Assessment of the Complementarity of Data from Multiple Analytical Techniques

James Stuart McKenzie

Submitted for the Degree of Doctor of Philosophy

University of York

Department of Chemistry

April 2013

Abstract

Whilst the capabilities of analytical techniques are ever-increasing, individual methods can provide only a limited quantity of information about the composition of a complex mixture. Interrogation of samples by multiple techniques may permit for complementary information to be acquired, and suitable data fusion strategies are required in order to optimally exploit such complementary information. A novel mid-level data fusion strategy has been implemented which uses two-stage genetic programming for feature selection and canonical correlation analysis such that highly discriminatory variables can be related together in a multivariate fashion. The approach offers an intuitive way to visualise variable interaction and their contributions to experimental trends.

Contents

List of Figures 13

List of Tables 25

Acknowledgements 29

Author's declaration 31

1 Introduction **33**

1.1 Scope of the thesis 38

2 Nuclear Magnetic Resonance **41**

2.1 Introduction 42

2.2 Approaches to experimental design 47

2.2.1 Sampling and collection procedures 48

2.3 Sample preparation 50

2.3.1 Matrix considerations 50

2.3.2 Target molecule considerations 50

2.3.3 Samples 52

2.3.3.1 Chemical reactions 52

2.3.3.2 Urine 52

2.3.3.3 Faecal matter 53

2.3.3.4 Plasma and serum 53

2.3.3.5 Amniotic fluid 54

2.3.3.6 Cultured cells 54

2.3.3.7 Animal tissues 54

2.3.3.8 Plant tissues 55

2.3.3.9 Honey 55

2.3.3.10 Oils 56

2.3.3.11 Environmental samples 56

2.4 Hardware and experimental setup 56

2.4.1 Availability of hardware 57

2.4.2 Speed or sensitivity? 58

2.4.3 NMR experiments 58

2.4.4 Homonuclear NMR 59

2.4.4.1 ^{1}H 59

2.4.4.2 ^{13}C 64

2.4.4.3 Diffusion ordered spectroscopy 64

2.4.5 Heteronuclear multi-dimensional techniques 66

2.4.5.1 HSQC 66

2.4.5.2 HMQC 67

2.4.5.3 HMBC 67

2.4.6 Solid state and magic angle spinning 67

2.4.7 Hyphenation 68

2.4.8 Technology selection 69

2.5 Data pre-processing 69

2.5.1 The free induction decay 70

2.5.2 Noise removal 70

2.5.3 Baseline correction 72

2.5.4 Peak alignment 74

2.5.5 Feature extraction 74

2.6 Data analysis 80

2.6.1 Univariate analysis 80

2.6.2 Multivariate analysis 81

2.6.2.1 Principal component analysis 82

2.6.2.2 Discriminant analysis 83

2.6.2.3 Partial least squares 84

2.6.2.4 Evolutionary computing 85

2.6.2.5 Kernel methods 88

2.6.2.6 Artificial neural networks 88
2.6.2.7 Multi-way decompositions 89
2.6.2.8 Decision trees 90
2.6.3 Assignment 92
2.7 Future directions 93

3 Mass Spectrometry and Liquid Chromatography 97
3.1 Introduction 98
3.2 Sample preparation 100
3.2.1 Standardisation 102
3.2.2 Normalisation 103
3.2.2.1 Internal standards 103
3.2.2.2 Quality control 103
3.2.3 Run order 105
3.2.4 Acquisition modes 106
3.2.5 Matrices 106
3.2.5.1 Urine 107
3.2.5.2 Blood 107
3.2.5.3 Animal tissue 109
3.2.5.4 Plant tissue 110
3.2.5.5 Food 111
3.2.5.6 Environmental samples 113
3.3 Hardware: the physical wherewithal 114
3.3.1 High performance liquid chromatography 114
3.3.1.1 Reversed-phase chromatography 114
3.3.1.2 Hydrophilic interaction chromatography 115
3.3.1.3 Mobile phase 115
3.3.1.4 Detection 116
3.3.2 Mass spectrometry 116
3.3.2.1 Ionisation source 116
3.3.2.1.1 Electrospray ionisation 117
3.3.2.2 Mass analysers 119
3.3.2.2.1 Collision induced dissociation 119

3.3.2.2.2 Quadrupoles 120
3.3.2.2.3 3D Paul traps 123
3.3.2.2.4 Linear ion traps 124
3.3.2.2.5 Orbitraps 125
3.3.2.3 Mass detection 128
3.3.2.4 Resolution, accuracy and other nomenclature 128
3.4 Data processing 129
3.4.1 File formats 129
3.4.2 Processing methods 130
3.4.2.1 Time-averaged 130
3.4.2.2 Time-averaged plus 132
3.4.2.3 Gaussian modelling 132
3.4.3 Missing values 133
3.4.4 Isotopic grouping 134
3.4.5 Normalisation 136
3.4.6 Assignment 136
3.4.6.1 Spectral databases 137
3.4.7 Software 138
3.5 Data analysis 138
3.5.1 Cross validation 139
3.5.1.1 Leave-one-out 139
3.5.1.2 *k*-fold 139
3.5.1.3 Partition 140
3.5.1.4 Bootstrapping 140
3.5.1.5 External validation 140
3.5.2 Univariate analysis 141
3.5.2.1 t-tests 141
3.5.2.2 Mann-Whitney test 141
3.5.2.3 Analysis of variance 142
3.5.2.4 Multiple testing problem 142
3.5.3 Multivariate analysis 143
3.5.3.1 Partial least squares 143
3.5.3.2 Principal components analysis 144

3.5.3.3 Independent components analysis 144

3.5.3.4 Multiple linear regression 146

3.5.4 Multiblock methods 146

3.5.4.1 Hierarchical methods 147

3.5.4.2 Canonical correlation analysis 148

3.6 Future directions 149

4 Quality Determinants in Pea Seeds 151

4.1 Data 152

4.1.1 Samples 152

4.1.2 Extraction procedure 154

4.1.2.1 LC-MS 154

4.1.2.2 NMR 154

4.1.3 Experimental conditions 155

4.1.3.1 LC-MS 155

4.1.3.2 NMR 155

4.2 Continuous classification 157

4.2.1 Distance measures 159

4.2.2 Results 161

4.2.3 Summary 162

5 Processing I: NMR 163

5.1 Abstract 164

5.2 Introduction 164

5.3 Methods 167

5.3.1 Materials 167

5.3.2 Standard metabolite samples 167

5.3.3 Bovine blood sample 167

5.3.4 Rat brain tissue samples 168

5.3.5 NMR data processing 168

5.3.6 Pre-processing of spectra 169

5.3.7 Peak modelling 169

5.3.8 Standard metabolite templates 171

5.3.9 Peak assignment in bovine blood spectrum 174
5.3.10 Variable extraction for metabolomics 175
5.4 Results 176
5.5 Conclusion 180
5.6 Acknowledgements 181

6 Processing II: LC-MS 183
6.1 Introduction 184
6.2 Data 185
6.3 Data processing 185
6.3.1 Conversion of profile data 186
6.3.2 Chromatographic peaks 189
6.3.3 Between-file comparisons 190
6.3.4 Reference spectrum 190
6.3.5 Summary 191
6.4 Isotopic information 193
6.5 Normalisation 194
6.5.1 Inherent variability 195
6.5.2 Traditional methods 196
6.5.2.1 Total ion current 196
6.5.2.2 Internal standard 198
6.5.3 Variable-by-variable methods: QC sample mean 198
6.5.4 Variable-by-variable methods: curve fitting 200
6.5.4.1 Batch size 202
6.5.4.2 Function 202
6.5.4.3 Weighting 203
6.5.5 Results and discussion 204
6.5.6 Conclusion 207
6.6 Summary 208

7 Data Fusion 211
7.1 Introduction 212
7.2 Experimental 214

7.2.1 Data pre-processing 214
7.2.1.1 Variable distribution 215
7.2.2 Single block methods 220
7.2.3 Concatenated methods 222
7.2.4 Multiblock approaches 222
7.2.5 Canonical correlation 225
7.2.6 Summary 228
7.2.7 Data 228
7.2.8 Method effectiveness 229
7.3 Results and discussion 229
7.3.1 Classification power 229
7.3.2 Maximum number of components 233
7.3.3 Most frequently selected variables 237
7.4 Conclusion 243

8 Feature Selection with Genetic Programming 245
8.1 Introduction 246
8.2 Experimental 248
8.2.1 Data 248
8.2.2 GP parameters 249
8.3 Results 252
8.3.1 GP selection frequencies 252
8.3.2 Selecting the number of variables 253
8.3.3 Seven variables 255
8.3.4 Comparison to previous fusion techniques 256
8.3.5 Complementary selection 257
8.3.6 Positive and negative ion mode LC-MS 258
8.3.7 Correlation threshold 261
8.3.8 Concatenated GP 263
8.4 Conclusion 265

9 Feature Assignment 267
9.1 Introduction 268

9.2 Data 270

9.3 Experimental 271

9.3.1 *m/z* and mass 271

9.3.1.1 Isotopic distribution 272

9.3.1.2 Coeluting ions 274

9.3.2 NMR spectral peaks 275

9.3.2.1 Related spectral variables 276

9.3.3 Limitations 278

9.4 Results and discussion 279

9.4.1 Alanine 279

9.4.2 Valine 282

9.4.3 $C_{24}H_{42}O_{21}$ 284

9.4.4 $C_{18}H_{32}O_{16}$ 285

9.4.5 $C_6H_{12}O_6$ 286

9.4.6 $C_6H_{11}NO_2$ 286

9.5 Summary 288

9.6 Conclusion 291

10 Summary **293**

10.1 Future work 296

A LC-MS File Processing **299**

B Assignment Tables **307**

Glossary **313**

References **319**

List of Figures

2.1 A flow-chart detailing the steps that should be considered as part of the design of an experiment. 47

2.2 A typical generic protocol that is suitable for the analysis of complex mixtures. 49

2.3 1D ^{1}H NMR spectrum of Corsican honey recorded at 500 MHz. The whole spectrum is shown in (a), and (b) and (c) show expanded regions from (a). Any peaks above the horizontal line in (a) are attributable to either fructose or glucose. 61

2.4 Two 2D ^{1}H homonuclear spectra of pea extracts recorded at 500 MHz, showing the spectral region between 1 and 6 ppm. (a) shows a correlation (COSY) spectrum with fewer correlations than the corresponding total correlation (TOCSY) spectrum shown in (b). The two techniques may be viewed as complementary to each other in terms of providing keys to molecular structure. 63

2.5 The *J*-resolved experiment can be useful removing splittings from couplings from resonances, as shown in a spectrum of fish embryo extract recorded at 500 MHz. (a) shows the fine structure of the resonances and (b) the projection of the spectrum. 63

2.6 (a) shows the non-corrected diffusion ordered (DOSY) spectrum of a three component mixture recorded at 400 MHz, with ambiguous regions at approximately 3.6ppm. (b) has undergone processing, resulting in better-separated peaks in the diffusional dimension, especially around 3.6ppm. 65

2.7 The application of Correlated Trace Denoising to a phase-cycled heteronuclear single quantum coherence (HSQC) spectrum, recorded at a ^{1}H frequency of 500 MHz (^{13}C frequency of 126 MHz). In the original spectrum (a) the arrowed peak is of a noise-like intensity. After denoising, in (b) the same arrowed peak is considerably more pronounced. 73

2.8 The application of (a) fixed-width and (b) adaptive binning showing overlaid spectra from multiple samples. Shaded regions in (a) represent bins of 0.04ppm, which can be seen to incorporate multiple peaks and to splice peaks into two bins. Each shaded region in (b) represents an individual peak. 76

2.9 A screenshot from rNMR, highlighting its Region of Interest (ROI) concept. Here, 6 ROIs are shown for multiple samples (in blue), and the red peaks represent standard spectra which corroborate the ROI assignment. Not Applicable (NA) is shown for chemical shifts greater than the recorded spectral width. 78

2.10 (a) shows a Lorentzian function with amplitude, A, of 100.0 and a width at half height, w, of 0.5 Hz. The corresponding modified Lorentzian is shown in (b). Here $I = 100.0$, so that $A = \frac{5}{4}I = 125.0$ and the width, w, of 0.5 Hz occurs at $\frac{A}{2} - \frac{A}{5} = 37.5$. 79

2.11 The partial least squares (PLS) model (a) can be used to separate classes, yet the variation between the two classes constitutes both PLS components. Similarly, the orthogonal partial least squares (O-PLS) model (b) separates the two classes but the first component represents inter-group variance. 85

2.12 (a) is a mathematical representation of a genetic programming (GP) solution, of which (b) is its graphical form. (c) represents a mutation of a parent solution by alteration of the encircled nodal operator. (d) shows an example of crossover where two selected solutions are formed into two new programs. The circles highlight the crossover nodes. 87

2.13 The data matrix, $\mathbf{X}$, is decomposed by parallel factor analysis (PARAFAC) into factors and a matrix of residual values, E. Red sections represent individual spectra, each of which represents a mixture of the three sugars at varying concentrations, as defined in the table. The green portions highlight the applied gradient strength, with the graph depicting the decay of signal intensity as the gradient strength increases. The bottom section, in blue, details the resolved, individual spectra of each of the three constituents. 91

3.1 The difference between the two methods is stark: the QuEChERS method is a reliable alternative to the more convoluted approaches that preceded it. 112

3.2 The electrospray ionisation source. The capillary inlet is maintained at a potential difference to the end plate. As the voltage is increased (see expansion), the droplet at the end of the capillary elongates, and eventually forms a Taylor cone from which smaller droplets are ejected. Their continual desolvation leads to the formation of ions. Ions are accelerated past the endplate, and towards the mass analyser. 118

3.3 The arrangement of the rods in a quadrupole is such that opposing rods carry the same charge. The application of an AC voltage results in the switching of positive and negative charges between the two pairs of rods. 120

3.4 The four MS/MS experiments that can be performed with a triple quadrupole instrument. 121

3.5 A 3D Paul trap, shown in cross-section. The ring electrode may be considered as a looped rod from a quadrupole, and to it an AC voltage is applied. The ring is sandwiched by two end cap electrodes, maintained at DC voltages, that contain passages for the entry and exit of ions. Ions weave figure-of-eight-like paths around the trap, due to the quadrupolar field exerted by the ring electrode. The ramping of the voltage that is applied to the ring electrode encourages the progressive excitation of ions, which are sequentially ejected towards both the detector and the ion source. 123

3.6 An illustration of the Orbitrap, showing the off-central injection of an ion stream. The motion of ions is detected as an image current, and Fourier transformation converts the free induction decay into a mass spectrum. 126

3.7 A schematic of a Thermo Scientific Exactive Plus Orbitrap mass spectrometer. The ions are injected in the bottom right-hand corner, and are transferred into the C-trap. 127

3.8 The *m/z* value of a peak is divided by its full width at half maximum (FWHM) to obtain an instrument's resolution. 128

3.9 A schematic showing the formation of a time-averaged mass spectrum. Individual scans are combined by rounding *m/z* values to an arbitrary precision, and summing their intensities. 131

3.10 Two observations show similar intensities for a peak, shown in the green box, in the time-averaged mass spectrum. Inspection of the extracted ion chromatogram (EIC) reveals that significant variance in one peak is masked by minor fluctuations of a much larger peak. . . 131

3.11 The original data is shown by the blue stems, and the Gaussian functions calculated to model the original data are shown by the solid red line. Each peak consists of at least 10 data points, whereas each Gaussian function is defined by only 3 parameters. 135

3.12 Each horizontal line represents an individual mass spectrum, with those sections that are coloured red grouped into a single chromatographic peak. The red sections are not contiguous, yet are grouped into a single peak. Nine individual peaks are also present, but these can be safely categorised as noise peaks as no other values with similar *m/z* appear in preceding or succeeding scans. 135

3.13 Representation of PLSR. The weights (W), loadings (L) and scores (S) are shown, for the response (X) and predictor variables (Y). 144

3.14 A schematic for two-stage PCA. Both A and B matrices are subjected to an individual PCA, which produces loadings (L) and scores (S). A subset of S_A and S_B are combined for a second analysis. Trends within the data can be visualised by plotting the fused scores (S_F). The fused loadings, L_F, indicate which principal components from the original analyses contribute to any observed trends. 148

4.1 PCA scores plot from analysis of the HSQC data from the pea study. It can be seen that both components exhibit changes in tenderness values, rather than it being confined to a single component. 158

4.2 The D^2 values for various distance metrics, calculated with up to 10 PCs. The inclusion of additional PCs tends to lower the D^2 value for all metrics, but most notably in the Mahalanobis measure. The city block implementation tends to mirror this effect, although to a lesser extent. Both the Euclidean and Chebychev distances stabilise with the inclusion of more components, but this point comes sooner for the Chebychev metric. 160

4.3 The ρ and D^2 correlation coefficients. The value of ρ show the correlation between a single component and the tenderness values, whilst the D^2 values calculate the distance matrix using all preceding components. 161

5.1 The Lorentzian line with an amplitude, A, of 50.0 and a width at half height, w, of 0.25 Hz is shown in (a). In (b) the corresponding modified Lorentzian is shown. Here $I = 50.0$ so that $A = 62.5$ and the width, w, of 0.25 Hz occurs at $A/2 - A/5 = 18.75$. 170

5.2 The radial width, w, through the point (x,y) of an ellipse with semimajor axis a and semiminor axis b, centred at (0,0). 172

5.3 A section of an HSQC spectra containing several overlapping peaks is shown in (a) with the corresponding modelled peaks shown in (b). 172

5.4 A glutamine peak in the bovine spectrum at 3.78 ppm (proton) and 56.85 ppm (carbon) swamped by a larger glutamate peak at 3.76 ppm (proton) and 57.39 ppm (carbon) so that the two are treated as a single peak. 175

5.5 Plot of the first two principal component scores. Solid symbols represent observations from rats injected with [U-^{13}C]-glucose and open symbols those injected with normal ^{12}C-glucose. The four post-injection time points are represented by different symbols with circles, diamonds, squares and triangles representing 0.5, 1, 2 and 4 hours respectively. With the exception of one early time-point [U-^{13}C]-glucose observation, there is clear separation of the two groups along the first principal component. It can be seen that the difference between samples from rats injected with [U-^{13}C]-glucose and those from rats injected with normal ^{12}C-glucose decreases gradually with time. 178

6.1 The calculation of a peak's half width is effected by linear interpolation. Where the linear interpolation and the intensity at half maximum (I/2) intersect, a peak's half-width is defined. The width at each side, where possible, is calculated to provide a FWHM. If peaks are resolved on one side only, then the FHWM is twice the value of the calculable half-width. 187

6.2 The widths of peaks, taken from 6 successive scans, plotted as a function of *m/z*. The data are recorded using an orbitrap analyser with the resolution set to 50,000 FWHM (*m/z* 400). It is clear that few points exhibit such a characteristic, and using it to filter the data is unwise. . 188

6.3 The peaks detected in negative mode HILIC LC-MS analyses of the pea samples, prior to the isotoping stage. The bespoke method only processes *m/z* values greater than 200 but detects approximately three times more features than the XCMS software. 192

6.4 PCA scores plot for the positive-mode RP-LC-MS variables, where observations are coloured according to the batch in which they were analysed. 197

6.5 The two-way ANOVA plot for the positive-mode HILIC variables, showing which variables are associated with which of the two factors. The p-values have been sorted so that it is clear that more variables exhibit run order-like dependency than vary according to the sample tenderness values. 197

6.6 The intensities of a single variable plotted as a function of run order. The six distinct batches can be observed: the first three show clear differences in intensities, whilst the intensity change in the final three batches is of a more continuous nature. 199

6.7 A variable's raw intensities are depicted in the top plot, and the interpolated correction function is depicted by the triangles. The bottom plot shows the effect of the correction function, such that now the QC sample intensities are considerably more constant. 201

6.8 Two sets of correction curves are shown for the same variable from positive mode RP-LC-MS. In (A) the weights for all observations are equal, with the weightings derived from the difference from the QC mean (see Equation 6.3). In (B) the weights have been multiplied by 100, such that in (B) the correction curves closely fit the QC sample intensities. 204

6.9 PCA scores plots of the same batch mean normalised data, where the observations have been coloured according to batch and tenderness. 206

7.1 The skewness values of the variables before and after each of the various transformations have been applied. The line indicates where no transformation has occurred. 218

7.2 The original and corrected skewness values are plotted for each variable. The optimal skewness is defined by the transformation that reduces its absolute magnitude closest to zero. 219

7.3 Principal components analysis of the original and corrected data variables. The scores for the original and corrected variables are shown in each plot. The second PC is most strongly related to tenderness. For both metrics, the agreement between the scores and tenderness has fallen as a result of the variable transformations. The percentage of variance explained by each component is given on the axes, with the former being for the original data and the latter for the corrected data. 219

7.4 The effect of logarithmically transforming all variables from the positive mode HILIC LC-MS data sets. Neither the D^2 value nor the correlation with tenderness (ρ) improved following the transformation. 220

7.5 A diagrammatic representation of the concatenation of three previously standardised data matrices (A, B, C) prior to multivariate analysis, which, in these methods, may be PCA or PLS. This example shows the formation of matrix S (scores, latent variables, etc...) and L (loadings, weights, etc...). S represents new variables, whilst L reflects the contributions made by the original variables. 222

7.6 A representation of a hierarchical fusion method, involving three data blocks (A, B, C). Each of these three is individually standardised and subjected to the same multivariate analysis, which results in new variables (S_1) which are then partially combined in a superscores matrix (S_f). This matrix is then standardised and subjected to a further round of multivariate analysis. 224

7.7 A flowchart detailing the various matrices involved in a two-stage CCA. The original data is transformed into new variables, S_x and S_y. Canonical correlation analysis (CCA) is performed, usually, on subsets of S_x and S_y. The new scores (U and V) are calculated such that the correlations (ρ) between each pair are maximised. Matrices C and D relate the contributions back to the original variables, and are calculated according to Equation 7.5. 227

7.8 Scores and loadings plots from hPCA using positive mode HILIC MS and HSQC data sets. Two PCs from each low level PCA have been used in the high level model. The high level scores are shown in (a), whilst in (b) and (c), respectively, the low level scores are shown for the MS and NMR data. After correction, the loadings for each block are shown in (d). 236

7.9 The loadings for the positive-mode HILIC LC-MS variables, which have been grouped according to the main multivariate technique. . . 241

7.10 The loadings for the HSQC variables, which have been grouped according to the main multivariate technique. 242

8.1 A typical GP tree. The variables (V) and constants (C) are shown by the green (leaf) nodes, whilst the operations are given in the brown (branch) nodes. The topmost branch node shows the subtraction of two MS variables, with other nodes showing addition (ADD), division (DIV), and sine (SIN). 250

8.2 The GP selection frequencies shown for each positive-mode HILIC LC-MS variable. Variables that have not been selected in any of the solutions are marked with × symbols. 252

8.3 The GP selection frequencies for the HSQC variables are shown, and those that are not selected are marked denoted by ×. 252

8.4 Correlation coefficients derived from CCA. The number of variables in each data set is successively increased until the data becomes rank deficient. The canonical correlation increases with each new variable pair, but the correlation of the scores to tenderness plateaus after the inclusion of seven variables from each data set. WIth more than 22 variable pairs, the tenderness correlation begins to fall. 253

8.5 Scatter plot of the first canonical pair, formed when seven variables from each dataset are used. 254

8.6 The loadings for the seven MS and seven NMR variables used in CCA. Only two variables have strong positive contributions in CP1. The most frequently selected variables by GP do not contrirbute most strongly to the trend seen in Figure 8.5. 254

8.7 Dendrogram of the fourteen variables, where the distance is measured according to one minus the correlation coefficient. It can be seen that the variables with large similar loadings in Figure 8.6 are clustered together. 254

8.8 The frequency of selection of MS and NMR variables plotted as a function of the correlation to tenderness. The top seven variables from each technique have been labeled and shown as circles. 256

8.9 The ranks of the GP-selected variables in other fusion techniques. A value of 1 indicates the largest absolute loading. 257

8.10 Plots showing how often other variables are selected with the first six variables, which are given by their indices on the y-axis. 258

8.11 The loadings for the positive and negative mode HILIC MS variables. In (B) the correlated variables within the same dataset have been removed. 259

8.12 The loadings from GP-CCA for the MS variables when paired with the HSQC variables. Variables correlated by greater than 0.9 to others within the same dataset are removed. Fourteen variables were found to be optimal for the positive mode MS variables, and thirteen for the negative mode variables. 261

8.13 The frequency of selection of MS and NMR variables plotted as a function of the correlation to tenderness. The two-stage GP was run with concatenated data, and then split according to MS and NMR variables. The variables with a selection frequency greater than 20 have been labeled. 264

8.14 The canonical and tenderness correlation values for GP-CCA, where the GP was run with concatenated MS and NMR data. Variables correlated within the same dataset with $\rho > 0.9$ have been removed. . . . 265

9.1 RANSY plot created for the encircled query variable at δH = 3.81, δC = 53.60 ppm. The intensity of the ratio for each variable to the query variable is given according to the colour shown in the colour bar. Only one other peak has a large positive value, which indicates that these peaks consistently vary in the same ratio across multiple observations. Their exclusive presence supports an assignment of alanine. 281

9.2 The structures for the ten candidates assignments in Table B.2. Clockwise, from top left, the molecules are: histamine, betaine, valine, 5-aminopentanoic acid, senecioic acid, propylene glycol, 2-ethylacrylic acid, tiglic acid, 2-pyrrolidinone and N-methyl-α-aminoisobutyric acid. Structure images taken from METLIN. 283

9.3 The experimental isotopic distribution for *m/z* 215.0322 is shown alongside three other distributions for ions shown in Table B.9. 287

A.1 A flowchart representing the various stages in the processing of mass chromatograms. Each file is initially processed to produce a list of its constituent peaks. Subsequently, all of these lists are compared in order that a data matrix may be extracted where each column represents one common observation. A reference spectrum can be used in order to remove noise-like variables, or those occurring in a small proportion of samples. The procedure then begins to interpret the isotopic information inherent to mass spectrometry. The reference to each stage of the process is given by §. 300

A.2 Flowchart detailing the procedure employed in calculating Gaussian functions for each peak in a mass spectrum. The procedure starts with the *m/z* and intensities for the particular scan, with the output containing representations of each slice in a much more condensed format. 301

A.3 Flowchart detailing the procedure used to group slices of a chromatographic peak together. The slices from the first scan form the beginnings of the peak list, with subsequent slices compared to the peak list. Where there is no valid comparison in the peak list, a slice begins its own peak which can be expanded in the subsequent scan. Peaks are 'closed' if they are not added to within a set number of scans, and remain unmodified. 302

A.4 A flowchart that details the procedure used to compare peaks across multiple peak lists. The aim is to produce a data matrix, whereby each column represents a comparable peak across all mass chromatograms. 303

A.5 A flowchart the shows the procedure used to calculated a 'group median' reference spectrum. Variables that are not found in at least half of a class' observations are removed from the data matrix. 304

A.6 A flowchart detailing the procedure used to group peaks into isotopic clusters. Variables that are highly correlated are likely to form part of the same isotopic cluster, and is therefore a good metric to use in addition to the distance. The charge state can be best estimated by looking at the clusters with peak increments of $\frac{1}{2}$ and 1 which correspond to *m/z* values of *m/2* and *m/1*, respectively. 305

List of Tables

4.1 Metadata for the pea samples. The TR value is the tenderness value for each sample, and serves a measure of maturity. 153

4.2 The solvent gradients for both LC types. Mobile phases A and B are described in the text. 155

5.1 Chemical shifts for the peaks that were assigned to the standard metabolites used in this study. 179

6.1 The D^2 values for the various correction methods when applied to the reversed-phase positive mode LC-MS variables. 205

6.2 The results of the various correction procedures that have been applied to remaining LC-MS datasets. 206

7.1 Various transformation methods applied to the data, on either a variable, block or observational basis. i refers to each observation, and j to each variable. The original data x_{ij} is transformed to the new value y_{ij} according to the equations below. $\bar{x_j}$ is the mean of the j^{th} variable and σ_j is its standard deviation. The definition of a vector norm is given in Equation 7.1. 216

7.2 The ladder of transformations that can be applied to variables to, for example, increase their linearity or reduce their non-normality of distribution. 217

7.3 The abbreviations used to refer to specific methods. The first three are the single-block methods, and are included so as to facilitate comparisons between the single- and equivalent multi-block models. 228

7.4 The D^2 values for the unsupervised fusion techniques. A maximum of 50 components from each dataset are considered. Values greater than 0.75 are coloured red. The concatenated methods are only scaled once according to the method indicated in the low level column. . . . 231

7.5 The ρ values for the supervised fusion techniques. The maximum number of components included from each dataset is 50. Values greater than 0.75 have been coloured red. cPLS has only had the low level standardisation applied. 232

7.6 The results when the maximum number of components from each dataset is changed. The data were autoscaled. 233

7.7 The effect of the number of components and the D^2 and ρ values. The maximum number of permitted components allowed in high level fusion is given in each column. For hPCA and hICA the changes are marginal, but mPCA uses these additional components to reduce the error and increase ρ. 234

7.8 The results of the single block methods. 234

7.9 The ranks of the top positive mode HILIC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values. 238

7.10 The ranks of the top HSQC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values. 239

7.11 The ranks of the top positive mode HILIC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values. 240

7.12 The ranks of the top HSQC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values. 240

8.1 GP parameters employed in the first and second stages. 251

8.2 The variables identified by the three GP-CCA routines that have the largest loadings. They have been grouped according to their correlation (ρ) with the tenderness values, and also according to the dataset from which they were taken. Variables that were removed from the CCA due to their high correlation are shown in italics. 263

9.1 An example entry of a molecule's record within JDB. All of the data is taken from HMDB. 271

9.2 The positive and negative mode adduct species that have been considered, along with their notation and charge. ACN, acetonitrile C_2H_4N; DMSO, dimethyl sulphoxide C_2H_6OS; IsoProp, 2-propanol C_3H_8O; FA, formic acid CH_2O_2; HAc, acetic acid $C_2H_4O_2$; TFA, trifluroroacetic acid $C_2HF_3O_2$. 273

9.3 The 18 elements and their compositions used in the determination of isotopic profiles. 273

9.4 HSQC hits to tentative assignments shown in Table B.1. The tolerances for the search were δH = ±0.1 ppm, δC = ±0.2 ppm. 280

9.5 Ions eluting within 10 seconds of the ion with *m/z* = 118.0863 at t / min = 6.427 282

9.6 HSQC hits to tentative assignments shown in Table B.2. The tolerances for the search were δH = ±0.1 ppm, δC = ±0.2 ppm. 283

9.7 Ions eluting within 10 seconds of the ion with *m/z* = 487.1654 at t / min = 13.8676 284

9.8 Ions eluting within 10 seconds of the ion with *m/z* = 503.1613 at t / min = 11.6193 286

9.9 Ions eluting within 10 seconds of the ion with *m/z* = 130.0862 at t / min = 6.7124 287

9.10 Ions eluting within 10 seconds of the ion with *m/z* = 84.0811 at t / min = 18.3632 288

9.11 Variables that are positively correlated to tenderness values, and their putative assignments. The confidence in assignment is given according to that defined by the MSI. 289

9.12 The putative assignments of the negatively correlated variables. The confidence in assignment is given according to that defined by the MSI.290

B.1 Potential candidate assignments for variable 16: t / min = 8.5684, *m/z* = 90.0552, z = 1. 308

B.2 Potential candidate assignments for variable 13: t / min = 6.427, *m/z* = 118.0863, z = 1. 308

B.3 Potential candidate assignments for variable 139: t / min = 13.8676, *m/z* = 487.1654, z = 1. 309

B.4 Potential candidate assignments for variable 133: t / min = 13.8814, *m/z* = 689.2114, z = 1. 309

B.5 Potential candidate assignments for variable 169: t / min = 14.0049, *m/z* = 684.2556, z = 1. 309

B.6 Potential candidate assignments for variable 67: t / min = 11.6193, *m/z* = 503.1613, z = 1. 310

B.7 Potential candidate assignments for variable 28: t / min = 11.6162, *m/z* = 549.1659, z = 1. 310

B.8 Potential candidate assignments for variable 60: t / min = 11.6165, *m/z* = 617.1545, z = 1. 310

B.9 Potential candidate assignments for variable 23: t / min = 5.241, *m/z* = 215.0322, z = 1. 311

B.10 Potential candidate assignments for variable 11: t / min = 6.7124, *m/z* = 130.0862, z = 1. 311

Acknowledgements

> No duty is more urgent than that of returning thanks.
>
> James Allen

Adrian Charlton, Jane Thomas-Oates, Julie Wilson
To my cohort of supervisors I owe an enormous debt of gratitude. For not only did they give me the opportunity but through inspiration, perseverance, encouragement and humour they made it enjoyable.

James Chisholm, Sarah Christmas, Matthew Collins, Mike Dickinson, James Donarski, Caryn Douglas, Helen Grundy, Mark Harrison, Karl Heaton, former and current members of the JTO group
Many others have helped me significantly throughout the course of my research, and without them I would have had either no data or no clue.

YCCSA, Department of Chemistry, Fera
My time at university has been split between three places, and I'm grateful and fortunate to have worked amongst some wonderfully stimulating and generally marvellous people.

And finally I would like to thank my family.

Author's declaration

With the exception of two chapters, all of the work contained in this thesis is original and has not been submitted for any other degree at this or any other institution. Chapters 2 and 5 have been previously published and are reproduced verbatim.

Chapter 2: McKenzie, J.S., Donarski, J.A., Wilson, J.C. and Charlton, A.J. (2011) Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics, *Progress in Nuclear Magnetic Resonance Spectroscopy*, **59**, 336. JSM was the main author of this paper.

Chapter 5: McKenzie, J. S., Charlton, A. J., Donarski, J. A., MacNicoll, A. D. and Wilson, J. C. (2010) Peak fitting in 2D ^{1}H-^{13}C HSQC NMR spectra for metabolomic studies, *Metabolomics*, **6**, 574. JSM and JCW co-wrote this paper.

Chapter 1

Introduction

> **complementary** *(adj)* **b.** of two (or more) things: mutually complementing or completing each other′s deficiencies.
>
> Oxford English Dictionary (Online)

In the field of non-targeted analysis, and for small molecules especially, the increasing capability of analytical hardware is resulting in an explosion of data. It is becoming more evident that the rate-limiting step in experiments is now related to the interpretation, rather than collection, of the data. The discipline of non-targeted analysis encompasses many potential applications, and offers a valuable alternative paradigm to the more constrained targeted analysis. Where the latter relies on prior knowledge of, for example, likely known toxicants, it is non-targeted analysis that is capable of being applied proactively and to systems where effects are currently unknown.

Metabolomics has been described by Fiehn as "a comprehensive analysis in which all the metabolites of a biological system are identified and quantified" [1]. Such studies, therefore, require a rigorous analysis of a system's metabolites which consequently must be as unbiased as possible. The use of multiple analytical techniques can help to enhance the analytical coverage, and potentially contribute to a more informative understanding of a system under investigation.

The generalised aim of a non-targeted study is to identify differences, which necessarily entails a comparison between observations of samples generated on subjecting the material being studied to different conditions. Suitable bases for comparison may be a reference material or the subjects of a control group in a medical study. With appropriate experimental design, differences between observations of the sample and the control group may be reasoned in terms of the differential treatment between the control and other samples. In medical studies especially, however, there are likely to be many confounding factors which, coupled with small sample cohorts, can combine and obscure informative trends [2, 3]. Non-targeted analysis is widely applied in, amongst others, toxicological [4–6], food [7–10], environmental [11, 12] and medical [13, 14] applications.

Non-targeted approaches rely on low analytical bias, with the aim of characterising as many components of a system as possible. The two most commonly used platforms in non-targeted analysis of small molecules are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), where MS is often coupled to liquid or gas chromatography (LC or GC). Whilst many approaches in the literature employ only a single technique, the complementary use of both has the potential to enhance the analytical coverage of samples, and to increase the quantity and value of information that may be extracted.

The combination of the two techniques builds on their complementarity. NMR spectroscopy may be considered to have no specificity of detection, as all molecules with hydrogen atoms induce an NMR signal. Its sensitivity, however, is typically much lower than that achievable with mass spectrometers which are generally regarded as having higher specificity than NMR due to the nature of the chromatographic elution and the mechanisms of ion formation and detection. Thus, the two techniques can be considered to be complementary. Furthermore, the nature of the information is different; MS provides information on the size and likely composition of an analyte, whilst NMR contributes structural information.

The very virtue of having collected more data does not, unfortunately, immediately equate to more information. The data are collected independently, but must be analysed in a concerted fashion so as to maximise the information recovery. Methods that employ multiple data sources are often termed data fusion. The aim of data fusion is to extract more information from the concerted use of data from multiple sources than could be achieved were only conventional single-source methods employed [15].

Of a plethora of multivariate statistical techniques, many of the most popular operate on a single data matrix. Techniques such as principal components analysis (PCA) [16] and partial least squares (PLS[1]) [17] are the most widely used multivariate methods for probing data from non-targeted analyses. Being single block methods, they are unsuitable for data fusion without adaptation of either the technique or the data. The simplest form of fusion (termed 'low level') involves the concatenation of multiple data matrices such that they are compatible with standard

[1] Perhaps more correctly referred to as 'projection to latent structures'.

multivariate techniques. Intermediate (or mid-) level fusion generally involves joining subsets of variables from multiple techniques, whilst high level fusion typically entails the combination of results from individual multivariate analyses. The aims of data fusion are typically either or both of the following: to enhance the classification or between-group discrimination of samples, or to assist with the assignment of molecules by relating spectral variables from multiple techniques.

Much of the chemical literature has focused on the development and implementation of high level fusion techniques, specifically multiblock or two-stage methods [18–22], which are typically variations of PCA and PLS. High level fusion has been shown to enhance the separation observed in low level approaches [20], and multiblock methods may be preferred as they maintain some of the structure inherent to the original data. Indeed, the grouping of spectral variables into distinct blocks allows the contribution of each block to be assessed and may prevent certain variables or blocks from exerting an excessive influence [23].

Many of the methods encountered in the literature focus on enhancing the classification power of the fused system in comparison to the more conventional single block approaches. High level fusion techniques relate trends in groups of variables, rather than variables themselves, across data sets. Whilst this assists with enhancing, for example, between-group separation, it may not highlight complementary variables across multiple techniques. Mid-level fusion techniques [24] involve a feature selection stage to determine discriminatory variables from each dataset. Applying a threshold to variable loadings allows obviously uninformative variables to be discarded, with subsets of variables from each dataset being combined. Unlike with low-level concatenation approaches, for mid-level fusion the observation : variable ratio becomes more favourable for multivariate techniques such as PCA and PLS.

The determination of suitable variable subsets is being increasingly commonly performed with evolutionary algorithms, where individual solutions evolve through processes designed to mimic the ‘survival of the fittest’ paradigm. Genetic programming (GP) [25] is one such example which is widely used for feature selection and classification applications [8, 26, 27]. Individual solutions in GP take the form of a simple computer program where variables are linked through a series of mathematical operations. Discriminatory variables are those that are picked frequently throughout many independent GP runs. Thus the application of GP to multiple datasets produces subsets of frequently picked variables which can be investigated by multivariate techniques.

Canonical correlation analysis (CCA [28]) is capable of analysing two sets of variables, and has been applied in geographical [29, 30] and gene expression studies [31]. Its application to high level data fusion in a chemical context has been demonstrated by Doeswijk et al. [32] in a study which initially used PLS analyses to reduce the complexity of GC- and LC-MS datasets prior to analysis with CCA. The high level fusion showed how trends within the individual datasets are related. The technique is incompatible with multivariate datasets due to their highly intercorrelated nature.

The combination of GP and CCA in a mid-level fusion approach allows their individual strengths to combine. GP is a highly capable feature selection procedure which can be used to identify subsets of discriminatory variables. CCA allows two such subsets to be related in a multivariate fashion rather than the more simplistic correlation based approaches (e.g. [33, 34]).

The wealth of information that exists within a dataset may, however, be insufficient to provide a confident spectral assignment. In electrospray ionisation mass spectrometry there is little fragmentation of ions, such that detailed structural information is often not available. The combination of accurate *m/z* and isotopic distributions is unlikely to allow a single candidate to be identified; Kind and Fiehn [35] have demonstrated that even sub-parts per million accuracy of high masses does not result in a unique molecular formula. Whilst interpretation of tandem MS fragmentation patterns may offer extra information regarding likely candidates, it is often not practical to acquire MS/MS spectra in non-targeted analyses. Akin to MS, one-dimensional NMR spectroscopy is sometimes unable to provide a definite assignment and is often hindered by the high density and superimposition of spectral peaks when analysing complex mixtures. Two-dimensional heteronuclear NMR spectroscopy has been demonstrated to be effective for metabolomics purposes [36], but is somewhat limited by its low sensitivity. The combination of MS and NMR spectroscopy, therefore, offers the possibility of using the strengths of each technique to mitigate deficiencies in the other. Whilst MS is generally unable to differentiate between isomeric candidates, the use of NMR spectroscopy may help to provide a single unique assignment. The assignment of spectral features is recognised as a major bottleneck in the data processing pipeline [37], and the combination of multiple spectroscopies may help somewhat in its widening.

1.1 Scope of the thesis

The generalised aim of this thesis is to describe the development and assessment of methods to extract and utilise information from multiple spectral resources. The extraction of spectral information allows comparisons between observations to be made, and the assignment of spectral features. The quantity of information available within spectra is large, and the extraction and use of such information is becoming more important as the ability of instruments to collect yet more data continues to rise. The applicability of NMR and LC-MS for the analysis of complex mixtures is reviewed in Chapters 2 and 3.

Development and application of a processing method for heteronuclear NMR spectra is presented in Chapter 5, and is based on the mathematical modelling of three-dimensional Lorentzian-like peaks. The processing allows the footprint of partially resolved peaks to be estimated such that accurate peak integrals can be calculated. In crowded spectral regions this approach confers significant advantages over methods that calculate peak integrals using rectangular footprints of unresolved peaks.

The wealth of information within an LC-MS dataset is enormous, and various processing procedures have been developed. These are presented in Chapter 6 and detail the extraction of peaks as a function of *m/z* and retention time, and their subsequent grouping into isotopic distributions. Also presented are various normalisation methods which are required in order to take account of drift in instrumental performance.

Whilst the individual datasets are highly informative, their complementary nature is such that a fused approach is likely to reveal more information than were the data from the two techniques to be analysed independently. Chapter 7 discusses and assesses many of the existing low and high level fusion approaches, whilst in Chapter 8 a novel mid-level fusion approach is described and its implementation presented. The mid-level approach selects complementary variables and shows in an intuitive fashion their inter-relationships.

Whilst identification of spectral features can be performed without prior statistical analysis, the application of fusion methods allows researchers to focus assignment efforts on variables that are highly discriminatory towards a specific property, for example class separation. Chapter 4 describes a study whose aim is to identify molecules which differentiate peas according to tenderness; identification of such molecules is presented and discussed in Chapter 9.

Chapter 2

Nuclear Magnetic Resonance

“Data! Data! Data!" he cried impatiently. “I can’t make bricks without clay."

The Adventures of Sherlock Holmes
Arthur Conan Doyle

2.1 Introduction

Minor variation in the composition of a mixture can have acute effects on its characteristic properties, such as flavour, efficacy or toxicity. Thus, the analysis of complex mixtures is of significant importance to the food industry [38] as well as in pharmaceutical [39], agricultural [40] and environmental chemistry [41]. In the age of the discerning consumer, food and drink producers are placing great emphasis on the quality and reproducibility of flavours and textures and compositional profiling plays a pivotal role in ensuring the chemical consistency of products between batches [42]. The European Union's Protected Geographical Status legislature and multiple national schemes are designed to protect regional foods and inform consumers. Foodstuffs with protected origins command higher prices, often associated with greater quality, and complex mixture analysis is paramount in the authentication of such products. Non-targeted screening can flag up unexpected or unwanted molecules, from contaminants in food products [43] to toxins in the watercourse [44]. Environmental scientists must monitor the effects of various pollutants on human health and the ecosystem and mixture analysis is involved in measuring the quality of air, soil and bodies of water [45]. Sentinel organisms are also helpful indicators of environmental health [46–48] often incorporating the use of metabolomics approaches to study complex biological preparations derived from indicators of environmental stressors such as drought [11]. The term metabolomics [1] is used to describe the non-biased analysis of small molecules present in complex mixtures, which can be used to identify biomarkers specific to certain stimuli. The metabolic changes may be responses to drugs, environmental changes or diseases, and metabolomics could lead to more efficient diagnosis and improved treatment.

Regulatory bodies are interested in assessing and minimising the impact of chemical waste and in detecting specific toxicants accidentally released into the environment. Perhaps the most critical application of complex mixture analysis is in the pharmaceutical industry, where the composition of the product is essential to efficacy and safety, making knowledge of the identity of formulation impurities paramount to both quality assurance and safety testing. In many countries, similar legislation and safety concerns drive the need for a complete understanding of the impurities present in agrochemical formulations. Such approaches are also used for the detection of unknown terrorist threats as applied to complex mixtures, such as the rapid determination of chemical agents and their degradants in the environment [44, 49].

New technology, advances in methodology and increased computational power have led to an enormous increase in the amount of data generated in analytical chemistry. The analysis, interpretation and presentation of chemical information underpin scientific research and the proliferation of data has resulted in the discipline of chemometrics. Developments in hardware have ensured the wide application of nuclear magnetic resonance (NMR) techniques [50], and when coupled with chemometrics the breadth of application of the technique increased markedly. Recording spectra from complex mixtures is no more challenging than from pure compounds although datasets can be highly complex. Advances in computational and statistical methods not only allow useful information to be extracted from such megavariate datasets, but also influence the development of instrumentation.

Careful planning of the sample preparation and experimental setup is of vital importance to obtain the best possible data for analysis and interpretation. Experimental design and sampling strategies need to be considered in relation to the questions to be answered and sample preparation will depend on the objectives of the study and be informed by the type of sample matrix. Chemical shift differences arising from different chemical structures are the fundamental reason for the use of NMR in the analysis of complex mixtures. However, experimental parameters such as the temperature, pH and ionic strength at which the spectra are recorded can also lead to changes in the dynamics of the sample and therefore peak positions. During data acquisition the temperature of the sample is tightly controlled leading to minimal chemical shift variation due to temperature fluctuations. The pH of an NMR sample can be controlled by the use of buffers or extraction solvents effectively limiting variation in chemical shifts. Control and monitoring of the experiment is critical to maintain data quantity, and working groups [51] have been set-up to encourage acquisition conditions to be reported alongside any spectral data, as a means to both encourage good practice and facilitate sharing and collaboration.

The alignment of data is particularly important in comparative studies but experimental parameters and sample conditions cannot always be adequately controlled. Salts present in the material under investigation affect the ionic strength of the sample and can only be controlled by introducing additional steps into the extraction methodology that will inevitably impact on the sample composition and reduce the comparability of the data obtained. Chemical interactions will inevitably occur in samples that contain many compounds and are often more pronounced in certain sample matrices. For example, urine has a wider range of ionic strengths and therefore produces larger chemical shift ranges for individual resonances than blood. It is now common practice to accommodate these spectral changes by averaging data points over a range of frequencies. Apodisation of the free induction decay (FID) to broaden the NMR peaks removes the effect of chemical shift changes at the cost of a significant loss in resolution. A common alternative approach involves binning the spectral data into a fixed number of user-defined ranges [52, 53]. The bin widths are quoted in terms of the chemical shift range that they encompass and typically 0.04 ppm bins are used. Recently, methods for adaptively binning the data to avoid data from a single NMR resonance being represented in adjacent bins have been developed [54–56].

The most appropriate NMR experiment may depend on factors such as speed or sensitivity. The ubiquity and naturally high isotopic abundance of the ^{1}H nucleus has ensured the important role of ^{1}H NMR in chemical analyses. The water resonance often dominates ^{1}H NMR spectra, often being positioned in the centre of the spectral region of interest. Developments in pulse sequences (e.g. WATERGATE [57], WET [58], CPMG [59], shaped pulses [60] etc. . .) allow for the influence of the water resonance and those from other solvents to be minimised. ^{1}H NMR remains competitive in relation to other analytical techniques for assessing the composition of complex mixtures, with the benefit of easy sample preparation. Furthermore, the availability of ^{1}H high-resolution magic angle spinning (^{1}H HR-MAS) NMR allows analysis of solid samples at high resolution. As whole samples can be analysed, the risk of sample degradation associated with extraction or solubilisation is avoided. Spectroscopy of ^{13}C nuclei was traditionally stymied by its low natural abundance, but new polarisation techniques have reported massive increases in sensitivity [61–63].

^{19}F spectroscopy is highly utilised in fluorochemical studies [64–67], such as drug candidate molecules, and ^{31}P experiments are important to researchers in the fields of biochemistry [68, 69], food analysis [70, 71] and environmental studies [49, 72–74].

Many different types of 2D NMR experiments exist, providing information about chemical shifts, *J*-couplings and diffusion coefficients that can be used in database searches to identify particular metabolites. The diffusion ordered spectroscopy (DOSY) experiment [75] resolves the NMR spectrum in two dimensions. One dimension is the diffusion rate (which can be roughly equated to molecular size) and the other dimension is resolved due to ^{1}H chemical shifts. In theory it should therefore be possible to obtain the 1D ^{1}H spectrum of a single molecule from within a complex mixture. However, diffusion experiments such as DOSY are currently not sufficiently well resolved in the diffusion dimension to reliably enable complete resolution of the 1D spectra of individual molecules. The use of heteronuclear experiments such as heteronuclear single quantum coherence (HSQC) [76] and heteronuclear multiple bond correlation (HMBC) [77, 78] give highly resolved spectra allowing molecular connectivities and more accurate ^{1}H chemical shifts to be determined.

At present, the major drawbacks to two-dimensional techniques are the significantly longer data acquisition times and the lack of chemometric methods to handle the huge datasets efficiently. However, sensitivity improvements due to polarisation techniques such as para-hydrogen sensitivity enhancement [62] and dynamic nuclear polarisation [79] allow spectra to be acquired with fewer scans. Such technological innovation has made data collection time considerably more practicable and two-dimensional data processing methods are being developed [36, 80]. Although techniques such as HSQC can suffer from systematic noise, sophisticated denoising algorithms now significantly reduce the effects of such artefacts [81]. Rapid multidimensional spectroscopy [82–87] is becoming a realistic goal, and could be another major milestone in the history of NMR innovation.

Typical one- and two-dimensional spectra may consist of tens of thousands and millions of data points, respectively. Although analysis with this many variables is now possible, the extraction of relevant features is not only computationally less expensive but also gives improved results due to the removal of noise. The aim of chemometrics is to reduce the quantity of data whilst maintaining its quality. Methods that use integrated peaks and ignore noise regions in the spectra can drastically reduce the number of variables to be considered in multivariate analyses.

Principal components analysis (PCA) is one of the most widely used chemometric tools. As an unsupervised method, PCA can be used in exploratory data analysis to visualise the data in a few characteristic dimensions and identify patterns and potential outliers. The combination of PCA with techniques to incorporate class information, such as principal components regression (PCR) or linear discriminant analysis (PC-LDA), provide a supervised method that can be used for the classification of samples. Partial Least Squares techniques (such as PLSR and PLS-DA) are commonly used supervised methods, and can provide good classification results. Cross-validation must, however, be applied to supervised methods to ensure that the model can generalise to independent data. Artificial intelligence has also been applied in chemometric analyses and computational learning techniques, such as neural networks, genetic algorithms and genetic programming, can provide results that are more readily interpreted in terms of the original data.

Experiments that involve the analysis of complex mixtures, perhaps previously monopolised by chromatographic techniques, are now entirely possible through the application of a range of NMR experiments in conjunction with sophisticated means of extracting valuable information from very large and highly complex datasets. NMR is a highly reproducible and information-rich technique and developments are set to continue. This article seeks to review the latest research in the use of NMR for the analysis of complex mixtures.

2.2 Approaches to experimental design

Good experimental design ensures that there are sufficient meaningful data to answer the specific questions that the experiment is intended to answer. It is therefore important to define the aims of the study and to identify the factors that can cause variability between samples. Any factors that cannot be fixed in the experiment but can be controlled can be dealt with by “blocking”, i.e. by considering the samples in blocks related to the factors. The number of samples required to ensure that the results are statistically relevant should be determined and will depend on the expected magnitude of the difference of interest and on the significance criterion. Samples should be selected to cover the experimental domain efficiently and a balanced sampling strategy needs to be established. Selection bias can overestimate or underestimate the significance of the results, as can any systematic differences during sample preparation and data collection. Analytical and biological variation need to be minimised and sample order should be randomised to prevent confounding with some aspect of the experimental protocol or analytical method. Trial experiments may be needed to optimise the parameters in experimental protocols, but statistical design of experiments can also be used to provide accurate, reproducible and interpretable data. Fig. 2.1 outlines the steps that should be considered as part of the design of experiments.

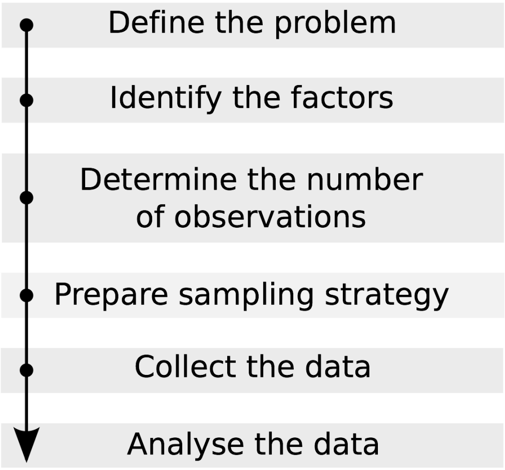


FIGURE 2.1: A flow-chart detailing the steps that should be considered as part of the design of an experiment.

It is recommended that statistical design of experiments is used throughout the whole process, from defining the aims of the experiment to the extraction of relevant information during data analysis [88]. In designing the sampling strategy and experimental protocol, the number of variables, or pertinent factors, can be varied systematically to ensure that maximum information is generated in the least number of experiments. Changing one variable at a time (OVAT) does not show the interaction between variables, whereas Design of Experiments (DoE) can determine the relationship between the output and the different factors affecting the process and has the ability to identify interaction between the factors. Response surface methodology (RSM) utilises experimental designs, such as factorial designs, central composite designs and artificial neural networks (ANNs), and has been applied to the optimisation of many analytical procedures [89].

In the final stage of the process, the analysis of complex data sets, involving many thousands of variables, requires multivariate methods to reduce the dimensionality. Projection-based methods, such as Principal Components Analysis (PCA) and Partial Least Squares (PLS), rely on the assumption that a few latent variables can account for most of the variance, and therefore most of the information, in the data. Genetic algorithms and genetic programming, based on the theory of evolution, were created to solve optimisation problems (for an excellent introduction, see Koza [25]) and can also be applied to feature selection in multivariate data sets [27].

2.2.1 Sampling and collection procedures

The errors that are associated with the sampling process may overwhelm those that are intrinsic to the analytical technique [90]. Thus, sampling is an important stage and should be considered less as a means to an end, and more as an essential part of a rigorously followed operating procedure, as outlined in Fig. 2.2. The most important part of this protocol needs to address the fundamental question: how does one obtain a representative sample?

It is important to ensure that any inhomogeneity in a material is considered. This is especially critical in solids, which are likely to consist of a spread of particle sizes [90]. This dispersion can point to differing compositions, and effective steps need to be taken to reduce the particles to a consistent size so that thorough mixing can ensure homogeneity. Liquids are relatively easy to homogenise with comprehensive mixing, although suspensions, slurries and highly viscous matrices require extra attention. Blood samples, which generally involve collecting a fraction of the total volume, are considered to be representative. Excreted biofluids can be homogenised as with any other sample. It is of paramount importance that the sample taken for analysis is a true representation of the whole.

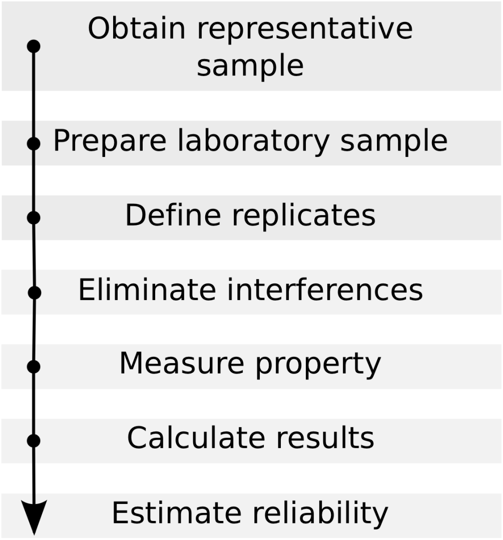


FIGURE 2.2: A typical generic protocol that is suitable for the analysis of complex mixtures.

Storage of samples post-collection is important for the sample to remain unchanged from the time that it is sampled to the time that it is analysed. Rapid freezing in liquid nitrogen will both attenuate (bio)chemical activity and preserve structural integrity until the sample is ready for preparation and analysis, and this snap-freezing method is commonly applied to biological samples. Metabolomic [1] studies, for example, need exceedingly stringent sampling strategies as any deviations from ideal collection and handling may introduce additional variation [91] and lead to erroneous biomarker identification.

2.3 Sample preparation

The method used to prepare samples for NMR analysis is of critical importance and warrants significant investigation prior to the adoption of a standard operation procedure (SOP) [92–94]. There are a number of considerations in relation to the sample preparation protocol, including for repeatability, sample stability and the range of compounds that are captured within the NMR sample. Although the sample preparation procedures employed in NMR analysis can be simple in comparison to other analytical techniques, the relatively large amounts of analyte required for analysis may mean that extra steps, such as pre-concentration, need to be performed.

2.3.1 Matrix considerations

The matrix can be liquid or solid. In fact, liquids may not be especially fluid, and highly viscous substances, such as oils and honey, will require different preparations to more fluid matrices, such as urine and blood. Solid matrices are unlikely to withstand dissolution without severe molecular changes, but magic-angle spinning allows for direct analysis of solids. If the solid matrix is merely acting as a carrier for something liquid, for example meat or cultured cells, then freeze/thawing may help to extract the liquid phase.

2.3.2 Target molecule considerations

The type of study may exact differing requirements on the preparation stage. Metabolomics studies may be broadly categorised as follows [95]:

1. Targeted: study of one or several previously identified compounds.

2. Fingerprinting: analysis that can provide characteristic metabolic profiles.

3. Footprinting: examination of released products in extracellular fluids.

4. Global profiling: investigation of a broad spectrum of molecules by one of more techniques.

An alternative interpretation of foot- and finger-printing is related to the information content within each. By means of analogy to those of a human, fingerprints are considered to be specific and have a high information content. The footprint left behind by a human is less informative than a fingerprint, and can disguise other footprints. Hence footprints are likely to have a lower information content, such as data which may have been recorded on low-resolution instruments.

Broadly speaking, molecules can be assigned as either hydrophilic or lipophilic. Extraction, and subsequent analysis, of both classes in a single stage is possible with a mixture of two or more solvents. Alternatively, individual fractions can be collected. The aliphatic region of the spectrum is generally populated by the broad signals of the fatty substances, upon which are superimposed the narrow signals of the water-soluble compounds. Such a superimposition is in effect a reduction in sensitivity, and may in turn obfuscate the spectral analysis. Use of acids and bases can cause the degradation of molecules, but may be advantageous in terms of denaturing enzymes that are unlikely to be altered by water or methanol. Perchloric acid is commonly used in plant tissue extractions [96], but its strength may limit the interpretability of a spectrum [97].

Lyophilisation, or freeze-drying, is an excellent procedure for the stabilisation of samples. The procedure essentially results in the sublimation of water from the sample, and the resultant solids are resistant to degradation through microbial growth. Furthermore, the removal of water allows for the sample to be resuspended in a deuterated solvent. Problems may be associated with lyophilisation and subsequent resuspension [98], but these should be weighed against the potential benefits.

The buffering of samples is necessary where the pH is not constant, as is typical with biological samples. Certain molecules, such as citrate, are known to have pH-dependent chemical shifts and sample buffering will help to remove this unnecessary variation.

2.3.3 Samples

2.3.3.1 Chemical reactions

NMR is fundamental in synthetic chemistry, as it can be used to identify, and quantify, any molecules that have formed during the reaction. Thus side-reactions, byproducts and purity can be assayed. Sample preparation is generally simple, although it may be necessary to seal off sample tubes to prevent oxidation.

2.3.3.2 Urine

A study by Lauridsen et al. [98] proposes various preparation and storage recommendations for human urine. In order to minimise the possibility of microbial contamination, a preservative can be added on sample collection: sodium azide causes no observable change to the NMR spectra, but citrate residues are found to shift with the use of sodium fluoride. The study recommends that samples are frozen to at least -25 °C before the work-up for analysis. Phosphate buffering (pH 7.4) to a concentration of 0.3 M has been found to be acceptable, although more concentrated samples require a stronger molarity. Freeze-drying altered the NMR profile of the sample, including the disappearance of a creatinine signal.

A study by Maher et al. [99] on the effects of storage temperature and time concludes that, post-collection, human urine is stable for 24 hours at room temperature. A large-scale study (INTERMAP), reports high analytical reproducibility for urine samples, although the majority of classification errors could be attributed to differences in the handling of samples [100]. A protocol for the preparation of urine samples may be found in Beckonert et al. [101], which recommends storage at -40 °C. In our experience, we find that -80 °C is a more appropriate temperature for storage so as to ensure that all metabolism is quenched.

2.3.3.3 Faecal matter

As a solid sample, faecal matter needs to be homogenised and this can be achieved by vortex mixing in a suspension with the solvent. Freeze/thaw cycles are efficient for disrupting cellular structure, and sonication can also be applied to achieve this aim. Research into the optimal extraction of low weight metabolites from faeces, found that the best results were obtained using 1mg of faeces per 10 μl of aqueous phosphate buffer (0.1 M, pH 7.4). Homogenisation was performed with an automated tissue-lyser, although manual sonication was also found acceptable. Two extractions from the same solid sample were pooled prior to analysis [102]. It has been suggested that base and acid addition be used in faeces preparation prior to lyophilisation as a means of stabilising the sample [103].

2.3.3.4 Plasma and serum

The collection of whole blood in tubes containing anti-coagulant (e.g. heparin) gives plasma supernatant after centrifugation. Plasma contains dissolved proteins and clotting factors. If whole blood is collected without an anti-coagulant, then the serum can be extracted by allowing the sample to clot before centrifugation, resulting in a serum supernatant. Beckonert et al. [101] describe a step-by-step procedure for the preparation of plasma and serum.

Differences in sample handling and treatment have been found to introduce bias into serum and plasma analyses [91]. The impact of clotting times, freeze/thaw cycles and temperature on the variation between samples was analysed and found to be exaggerated for molecules with high molecular weights. Sukumaran et al. [104] have developed a standard operating procedure (SOP) for high-throughput studies using spectrometers equipped with an auto-sampling micro-flow probe. The procedure is designed to minimise analytical variation and could be extended beyond its initial application to plasma and serum. The study also assesses the variation introduced by individuals executing the SOPs.

2.3.3.5 Amniotic fluid

Amniotic fluid should initially be centrifuged, prior to freezing for long-term storage, to remove any cellular debris. Graça et al. [105] have studied the effect of various handling and preparation techniques, including freeze/thaw cycles, lyophilisation and temperature.

2.3.3.6 Cultured cells

Rupture of the wall or membrane is an inherent step in the extraction of metabolites from cells. Mammalian cells can undergo osmotic lysis when immersed in pure water. The osmotic movement of water into the cell causes the membrane to burst as the volume of fluid exceeds the cellular capacity. Thus, the cell contents are liberated and can be separated from the cell debris by centrifugation. However, the dilution of the cellular soup may make further handling steps necessary in order to observe low concentration metabolites. Alternatively, freeze-thaw cycles are effective in rupturing cell membranes. A combination of fast and slow freeze/thaws, complemented by sonication, should ensure rupture of the cellular structure, which can be confirmed by observation through a microscope. An investigation into the intracellular metabolites of *Escherichia coli* concluded that the single best extraction solvent was very cold methanol, although complementary extraction solvents would ensure global coverage [106].

A different approach to cell monitoring, that has been termed footprinting, applies a metabolomics approach to cell culture media to monitor alterations in nutrient levels. The approach can be applied in the rational design of cell media and to better comprehend the nutrient uptake rate of cells.

2.3.3.7 Animal tissues

Prior to solvent extraction, animal tissues are usually thoroughly ground to ensure that extraction is carried out on a homogenised sample. It is important that the tissues are not allowed to thaw prior to extraction, so grinding should be carried out in a liquid nitrogen-cooled mill. Tissues will contain both hydrophilic and lipophilic components; hence polar and non-polar solvents will be needed to ensure global coverage. Beckonert et al. [101] provide complete procedures for the extraction of metabolites from tissues.

The use of ragworms to monitor estuarine ecosystems has also been proposed, and of the various extraction methods analysed, the most appropriate were 50% aqueous methanol and chloroform, for the extraction of polar and non-polar metabolites, respectively [48].

It is possible to analyse solids by use of high-resolution magic angle spinning NMR, as has been applied to, for example, worms [21], meat [107] and liver and kidney [108].

2.3.3.8 Plant tissues

Various extraction strategies exist for the preparation of plant tissue. A protocol for the perchloric acid extraction of metabolites is provided by Kruger et al. [96] and a comparison of the most common solvents is given in [97]. The assessment of drought stress in pea leaf was performed by analysis of D_2O extracts of lyophilised leaves [11].

2.3.3.9 Honey

The high viscosity of honey requires novel ways to ensure that samples are representative and homogenised. To this extent, the distribution of a UV-absorbing dye has been monitored as a means to measure the homogenisation in a study that seeks to identify botanical biomarkers in Corsican honey [109]. Schievano et al. [110] prepared honey samples by first performing a biphasic extraction using water and chloroform. Analysis of the chloroform extracts was used in the determination of the botanical origin.

2.3.3.10 Oils

Preparation may be as simple as mixing with a non-polar solvent, such as deuterated chloroform [111, 112] although other studies quantify sterols [71] and adulterants [113] present within oil using a more involved approach.

2.3.3.11 Environmental samples

The preparation of water samples is generally more trivial than for other matrices, but may include lyophilisation (freeze-drying) to remove the water and permit resuspension in a suitable solvent. Turner et al. [72] extracted phosphorus-containing compounds from soil by shaking 2 g of dried and ground soil with 40 ml of NaOH and Na_2EDTA before heating to 22 °C for 16 hours. After centrifugation of these extracts, they were freeze-dried and the solid remnants were ground. These extracts were then resuspended prior to ^{31}P NMR analysis. For a specialised review of the application of NMR to environmental research, see that by Simpson et al. [41].

2.4 Hardware and experimental setup

The range of applications utilising NMR spectroscopy has expanded in recent years due to significant technological advances in the field. New study areas, such as metabolomics, have created a vast amount of collectivised data. Online databases, such as the Human Metabolome Database (HMDB) [114, 115] and the BioMagResBank (BMRB) [116], provide interested parties with a vast amount of reference data, encouraging collaboration and the pooling of results. Groups aimed at standardising the reporting process of spectral acquisitions have been set up [51] to encourage the inclusion of the valuable metadata alongside the data itself; thus, due care and attention should be warranted for spectra where pH and other considerations are not reported. The inclusion of such metadata, however, does not address the question of whether or not NMR spectra are reliable information sources when considering, for example, variability of instrumentation and applied technologies.

2.4.1 Availability of hardware

Technological developments have seen the recent commercialisation of 23.4 T (1 GHz) magnets and the aspiration for greater field strengths will continue to be supported by further technological advances. Often the benefits of working at a higher field, such as greater spectral resolution and sensitivity are offset against the practicalities and budgets of scientific laboratories. As higher field strengths also lead to longer relaxation times and thus, for quantitative analyses, longer experiment times, the offset between the sensitivity improvements that can be obtained by simply acquiring more scans at a lower field strength should also be considered. Field strengths may be the most visible sign of spectrometric potential, but major innovations also arrive through the design of the probes. In cryogenically cooled probes, the electronics are cooled to liquid nitrogen temperature, such that the reduction in electronic noise leads to enhanced signal-to-noise ratios.

A 2007 study by Bertram et al. [117] assesses the influence of NMR parameters on urine samples, covering the reproducibility of data collected at 250, 500, 600 and 800 MHz field strengths. Using partial least squares-discriminate analysis (PLS-DA) of control and treatment groups, they identify the same discriminating spectral regions, regardless of the field strength. However, the study shows that the predictive capacity of the PLS-DA model increases significantly between data collected at 250 and 500 MHz, whilst a further increase of field strength to 800 MHz is less pronounced. Thus, the machines with magnets of either 500 (11.7 T) or 600 (14.1 T) MHz, used in the majority of metabolomic studies, are well suited for such applications.

An inter-laboratory comparison by Viant et al. [118] involves seven laboratories in an environmental metabolomics study with the emphasis on laboratory practices. The study focusses on the preparation of samples, as well as data collection, processing and analysis. Operating procedures for each of these phases were provided based on variables derived from both the hardware and software available in the various laboratories. The results show good reproducibility, across the seven laboratories, for both synthetic and biological samples. However, the authors propound that future exercises could be more prescriptive in terms of post-acquisition processing and temperature, and that acquisition could be extended to include two-dimensional, more complex, experiments. A similar exercise performed in the field of plant metabolomics [119] gives much the same results. Such studies show that NMR is a highly reproducible technique, in spite of variability in the acquisition parameters, probe and field strength.

The common theme of experiments studying instrumental variability is the focus on the 1D ^{1}H experiment. This is presumably due to the technique's wide range of applications and its suitability for quantitative studies. To date, little has been done to gauge the reproducibility of two-dimensional methods.

2.4.2 Speed or sensitivity?

Noise is a common, and unavoidable problem, but the co-addition of additional transients can increase the signal-to-noise ratio (SNR) of a spectrum. As the SNR is proportional to the square root of the number of acquired scans, it ultimately determines the number of scans required to give an acceptable noise level for each sample; this in turn depends on time, cost, stability, probe design, magnet, nucleus, and the experiment being performed. High-throughput studies need to strike a balance between sensitivity and speed, whilst considering sample stability.

The type of post-acquisition analysis may also dictate the number of scans that are required. Non-targeted analysis [43, 44], the identification of markers and metabolomics studies inevitably require many additional scans to allow molecules close to the limit of detection to be identified. If, however, a study focuses on targeted detection, the number of scans may be determined simply by the responses of the target molecules in question and other components of a mixture can be ignored.

2.4.3 NMR experiments

One-dimensional techniques are the most sensitive experiments and can be simple and quick to perform, whilst multidimensional experiments have the potential to provide more detailed information, albeit usually at the expense of significantly more time. Advances that reduce overall acquisition time by improving detection sensitivity may eventually make two-dimensional techniques a standard approach for complex mixture analysis.

^{1}H and ^{13}C are the obvious candidate nuclei for NMR studies of complex biological mixtures, with one-dimensional and two-dimensional homo- and hetero- nuclear techniques being the most informative in this context. Other innovations such as the use of diffusion ordered spectroscopy (DOSY) are also applied to obtain additional information for the identification of molecules in solution whilst magic angle spinning (MAS) spectroscopy has revolutionised the analysis of solids.

Heteronuclear experiments involving nuclei other than ^{13}C and ^{1}H are rarely used for the routine screening of mixtures, but can be applied to good effect when specific compounds or compound groups are targeted. For example, ^{31}P NMR has been applied in environmental studies, such as the release of phosphorus-containing compounds from lake sediments [73, 74], in land management studies [40] and the impact of phosphorus to soil formation [72]. ^{31}P NMR has also been applied in the profiling of phospholipids for cerebral tumour diagnosis [69], in the identification of polyphosphate-containing metabolites [68], and in the analysis of pesticides and chemical warfare agents [49]. Other nuclei such as ^{15}N which are used to study biological molecules have limited applicability without the adoption of isotopic labelling strategies or further technological advances such as polarisation technologies.

2.4.4 Homonuclear NMR

2.4.4.1 ^{1}H

^{1}H NMR spectroscopy has a broad range of applications, as it is the nucleus with the highest receptivity. However, the small spectral window of around 15 ppm means that the ^{1}H NMR spectrum can be crowded when analysing mixtures of even a few components.

One of several 1D ^{1}H NMR pulse sequences is typically used to acquire data for metabolomic studies. These are: standard 1D ^{1}H; Carr-Purcell-Meiboom-Gill (CPMG); and proton-decoupled projected 1D spectra from 2D *J*-resolved spectroscopy (p-JRES). Each pulse sequence has its own advantages and therefore their use is study dependent. Standard 1D NMR sequences have the advantage that they require minimal optimisation, can be acquired rapidly and are relatively sensitive. Although CPMG pulse sequences take slightly longer to acquire, they enable T_2 spectral editing. Molecules with short T_2 relaxation times give rise to broad resonances and CPMG pulse sequences are used to remove the influence of broad resonances from high molecular weight macromolecules and molecules bound to them.

p-JRES spectra yield a proton decoupled spectrum, i.e. a spectrum containing only singlet resonances. Therefore, p-JRES spectra are particularly useful where there is significant overlap of multiplet resonances. As p-JRES spectra are acquired using a 2D pulse sequence, spectral acquisition may be longer than that used for the other pulse sequences.

The removal of residual solvent resonances is accomplished by the use of specific pulse sequences. Potts et al. [120] assess the influence of the methods: presaturation; NOESY-presaturation; WATERGATE; and WET. They conclude that each introduces its own features on the spectrum, to the extent that the samples within the control group cluster with respect to the solvent suppression method. However, effects within the treatment group far outweigh the minor anomalies introduced by solvent suppression methods. They advise against the use of WATERGATE pulse sequences due to poor phase properties and suggest that NOESY-presaturation offers the most robust performance with little user optimisation. Solvent suppression using WET is also highlighted as producing spectra with very flat baselines.

A study in which ^{1}H spectroscopy is used for the identification and verification of Corsican honey [8], reports that honey consists of at least 200 molecular species. Fig. 2.3 shows that the spectrum of the predominantly sugar-containing matrix is dominated by the signals from glucose and fructose. The dynamic range of 1D NMR spectra is such that, even in the presence of such dominant signals, resonances with significantly lower intensities can be observed and can be identified by multivariate analysis as biomarkers of Corsican honey.

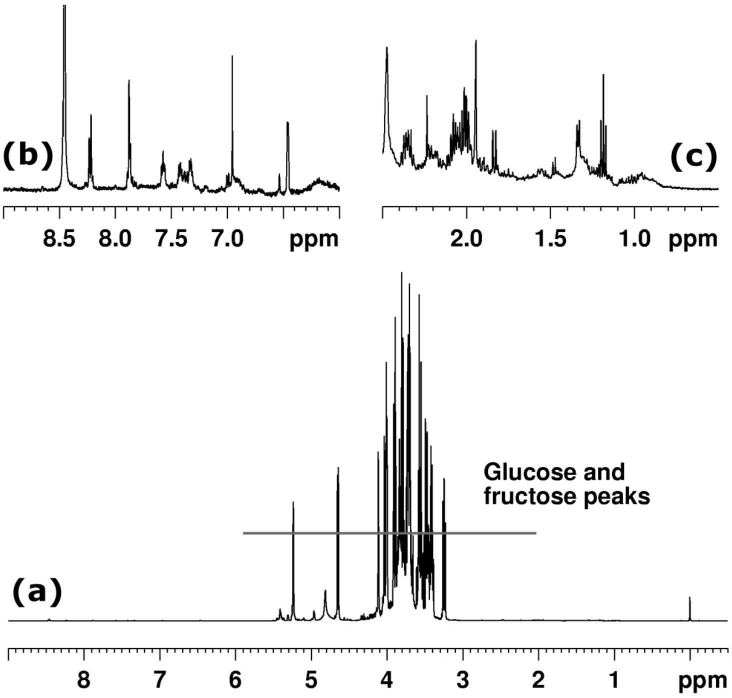


FIGURE 2.3: 1D ^{1}H NMR spectrum of Corsican honey recorded at 500 MHz. The whole spectrum is shown in (a), and (b) and (c) show expanded regions from (a). Any peaks above the horizontal line in (a) are attributable to either fructose or glucose.

