as indicated in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review October 27, 2009. Accepted January 12, 2010.

AC902443K
Appendix C

Paper II
The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping

Judith M. Fonville\textsuperscript{a}, Selena E. Richards\textsuperscript{a}, Richard H. Barton\textsuperscript{a}, Claire L. Boulange\textsuperscript{a}, Timothy M. D. Ebbels\textsuperscript{a}, Jeremy K. Nicholson\textsuperscript{a}, Elaine Holmes\textsuperscript{a} and Marc-Emmanuel Dumas\textsuperscript{a*}

Metabonomics is a key element in systems biology, and with current analytical methods, generates vast amounts of quantitative or qualitative metabolic data. Understanding of the global function of the living organism can be achieved by integration of ‘omics’ approaches including metabonomics, genomics, transcriptomics and proteomics, increasing the complexity of the full data sets. Multivariate statistical approaches are well suited to extract the characterizing metabolic information associated with each level of dynamic process. In this review, we discuss techniques that have evolved from principal component analysis and partial least squares (PLS) methods with a focus on improved interpretation and modeling with respect to biomarker recovery and data visualization in the context of metabonomic applications. Visualization is of paramount importance to investigate complex metabolic signatures, the power and potential of which is illustrated with key papers. Recent improvements based on the removal of orthogonal variation are discussed in terms of interpretation enhancement, and are supported by relevant applications. Flexibility of PLS methods in general and of O-PLS in particular allows implementation of derivative methods such as O2-PLS, O-PLS-variance components, nonlinear methods, and batch modeling to improve analysis of complex data sets, which facilitates extraction of information related to subtle biological processes. These approaches can be used to address issues present in complex multi-factorial data sets. Thus, we highlight the key advantages and limitations of the different latent variable applications for top-down systems biology and assess the differences between the methods available. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: metabonomics; multivariate; partial least squares; O-PLS; systems biology

1. INTRODUCTION

Metabonomics [1–6] was originally defined as ‘the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification’ [1]. In the same era, metabolomics was defined in plant biology and microbiology as ‘the systematic study of the metabolic complement of the cell’ [2], but has since broadened to encompass a wider characterization of metabolic responses, and often the two terms are used interchangeably.

The concept of metabonomics arose from pioneering work [3] on the application of nuclear magnetic resonance (NMR) spectroscopy to study the metabolic composition of biofluids, cells and tissues and was aimed at the augmentation and complementation of the information provided by measuring the genetic, and later proteomic, responses to xenobiotic exposure. Genomics, proteomics, and metabonomics [4]/metabolomics [5] now form a core part of the fundamental systems biology framework [7], and are required for an understanding of the integrated function of the living organism [8,9], with each ‘-omic’ platform generating large amounts of data.

To investigate the complex metabolic consequences of, for example, disease processes, toxic reactions or genetic manipulations, information-rich analytical approaches are required. Several methods in addition to NMR can produce metabolic signatures of biomaterials, including mass spectrometry (MS) [10], and optical spectroscopic techniques such as Raman, Fourier transform infrared (FTIR), near infrared (NIR) and ultraviolet–visible (UV–Vis) spectroscopies. Simplification of the mass spectra can be obtained by coupling MS to a separation technology such as gas chromatography (GC/MS) or high-performance liquid chromatography (HPLC/MS) [11]. NMR and MS are currently the most frequently used techniques for metabolic profiling [12], and the characteristic differences in their nature mean that they often provide complementary information for the analytical scientist. The obtained spectra of biofluids such as plasma and urine, or of tissues and tissue extracts, provide metabolic patterns corresponding to the metabolic status of the organism as a function of genetic, environmental or toxicological influence.

Correspondence to: M.-E. Dumas, Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK.

E-mail: m.dumas@imperial.ac.uk

Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK.
Ongoing analytical developments include development of cryogenic probes and increases in magnetic field strength for NMR as well as mass resolution for MS. The enhanced sensitivity and resolution gained by these developments increases the capability for biomarker detection, at the expense of increased spectral complexity. Spectral complexity was traditionally addressed by peak picking or by binning signals across the spectra in order to generate a smaller or more manageable data set. For NMR spectra typically 250 buckets of 0.04 ppm resolution would be generated [13,14]. With increased computational capacity, this procedure has largely been replaced by importing full resolution spectra (e.g. 32k data points) [15]. For this reason, the analyses are performed using the full data set rather than a peak-picked selection, which increases data dimensionality. This full spectral import can increase the information recovered, for example potential biomarkers, from the statistical modeling performed on the data. Similarly, mass resolution has now increased up to 10 000 for time-of-flight analyzers and up to 100 000 for Fourier-transform MS analyzers, which ensures access to accurate mass measurements and peak lists with ~1000–5000 peaks. The analysis of these data sets is now possible with the constant improvement of computing capabilities of both software and hardware for personal workstations. Additionally, increased throughput of samples and the move toward increased cross-platform integration have further increased the complexity and size of data sets and present a considerable challenge to the analyst.

In order to accommodate the enormity of the data, chemometric methods are often employed to extract the specific metabolic information characterizing each group or dynamic process, in order to form or confirm hypotheses associated with the expression of a metabolic profile. In this review the most commonly used latent variable methods for the extraction of statistically characterized metabolic phenotypes are discussed in semi-chronological order (Figure 1 and Textbox); we especially emphasize the improvement in interpretability of spectral data made possible through the use of orthogonal partial least squares (O-PLS) [16].

2. LATENT VARIABLE METHODS IN METABONOMICS

2.1. Univariate versus multivariate approaches

The application of the multivariate advantage to biological studies produces weighted combinations of the original variables (forming the latent variables) that delineate groupings, which are often not visible via any classical univariate analytical approach. Only multivariate methods can really take into account the relations between variables: multi-variable patterns can be significant, even if the individual variables are not. For example, classes of metabolically differentiated groups may only be evident using multivariate approaches. In an early paper on classification of brain tumors [17] it was noted that ‘...when these data are subjected to pattern recognition analyses, the possibility will be open for an objective differential diagnosis...’ based solely on differences in...biochemical composition', thus emphasizing the need for unbiased approaches in biochemical studies. Although initially, spectral analysis of metabolic profiles was performed visually [18,19] it was soon recognized that pattern recognition (PR) methods had great potential for metabolic studies [20]. The multivariate approach [21] results in an unbiased overview of the data set, as facilitated through the representation with latent variables. In such cases as NMR, latent variables may be visually presented in pseudo-spectral form as will be discussed later, a great aid to the experienced spectroscopist [15].

Two commonly used multivariate methods are principal component analysis (PCA [22], unsupervised) and partial least squares regression (PLS [23], supervised). In these models, a multivariate latent variable is constructed from the input (X matrix, e.g. spectral variables). Latent variables, as opposed to observed variables, are derived from modeling using the original variables. This latent vector maximizes total variance (in PCA), or the covariance between the X and response (Y) matrices, in PLS.

**Figure 1.** A timeline displaying the order of development of PLS and subsequent or related approaches with exemplar applications in metabolomic analyses. The applications listed here are discussed throughout the text and relate to typical examples, not the timeline.
The early use of multivariate analysis in the medical precursor work [17] was followed by the application of these techniques in metabonomics research [6,13,20,24]. An exemplary employment of PR methods in a metabonomics context was the classification and interrogation of the 1H NMR urinalysis data in a variety of experimental toxicity states in the rat [20]. PR reduced the potentially enormous spectroscopic data to a few interpretable latent variables and provided a means of classifying toxicological data.

Multivariate projection approaches are advantageous for a reliable and flexible analysis of NMR and MS data as they reduce the dimensionality of the data to improve data visualization and interpretation. This allows handling of the multicollinearity and information redundancy issues commonly observed in spectral data sets: the spectral signature of a given metabolite may contain multiple signals, for example isopolys, fragments, and adducts in MS and multiplets in NMR. Often, the number of variables is much larger than the number of observations, which is problematic for classical linear regression methods. All of these issues are appropriately addressed through the use of latent variable methods.

2.2. Principal component analysis

Early PR methods employed in metabonomics included nonlinear mapping, hierarchical cluster analysis (HCA), and PCA [13,14,20,24,25]. Based on 1H NMR urine spectra, HCA and PCA methods were found to give reliable clustering of renal proximal tubular toxins based on similarity of mechanism of toxicity. PCA is suited as a data representation because the decomposition retains a high proportion of the variance in the original data set. Many early metabonomic studies employing PCA were toxicity studies that involved large toxin-induced metabolic perturbations, which appeared in the first few principal components, and often displayed the separation of the different study groups on the first component. Clustering of liver, kidney, and testicular toxins could, for example, be obtained through such unsupervised PR methods applied to spectral biofluid data. Moreover, direct visualization of the prominent metabolic changes is possible by plotting the loadings of the PCA model [13,14,20,24,25].

2.3. PCA-based classification

Group classification and membership prediction was initially performed using unsupervised or ‘semi-supervised’ PCA-based approaches. The generated PCA model can be used to project new data and determine membership on the basis of scores projection. The PCA-based method called soft independent modeling of class analogy (SIMCA) [26] is based on class groupings: independent PCA models are generated for each of the given classes in a data set and new samples are projected onto each independent model; new sample membership is then assigned by proximity to the sub-models. Often, complementary information such as the unmodeled variance for a sample is incorporated into a Cooman’s plot, showing the proximity to the PCA scores confidence interval space (Hotelling’s T2) that provides a visual classification method for the two-class case. One of the earliest applications of the SIMCA modeling approach in metabolic profiling can be traced back to 1981, when Wold and co-workers [17] used GC/MS to study human cancer cells. Subsequently, applications in metabonomic toxicity, clinical pathologies, and molecular parasitology have been developed [27,28]. A disadvantage of SIMCA for data visualization is the presence of several different sub-models, rather than one overall model with clear-cut interpretation of the loadings. This is the main drawback in using SIMCA for metabonomic applications, as it is hard to ascertain what variables separate the classes. Alternative procedures providing this information include the use of density distributions on the latent variable [29] or the use of discriminant analysis (DA) version of PLS (PLS-DA).
2.4. Partial least squares

Unlike PCA, PLS regression is directed by a response data set \( Y \) to derive the components from the descriptor data set \( X \) that best describe the specified \( Y \) structure, as it maximizes the covariance expressing the common structure between \( X \) and \( Y \) \([23,30,31]\). PLS is sometimes subdivided into regression analysis and discriminant analysis (PLS-DA). In classification or DA, samples are allocated into the appropriate discrete classes, which are represented by using so-called ‘dummy’ variables (Booleans).

PLS and supervised approaches, in general, are often applied to tackle more subtle problems in metabonomics, in which large arrays of data require an approach that permits relationships buried in a background of other large and multiplexed effects to be uncovered within these data. For example in biological systems (e.g., environmental, nutritional, or epidemiological studies) it is often the case that several metabolic contributions control the composition of the same biosample and the main source of variation may not address the biological question; not all influences are relevant to the biological variation of interest and these should be disregarded or down-weighted during analysis. For example human urine can display large differences between individuals \([32]\), and different metabolic phenotypes have been suggested \([33]\), which can complicate studies of more subtle effects such as physiological challenges, the effect of probiotics and prebiotics, and nutritional status. A noteworthy application of PLS involved the introduction and validation of the pharmaco-metabonomics concept, in which the individual response to a toxic insult can be predicted prior to a drug treatment \([34,35]\). Such developments in the pursuit of personalized healthcare were direct consequences of careful and efficient application of PLS modeling. The main virtue of latent variable projection methods such as PLS is the transparency of the models with respect to scores (with possible sub-groupings), and weight vectors, which allow metabolic interpretation. This is more difficult with ‘black box’ methods such as support vector machines and other classifiers that need dimensionality reduction prior to nonlinear classification.

2.5. Validation

One of the main advantages of PLS is the predictive ability of the obtained models. However, without caution, any supervised modeling is prone to overfitting. Ideally, the availability of a second, independent data set would allow for validation of the calculated PLS models. Alternatively, there are a number of methods to perform cross-validation based on the data that are used for modeling. These include methods that leave either one or a fraction of the samples out (such as \( n \)-fold cross-validation), but preference should be given to the use of resampling techniques, such as Monte Carlo cross-validation \([36]\) and bootstrapping \([37]\). The use of cross-validation gives the possibility to calculate model statistics such as \( R^2 \) (proportion of explained variance) and \( Q^2 \) (proportion of predicted variance), as well as the sensitivity, specificity, and the receiver operating characteristic (ROC) \([38]\). The use of cross-validation also results in predictions from the left-out samples and one can advantageously use these cross-validated scores plots to get more realistic information on the predictivity and quality of the model.

Interestingly, the number of components is often determined based on the \( Q^2 \) value, but as the \( Q^2 \) is later used to report the predictive ability, some degree of ‘overfitting’ could result. Nested cross-validation is an approach to give a more unbiased optimization of the number of components while still reporting valid prediction results \([39]\), and a randomization test has been suggested \([40]\).

One sophisticated validation approach is to use computer-intensive model re-sampling under the null hypothesis: a population of models is generated by random permutations of the \( Y \) data (i.e., the correspondence between metabolic data \( X \) and class data \( Y \) is no longer valid) \([41]\). By evaluating the prediction performance of these ‘random’ models in comparison to the original model, the value of the predictive measures can be interpreted. This pragmatic approach enables the validation of the obtained prediction, resulting in increased confidence in the results obtained with this supervised approach.

3. ORTHOGONAL PARTIAL LEAST SQUARES FOR BIOMARKER RECOVERY IN METABONOMICS

3.1. From orthogonal signal correction to orthogonal partial least squares

A relatively recent development in ‘PLS evolution’ was the introduction of the concept of orthogonal filtration, which is the removal of variance orthogonal to the variation of interest. This concept of orthogonal ‘noise’ \([42]\) was developed on the basis of non-equivalence between the number of dimensions computed by PLS and the rank of the covariance matrix \( Y'X \). The operation deflates the information orthogonal to \( Y \) from the \( X \) and can be seen as a filter aimed at cleaning the \( X \) of the information unrelated to \( Y \), such as that caused by run order, temperature drift, or batch effects. Another interpretation is that this is equivalent to giving a different weight to the variables from \( X \). This makes orthogonal signal correction (OSC) a method for quantitative multivariate filtration, contrary to variable selection, which assigns a binary weight to the variables (preserved or rejected variable). Thus, the OSC filter enables variable filtration, without losing the spectral structure of the loadings, and consequently facilitates interpretation through identification of the metabolites which are the basis of the discrimination. Metabonomic applications of OSC to minimize confounding physical and biochemical variation with subsequent PLS modeling became popular in the early 2000s with successes in such areas as cardiovascular disease diagnosis \([43,44]\).

Originally, OSC was applied as a separate step in a data processing sequence, and, as such, was troublesome regarding validation and detection of overfitting \([42]\). Also, the proportions of the predictive and orthogonal variance in a data set were not easily ascertained. The use of OSC and PLS led quite naturally to their integration within a single-procedure decomposition, O-PLS, developed and refined by Trygg and Wold \([16]\). O-PLS performs a regression with orthogonal filtration of \( X \) on a vector \( Y \). This resolved the validation difficulties of OSC-PLS and continued to make a powerful addition to the repertoire of applied PLS methods because the variance in a data set can be split by O-PLS into predictive and orthogonal variance. The basis of an orthogonal component can be characterized by the structure and content of this latent variable, which may suggest or determine the basis of this spurious variance, such as temperature drift \([44]\). The ’structured noise’ described in the
orthogonal components can, if not filtered, decrease the interpretability of the resulting model, such as classical PLS models. Importantly, in O-PLS these components add an extra dimension in which the data can be interpreted: both predictive and orthogonal components display highly relevant information from the data set. Removal of orthogonal ‘structured noise’ from the data is very similar to assessing between-group variance and within-group variance, as indicated for MS-based metabolite data of wild type and transgenic poplar trees [45]. The orthogonal (non-predictive) scores and loadings can be used to interpret sub-group structures within the data: applications have included orthogonal filtration of diurnal variation from a data set, or the removal of gender effects across a study [46].

O-PLS allows an enhanced focus on variation of interest (such as disease state, diet, strain or genotype), while minimizing other variation, which can be of, for example, biological or analytical source. In the DA version of O-PLS, O-PLS-DA [47], the maximum separation between class samples $Y$ ($n$ dummy variables for $n$ classes) is explained using the input data $X$, by decomposing the covariance matrix ($Y^T X$) into $n - 1$ predictive components and a sufficient number of orthogonal components, to be determined by, for example, cross-validation. Subsequently, the model coefficients locate the variables associated with a specific class encoded in $Y$. This is of key importance for applications in metabolic profiling, as the basis of class separation is the biological information sought after. The possibility of direct interpretation of predictive and orthogonal components in relation with the experimental design (and other external variables such as sample run order) represents a great advantage in terms of simplicity and interpretability.

