Profiling the Metabolome using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry, Optimised Signal Processing, Noise Filtering and Constraints Methods

Tristan Graeme Payne

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Electronic, Electrical and
Computer Engineering
The University of Birmingham
26th April 2011

ABSTRACT

The main aim of this research study was to develop methods for metabolic profiling, a process involving the identification of metabolites present in biological samples.

This thesis focuses on the application of Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) to metabolic profiling. The first part of this work examined and optimised the processing of the data from the FT-ICR instrument into high quality mass spectra. In the second part, the optimised mass spectra were analysed to determine the molecular formulae of the compounds in a sample.

Three novel methods were developed to improve the quality of mass spectra and their interpretation. The first, SIM-stitching, combined multiple, narrow spectra into a single, wide spectrum; an approach newly adopted by several laboratories in the MS and metabolomics communities. The second method employed a three-stage noise filter. Finally, the mass spectra were analysed using constraints optimisation methods and utility theory. It was discovered that SIM-stitching increased the effective sensitivity of the instrument five-fold, allowing many more metabolites to be detected. The three-stage filter approach showed how filter parameters can be selected to optimise noise reduction, and yielded significant benefits over other methods typically used. It was found that constraints methods can robustly and systematically identify the molecular formulae of compounds in a spectrum.

A principal conclusion of this thesis is that optimised signal processing is essential to fully capture the metabolic content of samples. A further conclusion is that constraints methods can successfully be applied to metabolic profiling. Such methods have the potential to radically improve the quality of the metabolic results obtained from FT-ICR MS.

In loving memory of my Grandfather, Richard Langford.

ACKNOWLEDGEMENTS

I would like to thank my friends and family for their encouragement during the course of this research and while preparing this document. I especially thank Sophie, who has been very patient and understanding, especially during the last few months. My thanks also go to friends and colleagues with whom I have enjoyed university life, and who have provided much needed advice and help over the years. There are many, including John Easton, Alex Gibb (who first suggested utility theory), Greg Reynolds (for this LaTeX document template), Markus Saleh, Jie Hao, Yalda Danesh Sedigh, Andy Southam, Nadine Taylor, Ralf Weber and Olga Hrydziuszko.

I would like to thank Dr Theodoros Arvanitis for his help over the course of this research, and especially his encouragement which always seemed to appear when it was needed most. I am grateful to Dr Mark Viant (School of Biosciences, the University of Birmingham) and his research group, with whom I have worked closely from the beginning. Mark, as secondary supervisor, has shared experience and knowledge that has enabled achievements beyond my expectation. Members of Dr Viant's group have willingly provided data for this research, for which I am very grateful, and I have enjoyed collaborating with them to improve workflows. I also extend my thanks to Professor Roy Goodacre and Dr Rick Dunn for an interesting meeting that helped form some of the ideas in this work.

Finally, I thank the Engineering and Physical Sciences Research Council for funding me for three years — an essential contribution!

CONTENTS

1 Introduction **1**

1.1 Background 1
1.2 Challenges 4
1.3 Research Aims and Objectives 5
1.4 Contributions 5
1.5 Conclusions 7

2 FT-ICR Mass Spectrometry **8**

2.1 Mass Spectrometry as a Metabolomics Tool 8
2.2 Fourier Transform Ion Cyclotron Resonance Mass Spectrometry 10
2.3 Electrospray Ionisation 14
2.4 Automatic Gain Control and the Ion Trap 17
2.5 Quantification Performance 17
2.6 Processing of FT-ICR Data 18
2.6.1 Transient Acquisition 19
2.6.2 Conversion to Frequency Domain 20
2.6.3 Peak Centre Frequency 22
2.6.4 Peak Quantification 23
2.6.5 Noise Level Estimation 23
2.6.6 Signal-to-noise Level Estimation 24
2.6.7 Resolution 24
2.6.8 Conversion to m/z Domain and Internal Recalibration 25
2.7 Signal Processing Challenges 28
2.8 Conclusions 29

3 SIM-Stitching **30**

3.1 Introduction 30
3.2 Method 32
3.2.1 Overview 32
3.2.2 Sample Frequency Intensity Correction 33
3.2.3 Noise Level Estimation 33

3.2.4 SNR Estimation . . . 34
3.2.5 Resolution . . . 35
3.2.6 Noise Artefacts . . . 35
3.2.7 Intensity Correction . . . 36
3.2.8 Internal m/z Recalibration . . . 37
3.2.9 SIM Window Alignment . . . 39
3.2.10 Edge Effects and Concatenation . . . 41
3.3 Results and Discussion . . . 42
3.3.1 Mass Accuracy . . . 42
3.3.2 Sensitivity . . . 44
3.4 Conclusion . . . 50
3.5 Future Developments . . . 51

4 Noise Filtering **53**
4.1 Introduction . . . 53
4.2 Optimising Number of Scans . . . 56
4.3 Simulated Mass Spectra . . . 57
4.3.1 Assessing the Noise Filter . . . 63
4.4 Three Stage Noise Filtering . . . 63
4.4.1 Method . . . 63
4.4.2 Results and Discussion . . . 64
4.5 Conclusions . . . 69
4.6 Future Developments . . . 69

5 Interpreting the Mass Spectrum **72**
5.1 Profiling the Metabolome . . . 72
5.1.1 Accurate Mass . . . 73
5.1.2 Structural Information . . . 76
5.1.3 Biological Information . . . 77
5.1.4 van Krevelen Diagrams and Kendrick Mass Defect Analysis . . . 77
5.2 Search Strategy . . . 78
5.2.1 Current Solutions . . . 78
5.2.2 Model-Based Approaches . . . 80
5.2.3 Probabilistic Methods . . . 82
5.2.4 Constraints Optimisation Methods . . . 82
5.3 Challenges . . . 83
5.4 Conclusions . . . 84

6 Profiling the Metabolome as a Constraints Optimisation Problem **85**

6.1 Introduction . . . 85
6.1.1 Constraints Satisfaction . . . 86
6.1.2 Constraints Optimisation . . . 86
6.2 From Sample to Metabolite ID: a Model . . . 87
6.3 The Constraints Satisfaction Problem . . . 89
6.3.1 Viewpoint and Variables . . . 89
6.3.2 Constraints . . . 90
6.3.3 Solution Search . . . 91
6.4 The Constraints Optimisation Problem . . . 94
6.4.1 Optimisation Function . . . 94
6.5 Optimisation - Metrics and Utility Theory . . . 95
6.5.1 Mass Measurement Error . . . 96
6.5.2 Isotope Pattern Accuracy . . . 97
6.5.3 Presence of Adducts . . . 99
6.5.4 Likelihood of Molecular Formula . . . 100
6.5.5 Placing Value on a Molecular Formula Assignment using Utility Theory 103
6.6 Solution Search . . . 106
6.6.1 Branch and Bound Search . . . 106
6.6.2 Partitioning . . . 109
6.6.3 Redundant Constraints . . . 109
6.6.4 Dominance . . . 110
6.7 Results and Discussion . . . 111
6.7.1 Simulated Data . . . 111
6.7.2 Measured PEG Data . . . 115
6.7.3 Measured Biological Data . . . 117
6.7.4 Robustness to Noise . . . 122
6.8 Conclusions . . . 125
6.9 Future Developments . . . 126

7 Conclusions **128**
7.1 Future Developments . . . 130

LIST OF FIGURES

1.1 The stages involved in profiling the metabolome. The stages identified as a bottleneck are shaded. 2
1.2 An example mass spectrum. A narrow portion of the spectrum is enlarged, showing several peaks of varying abundance. 3

2.1 An example mass spectrum. 9
2.2 An enlarged portion of an example mass spectrum showing a compound peak. 10
2.3 FT-ICR MS block diagram 11
2.4 Schematic diagram of an ICR cell [1]. 12
2.5 The hypothetical transient signal for a single ion cloud. 14
2.6 Electrospray ionisation source, adapted from [2]. 15
2.7 Pre-Processing Data Flow 18
2.8 Transient Generation 19
2.9 The Hanning Window Function 21

3.1 Schematic of the optimized SIM-stitching method comprising 21 adjacent 30 m/z SIM windows, each overlapping by 10 m/z, covering a total scan range of 70–500 m/z. 31
3.2 Noise artefact present at ca. 101.9 m/z. (a) Spectrum over wide range. (b) A typical real peak in this region. (c) A series of anomalies (N.B. identical abscissa scales for (b) and (c)). 36
3.3 The intensities of several peaks in the range 352–354 m/z relative to their mean values within the central 10 m/z, measured across multiple SIM window m/z ranges. Considerable variation in peak intensity can be seen as the SIM window position moves across the peaks. The broken line is a visually-placed trend line. 38
3.4 The gradient of the measurement intensity error for a selection of peaks located between 70 and 500 m/z. The line of best-fit is shown. 38
3.5 Two overlapping SIM windows 40

3.6 The observed edge effect at the start of the SIM window (shown in blue with diamond markers) and the end of the SIM window (shown in green with asterisk markers). The representative edge effect region selected at the start and end of each SIM window is a function of the midpoint of the overlap with the adjacent SIM window, and is shown as a broken line for both the start and end of the SIM window (lower and upper lines, respectively). . . . 41

3.7 Location of internal calibrants and mass errors of all known non-calibrant peaks as a function of m/z for one replicate of the PEG&AA standard analysed by (a) WSR mode with an AGC target of 1×10^5, (b) WSR mode with an AGC target of 5×10^5, and (c) SIM-stitching method. SIM-stitched data was calibrated using 16 internal calibrants, while the WSR methods were calibrated using 14 of these; two were missing due to lack of sensitivity of the WSR method. 46

3.8 Comparison of the total number of peaks detected in the range 70–500 m/z for six different dab liver extracts using wide scan range (WSR) mode with an AGC target of 1×10^5 ions and 5×10^5 ions, and the optimised SIM-stitching method with an AGC target of 1×10^5 ions. 49

4.1 The three stage filtering schema, showing SNR threshold, replicate filter with parameter r, and sample filter with parameter s. 56

4.2 Median RSD of three spectra with mass ranges 110–140 m/z (a), 230–260 m/z (b), and 470–500 m/z (c) as the number of scans averaged to form each spectrum, n, is increased. The experiment was repeated three times each, shown as dotted lines, with the average as a solid line. The maximum number of scans in the range 110–140 m/z is limited by the second experiment repeat, which produced an insufficient number of total scans (66 obtained, 90 required). 58

4.3 Mean ($n = 10$) histogram of peak SNR in measured spectra (shaded bars) and simulated noise spectra (clear bars) and the difference between these relative to the mean number of peaks in the measured spectra, plotted as the thick line. The histograms show very good correspondence at low SNR (< 2.5). The error bars represent the range. 60

4.4 Empirical CDF of the population of real peaks (estimated from the difference between the simulated spectrum and the measured spectrum) showing the probability that the SNR is below the corresponding SNR on the x-axis. The broken line shows the linear interpolation for all SNR values. 61

4.5 Distribution of the intensity of replicate peaks relative to the mean intensity of the peak across all replicates. A normal distribution is shown fitted to the data. Two outliers, both at -1.4, are not shown. 62
4.6 Contour plots showing the quality of the simulated spectrum, after three-stage filtering has been applied for varying SNR threshold and sample filter, and fixed replicate filter, and quantified in terms of: (a) number of real peaks correctly identified, (b) probability that a detected peak was real, and (c) the product of (a) and (b). 68
4.7 The schema of SIM-stitch. The red boxes represent essential files, while the purple boxes represent temporary and output files. The '.dat' files are the transient files generated by the instrument. The '.raw' files are also generated by the instrument and contain essential information regarding the transient files, and instrument configuration. The stages in processing are shown, from summation (averaging) of the transient files, Fourier transformation, SIM-stitching and finally the three-stage filter. A separate module allows the spectral results at most stages of the processing to be viewed. . 70
4.8 The SIM-stitch graphical user interface. 71

5.1 Structural diagram of isomers sucrose (ID C00089) and maltose (ID C00208), as present in the KEGG database [3]. 73
5.2 Information sources of compound identification using MS. 73

6.1 Representation of the system as a data flow diagram. The asterisk (*) indicates the stage at which the peak list is generated. 87
6.2 Representation of mass spectrometry measurement as a series of transformations between components: molecular formulae, mf; monoisotopic ions plus adducts, $adion$; isotopic ions plus adducts, $isoion$; and mass spectral peaks, mz. 90
6.3 Example of representing the presence of tryptophan in a compound via a sodium adduct transformation, showing in full the presence of the [M+3] carbon-13 isotope and resulting in a peak occurring at 230.0892 m/z. . . . 92
6.4 An example system graph, showing four observed peaks, and three potential molecular formulae. 93
6.5 Isotope pattern accuracy, $isoacc$, for a peak pair as a function of measured ratio (mr) and expected ratio (er). 98
6.6 Example isotope pattern consisting of three peaks 98

6.7 Ratio of hydrogen to carbon atom count for appropriate compounds with unique molecular formulae in index range 00000001 to 10000001 in PubChem database [4], giving a total of 156,032 unique molecular formulae included. 101
6.8 Feature distributions of all unique molecular formulae in index range 00000001 to 10000001 in the PubChem database [4]. 102
6.9 Mahalanobis distance for appropriate compounds with unique molecular formulae in index range 10000001 to 10025001 in PubChem database [4], giving a total of 11,467 unique molecular formulae included. 103
6.10 Utility graphs for measurement mass error; isotope pattern; molecular likelihood and adduct pattern. 105
6.11 Representing a COP search as a constraints graph, for a simple example consisting of variable $X = (x_1, x_2)$. The numbers in parentheses indicate the assignments at each node to the two variables. Nodes where the variables are not yet fully assigned are shown as circular; 'failed' nodes in which the full or partial solution does not satisfy the constraints are triangular; and complete, valid solutions are represented by squares. 107
6.12 The search assignments when the order of the search variables is changed. . 108
6.13 Example search space split into two independent partitions A and B. . . . 109
6.14 A sub-graph showing two alternative potential solutions in mf_1 and mf_2: dominance rules can eliminate one of these solutions. 110
6.15 A sub-graph showing three alternative parallel solutions in mf_1, mf_2 and mf_3, with molecular likelihood utility x_1, x_2 and x_3, respectively. In the case that $x_1, x_3 < x_2$, dominance rules can again be used to eliminate the two solutions containing mf_1 and mf_3 from the search. 111
6.16 Simulated spectrum (noise-free), including isotopes. 113
6.17 Utility and abundance of all molecular formulae identified as present in the simulated spectrum, showing the effect of including only carbon-13 and potassium-41 isotopes in the solution search. 114
6.18 Utility and abundance of all molecular formulae identified as present in the spectrum of PEG sample. 118
6.19 Utility and abundance of all molecular formulae identified as present in the biological sample. Molecular formulae previously identified in the sample are distinguished from those that are unknown. 119
6.20 Histogram of utility values U_m for molecular formulae assignments, random data. 124
6.21 Histogram of utility values U_m for molecular formulae assignments, biological data. 124

LIST OF TABLES

3.1 Two-term vs three-term calibration. 39
3.2 Summary of observed edge effects 42
3.3 The identity, empirical formulae, type of adduct and exact mass of the 16 internal calibrants used to calibrate the PEG&AA standard data. 44
3.4 Summary of mass errors for SIM-stitching and WSR methods (with AGC targets of 1×10^5 and 5×10^5), determined from analysis of the PEG&AA standard. Each method was characterised in triplicate, and the numbers of internal calibrants (N_C) and other known peaks used to calculate mass errors (N_P) are shown. 45
3.5 The identity, empirical formula, form of ion and exact mass of the internal calibrants used to calibrate the dab liver extract data collected using the SIM-stitching method. 47

4.1 Metrics for simulated spectra, filtered by SNR threshold only: mean number of real peaks, mean number of noise peaks and mean probability that a detected peak is real, based on 10 repeats. 64
4.2 Metrics for simulated spectra, filtered by SNR threshold and replicate filtered: mean number of real peaks, mean number of noise peaks and mean probability that a detected peak is real, based on 10 repeats. 65
4.3 Metrics for simulated spectra and three stage filtering: number of real peaks, number of noise peaks and probability that a detected peak is real, based on a single repeat. 65

5.1 Proline and all naturally-occurring isotopes with relative abundance over 0.01%. Isotopic elements are underlined. 76

6.1 Adduct pattern classification. 100
6.2 Summary of molecular formula features. 102
6.3 Variable search order and order of value assignment. 108
6.4 Simulated spectrum composition. The total abundance was split approximately equally between the adducts. 112
6.5 Simulated spectrum profiling results. 114

6.6 Simulated spectrum noise peak assignments 115
6.7 Identification of PEG compounds in control sample. The table shows the number of peaks for each different adduct type observed. For example, values in the column '+H' indicate the number of $[M+H]^+$ ions found. The number in parentheses indicates the number of isotope peaks, with (1) representing a single carbon-13 isotope and (2) a double carbon-13 isotope. An asterisk (*) indicates that the peak, while present, has been assigned to a different molecular formula. 117
6.8 Compounds with mass $< 300\ \mathrm{Da}$ likely to be present in cancer cell-line experiment data. An asterisk (*) indicates that the peak, while present, has been assigned to a different molecular formula. 120

LIST OF COMMON ABBREVIATIONS AND SYMBOLS

Abbreviation / Symbol	*Meaning*
$\mathcal{F}$	The Fourier operator.
μ	Mean.
σ	Standard deviation.
AA	Amino acid.
ADC	Analogue to digital converter.
AGC	Automatic gain control.
AWGN	Additive white Gaussian noise.
C	The element carbon.
ca	Approximately.
CDF	Cumulative density function.
CSP	Constraints satisfaction problem.
COP	Constraints optimisation problem.
Da	Dalton, equivalent to the unified atomic mass unit. One carbon-12 atom has a mass of 12 Daltons.
DC	Direct current.
DI	Direct injection.
e	The charge on a single electron, measured in Coulombs.
ESI	Electrospray ionisation.
FFT	Fast Fourier transform, an implementation of the FT.
FT	Fourier transform.
FT-ICR MS	Fourier transform ion cyclotron resonance mass spectrometry.
GC	Gas chromatography.