^{1}H-^{1}H correlation spectroscopy (COSY) [121] experiments are structurally more informative and provide a visual link between coupled signals. The *J*-couplings between resonances are correlated and in a COSY spectrum the off-diagonal cross peaks link one coupling pattern to another.

Total correlation spectroscopy (TOCSY) [122] can be used to provide correlations between whole spin systems in a molecule. Although molecules may contain more than one spin-system, the technique has great application as it can be used to group spectral peaks as belonging to molecules, or parts thereof, allowing resonances to be assigned to different compounds in the same sample. Selective pulsing can also be used for mixture analysis; using a selective TOCSY sequence, irradiation of one spectral peak allows only the other resonances of that molecular spin-system to be observed. The complementarity of the techniques can be seen in Fig. 2.4 which contains both COSY and TOCSY spectra of the same sample. It can be seen that the resultant COSY and TOCSY spectra reveal different off-diagonal peaks, and this generates additional information that can be used for classification purposes.

Nuclear Overhauser effect spectroscopy (NOESY) experiments are two-dimensional techniques that correlate spins undergoing spin-lattice relaxation by means of through-space interactions. The relative intensities of the cross peaks can be used to calculate the distance between the two resonances. The use of this technique is perhaps more suited to the analysis of large molecules and is widely applied in protein and binding studies.

The two-dimensional ^{1}H *J*-resolved NMR technique [123] can be applied to separate *J*-couplings and chemical shifts and dramatically reduces the complexity of spectra. Fig. 2.5 shows both parts of a ^{1}H *J*-resolved spectrum, where Fig. 2.5a separates peaks as a function of their coupling constants, and the projection in Fig. 2.5b identifies the shifts of peaks without the complications caused by inter-resonance coupling. *J*-resolved NMR has been applied to the analysis of complex mixtures [124, 125], but is not common in online spectral databases. Ludwig and Viant [126] point out that in spite of potential quantification errors, *J*-resolved spectroscopy is theoretically field strength-independent, that it requires no manual processing, and that run-to-run variation between data sets is small in comparison to 1D ^{1}H experiments.

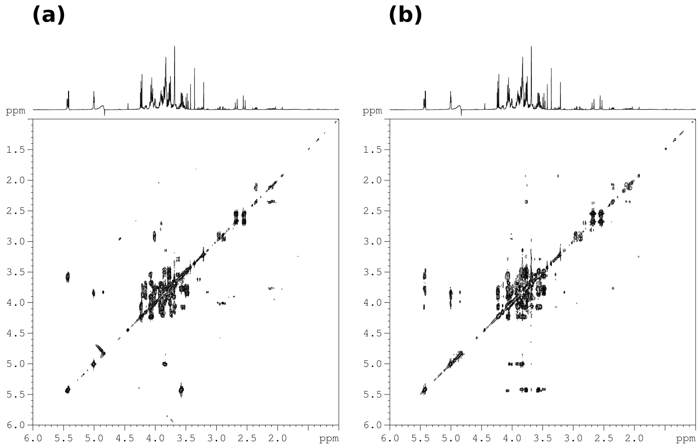


FIGURE 2.4: Two 2D ^{1}H homonuclear spectra of pea extracts recorded at 500 MHz, showing the spectral region between 1 and 6 ppm. (a) shows a correlation (COSY) spectrum with fewer correlations than the corresponding total correlation (TOCSY) spectrum shown in (b). The two techniques may be viewed as complementary to each other in terms of providing keys to molecular structure.

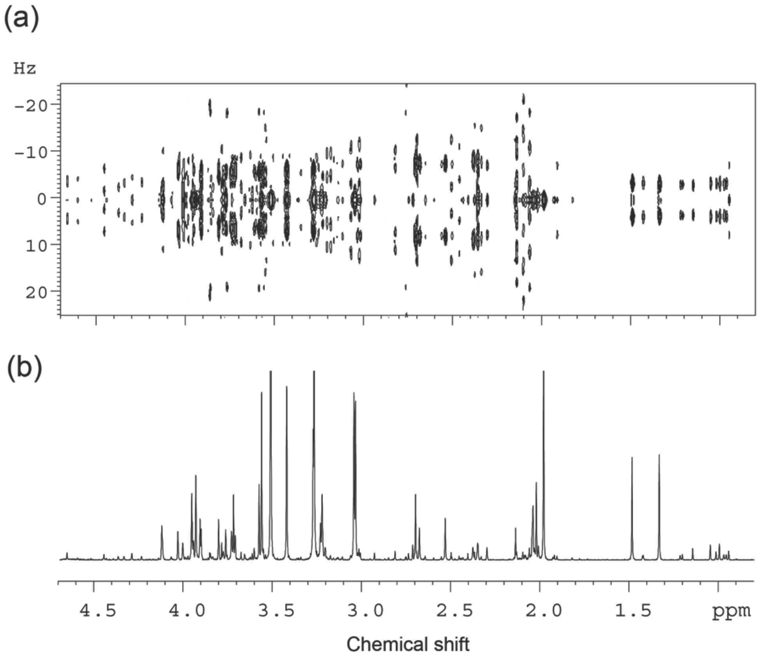


FIGURE 2.5: The *J*-resolved experiment can be useful removing splittings from couplings from resonances, as shown in a spectrum of fish embryo extract recorded at 500 MHz. (a) shows the fine structure of the resonances and (b) the projection of the spectrum. Reproduced with kind permission from Ref [126].

2.4.4.2 ^{13}C

^{13}C experiments suffer from the naturally low abundance of the nucleus, but the spectra are especially informative to those studying organic molecules. The low abundance of the nucleus precludes observation of any carbon-carbon coupling, as the probability of neighbouring ^{13}C nuclei is slim, hence ^{13}C-^{13}C correlation spectroscopy (e.g. INADEQUATE) in non-enriched samples is virtually impossible, particularly for mixtures containing low concentrations of some compounds. A standard ^{1}H-decoupled ^{13}C spectrum provides information about the backbone of the constituents, but the lack of coupling gives no clues about the construction of this backbone, except those inferred from their chemical shift. Distortionless enhancement through polarisation transfer (DEPT) experiments help provide some structural information, by differentiating between primary, secondary, tertiary and quaternary carbons.

2.4.4.3 Diffusion ordered spectroscopy

The application of a pulsed-field gradient (PFG) allows a mixture's constituents to be separated as a function of their translational diffusion coefficients, which are good approximations to molecular size. The method is similar to a chromatographic separation, and a DOSY [75] of a mixture has the potential to individually display the spectrum of each constituent. The reality, however, is that even simple mixtures do not provide instantly unambiguous results, and corrections may need to be implemented. The processing of DOSY data involves correcting the raw data to the experimental model as given by the Stejskal-Tanner equation [127]. This can be effected by use of software such as the DOSY Toolbox [128]. Fig. 2.6a shows an unprocessed DOSY spectrum, of a three component mixture. It can be seen that the separation afforded in the diffusion dimension is not optimal, due to the overlap and misplacement of resonances. Fig. 2.6b shows the DOSY spectrum after processing has been carried out, with the most dramatic changes being centred around 3.6ppm. It is clear that post-acquisition processing software, such as the DOSY Toolbox, can be used to dramatically increase the information quality of PFG experiments.

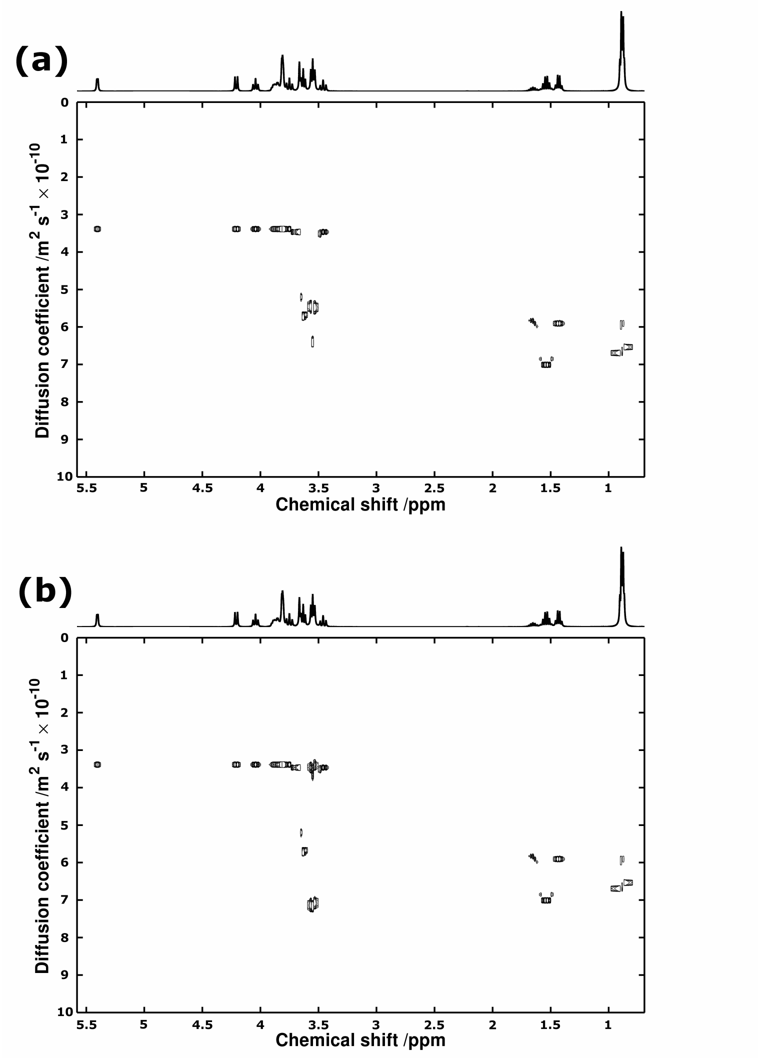


FIGURE 2.6: (a) shows the non-corrected diffusion ordered (DOSY) spectrum of a three component mixture recorded at 400 MHz, with ambiguous regions at approximately 3.6ppm. (b) has undergone processing, resulting in better-separated peaks in the diffusional dimension, especially around 3.6ppm. Reproduced with kind permission from Ref [128].

Pulsed field gradients have also been applied in conjunction with other NMR experiments, resulting in three-dimensional DOSY-TOCSY [129], DOSY-COSY [130], DOSY-HMQC [131], homodecoupled-DOSY [132, 133], *J*-resolved-DOSY [134] and HSQC-iDOSY [135].

2.4.5 Heteronuclear multi-dimensional techniques

Two-dimensional heteronuclear techniques generally involve the transfer of magnetisation from sensitive (e.g. hydrogen) nuclei to insensitive nuclei. This is most commonly applied as ^{1}H-^{13}C NMR experiments, but similar experiments involving coupling of hydrogen with nitrogen and phosphorous are also employed. Three-dimensional techniques are possible, and may be broadly categorised into two groups; the first involve the transfer of magnetisation across three nuclei, typically being ^{1}H-^{13}C-^{15}N for protein experiments, and the second involves the coupling of existing techniques, such as seen with the application of DOSY experiments to standard two-dimensional techniques.

2.4.5.1 HSQC

Heteronuclear single quantum coherence (HSQC) [76] is extensively used in the analysis of complex mixtures as a means of rapidly and simultaneously recording a range of ^{1}H and ^{13}C chemical shifts. The experiment couples, typically, ^{1}H nuclei with other NMR-active nuclei that are directly bonded to it, and uses magnetisation transfer to enhance the sensitivity of the secondary nucleus. HSQC is most commonly applied as a ^{1}H-^{13}C correlation experiment, but can also be used to determine single bond connectivity between ^{1}H and other nuclei such as ^{31}P and ^{15}N. HSQC spectra contain a great deal of structural information and thus with the correct supporting database can be used to unequivocally confirm the presence of a range of substances in complex mixtures. The ^{1}H-^{13}C HSQC is particularly useful and benefits from the chemical shift dispersion of the ^{13}C dimension providing excellent resolving power even in highly complex mixtures. Signals from nuclei that are directly bonded to protons are detected, i.e. chemical shifts of quaternary carbons are not represented in the spectra.

2.4.5.2 HMQC

Heteronuclear multiple quantum coherence (HMQC) experiments [136] yield similar resultant spectra to those obtained from HSQC. The experimental difference is that in HMQC the magnetisation of both nuclei evolves, which results in homonuclear proton *J*-coupling. This coupling results in peak broadening, and a subsequent loss in spectral resolution. Such broadening may be detrimental for the analysis of complex mixtures as, in general, larger line-widths make spectral interpretation more difficult.

2.4.5.3 HMBC

Heteronuclear multiple bond coherence (HMBC) [77, 78] may be considered as complementary to HSQC. The signals are shown for hydrogen-X nuclei interactions over two, three and possibly four bonds. Suppression of the HSQC-like H-X one-bond interaction is not always possible, and these may be visible as artefacts in HMBC spectra. The use of the technique for complex mixture analysis is low, in comparison to the more directly informative HSQC experiment.

2.4.6 Solid state and magic angle spinning

The application of high-resolution magic angle spinning (MAS) NMR spectroscopy has revolutionised the analysis of solid samples. The technique suppresses the line-broadening effects endemic to solids, which are caused by the anisotropy of chemical shifts and dipolar couplings. HR-MAS can be theoretically applied in one- and two- dimensional [137, 138] techniques, though at present the applications mostly focus on one-dimensional ^{1}H [21, 139–141], ^{13}C [142, 143], ^{19}F [144, 145] and ^{31}P [146] experiments.

2.4.7 Hyphenation

Coupling the NMR spectrometer to a chromatographic device by means of a flow-probe has the potential to further facilitate complex mixture analysis. In theory, a perfectly resolved liquid chromatographic (LC) separation has the potential to deliver each mixture component individually to the probe for spectral acquisition. In reality, perfectly resolved chromatograms of complex mixtures cannot be obtained. LC-NMR, similar in objective to DOSY-NMR, can be used to deliver chromatographically separated fractions of the sample to the magnet. The separation afforded by the column means that the spectra collected will display only parts of the mixture and hence be visually and computationally simpler to analyse. On-line acquisition of one-dimensional data is often compatible with respect to the LC timescale and sample volume. The acquisition of two-dimensional HSQC spectra requires the use of stopped-flow methodologies to collect the eluting analytes prior to analysis, yet the limited volume of sample that is compatible with LC columns may not immediately provide sufficient sensitivity for data acquisition. On-line acquisition of 2D spectra has been performed by Zhou et al. [147], whose study uses LC-NMR to record TOCSY spectra in real time. The technique, however, is presently limited by the fact that the method uses Hadamard encoding, which requires *a priori* knowledge of an analyte's spectrum. In general, the use of LC-NMR is also complicated by solvent resonances. These have to be removed, typically using either expensive deuterated solvents or by the incorporation of solvent suppression into a pulse sequence. LC-NMR has been applied to assist in the identification of metabolites [148, 149], the profiling of carotenoids in foods [150] and to analyse the anti-tumour properties of marine sponge [151].

Multiply hyphenated systems have also been proposed to link a range of analytical technologies together e.g. LC-NMR-MS [152]. These holistic approaches are able to provide vast amounts of data for structural identification, but the reality of these systems is that they are costly, for the most part impractical, and the added techniques introduce their own complexities to an already multifarious system. Furthermore, without adequate data management tools, spectra collected from such systems are liable to overwhelm.

2.4.8 Technology selection

It is recognised that not all researchers have access to the current state-of-the-art technology. Although gigahertz magnets and cryogenically cooled probes currently represent the pinnacle of spectrometer hardware, the analysis of complex mixtures does not always need the latest technology. The benefits of applied field strength are balanced by the longer acquisition times needed to acquire quantitative data. The use of 500 or 600 MHz machines has been highlighted [117] as more than adequate for the identification and quantification of components of complex mixtures. Indeed, much more information is deposited in public databases at these field strengths than for the higher field instrumentation.

LC-NMR is another possibility for the analysis of complex mixtures. Whilst multiple spectra provide a theoretical increase in a mixture's spectral resolution, this might be counteracted by computational and processing requirements. Current typical protocols involve the rapid collection of 1D ^{1}H spectra for a range of samples that describe natural variability and, in addition, a single or multiple two-dimensional spectrum of a representative sample. This is commonly a COSY, TOCSY or HSQC spectrum, and it is used to assist in the assignment of important spectral resonances to the responsible molecules. Currently, this strategy is generally the most time-efficient, but advances in reducing two-dimensional spectral acquisition times are set to make two-dimensional experiments the *de facto* method of choice.

2.5 Data pre-processing

Changes in experimental parameters such as temperature, pH and ionic strength lead to changes in the dynamics of the NMR measurement and therefore result in shifts in peak position. Instrumental parameters such as magnetic field homogeneity can cause changes in peak shape, and physicochemical parameters, such as exchange broadening and molecular mobility, also affect the distribution of NMR transition frequencies. Sources of variation must be controlled where possible and, for data sets consisting of multiple spectra, it is vital that all are processed in the same manner to ensure that additional variance is not introduced. Although artefacts may still remain, post-acquisition processing techniques can be applied in both the time-domain and the frequency-domain to limit the impact they impose on the analysis. For example, methods are available for baseline correction and the removal of noise and artefacts associated with solvent suppression.

2.5.1 The free induction decay

The application of various mathematical (window) functions to the free induction decay (FID) before it undergoes Fourier transformation (FT) can dramatically increase the quality of the spectrum obtained. However, there is generally a trade-off between the resolution of peaks and the signal-to-noise ratio (SNR).

Exponential (line broadening) functions are usually applied to the FID to effectively weight the decay and place more emphasis where the time-domain SNR is greatest. The function also forces the FID to decay to zero, which avoids the introduction of FT-induced artefacts. Although the SNR of the spectrum is increased, line broadening can introduce further problems, particularly in complex mixtures that will have many overlapping peaks. The reduction in resolution induced by line broadening may be sufficient to overwhelm peaks representing molecules at low concentrations. The trade-off between noise removal and loss of resolution must be considered; excessive alteration to improve one aspect of the FID before FT can lead to the deterioration of another. However, it has been proposed that the application of two functions, to weight the decay of both the real and imaginary parts of the FID, can result in simultaneous improvement of the SNR and the resolution [153].

2.5.2 Noise removal

Baseline noise results from the electronics and is unavoidable, although cryogenically cooled probes have helped to significantly reduce its contribution. As well as the impact of any processing of the FID on the noise in the spectrum, the SNR can be increased by the addition of extra scans. In theory, the combination of an infinite number of scans would make the mean noise level tend to zero. However, as the number of possible scans is obviously limited, noise can be reduced but not eliminated in this manner.

In fact, random noise has been shown to be a large source of variance between otherwise identical spectra [154]. Principal components analysis (PCA) performed on duplicate spectra of synthetic mixtures resulted in considerable separation solely as a result of the noise variables. The inter-spectrum variation in noise intensity was shown to be extremely small on an absolute scale, but dominating on a relative scale. A repetition of the analysis, excluding those regions identified as noise (being below a defined threshold), resulted in almost perfect clustering. The choice of threshold, that optimally excludes noise whilst retaining genuine information can fundamentally alter the variance between samples and requires judgement.

Although acquisition times for two-dimensional NMR have been greatly reduced, there are systematic noise problems associated with some techniques. Diffusion ordered spectroscopy is susceptible to inhomogeneities related to the maintenance of the local environment (for example, eddy currents will exacerbate molecular diffusion and field inhomogeneities will distort the experimental pulses that are applied to impart a diffusion gradient) and Huo et al. [155] have proposed a method for the diagnosis of such artefacts. The presence of t_1 noise artefacts in phase-cycled HSQC limits its use despite its superior sensitivity. This noise occurs as ridges parallel to the F_1 axis at the F_2 frequencies of intense peaks and has the potential to mask genuine peaks of low concentration compounds. Thus, the identification of genuine peaks is made considerably harder when using a threshold to pick peaks. However, the systematic nature of t_1 noise allows denoising techniques to be applied. The Correlated Trace Denoising (CTD) algorithm described by Poulding et al. [81] is an efficient way of denoising the spectra of complex mixtures and has been shown to significantly improve the SNR of small genuine peaks embedded within t_1 noise. Unlike other methods such as reference deconvolution [156–158] or the Cadzow procedure [159], that have specific pre-requisites, correlated trace denoising can be used regardless of complexity and the number of peaks in a spectrum, making it suitable for metabolomics studies. The application of CTD allows t_1 noise to be removed, and Fig. 2.7 shows that the signal-to-noise ratio of a benzoic acid peak is improved compared to that in the unprocessed spectrum.

2.5.3 Baseline correction

Both 1D and 2D NMR spectra can suffer from baseline distortions that can be much larger than either noise or small peaks. Definition of the spectral baseline is of paramount importance in the analysis of complex mixtures and, in quantitative analysis, these artefacts could be a major source of error. Many diverse causes of baseline distortion have been identified including the corruption of the first few data points of the time-domain signal [160], linear phase correction [161] and the application of filters to the FID [162]. Baseline correction can improve the quality of a spectrum and the accuracy of peak integration and most NMR processing software packages incorporate methods for baseline correction. As well as adjusting the acquisition parameters, oversampling and digital signal processing can be used to improve baselines [163]. Many methods for baseline correction have been proposed for both 1D and 2D spectra. These often involve the fitting of functions to regions in the spectrum identified as noise before their subsequent subtraction. Polynomial functions and cubic splines are often used for computational efficiency [164]. Whilst fluctuating noise is a baseline's dominating constituent, baselines also consist of the tails of intense peaks and the majority of the area of low intensity peaks. The procedure proposed by Chang et al. [165] uses a moving-average high pass filter to identify the signal, along with Lorentzian line-shape modelling to ensure that the filter does not destroy the Lorentzian-like tails of high intensity peaks. An alternative method proposed by Xi et al. [166] fits a curve to the lowest intensities of a spectrum and does not require the identification of data points as noise or signal. It is therefore suitable for application to the spectra of complex mixtures, where high signal density may preclude the identification of noise regions.

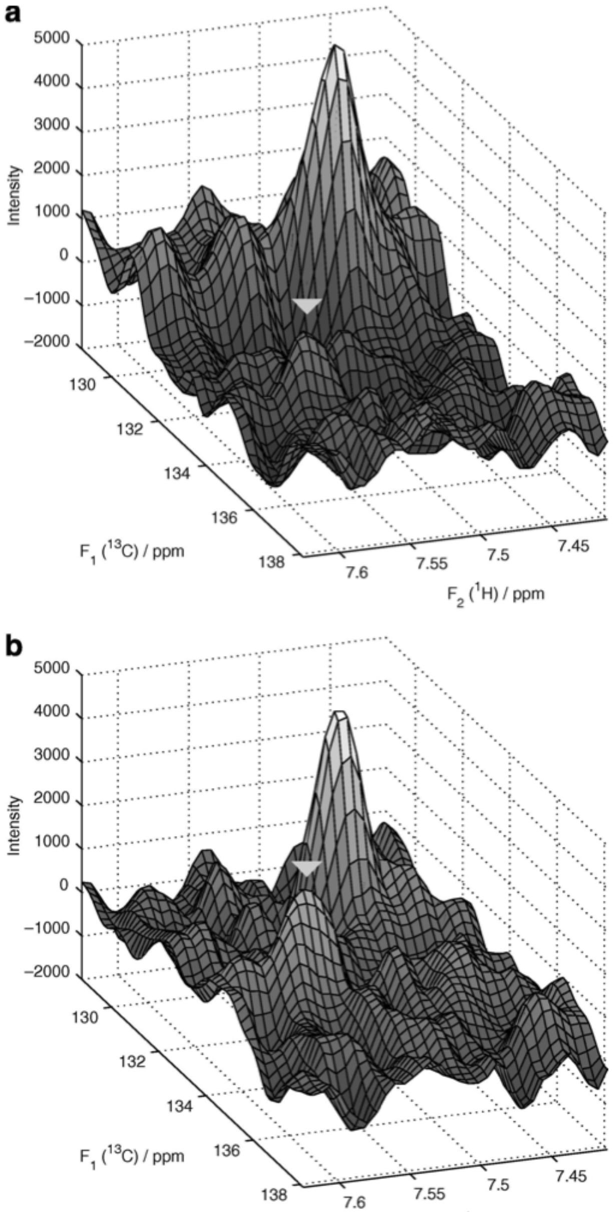


FIGURE 2.7: The application of Correlated Trace Denoising to a phase-cycled heteronuclear single quantum coherence (HSQC) spectrum, recorded at a $^{1}\mathrm{H}$ frequency of 500 MHz ($^{13}\mathrm{C}$ frequency of 126 MHz). In the original spectrum (a) the arrowed peak is of a noise-like intensity. After denoising, in (b) the same arrowed peak is considerably more pronounced.

2.5.4 Peak alignment

Chemical shift differences arising from different chemical structures are the fundamental reason for the use of NMR in the analysis of complex mixtures. However, changes in experimental parameters such as temperature, pH and ionic strength also lead to changes in the NMR measurements and therefore result in shifts in peak position. The temperature and pH of an NMR sample are tightly regulated, but the ionic strength of the NMR sample cannot be easily controlled. Thus, rigorous sample preparation procedures may prevent major changes in peak position, but irregularities between samples are still likely to occur. Any study that makes inferences based on inter-spectral differences requires comparison of the same variables from each spectrum. Alignment of peaks requires more than simply offsetting whole spectra as peak shifts may be uncorrelated. Pair-wise alignment methods require a target signal to be chosen as the standard to which the spectra are to be matched. Dynamic time warping (DTW) [167], originally used in speech recognition, and correlation optimised warping (COW), which uses the correlation coefficient as a measure of similarity between two signals [106], have been proposed as methods to align NMR spectra. Both methods are compared with a Baysian approach to the alignment of NMR spectra by Kim et al. [168]. Forshed et al. [169] have applied genetic algorithms to the alignment of NMR spectra in an iterative optimisation procedure that requires no prior processing or data reduction. Other algorithms that allow local alignment of sections of spectra include recursive segment-wise peak alignment (RSPA) [170] and the *i*coshift algorithm [171], which is most effective when the user manually selects the regions in need of alignment.

2.5.5 Feature extraction

Although, aligned data can be used directly in chemometric analyses, the large number of variables involved makes it computationally expensive. High resolution spectra can consist of tens of thousands of data points, making multivariate analysis inefficient. Furthermore, the information content can be improved by extracting only the relevant spectral features.

It is now common practice to accommodate small spectral shift changes by averaging data points over a range of frequencies. Uniform binning or bucketing involves integrating the spectral data over regions of equal length and the integrated values being used as variables in metabolomic studies. Typically, bins of length 0.04 ppm are used [53]. Although the method provides an easy means to extract features and account for minor peak misalignments, it suffers from several limitations. Using fixed-width bins can increase the variation in the dataset, as peak shifts occurring close to a bin borders result in the allocation of the same peak in different spectra to different bins. Bin ends often dissect NMR resonances and multiple peaks may be assigned to the same bin so that data interpretation can be difficult. An alternative method, termed adaptive binning, described by Davis et al. [54], overcomes the problems of fixed-width binning by assigning variable-width bins according to the peaks in a reference spectrum. The reference spectrum is obtained by taking the maximum value over all spectra in the dataset to be analysed at each data point. The resulting spectrum has a jagged appearance due to the shifts in peak maxima and requires smoothing before the minima are identified and used as bin ends. The smoothing is achieved using wavelet transforms to remove the small details from the reference spectrum. The level of smoothing depends on the resolution of the data and can be optimised to provide a reference spectrum in which the minima correspond to the starts and ends of peaks and can be used to define the bins. These bins are then applied to each spectrum in the dataset and the resulting values used as variables in further analysis. Thus the variables correspond to integrated peaks in the reference spectrum. The method also allows noise regions to be identified and excluded. Fig. 2.8 compares the effects of fixed-width and adaptive binning. Whilst fixed-width binning results in peak splitting, adaptive binning results in well-defined peak areas and excluded noise regions. The adaptive binning technique significantly reduces the intra-class variation with respect to the standard binning routine and facilitates interpretation as discriminating variables relate directly to peaks in the spectra.

An adaptive binning algorithm that does not require the creation of a reference spectrum has also been proposed. The AI-Binning [56] procedure recursively identifies new bins by subdivision of pre-existing bins. A single bin initially covers the entire

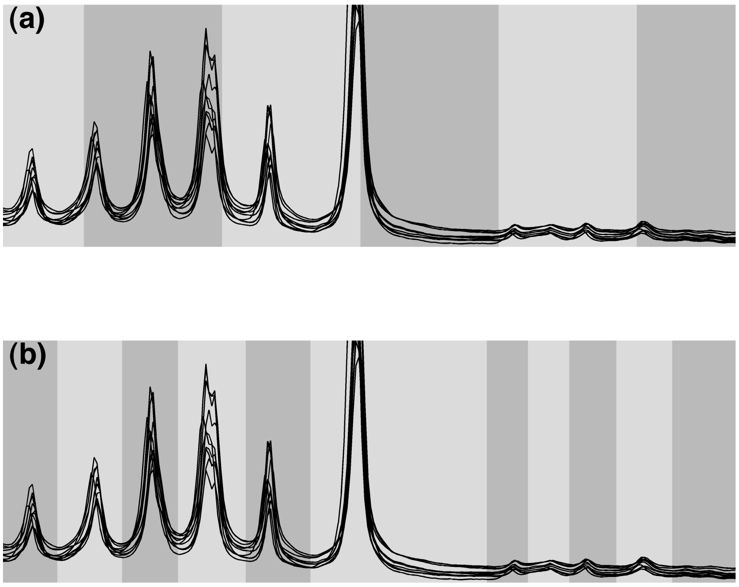


FIGURE 2.8: The application of (a) fixed-width and (b) adaptive binning showing overlaid spectra from multiple samples. Shaded regions in (a) represent bins of 0.04ppm, which can be seen to incorporate multiple peaks and to splice peaks into two bins. Each shaded region in (b) represents an individual peak.

spectral window and the algorithm proceeds to create two new bins by optimally dividing each existing bin. A metric is calculated to compare the quality of the potential new bins with that of the existing bin and the spectral noise. If the new bins improve the quality of the spectrum, they are accepted and the algorithm continues recursively testing new subdivisions until no more bins are accepted. The AI-Binning algorithm is shown to outperform fixed-width binning (0.04ppm) and use of the full resolution spectra in terms of predictive accuracy.

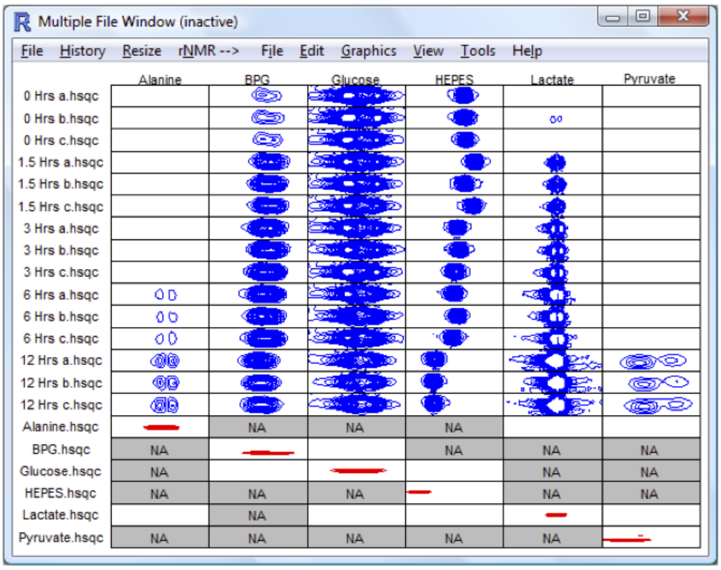
A binning technique that makes use of Gaussian distribution functions has been proposed [55] and avoids the problem of peak shifts close to bin boundaries by allowing bins to overlap. The technique does not impose fixed bin boundaries on spectra and uses a Gaussian function to weight the contributions of peaks in proportion to their distance from the bin centre. "Lorentzian Spectrum Reconstruction", proposed by Koh et al. [172], also considers overlapping peaks. The method deconvolutes an

NMR spectrum into a series of overlapping Lorentzian distribution functions and is therefore useful as a means of feature extraction and in facilitating spectral assignments.

The binning algorithms described provide feature extraction methods for one-dimensional spectroscopy, which at present dominates the analysis of complex mixtures. To date, few feature extraction methods have been developed for two-dimensional spectroscopy. Increased use of techniques, such as HSQC and HMBC, in the analysis of complex mixtures will depend on such algorithms as well as advances that reduce the acquisition times. At present, the major use of HSQC spectra is in the assignment of peaks, identified as being of interest from one-dimensional analyses. In an automated peak identification protocol proposed by Xi et al. [173], a database containing 21 HSQC metabolite spectra was used to identify peaks from both synthetic and complex biological spectra. The authors apply their method to the quantification of known metabolites although the potential for use in metabolomic studies is recognised.

The freely available multi-platform rNMR software [174] allows visualisation of multiple one- and two- dimensional spectra. Their region-of-interest (ROI) concept (see Fig. 2.9) is analogous to partitioning one-dimensional spectra into bins. Integrating within each ROI provides variables for multivariate analysis. The use of rectangular ROIs, however, limits the application of the programme to spectra that are well resolved and therefore cannot be used for feature extraction with complex two-dimensional spectra in metabolomic studies.

The method of Koh et al. [172] for the deconvolution of one-dimensional spectra using the Lorentzian distribution function, shown in Fig. 2.10a, can be extended to two dimensions. A feature extraction method for two-dimensional spectra that uses the Lorentzian-like properties of NMR peaks to bin ^{1}H-^{13}C HSQC spectra is described by McKenzie et al. [36]. A modified Lorentzian function is used to model peaks in a reference spectrum and provide footprints that can be used as elliptical bins. The one-dimensional modified Lorentzian function, in Fig. 2.10b, shows that this function does not suffer from the slow decay of the Lorentzian, which would not allow suitable footprints to be obtained for the peaks. The width of the modified function at three-eights of the original intensity is half that at the base. Thus, two-dimensional NMR spectral peaks can be modelled with only three parameters: the intensity and the width of the peak at three-eighths of this intensity for each dimension. Determining the widths at three-eighths of the intensity permits peaks that are not resolved at their base to be modelled. The technique is similar in application to the adaptive binning of Davis et al. [54] in that a reference spectrum is created from all spectra in the dataset and used to provide bins corresponding to the footprints of peaks. For each spectrum, the integrated values of each elliptical bin form the variables that can be used in metabolomic analyses. A threshold for the goodness of fit of the modified Lorentzian model excludes bins containing only spectral noise so that the set of variables is dramatically reduced in number and consists only of pertinent information. Peaks found to be of interest in multivariate analyses can be easily identified by cross-referencing with a database of known metabolites.



	Alanine	BPG	Glucose	HEPES	Lactate	Pyruvate
0 Hrs a.hsqc						
0 Hrs b.hsqc						
0 Hrs c.hsqc						
1.5 Hrs a.hsqc						
1.5 Hrs b.hsqc						
1.5 Hrs c.hsqc						
3 Hrs a.hsqc						
3 Hrs b.hsqc						
3 Hrs c.hsqc						
6 Hrs a.hsqc						
6 Hrs b.hsqc						
6 Hrs c.hsqc						
12 Hrs a.hsqc						
12 Hrs b.hsqc						
12 Hrs c.hsqc						
Alanine.hsqc		NA	NA	NA		
BPG.hsqc	NA			NA	NA	NA
Glucose.hsqc	NA				NA	NA
HEPES.hsqc	NA	NA	NA		NA	NA
Lactate.hsqc		NA				
Pyruvate.hsqc	NA	NA	NA	NA	NA	

FIGURE 2.9: A screenshot from rNMR, highlighting its Region of Interest (ROI) concept. Here, 6 ROIs are shown for multiple samples (in blue), and the red peaks represent standard spectra which corroborate the ROI assignment. Not Applicable (NA) is shown for chemical shifts greater than the recorded spectral width. Reproduced with kind permission from Ref [174].

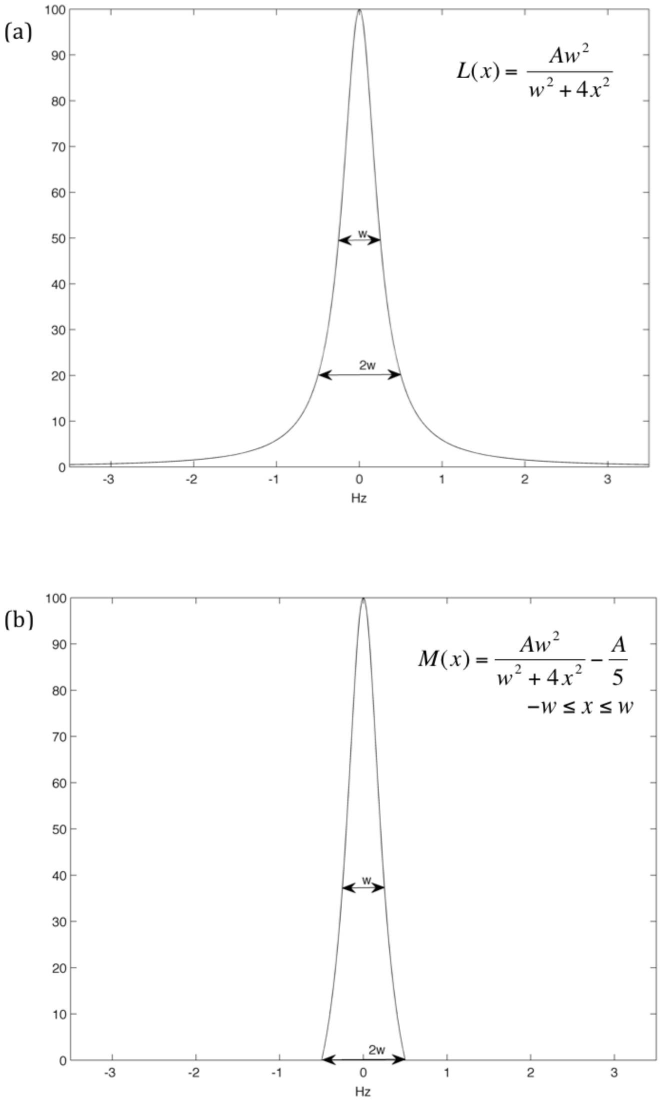


FIGURE 2.10: (a) shows a Lorentzian function with amplitude, A, of 100.0 and a width at half height, w, of 0.5 Hz. The corresponding modified Lorentzian is shown in (b). Here $I = 100.0$, so that $A = \frac{5}{4}I = 125.0$ and the width, w, of 0.5 Hz occurs at $\frac{A}{2} - \frac{A}{5} = 37.5$.

2.6 Data analysis

The analysis of complex mixtures requires the use of applied statistical techniques in order to extract meaningful information. Both univariate and multivariate methods may be used and the aims of the study influence the approach taken. By way of illustrative example, metabolic profiling usually focuses on a specific metabolic pathway or particular class of metabolites in a targeted, often quantitative approach, metabolomic fingerprinting involves exploratory data analysis to compare the patterns of metabolites in a hypothesis-generating approach.

2.6.1 Univariate analysis

Univariate methods are used to test hypotheses about a single variable. For multivariate data sets they can be used to analyse one variable at a time. The t-test can be used to test the significance of the difference between the means of two groups consisting of n_1 and n_2 samples. A test statistic is calculated as the difference between the two means, divided by their pooled variance and a p-value evaluated using a t distribution on $(n_1 + n_2) - 2$ degrees of freedom. For g groups, $g(g-1)/2$ pairwise t-tests must be calculated. It is assumed that the samples within each group are normally distributed and that the two groups have approximately equal variances.

ANalysis Of VAriance (ANOVA) assesses the effects and interactions of factors and can be used to test for a difference in means between more than two groups. The measure of separation is given by the *F*-statistic

$$F = \frac{V_b}{V_w} \tag{2.1}$$

where V_b is the between group variance and V_w is the within group variance.

Like t-tests, ANOVA relies on the assumption that the data are normally distributed within groups and that the variances are roughly equal. Although the tests are quite robust to skewness, outliers can seriously invalidate the results.

Whilst ANOVA can highlight a difference between groups, it does not identify the group or groups responsible for the difference. A combination of ANOVA and t-tests is therefore a good idea. The tests have been used as a method of feature selection in a number of metabolomics studies. Wu and Massart used the F-statistic to reduce the data set before using PCA-ANN [175], and found that the reduced data gave improved results. Similarly, Taylor et al. [26] used only the variables that changed most between groups in a genetic programming routine.

Alternative non-parametric tests can be used when the assumptions of normality or equal variances are violated. In the Mann-Whitney test, the data from the two groups are pooled and ranked according to size, but the group from which each sample came is tracked. The ranks are then summed for each group. Very large (or very small) values for the sum of one group suggest that the variables for the samples in that group are frequently greater (or smaller) than those in the other and therefore that the null hypothesis of no difference in means should be rejected.

2.6.2 Multivariate analysis

Multivariate techniques may be supervised or unsupervised. Unsupervised learning algorithms respond only to the input data presented in an $m \times n$ matrix, $\mathbf{X}$, where m is the number of observations, and n the number of variables. Supervised methods require an additional $m \times p$ matrix, $\mathbf{Y}$, where p is the number of outputs or responses, and learn to associate input variables with a particular output using training data. The response matrix, $\mathbf{Y}$, may contain quantitative data, for example assessments of biological activity, such that a relationship

$$\mathbf{Y} = \mathbf{aX} \tag{2.2}$$

for some vector, $\mathbf{a}$, can be formed. For classification, the output is the class of the observation and the training data consists of m observations of known class. In this case, the $m \times p$ matrix, $\mathbf{Y}$, is a dummy matrix where p is the number of distinct classes and

$$y_{jk} = \begin{cases} 1 & \text{if } k \text{ represents the class of observation } j \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Unsupervised techniques make no *a priori* assumptions about the observations and seek to highlight the structure and organisation inherent to a data set. All available data can therefore be included in the analysis. Supervised methods on the other hand are trained to associate input variables with a particular response and can easily be over-trained, particularly when the number of variables is much greater than the number of observations as is usually the case in metabolomic studies. It is therefore vital to retain an independent test set consisting of observations that have never been used during training in order to ensure that the model generalises and predicts the response for observations other than those with which it has been trained.

2.6.2.1 Principal component analysis

Principal component analysis (PCA) is one of the most widely used multivariate analysis techniques. The aim of PCA is to reduce the data to a few characteristic dimensions for visualisation and analysis. This is achieved by calculating a new set of variables, or principal components, each of which is a linear combination of the original variables. The principal components (PCs) are chosen so that the important information in the data is retained in just a few of these new variables, effectively summarising the samples or observations. The first PC is the linear combination that accounts for the maximum variation in the data and provides a one-dimensional approximation to the data. Better approximations are obtained by using more PCs, where each successive PC is uncorrelated with the previous PCs and expresses as much of the remaining variance as possible. Mathematically, this is achieved from the eigenvalue decomposition of the data covariance matrix with the coefficients for the *k*th principal component determined by the eigenvector of the covariance matrix corresponding to the *k*th largest eigenvalue. Data reduction is achieved by only keeping the first few PCs, which contain most of the information in the data. For k PCs we have

$$\mathbf{X} = \mathbf{P}\mathbf{A}^{\mathbf{T}} + \epsilon \tag{2.4}$$

where $\mathbf{X}$ is the $m \times n$ matrix of original variables, $\mathbf{P}$ is the $n \times k$ matrix of new variables or PCs, with $k \leq m$, $\mathbf{A}^{\mathbf{T}}$ is the $k \times n$ matrix of coefficients relating the original variables to the k principal components and ϵ is the residual matrix (with $\epsilon = 0$ if $k = n$). The values of the new variables for each observation, in the matrix $\mathbf{P}$, are known as component scores, or simply scores, and scores plots for just the first two PCs are often used to visualise patterns in the data. Graphical representation allows the structure and groupings inherent in the data to be visualised and potential outliers identified. The elements of the matrix $\mathbf{A}$ are known as the loadings and show how much each original variable contributes to each PC. Inspection of the loadings allows the original variables responsible for any patterns or trends to be identified.

As an unsupervised method, PCA is used in exploratory data analysis but the PCs obtained can also be used in further analysis. The data reduction allows techniques that cannot be used with the original variables due to the high dimensionality or correlation between variables to be applied to the PCs. When combined with discriminant analysis or regression, PCs can also be used as part of a supervised method.

2.6.2.2 Discriminant analysis

The objective of linear discriminant analysis (LDA) is to find the linear combination of original variables that maximises the difference between classes as measured by the F-statistic given in Eq. (1). This linear combination is the first discriminant function (DF) or canonical variable (CV). The separation between groups can be improved by using additional linear combinations, where the *k*th DF is chosen to separate the data in a manner that has not already been exploited. Generally, the covariance between DF_{k} and $\mathrm{DF}_{\mathrm{k-1}}, \ldots, \mathrm{DF}_1$ is zero.

LDA assumes normally distributed data and that the observation groups exhibit similar variances and covariances. Various types of discriminant analysis have been developed for cases where these assumptions are invalid. Generalised discriminant analysis algorithms, such as quadratic discriminant analysis (QDA), regularised discriminant analysis (RDA) and uncorrelated linear discriminant analysis (ULDA) are compared in Park and Park [176].

LDA has been applied to the analysis of ^{1}H NMR spectra of olive oils [111, 177], but is perhaps more commonly used in conjunction with other techniques, such as partial least squares (PLS-DA) [8, 110, 117, 178, 179] and principal component analysis (PCA-DA) [180].

2.6.2.3 Partial least squares

The most widely used supervised method in chemometrics is partial least squares regression (PLSR), based on Herman Wold's original non-linear partial least squares (NIPALS) algorithm [181]. PLSR applies regression analysis to a multivariate system to find components, or latent variables, that relate the variance in the input data matrix $\mathbf{X}$ to a response matrix $\mathbf{Y}$, whilst also modelling their structures. In fact PLS effectively performs PCA on the $\mathbf{X}$ and $\mathbf{Y}$ matrices simultaneously in an iterative procedure that retains the relationship between them.

A detailed review of PLSR and its use in chemometrics is given in Wold et al. [17]. PLSR is often applied in discriminant analysis mode (PLS-DA) [110, 111, 117, 178, 179, 182–186]. More recently, methods to remove the information unrelated to the $\mathbf{Y}$ variables, such as orthogonal signal correction [187] and orthogonal partial least squares (O-PLS [188] and O2-PLS [189]) have been used in chemometrics studies. Fig. 2.11 shows graphical representations of PLS and O-PLS, as applied to a two-group system. Separation of the two groups is achieved in both PLS and O-PLS models, yet the means of separation are different. In the PLS model (Fig. 2.11a, [190]) the between-group variance is spread throughout the new latent variables ($t1$ and $t2$). However, in the O-PLS model (Fig. 2.11b, [190]) the between-group variance is solely represented by the $t1_p$ latent variable, and the $t2_o$ variable expresses the intra-group variance.

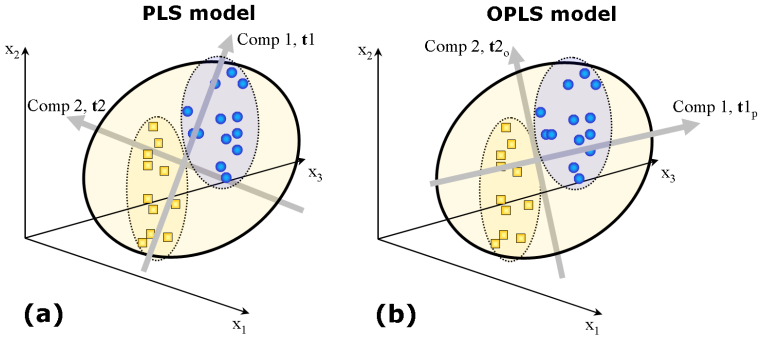


FIGURE 2.11: The partial least squares (PLS) model (a) can be used to separate classes, yet the variation between the two classes constitutes both PLS components. Similarly, the orthogonal partial least squares (O-PLS) model (b) separates the two classes but the first component represents inter-group variance. Reproduced with kind permission from Ref [190].

2.6.2.4 Evolutionary computing

Genetic algorithms (GA) [191] and genetic programming (GP) [25] are computational learning techniques that are based on the "survival of the fittest" principle. An initial population of random solutions to a problem evolves via rules that mimic reproduction and mutation until the best solution, in terms of some fitness function that depends on the type of problem, is achieved. In the case of classification, the fitness function will be related to the number of correctly identified samples. These techniques are gaining popularity in chemometrics, not least because they do not involve the creation of new variables as with other multivariate methods, making the results easier to interpret. Classification and feature selection are becoming common applications for GA and GP using the data at its full resolution or in a reduced or binned form [192].

In genetic algorithms, each potential solution is encoded as a symbolic string, or chromosome. The representation depends on the problem to be solved but may be a binary string consisting of ones and zeros or a string of real numbers. At each generation, the solutions are evaluated and a selection procedure used to identify those suitable for reproduction, which is achieved through an operation known as "crossover". Sections of two suitable parent chromosomes are then interchanged to produce new offspring and potentially obtain better solutions by combination of the parents' attributes. The mutation operator, which randomly changes some part of an offspring's chromosome, maintains genetic diversity and prevents premature convergence. The procedure of creation and evaluation of successive generations is repeated until a pre-specified fitness criterion is met or a maximum number of generations is reached. Genetic diversity can also be maintained by the formation of islands, known as demetic populations [193]. Solutions are generated within the island populations and combined every few generations by replacing the $n\%$ least fit solutions of one sub-population with the best $n\%$ of another.

Genetic programming (GP) is conceptually similar to GA, but rather than representing the solutions as symbolic strings, the solutions in GP are programs comprised of data variables and mathematical/logical operators. Fig. 2.12a and Fig. 2.12b shows how the relationship between variables and operators can be represented in a tree structure and highlights the interpretability of a GP approach, in which the importance of the original spectral variables is explicit. The initial population consists of trees built from randomly selected functions and variables, which evolve through the generations by the processes of mutation and crossover as demonstrated in Fig. 2.12c and Fig. 2.12d. Allowing trees to expand without control, a phenomenon known as bloating, could lead to over-fitting the training data. To prevent this, a maximum depth for the trees is allowed, which also limits the number of variables that can be selected.

The number of variables involved in most chemometric analyses result in an extremely large search space and the two-stage GP of Davis et al. [27] was designed for use with ^{1}H NMR datasets. The lower population strategy employed requires an increased mutation rate to prevent over-fitting. Computational efficiency is also significantly improved by limiting the number of generations in the first stage. This narrows down the search space by submitting only the most discriminatory variables to the second stage, in which the optimal classification solution is sought. Comparison showed the two-stage GP outperformed the traditional one-stage approach in both convergence and classification rates.

A similar two-stage GP achieved a high overall classification rate when applied to the geographical classification of honey, using ^{1}H NMR spectra [8]. However, the best results for discrimination between Corsican and non-Corsican honey samples were obtained from the novel application of PLS-GP, in which only those variables having significant variable importance of projection (VIP) scores in the PLS-DA analysis are used in the GP. A tree involving just 13 variables achieved a classification rate of 97.3%.

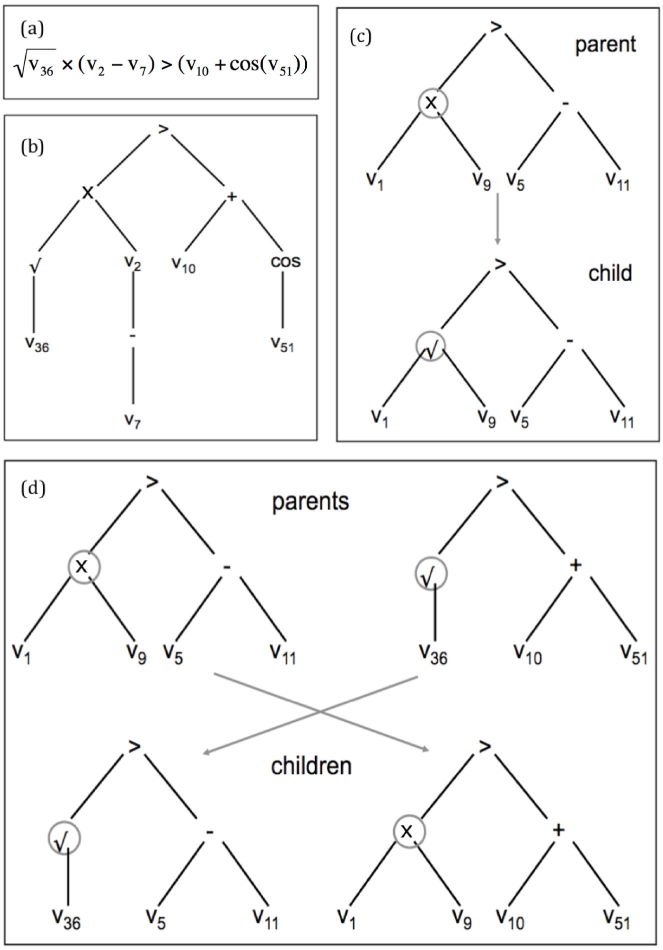


FIGURE 2.12: (a) is a mathematical representation of a genetic programming (GP) solution, of which (b) is its graphical form. (c) represents a mutation of a parent solution by alteration of the encircled nodal operator. (d) shows an example of crossover where two selected solutions are formed into two new programs. The circles highlight the crossover nodes.

2.6.2.5 Kernel methods

Kernel methods (KMs) allow non-linearly separable classes to be dealt with by applying a non-linear transformation of the input space to a higher dimensional feature space in which the classes are linearly separable. As the non-linear mappings or kernels allow computations to be performed in the input space, KMs are not computationally expensive. Kernel functions, commonly used to convert linear learning algorithms into non-linear methods, include polynomials and radial basis functions. The best known KMs are undoubtedly support vector machines (SVMs), but a wide range of algorithms have been used as KMs including K-PCA [194], K-LDA [195], K-PLS [196, 197] and K-OPLS [198]. Kernel-PCA can be used for data reduction and is computationally more efficient than standard PCA when there are significantly more variables than observations. K-OPLS has been applied to NMR data sets in metabolomics [199] and the freely available software can be used when linear methods are not sufficiently discriminating.

2.6.2.6 Artificial neural networks

Artificial neural networks (ANNs) are learning algorithms inspired by the processes in the brain. A network of processing units, or neurons, are trained in an iterative updating procedure. Self-organising maps (SOMs) [200] are the most common unsupervised ANN, which allow visualisation of multidimensional data, in a two-dimensional map. Each neuron, or node in the map, is a vector of length n, where n is the number of variables, representing typical input data. In each training cycle every sample, v say, in the training data set is presented to the network in turn and the elements (weights) of the node vector, w, identified as most similar are updated as

$$w_i = \alpha w_i + (1 - \alpha) v_i \tag{2.5}$$

for $i = 1, \cdots, n$. Physically close neurons are trained to recognise similar input patterns by updating not only the “winning" neuron but also those within a specified neighbourhood. SOMs can also be used as supervised methods for classification by assigning the class of the most similar vector in the training set to a node. In a semi-supervised mode [201], the classification data can be modified by weighting to prevent over-fitting and reduce the importance placed on the class assignment.

Other supervised networks are more complicated and involve hidden layers of neurons as well as an input layer and an output layer. Again, the networks learn from training set examples in an iterative process to change the weights of each neuron in order to minimise the difference between the actual output and the required output. The use of ANNs for the interpretation of spectroscopic data is much less common than other techniques such as PCA and PLS-DA [202]. One of the main problems with neural networks is that, whilst they may provide good results, they are difficult to interpret in terms of the original variables. However, ANNs have been used in the detection of adulterated olive oils [203], ^{13}C-based structure determination [204, 205], evaluation of sepsis in rats [206], differentiation of wines [207], and differentiation of the various stages of Parkinson's disease [185].

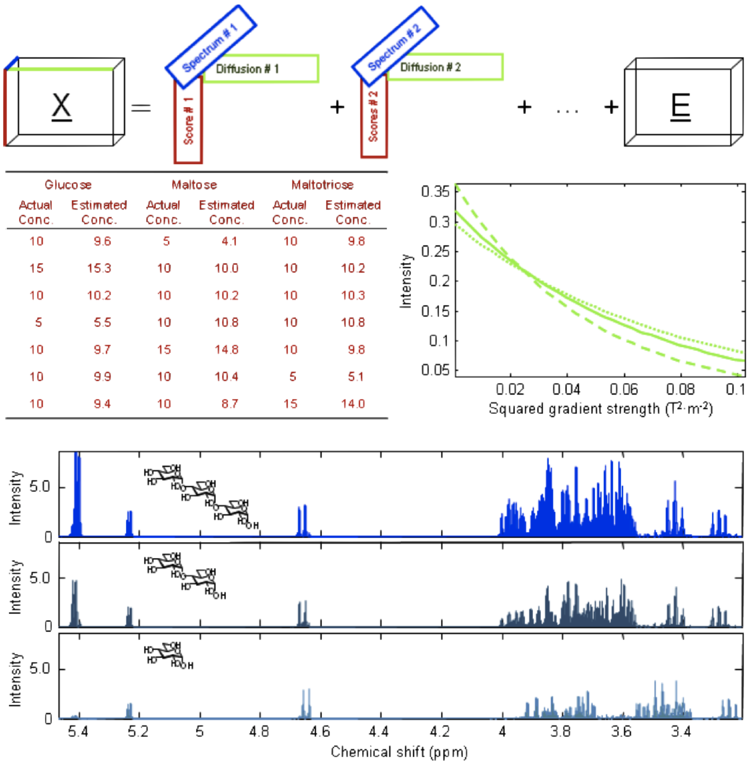
2.6.2.7 Multi-way decompositions

The extension of ANOVA to more than one variable is a technique known as multivariate-ANOVA (MANOVA). MANOVA is used to test for mean differences between groups based on linear combinations of the variables rather than individual variables. Like ANOVA it relies on the assumption of normality and, in this case, homogeneity of the covariance matrix is also required. Power in MANOVA depends on the relationship between the variables and the high degree of correlation between resonances means that MANOVA may not be best suited for NMR spectral interpretation. However, Smilde et al. [208] have shown that ANOVA-simultaneous components analysis (ANOVA-SCA or ASCA) can be applied to 1D ^{1}H NMR spectra, allowing the variation in the data to be assigned to the factors inherent in the experimental design.

Other multi-way decompositions include parallel factor analysis (PARAFAC [209, 210], known also as canonical decomposition, CANDECOMP [211]) and Tucker models. PARAFAC has been applied in metabolomics to separate differing time-course profiles and identify the variables responsible [212]. Smilde et al. [213] provide a tutorial into the analysis of temporal data. The combination of ASCA and PARAFAC, termed PARAFASCA by Jansen et al. [214], provides a model that can be interpreted in terms of experimental factors, such as groups, subjects, variables and time. Recently, PARAFAC has been applied to the analysis of poorly resolved NMR spectra in an approach termed "mathematical chromatography" [215]. The method was applied to a series of diffusion-edited 2D DOSY spectra of mixtures of glucose, maltose and maltotriose. The three sugar molecules have similar diffusion coefficients and their ^{1}H chemical shifts are tightly clustered, so that peaks in the spectra are not resolved. Fig. 2.13 shows the three-dimensional data array, composed of spectra collected as functions of intensity and gradient strength, together with its subsequent decomposition into constituent parts. The ability to identify and resolve individual components in highly overlapping spectra offers great promise for the analysis of complex mixtures.

2.6.2.8 Decision trees

Decision trees are computational learning algorithms, utilising a series of conditional if statements, that have been applied in classification and regression analyses (see, for example, Breiman et al. [216]). Every node in a decision tree tests the value of some classification variable, or combination of variables, and branches according to the possible values. The variables for a particular observation are sorted through the tree from a root node, down the appropriate branches to a leaf node that defines the class of the vector. At each branching point the choice of branch is determined by the response of the variable(s) to the condition at that node. The results of the technique are easily interpreted due to the diagrammatic nature of the solution and the fact that the original variables are involved. However, the construction of trees needs to be controlled to prevent the algorithm from bloating in the pursuit of marginally increased classification. One solution involves a pruning procedure



Glucose		Maltose		Maltotriose	
Actual Conc.	Estimated Conc.	Actual Conc.	Estimated Conc.	Actual Conc.	Estimated Conc.
10	9.6	5	4.1	10	9.8
15	15.3	10	10.0	10	10.2
10	10.2	10	10.2	10	10.3
5	5.5	10	10.8	10	10.8
10	9.7	15	14.8	10	9.8
10	9.9	10	10.4	5	5.1
10	9.4	10	8.7	15	14.0

FIGURE 2.13: The data matrix, $\mathbf{X}$, is decomposed by parallel factor analysis (PARAFAC) into factors and a matrix of residual values, E. Red sections represent individual spectra, each of which represents a mixture of the three sugars at varying concentrations, as defined in the table. The green portions highlight the applied gradient strength, with the graph depicting the decay of signal intensity as the gradient strength increases. The bottom section, in blue, details the resolved, individual spectra of each of the three constituents. Reproduced with kind permission from Ref [215].

in which the tree branches corresponding to the smallest improvements in classification are removed. The technique has been successfully applied to NMR data in various classification studies [9, 113, 217, 218]. As with any other supervised learning algorithm, external cross-validation is necessary to ensure that the decision tree will generalise to unseen data. Random forests [219] provide an extension of the technique that involves a collection of differently constructed trees, from which an overall result is obtained by common occurrence.

2.6.3 Assignment

Multivariate analysis may provide profiles indicative of certain states or conditions with which further samples can be classified. Whereas MS-based techniques rely, to a large extent, on targeted analyses of pre-defined compounds, NMR offers a global fingerprinting technique that can identify almost every molecule in a complex mixture. In food safety and authenticity studies, a valuable strategy is the comparison of samples with what is considered "normal". Quality control can be maintained via a database of acceptable NMR spectra, allowing abnormalities to be detected without having first to be identified as deleterious.

Identification of the variables, i.e. peaks in the spectra, responsible for differences between groups, such as those indicated by loadings or VIP scores, can identify individual biomarkers. Care must be taken to ensure that such biomarkers are related to the condition of interest. For example, in toxicity studies, peaks related to the dosing compound rather than the response must be recognised as such and not included as biomarkers for toxic stress. In order to understand the biochemical processes underlying the differences, potential biomarkers must be related back to the compounds from which they have arisen. Spectra from 1D ^{1}H NMR experiments are often crowded making peak selection more challenging and the selected data may not include all peaks associated with a biomarker. A database search may only provide partial assignments that would need to be complemented by further analysis. Alternatively, entire spectra can be submitted for assignment and software packages to analyse and assign peaks in one-dimensional spectra are available. The increased resolution allows HSQC and TOCSY spectra to be used in a complementary role to one-dimensional spectra and a single two-dimensional spectrum is often recorded to aid assignment. The interpretation of 2D spectra is supported by software such as MetaboMiner [80], which allows assignments of ^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC spectral peaks via a graphical user interface. Peak coordinates and intensities are imported and the program searches three metabolite databases (HMDB, BMRB and MRMD) to provide assignments. Similar software packages include MetaboAnalyst [220] and the COLMAR web server suite [221]. However, the use of on-line databases requires some caution as experimental conditions and acquisition parameters must be considered, particularly in the absence of widely-accepted “best practice" guidelines.

2.7 Future directions

The analysis of complex mixtures has traditionally been monopolised by chromatography techniques hyphenated with mass spectrometry. Although less sensitive, NMR spectroscopy is a highly reproducible technique that is now used extensively in metabolomic studies. Advances in both hardware, such as magnets and probes, and software, in terms of spectral processing and interpretation, have resulted in the much wider application of NMR spectroscopy. The mechanism of data acquisition is also being refined and the sensitivity of the method improved. Hyperpolarisation techniques, among them dynamic nuclear polarisation (DNP) and para-hydrogen enhancement, promise to drastically reduce acquisition times and vastly improve sensitivity.

Methods are also being developed to reduce the long acquisition times associated with the more informative multidimensional experiments, traditionally acquired by performing a series of experiments. A spatially encoded paradigm, proposed by Frydman et al. [222] divides a sample into multiple slices, each assigned a particular evolution time. The method allows the simultaneous acquisition of these slices so that multidimensional experiments can be performed using a single scan. A detailed review of the progress in single-scan spectroscopy is provided by Tal and Frydman [223].

Although the present trend in spectrometer design is the commercialisation of stronger and larger magnets, the development of low-field spectroscopy has the potential to both reduce the size of the instrumentation and, perhaps more importantly, increase the applicability of the technique. In particular, there is a large demand for lower cost non-invasive technologies such as NMR spectroscopy for application in manufacturing environments.The application of super-conducting quantum interference devices (SQUIDs) to NMR spectroscopy, coupled with prepolarisation, has allowed for the collection of spectra in microtesla-strength magnetic fields [224, 225]. SQUIDs are typically low-temperature super-conductors and require cooling to liquid helium temperatures. Although currently less sensitive, high temperature SQUIDs are now capable of functioning at liquid nitrogen temperatures.

Perhaps a more distant prospect is the routine utilisation of NMR technologies for high resolution three dimensional molecular imaging of intact systems [226]. To date molecular imaging has largely been focussed on clinical diagnosis in the biomedical area, but as technologies develop could have a major role to play providing solutions in areas as diverse as the detection of terrorist activities, to routine quality assurance monitoring in the manufacturing sector.

Reduction in acquisition times and increased sensitivity will ensure the routine use of NMR spectroscopy, but next-generation technologies require concomitant advances in software design. Without efficient data manipulation, analysis and interpretation, the huge increase in data quantity will not lead to a similar increase in information and knowledge. A major challenge will involve the combination of enhanced computer processing power with the rationalised logic inherent to humans.

The integration of different technologies forms the basis of systems biology and bioinformatics tools to combine data from multiple experimental sources are being developed. Rantalainen et al. [227] exploit the relationship between metabolite concentrations and protein abundances to model disease status and increase the chances of new biomarker discovery. Mass spectrometry is highly sensitive and, when used together with chromatography, allows the detection and quantification of low-level compounds. However, the technique is targeted to some extent due to column choice and mode of detection. Whilst sensitivity can be a problem in NMR spectroscopy, the technique allows all compounds above the limit of detection to be analysed. The techniques are therefore complementary and, when combined, can increase the information available. Although the use of more than one technique for analysis is common, the data from each are usually treated separately using distinct methodologies and the results interpreted individually. However, chemometric methods for the co-analysis of multiple data sets, known as data fusion, are being developed. For example, Crockford et al. [33] describe an approach to integrate metabolomic data from ^{1}H NMR and UPLC-TOFMS and Forshed et al. [20, 228] have pioneered methods for the fusion of NMR and LC-MS data sets. Data fusion methods offer increased information and understanding through the application of computational and statistical techniques.

Chapter 3

Mass Spectrometry and Liquid Chromatography

> Research is what I'm doing when I don't know what I'm doing.
>
> Wernher von Braun

3.1 Introduction

The power of LC-MS as an analytical technique is enormous; the range of instrumentation and and degree of customisation allows for the analysis of small and large molecules in various differing conditions. Whilst this Chapter focuses on the analysis of small molecules, LC-MS is certainly the most popular technique for the analysis of proteins and peptides. For small molecules, the range of ionisation techniques facilitates both targeted and non-targeted analyses alike, as LC-MS allows for initial separation of analytes, often according to polarity, before ionisation and mass analysis.

Although LC-MS is a powerful technique, its use must be considered as part of a larger experiment that is carefully designed around clear aims. Experimental planning necessitates cogent sampling and preparation strategies to ensure that the fidelity of the original sample is maintained. The long-term reproducibility of LC-MS remains a key challenge, and the adoption of quality control strategies is recommended in order to be able to correct for drifts in experimental performance [229, 230] but should not be used in place of good sample handling and instrumental maintenance practices.

The evolution of mass spectrometry has resulted in the ability to not only collect data more easily, but also to acquire more of it. The result is a mountain of better resolved data with highly accurate mass-to-chage (*m/z*) values. Without adequate data curation and management strategies, much of the information within mass chromatograms is likely to be insufficiently exploited. Techniques for adequate processing of spectrometric data [231] need to keep apace of the ability of the hardware to generate more, and better quality, data.

In non-targeted analysis the number of detected features throughout mass chromatograms may be enormous, and the application of multivariate statistical techniques helps to focus the experiment towards the key features that are highly discriminatory in terms of the experimental aims. The use of appropriate statistical techniques identifies trends within the datasets, alongside variables that contribute to such trends. The trends within a dataset may be the result of the up-regulation of molecules due to a certain stimulus (as in metabolomics studies [12, 14]) or more simply the presence of an unexpected molecule within, for example, food matrices [7, 232].

Assignment of spectral features represents one of the largest bottlenecks in an experiment [37]. The range of molecules found within complex matrices is enormous and considerably more varied than the structural elucidation required of proteins and peptides. MS provides (accurate) *m/z* values, possible molecular formulae and isotopic distributions but the so-called 'soft' ionisation techniques tend to provide little structural information. Instead, the use of secondary fragmentations is required but this is typically not achievable in non-targeted analyses. Thus, non-targeted studies typically rely on comparisons to spectral databases, which are either maintained in-house or those commercially or freely available.

Whilst the sharing of spectral data is to be encouraged, it must be provided with adequate metadata which describes the experiment and acquisition parameters. Initiatives aimed at harmonising the reporting of metadata [51, 233] have been proposed to facilitate more sharing. The ever-increasing sensitivity of LC-MS will result in the discovery of more constituents in a mixture and it is therefore unlikely that a database will ever be complete. The sharing of data may, therefore, be most effective for assignment of spectral features. The variability of LC-MS is such that metadata reporting is crucial for comparisons made under different conditions.

This Chapter details various aspects that should be considered when using LC-MS for the analysis of complex mixtures. The preparation of various matrices are discussed in relation to both targeted and non-targeted analyses, as is a selection of ionisation sources and mass analysers from amongst the plethora available in modern instrumentation. Various approaches to sample preparation are also discussed, as are multivariate techniques for the effective analysis of potentially megavariate datasets.

3.2 Sample preparation

Samples are key to any analysis, and their robust treatment is a precondition for a fair and unbiased analysis. Without defined procedures for sample collection and their preparation, measured responses will clearly be influenced by variability introduced during their handling. Thus the development of automated preparation strategies to minimise human handling may prove especially promising, not least in metabolomics (and other non-targeted) studies where the emphasis is placed on identifying differences in natural metabolic flux, that may be confounded by change introduced as a consequence of sample preparation. Of course, the quantification of such perturbations is non-trivial but can at least be estimated by the comparison of multiple preparation methods.

The preparation for samples used in a targeted study can afford to be less inclusive, but no less exacting. The detection and quantification of, for example, contaminants can employ preparation methods that are tailored towards high-recovery of certain analytes.

The storage of samples, both pre- and post-extraction is also a key consideration. The analysis of ambient temperature thermolabile compounds becomes especially challenging. Clearly storage of samples at low temperatures and in aseptic conditions will help inhibit enzymatic and microbial action, and the availability of cooled auto-sampling cabinets allows for the high-throughput of heat sensitive samples.

The matrix of a sample is a key preparation consideration. A matrix is technically defined as everything in a sample except for the analyte that is being studied. The over-arching 'matrix effect' is defined by IUPAC's Gold Book [234] as 'the combined effect of all components. . . on the measurement' of the analyte, and may be particularly prevalent in complex matrices. The process of standard additions can be used to gauge its effect on an analyte. The method involves successive spikings of the sample with a known quantity of analyte such that a relationship between an analyte‘s response and concentration can be determined. The dilution of a sample may also be considered. A comprehensive review of matrix effects, by Matuszewski et al. [235], studies the phenomenon for LC-MS, operating in both APCI and ESI modes.

In the experimental design, it is important to address what the requirements are in order for the experiment‘s aims to be achieved. This must include the identification of factors likely to introduce variability into the data, along with the development of methods to either nullify, where possible, or mitigate by 'blocking' samples according to the factors of variability. For example, a study may wish to assess the physiological effects of a certain stimulus on people. The subjects are variable, and come with a range of age, gender, diet etc. To mitigate their inherent variability, they should be 'blocked' such that the control and stimulus groups contain approximately equal mixes of people according to their sources of variability, i.e. age, gender and diet. By blocking the samples, the control-stimulus differences should not be related to any of the subjects‘ characteristics.

To ensure statistical significance, the number of replicates, the size of the expected difference, and the significance criterion need to be considered. For hypothesis-driven experiments this is relatively tangible; however, for hypothesis generating, such as non-targeted metabolomics experiments, there are no data on which prior assessments can be made. Determining the number of replicates relies on a 'best estimate' formed from previous studies, a priori assumptions or, more loosely, analytical intuition.

3.2.1 Standardisation

The integration of the 'omics' technologies has been heavily assisted by the availability of online databases, which necessarily involve the creation and management of metadata. This metadata, which describes the conditions under which the data were collected, is of singular importance relating to the broader applicability of the data to other studies. Certain good practice guides have been suggested, in the field of metabolomics, notably MIAMET (minimum information about a metabolomics experiment [233]) and MSI (metabolomics standard initiative [51, 236]).

An issue relating to open collaboration and sharing of data is structured around the reality of LC-MS data acquisition. Reproducibility is defined, by IUPAC's Gold Book [234], as "the closeness of agreement between independent results obtained with the same method on identical test material but under different conditions" and is an issue that needs resolving for methods using liquid chromatography-mass spectrometry.

In general, few studies [229, 237] have been presented in the literature that deal with reproducibility across hardware. Nuclear magnetic resonance (NMR) spectroscopy has been demonstrated to be highly reproducible regardless of the instrument's manufacturer [118, 119]. Whilst progress is being made [238], the poor reproducibility of LC-MS data is likely to remain the main stumbling block to the active sharing and use of communal data. However, this variability is in part due to the success of LC-MS. The ability to employ different column chemistries for separation, along with bespoke elution profiles clearly allows for a greater degree of characterisation than is associated with NMR. It is clearly such variety that helps to make LC-MS such a universal method.

3.2.2 Normalisation

The repeatability of an instrument is also a key consideration, and follows on directly from the discussion of reproducibility above. Analysis of replicates over a short period of time can reveal variations in response that need to be mathematically corrected. The variability in response may be derived from LC or MS. Alterations in retention time are possible, yet different molecules may be more susceptible to this than others. Equally, the response and mass accuracy may vary when detected by the mass spectrometer. And of course all of these factors can easily be influenced by minor variations in the sample pH, composition, etc, which may result in significant, and potentially unexplainable, variations.

3.2.2.1 Internal standards

Spiking a molecule at known concentration into a sample provides an internal standard, which can be used to gauge drifts in retention time, intensity of response, and *m/z* value. For quantitation experiments, such as multiple reaction monitoring (MRM), deuterium- and ^{13}C-labelled isotopomers of the analyte are often used. An internal standard should be exogenous to the sample, and a distinctive isotopic pattern (as is provided by, for example, multiple chlorine moieties) will facilitate its identification. For complementary studies, they should ionise well in positive and negative ion modes, and be retained by a plethora of column chemistries.