### 3.2. O-PLS and metabonomics

The use of PLS regression and DA in metabonomics has become standard over the past 15 years [48,49] (Figure 2). Moreover, the use of O-PLS, combined with increased data set resolution, has recently set new standards in metabonomic research; it appears that metabolic research is one of the main disciplines to have applied O-PLS, as demonstrated by the number of citations found for the original O-PLS article by Trygg and Wold [16]: an ‘ISI Web of Science’ search (11 February 2010) demonstrated that 92 out of 204 quotations contained the key ‘metabo *’. Interestingly, the only ‘psych* and O *PLS’ result was in fact from a metabolic study on the effect of stress stimuli on individuals [50].

The main reason for the success of O-PLS in metabonomics is that the interpretation of the model (even for a single response) is obtainable from just the first component, rather than an evaluation of many or all PLS components [51], and is easier than interpretation of the coefficients of a PLS model as this will also involve coefficients related to the $Y$-orthogonal variation. The possibility of direct interpretation of the relation between scores and loadings clearly simplifies interpretation of O-PLS-DA models (Figure 3). The typical data set used to exemplify this effect shown in Figure 3 was obtained from a nutri-genomic experimental design study of four different groups [22], consisting of two different genotypes (genetics and variation of interest) and two different nutritional statuses (diet and ‘orthogonal noise’). The direct separation of the scores in relation to the genetic background is possible with O-PLS-DA, whereas the PLS-DA model suffered from imperfect separation caused by variance due to the different diets. The ‘realignment’ of the scores

![Figure 2](https://example.com/figure2.png)

*Figure 2. A bar chart representing the results of an ISI web of Science search (11 February 2010), displaying the usage of multivariate methods in different areas of research. These do not include all applications of the methods, but serve to demonstrate to what extent the methods are used in different fields.*
results in the possibility to directly focus on the first latent variable, as the loadings change correspondingly between PLS and O-PLS, by maximizing the relation of the predictive component of \( \mathbf{X} \) and \( \mathbf{Y} \). The predictive ability of O-PLS models is inherited from PLS models, which enables building efficient and pragmatic predictive models. The incorporation of the OSC filter within the PLS algorithm enabled a straightforward validation and thus decreased the risk of undetected overfitting of the data when using O-PLS. The validation approaches discussed above are, without change, fully applicable to O-PLS models, and cautious modeling is always warranted to prevent over-interpretation of resultant models.

O-PLS became one of the most popular chemometric approaches in metabonomics following the development of an intuitive display of model coefficients. To represent the loadings of a unit variance scaled model with a spectral appearance, Cloarec et al. [15] introduced the concept of ‘back-scaling’. Here, the loadings are multiplied by the standard deviation of the corresponding variable, creating a pseudo-spectrum with a covariance-based appearance; subsequently the scaled O-PLS weights can be mapped onto this pseudo-spectrum using a color scale, resulting in a spectral appearance color-coded by O-PLS weights (Figure 4). Because the covariance in the backscaled plot is roughly proportional to the mean signal level, it appears like a normal spectrum, while the information from the loadings is retained. In fact, this representation of data can be implemented for any latent variable method, such as mean-centered and unit variance scaled PCA models [52]. A related approach is the S-plot [45], which plots the correlation versus the covariance for each variable, and thus contains the same information as backscaled loadings. However, one main disadvantage is that the S-plot does not display the information in a spectral shape, and hence there is a need to backtrack to determine the important variables in the spectrum, thus removing the easy visual spectroscopic interpretation is removed. Also, S-plots get cluttered when plotting full-resolution spectral variables. An additional advantage of the spectral presentation of loadings is the prevention of false interpretation of individual variables that appear important by chance, through interpretation of correlation patterns of the whole spectral peak.

The current implementation and visualization for O-PLS allows reliable multivariate models to be built, with intuitively interpretable loadings with a spectral appearance. This allows non-expert users to evaluate the results of O-PLS models, which is one of the key reasons for its success. Examples of applications of O-PLS as a generic method targeting between-group variance (in classification problems) include metabolic parasitology [53–55], nutritional metabonomics [56,57], the genetics of metabolic profiles [56,58], metabolomic toxicology, cancer metabonomics [59,60], and molecular epidemiology [61,62]. For example the discrimination between protein sources (meat and vegetables) in a large epidemiological study was validated through the application of O-PLS-DA [63]. In the area of genetical metabolomics [58,64], genomic variability and metabolic profiles are integrated, typically using a quantitative trait locus mapping. Recent developments introduced the concept of the metabolome-wide association study (MWAS) [65,66], where environmental and genomic influences are related, for example, through multivariate modeling, to phenotypic variations and disease risk factors in the general population.

In an O-PLS framework, if the data in \( \mathbf{X} \) (and \( \mathbf{Y} \)) are mean centered and scaled to unit variance, then the weight vectors of an O-PLS model on these data are directly related and scalable to the correlation between \( \mathbf{X} \) and \( \mathbf{Y} \) [51,66]. As a consequence, O-PLS can be instinctively described as a bridge between PLS modeling on the one hand, and univariate modeling on the other hand, as both approaches can be used to derive linear correlation coefficients. This property of O-PLS coefficients has been used in several applications, such as the computation of O-PLS-derived variance components for full factorial experimental designs with several factors (see section on O-PLS Variance Component.
models below) [64,67], or the use of univariate testing procedures to assess the significance of correlations of potential biomarkers identified by O-PLS modeling [68]. For interpretation of the latent variable in terms of original variables, it is important to know which parts are statistically reliable. Multiple testing correction for interpreting importance of individual variables in these situations is required, although there is currently no widely used objective criterion (e.g. \( p \)-value) for selecting significant variables from O-PLS models. Since multivariate modeling does not prevent inclusion of noisy variables in the latent variable, it is recommended to be cautious about false discoveries when considering validation of individual markers.

4. PLS-DERIVATIVE AND COMPLEMENTARY APPROACHES IN METABONOMICS

The expectations of systems biologists include the computational integration of, for example genetic, transcriptomic, proteomic,
and metabonomic data to give a holistic picture of the system, which is not available by other methods. This should ultimately enable a more fundamental understanding of biology. Moreover, the modeling of time-sequence systems biology as well as a more complete classification of the range of metabolic processes found in higher animals and how they may interact to determine the outcome of a disease process [69] or drug interaction [35] is required. In this section, some extensions of the PLS approach are discussed as examples, namely for the integration of different descriptor data sets, the analysis of dynamic biological processes, and the incorporation of nonlinear and probabilistic PLS approaches.

4.1. Use of O2-PLS for integration of multiple data sets

From a systems biology perspective, the integration of different data sets can serve to validate complex biological hypotheses and give a more complete insight into the response of an organism to induced perturbations. Rather than combining results from analysis of different biological data matrices post hoc [70] (sometimes known as conceptual level of data integration [71]), the prior integration of different ‘omics’ data sets has been achieved through the development of the sophisticated O2-PLS [72,73] method. Here, both the response and predictor matrices are iteratively orthogonalized and thus filtered for variation in Y, which is maximally uncorrelated to Y (and vice versa) before two-way modeling is carried out (Figure 5). This retains the multivariate advantage: cross-platform latent variables of the various ‘omics’ data sets are maintained in the final modeling.

Multivariate model-based data integration, achieved through O2-PLS, has been demonstrated through the combined modeling of transcript and metabolite data [74], metabolic and proteomic data [75], and metabolic NMR data combined with denaturing gradient gel electrophoresis-derived microbiome information [76]. A series of O2-PLS models was used to integrate transcript, protein, and metabolite data [77], thus increasing the capability for a holistic systems interpretation of biological data, although the need for sequential modeling is sub-optimal and a method to integrate all data sets simultaneously would be preferable. O2-PLS advances the integration of data from different analytical platforms, resulting in a synergistic effect in which the joint variation is highly informative, while the obtained orthogonal components display systematic variation that is unique to each data set. Compared to standard PLS, the predictions are identical, but the interpretation is greatly improved because the predictive latent vector contains the variation that is maximally related to the response. This simplification of the model interpretation makes it particularly suitable for top-down systems biology approaches.

4.2. O-PLS variance component models

Molecular epidemiology data sets and nutri-genomic experiments typically include several factors (genetics, diet, age and disease risk factors), which can be tested in a single experimental design. Variance components (VC) can be easily accessed using O-PLS-based multiple regressions on the design factors included in the data set. O-PLS-VC models correspond to an O-PLS regression model on a Y matrix constituted by full factorial experimental design (orthogonal by construction). The correlation coefficients between each variable and each design factor, derived from the model coefficients, are then squared, giving the proportion of variance. As a consequence, O-PLS-VC enables parallel quantification and visualization of the proportion of variance contributed by each design factor. VC could be accessed at the variable level from any PLS model, but only the O-PLS-VC can simultaneously provide variance components and the O-PLS scores. Thus the interpretation of the scores will relate directly to the design factors because the variation not related to the experimental design will have been removed by the orthogonal components (Figure 6). Blaise et al. [64] performed O-PLS-VC on a full factorial 2^4 experimental design to validate high-resolution magic angle spinning NMR spectroscopy in a functional genomics analysis of silent mutations in the model organism Caenorhabditis elegans. In a variance component model, it is possible to decipher the proportion of variance associated with each factor, for each input variable [64,67]. It is then possible to identify which part of the metabolome is associated with which factor, and assign the variance to the ‘influences’ incorporated in the experimental design. With such a strategy, the genetic and

![Figure 5](image-url)
biological variations of interest were shown to be one order of magnitude stronger than the potentially confounding technological variations linked to experimental protocol and instrumental irreproducibilities [64].

4.3. Correlation studies

Many approaches use correlation structure in the data in metabolic profiling. One method that is widely used, primarily to aid metabolite assignments, is statistical total correlation spectroscopy (STOCSY [78]), which is based on the correlation patterns emerging across a series of spectra. Although this method is not multivariate, it is often used in conjunction with PLS as it is a powerful tool for biomarker identification and structural elucidation. STOCSY is based on multicollinearity between spectral peaks arising from the same compound across a series of samples: the relative peak intensities of a given molecule remain constant. Other approaches resulting from methodological extensions of STOCSY include combining NMR and MS on the same sample set to facilitate structural assignment, which leads to Statistical Heterospectroscopy (SHY [79]). Visualization of STOCSY often involves producing pseudo-spectra with a spectral appearance, on which the correlation coefficient is then projected as a color. The peak-like displays are, for example, generated from sample or mean spectra (Figure 7), or by using the covariance between the X and Y data to generate a pseudo-spectral representation. Alternatively, STOCSY can be displayed in a 2D fashion, similar to the representation in multi-dimensional NMR spectroscopy [78,80].

STOCSY has proven useful for biomarker identification, especially in metabonomic studies, because the need for time-consuming two-dimensional experiments or separation techniques before spectral analysis is reduced. Moreover, in NMR, STOCSY actually provides a larger and more complete number of correlations than traditional pulse sequence approaches because these are limited by bond connectivities and distances. Exemplar studies using STOCSY included evaluation of analgesic usage in a large epidemiological study [81] and the study of drug metabolism kinetics [82]. The correlation-based approach of STOCSY highlights not only signals from the same molecule, but also other highly-correlated signals that can, for example, be pathway-based, for example connectivities in relation to metabolic effects of a renal toxin [75]. Direct metabolite interaction can also be identified, an example being where epidemiological plasma samples treated with EDTA were analyzed using STOCSY to detect chelation-based interactions with Ca-binding proteins within these samples [83].

4.4. Hierarchical methods

Hierarchical-PCA (H-PCA) or -PLS [84,85] can be used to simplify interpretations in applications involving megavariable data sets. In such a situation, plots and lists of loadings, weights and

Figure 6. An O-PLS-derived variance components analysis applied to the data set described by Dumas et al. [56]. The data set is constituted of NMR spectra (n = 273) with a typical nutrigenomic 2 x 2 design: 2 strains (genetic background effect) and 2 diets (environmental effect). This experimental design is retrieved in the scores plot (A). It is possible to derive variance proportions (r²) from O-PLS loadings (B) because the O-PLS models are related to data covariance. The colored area under the curve summarizes the amount of variance explained by each factor, for each NMR variable. The white area up to 100% corresponds to the residual of the O-PLS model (orthogonal components and model residual).

Figure 7. Statistical total correlation spectroscopy (STOCSY) plot performed on the data set described by Dumas et al. [56]. The unknown resonance at 7.83 ppm was used to generate a correlation profile across the NMR spectrum. Other peaks with strong correlations confirm the assignment of hippurate.
coefficients tend to become cluttered and the results are often difficult to overview. Instead of reducing the number of variables, a better alternative is to divide the variables into conceptually meaningful blocks of variables and apply H-PCA. In H-PCA each block of input data is modeled separately by a projection method, which accounts for the variation within each individual block. The block scores are then used as ‘super variables’ on a higher modeling level where a new multivariate model estimates the structure of the blocks, that is of the super variables. In this way H-PCA operates on two levels: on the lower levels the relationship between the variables in the same block is modeled by block models; on the higher levels the relationship between the blocks is modeled. In order to avoid overfitting, validation should be performed on both steps together, rather than for the individual steps only. When a valid model is generated, this provides two levels of interpretation: the top-level projection should display the entire data set, perhaps revealing the presence of clusters (i.e. overview of the data), while lower-level projections display block-specific structure (i.e. detail of the data), which might not be apparent in the higher-level projections. Hierarchical PCA can be easily extended to hierarchical PLS or O-PLS by adding a Y (response/discriminant) matrix for the application of PLS or O-PLS on either the lower or upper level [86]. H-PLS has been used for the visualization of multi-compartmental transgenomic metabolic interactions to study perturbations triggered by symbiotic microbiota; thus, the systems-level effects of microbial-host interactions could be integrated [87]. Similarly, clinical chemistry and expression data were integrated with metabolite data for toxicological class separation [88], allowing for the correlation of biological properties of the different types, improving interpretation of both data sources.

4.5. Multiway PLS

Multiway PLS (N-PLS [89]) is an extension of the ordinary regression model PLS to the case where the data array consists of more than two modes. For three-way PLS, a trilinear decomposition of the data cube \(X\) is completed. The trilinear constraint assumes that the signal intensity is, for example, a linear combination of time intervals, variables, and animals or individuals. The advantage of using N-PLS in metabonomics scenarios is the use of fewer parameters, which increases interpretability, and the models are more robust against the influence of noise in the data. Recently, the use of N-PLS for the quantification of lipoprotein fractions obtained by 2D diffusion-edited NMR spectra has been evaluated [90]. Longitudinal microbial metabolic data have been modeled [91] with N-PLS and parallel factor analysis (PARAFAC [92], a three-way PCA), to account for the multi-mode nature of the data. Another method of three-way PCA is through the Tucker3 model, which was used to study toxicity-related metabolic perturbations by providing interpretable time profiles [93]. An unsupervised multi-way approach provided by PARAFAC allowed the multivariate modeling of different compartments to make metabolic correlations between different organs [94], although direct interpretation of multi-way models can be complicated. Similarly, the effect of a genetic mutation on metabolic profiles of maize kernels was investigated [95] with N-PLS-DA: the different metabolic profiles as well as time responses could be evaluated from the model. Multilevel studies were also able to extract effects from a human nutritional intervention study [96], where data sets are typically affected by large amount of study-unrelated variation, and therefore profit from a multilevel design and supervised data analysis. However, due to trilinearity constraints (requiring signal robustness in several dimensions) and complexity of interpretation, such approaches have not been widely used. Alternatively, an approach that exploits the paired data structure, for example in human nutritional cross-over designs, has been suggested to improve the power of resulting multivariate models [97]. It remains to be seen to what extent novel methods such as paired data analysis become widely used in the metabolic profiling community.

4.6. Batch modeling

Analogous to industrial batch processes, biological processes are dynamic, such as the progressive metabolic influence of a toxic substance, aging, etc. The trajectories of metabolic profiles from such ‘batch’ studies contain information about the properties of the final state and how the biological perturbation affects the intermediate biological states. The main purpose of batch process monitoring is to detect and identify deviation from normal behavior. This can be accomplished using PLS [98,99] to find the relation between the dependent variables (e.g. spectroscopic variables) and dynamic independent variable such as time. The dependent variables are measured at a number of distinct time intervals throughout the process. Similar data will be collected on a number of individual objects (constituting a batch). The average scores trajectory for each component from a single PLS time regression is displayed with upper and lower limits (based on standard deviations) to indicate the normal dynamic trajectory for a batch. The control chart can be used for detecting deviations from normality. Batch modeling has been applied to monitor and interpret time dependent toxicological responses in metabonomics NMR studies with favorable results [100–102], as well as the time-dependent growth response of the poplar tree [103].

A different way to deal with time-series data in metabolic profiling, which retains the advantages of O-PLS, is piecewise multivariate modeling [104], demonstrated for modeling and prediction of the time response of HgCl$_2$ intoxication. This method allows the modeling of nonlinear affects, but is less applicable if the metabolic differences between two subsequent time points are small, for example when the frequency of sampling is high.

4.7. Nonlinear PLS methods

There are a number of nonlinear methods available for modeling metabonomic data [71], including nonlinear mapping, neural networks, genetic algorithms, etc. The complete description of these is beyond the scope of this review, and we focus on a method that is more closely related to PLS (kernel O-PLS), and self-organizing maps (SOMs), which with its simplicity and transparency provides easy interpretation of complex data sets.