Abbreviation / Symbol	*Meaning*
H	The element hydrogen.
Hz	Hertz.
ICR	Ion cyclotron resonance.
K	The element potassium.
LC	Liquid chromatography.
LNA	Low noise amplifier.
MME	Mass measurement error.
MS/MS	Two-stage mass spectrometry, where fragmentation occurs in between the stages.
MS^n	n-stage mass spectrometry, where fragmentation occurs in between the stages.
m/z	Mass-to-charge ratio, where mass is measured in Daltons and charge in multiples of e.
N	The element nitrogen.
nESI	Nano electrospray ionisation.
NMR	Nuclear magnetic resonance.
Na	The element sodium.
O	The element oxygen.
P	The element phosphorus.
PDF	Probability density function.
PEG	Polyethylene glycol.
ppm	Parts per million.
RF	Radio frequency.
RMS	Root mean square, a statistical measure of a quantity's magnitude.
RSD	Relative standard deviation.
S	The element sulfur.
Si	The element silicon.
SD	Standard deviation.
SIM	Selected ion monitoring, an FT-ICR MS measurement mode.
SWIFT	Stored wave inverse Fourier transform.
T	Tesla.
TIC	Total ion current.
WSR	Wide scan range, an FT-ICR MS measurement mode.

CHAPTER 1

INTRODUCTION

1.1 Background

The need to understand the state of health of an organism can be traced back to ancient times when, for example, the sweetness of urine was recognised as an indicator of the presence of diabetes in a person [5]. In those times, a human 'taste-test' was used to measure the sugar level in the urine of a person, and from that deduce the ability of the person's body to control sugar levels; poor control can indicate that they are diabetic. The human body's control of sugar levels in the blood is just one example of the many chemical processes that exist within organisms to support life.

The small-molecule compounds involved in such chemical processes are called *metabolites* [6, 7, 8, 9, 10], of which around 3,000 are thought to exist in the human body [11], and up to 200,000 are estimated to exist in the plant kingdom [12]. They include lipids, amino acids, sugars, vitamins and hormones [13, 8].

When the organism processes are disturbed, for example as the result of external stress or a disease, the chemical processes are altered and so the quantities of metabolites present in the organism are affected. In the case of diabetes, an increase or decrease in the quantity of sugar molecules in urine is one such effect, as discovered many years ago. Metabolomics is the discipline of studying the *metabolome*, the collection of metabolites present in a cell, organ or organism, in order to gain knowledge about the organism and its reaction to external stressors, for example. Metabolites provide a direct and fast-changing indication of the current chemical 'situation' within an organism, and hence metabolomics holds an advantage over genomic, transcriptomic and proteomic approaches in assessing the current phenotype of an organism [11]. The discussion of these methods is beyond the scope of this thesis, suffice to say that they are all of great value in understanding biological organisms to the extent that the rapidly developing field of 'systems biology' [14] is dedicated to the

integration of these 'omic' methods together with advanced computing and mathematical techniques.

Profiling the metabolome describes the process of identifying and quantifying all metabolites of a certain class (e.g. lipids or carbohydrates) within a sample [12, 15]. Focussing on a particular class of metabolites offers the benefit that the sample preparation can be optimised for these compounds of interest [12]. Several different techniques are available to measure metabolites, from simple chemical assays to highly complex *mass spectrometry* (MS) instruments. The type of MS used in this work, Fourier transform ion cyclotron resonance (FT-ICR) MS, is discussed in detail in Chapter 2. The stages involved in using MS to profile the metabolome are shown in Figure 1.1. A mass spectrometer generates a signal that, once processed into a mass spectrum, shown in Figure 1.2, allows the molecular composition of the sample to be known in terms of molecular mass and a corresponding intensity. Figure 1.2 shows a portion of the spectrum containing potentially three real peaks, with corresponding m/z and intensity values for each. The mass spectra contain information regarding the composition of the sample. The aim of the data mining stage is to interpret the peaks in the spectra as molecules, and identify the composition of each molecule. From the quantity of material published concerning analysis of MS data, it is apparent that these data processing and mining stages are a bottleneck, as indicated in the Figure 1.1.

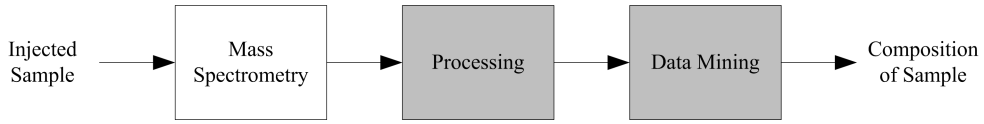

Figure 1.1: The stages involved in profiling the metabolome. The stages identified as a bottleneck are shaded.

Although this thesis uses FT-ICR MS to profile the metabolome, the methods developed are more widely applicable. There are other types of MS that produce high resolution mass spectra, including the orbitrap. The signal processing methods described in Chapters 3 and 4 are broadly applicable to such instruments. These techniques are also relevant when measuring molecules other than metabolites, such as proteins. Proteins are generally larger than metabolites, but mass spectrometry can be applied in the same manner, and consequently similar problems arise. Furthermore, the profiling methods discussed in Chapters 5 and 6 are not exclusively applicable to metabolomics data, but can be applied to the problem of mining mass spectra of other, larger, molecule types; the task is essentially the same.

The broad experimental challenges tackled in this work are three-fold. The first two challenges are to increase the sensitivity of the mass spectrometer and to reduce noise,

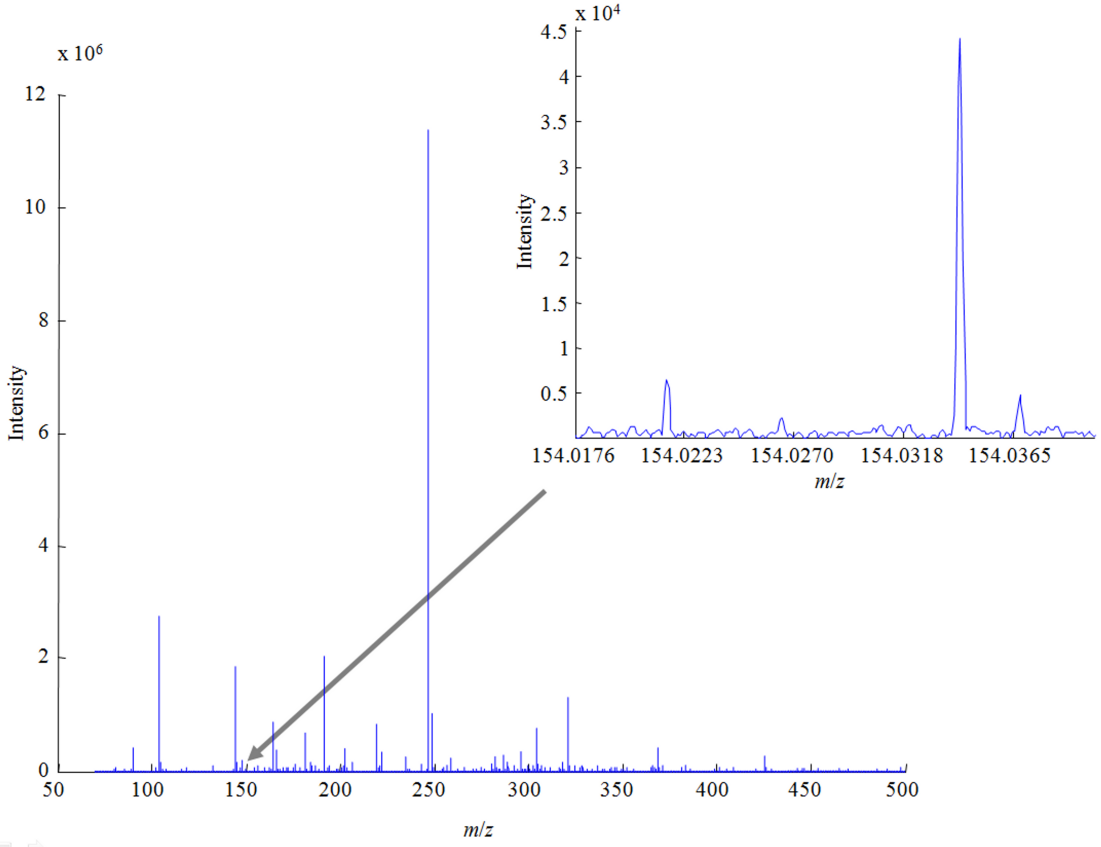

Figure 1.2: An example mass spectrum. A narrow portion of the spectrum is enlarged, showing several peaks of varying abundance.

as described in Section 1.2, and are applicable wherever MS is used in chemistry and biochemistry. The contributions of this thesis, as described in Section 1.4, result in high quality spectra that provide the chemist and biochemist with more information and less noise. The third challenge addressed in this thesis is to identify the patterns present in the complex, and even with the above improvements, noisy, mass spectra. As described in Section 1.4, knowledge from the expert mass spectroscopist is captured and used to mine the data contained in the spectra and identify the composition of the sample molecules.

In summary, the methods developed in this thesis aim to optimise the processing and mining of mass spectrometry data, and consequently assist scientists in answering experimental questions by providing more information about the samples that they are investigating. For example, they have been applied by Taylor *et al.* [16], who demonstrated the use of metabolomics in determining the mode of action of copper on the water flea species *Daphnia magna*, a widely used subject for toxicity studies. After exposure to the toxin, metabolites were extracted from the *Daphnia magna* using the protocol described in the paper. FT-ICR MS was used to detect and quantify the metabolites in the extract. A subset of metabolites was identified for which the concentration present in the extract changed significantly after exposure to high concentrations of copper. Within the set of metabolites identified as being affected by copper toxicity, a novel biological indicator for copper toxicity was identified, *N*–acetylspermidine. This experiment thus showed how FT-ICR MS metabolomics can be used to screen for *Daphnia magna* chemical toxicity, and reveal new, valuable information to the biologist.

1.2 Challenges

Three major challenges are identified in the analysis of metabolomics data, as listed below. From these challenges, the objectives of this thesis are drawn as listed in section 1.3.

1. As discussed in Chapter 2, the design of an FT-ICR mass spectrometer is such that there is a compromise between mass accuracy and the sensitivity of the instrument in detecting low abundance molecules. This is an important issue for two reasons. Firstly, the quantity of some metabolites in the sample can be very low; indeed Markley *et al.* [17] found that the variety of metabolites observed in a biological sample increases significantly as the detection sensitivity increases. To profile the metabolome as completely as possible, it is necessary to detect as many metabolites as possible. The second benefit of increased sensitivity is to improve the identification of metabolites from the spectrum. As discussed in Chapter 5, it is generally insufficient to observe a compound in isolation; additional evidence within the spectrum is necessary to distinguish between alternatives. Therefore, the more detail

in the spectrum, the more confidence that can be placed in compound identification. The challenge this presents is to maximise sensitivity of FT-ICR MS, without compromising mass accuracy.

2. Both variability and noise are present in the measurements. Variability occurs between samples as a result of phenotypical variations in the biology of samples, for example between two distinct fish samples. Noise is present in the signal as a result of both random and systematic influences on the instrument. One effect of the variability and noise is to reduce the reliable detection of real signal and increase the false detection of non-real signal. The challenge is to distinguish signal from noise without negatively impacting on the increased sensitivity discussed above.

3. Arguably the 'holy grail' of MS-based metabolomics is metabolic profiling — the identification of metabolites from the data acquired by mass spectrometry. As shown in Chapter 5, the signal is complex, noisy and, at 'ultra-high' resolution, extremely dense in terms of the amount of information it contains. This presents a challenge to the data mining process, which must relate this signal to unique metabolites without being obfuscated by the noise in the signal, nor missing the low abundance metabolites that are known to be prevalent.

All of these challenges need to be addressed in the context of high throughput analysis of biological samples.

1.3 Research Aims and Objectives

The aim of this thesis is to present novel, improved methods for the processing and subsequent analysis of metabolomics-based FT-ICR MS data. The objectives are three-fold:

1. To improve the signal from the mass spectrometer in terms of sensitivity, while retaining mass accuracy and high throughput. This is important to maximise the potential for extracting as much information as possible from the spectrum, and subsequently, interpreting the data.

2. To reduce noise that is present in the spectrum and hence improve the data for the subsequent interpretation stage.

3. To 'profile the metabolome' by extracting information relating to the metabolites present in a sample, using constraint satisfaction methods applied to processed mass spectrometry data.

1.4 Contributions

This thesis offers several novel contributions to metabolomics based FT-ICR MS data. The first is an optimised signal-processing method named 'SIM-stitching' [18, 19, 20], described in Chapter 3. Compared to the mode typically used to acquire spectra, SIM-stitching is shown to offer a 5.3-fold increase in metabolite sensitivity and a 1.3-fold improvement in mass accuracy. While such improvements typically require extended acquisition time, it is shown that SIM-stitching facilitates high throughput experiments, with typical analyses having a duration of 5.5 minutes on a standard desktop computer. The algorithm is implemented in the MATLAB* mathematical programming environment. Together with a graphical user interface (GUI), the software has been made publicly available, with requests received from several international laboratories.

The second contribution is a novel noise-filtering method [21] that enables MS measurement noise to be reduced. Often, a hard threshold is applied to the measured intensity of metabolites in order to classify them as signal or noise. However in many experiments, replicates of the same sample as well as multiple similar samples are measured. It is shown in Chapter 4 that these *inter*-sample as well as *intra*-sample spectra are valuable in improving noise filtering. The noise filtering algorithm is also implemented in MATLAB and integrated with the SIM-stitch methods developed.

The third contribution offers the novel application of constraints satisfaction and optimisation methods to metabolic profiling. As shown in Chapter 6, using this approach provides a framework for accurately capturing current knowledge about metabolic based mass spectrometry data that is not currently possible. By applying the knowledge to interpret the data in an optimal manner, the results show that the method yields accurate information about the composition of a sample.

Two distinct stages in handling the FT-ICR MS data are presented, and so this thesis is broadly split as outlined below:

1. Chapters 2 to 4 cover the processing of the raw data produced by FT-ICR MS. In particular, Chapter 2 provides a description of the instrument, its theory of operation, methods used to process the data, including calibration and extraction of the salient features from the spectrum. Chapter 3 discusses shortcomings in the instrument manufacturer's software that introduce additional noise in the data and describes methods to overcome them. The rest of the chapter addresses the first objective through the development and application of the SIM-stitching method to the generation of mass spectra. In Chapter 4, the second objective relating to noise

*MATLAB 7.4.0 (R2007a), The MathWorks, Natick, MA

filtering is met through the use of a three-stage noise filter that incorporates cross-replicate and cross-sample filtering as well as threshold noise filtering.

2. Chapters 5 and 6 relate to the mining of the FT-ICR MS data. In Chapter 5, the relation between the measured data and the metabolic content of a sample is explored, and it is shown how metabolites are identified in mass spectra. Chapter 6 meets the third objective by showing how constraints satisfaction methods can be applied to the data mining problem described in Chapter 5.

3. Chapter 7 presents concluding remarks relating to both the above, and discusses how the integration of these achievements contributes to the thesis aim.

1.5 Conclusions

The aim of this thesis is to improve metabolic profiling, which is the process of measuring and identifying metabolites of a certain class within a sample. This can be achieved using FT-ICR mass spectrometry, which is described in Chapter 2. It is important to maximise the sensitivity of FT-ICR MS, and the SIM-stitching method is described in Chapter 3 that achieves this without compromising mass accuracy. Furthermore, there is noise present in the spectrum that can obscure the signal. Chapter 4 presents a three-stage noise filtering method to reduce the spectral noise. Mining the data in the mass spectra is a difficult task due to the complexity of the spectra, as described in Chapter 5. A new approach to solving this problem using constraints optimisation is described in Chapter 6.

CHAPTER 2

FT-ICR MASS SPECTROMETRY

The first two objectives presented in Section 1.3 relate to improving the signal from the mass spectrometer in terms of sensitivity, accuracy and the signal-to-noise ratio (SNR). In order to meet these objectives, it is necessary to understand the measurement process right through from sample to signal, and at each stage identify opportunities for improving the final signal. To this end, this chapter explores mass spectrometry (MS), its relevance to metabolomics, the hardware and data processing associated with Fourier transform ion cyclotron resonance (FT-ICR) MS, and concludes by highlighting the specific challenges in the signal processing of metabolomics-based FT-ICR MS data which should be met. In Chapters 3 and 4, novel data processing methods will be presented that significantly improve FT-ICR MS data.