3.2.2.2 Quality control

The use of quality control (QC) samples can provide analysts with a capability to correct spectral variables on a case-by-case basis. QC samples are often formed as pooled samples, in that a fraction of each sample is mixed into a global QC sample. As the QC is a pooled sample, each variable is present (although subsequent dilution may put the concentration below the instrumental detection limits) and can be used to show how a single variable responds over the course of an analytical run. QC samples are wisely used at the beginning, end and regularly spaced throughout a run, and are especially highly recommended in non-targeted experiments. It should be noted that the use of QC samples does not obviate the need for good experimental design and operating procedures. QC samples cannot be considered as the saviour of experiments with badly controlled sources of variability. Any such factors should be identified and addressed prior to the experiment, such that the methods are both repeatable and reproducible.

Zelena et al. [229] detail their method development for a 3-year-long metabolomics study. They analysed the reproducibility of two identical analytical columns from different production batches, and compared the resultant analyte intensities by principal components analysis (PCA). The results from PCA revealed no significant difference between the columns. An analytical block may be considered as a group of samples run consecutively, without human interference. In order to gauge an appropriate size for such a block, an experiment was carried out where 120 QC samples were sequentially injected and analysed by PCA. Runs one to 60 were found to cluster randomly, whilst 61 to 120 were found to vary in relation to their run order. The results were compared to another experiment, which consisted of two blocks of 60 QC samples with an instrumental cleaning stage in between. As neither of the two smaller batches showed a run order dependency, it was concluded that 60 samples was a more appropriate block size.

Aside from reproducibility measurements, QC samples can also be used as a measure to correct for order-of-injection-dependent variations, as is reported by in the protocol developed by Dunn et al. [230]. It relies on periodic analysis of QC samples throughout an analytical block, and is applied on a variable-by-variable basis. Called locally estimated smoothing (LOESS), the method fits polynomial functions between subsets of adjacent QC sample intensities. A smoothing parameter is used to determine the number of adjacent QC intensities, and a linear or quadratic function is fit to adjacent sets by least squares regression. Leave-one-out cross validation is applied to each function to prevent the inclusion of excessive random error. A single correction curve is generated by applying cubic spline interpolation to the individual polynomial functions, and all intensities are normalised to this single curve.

3.2.3 Run order

The choice of sample run order largely depends on the type of study being carried out. For targeted studies, it is good practice to analyse lower concentration samples before those of higher concentration. This way the effects of sample carry-over can be reduced and fewer column washes are needed between samples to ensure that each run is unbiased by the previous one. For non-targeted studies, the importance of sample randomisation must be emphasised. However, for long experiments the use of complete randomisation must be inherently avoided due to the finite chance of technical replicates (repeated injections of the same sample) being run sequentially.

Instead, technical replicates should be evenly spaced throughout the full analytical run so that variability in instrumental conditions is reflected in of the data from the different replicates. Whilst it may sound counter-intuitive to introduce additional variance within replicates, it does allow, as an example, multivariate statistical techniques to identify genuine differences between variables. If sample A is run and replicated on day 1, and the same for sample B on day 2, observable differences may appear to be related to their expected (class) difference. But other sources of variability may be influencing any eventual differences. The randomisation of the run order (such as AB on day 1 and BA on day 2) can assist with separating day-to-day variability from A-B variability. Without such separation, it is likely to lead to the conflation of the two sources of variability.

As an example, for 30 samples each injected in triplicate (technical replicates A, B, and C), the experiment could be run in three distinct batches. Each of these batches contains either A, B or C, and for each batch their position in the run order is random.

3.2.4 Acquisition modes

The choice of acquisition mode is dictated largely by the aims of the experiment. Non-targeted and metabolomics studies of small molecules typically perform full scan experiments encompassing the *m/z* range of 50 to 1000. Full scan acquisition makes no *a priori* assumptions, is the least wasteful of sample, and allows for maximum information recovery about the constituents of the system. Recording the full range also has the advantage that it permits additional hypotheses to be formulated and tested later, so that the data can be used for purposes beyond the original experimental aims. However, the collection of 'excess' data is not without drawbacks; data files can becomes very large, and require storage and high-performance computers for processing.

Targeted experiments are clearly more selective and the use of MS/MS can allow for specific quantification (using multiple reaction monitoring) or deeper characterisation of analytes (product ion analyses). The application of automated MS/MS allows the combination of full scan acquisition along with fragmentation of selected components. Thus full scan and structural information can be extracted from a sample, but the high-intensity selection method suffers from the assumption that the interesting components are at a high intensity.

3.2.5 Matrices

A standard operating procedure (SOP) for specific matrices should be adopted. It helps to ensure that preparation is consistent across all samples, by helping to minimise potential variability caused by long time delays between batch preparations, and in homogenising the lab habits of various analysts. A non-exhaustive guide to sample handling for common matrices is presented below.

3.2.5.1 Urine

The preparation required for urine is often minimal, and samples may be injected without dilution, due to the relative simplicity of the matrix. Due to the ease of sample collection, in part due to its non-invasive nature, urinalysis is widely employed in metabolomic studies and has great prognostic (evolution) and diagnostic (identity) capabilities.

A recommended sample preparation strategy, presented in work by Cubbon et al. [14], analysed the metabolic profile of human urine and compared HILIC and reversed-phase separations. The method recommends that samples are frozen within 2 hours of collection, and remain frozen for at least a week to permit any flux in components to be consistent across all samples [239].

3.2.5.2 Blood

Where blood samples have been collected into tubes with an anti-coagulation agent (such as heparin), a plasma supernatant forms after centrifugation. In the absence of an agent, serum is the resultant supernatant liquid after centrifugation of clotted blood. The difference between the two is that plasma contains proteins and factors associated with blood clotting, whilst serum does not. However, there remains a large quantity of protein which can be removed by a 'protein crash' [240] as part of the preparation stage. Such an approach, however, is problematic to non-targeted analyses due to the possibility of small molecules binding to proteins, which are accordingly lost during the protein precipitation.

The choice of anticoagulant is also important, as its residues may overlap with endogenous metabolites, and the increase in metal cations may lead to the formation of a large range of charge-carrying species.

The preparation of serum necessarily involves the clotting of blood. The time required for the clotting is not fixed, and would clearly depend on, amongst other factors, the temperature at which the sample is maintained. However, without the characterisation of such an effect, it would seem that plasma is an easier sample for which to standardise a collection protocol. The blood collected into an anticoagulant-containing tube can be stored, centrifuged and frozen all within a fixed time and under standard conditions. Plasma is less suitable than serum for small molecule analysis due to the larger quantity of peptides and proteins within the sample. Phospholipids are known to cause ionisation suppression in serum samples, and their removal is often desirable such that lower intensity metabolites can be observed. In spite of this, phospholipids can be useful as part of a targeted metabolomics study: Ferreiro-Vera et al. [241] have compared various sample preparation procedures for the profiling of phospholipids in human serum. The numbers of features in RP-LC-TOF spectra (positive and negative ion modes) were compared for three distinct methods: single-solvent protein precipitation, liquid-liquid extraction (LLE) and solid-phase extraction (SPE). Each of the methods was assessed with different solvent combinations, and the optimal solution was found to be SPE. The method is applicable to those wishing to remove phospholipids, and also those wishing to analyse them; phospholipids were retained in the SPE cartridge, and subsequent elution with methanol facilitated their analysis.

Denery et al. [242] and Wedge et al. [243] have investigated the difference, in terms of metabolites, between plasma and serum, and report that one sample type may be more suitable than the other. Both studies found that the number of detected features was broadly comparable across serum and plasma, and that there was no evidence suggesting the clear superiority of one sample or the other.

The collection of blood from capillary veins in the finger was examined by Denery et al. [242], due to its potential for a greater surgical throughput and lower cost, such that it may be easier to translate the practice into the developing world. A comparison between capillary blood and venipuncture blood compared the number of detected features in ESI-TOF-MS spectra, in both positive and negative ion modes. Whilst the difference in feature quantity was small, further investigation of the significantly different features revealed that many were attributable to the antiseptic wipe used to prepare the finger prior to puncture. As none of the differences were attributable to endogenous molecules, it is clear that the less invasive capillary method could be widely applied. However, it is obvious that the materials used need to be properly reported, analysed and characterised.

Whilst Wedge et al. [243] found no *de facto* reason to favour either plasma or serum, some prognostic biomarkers related to small cell lung carcinoma (around which the study was designed) were found uniquely in plasma samples. The study raises the obvious point that researchers may find either plasma or serum to be more informative due to the inability to convincingly measure the entire range of molecules in either medium. Therefore it is clear that without initial characterisation studies of both serum and plasma, it is possible for key metabolite fluxes to be missed by solely analysing one fluid.

3.2.5.3 Animal tissue

The non-targeted analysis of animal tissues, for applications such as the assessment of meat quality and geographical origin, is dominated by NMR [107] and proteomic [244] approaches. Suroweic et al. [245] have presented a GC-MS metabolomics approach for meat authenticity, in relation to the method of meat recovery. The approach allows for the successful discrimination of mechanically recovered meat from other meat harvesting methods. The distinction is important, due to the EU's regulations regarding the distinct labelling of mechanically recovered meat. The authors detail a range of tentatively assigned molecules that successfully discriminate meat types, but suggest that non-targeted multivariate, as opposed to targeted, classification is better suited to the discrimination of meat types.

The majority of small molecule analyses of animal tissue in the literature are targeted studies, typically focused on the detection of regulated species [246–249] or determinants of sample quality [250, 251]. The analysis of vitamin D and its derivatives is problematic due to the labile nature of the derivatives with regards to both ultraviolet radiation and oxidising chemicals. A sample preparation procedure, reported by Strobel et al. [251], recommends the use of subdued yellow wavelength light, and high purity nitrogen for drying and as an MS inlet gas so as to prevent the reaction of labile species with oxygen.

3.2.5.4 Plant tissue

The extraction of small molecules from plant cells is considerably more involved than that needed for animal cells. The presence of a cell wall prevents rupture by osmotic lysis, and the use of snap freezing, lyophilisation and cryogenic mills are common to encourage the necessary cellular rupture. Whilst sample preparation is best reduced to as few stages as necessary, it is also important to understand that extraction efficiency is clearly dependent on the matrix. Accordingly, there is unlikely to be a single procedure that is optimal for all plant material.

The non-targeted analysis of grapes by Theodoridis et al. [252] applied experimental design in the determination of an optimal extraction solvent, consisting of varying the proportions of methanol, water and chloroform used. Grapes were homogenised under liquid nitrogen, and then vortexed with the extraction solvent before being centrifuged. Where the solvent contained chloroform, the two phases were then separated. The aqueous layer was further separated by solid-phase extraction to separate low concentration secondary metabolites from the dominant primary metabolites. The number of features detected in positive mode ESI, presented as a response surface [89], shows that the optimal extraction conditions for reversed-phase separations used 40-60% methanol and chloroform and up to 20% water. For HILIC, the maximal number of features was extracted with higher methanol content than for RP, and 40% being the upper limit for chloroform.

Using porous graphitic carbon (PGC) as a stationary phase can provide an enhanced separation, in comparison to HILIC, of highly polar metabolites. This is especially true for carbohydrates. Developed by Antonio et al. [253] a water, methanol and chloroform extraction is performed on previously lyophilised plant stems. The procedure combines two sequential aqueous extracts, and separation of the carbohydrates of interest can be effected within 10 min using a PGC column.

3.2.5.5 Food

The analysis of food can be regulation- and quality-driven. The former approach is focused on the detection and quantification of restricted substances, such as pesticides and intentional dopants. The QuEChERS method, defined, by Anastassiades et al. [254], as 'quick, easy, cheap, effective, rugged, safe', is used as the *de facto* multi-residue method for pesticide analysis. Its wide adoption is likely due in part to its relative simplicity: the comparison made by Soler and Picó [232], reproduced in Figure 3.1, aptly summarises the simplicity of QuEChERS compared to a previous multi-residue method. Their review details alternative extraction procedures, such as matrix-solid phase dispersion (MSPD) and pressurised liquid extraction (PLE). MSPD is well suited to small sample volumes, and is cheaper on a per-sample basis than PLE and solvent extraction methods. PLE is recommended for analysis of matrices with low water contents, but the high temperatures that are required may degrade some components. The benefits and drawbacks of various MS/MS configurations, in the context of pesticide analysis, are also discussed. Whilst quadrupole time of flight (QqTOF) systems offer high sensitivity and mass accuracy, the amount of information regarding their use for pesticide analysis is rather limited. The ability to perform MS^n with quadrupolar ion traps (QIT) is offset by the poor dynamic range, and triple quadrupoles (QqQ) exhibit a strong quantitative capacity but the number of residues that can be analysed is limited by the rate at which the quadrupoles can be switched.

Eighteen organophosphate pesticides have been analysed in multiple reaction monitoring (MRM) mode by Sinha et al. [255] using a QuEChERS methodology. An initial full spectral scan was recorded of the individual pesticides, with the first quadrupole set to transmit the protonated form. At least two product ions were identified, such that one was used for quantification and another to act as a confirmatory peak.

Quality-driven approaches, more likely to be non-targeted, are widely used to differentiate samples, such as by age, type, location and production batch. Non-targeted methods are routinely used to differentiate samples with protected statuses, and also to characterise the nutritional content of various foodstuffs.

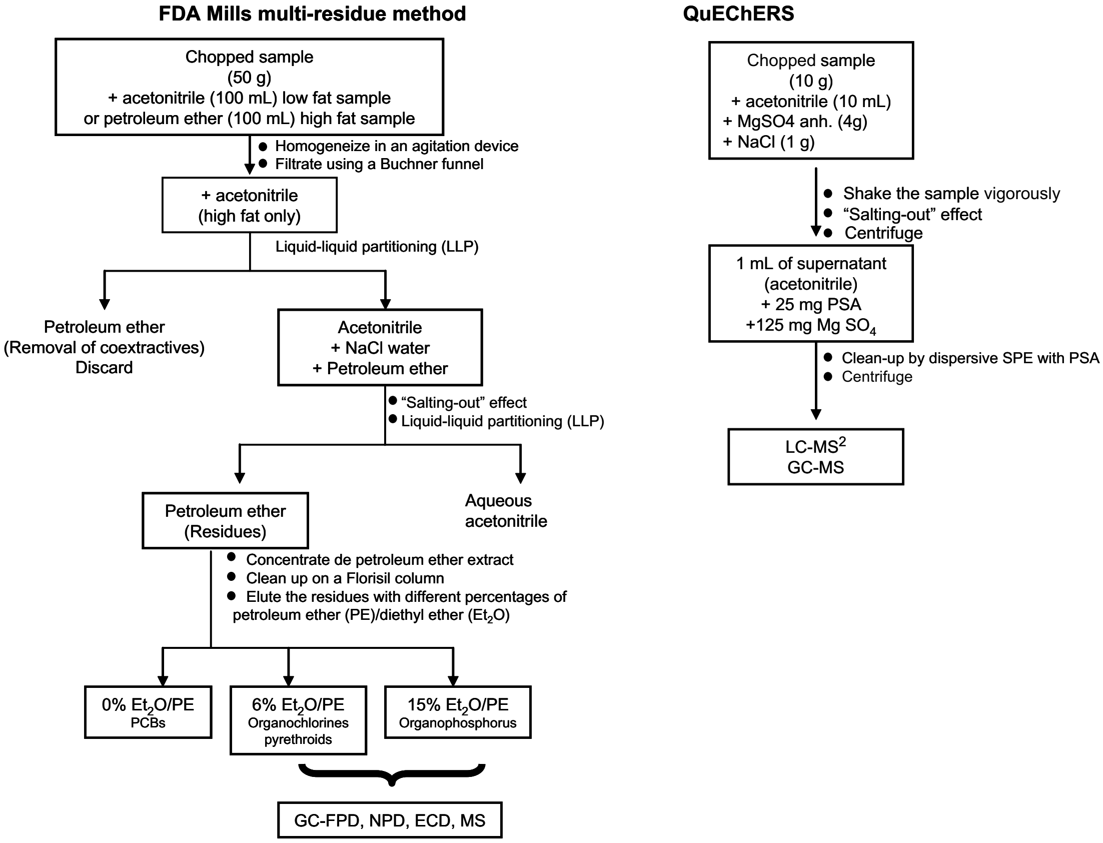

FDA Mills multi-residue method
Chopped sample (50 g) + acetonitrile (100 mL) low fat sample or petroleum ether (100 mL) high fat sample
● Homogeneize in an agitation device
● Filtrate using a Buchner funnel
+ acetonitrile (high fat only)
Liquid-liquid partitioning (LLP)
Petroleum ether (Removal of coextractives) Discard
Acetonitrile + NaCl water + Petroleum ether
● "Salting-out" effect
● Liquid-liquid partitioning (LLP)
Petroleum ether (Residues)
Aqueous acetonitrile
● Concentrate de petroleum ether extract
● Clean up on a Florisil column
● Elute the residues with different percentages of petroleum ether (PE)/diethyl ether (Et_2O)
0% Et_2O/PE PCBs
6% Et_2O/PE Organochlorines pyrethroids
15% Et_2O/PE Organophosphorus
GC-FPD, NPD, ECD, MS
QuEChERS
Chopped sample (10 g) + acetonitrile (10 mL) + MgSO4 anh. (4g) + NaCl (1 g)
● Shake the sample vigorously
● "Salting-out" effect
● Centrifuge
1 mL of supernatant (acetonitrile) + 25 mg PSA +125 mg Mg SO_4
● Clean-up by dispersive SPE with PSA
● Centrifuge
LC-MS^2
GC-MS

FIGURE 3.1: The difference between the two methods is stark: the QuEChERS method is a reliable alternative to the more convoluted approaches that preceded it. Reproduced from Soler and Picó [232].

The purported health benefits to tea drinkers include, amongst many others, reduced risks of cancer and cardiovascular events. Many studies have been performed on both infusions and raw leaves, for the determination of, for example, metals [256], antioxidants [257] and amino acids [258], with the ability to classify samples according to type or geographical origin. A non-targeted analysis of tea infusions by Fraser et al. [259], using HILIC-ESI-MS, allows for a very simple sample preparation. Infusions, spiked with an internal standard, were filtered prior to 1:1 dilution with MeCN. The use of HILIC allowed for the identification of highly polar compounds, notably amino and organic acids, and also carbohydrates.

3.2.5.6 Environmental samples

The use of mass spectrometry in environmental metabolomics studies has recently been reviewed by Viant and Sommer [12]. It details GC, LC and direct infusion approaches that have been applied to the analysis of a number of organisms, such as tree, squirrel, and fish. So-called sentinel organisms are widely studied in order to gauge the effect of perturbations in a natural environment; the contamination of water may be directly detectable, but any potential effects cannot be understood without analysis of an ecosystem‘s inhabitants.

One of the successes of metabolomics is that it can be used as a direct method to link an organism‘s phenotype with its genotype. The combination of metabolite flux and genetic sequencing allows for a greater understanding of the effect of, for example, a gene knockout on an organism. Such experiments are difficult to perform in isolation, and collaboration is often needed. Where collaboration is needed, it is clear that standardisation is required. Such a framework, in the context of environmental studies, has been proposed by Morrison et al. [260] for the reporting of environmental conditions in metabolomics experiments. The suggestions involve reporting the more obvious necessities, such as the binomial name of the studied organisms and any perturbations to which they were subjected. Also important may be the weather, and other environmental conditions, such as soil or water pH and temperature. For studies involving nurtured organisms, such as in a laboratory, their location (plant pot, cage etc.) and feeding regime are also important to detail. It is clear that with the increasing importance of environmental issues, the need for adequate reporting and dissemination will only increase.

3.3 Hardware: the physical wherewithal

As with many other technologies, the hardware used in LC-MS is evolving apace. New technologies are being developed that are capable of offering yet greater efficiencies, be they in terms of chromatographic resolution, ionisation, ion transmission, or detection.

3.3.1 High performance liquid chromatography

Chromatographic techniques can be used to perform the separation of a mixture's components, and liquid chromatography (LC) is applicable to liquid-soluble molecules. The principles of separation relate to the differential interactions of each constituent with both a stationary and a mobile phase.

3.3.1.1 Reversed-phase chromatography

Reversed-phase (RP) separations employ non-polar stationary phases. These are often silica particles that have been functionalised with, most commonly, octadecyl carbon chains (C_{18}), but shorter (e.g. C_8) chains and phenyl groups may be used. In spite of superseding normal phase chromatography, RP separations have become *de rigueur* not least due to their ability to directly hyphenate to mass spectrometers, through use of atmospheric pressure ionisation (API) sources.

3.3.1.2 Hydrophilic interaction chromatography

Normal phase chromatography separates polar molecules according to the strength of their interaction with an often unmodified silica stationary phase. Molecules are retained by an adsorption mechanism, such as through hydrogen bonding to the surface, with the strength of such interactions dictating the rate of passage through the column. The poor reproducibility of retention times has largely seen normal phase chromatography superseded by hydrophilic interaction chromatography (HILIC) [261].

The mode of separation in HILIC is, like in RP, partition based, where polar molecules are detained in polar solvent 'pockets'. Initial equilibration of the HILIC column with an aqueous solvent leads to the sequestration of water around the stationary phase. It is with this water, the so-called 'liquid-stationary' phase that the polar molecules in the eluting 'liquid-liquid' phase interact. An early study showing the applicability of HILIC in the analysis of complex biological molecules is presented by Tolstikov and Fiehn [262], and HILIC offers good separation for acids and bases alike. Cubbon et al. [263] provide a comprehensive review of the application of HILIC to the study of small molecules.

3.3.1.3 Mobile phase

Lengthy chromatographic separations result in broad peaks for the late-eluting compounds. The use of gradient elution, as opposed to single-solvent, isocratic elution, encourages the more timely passage of strongly retained species, and this reduction in retention time ameliorates peak broadening. A typical RP-LC mobile phase gradient would start with a 95:5 (v:v) mixture of water and organic solvent. The proportion of organic solvent is increased gradually throughout the run until a 5:95 ratio is achieved. The high proportion of organic solvent towards the end of the separation helps to remove stubbornly interacting molecules, effectively cleaning the column ready for the next separation. The stationary phase is then re-equilibrated with a 95:5 aqueous mixture in preparation for another separation.

3.3.1.4 Detection

Detection methods in chromatography can be either destructive or non-destructive. Whilst preparative separations employ non-destructive methods, such as ultraviolet (UV) diode arrays, or flow splitters, analytical techniques can use destructive methods such as atomic emission and, more commonly, mass spectrometry (MS).

The seeds of LC hyphenation to MS were sown in 1968 by Tal′rose [264], and the coupling of both was further advanced by Baldwin and McLafferty in 1973 [265]. The major challenge facing the pioneers of LC-MS was in interfacing an incoming liquid at a sufficient flow rate with the need to maintain a strong vacuum. The development of atmospheric pressure ionisation (API) sources continued throughout the 1980s, and manifested itself in two, now almost universal, ion sources: electrospray ionisation (ESI) and atmospheric pressure chemical ionisation (APCI). For a more comprehensive review of the history related to coupling chromatography to mass spectrometry, see that by Abian [266].

3.3.2 Mass spectrometry

Mass spectrometers are powerful analytical instruments. Whilst they vary in their capabilities, their common aim is the determination of the mass-to-charge ratio (*m/z*) of analyte ions. A mass spectrometer may be broadly divided into an ionisation source, mass analyser, and a detector, each of which is discussed in more detail below.

3.3.2.1 Ionisation source

In order for analyte molecules to be transmitted and separated by the spectrometer, they need to be ionised. Early spectrometers ionised gaseous analyte molecules following their interaction with a stream of electrons (electron ionisation, EI), whereby such interactions between molecule and electron typically led to the formation of a charged species. The energy transmitted from electron to molecule is typically chosen to be greater than that required to induce ionisation alone. The surplus energy may result in the ion‘s fragmentation due to its relative instability. However, the drawback to EI is the requirement for molecules to be in the gas phase. For non-volatile molecules this can be accomplished by heating, which risks pyrolysis, or derivatisation and the associated heating, which is not applicable to all molecules. The analysis and assignment of water soluble and fragile molecules is not always possible with fragmenting techniques, such as EI, and various attempts were made to create a 'softer' ionisation mechanism by which molecules would be unlikely to suffer extensive fragmentation. Fast atom bombardment [267] was aimed at such ionisations, and its use as a 'soft' ionisation technique has largely been superseded by the more capable matrix-assisted laser desoprtion/ionisation (MALDI) and electrospray ionisation (ESI), the latter of which is now most commonly used in combination with liquid chromatography for the analysis of small and large molecules alike.

3.3.2.1.1 Electrospray ionisation

Originally developed by Dole [268], electrospray ionisation (ESI) was applied by Fenn to the mass spectrometric analysis of large molecules [269–271]. This was made possible by the ability of the method to multiply charge analytes, such that their *m/z* values are compatible with the usual *m/z* ranges of mass analysers. A schematic of an ESI source is shown in Figure 3.2; a liquid sample containing dissolved analyte is introduced into the source by means of a capillary, which is maintained at a large potential difference with respect to an endplate. The production of the electrospray occurs at atmospheric pressure, and lenses and skimmers are arranged to help with sequential pumping to reduce the pressure to a level compatible with mass analysers.

With no potential difference applied between the capillary and endplate, an eluting stream of liquid will drip out the end of the capillary. As the voltage difference is increased, the formation of ions results in the creation of droplet at the capillary's

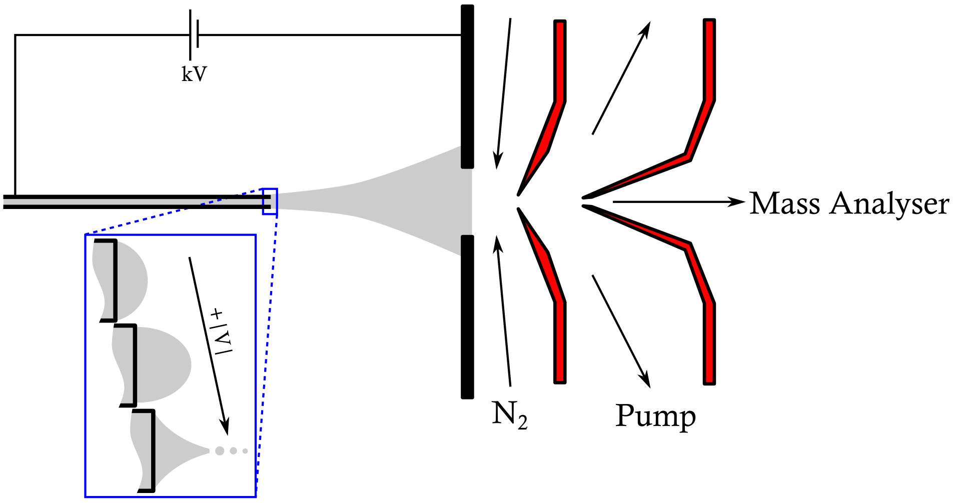


FIGURE 3.2: The electrospray ionisation source. The capillary inlet is maintained at a potential difference to the end plate. As the voltage is increased (see expansion), the droplet at the end of the capillary elongates, and eventually forms a Taylor cone from which smaller droplets are ejected. Their continual desolvation leads to the formation of ions. Ions are accelerated past the endplate, and towards the mass analyser.

tip, as is shown in the inset of Figure 3.2. A Taylor cone will form at the tip once the opposing force due to the electrostatic repulsion of ions overcomes the liquid's surface tension. It is from the Taylor cone that droplets are emitted towards the endplate, and from these the electrospray is formed.

Emitted droplets undergo desolvation such that free ions are formed prior to injection into the mass analyser. With high solvent flow rates, the application of a counter-flowing current of, sometimes heated, gas helps with the removal of solvent molecules, which are necessarily injected into the ionisation source. Ions are accelerated out of the source, towards the mass analyser. The pressure differential between the two is maintained by apertures and sequential pumping.

An electrospray source may be considered to be analogous to a constant current electrolytic (CCE) cell [272] where the current can be considered as the flow rate of charge. The current is essentially dependent on the solution flow rate through the capillary, the constituents of the solution, the applied voltage and the physical size of the device. The process of charge formation can essentially be described as an electrolytic reaction, such that the rate at which it occurs is proportional to the concentration, rather than absolute amount, of an analyte. Furthermore, the reduction in flow rates leads to enhanced sensitivity due to the increased contact time between analytes and capillary. Micro-ESI (μESI) [273] and nano-ESI (nESI) [274] have capitalised on these factors, and consequently employ lower flow rates and a narrower capillary. Indeed, nESI does not require a solvent pump as the ESI process alone induces sufficient flow. The size of droplets produced from the Taylor cone is proportional to the flow rate, such that at suitable flow/concentration combinations, the number of analyte molecules per droplet progeny can average one. The sensitivity enhancement seen in nESI is due in part to the fact that ions find it considerably easier to 'escape' from its smaller droplets [275].

3.3.2.2 Mass analysers

Ions are ejected from the source into a mass analyser en route to detection. There are various forms of mass analyser, and instruments may contain multiple identical analysers, such as triple quadrupolar systems, or a mixture in hybrid instruments, such as quadrupoles and time-of-flight. The focus here is on analysers that are applicable to the analysis of small molecules resulting from soft ionisation sources.

3.3.2.2.1 Collision induced dissociation

Soft ionisation techniques produce mostly intact molecular species. These provide information related to the molecular mass, but bestow little in terms of structural information: this can be gleaned by inducing fragmentation.

An ion stream can be directed through an inert gas, which is maintained at a pressure high enough to force gas-ion interactions. Such an interaction results in the conversion of a fraction of the ion's kinetic energy into internal energy. This energy is subsequently distributed through vibrations of the ion's bonds. The weakest bonds within a molecule are the most likely to cleave, and the resultant ionic fragments can be transmitted and detected. The ability to detect both the products of soft ionisation and their subsequent fragmentation products is afforded by the arrangement of multiple mass spectrometric analysers. The most common instrumental setups are discussed with reference to each analyser.

3.3.2.2.2 *Quadrupoles*

Quadrupoles, as depicted in Figure 3.3, consist of four parallel rods where opposite pairs are wired to carry the same charge. The charge is oscillated by the application of an alternating current (AC) voltage which is superimposed on a direct current (DC) voltage. Quadrupoles are used to act as filters which transmit a small range of *m/z* values, or as focussing devices to reduce the spread of an incoming stream of ions. Specific combinations of AC and DC voltages determine the stability of individual *m/z* values throughout the length (z) of the quadrupole.

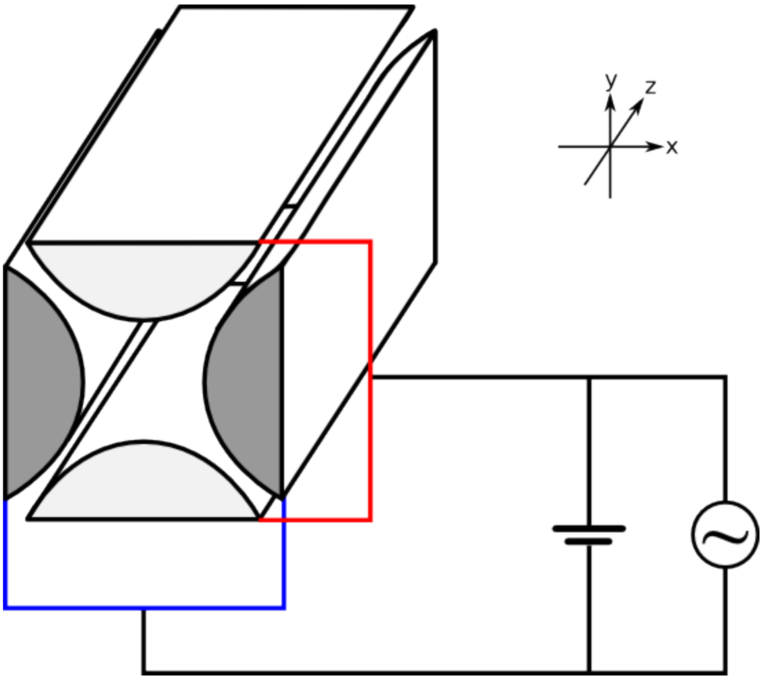


FIGURE 3.3: The arrangement of the rods in a quadrupole is such that opposing rods carry the same charge. The application of an AC voltage results in the switching of positive and negative charges between the two pairs of rods.

Quadrupoles may be arranged in a triumvirate in tandem instruments, and can be used in four distinct MS/MS modes. The central quadrupole acts as a collision cell that contains gas at a pressure liable to induce ion-gas collisions; subsequent conversion of the kinetic energy into, e.g. vibrational, energy provides an opportunity for fragmentation. The instrument can be set up to permit the fragmentation, transmission, or discovery of specific *m/z* values, where such information may prove useful in the identification of targeted ions. The various modes are diagrammatically represented in Figure 3.4, and discussed below.

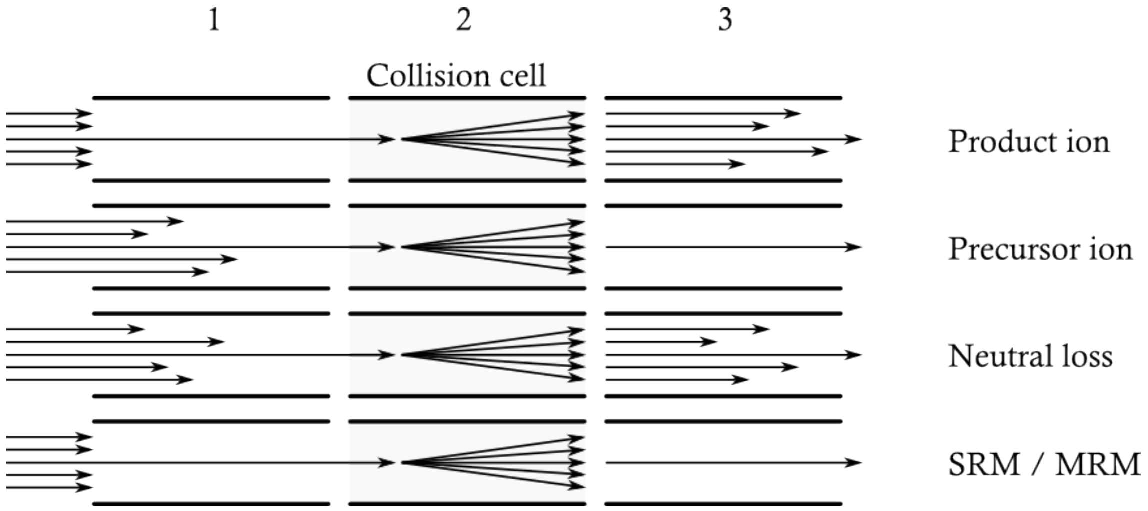


FIGURE 3.4: The four MS/MS experiments that can be performed with a triple quadrupole instrument.

The product ion experiment may be considered as the classic MS/MS experiment; a fragmentation spectrum is formed for each ion transmitted through the first quadrupole. Q1 is set up to transmit a single *m/z* value into the collision cell, and the final quadrupole is scanned to generate a spectrum of the product ions. The scanning nature of the instrument, however, means that only a fraction of ions of each *m/z* value are transmitted at any one time. This is often referred to as the 'duty cycle', which represents for how long an instrument is usefully employed. Instruments with poor duty cycles are less efficient.

Only fragments of specific *m/z* value are detected in the precursor ion experiment. The first quadrupole sequentially transmits all ions into the collision cell, and only those ions that fragment to produce the pre-specified *m/z* product ion cause a signal at the detector. The experiment can be applied when searching for charged ions, which are also diagnostic of an ion's identity.

Molecules with common functional and structural groups may fragment similarly, such that the loss of a common moiety may be diagnostic of a molecule's class. For example, alpha-amino acids are commonly observed to fragment via the loss of methanoic (formic) acid, a neutral molecule with a mass of 46 Da. In this case, the third quadrupole can be set to sequentially scan to transmit ions that are 46 Da lower than those that are sequentially passed through the first quadrupole as it scans. Thus each *m/z* value is transmitted, fragmented, and only those ions that fragment and lose the pre-specified mass cause a detector signal. The pre-requisite for the method is the loss of a neutral fragment, hence the technique's name of 'neutral loss'.

Selected reaction monitoring (SRM) is applicable to studies that are targeted at a specific component with a particular *m/z* value that fragments to form a product ion of a specific *m/z* value. The first quadrupole transmits a single *m/z* that is fragmented in the collision cell. The final quadrupole again transmits only a single predetermined *m/z* value. This experiment is most suitable for targeting known compounds with an understood fragmentation pathway. As the scanning ability of a quadrupole matches the chromatographic timescale, many specific fragmentations can be probed, and for this reason the technique is also referred to as multiple reaction monitoring (MRM).

Product and precursor modes are best used in conjunction with information about the sample; transmitting, fragmenting and then detecting every incident ion is costly, in terms of both sample quantity and time. The use of SRM is most applicable to targeted studies looking to screen for and often quantitate known compounds, such as the analysis of pesticides and veterinary drugs in food products.

The range of applications of triple quadrupolar instruments is vast, with the MS/MS experiments listed in Figure 3.4 being routinely used for the identification of molecules across a range of fields [232, 248, 276–278].

3.3.2.2.3 3D Paul traps

Ion traps, conceivably opposite in principle to quadrupoles, can be used to trap all ions within them, and these can subsequently be expelled to obtain a mass spectrum. 3D Paul traps [279–281], as depicted in Figure 3.5, consist of a ring electrode arranged in the xy-plane, above and below which sit two end cap electrodes with holes serving as ingress and egress routes for ions. A DC potential is applied to the end caps, whilst to the ring electrode an oscillating electric potential (AC) is applied, and this exerts a quadrupolar field on the ions.

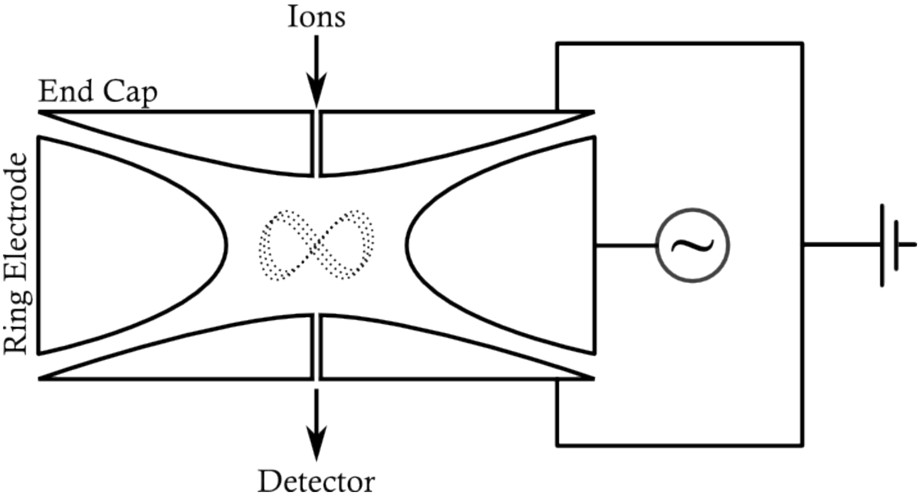


FIGURE 3.5: A 3D Paul trap, shown in cross-section. The ring electrode may be considered as a looped rod from a quadrupole, and to it an AC voltage is applied. The ring is sandwiched by two end cap electrodes, maintained at DC voltages, that contain passages for the entry and exit of ions. Ions weave figure-of-eight-like paths around the trap, due to the quadrupolar field exerted by the ring electrode. The ramping of the voltage that is applied to the ring electrode encourages the progressive excitation of ions, which are sequentially ejected towards both the detector and the ion source.

Ion trajectories are dependent on their *m/z* ratio, the size of the trap, DC and AC magnitudes, and the AC oscillating frequency. In order to maximise the range of *m/z* values that can be contained, the DC voltage can be set to zero. This is known as the mass-selective instability mode. Trapped ions can be scanned out sequentially to generate a spectrum. However, the inclusion of too many ions in the trap results in space-charge effects, whereby the quadrupolar field is distorted such that the paths of ions become erratic and they collide with the electrodes.

In order to form a mass spectrum, ions need to be ejected from the trap. To achieve this, an increasing voltage can be applied to the ring electrode. This destabilises the path of the ions, which are then ejected from the trap along the z-axis. As a result of the geometric arrangement, only half of the ions exit towards the detector.

The ejection of ions can be tailored to empty the trap leaving all but ions of some desired *m/z* value within the trap by ejecting those heavier and those lighter ions. After ejection of all undesired ions, the remaining ions can be excited. The excess kinetic energy of the ions can be converted into potential energy through collisions with an inert gas, generally the helium also used as a damping gas within the trap. This collision-induced dissociation (CID) forms fragments that can be maintained within the trap. All ions can be ejected, in this case forming a product ion spectrum for a single precursor. Alternatively, particular ions can be retained within the trap, and forced to undergo additional dissociations. This approach is known as MS^n. The value for n is effectively limited by the number of ions that can be put into the trap, which is stymied by the fact that half of the fragments are ejected away from the detector.

3.3.2.2.4 Linear ion traps

Linear ion traps (LIT) are based on a quadrupolar design, with electrostatic lenses at the extremities [282]. These act to confine the ions in the axial dimension, and the quadrupolar field maintains the ions within the trap‘s radial dimension. LITs have a greater ion capacity than Paul traps, and, therefore, are less susceptible to space-charge effects due to the linear, rather than confined, spread of the ions. Additionally, LITs are more efficient in terms of the number of incident ions that can be successfully trapped.

As with Paul traps, ions can be selectively ejected, but linear traps allow the ejection in either axial or radial dimensions [282, 283]. Mass-selective axial ejection (MSAE) can be achieved by providing ions with a degree of radial excitation, which can impart sufficient kinetic energy to allow them to overcome the repulsive potential of the axial end plate [283]. The amount of energy provided to the ion is much less than would be required to effect an ejection in the radial dimension.

For radial ejection, slits in two opposing quadrupoles allow passage of the ejected ions, and placing a detector in line with each slit allows for detection of the ions. Schowalter et al. [284] describe the development of a novel radial ejection method, in combination with a time-of-flight drift tube. Whilst the development is geared towards the study of atomic and small molecular species, radial ejection into a TOF tube provides sufficient mass resolution without the additional technicalities of axial ejection and orthogonal acceleration.

3.3.2.2.5 Orbitraps

Orbitraps [285, 286], or, more generically, electrostatic traps, contain a central spindle-like electrode which is maintained at an attractive potential to the ions being analysed (Figure 3.6). Around the outside sits a larger electrode. This is split in two, and is used to record the oscillations of the ions as they weave around the central spindle. Ions are injected into the trap either off-centrally or tangentially to the central spindle, at a velocity that balances the inward attractive force and the outward centrifugal force. This process is likened to the act of pulling back on a pendulum, and the ions weave around the central spindle electrode at a frequency proportional to their *m/z* value.

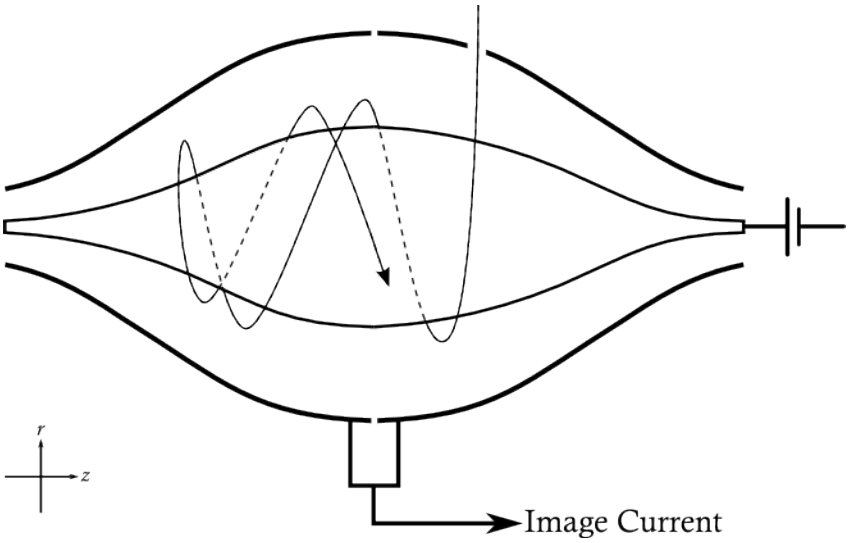


FIGURE 3.6: An illustration of the Orbitrap, showing the off-central injection of an ion stream. The motion of ions is detected as an image current, and Fourier transformation converts the free induction decay into a mass spectrum.

Orbitraps are similar to quadrupolar ion traps (QITs), in as much as the motion of ions within them is subject to quadrupolar electrostatic potentials. The main difference between the two is that orbitraps operate only with applied DC voltages, whereas AC voltages are applied to QITs. The electrostatic potential, U, of an orbitrap is given by Equation 3.1,

$$U(r,z) = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2} \times (R_m)^2 \times ln\left[\frac{r}{R_m}\right] + C \tag{3.1}$$

where r and z are the axes defined in Figure 3.6, k represents the field curvature, R_m the characteristic radius and C is a constant. Ions move around and along the central electrode, and Equation 3.1 reveals that these two movements are independent of each other. The frequency, ω, of the oscillations along the length of the electrode is given by Equation 3.2 where

$$\omega = \sqrt{\frac{z}{m}k} \tag{3.2}$$

As is shown, the frequency of an ion‘s oscillation along the z-axis is inversely proportional to its *m/z* ratio, and the outer electrode is used to record the current that is induced by the movement of the ions. The signal resembles a series of superimposed oscillations (one for each ion packet) as a function of time, which is typically around 1 s. Conversion from the time- to the frequency-domain is achieved by a Fourier transform (FT), which yields intensities as a function of frequency, which can be converted into a mass spectrum.

Orbitraps operate in a pulsed fashion, and require the use of traps such that they may be coupled to continuous ion streams, such as from atmospheric pressure ionisation techniques [288]. Furthermore, ions need to be injected into the trap in a well-defined bunch; as ions become more evenly distributed along the axial dimension, their average current tends to zero, and as a result of this they can no longer be detected. Other reductions in signal intensity are caused by space-charge effects, collisions due to the imperfect vacuum, and inhomogeneities in the field due to minor imperfections in the electrodes. Current instruments, of which an example is shown in Figure 3.7, use a C-trap to radially eject short packets of ion bunches towards the orbitrap. The spread of ions achieved by such rapid ejection is insignificant in relation to the time taken for a packet of ions to oscillate half the length of the trap. The novel use of lenses within the instrument helps to achieve such short ion pack-

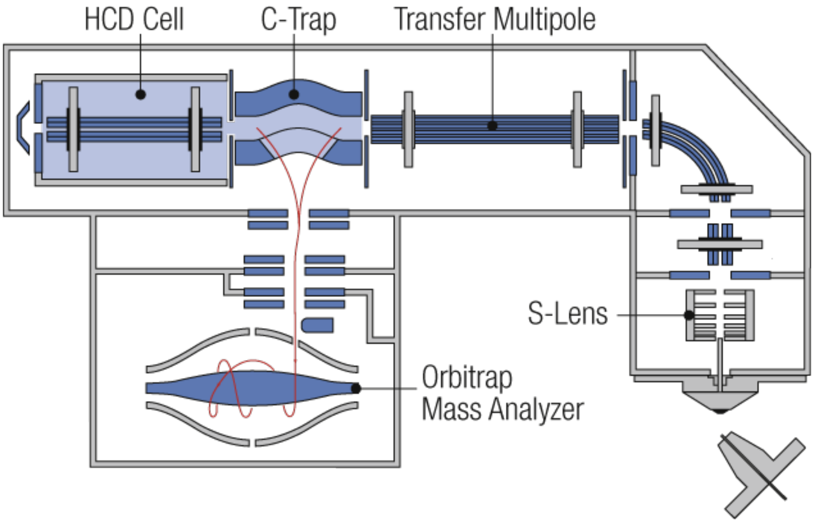


FIGURE 3.7: A schematic of a Thermo Scientific Exactive Plus Orbitrap mass spectrometer. The ions are injected in the bottom right-hand corner, and are transferred into the C-trap. Reproduced from Ref [287].

ets, which are at least 2-3 orders of magnitude shorter than can be achieved with standard linear traps [286]. Of the various MS technologies, only ion cyclotron resonance (ICR), another FT technique, can surpass the resolutions (FWHM) achievable with an orbitrap. Current orbitrap instruments can achieve resolutions in excess of 60,000 FWHM, with mass accuracies often better than within 2 ppm. By comparison, the resolution for ICR can be in excess of 100,000 FWHM with mass accuracies within 1 ppm.

3.3.2.3 Mass detection

Once ions have passed through the mass analyser, they need to be detected. The exception is for FT-based methods, that detect *in situ*. The number of ions leaving the analyser is typically too low for direct detection, and requires amplification that is often performed to generate a cascade of secondary electron ejections. For a comprehensive review of detectors, albeit from 2005, see that by Koppenaal et al. [289].

3.3.2.4 Resolution, accuracy and other nomenclature

The resolution of a mass spectrometer is commonly defined by the full width half maximum (FWHM) measure, as is pictorially shown in Figure 3.8. The resolution is *m/z* dependent, and an instrument‘s quoted resolution is therefore given at some arbitrary *m/z* value.

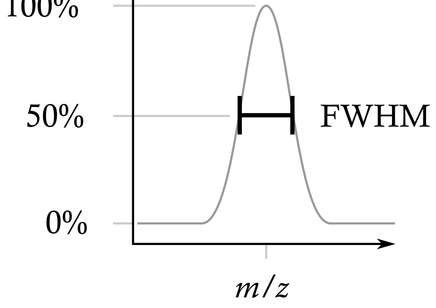


FIGURE 3.8: The *m/z* value of a peak is divided by its full width at half maximum (FWHM) to obtain an instrument's resolution.

The accuracy (or error) of a spectrometer is defined as the difference between the experimentally measured value (m_e) and the true value (m_t). It is often expressed in milli-mass units (mmu) or parts per million (Equation 3.3), which will be used from this point forwards.

$$\text{Error/ppm} \quad = \quad \frac{m_e - m_t}{m_t} \times 10^6 \tag{3.3}$$

3.4 Data processing

The processing requirements for targeted and non-targeted studies are very different. Most of this section details approaches for the extraction of all spectral variables. Targeted studies can simply focus on the ions of interest, and the processing requirement for this is much lower.

Full scan data can be collected in either profile or centroid mode. In profile data, multiple data points represent a single peak, whilst in centroid data these individual points have been merged and are presented as one. The advantage of profile data is that it allows for a value judgement to be made with regards to noise peaks, but at the cost of a much larger file size.

3.4.1 File formats

Whilst many instrument manufacturers favour their own proprietary formats, many offer tools for conversion into more generic formats. mzXML and netCDF are perhaps the two most common alternatives to instrument-specific formats, and are accepted by a wide range of open source data-processing software.

3.4.2 Processing methods

For studies involving multiple samples, the objective of data processing is to create an $m \times n$ data matrix of peak intensities, whereby each column (n) represents a single variable and each row (m) an observation. Processing necessarily begins on a file-by-file basis, where each produces a list of its variables. These lists must then be merged such that common variables are grouped together. It is often assumed that *m/z* values remain constant, whilst drifts in retention time require for variables to be aligned. Without such correction, a data matrix may contain multiple entries for a single ion, with a large proportion containing zero-intensity values. Periodic analysis of a QC sample can help in the identification of chromatographic drift.

There are various approaches to data processing, but common to all of them is the time required to perform the data manipulation. High-throughput capabilities, low limits of detection and high-resolution instrumentation all help to contribute to the enormous wealth of data that can be collected. Without the concurrent development of capable data routines, it is clear that the vast quantity of data is likely to overwhelm.

3.4.2.1 Time-averaged

A mass chromatogram is a series of mass spectra recorded at a specific frequency, for a specific time period. By taking the average of each *m/z* value over the full chromatographic period, a time-averaged mass spectrum can be formed, as is shown in Figure 3.9. This effectively negates the chromatography, producing a result akin to that expected from direct infusion mass spectrometry (DIMS). The rounding of *m/z* values must be performed in line with the instrument's resolution so as to avoid dissecting peaks into multiple bins.

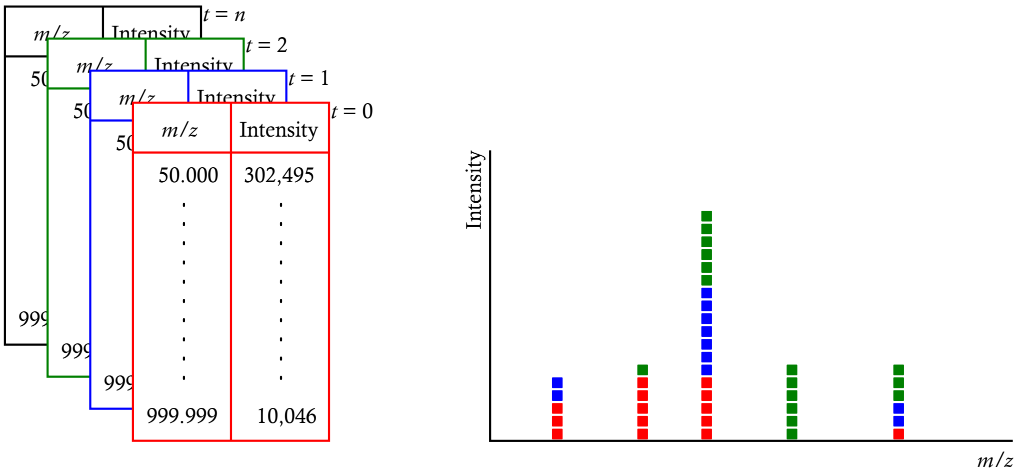
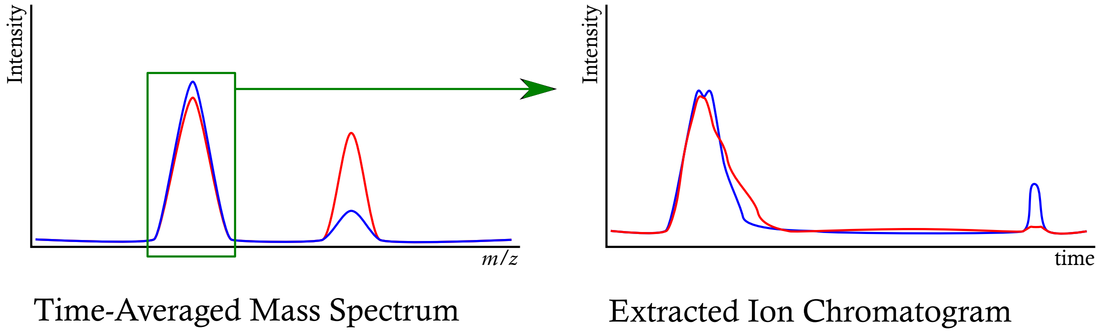


FIGURE 3.9: A schematic showing the formation of a time-averaged mass spectrum. Individual scans are combined by rounding *m/z* values to an arbitrary precision, and summing their intensities.



Time-Averaged Mass Spectrum

Extracted Ion Chromatogram

FIGURE 3.10: Two observations show similar intensities for a peak, shown in the green box, in the time-averaged mass spectrum. Inspection of the extracted ion chromatogram (EIC) reveals that significant variance in one peak is masked by minor fluctuations of a much larger peak.

3.4.2.2 Time-averaged plus

The time-averaged approach, above, is a simple procedure and can be used in multivariate analyses to identify *m/z* values that discriminate between groups. Its major drawback is that it assumes that only one analyte is represented by each *m/z* value.

Full scan mass chromatographic data can be quite sparse, in that little of the time-*m/z* space contains a signal above the detection limit. A time-averaged mass spectrum can initially assess which *m/z* bins are 'full', and each of these can be re-expanded such that the chromatographic information is recovered. This process essentially involves the formation of an extracted ion chromatogram (EIC or XIC) over the width of an *m/z* bin. This EIC may help in the determination of noise peaks, and can reveal the presence of multiple isobaric ions, assuming that they have different retention times.

The characterisation of peaks as both a function of *m/z* and time helps to avoid the pitfall of (not) identifying significant between-group differences, as is shown in Figure 3.10. It is entirely plausible for two isobaric ions, individually exhibiting between-group differences, to combine such that together they demonstrate no remarkable group-to-group variation.

3.4.2.3 Gaussian modelling

The methods above rely on the adequate rounding of data so as to decide into which bin to place an ion's intensity value. This allows for the construction of a mass chromatogram, but at high resolution becomes unmanageable due to the sheer number of mass and time points. Data files are structured as a series of lists, each representing a scan, containing *m/z* and intensity pairs for spectral regions where an ion was detected. The list is discontinuous such that compression techniques, for example wavelet transformations, cannot be performed without first placing the data into a continuous vector with equally spaced elements; such an approach relies on rounding *m/z* values.

The compression of data is a crucial point, as file sizes may run into the many gigabytes. Compression can be achieved by fusing together those data points that form part of a single peak. Gaussian distribution functions can be used to map a peak, and hence reduce the number of data points (Figure 3.11). Each function can be represented by location (*m/z*), width, intensity as well as the sum of the original data points that it encompasses.

These Gaussian peaks, determined from each scan, can be compared to preceding peaks such that an ion's chromatographic profile can be developed without the need to arbitrarily round *m/z* values (native resolution is retained). Peaks that exhibit either too long or too short a retention time, along with those with an abnormal Gaussian profile, can be filtered out on the assumption that they are noise. This is shown in Figure 3.12.

3.4.3 Missing values

Missing intensity values for spectral peaks may be due to either technical issues, such as sample or instrumental inhomogeneity, or the real possibility that a peak's intensity in a sample falls below the instrumental limit of detection. The comparison of technical replicates – i.e. repeated injections of the same sample – should help in determining whether the missing value is expected.

The appearance of missing values in a data matrix may compromise uni- and multivariate analyses. Whilst for larger data sets the presence of a few missing values may be inconsequential, smaller-scale metabolomics studies may be adversely affected as a reduction in sample size might render redundant any significance tests. The estimation of missing values has been proposed as a method to mitigate their impact. However, it should be noted that whilst missing values may themselves bias any analyses, their estimation is likely to introduce an additional source of bias.

Hrydziuszko and Viant [290] note that missing values can be extensive throughout a metabolomics data set, with their occurrence generally linked to *m/z* and signal intensity. They analysed the efficacy of eight different methods for the estimation of zero intensity values, with complexity ranging from a simple substitution with a small value (0.01), to a more complex three-stage approach involving Bayesian PCA. The favoured method of the eight was based on a k-nearest neighbours methodology, which garners an estimate for the missing value from a combination of the k most similar observations.

3.4.4 Isotopic grouping

The presence of heavy elemental isotopes in molecules results in the formation of an isotopic distribution, whereby multiple ions represent the same molecule. Coupled with elemental isotopic profiles, this distribution can be used to provide structural clues regarding an ion's identity. Isotopic peaks can be grouped on a file-by-file basis, where each peak is examined for neighbours with the same retention time and at logical *m/z* increments (for example, ±0.5, ±1). Where such peaks are found, they can be grouped together. The disadvantage of this method is that it is hard to be sure that they should be grouped together; relative intensities of peaks can provide only minor clues regarding the sensibleness attached to a grouping.

A more intelligent grouping of isotopic peaks can be effected if the all of the data files are considered together. Multiple peaks that are found across multiple observations, in the same spatial proximity, should be correlated to each other if they form part of the same isotopic cluster; the absolute intensities of the M and M+1 peaks will vary from sample to sample, yet the M:M+1 ratio should remain constant. Peaks that do not follow such a pattern are likely to belong to another cluster.

Peaks derived from chemical and electrical noise can also be removed during the isotoping stage; features without an isotopic distribution can be removed as a good way to reduce the number of features that are carried forward. Whilst most metabolites can be expected to contain carbon atoms and hence have an isotopic distribution reflecting the presence of ^{13}C, this way of peak filtering will discriminate against the, albeit few, low molecular weight molecules and those whose M+1 peak falls below the limit of detection. However, for experiments that record *m/z* data starting from *m/z* 100, few molecules are likely to be 'observably isotope free'.

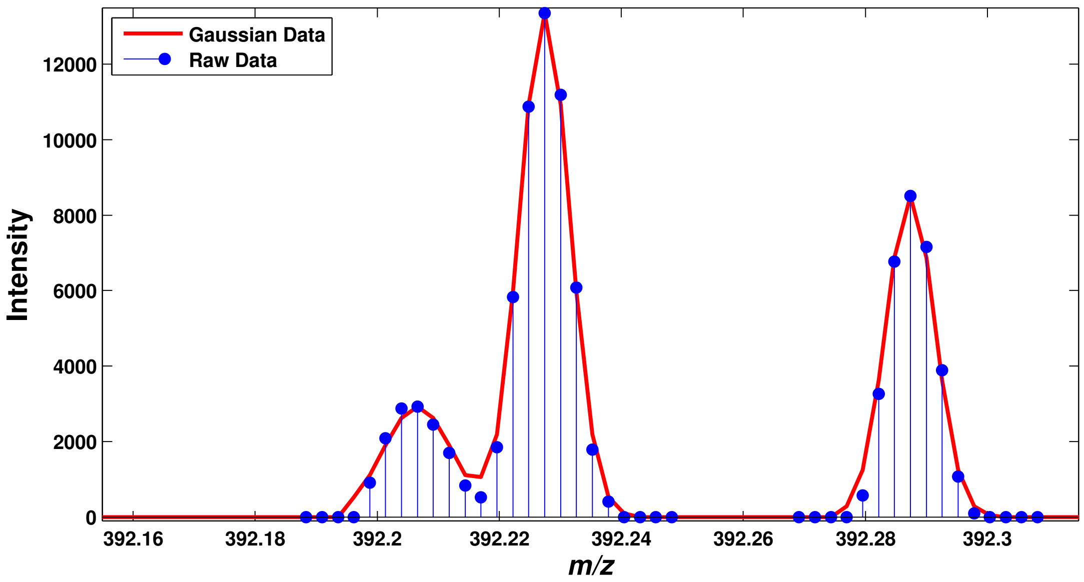


FIGURE 3.11: The original data is shown by the blue stems, and the Gaussian functions calculated to model the original data are shown by the solid red line. Each peak consists of at least 10 data points, whereas each Gaussian function is defined by only 3 parameters.

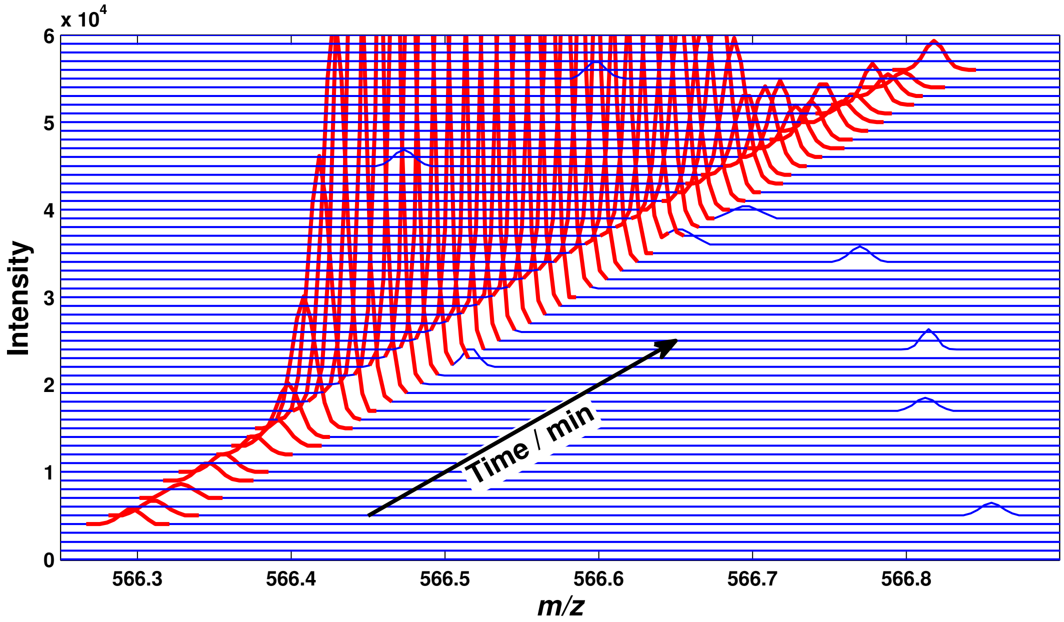


FIGURE 3.12: Each horizontal line represents an individual mass spectrum, with those sections that are coloured red grouped into a single chromatographic peak. The red sections are not contiguous, yet are grouped into a single peak. Nine individual peaks are also present, but these can be safely categorised as noise peaks as no other values with similar *m/z* appear in preceding or succeeding scans.

3.4.5 Normalisation

The normalisation of data is a crucial consideration prior to any statistical analysis. The total ion count is a common metric with which to scale data globally, as is the intensity of an exogenous compound spiked into a sample as an internal standard. In studies of urine, it is often reported that creatinine acts as a good proxy for an internal standard. This must, however, be gauged against the possibility that levels may have been unintentionally perturbed [14].

3.4.6 Assignment

Feature assignment in non-targeted analysis can make use of *m/z* and isotopic information, whilst targeted studies will often have the advantage of additional fragmentations. The number of features that are detected in non-targeted analysis is likely to be plentiful, and feature assignment is more judiciously performed alongside statistical techniques that identify highly discriminatory variables, i.e. those that show significant between-group differences.

Accurate mass measurements, even to within a few ppm, seldom provide a single suitable elemental composition. The incorporation of isotopic abundances can help to reduce the list of elemental combinations from many hundreds to a few tens. Kind and Fiehn [35] summarise the extent of the issue, along with the presentation of seven rules designed to facilitate elemental elucidation. The anti-cancer agent taxol has a monoisotopic mass of 853.906 Da, and 1418 formulae can be matched to this mass within a 2 ppm tolerance and a palette limited to C, H, N, O, P, S, F, Cl and Br. This quantity can be whittled down to 29 candidates after inspection of the isotopic information, where the correct formula was ranked 25th. In spite of it not being the single best result, none of the other suggestions appeared in the PubChem database [291].

ChemSpider [292], the freely available database, has been adapted by Little et al. [293] to search for what they term ‘known unknowns’. These are known to the chemical community that drives ChemSpider yet are not known to the researcher performing the investigation. Their adaptation of the user interface returns relevant results to the query, which are sorted according to the number of references that are associated with the entries. Whilst the number of references is no guarantee of success, the authors conclude that the modification is a useful addition to the database.

Both of the methods presented above rely on databases to enhance or reduce the likelihood of specific assignments. It should be stressed that the appearance, and indeed absence, of a molecule in a database query result is not a suitable criterion for outright acceptance or rejection, but rather a contribution to an overall assignment probability.

3.4.6.1 Spectral databases

In spite of likely differences in acquisition conditions, spectral databases can prove invaluable for assignment purposes. The METLIN [294] database is an extremely comprehensive service that contains upwards of 60,000 metabolite spectra, where 10,000 of these have high-resolution MS/MS data. Results from METLIN are cross-referenced, where available, to other databases such as KEGG [295], HMDB [114] and Lipid Maps [296]. The capability of METLIN to perform batch searches for positive and negative ions, and neutral molecules is invaluable to any researcher with a lot of data.

KEGG, or Kyoto Encyclopaedia of Genes and Genomes, provides reaction and pathway information for small molecules. It is an excellent 'systems biology' resource for, amongst others, linking metabolites to genes. The Human Metabolome Database (HMDB), along with the Biological Magnetic Resonance Data Bank (BMRB) [116] are two online resources with both MS and NMR data. Combinations of multiple spectral resources can also be found in SDBS (Spectral Database System) [297], but unlike in HMDB and BMRB, the data cannot be downloaded for offline use.

3.4.7 Software

There are various open source software packages that are capable of processing series of raw mass chromatograms. Perhaps the most widely used is XCMS, which is available as both a command line-driven R package (platform independent) [298] and a web version, which offers a user interface [299]. A useful tutorial to the R version has been provided [300], which covers processing stages from importing data files to the visualisation of peaks. The command line-driven approach, whilst potentially off-putting to many users, is advantageous in its simplicity. Furthermore, the batch processing of large data sets can be effected by a simple script, as well as easy access to multivariate statistics.

MZmine [301] is a Java application, (also platform independent) run through a user interface, although a batch mode is configurable through a series of drop-down menus. A disadvantage of MZmine compared to XCMS is that some parameters are provided with no default values, nor a reasonable suggestion. Whilst defaults are unlikely to be optimal for all data, they do provide an initial starting point against which subsequent changes can be made. Various other software packages are available, and a more comprehensive analysis of features has been made by Castillo et al. [231].

3.5 Data analysis

The advance in measurement technologies has resulted in the generation of larger quantities of data. Such megavariate data has spawned new chemometric approaches that are capable of handling these vast matrices. Whilst the mathematics is often hidden to the user, it is important to gain an understanding of the mechanism of the techniques, such that they are used appropriately. Methodologies may broadly be divided into two groups: supervised and unsupervised. Techniques in the former case use class information to discriminate between observations, whilst the latter assume no *a priori* knowledge regarding an observation.

3.5.1 Cross validation

Cross validation methods are required in the correct application of supervised techniques, where they are used in order to gauge the error rate of a particular model. Supervised techniques can be over-fitted, which thus renders them incapable of being able to adequately generalise to unseen, or independent, data. There are various methods in which a data set can be divided into the testing and training data sets.

3.5.1.1 Leave-one-out

Perhaps the most suitable method for small data sets, leave-one-out cross validation (LOOCV) involves the averaging of k results, where k is the number of observations. Each training data set consists of $k - 1$ samples, and the corresponding test sets are formed from the singly omitted observation. The predictive capabilities and errors of each of the k models are tested by reference to the omitted sample.

3.5.1.2 *k*-fold

An extension of LOOCV, the full data set is (randomly) divided into k equally sized groups, where k is less than the number of observations. Each of the groups is used once in a testing capacity, and $k - 1$ times to train the models. Where there are technical replicates (essentially the same sample measured at a different time) in a data set, it is important to consider the effect of bias that may arise due to broadly similar samples being in both training and testing data sets.

3.5.1.3 Partition

The partition approach to cross validation uses n repetitions, where each of the n training data sets is formed from a random subset of the full data. The random nature of subset selection leads to the possibility that not all observations are used equally in each data set: some samples may always appear in the testing data set, whilst others may be used exclusively to train the n models.

3.5.1.4 Bootstrapping

Bootstrapping is a resampling method that allows for error estimates to be made in cross validation. The method is based on repetitive resampling of observations, with replication of observations within a training set being possible (unlike in the partition method). For datasets with N observations, a training set is created consisting of N randomly selected observations, where the same observation can be selected many times. The test set, created from all observations that do not feature in the training set, is used to estimate the error. The method is repeated multiple times, such that error values are derived for each resampling, from which a global error can be derived.

3.5.1.5 External validation

The above methods refer to the application of internal cross validation, whereby all data from a study is used to validate a predictive model. The alternative to this is external validation, where a single model is produced from a training set, and gauged by the test set. As only one iteration is performed the procedure is quicker, and the test set it wholly independent as it was not used to create the model. However, the model is essentially ephemeral, and may be due to a 'lucky' choice in choosing the training set. The use of external validation is perhaps best suited to data sets with a large quantity of observations; where this is not the case, the predictive power of the model is diminished [302].

3.5.2 Univariate analysis

Univariate methods are applied to multiple observations of a single variable. They are often formulated around a null hypothesis, and the data may either reject this assertion, or fail to reject it (distinct from 'pass') whereby the data was insufficiently convincing.

Statistical tests are categorised into two groups: parametric and non-parametric. The former applies to tests that make an inherent assumption about the statistical distribution of the data, whilst for the latter there is no such assumption, and these may be considered to be distribution-free statistics.

3.5.2.1 t-tests

Student's t-test is applied to normally distributed data, and can be used to compare the means of two groups of independent observations. In practice, it assesses if the mean of n_A observations from group A differs significantly from the mean of n_B observations from group B. Each of the observations should be independent, which precludes measurements in a 'before-and-after' study (e.g. pre- and post-intervention observations), as well as those where the observations may be grouped into, for example, class.

Where such conditions are not true, the repeated measures (also paired) t-test is applied. As well as the normality assumption, Student's t-test relies on approximately equal variances. Where such a condition cannot be reasonably assumed, then Welch's t-test should be applied instead.

3.5.2.2 Mann-Whitney test

The Mann-Whitney U test is the non-parametric equivalent of a t-test. Values from both groups are ranked, and the ranks from each group are summed: a large disparity in the sums indicates a rejection of the null hypothesis. The paired equivalent is the Wilcoxon signed-rank test.

3.5.2.3 Analysis of variance

The analysis of variance (ANOVA) is a statistical procedure to identify sources of variance. Its most simple form, one-way ANOVA assesses, for example, whether analyses carried out by different people vary significantly. It is limited in that it is likely that multiple sources of variation are in operation, and thus two- and n-way ANOVA may be more appropriate to assess the contribution of multiple events, either individually or cooperatively. The Kruskal-Wallis test is the non-parametric equivalent of ANOVA, and has been applied to identify differences in amino acid levels in black, oolong and green teas [259]. Zand et al. [303] have applied ANOVA to assess the variation between groups and replicates, whilst Ricard et al. [304] used ANOVA to assess the inter- and intra-day variation of QC samples.

3.5.2.4 Multiple testing problem

A difference between, for example, the means of two groups may be calculated as significant at a p-value of 0.05. As such, there is a 5% probability that the difference is due only to chance, i.e. a false positive. If multiple tests are carried out, then it can be rationalised that 5% of all tests will result in falsely positive results.

In order to overcome the multiple testing problem, adjusted p-values can be derived so as to be able to account for the expected proportion of false discoveries. The Bonferroni correction adjusts the confidence interval by dividing it by the number of tests being performed, such that now fewer significant differences will be identified and the likelihood of a false discovery is therefore lower. The Bonferroni correction typically increases the false negative rate, as the significance criterion is set too unrealistically. The false discovery rate (FDR) method calculates a q-value based on the distribution of p-values of all of the tests. This q-value expresses the proportion of significant tests resulting in false discoveries. Whilst FDR is less strict than the Bonferroni method (i.e. more false positives with FDR), crucially the number of false negatives is also lower.

3.5.3 Multivariate analysis

Multivariate analysis can provide information about multiple variables and how they are inter-related. Perhaps the simplest form of multivariate analysis is bivariate regression, where the change in one variable can be compared to that of another.

3.5.3.1 Partial least squares

Partial least squares (PLS, Section **??**), or projection to latent structures [17], is a supervised technique that relates variance in X to that in Y. Figure 3.13 shows the scores (S), weights (W) and loadings (L) for PLS. The X scores ($\mathrm{S_X}$) are calculated according to Equation 3.4, where $\mathrm{W_X}$ denotes the X weights. The product of the scores and loadings should minimise the error (ϵ) in the approximation of the original X variables (Equation 3.5). The Y variables can be approximated by their respective scores and weights, but the X scores should approximate well to the Y scores, such that their combination (Equation 3.6) reduces the error. By substitution of the X scores, a regression-like equation can be derived, as is shown in Equation 3.7.

$$S_X = X \times W_X \tag{3.4}$$

$$X = S_X \times L_X + \epsilon \tag{3.5}$$

$$Y = S_X \times W_Y + \epsilon \tag{3.6}$$

$$Y = X \times (W_X \times W_Y) + \epsilon \tag{3.7}$$

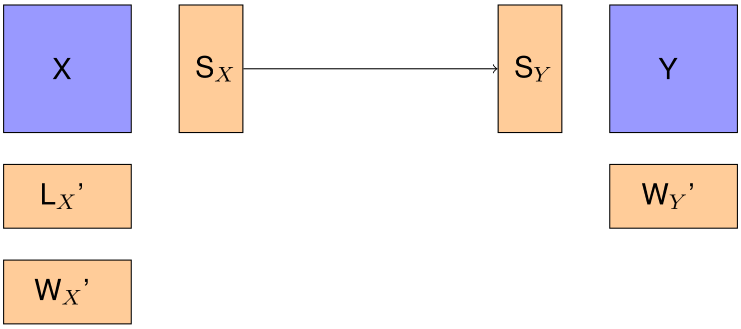


FIGURE 3.13: Representation of PLSR, adapted from Wold et al. [17]. The weights (W), loadings (L) and scores (S) are shown, for the response (X) and predictor variables (Y).