4.8. Kernel methods

The aforementioned methods are all based on the assumption that the relation between descriptor and response variables is linear, which is not necessarily correct for all data, particularly biological data. Although the (linear) PLS model has the ability to model mild nonlinearities, in instances where the nonlinear relationship between the descriptor variables and the response variables is substantial, the accuracy and predictive performance...
of the model will degrade or it may fail entirely. A general solution for modeling nonlinear data is provided by kernel-based methods, such as the kernel-PLS (KPLS) ([105,106]) model. In this approach, a kernel, which expresses the ‘core’ nature of the proposed nonlinearity, is introduced into the modeling and the regression is then matched to this structure. The choice of the value of the kernel parameter has great influence on the final results.

A limited number of applications of KPLS to metabonomic data sets have been found so far [107–109]. After the development of KPLS, the addition of an OSC filter prior to KPLS was suggested [110], and kernel-based OSC algorithms have also been published [111]. A kernel-based implementation of the O-PLS model has recently been developed (K-O-PLS [108,112]), enabling the separate modeling of Y-predictive and Y-orthogonal components in the feature space. This approach has been shown to result in improved predictions for a metabolic profiling data set [108], as well as improved model interpretation when Y-orthogonal variation is present [108,112]. For example metabolic data from aspen trees, acquired by magic angle spinning NMR, demonstrated different properties between mutant and wild-type genetics, with orthogonal variation and nonlinearities that were efficiently modeled with the K-O-PLS approach. Improved predictive abilities and interpretation of orthogonal variation can be expected with use of an appropriate kernel.

4.9. Self-organizing maps

Traditional PR methods such as PCA and PLS explicitly assume that the data behave linearly, and only perform well if the data do not contain outliers. Self-organizing maps (SOMs), developed by Kohonen, provide an elegant mapping technique for examining the structure of high-dimensional nonlinear data sets [113,114]. Such a Kohonen map incorporates in an unsupervised way the topology present in the data; usually a two-dimensional, topographical plane is sufficient to elucidate most specific features characterizing the nature of the high-dimensional data. The SOMs can be prone to overfitting, acting as a ‘memory’ of the data, and hence a rigorous (external) validation as well as a large number of samples is advisable. The created map is used to visualize and interpret the data and to link the metabolic profiles to the underlying factors, for example, environmental, physiological, or gut microbial. SOMs have been used to study transcriptome and metabolome data in plant metabolism [115], and to evaluate human blood plasma 1H NMR lipoprotein lipid profiles [116] and were recently used to study subtle metabolic changes in diabetics to evaluate the different clinical phenotypes, vascular complications, and mortality [117]. A demonstration of the application of SOM to the analysis of human saliva samples analyzed by NMR spectroscopy demonstrated the effectiveness of the approach in the study of an oral healthcare product [118]. Additionally, there is a plethora of supervised versions of the SOM [113,119,120], including SOM-PLS [121], although the latter has not yet been applied in metabonomic studies.

5. Conclusion

PLS and its derivative methods, in particular O-PLS, enable the efficient analysis of metabolic profiles and the generation of biological hypotheses at the metabolite, gene, and protein level. PLS and O-PLS methods are now playing a critical role in statistical integration of systems biology data and predictive modeling in personalized healthcare. The use of PLS and derived methods for model data reduction will become increasingly relevant to handling the current explosion of the size of analytical data sets obtained from any biological system. A progressive move was observed from PCA to PLS, and then via OSC followed by PLS and then on to O-PLS (and O2-PLS) as part of meta-bonomic ongoing evolution, together with their appropriate linked DA methods. In this review, we have illustrated how developments in PLS modeling in metabolic profiling are driven both by fundamental methodological improvement in the chemometrics community and by new challenges posed by applications in the systems biology arena.

Acknowledgements

The authors would like to thank Mattias Rantalainen for helpful discussions. This work was supported by the award of an RSC PhD studentship to JMF; MED holds a Young Investigator Award from Agence Nationale de la Recherche (ANR-07-JCJC-0042-01); CLB, SER and MED are funded by Nestlé (RDLS015375); RHB is funded by BBSRC at the Centre for Integrated Systems Biology, Imperial College (CISBIC, PS0283) and EC-FP7 project Florinash (P22634).

REFERENCES


Partial least squares models in metabonomics


Appendix D

Paper III
Non-linear modeling of $^1$H NMR metabonomic data using kernel-based orthogonal projections to latent structures optimized by simulated annealing

Judith M. Fonville, Max Bylesjö, Muireann Coen, Jeremy K. Nicholson, Elaine Holmes, John C. Lindon, Mattias Rantalainen

Abstract

Linear multivariate projection methods are frequently applied for predictive modeling of spectroscopic data in metabonomic studies. The OPLS method is a commonly used computational procedure for characterizing spectral metabonomic data, largely due to its favorable model interpretation properties providing separate descriptions of predictive variation and response-orthogonal structured noise. However, when the relationship between descriptor variables and the response is non-linear, conventional linear models will perform sub-optimally. In this study we have evaluated to what extent a non-linear model, kernel-based orthogonal projections to latent structures (K-OPLS), can provide enhanced predictive performance compared to the linear OPLS model. Just like its linear counterpart, K-OPLS provides separate model components for predictive variation and response-orthogonal structured noise. The improved model interpretation by this separate modeling is a property unique to K-OPLS in comparison to other kernel-based models. Simulated annealing (SA) was used for effective and automated optimization of the kernel-function parameter in K-OPLS (SA-K-OPLS).

Our results reveal that the non-linear K-OPLS model provides improved prediction performance in three separate metabonomic data sets compared to the linear OPLS model. We also demonstrate how response-orthogonal K-OPLS components provide valuable biological interpretation of model and data. The metabonomic data sets were acquired using proton Nuclear Magnetic Resonance (NMR) spectroscopy, and include a study of the liver toxin galactosamine, a study of the nephrotoxin mercuric chloride and a study of Trypanosoma brucei brucei infection. Automated and user-friendly procedures for the kernel-optimization have been incorporated into version 1.1.1 of the freely available K-OPLS software package for both R and Matlab to enable easy application of K-OPLS for non-linear prediction modeling.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Multivariate projection methods are frequently applied for modeling of spectroscopic biological data [1–3]. This category of methods provides predictive and interpretable multivariate models well suited for handling the multi-collinearity of variables encountered in spectral data and in other high-dimensional ‘-omics’ data, where the number of observations is much smaller than the number of variables. Two commonly used multivariate prediction models are partial least squares (PLS) [4] and the related orthogonal PLS (OPLS) [5–7]. The OPLS algorithm enables separate modeling of $Y$-predictive (response-related) and systematic $Y$-orthogonal (response-orthogonal) variation in data, “structured noise” [5,8–11], providing benefits in terms of model interpretation compared to PLS. The concept of $Y$-orthogonal variation can be thought of as systematic effects that are needed to characterize the system but are unrelated to the question at hand, i.e. the model predictions. For instance, when aiming to classify a group of responders versus non-responders to a particular treatment, the structured noise could be composed of inter-sample differences that are needed to describe variability of the system but are not useful for separating responders from non-responders.

OPLS has been successfully employed in metabonomics [12] and the related field of metabolomics, branches of systems...
biology studying the metabolic responses under various conditions and perturbations [13–15]. For metabolomic profiling of biological samples, analytical methods such as mass spectrometry and proton (1H) nuclear magnetic resonance (NMR) spectroscopy are commonly employed [16]. Metabonomics has successfully been applied in a number of areas including toxicology [17,18], biomarker and drug discovery [19,20], personalized medicine [21,22], nutrition [23,24] and parasitology research [25–27].

Most of the commonly used multivariate prediction models in metabonomics assume a linear relationship between the descriptor and Y (response) variables. However, many biological systems display non-linear characteristics in response to a perturbation. Under such conditions, non-linear methods are expected to provide improved models, which is particularly important in predictive applications such as disease diagnostics [28], assessment of toxicity [29] and characterization of variable response of individuals to drugs in personalized healthcare [22].

A particular class of non-linear models are kernel-based models [30], with an early chemometric application in the form of radial basis functions – PLS [31,32]. Other examples of kernel-based models include support vector machines [33–35], kernel-based partial least squares (KPLS [36,37]) and kernel-based least squares regression [38]. Kernel-OPLS (K-OPLS) [39–42] is the non-linear extension of the OPLS model, a commonly used multivariate model in metabolic studies [1,12]. In contrast to separate (linear) orthogonal signal correction (OSC) followed by KPLS modeling [43], or kernel-OSC followed by KPLS modeling [44], K-OPLS provides an integrated orthogonal signal correction property that allows for separate modeling of predictive and Y-orthogonal variation in the feature space, removing drawbacks associated with multi-step solutions, such as separate (K)OSC and (K)PLS steps [5].

Kernel-based models require an optimization of the kernel function parameter, which may be challenging for the non-specialist, as the kernel parameter is a continuous parameter with an undefined upper limit that may have multiple local optima. The optimization step is essential to produce a model with optimal predictive performance [41,43,45]. Here we have implemented an automated procedure for optimization of the kernel parameter based on simulated annealing (SA), a stochastic optimization method [46,47], which greatly simplifies the use of K-OPLS. This optimization step has been incorporated into the freely available K-OPLS package for both R and Matlab (http://sourceforge.net/projects/kopls/).

Our main objective in this study is to evaluate if non-linear prediction models provide an advantage compared to linear alternatives in two common application areas of metabonomics: toxicology and disease diagnostics. Using the non-linear SA-K-OPLS method, we demonstrate how prediction performance can be improved in comparison to the linear OPLS model for three separate spectral NMR metabolic data sets. We focus in particular on problems where prediction is of paramount importance, and we also show how structured Y-orthogonal variation can be interpreted to gain further insight into the data.

2. Methods

2.1. Notation

Let \( \mathbf{X} \) be the matrix of descriptor variables of dimensionality \( [N \times k] \) for \( N \) observations and \( k \) variables and \( \mathbf{Y} \) be the matrix of response variables of dimensionality \( [N \times m] \) for \( m \) variables in \( \mathbf{Y} \). In the case of classification, \( \mathbf{Y} \) consists of ones and zeros forming a design matrix indicating the class of each observation; the matrix transpose is indicated as \( ^T \).

2.2. Orthogonal projections to latent structures (OPLS)

The OPLS algorithm [5] decomposes the descriptor data matrix \( \mathbf{X} \) into a set of predictive (response-related) and a set of Y-orthogonal (response-orthogonal) latent variables. Let \( \mathbf{T}_p \) \( [N \times A_p] \) and \( \mathbf{T}_o \) \( [N \times A_o] \) represent the predictive and Y-orthogonal score matrices respectively, with \( A_p \) predictive components and \( A_o \) orthogonal components (latent variables). \( \mathbf{P}_p \) \( [k \times A_p] \) and \( \mathbf{P}_o \) \( [k \times A_o] \) are the corresponding loading matrices. The OPLS model of \( \mathbf{X} \) is outlined in Eq. (1), where \( \mathbf{E} \) \( [N \times k] \) is the residual matrix.

\[
\mathbf{X} = \mathbf{T}_p \mathbf{P}_p^T + \mathbf{T}_o \mathbf{P}_o^T + \mathbf{E}. \tag{1}
\]

2.3. Kernel-OPLS

The K-OPLS algorithm is a kernel reformulation of the OPLS algorithm, utilizing the computationally efficient ‘kernel-trick’, allowing the elements of the kernel Gram matrix \( \mathbf{K} \) \( [N \times N] \) to be treated as dot-products in the higher-dimensional feature space \( F \) to which the data are (non-linearly) mapped [48]. K-OPLS [39,40] has a set of \( \mathbf{Y} \)-predictive latent variables represented by the predictive scores matrix \( \mathbf{T}_p \) \( [N \times A_p] \) and a set of Y-orthogonal latent variables represented by the scores matrix \( \mathbf{T}_o \) \( [N \times A_o] \), in analogy to the OPLS model. The kernel-trick is common among all kernel-based pattern recognition methods, and is computationally efficient since no explicit transformation of the data is needed. Estimating a kernel-based model is a two-step procedure. First, \( \mathbf{K} \) is calculated using a suitable kernel function \( k(\cdot,\cdot) \), circumventing the explicit mapping of the descriptor data to the (higher dimensional) feature space to facilitate modeling of non-linear variation, see Eq. (2), where \( x_i \) and \( x_j \) represent rows \( i \) and \( j \) in \( \mathbf{X} \).

\[
\mathbf{K}_{ij} = k(x_i, x_j). \tag{2}
\]

Second, K-OPLS is used to establish a regression or discriminant analysis model of the potentially non-linear, relationship between the descriptor data, now represented by the kernel-matrix, \( \mathbf{K} \), and the response variable(s), \( \mathbf{Y} \).

2.4. Kernel function optimization

The kernel matrix (\( \mathbf{K} \)) is required to be positive semi-definite, and there are many kernel functions available fulfilling this requirement [49], here we focus on the commonly used Gaussian kernel function (Eq. (3)).

\[
k(x, y) = e^{-\frac{(x-y)^2}{2\sigma^2}} \tag{3}
\]

This kernel function has a tuning parameter related to the size of the Gaussian kernel (\( \sigma \)). The choice of kernel parameter greatly influences the predictive performance of the resulting model (see the gray points in Fig. 1A, showing the predictive ability as a function of the kernel parameter), emphasizing the importance of optimization of this parameter. An exhaustive search over all possible values is not computationally tractable due to the continuous nature of the parameter. It is common to apply a grid search strategy to find a reasonable parameter value [41], but this approach does not necessarily provide an optimal value as it may find local minima and is also dependent on the selected grid range and resolution. Here, we use simulated annealing (SA) [50,51] to automatically optimize the kernel parameter in the K-OPLS model in order to achieve good prediction performance.

2.5. Simulated annealing

In analogy with thermodynamic principles [52], where an infinitely slowly cooled ‘sample’ will occupy the lowest energy
state, simulated annealing enables stochastic sampling of the kernel parameter space as a function of the decreasing 'temperature' [46]. The temperature influences the probability of temporarily moving to sub-optimal parameter values in order to overcome local minima, and with decreasing temperature the search range of the kernel parameter is slowly narrowed down (see the bars in Fig. 1B). The pseudo code for the SA algorithm is provided in Supplementary data 1 (also see Corana et al. [47] for further details). Here, we present results from SA-K-OPLS using a Gaussian kernel function, however, the simulated annealing approach is general and applicable to other kernel functions with a continuous parameter to optimize. The kernel parameter (σ) in the Gaussian kernel function defined in Eq. (3) is optimized for values > 0.

2.6. Evaluation of models

2.6.1. Performance metrics

Prediction performance of regression models is evaluated using the $Q^2$ metric, commonly referred to as the ‘goodness of prediction’, which is inversely proportional to the generalization error (Eq. (4)).

$$Q^2 = 1 - \frac{\sum_{i}(\hat{Y}_{cv,i} - Y)^2}{\sum_{i}(Y_{mean} - Y)^2}$$  \hspace{1cm} (4)

Here, $\hat{Y}_{cv,i}$ is the vector (or matrix) of predicted values for $Y$ from cross-validation and the summation is over all elements of $Y$. In discriminant analysis, class membership is determined by the decision function $f_{max}(\hat{Y}_{cv,i})$, returning the class label (or class index) for which $\hat{Y}_{cv,i}$ has the highest value. $\hat{Y}_{cv,i}$ is a vector of the predicted values for each class for observation $i$. Prediction performance of discriminant analysis models was evaluated by two different metrics. The ‘mean sensitivity’ is calculated across all $k$ classes with Eq. (5), where $TP_i$ (the number of true positives, i.e. number of detected cases for class $i$) is divided by the total number of cases for class $i$, calculated as the sum of the number of false negatives (FN$_i$) and true positives. A mean sensitivity equal to one implies perfect discrimination between the classes.

$$\text{mean sensitivity} = \frac{\sum_{i=1}^{k} TP_i}{(TP_i + FN_i)}$$  \hspace{1cm} (5)

In the special case of two classes, an alternative performance metric for discriminant analysis models is the area under the receiver operating characteristic curve (AUC), which was calculated using a standard trapezoid approximation [53]. AUC is interpreted as the probability that the model will rank a random true positive observation higher than a random true negative observation and is related to the non-parametric Wilcoxon rank sum test as well as the Mann–Whitney U test. An AUC value of 1 corresponds to perfect classification, a value of 0.5 indicates random performance. The AUC metric has favorable properties in classification scenarios as it is independent of the class frequencies and does not depend on a cut-off point between the classes used for dichotomization.