2.1 Mass Spectrometry as a Metabolomics Tool

A mass spectrometer is an instrument that resolves the exact mass of a substance's constituent molecules [22]. The measured mass of the molecules, if observed with sufficiently high accuracy and resolution, allows the empirical molecular formula, i.e. the chemical composition, of the molecule to be deduced. In some cases isomers exist, which are molecules with the same composition but different structures, and so additional information may be needed to identify certain compounds.

Common to the variety of mass spectrometry methods is an initial sample preparation stage, in which the molecules of interest are isolated from other entities present in the sample. This is followed by the process of ionising the molecules to form charged ions. The ions are then detected to yield a spectrum plotting the quantity of each unique molecular ion as a function of their mass-to-charge ratio, an example of which is shown in Figure 2.1. In general, the quantity of each unique molecular ion is approximated from the magnitude

of the current induced on a conductor. The mass-to-charge ratio is the ion's mass number, m, in units of Dalton (Da), divided by the number of elementary charges associated with the ion, z [23]. 1 Da is equivalent in mass to $\frac{1}{12}$ of a carbon atom; metabolites are small molecules and have a molecular weight of up to ca. 1000 Da [6, 24, 13].

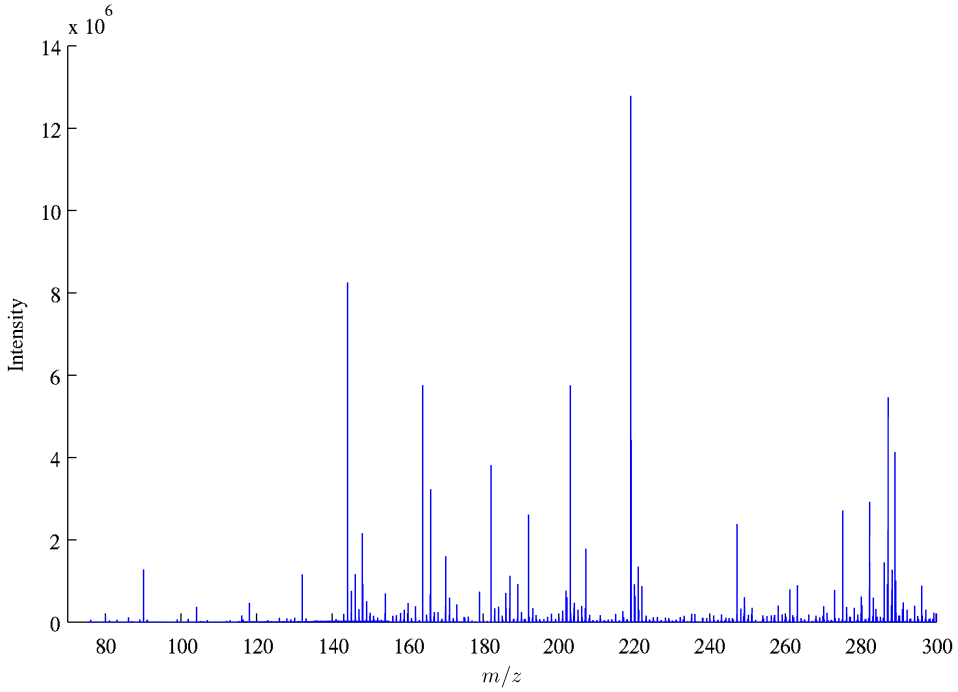

Figure 2.1: An example mass spectrum.

The mass spectrum allows the composition of molecules to be identified as a result of the known exact masses of the elements that constitute them. As an example, consider the case of a peak present in a spectrum at location 181.0739 m/z, as illustrated in Figure 2.2. Small molecular ions measured using FT-ICR MS are almost exclusively singly-charged [25], i.e. it is safe to assume that $z = 1$. This can be confirmed by inspecting the location of isotope peaks, which will be at half the expected distance if $z = 2$, for example. Therefore, the ions have a mass of 181.0739 Da. A search is carried out to identify all combinations of the common elements present in biological samples that, when combined, form a molecule of mass 181.0739 Da. In this case, one close match is the empirical formula $C_9H_{11}NO_3$. This formula is found from a compound database to be the molecule tyrosine, a metabolite commonly detected in metabolic profiling experiments [26]. In general there are multiple different combinations of elements that could correspond to a single peak in the spectrum, and so additional information in the mass spectrum is required. Typically, this information takes the form of peaks present in the spectrum that indirectly relate to other compounds. This is discussed in more detail in Chapter 5.

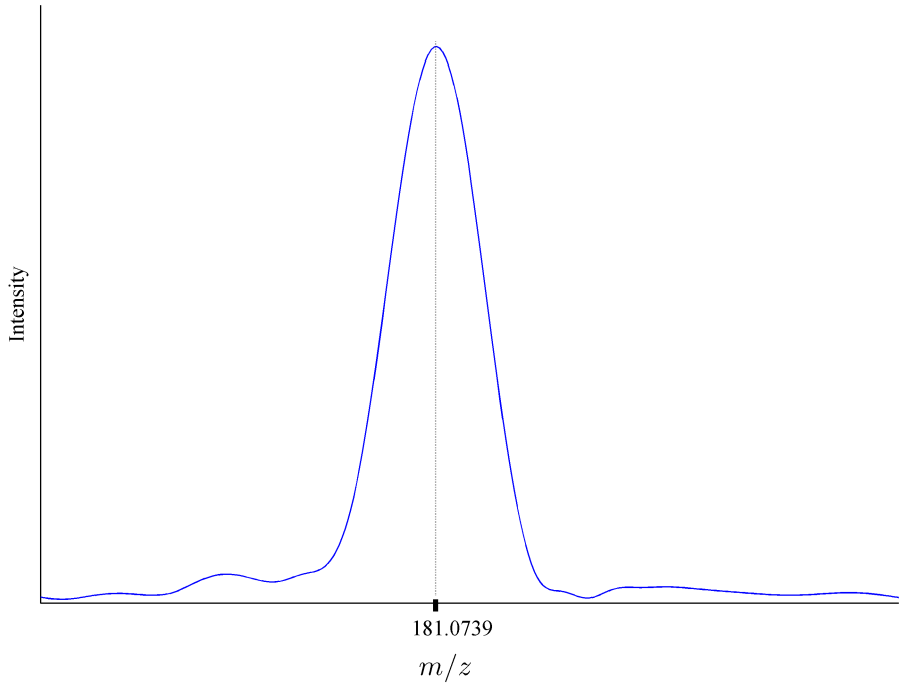

Figure 2.2: An enlarged portion of an example mass spectrum showing a compound peak.

2.2 Fourier Transform Ion Cyclotron Resonance Mass Spectrometry

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) first appeared in 1968, with a paper in the journal Science by Baldeschwieler [27] describing an instrument with the same operating principles as modern FT-ICR instruments. Although not Baldeschwieler's main use for the instrument in his work, he noted that the instrument could also be used as a "rather good" mass spectrometer. This was followed up in 1974 by Comisarow and Marshall [28, 29], who promoted the use of FT-ICR MS and contributed greatly to the early development of the instrument.

In FT-ICR MS, ions are held in a circular orbit by a strong magnetic field. Provided that all ions have been given the same energy in the form of excitation by an electric field, each ion will orbit with a frequency that is related to their mass-to-charge ratio. Smaller, or more charged, ions will travel faster in the magnetic field than larger, or less charged, ions. The frequency of the orbit, known as the 'cyclotron' frequency, is measured. A Fourier transform, when applied to the signal from multiple ions of different mass-to-charge ratio, allows the signal to be decomposed into a spectrum. From the frequency measurements, the mass-to-charge ratio of the ions can be determined. The relationship between cyclotron

frequency and mass-to-charge ratio is calculated from a calibration equation, as described in Section 2.6.8.

Mass spectrometry consists of two main sequential components: an ionisation stage, in which molecules are charged; and a detection stage, in which the mass and intensity of the ions are measured. Figure 2.3 is a diagram of an FT-ICR MS, showing the electrospray ionisation (ESI) part followed by the two detection components, which are separated by an ion 'gate'.

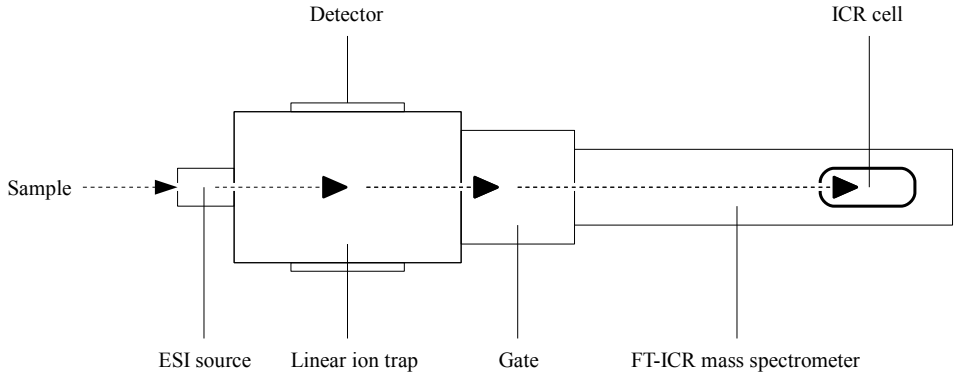

Figure 2.3: FT-ICR MS block diagram

In Figure 2.3, the flow of molecules is shown with arrows. The sample enters the instrument as a liquid and is simultaneously ionised and evaporated into a gas by the ESI source, which is described in detail in Section 2.3. After ESI, the molecules are now either positively or negatively charged ions, corresponding to the removal or addition of an electron, respectively. The ions now enter the first detector stage, the 'linear ion trap', which is a mass spectrometer in its own right. The principle of operation is similar to that of the FT-ICR detector as discussed below, except that detection is destructive and involves detectors positioned on the outside walls of the detector; the specific details are beyond the scope of this thesis. Importantly, when not in detection mode, the ion trap can be used as discussed in Section 2.4 to control the flow of ions, via the gate, to the ion cyclotron resonance (ICR) cell for measurement.

In the ICR cell, each ion, travelling with velocity v, and having charge q, is subjected to a strong magnetic field, B. Each ion experiences the Lorentz force F:

$$F = qvB, \tag{2.1}$$

which according to the right-hand rule, acts in a direction perpendicular to the direction of travel of the ion, thus causing the ion to travel in a circular orbit. Also acting on the ion is the centrifugal force, of equal magnitude but acting away from the centre of the

orbit, with magnitude mv^2/r, and so

$$mv^2/r = qvB, \tag{2.2}$$

where m is the mass of the ion and r is the radius of the orbit. By combining equations (2.1) and (2.2),

$v = qrB/m$,

and since the angular frequency of orbit $\omega = v/r$ [1], the relationship between ion cyclotron frequency, mass and charge is revealed in equation (2.3), known as the 'cyclotron equation'.

$$\omega = qB/m \tag{2.3}$$

This equation shows that ions with the same mass-to-charge ratio will exhibit the same frequency of motion, notably *independent of their velocity*. The purpose of the ICR cell, therefore, is to contain, excite and detect the ions so that the cyclotron equation can be applied to determine their mass-to-charge ratios. A simple schematic of the cell is shown in Figure 2.4, which is continuously vacuum-pumped to maintain very low pressure, and therefore reduce the number of collisions between ions of interest and foreign matter.

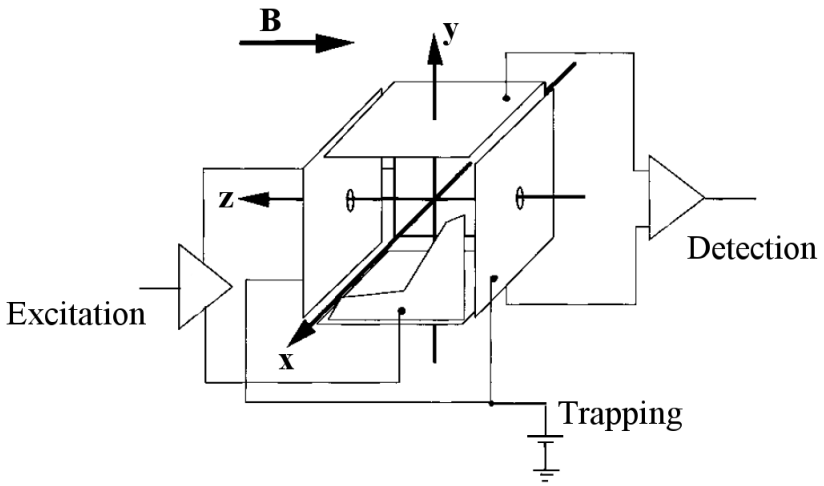

Figure 2.4: Schematic diagram of an ICR cell [1].

On entering the ICR cell, ions of the same mass-to-charge ratio will initially be moving non-coherently, i.e. with random phases [30], and in a small orbit. Without coherent motion, ions would be undetectable since there would be similar amounts passing each detector plate simultaneously, and hence there would be no net current induced in the detector

plates. In order to excite ions of the same mass-to-charge ratio into a coherent orbit, a radio frequency (RF) signal is applied with direction perpendicular to the magnetic field B, i.e. along the x axis in Figure 2.4. Known as the *excitation waveform*, the frequencies in this signal are important in determining the final radius of orbit of each ion [31], since it adds energy to ions according to their resonant frequency. The excitation waveform is shaped such that ions within a m/z range of interest are excited into a detectable orbit, and any ions outside this range are removed by being excited into orbits beyond the cell dimensions. One successful technique to achieve this ion selection and excitation is to use a stored waveform inverse Fourier transform (SWIFT) waveform [31], where the time domain RF waveform is created by inverse FT of the spectrum for the band-pass filter required. The end result of this excitation is that ions will now be travelling in coherent 'clouds' of the same mass-to-charge ratio.

To prevent ions drifting along the length of the ICR cell, a further set of plates maintain a low intensity steady electric field, known as the 'trapping potential', along the z axis, as shown in Figure 2.4.

In order to measure the frequency of orbit of the ions, the ICR cell contains parallel detector plates, positioned at opposite ends of the cell, as shown in Figure 2.4. When an ion cloud passes close to either of the detector plates, a net image current is induced in the plate [1]. By measuring the induced current in the detector, a time domain 'transient' signal is obtained. For the hypothetical case of an isolated ion cloud, the transient has the form of an exponentially decaying sinusoid, as shown in Figure 2.5. This decay is due to collisions of the ions with other molecules in the ICR cell, present since a perfect vacuum is not possible within the cell. These collisions cause the ions to fall out of orbit at a rate proportional to the number of ions still in orbit, causing an exponentially diminishing signal from the detector plates. Hence, ions can be measured within the ICR cell for a limited time of typically one second.

When multiple ion clouds are present in the ICR cell, by the Principle of Superposition [32], the induced signal $y(t)$ is the summation of the induced signals for each ion cloud, as shown in equation (2.4):

$$y(t) = \sum_{i=1}^{N} A_i \sin\left(2\pi f_i t T + \Phi\right) \exp\left(-\lambda t T\right), \tag{2.4}$$

where N is the number of ion clouds in the cell; A and f are the observed ion cloud intensity and cyclotron frequency, respectively; Φ and λ are the phase and decay constant, respectively (both assumed constant and independent of ion mass); and T is the duration of the transient sampling.

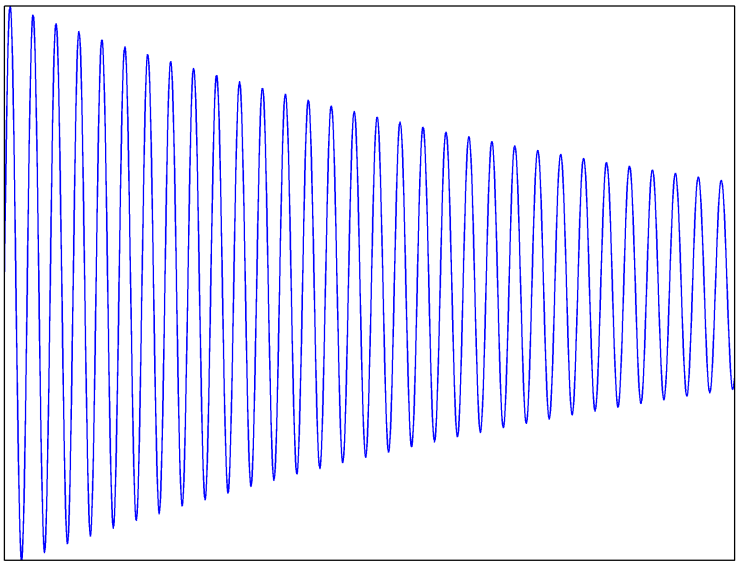

Figure 2.5: The hypothetical transient signal for a single ion cloud.

As evident from equation (2.3), the cyclotron equation, the frequency of each component i in equation (2.4) is proportional to the m/z of the ion cloud, and the intensity is proportional to the number of ions in the cloud, assuming the orbit radii are equal. In this way, the quantity and m/z of an ensemble of ions can be determined simultaneously from a single transient [30].

2.3 Electrospray Ionisation

Electrospray ionisation (ESI) provides a short-term steady flow of charged molecules (ions) to the ICR cell. Different types of injection and ionisation are available, with the simplest being direct injection (DI-) ESI, in which the sample is constantly sprayed directly into the mass spectrometer at a very low volume rate, while simultaneously being ionised. Alternatively, chromatography, such as liquid chromatography (LC), can be used, where ions are passed to the ESI source as they elute from a column [6, 26]. Chromatography separates the molecules in the sample according to specific characteristics such as affinity with the chromatography matrix, thus providing additional information about the molecules according to the time at which they elute. However, this benefit is at the expense of increased sample preparation effort and reduced throughput [6]. DI-ESI also offers a simpler data set and a significantly more reproducible m/z axis [18] and is consequently the injection method used throughout this work.