3.5.3.2 Principal components analysis

Principal components analysis (PCA, Section 2.6.2.1) [16] is an unsupervised method that takes linear combinations of existing variables, such that the variance is maximised. The first principal component (PC) is formed from the linear combination that explains the maximum amount of variance. Subsequent PCs are formed to explain as much of remaining variance whilst being uncorrelated to preceding PCs. As PCA is a variance-based technique, the scaling applied to the data set can influence subsequent results. Raw LC-MS data variables are likely to contain a large range of variances; whilst high-variance variables will dominate a subsequent PCA, they may be overwhelming more informative variables with lower variabilities. Initial unit-variance scaling, by dividing each variable by its variance, gives each element an equal influence in PCA. The disadvantage is that low-intensity noisy variables have a greater influence in PCA; hence without sufficient denoising, unit-variance scaled data may only reveal trends in noise levels.

3.5.3.3 Independent components analysis

In PCA, each of the components is determined such that they are uncorrelated with each other. In independent components analysis (ICA), the constraint is extended such that now they must be independent. As independence is a stronger mathematical concept than non-correlation, the independent components (ICs) potentially contain complementary information about the system. The disadvantage of ICA is that it is only practically applied to low-dimensional data sets, such that spectral or chromatographic data is singularly unsuitable due to the large number of variables. A suitable approach would be to use an initial data reduction method, such as PCA, to allow inclusion of only the most significant variables in ICA.

A two-stage PCA-ICA has been performed by Scholz et al. [305], and is shown to offer an improved interpretation of the results than that offered by PCA alone. Estimating the number of PCs to take into the second stage is crucial; too many and ICA is working with insignificant and noisy variables, whilst too few will result in key information being omitted. The optimal number of PCs to include was determined by the number of resultant ICs that were found to be negatively kurtotic[1]. ICs with negative kurtoses are more informative as they potentially indicate, for example, a variable composed of high and low concentrations. In terms of separation, ICs 1 and 2 are shown to exhibit a better between-class discrimination than is seen in PCA, whilst IC 3 reveals information related to the sample run time.

Krumsiek et al. [306] have applied Bayesian ICA to metabolomics data. The technique differs from standard ICA in that Bayesian techniques are used to enforce biologically intelligent assumptions (e.g. non-negative concentrations), and to estimate the optimal number of components to adequately describe the data. The study uses well-annotated metabolomics data, where the identity and biological pathways of the molecules are fully characterised. In this study, the number of variables is significantly less than observations, such that ICA can be applied without prior data reduction. Upon inspection of the contributions of individual variables to each independent component, the eight ICs are demonstrably attributable to either a class of molecule, such as amino acids, or to a group of molecules that is linked to a specific process. The results from Bayesian ICA were found to be more biologically consistent than those from PCA and k-means clustering approaches. Whilst not observed in PCA, the second IC, dominated by branched amino acids, was found to

[1] Kurtosis defines the degree of sharpness around a frequency distribution. It is often defined as the fourth moment of a distribution; the preceding three are mean, variance and skewness.

be negatively correlated to high-density lipoprotein (HDL) measures, which may help to augment understanding of obesity and diabetes. The authors conclude that Bayesian ICA has the potential to supersede the omnipresent PCA. However, it is unclear how well the approach can work for partially annotated data.

3.5.3.4 Multiple linear regression

Multiple linear regression (MLR) models a response variable (y) to a series of predictor variables (X). It is an extension of simple linear regression that attempts to find a relationship between two variables, according to Equation 3.8. MLR introduces additional explanatory variables where the solution attempts to satisfy Equation 3.9.

$$y = ax + c \tag{3.8}$$

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + c \tag{3.9}$$

MLR is not suitable for highly correlated data, as is likely to be true of an LC-MS variable list. The method is best applied to uncorrelated variables, either the output from PCA or a subset of original variables chosen by a feature selection method such as genetic programming. The logical extension of MLR is canonical correlation analysis, which maximises the correlation between multiple predictor and multiple response variables. It is discussed amongst the multiblock methods below.

3.5.4 Multiblock methods

Positive- and negative-mode mass spectrometry data that describes the same samples can be considered as complementary. The same applies to the use of multiple columns, and to different instrumentation, such as nuclear magnetic resonance

(NMR) spectroscopy. Analysing each of these individually will not reveal inter-block relationships and so the full complementarity of the data cannot be exploited. So-called high level data fusion methods allow all variables to contribute to a single multivariate model.

3.5.4.1 Hierarchical methods

Hierarchical, or two stage, PCA and PLS are multi-block methods capable of analysing multiple matrices [18, 19], which may be originally or artificially distinct. An example of the former would be be data collected under different instrumental conditions, such as positive and negative mode, or with different instruments entirely. Artificially distinct data can be created by division of a dataset from a single source into more meaningful blocks. An example of this, applied to infrared spectra, is given by Janné et al. [307] where they divide spectra according to likely functional groups. Whilst the study does not use MS data, it serves as a well-written introduction to multiblock methods.

Two-stage hierarchical models essentially involve the combination of individual multivariate analyses, as is shown in Figure 3.14. A series of data blocks undergoes an initial data reduction stage, such as PCA, to produce a series of new variables. Subsets of these variables are then combined, and subjected to another analysis. This second stage allows for a trend in the data to be evaluated with reference to each block. Forshed et al. [20] have applied such approaches to the fusion of LC-MS and ^{1}H-NMR data. They found that the two-stage approach improved the classification success and class separation, in comparison to low level and individual analyses of the data matrices. The use of various scaling regimes is also discussed.

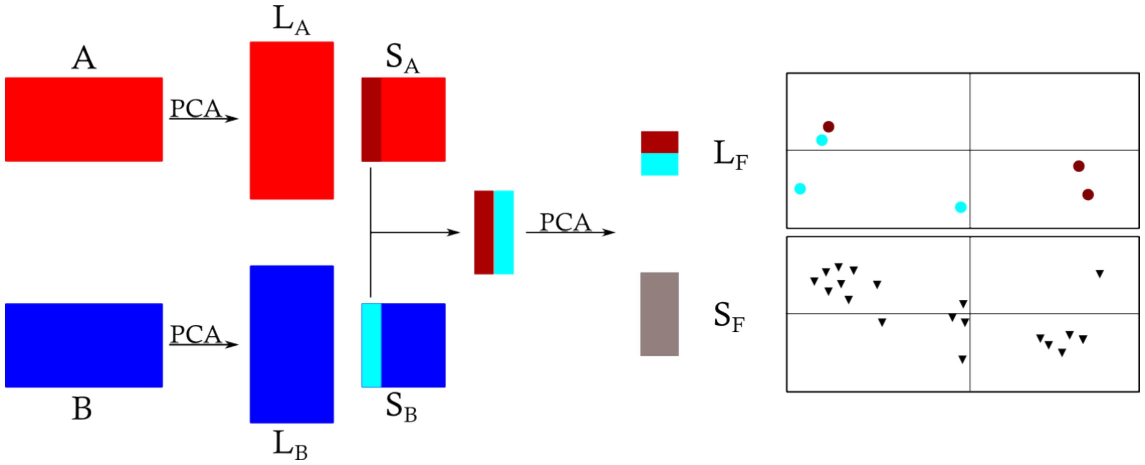


FIGURE 3.14: A schematic for two-stage PCA. Both A and B matrices are subjected to an individual PCA, which produces loadings (L) and scores (S). A subset of S_A and S_B are combined for a second analysis. Trends within the data can be visualised by plotting the fused scores (S_F). The fused loadings, L_F, indicate which principal components from the original analyses contribute to any observed trends.

3.5.4.2 Canonical correlation analysis

Canonical correlation analysis (CCA [28]) seeks to find linear combinations of correlated variables across two data matrices. For two matrices X and Y, the first canonical pair (CP) is formed from two linear combination of x and y variables, where this combination maximises the correlation, r; this is expressed in Equation 3.10.

$$\begin{aligned} u &= a_1x_1 + a_2x_2 + \cdots + a_px_p \\ v &= b_1y_1 + b_2y_2 + \cdots + b_qy_q \\ r(u,v) &\rightarrow max \end{aligned} \tag{3.10}$$

Subsequent CPs are formed similarly, subject to the constraint that they are uncorrelated with previous ones. The number of CPs that can be formed is dictated by the minimum number of variables in X or Y. The technique is incompatible with rank deficient matrices, such as those that contain many correlated variables. Thus, a two-stage approach is required, such as through the use of PCA or other data reduction techniques.

Doeswijk et al. [32] applied a two-stage approach to the fusion of GC- and LC-MS data matrices. They used partial least squares regression as the initial data reduction method, and performed CCA on the combined scores. The results from CCA are used to identify molecules that are related to the original regressor variable used in PLS-R.

Regularised canonical correlation analysis (RCCA) differs from CCA in that it can operate on rank deficient matrices, and is hence suitable for non-targeted MS data. Yamamoto et al. [308] have applied RCCA to a study of green tea using GC-MS. By means of comparison to PLS-R, RCCA requires the inclusion of fewer latent variables into a predictive model, which suggests that the redundancy in the PLS model is high. Whilst the obvious drawback to RCCA is the calculation time, the adaptation of the method to involve kernels has been demonstrated to reduce the analysis time from hours to seconds.

3.6 Future directions

The evolution of liquid chromatography and mass spectrometry has resulted in a powerful analytical platform with exceptional limits of detection. Advances in instrumentation are resulting in more accurate and better resolved *m/z* peaks and greater sensitivity leads to enhanced detection such that the analytical coverage of a mixture is constantly increasing. These continual improvements rely on the commensurate development of data processing and interpretive tools, such that the maximal amount of spectral information can be extracted and exploited.

The exploitation of spectral information typically depends on the assignments that can be made for specific variables. Whilst comparisons to certified reference materials provide the most suitable proof of confirmation, this approach is neither afford-able nor achievable for all analyses. Many studies, therefore, rely on commercial or freely-available spectral databases with which to make assignments. Communal pooling of resources may also help, and techniques aimed at enhancing the reporting of metadata have been proposed.

The interpretation of megavariate datasets, and the identification of pertinent features, relies on multivariate statistical techniques. These help to focus the analysis and find features that are deemed to be of interest to the experiment. Whilst LC-MS is recognised for its excellent sensitivity, it is essentially a biased technique as choices between chromatographic column and ion detection mode limit the range of analytes that can be discovered. Complementary analyses involving multiple columns and positive and negative ion modes help to enhance the coverage, but also add to the data burden. Independent analyses of the various complementary datasets may fail to fully exploit the inherent information present. Instead, the analysis through data fusion offers the possibility of combining the datasets such that a more holistic conclusion about the samples can be reached.

Chapter 4

Quality Determinants in Pea Seeds

How luscious lies the
pea within the pod.

Emily Dickinson

4.1 Data

The data used throughout much of the thesis is taken from a larger study, the aim of which is to enhance sustainable agriculture through a better understanding of the quality traits within pea seeds. A more in-depth understanding of what makes high quality peas enhances the marketability of certain pea cultivars, which benefits the growers. One of the main benefits of pea plants, and other legumes, is their nitrogen fixing ability which provides a natural way to restore biologically available nitrogen to the soil. In days of heavy agriculture, a sensible crop rotation strategy is as important as ever.

The current measure of a pea's maturity is given by its tenderometer (TR) measurement, which is a proxy for a seed's water content. Being able to define a pea by its contents (i.e. other than water) would be useful in being able to assess the quality of the sample. The study employs genetic, proteomic and metabolomic approaches in order to identify specific determinants of quality within pea seeds in order to further the understanding of the plants. Only the metabolomics data sets have been used here, and the relevant details are provided for the samples, the extraction procedure, and experimental conditions.

4.1.1 Samples

The number of pea samples used in this study is 54. These consist of various cultivars, and each sample has its own TR measurement. An in-house reference (IHR) standard was periodically analysed throughout the experiments to serve as a QC sample, for both NMR and LC-MS.

TABLE 4.1: Metadata for the pea samples. The TR value is the tenderness value for each sample, and serves a measure of maturity.

Cultivar	Run order	TR	Cultivar	Run order	TR
Oasis-1	4, 17	89.5	KirosUNT-2	88, 101	100
Yoda-4	5, 18	127.5	Oracle	89, 102	97.6
Bikini-1	11, 25	90	Peregrine	90, 103	79.2
Oasis-3	14, 28	116	Peregrine	91, 104	80
Mondial-2	30, 43	97.5	Bikini-4	108, 121	123
Yoda-2	31, 44	104.5	Mondial-1	109, 122	88.5
Recital-1	32, 45	83.5	Mondial-4	110, 123	123
Yoda-3	33, 46	114.5	Recital-2	111, 124	110
Bikini-3	34, 47	120	Yoda-1	112, 125	89
Anubis-2	35, 48	101.3	KirosUNT-1	113, 126	88
Anubis-3	36, 49	101.6	KirosUNT-3	114, 127	106
Peregrine	37, 50	96.2	KirosUNT-4	115, 128	115
Peregrine	38, 51	87.2	Anubis-1	116, 129	99.6
Peregrine	39, 52	97.6	Avola-1	117, 130	97
Oracle	56, 69	144.4	Avola-2	134, 147	103.3
Oracle	57, 70	125.6	Avola-3	135, 148	103.6
Peregrine	58, 71	108.4	Bikini-1	136, 149	101.6
Peregrine	59, 72	97.6	Bikini-2	137, 150	111
Oracle	60, 73	115.2	Oasis-1	138, 151	96.6
Oracle	61, 74	88.4	Oasis-2	139, 152	102.3
Peregrine	62, 75	103.6	Tendrilla-1	140, 153	105.6
Oracle	63, 76	105.6	Tendrilla-2	141, 154	108.6
Peregrine	64, 77	86.4	Waverex-1	142, 155	117
Peregrine	65, 78	91.2	Zephyr-1	143, 156	100.6
Bikini-2	82, 95	99.5	IHR	2, 22	105
Mondial-3	83, 96	108	IHR	30, 43	105
Oasis-2	84, 97	101.5	IHR	56, 69	105
Oasis-4	85, 98	125	IHR	82, 94	105
Recital-3	86, 99	113.5	IHR	108, 121	105
Recital-4	87, 100	134	IHR	134, 147	105

4.1.2 Extraction procedure

Upon delivery, samples were stored at -20 °C until ready for preparation. All samples were freeze dried for approximately 48 h, and subsequently ground into a fine powder. 1.5 mL of 1:1 methanol:water was added to 150 mg of this powder, and the mixture was vortexed for 30 min. Following this, the sample was centrifuged at 20817 *g* for 10 min, and separate aliquots of the resulting supernatant used for NMR and LC-MS analyses.

4.1.2.1 LC-MS

100 μL of supernatant was diluted one part in ten with 1:1 methanol:water. Triclabendazole ($C_{14}H_9N_2OSCl_3$) is used as an internal standard, spiked into all samples at a concentration of 1 μg / mL.

4.1.2.2 NMR

The liquid from a 900 μL sample of the supernatant was removed by drying with a stream of nitrogen gas for 1 h. The remaining sample was lyophilised to remove residual moisture and the dried sample was reconstituted in 700 μL of phosphate buffer with added internal standard (250 mM potassium phosphate, pH = 7.0, and 0.5 mM trimethylsilyl propanoic acid, TSP, both dissolved in D_2O). Following resolvation, the sample was centrifuged at 2300 *g* for 5 min and 540 μL of the supernatant was transferred to an NMR tube, to which 60 μL of 10 mM sodium azide (dissolved in D_2O) was added.

4.1.3 Experimental conditions

4.1.3.1 LC-MS

For reversed-phase LC a Waters Sunfire C_{18} HPLC column (150 mm × 2.1 mm × 3 μm) was used and for HILIC the column employed was a SeQuant ZIC-HILIC (100 mm × 2.1 mm × 3.5 μm). For both columns the flow rate was 250 μL / min, 20 μL of sample was injected, and the column temperature was set to 30 °C. The two mobile phases were water (A) and acetonitrile (B). To both, 0.1% formic acid was added to encourage ionisation. The gradients employed are summarised in Table 4.2. Both positive and negative ion modes were recorded for electrospray ionisation (ESI) using a Thermo Exactive. Data were acquired over an *m/z* range of 50–1000, at a resolution of 50,000 FWHM. All samples were analysed in duplicate, and Chapter 6 details the implemented processing procedures.

TABLE 4.2: The solvent gradients for both LC types. Mobile phases A and B are described in the text.

(A) Reversed-phase

Time / min	% A	% B
0.0	90	10
5.0	90	10
20.0	0	100
25.0	0	100
25.1	90	10
30.0	90	10

(B) HILIC

Time / min	% A	% B
0.0	10	90
35.0	70	30
40.0	70	30
41.1	90	10
50.0	10	90

4.1.3.2 NMR

All spectra were acquired using an 11.7 T Bruker 500 MHz NMR spectrometer, equipped with a 16 bit digitiser (maximum sampling rate of 2 MHz) capable of generating Z magnetic field gradients of up to 50 G cm^{-1}. A 5 mm coldprobe with cooled ^{13}C and ^{1}H preamplifiers (Bruker TCI cryoprobe) was tuned to detect ^{1}H resonances at 500.13 MHz and ^{13}C resonances at 125.76 MHz. The probe was manually tuned and matched and the magnetic field homogeneity optimised using up to 34 shims. Manual data acquisition was used throughout.

For one-dimensional ^{1}H NMR, spectra were acquired using the Bruker pulse program zgpr. The following acquisition parameters were used for data collection: 9.2 μs 90° observation pulse length; 15.01 ppm spectral width; 128 recorded FIDs; 2 unrecorded FIDs; 32768 data points in FID (complex); 10 s relaxation delay; digital quadrature detection (DQD); 4.647 ppm offset frequency. These parameters gave a total experiment time of 26m 32s. A sine-bell shaped window function phase shifted by 90° was applied over all points prior to Fourier transformation, phase and baseline correction.

Two-dimensional phase cycled ^{13}C-^{1}H heteronuclear single quantum coherence (HSQC) spectra were acquired using the Bruker pulse sequence hsqcphpr. This experiment performs on-resonance presaturation of the ^{1}H signal from residual water and an INEPT sequence for spectral editing. The following acquisition parameters were used: indirect nucleus (F1) ^{13}C; direct nucleus (F2) ^{1}H; GARP decoupling method; decoupled nuclei ^{13}C-^{1}H; ±15 kHz decoupling bandwidth; 9.2 μs 90° ^{1}H pulse length; 16.5 μs 90° ^{13}C pulse length; J_{C-H} 145 Hz; ^{1}H spectral width 13.330 ppm; ^{13}C spectral width 179.990 ppm; 2 s relaxation delay; 64 recorded FIDs per t1 increment; 16 unrecorded FIDs; 395 t1 increments; 1536 data points per FID (complex); 4.647 ppm ^{1}H offset frequency; 90 ppm ^{13}C offset frequency; DQD acquisition mode; t1 3 μs; t1 increment 22.09 μs; States-TPPI quadrature detection. These parameters gave a total experiment time of 15h 24m 30s.

Two-dimensional phase cycled ^{1}H-^{1}H total correlation spectroscopy (TOCSY) spectra were acquired using the Bruker pulse sequence mlevphpr. This experiment performs on-resonance presaturation of the ^{1}H signal from residual water. The following acquisition parameters were used: indirect nucleus (F1) ^{1}H; direct nucleus (F2) ^{1}H; 10.13 μs 90° ^{1}H pulse length; ^{1}H spectral width (F1 and F2) 13.330 ppm; 1.5 s relaxation delay; 32 recorded FIDs per t1 increment; 16 unrecorded FIDs; 512 t1 increments; 4096 data points per FID (complex); 4.709 ppm ^{1}H offset frequency; DQD acquisition mode; t1 3 μs; t1 increment 150 μs; spinlock duration 100 ms; trim pulse length 2 ms; spinlock field strength 7.142 kHz; States-TPPI quadrature detection. These parameters gave a total experiment time of 8h 55m 26s. HSQC and TOCSY spectra were processed using a sine-bell shaped window function phase shifted by 90° over all points. These data were zero filled to give a real data matrix size of 4096 × 2048 points prior to Fourier transformation, phase and baseline correction.

NMR spectral processing was carried out as described by Davis et al. [54] for 1D spectra, and 2D spectra were processed by the method described in Chapter 5.

4.2 Continuous classification

In order to test the classification power of a model, it is necessary that observations have a defined class, and that there are multiple observations per group. The within- and between-class scatter of groups can be used to classify new samples from a test set, or to gauge how well a multivariate method separates observations between classes. When the response variable is continuous and cannot be defined by discrete classes, a different measure of how well the response is related to the measured data is required.

The nature of supervised techniques, for example PLS, is such that the correlation between the response and scores is a valid approach; for unsupervised, variance-based methods this is an inappropriate way to determine how well the scores approximate to the variance of interest. In regression and discriminant analysis mode, it is expected that the first PLS component is most related to the classification variable, and the measure of correlation is appropriate to techniques where much of the variance is captured by a single component.

In unsupervised techniques it is often the case that the experimental variance is spread across more than one score, as is shown in Figure 4.1. The calculation of Pearson's product moment correlation (ρ) would reveal two components with moderate values, but neither one of these effectively describes the strength of the relationship. The same issue arrises when analysing scores produced by canonical correlation analysis (CCA, Section 3.5.4.2), as whilst the agreement within each canonical pair (CP) is given by the canonical correlation coefficient, this does not reflect how well the CP is related to an external variable.

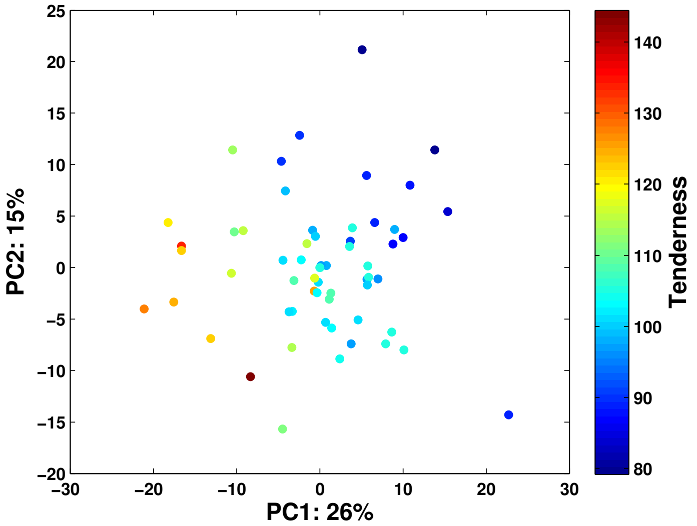


FIGURE 4.1: PCA scores plot from analysis of the HSQC data from the pea study. It can be seen that both components exhibit changes in tenderness values, rather than it being confined to a single component.

For unsupervised methods, therefore, an alternative metric would be more appropriate, such as one capable of expressing the agreement between a continuous classification variable and several scores produced by a multivariate analytical technique. A new method is proposed, termed the D^2 value, that is calculated as the Pearson product moment correlation between two distance/dissimilarity matrices. One matrix is calculated using several scores, and the other from the classification variable(s).

The Mantel test [309] is often applied as the *de facto* metric for the comparison of distance matrices (X and Y). It is commonly applied in bio- and socio-logical contexts, and a few instances in the chemical literature have been reported [238, 310–312]. The statistic takes into account the fact that the pairwise distances are not independent of each other, as alteration in one observation will have a knock-on effect on many distance values. In order to mitigate such dependence, one distance matrix (e.g. X) is randomly permuted many times (where 'many' helps to define the statistic's associated p-value) and the Pearson product moment correlation between Y and each permutation of X is calculated. The random permutations help to test for the null hypothesis, which states that there is no correlation between the two matrices. If there is no similarity, then it is possible that the permutation's correlation is greater than the correlation achieved using the full data. The Mantel test returns a significance of the correlation, and is determined by the frequency at which random permutations achieve greater correlations than that from the non-permuted form. Low p-values allow rejection of the null hypothesis and acceptance of the alternate hypothesis which states that there is a correlation between the two matrices. The test has been implemented using 10,000 permutations, and all calculations returned a significance value of p = 0.0001.

The D^2 value reveals how well the pattern seen in scores is related to the classification variable. The optimal number of scores to include is given by the combination with the greatest D^2 value.

4.2.1 Distance measures

There are various ways to measure the multidimensional difference between observations. The Euclidean distance between two points in n dimensions is given by Equation 4.1. It is the true difference defined in terms of multidimensional space.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.1}$$

The city block difference measure does not measure the shortest distance, but rather the sum of the absolute differences between two coordinates. Where the Euclidean, or Pythagorean, distance measures the hypotenuse of a right-angled triangle (in two dimensions), the city block distance is the sum of the two other sides of the triangle. The Mahalanobis distance is calculated according to the correlation between variables. By definition, PCs are uncorrelated with each other, so the metric may not be the most appropriate when applied to PCA. The Chebychev distance for two observations is the maximum distance in any of the various dimensions. As such, the method seeks to accentuate differences between observations, and is susceptible to outliers.

In order to determine the most appropriate distance metric, all of the listed metrics were used to calculate the D^2 values. Successively larger number of PCs were included in the calculations, up to a maximum of 10. Whilst it is acknowledged that the first PC may be detrimental to the metric, the aim is to create a metric that is as unsupervised as possible; therefore, the only consideration is how many, rather than which, components to include.

Figure 4.2 shows the D^2 values obtained for the various distance metrics, when up to 10 PCs have been included. The decay of the Mahalanobis distance with increasing PCs is unsurprising due to the fact that PCs are uncorrelated with each other. The city block metric also deteriorates with more components. The Chebychev metric stabilises with the first five components, where additional PCs do little to affect the distances between observations. The Euclidean distance also stabilises, but after many more components (not shown). The highest D^2 value, albeit marginally, is obtained when using 2 components and measured by the Euclidean distance. In terms of the best metric, the choice is between the Chebychev and Euclidean measures. The advantage of the Euclidean measure is that it is the actual distance in terms of all dimensions, rather than the Chebychev metric where only the largest distance is used. For this reason, and its greater resilience to outliers, the Euclidean measure is used in favour of the others. As, in this instance, there is only one response variable, the choice of metric for the second distance matrix is unimportant.

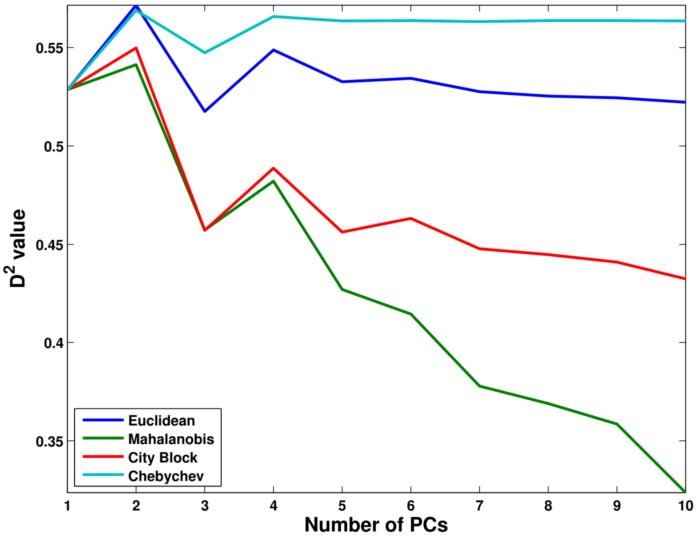


FIGURE 4.2: The D^2 values for various distance metrics, calculated with up to 10 PCs. The inclusion of additional PCs tends to lower the D^2 value for all metrics, but most notably in the Mahalanobis measure. The city block implementation tends to mirror this effect, although to a lesser extent. Both the Euclidean and Chebychev distances stabilise with the inclusion of more components, but this point comes sooner for the Chebychev metric.

4.2.2 Results

Figure 4.3 shows the correlation coefficients up to the tenth PC. The ρ values are the Pearson measure of correlation between each PC and the tenderness value (i.e. classification variable), and the correlation between the two Euclidean distance matrices is given by the D^2 values. The D^2 metric refers to using the first n PCs, whilst ρ is calculated with only a single PC. Although the two series are not entirely comparable, it can be seen that the first two PCs are much more highly correlated to tenderness than the subsequent PCs. Whilst the ρ value for the second PC is lower than the first, its inclusion in the distance matrix calculation does improve the model validity. The inclusion of the second component can be warranted by inspection of the PCA scores plot in Figure 4.1, which shows that the component varies somewhat with the tenderness values.

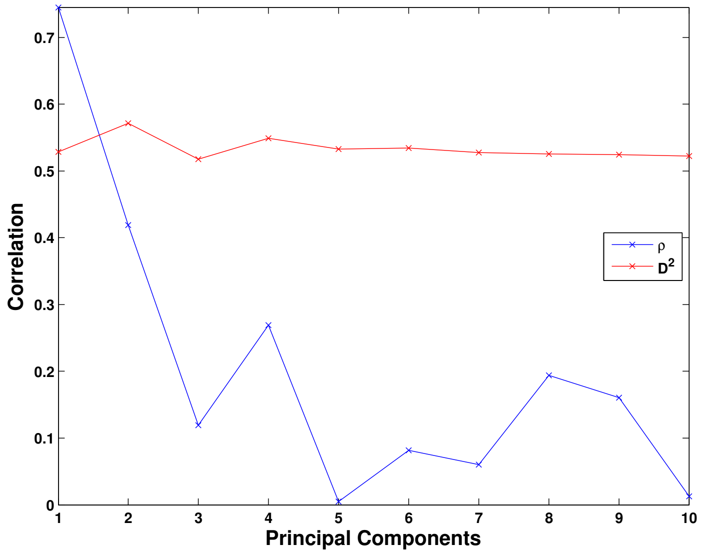


FIGURE 4.3: The ρ and D^2 correlation coefficients. The value of ρ show the correlation between a single component and the tenderness values, whilst the D^2 values calculate the distance matrix using all preceding components.

4.2.3 Summary

The calculation of D^2 values is more appropriate for continuous classification in pattern recognition techniques. Although it is typically lower than the ρ value from the most highly correlated principal component, it describes the relevant variance throughout multiple components. It will be used as a comparative metric when determining how well various unsupervised multivariate methods approximate tenderness. The optimal number of components to include is given by the combination resulting in the highest D^2 value.

Chapter 5

Processing I: NMR

Then there is the man who drowned crossing a stream with an average depth of six inches.

W. I. E. Gates

5.1 Abstract

A modified Lorentzian distribution function is used to model peaks in two-dimensional (2D) ^{1}H-^{13}C heteronuclear single quantum coherence (HSQC) nuclear magnetic resonance (NMR) spectra. The model fit is used to determine accurate chemical shifts from genuine signals in complex metabolite mixtures such as blood. The algorithm can be used to extract features from a set of spectra from different samples for exploratory metabolomics. First a reference spectrum is created in which the peak intensities are given by the median value over all samples at each point in the 2D spectra so that ^{13}C-^{1}H correlations in any spectra are accounted for. The mathematical model provides a footprint for each peak in the reference spectrum, which can be used to bin the ^{13}C-^{1}H correlations in each HSQC spectrum. The binned intensities are then used as variables in multivariate analyses and those found to be discriminatory are rapidly identified by cross referencing the chemical shifts of the bins with a database of ^{13}C and ^{1}H chemical shift correlations from known metabolites.

5.2 Introduction

Metabolomics involves the study of the complete set of small molecules present in a biological sample and like other '-omics' studies requires the analysis of complex data sets. The use of various spectroscopic methods allows the simultaneous identification of a wide range of these metabolites, providing characteristic profiles, or biochemical 'fingerprints' that detail the relative concentrations of compounds present in a sample. Techniques, often used in combination in metabolomic studies, include Fourier transform infrared spectroscopy [313] and mass spectrometry coupled to a separation technique such as gas chromatography [314] or liquid chromatography [14] Although less sensitive than mass spectrometry, the inherent reproducibility and non-destructive nature of high-resolution nuclear magnetic resonance spectroscopy (NMR), in particular ^{1}H NMR, without the need for prior separation has made this the method of choice for many metabolic analyses [315–317]. ^{1}H nuclear magnetic resonance (NMR) spectroscopy has been applied to great effect in the elucidation of metabolomes in a wide range of applications, including disease biomarker discovery [318, 319], plant metabolomics [11, 320–322], process chemistry [323] and even estimation of postmortem interval [324].

Examination of the entire metabolite content requires multivariate data analysis techniques in order to reduce the high dimensionality of the data. The most widely used methods are principal components analysis [16] and partial least squares [325] although artificial intelligence techniques, including neural networks [326], genetic algorithms [169] and genetic programming [27] have also been used to extract discriminatory features from ^{1}H NMR data. The methods can provide characteristic profiles for particular biological states that can be used for classification, but identification of the metabolites involved often requires the use of 2D NMR experiments. Many different types of 2D NMR experiments exist, providing information about chemical shifts, J-couplings and diffusion coefficients that can be used in database searches to identify particular metabolites. However, prohibitive data acquisition times have meant that such methods have not been fully exploited in non-targeted approaches. Recent advances have significantly helped to reduce the data collection time needed to obtain 2D spectra [327, 328] and to minimise the effect of spectral noise [81].

Where 1D ^{13}C NMR spectroscopy suffers due to the naturally low ^{13}C abundance (hence a poor signal-to-noise ratio) and also from the overlap of different resonances of interest, 2D heteronuclear experiments effectively overcome these problems. Heteronuclear Single Quantum Coherence (HSQC) pairs ^{1}H nuclei with other spin-$\frac{1}{2}$ nuclei (usually ^{13}C or ^{15}N) and the resultant signals identify the protons that are attached to these spin-$\frac{1}{2}$ nuclei. Thus 2D ^{1}H-^{13}C HSQC NMR correlates ^{1}H and ^{13}C NMR resonances from the same molecule, detecting only those carbons that are attached to protons (hence most of those in organic molecules such as metabolites). The magnetisation is transferred from the ^{13}C to the ^{1}H before detection, vastly improving the ^{13}C sensitivity. However, the main advantage of using the heteronuclear experiments such as HSQC is that the carbon chemical shift range extends to about 250 ppm so that these spectra are much better resolved allowing accurate and concurrent assignment of ^{1}H and ^{13}C chemical shifts. This is reason enough to seek to utilise two-dimensional spectra in metabolomic studies, though presently many studies employ HSQC as a technique to aid peak assignment, rather than as a stand-alone technique. This is predominantly because HSQC NMR is not a quantitative technique. However, this has recently been overcome with the use of adiabatic pulsing [49] generating fully quantitative HSQC spectra. Metabolomics is based on the observation of relative changes in concentration of metabolites when treatment and control groups are compared. In this case it is therefore reasonable to use the semi-quantitative gHSQC experiment [329] as this allows the direct comparison of the concentration of metabolites in different samples. Similarly, the use of 2D HSQC NMR allows metabolites that are unique to a particular sample or sample subgroup to be rapidly identified.

Protocols for quantitative metabolomics have been described in which the concentrations of particular metabolites in biological samples are determined by comparison of peaks in the 2D ^{1}H-^{13}C HSQC spectra obtained for the samples with those obtained for known concentrations of the metabolites in question [330, 331]. This requires the metabolites of interest to be known in advance and is therefore more suited to confirming or rejecting existing hypotheses rather than exploratory data analysis. Here we describe a protocol for data reduction and qualitative analysis that can be used in a hypothesis generating rather than hypothesis driven approach to metabolomics. An automated peak picking routine recognises true peaks in the 2D spectrum obtained for a complex biological sample by comparison with a modified Lorentzian distribution function. The integrated peak volumes are used as variables in multivariate statistical analysis and only those identified as potential biomarkers are subjected to further analysis to determine the metabolites concerned. As a demonstration, the method is applied to the analysis of 2D ^{1}H-^{13}C HSQC spectra from rat brain tissue following intra-peritoneal injection with [U-^{13}C]-glucose and those injected with normal ^{12}C-glucose.

5.3 Methods

5.3.1 Materials

Deuterated solvents were obtained from Goss Scientific (Cambridge, U.K.). Standard laboratory chemicals were obtained from reputable UK suppliers. Fetal bovine serum was obtained from JRH Biosciences (Lenexa, Kansas, USA).

5.3.2 Standard metabolite samples

Metabolite standards, where solubility allowed, were accurately prepared to approximately 100 mM with the exception of citric acid and tyrosine. Citric acid was prepared to approximately 10 mM to ensure that the buffer was still capable of maintaining the solution at pH 7.0. Tyrosine was prepared at approximately 5 mM because of its limited solubility.

All standards were dissolved into phosphate buffer (100 mM K_2HPO_4/KH_2PO_4, pH 7.0, 0.5 mM TSP) in ^2H_2O. Standards were centrifuged at 1000 g for 10 min and 600 μL of the solution transferred to a 5 mm NMR tube.

5.3.3 Bovine blood sample

Fetal bovine serum (1 mL) was deproteinated by centrifugation (3622 g, 60 min) through a 10 kDa molecular weight cut off centrifugal filter (Millipore). The filtrate was lyophilised to dryness and then dissolved into 600 μL of a phosphate buffer solution (90 mM K_2HPO_4/KH_2PO_4, pH 7.0, 0.55 mM TSP, 1 mM NaN_3) in ^2H_2O, mixed in a vortex mixer and transferred to a 5 mm NMR tube.

5.3.4 Rat brain tissue samples

Ice-cold acetonitrile (5 mL, 2:1 v/v) in water was added per gram of frozen wet tissue and homogenised for two 30 second periods separated by 1 minute, with the sample in an ice bath, using an Ultraturrax (11000 RPM). Following homogenisation, residual cells were disrupted by four treatments with a sonic probe (Ultrasonics Ltd), each lasting 30 seconds, separated by 2 minute intervals. The sample was centrifuged at 2300 *g* for 5 minutes and the supernatant retained. Acetonitrile was removed under a stream of dry nitrogen (4 hours, 20 °C) prior to lyophilisation to dryness. The lyophilised extract was reconstituted into 600 μL of a phosphate buffer (90 mM K_2HPO_4/KH_2PO_4, pH 7.0, 1 mM TSP, 1mM NaN_3) in ^2H_2O using a vortex mixer and was transferred to a 5 mm diameter NMR tube. The volume of buffer added to the sample was determined by the original mass of the brain. 1 mL of buffer was added to the organ with the lowest mass and the volume added to other samples was adjusted so that the mass to volume ratio was maintained. Any residual insoluble matter was removed by ultracentrifugation at an average of 511,258 *g* and 4 °C for 30 minutes.

5.3.5 NMR data processing

HSQC experiments (^{13}C-^{1}H) were carried out using liquid state high resolution NMR spectroscopy and a Bruker Avance 500 MHz NMR spectrometer equipped with a 5 mm TCI cryoprobe, variable temperature supply and pulsed field z gradients.

The ^{13}C-^{1}H gradient enhanced heteronuclear single quantum coherence (HSQC) NMR spectra were acquired using the standard library gradient enhanced pulse sequence [329] at the central frequencies of 500.1323541 MHz (^{1}H, F2 dimension) and 125.7691082 MHz (^{13}C, F1 dimension). A carbon-proton coupling constant of 145 Hz, gradient ratios, all using a sine shaped gradient, of 80, 20.1, 11 and -5% for gradients 1, 2, 3 and 4, respectively, a homospoil gradient pulse length of 1 ms, a gradient pulse length of 600 μs, a homospoil gradient pulse recovery delay of 200 μs, a delay of 862.07 μs to select for all carbon proton multiplicities, an interscan delay of 1.5 s, and an acquisition mode of echo-antiecho were used. 90°) pulse lengths of 11.1 and 13.0 ms were used for ^{1}H and ^{13}C, respectively, 1536 complex data points were collected in the F2 dimension with a spectral width of 13.33 ppm giving an acquisition time of 0.11532 s and proton carbon decoupling using a waltz decoupling sequence with a decoupling power of 10 dB was applied during the acquisition period. 384 increments were collected in the F1 dimension over a spectral width of 180.0 ppm giving a total experimental time of approximately 27 min 30 s.

All data acquisition was performed at a temperature of 300 K without sample rotation, using the software package Topspin v1.3 (Bruker, Germany). A sine-bell shaped window function phase shifted by 90 °C was applied over all data points and the data were zero filled to 2048 and 1024 data points in the F2 and F1 dimensions, respectively, prior to Fourier transformation, phase and baseline correction. The chemical shifts of all data were referenced to the resonance of the TSP peak at 0 ppm in both the ^{1}H and the ^{13}C dimensions.

5.3.6 Pre-processing of spectra

Phase-cycled ^{1}H-^{13}C HSQC NMR experiments suffer from t1 noise ridges parallel to the F1 axis at the F2 frequencies of intense peaks due to instrumental imperfections and external disturbances [332]. The largest t1 noise ridges can be higher in intensity than genuine peaks associated with low concentration compounds, causing problems for automated peak identification. However, the fact that t1 noise ridges are highly correlated for different F2 values, allows these artifacts to be removed and Correlated Trace Denoising [81] was applied to all spectra to reduce the effects of t1 noise before analysis.

5.3.7 Peak modelling

As the Fourier transform of a sum of exponentially decaying sinusoids, the natural lineshape in solution NMR spectroscopy is known to be Lorentzian [333]. However, as the ideal lineshape of NMR signals is often distorted due to sample inhomogene-

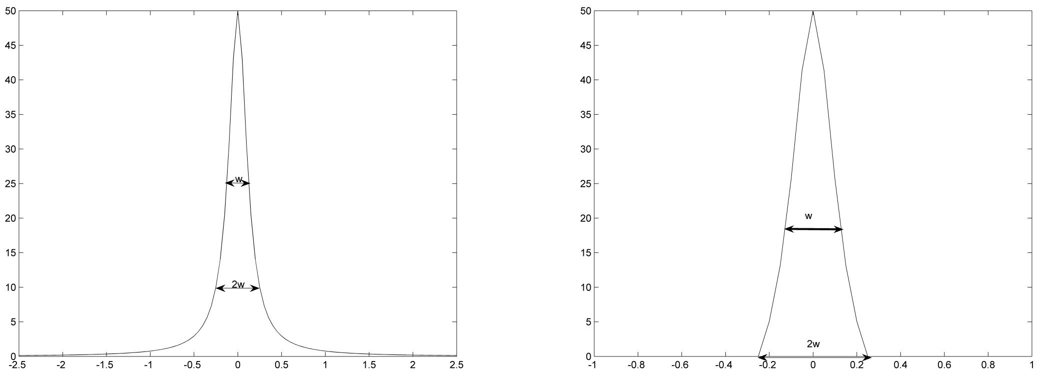


FIGURE 5.1: The Lorentzian line with an amplitude, A, of 50.0 and a width at half height, w, of 0.25 Hz is shown in (a). In (b) the corresponding modified Lorentzian is shown. Here $I = 50.0$ so that $A = 62.5$ and the width, w, of 0.25 Hz occurs at $A/2 - A/5 = 18.75$.

ity and experimental errors, a combination of Gaussian and Lorentzian lineshapes is often used for modelling NMR spectra from complex mixtures of metabolites. Figure 5.1 shows the Lorentzian function, $L(x)$, given by

$$L(x) = \frac{Aw^2}{w^2 + 4(x - x_0)^2} \tag{5.1}$$

where A is the amplitude, w is the peak width at half height, in Hertz, and x_0 is the peak position in Hertz. In two dimensions, we have

$$L(x, y) = \frac{Aw^2}{w^2 + 4((x - x_0)^2 + (y - y_0)^2)} \tag{5.2}$$

for a peak at (x_0, y_0). The greater resolution in the proton dimension results in peaks in the HSQC spectra covering more data points in the proton dimension than in the carbon dimension and w changes as the radius of an ellipse. For an ellipse, positioned at $(0, 0)$, with semimajor axis a and semiminor axis b, as in Figure 5.2 the radius w passing through the point (x, y) meets the ellipse at (x', y') where

$$\left(\frac{x'}{a}\right)^2 + \left(\frac{y'}{b}\right)^2 = 1 \tag{5.3}$$

Thus, as $\frac{x'}{w} = \frac{x}{r}$ and $\frac{y'}{w} = \frac{y}{r}$, where $r = \sqrt{x^2 + y^2}$, we have

$$\left(\frac{wx}{ar}\right)^2 + \left(\frac{wy}{br}\right)^2 = 1 \tag{5.4}$$

Rearragning Equation 5.4 gives the radius through (x, y) as

$$w = \frac{abr}{\sqrt{b^2x^2 + a^2y^2}} \tag{5.5}$$

The problem in using a 2-D Lorentzian model is that the slow decay of the function does not allow a suitable footprint at the base of a peak to be identified. However, we have found that a modified Lorentzian function can be used to model peaks in HSQC spectra. Taking $x_0 = y_0 = 0$ in Equation 5.2 gives

$$L(x', y') = \frac{Aw^2}{w^2 + 4w^2} = \frac{A}{5} \tag{5.6}$$

where again $w = \sqrt{x'^2 + y'^2}$, so that the function reaches twice the width at half height at one-fifth of the amplitude. The modified Lorentzian function, M(x,y), is calculated as

$$M(x, y) = L(x, y) - \frac{A}{5} = \frac{Aw^2}{w^2 + 4((x - x_0)^2 + (y - y_0)^2)} - \frac{A}{5} \tag{5.7}$$

where $A = \frac{5I}{4}$ and I is the intensity at the peak maximum. Peaks in the HSQC spectra can be modelled using just three parameters, the intensity at the peak maximum, I, and the widths at $\frac{A}{2} - \frac{A}{5} = \frac{3I}{8}$ in both the proton and carbon dimensions. Figure 5.3 shows a section of an HSQC spectrum together with the modelled version.

5.3.8 Standard metabolite templates

The HSQC spectra for 15 standard metabolites (alanine, betaine, citric acid, creatine, creatinine, fructose, glucose, glutamate, glutamine, lactate, leucine, myoinositol, phenylalanine, tyrosine and valine) were used to test the peak modelling and determine the criterion for the fit to represent a genuine peak. In addition, the

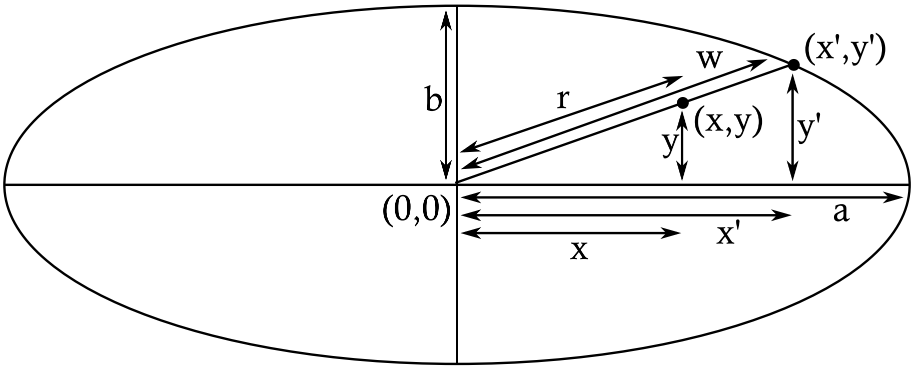


FIGURE 5.2: The radial width, w, through the point (x,y) of an ellipse with semimajor axis a and semiminor axis b, centred at (0,0).

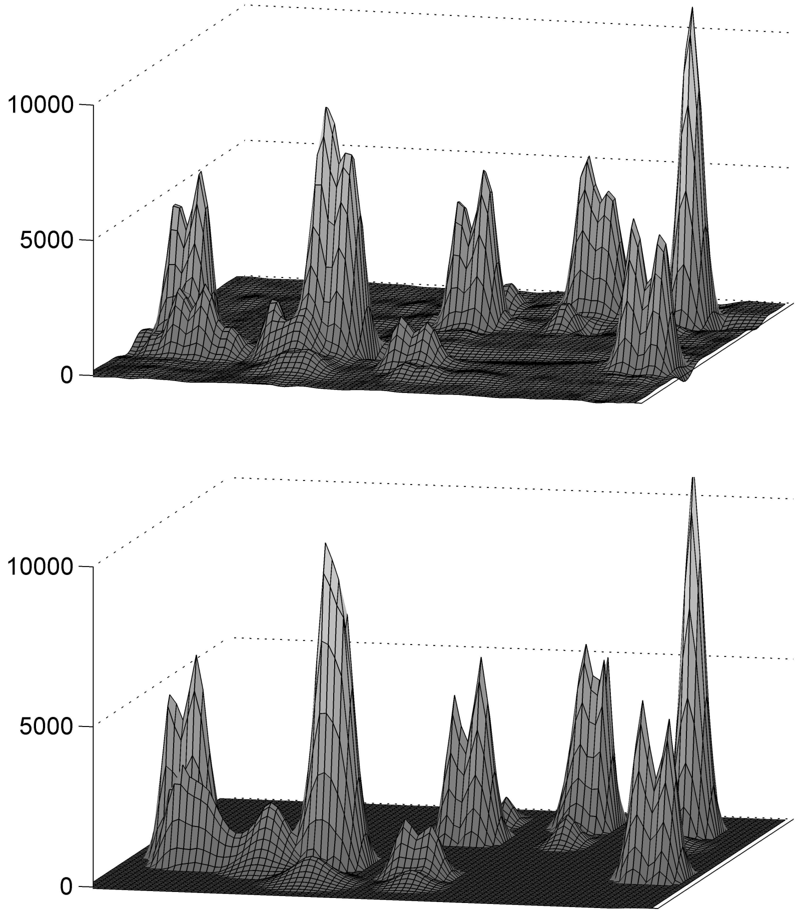


FIGURE 5.3: A section of an HSQC spectra containing several overlapping peaks is shown in (a) with the corresponding modelled peaks shown in (b).

HSQC spectrum for γ-aminobutyric acid (GABA) was obtained from the Biological Magnetic Resonance Data Bank (BMRB [116]), and processed in TOPSPIN in a comparable way to that used for the other spectra. For each metabolite, all maxima (i.e. greater in intensity than all eight neighbouring data points) above a threshold determined by the standard deviation, σ, of the intensities over the spectrum were identified. A threshold of 3σ proved useful for most spectra but was increased for citric acid and tyrosine as the spectra for these metabolites had low signal to noise levels.

For each potential peak, the proton and carbon peak widths at 3/8 of the intensity were determined. As the twin peaks of doublets in the proton dimension are close enough to overlap at this height, the minimum of the positive and negative radial widths was used to provide the axes for the ellipse. Any maxima with both positive and negative radial widths beyond a pre-defined maximum were eliminated at this stage. The position, widths and intensity of each of the remaining maxima were stored along with the details of any maxima close enough to overlap. For each potential peak, a modified Lorentzian model of appropriate widths and a height of 1.0 was calculated and the contribution from any maxima that would overlap the elliptical footprint added using their relative peak heights. This model peak was then rescaled to unit height and a residual score calculated as the sum of the absolute differences between the model and the experimental data (also rescaled to a peak height of unity) over all data points within the ellipse.

As the positions of peaks related to the standard metabolite signals are known, an estimate of the acceptable error for a real peak could be obtained. Although the contribution from close peaks was included in the model when calculating residuals, the scores for peaks overlapped by much larger peaks still tend to be significantly higher than those of single peaks. Scores less than 1.0 were obtained for most of the expected peaks, whereas all scores greater than 2.0 corresponded to spurious maxima, usually due to t1 noise. A cut-off of 1.5 allowed all expected peaks, including those with several large neighbours (as can be found in fructose, for example) to be identified as genuine. Up to five extra peaks were also classed as real peaks using this cut-off, including the small cross-peaks seen in sensitivity-enhanced HSQC spectra [334]. As the inclusion of a few extra peaks as variables would not be detrimental to multivariate analyses, the cut-off of 1.5 was chosen as the criterion to determine a genuine peak. For each standard metabolite, the relative intensities of every such peak and their chemical shifts (in ppm) provide a pattern for that metabolite. A database of ^{13}C and ^{1}H chemical shifts correlations from known metabolites will be constructed but the inclusion of a spectrum obtained from the BMRB database demonstrates that existing resources can be utilised.

5.3.9 Peak assignment in bovine blood spectrum

The peak-picking algorithm was applied to the ^{13}C-^{1}H HSQC spectrum obtained from bovine blood. Most peaks in the spectra obtained for the initial 15 standard metabolites were identified by eye in the bovine spectrum although none of the peaks associated with GABA were observed. A threshold of 3σ resulted in a total of 110 peaks being identified as genuine. Where possible, these peaks were assigned to one of the 16 standard metabolites, allowing for minor shifts caused by fluctuations in experimental conditions, such as temperature and pH. As expected, there were no peaks assigned to GABA but, with the threshold at this level, 19 other peaks from the standard metabolites could be identified, including all of the peaks corresponding to the aromatic amino acids, phenylalanine and tyrosine. In fact the threshold had to be lowered to 1σ before all of the valine and leucine peaks could be found. Four peaks from the 15 standard spectra could still not be identified. Of these, two correspond to phenylalanine and one to tyrosine. All three are relatively small in the standard spectra and cannot be identified by eye in the bovine spectrum. The other missing peak is the largest peak found in the glutamine spectrum at 3.78 ppm (proton) and 56.85 ppm (carbon). Inspection by eye shows this peak to be swamped by a larger glutamate peak at 3.76 ppm (proton) and 57.39 ppm (carbon) as can be seen in Figure 5.4. The two are treated as a single peak as there is no maximum associated with the glutamine peak. Although the fitness score is lower than some obtained for peaks recognised as belonging to multiplets, it is significantly higher than any other score for a single peak. Therefore, such problems could potentially be recognised and dealt with.

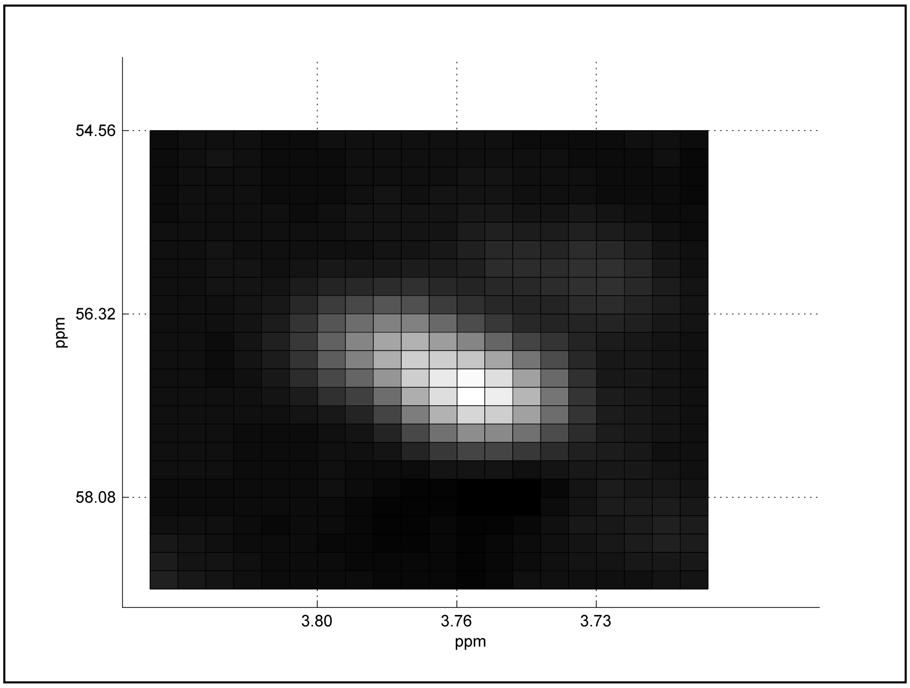


FIGURE 5.4: A glutamine peak in the bovine spectrum at 3.78 ppm (proton) and 56.85 ppm (carbon) swamped by a larger glutamate peak at 3.76 ppm (proton) and 57.39 ppm (carbon) so that the two are treated as a single peak.

5.3.10 Variable extraction for metabolomics

Misalignment of data can lead to erroneous interpretation of results in statistical analysis and pre-processing methods for peak alignment and data averaging (binning or bucketing) have been used for correcting variation in chemical shifts in 1D NMR spectroscopy data [335]. Uniform binning involves integrating the spectral data over regions of equal length, typically 0.04 ppm and can result in peaks being split between bins or more than one peak being assigned to a bin. This reduces data interpretability and can increase the variation in a dataset when peak shifts occur close to a bin border. To overcome these problems, Davis et al. [54] introduced adaptive binning in which the bins correspond to peaks in a reference spectrum. Here we extend the method to two dimensions and apply the peak picking routine to a reference spectrum. Although novel alignment methods have been developed for 2D data [336], the considerably lower resolution of 2D NMR experiments together with careful data collection makes such methods unnecessary. The main purpose of the reference here is to provide a single spectrum incorporating the peaks present in any of the spectra rather than for alignment. Thus the peak fitting only has to be performed once using this reference spectrum and the peak footprints obtained applied to each sample spectrum in the data set.

A reference spectrum created by taking the maximum, over every sample to be used in the analysis, at each point in the spectra would allow peaks occurring in any individual sample to be represented. However, this gives a very noisy reference spectrum and any smoothing at this resolution results in some close peaks being amalgamated. We therefore use the median value at each point to obtain a reference spectrum in which the peaks in any sample class are accounted for but which requires no smoothing. For each peak identified in the reference spectrum, the ellipse with major and minor axes corresponding to the carbon and proton peak widths provide a footprint that can be used to bin the ^{1}H-^{13}C correlations for each sample in the data set. The binned intensities can then be used as variables in multivariate analyses and those found to be important for classification identified by cross referencing with the ^{13}C and ^{1}H chemical shift correlations from known metabolites.

5.4 Results

Using a threshold of 2σ, a total of 105 peaks were identified as genuine in the reference spectrum obtained from the HSQC spectra of rat brain extracts. Principal components analysis (PCA) was performed on the resulting variables after scaling to unit variance. The plot of the first two principal component scores is shown in Figure 5.5. There is clear separation between the rats injected with [U-^{13}C]-glucose and those with normal ^{12}C-glucose. There is clear separation of the two groups along the first principal component and, with the exception of one early time-point [U-^{13}C]-glucose observation, the difference between samples from rats injected with [U-^{13}C]-glucose and those from rats injected with normal ^{12}C-glucose decreases gradually with time. The outlier is from a rat injected with [U-^{13}C]-glucose and killed after 0.5

hours that clusters with the control group. It is probable that the injection of glucose into the peritoneum of this animal was unsuccessful. The loadings for the first principal component were used to identify the peaks responsible for the difference between groups. A number of peaks were found to have increased volumes for the [U-^{13}C]-glucose observations relative to those corresponding to normal ^{12}C-glucose. The early time-points were found to have the greatest difference in volume with a gradual decrease in volume over time. Cross referencing with the ^{13}C and ^{1}H chemical shift correlations obtained for the 16 standards allowed 40 of the 105 peaks used in the analysis to be associated with a known metabolite. Table 5.1 lists the chemical shifts for these peaks and the metabolites to which they were assigned. Peaks found to be different between the two groups were assigned to glutamate, GABA, alanine, lactate and glutamine whereas peaks for which no consistent change was found were assigned to myoinositol, betaine, creatine, glucose, fructose, valine and phenylalanine. No peaks related to creatinine, citric acid, tyrosine or leucine were identified. Four discriminatory peaks were not assigned to one of the metabolites included in this study but further investigation showed that two were cross peaks related to the largest glutamate peaks and two were cross peaks related to lactate. As none of the other unassigned peaks showed a significant difference between the two groups no further investigation into their identity was carried out.

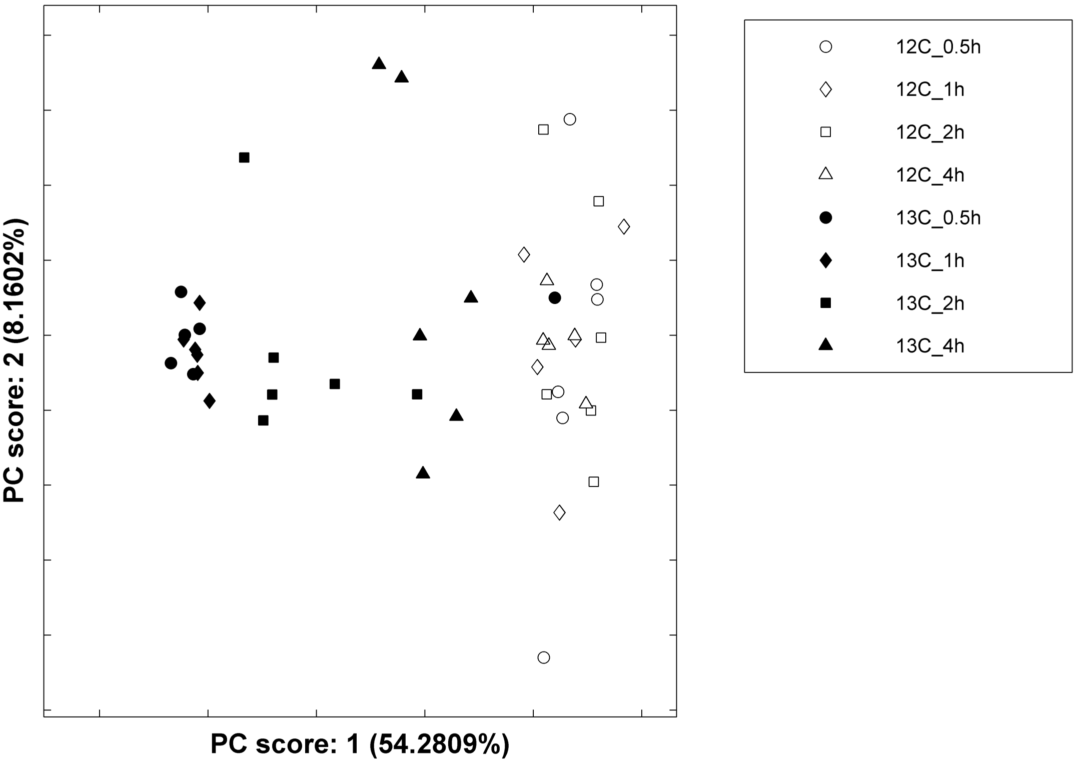


FIGURE 5.5: Plot of the first two principal component scores. Solid symbols represent observations from rats injected with [U-^{13}C]-glucose and open symbols those injected with normal ^{12}C-glucose. The four post-injection time points are represented by different symbols with circles, diamonds, squares and triangles representing 0.5, 1, 2 and 4 hours respectively. With the exception of one early time-point [U-^{13}C]-glucose observation, there is clear separation of the two groups along the first principal component. It can be seen that the difference between samples from rats injected with [U-^{13}C]-glucose and those from rats injected with normal ^{12}C-glucose decreases gradually with time.

TABLE 5.1: Chemical shifts for the peaks that were assigned to the standard metabolites used in this study.

^{1}H chemical shift (ppm)	^{13}C chemical shift (ppm)	Metabolite assigned
3.93	56.68	creatine
1.33	22.89	lactate
3.76	57.38	glutamate
2.36	36.27	glutamate
3.04	39.78	creatine
4.06	74.97	myoinositol
4.12	71.28	lactate
3.54	73.92	myoinositol
2.14	29.58	glutamate
3.23	56.85	betaine
3.63	75.15	myoinositol
2.07	29.76	glutamate
3.01	42.07	GABA
3.28	77.26	myoinositol
1.91	26.24	GABA
2.30	37.15	GABA
2.45	33.63	glutamine
3.54	75.33	myoinositol
2.40	36.97	glutamate
3.90	64.07	glucose
3.85	64.24	glucose
3.67	65.12	fructose
3.63	65.12	fructose
3.62	73.92	myoinositol
1.48	19.02	alanine
3.83	64.24	fructose
2.40	32.57	glutamine
4.02	66.35	fructose
3.86	74.09	glucose
3.59	63.19	valine
3.78	53.33	alanine
3.92	73.39	fructose
2.53	34.15	glutamine
3.57	65.30	fructose
3.95	69.34	betaine
3.88	69.34	betaine
3.28	38.20	phenylalanine
2.23	36.79	GABA
3.35	76.38	myoinositol
3.87	65.65	fructose

5.5 Conclusion

There has been an increasing prevalence in the literature regarding the potential of using HSQC spectra in metabolomics studies, though this has not been fully exploited in exploratory chemometrics. The freely available rNMR software [174] allows visual representation of multiple spectra. Complex spectra are interpreted in terms of 'regions of interest' (broadly analogous to our footprint-based approach), with the potential for manual changes to account for discrepancies caused by, for example, variations in pH. However, the use of rectangles means overlapping peaks can create large 'regions of interest' so that peaks cannot be considered individually. Xi et al. [173] have investigated the use of HSQC analyses to identify the presence or absence of known metabolites in biological matrices using an algorithm to determine individual peaks. The method is applied to the quantification of known metabolites although the potential for use as a feasible alternative to 1D NMR in metabolomics is recognised.

We have shown that a modified Lorentzian distribution function can be used to identify peaks in HSQC spectra, allowing accurate chemical shift information to be extracted without prior knowledge of the chemical composition. The determination of peak widths at 3/8ths of the intensity rather than at the base of a peak allows individual peaks to be resolved and their integrated volumes used as variables in multivariate analyses. Discriminatory peaks can be identified and assigned to particular metabolites by comparison of the ^{13}C and ^{1}H chemical shift correlations with those of known metabolites. Such assignments require a database of ^{13}C and ^{1}H chemical shifts correlations from known metabolites and the use of those from a spectrum obtained from the BMRB database demonstrates that existing resources can be utilised. The number of 2D metabolite spectra available in online databases is increasing and can provide the chemical shift correlations for many more metabolites. Although chemical shift variation due to experimental factors, such as pH and solvent, must be taken into account, this will allow more peaks identified by multivariate analysis to be allocated to metabolites.

The continuing developments in NMR spectroscopy could soon make the routine use of 2D HSQC experiments for non-targeted metabolomics a reality. Sensitivity improvements due to polarisation techniques such as parahydrogen enhancement [337] and dynamic nuclear polarisation [338] allow spectra to be acquired with fewer scans. Recently, a DNP enhanced HSQC spectrum of 0.15 mg of pyridine was acquired in 0.13 s [339]. Clearly these revolutionary NMR technologies will enable 2D spectra to be recorded from complex mixtures on timescales that are amenable to high-throughput metabolomics. Automated peak selection and matching as described here represents a massive data reduction with a small number of pertinent variables extracted from these highly complex spectra. As demonstrated, these variables can be used in multivariate analyses and potential biomarkers assigned directly to specific metabolites.

5.6 Acknowledgements

The authors would like to thank the Defra seedcorn programme for funding this work.

Chapter 6

Processing II: LC-MS

> If it wasn't for Venetian blinds, it would be curtains for all of us.
>
> Eric Morcambe

6.1 Introduction

The processing of LC-MS data is required in order to be able to effect subsequent multivariate analysis. Processing the data also aids interpretation, as the data are transformed from a series of intensities recorded as functions of retention time and *m/z*, to a series of peaks and their corresponding isotopic distributions. There is a wealth of information to be gleaned from LC-MS, and effective processing will achieve this. The ever-enhancing capabilities of hardware allow for more and more data to be collected, and the bottleneck is often associated with data processing and its subsequent interpretation [37]. The development of the data processing toolbox must, therefore, keep abreast of hardware developments.

Liquid chromatography-mass spectrometry is the most widely used analytical technique for the the analysis of biological specimens, and non-targeted and metabolomics studies have the capabilities of generating vast quantities of data in both long- and short-term studies. Whilst the sensitivity of LC-MS surpasses that of NMR spectroscopy, the reproducibility of data acquired using LC-MS does not yet match the consistency which is achievable with NMR spectroscopy. The same samples analysed on different NMR spectrometers give highly comparable data [118, 119]; the reality with LC-MS is that such comparisons are fraught with difficulties, and the routine use of quality control (QC) samples [229, 230] or the process of calibration transfer [238] is required in experiments in order to be able to assess, and correct for, any drift observed in experiments.

Spectral processing can be easily performed by use of proprietary instrumental software or one of many freely available packages (Section 3.4.7). These typically result in the generation of a peak table, showing the intensities of peaks as functions of retention time and *m/z*, and many allow for visualisation of data using statistical techniques such as principal components analysis (PCA). In non-targeted analysis, data often needs additional processing beyond that available in many software packages; the poor reproducibility of LC-MS typically requires correction of intensities based on a quality control (QC) strategy. This Chapter details the processing of LC-MS datasets, from the raw files through to a series of peaks.

6.2 Data

The data used to develop the LC-MS processing method is taken from an exploratory study looking to assess the practicalities of using serum for the diagnosis of tuberculosis in badgers. Data were recorded in profile mode using a high-resolution Thermo Exactive with a resolution of 50,000 FWHM. Data files were exported as netCDF, which is a common scientific data format.

6.3 Data processing

Akin to HSQC data, as discussed in Chapter 5, liquid chromatography-mass spectrometry (LC-MS) is a three-dimensional data form, and methods to model these 3D peaks are necessary for the processing of the mass chromatograms. LC-MS peaks are not as mathematically regular as those from NMR, as the LC elution profile cannot be assumed to always be regular. LC-MS data files are typically organised into a series of individual mass spectra (or scans), each recorded at a specific elution time. The quantity of data within scans varies and is typically discontinuous in *m/z* value. The data are presented as a series of *m/z* and intensity measurements. There are various approaches to file processing, and some have been outlined in Chapter 3. In the development of this method, the emphasis was to minimise the need for user input, and to avoid the rounding of *m/z* values such that peaks may be artificially split.

The first stage of this data processing method involves converting each scan from a series of data points into a series of peaks, rather than the line spectral typically seen in centroid mode data. The distinction is necessary, as a line has no width. The use of mathematical functions to model each peak permits that a width is associated with each peak. Subsequently, these series of peaks across the retention window are grouped into chromatographic, three-dimensional, peaks if their retention times overlap. The peaks lists from each observation need to be sorted into an $m \times n$ matrix, such that each column describes the same variable. For instances where a peak is identified in a few samples, a filter may be applied to remove these. The final processing stage involves the grouping of separate peaks that are attributable to various isotopologues into an isotopic 'cluster'. This stage is beneficial for identification purposes and it also removes redundant variables from subsequent multivariate analysis. Each of the various stages is discussed in more detail below, and flow charts for each stage are presented in Appendix A.

6.3.1 Conversion of profile data

Data that are collected in profile mode are potentially more informative than centroid data, as an estimation of noise levels can be accomplished. The aim of conversion to centroid data is to reduce the quantity of data, whilst retaining the level of information. On its own, centroid data must be taken at face value: each and every 'line' must be considered as a genuine peak.

Each mass chromatogram is composed of a series of mass spectra, or scans. Each of these scan is individually processed to convert its data points into a series of peaks. Each of these peaks can be modelled by a Gaussian distribution function, which takes the form as shown in Equation 6.1. The function depends on only three parameters, namely the intensity at the centre (a), the peak's position or *m/z* (μ) and the width of the function (σ). The full width at half-maximum (FWHM) of a Gaussian function is calculated according to Equation 6.2, and peaks that are resolved at 50% of the maximum can be modelled by the function. By modelling each peak to a distribution function, the number of data points that comprise each spectrum can be dramatically reduced.

$$f(x) = a \times e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \tag{6.1}$$

$$FWHM = 2\sqrt{2 \times ln(2)} \times \sigma \tag{6.2}$$

Figure 3.11 shows the calculated Gaussian distribution functions for three peaks in a mass spectrum. In the original data, the right most peak is composed of 16 discrete data points. A modelled function is represented by just three, namely the position, width, and the intensity at the centre. The central intensity is only used for reconstructions of the peak (e.g. for diagrammatic purposes), whilst the sum of the individual data points is used to represent the abundance. As such, the magnitude of a chromatographic peak is calculated by the sum, or integral, of its individual data points.

Each data point is evaluated consecutively, and if its intensity is greater than its four nearest neighbours, then it is considered as a local maximum. For this maximum, the limits at each side are found by searching for either zero intensities (fully resolved peak) or local minima (partially resolved). These boundaries define the extent of the peak. A peak's FHWM is calculated by interpolating the intensities at 50% of the peak's magnitude (Figure 6.1). The FWHM value is then used to calculate σ according to Equation 6.2.

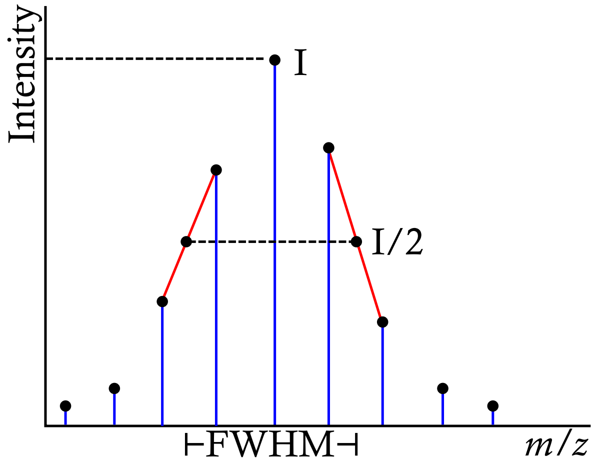


FIGURE 6.1: The calculation of a peak's half width is effected by linear interpolation. Where the linear interpolation and the intensity at half maximum (I/2) intersect, a peak's half-width is defined. The width at each side, where possible, is calculated to provide a FWHM. If peaks are resolved on one side only, then the FHWM is twice the value of the calculable half-width.

The resolution of MS instrumentation is often quoted in terms of the peak widths, for example the resolution of the orbitrap analyser is noted as 50,000 FWHM at *m/z* 400. In theory, if peaks are observed to deviate strongly from this resolution, i.e. width, then they may be considered as spurious and subsequently removed. The reality of this, however, is that resolution depends on *m/z* as well. This relationship is shown in Figure 6.2, where it can be seen that FWHM increases with decreasing *m/z*. The ballooning of FHWM at low *m/z* values is such that the modelling of any relationship would need to provide wider confidence limits at low *m/z* values in order to avoid the rejection of too many genuine peaks. Peak filtering of this nature requires for real and spurious peaks to be carefully characterised in order that a model can be generated. No maximum peak width has been employed here, but a minimum width requirement of 3 points (i.e. a local maximum) has been implemented to permit the modelling of a Gaussian function.