2.6.2. Nested cross-validation

A nested cross-validation procedure [54,55] was used to evaluate prediction performance during optimization of the kernel parameter, and the optimization was performed separately for models with different numbers of Y-orthogonal components (see Section 2.7). In each ‘outer cross-validation loop’, the data were split into a training set and a test set, and the kernel parameter was optimized in the ‘inner cross-validation loop’ (using only the training data from the outer cross-validation loop). The prediction performance was then evaluated using the test set from the outer cross-validation loop. Thus, the outer loop test set is never used for model optimization, but only for evaluation of overall prediction performance. This nested cross-validation procedure is repeated for $n$ outer cross-validation rounds, and Fig. 2 provides a schematic overview of the procedure. For evaluating the prediction in regression, the $Q^2$ metric was used. In discriminant analysis, the mean sensitivity and the area under the receiver operating characteristic curve (AUC, for two classes) were evaluated. For the presented results, we used Monte Carlo cross-validation (‘mccv’) for regression problems and the class-balanced mccv approach (‘mccv$b$’) for discriminant analysis, with 10 and 100 repetitions in the inner and outer loop, respectively and 75% of the data in the training and 25% in the test set for both the inner and the outer loop. Identical training and test sets were used to generate SA-K-OPLS and OPLS models in order to achieve a fair comparison. The score plots and reported kernel parameter values (σ, Table 1) were generated based on a kernel-optimization and K-OPLS model using the full data set, as this model would be the one that could be used for prediction of (future) new data (see Supplementary data 2 for an example of external validation). The convergence plots (see Supplementary data 3) demonstrate that the prediction performances stabilize.
within the 100 outer cross-validation rounds to provide reliable estimates of mean prediction performance.

2.7. Number of Y-orthogonal components

The number of components for the OPLS and SA-K-OPLS models presented was chosen by evaluating results for models calculated with 0 up to 10 orthogonal components (OC); all models required the use of 1 predictive component (because Y was a single vector or a two-class dummy matrix). The kernel parameter is optimized separately for each number of calculated components in the model (e.g. with 2 OC, the kernel parameter will be slightly different than with 1 OC). The final number of orthogonal components was determined as follows: additional components were included if $Q^2_Y$ increased by more than 1% per component for regression analysis, and in discriminant analysis the number of components corresponding to the maximum AUC was chosen. Table 1 lists the final number of components used for each data set. The optimal number of components is determined separately for SA-K-OPLS and the linear OPLS models, as the models are different in nature.

2.8. Software implementation

The simulated annealing K-OPLS method is available as a software package, which can be downloaded from [http://sourceforge.net/projects/kopls/](http://sourceforge.net/projects/kopls/) and is freely provided under a GPL v2 license. The updated K-OPLS [40] software (version 1.1.1) is available for Matlab and R on Windows, Mac OS X and Linux operating systems and includes functionality for nested cross-validation and kernel parameter optimization using simulated annealing and grid search. All results presented here were obtained using Matlab (The Mathworks, Natick, MA, USA).

3. Data sets

3.1. Galactosamine toxicity

The galactosamine toxicity data set comprises $^1$H NMR urinary spectra from animals treated with galactosamine hydrochloride ($n = 40$, 500 mg kg$^{-1}$). Galactosamine has been shown to induce a highly variable pathological response in a given cohort of rats (Ref. [56] and within): it is common that a proportion of animals show no adverse effects (non-responders) and responders display varied severities of hepatic necrosis. Independent histopathological and clinical chemistry analyses were used to determine the degree of response to galactosamine. Urine samples were obtained at 8 and 16 h collection intervals representing $−48$ to $−40$ (8 h), $−40$ to $−24$ (16 h), $−24$ to $−16$ (8 h) and $−16$ to 0 (16 h) h before dosing (samples courtesy of the COMET-2 project), as described previously in Coen et al. [56]. For this data set, the goal was the classification of non-responders ($n = 10$) versus responders ($n = 30$) from the presto urinary $^1$H NMR spectra (acquired at $−40$, $−24$, $−16$ and 0 h before dosing), and this was performed using discriminant analysis with SA-K-OPLS.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Biofluid type</th>
<th>Model type</th>
<th>Size $[N \times k]$</th>
<th>Response variable</th>
<th>$\sigma$</th>
<th>Prediction performance</th>
<th>#OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GalN - resp.</td>
<td>Urine</td>
<td>DA</td>
<td>$159 \times 830$</td>
<td>2 classes</td>
<td>20.06</td>
<td>OPLS</td>
<td>0.834</td>
</tr>
<tr>
<td>HgCl$_2$ - high dose</td>
<td>Urine</td>
<td>RE</td>
<td>$60 \times 830$</td>
<td>9 time points</td>
<td>21.62</td>
<td>OPLS</td>
<td>0.876</td>
</tr>
<tr>
<td>HgCl$_2$ - two doses</td>
<td>Urine</td>
<td>RE</td>
<td>$123 \times 830$</td>
<td>9 time points</td>
<td>24.28</td>
<td>OPLS</td>
<td>0.615</td>
</tr>
<tr>
<td>TBB - class.</td>
<td>Plasma</td>
<td>DA</td>
<td>$108 \times 950$</td>
<td>2 classes</td>
<td>20.29</td>
<td>OPLS</td>
<td>0.666</td>
</tr>
<tr>
<td>TBB - time</td>
<td>Plasma</td>
<td>RE</td>
<td>$64 \times 950$</td>
<td>6 time points</td>
<td>22.66</td>
<td>OPLS</td>
<td>0.584</td>
</tr>
</tbody>
</table>

The prediction of responders in galactosamine toxicity (GalN), the prediction of time after intoxication with mercuric chloride (HgCl$_2$) for a ‘high dose’ level and two dose levels and the classification of infection with *T. brucei* brucei parasite (TBB) and the time after infection with TBB were modeled (using discriminant analysis (DA) or regression (RE) models). The Gaussian kernel parameter ($\sigma$) of the full model is shown; this value is greatly dependent on data set and scaling, and should therefore not be used as a guideline. The prediction performance is given in terms of AUC for discriminant analysis and $Q^2_Y$ for regression. The number of orthogonal components (#OC) is indicated for both models.
3.2. Mercuric chloride toxicity

Animals were administered mercuric chloride, full details are given in Lindon et al. [29]. Mercuric chloride and its possible metabolic derivatives are NMR-invisible, and thus all observed changes can be attributed to endogenous metabolites. Doses were given at two levels (n = 10 ‘low’ dose, 0.5 mg kg\(^{-1}\), and n = 10 ‘high’ dose, 2.0 mg kg\(^{-1}\)) and urinary samples were obtained at time points before (0 h) and after dosing (8, 24, 48, 72, 96, 120, 144 and 168 h post-dose) and \(^1\)H NMR spectra were measured. Half of the animals at each dose level were sacrificed at 24 h after dosing, the other animals were sacrificed at 168 h. The time course (9 time points) is modeled with SA-K-OPLS regression.

3.3. Trypanosoma brucei brucei parasitic infection

To evaluate the effect of \(T.\) brucei brucei parasite infection in mice, plasma samples were collected from an infected group (n = 12), and a control group (n = 12) of animals. Animals in both groups provided samples (108 in total) at a series of time points pre- and post-infection (−2, 1, 7, 14, 21, 28 and 33 days after infection), as described by Wang et al. [23]. The presence of infection was established using discriminant analysis of pre-infection and control animals versus infected animals. This was followed by a regression analysis, on the infected animals only (64 samples), to predict the time after infection.

3.4. Data processing

Proton (\(^1\)H) NMR spectra were acquired as previously detailed [25,29,56]. As is standard, spectra were imported with \(\delta\) 0.0005 resolution. Spectral regions, excluding the water resonance for all spectra and the urea peak for urine samples, were uniformly binned with \(\delta\) 0.01 resolution, where a summed spectral intensity is calculated for each bin. For plasma NMR spectra we used the spectral regions \(\delta\) 0.00–4.60 and 5.10–10.00 and for urine NMR spectra, this was \(\delta\) 0.60–4.70, 5.05–5.50 and 6.25–10.00. Data quality control was performed by identifying and removing outliers from an initial investigation with principal component analysis. Samples were only omitted if they occupied an extreme value in score space, for example for technical reasons such as poor water peak suppression and lower than acceptable signal-to-noise ratio. The resulting spectra were subsequently normalized with probabilistic quotient normalization [57], and variables were mean-centered and scaled to unit variance. The data set sizes and SA-K-OPLS model details can be found in Table 1.

4. Results

4.1. Prediction performance and optimization of the kernel parameter

A summary of prediction performances and optimized Gaussian kernel parameter (\(\sigma\)) values for each data set can be found in Table 1. See Supplementary data 3 for figures displaying the convergence of the prediction performance metrics during cross-validation. Detailed results for each data set are presented in the following sections.

4.2. Prediction of non-responders in galactosamine toxicity

Galactosamine has been widely used as a hepatotoxin to model hepatitis, although numerous studies have identified significant variability in response to the toxin [56]. In drug discovery, the ability to prospectively identify and account for non-responders would greatly reduce the number of patients needed in clinical trials [58]. Attempts have been made to predict non-responders from their urinary metabolic profiles prior to [22,59], or soon after dosing [60] in pharmacometabonomics.

Non-responders (n = 40) and responders (n = 119) to galactosamine, as determined by histopathological and clinical chemistry markers of necrosis, were classified from the baseline (i.e. pre-dose) urinary \(^1\)H NMR spectra using SA-K-OPLS discriminant analysis. An increased prediction performance was observed with respect to the linear OPLS model: the mean sensitivity increased from 0.75 to 0.79, and the AUC from 0.84 to 0.88, using 5 and 3 orthogonal components for OPLS and SA-K-OPLS, respectively, see also Fig. 3. The ROC curves for OPLS and SA-K-OPLS are shown in Supplementary data 4.

This data set was also used to evaluate the prediction performance of SA-K-OPLS on an (external) test data set, generated by splitting the data into a modeling and an (external) validation set. Thus the nested cross-validation was used to create a model based on 120 observations, and the 39 removed samples from the validation set were predicted using this model. Also in this instance, the prediction improved for SA-K-OPLS compared to OPLS, further details and results can be found in Supplementary data 2. The discrimination between the classes is clear in Fig. 4, and an interesting, initially unexpected source of variation is visible, which could be associated with the collection period (urine was collected from −40 to −40 h and −24 to −16 h, whereas two other fractions represented collections from −40 to −24 and −16 to 0 h). It is visible that the 16 h (night) collections have lower values on the first orthogonal component (OC 1), and the 8 h (day) collections appear to have a higher value on OC 1. The orthogonal variance is therefore likely to reflect diurnal variation, which has been shown to exert a subtle impact on metabolite profiles of urine and other biofluids [61]. The score plots of KPLS (using the same kernel parameter and number of components as SA-K-OPLS), shown in Supplementary data 5A, do not allow this observation and separation of orthogonal variation. This demonstrates one of the unique features of K-OPLS compared to other kernel-based methods: the ability to capture and interpret non-linear structured variation in a data set that is not related to the response variable (Y) to provide further interpretation and understanding of model and data. The score plots of OPLS models based on this data set are shown in Supplementary data 6A.

4.3. Time-course modeling of mercuric chloride toxicity

Mercuric chloride is a well-studied nephrotoxin targeting the proximal tubules [62,63]. A single dose induces a reversible lesion demonstrating a clear dynamic profile of metabolic sequelae. Urinary \(^1\)H NMR data were used to evaluate the prediction of time after dosage with mercuric chloride (high dose level). This is of relevance for toxicological studies, because a good delineation of the time course describing the onset, evolution and regression of a toxic lesion is of value for comparison with the gold standard histopathology approach, to identify fast and slow responding animals, and for selecting the most appropriate sampling times for other assays such as transcriptomics [64]. The prediction of time after intoxication was improved using kernel methods compared to linear OPLS: Q\(^2\) was 0.67 compared to 0.61 (4 orthogonal components for both OPLS and SA-K-OPLS), see Fig. 3. Interpretation of the orthogonal component also revealed inter-animal variation in response: Fig. 5A shows that animal 28 exhibits a different response compared to the other animals, and animal 29 also diverges from the other animals at later time points. This was confirmed independently in the serum clinical chemistry data, where at 24 and 168 h after dosing animal 28 showed lower levels of typical markers such as creatinine, gamma-glutamyl transpeptidase, urea nitrogen and phosphate than the rest of the high dose group, and similar effects were observed for animal 29 at 168 h after dosing. Thus, these
animals were correctly identified as having a different outcome, without prior knowledge or consultation of the metadata.

A regression model was also constructed where data from the low and high dose levels of mercuric chloride were combined. Thus, this model is focused on the common time-related metabolic response across toxin dose-levels, allowing prediction of the time from dosing, in analogy with clinical situations where the dose level and time after exposure may not be known. Also for this model, the prediction performance increased for SA-K-OPLS compared to OPLS (see Fig. 3, $Q^2$ is 0.66 compared to 0.58). Moreover, the two aberrant animals 28 and 29 in the high dose group discussed above now cluster with the low dose animals in the score plot (see Fig. 5B).

**Fig. 3.** Bar graph of prediction performance. The difference of the prediction performance results for SA-K-OPLS and OPLS models was evaluated using 100 outer cross-validation rounds (mean and the standard error of the mean, representing the variability in the 100 cross-validation results, are indicated as bars with error bars). Regression analysis results (mercuric chloride (HgCl$_2$) data sets, and the *T. brucei brucei* (TBB) time after infection data) were evaluated with $Q^2$. Discriminant analysis results are presented with both the mean sensitivity and the area under the receiver operating characteristic curve (AUC), for classification of non-responders versus responders in the galactosamine toxicity data set (GalN resp.) and the separation of infected versus non-infected samples from the *T. brucei brucei* infection study (TBB class). The prediction improvement of SA-K-OPLS over its linear equivalent OPLS is visible for all these models.

**Fig. 4.** Score plot showing discriminant analysis of non-responders and responders to the toxin galactosamine. A score plot of the predictive and first orthogonal component of the galactosamine toxicity data set shows a clear discrimination in the predictive component (x-axis) between responders and non-responders (circles and squares, respectively). The first orthogonal component (y-axis) shows a trend related to diurnal variation (dark versus light periods) with respect to the urine collection time period (either −48 to −40 and −24 to −16 h; 8 h (day) collection; and −40 to −24 and −16 to 0; 16 h (night) collection).
indicating that they show a less severe response to the toxin, in agreement with the histology results. This example therefore confirmed that SA-K-OPLS not only improved the predictive ability of the data set compared to OPLS, but moreover, unlike other non-linear kernel-based methods, provided a visual interpretation of the variation in the data set, related to the predictive variation as well as structured $Y$-orthogonal variation, which was not observed for the corresponding KPLS models (see Supplementary data 5B and C), and OPLS models (Supplementary data 6B and C). The two presented toxicology data sets clearly demonstrate the value of using non-linear methods in these metabonomic studies, and in particular the value of K-OPLS for improved model interpretation.

### 4.4. Classification of T. brucei brucei parasitic infection

Here, we evaluated the $^1$H NMR data of plasma samples from a T. brucei brucei parasitic infection (African sleeping sickness) in mice [25]. Spectra were acquired from the animals before and at various stages during the infection and similarly from control animals. Data from infected animals formed the ‘infected group’, the ‘uninfected group’ consisted of the combined pre-infection and control animal data. The discrimination between the infected and uninfected groups resembles the situation in a clinical setting, where it is crucial to establish the occurrence of infection, irrespective of the elapsed time after initial infection. Improved classification accuracy between ‘infected’ and ‘uninfected’ individuals with SA-K-OPLS was observed compared to linear modeling, see Fig. 3 and Table 1, with an increased mean sensitivity (SA-K-OPLS: 0.92 with 1 OC, OPLS: 0.88 with 2 OC), as well as a larger AUC (0.97 compared to 0.96), the ROC curves are plotted in Supplementary data 4. In the K-OPLS score plot (Fig. 6A), it is visible that the early time point of infection (day 1) is more difficult to discriminate from the uninfected spectra than later time points, which can be expected from the development of a parasitic infection, see next section. KPLS and OPLS score plots are shown in Supplementary data 5D and 6D, respectively.

### 4.5. Prediction of T. brucei brucei parasitic infection time-course

The status of the T. brucei brucei infection in terms of the time after initial infection is highly relevant as the nature of therapeutic intervention and the prognosis for recovery are greatly dependent on the event of blood–brain barrier crossing of the parasite that occurs in this infection. As with the mercuric chloride toxicity data set, a metabolic trajectory reflecting the progression of the infection can be characterized using spectral profiling [25]. To stage the infection time point, the time after infection was modeled non-linearly as a regression against time, as displayed in Fig. 6B; the corresponding OPLS and KPLS models are shown in Supplementary data 5E and 6E. It was found that the prediction was improved for the SA-K-OPLS model compared to the results from the linear OPLS model: $Q^2$ increased from 0.64 in OPLS to 0.68 in SA-K-OPLS, both with 3 orthogonal components, see also Fig. 3 and Table 1.

### 5. Discussion

For certain metabonomic applications, predictive performance is of paramount importance, such as in toxicology studies and disease diagnostics. Although linear models often provide a basic level of prediction performance for multivariate data sets, in the presence of substantial non-linear variability in data, non-linear models are expected to provide improved predictions compared to linear models. This might be explained by the potential ability of a more flexible non-linear model to incorporate additional predictive information if the predictor and response variables are related non-linearly, but also by the aggregation of information across many weakly associated descriptor variables, linearly and non-linearly related to $Y$, which contributes to an improved, and potentially also more robust, prediction model.