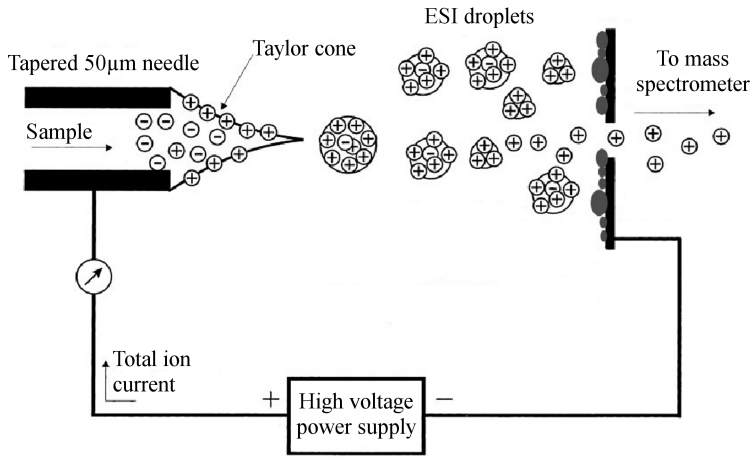

Figure 2.6: Electrospray ionisation source, adapted from [2].

Ions are formed in stages [33, 23]. In the first stage, a potential difference (shown in Figure 2.6) between a fine needle and a metal plate causes the molecules to leave the needle. This creates a cone-shaped spray, known as the Taylor cone [34], which forms a continuous jet. Next, the jet turns into a plume of droplets as it evaporates and breaks apart. These small, charged droplets continue to break apart while travelling towards, and into, the mass spectrometer. Finally, ions are separated from the droplets, either during ion evaporation in which ions escape the droplets, or during solvent evaporation which leaves ions behind. The total ion current (TIC) measures the total number of ions that are being created, as shown in Figure 2.6.

As illustrated in Figure 2.6, the ions leave the tapered needle and form a Taylor cone. The point at which the ions leave the needle can become 'dirty' as ions stick to the needle, much like paint clogging the jet of a spray can, and consequently the needle must be changed or cleaned frequently. One solution is to use disposable needles such as nanoESI, as described below.

FT-ICR MS can operate in either positive or negative ion mode, corresponding to the ESI source producing either positive or negative ions. This is achieved by providing a surplus of protons or electrons, respectively, to the sample within the needle. Both modes have their advantages and ideally both should be acquired [35].

ESI is a relatively 'soft' ionisation method, meaning that during the process, molecules are less likely to fragment than with other techniques [33]. This results in 'cleaner' mass spectra with less signal due to molecular fragments, which thus simplifies the interpretation of the spectra. It is for this appealing feature that John B. Fenn was awarded a share of the Nobel Prize in Chemistry in 2002, "for their development of soft desorption ionisation methods for mass spectrometric analysis of biological macromolecules" [36]. The

overall process is not well modelled or understood and involves many processes including evaporation, collisions, ionic interactions, thermal effects and some fragmentation. Two competing models that attempt to explain the process are the ion evaporation model and the charge residual model, discussed in detail by Kebarle [37]. Their discussion here is beyond the scope of this thesis.

The effect of using ESI is that the population of ions formed does not accurately reflect the sample molecular composition. While this leaves ESI FT-ICR MS poorly suited to the quantitation of compounds, the use of ESI is highly suitable for the *identification* of compounds, mainly due to the soft ionisation that occurs. Experiments reveal that during the ESI process, up to four main classes of ion may be formed for a molecule [M]:

1. The molecular ion $[M]^{\pm}$ — resulting from the modification of the molecule by the loss or addition of an electron;
2. (De-)protonated ion $[M{\pm}H]^{\pm}$ — the result of the addition to, or loss from, the molecule, of a proton, which is a hydrogen atom minus the electron;
3. Adducts $[M{\pm}A]^{\pm}$ — the result of the addition of an element, or combination of elements, to the molecule. These are referred to as adducts, and the resulting ion may be positively or negatively charged;
4. Fragments — even with soft ionisation methods, the neutral molecule may fragment into multiple smaller molecules, which subsequently ionise. This can occur while still in solution, during ionisation or as a result of gas-phase collisions, and occurs when the intra-molecular bonds break [38, 35].

A significant drawback with ESI is *ionisation suppression*, where ions that more readily accept, or lose, electrons 'steal' charge from other ions. This results in incorrect relative quantities of ions compared to the relative abundances of the molecules in the sample. While chromatographic methods help to reduce ionisation suppression by presenting a smaller number of different molecules to the ionisation system at once, the problem is still significant.

An additional technique to reduce ionisation suppression, and yield a more accurate quantitation measurement, is by using a nano-electrospray ionisation (nESI) source. In nESI, the needle has a bore of 1–4 μm diameter, compared to the 50 μm for conventional ESI. This has the effect of reducing the initial droplet size from 1–2 μm diameter to <200 nm, and consequently the sample is delivered with a flow rate as low as 20–50 nL/min [39]. The benefit of such a low flow rate is to provide a larger surplus of charge, which reduces the preferential ionisation effects [40].

2.4 Automatic Gain Control and the Ion Trap

The ion trap is an independent mass spectrometer that is located between the ion source and the main ICR cell. One function of the ion trap is to facilitate fragmentation experiments [41], where an ion of interest, the 'parent ion', is purposely fragmented. The pattern of fragments that form is used to draw conclusions about the structure of the ion that was fragmented, and thus help to confirm the identity of the parent ion. Since this process is not suitable for high throughput experiments due to the time implications, it is not considered further in this thesis.

Another important function of the ion trap is to filter the m/z range of ions that enter the ICR cell for analysis [41], and thus reduce the number of unnecessary ions in the ICR cell. The number of ions present in the ICR cell is a very important parameter in terms of the quality of the mass spectrum. As the density of ions in the ICR cell increases, ions of the same, and different, m/z interact and mutually alter their cyclotron frequency. This has a negative effect on the mass measurement error [18, 42]. As well as filtering the mass range of ions, the ion trap regulates the total number of ions passed from the ion trap to the ICR cell. This is achieved with the use of an automatic gain control (AGC), which controls the ion trap and allows a user-definable quantity of ions into the ICR cell. The ion trap can detect ions by exciting them into detectors, and by integrating this signal the AGC estimates the ion flow rate. From this, the AGC calculates the duration that the gate to the ICR cell should be open, thus controlling the number of ions that enter the ICR cell.

The result is that the number of ions can be controlled to some extent by varying the gate-open time, although the exact number of ions is still not constant [43], most likely because the flow rate from the ion source is prone to vary after the static flow sample is measured by the ion trap.

2.5 Quantification Performance

The ICR cell itself is fundamentally linear in terms of measuring the abundance of ions [44]; intuitively, the amplitude of the image current induced in the detector plates by the passing ions is proportional to the total charge carried by the ions, which is proportional to the number of ions in the cloud. Indeed, Marshall *et al.* have shown that, after determining the equivalent resistance and capacitance of the ICR cell, together with its internal geometry, absolute quantification can be obtained if required [45]. However in practise, it is unnecessary to determine the absolute number of ions in the ICR cell, due to the ionisation suppression that occurs in ESI, as described in Section 2.3. This means that

there is not a direct relationship between the abundance of molecules in the sample and the number of ions that are formed in the instrument; the number of ions formed is also dependent upon the ion species.

As a result, metabolomics experiments generally compare the relative abundance change between the same metabolites across different samples. Inter-sample comparisons are only valid when the overall composition of the samples is kept as similar as possible, which will result in relatively consistent ionisation suppression effects across all spectra [46].

2.6 Processing of FT-ICR Data

Section 2.2 describes how ions are captured within the ICR cell, are made to orbit with a frequency that relates to their mass-to-charge ratio (m/z) by the cyclotron equation (2.3), and can be detected and quantified by a pair of detector plates. This section describes how, in a standard FT-ICR MS experiment, the time domain signal from the detector plates is processed into a frequency domain spectrum using a Fourier transform. The spectrum quantifies the presence of ions at a range of orbital frequencies. Where a significant number of ions are detected, i.e. a signal is measured above the noise, a 'peak' is observed at the orbital frequency of the ion cloud. Thus, each peak, or 'feature', relates to a distinct ion cloud. These features of interest are measured and their location in the frequency domain converted into m/z measurements. The main stages of processing detailed in this section are summarised in Figure 2.7, and listed below.

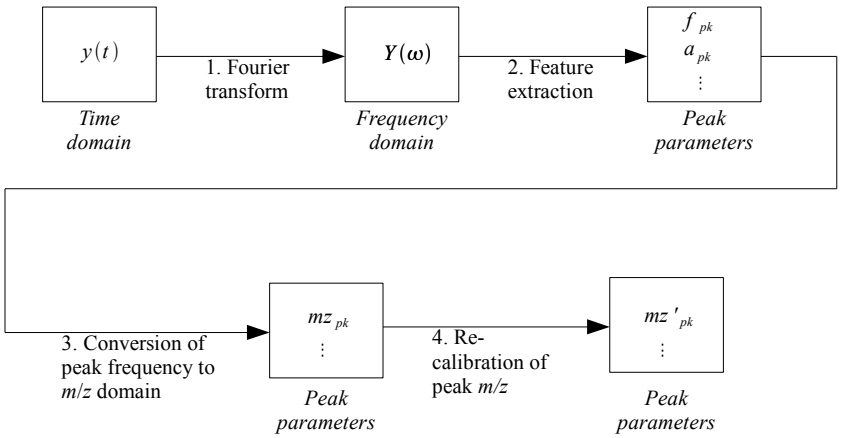

Figure 2.7: Pre-Processing Data Flow

1. The analogue signal from the detector plates is sampled to produce a digital 'transient' time domain signal, $y(t)$, as described in Section 2.6.1.

2. The time domain transient signal is converted into a frequency spectrum as a function

of angular frequency, $Y(\omega)$, using Fourier transform methods. This is described in Section 2.6.2.

3. Features, i.e. peaks, of interest within the spectrum are located, and required parameters of those features are quantified, including: peak frequency, f_{pk}; peak amplitude, a_{pk}; signal-to-noise ratio (SNR); and, if required, the resolution of the peak. These parameters are described in Sections 2.6.3, 2.6.4, 2.6.5 and 2.6.7, respectively.

4. Each peak frequency is mapped to the m/z domain using a 'calibration equation', to yield the peak m/z measurements, mz_{pk}. The calibration process is described in Section 2.6.8.

5. If appropriate, a second calibration stage can be used to further increase the accuracy of the m/z measurements. This 'recalibration' process uses compounds which are known to be present in the spectrum, and for which an exact m/z can be calculated, to improve overall mass accuracy. The result is a more accurate set of peak m/z locations, mz'_{pk}.

Each of these stages is briefly reviewed below.

2.6.1 Transient Acquisition

As shown by the cyclotron equation (2.3), the mass-to-charge ratio of the ions in the ICR cell is calculated from their frequency of orbit. As described in Section 2.2, the ions are measured as they pass detector plates, inducing a small current. Figure 2.8 shows how the signal is amplified by a low noise amplifier (LNA), digitally sampled by an analogue to digital converter (ADC) and stored in a transient file.

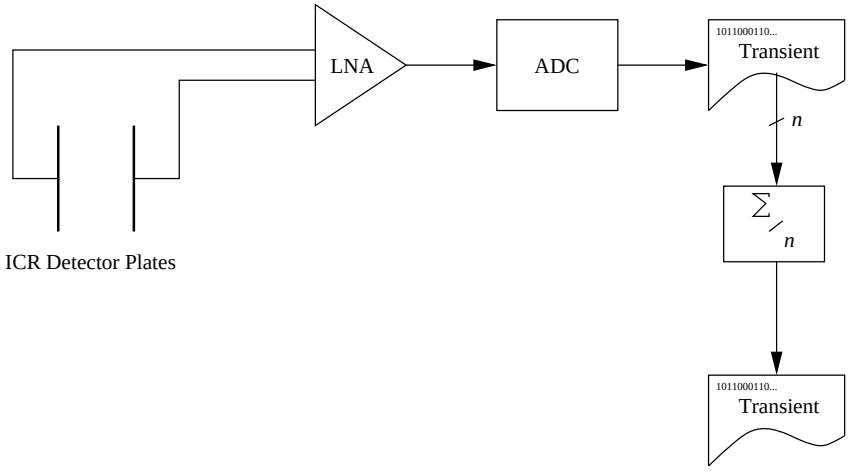

Figure 2.8: Transient Generation

As discussed in Section 2.2, the transient signal decays exponentially, and consequently,

the transient signal has limited duration, typically one second. One method to increase the SNR without increasing the length of the transient is to average multiple transients together [46, 47], and is possible by acquiring multiple transients of the same duration and sampling frequency. By calculating the mean intensity across all transients on a sample point-by-sample point basis, a 'mean transient' is generated that yields improved SNR characteristics. The overall improvement in SNR is $\sqrt{n}$, where n is the number of transient signals [48].

2.6.2 Conversion to Frequency Domain

As discussed above, converting the transient data to a frequency spectrum is necessary in order to locate and measure the features of interest, where each feature relates to a specific ion cloud. A Fourier transform (FT) can be applied to the transient data to achieve this. It is important that the spectrum contains as few artefacts as possible, and one such artefact present after FT of a signal of finite duration is Gibbs oscillations, which occur on either side of a frequency-domain peak as a result of the abrupt start and end of the transient signal [49]. These unwanted frequency components can be reduced by applying apodisation (literally 'removing the feet'). The transient signal has an envelope, i.e. an overall shape of the signal, that abruptly starts and ends. It is these step changes in the envelope that, after applying the Fourier transform, cause the oscillatory phenomenon either side of peaks in the spectrum. By smoothing the envelope of the signal, the Gibbs oscillations are reduced. This is achieved by multiplying the transient signal by a 'window function'. There are several different window functions, the most common of which is the Hanning function [50] which has the form shown in Figure 2.9. Studies have shown that there is no obvious 'best choice' for the window function, with the resultant mass measurement error being very similar regardless of which window function is applied [50, 51].

The signal is then zero-filled [52], and a fast Fourier transform (FFT) applied. The zero-filling simply increases the number of data points in the transient by padding the end of the signal with zeros. This effectively causes the resulting spectrum to contain more data points, which decreases the granularity of the data points in the spectrum, albeit without increasing the amount of information contained. This increases the effectiveness of fitting functions applied to extract feature information such as peak location and intensity. In this work, a single zero-fill, i.e. a doubling of the length of the transient, yields the required accuracy.

Since the location on the frequency axis of the points in the spectrum can also be controlled by applying zero-filling, an interesting suggestion by Comisarow in 1979 [53] was, for each peak, to increase the number of zero-fills until the data points in the spectrum coincide with

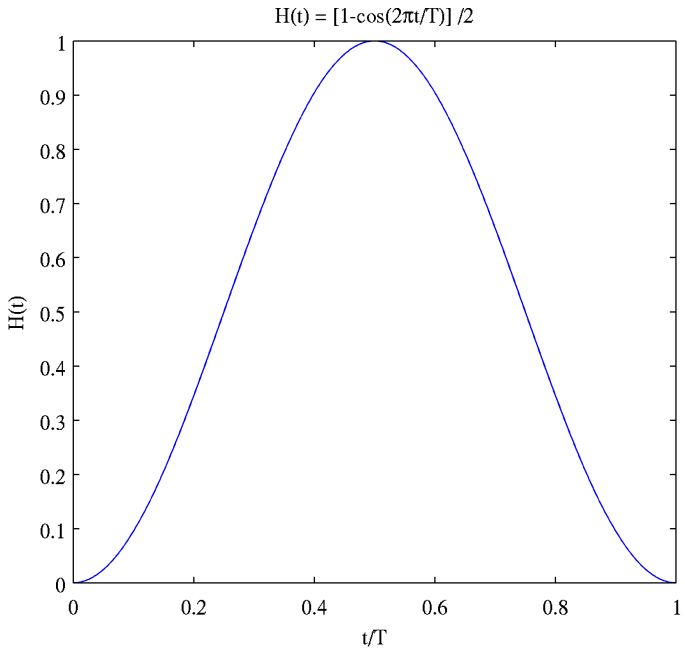

Figure 2.9: The Hanning Window Function

the centre of the peak. While this has the advantage of allowing the precise frequency location of any peak to be measured regardless of the shape of the peak, it was found unnecessary in the experiments reported here due to the very high resolution of the raw data acquired on modern instrumentation. A more recent method is reported by Savitski *et al.* [34], which avoids increasing the number of data points in the FFT by repeatedly constructing a frequency spectrum with data points shifted by a small, varying amount ϵ. The full FFT is calculated for each value of ϵ, and by determining the optimal ϵ for each peak where the data points coincide with the peak crest, the frequency and intensity of each peak is produced. Due to the large overhead associated with the many FFT's involved, this solution is not appropriate for a high throughput environment and is not pursued in this thesis. However, optimisations in the FFT implementation, such as hardware implementation, may merit investigation, but are beyond the scope of this thesis.