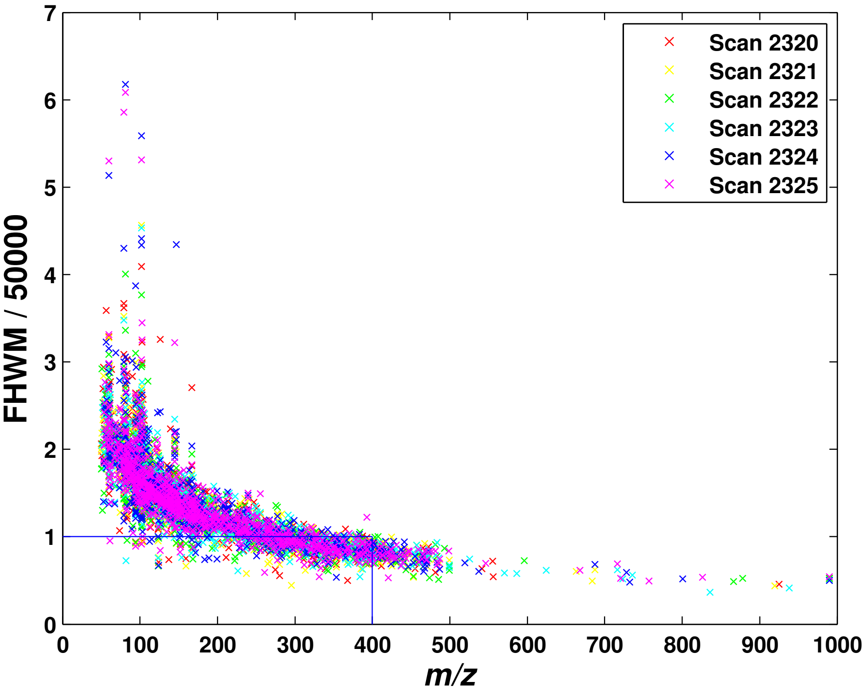


FIGURE 6.2: The widths of peaks, taken from 6 successive scans, plotted as a function of *m/z*. The data are recorded using an orbitrap analyser with the resolution set to 50,000 FWHM (*m/z* 400). It is clear that few points exhibit such a characteristic, and using it to filter the data is unwise.

The conversion of data from profile to centroid mode helps to dramatically reduce the quantity of data points, and has been effected without defining thresholds (for example, noise) for each data file. The advantage of using Gaussian functions, rather than other centroid-like approaches, is that a Gaussian function has a defined width. This is contrasted to typical centroid spectra, where a peak is typically represented as a single point.

6.3.2 Chromatographic peaks

The series of two-dimensional Gaussian-modelled peaks need to be grouped according to their retention time. This stage combines peaks with similar *m/z* values and retention times, and forms three-dimensional peaks.

The algorithm uses peaks from the first scan to form the template against which peaks from subsequent scans are compared. If a peak from the next scan matches the *m/z* of a peak in the template, then these two peaks are combined. Otherwise, where no match can be made, the new peak is added to the template so that any peaks from subsequent scans can be compared against it. Each Gaussian-modelled peak, therefore, is either matched to an existing chromatographic peak, or forms the beginning of a new chromatographic peak. Peaks that do not match any in the template may well be noise peaks, but they cannot be discounted without reference to subsequent scans.

As genuine, i.e. non-noise, chromatographic peaks do not expand indefinitely, a peak can be considered to have fully eluted when no further Gaussian-modelled peaks are appended to it. These 'fully eluted' peaks cannot be added to, and subsequent peaks which overlap at FWHM are not included in the peak; instead they form the beginning of a new chromatographic peak. Peaks are considered to have fully eluted when there are no further additions after three consecutive scans. This allows for the heads and tails of peaks to be included even though they may not be contiguous with the body of the peak. Figure 3.12 (p. 135) shows each of the individual peaks, in red, that form part of a single chromatographic peak. At early retention times the individual peaks are not contiguous yet should clearly be included as part of the main peak. Likewise, a single non-contiguous slice at the end of the peak is included. The presence of such gaps is likely to the fact the intensity at that time fell below the instrument's limit of detection. Peaks with a narrow elution profile are unlikely to be genuine. Here, any peak eluting over fewer than 5 scans (approximately 5 s) was excluded; Figure 3.12 shows the presence of nine such examples.

6.3.3 Between-file comparisons

In order for comparisons to be effected across multiple files, all of the variables need to be arranged into an $m \times n$ matrix, where each column, n, represents a common variable. The process is achieved by using the first observation as a template, against which all subsequent observations are compared. Where chromatographic peaks overlap in both time and *m/z* value, they are grouped together. If there are no matches, then the peak is added to the template for subsequent comparisons.

Drift in a peak's retention time is relatively common in large LC-MS studies, and is typically unpredictable (i.e. non linear). The variation in drift is different for individual peaks, such that blanket transformation is not possible. Instead, each peak must be corrected individually. If drifts in retention time are assumed to be relatively small (i.e. less than the peak width) and to vary in a sequential manner, then it may be considered that there is a continuous chain of overlap in the time dimension. By defining a peak's time boundary as the extremities of its overlapping parts, any drifts in retention time can be nullified. Such a flexible approach to drifting peaks is not applied to the *m/z* value. Whilst there are likely to be minor fluctuations in *m/z* across multiple observations, the boundaries for a series of grouped peaks are defined by the mean FWHM around the mean *m/z* value.

6.3.4 Reference spectrum

A reference spectrum is a device that allows the consistency of variables to be analysed across all of the observations. Persistent noise peaks are unlikely to appear consistently across multiple observations, so peaks that appear in a small proportion of files may be removed.

A median reference spectrum can be calculated by taking the median value for each variable. Where the median for a particular variable is zero, then less than half of the observations have a recorded intensity for that particular variable. Using this approach is an effective method to trim the number of variables. It may, however, be too proscriptive in that a variable appearing exclusively, and consistently, in a single class of observations may be excluded simply because the majority of the observations contained no trace of that particular variable.

To combat this problem, a group median reference spectrum can be applied instead. The spectrum requires *a priori* knowledge of the observations' classes, so that the median of the variables across each class can be calculated. Variables that appear in at least half of the observations of any class are included. Thus, the emphasis is placed on peaks that appear consistently throughout the classes. One drawback of this method is the 'supervised' nature of the reference spectrum creation; classes which are not easily segregated (e.g. if there is a continuous rather than discrete classifier) may suffer as too many peaks are removed from the analysis.

6.3.5 Summary

The processing method developed proved successful in extracting variables from the badger serum data. However, the principal limitation of the method is the time taken to process the data files. For smaller datasets the number of peaks generated is fewer and the system memory requirements are reduced. The procedure was tested with the pea data files (Chapter 4) but satisfactory processing times (<1 day) were achievable when only *m/z* values greater than 200 were considered, due to the high density of detected signals at values lower than this.

All of the pea data files have instead been processed using XCMS [298]. With a 2.16 GHz processor with 3 Gb memory, the processing times were ~16 h for HILIC datasets and ~8 h for the RP datasets. Figure 6.3 shows which variables were detected by the two techniques. Whilst XCMS offers the considerable advantage of the full *m/z* range, it clearly misses many features. A full comparison of the two techniques is beyond the scope of this Chapter, but the clear difference between the two approaches merits an in-depth investigation into the capabilities of both techniques.

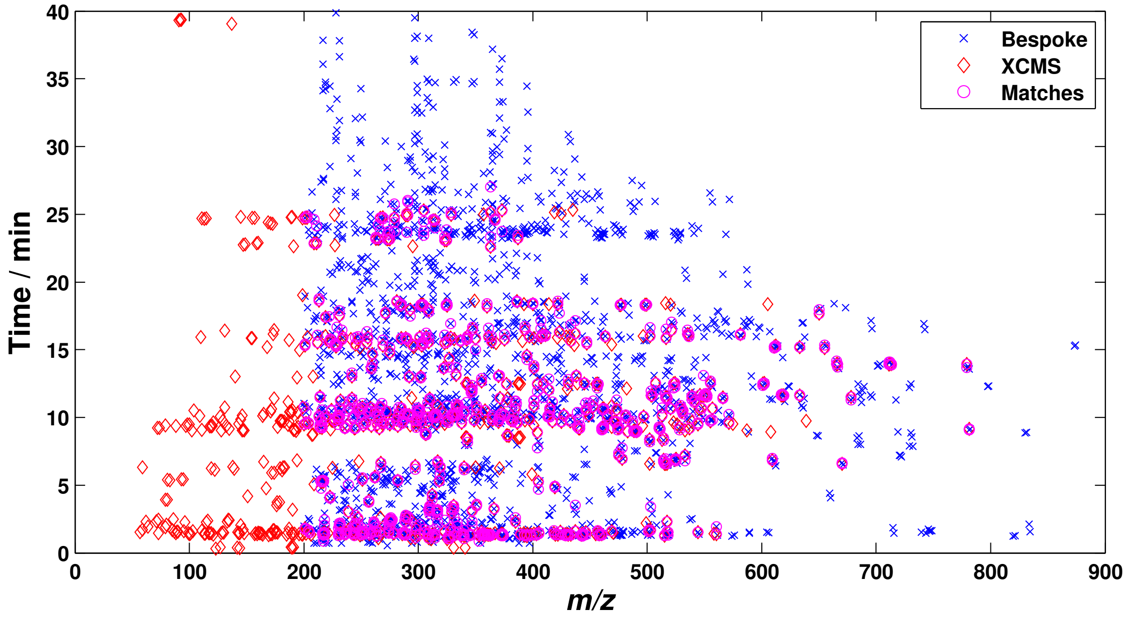


FIGURE 6.3: The peaks detected in negative mode HILIC LC-MS analyses of the pea samples, prior to the isotoping stage. The bespoke method only processes *m/z* values greater than 200 but detects approximately three times more features than the XCMS software.

6.4 Isotopic information

For many elements there exist multiple isotopes which, in various combinations, result in a molecule existing as a series of isotopologues. These isotopologues differ in isotopic composition, and their relative abundances are reflected in the distribution of *m/z* and intensity values.

A molecule without any isotopologues will appear in a mass spectrum as a single peak. As the isotopic composition of a (natural) molecule depends on the relative abundance of the isotopes, a mass spectrum will invariably reveal the presence of multiple isotopes unless, of course, the molecule contains purely monoisotopic elements[1]. As the natural abundances of isotopes are known, the isotopic distribution has the potential to reveal the elemental composition of an analyte. Additionally, the grouping of peaks attributable to the same analyte will reduce the number of variables: the various isotopic peaks of a single molecule are highly correlated, as and they describe the same thing they are essentially redundant.

The isotoping procedure is designed to account for singly and doubly charged ions, with the assumption that small molecules are unable to support three charges. Each individual peak is compared to others that have not been grouped into an isotopic feature. Peaks found within approximately 0.5 and 1 *m/z* units, and which overlap in their elution time are identified, and the correlation between all of these peaks is calculated. Although peak intensities are unrelated across multiple observations, the ratio of intensities of two isotopic peaks, e.g. [M]:[M+1], across a group of observations should remain constant. Thus, for peaks that are spatially close, those forming part of the same isotopic distribution should be strongly positively correlated.

Without the use of correlation analysis, the only discriminating factors for peak grouping can be spatial proximity, which may not be sufficient to distinguish between two overlapping features. The use of correlation coefficients helps to impart extra confidence in grouping features together.

[1] Of the few monoisotopic elements (Be, F, Na, Al, P, Sc, Mn, Co, As, Y, Nb, Rh, I, Cs, Au), phosphorus is the only polyvalent element capable of forming large molecules. Phosphates are perhaps, therefore, the only natural molecules without an appreciable distribution.

6.5 Normalisation

The reproducibility (i.e. long-term) of data acquired using LC-MS instruments is typically poor, especially in relation to data obtained by NMR spectroscopy [118, 321]. Studies that are carried out over the course of many months or years are subject to variations in the quality of recorded data. Even over the course of a day, it is possible that the dominant source of variance seen in spectral intensities is related to the sample run order. One likely cause is related to contamination of the chromatographic column, or the ion source; the degradation in quality is gradual, and a cleaning regime will likely restore the capabilities but doing this between each injection is clearly not possible in large-scale studies [229]. Instead, the capabilities of a system should be addressed such that reproducibility is not entirely compromised in the name of high-throughput studies.

In smaller and targeted studies, the spiking of isotopically labelled standards into the matrix allows for the reproducibility issue to be effectively negated. This is because each analyte being investigated can be scaled according to the intensity of the labelled compound, which was added at a known concentration. This course is not practical for larger studies in which hundreds or thousands of analytes are recorded. A method often propounded corrects the data according to the sum of the ion count, which assumes that constant volumes of injected samples produces the same quantity of ions. In scaling the data to this metric, injection-based failures can be mitigated, such as where it is apparent that not all of a sample has been injected. Alternative methods involve the use of a single internal standard, which is often an exogenous molecule; drifts in chromatographic retention and intensity can be observed and, where possible, corrected. Internal standards rely on the assumption that all analytes are equally affected, which is not always an appropriate assumption.

Some variables exhibit drifts that do not match the internal standard. In these instances, a more localised correction method is necessary. QC samples injected periodically throughout the course of an experiment provide clues regarding how variables respond over time. If QC samples are maintained as identical, then monitoring each variable is possible, and deviations from normality can be corrected. Although the true intensity of an analyte is never known, repeated measurements serve to reduce the estimation error, such that correction to the mean intensity of the QC samples is acceptable. Correction on a variable-by-variable basis is, however, only possible if a variable in experimental samples is also found in the QC samples; consequently, the best composition of a QC sample is taken from aliquots of all analytical samples, such that it is wholly representative of the experiment.

The use of various normalisation strategies is necessary due to the unpredictability of LC-MS data collection. All of the methods discussed above have been implemented on the pea data (Chapter 4) with the aim of reducing the impact that time imparts on variable intensities; until these deleterious effects are removed, inferences from multivariate analysis are likely to be flawed.

6.5.1 Inherent variability

Principal components analysis (PCA) is a multivariate statistical technique that allows for inspection of trends within data sets. Linear combinations of original variables are produced that explain as much of the original variance within the data. The first new variable, or score, explains as much of the original variance, whilst successive scores account for as much of the remaining variance as possible. Inspection of the scores reveals, therefore, the dominant sources of variance which may be identified with, for example, the class of an observation.

Figure 6.4 shows the PCA scores plot of the positive mode reversed-phase (RP) LC-MS variables, and the colouring denotes the batches in which the samples were run. It can be clearly seen that the overarching source of variance is associated with sample batch rather than a factor that is inherent to the samples. Until this undesirable source of variance is removed, any analysis of the data will be highly flawed.

Two-way univariate analysis of variance (ANOVA) was also performed to illustrate the extent to which the variance in variable intensities is dominated by the sample batch. The technique is used to gauge if the variance of a single variable is attributable to specific factors. Each variable is assessed individually, and the p-value indicates the probability that the variance is not due to that factor: a low p-value (e.g. < 0.05) suggests that the variance is linked to that factor. The two factors used in this analysis were the tenderness value and the sample run order. The run order is distinct from batch, and is a continuous variable, such that it takes account of the number of injections and analyses performed on the instrument.

Each variable was subjected to a 2-way ANOVA, with a significance threshold of $\alpha = 0.01$. The results, illustrated in Figure 6.5, show that the vast majority of variables show significant variance associated with the run order of observations (i.e. a high quantity of low p-values). The experimental trait, tenderness, is clearly an auxiliary source of variance due to the paucity of variables exhibiting a low p-value. Both PCA and ANOVA confirm that sample run order is the dominant factor that underlies variation in the variable intensities.

6.5.2 Traditional methods

Figures 6.4 and 6.5 reveal that variable intensities vary as a function of the time at which the samples were analysed. Traditional normalisation methods, such as by using the total ion current (TIC) or an internal standard, will only work if the intensity drift for all variables is the same over time.

6.5.2.1 Total ion current

Normalisation of each observation to its total ion current (TIC) is perhaps the most basic of methods to correct for variations in intensities as a function of time. It relies on the assumption that the total amount of injected sample is constant for all observations, and that the mobile phase composition is also constant. It does not, however, account for external factors which may influence instrumental performance. An example of such a factor would be a cleaning regime which may be performed between batches, especially in the case of long-term studies.

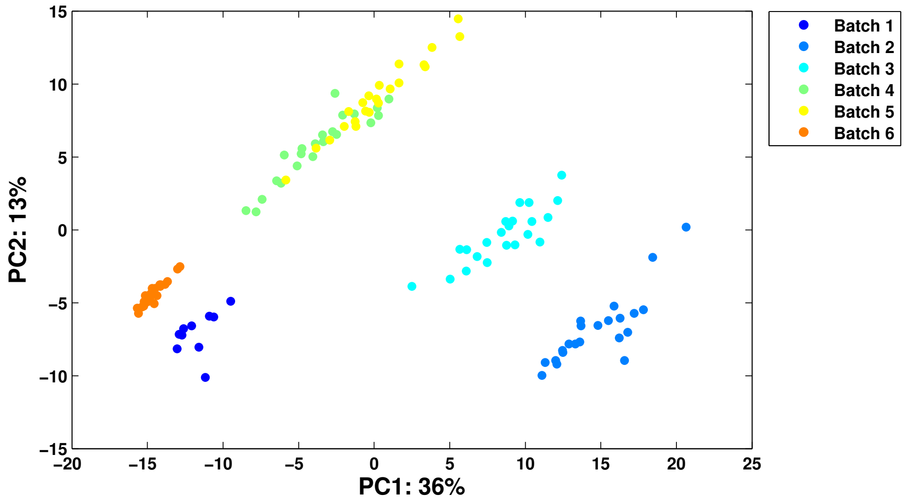


FIGURE 6.4: PCA scores plot for the positive-mode RP-LC-MS variables, where observations are coloured according to the batch in which they were analysed.

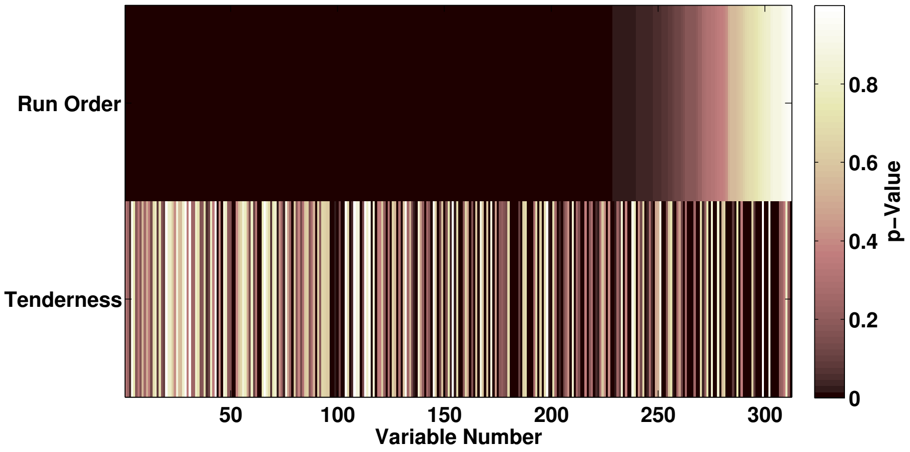


FIGURE 6.5: The two-way ANOVA plot for the positive-mode HILIC variables, showing which variables are associated with which of the two factors. The p-values have been sorted so that it is clear that more variables exhibit run order-like dependency than vary according to the sample tenderness values.

6.5.2.2 Internal standard

The addition of an internal standard can help to identify issues with individual samples. Correction is performed by dividing the intensity of each variable in an observation by the intensity of that observation's internal standard. The method assumes that the variability in intensities of the internal standard mirrors any such drifts in all other analytes. Whilst this might be true in certain situations, the normalisation of variables according to the internal standard failed to meaningfully reduce the batch-dominance trend observed by PCA and ANOVA.

6.5.3 Variable-by-variable methods: QC sample mean

In order to effect the correction of variable intensities on an individual basis quality control (QC) samples are required. These must be factored in as part of the experimental design and, ideally, a QC sample is a pool of aliquots from all analytical samples. In this way, it contains a fraction of each sample so that it is representative of the entire experiment. When run periodically throughout the batch, variation in intensities of features in the different data from the QC samples can be used to correct for batch-to-batch differences in intensities for each variable. Assuming that the components of the QC sample remain constant over the length of the experiment, drifts in individual variable intensities can be corrected by comparison to the corresponding intensities from the QC samples.

Figure 6.6 shows the intensities for a single variable across all observations (QC samples coloured red). QC samples have been periodically analysed throughout the entire experiment (two in each batch), and correction of the intensities towards a more consistent value should help to reduce the observational dependence on time. The six batches can be observed in Figure 6.6: the first three show step-changes in the QC sample intensities, whilst in the latter three batches the intensities arc upwards as the sample run order increases. As it should be expected that the QC sample intensities remain approximately constant, all of the intensities may be corrected by modelling the drift observed in the QC samples. The IHR samples, as described in Section 4.1.1, are used as the QC samples.

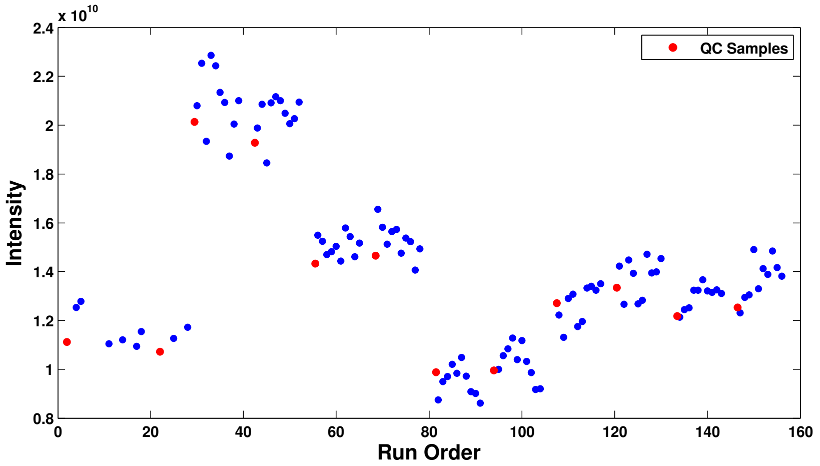


FIGURE 6.6: The intensities of a single variable plotted as a function of run order. The six distinct batches can be observed: the first three show clear differences in intensities, whilst the intensity change in the final three batches is of a more continuous nature.

The correction of sample intensities according to those from QC samples requires, of course, that the variable has non-zero intensities throughout the QC samples. Where a variable is absent from the QC samples, it cannot be corrected and must be excluded. This approach penalises analytes with an insufficiently high concentration to withstand the dilution imparted by the pooling of aliquots from all samples.

The correction of a variable's intensity can be achieved simply by using the mean intensity of the QC samples. So for each discrete batch, the intensities of the observations within the batch are divided by the mean of the intensities from the QC samples. The procedure is summarised below:

1. Calculate mean of intensities of QC samples in each batch.
2. If the median of means is equal to zero, then delete the variable (as is does not appear consistently throughout the experiment).
3. Divide all intensities in a batch by the value determined in step 1.
4. Replace NaN and infinite values with an arbitrary small number.
5. Multiply all corrected intensities by the mean of the uncorrected variable intensities across all observations.

The first stage calculates the mean of the QC sample intensities across all of the batches. In stage 2, these are assessed to gauge whether a variable appears consistently throughout all of the batches. If, for example, only one batch has non-zero QC sample intensities then it is likely that this is not a genuine analyte, nor one on which any statistical inference can be made. Furthermore, a zero value for the mean of the intensity of the QC samples does not allow for correction as division by zero is ineffective; thus by demanding that the median is non-zero, then at least half of the batches can be corrected. The value of including variables which appear in less than half of the batches is very low, and for this reason they are excluded at the beginning. The third stage performs the correction by dividing each intensity by the corresponding QC sample mean, and the fourth stage replaces any non-numeric values which arise due to zero values in either QC sample mean or observation mean. All of the corrected variables are centred around a value of one, and provide no information regarding their actual scale, or abundance. The final, optional, stage restores the overall size of the variable, by multiplying the corrected variable by the mean, as calculated from all observations from the uncorrected data.

6.5.4 Variable-by-variable methods: curve fitting

The relationship between a variable's intensity and the sample run order can be modelled, such that a correction function can be applied to the data on a variable-wise basis. The most simple example of such a function would be a linear interpolation of intensities between QC samples. A linear relationship, however, is unlikely to be valid in all cases and polynomial functions may be more appropriate. An approach that fits linear or quadratic functions between subsets of QC sample intensities has been presented by Dunn et al. [230]. Their procedure determines a series of local regression functions, which are fitted to the data via a weighted least-squares algorithm, and finally stitched together to produce a smoothed correction factor. A 'before and after' illustration for their method is shown in Figure 6.7. The QC samples have been approximately normalised to an intensity of one, and the correction curve broadly follows the drifts in QC sample intensities. The method is suitable for intensities that exhibit only minor drift; variables with larger, batch-to-batch intensity shifts are not suited to this method, as the smoothing of individual regression functions assumes that neighbouring QC sample intensities are on the same scale. A modified approach to that summarised above has been developed for application to data that exhibit significant and abrupt changes in intensity values, as depicted in Figure 6.6.

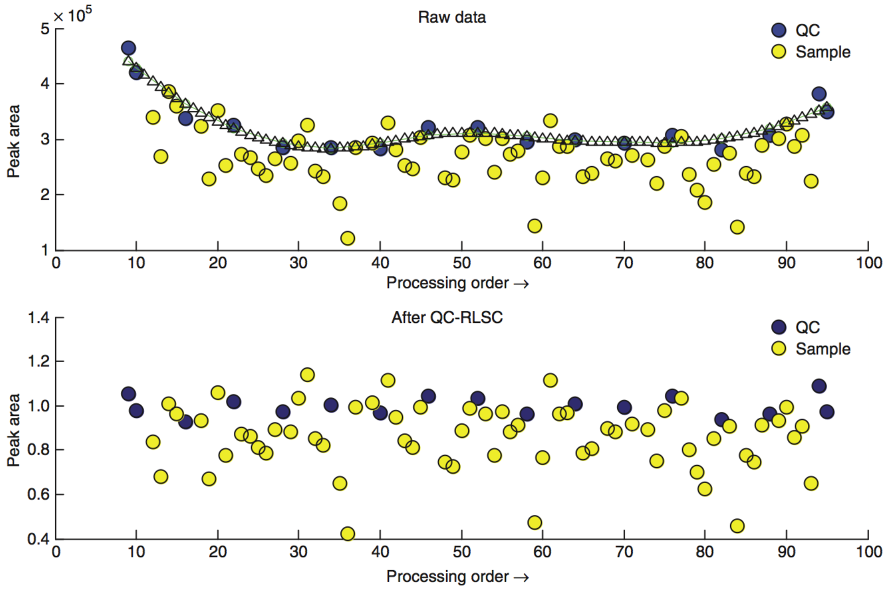


FIGURE 6.7: A variable's raw intensities are depicted in the top plot, and the interpolated correction function is depicted by the triangles. The bottom plot shows the effect of the correction function, such that now the QC sample intensities are considerably more constant. Reproduced from [230].

In order to fit a function to QC sample intensities, the nature of the polynomial must be determined, along with the span of the function, i.e. how many QC samples the function covers. The most localised solution would be fit to pairs of intensity values, whilst the most globalised form would have a single function between the first and last intensity. It is likely that the optimal solution lies somewhere between the two.

The appropriateness of a regression function is often determined by a least-squares difference between function and data. Fitting the non-QC samples to a function is undesirable as, at its most extreme, all intensities will be normalised to the same value. Clearly there is inherent variation in the data which needs to be preserved.

Inherent variation in an observation may manifest itself as an ‘outlier’ from a logical trend of intensity values between QC samples. These ‘outliers’ should not be forced to fit the function to such an extent as those whose intensities drift in a manner more related to that of the QC samples. Consequently, the more removed from the function an observation is, the less significance it should have in determining a function’s validity. Furthermore, to minimise the possibility of an ‘outlier’ influencing the curve fitting, leave-one-out (LOO) cross validation can be implemented (QC samples are never omitted).

6.5.4.1 Batch size

The samples were analysed by LC-MS over six days, and inspection of Figure 6.6 reveals that the first three batches exhibit changing intensities, whilst the latter three batches show a gradual change from one QC sample to the next. Whilst this plot shows the pattern for a single variable, a similar trend is visible throughout many others (not shown). As was the case for correction according to the QC sample mean from each batch (section 6.5.3), each of the six batches can be processed individually; this instance will fit six independent functions to the six batches of data. Alternatively, the first three batches can be treated individually, whilst the latter three have a single applied function. The first of these two alternatives was initially implemented, and it allows a direct comparison to the more simple batch mean normalisation method.

6.5.4.2 Function

The choice of function that is fitted to intensity values is practically limited by the number of data points. A best-fit linear solution is possible with two data points, a quadratic with three, and so on. It is probable that a variable’s intensities drift either up- or down-wards over the course of a batch and, as such, the use of higher order polynomials represents an over-fitting. Consequently, linear and quadratic functions are the most appropriate. If there were more than two QC samples per batch, then perhaps an extension to higher order polynomials may be appropriate.

6.5.4.3 Weighting

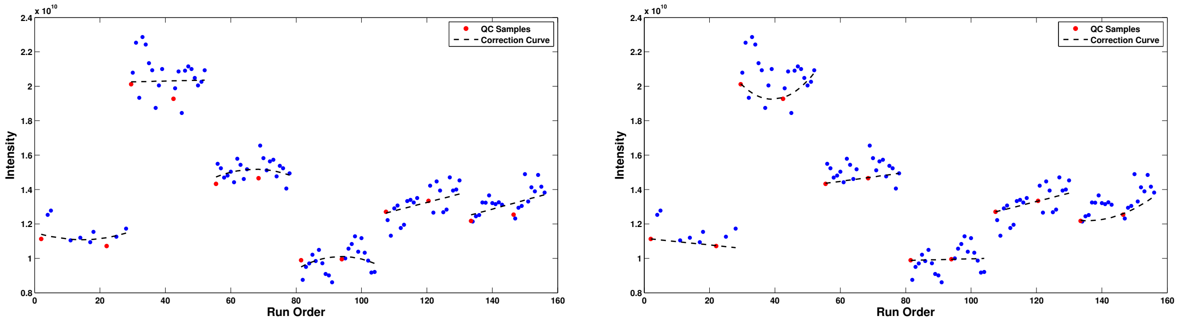
The 'line of best fit' is usually that which reduces the residual sum of squares to the smallest value. Essentially it represents the error between a function and the experimental data. A weighted approach to the sum of squares is more appropriate in this method, as not all variable intensities should contribute equally when determining the best function. Indeed, the aim is to fit a line between the QC sample intensities, but the other intensities must influence the line to some extent. Intensities that differ greatly from the QC sample values should be less influential that those that are closer; the reason for this is that the variable with a greater difference is exhibiting signs of biological variation, rather than, for example, a drift attributable to time. By reducing the influence of observations with a greater difference to the QC sample mean, a function can be better fit to resemble the time-dependent drift, as opposed to factors inherent to the observations.

The determination of the best function depends on the difference between observations and the mean of the QC samples. A weight for each observation is calculation, which represents the relative contribution made in the determination of the line of best fit. The residual sum of squares for each observation, a_i, between its intensity value (x) and the mean of the QC sample intensities ($\bar{x}_{QC}$) is used to calculate the weight, w_i, for each observation. The calculation of the weights is shown in Equation 6.3, where they are scaled between 0 and 1. The procedure is known as a weighted sum-of-squares regression.

$$\begin{aligned} a_i &= (x_i - \bar{x}_{QC})^2 \\ w_i &= 1 - \frac{a_i}{max(a)} \end{aligned} \tag{6.3}$$

Although the weighting of variables limits the influence of more extreme intensity values, the function should be significantly biased towards the QC samples. By augmenting the QC sample weights by a specific factor, the determination of the best-fitting function will place a greater emphasis on QC samples, and reduce the effect of other observations with a QC-like intensity. After all, the aim is to correct the drift observed in the QC sample intensities, rather than to strip out any inherent variation between observations.

Two exemplar correction curves are shown in Figure 6.8, which demonstrate the effect of over-weighting the QC sample intensities for the purpose of curve fitting. Figure 6.8a shows the curve where no additional bias is imparted on the QC samples (i.e. QC factor equals 1), whilst the QC factor in Figure 6.8b is set to 100. The difference between the two is clear, in that the additional QC sample bias forces the lines of best fit through the QC sample intensities.



(A) Here the value for the QC factor is 1. **(B)** The QC factor has been raised to 100.

FIGURE 6.8: Two sets of correction curves are shown for the same variable from positive mode RP-LC-MS. In (A) the weights for all observations are equal, with the weightings derived from the difference from the QC mean (see Equation 6.3). In (B) the weights have been multiplied by 100, such that in (B) the correction curves closely fit the QC sample intensities.

6.5.5 Results and discussion

The QC sample curve fitting method described above has been implemented to the reversed-phase positive mode LC-MS pea data. The objective is to remove the strong dependency between variance and time, which was demonstrated in Figures 6.4 and 6.5. PCA is used to visualise trends in data, and it is often the case that inherent variance is expressed in the new variables produced by PCA. The D^2 metric, introduced in Chapter 4, is used to express the similarity between the PCA scores and the tenderness values.

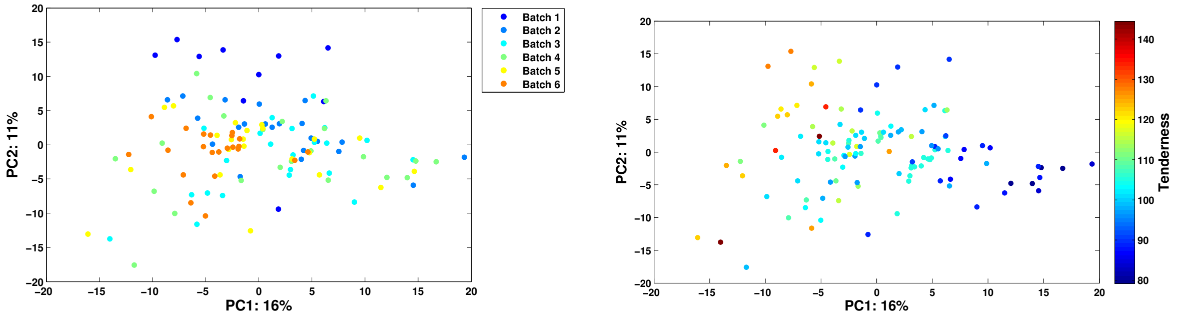
The D^2 values obtained for each method and parameter are shown in Table 6.1. It can be seen that correction according to the internal standard is clearly ineffective, and the results from PCA and ANOVA (not shown) confirm that the method fails to effectively remove the inherent run order dependency.

TABLE 6.1: The D^2 values for the various correction methods when applied to the reversed-phase positive mode LC-MS variables.

Method	Cross Validation	QC Factor	D^2
Raw data	-	-	0.126
Internal standard	-	-	0.084
Batch mean	-	-	0.465
Curve fitting	None	1	0.518
"	"	2	0.519
"	"	5	0.514
"	"	10	0.507
"	"	20	0.500
"	"	100	0.482
"	Leave one out	1	0.515
"	"	2	0.513
"	"	5	0.506
"	"	10	0.495
"	"	20	0.491
"	"	100	0.471

The use of the QC samples has, however, dramatically improved the relationship between the variables and the tenderness values. The most simplistic method, by correction according to mean QC sample intensity from each batch, is not significantly worse than the more involved methods that calculate a correction function using local regression. Two PCA scores plots of the batch mean corrected data are shown in Figure 6.9, where the former is coloured according to the batch and the latter to the tenderness values. A comparison of Figures 6.4 and 6.9a highlights the extent to which the run order dependency has been removed. Figure 6.9b shows a clear relationship between PC1 and the tenderness values.

Table 6.1 reveals that the D^2 values from the more complicated curve fitting method are higher than that obtained from the batch mean method. There is little difference in the D^2 as a result of applying cross validation; this demonstrates that outliers are not overly influencing the curve fitting.



(A) The observations are coloured according to batch, and the trends seen in Figure 6.4 no longer dominate.

(B) The tenderness values are used to colour the observations, which are given in the colourbar.

FIGURE 6.9: PCA scores plots of the same batch mean normalised data, where the observations have been coloured according to batch and tenderness.

TABLE 6.2: The results of the various correction procedures that have been applied to remaining LC-MS datasets.

Method	Cross Validation	QC Factor	RP-	HIL+	HIL-
Raw data	-	-	0.222	0.212	0.251
Internal standard	-	-	0.245	0.206	0.132
Batch mean	-	-	0.391	0.570	0.497
Curve fitting	None	1	0.458	0.586	0.536
"	"	2	0.462	0.596	0.538
"	"	5	0.461	0.604	0.538
"	"	10	0.462	0.603	0.533
"	"	20	0.430	0.596	0.524
"	"	100	0.388	0.584	0.511
"	Leave one out	1	0.455	0.593	0.534
"	"	2	0.465	0.591	0.534
"	"	5	0.459	0.599	0.540
"	"	10	0.442	0.593	0.532
"	"	20	0.416	0.586	0.501
"	"	100	0.372	0.572	0.497

As the QC multiplier (i.e. by what factor the weights of the QC sample observations should be multiplied) is increased, such that the QC samples hold greater dominance over the regression, the D^2 values tends to decrease. This suggests that higher relative weightings for non-QC samples artificially enhance the relationship between tenderness and the principal components. Although the difference in D^2 between small and large QC factors is relatively small, it is clear that as the QC sample weights dominate, the results approach that of the batch mean normalised data.

Whilst the curve fitting procedure produces the best results (by one metric at least), it is offset by the parsimony of the batch mean normalisation method. This latter method is computationally much quicker to implement, and produces results of a comparable nature. Furthermore, there are no parameters (e.g. weightings) which need to be optimised. The selection of the 'best' method may, therefore, be a question of simplicity. The focus has, so far, been placed on only one out of the four LC-MS data sets. The results from the remaining three are presented in Table 6.2.

6.5.6 Conclusion

It has been shown the some LC-MS data exhibit significant drifts in intensity values with respect to time. This variation in signal amplitude cannot always be corrected by the use of an internal standard. Instead, the periodic analysis of a QC sample is a more appropriate method that can be used to track, and model, the changes in intensities of individual variables with respect to the run order. The approach is limited, however, to analytes that are not deleteriously diluted in the preparation of a pooled QC sample. Analytes that are not detected in the QC samples cannot be corrected, and thus must be removed regardless of their information content.

The use of regression functions to correct drifting intensities is essentially a supervised technique as the non-QC samples must necessarily be included when determining the line of best fit. The influence should, however, be minimal with most of the influence, or weighting, placed on the QC samples. The weightings, QC multiplying factor, and the use of leave-one-out cross validation have all been tested in order to gauge the effect of bias imparted by non-QC samples. Giving an equal weighting to QC and non-QC samples results in a biased correction curve; choosing the correct weightings requires careful investigation in order to avoid artificially enhancing the data. The results from the simpler method involving correction according to the mean intensity of the QC samples are not altogether inferior to those from the curve fitting approach; due to the simplicity of the mean correction method, this method has been used to correct the run-order dependent intensity drift in all four of the LC-MS data sets.

6.6 Summary

The processing of LC-MS data files is a crucial part of any analysis. The quantity of data within mass chromatograms is enormous, and adequate processing tools facilitate the conversion of the data into information. Coeluting and fragment ions along with isotopic distributions allow for a greater understanding of the sample composition to be achieved.

A bespoke peak processing routine has been developed, but it requires improvements to enhance the efficiency. The method converts profile data by modelling each peak with a Gaussian function, hence allowing for extensive data reduction and spectral reconstruction. Furthermore, the *m/z* values are not arbitrarily rounded which prevents peaks from being split and maintains the inherent precision of the data.

Whilst the extraction of meaningful peaks from mass chromatograms is important, it does not represent the final processing stage. The reproducibility of LC-MS is as yet insufficient and a major source of experimental variance is likely to be observable between samples analysed at different times. Spectral intensities may show considerable drift, and the development and use of a quality control system is essential in ensuring that variance as a result of sample acquisition time does not influence subsequent conclusions. Various normalisation methods have been implemented for the effective removal of run order dependency. The most appropriate method for the pea data was found to be correction according to the mean intensity of the QC samples in each batch.

The LC-MS variables produced by XCMS and subsequently corrected by the procedures outlined in this Chapter are used subsequently to assess the complementarity that exists between multiple analytical datasets.

Chapter 7

Data Fusion

> However beautiful the strategy, you should occasionally look at the results.
>
> Winston Churchill

7.1 Introduction

Data fusion is a term that describes the process of using multiple datasets in a concerted fashion. The aim of fusion is that the output from the fused system provides more information than would have been achieved were each dataset analysed individually [15]. The combination of datasets can be effected in one of many ways, and data fusion methods may be further sub-divided according to the way in which variables are combined.

Low level fusion is conceptually the most simple, in which variables from multiple datasets are combined by concatenation of data matrices. Intermediate fusion techniques generally combine subsets of original variables across multiple techniques, whilst high level techniques combine datasets after they have been independently analysed. Recently, an additional class of fusion has been developed: Smolinska et al. [340] have applied non-linear kernel approaches to data with non-linear responses. One of the major considerations in any fusion method relates to the scaling and standardisation of variables, so that one dataset does not dominate subsequent analyses [20, 341].

As many of the commonly applied statistical techniques analyse only a single data matrix, it is the development of multiblock approaches that allows for techniques such as principal components analysis (PCA) and partial least squares (PLS) to be applied in high level fusion approaches. Multiblock PCA (or PLS) [18, 342] analyses separate blocks of related variables in order to find a consensus, or common trend, between the blocks. A data block may consist of variables from a single analytical technique, but could also take account of other sources of variance that might otherwise confound a single block approach. For example, in a study by Biais et al. [23] the data were blocked according to both analytical platform and the sample cultivar, whereas Janné et al. [307] separated infrared spectra into regions characteristic of specific functional groups.

Low and high level approaches have been applied to the fusion of LC-MS and ^{1}H NMR datasets by Forshed et al. [20], who also assessed various scaling and standardisation metrics. The high-level approaches involved the two-stage use of multivariate analyses, where individual analysis of each dataset was performed, and the results then concatenated prior to a second analysis. Both principal components analysis (PCA) and partial least squares (PLS) were used, and the best between-group separation of observations was achieved when discriminant analysis mode PLS was performed on concatenated subsets of PCA scores.

Doeswijk et al. [32] applied canonical correlation analysis (CCA) in another high level fusion approach. CCA differs from many other multivariate techniques as it operates on two distinct data matrices. The PLS scores from the analysis of GC- and LC-MS data produced two sets of new, latent variables which were analysed by CCA. This method seeks to maximise the correlation between pairs of linear combinations, one from each data matrix. The study identified related spectral trends that were observable across both datasets.

The technique developed by Crockford et al. [33] employs a simple approach to the fusion of MS and NMR data. Their technique, termed statistical heterospectroscopy (SHY) calculates the pair-wise correlation between variables from two complementary datasets. A similar approach by Moco et al. [34] has also been applied to MS and NMR data. The technique is computationally demanding as hundreds of variables in each dataset require the calculation of upwards of 10,000 correlation coefficients. These correlation approaches aim to help the assignment of features, rather than to enhance the between-group discrimination.

The key aim of data fusion is to improve the information recovery by virtue of combining multiple data sources. In relation to multivariate approaches, one common measure of success relates to the enhancement in the between-group discrimination in fused systems, compared to their single block equivalents. The enhancement in between group separation, or more generally the agreement between multivariate results and an experimental outcome, may help to identify specific variables that are highly discriminatory and those with strong relationships between blocks. Thus data fusion may identify relevant variables that would otherwise have been missed were only single block approaches employed, and these variables may assist with feature assignment. Within this Chapter various existing fusion methods have been implemented and modified, with the aim of assessing their applicability and merit. This Chapter also details the use of data pre-treatment, and how these influence the various techniques.

7.2 Experimental

7.2.1 Data pre-processing

The large dynamic range of analytical instrumentation allows the detection of molecules across a large concentration range, but is not necessarily related to the significance of a molecule. Many biologically necessary molecules exist at relatively low concentrations, and their importance will most certainly be overlooked without appropriate pre-processing. For example, high intensity variables may have a greater influence than smaller ones, but this difference in magnitude may inhibit the smaller, yet more informative, variables from effectively contributing. The use of pre-processing techniques may, therefore, be applied to data sets such that the influence of variables is not necessarily determined by, for example, their magnitude.

Scaling and standardisation of variables are terms that are occasionally used to represent similar mathematical operations. Scaling is generally used to refer to operations on variables that act to convert the scale, such as dividing a variable by its mean. This process may be considered as analogous to unit conversion whereby, for example, metres are converted into kilometres. Variable standardisation uses a measure of the variable's spread, such as its standard deviation, with which to convert the original data. Normalisation, in mathematical terms at least, refers to the process of dividing a vector by its norm. A vector norm, $||x||_n$, is determined by Equation 7.1.

$$||x||_n = \sqrt[n]{\sum_j |x_j|^n} \tag{7.1}$$

Some of the more commonly applied scaling and standardisation methods are listed in Table 7.1. Mean centering does not alter the information content of a variable, and merely focuses all of the variables' fluctuations around zero. The effect of unit variance-scaling (UV, also termed autoscaling) is to give every variable a mean of zero and standard deviation of one, so that each variable has an equivalent influence. Variable stability, or vast, scaling emphasises variables which undergo smaller changes and downscales the influence of variables with larger standard deviations. In Pareto scaling the larger variances are reduced by a greater proportion than smaller variances, and it may be considered as less extreme version of autoscaling. Vector normalisation differs in that it is applied to observations rather than variables. The effect is to keep constant the ratios between an observation's variables, and the effect is to group correlated observations together (see, for example, Scholz et al. [305]). The block scaling methods [20] may help to account for differences between two analytical techniques: scaling the results from the low level MVA according to the sum of its standard deviations prevents one block from dominating, and block mean scaling corrects for intensity differences between the multiple analytical techniques.

7.2.1.1 Variable distribution

The processes described above deal only with magnitude and variance, and do little to alter the distribution of variables. Some statistical techniques, for example the t-test, rely on the assumption that the variables are normally distributed. Of course a non-parametric equivalent, i.e. one that assumes no underlying distribution, could be applied where such assumptions are invalid (in this example, the equivalent would be the Mann-Whitney U test).

TABLE 7.1: Various transformation methods applied to the data, on either a variable, block or observational basis. i refers to each observation, and j to each variable. The original data x_{ij} is transformed to the new value y_{ij} according to the equations below. $\bar{x_j}$ is the mean of the j^{th} variable and σ_j is its standard deviation. The definition of a vector norm is given in Equation 7.1.

Method	Type	Operation
None (mean centering)	Scaling	$y_{ij} = x_{ij} - \bar{x_j}$
Block standard deviation	Scaling	$y_{ij} = x_{ij} \div \sum_{j=1}^{J} \sigma_j$
Block mean	Scaling	$y_{ij} = x_{ij} \div \sum_{j=1}^{J} \bar{x_j}$
Unit variance/autoscaling	Standardisation	$y_{ij} = (x_{ij} - \bar{x_j}) \div \sigma_j$
Vast	Standardisation	$y_{ij} = \frac{\bar{x_j}}{\sigma_j} \times \frac{x_{ij} - \bar{x_j}}{\sigma_j}$
Pareto	Standardisation	$y_{ij} = (x_{ij} - \bar{x_j}) \div \sqrt{\sigma_j}$
Vector	Normalisation	$y_{ij} = x_{ij} \div \lvert\lvert x_i \rvert\rvert_n$

Variables with skewed distributions can be corrected by a series of functions often referred to as either the ladder of powers or transformations [343]. Tukey's original aim was to apply a function to data such that it becomes linearised in the form of $f(x) = mx^{\lambda} + c$, where λ represents the power that produces the most linear function. The approach is equally valid to the transformation of variables with a non-normal distribution. A simple measure of a variable's normality is its skewness; negatively skewed variables have a few unusually low intensities, whilst those with positive skewness exhibit a few unusually large values. Table 7.2 details the various functions which can be used to transform data, alongside a qualitative indication of the extent to which the skewness can be corrected. It should of course be noted that outliers can have a large effect on a variable's skewness; an otherwise normally distributed variable may appear to be highly skewed as a result of a single outlier.

In order to gauge the validity of increasing the overall normality of a dataset, the various functions listed in Table 7.2 were individually applied to each of the variables from positive-mode HILIC LC-MS. Figure 7.1 plots the new skewness values

TABLE 7.2: The ladder of transformations that can be applied to variables to, for example, increase their linearity or reduce their non-normality of distribution.

λ	$f(x)$	Skewness
-2	$-1/x^2$	Negative
-1	$-1/x$	
-1/2	$-1/\sqrt{x}$	
0	$log(x)$	
1	x	None
2	x^2	
3	x^3	
4	x^4	Positive

as a function of the original values. After determining the post-correction skewness value, the function that reduced the magnitude closest to zero was selected as the optimal transformation. The corrected skewness values are shown in Figure 7.2, and it can be seen that the largest original skewness has been reduced by a factor of 10.

To determine the effectiveness of optimally transforming the variables, the data were first unit variance-scaled and then analysed with PCA. The plot of the first two PCs is shown in Figure 7.3a, together with a plot of the sample tenderness values given as a function of the second PC (Figure 7.3b). On face value, little in the first two components has changed, save for a general reduction in the spread of the scores. The correlation between the second component and the tenderness values remains essentially unchanged. This is also reflected in the D^2 values which are given for the corrected and unchanged variables.

PCA may not be the most appropriate test of the validity of the normality-correcting functions. The number of variables that are highly correlated to the observational tenderness values may provide an indicative measure of whether variable skewness adversely impacts the relationship with tenderness. In the original data 15 variables achieve a correlation coefficient greater than 0.6, and for the corrected data the number is 17. The addition of just two variables demonstrates that no significant advantage is conveyed by such transformations.

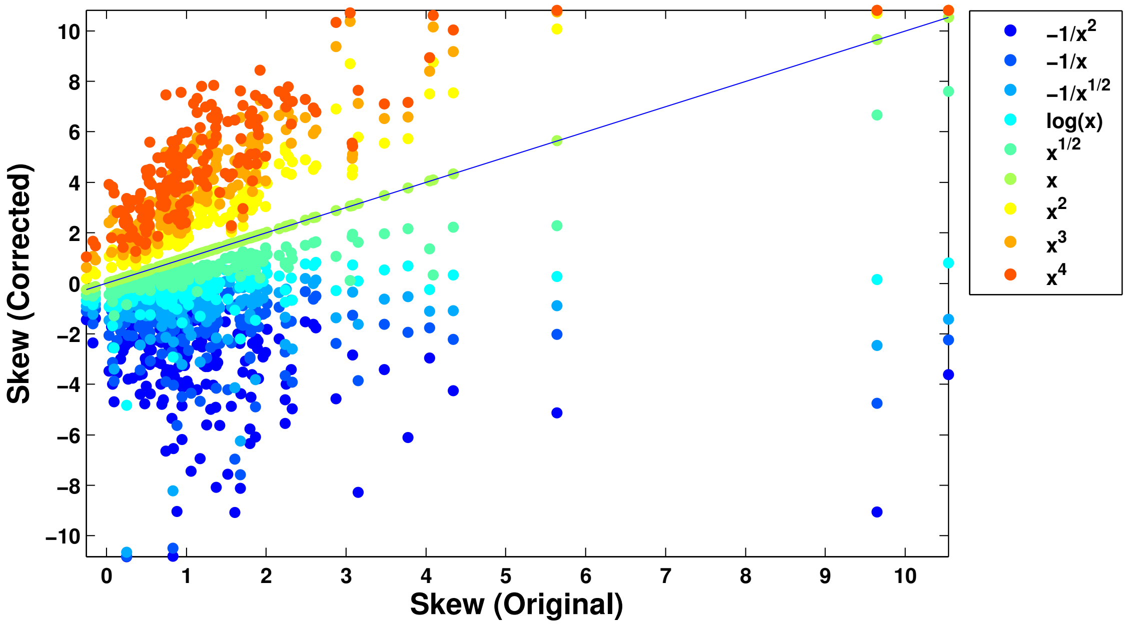


FIGURE 7.1: The skewness values of the variables before and after each of the various transformations have been applied. The line indicates where no transformation has occurred.

Nevedomskaya et al. [344] logarithmically transformed metabolomics data from both LC-MS and NMR, prior to unit variance scaling and PCA. The effects of a blanket transformation to the HILIC data are shown in Figure 7.4, where neither the D^2 or ρ values have increased as a result. The blanket application of such a transformation relies on all variables exhibiting approximately the same skewness values. van den Berg et al. [341] have also assessed the effects of log scaling, noting that as intensities tend to 0 their logarithm tends to $-\infty$, hence accentuating variables with high variance but low intensities. Whilst all of the variables in Figure 7.1 have a positive skewness, it is clear that the $log(x)$ function is not always the most appropriate at reducing the value as close to 0 as possible.

Although skewness is not considered as the most robust measure of non-normality, its minimisation is demonstrated to have little effect on the information content within the first two principal components. The D^2 and ρ values also fail to show any improvements, and as such, no transformations are implemented here prior to multivariate analysis.

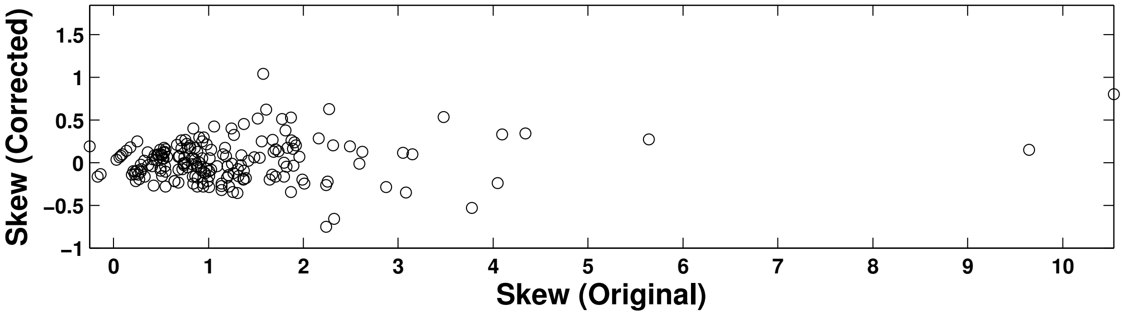
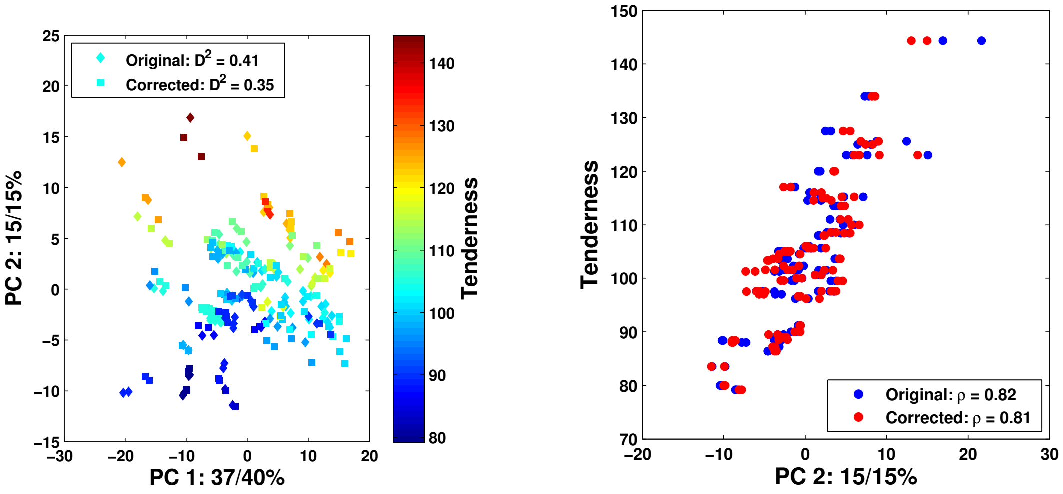


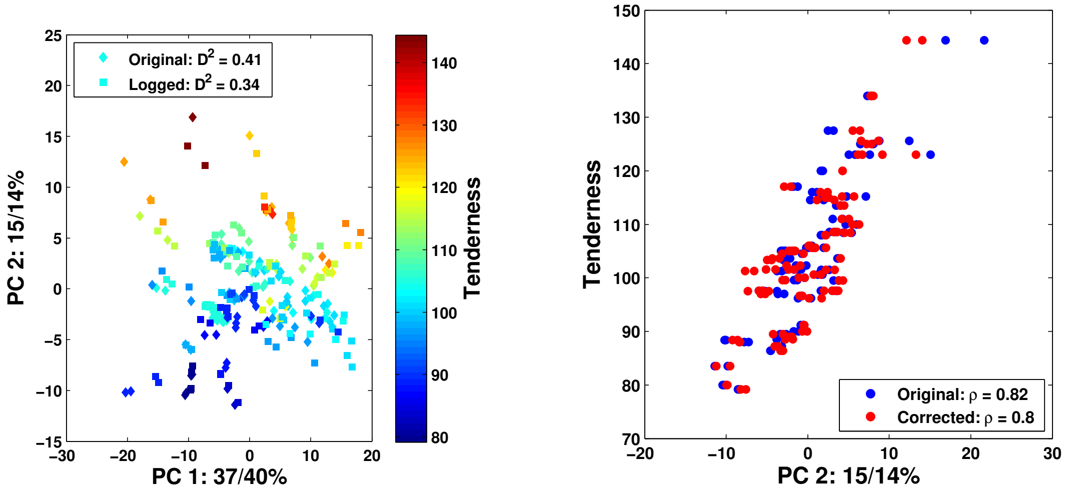
FIGURE 7.2: The original and corrected skewness values are plotted for each variable. The optimal skewness is defined by the transformation that reduces its absolute magnitude closest to zero.



(A) The first and second principal components of the original and corrected data.

(B) The second PC correlates best with the tenderness values.

FIGURE 7.3: Principal components analysis of the original and corrected data variables. The scores for the original and corrected variables are shown in each plot. The second PC is most strongly related to tenderness. For both metrics, the agreement between the scores and tenderness has fallen as a result of the variable transformations. The percentage of variance explained by each component is given on the axes, with the former being for the original data and the latter for the corrected data.



(A) The first and second principal components of the original and logged data, showing a decrease in the D^2 value.

(B) The second PC correlates best with the tenderness values, but there is no improvement when all variables are logged.

FIGURE 7.4: The effect of logarithmically transforming all variables from the positive mode HILIC LC-MS data sets. Neither the D^2 value nor the correlation with tenderness (ρ) improved following the transformation.

7.2.2 Single block methods

Three single block multivariate analysis (MVA) methods have been implemented here, and form the root of many of the fusion methods. More detailed explanations of the methods have been given in previous Chapters, but they are briefly summarised here. All of the methods are common in their output; from an original matrix of data variables, each produces a series of new variables (scores) alongside the loadings. These loadings represent the contributions made by the original variables to the scores.

Principal components analysis (PCA, Section 2.6.2.1) is an unsupervised multivariate technique. It takes linear combinations of original variables that maximise the inherent variability within the data. Data reduction is achieved by accounting for most of the variance (information in the data) in just a few components. It is often used to identify trends within a data set, but its unsupervised nature means that expected trends may not be visible in the first few components.

Independent components analysis (ICA, Section 3.5.3.3) is, like PCA, an unsupervised variance-based method. Whilst PCA creates new variables that are uncorrelated with each other, ICA seeks to maximise the independence between these new variables. Independence is a stronger measure of dissimilarity than non-correlation, and it is therefore expected that the results between ICA and PCA should differ. The main disadvantage of ICA is that it does not function on rank deficient matrices and, in metabolomics, requires an initial data reduction stage, which is typically PCA. As such, ICA essentially maximises the independence between principal components rather than the original variables.

Partial least squares regression (PLSR) is a supervised technique that uses *a priori* class information to form linear combinations of original variables that simultaneously maximise the variance between the variables and the class information. As shown in Section 2.6.2.3, it is the weights of X, rather that its loadings, that reveal the relationship between X and Y. As such, variable contributions in PLSR are defined by their weights and not their loadings. The technique's supervised nature renders the solution at risk of being over-fit, and cross validation can help to mitigate such effects. External, or hold-out validation uses a purely independent test set, but this is subject to the vagaries of arbitrary (including arbitrarily random) choice. For all approaches used henceforth, 10-fold internal cross validation has been applied in order to estimate the error; the number of components is selected in order to minimise the predictive error.

7.2.3 Concatenated methods

Concatenation may be regarded as the simplest conceptual form of data fusion, and is often described as a 'low-level' technique. It is appropriate for two or more data blocks, which are individually standardised and arranged into a 'supermatrix' by concatenation of the variables. Figure 7.5 shows the concatenation of three data blocks into a single supermatrix. A subsequent multivariate analysis produces, generally, a series of new variables (scores), along with the contributions made by the original variables to the new scores. As with the single block approach, the optimal PLS model is chosen to minimise the error, whilst for PCA and ICA the number of components is varied to optimise the D^2 value.

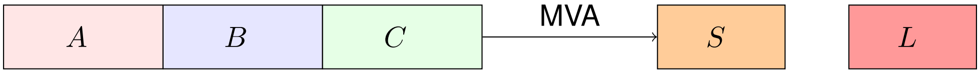


FIGURE 7.5: A diagrammatic representation of the concatenation of three previously standardised data matrices (A, B, C) prior to multivariate analysis, which, in these methods, may be PCA or PLS. This example shows the formation of matrix S (scores, latent variables, etc...) and L (loadings, weights, etc...). S represents new variables, whilst L reflects the contributions made by the original variables.

7.2.4 Multiblock approaches

Multiblock, or 'high-level', fusion involves two multivariate analysis stages. The first is performed on each individual data block, whilst the second is performed on a matrix formed from the concatenation of scores from the preceding multivariate analyses. Multiple combinations of analytical technique can be used, and each of these is discussed below. Each dataset is independently standardised, as are the concatenated scores prior to the second MVA.

Perhaps the earliest application of multiblock PCA to the fusion of NMR and LC-MS data was that of Forshed et al. [20], who employed a two block approach to enhance the between-group discrimination of observations. Multiblock methods are often employed to highlight a common trend across data blocks. Such trends may be missed if the data blocks are concatenated as other sources of variance may dominate. This is highlighted in the assessment of metabolite profiles at various spatial locations in melons [23]. The between-cultivar variance of various observations was demonstrated to be 'obscuring' the experimental variance of interest. The data were, therefore, separated further into six blocks: two analytical platforms (NMR and GC-MS) and three cultivars. The multiblock PCA was therefore able to accentuate commonalities across the blocks. The variance between pea cultivars in this experiment does not dominate any tenderness trends; as such, the extra blocking of observations has not been implemented.

A 'pure' hierarchical method has been considered as one which uses the same MVA technique in both stages. In PCA-PCA the initial MVA is performed on each original data block, and a subset of the scores from each of these analyses is concatenated prior to the second PCA. In order to determine the optimal number of scores to combine, this implementation employs an iterative procedure whereby the model is assessed each time a new PC is added from each initial MVA; the subset of PCs with the highest D^2 value is taken as the optimal number. A similar approach is employed in hierarchical ICA, except that the second MVA is ICA rather than PCA.

The PLS-PLS method combines all of the scores from the multiple error-minimised single block approaches. These concatenated PLS scores are then subjected to the second PLS, and the number of components is selected to again minimise the predictive error. The method is not optimised to find the most predictive model, solely to minimise the error.

In the mixed-mode hierarchical methods, two different MVA techniques are used. In PLS-PCA all of the single-block PLS scores are concatenated, and these are analysed by PCA. In PCA-PLS the unsupervised nature of the initial MVA allows for a varying number of PCs to be used in successive analyses by PLS. Of the various permutations, the best PCA-PLS method is that with the lowest error.

Figure 7.6 outlines the two-stage procedure, as applied to three data matrices. Each of A, B and C are subjected to an initial multivariate analysis, where subsets of the new variables, S_1, are combined for the final statistical analysis. The results from this stage are not expressed in terms of the original variables, such that any observed trends may not be immediately obvious. A transformation is required to express the second set of loadings (L_2) as functions of the first sets (L_1)[1].

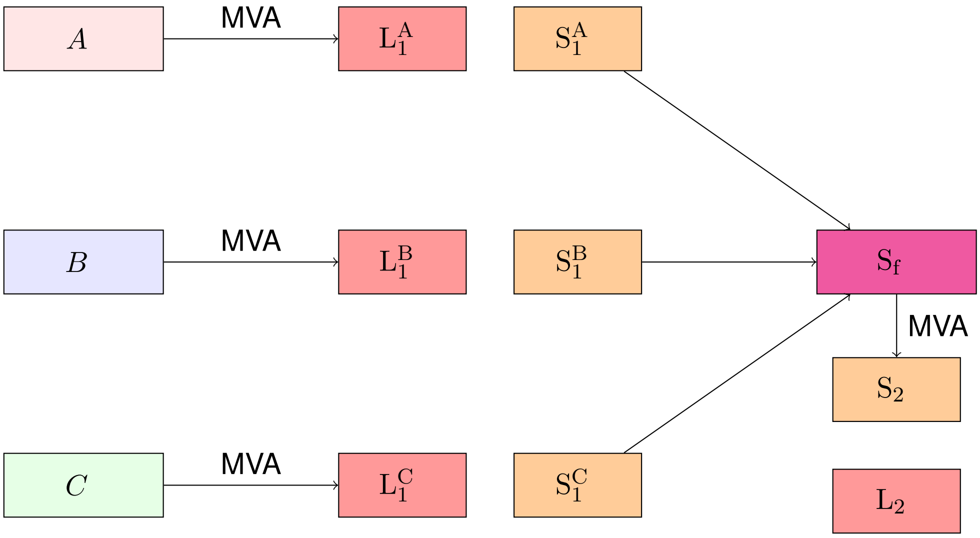


FIGURE 7.6: A representation of a hierarchical fusion method, involving three data blocks (A, B, C). Each of these three is individually standardised and subjected to the same multivariate analysis, which results in new variables (S_1) which are then partially combined in a superscores matrix (S_f). This matrix is then standardised and subjected to a further round of multivariate analysis.

[1] For clarity, references to PLS loadings actually refer to the weights.

As many multivariate techniques take linear combinations of original variables, the relationship between original (A) and new (S_1^A) variables can be expressed as

$$S_1^A = A \times L_1^A \tag{7.2}$$

whereby L_1^A is the loadings matrix. For the second MVA, scores matrix S_1^A is related to the final set of scores S_2, where

$$S_2 = S_1^A \times L_2 \tag{7.3}$$

Thus the relationship between A and S_2 is shown in Equations 7.4 and 7.5.

$$S_2 = A \times L_s \tag{7.4}$$

$$L_s = L_1^A \times L_2 \tag{7.5}$$

7.2.5 Canonical correlation

Canonical correlation analysis (CCA, Section 3.5.4.2) is a multivariate statistical technique that can be used to find correlated variables across two data matrices. It may be considered as an extension of multiple linear regression. Simple linear regression seeks to explain the relationship between a single predictor (x) and a single response variable (y), and is shown in Equation 7.6. Multiple linear regression, as shown in Equation 7.7, maintains a single predictor variable, but has more than one response variable. CCA expands this further by having multiple variables in each of the X and Y data matrices. They are not considered as predictor and response variables (nor dependent and independent variables), and their order of usage in the analysis does not affect the results. A representation of CCA is shown in Equation 7.8.

$$y = ax + b \tag{7.6}$$

$$y = b + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{7.7}$$

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b_1y_1 + b_2y_2 + \cdots + b_my_m \tag{7.8}$$

CCA uses two data matrices, X and Y, which by default are autoscaled, with sizes $m \times q_x$ and $m \times q_y$, whereby m equals the numbers of observations (rows) and q the number of variables. Whilst the number of variables in each matrix may vary, the number of canonical pairs (CP) is limited by the minimum of q_x and q_y. The first CP, as shown in Equation 7.9, is formed by taking linear combinations of X and Y, such that the correlation, ρ, between these combinations is maximised.

$$\left.\begin{aligned} u_1 &= X \times a_1 \\ v_1 &= Y \times b_1 \end{aligned}\right\} \rho(u_1, v_1) \rightarrow max \tag{7.9}$$

Subsequent CPs are calculated such that they are uncorrelated with the preceding CPs. CCA can only be applied to data blocks which are not 'rank deficient'. In the context of metabolomics data, such a situation arises due to the large number of inter-correlated variables, along with the fact that there may be many more variables than observations. The deficiency in the data can be obviated either by variable selection, or by the use of a multivariate technique that produces new, unrelated variables.

Where variables are to be removed, feature selection (FS) algorithms may be applied. These may run in either forward or backward mode, where the objective of the former is to add descriptive variables and that of the latter is to remove those considered as redundant. The definitions, however, of descriptive and redundant variables often depend on their correlation with other variables, such that a small subset of uncorrelated variables are often found to be the most representative combination. It is plausible that only one highly correlated variable is selected which, along with other uncorrelated variables, provides an ample description of the full dataset. This description, however, excludes too many 'interesting' variables, i.e. those that are related to the variance of interest.

The impracticality of using stepwise FS with the number of variables in metabolomics means that the use of methods such as PCA to produce a series of new and uncorrelated variables is preferable. Thus, in order to ameliorate the effects of rank deficiency, a two-stage CCA has been adopted. In order to effectively reduce intercorrelation, PCA and PLS were used to produce a series of uncorrelated variables, which are then used in CCA. The inputs for PLS-CCA are all of the components from the error-minimised single-block method. The optimal number of components used in PCA-CCA is is that which produces the best D^2 value. A similar two stage approach has been implemented by Doeswijk et al. [32] who use PLS to provide CCA with scores related to the experimental variance under investigation.

Due to the large number of input and output matrices involved in two-stage CCA, the procedure is summarised in Figure 7.7. The raw data are initially processed by data reduction methods, which produce a series of uncorrelated new variables. These are used in CCA to produce linear combinations that maximise the correlations between variables. The loadings from CCA are then multiplied by the loadings from the original data MVA to produce a series of loadings that show the contributions of the original variables.

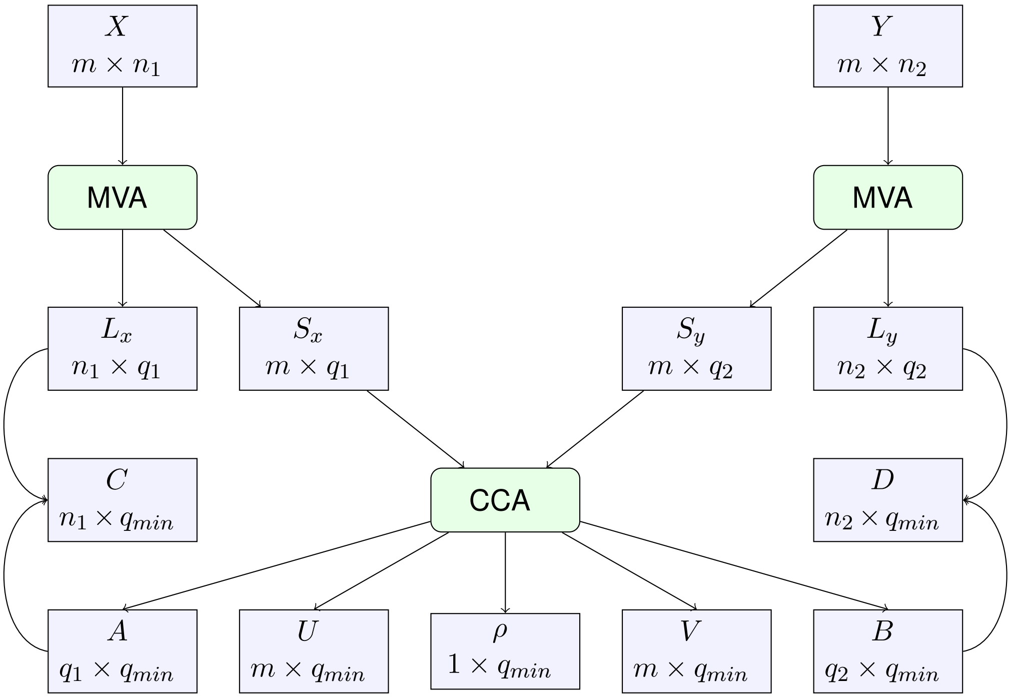


FIGURE 7.7: A flowchart detailing the various matrices involved in a two-stage CCA. The original data is transformed into new variables, S_x and S_y. Canonical correlation analysis (CCA) is performed, usually, on subsets of S_x and S_y. The new scores (U and V) are calculated such that the correlations (ρ) between each pair are maximised. Matrices C and D relate the contributions back to the original variables, and are calculated according to Equation 7.5.

7.2.6 Summary

The various single block and fusion methods implemented in this Chapter are summarised in, along with shorter notations, in Table 7.3. Independent components analysis, like canonical correlation analysis, cannot be applied to data that is rank deficient; PCA is used as the *de facto* data reduction method to provide uncorrelated variables for ICA. In the multi-block approaches employed here, cICA performs ICA on the PCA scores of the concatenated data (essentially this is cPCA-ICA), whilst in hICA the PCs from each data set are concatenated and subjected to ICA. ccICA has not been implemented, due to the the requirement to do PCA-ICA-CCA.

TABLE 7.3: The abbreviations used to refer to specific methods. The first three are the single-block methods, and are included so as to facilitate comparisons between the single- and equivalent multi-block models.

Method	Abbreviation
Principal components analysis	PCA
Partial least squares regression	PLS
Independent components analysis	ICA
Concatenated PCA	cPCA
Concatenated PLS	cPLS
Concatenated ICA	cICA
Hierarchical PCA (PCA-PCA)	hPCA
Hierarchical PLS (PLS-PLS)	hPLS
Mixed mode PCA (PCA-PLS)	mPCA
Mixed mode PLS (PLS-PCA)	mPLS
Hierarchical ICA (PCA-ICA)	hICA
PCA canonical correlation analysis	ccPCA
PLS canonical correlation analysis	ccPLS

7.2.7 Data

Whilst the concatenated and hierarchical methods are not limited to the number of data blocks, CCA is restricted to two. As such, the focus will be initially placed on the fusion of positive-mode HILIC LC-MS variables with those from HSQC NMR (Section 4.1).

7.2.8 Method effectiveness

In order to gauge a fusion technique's effectiveness, it needs to be compared to how well it approximates the study's variance. For applications looking to maximise the between-class variance, a metric of class separation would be most appropriate (e.g. the distance between class centroids). Due to the continuous nature of the classification variable (tenderness), the supervised and unsupervised techniques will be compared differently. The D^2 metric will be used to compare the fusion techniques that use unsupervised MVA in the first stage, and the Pearson correlation coefficient, ρ, for those where supervised methods are used as the initial data reduction stage:

D^2	PCA, ICA, cPCA, cICA, hPCA, mPLS, hICA, ccPCA
ρ	PLS, cPLS, hPLS, mPCA, ccPLS

For the unsupervised methods, the optimal number of components to include is defined by the subset that maximises the D^2 value. Whilst it is expected that the best fitting component in supervised methods is the first, the over-riding factor in these methods was on the minimisation of the predictive error. As such it may have been possible to achieve a higher ρ value but at the expense of (further) overfitting the analysis. Whilst the minimisation of the predictive error reduces the chance of overfitting, the cross validation involved does not eliminate the risk entirely.

7.3 Results and discussion

7.3.1 Classification power

One of the aims of data fusion is to enhance the discriminating power of a classification model by, for example, exhibiting better between-group discrimination. Such enhancements generally allow for greater confidence to be placed in the success of the model, and may facilitate interpretations of the variable loadings that ultimately underpin the trend.