Prediction performances reported here were calculated by nested cross-validation, which reduces the risk of over-fitting when optimizing the kernel parameter [54]. If obtainable, an external test set should ideally be used to evaluate prediction performance, however the nested cross-validation is expected to give an unbiased estimate of predictive ability if no external test data are available. We note that simulated annealing is a stochastic optimization
method, and therefore the end result of every optimization will vary slightly. To ensure that the prediction performance estimates were reliable for comparison, and not unduly affected by random variability relating to the cross-validation procedure, we monitored the convergence of the 100 cross-validation rounds, seen in Supplementary data 3. For the data sets analyzed here, we observed that predictions by SA-K-OPLS were better than, or equivalent to, the linear OPLS model. However, the additional assumption of linearity in the OPLS model may result in better prediction performance under certain specific conditions, for example if the problem is truly linear and the available training data are very limited in the number of observations.

Although the K-OPLS method does not necessarily provide improved prediction performance compared to other kernel-based models [30], K-OPLS facilitates an improved model interpretation compared to alternative models, which can aid quality control and further understanding of the model and data (see Supplementary data 5). The difference between K-OPLS and KPLS is analogous to the difference between OPLS and PLS: SA-K-OPLS provides a separate set of predictive and Y-orthogonal model components, a property that caused OPLS to gain great attention for analysis of metabonomic data [12]. Y-orthogonal variation may be interpreted as multivariate latent variables that describe unknown covariate variation, for example unknown batch effects, animal variability or drift in analytical instruments over time, thus providing additional means for quality control and understanding of data. The score plots presented in Figs. 4 and 5 clearly demonstrate the interpretation of Y-orthogonal model components, in addition to the strength of prediction using a non-linear modeling approach.

To find a good parameter value for the kernel function, grid search or gradient minimization could be applied [35,41], but these approaches may be computationally and labor intensive or inefficient and can get trapped in local minima. There is a multitude of global optimization methods available, such as simplex-based methods [65] and genetic algorithms [51,66], previously reported in the context of support vector machine applications. We have demonstrated how simulated annealing can be used to achieve good optimization performance in a user-friendly manner, but other approaches could also have been applicable to this optimization problem. We note that the kernel parameters for the different analyzed data sets had comparable magnitude (Table 1). This cannot, however, be assumed in general and is possibly related to the fact that these data sets were all acquired by comparable 1H NMR spectroscopy procedures and pre-processed in a highly similar manner. Our goal here was not to compare different optimization methods, but to provide and demonstrate an automated and easy-to-use procedure. Grid search optimization has also been implemented as an option in the K-OPLS software package.

In contrast to linear methods, there are no model loadings accessible for interpretation for kernel-based methods, due to the usage of the kernel trick. This may be a disadvantage of the kernel-based non-linear modeling, depending on the application, and therefore the method is particularly suited to studies where maximum predictive performance is imperative, for example in toxicology applications and disease diagnostics using multivariate biomarker panels.

We have demonstrated that SA-K-OPLS can provide an improved prediction performance compared to the linear OPLS method in the context of 1H NMR spectral data sets of two biofluids: blood plasma and urine. The presented method is expected to be applicable to other data sets, not restricted to analytical method or study type. SA-K-OPLS is especially appropriate to be used for predictive modeling if the data show presence of non-linear variability and could potentially contain systematic Y-orthogonal variation.

6. Conclusions

Predictive modeling in metabonomic studies is currently almost exclusively performed using linear methods. We demonstrate here that a non-linear kernel-based model can provide valuable prediction performance improvements compared to the linear model in three typical metabonomic data sets. The kernel-based OPLS (K-OPLS) model was used in this study, which is analogous in its model structure to the commonly used linear OPLS model. We found that the non-linear (K-OPLS) prediction performance is better than or equal to the linear model (OPLS). K-OPLS provides added value compared to alternative kernel-based models by enabling separate modeling of predictive and Y-orthogonal model components, facilitating improved model interpretation and providing a means of quality control.

A complicating factor for non-expert users in the application of kernel-based models can be the optimization of the kernel function parameter. We demonstrated how simulated annealing can be applied successfully to automate optimization of the kernel parameter, providing a user-friendly procedure based on nested cross-validation. Thus, SA-K-OPLS has been shown to provide a flexible framework for non-linear classification and regression modeling and we anticipate that the method can also be employed successfully in studies with data collected by other analytical methods and for modeling of other ‘omics’ data types, as well
as high-dimensional and potentially non-linear data in general [39,41].

We hope that our results will facilitate increased utilization of non-linear prediction models in metabolic profiling studies, especially in application areas such as disease diagnostics and toxicology, where prediction performance is of utmost importance. The K-OPLS procedures applied here are available in the K-OPLS software packages for R and Matlab, including cross-validation procedures, kernel-optimization and prediction performance metrics.

Acknowledgements

The authors are grateful to COMET-1, COMET-2 and Y. Wang for provision of the metabolic NMR spectroscopic data. This work was supported by the award of an RSC PhD studentship to JMF. The MRC Integrative Toxicology Training Partnership (ITTP) is acknowledged for financial support to MC as a career development fellow.

Appendix A. Supplementary data


References

Appendix E

Paper IV
Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework

Selena E. Richards⁎, Marc-Emmanuel Dumas, Judith M. Fonville, Timothy M.D. Ebbels, Elaine Holmes, Jeremy K. Nicholson

Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington London, SW7 2AZ, UK

A R T I C L E   I N F O

Article history:
Received 19 February 2010
Received in revised form 28 June 2010
Accepted 22 July 2010
Available online 3 August 2010

Keywords:
Chemometrics
Integration
Fusion
Metabonomics
Metabolomics
Systems biology

A B S T R A C T

The explosion of -omics technologies in the characterization and prediction of defined physiological or pathological states necessitates a parallel development in data integration and visualization tools in order to display and interpret vast amounts of data in an efficient manner. Here we summarize some of the key achievements in this area and compare the strengths and limitations of each method with respect to their use in representing biological processes in a systems biology environment.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Metabonomics is a metabolism driven top-down systems biology approach for detecting, identifying, analyzing and cataloguing time-related metabolic changes in complex organisms that have been subjected to a biological challenge [1]. The approach is defined as “the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification” [1]. This concept arose mainly from the work on the application of nuclear magnetic resonance (NMR) spectroscopy to study the metabolic composition of biofluids, cells and tissues. A parallel approach, metabolomics arose in the late 1990s, mainly from plant and microbial sciences and originally from the study of in vitro cellular systems [2]. Historically, metabolomics was defined on the basis of analytical rather than biological criteria, referring to the total sum of measurable metabolites within a cell or tissue type. In general metabolomic experiments require comprehensive analysis in which all the metabolites of a biological system are identified and quantified, whereas metabonomic experiments require the identification and characterization of the metabolic response to biological stimuli. The analytical methods and multivariate statistical approaches used for cells, plants, and animals are now highly convergent and for the purpose of this review we will refer to the combined field as metabonomics [3,4]. Powerful analytical strategies, in combination with advanced multivariate statistical tools, are required to provide maximum relevant information and knowledge about the complex biological system to enable the detection and identification of endogenous and exogenous metabolites responsible for mediating the phenotypic expression of altered metabolism resulting from a biological challenge [5,6]. Genomics, proteomics, metabonomics form the fundamental systems biology framework and can give an integrated understanding of the cellular functions in living organisms. In these -omics fields of study vast amounts of quantitative or semi-quantitative data are derived covering a variety of levels of biomolecular organization. One of the expectations of the systems biologist is that these datasets can be integrated to give a holistic picture of the state of the system, e.g. development, ageing, health or disease that provides insights that enable a more fundamental understanding network connections of interactions at the molecular level [7–14].

In this review we discuss the various efforts that have been made to integrate information from different -omics datasets. Although there are multiple ways to classify levels of data integration, we recognize three main levels of integration [15]: conceptual integration, the integrative interpretation of results from multiple experiments; statistical integration, the simultaneous multivariate analysis of data from multiple sources; and model-based integration, the construction of predictive functional networks or models from multiple experimental components. Here, we focus on the recent developments and applications at the statistical level of data integration for metabonomic applications. The advantages and disadvantages of the statistical methodologies for each application...
are discussed, initially for statistical integration strategies using different analytical techniques for the same biological sample. This is followed by the integration of metabolic data from different biological samples taken from the same comprehensive metabolic studies. These include, for example, analysis of various tissues, tissue extracts and biofluids, in what has been termed ‘integrative metabonomic’ investigations. In the second section of the review we discuss the ‘holy grail’ of systems biology, that is the intelligent integration of data from different -omic studies at different levels of molecular biology, such as transcriptomics, proteomics and metabolomics, such that the data accurately reflect the biology of the whole organism, and we present and review where this has been achieved at the level of interpretation and data construction. Finally the complexities of modeling biological systems are discussed, and methods for achieving full systems biology integration are proposed.

1.1. Overview of current analytical technology used in metabolic profiling

Bio-analytically, nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the most powerful and widely used spectroscopic techniques for both untargeted and targeted metabolic profiling, which is the compositional analysis of low molecular weight species in complex biological samples. The metabolic profiles can be used for two purposes: either to characterize the metabolome of an organism or to identify metabolites or ‘candidate’ biomarkers (whose identity may or may not be known) that are associated with a particular biological state in biofluids and tissue samples [8,16–22]. $^1$H NMR spectroscopy has several advantages for measuring metabolite levels including minimal sample preparation, data acquisition typically takes only a few minutes per sample and so high-throughput and automated analysis is possible. Furthermore, it is non-destructive and cost effective. Undoubtedly the largest disadvantage of NMR spectroscopy is its poor sensitivity in comparison to MS techniques, which limits the number of observable molecular species. However, the high degree of connectivity within metabolic networks can somewhat reduce the problem of comparatively low sensitivity, as changes in low concentration (invisible or latent to NMR spectroscopy) metabolites may lead to indirect changes in higher concentrations (visible to NMR spectroscopy) metabolites. Additionally, NMR spectroscopy has been repeatedly shown to be highly reproducible with very low coefficients of variation in metabolic studies [23,24]. Alternatively, MS-based metabonomics offers quantitative analyses with high selectivity and sensitivity. Combination with a separation technique such as gas chromatography/MS (GC/MS) [25–27], high performance liquid chromatography/MS (HPLC/MS) [28–30] and capillary electrophoresis/MS (CE/MS) [31–33] reduces the complexity of the mass spectra due to metabolite separation in a time dimension, reduces ion suppression, can provide isobar separation, i.e. the differentiation between molecules with similar nominal mass and different molecular formula, and delivers additional information on the physico-chemical properties of the metabolites. However, mass spectrometry-based techniques usually requires a sample preparation step, which can cause metabolite losses, and based on the sample introduction system and depending on the ionization technique used, specific metabolite classes may be discriminated or metabolites differentially ionized. The largest disadvantage of MS in comparison with NMR spectroscopy is its lower reproducibility. Nevertheless, both NMR spectroscopy and MS provide complementary information to the analytical scientist. Other spectroscopic methods that can produce metabolic signatures of biomaterials include optical spectroscopic techniques such as Raman [34,35], Fourier transform infrared (FTIR) [36,37], near infrared (NIR) [38] and ultraviolet–visible (UV–VIS) [39] spectroscopy, for a more comprehensive review see reference [3].

1.2. Properties of the various biological samples used in metabolic profiling

The various physiological properties of mammalian biofluids have been reviewed in relation to spectroscopic information recovery [177,178]. The spectroscopic measurement of biofluids or tissues provides information on chemical structure, quantitative composition, metabolic pathway relationships as well as latent information on the status of the organism as a function of the genetic background, environment or gene–environment interaction [40]. Each biofluid yields a characteristic spectroscopic fingerprint in which the spectral intensity distribution is determined by the relative concentration of the solutes, and in some cases by their intermolecular interactions, an example of an 800 MHz $^1$H NMR spectrum of untreated human urine is given in Fig. 1. Although metabolite concentrations can vary dramatically between individuals and between metabolites in a single individual, significant differences between groups can be optimized through appropriate scaling of the data [41]. Most mammalian metabonomic analyses are performed on plasma, serum and urine because they are easily collected and reflect instantaneously sampled states in dynamic equilibrium with multiple cells and tissues of the body and have traditionally been used for prognosis or diagnosis of many diseases. Other biofluids used for metabolic profiling include, saliva [42], amniotic fluid [43,44], cerebral spinal fluid [45,46], breast milk, synovial fluid, seminal plasma [47], gastric, bile and digestive fluids. Biological tissue samples used for metabolic profiling include liver [48–50], kidney [51,52], feces [53–55], white adipose tissue (WAT) [56] and brown adipose tissue (BAT). Tissues are ideal for biomarker discovery because they are under tight homeostatic regulation and can provide highly consistent organ specific metabolic measurements between individuals leading to a precise metabolic characterization of groups or dynamic processes. The difficulty of using tissue samples are that sample acquisition is invasive and samples generally require more preparation steps in comparison to biofluids prior to analysis. The metabolic information provided by biofluids and tissues can be used to characterize multicompartmental consummate events such as toxicity and disease [57,58], thus giving a more complete description of the mechanistic consequences than can be obtained from one fluid or tissue alone. There are a large variety of samples used for plant metabolomics and include leaves, roots, sap, fruits, stalks, tubers, flowers, derived materials (e.g., oil, wine, resins), reflected in the wide number of reported studies in this field [59–62] as well as cells in microbial metabolomics [63], metabolomic analysis of nematodes [64,65] and non-model organism [66], etc.

1.3. Chemometric analysis of spectral data

A limiting factor in understanding the biochemical information from both spectral profiles of tissues and biofluids is their complexity; high resolution spectra of biofluids may contain several thousand peaks from potentially 100s or 1000s of metabolites (Fig. 1). Therefore multivariate statistical chemometric techniques are required to reduce the dimensionality of the data in order to extract the defining metabolic information associated with each group or dynamic process. Multivariate statistical methods include unsupervised pattern recognition methods such as principal component analysis (PCA) [67–69] that reveals patterns, trends and outliers in these data, via an orthogonal decomposition. Self modeling curve resolution (SMCR) [70–73] resolves the underlying metabolic signatures (spectrotypes) and their relative contributions which are associated with, for example, dynamic biological processes, such as aging, via linear decomposition followed by alternating least-squares refinement. Supervised pattern recognition methods, such as partial least-squares (PLS) regression [74], try to maximize the joint covariance structure between a measurement matrix, $X$ and a second matrix, $Y$ containing response, compositional or class identification information, with the
goal of identifying biomarkers that are associated with particular classes or responses. The implementation of filtration steps are often incorporated into the multivariate methods, using for example orthogonal partial least-squares methods OPLS [79] and O2PLS [80]. Other commonly employed methods include neural networks [81,82], kernel methods [83–85] or probabilistic approaches [86,87], which offer comparable or even superior classification potential but do not provide transparency in terms of the discriminating signals.

Fig. 1. 800 MHz $^1$H NMR spectrum of untreated human urine. The $^1$H NMR spectrum was recorded on a Bruker Avance II 800 NMR spectrometer (Bruker, Germany), operating at 800.00 MHz for $^1$H. A standard one dimensional $^1$H NMR spectrum with water peak presaturation was acquired using the pulse sequence (RD-90-$t_1$-90-$t_m$-90-acq), where RD is a relaxation delay and $90$ represents 90° high power pulses. Here, the inter-pulse delay $t_1$ was 3 μs and the mixing time $t_m$ was 100 ms. A weak irradiation field was applied at the water resonance frequency during both the mixing time and the recycle delay. A total of 1024 scans were collected into 64 k data points with a spectral width of 16,025 Hz and relaxation delay of 3 s. The acquisition time was 2.045 s and the total pulse recycle time was 5.2 s. $^1$H NMR spectra were corrected for phase and baseline distortion using Topspin 2.0 (Bruker). The spectra were referenced to the chemical shift of 3-(trimethylsilyl)-2,2,3,3-tetradeuteropropionic acid (TSP) resonance at 0.00 ppm.

There are several multivariate statistical tools available for the co-analysis of multiple data sets using either (a) multiblock/multiset model, applied to either the individual block data or the two-way matrices; $X_i(n \times mq)$ or $X_j(nq \times m)$ and (b) three-way methods applied to the three-way matrix $X_{i \times j \times q}$, when the length of at least two dimensions are common, see reference [90] for extensive reviews of three-way multivariate analysis methods. In this review various multivariate statistical approaches applied and developed for the simultaneous analysis of metabolomics and multi-omics datasets and the statistical analysis of the signal amplitude co-variation are discussed.

2. Methods for organizing -omics data prior to data integration

For simultaneous analysis of data derived from multiple sources, e.g. NMR spectroscopy and MS, multiple NMR nuclei ($^1$H, $^{13}$C, $^{19}$F, $^{31}$P), multiple biological samples (urine, liver), or -omics experiments (DGGE, NMR spectroscopy) can be arranged in different directions [88]. In typical metabolomic applications, rows in the data matrix represent mixture spectra recorded for each participant in the study (n), columns represent measurements on multiple variables, i.e. chemical shift, mass to charge ratio, etc. (m), and the third dimension, sometimes called tubes, represent the multiple experiments, i.e. different analytical spectroscopic measurements, different NMR nuclei, different biological samples, different -omics experiments, or different time points (time-series or batch analysis) [89] etc. Multiple datasets with at least one order in common can be arranged in three different directions: along the row space (n), along the column space (m), or along the tube space (q). The three procedures give a row-wise augmented matrix $X_i(n \times mq)$, a column-wise augmented matrix $X_j(nq \times m)$, and a tube-wise augmented matrix $X_{i \times j}(nm \times q)$ respectively.