The FFT yields a complex frequency spectrum with both real and imaginary components. For the case of only two detector plates, the phase information is not required and so the absolute, 'magnitude-mode', spectrum is used:

$$Y(\omega) = |\mathcal{F}(y(t))|,$$

where $Y(\omega)$ is the intensity in the spectrum as a function of frequency, ω, $\mathcal{F}$ is the Fourier transform, and $y(t)$ is the time domain transient signal. At a static magnetic field of 7T, as present in the Finnigan LTQ FT instrument used in this work, the observed frequencies

range between 107.5 kHz and 21.5 MHz, representing ions in the range 50–1000 m/z.

2.6.3 Peak Centre Frequency

The analytically-derived line shape $Y(\omega)$, resulting from the Hanning-apodisation and Fourier transform of the transient signal, is described by a rather large formula [54]. It is not feasible to find the best-fit of this line shape, and hence determine ω_0, for each peak due to the complexity and high degrees in the equation. An alternative to fitting the exact line shape in the frequency domain is to fit, in the time domain, to ion signals within the transient. This can be achieved, for example, by using an algorithm presented by Umesh and Tufts [55], which has been successfully applied to NMR data [56, 57]. This method is able to fit even when peaks overlap, and would obviate the need to process a frequency-domain spectrum. However in exploratory experiments, it was found to be untenable when applied to high resolution FT-ICR MS data due to the high number of data points and number of peaks compared to NMR data.

Peak fitting, using empirically-proven functions, is therefore pursued as the viable alternative. Since the apodised line shape is symmetrical and approximately Lorentzian [58], a good solution to locating the peak m/z centre, x_0, and the method adopted here, is to assume the parabolic function and apply quadratic interpolation from three data points (x_1, y_1), (x_2, y_2), (x_3, y_3):

$x_0 = x_2 - \frac{\Delta x}{2}\left[\frac{y_3 - y_1}{y_1 - 2y_2 + y_3}\right]$,

where Δx is the spacing between the monotonic abscissae.

While other methods, such as Keefe and Comisarow's exponent (KCe) interpolation [58] and Goto's generalised interpolation (GIα) [50], perform better under certain conditions of apodisation window, decay constant and zero-filling, parabolic interpolation performs comparably and with significantly less computational expense. Such interpolation has also been shown to perform comparably with Lorentzian-based models [51]. Simple parabolic interpolation of a once zero-filled spectrum introduces a systematic frequency error of ca. 6.5% of the frequency spacing [58]. In the worst case, at 500 m/z and at a typical frequency spacing of 0.651 Hz, this equates to a 0.0214 ppm error on the estimated m/z: below the ca. 0.5 ppm random mass measurement error typically observed in an FT-ICR MS experiment (see Chapter 3). This method is used here for the low-cost processing advantage, however as the mass measurement error of FT-ICR MS instruments continues to decrease, the additional processing cost will become justified.

2.6.4 Peak Quantification

As discussed in Section 2.5, the measurement of ion abundance is relative in nature, and is typically measured from peak height or peak area. While peak height provides a fast estimate of quantity, it is peak area that is directly proportional to the initial amplitude of the frequency component in the transient signal [59]. The area measurement accounts for significant variation in peak widths [59], which arise as the result of ion-ion interactions, discussed in Section 2.4.

An obvious approach to calculate the peak area is to numerically integrate the peak, however this suffers from two major drawbacks. Firstly, an appropriate method of determining the start and end point for each peak would need to be found that takes into account the width of the peak [59]. Secondly, and more importantly, such a method would have to take into account peaks that are overlapping other peaks; integration of the signal would not return an accurate peak area.

An alternative approach is to measure peak area indirectly, by fitting the observed data points to the theoretical peak shape in a manner similar to that described above for peak frequency location, and then calculating peak area from the fitted parameters. This is an approach adopted by Goodner *et al.* [60] *, who use a quadratic fitting function developed previously by Keefe and Comisarow [58] to determine peak area from a least-squares fit to each peak individually. The three data points nearest the apex of each peak are used for the fit, thus reducing the effect of overlapping peaks on the area measurement. They achieve superior average mass measurement error than three other peak area measures and five peak height measures over zero to three zero-fills and with 1,100 and infinite scans. Therefore it is the method adopted in this thesis.

2.6.5 Noise Level Estimation

There are many random and independent sources of noise in the acquisition and measurement systems of the FT-ICR instrument, including electrical and thermal causes. The central limit theorem states that the overall distribution of many independent and random noise sources, when summed, is Gaussian [49]. The resulting Gaussian-distributed noise is observed in equal measure in both the real and imaginary components of the frequency spectrum, which is obtained after a Fourier transform of the transient signal. A common measure of the noise level of Gaussian random variables [45, 61, 56], that is used in this thesis, is:

*Note an error on p1208 of Goodner *et al.* [60], equation 9: $k = 4a/q$ is incorrectly printed as $k = 4c/q$

$$\text{noise level} = \text{standard deviation}, \sigma, \text{of the noise in signal-free region}, \tag{2.5}$$

and is applicable to the noise present in the real and imaginary components of the frequency spectra only. When the *magnitude*-mode frequency spectrum is calculated, the noise floor is not Gaussian, but Rayleigh-distributed [59, 45, 62]. Visually, this noise in an FT-ICR mass spectrum appears to be a noise 'floor' that is present at all m/z's, and is relatively constant throughout the spectrum. Consequently, low abundance peaks from real signal are distorted or completely masked. The probability distribution function (PDF), $f(y)$, and cumulative distribution function (CDF), $F(y)$, of the Rayleigh distribution are shown in equation 2.6 and 2.7, respectively. Consequently, σ can be measured directly from the magnitude-mode spectrum by best-fit of the data to the Rayleigh distribution.

$$f(y) = \frac{y}{\sigma^2} e^{-\frac{y^2}{2\sigma^2}} \tag{2.6}$$

$$F(y) = 1 - e^{\frac{-y^2}{2\sigma^2}} \tag{2.7}$$

2.6.6 Signal-to-noise Level Estimation

The calculation of SNR is sometimes unclear and inconsistent, however the most common is:

$$\text{SNR} = \frac{\text{height of signal peak in magnitude spectrum}}{\text{noise level},} \tag{2.8}$$

where calculation of the noise level is described in Section 2.6.5.

2.6.7 Resolution

The resolution of a peak is typically defined as $mz_0/\Delta mz$, where mz_0 is the m/z at peak centre, found in the frequency domain as described in Section 2.6.3 and converted to the m/z domain as described below in Section 2.6.8, and Δmz is the m/z width of the peak at some fraction of peak height [63]. A commonly used fraction is 50%, defining the full width at half maximum (FWHM) resolution [23].

2.6.8 Conversion to m/z Domain and Internal Recalibration

The frequency spectrum and peak features are converted to a mass spectrum measured in m/z, by means of a *mass calibration equation* and associated parameters that are selected to provide an accurate conversion. The most basic calibration equation is derived directly from the cyclotron equation, equation (2.3), which can be rewritten as

$(m/z) = a/f,$

where f is the observed frequency in Hz, m/z is the mass-to-charge ratio in Daltons, and a is the 'calibration parameter'. In theory, then, the calibration parameter is

$$a = \frac{eB}{2\pi\mu}, \tag{2.9}$$

where e is the charge of an electron in Coulombs, B the magnetic field strength within the ICR cell in Tesla, and μ is the weight of one Dalton in kilograms.

However, it was found that the magnetic field strength alone is insufficient to explain the relationship between ω and m/z [22], and this is attributed to two main causes. Firstly, the electrostatic field holding the ions inside the ICR cell perturbs the resonant frequencies of the ions [22, 1]. Secondly, ion clouds, being charged and in close proximity to other ion clouds, are subject to Coulomb interactions as the result of electrostatic forces that repel clouds apart. The observed result is the '*space-charge*' effect, which distorts the observed frequencies, depending upon the proximity in frequency, and abundance of ion clouds inside the ICR cell [22, 45, 64]. Such ion-ion interactions within an ICR cell, while understood, cannot, to date, be modelled [30]. For these reasons, all calibration equations to date are empirically formulated and many different calibration equations have been proposed. However, there is currently no 'gold standard' for the calibration equation. An informative review by Gross *et al.* [65] summarises their development.

The most significant improvement to the calibration equation is the addition of a second term as shown in equation 2.10. First proposed by Ledford *et al.* in 1984 [64], it showed significantly improved mass accuracy over a wide mass range, explaining its commonplace use and adoption within this thesis.

$$(m/z) = a/f + b/f^2, \tag{2.10}$$

where the parameter a is defined in equation (2.9) above and b is defined as

$b = -eG_T V_{\text{eff}}/2\pi^2,$

where e is the charge of an electron in Coulombs, G_T is a factor depending upon the trap geometry, and V_{eff} is the effective trap voltage in Volts. The calibration parameter b applies a correction for the offsets in observed frequency that result from the electrostatic field, and also applies some correction for the space-charge effects. It is important to note that all calibration equations to date are empirically derived [65], i.e. they are best-fit solutions to observations, and not formed completely from theory. This is due to an incomplete modelling of the physical processes inside the ICR cell. In practise, the a parameter, being related to constants and the field strength of the superconducting magnet, is much more stable than the b parameter, which can be used to take account of variations including the total number of ions in the cell and even the cleanliness of the cell [51]. Changes in the a parameter have the effect of 'stretching' the m/z spectrum, while variations in b result in a near-linear m/z shifting of the mass spectrum.

Other forms of the equation have been proposed since, including reformulations of the two-term equation and in particular a more statistically-correct form that shows a slight improvement in measurement accuracy and precision as a result [66]. Of other suggestions, the introduction of a third term, c, dependent upon the intensity of the peak appears the most promising [42], and has been previously used [51]:

$$(m/z)_i = a/f + b/f^2 + cI_i/f^2, \tag{2.11}$$

where a, b and f are as defined previously, $(m/z)_i$ and I_i are the m/z and intensity of peak i, respectively. Masselon *et al.* observed that while small in comparison with the space-charge effect (parameter b), the root mean square (RMS) mass measurement error of peaks between 0 and 2000 Da can be reduced by between 35 and 47% (depending upon the radius of the excited ions). Using this approach, several papers show a significant improvement over the standard two-term equation [66, 67]. However, at least three reference peaks are required to estimate the parameters a, b and c. Since there is generally a lack of reference peaks in mass spectra, and three-term calibration is not routinely used, it is also not adopted in this thesis.

Recently, attention has focussed on a more complete accounting for the space-charge effect in the calibration equation. Masselon's three-term equation has been developed by Zhang *et al.* [65] to include additional terms to account for other factors relating to the ion, other than its abundance. There is no empirical evidence that this equation is an improvement on the three-term equation.

The parameters for the chosen calibration equation are either provided by the instrument software, which is termed 'external' calibration [23], or are found by comparing the calculated location of compounds known to be present in the spectrum with the observed

location. This is known as 'internal' calibration [23], and the compounds used to calibrate the spectrum are 'internal calibrants'. The external calibration parameters are calculated by the instrument during a calibration procedure, where a chemical standards mixture is analysed and the measured spectrum is compared with the expected spectrum from the standard. Such parameters do not account for instrument drift since the last calibration, or variations in the measured spectrum caused by the content of the sample being analysed. Therefore, a significant improvement in mass accuracy can be achieved if internal calibration is used to correct for these variations. Actual gains depend upon the number and spread within the spectrum of the internal calibrants; however, even with a single calibrant, the absolute mean error has been shown to be reduced from 5.2 parts per million (ppm) to 0.8 ppm, an improvement of 85% [51]. Eyler *et al.* [51] show that multiple internal calibrants can further improve the absolute mean error to 0.7 ppm. The calibrants are typically either ions that have been added to the sample for the sole reason of calibrating, or they are peaks relating to compounds that are strongly believed to exist in the sample under test, and for which an exact mass is known. It is preferable to avoid adding compounds to the sample since this will inevitably result in increased ionisation suppression and ion-ion interactions, both of which are detrimental to mass measurement accuracy and quantification performance. Additionally, 'background' ions are present in the ICR cell as undesirable, but unavoidable, contaminants. Background ions come from the sample preparation method, carry-over within the cell itself or atmospheric contaminants that enter the cell with the sample. Since such ions can be expected to be largely present in all samples, they have successfully been used as internal calibrants themselves [68].

Other techniques to improve calibration, while avoiding the use of an internal calibrant, use a step-wise approach [67], which effectively applies a better *external* calibration. Such methods first use an internal calibrant, optimise the instrument parameters accordingly, then inject the analyte and use the optimised calibration parameters to externally calibrate. This method is suitable when a narrow m/z range is being analysed, so that the spectrum mass accuracy can be optimised over a small area. Otherwise, no benefit would be seen over standard external calibration. An alternative approach to mitigate the problems associated with the use of an internal calibrant is to use multiple ionisation techniques simultaneously [51], or dual ESI sources [69], with the aim of isolating the analyte and the internal calibrants from each other and so reduce the overall ion-ion interactions.

A new technique, which is applicable for a variety of calibration equations, is multidimensional recalibration, proposed by Smith *et al.* [70]. Here, the calibration parameters of the two-term equation are determined for a set of variable ranges, including peak intensity, LC separation time and m/z. A model is empirically determined for the parameters as

a function of those variables, and can comprise up to 200 discrete regions, each with a unique range of variable values. The model is then applied to each m/z point, and Smith *et al.* report a significant increase in mass accuracy. However, since it is statistical, over 100 potential calibrants must exist in each region, which makes this an unsuitable choice for this thesis.

In a similar vein, Kearney *et al.* report a good estimate of systematic errors can be calculated by using a Bayesian approach [71]. Again, for this method, the number of potential internal calibrants must be high.

2.7 Signal Processing Challenges

In order to meet the aim of profiling the metabolome using FT-ICR MS, several challenges need to be overcome. These challenges relate to all data processing from the current waveform detected in the ICR cell to the list of peaks detected (and quantified) in the resulting spectrum. In particular, three aspects are currently targeted: optimised pre-processing; maximised sensitivity; and noise filtering.

The particular challenges in each of these aspects are described below:

1. The processing of raw ESI FT-ICR MS data is not optimal when using the instrument manufacturer's software. It has been found, for example, that regions of the spectrum contain 'bursts' of noise that appears above the nominal noise floor. Additionally, methods used in the peak detection and parameterisation stages are unpublished and not under the user's control. These issues represent a need to implement a pre-processing method that optimises the quality of the spectrum, which is addressed in Chapter 3.

2. The sensitivity of the instrument describes the minimum number of ions required to generate a detectable signal. One solution to the problem of measuring low abundance ions is to inject more into the ICR cell. However, as discussed above, the trade-off with analysing higher numbers of ions is an increase in ion-ion interactions, which decreases the achievable mass accuracy. In order to allow sufficient quantities of low abundance molecules into the ICR cell for detection, a method is presented in Chapter 3 that allows higher numbers of low prevalence metabolites into the ICR cell without overloading the cell or compromising mass accuracy.

3. The third challenge tackled is the problem of noise filtering. Even with the above optimisations, the signal from some metabolite species will be at, or below, the noise floor due to the hardware limitations of the instrument. A simple hard threshold is common but not optimal. In many cases, multiple acquisitions of a sample are

made, and in Chapter 4, it is shown how, by using this additional information, noise filtering can be optimised, thus minimising the unnecessary loss of low intensity, but relevant peaks.

2.8 Conclusions

This chapter has described the operation of FT-ICR MS and, in particular, the relationship between the resonant frequency of orbiting ions and their mass to charge ratio, m/z. It has described the ion trap, and how electrospray ionisation introduces the sample and creates ions, ready for detection. Also discussed is the processing of the instrument data from the time domain to a mass spectrum. The limitations of the instrument include noise, and space-charge effects, which restrict the number of ions that can be present in the ICR cell simultaneously. Furthermore, the software supplied by the manufacturer does not remove regions of high noise, which could degrade the performance of subsequent processing and interpretation stages. From these issues arise three specific challenges. The first two challenges are to overcome the limitations in the manufacturer's software and the space-charge effects, while maximising sensitivity. These challenges are addressed in the SIM-stitching algorithm presented in Chapter 3. The third challenge is to reduce the noise level in the spectra, and Chapter 4 meets this challenge with the development of a three-stage filter. Chapters 5 and 6 build on this work of optimising the spectra, by describing the data mining of mass spectra using constraints optimisation.

CHAPTER 3

SIM-STITCHING

3.1 Introduction

The dynamic range of the metabolome describes the range of prevalence of metabolites in a sample; a sample with a vast difference between the most and least abundant metabolites has a high dynamic range. Therefore, one of the key performance parameters of a mass spectrometer is the instrument's *dynamic range*, which describes the lowest intensity ion population that can be detected, relative to the intensity of the most abundant. A typical value for the dynamic range of a FT-ICR mass spectrometer is 5,000 [72], i.e. for a maximum signal intensity of 100%, the detection limit is 0.02%, below which ions cannot be detected. Newer designs offer an improved dynamic range, and the benefit of this is considerable in terms of the number of features observed and consequently the number of low-abundance ions that can be measured; as discussed in Section 1.2, it is very important to measure such low intensity ions, since they form a significant portion of any biological sample. Additionally, many isotopes will exist with very low natural abundance, and the ability to measure more of these will provide further relevant data with which to identify the empirical formulae of ions.

This chapter describes a strategy for direct injection nano-electrospray ionisation (DI nESI) FT-ICR MS that satisfies objective 1 in Section 1.3, by effectively increasing the overall dynamic range of the instrument, thus increasing its sensitivity to low intensity ions. The content of this chapter is mainly work originally published in 2007 [18] with modifications and additions. The approach is based upon the collection of multiple narrow, overlapping spectra (or 'windows') that are subsequently combined together using a novel 'stitching' algorithm.