Table 7.4 shows the D^2 values for the unsupervised techniques. Various low and high level scalings have been applied in order to assess their validity. The concatenated methods have only been treated with the low level standardisations, and autoscaling for these is the method that gives the best classification. Autoscaling produces the highest results, and values greater than 0.75 are shown in red. The high level approaches generally have higher D^2 values than the concatenated methods. The highest values arise from mPCA (PCA-PLS), which is partly due to the supervised nature of the higher level technique. The values for mPCA can be compared to those for hPCA; much of the difference in values is attributable to the effect of the PLS which tends to add at least 0.2 to the classification value. Where autoscaling has been applied as the high level standardisation method, the results are typically highest within each low level method. In few instances, most notably for low level autoscaled hICA, the effect of high level block scaling produces the highest D^2 values. Whilst the block scaling methods may prevent one analytical technique dominating, it is clear that a more appropriate high level method is autoscaling as it ensures each component contributes equally. Of the lower level methods, autoscaling also appears to be the most appropriate. The use of vector normalisation for ICA, as suggested in Scholz et al. [305], appears to impart no advantage.

The supervised techniques produce much higher classification values (here ρ) than the unsupervised methods (Table 7.5). The values for concatenated PLS exhibit a large range for the low level standardisation methods, which is in direct contrast to the doubly supervised hPLS which results in perfect matches for all combinations of scalings and standardisations. mPLS and ccPLS also produce high values. The values for cPLS, hPLS and mPLS are the same regardless of the scaling/standardisation methods applied. For ccPLS, the situation is slightly different as CCA mandatorily autoscales the data. The high similar values increase the possibility of overfitting, as perfect values for hPLS are achieved regardless of the scaling/standardisations applied. However, the values for mPLS show that the effect of the second PLS in mPLS is not drastic. In summary, the most appropriate scaling/standardisation method for this data involves two autoscalings, as this results in high values for all techniques.

TABLE 7.4: The D^2 values for the unsupervised fusion techniques. A maximum of 50 components from each dataset are considered. Values greater than 0.75 are coloured red. The concatenated methods are only scaled once according to the method indicated in the low level column.

Low level	High Level	cPCA	cICA	hPCA	mPCA	hICA	ccPCA
Centred	Centred	*0.18*	0.36	*0.18*	0.34	0.34	0.64
	BlockMean	*0.18*	0.36	*0.19*	0.42	0.36	0.64
	BlockStdDev	*0.18*	0.36	*0.23*	0.48	0.39	0.64
	Autoscale	*0.18*	0.36	0.64	**0.99**	0.35	0.64
Autoscale	Centred	0.59	0.66	0.59	**0.78**	0.61	0.67
	BlockMean	0.59	0.66	0.57	**0.77**	0.72	0.67
	BlockStdDev	0.59	0.66	0.57	**0.77**	0.72	0.67
	Autoscale	0.59	0.66	0.67	**0.99**	0.60	0.67
Vast	Centred	0.26	0.61	0.26	0.67	0.51	0.67
	BlockMean	0.26	0.61	0.25	0.64	0.56	0.67
	BlockStdDev	0.26	0.61	0.25	0.64	0.56	0.67
	Autoscale	0.26	0.61	0.67	**0.97**	0.55	0.67
Pareto	Centred	0.26	0.50	0.26	0.53	0.49	0.67
	BlockMean	0.26	0.50	0.55	0.66	0.53	0.67
	BlockStdDev	0.26	0.50	0.55	0.66	0.50	0.67
	Autoscale	0.26	0.50	0.67	**0.99**	0.52	0.67
Vector2	Centred	0.26	0.39	0.26	0.45	0.36	0.68
	BlockMean	0.26	0.39	0.56	0.64	0.47	0.68
	BlockStdDev	0.26	0.39	0.56	0.64	0.48	0.68
	Autoscale	0.26	0.39	0.68	**0.99**	0.42	0.68

TABLE 7.5: The ρ values for the supervised fusion techniques. The maximum number of components included from each dataset is 50. Values greater than 0.75 have been coloured red. cPLS has only had the low level standardisation applied.

Low level	High level	cPLS	hPLS	mPLS	ccPLS
Centred	Centred	0.37	**1.00**	**0.96**	**0.90**
	BlockMean	0.37	**1.00**	**0.96**	**0.90**
	BlockStdDev	0.37	**1.00**	**0.96**	**0.90**
	Autoscale	0.37	**1.00**	**0.96**	**0.90**
Autoscale	Centred	**0.90**	**1.00**	**1.00**	**1.00**
	BlockMean	**0.90**	**1.00**	**1.00**	**1.00**
	BlockStdDev	**0.90**	**1.00**	**1.00**	**1.00**
	Autoscale	**0.90**	**1.00**	**1.00**	**1.00**
Vast	Centred	**0.83**	**1.00**	**1.00**	**0.99**
	BlockMean	**0.83**	**1.00**	**1.00**	**0.99**
	BlockStdDev	**0.83**	**1.00**	**1.00**	**0.99**
	Autoscale	**0.83**	**1.00**	**1.00**	**0.99**
Pareto	Centred	0.72	**1.00**	**1.00**	**0.99**
	BlockMean	0.72	**1.00**	**1.00**	**0.99**
	BlockStdDev	0.72	**1.00**	**1.00**	**0.99**
	Autoscale	0.72	**1.00**	**1.00**	**0.99**
Vector2	Centred	0.58	**1.00**	**0.89**	**0.75**
	BlockMean	0.58	**1.00**	**0.89**	**0.75**
	BlockStdDev	0.58	**1.00**	**0.89**	**0.75**
	Autoscale	0.58	**1.00**	**0.89**	**0.75**

7.3.2 Maximum number of components

The two stage methods have been allowed to include up to 50 components (from each dataset) in the second stage. In attempting to reduce the predictive error, mPCA (PCA-PLS), for example, uses many components which on their own are not obviously linked to the tenderness values. Table 7.6a shows the D^2 or ρ results of the various fusion techniques. Where appropriate, the number of PCs selected by the methods are shown ('Picked'), and also the predictive error as determined when PLS is the final technique. For example, in concatenated ICA (cICA), three components from each PCA were concatenated prior to the ICA, and for hPCA this value was 35. The difference between hPCA and mPCA is the second MVA, which is PLS in the latter. The difference in the number of picked components is not large, but Table 7.7 demonstrates that mPCA utilises these additional components in order to reduce its predictive error.

TABLE 7.6: The results when the maximum number of components from each dataset is changed. The data were autoscaled.

(A) A maximum of 50 components from each dataset were included.

Technique	D^2/ρ	Picked	Error
cPCA	0.59		
cPLS	0.90		0.0240
cICA	0.66	3	
hPCA	0.67	35	
hPLS	1.00		0.0002
mPCA	0.99	48	0.0258
mPLS	1.00		
hICA	0.60	9	
ccPCA	0.67	2	
ccPLS	1.00		

(B) The component limit is reduced to 5, without drastically altering the results.

Technique	D^2/ρ	Picked	Error
cPCA	0.59		
cPLS	0.90		0.0128
cICA	0.66	3	
hPCA	0.67	2	
hPLS	1.00		0.0000
mPCA	0.87	5	0.1071
mPLS	0.96		
hICA	0.59	2	
ccPCA	0.67	2	
ccPLS	0.91		

Table 7.7 shows the effect of limiting the number of components that may be used from PCA. The results for hPCA remain unchanged, in spite of the component numbers being severely limited. It can be seen that the predictive error in mPCA decreases as the number of permitted components increases. This is commensurate with an increase in its ρ value. For hICA, the effect of including additional components is negligible. With regards to these various results, it would appear that 50 components is unnecessarily high. A more appropriate limit of 5 would suffice to enhance the parsimony without inhibiting the results. Table 7.6b shows the results obtained when the maximum number of components has been limited to 5 (from each of the first round analyses).

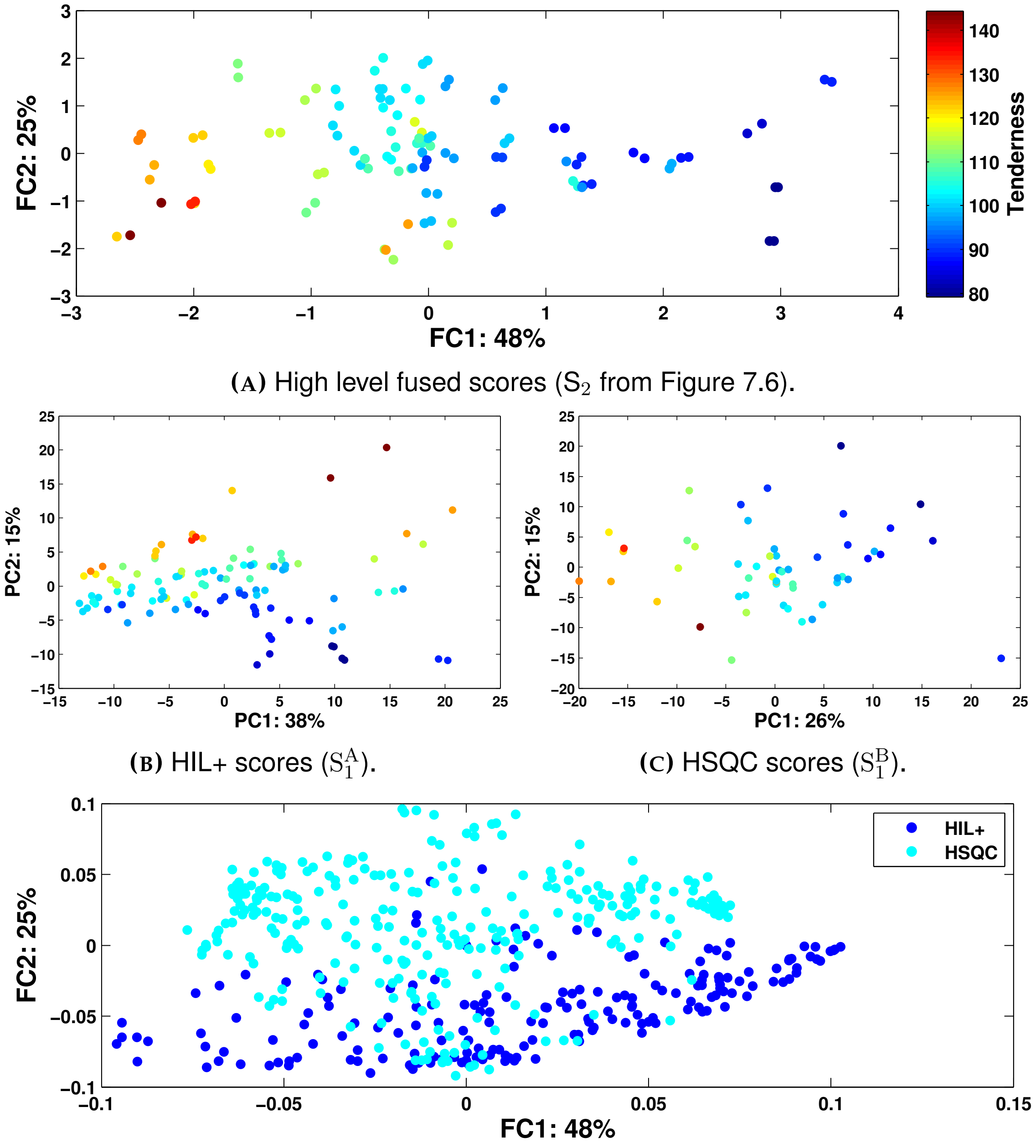
TABLE 7.7: The effect of the number of components and the D^2 and ρ values. The maximum number of permitted components allowed in high level fusion is given in each column. For hPCA and hICA the changes are marginal, but mPCA uses these additional components to reduce the error and increase ρ.

	2	5	10	25	50
hPCA D^2	0.67	0.67	0.67	0.67	0.67
mPCA ρ	0.77	0.87	0.90	0.95	0.99
mPCA Error	0.16	0.11	0.09	0.07	0.03
hICA D^2	0.59	0.59	0.60	0.60	0.60

TABLE 7.8: The results of the single block methods.

Data	PCA D^2	ICA D^2	PLS ρ
HILIC Positive	0.41	0.48	0.84
HSQC	0.57	0.57	0.88

Table 7.8 shows the D^2 and ρ values for the single block MVA techniques for the HSQC and MS datasets. A comparison of these single block results and the multiblock results in Table 7.6b reveals that data fusion can be used to enhance the discriminatory power for multiple datasets. This is demonstrated in Figure 7.8 which shows the results from hPCA. The fused scores are presented in Figure 7.8a which, on comparison to the single block scores in Figures 7.8b and c, show a marked improvement in their relationship to the tenderness values. The visual improvement in the scores is quantified by the D^2 metric in Tables 7.7 and 7.8. The range in such values for the various fusion techniques suggests that certain methods may be more appropriate than others. However, the variables that contribute to such trends (hPCA loadings are shown in Figure 7.8d) must also be considered, and the next section details the most influential variables.



(A) High level fused scores (S_2 from Figure 7.6).

(B) HIL+ scores (S_1^A).

(C) HSQC scores (S_1^B).

(D) Individual variable loadings (L_2).

FIGURE 7.8: Scores and loadings plots from hPCA using positive mode HILIC MS and HSQC data sets. Two PCs from each low level PCA have been used in the high level model. The high level scores are shown in (a), whilst in (b) and (c), respectively, the low level scores are shown for the MS and NMR data. After correction, the loadings for each block are shown in (d).

7.3.3 Most frequently selected variables

In order to demonstrate which variables are being selected by which technique (fusion or single block), the top 5 variables, i.e. those with the largest absolute loadings, from each of the methods are presented in four tables below, according to the data set and the nature of the technique. The loadings are taken from the single component that best correlates with tenderness.

Tables 7.9 and 7.11 contain the unsupervised and supervised rankings for the MS variables, and Tables 7.10 and 7.12 display those for the HSQC variables. Variable indices are displayed in the first column, and their relative rank is given for each technique, with 1 being the largest loading value. The final column of each table shows the correlation coefficient between the variable and the tenderness values, and is included as a guide to those variables which are strongly correlated to the variance of interest.

In Table 7.9 the top variables in cPCA and hICA are poorly correlated to tenderness, and not selected by other techniques. The similarity of PCA and ICA is expected, as ICA uses the principal components. The top variables for hPCA and ccPCA are similar. For the HSQC variables in Table 7.10, the top cPCA variables are also poorly correlated. hPCA and ccPCA exhibit the same similarity as with the MS variables.

The range of variables selected by the unsupervised methods is relatively large, especially in comparison to the quantity that feature in the top five in the supervised methods. Only six HSQC different variables are selected by the supervised methods (Table 7.12), whilst for MS this number is 10 (Table 7.11). The variables picked by the supervised fusion methods are the same; there is no appreciable difference between the variables picked in the complicated high level fusion methods in comparison to the more simplistic concatenated PLS approach. There are, however, differences between the single and concatenated method, but it is clear that no complementary information is gleaned as a result of performing multiblock PLS methods.

TABLE 7.9: The ranks of the top positive mode HILIC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values.

#	PCA	ICA	cPCA	cICA	hPCA	mPCA	hICA	ccPCA	ρ
54	**3**	**4**	151	**2**	10	**3**	144	10	**0.77**
139	6	9	180	**1**	11	7	157	11	**0.77**
133	**2**	**2**	133	**4**	7	**2**	129	7	**0.74**
81	**1**	**1**	115	17	16	**1**	78	16	**0.73**
189	9	10	53	41	34	**4**	97	34	**0.68**
14	16	42	51	6	**1**	15	150	**1**	**-0.67**
58	19	47	48	13	**3**	6	148	**3**	**-0.65**
109	**5**	**5**	162	10	21	12	126	21	**0.65**
169	**4**	6	116	20	15	8	139	15	**0.64**
95	11	11	73	49	49	**5**	77	49	**0.62**
89	21	51	136	**3**	**4**	18	147	**4**	**-0.62**
68	27	71	159	**5**	**5**	16	130	**5**	**-0.62**
147	7	**3**	9	66	36	11	55	36	**0.57**
87	18	45	177	8	**2**	19	149	**2**	**-0.55**
25	33	20	**4**	155	107	62	35	107	**0.30**
85	58	74	**1**	119	74	80	175	74	**0.29**
69	78	82	**5**	58	103	110	167	103	**0.29**
188	80	89	**2**	141	93	115	176	93	**0.21**
47	50	28	**3**	181	137	100	15	137	**0.19**
48	81	38	66	153	181	95	**4**	181	**0.16**
80	106	55	13	123	155	129	**3**	155	**0.12**
39	119	64	81	120	143	134	**5**	143	**0.07**
96	115	59	114	137	148	148	**2**	148	**0.07**
45	129	75	129	125	131	166	**1**	131	**0.03**

TABLE 7.10: The ranks of the top HSQC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values.

#	PCA	ICA	cPCA	cICA	hPCA	mPCA	hICA	ccPCA	ρ
43	6	37	220	**3**	8	15	14	8	**0.79**
169	**4**	48	215	**2**	9	16	11	9	**0.79**
45	15	54	227	**4**	16	11	20	16	**0.79**
89	**2**	33	209	**1**	**4**	22	**5**	**4**	**0.78**
49	9	30	229	**5**	**5**	17	22	**5**	**0.78**
66	49	92	254	40	74	**3**	66	74	**0.75**
124	**5**	60	170	11	15	39	6	15	**0.72**
84	**3**	67	158	9	24	32	**4**	24	**0.72**
29	**1**	82	154	6	29	33	**1**	29	**0.72**
44	28	9	222	28	**3**	36	54	**3**	**-0.71**
72	83	11	19	80	45	**1**	177	45	**-0.70**
16	30	**1**	189	31	**1**	31	80	**1**	**-0.70**
50	26	13	248	25	**2**	52	49	**2**	**-0.69**
207	71	7	31	65	37	**2**	173	37	**-0.69**
67	89	6	18	84	41	**4**	183	41	**-0.69**
160	63	21	98	53	38	**5**	145	38	**-0.68**
211	96	**4**	33	76	42	9	192	42	**-0.68**
61	100	**5**	26	85	46	8	206	46	**-0.68**
219	12	89	141	27	43	66	**3**	43	**0.67**
256	19	93	139	22	53	58	**2**	53	**0.67**
75	111	**2**	41	101	34	72	223	34	**-0.60**
86	79	**3**	55	88	36	75	180	36	**-0.53**
77	196	70	**2**	183	115	101	269	115	**-0.46**
242	225	90	**3**	214	134	90	246	134	**-0.44**
224	250	107	**5**	236	154	102	227	154	**-0.39**
215	232	135	**4**	274	201	173	103	201	**-0.22**
91	230	132	**1**	268	198	186	100	198	**-0.18**

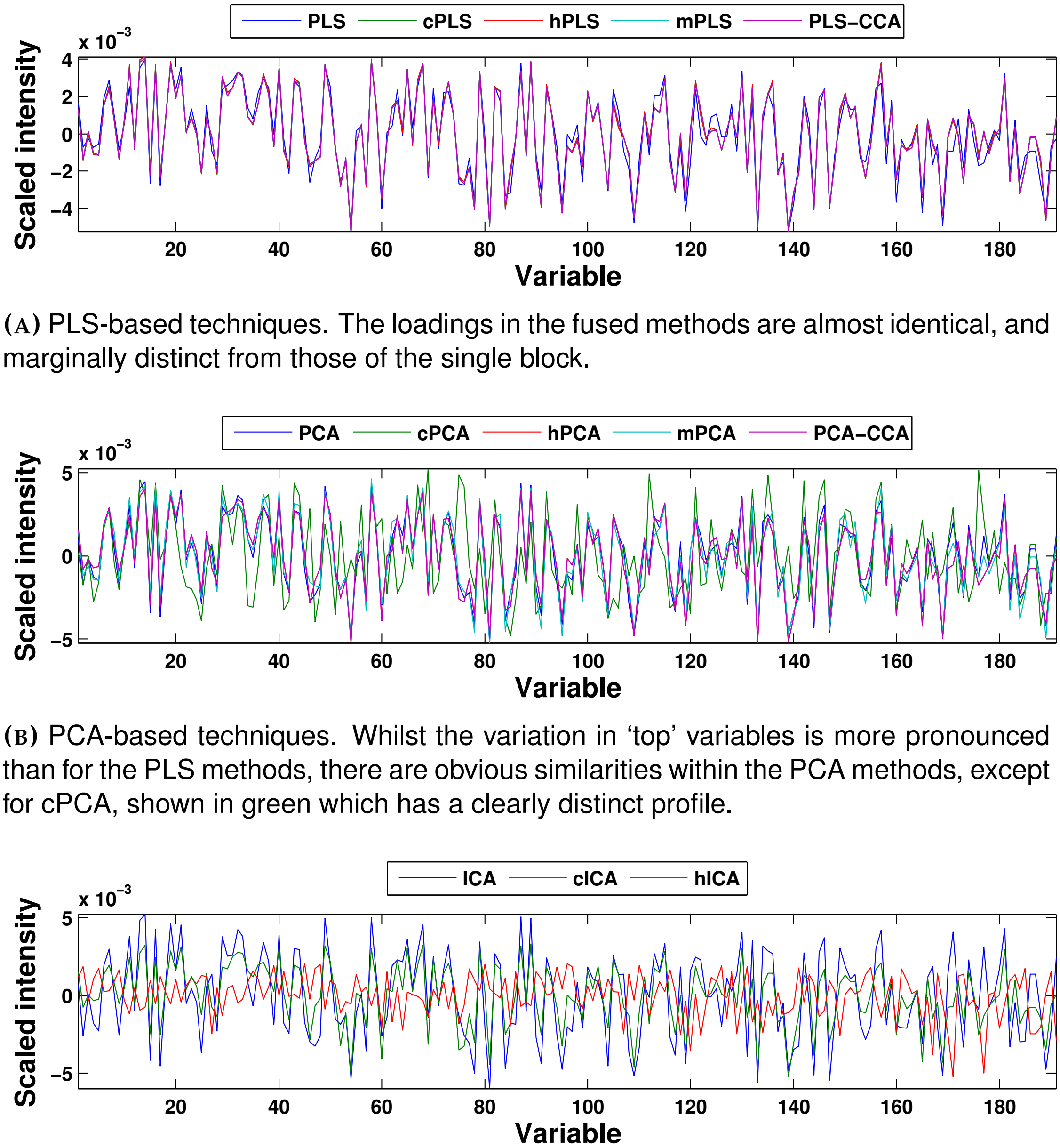
TABLE 7.11: The ranks of the top positive mode HILIC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values.

#	PLS	cPLS	hPLS	mPLS	ccPLS	ρ
54	16	**1**	**1**	**1**	**1**	**0.77**
139	13	**2**	**2**	**2**	**2**	**0.77**
133	15	**3**	**3**	**3**	**3**	**0.74**
81	23	**4**	**4**	**4**	**4**	**0.73**
189	45	**5**	**5**	**5**	**5**	**0.68**
14	**1**	6	6	8	8	**-0.67**
58	**2**	8	8	9	9	**-0.65**
89	**3**	14	14	11	11	**-0.62**
68	**5**	15	15	17	17	**-0.62**
87	**4**	27	27	26	26	**-0.55**

TABLE 7.12: The ranks of the top HSQC variables from each technique are. The final column shows the correlation coefficients between the variables and the tenderness values.

#	PLS	cPLS	hPLS	mPLS	ccPLS	ρ
43	**2**	**2**	**2**	**3**	**3**	**0.79**
169	**4**	**4**	**4**	**4**	**4**	**0.79**
45	**5**	**3**	**3**	**1**	**1**	**0.79**
82	6	**1**	**1**	**2**	**2**	**0.79**
89	**1**	6	6	6	6	**0.78**
49	**3**	**5**	**5**	**5**	**5**	**0.78**

The scaled variables loadings are presented in Figures 7.9 and 7.10, and help to illustrate the similarities and differences in loading magnitudes for MS and NMR variables. The loadings for all fused PLS methods are the same, with only minor differences to the equivalent single block PLS loadings. The similarities amongst the PCA and ICA loadings are more obvious than are detailed by the tables of variable rankings. cPCA is perhaps the most distinct of all methods, as it differs greatly from all other PCA methods. In spite of the enhanced classification of the data fusion methods, there are no significant changes in variable loadings in comparison to the single block methods. As is seen in Figures 7.9 and 7.10 there are many variables with equivalently large loadings; trying to associate variables across the multiple techniques is therefore hindered as there are too many potential candidates to consider.

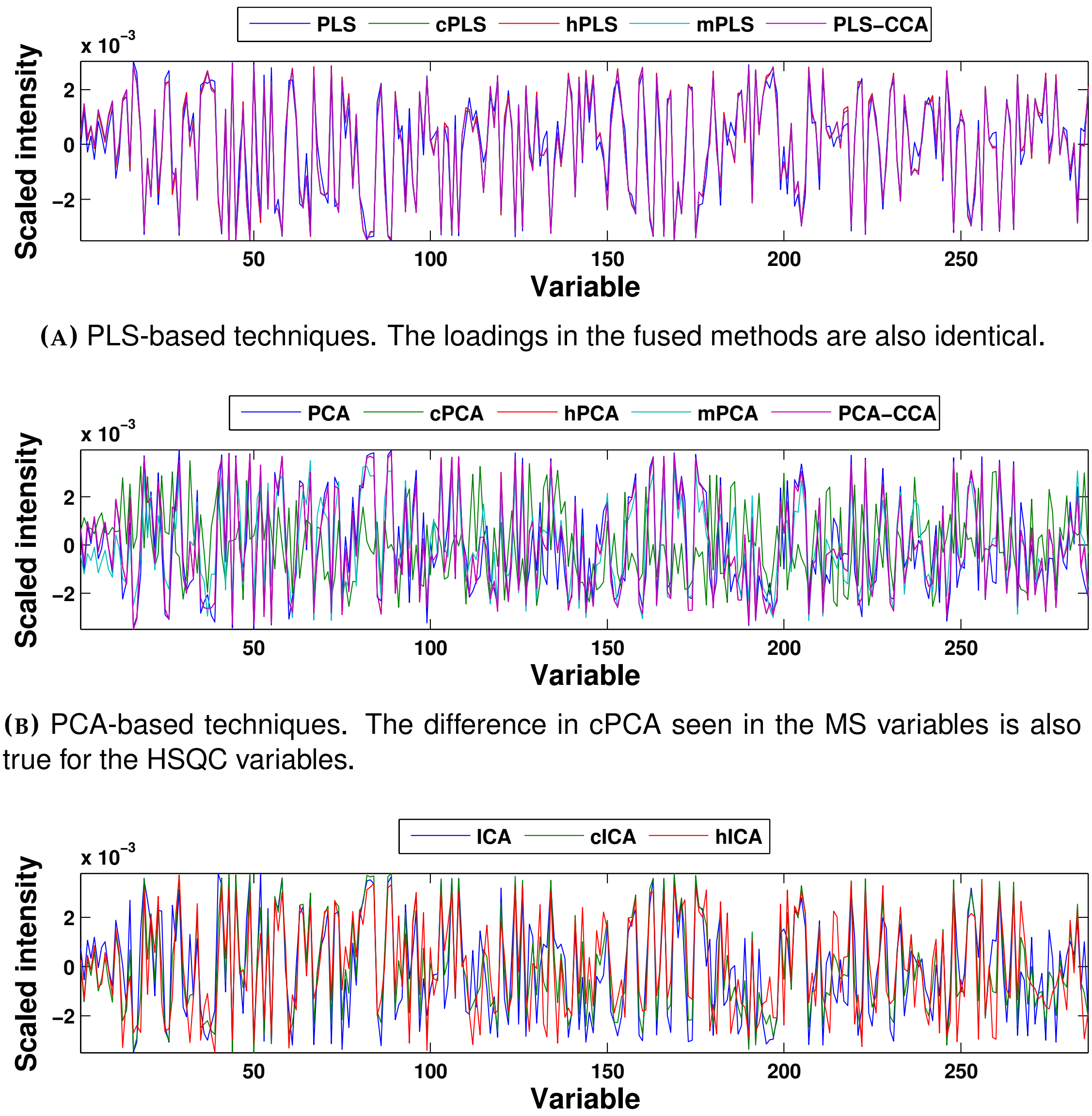


(A) PLS-based techniques. The loadings in the fused methods are almost identical, and marginally distinct from those of the single block.

(B) PCA-based techniques. Whilst the variation in 'top' variables is more pronounced than for the PLS methods, there are obvious similarities within the PCA methods, except for cPCA, shown in green which has a clearly distinct profile.

(C) ICA-based techniques. The single block and concatenated variants show similar trends to those seen the the PCA-based methods. The exception is for hICA.

FIGURE 7.9: The loadings for the positive-mode HILIC LC-MS variables, which have been grouped according to the main multivariate technique.



(A) PLS-based techniques. The loadings in the fused methods are also identical.

(B) PCA-based techniques. The difference in cPCA seen in the MS variables is also true for the HSQC variables.

(C) ICA-based techniques. Again, these broadly follow the trends seen for the PCA-based methods.

FIGURE 7.10: The loadings for the HSQC variables, which have been grouped according to the main multivariate technique.

7.4 Conclusion

The correct processing of spectral data is crucial to any multivariate analysis. Variable transformation methods, such as those suggested by Tukey [343], can be applied to data in order to reduce any deviations that might upset statistical techniques, especially those that rely on normally distributed data. The use of log-scaling of variables has been encountered in the literature [341, 344], but this approach should be considered in terms of any deleterious effects that may be imparted as a consequence of blanket transformations. An optimised approach has been implemented, which applied one of various functions that best reduced the skewness of variables. However, subsequent statistical analysis showed little improvement, and the method was not applied.

Many of the fusion methods demonstrated an enhanced agreement with the tenderness values, in comparison to the single block methods. The combination of datasets allowed for the variables to work together and improve the relationship between the multivariate scores and the tenderness values. However, as shown in Figure 7.9 and 7.10 the loadings profiles across the various techniques are highly influenced by the loadings from the lower level multivariate analysis. Whilst this is to be expected considering the nature of the techniques, these fusion methods do little to maximise the datasets' complementarity. The variable loading profiles show variables with approximately the same loading magnitude. As such, there is little interaction between the datasets which may have assisted with feature assignment by identifying potentially related variables.

Whilst the fusion methods implemented in this Chapter have been successfully applied in other situations, none of the low and high level fusion methods have proved successful here. In the chemical literature, no implementations of intermediate level fusion have been encountered. This methodology uses feature selection to reduce large quantities of variables from multiple data sets, and these key variables can then be analysed using single block multivariate methods. Whilst stepwise feature selection is not appropriate for metabolomics and non-targeted data, evolutionary algorithms, which are discussed in the next Chapter, are being used increasingly for feature selection and classification. Canonical correlation analysis is a multivariate and, crucially, multiblock technique, and using subsets of variables circumvents the rank deficiency issue, and offers the prospect of showing how variables across multiple platforms are related.

Chapter 8

Feature Selection with Genetic Programming

In the process of natural selection, then, any device that can insert a higher proportion of certain genes into subsequent generations will come to characterise the species.

Edward O. Wilson

8.1 Introduction

As was demonstrated in the previous Chapter, the use of various multi-block methods can assist with feature identification. The drawback to such techniques is that the use of multiple analyses involves a complex process of associating trends back to the original variables. Furthermore, the number of components to include is also important, and relies on users making judgements on the value of including each additional principal component into the second round of multivariate analysis.

Whilst stepwise feature selection (FS) methods can be used to produce a subset of variables that are capable of adequately describing the original data set, they are not particularly appropriate in metabolomics studies [345, 346] which typically contain limited observations and many variables. Thus the size of the search space and the numerous calculations entailed in stepwise FS is likely to render overfit solutions, which at their most extreme are liable to include only one variable that is correlated to the variance of interest. Feature selection can reduce, therefore, the useful information in a data set by too much, and limit the use of many biologically interesting variables in multivariate analyses. By including just a few variables, the method in essence ensures that any results are almost univariate in nature. Fusion strategies that involve feature selection are typically referred to as mid-level strategies.

Ramos and Ruisánchez [24] used variable selection from two complementary spectroscopic techniques (Raman spectroscopy and x-ray fluorescence) to enhance the classification of pigments encountered in works of art. The mid-level fusion method involved the use of PLS-DA to select discriminatory spectral regions which were then concatenated to give so-called 'metaspectra'. This block was then subjected to another PLS-DA analysis, with the class of test-set data being estimated. In terms of classification, the fused system was found to outperform the models achieved with the individual spectra. The fusion of NMR spectral variables of two biofluids has been performed by Smolinska et al. [347]. They applied a support vector machine recursive feature elimination method to select discriminatory variables from each of the biofluids which are then concatenated into metaspectra.

Alternative variable selection procedures include evolutionary algorithms [25], which are explained in more detail in Section 2.6.2.4 (p. 85). Evolutionary approaches uses the 'survival of the fittest' principle to select the better solutions to a problem, which is derived, from a random starting position, through a series of mutation and crossover operations designed to mimic the natural process of evolution. In genetic programming (GP), the solutions are combinations of small variable subsets which are related together by a series of mathematical and/or logical operations (Figure 2.12a). These can be depicted as a tree, as is demonstrated in Figure 2.12b, and examples of mutation and crossover are given, respectively, in Figures 2.12c and 2.12d.

A single run of a GP starts from a random position with a subset of variables, and subsequent generations are formed through evolutionary processes. Each generation is assessed against a defined fitness function, which for classification purposes is typically how well the programme predicts the training data. When the optimal solution has been found, or the maximum number of generations has been reached, the analysis is halted. The variables included in the most apposite solution each receive a vote. Many independent GP runs are performed, and the votes received by each variable can be taken as an indication of their importance (c.f. the loadings from multivariate techniques).

Evolutionary approaches are becoming more common for chemical data [26, 192, 348–351] and are generally superior to stepwise FS methods. As the extremely large search space produced by such spectra is often inhibitive to optimal classification, Taylor et al. [26] reduced the number of input variables by including only those with a high characteristicity. This is similar to Fisher's F ratio, and selects only the most informative variables to include in the GP. A two-stage GP was presented by Davis et al. [27] to reduce the high dimensionality of ^{1}H NMR datasets. The first stage identified the most discriminatory variables, which were then used in the second stage to enhance the classification rate. The two-stage approach produced superior results to the stand-alone GP. Donarski et al. [8] used PLS in the first stage to reduce the number of ^{1}H NMR spectral variables prior to using GP for the final classification of samples.

The main benefit of using GP is that the results are expressed explicitly in terms of original variables (rather than linear combinations). They are, therefore, easy to interpret and the method has the capability for both feature selection and classification. A novel use of GP is presented here, specifically in relation to its use with canonical correlation analysis (CCA, Section 7.2.5). CCA maximises the correlation between linear combinations of two data matrices, but is not directly applicable to data that can be described as rank deficient. Metabolomics datasets, such as from NMR or LC-MS, contain many variables, and the inter-correlated nature of the variables is generally sufficient to render them rank deficient. This problem was overcome in Chapter 7 by using principal components analysis (PCA) and partial least squares (PLS) regression for data reduction prior to CCA. The scores from PCA and PLS are, by definition, uncorrelated but the approach means that the results from CCA are not directly interpretable in terms of original variables. Using GP as an FS method provides the opportunity to use CCA with subsets of original variables, such that the results are expressed in terms of the original variables.

8.2 Experimental

8.2.1 Data

The data used in this section are those described in Section 4.1. After assessment of the scaling and standardisation methods, all data is autoscaled. The previous Chapter focused on the HSQC and HILIC LC-MS data, and for purposes of comparison the GP-CCA approach is also focused on these datasets.

8.2.2 GP parameters

The genetic program implemented is a two-stage approach based on that developed by Davis et al. [27] for the classification of discrete classes. Here a modification has been implemented to account for the continuous nature of the classification variable. The first stage runs the GP with all of the spectral variables, and only those that are frequently selected are used in the second stage.

GP is not an optimisation routine and as such its results can be considered as the 'local' best rather than the 'global' best. Therefore multiple iterations of GP are required in order to ensure that as much of the search space is covered. Using only select variables in the second stage has the effect of reducing the search space, such that a more global solution can be found.

An individual solution in GP, such as that shown in Figure 8.1 is a simple program comprising of mathematical operators, variables and constants. As analysing all possible trees is impossible, each iteration of GP starts with a random population of trees. These individuals are then assessed, using an independent training dataset, in terms of their fitness which in this case is how well the solutions estimate the tenderness values. These individuals are then ranked, and the ranking affects the chances of an individual solution undergoing various genetic operations; this is known as the 'survival of the fittest' principle. The fitness function is calculated as the sum (over all observations) of the absolute difference between predicted and expected tenderness values.

Each program is assigned a weight depending on its fitness; this weight determines the probability of it being selected for reproduction. Weights are calculated as 1 ÷ (1 + fitness) such that less fit solutions (i.e. higher fitness values) are penalised more. After the first generation has been assessed, a new generation is created from it. In order to prevent the loss of the best solutions, the top $x\%$ can be automatically replicated from one generation to the next; equally, the worst $y\%$ can be discarded. The mutation operation randomly selects a node in a solution and replaces it with another randomly chosen node. Should the number of arguments for the new node differ to the old node, then this is accounted for in the mutation process.

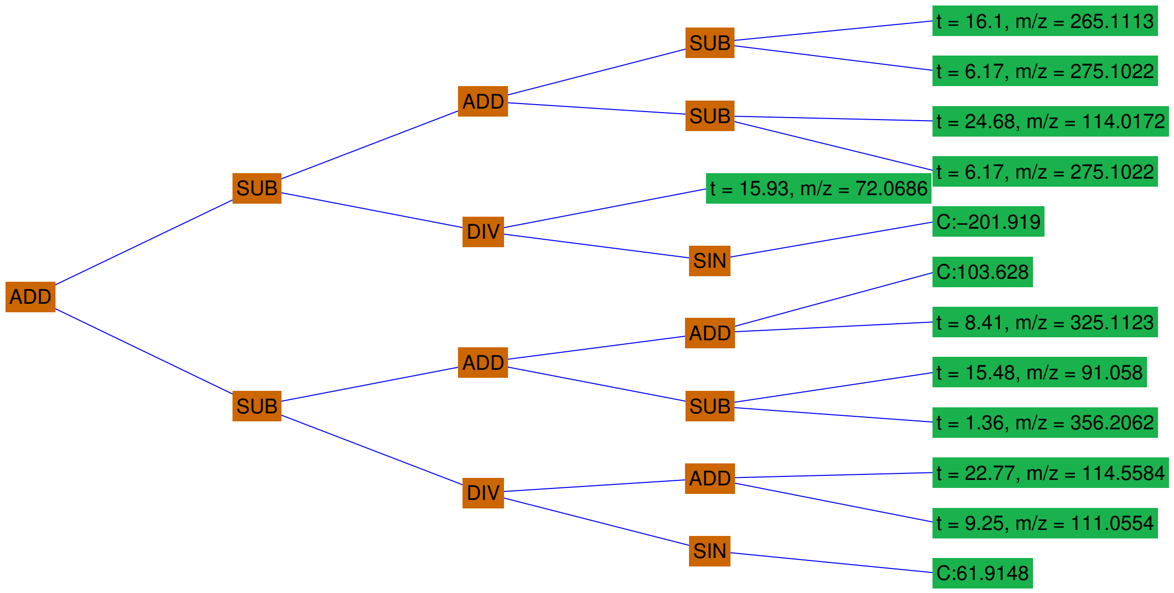


FIGURE 8.1: A typical GP tree. The variables (V) and constants (C) are shown by the green (leaf) nodes, whilst the operations are given in the brown (branch) nodes. The topmost branch node shows the subtraction of two MS variables, with other nodes showing addition (ADD), division (DIV), and sine (SIN).

Alternatively, the crossover operation can be applied to two nodes, where random subsections from each is swapped; in genetic terms this creates two children with characteristics of both parents.

In order to guard against the possibility of becoming trapped in a local minimum, island or demetic populations can be used [193]. Each island contains its own solutions, which evolve independently of those in the other islands. After a certain number of generations, some of the best solutions are migrated into neighbouring islands. The islands allow for various different starting positions to be probed, whilst also occasionally adding genetic diversity into neighbouring populations.

There is a propensity for trees to bloat over time, as more nodes are included. Larger trees reduce their interpretability, slow down the calculations and lead to the possibility of overfitting the data. The effect of bloating can be limited by restricting the tree to a specific depth.

Each individual GP iteration starts with a random selection of solutions, of which the most fit survive into subsequent generations. The second GP stage is triggered only when a pre-specified fitness is achieved in the first stage. All variables present in the top $z\%$ of solutions are carried forward and used in the second stage, which begins with a random population of starting solutions.

The evolutionary process in stage two continues until a defined fitness is achieved, or the number of generations has reached the stop value. The variables that are found in the best solution score a value of one, which is added to their score over all iterations of the GP. Thus, variables with high frequencies are those that appear in a large proportion of the fittest solutions. All relevant parameters are detailed in Table 8.1.

TABLE 8.1: GP parameters employed in the first and second stages.

Parameter		Stage 1 & 2
Islands		5
Migration interval	generations	50
Starting population	trees	200
Discard	trees	20
Replication	trees	20
Crossover	trees	150
Mutation	trees	30
Tree depth	maximum	5
Stop	generations	700
	mean fitness	1
Operations	mathematical	ADD SUB MUL DIV MAX MIN AVG ABS SIN COS TAN SQRT SQR INV
	logical	GT LT EQ NEQ

8.3 Results

8.3.1 GP selection frequencies

The GP selection frequencies for the positive-mode HILIC MS variables are shown in Figure 8.2, and those from HSQC are shown in Figure 8.3. These plots show the frequencies of variable selection in the second stage of the GP. The selection of MS variables appears to be quite uniform in terms of the range of *m/z* and time values that are selected. This can be contrasted with much more localised selection observed for the HSQC variables, where most of the selected variables occupy the traditional sugar and aliphatic regions with few aromatic shifts being selected.

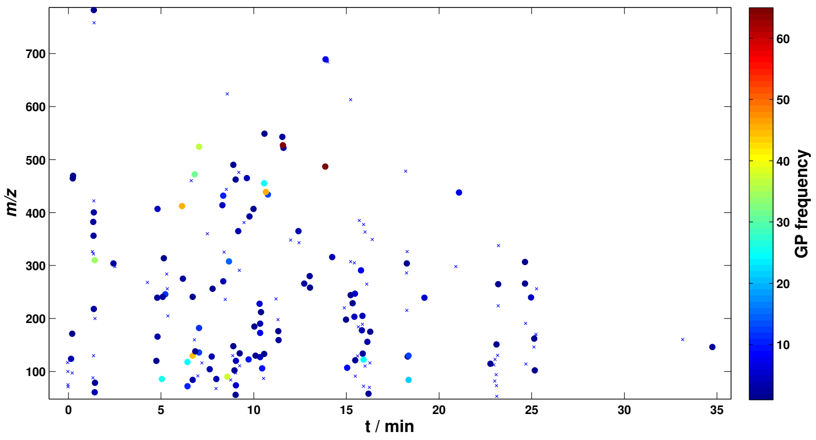


FIGURE 8.2: The GP selection frequencies shown for each positive-mode HILIC LC-MS variable. Variables that have not been selected in any of the solutions are marked with × symbols.

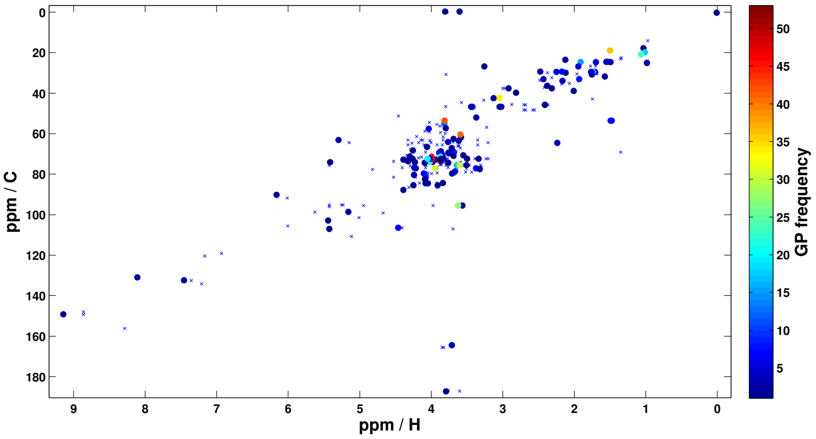


FIGURE 8.3: The GP selection frequencies for the HSQC variables are shown, and those that are not selected are marked denoted by ×.

8.3.2 Selecting the number of variables

The variables with the highest second stage frequencies were used to define the subsets on which the CCA was performed. The optimal number of variables was determined by the successive addition of variables to each subset. Figure 8.4 shows the correlations achieved by CCA as additional variables were included in the two data matrices.

The agreement between the tenderness values and the canonical scores are calculated using the D^2 metric. The canonical pair with the highest D^2 value is used, and the canonical correlation for this CP is shown in Figure 8.4. It can be seen that the successive addition of variables raises the canonical correlation, but with too many variables this is often at the expense of the D^2 value. The HSQC data becomes rank deficient with the addition of the 54th variable, and the analysis is halted.

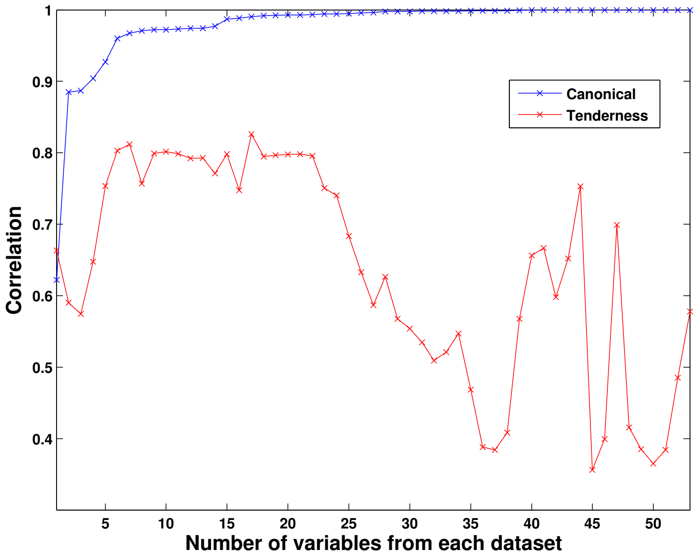


FIGURE 8.4: Correlation coefficients derived from CCA. The number of variables in each data set is successively increased until the data becomes rank deficient. The canonical correlation increases with each new variable pair, but the correlation of the scores to tenderness plateaus after the inclusion of seven variables from each data set. WIth more than 22 variable pairs, the tenderness correlation begins to fall.

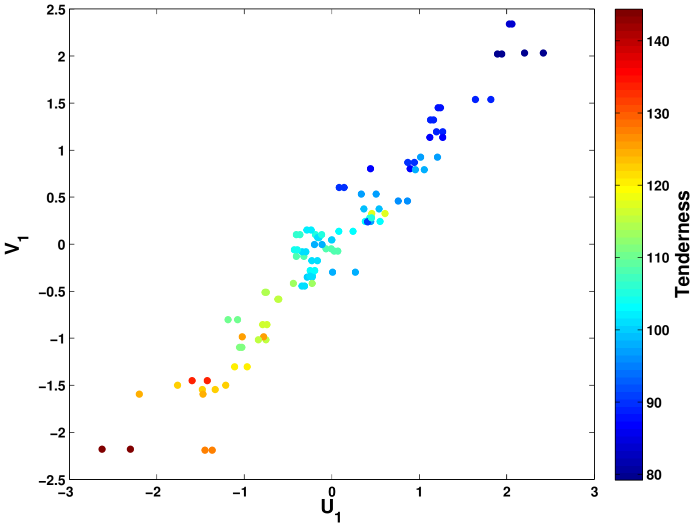


FIGURE 8.5: Scatter plot of the first canonical pair, formed when seven variables from each dataset are used.

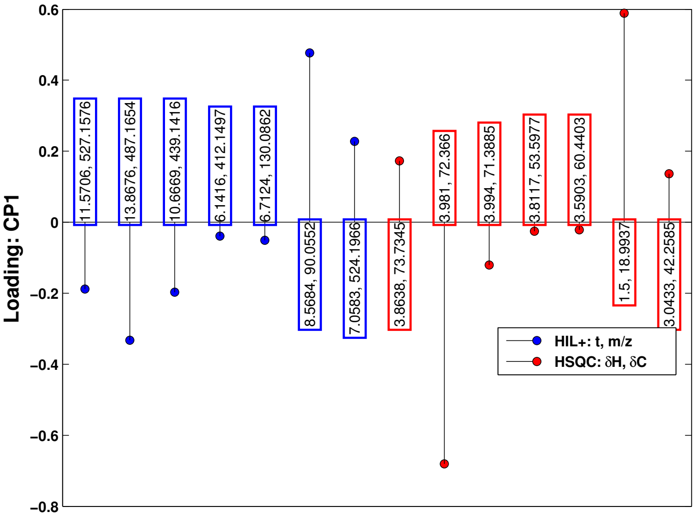


FIGURE 8.6: The loadings for the seven MS and seven NMR variables used in CCA. Only two variables have strong positive contributions in CP1. The most frequently selected variables by GP do not contrirbute most strongly to the trend seen in Figure 8.5.

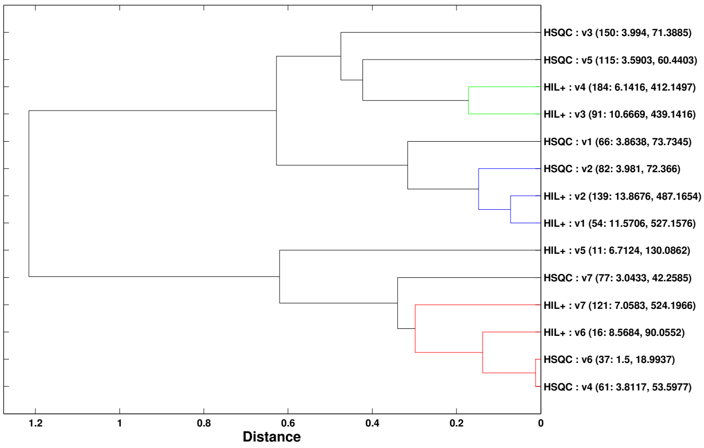


FIGURE 8.7: Dendrogram of the fourteen variables, where the distance is measured according to one minus the correlation coefficient. It can be seen that the variables with large similar loadings in Figure 8.6 are clustered together.

Figure 8.4 demonstrates that the canonical correlations from CCA increase with the addition of more variables. However, these variables do not enhance the agreement with the variance of interest (i.e. tenderness). If, instead, the number of variables is determined by the tenderness correlation, then 17 variables from each dataset would be chosen. However, as the improvement from seven variables is negligible, this might suggest that the additional 10 variables from each data set contribute little of value to the model. The results obtained using seven variables from each dataset are presented below.

8.3.3 Seven variables

The scatter plot of the first CP (U_1 and V_1) is shown in Figure 8.5, with the data points coloured according to tenderness. As well as the expected strong correlation between U_1 and V_1, the strong correlation with tenderness is obvious. The loadings for each of the fourteen variables in CP1 are shown in Figure 8.6. There are two large positive contributions, corresponding to one from each of the datasets. Of the negative loadings, there is one of a large magnitude, and three more with moderate values. The correlation dendrogram for these twelve variables is shown in Figure 8.7, with the distance between variables being calculated as one minus their correlation coefficient. It demonstrates that, as would be expected, correlated variables contribute to the same trends in CCA.

The value of this GP-CCA approach in assignment is neatly demonstrated by the grouping of two variables. MS variable number 16 (t = 8.6 min, *m/z* = 90.0552) and the 37th HSQC variable (δH = 1.50 ppm, δC = 18.99 ppm), have strong positive loadings (Figure 8.6) and are positively correlated (Figure 8.7). Both of these variables may be attributable to the amino acid, alanine. A more detailed assignment of variables is, however, deferred until Chapter 9.

The top seven variables from each technique are shown in Figure 8.8, where the GP selection frequency has been plotted as a function of the correlation to tenderness. The spread of variables broadly resembles that of a bathtub distribution, with highly correlated variables being more frequently selected. Although more of the highly correlated variables are selected by PLSR, for example, these additional variables are easily identified subsequently by virtue of their inherent inter-correlation.

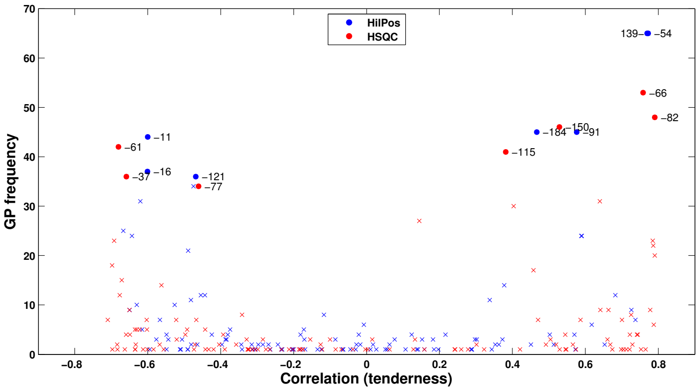
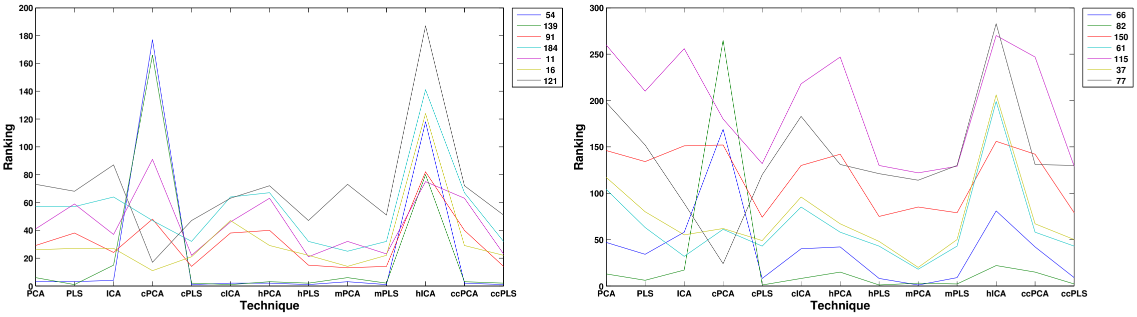


FIGURE 8.8: The frequency of selection of MS and NMR variables plotted as a function of the correlation to tenderness. The top seven variables from each technique have been labeled and shown as circles.

8.3.4 Comparison to previous fusion techniques

Chapter 7 described various fusion techniques, which were implemented with the positive-mode HILIC MS variables and those from HSQC. The top variables from each of the techniques were presented according to their relative magnitudes (ranks). For the seven variables identified by GP in each of the datasets, their rankings in each of the various fusion methods are shown in Figure 8.9. Few of the GP-selected variables consistently have high ranks, where a value of 1 indicates the highest rank and largest absolute loading. Variables 82 (HSQC), 139 and 54 (both MS) do appear with consistently high ranks with the exception of cPCA and hICA. The 16th MS variable and the 37th NMR variable, putatively assigned to alanine, do not feature amongst the top variables of any technique. NMR variable 115 is frequently selected by GP, yet is ranked poorly by all other fusion techniques. Whilst its selection by GP may appear anomalous, its approximately zero loading in CP1 indicates that GP-CCA is capable of withstanding apparently uninformative variables. It is clear that GP-CCA is a useful technique which is capable of identifying variables that may be overlooked by other approaches.



(A) Positive-mode HILIC MS.

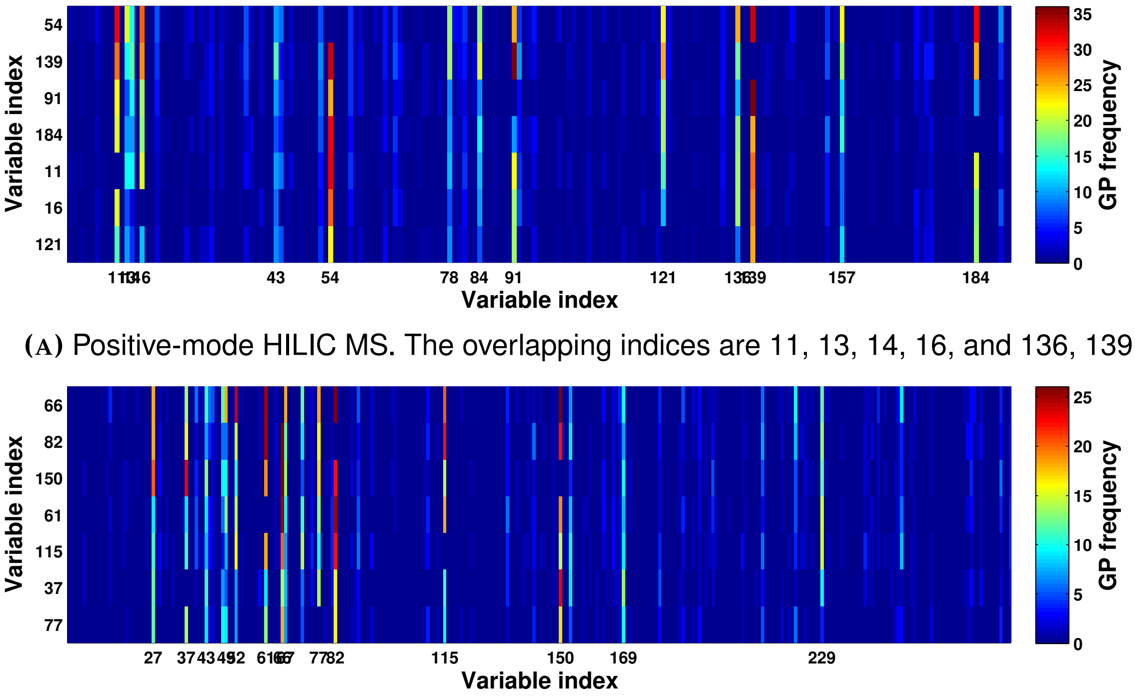
(B) HSQC

FIGURE 8.9: The ranks of the GP-selected variables in other fusion techniques. A value of 1 indicates the largest absolute loading.

8.3.5 Complementary selection

When considering the high frequency and low correlation of, for example, HSQC variable 115, it is necessary to consider that its high selection rate might depend on other variables; in short, variable 115 does not classify alone, so what else is frequently selected alongside it? The plots in Figure 8.10 show how often other variables are picked when each of the seven top variables are also chosen. Each plot is shown as a heat-map with the selection frequency given by the colour bar. In Figure 8.10b it can be seen that HSQC variable 115 is selected most frequently with variable 82, and to a lesser extent with 61 and 67. After inspection of various GP trees, reasons for its repeated selection remain unclear, but further GP studies may help in rationalising its high frequency.

For the MS variables in Figure 8.10a, there are fourteen that are consistently high, whilst for the HSQC variables (Figure 8.10b) the distinction between high and middle frequency is less clear. The distinction is not as important as the fact that the top seven variables from each technique are included in the GP trees consistently with other variables. Excluding these 'other variables' may, therefore, not be justifiable.



(A) Positive-mode HILIC MS. The overlapping indices are 11, 13, 14, 16, and 136, 139.

(B) HSQC. The congested variable indices are 37· · · 43, 49, 52, 61, 66, 67, 77, 82· · · 115

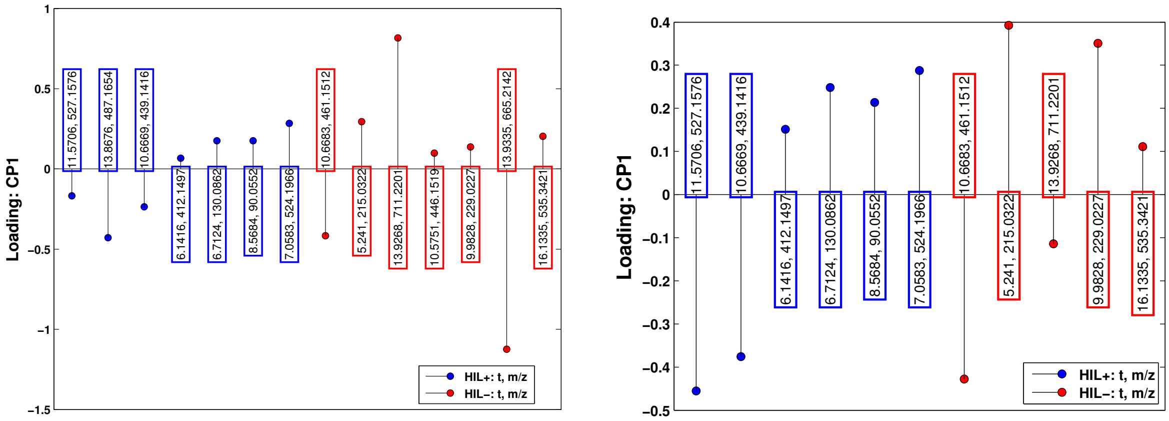
FIGURE 8.10: Plots showing how often other variables are selected with the first six variables, which are given by their indices on the y-axis.

With reference back to Figure 8.4, it may be more appropriate to permit the inclusion of seventeen variables from each dataset for analysis by CCA. Whilst there is little improvement in the D^2 value compared to seven, these two subsets provide, technically, the greatest correlation to tenderness, and it might be better to include those variables in CCA that consistently work together in the GP solutions.

8.3.6 Positive and negative ion mode LC-MS

The previous section was concerned with the fusion of NMR and MS data. These two techniques can provide a vast amount of complementary information due to their different detection methods. However, not all samples are appropriate for analysis by two techniques, for example those with molecules at very low concentrations. Furthermore, it is more common for analysts to have recorded multiple LC-MS experiments under different parameters, notably the ion mode and the chromatographic column. Running positive- and negative-mode LC-MS is relatively simple as the same sample can be used, rather than the more involved NMR/MS sample preparation. The analysis of both ion modes is very likely to reveal complementary information, as not all molecules are likely to form cationic and anionic species; the use of both methods can help to enhance the analytical coverage of a sample. Also, for molecules that do ionise in both forms, then these can be directly compared as the retention times should be the same (± experimental drift).

The optimal number of variables to include was determined by reference to the D^2 metric. In this instance, seven variables from each dataset were chosen, and the loadings for these variables are shown in Figure 8.11a. There are two negative mode variables with large absolute loadings but different signs. These have approximately the same retention time (13.9 min), and are strongly positively correlated ($\rho = 0.99$). Additionally, there are further high correlations between HILIC negative variables two and four ($\rho = 0.93$) and the first two HILIC positive variables ($\rho = 0.93$).



(A) All 14 variables are present, and the two major contributions derive from positively correlated variables.

(B) Variables correlated by greater than 0.90 to others within their dataset have been removed.

FIGURE 8.11: The loadings for the positive and negative mode HILIC MS variables. In (B) the correlated variables within the same dataset have been removed.

As the variables are strongly correlated to each other, and to tenderness, it might be expected that they should contribute in a concerted manner. Their difference is, in fact, due to their high colinearity and these opposite magnitude loadings are commonly observed in regression-like techniques (e.g. multiple linear regression) for highly correlated variables. This contradiction in the loadings suggests that whilst GP may be reducing the rank deficiency, CCA is still affected by colinear variables. Clearly, two correlated variables should be contributing in a concerted manner; as the variables are correlated, no information would be lost were one of the variables to be removed from the CCA. For assignment purposes any correlated variables can be subsequently referred to.

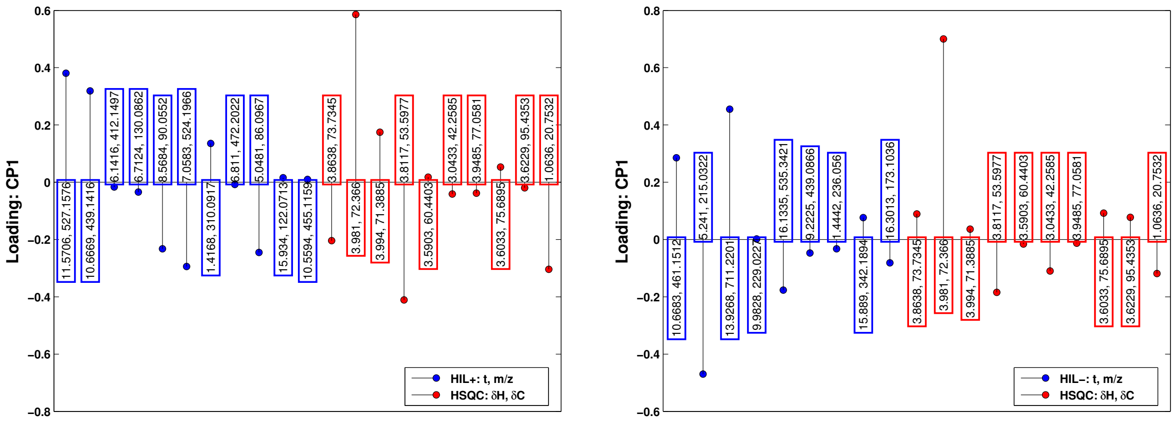
The three variables identified above as being highly correlated ($\rho > 0.9$) have been removed prior to CCA, and the loadings for the remaining 11 variables are shown in Figure 8.11b. There are four variables with negative loadings, and all but one of these are of relatively large magnitudes. Furthermore, two have experimentally identical retention times of 10.67 min and it may be the case that these positive and negative mode features represent the same analyte. The other two features have distinct retention times and are clearly distinct species. The manual interpretation of the positively loaded variables is a little less straightforward due to the lack of coeluting ions. Many of these, however, have lower retention times than the negatively loaded variables, which suggests a different class of molecule in terms of, for example, size.

The use of the two LC-MS datasets made clear the effects of including variables in CCA that are colinear. All variables are, to some extent, correlated with others and the difficulty arises in setting a threshold at which variables are considered colinear. As the loadings in Figure 8.11b make intuitive sense, a threshold of 0.9 is chosen here.

8.3.7 Correlation threshold

Figure 8.12 shows the loadings for the analyses performed with the MS variables (positive and negative modes) and the HSQC data. Again, a threshold of $\rho > 0.9$ was applied in order to filter out the obviously colinear variables. These two plots complement that shown in Figure 8.11b and the three of them complete the relationship 'triangle' between the three datasets.

Figure 8.12a reveals two strong relationships between positive mode HILIC and HSQC variables. In comparison to the seven variable results (Figure 8.6), the agreement between the first two MS variables and the second NMR variable has seemingly strengthened. This is most likely due to the removal of the MS variable with m/z = 487.1654 that was highly correlated to the first. In Figure 8.6 the HSQC variable δH = 3.8117, δC = 53.5977 ppm has an insignificant loading, but in Figure 8.12a it is the largest negatively loaded variable. The change in variable significance is due to the removal of the HSQC peak (δH = 1.50 ppm, δC = 18.99 ppm) with which it is highly correlated.



(A) Positive mode HILIC MS and HSQC

(B) Negative mode HILIC MS and HSQC

FIGURE 8.12: The loadings from GP-CCA for the MS variables when paired with the HSQC variables. Variables correlated by greater than 0.9 to others within the same dataset are removed. Fourteen variables were found to be optimal for the positive mode MS variables, and thirteen for the negative mode variables.

Figure 8.12b shows three variables with strong positive loadings, and four with moderately negative values. The HQSC variable with the largest loading also appears strongly in Figure 8.12a, showing potentially how variables are related across the three techniques. A summary of the 'top' variables is given in Table 8.2, and the variables have been grouped according to their correlation with tenderness.

Twenty-eight variables are shown in Table 8.2, and these are taken from the three analyses using the HILIC LC-MS and HSQC data. These variables contribute most strongly to the tenderness observed in each of the analyses' canonical pairs. Some of the LC-MS variables are coeluting which increases the probability of them being the same analyte. For the HSQC variables, their relationship to other HSQC variables may be assessed by the use of RANSY [352] (Section 9.3.2.1) which can be used to group peaks that exhibit similar intensity ratios throughout multiple observations. The assignment of these features is discussed in more detail in Chapter 9.

TABLE 8.2: The variables identified by the three GP-CCA routines that have the largest loadings. They have been grouped according to their correlation (ρ) with the tenderness values, and also according to the dataset from which they were taken. Variables that were removed from the CCA due to their high correlation are shown in italics.

$\rho > 0$		$\rho < 0$	
t / min	*m/z* (+)	t / min	*m/z* (+)
11.57	527.1576	8.57	90.0552
10.67	439.1416	7.06	524.1966
13.87	*487.1654*	6.71	130.0862
		5.05	86.0967
		6.43	*118.0863*
		18.36	*84.0811*
t / min	*m/z* (-)	t / min	*m/z* (-)
10.69	461.1512	5.24	215.0322
13.93	711.2201	16.13	535.3421
13.93	*665.2142*	9.98	229.0227
11.62	*617.1545*	*10.58*	*446.1519*
11.62	*503.1613*		
δH	δC	δH	δC
3.98	72.37	3.86	73.73
3.99	71.39	1.06	20.75
3.87	*71.39*	3.81	53.60
4.02	*73.93*	*1.50*	*18.99*
4.05	*72.37*		

8.3.8 Concatenated GP

The use of GP has so far been limited to individual datasets. The previous Chapter detailed various concatenated approaches to 'standard' multivariate techniques, such as cPCA where all of the variables from multiple datasets are concatenated together prior to a single principal components analysis. In grouping all of the variables together, a different linear combination of variables is achieved and this may result in different variables contributing to specific trends seen in the scores. The combination of the variables from the outset may, therefore, capitalise on the inherent complementarity that exists between the datasets.

A single GP was performed with the concatenated and unit-variance standardised variables from positive mode HILIC LC-MS and HSQC. Figure 8.13 shows the variable selection frequency plotted as a function of the correlation to tenderness. The variables selected in the concatenated form are different to those picked by the two individual GP analyses (see Figure 8.8). Only four variables are found with high frequencies in both analyses (MS variable 91, and 37, 61 and 66 from HSQC), and those that are picked in the concatenated GP tend to have greater correlations with tenderness than those picked in the individual GP analyses.

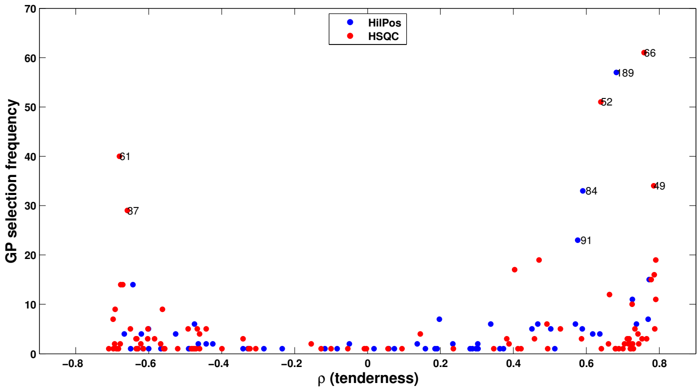


FIGURE 8.13: The frequency of selection of MS and NMR variables plotted as a function of the correlation to tenderness. The two-stage GP was run with concatenated data, and then split according to MS and NMR variables. The variables with a selection frequency greater than 20 have been labeled.

As the variables selected in this concatenated form of GP are highly inter-correlated, it is again crucial that the colinearity of variable subsets is reduced prior to CCA. As was implemented above, variables within the dataset that correlate strongly to others are not included. Although the data was concatenated prior to GP, the removal of variables correlated to others in the opposing dataset would not allow for the visualisation of variables that are related across the two datasets. Therefore, the removal of variables is independent of the other dataset. A threshold of 0.9 was applied, and the effect of enlarging each subset with more variables is shown in Figure 8.14.

As expected the canonical correlation rises continually with the inclusion of more variables, but the tenderness correlation is best when 26 pairs of variables are considered for CCA. With the inclusion of the fourth pair, however, the agreement suffers considerably and does not recover until the 23rd variable pair is included. The intervening variables do not contribute to the agreement with tenderness, so whilst

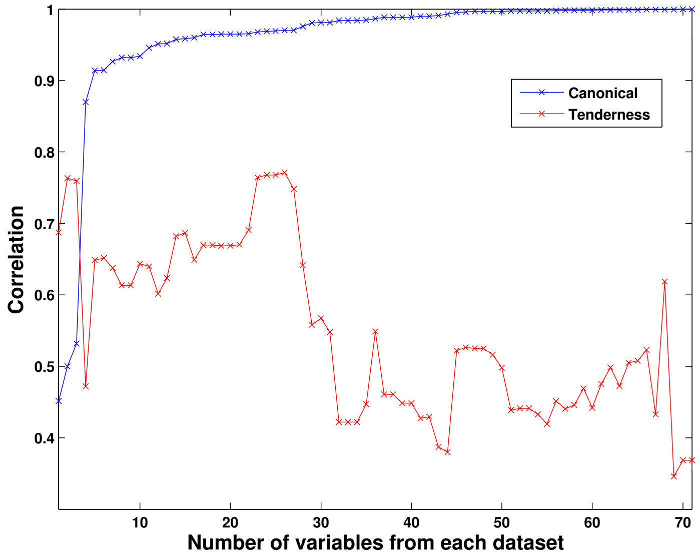


FIGURE 8.14: The canonical and tenderness correlation values for GP-CCA, where the GP was run with concatenated MS and NMR data. Variables correlated within the same dataset with $\rho > 0.9$ have been removed.

these variables have high selection frequencies they appear to not be informative. Hence the use of concatenated data within GP appears to be ineffectual, with individual analyses being preferred.

It is speculated that the poor performance may due to the inclusion of too many variables. Extra variables enlarge the search space, with the effect that it may not have been adequately sampled by the repeated GP runs. Optimisation of the various GP parameters may help in improving the concatenated approach.

8.4 Conclusion

The use of multiple data sources provides great scope in enhancing the analytical coverage of chemical samples. In an ideal case, each technique's weakness is mitigated by the strengths of a different technique; for instance, the lower sensitivity of NMR is offset by LC-MS whose selectivity is countered by the inclusivity of NMR. Using multiple techniques together requires the adoption of statistical approaches that allow variables from multiple datasets to work in a concerted fashion.

The development of the two-stage GP-CCA in this Chapter is a direct improvement on the other two-stage CCA approaches that involved initial data reductions using either PCA or PLS. These data reduction stages were necessary to reduce the rank deficient nature of the data, but they confounded the interpretation of the results, as CCA was maximising the correlation between linear combinations of linear combinations. In GP-CCA, however, the variables used by CCA are 'original', so that the interpretation of the scores produced is significantly simpler and more chemically intuitive.

Whilst the results from GP-CCA may be simpler to interpret, they may also be considered as complementary to those from other fusion methods. The plots in Figure 8.9 show the top GP variables do not always feature strongly in the more traditional multivariate methods, such that variables which may easily be assigned, and which show good agreement with the experimental aims (in this case the sample tenderness), are insufficiently emphasised.

Whilst the variable subsets selected by GP are not rank deficient, they do contain colinear variables that together inhibit the interpretability of CCA. As with other regression-like techniques, the presence of colinear variables often results in them having loadings of opposite signs where in fact they contribute to the same global trend. The use of a $\rho > 0.9$ correlation threshold removes the highly positively correlated variables without reducing the information content of the variable subset, and increases the interpretability of the loadings as variables within the same data set are no longer competing against each other.

The overall aim of using CCA is to identify variables from different datasets that are complementary. This should help with assignment purposes, and bring together the strengths of multiple techniques, whether they are both ion modes in LC-MS, or the use of NMR to complement an individual LC-MS analysis. The variables listed in Table 8.2 may help to reveal molecules that contribute to the tenderness trend, and their assignments are discussed in more detail in Chapter 9.

Chapter 9

Feature Assignment

There are ′known knowns′; these are things we know that we know. There are ′known unknowns′; that is to say, there are things that we now know we don′t know. But there are also ′unknown unknowns′ - there are things we do not know we don′t know.