2.1. Integration of data from different spectroscopic platforms: inter-platform integration

One of the most useful recent advances in data modeling is the statistical integration of molecular profiles of the same sample using different spectroscopic platforms. This enables molecular biomarker identification and information recovery on metabolic pathway connectivities by the examination of different levels of metabolic information from the statistical integration of analytical platform–platform variation. Various multivariate methods for inter-platform integration have been proposed. These methods range from extraction of simple correlations to trilinear projection-based approaches. Selected examples of currently applied integration methods are given below.

A correlation-based multiblock hierarchical PCA (HPCA) [91,92] has been used for the simultaneous analysis of $^1$H NMR spectroscopy combined with gas chromatography-electrospray ionization time-of-flight MS ($^1$H NMR spectroscopy and GC-EI-TOF-MS) for the global compositional analysis of melon fruit flesh extracts [62]. Appropriate scaling of the data is required to ensure the variation from one platform does not dominate or mask the variation from a second...
platform. HPCA and hierarchical PLS (HPLS) have been applied for the analysis of LC/MS and NMR spectroscopy for the classification of rats dosed with citalopram (positive control for phospholipidosis). In this study several different variable and block scaling methods were assessed. Variable scaling methods included division by the mean, unit variance scaling, Pareto scaling, variance scaling and vast scaling. Block scaling methods included those that ensured each block (LC/MS or NMR spectroscopy) was scaled to give equal mean intensity or each block was scaled to give equal sum of standard deviation. The optimization of the different scaling was performed by cross-validation. In this application the optimal block-scaling method for the row-wise augmented data was scaling each block to give equal sums of standard deviation [93]. A more generalized framework for the evaluation of sequential multiblock methods, found that consensus PCA (CPCA) [94], (which ensures that both blocks are considered equal by weighting each block prior to the PCA decomposition of the row augmented data) was superior in a separate evaluation [94]. Application of CPDA (using blocks weights equal to the square root of sum of squares) and comparison with canonical correlation analysis (CCA) [95] for the integration of microbial metabolomics datasets (GC/MS and LC/MS) from Pseudomonas putida S12 and Escherichia coli found CPDA focused on the direction describing the largest source of variance (i.e. the identification of the large common effects possibly associated with relations hips between metabolome and the specified metabolites) whilst CCA identified trends between the two datasets that were strongly correlated (i.e. possibly biochemically related metabolite associations) [96]. In a similarly designed experiment using LC/MS and GC/MS the proportion of the overlapping sources of variance from both datasets was quantified using CPDA; validated against the matrix correlation coefficient, (the RV-coefficient) [97]. The RV-coefficient was modified in a subsequent study to circumvent problems of high dimensional data [98].

Multiple factor analysis (MFA) [99,100], a similar method to CPDA, in which each block is weighted to the square root of the first eigenvalue derived from PCA of each block [99,100] rather than the square root of the sum of squares, followed by PCA decomposition of the row augmented matrix, has been applied for the co-analysis of $^1$H-$^{13}$C Heteronuclear Multiple Bond Connectivity NMR spectroscopy ($^1$H-$^{13}$C HMBC NMR spectroscopy) and Pyrolysis Metastable Atom Bombardment Time-of-Flight MS (Py-MAB-TOF-MS) to perform metabolic fingerprinting on cattle treated with anabolic steroids [101]. The advantage of the MFA weighting procedure is it accounts for dominant scaling effects. Furthermore, Van Deun et al. [102], in a separate study showed simultaneous component analysis (SCA) methods such as MFA, were largely dependent upon the pre-processing, weighting [102] and noise structure [103].

Three-way analysis methods that have been used for the full integrative modeling of multiple platform data include parallel factor analysis (PARAFAC) [104]. The requirement and hence limitation of the PARAFAC model is that the data, must be approximately low rank and trilinear to provide physically meaningful loadings (e.g. pure spectra) which limits the application of PARAFAC models to metabolomics spectroscopic datasets which are generally rank-deficient, containing high numbers of metabolites which are both biologically and statistically correlated and have overlapped spectroscopic peaks. Tucker decomposition [105–107], a generalized PCA on multiway data, has the same capabilities of PCA i.e. to compress variation, extract features and explore data. In addition to these capabilities, the presence of the core array in the Tucker decomposition allows for interpretation of the interaction between the different components, describing the variation between the samples, measurement on multiple variables and experiments. Furthermore, the three-way structure of the original data may allow for more direct interpretation of features common to the data [108,109]. Dyrby et al. [110] showed how a traditional dose–response experiment could be properly explored using Tucker3 models and PARAFAC was used for similar data [111]. Idborg et al. [112] exemplified the use of PARAFAC and N-way partial least squares (NPLS) [113] in handling metabolite screening using liquid chromatography/electrospray ionization MS (LC/El-MS). Lipoprotein characterization using 2D diffusion-edited NMR spectroscopy was analyzed with PARAFAC providing chemically meaningful spectra [114]. Visualization of important variables in each mode (samples, measurement on multiple variables and experiments) can sometimes be difficult with these models (i.e. NPLS, Tucker and PARAFAC). An analogy to the two-way correlation loading plot to assess variable significance was considered by Lorho et al. [115], and a generalization to the multiway case was developed using congruence loadings and applied to a metabonomics experiment using 2D diffusion-edited NMR spectroscopy [115]. ANOVA-simultaneous component analysis (ASCA) [116] has been developed to increase the interpretability of a multivariate dataset in terms of the experimental design (i.e. providing a more direct answer to the experimental question at hand) and extended to the multiway case by combining ASCA with PARAFAC (PARAFASCA) [117]. This approach was used to study the time-effect of hydrazine toxicity in the urinary composition of rats evaluated by $^1$H NMR spectroscopy.

Structural and biological connectivity between metabolites measured on multiple spectroscopic platforms have been achieved using correlation-based approaches, such as Statistical Total Correlation Spectroscopy (STOCSY [118]), which was primarily designed to aid peak assignments in NMR spectroscopic metabolic studies. The
The assumption underlying the STOSCY approach is that spectral peaks arising from the same compound across a series of samples are multicollinear, i.e., the relative signal intensities of a given molecule are constant. Thus the correlation of two non-overlapped peaks arising from the same compound throughout the array is high, whereas unrelated compounds will generally show a low correlation unless there is a biological linkage producing correlation. The ease with which STOSCY can be implemented resulted in valuable applications, including studies on analgesic usage in a large epidemiological study [119], pathway connectivities in relation to a renal toxin [120] and drug metabolism kinetics [121]. Both structural and novel toxicological connectivities between xenobiotic and endogenous metabolite signals were detected in the rat [122]. It should be noted, though, that basis for an observed correlation, is far from unambiguously decided, and a comparison of metabolite concentrations across various experimental conditions is suggested to evaluate and identify invariant features of metabolic profiles [123]. Integrative analysis using HET-STOCSY (heteronuclear statistical total correlation spectroscopy) based on STOCSY principles, can be performed across a set of samples by using two NMR spectroscopic data sets of different nuclei, for example 1H and 31P [124,125], Fig. 2, or 19F and 1H [126]. Outcomes of the analysis, contribute to confirmation of structural elucidation or pathway elucidation determined from other single platform analyses allowing an integrated overview of biological response to physiological or pathological stimuli.

The correlations between two different matrices derived from the sample set, e.g. arising from two analytical platforms, can be calculated. The first implementation demonstrated was the integration of 1H NMR spectroscopy and MS data, and resulted in the successful demonstration of Statistical Heterospectroscopy (SHY) [127,128]. As most analytical platforms contain highly complementary information, the cross-correlation of spectral parameters has a synergistic effect with increased information extraction, as bidirectional and unidirectional correlations between identified and unidentified signals can lead to mutually increased information content and significantly improved interpretability. The approach is not limited to these two platforms, and other examples include the integration of NMR spectroscopy and capillary electrophoresis [129]. Thus the statistical integration of molecular profiles of the same sample using different spectroscopic platforms enables molecular biomarker identification, information recovery on metabolic pathway and connectivities by the examination of different levels of the analytical platform–platform correlation matrices. A summary of published inter-platform statistical integration methods is given in Table 1.

### 2.2. Inter-sample type integration: integrative metabonomics

Integrative metabonomics [8,130] can be used to describe changes in metabolic chemistry in different body compartments that are caused by exposure to a biological intervention. Such timed profiles in multiple compartments are themselves characteristic of particular types and mechanisms and the statistical integration of multiple tissue samples can reveal metabolic information associated with e.g. the Cori cycle between muscle and liver or insight about renal clearance or retention from the co-analysis of plasma and urine, thus giving a more complete description of the mechanistic consequences than can be obtained from one fluid or tissue alone. There are many more examples of conceptual integrative metabonomics methods than statistical integration for different sample types. Typically, conceptual integrative models are constructed on each biofluid and (or) tissue using either unsupervised or supervised pattern recognition methods such as PCA, PLS-DA and OPLS-DA, followed by an integrated interpretation of the results from the different biological compartments. Examples of conceptual integrative metabolic include studies of toxicity [131–134], the effects of various pre- and probiotics on gut microbiota [58,135], nutritional studies [138] etc. The disadvantage of this approach is information relating to the common systematic variance across the different compartments and the detection of intercompartment functional relationships are hard to recover. There are few published studies that use true statistical integration for different sample types. Recently a systems-based approach relating observed biomolecular changes in biological fluids to drug induced toxicity effects in liver tissues for in vivo models, using a

<table>
<thead>
<tr>
<th>Integration strategy</th>
<th>Analytical components</th>
<th>Chemometric method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-platform</td>
<td>1H NMR and GC–EI–TOF–MS</td>
<td>HPCA [62]</td>
</tr>
<tr>
<td></td>
<td>1H NMR and LC–MS</td>
<td>HPCA and HPLS [93]</td>
</tr>
<tr>
<td></td>
<td>GC–MS and LC–MS</td>
<td>CPCA [95,96] and CCA [95]</td>
</tr>
<tr>
<td></td>
<td>1H–13C HMBC NMR and Py–MAB–TOF–MS</td>
<td>MFA [101]</td>
</tr>
<tr>
<td></td>
<td>LC–EI–MS</td>
<td>PARAFAC and NPLS [112]</td>
</tr>
<tr>
<td></td>
<td>2D diffusion-edited 1H NMR</td>
<td>PARAFAC [114] and congruence loadings [115]</td>
</tr>
<tr>
<td></td>
<td>1H and 31P NMR</td>
<td>HET-STOCSY [124,125]</td>
</tr>
<tr>
<td></td>
<td>19F and 1H NMR</td>
<td>HET-STOCSY [126]</td>
</tr>
<tr>
<td></td>
<td>1H NMR and UPLC–MS</td>
<td>SHY [127,128]</td>
</tr>
<tr>
<td></td>
<td>1H NMR and CE</td>
<td>SHY [135]</td>
</tr>
<tr>
<td>Inter-sample type integration:</td>
<td>Plasma and liver tissue</td>
<td>Correlation [139]</td>
</tr>
<tr>
<td>Integrative metabonomics</td>
<td>Urine, plasma, liver, pancreas and kidney cortex</td>
<td>Correlation [136]</td>
</tr>
<tr>
<td></td>
<td>Liver and feces</td>
<td>Correlation [137]</td>
</tr>
<tr>
<td></td>
<td>Jejunum, ileum, colon, liver, spleen and kidney</td>
<td>OSC-PLS and HPCA [140]</td>
</tr>
<tr>
<td></td>
<td>Plasma, liver, pancreas, adrenal gland and kidney cortex</td>
<td>NPCA, MCR-ALS, PARAFAC [141]</td>
</tr>
<tr>
<td>Inter-omic integration</td>
<td>Metabonomics and genomics</td>
<td>Correlation networks [160]</td>
</tr>
<tr>
<td></td>
<td>Proteomics and genomics</td>
<td>Correlation networks [159]</td>
</tr>
<tr>
<td></td>
<td>Transcriptomics and metabolomics</td>
<td>Correlation [158]</td>
</tr>
<tr>
<td></td>
<td>Proteomics, genomic and metabolomics</td>
<td>Correlation [139]</td>
</tr>
<tr>
<td></td>
<td>Functional genomics and metabolomics</td>
<td>PLS [161]</td>
</tr>
<tr>
<td></td>
<td>Metabonomics and proteomics</td>
<td>O2PLS [162]</td>
</tr>
<tr>
<td></td>
<td>Transcriptomics and metabolites</td>
<td>O2PLS [163]</td>
</tr>
<tr>
<td></td>
<td>Proteomics, genomic and metabolomics</td>
<td>O2PLS [164]</td>
</tr>
<tr>
<td></td>
<td>Genomics and metabolomics</td>
<td>msQTL [168–170]</td>
</tr>
</tbody>
</table>
cross-tissue correlation network-based integration of molecular profiling data has been employed [139]. Correlation analysis has also been used to summarize intra- and inter-compartmental metabolic correlations derived from pair-wise statistical analysis to identify latent metabolic links between several biological compartments, such as plasma, liver and kidney cortex [136,137]. Other methods include, OSC-PLS and HPCA which have been used for the full statistical integration of jejunum, ileum, colon, liver, spleen and kidney samples, which enhanced and increased understanding of physiological and pathological consequences of a patent *Schistosoma mansoni* infection in mice [140]. HPCA has also been used to integrate multicompartment metabolic data in a probiotic study [F-P.J. Martin, Y. Wang, N. Sprenger et al, Molecular Systems Biology 4 (2008), pp. 1–15]. Spatiotemporal inter-organ metabolic control functions were modeled by Montoliu et al. [141], with various 2- and 3-way multivariate statistical approaches including PCA, multi-way PCA (N-PCA) [142], multivariate curve resolution-alternating least squares (MCR-ALS) [72] and PARAFAC to abstract the inter-compartmental functional relationships between healthy mice plasma, liver, pancreas, adrenal gland and kidney cortex. Of the multivariate approaches MCR-ALS and PARAFAC appeared to be better adapted for stepwise variable and compartment selection for further correlation analysis because both methods provided an overview of functional relationships across matrices and enabled the characterization of compartment-specific metabolite signatures (spectrotypes [71]).

**Fig. 3.** PARAFAC analysis. The first mode loading describes the relative contribution of the each spectrotpe associated with each individual. The second mode loading shows the four spectrotypes obtained from the PARAFAC analysis. The third mode loading describes the relative contribution of each spectrotpe in the individual compartments [141].
differentiation between both specific and common biochemical profiles associated with the different biological compartments [141]. A summary of published inter-sample type integration methods is given in Table 1.

2.3. Integration of data from different -omic platforms: inter-omic integration

In addition to characterizing inter-compartmental correlations and integrating multiple analytical data types, it is also desirable to integrate data from various -omics platforms. Bio-analytical measurements of metabolites and the expression levels of genes and proteins may give profound new insights into mechanisms of toxicology, aging, disease etiology etc. and provide new surrogate biomarkers of such processes. Systems biology aims to combine the molecular components (transcripts, proteins, and metabolites) of an organism and incorporate them into functional networks or models designed to describe its molecular connectivities and dynamic activities. While many of the functions of individual parts are unknown or not well defined, their biological role can sometimes be inferred through association with other known parts, providing a better understanding of the biological system as a whole. On a system-wide scale the description requires three levels of information [13,143]: (A) identification of the components (structural annotation) and characterization of their identity (functional annotation); (B) identification of molecules that interact with each component; and (C) characterization of the behaviors of the transcripts, proteins, and metabolites under various conditions [144].

Bioinformatic approaches in systems biology include biomarker discovery and reconstruction of regulatory signals in biological networks by, for example, mapping molecular profiles onto public databases of functional and pathway information, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [145], Gene Ontology [146], Interactome [147], etc. Metabolic pathway annotation is now part of the post-genomic knowledge compilation effort [145]. Such databases enable the rapid development of knowledge-based methods for uncovering higher order systemic operation of the cell and the organism from genomic and molecular information.

There are several statistical issues that may directly impact on -omic data merging and the interpretability of the results, one of the main factors being the differences in dimensionality between each -omic platform, leading to unbalanced integration i.e. metabolomics vs. transcriptomics (100s metabolites vs. 10,000s genes). Sparsity refers to the percent coverage of the metabolome using analytical instrumentation. Typically 10–100s of metabolites can be detected with NMR spectroscopy, whereas 100–1000s of metabolites can be detected using MS. Furthermore, the noise characteristics and level associated with the analytical instruments and biological sample will directly impact the choice of statistical method and data pre-processing procedures. When different biological experiments have been conducted for each -omic under identical conditions the replicates cannot be matched between -omic data blocks. In such ‘unmatched replication’ experimental designs only group-wise statistics can be compared. For example, one could compare the means and standard deviations of gene expression and metabolite levels (but not individual replicate values) when different biological replicates have been used for transcriptomics and metabolomics. Finally merging data from different biological experiments for different -omics may be difficult due to the timing of sample collections e.g. urine and faeces can be measured much more frequently than tissue samples which are usually collected once after sacrifice. In such ‘asynchronous’ experimental designs, replicates and even group-wise statistics cannot easily be compared meaningfully across -omic data blocks.