Typically an FT-ICR instrument operates in 'full scan' mode [73], where a wide m/z range is scanned (e.g. 50–1000 m/z) and all ions within that mass range are concurrently

measured. The total number of ions that can be measured simultaneously is limited by the automatic gain control (AGC) in order to mitigate space-charge effects, as discussed in Section 2.4. Consequently, many ions of low intensity will not be detected. In the approach adopted here, narrow spectral 'windows' of 30 m/z width are analysed using the selected ion monitoring (SIM) feature of the mass spectrometry. Since ions beyond the window boundaries are ejected by the ion trap, those within the window can be accumulated in greater concentration than would otherwise be possible in a wide scan. Thus, ions of low abundance that would be below the detection limit in a normal wide range scan become detectable, without increasing the concentration of ions in the ICR cell. As discussed in Section 2.4, to increase the number of ions in the cell compromises mass accuracy. The SIM windows are then combined, with a 10 m/z overlap, as illustrated in Figure 3.1, to produce a 'SIM-stitched' mass spectrum that covers a wide m/z range with significantly greater dynamic range. This enables many more metabolites to be detected, and with higher mass accuracy than is possible with existing methods. A similar acquisition strategy has been previously reported by Venable *et al.* [74], although applied to the detection of peptides and with the focus on liquid chromatography and fragmentation experiments.

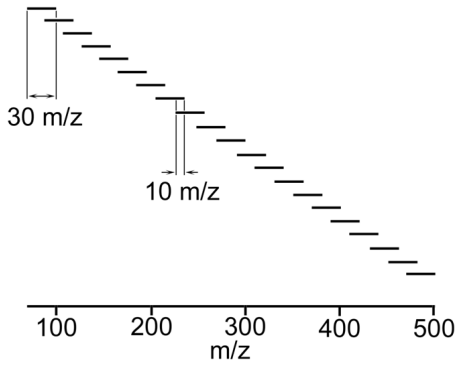

Figure 3.1: Schematic of the optimized SIM-stitching method comprising 21 adjacent 30 m/z SIM windows, each overlapping by 10 m/z, covering a total scan range of 70–500 m/z.

Stitching the multiple SIM windows together to produce a single contiguous, wide-scan spectrum is important for several reasons. First, the multivariate analysis of mass spectral fingerprints using widely used methods in metabolomics, such as principal components analysis, requires a single spectral fingerprint per biological sample. Second, methods that search for characteristic peak spacings across the entire spectral range, for example as used by Breitling *et al.* [75], also require one contiguous data set. Third, the stitching algorithm is used to mass-calibrate SIM windows that do not contain internal calibrants. Finally, stitching multiple data sets together substantially aids data handling, storage and visual inspection of the mass spectral measurements.

3.2 Method

3.2.1 Overview

Given a set of SIM windows overlapping as described in Section 3.1, the aim is to produce an optimally internally-calibrated single spectrum and list of features that the spectrum represents. This is contained within a software named SIMStitch [18, 19, 20], which also enables the user to configure many different parameters and establish a work flow by defining the input and output files at each stage of the processing, including transient file averaging, pre-processing, stitching and peak noise filtering (see Chapter 4).

The input to this stage is, for each overlapping SIM window spectrum, the frequency and intensity data points of the frequency spectrum, together with the external calibration parameters relating to the calibration equation (2.10). The work flow of processing on these SIM windows is:

1. Correct the intensity values of each SIM window due to varying sample frequencies, as described in Section 3.2.2.
2. Extract all peak features from each SIM window: peak center frequency, peak height and area, noise level, peak SNR and peak resolution. The calculation of these features is described in Chapter 2, Section 2.6. Details specific to this new method are to be found below, in Sections 3.2.3 to 3.2.5.
3. Remove peaks that are identified as noise artefacts, as described in Section 3.2.6.
4. Apply intensity correction to each SIM window, as described in Section 3.2.7.
5. Recalibrate each SIM window where possible to obtain a new set of internal calibration parameters, as detailed in Section 3.2.8.
6. Align externally calibrated SIM windows with neighbouring and overlapping internally calibrated SIM windows, where applicable, to obtain adjusted frequency values, as described in Section 3.2.9.
7. Adjust frequency values such that calibration parameters are uniform across all SIM windows.
8. Remove edge effects, and concatenate SIM windows to give final spectrum frequency and intensity values and associated list of features, as described in Section 3.2.10.

3.2.2 Sample Frequency Intensity Correction

It was observed that the sample frequency, f_s, of the instrument varied according to the starting m/z of the SIM scan. This occurs since, according to Nyquist theory [52], the sample frequency is required to be at least twice the highest frequency, corresponding to the smallest m/z, being measured. The sampling frequency does not affect the decay rate of the measured signal, since this is independent of the electronic measurement hardware, and therefore the peak shape is also unchanged. However, the effect of the additional data points in the time domain, as a result of the increased sampling frequency, is to proportionally scale the intensities in the frequency spectrum post-FFT. Consequently, a peak area estimation would be incorrect unless the correction in the following equation is applied to intensity points Y for each SIM window:

$$Y'_s = Y \frac{f_{max}}{f_s},$$

where Y_s is the corrected intensity values for the SIM window, f_{max} is the maximum sampling frequency across all SIM windows and f_s is sampling frequency for the SIM window.

3.2.3 Noise Level Estimation

As discussed in Section 2.6.5, noise in a signal-free region of the magnitude-mode spectrum has a *Rayleigh* distribution, and the standard deviation of the noise, or 'noise level', can be measured directly from the magnitude-mode spectrum.

In order to discover a suitable signal-free region from which the noise level can be measured, the degree to which the spectral range is signal-free is quantified by the fit of the data in the range to the Rayleigh noise model. Yeh and Röbel [76] adopted a similar approach, by iteratively classifying peaks in the test region as signal, until the distribution of the remaining peaks matched well with the noise model. However, in the case of high resolution data, it is desirable to measure the noise using data points and not peak heights, in order to maintain high throughput and reduce the impact of irregular peak shapes. Therefore, the method adopted is to instead iteratively move and resize the test region until a signal-free region is found, as shown in Algorithm 1. The mass spectrum data are generally sufficiently sparse that several hundred consecutive signal-free data points can easily be found.

To determine if a selected region can be classed as signal-free ('Rtest' in Algorithm 1), firstly the estimated Rayleigh parameter $\hat{\sigma}$ of the range is determined by maximum likelihood fit of the data distribution to the Rayleigh distribution (equation 2.6). Public domain code [77] is used to determine the Rayleigh parameter estimate $\hat{\sigma}$. Secondly, the

maximum intensity permissible, y_{max}, within the range is calculated, based upon $\hat{\sigma}$ and an arbitrarily low probability of any of the Rayleigh-distributed data points randomly exceeding y_{max} [18]. If no data point has intensity above y_{max}, the test range is classed as signal-free, otherwise it is classed as containing signal.

Algorithm 1 To locate signal-free regions within Rayleigh-distributed noise

```
i_start ← 0 {starting index}
l ← 10000 {test region length}
n ← lengthof(input) {length of spectrum}
while Rtest(i_start, i_start + l) = false do
  i_start ← i_start + l/100
  if i_start > n − 2 then
    l ← l/1.01
    i_start ← 0
    if l < l_min then
      error() {no signal-free region found}
    end if
  end if
end while
return i_start, l
```

The minimum number of data points for reliable estimation of the noise level l_{min} in the algorithm was estimated from analysis of the variation of noise level as a function of the number of data points used. The point at which the noise level measure stabilised was taken as l_{min} and is typically ca. 5000.

3.2.4 SNR Estimation

The SNR definition shown in equation (2.8) is used. The noise level is calculated as described in Section 3.2.3. The peak height is determined using an interpolation function derived by Keefe and Comisarow, which they termed 'KC*e* interpolation' [58]. It is based on parabolic interpolation of the line $\hat{y}(\omega)$, but with the addition of a constant exponent, e, that alters the shape of the curve as shown in equation 3.1:

$$\hat{y}(\omega) = \left(a\omega^2 + b\omega + c\right)^e, \tag{3.1}$$

where the parameters a, b and c are to be found from the observed data, and this is achieved by solving the set of equations:

$$y_1 = \left(a(\omega_1)^2 + b\omega_1 + c\right)^e,$$
$$y_2 = \left(a(\omega_2)^2 + b\omega_2 + c\right)^e,$$
$$y_3 = \left(a(\omega_3)^2 + b\omega_3 + c\right)^e,$$

where $y_{1,2,3}$ are the intensities of the data points corresponding to the maximum point of the observed peak (y_2), plus the data points either side of that maximum, and where $\omega_{1,2,3}$ are the frequencies of the data points $y_{1,2,3}$, respectively.

The constant e is selected to shape the function to the observed peaks, and is consequently dependent upon the apodisation method [58]. For Hanning apodisation with 0 to 3 zero fillings, Keefe and Comisarow find the optimal value of e to be ca. 5.5. Differentiating and solving equation 3.1 allows the peak height, $\hat{y}_0$, to be estimated [60]:

$$\hat{y}_0 = \left(-b^2/4a + c\right)^e \tag{3.2}$$

3.2.5 Resolution

As described in Section 2.6.7, the resolution of a peak is defined as $mz_0/\Delta mz$, where mz_0 is the m/z at peak centre, found in the frequency domain as described in Section 2.6.3 and converted to the m/z domain as described in Section 2.6.8, and Δmz is the m/z width of the peak at some fraction of peak height [63].

The peak width Δmz can be found by solving equation 3.1 for the case of $\hat{y}(\omega) = \hat{y}_0/2$, where $\hat{y}_0/2$ is calculated from equation 3.2, to yield the frequency locations at each half maximum point, ω_1 and ω_2, respectively. Application of the calibration equation described in Section 2.6.8 yields the m/z difference, Δmz, between ω_1 and ω_2.

3.2.6 Noise Artefacts

A significant anomaly of unknown origin has been observed in the mass spectrum obtained from the Finnigan LTQ FT instrument used in this work, and occurs at specific (and generally constant) locations on the mass axis [21]; an example occurring between ca. 101.8 and 102.0 m/z is shown in Figure 3.2. Since this signal occurs at much higher resolution than 'real' peaks, it is not thought to be chemical noise but may be due to unexpected and non-random noise on the power supply [78]. Whatever the cause, these peaks should not be classified as real peaks, since they are certainly not the result of ions inside the ICR cell, and should be identified and removed before further processing. However, automatic detection based upon local peak density or the resolution of peaks failed to provide a robust method of selecting these regions of noise. Instead, 'exclusion

regions' are manually defined and automatically removed from spectra throughout the data in this thesis. It is notable that commercial instrument software (Xcalibur v2.1.0, Thermo Scientific, Bremen Germany) treats these regions as signal, thus falsely identifying large numbers of peaks within these very narrow regions.

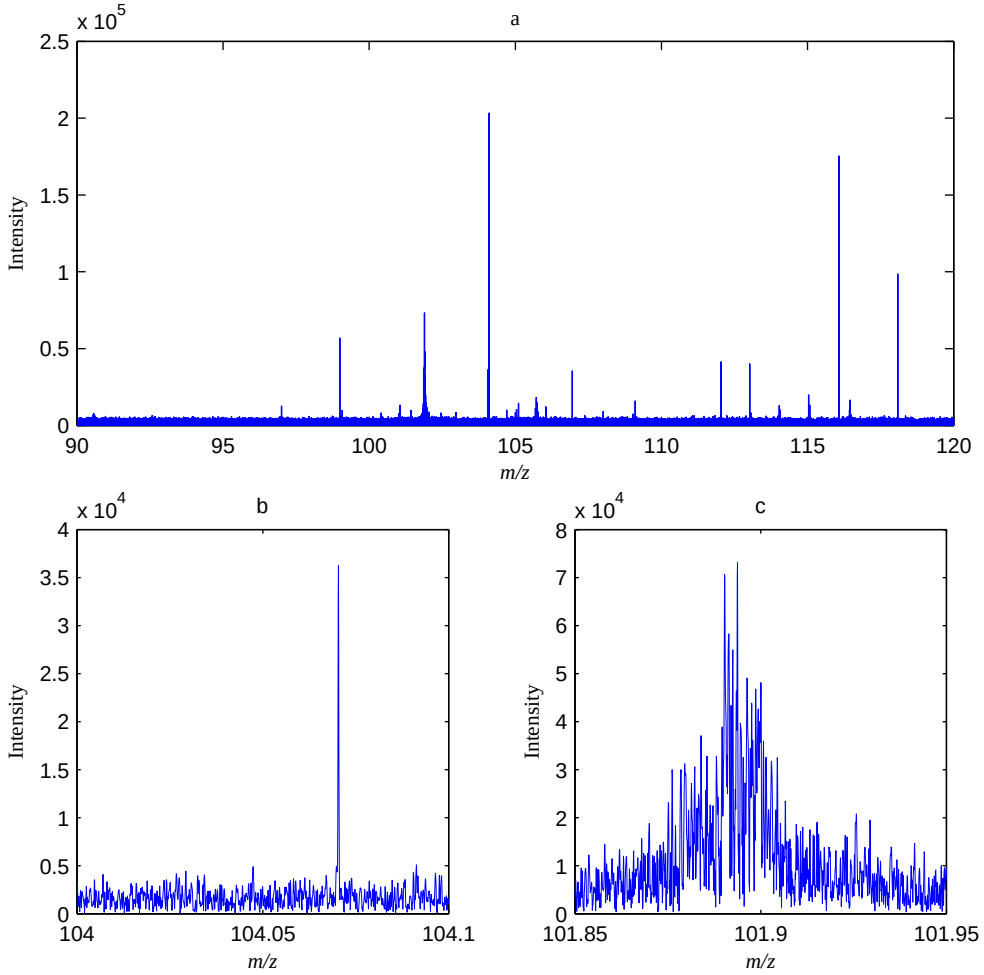

Figure 3.2: Noise artefact present at ca. 101.9 m/z. (a) Spectrum over wide range. (b) A typical real peak in this region. (c) A series of anomalies (N.B. identical abscissa scales for (b) and (c)).

3.2.7 Intensity Correction

Analysis of SIM windows acquired for the SIM-stitching algorithm revealed a systematic error in the intensity measurements, where the intensity of a peak is dependent upon its location within the SIM scan range being measured [21]. It is assumed this is a characteristic of the instrument used in this work. To characterise and correct the intensity error, a liver extract was analysed over a 'sliding' SIM window of fixed width 30 m/z and starting location from 50 to 430 m/z in 2 m/z increments, i.e. a total of 191 staggered

SIM windows were recorded. All available scans were used, regardless of the variation in the number of scans the instrument acquired within the allocated acquisition time, as described in Section 2.6.1. Peaks consistently present at high intensity in the range 80–430 m/z, and that did not neighbour other high intensity peaks, were selected. The distance from peaks of high intensity was important in order to reduce mutual space-charge effects on the measured peak intensities. As the SIM window was moved, the intensities of the peaks of interest were tracked.

The intensity of peaks, relative to the mean peak intensity within the central 10 m/z, is shown to vary as the SIM window location defining the m/z region being scanned is varied, as shown in Figure 3.3. The figure clearly shows that the measured signal intensity consistently increases as the peak moves across the SIM window. This instrumental measurement error is large, with a measured peak area at the high m/z end of a SIM window approximately 3 times more intense than if the identical peak is measured at the low m/z end of an adjacent SIM window. To quantify this effect, the gradient of the trend line, for example that shown in Figure 3.3, was measured for several different peaks over the range 70–500 m/z, and the results are plotted in Figure 3.4. This reveals an albeit noisy trend in the gradient, which is parameterised as a linear function in the figure, and consequently allows the intensity of each peak in the mass spectrum to be corrected based on its location from the start of the SIM window. Also evident in Figure 3.3 is the 'edge effect', which is apparent as a significant decrease in peak intensity (or disappearance of the peak altogether) as peaks near the extremities, particularly at the high m/z end of the SIM window. A more complete experimental characterisation of these edge effects, which are dealt with during spectral processing, is in Section 3.2.10.

3.2.8 Internal m/z Recalibration

The use of three-term instead of two-term calibration was investigated, over n=35 calibrants within a scan with range 70–500 m/z and with an AGC target of 5×10^5. The sample consisted of polyethylene glycol (PEG), a compound that results in a series of peaks present at known m/z values. The experiment was repeated three times. By comparing the m/z values of the PEG peaks after calibration using both calibration equations, the results in Table 3.1 show a slight improvement in mean measurement mass error of ca. 5% and in root mean square (RMS) mass measurement error of ca. 4%. The maximum error decreased by ca. 2%. These improvements are lower than shown by Muddiman *et al.* [66]; however, in that work the AGC setting was not defined, and assuming the number of ions in the cell were higher than used here, space-charge effects would be more significant, thus increasing the effect of the third term of equation (2.11).