Donald Rumsfeld

9.1 Introduction

The range of information available in spectral techniques varies, but one of the aims of data fusion is to combine complementary data sources. Not only do MS and NMR complement each other in terms of sensitivity and specificity, they are also complementary in terms of spectral interpretation: NMR is better than MS at differentiating between isomers, whilst MS outperforms NMR in terms of mass and composition.

Without extensive additional fragmentation spectra, then interpretation of MS spectral features relies on (accurate) *m/z* values, coeluting and fragment ions, and their corresponding isotopic distributions. Retention time is not a fixed molecular property and its use for comparative purposes is only recommended when the operating conditions between reference and experimental samples are identical. The greater the accuracy of the *m/z* value the fewer the number of molecular formulae that can be matched to it. However, even at accuracies at less than one part per million (ppm) larger *m/z* values will not result in a single candidate. The combination with isotopic information may reduce the number of candidates, as may the application of other so-called 'golden rules' [35] which are aimed at assessing the structural and elemental likelihood of a feature (e.g. carbon : hydrogen ratio and the number of heteroatoms).

There are many available online repositories that contain spectral data, and they serve as an excellent resource for many of the more generic molecules. Where databases are incomplete, and in-house reference spectra are unavailable, then *de novo* structural interpretation is the only remaining solution. However, without fragmentation spectra for MS, the ability to assign putative candidates is limited to the *m/z* accuracy and the ability to reduce the list of possible molecular formulae. With NMR, interpretation is typically not possible without library spectra of reference materials. *In silico* advances in prediction of NMR spectra [353] may assist with this, but should perhaps be used only when the differences between theoretical and experimental shifts for known molecules have been quantified. Knowledge of the biological processes in systems may also prove useful for assignments. Where genomic or proteomic approaches have also been conducted, then reference to specific genetic pathways may further help with assignments.

The assignment of spectral features is typically never certain unless a pure standard is analysed using the same experimental conditions. The metabolomics standards initiative (MSI) has proposed a scheme to harmonise the reporting of assignments [236]. Where multiple properties have been compared under identical conditions then an assignment value of 1 can be awarded. A value of 2 is given to compounds where the best match is derived from external data sources, e.g. online databases. Where only the compound class (e.g. sugar) can be differentiated a value of 3 is given, and the fourth and final classification is given to spectral features where nothing can be inferred relating to their identity.

This Chapter details various procedures that have been implemented in order to facilitate the semi-automated assignment of spectral features, which have been demonstrated as being highly discriminatory in Chapter 8.

9.2 Data

Of the various online spectral databases, very few allow for the data to be downloaded such that it can be used in local databases. SDBS [297] allows for only website access, which is limited to performing single queries. These, however, are not suitable for MS applications as only masses, instead of *m/z* values, can be searched. The Metlin [294] database is exclusively a mass spectrometry database, and allows multiple types of search to be carried out; queries can be in positive, negative and neutral mode, and even specific fragments can be probed for MS/MS assignments. Whilst downloading the spectral data is not possible, the database allows access through an application programming interface (API) such that custom scripts can be written by researchers for use outside of web browsers. The human metabolome database (HMDB [114, 115]) allows its data to be downloaded, and many of its records contain NMR spectral information in either or both of 1D ^{1}H NMR or 2D ^{1}H-^{13}C HSQC formats. This is an incredibly rich database, and in spite of it being primarily focused at human metabolites, many of its spectra are for those molecules that are universal to metabolism. Due to its contents and its availability, the HMDB [114, 115] data has been downloaded and incorporated into an in-house database, which is used for all of the subsequent feature assignment.

Whilst the database contains many thousands of entries (approximately 8000), only about 10% of these contain NMR spectral information. Each of the entries contains the molecule's name, its chemical formula and monoisotopic mass. A range of other information relating to its known roles in metabolism is also present, but this has not been included in the customised version (referred to henceforth as JDB). The record for each metabolite thus contains: JDB record number, HMDB code, name, formula, mass and each NMR chemical shift where available. Table 9.1 shows the entry for a typical metabolite.

TABLE 9.1: An example entry of a molecule's record within JDB. All of the data is taken from HMDB [114, 115].

Code	HMDB00044	**Ascorbic acid**
Formula	$C_6H_8O_6$	
Monoisotopic mass	176.032089	
1D ^{1}H / ppm	HSQC ^{1}H / ppm	HSQC ^{13}C / ppm
4.5096	4.5046	81.0846
4.0096	4.0111	72.2901
3.7446	3.7296	65.3719

9.3 Experimental

Each JDB entry contains NMR spectral peaks, monoisotopic mass and the chemical formula, and all three of these sections are used in the assignment of features. The database can be searched with either a monoisotopic mass, or a chemical shift (1D and 2D). The results returned are for those where a match has been found.

9.3.1 *m/z* and mass

Each ion from a mass spectrum may exist in one of many forms, which are commonly, in positive mode, protonated or sodiated. In order to search a database of masses, each of the possible ion forms must be considered. As shown in Table 9.2, for positive mode, 32 ions have been considered, whilst for negative mode this number is 14. It is with each of these possible masses that the database is, in turn, searched. Any hits for each of the masses are presented, along with an accuracy value, in parts per million (ppm), the ion form (e.g. M+Na$^+$), and an isotopic score which details the goodness of fit between observed and theoretical isotopic distributions.

9.3.1.1 Isotopic distribution

For each of the hits returned from the original query, the isotopic distributions are compared. As the experimentally measured distribution contains the contributions from the charge-carrying species, there cannot be a fair comparison until the effects of this are negated. By adding (or subtracting) the charge-carrying elements to the candidate molecular formula, the effective chemical formula is determined, from which the theoretical distribution can be calculated. For example, an ion is found to match the database entry of valine, with a formula of $\mathrm{C_5H_{11}O_2}$. If the charge is assumed to derive from a potassium ion, then the distribution of M+K, i.e. $\mathrm{C_5H_{11}O_2K}$ must be calculated. This approach also allows for easy interpretation of multiply charged ions or where M is greater than 1 (e.g. $[2M+H]^+$).

Once the 'correct' elemental composition has been determined, the theoretical distribution can be calculated from it. An algorithm that uses Fourier transformation [354] has been used to calculate the isotopic distributions from the effective permutations that arise from a series of elements with differing coefficients (compositions). The method has been extended here to allow for 18 elements, such that many of the elements that have been observed in metabolites are included. The elements, along with their compositions are shown in Table 9.3.

The comparison is made by calculating pair-wise intensity differences for both distributions, which are scaled so that the largest peak has an intensity of 100. The difference between the two initial peaks is often zero. The sum of the absolute differences between each successive pair of peaks is calculated, and normalised to the sum of the theoretical distribution. As the number of peaks for a theoretical distribution is likely to outstrip the number in an experimentally observed distribution, only the first n peaks are considered, where n is the minimum number of peaks between the two distributions.

TABLE 9.2: The positive and negative mode adduct species that have been considered, along with their notation and charge. ACN, acetonitrile C_2H_4N; DMSO, dimethyl sulphoxide C_2H_6OS; IsoProp, 2-propanol C_3H_8O; FA, formic acid CH_2O_2; HAc, acetic acid $C_2H_4O_2$; TFA, trifluroroacetic acid $C_2HF_3O_2$.

Mode	z	Notation
Positive	1	M+H, M+NH_4, M+Na, M+CH_3OH+H, M+K, M+ACN+H, ...
		... M+2Na-H, M+IsoProp+H, M+ACN+Na, M+2K+H, ...
		... M+DMSO+H, M+2ACN+H, M+IsoProp+Na-H, 2M+H, ...
		... 2M+NH_4, 2M+Na, 2M+K, 2M+ACN+H, 2M+ACN+Na
	2	M+H+NH_4, M+H+Na, M+H+K, M+ACN+2H, M+2Na, ...
		... M+2ACN+2H, M+3ACN+2H, 2M+3H2O+2H
	3	M+3H, M+2H+Na, M+H+2Na, M+3Na
Negative	1	M-H, MNa-2H, MCl, MK-2H, MFA-H, MHAc-H, MTFA-H, ...
		... M-H2O-H, 2M-H, 2M+FA-H, 2M+HAc-H, 3M-H
	2	M-2H
	3	M-3H

TABLE 9.3: The 18 elements and their compositions used in the determination of isotopic profiles. All values are taken from the Bruker Almanac 2011 [355].

Element	**Mass Number** (Composition)
H	**1** (0.999885), **2** (0.000115)
Li	**6** (0.0759), **7** (0.9241)
C	**12** (0.9893), **13** (0.0107)
N	**14** (0.99636), **15** (0.00364)
O	**16** (0.99757), **17** (0.00038), **18** (0.00205)
F	**19** (1)
Na	**23** (1)
Mg	**24** (0.7899), **25** (0.100), **26** (0.1101)
Si	**28** (0.92223), **29** (0.04685), **30** (0.03092)
P	**31** (1)
S	**32** (0.9499), **33** (0.0075), **34** (0.0425), **36** (0.0001)
Cl	**35** (0.7576), **37** (0.2424)
K	**39** (0.932581), **40** (0.000117), **41** (0.067302)
Ca	**40** (0.96941), **42** (0.00647), **43** (0.00135), **44** (0.02086), **46** (0.00004), **48** (0.00187)
As	**75** (1)
Se	**74** (0.0089), **76** (0.0937), **77** (0.0763), **78** (0.2377), **80** (0.4961), **82** (0.0873)
Br	**79** (0.5069), **81** (0.4931)
I	**127** (1)

Therefore, the lower the score, the closer the match between experiment and theory. The isotopic score, alongside the mass fidelity, is a useful metric in determining the goodness of the proposed assignment. The main issue, however, is that neither of these two metrics is able to discriminate between isomeric assignments. This may be achieved by inspection of any coeluting ions (or those from MS/MS), or techniques capable of discriminating between isomers (e.g. NMR spectroscopy).

9.3.1.2 Coeluting ions

Ions that coelute may or may not be related; identical partition coefficients may result from unrelated ions with the same polarity. Related ions may either be fragments of an analyte, or an intact analyte which has been formed as an alternative charge-carrying species. Where multiple charge-carrying species are found to coelute, it may be possible to determine their common mass, by considering these as two (or more) unknowns. For each of the various ion forms (listed in Table 9.2), a series of masses can be determined and where a common mass is found then a putative mass can be assigned. For example, if the difference in *m/z* values between two coeluting ions is 15.974[1] then it can be assumed that these ions are the potassiated and sodiated species, and from these a mass for the analyte can be determined with reasonable confidence.[2]

The alternative category of related ions is that of a fragment if the mass is lower. If the reverse is true then it appears that the ion in question is a fragment of the heavier ion. Fragmentation patterns require chemical interpretation or a database of MS/MS spectra. Unfortunately, HMDB v2.5 does not contain this information such that the automated interpretation of fragmentation spectra is limited. The recently-released version of HMDB does indeed contain various MS/MS spectra, and their inclusion will clearly assist for assignment purposes.

[1] The mass of ^{16}O is 15.995 Da, which can be differentiaited using high-resolution MS.

[2] Of course, there remains a finite possibility that these ions are not the same analyte.

As it is incorrect to state that all coeluting ions derive from the same molecule, no firm assignments should be made exclusively with this information. Instead, fragments may be useful in determining the functional group of a molecule, or its class. For example, α-amino acids are often observed to fragment by loss of formic acid (CH_2O_2) and the corresponding mass loss of 46 Da can be observed.

9.3.2 NMR spectral peaks

Chemical shifts result from an intransigent property of the molecule, and are unlike *m/z* values which depend on the nature of their formation. Searching a database with chemical shifts is therefore, significantly easier, but is subject to the fact that chemical shifts of a molecule are dependent on the sample conditions such as, for example, the pH. A certain tolerance must, therefore, be allowed for. Whilst it may be expected that multiple chemical shifts from a single molecule might vary equally, experience in both one and two dimensional spectroscopy shows that peaks from the same molecule do not always exhibit coordinated variations, i.e. they do not all move the same amount in the same direction.

As most molecules have more than one chemical shift, it is possible to be more stringent when searching NMR databases than with MS data. ‘Soft’ ionisation MS provides, generally, only a single *m/z* for a molecule, whilst in NMR there are likely to be multiple chemical shifts. As such, an AND query may be developed, where multiple shifts are used in a single query and only results which can satisfy all of the criteria are returned. It may, therefore, be easier to assign features, but this assumes that all of the chemical shifts are, in fact, attributable to the same molecule. For 1D spectra, this assertion is difficult to prove, as spectral overlap is extremely likely. Indeed, the same may be said for 2D heteronuclear spectra, such as those from HSQC. With these two types of spectra, the only way to probe the relationship between multiple peaks is through correlation. Strongly positive correlations do not, however, prove that multiple peaks are attributable to the same molecule.

NMR spectrometers do, however, allow for 2D correlation experiments to be run and these are capable of identifying which chemical shifts are structurally related. In COSY and TOCSY spectra off-diagonal peaks reveal chemical shifts that are due to the same spin system (COSY) or molecule (TOCSY). The use of these spectra, therefore, may allow for a more discriminating analysis as candidates may be filtered on their ability to match multiple peaks.

Querying the database with a single spectral peak is perhaps the common way in the assignment of spectral features. This is not, however, the most efficient method as it does not include the other peaks in the spectrum, which may assist in reducing the potential candidate list. As any database is of finite size, it may be better to query a spectral database with all chemical shifts, such that a score for each database record may be generated. In this way, the number of peaks that have been found for each molecule can be presented.

Clearly it is possible for a single experimental variable to be assigned to multiple molecules, but the generation of a score does provide a confidence to be placed on some assignments. For example, molecules where only one of, say, 10 shifts have been identified are less plausible than records that have identified a greater proportion. Again, a caveat must apply: only one peak might be detected from a single molecule, but all the others may be swamped by larger peaks, or be of sufficiently low intensity to be considered noise.

9.3.2.1 Related spectral variables

The use of post-experimental statistical techniques to assist with the identification of linked variables was proposed by Noda [356] for infrared spectra, and techniques such as statistical total correlation spectroscopy (STOCSY) [357] use correlation approaches to identify potentially linked 1D ^{1}H NMR spectral variables. A corollary to this approach has been developed by Wei et al. [352], with what they term ratio analysis spectroscopy (RANSY). The technique differs from correlation spectroscopies in that it uses the ratios of peaks as a means to identify related features.

The intensities of multiple spectral peaks from the same molecule would be expected to remain in the same proportions to each other, regardless of the molecule's concentration. If the same molecule is found across multiple spectra, these ratios should also remain constant, subject to the influences of spectral overlap and noise. Therefore, the same molecule, across multiple spectra, should exhibit very similar ratios. The similarity imbues a low standard deviation, and the inverse of the coefficient of variation ($\bar{x} \div \sigma$) gives a larger value to those peaks displaying unchanging ratios.

The procedure is applied to a single variable, q, and reveals which other peaks vary in the same proportions. The ratios are calculated according to Equation 9.1, where the value for variable j of observation i is denoted by $R_{i,j}$. The ratio matrix, R, is of the same size as the original data matrix, X.

$$R_{i,j} = \frac{X_{i,j}}{X_{i,q}} \tag{9.1}$$

The RANSY spectrum, P, is calculated according to Equation 9.2, whereby the mean of the ratio R_j is divided by its standard deviation.

$$P_j = \frac{\bar{R_j}}{\sigma_{R_j}} \tag{9.2}$$

As the value for P_q is infinite, it is convenient to set it to, for example, 110% of the maximum value such that it is obviously the query peak. Large values indicate a strong likelihood that that a peak is strongly related to the query peak.

The power of RANSY is that it can be applied to both one- and two-dimensional spectra. Furthermore, whilst the technique is capable of being applied to both raw and binned data, the computational requirements, especially to raw 2D NMR, render the technique most easily applied to binned data. The applicability of RANSY plots are shown in Section 9.4.1.

9.3.3 Limitations

The greatest limitation in terms of spectral assignment derives from the range of the database, and the molecules that encompass it. Only with an 'infinite' database can an assignment be confidently made. The database is designed such that it can be easily enlarged.

The advantage of MS is that only a finite number of elemental compositions exist for any given mass (and resultant isotopic distribution), such that with sufficient computational resources molecular formulae can be calculated to match a given mass within a specific mass tolerance. However, the number of possibilities increases with mass [35]. Whilst it may not be efficient to derive all realistic formulae for any given ion, it may become a possibility for unidentified ions when the advance of computation makes such an undertaking feasible.

NMR is significantly more hamstrung, in that unknown features cannot be so easily assigned a putative identity. Even though spectral windows can broadly be divided into regions associated to functional groups, these approximations can never truly be used on their own. As such, the growth of databases with additional spectra is perhaps the most certain way of increasing the opportunity for assignments. NMR spectral predictors, such as nmrdb.org [358], allow a user to draw a molecule, for which the proton NMR spectrum is returned. Although these tools provide only approximations, they may prove useful to those seeking to confirm or refute any potential assignments for which standards are unavailable.

It is clear that any assignment capability depends on the breadth of the database that underpins the process. Many of the online databases provide an excellent starting point, to which many laboratories can add their own records. A key point is that experimental conditions are crucial to all spectra, and that the standardisation of solvents, and operating instrumental conditions are required in order that a database is widely applicable. Not all experiments can be run under the same conditions, perhaps due to constraints on analyte solubility, but metabolomics-like experiments, of primarily water-soluble molecules, should be conducted where possible under standardised conditions.

9.4 Results and discussion

Although the database is relatively small, it serves as a starting point which can be subsequently expanded. Its small size, however, is beneficial in the testing stages, as the number of potential assignments and permutations is more manageable. The variables identified by GP-CCA as being most discriminating with regards to the tenderness values will be subjected to a concerted assignment effort, that is detailed below. These variables are summarised in Table 8.2.

9.4.1 Alanine

An ion with *m/z* = 90.0552 and a retention time of 8.57 min, recorded in positive-mode HILIC MS was considered to exist as one of 32 ions (see Table 9.2). Of these various mass combinations, four were found to match entries in the database. Two of these masses resulted in multiple isomeric hits, and all are shown in Table B.1 in Appendix B. The results are sorted according to increasing ΔM, which is, in parts per million (ppm) the difference between the effective mass of the ion, M, and the mass of the database entry. The match between the experimental and theoretical isotopic distributions is given in the column labelled 'Iso'. The first three candidates in the table all have relatively poor isotopic matches coupled with a generally poor mass agreement, and for the combination of these factors can potentially be excluded. Of the remaining candidates, four are isomeric and are hence inseparable by the two measures. Whilst JDB 265 has the best mass agreement, its poorer isotopic match discounts it from being the most likely candidate. The four $[M+H]^+$ candidates are, therefore, the most likely assignments and the differentiation of these may be effected by NMR. The 'HSQC' column shows how many of a candidates peaks have been found throughout the entirety of the HSQC spectra. This allows for all peaks to contribute, rather than limiting it to only those peaks that are deemed to be statistically significant. None of the candidates have a large ratio of found peaks to total peaks, with the exception of L-alanine, where two of its three database peaks have been found within the experimental HSQC spectra.

Each row of Table 9.4 represents an HSQC variable, given by its index (#), its ppm values and its correlation coefficient to the MS variable. The remaining columns represent the JDB entries, and values of one indicate a match between experimental data and the database. Two variables that exhibit negative correlation are unrelated for assignment purposes, and this premise can be used to discount JDB entries 345 and 352. The most convincing candidates according to HSQC are JDB 99, 265 and 537 as all of these exhibit moderate to strong correlations with the MS variable in question[3]. JDB entries 99 and 537 represent the optical isomers of the amino acid alanine, which are indistinguishable to HSQC and MS. Therefore, the assignment can only focus on the difference between alanine and aminoadipic acid.

TABLE 9.4: HSQC hits to tentative assignments shown in Table B.1. The tolerances for the search were δH = ±0.1 ppm, δC = ±0.2 ppm.

				JDB (Da)					
HSQC #	δH	δC	ρ	99	212	265	345	352	537
				(89)	(245)	(161)	(245)	(137)	(89)
16	3.80	57.31	0.69			1			
37	1.50	18.99	0.86	1					1
61	3.81	53.60	0.87	1					
70	4.46	51.25	-0.27					1	
95	3.23	56.92	0.19		1				
110	1.93	33.07	0.44			1			
117	3.87	59.46	0.65			1			
134	3.89	69.43	-0.41				1		
157	8.11	131.02	-0.38					1	
163	3.75	69.63	-0.44				1		
166	3.76	69.43	-0.44				1		
248	3.69	71.00	-0.43		1				

HSQC variable 61 (δH = 3.81, δC = 53.60) is one of the variables identified by GP-CCA as being related to tenderness, and its appearance in Table 9.4 shows that it can be assigned to alanine and not aminoadipic acid. Furthermore, the power of the RANSY approach (Section 9.3.2.1) is illustrated in Figure 9.1; the plot shows the RANSY spectrum when δH = 3.81, δC = 53.60 is used as the query peak, and the

[3] Again, correlation and causation should be separated, but correlation is a good indicator of whether variables can be considered related.

only other strongly related peak occurs at δH = 1.50, δC = 18.99. The third JDB peak forms part of a doublet which has not been detected in the experimental data. As such, there is unlikely to be a match for it. The presence of multiplet structures in the database artificially alter the database score, and future improvements of the database must ensure that multiplets with low J-values are removed. High J-values should not be so problematic, as these are likely to be resolved sufficiently such that they are considered as two, or more, peaks.

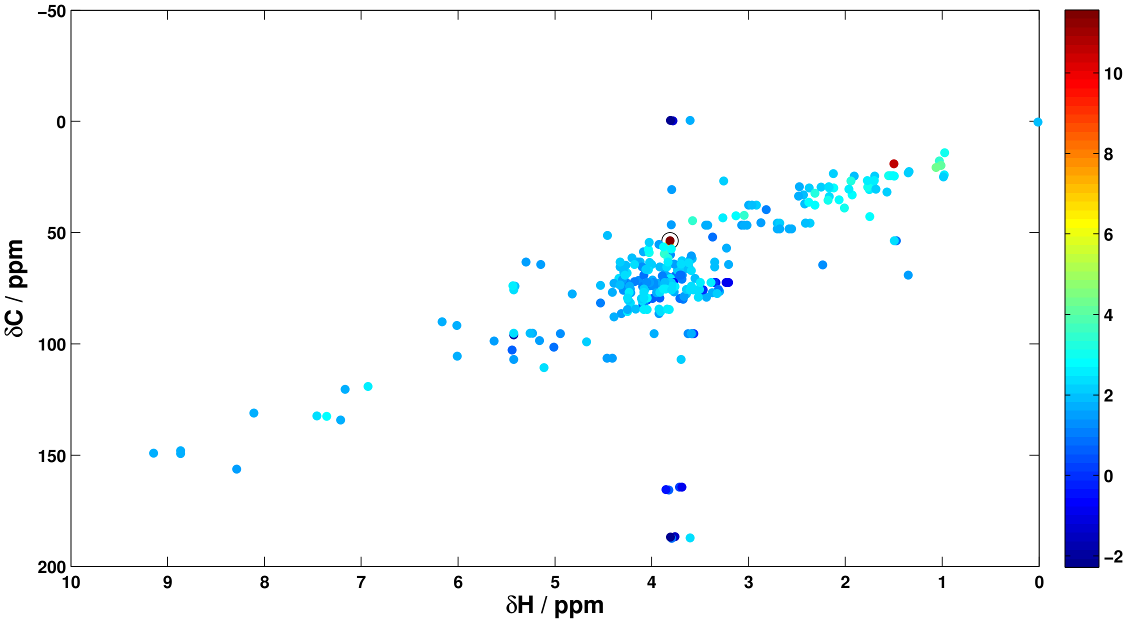


FIGURE 9.1: RANSY plot created for the encircled query variable at δH = 3.81, δC = 53.60 ppm. The intensity of the ratio for each variable to the query variable is given according to the colour shown in the colour bar. Only one other peak has a large positive value, which indicates that these peaks consistently vary in the same ratio across multiple observations. Their exclusive presence supports an assignment of alanine.

All of the evidence supports the assignment of *m/z* = 90.0552 as alanine; both the *m/z* and HSQC variables identified by GP-CCA are attributable to the same molecule, which demonstrates that the two-stage fusion approach is valid in grouping related variables together. The applicability of RANSY to binned HSQC data has also been demonstrated, which may prove a powerful addition to any complex mixture analysis. The high-level fusion methods implemented were not so successful at highlighting the significance of these two variables.

9.4.2 Valine

The putative candidates for the positive mode ion with *m/z* = 118.0863 and a retention time of 6.43 min are listed in Table B.2. All but one of the candidates has a mass accuracy within 1 ppm and an equally high isotopic match. The similarity of the values is due to the fact that all candidates have the same effective molecular formula when considering the molecule with the charge carrying species. Differentiation of the species is therefore not possible without inspection of NMR or fragment ions. Table 9.5 shows the coeluting ion, and it is strongly correlated to the main ion. The *m/z* difference between the two ions is characteristic of the neutral loss of formic (methanoic) acid (CH_2O_2).

TABLE 9.5: Ions eluting within 10 seconds of the ion with *m/z* = 118.0863 at t / min = 6.427

MS #	Δt	t	*m/z*	Δ*m/z*	ρ
13		6.4	118.0863		1.00
19	0.00	6.4	72.0811	-46.0052	0.92

This single fragment is significant, in that it allows the candidate list in Table B.2 to be reduced by excluding those that are not known to be capable of fragmenting to lose CH_2O_2. The structures of the candidates are shown in Figure 9.2, and it can be seen that only 2-pyrrolidinone does not natively contain the appropriate elements, and the others are theoretically capable of such a neutral loss. The use of fragmentation spectra from Metlin [294] is sufficient to reduce the list to two candidates. α-amino acids are known to fragment by loss of formic acid, of which L-valine (presumably also the absent D-valine) and N-methyl-α-aminoisobutyric acid are exclusively capable amongst the candidates. Where MS/MS spectra for a candidate were unavailable, the structure was sufficient to suggest the unlikelihood of such a fragmentation.

RANSY plots (not shown) do not provide a conclusive match for either of the two remaining candidates. However, the presence of an experimental chemical shift (δH = 1.06, δC = 20.75 ppm) amongst the top GP-CCA variables that matches one of valine's (Table 9.6) helps to make the case for an assignment of the valine.

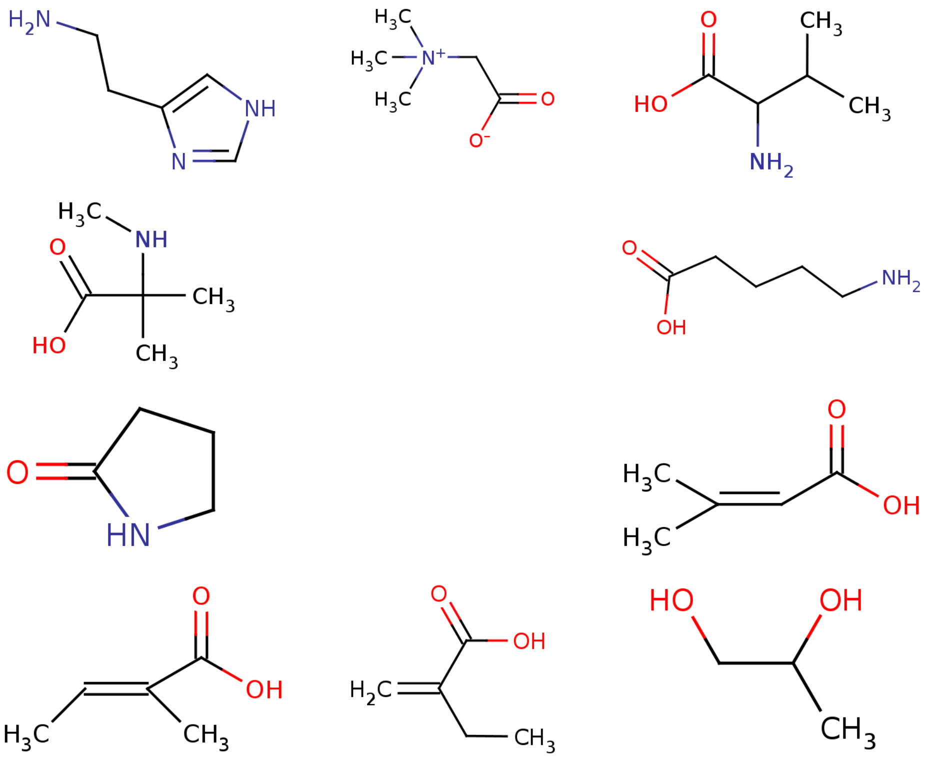


FIGURE 9.2: The structures for the ten candidates assignments in Table B.2. Clockwise, from top left, the molecules are: histamine, betaine, valine, 5-aminopentanoic acid, senecioic acid, propylene glycol, 2-ethylacrylic acid, tiglic acid, 2-pyrrolidinone and N-methyl-α-aminoisobutyric acid. Structure images taken from METLIN [294].

TABLE 9.6: HSQC hits to tentative assignments shown in Table B.2. The tolerances for the search were δH = ±0.1 ppm, δC = ±0.2 ppm.

				JDB (Da)					
HSQC #	δH	δC	ρ	28 (117)	445 (117)	636 (76)	694 (85)	712 (117)	802 (117)
67	1.06	20.75	0.78		1				
77	3.04	42.26	0.66						1
97	3.88	56.33	0.60					1	
122	3.81	70.61	0.52			1			
158	3.86	68.65	-0.31	1					
246	2.43	33.07	0.62				1		

9.4.3 $C_{24}H_{42}O_{21}$

The feature with *m/z* = 487.1654 was found to have only a single hit, as is shown in Table B.3. Whilst the mass and isotopic agreements are good, the assignment of amlodipine is unlikely due to its use as an anti-hypertensive drug. The coeluting ions are shown in Table 9.7. The two ions shown are strongly correlated, and are heavier which suggests that *m/z* = 487.1654 is a fragment of one or both or these ions. The assignments for the heavier ions of *m/z* = 689.2114 and *m/z* = 684.2556 are shown in Table B.4 and Table B.5, respectively.

TABLE 9.7: Ions eluting within 10 seconds of the ion with *m/z* = 487.1654 at t / min = 13.8676

MS #	Δt	t	*m/z*	$\Delta m/z$	ρ
133	0.01	13.9	689.2114	202.0460	0.91
169	0.14	14.0	684.2556	197.0902	0.79
139		13.9	487.1654		1.00

Both of these ions have a common mass of 666.222, when considered as $[M+Na]^+$ and $[M+NH_4]^+$, for which three isomers in the database have been identified. By virtue of clotrimazole being a synthetic drug, and having poorer mass and isotopic agreements it can be discounted[4]. Hesperidin is a flavonone glycoside found in citrus fruits and cannot be as easily excluded as with synthetic drugs. The absence of other coeluting ions with effective masses of 610.1902 and its relatively few peaks identified by NMR, are two factors in favour of excluding it as a likely candidate assignment. Glycogen is also isomeric with these, and whilst the term is often used exclusively with reference to animals, glycogen and its plant equivalent (starch) refer to a molecule with four glucose residues. The differentiation of isomers is again the major issue, but the limited size and scope of the database make it appear that there are only three molecules from which to chose. The reality is that there are many more plant-based molecules with a mass of 666.22 Da which are absent from the human-centric database.

[4] Clotrimazole is a drug for the treatment of fungal infections.

A search in Metlin, which is a more plant oriented database, reveals 16 isomeric structures with a molecular formula of $C_{24}H_{42}O_{21}$, of which three have MS/MS spectra: maltotetraose, stachyose and cellotetraose. The fragmentation patterns of cellotetraose and maltotetraose reveal an *m/z* fragment consistent with the loss of $C_6H_{11}O_6$. Whilst stachyose has the potential to fragment similarly, no positive ion mode spectra have been found to support this. Table 8.2 shows two negative mode ions with a retention time of 13.93 min; their elution is within 4 s of the two coeluting positive mode variables, and with *m/z* values of 711.2201 and 665.2142 they provide a good mass match to 666.222 when considered as formylated and deprotonated species, respectively. Unfortunately, there are no additional coeluting negative mode ions, such that the fragmentation of the features cannot be probed. Due to the high structural similarity of the candidates, it may not be possible to distinguish them. There is no record for cellotetraose in HMDB, which prevents the comparison of the peaks of the sugar molecules.

9.4.4 $C_{18}H_{32}O_{16}$

There are four coeluting ions in negative mode HILIC LC-MS with a retention time of approximately 11.62 min (Table 9.8). Three of these have a common mass of 504.17 Da, with a proposed molecular formula of $C_{18}H_{32}O_{16}$. These three features are all strongly positively correlated and their assignments are shown in Tables B.6–B.8. All have good mass accuracies and isotopic fits alongside multiple NMR hits. Whilst deoxyinosine is also a candidate, it does not have as good statistics and does not provide a match for all three ions. The fourth ion in Table 9.8 is poorly correlated and has a minor retention time drift. Both of these facts suggest that it is attributable to a distinct analyte; however, no matches are found in HMDB. Additionally, the positive mode ion with *m/z* of 526.1576 and a similar retention time gives a mass of 504.17 Da when considered as a $[M+Na]^+$.

Although JDB contains only two isomeric molecules with a formula of $C_{18}H_{32}O_{16}$, Metlin lists upwards of thirty compounds with the same molecular formula. Without comparison to in-house sugar standards, it is likely that NMR will remain unable to differentiate between multiple sugars.

TABLE 9.8: Ions eluting within 10 seconds of the ion with *m/z* = 503.1613 at t / min = 11.6193

MS #	Δt	t	*m/z*	Δ*m/z*	ρ
60	-0.00	11.6	617.1545	113.9931	0.98
28	-0.00	11.6	549.1659	46.0045	0.99
87	-0.15	11.5	516.2046	13.0433	-0.39
67		11.6	503.1613		1.00

9.4.5 $C_6H_{12}O_6$

The JDB assignments of *m/z* = 215.0322 are shown in Table B.9, and it can be seen that many of them are for chlorinated forms of $C_6H_{12}O_6$. The mass accuracy is high, but the isotopic match is poor, and is approximately equally poor for all assignments. The experimental and theoretical isotopic distributions are shown in Figure 9.3. The large intensity M+2 peak is highly indicative of the presence of a chlorine atom. There is little difference between the chlorine-containing theoretical distributions, but perhaps the most telling difference is that between the [M+1] peaks where a better match is obtained for the sugars rather than theobromine/theophylline. Furthermore, the lack of HSQC hits for the non-sugar candidates seems sufficient to relegate their assignment chances. Many of the sugar candidates have high HSQC selection frequencies, and glucose is unique amongst those as all of its database peaks have been found in the experimental data. The sugar region in NMR is generally very crowded, and with the high quantities of sugars, many isomeric, in plant samples the differentiation is likely to be speculative at best. NMR does, however, provide a starting point in isomeric differentiation, which is one that cannot be provided by 'one-dimensional' MS.

9.4.6 $C_6H_{11}NO_2$

Pipecolic acid is the most likely assignment candidate for the positive mode ion with *m/z* of 130.0862 (Table B.10). It has a coeluting fragment (Table 9.9) of *m/z* 84.0811 which is indicative of the loss of CH_2O_2 in α-amino acids, alongside full HSQC spectral peak matching and good mass and isotopic agreement. Another ion of *m/z*

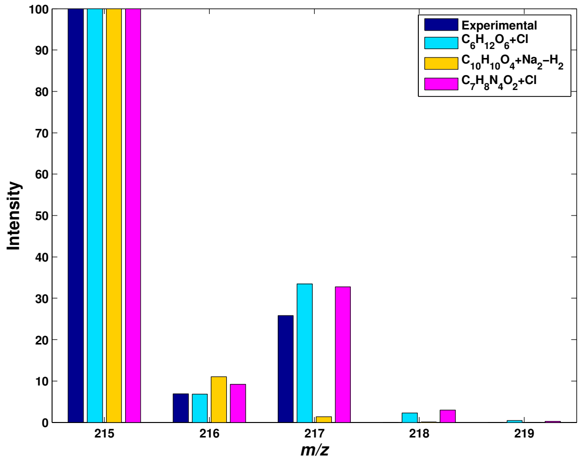


FIGURE 9.3: The experimental isotopic distribution for *m/z* 215.0322 is shown alongside three other distributions for ions shown in Table B.9.

= 84.0811 is amongst the top variables from GP-CCA. This ion elutes after 18.36 minutes and also has a coeluting ion with *m/z* = 130.0862. Other coeluting ions are shown in Table 9.10.

There are two sets of eluting ions each which two ions which potentially characteise the loss of CH_2O_2 from $C_6H_{11}NO_2$. No hits within JDB have been found to the heavier ions, but these may be the precursors for $C_6H_{11}NO_2$. Of course, the other coeluting ions may be entirely distinct, having only the same retention time by chance. Further investigations of the identity of the heavier ions may demonstrate if the lighter ions are fragements of the heavier ones, or discrete entities which coelute.

TABLE 9.9: Ions eluting within 10 seconds of the ion with *m/z* = 130.0862 at t / min = 6.7124

MS #	Δt	t	*m/z*	Δ*m/z*	*ρ*
157	0.10	6.8	472.2022	342.1160	0.85
132	-0.08	6.6	460.2024	330.1162	0.28
125	0.00	6.7	241.1540	111.0678	0.45
88	0.10	6.8	160.0366	29.9504	-0.02
6	0.13	6.8	138.0547	7.9685	0.40
11		6.7	130.0862		1.00
29	0.00	6.7	84.0811	-46.0052	0.71

TABLE 9.10: Ions eluting within 10 seconds of the ion with *m/z* = 84.0811 at t / min = 18.3632

MS #	Δt	t	*m/z*	Δ*m/z*	ρ
161	-0.16	18.2	478.2613	394.1802	0.36
131	-0.09	18.3	326.1318	242.0508	0.37
45	-0.10	18.3	304.1499	220.0688	0.58
96	-0.11	18.3	286.1394	202.0583	0.52
110	-0.10	18.3	215.1024	131.0213	0.28
44	0.00	18.4	130.0862	46.0051	0.96
128	-0.05	18.3	128.0707	43.9897	0.87
43		18.4	84.0811		1.00

9.5 Summary

The various features identified by GP-CCA in Table 8.2 are again presented in Tables 9.11 and 9.12. Any putative assignments for each feature have been noted, whether it is a specific molecule, a molecular formula or a molecular class. Much of the ability to provide single candidate assignments depends on the quality and availability of NMR data. Overlapping variables across the two techniques have been demonstrated to merit the acquisition of multiple datasets. An additional reason for multiple acquisitions is that they enhance the analytical coverage (rather than just mirroring the information). Whilst it is difficult to prove without conclusive assignments, it is plausible that many of the unidentified MS analytes were undetected by NMR.

TABLE 9.11: Variables that are positively correlated to tenderness values, and their putative assignments. The confidence in assignment is given according to that defined by the MSI [236].

$\rho > 0$		MSI	Assignment
t / min	*m/z* (+)		
11.57	527.1576	3	$C_{18}H_{32}O_{16}$
10.67	439.1416	4	–
13.87	487.1654	3	$C_{24}H_{42}O_{21}$
t / min	*m/z* (-)		
10.69	461.1512	4	–
13.93	711.2201	3	$C_{24}H_{42}O_{21}$
13.93	665.2142	3	$C_{24}H_{42}O_{21}$
11.62	617.1545	3	$C_{18}H_{32}O_{16}$
11.62	503.1613	3	$C_{18}H_{32}O_{16}$
δH	δC		
3.98	72.37	3	Sugar
3.99	71.39	3	Sugar
3.87	71.39	3	Sugar
4.02	73.93	3	Sugar
4.05	72.37	3	Sugar

TABLE 9.12: The putative assignments of the negatively correlated variables. The confidence in assignment is given according to that defined by the MSI [236].

$\rho < 0$		MSI	Assignment
t / min	*m/z* (+)		
8.57	90.0552	2	Alanine
7.06	524.1966	4	–
6.71	130.0862	3	$C_6H_{11}NO_2$
5.05	86.0967	4	–
6.43	118.0863	2	Valine
18.36	84.0811	3	C_5H_9N
t / min	*m/z* (-)		
5.24	215.0322	3	$C_6H_{12}O_6$
16.13	535.3421	4	–
9.98	229.0227	4	–
10.58	446.1519	4	–
δH	δC		
3.86	73.73	3	Sugar
1.06	20.75	2	Valine
3.81	53.60	2	Alanine
1.50	18.99	2	Alanine

9.6 Conclusion

The assignment of spectral features may form part of an analysis, whether its aim is screening for adulterants in food or in assessing which molecules are affected under certain stimuli. Assignment of features is essentially a comparison between an unknown in an experiment and a database of compounds showing how each responds under similar conditions.

Whilst targeted analysis is limited by being focused on pre-specified molecules, the aim of non-targeted analysis is to circumvent this restriction. This can only be achieved effectively if the quality of the spectral database is sufficient such that it contains ‘enough’ compounds. Perhaps the largest issue with such analyses is the size of the database with which to make comparisons, with the stark reality being such that it is likely that no database is ever large enough.

The adoption here of the spectral data available from HMDB shows that molecules can be tentatively assigned (MSI level 2) through the concerted use of NMR and MS data. Whilst high-resolution MS data is often sufficient to provide a single molecular formula, without MS/MS data then structural assignments are often not possible for more complex molecules. One of the strengths of NMR is its ability to differentiate between isomeric structures. The combination of a molecular formula and mass with a series of chemical shifts highlights the complementarity between the datasets. In some instances, however, the NMR information is incomplete, such as with the interpretation of sugar molecules in sugary matrices. The heavy spectral overlap in this region is sufficient to obfuscate the signals and make comparisons difficult, especially when the database and experimental spectra have been recorded under different conditions. Work is ongoing to acquire pure samples for many of the isomeric sugars identified in the course of this Chapter. This data should provide a more complete resource for the interpretation and assignment of sugars.

External data sources provide an excellent initial starting point with which to make assignments, or at least to narrow down the list of candidates. Absolute assignments, however, should not be made using data acquired under different conditions. Consequently, assignments to MSI level 1 (confidently identified compounds) cannot be justified. Whilst the highest level achieved throughout these assignments is level 2, the validity of using two analytical techniques has been demonstrated by the ability to turn assignments from level 3 (compound class) to 2 (putative assignments).

Through the series of assignments detailed in this Chapter, the validity of using a two-stage GP-CCA mid-level fusion approach has been demonstrated. Both approaches work well with small subsets of variables, and provide an intuitive way to see how variables relate within (GP) and across (CCA) datasets. The small size of the database employed has limited the capabilities for assignment, but the method's proof of principle has been clearly demonstrated: GP-CCA helps with classification and feature assignment purposes.

Chapter 10

Summary

Finally, in conclusion, let me just say this.

Peter Sellers

Throughout the preceding Chapters, the validity of using NMR and MS for the non-targeted analysis of complex mixtures has been reviewed. Whilst both techniques are highly capable individually, their complementary nature allows a more holistic analysis through fusion of the data sets. Data fusion, however, represents just one of the end stages of a non-targeted analysis involving sample preparation, data acquisition and data processing.

Adequate processing of data is crucial in order that trends between observations can be detected and ultimately traced back to groups of or individual variables. A modelling process was developed for HSQC spectral peaks such that accurate integrals could be extracted even if the peaks were not resolved at the baseline. Although spectral overlap is less dominant in 2D than 1D NMR, spectra of samples in complex matrices undoubtedly contain unresolved peaks. The processing routine models peaks based on a modified Lorentzian distribution function, and the method was demonstrated to be effective by mapping the change in metabolite intensities in rat brains after injection with uniformly labelled ^{13}C-glucose.

A routine was devised for the processing of LC-MS files which converts profile data into a series of peaks as a function of retention time and *m/z* value. The method involved no explicit rounding of *m/z* values such that the inherent mass resolution was maintained and peaks were not split across multiple bins. Further optimisation of the method is required, however, due to the poor speed at which large datasets are processed. The grouping of isotopologue peaks was performed in order to reduce the degree of redundancy within spectra, and to make subsequent spectral assignments easier. The poor long-term reproducibility of LC-MS is a significant issue, and an effective correction strategy based on the use of quality control sample data was implemented in order to negate the dependency of variable intensities on run order.

Both processing regimes have been applied to the LC-MS and HSQC data used throughout much of this thesis. The data were obtained from aqueous methanolic extracts of pea seeds which were analysed in order to garner a greater understanding of the composition of peas. Improvements in the knowledge of peas, with specific regard to the nitrogen fixing ability of the plant, may prove useful in assisting with the development of agricultural policies that are both sustainable and commercially viable. The metabolite profiles of pea seeds have been analysed, and certain traits that are characteristic of quality have been identified through the novel use of a mid-level fusion approach, and subsequently assigned by reference to freely-available spectral databases. The concerted use of genetic programming (GP) and canonical correlation analysis (CCA) has been demonstrated to be effective, in spite of the challenges posed by the low sensitivity of HSQC and the limited size of the database used for assignments.

Many of the existing data fusion methods implemented in the scientific literature focus on the development of high level techniques. These approaches typically enhance the classification power compared to low level and single block approaches. Many such methods were implemented in Chapter 7, but the variable loadings from these techniques were found to be broadly similar to the loadings from the single block methods. This similarity failed to demonstrate any complementarity between variables across the datasets. In studies involving fewer spectral variables [23, 340], those variables that are attributable to the same molecule exhibited similar loadings profiles across two analytical platforms. Whilst this may be true for the high level approaches, here the quantity of variables is inhibitory and the lack of improvement compared to the single block methods shows that high level fusion is not effective in relating spectral variables.

Mid-level fusion strategies are characterised by variable selection, such that discriminatory variables are selected across multiple datasets. These are typically combined into a single ‘metaspectrum’ and analysed by techniques like PCA and PLS. The use of CCA avoids the need to concatenate variable subsets, and allows the direct relationship between variables across two techniques to be probed. Whilst CCA has been previously applied in high level data fusion studies [32], latent variables from PLS were related together rather than spectral variables. The inspection of CCA loadings revealed, therefore, related trends within the dataset rather than the relationship between discrete variables. The power of CCA when used with individual variables has been demonstrated here in a mid-level fusion approach.

Spectral assignment is limited by the scope of the database used, and relatively few of the 'top' variables have been putatively assigned. However, the use of GP and CCA has been demonstrated in helping to focus assignment efforts on fewer variables. The loadings from PCA and PLS shown in Chapter 7 identify many variables with broadly similar values; these multivariate techniques are unable to provide a small clutch of highly discriminatory variables on which assignment efforts can be focused. GP, however, has shown that relatively few spectral variables are repeatedly selected. Whilst the reason for the selection of certain variables by GP is unclear, it has been demonstrated that CCA acts a second filter for variables that are seemingly unrelated to the response variables. Many of the unassigned features can be attributed to isomeric substances, especially sugars. Whilst the reduced spectral overlap in 2D NMR facilitates spectral assignment, the paucity of available HSQC data for many of the commonly occurring sugars in plants hinders their assignment. Many masses and chemical shifts characteristic of sugars have been identified, but combining the two relies on an enhanced database.

10.1 Future work

The ability to assign spectral features relies on a good database. The most recent version of HMDB [114] contains entries for more metabolites than the version used here, and many of these entries have associated MS/MS information. Use of this enlarged database will improve the probability of positive assignments, but the limited number of HSQC spectra within the database poses the most significant challenge. Future database versions will also aim to include the range of HSQC spectra available in BMRB [116]. In addition, the inclusion of user-generated spectra (especially of sugars) will facilitate spectral assignment, and allow a more semi-automated process which has the capability of learning from previous assignments.

Additionally, one- and two-dimensional proton NMR spectra were acquired which may prove useful due to the higher sensitivity of the techniques in comparison to HSQC. The processing of TOCSY spectra has been performed successfully with the same procedure as for HSQC peaks. However, the nature of spectral noise in TOCSY differs from that in HSQC, such that the denoising process employed [81] is ineffective. Further developments in denoising of COSY and TOCSY spectra may permit their use in metabolomics studies. TOCSY spectra may also prove useful in feature assignment, as chemical shifts attributable to the same molecule can be easily identified.

Although the usefulness of HSQC spectra is limited by the sensitivity of the technique, the merit of its combination with MS has been demonstrated. Improvements in the sensitivity and acquisition time [223, 339] of HSQC will clearly prove beneficial, not least due to extra structural information available from HSQC spectra compared to 1D ^{1}H spectra.

Principally due to the use of an aqueous solvent in NMR, much of the focus of this research has been placed on the analysis of the LC separations employing a HILIC column. Most of the assignments have been of noticeably polar molecules, especially amino acids and sugars. Future interpretation of the reversed-phase data is likely to reveal which non-polar molecules are related to tenderness trends, such that a more complete analytical coverage can be achieved. The LC-MS processing method presented in Chapter 6 requires further optimisation such that it can be applied efficiently over the full *m/z* range. The number of features detected by XCMS was demonstrated to be lower than the bespoke method and a more in-depth comparison of the two techniques is warranted.

The variable selection process in GP resulted in a few key variables with high frequencies, and some variables were almost never selected in spite of being highly correlated to those that were frequently picked. Over the course of 100 GP runs, these correlated variables may have been expected to be selected equally frequently. Whilst it is not expected that GP should provide results similar to uni- or multivariate approaches, more investigations into variable selection may prove insightful.

The results from the concatenated GP approach in Chapter 8 were less intuitive than those obtained from individual analyses, which may have been a consequence of enlarging the search space due to the increased number of variables. Varying the GP parameters, such as the maximum number of generations, or the development of a three-stage GP may help with the interpretation of results from concatenated GP approaches.

The use of concatenated datasets may allow pairs of datasets to be analysed by CCA. For example, using GP with concatenated positive and negative mass spectrometric mode variables may reveal features that are discriminatory when using a specific column, and the complementarity between two columns could be probed by CCA. This may help to overcome the major limitation of CCA in that the present implementation is capable of analysing only two distinct blocks of data. The nature of CCA, however, does not preclude the inclusion of more. Maximising the pairwise correlations between three or more linear combinations of original variables may prove useful in revealing more complete interactions between variables [359, 360].

Appendix A

LC-MS File Processing

> If you build a better mousetrap, you will catch better mice.
>
> George Gobel

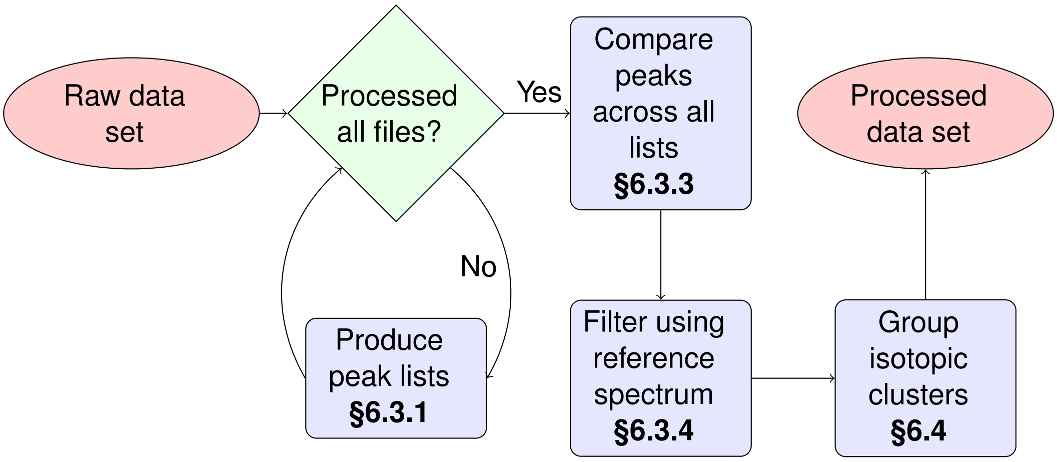

Raw data set
Processed all files?
Yes
No
Produce peak lists §6.3.1
Compare peaks across all lists §6.3.3
Filter using reference spectrum §6.3.4
Group isotopic clusters §6.4
Processed data set

FIGURE A.1: A flowchart representing the various stages in the processing of mass chromatograms. Each file is initially processed to produce a list of its constituent peaks. Subsequently, all of these lists are compared in order that a data matrix may be extracted where each column represents one common observation. A reference spectrum can be used in order to remove noise-like variables, or those occurring in a small proportion of samples. The procedure then begins to interpret the isotopic information inherent to mass spectrometry. The reference to each stage of the process is given by §.

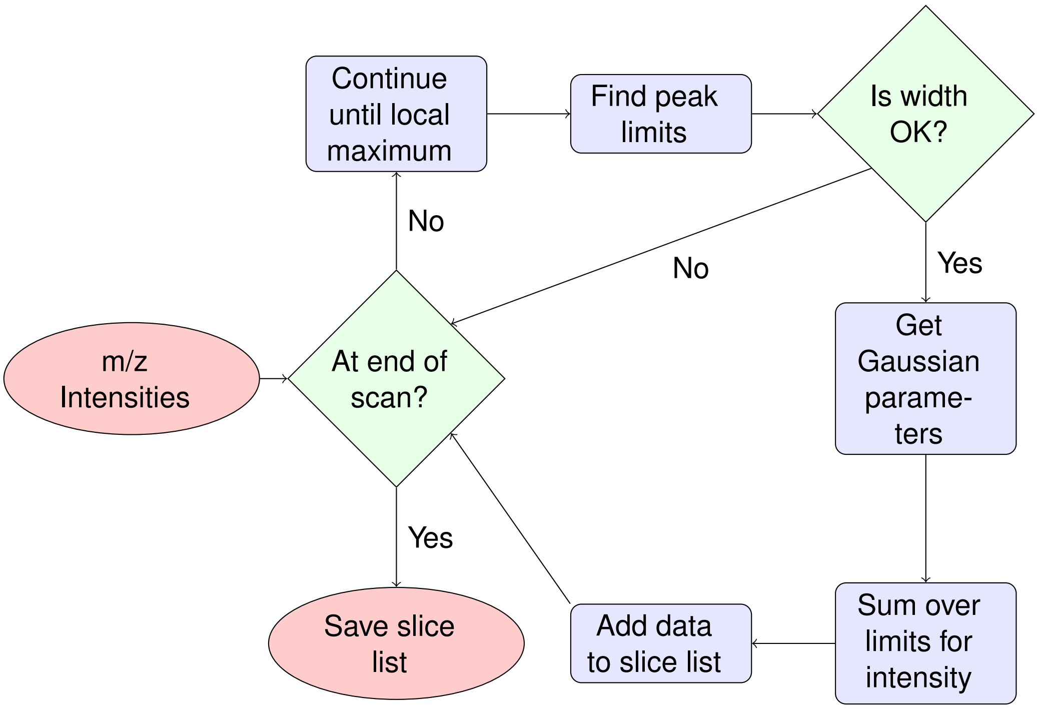


FIGURE A.2: Flowchart detailing the procedure employed in calculating Gaussian functions for each peak in a mass spectrum. The procedure starts with the *m/z* and intensities for the particular scan, with the output containing representations of each slice in a much more condensed format.

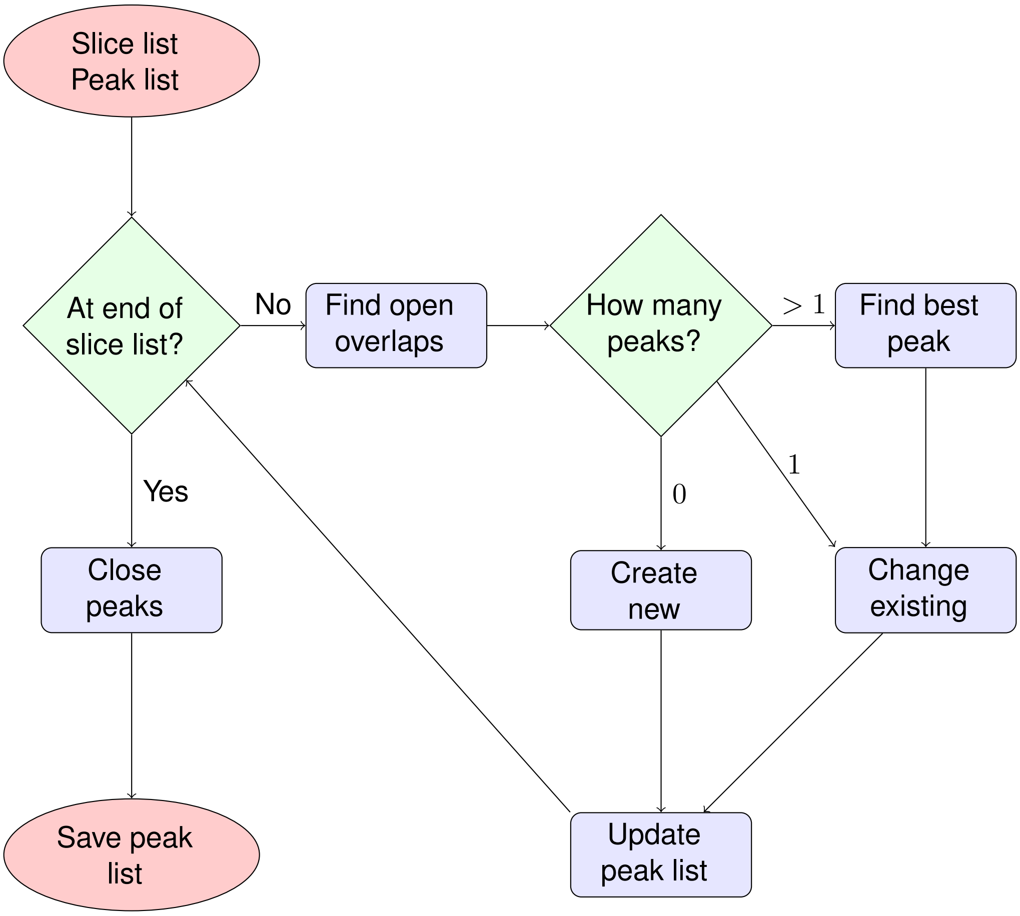


FIGURE A.3: Flowchart detailing the procedure used to group slices of a chromatographic peak together. The slices from the first scan form the beginnings of the peak list, with subsequent slices compared to the peak list. Where there is no valid comparison in the peak list, a slice begins its own peak which can be expanded in the subsequent scan. Peaks are ‘closed’ if they are not added to within a set number of scans, and remain unmodified.

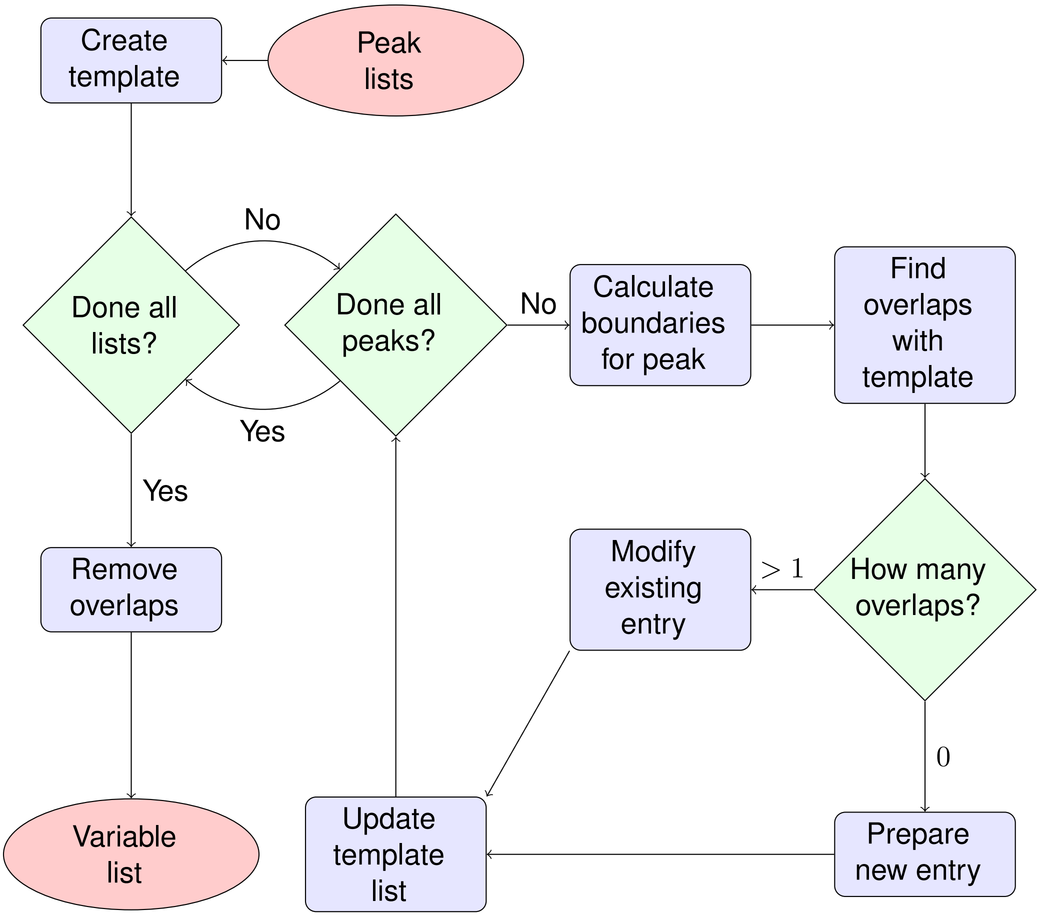


FIGURE A.4: A flowchart that details the procedure used to compare peaks across multiple peak lists. The aim is to produce a data matrix, whereby each column represents a comparable peak across all mass chromatograms.

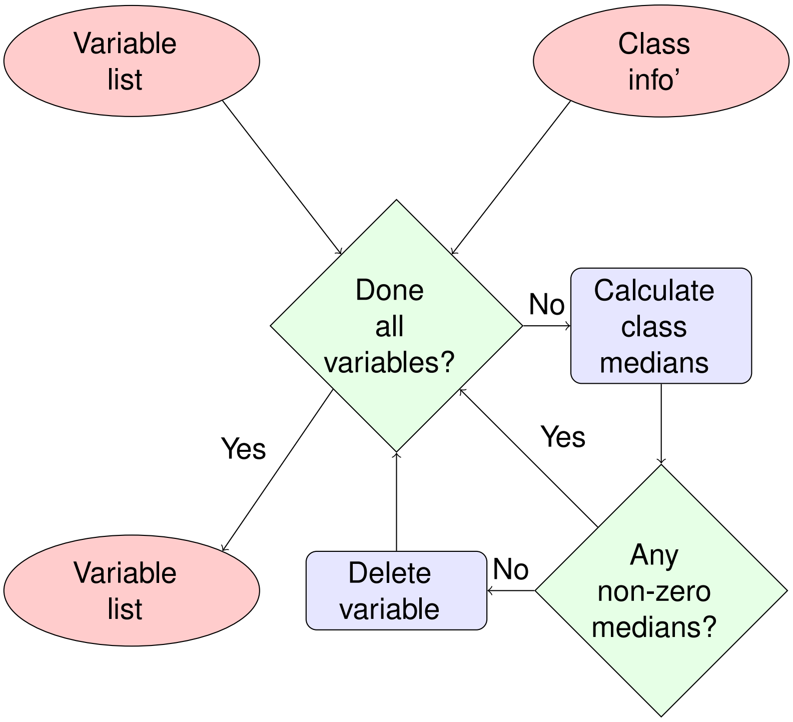


FIGURE A.5: A flowchart the shows the procedure used to calculated a 'group median' reference spectrum. Variables that are not found in at least half of a class' observations are removed from the data matrix.

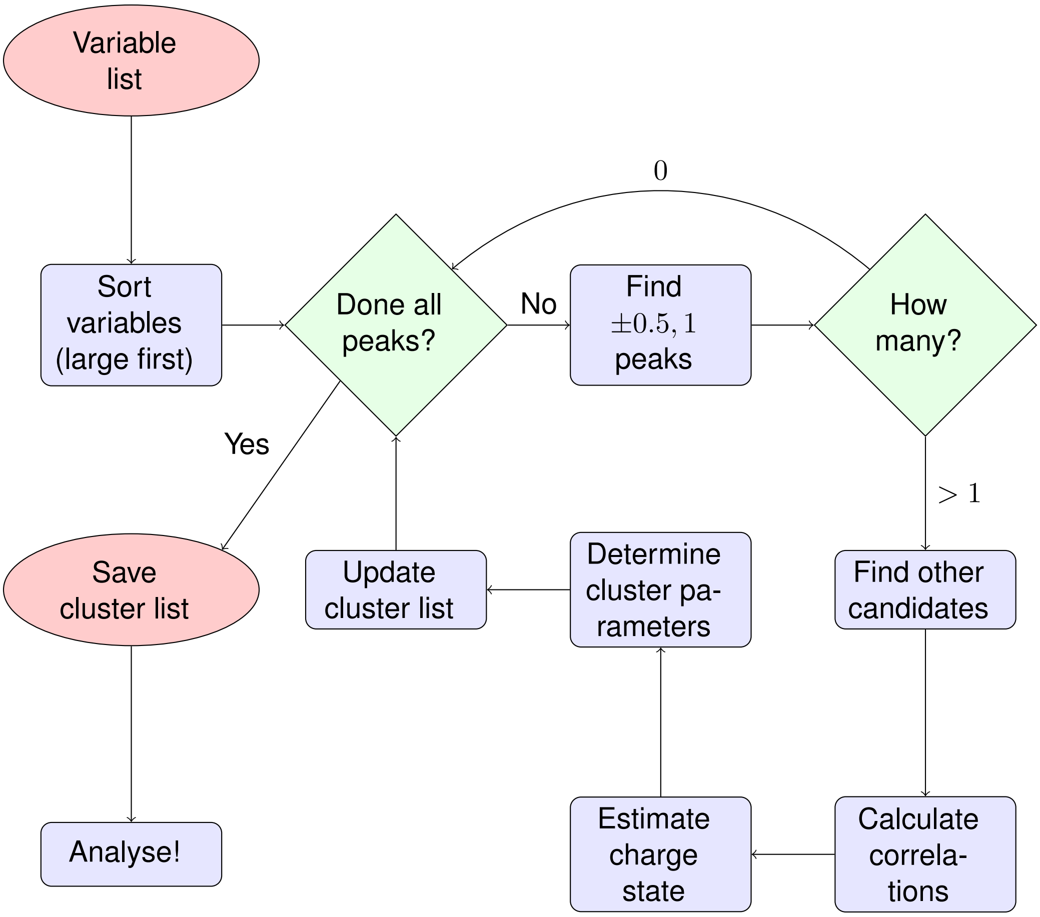


FIGURE A.6: A flowchart detailing the procedure used to group peaks into isotopic clusters. Variables that are highly correlated are likely to form part of the same isotopic cluster, and is therefore a good metric to use in addition to the distance. The charge state can be best estimated by looking at the clusters with peak increments of $\frac{1}{2}$ and 1 which correspond to *m/z* values of *m/2* and *m/1*, respectively.

Appendix B

Assignment Tables

Errors using inadequate data are much less than those using no data at all.

Charles Babbage

TABLE B.1: Potential candidate assignments for variable 16: t / min = 8.5684, *m/z* = 90.0552, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
352	M+ACN+2H	137.0692	-16.84	7.06	2/5	C7H9N2O	1-Methylnicotinamide
212	M+2H+Na	245.1618	-3.82	9.31	2/12	C12H23NO4	2-Methylbutyroylcarnitine
345	M+2H+Na	245.1618	-3.82	9.31	3/8	C12H23NO4	Isovalerylcarnitine
265	M+H+NH4	161.0692	2.33	4.06	3/9	C6H11NO4	Aminoadipic-acid
37	M+H	89.0479	2.39	0.56	0/2	C3H7NO2	Beta-Alanine
99	M+H	89.0479	2.39	0.56	2/3	C3H7NO2	L-Alanine
170	M+H	89.0479	2.39	0.56	0/2	C3H7NO2	Sarcosine
537	M+H	89.0479	2.39	0.56	1/2	C3H7NO2	D-Alanine

TABLE B.2: Potential candidate assignments for variable 13: t / min = 6.427, *m/z* = 118.0863, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
438	M+3ACN+2H	111.0784	-11.15	7.96	0/4	C5H9N3	Histamine
28	M+H	117.0790	0.20	0.72	1/2	C5H11NO2	Betaine
445	M+H	117.0790	0.20	0.72	1/4	C5H11NO2	L-Valine
712	M+H	117.0790	0.20	0.72	1/2	C5H11NO2	N-Methyl-a-aminoisobutyric-acid
802	M+H	117.0790	0.20	0.72	1/4	C5H11NO2	5-Aminopentanoic-acid
694	M+CH3OH+H	85.0528	0.44	0.72	1/3	C4H7NO	2-Pyrrolidinone
264	M+NH4	100.0525	0.73	0.72	0/3	C5H8O2	Senecioic-acid
571	M+NH4	100.0525	0.73	0.72	0/3	C5H8O2	Tiglic-acid
624	M+NH4	100.0525	0.73	0.72	0/4	C5H8O2	2-Ethylacrylic-acid
636	M+ACN+H	76.0525	0.96	0.72	1/4	C3H8O2	Propylene-glycol

Table B.3: Potential candidate assignments for variable 139: t / min = 13.8676, *m/z* = 487.1654, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
865	M+DMSO+H	408.1442	-2.41	6.94	1/13	C20H25ClN2O5	Amlodipine

Table B.4: Potential candidate assignments for variable 133: t / min = 13.8814, *m/z* = 689.2114, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
658	2M+H	344.1021	-17.33	18.73	0/9	C22H17ClN2	Clotrimazole
391	M+Na	666.2222	0.56	3.16	8/10	C24H42O21	Glycogen
536	M+Na	666.2222	0.56	3.16	10/18	C24H42O21	Maltotetraose
824	M+Na	666.2222	0.56	3.16	7/23	C24H42O21	Stachyose
791	M+DMSO+H	610.1902	0.78	8.26	2/22	C28H34O15	Hesperidin

Table B.5: Potential candidate assignments for variable 169: t / min = 14.0049, *m/z* = 684.2556, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
391	M+NH4	666.2218	-0.04	5.59	8/10	C24H42O21	Glycogen
536	M+NH4	666.2218	-0.04	5.59	10/18	C24H42O21	Maltotetraose
824	M+NH4	666.2218	-0.04	5.59	7/23	C24H42O21	Stachyose

TABLE B.6: Potential candidate assignments for variable 67: t / min = 11.6193, *m/z* = 503.1613, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
47	2M-H	252.0843	-6.12	0.19	1/10	C10H12N4O4	Deoxyinosine
531	M-H	504.1686	-0.80	2.48	10/21	C18H32O16	Maltotriose
784	M-H	504.1686	-0.80	2.48	7/19	C18H32O16	Raffinose

TABLE B.7: Potential candidate assignments for variable 28: t / min = 11.6162, *m/z* = 549.1659, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
47	2M+FA-H	252.0838	-8.05	3.78	1/10	C10H12N4O4	Deoxyinosine
531	MFA-H	504.1677	-2.72	4.18	10/21	C18H32O16	Maltotriose
784	MFA-H	504.1677	-2.72	4.18	7/19	C18H32O16	Raffinose

TABLE B.8: Potential candidate assignments for variable 60: t / min = 11.6165, *m/z* = 617.1545, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
531	MTFA-H	504.1689	-0.32	2.28	10/21	C18H32O16	Maltotriose
784	MTFA-H	504.1689	-0.32	2.28	7/19	C18H32O16	Raffinose

TABLE B.9: Potential candidate assignments for variable 23: t / min = 5.241, *m/z* = 215.0322, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
804	2M-H	108.0197	-12.89	26.23	0/1	C6H4O2	Quinone
622	MCl	180.0628	-10.89	24.81	0/4	C7H8N4O2	Paraxanthine
640	MCl	180.0628	-10.89	24.81	0/3	C7H8N4O2	Theophylline
757	MCl	180.0628	-10.89	24.81	0/3	C7H8N4O2	Theobromine
72	MCl	180.0628	-3.43	22.54	10/10	C6H12O6	D-Glucose
87	MCl	180.0628	-3.43	22.54	7/11	C6H12O6	D-Galactose
105	MCl	180.0628	-3.43	22.54	5/8	C6H12O6	D-Mannose
131	MCl	180.0628	-3.43	22.54	1/4	C6H12O6	Myoinositol
201	MCl	180.0628	-3.43	22.54	2/6	C6H12O6	3-Deoxyarabinohexonic-acid
326	MCl	180.0628	-3.43	22.54	2/7	C6H12O6	D-Fructose
510	MCl	180.0628	-3.43	22.54	0/7	C6H12O6	Allose
533	MCl	180.0628	-3.43	22.54	3/7	C6H12O6	L-Sorbose
799	MCl	180.0628	-3.43	22.54	9/13	C6H12O6	Alpha-D-Glucose
815	MCl	180.0628	-3.43	22.54	4/7	C6H12O6	D-Tagatose
480	MNa-2H	194.0576	-1.75	25.43	0/6	C10H10O4	trans-Ferulic-acid
481	MNa-2H	194.0576	-1.75	25.43	0/6	C10H10O4	Isoferulic-acid
141	MHAc-H	156.0183	7.45	23.13	0/1	C5H4N2O4	Orotic-acid

TABLE B.10: Potential candidate assignments for variable 11: t / min = 6.7124, *m/z* = 130.0862, z = 1.

JDB	Adduct	M	ΔM / ppm	Iso	HSQC	Formula	Name
46	M+H	129.0789	-0.35	0.48	6/6	C6H11NO2	Pipecolic-acid
363	M+H	129.0789	-0.35	0.48	7/9	C6H11NO2	L-Pipecolic-acid
26	M+ACN+H	88.0524	0.05	0.48	0/4	C4H8O2	Butyric-acid
631	M+ACN+H	88.0524	0.05	0.48	0/2	C4H8O2	Isobutyric-acid
788	M+ACN+H	88.0524	0.05	0.48	0/3	C4H8O2	Acetoin
462	M+2Na	214.1941	3.66	6.89	4/7	C13H26O2	Tridecanoic-acid

Glossary

I'm so clever that sometimes I don't understand a single word of what I'm saying.