Tavazoie et al. [148] put forward a method now known as over representation analysis and discovered distinct expression pattern clusters in mRNA that were significantly enriched for genes with similar function [148]. Mootha et al. [149], introduced the concept of Gene-Set Enrichment Analysis [149,150], in which a set of genes involved in a signature are compared to existing knowledge about gene pathways [151]. This allows a transition between simple biological entities, such as genes, to a higher order description, such as gene pathways, and lead to a synthetic view of the relevant biological pathways and functions affected. Gene-Set Enrichment Analysis is a generic method that can be easily extended to proteomic data and metabolic profile data, or Metabolite-Set Enrichment Analysis. To our knowledge, no real attempt has been made to integrate metabolomic markers to the known interactome — the rapidly growing knowledge of protein–protein interaction (PPI) networks for human, model organisms and host–pathogen which may begin to provide network-based models for diseases [147,152,153]. According to Sanchez et al. [153], the total number of genes of an organism is less important than the complete repertoire of interactions potentially encoded by its genome (i.e., the interactome) [153]. Several interactions (protein–DNA, protein–RNA or protein–protein) were initially included, but due to the presence of double-hybrid and co-immunoprecipitation assays, most of the work has focused on protein–protein interactions. Topological properties of the entire metabolic network exhibit robustness against isolated random genetic mutations and/or metabolic perturbations. Robustness of the metabolic reaction network also depends on its high-level of plasticity and regulation, that is at least partly mediated by gene variants, transcriptional variation, protein abundance and modulation activity through direct molecular interactions and covalent binding. Full sequencing of various organisms now allows genome-scale metabolic reconstruction approaches. The metabolic network of organisms or tissues can then be inferred from the genome sequence. Such strategies involving network-based predictions of metabolism have been successfully applied to yeast [154] and a range of other microorganisms, as well as human metabolism are now complementary to metabolic profiling [155,156]. A comprehensive review on integrated -omics approaches for modeling of cellular networks is given by Joyce et al. [157].

Metabolic phenotypes, or metabolotypes [76] can be analyzed in a statistical framework linking either the genomic, transcriptomic or proteomic variation information with the metabolomic variation information using appropriate statistical tools, such as correlation networks [139,158–160] PLS [161], O2PLS [162–164] and PARAFAC [165]. These statistical tools and concepts have been extended to transgenic analyses linking the microbiome and metabolic phenotype (metabolotype), for a comprehensive review of such approaches see reference [176]. The concept of ‘genetical genomics’ [166] was introduced by Jansen and Nap [166], aiming at merging genome-wide expression profiling and genome-wide genotyping in segregating populations, using the statistical genetics used in the analysis of quantitative trait loci (QTL). A subsequent paradigm shift was introduced, know as expression QTL (eQTL) approaches [167]. The genetical genomics paradigm can be extended to a metabolomic context. When comparing the metabolome of different conditions (genetics, toxicology pathology), it is impossible to assign a gene or a genomic region to a metabolic feature. The use of Quantitative Trait Locus mapping of metabolomic traits, or ‘mQTL,’ provides a very good conceptual, analytical and mathematical framework to identify genomic regions influencing metabolic profiles. This approach was introduced in plants [168,169] and in rodent genetic intercrosses targeting type 2 diabetes [170]. Altogether these studies apply the genetical genomics concept to metabolomic traits, or genetical metabolomics (Fig. 4). The resulting linkage maps (Fig. 4) allow the identification of associations between genomic sequence variants and plasma metabolites. The horizontal axis summarizes metabolome-wide 1H NMR spectrum variation. The vertical axis shows the genomic position of >2000 microsatellite and SNP markers.
Significant associations are with a LOD score >3 (P<10^{-3}) are reported (from Dumas et al.[170]). The mQTL approach integrates variations obtained from metabolome-wide profiling and genome-wide genotyping to identify genetic variants or loci influencing metabolite levels: in tissues and biofluids, if a genetic variant affects the activity of a specific enzyme, a variation in the levels of metabolic substrates or products is expected. The concept of metabolome-wide association study (MWAS) [171] was introduced to highlight the potential of metabolomic markers in molecular epidemiology [171], just like genome-wide association studies (GWAS) with the INTERMAP consortium[24]. The mQTL, GWAS and MWAS approaches have now merged in human epidemiology with several successful clinical applications[172–174]. In terms of validation of loci, the idea that a genome sequence variant (QTL) could directly influence simultaneously the expression of its gene (eQTL), its translation into a protein (protQTL), its effect on metabolism (mQTL) and phenotypic outcome (pQTL) provides a powerful self-validation tool [175]. However, occurrence of co-localization is very rare (<10 loci in Arabidopsis thaliana genome), suggesting the limits of cis-QTL effects. Ultimately, genetical genomic and mQTL approaches require an independent validation for any locus of interest identified, whether it is in an independent experimental cross for animal models [170], or in an independent cohort for clinical applications [173]. A summary of published inter-omic type integration methods is given in Table 1.

3. Conclusions and future requirements

Understanding the relationships between human genetic factors, the risk of developing major diseases, the molecular basis of drug efficacy and toxicity as well the parameters defining health will lead to practical innovations in medicine, drug discovery and engineering. Such an understanding requires intra- and inter-omic integration of experimental data between different assay types, sample types and across multiple levels of biological organization. It is clear that this field, though comparatively young, is developing fast and will yield vital tools and techniques to increase our understanding of biological processes at the systems level.

Acknowledgements

The authors would like to acknowledge Olaf Beckonert for the acquisition of the 800 MHz 1H NMR spectrum of human urine. JMF holds a Royal Society of Chemistry PhD studentship award. MED holds a Young Investigator Award from the Agence Nationale de la Recherche (ANR-07-JCJC-0042-01), SER and MED are funded by Nestlé (RDLS015375).

References


Appendix F

Paper V
We characterize the integrated response of a rat host to the liver fluke *Fasciola hepatica* using a combination of $^1$H nuclear magnetic resonance spectroscopic profiles (liver, kidney, intestine, brain, spleen, plasma, urine, feces) and multiplex cytokine markers of systemic inflammation. Multivariate mathematical models were built to describe the main features of the infection at the systems level. In addition to the expected modulation of hepatic choline and energy metabolism, we found significant perturbations of the nucleotide balance in the brain, together with increased plasma IL-13, suggesting a shift toward modulation of immune reactions to minimize inflammatory damage, which may favor the co-existence of the parasite in the host. Subsequent analysis of brain extracts from other trematode infection models (i.e. *Schistosoma mansoni*, and *Echinostoma caproni*) did not elicit a change in neural nucleotide levels, indicating that the neural effects of *F. hepatica* infection are specific. We propose that the topographically extended response to invasion of the host as characterized by the modulated global metabolic phenotype is stratified across several bio-organizational levels and reflects the direct manipulation of host–nucleotide balance.

**Introduction**

*Fascioliasis* is a re-emerging zoonotic disease caused by two liver fluke species, *Fasciola hepatica* and *F. gigantica*. As many as 17 million people might be infected with *Fasciola* spp. (Keiser and Utzinger, 2009) and the disease is of considerable economic and public health importance. At present, triclabendazole is the only drug available for treatment of infected human beings and livestock and worryingly, parasite resistance to triclabendazole is already widespread in farm animals (Fairweather, 2009). Hence, a deeper understanding of the biology of this parasite is warranted to uncover novel therapeutic targets. Fundamental investigations in well-controlled host–parasite models hold promise for recovery of metabolic biomarkers, at different bio-organizational levels. This may promote the mechanistic understanding of the host response to infection at the systems level with a view to identifying drug targets.

The regulation of any mammalian system is such that events originating in a discrete tissue or organ can trigger a cascade of consequent events throughout the host system as the host attempts to maintain homeostatic equilibrium (Lederberg,
2000; Nicholson et al, 2002). Thus, the effects of a parasitic infection are rarely confined to a single target tissue; instead a network of molecular events can generally be detected throughout the host system. Top-down systems biology driven by metabolic phenotyping and metabonomics (Nicholson et al, 1999; Holmes et al, 2008a; Li et al, 2008a, b; Tsang et al, 2009) has been shown to be a useful tool for studying host–parasite interactions (Nicholson et al, 2005; Martin et al, 2007; Holmes et al, 2008b), and it is clear that any transgenicomic interaction can only be studied in vivo, as the effects of microbial or parasitic modulations are panorganismal (Wang et al, 2008; Saric et al, 2009).

In this study, we characterize and interpret the system-wide effect of F. hepatica in the rat, using a metabolic profiling strategy to develop mechanistic hypotheses. A multilevel statistical approach was applied to the analysis of 1H nuclear magnetic resonance (NMR) spectroscopy-generated data to obtain an integrative profile of the global response of the host based on differential metabolic, immunological, and biometric responses across multiple tissues. We report novel findings regarding the effect of F. hepatica on the neural metabolite profile of the rat and show metabolic connections between the liver, gut, and brain, which we compare with the metabolic effects elicited by two separately assessed trematode infections, namely Schistosoma mansoni and Echinostoma caproni in a mouse model. Although F. hepatica-induced liver damage has been described in detail before (Lim et al, 2007; Marcos et al, 2008), no direct association with the central nervous system has been found in the extant literature. The findings from this study expose a potential mechanism of parasite-induced immune modulation and exemplify the value of top-down systems approaches based on metabolic profiling for recovering mechanistic information from a system without a priori knowledge.

Results and discussion
Parasite burden, physiological monitoring, and histology
Patent infection of rats with F. hepatica resulted in a mean parasite burden of 5.5 (range=2–10; s.d.=2.6) on day 77 postinfection and a significant reduction of the mean packed cell volume (PCV) at day 71 postinfection (uninfected control rats: mean=50.8%, s.d.=2.8%; infected: mean=42.0%, s.d.=3.8%). The mean bodyweight did not significantly differ between infected and uninfected rats at any of the investigated time points. Clear evidence of hepatic necrosis was found in F. hepatica-infected rats (Supplementary Figure S1) together with follicular hyperplasia of splenic white pulp and interstitial lymphoplasmocytic inflammatory infiltrates in the kidney.

Multivariate statistical modeling of F. hepatica infection
Assignment of metabolic components in the tissues and biofluid profiles for both uninfected and infected rats were made on the basis of extant literature (Coen et al, 2003; Holmes et al, 2006; Beckonert et al, 2007; Li et al, 2009; Saric et al, 2009; Tsang et al, 2009), and in the case of the brain extract spectra, from addition of authentic standards (assignments provided in Supplementary Figure S2; Supplementary Table S1).

The metabolic changes in the rat caused by patent infection with F. hepatica were modeled separately for each biological matrix through principal component analysis (PCA) and projection to latent structure discriminant analysis (PLS-DA). In brief, the metabolic profiles of F. hepatica-infected rats could be differentiated from those of uninfected control rats through PCA for all assessed biological compartments except the ileum and the renal medulla. A list of important infection-discriminatory metabolites is given in Table I together with the P-values, validated by permutation testing, Figure 1.

Consistent changes in lipid metabolism, including elevated choline and/or choline derivatives such as betaine, phosphocholine, and glycerophosphocholine (GPC) occurred throughout most tissues (Figure 1; Table I). This indicates a generally increased usage of choline-derived metabolites that can be further converted to phosphocholine (Zeisel et al, 2003; Li and Vance, 2008), a basic component of membrane anabolism, or used for the production of polysaturated fatty acids such as arachidonic acid, which gives rise to pro-inflammatory eicosanoid mediators (Calder, 2008, 2009). Prostaglandins, for instance, exert multiple pro-inflammatory effects, such as chemotaxis of neutrophils, which are typically the first infiltrating cells at sites of tissue damage. Evidence of inflammatory infiltrates was detected in the histological and metabolic profiles of the liver, spleen, and kidney (Supplementary Figure S1) and inflammation was implicated biochemically in the colon and plasma. Increased plasma intensities of O- and N-acetylated glycoprotein signals were noted in F. hepatica-infected rats. Increased concentrations of acetylated glycoproteins have been earlier identified in Trichinella spiralis- and Trypanosoma brucei-brucei-infected mice (Martin et al, 2006; Wang et al, 2008) as inflammatory markers and include acute phase reaction proteins, such as α1 acid glycoprotein, haptoglobin, and transferrin, which are produced in the liver and which show markedly increased serum levels during such systemic responses to inflammation (Schreiber et al, 1982; Bell et al, 1987a,b) (Figure 1; Table I).

Other features of F. hepatica infection were specific to a single biological compartment; for example hippurate, which is only observed in the urine, was found to be decreased in infected rats, reflecting a parasite-induced disturbance of the gut-microbial composition or activity (Nicholls et al, 2003; Li et al, 2008a, b). The liver, spleen and plasma showed the greatest changes at the low molecular weight level, consistent with the direct damage caused by the migrating larvae of F. hepatica during the acute stage of an infection in the hepatic tissue and the immunological activity in the spleen (Supplementary Figure S1). The proliferation of B cells in the spleen primary follicles as response to blood-borne antigens uses amino acids and lipid fractions for the new cellular membrane bilayers and organelles and may account for the major changes in the metabolic profile of the spleen, and perhaps in other tissue compartments. The depletion in hepatic glucose is more likely to be related to the direct damage of the energy stores and is consistent with elevated glucose levels in plasma.
One of the strongest effects of infection with *F. hepatica* manifested in the neurochemical profiles, further highlighting the ability of the parasite to impact on remote tissue. Clear visual separation of the infected and uninfected animals was achieved in the PCA and PLS-DA scores plots (Figure 2A and B) of the neural profiles, with a model predictivity ($Q^2_Y$) of 0.46. The pattern of differentiating metabolites in the neurochemical profiles as extracted from the orthogonal PLS-DA (O-PLS-DA) coefficients was composed of a significant increase in the relative levels of inosine, tyrosine, and phenylalanine. Conversely, the relative tissue concentrations of GPC, succinate, inosine mono-, di-, and triphosphate, adenosine, and adenosine mono-, di-, and triphosphate were lower in the brains of infected animals (Figure 2C and D).

Secretion of nucleotide-degrading enzymes occurs in a variety of ecto- and endoparasitic organisms, such as ticks, blood sucking insects, and helminths and serves the primary purpose of minimizing immune reactions in the host organism to prolong their survival (Parshad and Guraya, 1977; Ribeiro and Francischetti, 2003). For example, adenosine deaminase and 5'-nucleotidase have been found in secretory channels of *T. spiralis* and *F. gigantica* (Gounaris, 2002; Ali, 2008). The former catalyzes the intracellular conversion of adenosine to inosine, whereas the latter hydrolyzes inosine monophosphate (Figure 2E).

We observed a substantial shift of adenosine and phosphorylated nucleotides to inosine in the host brain, which indicates an attenuated inflammatory response and may be suggestive that *F. hepatica*, similarly to *F. gigantica*, secretes nucleotide-degrading enzymes. Both inosine and adenosine induce mast cell degranulation, an event that has a central function in localized inflammatory response, an important mediator of type I hypersensitivity reactions, with adenosine being the more potent activator (Marquardt et al, 1978; Tilley et al, 2000). Furthermore, inosine, which was believed to be inert for a long time, has been proven to effectively suppress pro-inflammatory cytokines such as IFN-$\gamma$, TNF-$\alpha$, and IL-12 (Hasko et al, 2000, 2004; Mabley et al, 2003) in vitro and in vivo.

We assessed the concentrations of selected plasma cytokines and found the levels of both IL-13 and IL-5 to be significantly higher in *F. hepatica*-infected rats compared with the control group on day 22 postinfection. IL-13 maintained a significant increase on day 43 postinfection (Figure 2E). An enhanced Th2 immune response such as that suggested by the increase in both IL-5 and IL-13 would serve to counteract mechanisms of inflammation. IL-13 is also an important direct negative regulator of inflammatory cytokines in macrophages (Minty et al, 1993). Minimization of such an intense immune reaction, and hence prolongation of the period in which the parasite can remain undetected by the host at earlier stages of infection, is clearly beneficial for the survival of the worms.

**Comparison of the systemic effects of *F. hepatica* with *S. mansoni* and *E. caproni***

To ascertain the specificity of the global metabolic response of the rat to *F. hepatica* across different tissue compartments, particularly with respect to the observed modulation of the neural nucleotide balance, two further rodent-trematode models were compared, namely *S. mansoni* and *E. caproni*, in a murine host. Both trematodes induced a biochemical response across a range of biological compartments, but whereas *S. mansoni* caused pathology in the spleen and liver (Li et al, 2009), no overt necrosis was found in any tissue after *E. caproni* infection (Saric et al, 2009). At a global level, *E. caproni* infection was predominantly characterized by...
malabsorption of amino acids, which manifested predominantly in the biochemical profiles of the small intestine and liver. Altered amino-acid levels and energy metabolism were also a core signature of S. mansoni infection with disturbed levels of choline-containing membrane components in the liver and intestinal tissues. As the altered neural nucleotide signature was one of the strongest metabolic responses in rats infected with F. hepatica, we directly compared the metabolic profiles of brain extracts with those from S. mansoni and E. caproni in the mouse. S. mansoni infection induced obvious changes in the metabolic composition of the host brain, as indicated by the PCA and PLS-DA scores plots (Supplementary Figure S3), with the PLS-DA model indicating a predictive value ($Q^2$Y) of 0.61. Infection with S. mansoni was characterized by higher levels of glutamine, but lower concentrations of γ-aminobutyric acid (GABA), choline, phosphocholine, and scyllo-inositol (Supplementary Figure S3C). The E. caproni-infected animals could not be statistically differentiated from the control group based on the brain extract spectra ($Q^2$Y=−0.21, data not shown), indicating that the infection did not systematically alter the brain biochemistry of the host.