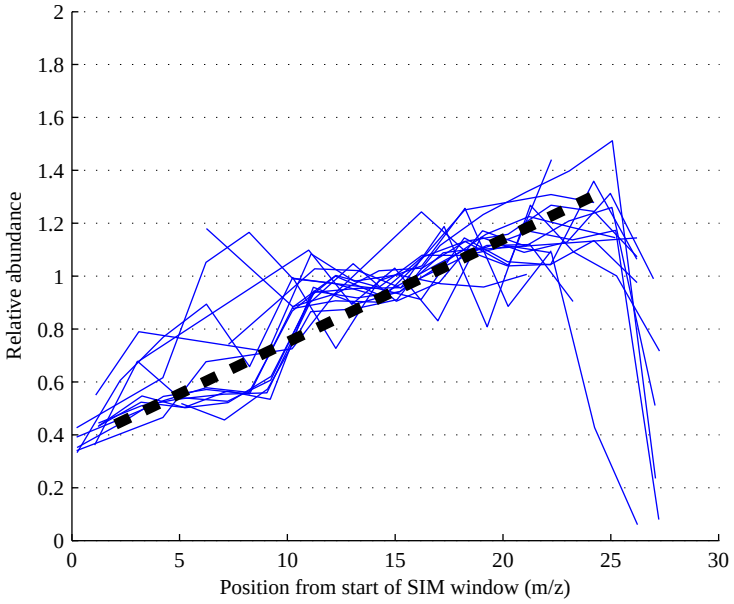

Figure 3.3: The intensities of several peaks in the range 352–354 m/z relative to their mean values within the central 10 m/z, measured across multiple SIM window m/z ranges. Considerable variation in peak intensity can be seen as the SIM window position moves across the peaks. The broken line is a visually-placed trend line.

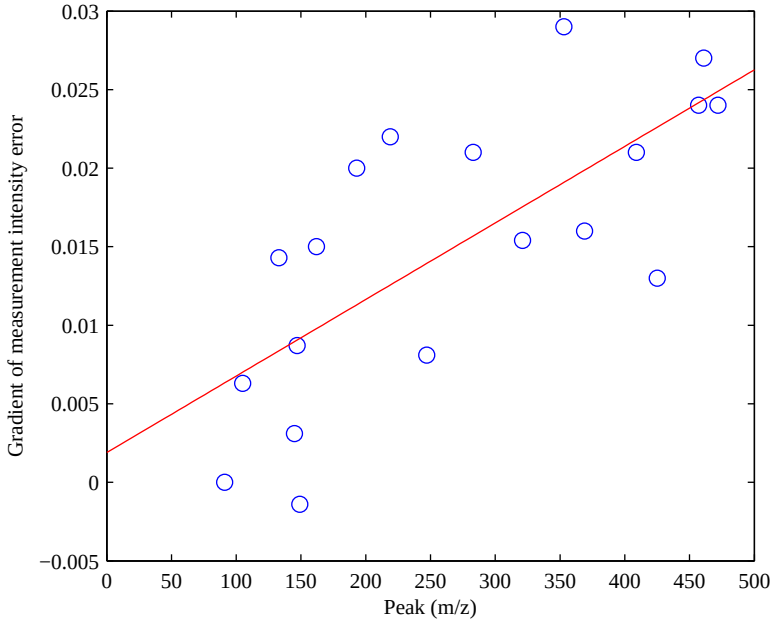

Figure 3.4: The gradient of the measurement intensity error for a selection of peaks located between 70 and 500 m/z. The line of best-fit is shown.

Statistic	Experiment	Two-Term (ppm)	Three-Term (ppm)	Improvement (%)	Overall Mean Improvement (%)
mean	A	0.44	0.42	4.5	5.3
	B	0.42	0.40	4.8	
	C	0.41	0.38	6.7	
rms	A	0.55	0.54	0.2	3.8
	B	0.48	0.45	6.3	
	C	0.48	0.46	4.8	
max	A	1.89	1.88	0.5	1.7
	B	1.02	1.01	1.0	
	C	1.25	1.21	3.5	

Table 3.1: Two-term vs three-term calibration.

Where internal calibrants are provided, they are used to recalibrate the respective SIM window(s). The two-parameter or three-parameter equations (equations (2.10) and (2.11), respectively) are used, as selected by the user. In order to increase the robustness of the recalibration stage, the inclusion in recalibration of the a parameter can be selected to be dependent upon the total m/z range of the calibrants: this reduces the maximum error that may be propagated due to the m/z distance from the calibrants. Additionally, peaks used as calibrants are required to meet a minimum SNR level (typically 6.5), in order to increase the likelihood of a correct assignment. Peaks are identified as calibrants based upon their externally-calibrated distance from the exact mass of the calibrants.

The new parameters, b_{int} and a_{int}, are determined from a least-squares fit of the peak frequencies to the calibrant exact masses using the appropriate calibration equation.

3.2.9 SIM Window Alignment

In the case of SIM windows where no calibrant features can be identified, adjacent and overlapping SIM windows that have been internally calibrated are used to infer calibration parameters. The assumption is that the overlapping portions of the SIM window are calibrated by virtue of the calibration applied to the entire window. Consider two overlapping SIM windows s_1 and s_2 as shown in Figure 3.5.

Consider window s_1 is internally calibrated and s_2 is a SIM window with no identifiable internal calibrants. Alignment is achieved by modifying the *frequency* points in s_2 such that the frequency of peaks in the overlap region have a minimal overall difference. The model used for frequency alignment is either

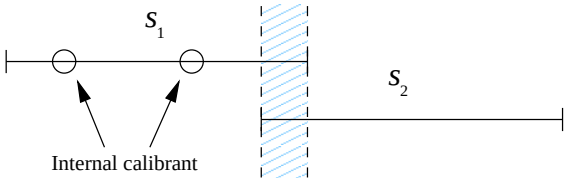

Figure 3.5: Two overlapping SIM windows

$$F'_p = mF_p + c, \tag{3.3}$$

or

$$F'_p = F_p + c, \tag{3.4}$$

where F_p and F'_p are the centre frequencies of the peaks in the overlap region of s_2 before and after alignment, respectively, and m and c are unknown multiplicative and additive adjustments, respectively, to the frequency axis. Equation (3.3) is used where the range of the overlapping peaks exceeds 20% of the SIM window width, otherwise equation (3.4) is used. The overall difference, d, is expressed as a sum of squares over all frequency points f, i.e.

$$d = \sum_f |F_p - F'_p|, \tag{3.5}$$

where F_p is the frequencies of the common peaks in the overlap region. To determine the optimal value of m and c, the simplex search method is used to minimise d using equations (3.5), (3.3) and (3.4). In order to identify F_p, the externally calibrated forms of windows s_1 and s_2 are used as the most non-biased estimate of m/z without internal calibration. Also, to be labelled as common, peaks in both spectra are required to meet a minimum SNR of 6.5 and have a m/z difference of better than 1.5 ppm.

Before combining the SIM windows into a single spectrum, a single calibration parameter that can be applied to all SIM windows is required. This is achieved by simply converting each SIM window to the m/z domain using its unique set of calibration parameters and equation (2.10), then un-calibrating the m/z spectrum to a new frequency spectrum using the required common calibration parameters and the inverse of the calibration equation, equation (3.6):

$$f = \frac{a + \sqrt{a^2 + 4b(m/z)}}{2(m/z)} \tag{3.6}$$

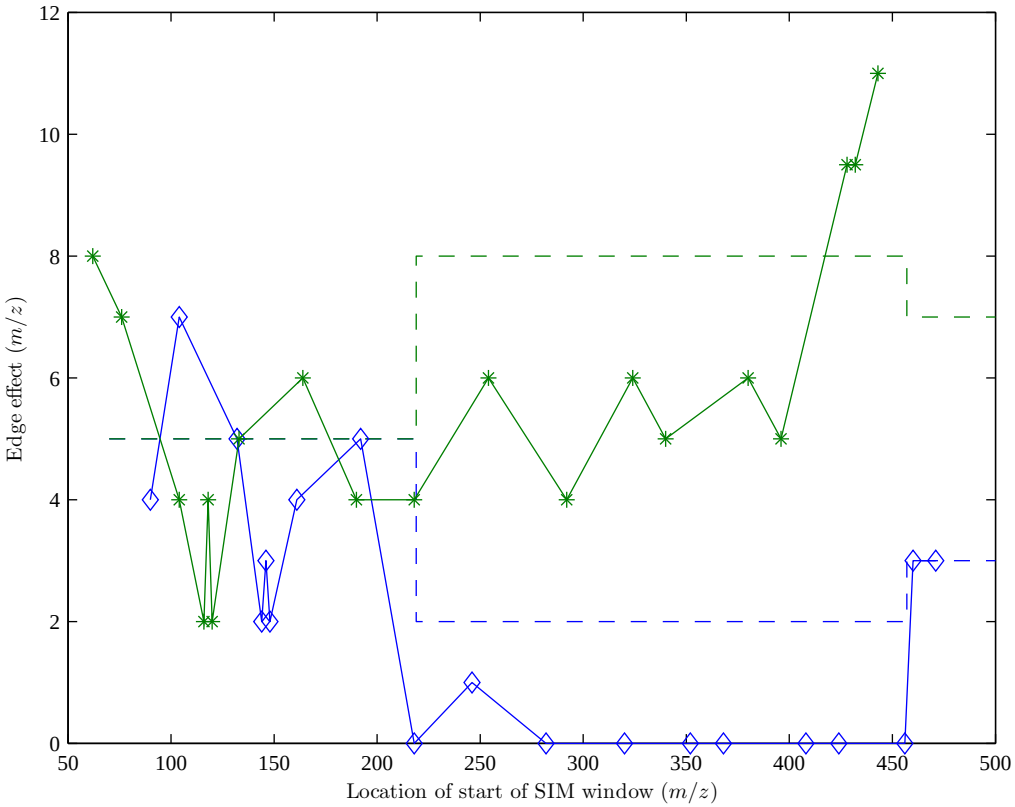

Figure 3.6: The observed edge effect at the start of the SIM window (shown in blue with diamond markers) and the end of the SIM window (shown in green with asterisk markers). The representative edge effect region selected at the start and end of each SIM window is a function of the midpoint of the overlap with the adjacent SIM window, and is shown as a broken line for both the start and end of the SIM window (lower and upper lines, respectively).

3.2.10 Edge Effects and Concatenation

The results obtained during the previous section pertaining to the SIM window intensity trend provide evidence of the 'edge effect' that manifests itself as the gross reduction or disappearance altogether of peaks as they near the edge of the SIM window. The 19 trend plots similar to Figure 3.3 were visually inspected, and the observed edge effect both at the start and end of the SIM window was recorded. For example, in Figure 3.3, the start of the edge effect is at 0 m/z in the SIM window and the end edge effect is at 5 m/z. The results are shown in Figure 3.6, as a function of the location of the start of the SIM window.

The edge effect varies considerably between each case, depending upon the starting m/z of the SIM window. This reflects the high variation in the trend observed between peaks in the same vicinity, and the somewhat subjective nature of the classification of the edge effect. However, there is sufficient evidence to extract a meaningful quantitation of the edge effect as a function of SIM window location, as shown in Figure 3.6 and summarised in Table 3.2.

Mid-point of SIM overlap at the at the relevant SIM window edge, mp (m/z)	Edge effect (start) (m/z)	Edge effect (end) (m/z)
$70 < mp \leq 219$	5	5
$219 < mp \leq 457$	2	8
$457 < mp \leq 500$	3	7

Table 3.2: Summary of observed edge effects

The boundaries ensure that an overlap of 10 m/z is required in all cases; this simplifies the acquisition and removes the majority of the edge effect while retaining a high throughput.

Finally, the edge effect regions are removed from the SIM windows, which are then truncated and concatenated to obtain a single spectrum. The SIM windows may still be overlapping, and so the truncation point is moved towards the centre of the SIM window such that the spectra are truncated in a signal-free region.

3.3 Results and Discussion

As described in Section 3.1, the stitching process provides two major benefits to the quality of the mass spectrum obtained using a DI-nESI FT-ICR MS instrument: improved mass accuracy and improved sensitivity. The benefits realised when analysing data obtained from an actual sample are presented and discussed in this section.

3.3.1 Mass Accuracy

The mass accuracy of the spectrum obtained using the SIM-stitching method is compared to that obtained using the leading commercial software package Xcalibur (Xcalibur v2.1.0, Thermo Scientific, Bremen Germany). The sample used is a standard mix comprising two polyethylene glycol and ten amino acids, 'PEG&AA', prepared using a standard protocol [18]. This mix is designed to produce a spectrum with a number of peaks spread evenly over the range 70–500 m/z. The peaks relating to 'known' compounds can be used either as internal calibrants during the stitching process, or as features to validate the resulting spectrum.

The SIM-stitching method was assessed against wide scan range (WSR) mode to ensure that high mass accuracy was being achieved. The optimised SIM-stitching method comprised of 21 adjacent 30 m/z windows between 70–500 m/z, each overlapping by 10 m/z to facilitate m/z-based stitching and to remove deleterious edge effects. Each SIM window was acquired for 15 seconds, with a single 15 second delay post electrospray initiation, giving a 5.5 minute total analysis time; the AGC target was 1×10^5. The PEG&AA standard was analysed in triplicate over the range 70–500 m/z by SIM-stitching and by WSR mode using AGC targets of 1×10^5 and 5×10^5, the lower setting comparable the SIM-stitching method and the higher comparable to the Thermo recommendation. A total acquisition time of 5.5 minutes was used for all three methods. For the SIM-stitching method, transient data for the 21 SIM windows were acquired, processed and each internally calibrated using a single calibrant. In total, 16 calibrants were used, with 5 used twice as they occurred in adjacent windows, see Table 3.3. The SIM windows were then stitched together along the m/z and intensity-axes. For WSR mode, transient data was acquired, processed and internally calibrated using 10 to 14 of the 16 calibrants used previously. It was necessary to use less calibrants due to the reduced sensitivity of WSR as shown in Table 3.4.

The mass accuracy results summarised in Table 3.4 were calculated using all known non-calibrant peaks, where a peak was only considered real if, within a sliding window of 1 ppm, exactly one peak appeared in each of the three spectra. Figure 3.7 shows, for a single replicate, the errors associated with each non-calibrant peak as well as the location of the internal calibrant peaks. Considering the three replicates for each method, SIM-stitching yielded the smallest average RMS error, smallest maximum absolute error and smallest average standard deviation. In WSR mode using an AGC target of 5×10^5, the largest errors were found, almost certainly resulting from increased space-charge effects in the ICR cell. The small increase in error in the WSR method using an AGC target of 1×10^5 versus SIM-stitching could be due to the lack of calibrants. However, replicate 3 of the WSR mode experiment with an AGC target of 1×10^5 and using 10 calibrants achieved similar mass errors to replicate 1, which used 14 calibrants. This suggests that increasing the number of calibrants above 10 has little benefit. Therefore it is unlikely that addition of two more calibrants, to equal the 16 used in SIM-stitching, would further improve the mass accuracy. This suggests that the algorithm used to calibrate SIM-stitched data could be improving the mass accuracy. Each window comprising the SIM-stitched wide scan was calibrated using unique calibration parameters allowing for precise correction of local m/z shifts, whilst in WSR mode just one set of calibration parameters is used for the whole spectrum, likely explaining the slightly higher mass errors of the WSR mode. The effect of this can be seen in Figure 3.7, in which the post calibration mass errors associated with the WSR mode spectra can be seen to vary considerably along the spectrum, whereas the

SIM-stitched spectrum allows for a more stable mass error.

Compound	Formula	Adduct	Exact Mass (Da)
Alanine	$C_3H_7NO_2$	H^+	90.054954
Proline	$C_5H_9NO_2$	H^+	116.070605
Proline	$C_5H_9NO_2$	Na^+	138.052549
Arginine	$C_6H_{14}N_4O_2$	H^+	175.118952
PEG-3	$C_2H_6O_2(C_2H_4O)_3$	Na^+	217.104644
PEG-4	$C_2H_6O_2(C_2H_4O)_4$	H^+	239.148915
PEG-4	$C_2H_6O_2(C_2H_4O)_4$	Na^+	261.130859
PEG-5	$C_2H_6O_2(C_2H_4O)_5$	H^+	283.175129
PEG-5	$C_2H_6O_2(C_2H_4O)_5$	Na^+	305.157074
PEG-6	$C_2H_6O_2(C_2H_4O)_6$	H^+	327.201344
PEG-6 (with one ^{13}C)	$C_2H_6O_2(C_2H_4O)_6$	Na^+	350.186644
PEG-7	$C_2H_6O_2(C_2H_4O)_7$	Na^+	393.209504
PEG-8	$C_2H_6O_2(C_2H_4O)_8$	H^+	415.253774
PEG-8	$C_2H_6O_2(C_2H_4O)_8$	Na^+	437.235718
PEG-9	$C_2H_6O_2(C_2H_4O)_9$	H^+	459.279988
PEG-9	$C_2H_6O_2(C_2H_4O)_9$	Na^+	481.261933

Table 3.3: The identity, empirical formulae, type of adduct and exact mass of the 16 internal calibrants used to calibrate the PEG&AA standard data.

3.3.2 Sensitivity

In this section, the effective dynamic range of the spectrum obtained using the SIM-stitching method is compared to that obtained using the leading commercial software package Xcalibur. The goal is to identify if the number of detected peaks and the dynamic range could be increased using SIM-stitching compared to WSR. The sample used is a liver extract from a wild-caught dab, a marine flatfish, and is prepared using a standard protocol [18]. The sample was kindly provided by the Centre for Environmental, Fisheries and Agricultural Sciences (Cefas, Weymouth, UK). As a complex biological sample, this liver extract is representative of the type of material used in a typical environmental toxicology study.