Oscar Wilde

AC	Alternating Current
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
APCI	Atmospheric Pressure Chemical Ionisation
ASCA	ANOVA Simultaneous Components Analysis
BMRB	BioMagRes Bank
CANDECOMP	Canonical Decomposition
CCA	Canonical Correlation Analysis
CID	Collision Induced Dissociation
COSY	Correlation Spectroscopy
COW	Correlation Optimised Warping
CP	Canonical Pair
CPMG	Carr-Purcell-Meiboom-Gill
CTD	Correlated Trace Denoising
DC	Direct Current
DEPT	Distortionless Enhancement through Polarisation Transfer
DNP	Dynamic Nuclear Polarisation
DoE	Design of Experiments
DOSY	Diffusion Ordered Spectroscopy
DQD	Digital Quadrature Detection
DTW	Dynamic Time Warping
ESI	Electrospray Ionisation
EU	European Union
FID	Free Induction Decay
FS	Feature Selection
FT	Fourier Transform
FWHM	Full Width Half Maximum
GA	Genetic Algorithms
GARP	Globally Optimised Alternating Phase Rectangular Pulse
GC	Gas Chromatography
GP	Genetic Programming
HILIC	Hydrophilic Interaction Chromatography
HMBC	Heteronuclear Multiple Bond Correlation

HMDB	Human Metabolome Database
HMQC	Heteronuclear Multiple Quantum Coherence
HPLC	High Performance Liquid Chromatography
HR	High Resolution
HSQC	Heteronuclear Single Quantum Coherence
ICA	Independent Components Analysis
ICR	Ion Cyclotron Resonance
INADEQUATE	Incredible Natural Abundance Double Quantum Transfer Experiment
IUPAC	International Union of Pure and Applied Chemistry
JDB	James Database
JRES	*J*-Resolved Spectoscopy
KM	Kernel Methods
LC	Liquid Chromatography
LDA	Linear Discriminant Analysis
LIT	Linear Ion Trap
LLE	Liquid-Liquid Extraction
LOESS	Locally Estimated Smoothing
LOO	Leave One Out
SPE	Solid Phase Extraction
m/z	Mass to Charge Ratio
MANOVA	Multivariate Analysis of Variance
MAS	Magic Angle Spinning
MIAMET	Minimum Information About a Metabolomics Experiment
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MSI	Metabolomics Standard Initiative
MSPD	Matrix-Solid Phase Dispersion
MVA	Multivariate Analysis
(net)CDF	(Network) Common Data Form
NIPALS	Non-Linear Iterative Partial Least Squares
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy

O-PLS	Orthogonal Partial Least Squares
OVAT	One Variable At a Time
PARAFAC	Parallel Factor Analysis
PARAFASCA	Parallel Factor Analysis-ANOVA Simultaneous Components Analysis
PC	Principal Component
PCA	Principal Components Analysis
PCR	Principal Components Regression
PFG	Pulsed-Field Gradient
PGC	Porous Graphitic Carbon
PLE	Pressurised Liquid Extraction
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
ppm	Parts Per Million
QC	Quality Control
QDA	Quadratic Discriminant Analysis
QIT	Quadrupolar Ion Trap
QIT	Quadrupole Ion Trap
QqQ	Triple Quadrupole
QqTOF	Quadrupole Time of Flight
QuECHERS	Quick East Cheap Effective Rugged Safe
RANSY	Ratio Analysis Spectroscopy
RDA	Regularised Discriminant Analysis
ROI	Region of Interest
RP	Reversed Phase
rpm	Revolutions Per Minute
RSM	Response Surface Methodology
RSPA	Recursive Segment-Wise Peak Alignment
SNR	Signal to Noise Ratio
SOM	Self-Organising Map
SOP	Standard Operating Procedure
SQUID	Super-Conducting Quantum Interference Device
SRM	Single Reaction Monitoring

STOCSY	Statistical Total Correlation Spectroscopy
SVM	Support Vector Machine
TIC	Total Ion Current
TOCSY	Total Correlation Spectroscopy
TOF	Time of Flight
TPPI	Time Proportional Phase Incrementation
TR	Tenderometer
TSP	Trimethylsilyl Propanoic Acid
U(H)PLC	Ultra (High) Performance Liquid Chromatography
ULDA	Uncorrelated Linear Discriminant Analysis
VIP	Variable Importance of Projection
WATERGATE	Water Supression through Gradient Tailored Excitation

References

Most people are other people. Their thoughts are someone else's opinions, their lives a mimicry, their passions a quotation.

Oscar Wilde

[1] Fiehn, O. (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology*, **48**, 155.

[2] Brindle, J. T., et al. (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature Medicine*, **8**, 1439.

[3] Kirschenlohr, H. L., Griffin, J. L., Clarke, S. C., Rhydwen, R., Grace, A. A., Schofield, P. M., Brindle, K. M., and Metcalfe, J. C. (2006) Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nature Medicine*, **12**, 705.

[4] Coen, M., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2008) Nmr-based metabolic profiling and metabonomic approaches to problems in molecular toxicology. *Chemical Research in Toxicology*, **21**, 9.

[5] Beger, R. D., Sun, J., and Schnackenberg, L. K. (2010) Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity. *Toxicology and Applied Pharmacology*, **243**, 154.

[6] Robertson, D. G., Watkins, P. B., and Reily, M. D. (2011) Metabolomics in toxicology: preclinical and clinical applications. *Toxicological Sciences*, **120**, S146.

[7] Wishart, D. S. (2008) Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, **19**, 482.

[8] Donarski, J. A., Jones, S. A., and Charlton, A. J. (2008) Application of cryoprobe ^{1}H nuclear magnetic resonance spectroscopy and multivariate analysis for the verification of Corsican honey. *Journal of Agricultural and Food Chemistry*, **56**, 5451.

[9] Charlton, A. J., Wrobel, M. S., Stanimirova, I., Daszykowski, M., Grundy, H. H., and Walczak, B. (2010) Multivariate discrimination of wines with respect to their grape varieties and vintages. *European Food Research and Technology*, **231**, 733.

[10] Ali, K., Maltese, F., Toepfer, R., Choi, Y. H., and Verpoorte, R. (2011) Metabolic characterization of Palatinate German white wines according to sensory attributes, varieties, and vintages using NMR spectroscopy and multivariate data analyses. *Journal of Biomolecular NMR*, **49**, 255.

[11] Charlton, A. J., et al. (2008) Responses of the pea (*Pisum sativum L.*) leaf metabolome to drought stress assessed by nuclear magnetic resonance spectroscopy. *Metabolomics*, **4**, 312.

[12] Viant, M. and Sommer, U. (2013) Mass spectrometry based environmental metabolomics: a primer and review. *Metabolomics*, **9**, 144.

[13] Waybright, T. J., Van, Q. N., Muschik, G. M., Conrads, T. P., Veenstra, T. D., and Issaq, H. J. (2006) Lc-ms in metabonomics: Optimization of experimental conditions for the analysis of metabolites in human urine. *Journal of Liquid Chromatography & Related Technologies*, **29**, 2475.

[14] Cubbon, S., Bradbury, T., Wilson, J., and Thomas-Oates, J. (2007) Hydrophilic interaction chromatography for mass spectrometric metabonomic studies of urine. *Analytical Chemistry*, **79**, 8911.

[15] Hall, D. and Llinas, J. (1997) An introduction to multisensor data fusion. *Proceedings of the IEEE*, **85**, 6.

[16] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417.

[17] Wold, S., Sjostrom, M., and Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**, 109.

[18] Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998) Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, **12**, 301.

[19] Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J. C., and Jellema, R. H. (2005) Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, **77**, 6729.

[20] Forshed, J., Idborg, H., and Jacobsson, S. P. (2007) Evaluation of different techniques for data fusion of LC/MS and ^{1}H NMR. *Chemometrics and Intelligent Laboratory Systems*, **85**, 102.

[21] Blaise, B. J., Giacomotto, J., Triba, M. N., Toulhoat, P., Piotto, M., Emsley, L., Segalat, L., Dumas, M. E., and Elena, B. (2009) Metabolic profiling strategy of *Caenorhabditis elegans* by whole-organism nuclear magnetic resonance. *Journal of Proteome Research*, **8**, 2542.

[22] Vera, L., Aceña, L., Guasch, J., Boqué, R., Mestres, M., and Busto, O. (2011) Discrimination and sensory description of beers through data fusion. *Talanta*, **87**, 136.

[23] Biais, B., et al. (2009) 1H NMR, GC-EI-TOFMS, and data set correlation for fruit metabolomics: application to spatial metabolite analysis in melon. *Analytical Chemistry*, **81**, 2884.

[24] Ramos, P. M. and Ruisánchez, I. (2006) Data fusion and dual-domain classification analysis of pigments studied in works of art. *Analytica Chimica Acta*, **558**, 274.

[25] Koza, J. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT press, Cambridge, MA.

[26] Taylor, J., Goodacre, R., Wade, W. G., Rowland, J. J., and Kell, D. B. (1998) The deconvolution of pyrolysis mass spectra using genetic programming: application to the identification of some *Eubacterium* species. *FEMS Microbiology Letters*, **160**, 237.

[27] Davis, R. A., Charlton, A. J., Oehlschlager, S., and Wilson, J. C. (2006) Novel feature selection method for genetic programming using metabolomic ^{1}H NMR data. *Chemometrics and Intelligent Laboratory Systems*, **81**, 50.

[28] Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321.

[29] Larson, M., Capobianco, M., and Hanson, H. (2000) Relationship between beach profiles and waves at Duck, North Carolina, determined by canonical correlation analysis. *Marine Geology*, **163**, 275.

[30] Ekanade, O. and Orimoogunje, O. O. I. (2012) Application of canonical correlation for soil-vegatation interrelationship in the cocoa belt of south western Nigeria. *Resources and Environment*, **2**, 87.

[31] Naylor, M. G., Lin, X., Weiss, S. T., Raby, B. A., and Lange, C. (2010) Using canonical correlation analysis to discover genetic regulatory variants. *PLoS ONE*, **5**, e10395.

[32] Doeswijk, T., Hageman, J., Westerhuis, J., Tikunov, Y., Bovy, A., and van Eeuwijk, F. (2011) Canonical correlation analysis of multiple sensory directed metabolomics data blocks reveals corresponding parts between data blocks. *Chemometrics and Intelligent Laboratory Systems*, **107**, 371.

[33] Crockford, D. J., Holmes, E., Lindon, J. C., Plumb, R. S., Zirah, S., Bruce, S. J., Rainville, P., Stumpf, C. L., and Nicholson, J. K. (2006) Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Analytical Chemistry*, **78**, 363.

[34] Moco, S., Forshed, J., De Vos, R. C. H., Bino, R. J., and Vervoort, J. (2008) Intra- and inter-metabolite correlation spectroscopy of tomato metabolomics data obtained by liquid chromatography-mass spectrometry and nuclear magnetic resonance. *Metabolomics*, **4**, 202.

[35] Kind, T. and Fiehn, O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

[36] McKenzie, J. S., Charlton, A. J., Donarski, J. A., MacNicoll, A. D., and Wilson, J. C. (2010) Peak fitting in 2D ^{1}H-^{13}C HSQC NMR spectra for metabolomic studies. *Metabolomics*, **6**, 574.

[37] Dunn, W., et al. (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, **9**, S44.

[38] Charlton, A. J. (2009) High resolution NMR analysis of complex mixtures. Guðjónsdóttir, M., Belton, P. S., and Webb, G. A. (eds.), *Magnetic Resonance in Food Science: Challenges in a Changing World*, pp. 3–11, RSC, London.

[39] Trefi, S., Gilard, V., Balayssac, S., Malet-Martino, M., and Martino, R. (2009) The usefulness of 2D DOSY and 3d DOSY-COSY ^{1}H NMR for mixture analysis: application to genuine and fake formulations of sildenafil (viagra). *Magnetic Resonance in Chemistry*, **47**, S163.

[40] McDowell, R. W. and Stewart, I. (2006) The phosphorus composition of contrasting soils in pastoral, native and forest management in Otago, New Zealand: sequential extraction and ^{31}P NMR. *Geoderma*, **130**, 176.

[41] Simpson, A. J., McNally, D. J., and Simpson, M. J. (2011) NMR spectroscopy in environmental research: from molecular interactions to global processes. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **58**, 97.

[42] Charlton, A. J., Farrington, W. H., and Brereton, P. (2002) Application of ^{1}H NMR and multivariate statistics for screening complex mixtures: quality control and authenticity of instant coffee. *Journal of Agricultural and Food Chemistry*, **50**, 3098.

[43] Charlton, A. J., Robb, P., Donarski, J. A., and Godward, J. (2008) Non-targeted detection of chemical contamination in carbonated soft drinks using NMR spectroscopy, variable selection and chemometrics. *Analytica Chimica Acta*, **618**, 196.

[44] Charlton, A. J., Donarski, J. A., Jones, S. A., May, B. D., and Thompson, K. C. (2006) The development of cryoprobe nuclear magnetic resonance spectroscopy for the rapid detection of organic contaminants in potable water. *Journal of Environmental Monitoring*, **8**, 1106.

[45] Wenning, R. J. and Erickson, G. A. (1994) Interpretation and analysis of complex environmental data using chemometric methods. *TrAC - Trends in Analytical Chemistry*, **13**, 446.

[46] Bundy, J. G., Davey, M. P., and Viant, M. R. (2009) Environmental metabolomics: a critical review and future perspectives. *Metabolomics*, **5**, 3.

[47] Viant, M. R., Rosenblum, E. S., and Tjeerdema, R. S. (2003) NMR-based metabolomics: a powerful approach for characterising the effects of environmental stressors on organism health. *Environmental Science & Technology*, **37**, 4982.

[48] del Carmen Alvarez, M., Donarski, J., Elliott, M., and Charlton, A. J. (2010) Evaluation of extraction methods for use with NMR-based metabolomics in the marine polychaete ragworm, *Hediste diversicolor*. *Metabolomics*, **6**, 541.

[49] Koskela, H. (2010) Use of NMR techniques for toxic organophosphorus compound profiling. *Journal of Chromatography B*, **878**, 1365.

[50] Emsley, J. W. and Feeney, J. (2007) Forty years of progress in nuclear magnetic resonance spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **50**, 179.

[51] Fiehn, O., et al. (2007) The metabolomics standards initiative (MSI). *Metabolomics*, **3**, 175.

[52] Holmes, E., et al. (1994) Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Analytical Biochemistry*, **220**, 284.

[53] Spraul, M., et al. (1994) Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of Pharmaceutical Biomedicine*, **12**, 1215.

[54] Davis, R. A., Charlton, A. J., Godward, J., Jones, S. A., Harrison, M., and Wilson, J. C. (2007) Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, **85**, 144.

[55] Anderson, P. E., Reo, N. V., DelRaso, N. J., Doom, T. E., and Raymer, M. L. (2008) Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics*, **4**, 261.

[56] De Meyer, T., Sinnaeve, D., Van Gasse, B., Tsiporkova, E., Rietzschel, E. R., De Buyzere, M. L., Gillebert, T. C., Bekaert, S., Martins, J. C., and Van Criekinge, W. (2008) NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry*, **80**, 3783.

[57] Piotto, M., Saudek, V., and Sklenar, V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *Journal of Biomolecular NMR*, **2**, 661.

[58] Ogg, R. J., Kingsley, P. B., and Taylor, J. S. (1994) WET, a T-1 insensitive and B-1 insensitive water-suppression method for *in vivo* localized ^{1}H NMR spectroscopy. *Journal of Magnetic Resonance B*, **104**, 1.

[59] Meiboom, S. and Gill, D. (1958) Modified spin-echo method for measuring nucelar relaxation times. *Review of Scientific Instruments*, **29**, 688.

[60] Kessler, H., Mronga, S., and Gemmecker, G. (1991) Multidimensional NMR experiments using selective pulses. *Magnetic Resonance in Chemistry*, **29**, 527.

[61] Adams, R. W., Aguilar, J. A., Atkinson, K. D., Cowley, M. J., Elliott, P. I. P., Duckett, S. B., Green, G. G. R., Khazal, I. G., Lopez-Serrano, J., and Williamson, D. C. (2009) Reversible interactions with para-hydrogen enhance NMR sensitivity by polarization transfer. *Science*, **323**, 1708.

[62] Aguilar, J. A., Elliott, P. I. P., Lopez-Serrano, J., Adams, R. W., and Duckett, S. B. (2007) Only para-hydrogen spectroscopy (OPSY), a technique for the selective observation of para-hydrogen enhanced NMR signals. *Chemical Communications*, p. 1183.

[63] Duckett, S. B. and Wood, N. J. (2008) Parahydrogen-based NMR methods as a mechanistic probe in inorganic chemistry. *Coordination Chemistry Reviews*, **252**, 2278.

[64] Battiste, J. L., Jing, N. Y., and Newmark, P. A. (2004) 2D ^{19}F/^{19}F NOESY for the assignment of NMR spectra of fluorochemicals. *Journal of Fluorine Chemistry*, **125**, 1331.

[65] Cobb, S. L. and Murphy, C. D. (2009) ^{19}F NMR applications in chemical biology. *Journal of Fluorine Chemistry*, **130**, 132.

[66] Duckett, C. J., Lindon, J. C., Walker, H., Abou-Shakra, F. R., Wilson, I. D., and Nicholson, J. K. (2006) Metabolism of 3-chloro-4-fluoroaniline in rat using [^{14}C]-radiolabelling, ^{19}F NMR spectroscopy, HPLC-MS/MS, HPLC-ICPMS and HPLC-NMR. *Xenobiotica*, **36**, 59.

[67] Duckett, C. J., Wilson, I. D., Douce, D. S., Walker, H. J., Abou-Shakra, F. R., Lindon, J. C., and Nicholson, J. K. (2007) Metabolism of 2-fluoro-4-iodoaniline in earthworm eisenia veneta using ^{19}F NMR spectroscopy, HPLC-MS, and HPLC-ICPMS (I-127). *Xenobiotica*, **37**, 1378.

[68] Majumdar, A., Sun, Y., Shah, M., and Meyers, C. L. F. (2010) Versatile ^{1}H-^{31}P-^{31}P COSY 2D NMR techniques for the characterization of polyphosphorylated small molecules. *Journal of Organic Chemistry*, **75**, 3214.

[69] Solivera, J., Cerdan, S., Pascual, J. M., Barrios, L., and Roda, J. M. (2009) Assessment of ^{31}P NMR analysis of phospholipid profiles for potential differential diagnosis of human cerebral tumors. *NMR in Biomedicine*, **22**, 663.

[70] Hatzakis, E., Agiomyrgianaki, A., and Dais, P. (2010) Detection and quantification of free glycerol in virgin olive oil by ^{31}P NMR spectroscopy. *Journal of the American Oil Chemists Society*, **87**, 29.

[71] Hatzakis, E., Dagounakis, G., Agiomyrgianaki, A., and Dais, P. (2010) A facile NMR method for the quantification of total, free and esterified sterols in virgin olive oil. *Food Chemistry*, **122**, 346.

[72] Turner, B. L., Condron, L. M., Richardson, S. J., Peltzer, D. A., and Allison, V. J. (2007) Soil organic phosphorus transformations during pedogenesis. *Ecosystems*, **10**, 1166.

[73] Liu, J. Y., Wang, H., Yang, H. J., Ma, Y. J., and Cai, O. C. (2009) Detection of phosphorus species in sediments of artificial landscape lakes in China by fractionation and phosphorus-31 nuclear magnetic resonance spectroscopy. *Environmental Pollution*, **157**, 49.

[74] Ahlgren, J., Reitzel, K., De Brabandere, H., Gogoll, A., and Rydin, E. (2011) Release of organic P forms from lake sediments. *Water Research*, **45**, 565.

[75] Morris, K. F. and Johnson Jr, C. S. (1992) Diffusion-ordered 2-dimensional nuclear magnetic resonance spectroscopy. *Journal of the American Chemical Society*, **114**, 3139.

[76] Bodenhausen, G. and Ruben, D. J. (1980) Natural abundance ^{15}N NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, **69**, 185.

[77] Bax, A., Aszalos, A., Dinya, Z., and Sudo, K. (1986) Structure elucidation of the antibiotic desertomycin through the use of new two-dimensional NMR techniques. *Journal of the American Chemical Society*, **108**, 8056.

[78] Summers, M. F., Marzilli, L. G., and Bax, A. (1986) Complete ^{1}H and ^{13}C assignments of coenzyme B12 through the use of new two-dimensional NMR experiments. *Journal of the American Chemical Society*, **108**, 4285.

[79] Mishkovsky, M. and Frydman, L. (2008) Progress in hyperpolarized ultrafast 2D NMR spectroscopy. *Chemphyschem*, **9**, 2340.

[80] Xia, J. G., Bjorndahl, T. C., Tang, P., and Wishart, D. S. (2008) Metabominer - semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507.

[81] Poulding, S., Charlton, A. J., Donarski, J., and Wilson, J. C. (2007) Removal of t1 noise from metabolomic 2D ^{1}H-^{13}C HSQC NMR spectra by correlated trace denoising. *Journal of Magnetic Resonance*, **189**, 190.

[82] Shrot, Y. and Frydman, L. (2008) Single-scan 2D DOSY NMR spectroscopy. *Journal of Magnetic Resonance*, **195**, 226.

[83] Giraudeau, P. and Akoka, S. (2007) A new detection scheme for ultrafast 2D J-resolved spectroscopy. *Journal of Magnetic Resonance*, **186**, 352.

[84] Giraudeau, P., Remaud, G. S., and Akoka, S. (2009) Evaluation of ultrafast 2D NMR for quantitative analysis. *Analytical Chemistry*, **81**, 479.

[85] Giraudeau, P., Shrot, Y., and Frydman, L. (2009) Multiple ultrafast, broadband 2D NMR spectra of hyperpolarized natural products. *Journal of the American Chemical Society*, **131**, 13902.

[86] Giraudeau, P. and Akoka, S. (2010) A new gradient-controlled method for improving the spectral width of ultrafast 2D NMR experiments. *Journal of Magnetic Resonance*, **205**, 171.

[87] Gal, M., Kern, T., Schanda, P., Frydman, L., and Brutscher, B. (2009) An improved ultrafast 2D NMR experiment: towards atom-resolved real-time studies of protein kinetics at multi-Hz rates. *Journal of Biomolecular NMR*, **43**, 1.

[88] Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nystrom, A., Pettersen, J., and Bergman, R. (1998) Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*, **42**, 3.

[89] Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., and Escaleira, L. A. (2008) Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*, **76**, 965.

[90] Petersen, L., Minkkinen, P., and Esbensen, K. H. (2005) Representative sampling for reliable data analysis: theory of sampling. *Chemometrics and Intelligent Laboratory Systems*, **77**, 261.

[91] Teahan, O., Gamble, S., Holmes, E., Waxman, J., Nicholson, J. K., Bevan, C., and Keun, H. C. (2006) Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry*, **78**, 4307.

[92] Pawliszyn, J. (2003) Sample preparation: quo vadis? *Analytical Chemistry*, **75**, 2543.

[93] Alvarez-Sanchez, B., Priego-Capote, F., and Luque de Castro, M. D. (2010) Metabolomics analysis I. selection of biological samples and practical aspects preceding sample preparation. *TrAC - Trends in Analytical Chemistry*, **29**, 111.

[94] Alvarez-Sanchez, B., Priego-Capote, F., and Luque de Castro, M. D. (2010) Metabolomics analysis II. preparation of biological samples prior to detection. *TrAC - Trends in Analytical Chemistry*, **29**, 120.

[95] Dunn, W. B. and Ellis, D. I. (2005) Metabolomics: current analytical platforms and methodologies. *TrAC - Trends in Analytical Chemistry*, **24**, 285.

[96] Kruger, N. J., Troncoso-Ponce, M. A., and Ratcliffe, R. G. (2008) ^{1}H NMR metabolite fingerprinting and metabolomic analysis of perchloric acid extracts from plant tissues. *Nature Protocols*, **3**, 1001.

[97] Kaiser, K. A., Barding, G. A., and Larive, C. K. (2009) A comparison of metabolite extraction strategies for ^{1}H NMR-based metabolic profiling using mature leaf tissue from the model plant *Arabidopsis thaliana*. *Magnetic Resonance in Chemistry*, **47**, S147.

[98] Lauridsen, M., Hansen, S. H., Jaroszewski, J. W., and Cornett, C. (2007) Human urine as test material in ^{1}H NMR-based metabonomics: recommendations for sample preparation and storage. *Analytical Chemistry*, **79**, 1181.

[99] Maher, A. D., Zirah, S. F. M., Holmes, E., and Nicholson, J. K. (2007) Experimental and analytical variation in human urine in ^{1}H NMR spectroscopy-based metabolic phenotyping studies. *Analytical Chemistry*, **79**, 5204.

[100] Dumas, M. E., et al. (2006) Assessment of analytical reproducibility of ^{1}H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry*, **78**, 2199.

[101] Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J. G., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, **2**, 2692.

[102] Wu, J. F., An, Y. P., Yao, J. W., Wang, Y. L., and Tang, H. R. (2010) An optimised sample preparation method for NMR-based faecal metabonomic analysis. *Analyst*, **135**, 1023.

[103] Jacobs, D. M., Deltimple, N., van Velzen, E., van Dorsten, F. A., Bingham, M., Vaughan, E. E., and van Duynhoven, J. (2008) ^{1}H NMR metabolite profiling of feces as a tool to assess the impact of nutrition on the human microbiome. *NMR in Biomedicine*, **21**, 615.

[104] Sukumaran, D. K., Garcia, E., Hua, J., Tabaczynski, W., Odunsi, K., Andrews, C., and Szyperski, T. (2009) Standard operating procedure for metabonomics studies of blood serum and plasma samples using a ^{1}H NMR micro-flow probe. *Magnetic Resonance in Chemistry*, **47**, S81.

[105] Graça, G., Duarte, I. F., Goodfellow, B. J., Barros, A. S., Carreira, I. M., Couceiro, A. B., Spraul, M., and Gil, A. M. (2007) Potential of NMR spectroscopy for the study of human amniotic fluid. *Analytical Chemistry*, **79**, 8367.

[106] Winder, C. L., Dunn, W. B., Schuler, S., Broadhurst, D., Jarvis, R., Stephens, G. M., and Goodacre, R. (2008) Global metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. *Analytical Chemistry*, **80**, 2939.

[107] Sacco, D., Brescia, M. A., Buccolieri, A., and Jambrenghi, A. C. (2005) Geographical origin and breed discrimination of Apulian lamb meat samples by means of analytical and spectroscopic determinations. *Meat Science*, **71**, 542.

[108] Waters, N. J., Garrod, S., Farrant, R. D., Haselden, J. N., Connor, S. C., Connelly, J., Lindon, J. C., Holmes, E., and Nicholson, J. K. (2000) High-resolution magic angle spinning ^{1}H NMR spectroscopy of intact liver and kidney: optimization of sample preparation procedures and biochemical stability of tissue during spectral acquisition. *Analytical Biochemistry*, **282**, 16.

[109] Donarski, J. A., Jones, S. A., Harrison, M., Driffield, M., and Charlton, A. J. (2010) Identification of botanical biomarkers found in Corsican honey. *Food Chemistry*, **118**, 987.

[110] Schievano, E., Peggion, E., and Mammi, S. (2010) ^{1}H nuclear magnetic resonance spectra of chloroform extracts of honey for chemometric determination of its botanical origin. *Journal of Agricultural and Food Chemistry*, **58**, 57.

[111] Alonso-Salces, R. M., Moreno-Rojas, J. M., Holland, M. V., Reniero, F., Guillou, C., and Heberger, K. (2010) Virgin olive oil authentication by multivariate analyses of ^{1}H NMR fingerprints and δ ^{13}C and δ ^{2}H data. *Journal of Agricultural and Food Chemistry*, **58**, 5586.

[112] Mannina, L., D'Imperio, M., Capitani, D., Rezzi, S., Guillou, C., Mavromoustakos, T., Vilchez, M. D. M., Fernandez, A. H., Thomas, F., and Aparicio, R. (2009) ^{1}H NMR-based protocol for the detection of adulterations of refined olive oil with refined hazelnut oil. *Journal of Agricultural and Food Chemistry*, **57**, 11550.

[113] Agiomyrgianaki, A., Petrakis, P. V., and Dais, P. (2010) Detection of refined olive oil adulteration with refined hazelnut oil by employing NMR spectroscopy and multivariate statistical analysis. *Talanta*, **80**, 2165.

[114] Wishart, D. S., et al. (2007) HMDB: the human metabolome database. *Nucleic Acids Research*, **35**, D521.

[115] Wishart, D. S., et al. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, **37**, D603.

[116] Ulrich, E. L., et al. (2008) BioMagResBank. *Nucleic Acids Research*, **36**, D402.

[117] Bertram, H. C., Malmendal, A., Petersen, B. O., Madsen, J. C., Pedersen, H., Nielsen, N. C., Hoppe, C., Molgaard, C., Michaelsen, K. F., and Duus, J. O. (2007) Effect of magnetic field strength on NMR-based metabonomic human urine data: comparative study of 250, 400, 500, and 800 MHz. *Analytical Chemistry*, **79**, 7110.

[118] Viant, M. R., et al. (2009) International NMR-based environmental metabolomics intercomparison exercise. *Environmental Science & Technology*, **43**, 219.

[119] Ward, J. L., et al. (2010) An inter-laboratory comparison demonstrates that ^{1}H NMR metabolite fingerprinting is a robust technique for collaborative plant metabolomic data collection. *Metabolomics*, **6**, 263.

[120] Potts, B. C. M., Deese, A. J., Stevens, G. J., Reily, M. D., Robertson, D. G., and Theiss, J. (2001) NMR of biofluids and pattern recognition: assessing the impact of NMR parameters on the principal component analysis of urine from rat and mouse. *Journal of Pharmaceutical Biomedicine*, **26**, 463.

[121] Aue, W. P., Bartholdi, E., and Ernst, R. R. (1976) Two-dimensional spectroscopy - application to nuclear magnetic resonance. *Journal of Chemical Physics*, **64**, 2229.

[122] Braunschweiler, L. and Ernst, R. R. (1983) Coherence transfer by isotropic mixing - application to proton correlation spectroscopy. *Journal of Magnetic Resonance*, **53**, 521.

[123] Aue, W. P., Karhan, J., and Ernst, R. R. (1976) Homonuclear broad-band decoupling and 2-dimensional J-resolved NMR spectroscopy. *Journal of Chemical Physics*, **64**, 4226.

[124] Viant, M. R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, **310**, 943.

[125] Wang, Y. L., Bollard, M. E., Keun, H., Antti, H., Beckonert, O., Ebbels, T. M., Lindon, J. C., Holmes, E., Tang, H. R., and Nicholson, J. K. (2003) Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning ^{1}H nuclear magnetic resonance spectroscopy of liver tissues. *Analytical Biochemistry*, **323**, 26.

[126] Ludwig, C. and Viant, M. R. (2010) Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, **21**, 22.

[127] Stejskal, E. O. and Tanner, J. E. (1965) Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *Journal of Chemical Physics*, **42**, 288.

[128] Nilsson, M. (2009) The DOSY toolbox: a new tool for processing PFG NMR diffusion data. *Journal of Magnetic Resonance*, **200**, 296.

[129] Viel, S. and Caldarelli, S. (2008) Improved 3D DOSY-TOCSY experiment for mixture analysis. *Chemical Communications*, **17**, 2013.

[130] Balayssac, S., Trefi, S., Gilard, V., Malet-Martino, M., Martino, R., and Delsuc, M. A. (2009) 2D and 3D DOSY ^{1}H NMR, a useful tool for analysis of complex mixtures: application to herbal drugs or dietary supplements for erectile dysfunction. *Journal of Pharmaceutical Biomedicine*, **50**, 602.

[131] Barjat, H., Morris, G. A., and Swanson, A. G. (1998) A three-dimensional DOSY-HMQC experiment for the high-resolution analysis of complex mixtures. *Journal of Magnetic Resonance*, **131**, 131.

[132] Cobas, J. C. and Martin-Pastor, M. (2004) A homodecoupled diffusion experiment for the analysis of complex mixtures by NMR. *Journal of Magnetic Resonance*, **171**, 20.

[133] Nilsson, M. and Morris, G. A. (2007) Pure shift proton DOSY: diffusion-ordered ^{1}H spectra without multiplet structure. *Chemical Communications*, p. 933.

[134] Lucas, L. H., Otto, W. H., and Larive, C. K. (2002) The 2D-j-DOSY experiment: resolving diffusion coefficients in mixtures. *Journal of Magnetic Resonance*, **156**, 138.

[135] McLachlan, A. S., Richards, J. J., Bilia, A. R., and Morris, G. A. (2009) Constant time gradient HSQC-iDOSY: practical aspects. *Magnetic Resonance in Chemistry*, **47**, 1081.

[136] Bax, A. and Subramanian, S. (1986) Sensitivity-enhanced two-dimensional heteronuclear shift correlation NMR spectroscopy. *Journal of Magnetic Resonance*, **67**, 565.

[137] Brown, S. P. and Emsley, L. (2004) The 2D MAS NMR spin-echo experiment: the determination of ^{13}C-^{13}C J couplings in a solid-state cellulose sample. *Journal of Magnetic Resonance*, **171**, 43.

[138] Nadaud, P. S., Helmus, J. J., Sengupta, I., and Jaroniec, C. P. (2010) Rapid acquisition of multidimensional solid-state NMR spectra of proteins facilitated by covalently bound paramagnetic tags. *Journal of the American Chemical Society*, **132**, 9561.

[139] Wind, R. A. and Hu, J. Z. (2006) *In vivo* and *ex vivo* high-resolution ^{1}H NMR in biological systems using low-speed magic angle spinning. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **49**, 207.

[140] Wilson, M., Davies, N. P., Brundler, M. A., McConville, C., Grundy, R. G., and Peet, A. C. (2009) High resolution magic angle spinning ^{1}H NMR of childhood brain and nervous system tumours. *Molecular Cancer*, **8**.

[141] Zietkowski, D., Davidson, R. L., Eykyn, T. R., De Silva, S. S., deSouza, N. M., and Payne, G. S. (2010) Detection of cancer in cervical tissue biopsies using mobile lipid resonances measured with diffusion-weighted ^{1}H magnetic resonance spectroscopy. *NMR in Biomedicine*, **23**, 382.

[142] Martinez-Richa, A. and Joseph-Nathan, P. (2003) Carbon-13 CP-MAS nuclear magnetic resonance studies of teas. *Solid State Nuclear Magnetic Resonance*, **23**, 119.

[143] Han, O. H., Bae, Y. K., and Jeong, S. Y. (2008) Carbon-13 CP MAS NMR study on structures of octadecyl chains influenced by co-presence of 3-aminopropyl chains on SBA-15. *Bulletin of the Korean Chemical Society*, **29**, 405.

[144] O'Donnell, M. D., Hill, R. G., Law, R. V., and Fong, S. (2009) Raman spectroscopy, ^{19}F and ^{31}P MAS-NMR of a series of fluorochloroapatites. *Journal of the European Ceramic Society*, **29**, 377.

[145] Hill, R. G., Law, R. V., O'Donnell, M. D., Hawes, J., Bubb, N. L., Wood, D. J., Miller, C. A., Mirsaneh, M., and Reaney, I. (2009) Characterisation of fluorine containing glasses and glass-ceramics by ^{19}F magic angle spinning nuclear magnetic resonance spectroscopy. *Journal of the European Ceramic Society*, **29**, 2185.

[146] He, Z. Q., Honeycutt, C. W., Xing, B., McDowell, R. W., Pellechia, P. J., and Zhang, T. Q. (2007) Solid-state fourier transform infrared and ^{31}P nuclear magnetic resonance spectral features of phosphate compounds. *Soil Science*, **172**, 501.

[147] Zhou, Z. M., Lan, W. X., Zhang, W. N., Zhang, X., Xia, S. A., Zhu, H., Ye, C. H., and Liu, M. L. (2007) Implementation of real-time two-dimensional nuclear magnetic resonance spectroscopy for on-flow high-performance liquid chromatography. *Journal of Chromatography A*, **1154**, 464.

[148] Graça, G., Duarte, I. F., Goodfellow, B. J., Carreira, I. M., Couceiro, A. B., Domingues, M. D., Spraul, M., Tseng, L. H., and Gil, A. M. (2008) Metabolite profiling of human amniotic fluid by hyphenated nuclear magnetic resonance spectroscopy. *Analytical Chemistry*, **80**, 6085.

[149] Akira, K., Mitome, H., Imachi, M., Shida, Y., Miyaoka, H., and Hashimoto, T. (2010) LC-NMR identification of a novel taurine-related metabolite observed in ^{1}H NMR-based metabonomics of genetically hypertensive rats. *Journal of Pharmaceutical Biomedicine*, **51**, 1091.

[150] Tode, C., Maoka, T., and Sugiura, M. (2009) Application of LC-NMR to analysis of carotenoids in foods. *Journal of Separation Science*, **32**, 3659.

[151] Dias, D. A. and Urban, S. (2009) Application of HPLC-NMR for the rapid chemical profiling of a southern australian sponge, *Dactylospongia sp. Journal of Separation Science*, **32**, 542.

[152] Wilson, I. D. and Brinkman, U. A. T. (2007) Hype and hypernation: multiple hyphenation of column liquid chromatography and spectroscopy. *TrAC - Trends in Analytical Chemistry*, **26**, 847.

[153] Traficante, D. D. and Rajabzadeh, M. (2000) Optimum window function for sensitivity enhancement of NMR signals. *Concepts in Magnetic Resonance*, **12**, 83.

[154] Halouska, S. and Powers, R. (2006) Negative impact of noise on the principal component analysis of NMR data. *Journal of Magnetic Resonance*, **178**, 88.

[155] Huo, R., van de Molengraaf, R. A., Pikkemaat, J. A., Wehrens, R., and Buydens, L. M. C. (2005) Diagnostic analysis of experimental artefacts in DOSY NMR data by covariance matrix of the residuals. *Journal of Magnetic Resonance*, **172**, 346.

[156] Gibbs, A., Morris, G. A., Swanson, A. G., and Cowburn, D. (1993) Suppression of t1 noise in 2D NMR spectroscopy by reference deconvolution. *Journal of Magnetic Resonance*, **101**, 351.

[157] Horne, T. J. and Morris, G. A. (1996) Combined use of gradient-enhanced techniques and reference deconvolution for ultralow t1 noise in 2D NMR spectroscopy. *Journal of Magnetic Resonance*, **Ser A**, 246.

[158] Morris, G. A., Barjat, H., and Horne, T. J. (1997) Reference deconvolution methods. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **31**, 197.

[159] Brissac, C., Malliavin, T. E., and Delsuc, M. A. (1995) Use of the Cadzow procedure in 2D NMR for the reduction of t1 noise. *Journal of Biomolecular NMR*, **6**, 361.

[160] Otting, G., Widmer, H., Wagner, G., and Wuthrich, K. (1986) Origin of t1 and t2 ridges in 2D NMR spectra and procedures for suppression. *Journal of Magnetic Resonance*, **66**, 187.

[161] Plateau, P., Dumas, C., and Gueron, M. (1983) Solvent peak suppressed NMR - correction of baseline distortions and use of strong-pulse excitation. *Journal of Magnetic Resonance*, **54**, 46.

[162] Hoult, D. I. and Richards, R. E. (1975) Critical factors in design of sensitive high-resolution nuclear magnetic resonance spectrometers. *Proceedings of the Royal Society of London Series A - Mathematical and Physical Sciences*, **344**, 311.

[163] Cobas, J. C., Bernstein, M. A., Martin-Pastor, M., and Tahoces, P. G. (2006) A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance*, **183**, 145.

[164] Brown, D. E. (1995) Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials. *Journal of Magnetic Resonance*, **114**, 268.

[165] Chang, D., Banack, C. D., and Shah, S. L. (2007) Robust baseline correction algorithm for signal dense NMR spectra. *Journal of Magnetic Resonance*, **187**, 288.

[166] Xi, Y. X. and Rocke, D. M. (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, **9**, 324.

[167] Pravdova, V., Walczak, B., and Massart, D. L. (2002) A comparison of two algorithms for warping of analytical signals. *Analytica Chimica Acta*, **456**, 77.

[168] Kim, S. B., Wang, G., and Duran, C. M. (November 2006) *A Bayesian approach for the alignment of high-resolution NMR spectra*. Pittsburgh, PA.

[169] Forshed, J., Schuppe-Koistinen, I., and Jacobsson, S. P. (2003) Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, **487**, 189.

[170] Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., Davies, D. B., and Nicholson, J. K. (2009) Recursive segment-wise peak alignment of biological ^{1}H NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, **81**, 56.

[171] Savorani, F., Tomasi, G., and Engelsen, S. B. (2010) icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, **202**, 190.

[172] Koh, H. W., Maddula, S., Lambert, J., Hergenroder, R., and Hildebrand, L. (2009) An approach to automated frequency-domain feature extraction in nuclear magnetic resonance spectroscopy. *Journal of Magnetic Resonance*, **201**, 146.

[173] Xi, Y. X., de Ropp, J. S., Viant, M. R., Woodruff, D. L., and Yu, P. (2008) Improved identification of metabolites in complex mixtures using HSQC NMR spectroscopy. *Analytica Chimica Acta*, **614**, 127.

[174] Lewis, I. A., Schommer, S. C., and Markley, J. L. (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance in Chemistry*, **47**, S123.

[175] Wu, W. and Massart, D. L. (1996) Artificial neural networks in classification of NIR spectral data: selection of the input. *Chemometrics and Intelligent Laboratory Systems*, **35**, 127.

[176] Park, C. H. and Park, H. (2008) A comparison of generalized linear discriminant analysis algorithms. *Pattern Recognition*, **41**, 1083.

[177] D'Imperio, M., Mannina, L., Capitani, D., Bidet, O., Rossi, E., Bucarelli, F. M., Quaglia, G. B., and Segre, A. L. (2007) NMR and statistical study of olive oils from Lazio: a geographical, ecological and agronomic characterization. *Food Chemistry*, **105**, 1256.

[178] Mannina, L., Marini, F., Gobbino, M., Sobolev, A. P., and Capitani, D. (2010) NMR and chemometrics in tracing European olive oils: the case study of Ligurian samples. *Talanta*, **80**, 2141.

[179] Boffo, E. F., Tavares, L. A., Ferreira, M. M. C., and Ferreira, A. G. (2009) Classification of Brazilian vinegars according to their ^{1}H NMR spectra by pattern recognition analysis. *LWT - Food Science and Technology*, **42**, 1455.

[180] Rezzi, S., Axelson, D. E., Heberger, K., Reniero, F., Mariani, C., and Guillou, C. (2005) Classification of olive oils using high throughput flow ^{1}H NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks. *Analytica Chimica Acta*, **552**, 13.

[181] Wold, H. (1966) *Non-linear estimation by iterative least squares procedures*, pp. 411–444. Wiley.

[182] Young, S. P., Nessim, M., Falciani, F., Trevino, V., Banerjee, S. P., Scott, R. A. H., Murray, P. I., and Wallace, G. R. (2009) Metabolomic analysis of human vitreous humor differentiates ocular inflammatory disease. *Molecular Vision*, **15**, 1210.

[183] Ren, Y. F., Wang, T., Peng, Y. F., Xia, B., and Qu, L. J. (2009) Distinguishing transgenic from non-transgenic *Arabidopsis* plants by ^{1}H NMR-based metabolic fingerprinting. *Journal of Genetics and Genomics*, **36**, 621.

[184] Mao, H. L., Wang, H. M., Wang, B., Liu, X., Gao, H. C., Xu, M., Zhao, H. S., Deng, X. M., and Lin, D. H. (2009) Systemic metabolic changes of traumatic critically iII patients revealed by an NMR-based metabonomic approach. *Journal of Proteome Research*, **8**, 5423.

[185] Ahmed, S. S. S. J., Santosh, W., Kumar, S., and Christlet, H. T. T. (2009) Metabolic profiling of Parkinson's disease: evidence of biomarker from gene expression analysis and rapid neural network detection. *Journal of Biomedical Science*, **16**.

[186] Giskeodegard, G. F., Grinde, M. T., Sitter, B., Axelson, D. E., Lundgren, S., Fjosne, H. E., Dahl, S., Gribbestad, I. S., and Bathen, T. F. (2010) Multivariate modeling and prediction of breast cancer prognostic factors using MR metabolomics. *Journal of Proteome Research*, **9**, 972.

[187] Beckwith-Hall, B. M., Brindle, J. T., Barton, R. H., Coen, M., Holmes, E., Nicholson, J. K., and Antti, H. (2002) Application of orthogonal signal correction to minimise the effects of physical and biological variation in high resolution ^{1}H NMR spectra of biofluids. *Analyst*, **127**, 1283.

[188] Trygg, J. and Wold, S. (2002) Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16**, 119.

[189] Trygg, J. (2002) O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics*, **16**, 283.

[190] Trygg, J., Holmes, E., and Lundstedt, T. (2007) Chemometrics in metabonomics. *Journal of Proteome Research*, **6**, 469.

[191] Holland, J. (1975) *Adaption in Natural and Artificial Systems*. MIT press, Cambridge, MA.

[192] Ramadan, Z., Jacobs, D., Grigorov, M., and Kochhar, S. (2006) Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta*, **68**, 1683.

[193] Whitlock, M. C. and Barton, N. H. (1997) The effective size of a subdivided population. *Genetics*, **146**, 427.

[194] Schölkopf, B., Smola, A., and Müller, K.-R. (1997) *Kernel principal component analysis*, vol. 1327 of *Lecture Notes in Computer Science*, pp. 583–588. Springer Berlin / Heidelberg.

[195] Baudat, G. and Anouar, F. E. (2000) Generalized discriminant analysis using a kernel approach. *Neural Computation*, **12**, 2385.

[196] Rosipal, R. and Trejo, L. J. (2002) Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, **2**, 97.

[197] Embrechts, M. J. and Ekins, S. (2007) Classification of metabolites with kernel-partial least squares (K-PLS). *Drug Metabolism and Disposition*, **35**, 325.

[198] Rantalainen, M., Bylesjo, M., Cloarec, O., Nicholson, J. K., Holmes, E., and Trygg, J. (2007) Kernel-based orthogonal projections to latent structures (K-OPLS). *Journal of Chemometrics*, **21**, 376.

[199] Bylesjo, M., Rantalainen, M., Nicholson, J. K., Holmes, E., and Trygg, J. (2008) K-opls package: kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinformatics*, **9**.

[200] Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biolical Cybernetics*, **43**, 59.

[201] Wongravee, K., Lloyd, G. R., Silwood, C. J., Grootveld, M., and Brereton, R. G. (2010) Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling. *Analytical Chemistry*, **82**, 628.

[202] Rezzi, S., Giani, I., Heberger, K., Axelson, D. E., Moretti, V. M., Reniero, F., and Guillou, C. (2007) Classification of gilthead sea bream (*Sparus aurata*) from ^{1}H NMR lipid profiling combined with principal component and linear discriminant analysis. *Journal of Agricultural and Food Chemistry*, **55**, 9963.

[203] Garcia-Gonzalez, D. L., Mannina, L., D'Imperio, M., Segre, A. L., and Aparicio, R. (2004) Using ^{1}H and ^{13}C NMR techniques and artificial neural networks to detect the adulteration of olive oil with hazelnut oil. *European Food Research and Technology*, **219**, 545.

[204] Gerbst, A. G., Grachev, A. A., Ustuzhanina, N. E., Nifantiev, N. E., Vyboichtchik, A. A., Shashkov, A. S., and Usov, A. I. (2010) Application of ar-

tificial neural networks for analysis of ^{13}C NMR spectra of fucoidans. *Journal of Carbohydrate Chemistry*, **29**, 92.

[205] Meiler, J. and Kock, M. (2004) Novel methods of automated structure elucidation based on ^{13}C NMR spectroscopy. *Magnetic Resonance in Chemistry*, **42**, 1042.

[206] Lin, Z. Y., Xu, P. B., Yan, S. K., Meng, H. B., Yang, G. J., Dai, W. X., Liu, X. R., Li, J. B., Deng, X. M., and Zhang, W. D. (2009) A metabonomic approach to early prognostic evaluation of experimental sepsis by ^{1}H NMR and pattern recognition. *NMR in Biomedicine*, **22**, 601.

[207] Masoum, S., Bouveresse, D. J. R., Vercauteren, J., Jalali-Heravi, M., and Rutledge, D. N. (2006) Discrimination of wines based on 2D NMR spectra using learning vector quantization neural networks and partial least squares discriminant analysis. *Analytica Chimica Acta*, **558**, 144.

[208] Smilde, A. K. (2007) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **23**, 3415.

[209] Harshman, R. A. (1970) Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1.

[210] Bro, R. (1997) PARAFAC. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, **38**, 149.

[211] Carroll, J. D. and Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, **35**, 283.

[212] Alm, E., Torgrip, R. J. O., Aberg, K. M., Schuppe-Koistinen, I., and Lindberg, J. (2010) Time-resolved biomarker discovery in ^{1}H NMR data using generalized fuzzy hough transform alignment and parallel factor analysis. *Analytical and Bioanalytical Chemistry*, **396**, 1681.

[213] Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., Jellema, R. H., Pijl, H., Roelfsema, F., and van der Greef, J. (2010) Dynamic metabolomic data analysis: a tutorial review. *Metabolomics*, **6**, 3.

[214] Jansen, J. J., Bro, R., Hoefsloot, H. C. J., van den Berg, F. W. J., Westerhuis, J. A., and Smilde, A. K. (2008) PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. *Journal of Chemometrics*, **22**, 114.

[215] Bro, R., Viereck, N., Toft, M., Toft, H., Hansen, P. I., and Engelsen, S. B. (2010) Mathematical chromatography solves the cocktail party effect in mixtures using 2D spectra and PARAFAC. *TrAC - Trends in Analytical Chemistry*, **29**, 281.

[216] Breiman, L., Freidman, J., Stone, C. J., and Olshen, R. A. (1984) *Classification and regression trees*. Chapman and Hall//CRC.

[217] Petrakis, P. V., Agiomyrgianaki, A., Christophoridou, S., Spyros, A., and Dais, P. (2008) Geographical characterization of greek virgin olive oils (cv. koroneiki) using ^{1}H and ^{31}P NMR fingerprinting with canonical discriminant analysis and classification binary trees. *Journal of Agricultural and Food Chemistry*, **56**, 3200.

[218] Thomas, M. A., et al. (2009) Investigation of breast cancer using two-dimensional MRS. *NMR in Biomedicine*, **22**, 77.

[219] Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5.

[220] Xia, J. G., Psychogios, N., Young, N., and Wishart, D. S. (2009) Metaboanalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, **37**, W652.

[221] Robinette, S. L., Zhang, F., Brüschweiler-Li, L., and Brüschweiler, R. (2008) Web server based complex mixture analysis by NMR. *Analytical Chemistry*, **80**, 3606.

[222] Frydman, L., Scherf, T., and Lupulescu, A. (2002) The acquisition of multidimensional NMR spectra within a single scan. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15858.

[223] Tal, A. and Frydman, L. (2010) Single-scan multidimensional magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **57**, 241.

[224] McDermott, R., et al. (2004) SQUID-detected magnetic resonance imaging in microtesla magnetic fields. *Journal of Low Temperature Physics*, **135**, 793.

[225] Trabesinger, A. H., McDermott, R., Lee, S. K., Muck, M., Clarke, J., and Pines, A. (2004) SQUID-detected liquid state NMR in microtesla fields. *Journal of Physical Chemistry A*, **108**, 957.

[226] Jaffer, F. A. and Weissleder, R. (2005) Molecular imaging in the clinical arena. *JAMA - Journal of the American Medical Association*, **293**, 855.

[227] Rantalainen, M., et al. (2006) Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *Journal of Proteome Research*, **5**, 2642.

[228] Forshed, J., Stolt, R., Idborg, H., and Jacobsson, S. P. (2007) Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and ^{1}H NMR data. *Chemometrics and Intelligent Laboratory Systems*, **85**, 179.

[229] Zelena, E., et al. (2009) Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Analytical Chemistry*, **81**, 1357.

[230] Dunn, W. B., et al. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, **6**, 1060.

[231] Castillo, S., Gopalacharyulu, P., Yetukuri, L., and Orešič, M. (2011) Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, **108**, 23.

[232] Soler, C. and Picó, Y. (2007) Recent trends in liquid chromatography-tandem mass spectrometry to determine pesticides and their metabolites in food. *TrAC - Trends in Analytical Chemistry*, **26**, 103.

[233] Bino, R. J., et al. (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, **9**, 418.

[234] McNaught, A. D. and Wilkinson, A. (eds.) (1997) *IUPAC. Compendium of Chemical Terminology*. Blackwell Scientific Publications, Oxford, 2nd edn.

[235] Matuszewski, B. K., Constanzer, M. L., and Chavez-Eng, C. M. (2003) Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Analytical Chemistry*, **75**, 3019.

[236] Sumner, L. W., et al. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211.

[237] Gika, H. G., Theodoridis, G. A., Wingate, J. E., and Wilson, I. D. (2007) Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: application to human urine. *Journal of Proteome Research*, **6**, 3291.

[238] Vaughan, A. A., Dunn, W. B., Allwood, J. W., Wedge, D. C., Blackhall, F. H., Whetton, A. D., Dive, C., and Goodacre, R. (2012) Liquid chromatography–mass spectrometry calibration transfer and metabolomics data fusion. *Analytical Chemistry*, **84**, 9848.

[239] Saude, E. and Sykes, B. (2007) Urine stability for metabolomic studies: effects of preparation and storage. *Metabolomics*, **3**, 19.

[240] Want, E. J., O'Maille, G., Smith, C. A., Brandon, T. R., Uritboonthai, W., Qin, C., Trauger, S. A., and Siuzdak, G. (2006) Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Analytical Chemistry*, **78**, 743.

[241] Ferreiro-Vera, C., Priego-Capote, F., and de Castro, M. L. (2012) Comparison of sample preparation approaches for phospholipids profiling in human serum by liquid chromatography–tandem mass spectrometry. *Journal of Chromatography A*, **1240**, 21.

[242] Denery, J. R., Nunes, A. A. K., and Dickerson, T. J. (2011) Characterization of differences between blood sample matrices in untargeted metabolomics. *Analytical Chemistry*, **83**, 1040.

[243] Wedge, D. C., et al. (2011) Is serum or plasma more appropriate for intersubject comparisons in metabolomic studies? An assessment in patients with small-cell lung cancer. *Analytical Chemistry*, **83**, 6689.

[244] Sentandreu, M. A. and Sentandreu, E. (2011) Peptide biomarkers as a way to determine meat authenticity. *Meat Science*, **89**, 280.

[245] Surowiec, I., Fraser, P. D., Patel, R., Halket, J., and Bramley, P. M. (2011) Metabolomic approach for the detection of mechanically recovered meat in food products. *Food Chemistry*, **125**, 1468.

[246] Martinez-Villalba, A., Moyano, E., and Galceran, M. T. (2010) Analysis of amprolium by hydrophilic interaction liquid chromatography-tandem mass spectrometry. *Journal of Chromatography A*, **1217**, 5802.

[247] Dasenaki, M. E. and Thomaidis, N. S. (2010) Multi-residue determination of seventeen sulfonamides and five tetracyclines in fish tissue using a multistage LC-ESI-MS/MS approach based on advanced mass spectrometric techniques. *Analytica Chimica Acta*, **672**, 93.

[248] Ho, C., Lee, W.-O., and Wong, Y.-T. (2012) Determination of N-methyl-1,3-propanediamine in bovine muscle by liquid chromatography with triple quadrupole and ion trap tandem mass spectrometry detection. *Journal of Chromatography A*, **1235**, 103.

[249] Merou, A., Kaklamanos, G., and Theodoridis, G. (2012) Determination of carbadox and metabolites of carbadox and olaquindox in muscle tissue using high performance liquid chromatography–tandem mass spectrometry. *Journal of Chromatography B*, **881**, 90.

[250] Xiong, Y., Zhao, Y.-Y., Goruk, S., Oilund, K., Field, C. J., Jacobs, R. L., and Curtis, J. M. (2012) Validation of an LC-MS/MS method for the quantification of choline-related compounds and phospholipids in foods and tissues. *Journal of Chromatography B*, **911**, 170.

[251] Strobel, N., Buddhadasa, S., Adorno, P., Stockham, K., and Greenfield, H. (2012) Vitamin D and 25-hydroxyvitamin D determination in meats by LC-IT-MS. *Food Chemistry*, **138**, 1042.

[252] Theodoridis, G., Gika, H., Franceschi, P., Caputi, L., Arapitsas, P., Scholz, M., Masuero, D., Wehrens, R., Vrhovsek, U., and Mattivi, F. (2012) LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. *Metabolomics*, **8**, 175.

[253] Antonio, C., Pinheiro, C., Chavez, M. M., Ricardo, C. P., Ortuño, M. F., and Thomas-Oates, J. E. (2008) Analysis of carbohydrates in lupinus albus stems on imposition of water deficit, using porous graphitic carbon liquid chromarography-electrospray ionisation mass spectrometry. *Journal of Chromatography A*, **1187**, 111.

[254] Anastassiades, M., Lehotay, S. J., Stajnbaher, D., and Schenck, F. J. (2003) Fast and east multiresidue method employing acetonitrile extraction/partitioning and dispersive solid-phase extraction for the determination of pesticide residues in produce. *Journal of AOAC International*, **86**, 412.

[255] Sinha, S. N., Vasudev, K., and Rao, M. V. V. (2012) Quantification of organophosphate insecticides and herbicides in vegetable samples using the "quick easy cheap effective rugged and safe" (QuEChERS) method and a high-performance liquid chromatography–electrospray ionisation–mass spectrometry (LC-MS/MS) technique. *Food Chemistry*, **132**, 1574.

[256] McKenzie, J. S., Jurado, J. M., and de Pablos, F. (2010) Characterisation of tea leaves according to their total mineral content by means of probabilistic neural networks. *Food Chemistry*, **123**, 859.

[257] Celik, S. E., Ozyurek, M., Guclu, K., and Apak, R. (2010) Determination of antioxidants by a novel on-line HPLC-cupric reducing antioxidant capacity (CUPRAC) assay with post-column detection. *Analytica Chimica Acta*, **674**, 79.

[258] Alcázar, A., Ballesteros, O., Jurado, J. M., Pablos, F., Martín, M. J., Vilches, J. L., and Navalón, A. (2007) Differentiation of green, white, black, oolong, and pu-

erh teas according to their free amino acids content. *Journal of Agricultural and Food Chemistry*, **55**, 5960.

[259] Fraser, K., Harrison, S. J., Lane, G. A., Otter, D. E., Hemar, Y., Quek, S.-Y., and Rasmussen, S. (2012) Non-targeted analysis of tea by hydrophilic interaction liquid chromatography and high resolution mass spectrometry. *Food Chemistry*, **134**, 1616.

[260] Morrison, N., et al. (2007) Standard reporting requirements for biological samples in metabolomics experiments: environmental context. *Metabolomics*, **3**, 203.

[261] Alpert, A. J. (1990) Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography A*, **499**, 177.

[262] Tolstikov, V. V. and Fiehn, O. (2002) Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry*, **301**, 298.

[263] Cubbon, S., Antonio, C., Wilson, J., and Thomas-Oates, J. (2010) Metabolomic applications of HILIC-LC-MS. *Mass Spectrometry Reviews*, **29**, 671.

[264] Tal'rose, V. L., Karpov, G. V., Gorodetskii, I. G., and Skurat, V. E. (1968) Analysis of mixtures of organic substances in a mass spectrometer with a capillary system for introducing liquid samples. *Russian Journal of Physical Chemistry (English Translation)*, **42**, 1658.

[265] Baldwin, M. A. and McLafferty, F. W. (1973) Direct chemical ionization of relatively involatile samples. application to underivatized oligopeptides. *Organic Mass Spectrometry*, **7**, 1353–1356.

[266] Abian, J. (1999) The coupling of gas and liquid chromatography with mass spectrometry. *Journal of Mass Spectrometry*, **34**, 157.

[267] Morris, H. R., Panico, M., Barber, M., Bordoli, R. S., Sedgwick, R. D., and Tyler, A. (1981) Fast atom bombardment: A new mass spectrometric method

for peptide sequence analysis. *Biochemical and Biophysical Research Communications*, **101**, 623.

[268] Gieniec, J., Mack, L. L., Nakamae, K., Gupta, C., Kumar, V., and Dole, M. (1984) Electrospray mass spectroscopy of macromolecules: application of an ion-drift spectrometer. *Biological Mass Spectrometry*, **11**, 259.

[269] Wong, S. F., Meng, C. K., and Fenn, J. B. (1988) Multiple charging in electrospray ionization of poly(ethylene glycols). *The Journal of Physical Chemistry*, **92**, 546.

[270] Fenn, J., Mann, M., Meng, C., Wong, S., and Whitehouse, C. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**, 64.

[271] Fenn, J. B. (2003) Electrospray wings for molecular elephants (Nobel lecture). *Angewandte Chemie International Edition*, **42**, 3871.

[272] Van Berkel, G. J. and Zhou, F. (1995) Characterization of an electrospray ion source as a controlled-current electrolytic cell. *Analytical Chemistry*, **67**, 2916.

[273] Emmett, M. and Caprioli, R. (1994) Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. *Journal of the American Society for Mass Spectrometry*, **5**, 605.

[274] Wilm, M. and Mann, M. (1996) Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry*, **68**, 1.

[275] Juraschek, R., Dülcks, T., and Karas, M. (1999) Nanoelectrospray - more than just a minimized-flow electrospray ionization source. *Journal of the American Society for Mass Spectrometry*, **10**, 300.

[276] Leandro, C. C., Hancock, P., Fussell, R. J., and Keely, B. J. (2006) Comparison of ultra-performance liquid chromatography and high-performance liquid chromatography for the determination of priority pesticides in baby foods by tandem quadrupole mass spectrometry. *Journal of Chromatography A*, **1103**, 94 – 101.

[277] Soler, C., James, K., and Picó, Y. (2007) Capabilities of different liquid chromatography tandem mass spectrometry systems in determining pesticide residues in food: Application to estimate their daily intake. *Journal of Chromatography A*, **1157**, 73 – 84.

[278] Khayoon, W. S., Saad, B., Salleh, B., Ismail, N. A., Manaf, N. H. A., and Latiff, A. A. (2010) A reversed phase high performance liquid chromatography method for the determination of fumonisins b1 and b2 in food and feed using monolithic column and positive confirmation by liquid chromatography/tandem mass spectrometry. *Analytica Chimica Acta*, **679**, 91 – 97.

[279] Stafford Jr., G., Kelley, P., Syka, J., Reynolds, W., and Todd, J. (1984) Recent improvements in and analytical applications of advanced ion trap technology. *International Journal of Mass Spectrometry and Ion Processes*, **60**, 85.

[280] Paul, W. (1990) Electromagnetic traps for charged and neutral particles (Nobel lecture). *Angewandte Chemie International Edition in English*, **29**, 739.

[281] Jonscher, K. R. and Yates III, J. R. (1997) The quadrupole ion trap mass spectrometer - a small solution to a big challenge. *Analytical Biochemistry*, **244**, 1.

[282] Douglas, D. J., Frank, A. J., and Mao, D. (2005) Linear ion traps in mass spectrometry. *Mass Spectrometry Reviews*, **24**, 1.

[283] Londry, F. and Hager, J. W. (2003) Mass selective axial ion ejection from a linear quadrupole ion trap. *Journal of the American Society for Mass Spectrometry*, **14**, 1130.

[284] Schowalter, S. J., Chen, K., Rellergert, W. G., Sullivan, S. T., and Hudson, E. R. (2012) A novel time-of-flight mass spectrometer using radial extraction from a linear quadrupole trap for atomic, molecular, and chemical physics. *ArXiv e-prints*.

[285] Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, **72**, 1156.

[286] Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005) The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, **40**, 430.

[287] ThermoScientific (2013), http://planetorbitrap.com/exactive-plus.

[288] Hardman, M. and Makarov, A. A. (2003) Interfacing the orbitrap mass analyzer to an electrospray ion source. *Analytical Chemistry*, **75**, 1699.

[289] Koppenaal, D. W., Barinaga, C. J., Denton, M. B., Sperline, R. P., Hieftje, G. M., Schilling, G. D., Andrade, F. J., Barnes, J. H., and IV, I. (2005) MS detectors. *Analytical Chemistry*, **77**, 418A.

[290] Hrydziuszko, O. and Viant, M. (2012) Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, **8**, 161.

[291] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008) PubChem: integrated platform of small molecules and biological activities. vol. 4 of *Annual Reports in Computational Chemistry*, pp. 217 – 241, Elsevier.

[292] Pence, H. E. and Williams, A. (2010) ChemSpider: An online chemical information resource. *Journal of Chemical Education*, **87**, 1123.

[293] Little, J., Cleven, C., and Brown, S. (2011) Identification of known unknowns utilizing accurate mass data and chemical abstracts service databases. *Journal of the American Society for Mass Spectrometry*, **22**, 348.

[294] Smith, C., O'Maille, G., Want, E., Qin, C., Trauger, S., Brandon, T., Custodio, D., Abagyan, R., and Siuzdak, G. (2005) METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, **27**, 747.

[295] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27.

[296] Sud, M., et al. (2007) LMSD: lipid maps structure database. *Nucleic Acids Research*, **35**, D527.

[297] SDBSWeb, http://riodb01.ibase.aist.go.jp/sdbs/, National Institute of Advanced Industrial Science and Technology.

[298] Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, **78**, 779.

[299] Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012) XCMS online: a web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, **84**, 5035.

[300] Smith, C. A. (2008) LC/MS preprocessing and analysis with XCMS. *http://bioconductor.wustl.edu/bioc/vignettes/xcms/inst/doc/xcmsPreprocess.pdf*.

[301] Katajamaa, M. and Oresic, M. (2005) Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, **6**, 179.

[302] Steyerberg, E. W., Bleeker, S., Moll, H., Grobbee, D., and Moons, K. (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, **56**, 441.

[303] Zand, N., Chowdhry, B. Z., Pullen, F. S., Snowden, M. J., and Tetteh, J. (2012) Simultaneous determination of riboflavin and pyridoxine by UHPLC/LC-MS in UK commercial infant meal food products. *Food Chemistry*, **135**, 2743.

[304] Ricard, F., Abe, E., Duverneuil-Mayer, C., Charlier, P., de la Grandmaison, G., and Alvarez, J. C. (2012) Measurement of atropine and scopolamine in hair by LC-MS/MS after *Datura stramonium* chronic exposure. *Forensic Science International*, **223**, 256.

[305] Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., and Selbig, J. (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics*, **20**, 2447.

[306] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2012) Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *Journal of Proteome Research*, **11**, 4120.

[307] Janné, K., Pettersen, J., Lindberg, N., and Lundstedt, T. (2001) Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration. *Journal of Chemometrics*, **15**, 203.

[308] Yamamoto, H., Yamaji, H., Fukusaki, E., Ohno, H., and Fukuda, H. (2008) Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochemical Engineering Journal*, **40**, 199.

[309] Mantel, N. (1967) The detection of disease clustering and a generalised regression approach. *Cancer Research*, **27**, 209.

[310] Laurentin, H., Ratzinger, A., and Karlovsky, P. (2008) Relationship between metabolic and genomic diversity in sesame (sesamum indicum l.). *BMC Genomics*, **9**, 250.

[311] Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., and Schmitt-Kopplin, P. (2009) Metabolomics reveals metabolic biomarkers of crohn's disease. *PLoS ONE*, **4**, e6386.

[312] Houshyani, B., Kabouw, P., Muth, D., de Vos, R. C. H., Bino, R. J., and Bouwmeester, H. J. (2012) Characterisation of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomcs*, **8**, S131.

[313] Goodacre, R., Shann, B., Gilbert, R. J., Timmins, E. M., McGovern, A. C., Alsberg, B. K., Kell, D. B., and Logan, N. A. (2000) Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Analytical Chemistry*, **72**, 119.

[314] Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M., and Moritz, T. (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical Biochemistry*, **331**, 283.

[315] Bailey, N. J. C., Oven, M., Holmes, E., Nicholson, J. K., and Zenk, M. H. (2003) Metabolomic analysis of the consequences of cadmium exposure in *Silene cucubalus* cell cultures via H-1 NMR spectroscopy and chemometrics. *Phytochemistry*, **62**, 851.

[316] Bollard, M. E., Stanley, E. G., Lindon, J. C., Nicholson, J. K., and Holmes, E. (2005) NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR in Biomedicine*, **18**, 143.

[317] Ali, K., Maltese, F., Zyprian, E., Rex, M., Choi, Y. H., and Verpoorte, R. (2009) NMR metabolic fingerprinting based identification of grapevine metabolites associated with downy mildew resistance. *Journal of Agricultural and Food Chemistry*, **57**, 9599.

[318] Holmes, E., Nicholson, J. K., Nicholls, A. W., Lindon, J. C., Connor, S. C., Polley, S., and Connelly, J. (1998) The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine. *Chemometrics and Intelligent Laboratory Systems*, **44**, 245.

[319] Jankevics, A., Liepinsh, E., Liepinsh, E., Vilskersts, R., Grinberga, S., Pugovics, O., and Dambrova, M. (2009) Metabolomic studies of experimental diabetic urine samples by H-1 NMR spectroscopy and LC/MS method. *Chemometrics and Intelligent Laboratory Systems*, **97**, 11.

[320] Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L., and Bino, R. (2002) Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, **14**, 1437.

[321] Ward, J. L., Harris, C., Lewis, J., and Beale, M. H. (2003) Assessment of (1)H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*, **62**, 949.

[322] Charlton, A., Allnutt, T., Holmes, S., Chisholm, J., Bean, S., Ellis, N., Mullineaux, P., and Oehlschlager, S. (2004) NMR profiling of transgenic peas. *Plant Biotechnology Journal*, **2**, 27.

[323] Choi, H.-K., Yoon, J.-H., Kim, Y.-S., and Kwon, D. Y. (2007) Metabolomic profiling of Cheonggukjang during fermentation by H-1 NMR spectrometry and principal components analysis. *Process Biochemistry*, **42**, 263.

[324] Hirakawa, K., Koike, K., Uekusa, K., Nihira, M., Yuta, K., and Ohno, Y. (2009) Experimental estimation of postmortem interval using multivariate analysis of proton NMR metabolomic data. *Legal Medicine*, **11**, S282.

[325] Wold, H. (1966) *Estimation of principal components and related models by iterative least squares*, p. 391. NY: Academic Press.

[326] Alsberg, B. K., Goodacre, R., Rowland, J. J., and Kell, D. B. (1997) Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, k-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta*, **348**, 389.

[327] Tiziani, S., Schwartz, S. J., and Vodovotz, Y. (2006) Profiling of carotenoids in tomato juice by one- and two-dimensional NMR. *Journal of Agricultural and Food Chemistry*, **54**, 6094.

[328] Giraudeau, P., Shrot, Y., and Frydman, L. (2009) Multiple ultrafast, broadband 2D NMR spectra of hyperpolarized natural products. *Journal of the American Chemical Society*, **131**, 13902.

[329] Schleucher, J., Schwendinger, M., Sattler, M., Schmidt, P., Schedletzky, O., Glaser, S. J., Sorensen, O. W., and Griesinger, C. (1994) A general enhancement scheme in heteronuclear multidimensional NMR employing pulsed-field gradients. *Journal of Biomolecular NMR*, **4**, 301.

[330] Lewis, I. A., Schommer, S. C., Hodis, B., Robb, K. A., Tonelli, M., Westler, W. M., Suissman, M. R., and Markley, J. L. (2007) Method for determining molar concentrations of metabolites in complex solutions from two-dimensional H-1-C-13 NMR spectra. *Analytical Chemistry*, **79**, 9385.

[331] Gronwald, W., Klein, M. S., Kaspar, H., Fagerer, S. R., Nuernberger, N., Dettmer, K., Bertsch, T., and Oefner, P. J. (2008) Urinary metabolite quantification employing 2D NMR spectroscopy. *Analytical Chemistry*, **80**, 9288.

[332] Mehlkopf, A. F., Korbee, D., Tiggelman, T. A., and Freeman, R. (1984) Sources of t1 noise in two-dimensional NMR. *Journal of Magnetic Resonance*, **58**, 315.

[333] Mierisova, S. and Ala-Korpela, M. (2001) MR spectroscopy quantitation: a review of frequency domain methods. *NMR in Biomedicine*, **14**, 247.

[334] Turner, C. J., Connolly, P. J., and Stern, A. S. (1999) Artifacts in sensitivity-enhanced HSQC. *Journal of Magnetic Resonance*, **137**, 281.

[335] Griffin, J. L., Williams, H. J., Sang, E., Clarke, K., Rae, C., and Nicholson, J. K. (2001) Metabolic profiling of genetic disorders: a multitissue H-1 nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. *Analytical Biochemistry*, **293**, 16.

[336] Kaczmarek, K., Walczak, B., de Jong, S., and Vandeginste, B. G. M. (2002) Feature based fuzzy matching of 2D gel electrophoresis images. *Journal of Chemical Information and Computer Sciences*, **42**, 1293.

[337] Wood, N. J., Brannigan, J. A., Duckett, S. B., Heath, S. L., and Wagstafft, J. (2007) Detection of picomole amounts of biological substrates by para-hydrogen-enhanced NMR methods in conjunction with a suitable receptor complex. *Journal of the American Chemical Society*, **129**, 11012.

[338] Day, I. J., Mitchell, J. C., Snowden, M. J., and Davis, A. L. (2008) Investigation of the potential of the dissolution dynamic nuclear polarization method for general sensitivity enhancement in small-molecule NMR spectroscopy. *Applied Magnetic Resonance*, **34**, 453.

[339] Frydman, L. and Blazina, D. (2007) Ultrafast two-dimensional nuclear magnetic resonance spectroscopy of hyperpolarized solutions. *Nature Physics*, **3**, 415.

[340] Smolinska, A., Blanchet, L., Coulier, L., Ampt, K. A. M., Luider, T., Hintzen, R. Q., Wijmenga, S. S., and Buydens, L. M. C. (2012) Interpretation and visualization of non-linear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis. *PLoS ONE*, **7**, e38163.

[341] van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., and van der Werf, M. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

[342] Wold, S., Kettaneh, N., and Tjessem, K. (1996) Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, **10**, 463.

[343] Tukey, J. W. (1977) *Exploratory data analysis*. Adison-Wesley, Reading MA, USA.

[344] Nevedomskaya, E., Mayboroda, O. A., and Deelder, A. M. (2011) Cross-platform analysis of longitudinal data in metabolomics. *Molecular Biosystems*, **7**, 3214.

[345] Thompson, B. (1995) Stepwise regression and stepwise discriminant analysis need not apply here: a guidlines editorial. *Educational and Pstchological Measurement*, **55**, 525.

[346] Babyak, M. A. (2004) What you see may not be what you get: a brief, non-technical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, **66**, 411.

[347] Smolinska, A., et al. (2012) Simultaneous analysis of plasma and CSF by NMR and hierarchical models fusion. *Analytical and Bioanalytical Chemistry*, **403**, 947.

[348] Worzel, W. P., Yu, J., Almal, A. A., and Chinnaiyan, A. M. (2009) Applications of genetic programming in cancer research. *The International Journal of Biochemistry & Cell Biology*, **41**, 405.

[349] Ahmed, S., Zhang, M., and Peng, L. (2012) Genetic programming for biomarker detection in mass spectrometry data. Thielscher, M. and Zhang, D. (eds.), *AI 2012: Advances in Artificial Intelligence*, vol. 7691 of *Lecture Notes in Computer Science*, pp. 266–278, Springer Berlin Heidelberg.

[350] Argyri, A. A., Jarvis, R. M., Wedge, D., Xu, Y., Panagou, E. Z., Goodacre, R., and Nychas, G.-J. E. (2013) A comparison of Raman and FT-IR spectroscopy for the prediction of meat spoilage. *Food Control*, **29**, 461.

[351] Bahado-Singh, R. O., Akolekar, R., Mandal, R., Dong, E., Xia, J., Kruger, M., Wishart, D. S., and Nicolaides, K. (2013) First-trimester metabolomic detection of late-onset preeclampsia. *American Journal of Obstetrics and Gynecology*, **208**, 58.e1.

[352] Wei, S., Zhang, J., Liu, L., Ye, T., Nagana Gowda, G. A., Tayyari, F., and Raftery, D. (2011) Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples. *Analytical Chemistry*, **83**, 7616.

[353] Binev, Y. and Aires-de Sousa, J. (2004) Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *Journal of Chemical Information and Computer Sciences*, **44**, 940.

[354] Palmblad, M. (1999), http://www.ms-utils.org/isotop.html.

[355] Bruker (2011) *Almanac: analytical tables and product overview*. Bruker.

[356] Noda, I. (1989) Two-dimensional infrared spectroscopy. *Journal of the American Chemical Society*, **111**, 8116.

[357] Cloarec, O., et al. (2005) Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Analytical Chemistry*, **77**, 1282.

[358] Patiny, L., http://www.nmrdb.org/.

[359] Kettenring, J. R. (1971) Canonical analysis of several sets of variables. *Biometrika*, **58**, 433.

[360] McVicar, M. and de Bie, T. (2012) CCA and a multi-way extension for investigating common components between audio, lyrics and tags. *Proceedings of the 9th International Symposium on Computer Modelling and Retrieval*.