Although both S. mansoni and F. hepatica induced a marked perturbation of the neurochemical profiles in their host animals, the patterns of differentiating metabolites were distinct for these two trematode-rodent models, with the exception of the common effect of decreasing concentrations of lipid degradation products, such as choline, phosphocholine, and GPC, all of which have a function in cell membrane formation. In contrast to the perturbed immunological function suggested by the altered nucleotide levels in the brains of F. hepatica-infected animals, modulation of the neural metabolite profile by S. mansoni infection suggested the dominant effect related to the neurotransmitters, as supported by decreased levels of GABA and increased tissue concentrations of glutamine, which is a substrate for both excitatory and inhibitory neurotransmitters, including GABA. However, as neural GABA decreased in infected animals, the increased concentrations of glutamine are more likely to be indicative of a detoxification process. Compensatory locations for ammonium detoxification are muscle cells and the brain, combining ammonia and glutamate to glutamine (Kreis et al., 1991; Olde Damink et al., 2002), which could also contribute to the relatively higher levels of glutamine in the brain extracts of S. mansoni-infected mice documented in the current investigation.

![Figure 1](https://example.com/figure1)

**Figure 1** Summary of the systemic involvement of choline-containing species and their potential relationship with inflammatory processes. Increased immune activity in the liver- and gut-associated lymphoid tissue (GALT) in the colon on F. hepatica infection may lead to cell-intestinal lipid degradation to arachidonic acid, which is the substrate of many eicosanoid pro-inflammatory mediators. The spleen and the kidney, in which inflammatory activity is less extensive, respond to the infection with increased betaine, which shares the same pathway. AA, amino acids; CDP-Cho, CDP-choline; Cho, choline; GPC, glycerophosphocholine; HE, hydroxyeicosatetraenoic acid; Hp, hydroperoxy-eicosatetraenoic acid; LA, lipoxin; LT, leukotrienes; PC, phosphatidylcholine; PG, prostaglandine; PhCho, phosphocholine; PL, phospholipase; SM, sphingomyelin; TX, thromboxane; red, increased relative metabolic levels; *pro-inflammatory effects.
Figure 2  Multivariate analysis of brain spectra from the *F. hepatica*-rat infection model. (A) PCA scores plot and (B) PLS-DA scores plot showing differences between *F. hepatica*-infected (red; c1–c12) and uninfected animals (blue; t1–t12). *P*-value back projection to O-PLS-DA analysis of *F. hepatica*-infected (upward oriented peaks) and uninfected animals (downward oriented peaks) for the aliphatic region (C) and aromatic region (D), whereby the metabolic regions differentiating between infected and uninfected animals (*P* < 0.05) are color coded according to the power of discrimination. (E) Schematic of intra- and extracellular adenosine–inosine interconversion in the brain by two main enzymes, 1adenosine dehydrogenase and 25'-nucleotidase. The shift from adenosine toward inosine observed in the coefficient plots in (C) and (D) may lead to emphasized Th2-mediated immune mechanisms reflected by increased levels of IL-5 and IL-13. Plasma cytokine assessment on days 1, 8, 22, 43, 57 and 71 (d) post-infection confirmed significantly higher levels of IL-5 and IL-13 in the *F. hepatica*-infected animals (cyan and green, respectively) at day 22 postinfection in the case of IL-5, and days 22 and 43 postinfection for IL-13, compared with the uninfected animals (blue). A, adenosine; I, inosine.
Systems impact of the modulation of the neural metabolic phenotype of *F. hepatica* infection: multivariate modeling and integration of physiological compartments

Metabolic correlates of the six important discriminatory spectral regions containing inosine, GPC, succinate, tyrosine, phenylalanine, and the overlapped regions containing adenosine and the phosphorylated nucleotides in the brain tissue of *F. hepatica*-infected rats were identified in each of the other tissue or biofluid matrices. Sequential correlations between the integrated-discriminatory signals in the brain spectra tissue were made with the whole spectra from each tissue and with the spectra from biofluids obtained at day 71 postinfection (Figure 3; Supplementary Table SII). Here, univariate correlation of single signals was preferred over a multivariate approach to maximize information recovery and to infer direct inter-compartmental links between the metabolites. A comprehensive list of compartment-specific correlations with the neural metabolites is provided in Supplementary Table SII. The liver, urine, and spleen showed the highest number of components correlated with the selected brain metabolites. The high degree of correlation between the metabolic markers in the brain and the liver spectra across the tissue compartments (Figure 3A) may simply relate to the magnitude of spectral changes in both organs and does not necessarily imply any causal relationships between metabolites. The gross changes in the hepatic metabolite profile are reflective of direct mechanical damage of the liver tissue by the migration and feeding of the juvenile fluke, whereas the effect of the parasite on remote organs such as the brain is more difficult to rationalize. Hepatic dysfunction in *F. hepatica* infection is known to result in increased circulating toxins, such as ammonia, thiols, and phenols (Zaki et al., 1983), which are typically released after hepatic failure, because of reduced ability to degrade aromatic amino acids and ammonia. Animals infected with *F. hepatica* showed increased neural concentrations of phenylalanine and tyrosine. Indeed, it has been shown that hepatic failure induces an increased permeability of the blood–brain barrier for several substances, among which phenylalanine and tyrosine were found to increase up to 30% in the host brain (Zaki et al., 1984).

A recurrent theme across multiple levels of this host–parasite system was the inflammatory response, which manifested at both the level of structural damage and that of the metabolite signature. The increased signal intensities of plasma acetyl glycoprotein fragments were statistically associated with all six cerebral markers of infection underscoring the presence and global effect of a strong infection-induced inflammatory response (Figure 3B; Supplementary Table SII). Interestingly, a shift has been observed in the brain nucleotides inosine and adenosine in favor of inosine, which has been shown to have anti-inflammatory effects at the posttranscriptional level (Hasko et al., 2000), including suppression of pro-inflammatory cytokines (e.g. IFN-γ and TNF-α) and minimizing macrophage-mediated mechanisms of inflammation (Hasko et al., 2000, 2004; Lüdert et al., 2002). This shift may be directly induced by nucleotide-degrading secretory enzymes of *F. hepatica*. Significant direct correlations between the selected neural indices of *F. hepatica* infection and compounds involved in lipid metabolism were detected.

In addition to the correlations driven by the liver pathology, direct metabolic interactions such as plasma glucose being anti-correlated with brain succinate and 2-ketoglutarate, suggestive of an interaction through glycolysis and the tricarboxylic acid cycle, were found. Another example of such co-variation is the inverse correlation between cerebral tyrosine and urinary hippurate, deriving from gut-microbial/mammalian co-metabolism. Direct alteration of urinary metabolite signatures through gut-microbial species seems to be a common feature of many host–parasite models (Martin et al., 2006; Li et al., 2008a; Saric et al., 2008; Wang et al., 2008). Here, the negative correlations of cerebral adenosine with urinary dimethylglycine and the bile acids, which are co-metabolized by gut microbiota, strengthens the notion of a bidirectional communication between gut and brain (gut–brain axis). This leads to the hypothesis that nucleotide-initiated immunoactivity has a function in the changed gut-microbial dynamics observed in parasitic infections.

In conclusion, we have characterized the global metabolic phenotype of a host–parasite system and have shown a clear effect of a trematode infection on the biochemical composition of the host brain using a metabolic profiling strategy to develop and pursue hypotheses relating to the observed neurochemical changes in the rat. *F. hepatica* induced a focussed response, primarily associated with the hypothesized worm-induced shift from adenosine toward inosine and the subsequent induction of anti-inflammatory cytokines. The coherence of the histological, metabolic, and cytokine data further facilitated elucidation of general and specific metabolic events and provided a means of probing inter-compartmental co-variation of metabolites. Thus, the application of a top-down systems approach has been shown to be of value in driving the articulation of novel mechanistic hypotheses relating to parasitic invasion and has wide application in molecular parasitology.

Materials and methods

**F. hepatica**-rat model and experimental design

Experimental procedures were carried out at the Swiss Tropical and Public Health Institute (Swiss TPH; Basel, Switzerland), adhering to local and national guidelines of animal welfare (permission no. 2070 and 2081). A total of 24 Wistar female rats were purchased from RCC (Itingen, Switzerland) and kept under environmentally controlled conditions (temperature: 25°C; relative humidity: 60–70%; light/dark cycle: 12/12 h). Rats were acclimatized for 1 week and all animals had free access to commercially available rodent diet obtained from Nafag (Gossau, Switzerland) and community tap water.

Rats were individually marked and group housed with four animals per cage. Twelve rats were orally infected with 20–25 *F. hepatica* metacercariae each (Cullompton isolate) obtained from Mr G Graham (Addlestone, UK). The remaining 12 rats were left uninfected and served as controls.

The body weight of each animal was measured throughout the experiment. On day 1 pre-infection and days 1, 4, 8, 15, 22, 28, 36, 43, 57, and 71 postinfection, urine and feces were collected into Petri dishes by rubbing the abdomen of the rats gently. Approximately 50 µL of blood was collected from the tail tips collected into hematocrit capillaries (Sodium [Na] heparin coated) and spun at 10,500 g for 4 min. The PCV was calculated and expressed as a percentage of the plasma ratio to red blood cells (Li et al., 2008a).
All animals were euthanized on day 77 postinfection using CO2, and the worm burden was determined in each infected animal by removing adult flukes from the livers and bile ducts on dissection. The whole brains of rats, regardless of their infection status, were removed on dissection and the left hemispheres were transferred into cryo-tubes, snap frozen in liquid nitrogen and stored at −80°C pending 1H NMR spectroscopic data acquisition. In addition, the left lateral lobe of the liver, the left kidney, spleen, and three parts of the intestine (colon, ileum, and jejunum) were removed and stored in the same manner as the brain for 1H NMR spectroscopic data acquisition. The right brain hemispheres, the right kidney, and parts of the spleen, liver, and intestines were transferred into separate Eppendorf tubes containing 4% buffered formalin for subsequent histological examination.

Figure 3  Correlation plots showing significant associations between the 1H NMR spectra of (A) liver and (B) plasma each as the X-matrix with the integrals of the control/infection-differentiating metabolites of the brain extracts (AIP, adenosine and phosphorylated nucleotides; BCAA, branched chain amino acids; GPC, glycerophosphocholine; Ino, inosine; Phe, phenyalanine; Suc, succinate; Tyr, tyrosine) as the Y-matrix are color coded according to the correlation coefficient.
Methodology relating to the S. mansoni- and E. caproni-mouse models has been described earlier (Saric et al, 2008, 2009; Li et al, 2009) and followed a similar protocol to that described for F. hepatica infection in the rat.

Sample preparation for 1H NMR spectroscopy
Each brain sample was placed in a mortar, mashed using a 1 ml mixture of H2O and acetonitrile (1:1, v/v) and transferred into a glass tube. Another 2 ml of the solvent mixture was used to rinse the mortar twice and transferred into the same glass tube. The brain homogenate was centrifuged for 6 min at 10 000 g. The supernatant was collected into a new glass vial, evaporated overnight and lyophilized. The resultant dry mass was resuspended into 0.55 ml of D2O until completely dissolved and transferred into a 5 mm-inner-diameter NMR tube for subsequent analyses. Urine, plasma, and fecal pellets were prepared for conventional high-resolution 1H NMR data acquisition as described earlier (Saric et al, 2008), whereas tissue metabolic fingerprints were acquired through magic angle spinning (MAS) 1H NMR (Saric et al, 2009).

The extraction protocol for the S. mansoni-mouse model was slightly extended to gain additional information regarding potential changes in the lipid metabolic profile, as a pharmacological intervention was involved in the original study protocol. Each brain sample obtained from mice infected with S. mansoni and the corresponding uninfected control mice was transferred into a 2 ml Eppendorf tube containing a metal bead, 0.75 ml of water, and 0.75 ml of methanol. The Eppendorf tube was placed in the tissue lyser and shaken for 5 min at the speed of 22 Hz. The resulting brain homogenate was transferred into a glass tube. A further 0.75 ml of each liquid was used to rinse the Eppendorf tube and transferred into the same glass tube. A total of 1.5 ml of chloroform was added into the mixture and centrifuged at 2500 g for 30 min. The aqueous and the chloroform phases were transferred into a new glass tube each and both were left to evaporate overnight and lyophilized. Before 1H NMR analysis, the powder obtained from the aqueous phase was dissolved in 0.55 ml phosphate buffer (D2O:HO;O = 9:1, v/v, 0.01 % of sodium 3-(trimethylsilyl) propionate-2,2,3,3-d4 ([TSP)], pH = 7.4), whereas the dry mass of the chloroform fraction was dissolved in deuterated chloroform (CDCl3).

Acquisition of 1H NMR spectral data
All 1H NMR spectra from rat and mouse brain extracts, urine, plasma, and fecal water were recorded on a Bruker Avance 600 NMR spectrometer (Bruker; Rhein restetten, Germany), operating at 600.13 MHz for proton frequency. A 5-mm triple resonance probe spectrometer (Bruker; Rheinstetten, Germany), operating at 2.10013 MHz for proton frequency. A 5-mm triple resonance probe was used for the brain extracts of both the S. mansoni- and E. caproni-mouse model. A total of 128 scans were acquired for each sample into -90 -90 -90 -90 acquisition time). Optimal water suppression was achieved by irradiating the water frequency during the O-PLS-DA plot, whereby regions that significantly differentiate the O-PLS-DA model was built by using an X-matrix containing all spectral information and a binary dummy matrix, as Y-determining class affiliation, for example infected or controls. Sevenfold cross-validation was applied (Bro et al, 2008) to validate the models and calculate the goodness of prediction Q2. To validate the statistical power of the discriminatory metabolites, the P-values were calculated for each data point using 10 000 permutations and back projected to the O-PLS-DA plot, whereby regions that significantly differentiate infected from non-infected animals at a level of P<0.05 are coded in red (Pitman, 1938).

For assessing the direct impact of the cerebral metabolic changes caused by F. hepatica infection, a further development of the STOCSY (Cloarec et al, 2005) and the statistical heterospectroscopy (Crockford et al, 2006) method has been applied between two data matrices whereby the selected regions occupied by each of the eight cerebral biomarkers were integrated in the s.d. calculated from the first 500 data points, which consist of noise only. A 10 000-fold permutation was embedded in the correlation script and only correlations with P<0.05 are displayed.

Cytokine multiplex quantification
Plasma from five F. hepatica-infected rats and five uninfected rats from earlier NMR spectral preparation (e.g. 0.9% saline (NaCl) in
D₂O:H₂O=1:1), over six different time points, were tested by a 'Mesoscale Multiplex Assay' (MS6000 Rat Demonstration 7-Plex Ultra-Sensitive kit, Meso Scale Discovery). The rat demonstration ‘7-Plex Ultra-Sensitivity kit’ was used for this purpose, which includes IFN-γ, IL-1β, IL-4, IL-5, IL-13, KC/GRO, and TNF-α. Each sample was assessed using duplicates and split in 15 µl per well in a 96-well plate. The method was used according to the manufacturer’s specifications (Meso Scale Discovery, Rat cytokine assays: Rat demonstration 7-Plex Ultra-Sensitive kit), with the exception of the incubation time, which was extended to 4 h to counteract the 1:1 dilution. An MSD Sector Imager was used to read the plates (Meso Scale Discovery, Sector Imager 6000).

Supplementary information

Supplementary information is available at the Molecular Systems Biology website (http://www.nature.com/msb).

Acknowledgements

We thank Dr K Veselkov, S Robinette, and Dr I Yap for providing in-house software and G Calvert from Meso Scale Discovery for providing expertise on the cytokine assay. This study received financial support from the Swiss National Science Foundation to JS, JVL, JU (project no. PPPOB-102883 and PPPOOB-119129), and JK (project no. PPPOOA-114941), and from Chinese Academy of Sciences to YW (project no. PPOOB-102883 and PPOOB-119129), and from the Sir Henry Wellcome Fellowship (Wellcome Trust award number P23665).

Conflict of interest

The authors declare that they have no conflict of interest.

References

© 2010 EMBO and Macmillan Publishers Limited

Systemic metabolic impact of F. hepatica in the rat
J Saric et al

Molecular Systems Biology 2010 9


Petman EG (1938) Significance tests which may be applied to samples from any populations III. The analysis of variance test. *Biometrika* **29**: 322–335


**Molecular Systems Biology** is an open-access journal published by European Molecular Biology Organization and Nature Publishing Group. This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License.