Six liver extracts were analysed using SIM-stitching, WSR mode with an AGC target of 1×10^5 and WSR mode with an AGC target of 5×10^5 between 70–500 m/z and acquired in triplicate. In each case the total acquisition time was 5.5 minutes, including a 15 sec data acquisition delay post electrospray initiation, to facilitate a direct comparison

Scan Type (AGC Target)	Replicate	N_C	N_P	RMS Error (ppm)	Max Absolute Error (ppm)	SD of Error (ppm)	Summary Mean RMS Error (ppm)	Summary Max Absolute Error (ppm)	Summary Mean SD of Error (ppm)
WSR	1	14	26	0.254	0.617	0.257			
(1×10^5)	2	14	26	0.199	0.461	0.194	0.234	0.617	0.228
	3	10	20	0.248	0.613	0.232			
WSR	1	14	35	0.612	2.041	0.593			
(5×10^5)	2	14	35	0.519	0.915	0.441	0.565	2.041	0.492
	3	14	35	0.564	1.408	0.443			
SIM	1	16	39	0.179	0.423	0.161			
-stitched	2	16	39	0.193	0.459	0.181	0.181	0.475	0.165
(1×10^5)	3	16	39	0.171	0.475	0.154			

Table 3.4: Summary of mass errors for SIM-stitching and WSR methods (with AGC targets of 1×10^5 and 5×10^5), determined from analysis of the PEG&AA standard. Each method was characterised in triplicate, and the numbers of internal calibrants (N_C) and other known peaks used to calculate mass errors (N_P) are shown.

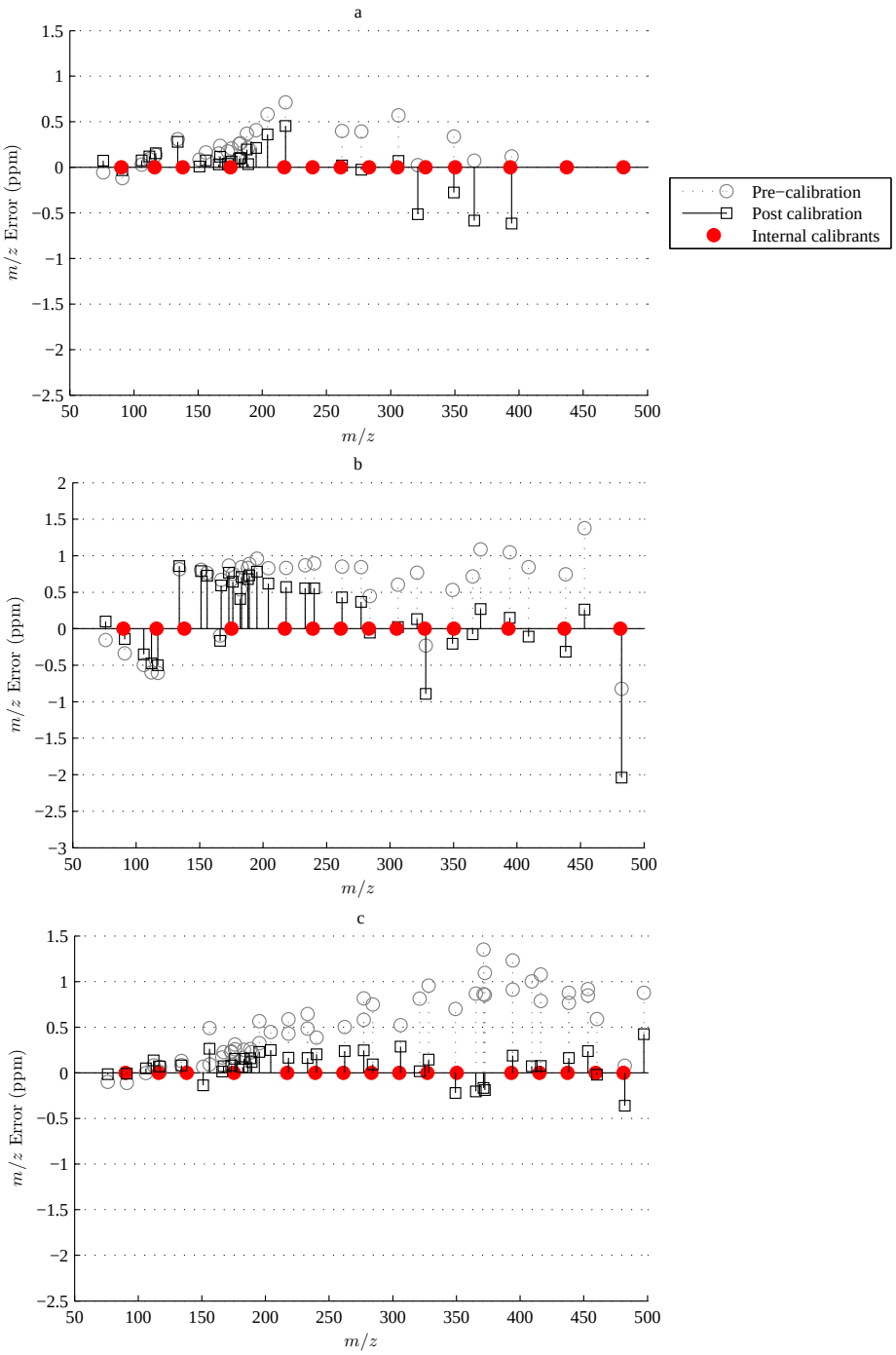

Figure 3.7: Location of internal calibrants and mass errors of all known non-calibrant peaks as a function of m/z for one replicate of the PEG&AA standard analysed by (a) WSR mode with an AGC target of 1×10^5, (b) WSR mode with an AGC target of 5×10^5, and (c) SIM-stitching method. SIM-stitched data was calibrated using 16 internal calibrants, while the WSR methods were calibrated using 14 of these; two were missing due to lack of sensitivity of the WSR method.

between the methods. This m/z range of 70–500 corresponds to the highest density of low molecular weight endogenous metabolites. For all methods, transient data was collected and processed as described for the PEG&AA studies above, except that known endogenous metabolites were used as calibrants, as shown in Table 3.5. Visual inspection of the spectra, which were normalised to the largest peak, with a five-fold zoom along the y-axis revealed that both WSR with an AGC target of 5×10^5 and SIM-stitching detected a more dense spread of peaks throughout the entire 70–500 m/z range than WSR with an AGC target of 1×10^5. A further 200-fold zoom within 355–360 m/z revealed that SIM-stitching detects significantly more peaks than WSR with an AGC target of 5×10^5, whilst WSR with AGC target of 1×10^5 failed to detect any peaks.

Metabolite, M	Formula	Ion Form	Exact Mass (Da)
Alanine	$C_3H_7NO_2$	$[M+H]^+$	90.05495
Alanine	$C_3H_7NO_2$	$[M+Na]^+$	112.03690
Proline	$C_5H_9NO_2$	$[M+H]^+$	116.07060
Valine	$C_5H_{11}NO_2$	$[M+H]^+$	118.08625
Alanine	$C_3H_7NO_2$	$[M+K]^+$	128.01084
Proline-betaine	$C_7H_{13}NO_2$	$[M+H]^+$	144.10190
Taurine	$C_2H_7NSO_3$	$[M+Na]^+$	148.00389
Proline	$C_5H_9NO_2$	$[M+K]^+$	154.02649
Valine	$C_5H_{11}NO_2$	$[M+K]^+$	156.04214
Taurine	$C_2H_7NSO_3$	$[M+K]^+$	163.97782
Proline-betaine	$C_7H_{13}NO_2$	$[M+Na]^+$	166.08385
Proline-betaine	$C_7H_{13}NO_2$	$[M+K]^+$	182.05779
Glucose, Fructose, Galactose or Mannose	$C_6H_{12}O_6$	$[M+Na]^+$	203.05261
Glucose, Fructose, Galactose or Mannose	$C_6H_{12}O_6$	$[M+K]^+$	219.02655
D-glycerophosphocholine	$C_8H_{20}NO_6P$	$[M+H]^+$	258.11010
Inosine	$C_{10}H_{12}N_4O_5$	$[M+H]^+$	269.08805
D-glycerophosphocholine	$C_8H_{20}NO_6P$	$[M+Na]^+$	280.09205
D-glycerophosphocholine	$C_8H_{20}NO_6P$	$[M+K]^+$	296.06598
Inosine mono-phosphate	$C_{10}H_{13}N_4O_8P$	$[M+H]^+$	349.05438
Inosine mono-phosphate	$C_{10}H_{13}N_4O_8P$	$[M+Na]^+$	371.03632
S-Adenosyl-homocysteine	$C_{14}H_{20}N_6O_5S$	$[M+H]^+$	385.12887
S-Adenosyl-homocysteine	$C_{14}H_{20}N_6O_5S$	$[M+Na]^+$	407.11081

Table 3.5: The identity, empirical formula, form of ion and exact mass of the internal calibrants used to calibrate the dab liver extract data collected using the SIM-stitching method.

For each method, the total number of peaks was counted between 70–500 m/z for each of the six liver extracts as shown in Figure 3.8, and the dynamic ranges were determined. The average peak counts for the six liver extracts were 3046 for SIM-stitching, 575 for WSR mode with an AGC target of 1×10^5 and 1719 for WSR mode with an AGC target of 5×10^5. SIM-stitching detected on average 5.3 and 1.8 times more peaks than both WSR methods with AGC targets of 1×10^5 and 5×10^5, respectively, even though each SIM window received 21 times less signal averaging than the full scan WSR spectra. Even when compared to an experiment, where five times as many ions enter the ICR detector cell (i.e. WSR mode with an AGC target of 5×10^5), the SIM-stitching method still detected around 44% additional ion species. The relative standard deviation (RSD) of the number of peaks detected across the six fish livers was 7% for SIM-stitching, 5% for WSR with an AGC target of 1×10^5 and 7% for WSR with an AGC target of 5×10^5. This showed that each method was able to generate consistent peak counts when analysing genetically different fish from the same species, and highlights the reproducibility of all methods. RSDs of peak intensities were calculated for all 22 known internal calibrants, as shown in Table 3.5, across triplicate analyses of one fish. This was repeated for each of the three methods. The mean and maximum RSD values were 8.1% and 16.5% for the SIM-stitching method, 11.4% and 23.2% for WSR with an AGC target of 1×10^5, and 11.3% and 17.9% for WSR with an AGC target of 5×10^5. These results further emphasise the reproducibility of all three methods, but more importantly highlight another advantage of the SIM-stitching method, specifically that by acquiring the metabolic data using narrow SIM windows, the reproducibility of the intensity profiles are increased relative to wide scan methods. The mean dynamic range of 16,061 for the SIM-stitched data was ca. 22-fold greater than for WSR with an AGC target of 1×10^5, which achieved a measured mean dynamic range of 726, and ca. 4.3-fold greater than for WSR with an AGC target of 5×10^5, which achieved a measured mean dynamic range of 3,684.

When acquiring data in a single large window (e.g. 70–500 m/z), which for a complex biological mixture will comprise a large number of different ion species, a considerable percentage of the AGC target value will be occupied by the most highly abundant ions, and there is less chance that low abundance species will achieve the detection threshold of ca. 200 ions [1]. However, when analysing a 30 m/z SIM window, only ion species within the specified range are transferred to the ICR detector and all other ions are excluded by isolation waveforms. This greatly reduces the number of individual ion species, meaning that the AGC target value is no longer dominated to the same extent by highly abundant ions, which in turn allows low abundance species to reach the detection threshold. In effect, SIM-stitching allows more ions per nominal mass unit to transfer into the ICR detector; e.g. when using an AGC target of 1×10^5 and scanning a 30 m/z window, ca. 3334 ions per nominal mass unit are measured, whereas the higher AGC target of 5×10^5 when scanning

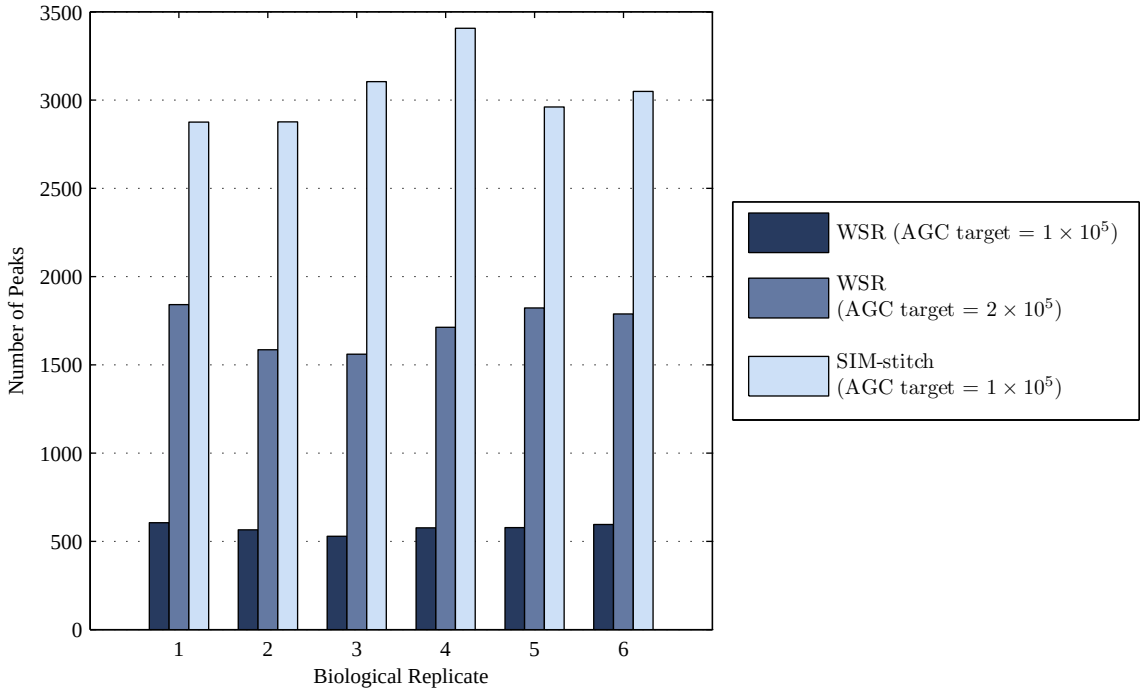

Figure 3.8: Comparison of the total number of peaks detected in the range 70–500 m/z for six different dab liver extracts using wide scan range (WSR) mode with an AGC target of 1×10^5 ions and 5×10^5 ions, and the optimised SIM-stitching method with an AGC target of 1×10^5 ions.

70–500 m/z only ca. 1163 ions per nominal mass unit are measured. Therefore, compared to WSR with an AGC target of 5×10^5, SIM-stitching increases the number of ions scanned per nominal mass unit by ca. 3-fold. This yields greater sensitivity and dynamic range, and consequently the number of peaks detected is increased by ca. 1.8-fold. The increase is significantly larger when compared to using WSR with an AGC target of 1×10^5. In this case, SIM-stitching increases the number of ions scanned per nominal mass unit by ca. 15-fold, yielding a ca. 5.3-fold increase in peak detection. The reduced signal averaging, inherent to the SIM-stitching method, apparently does not offset this sensitivity increase too significantly.

3.4 Conclusion

Objective 3 of this thesis was to improve the sensitivity and accuracy of the mass spectrum. In this chapter, it has been shown how a SIM-stitching method [18], never before applied to metabolomics data, allows significant improvement to both sensitivity and mass accuracy in mass spectra acquired using DI nESI FT-ICR. This is achieved by intelligently combining several narrow spectra into a single, wide spectrum. Without the use of SIM-stitching, a compromise is required between sensitivity and mass accuracy, whereas this approach offers vastly improved sensitivity without reducing the mass accuracy obtained; indeed, mass accuracy is also improved. Sensitivity is key to the complete analysis of the metabolome since, as shown in Section 1.2, many metabolites exist with low abundance. Without using the SIM-stitch method, it is inevitable that many metabolites are not detected, and it is for this reason that the metabolomics community has shown substantial interest in the published work, with 32 citations up to September 2010.

In order to maximise mass accuracy, the AGC target and hence ion count is kept low at 1×10^5 ions, and by acquiring multiple, overlapping SIM windows, the sensitivity is also optimised. The results have shown superior performance in both sensitivity and mass accuracy when compared to spectra acquired using the typical WSR mode, even when WSR acquisition is optimised for sensitivity with an AGC target of 5×10^5 ions, or mass accuracy with an AGC target of 1×10^5 ions. By using SIM-stitching and the same AGC setting, sensitivity can be increased by a factor of 5.3, and mass error can be decreased by a factor of 1.3. If the manufacturer-recommended AGC setting is used in WSR mode, sensitivity can be increased by a factor of 1.8 and mass error can be improved by a factor of 4.3.

Using the instrument's SIM mode, which allows spectra to be acquired over a narrow mass range, detection sensitivity was enhanced by collecting the data as a series of narrow, overlapping windows of width 30 m/z. The increased detection sensitivity allowed

ion transmission to the ICR cell to be reduced to 1×10^5 ions, thus reducing space-charge effects and allowing high mass accuracy. Mass accuracy was enhanced further by the calibration method, where each of the 21 overlapping SIM windows was individually, internally calibrated prior to stitching, which allowed correction of local m/z shifts. The maximum absolute mass error was shown to be 0.48 ppm. When compared to conventional DI-ESI MS methods, the SIM-stitching method can provide a significant increase in the number of metabolites detected, thereby increasing coverage of the metabolome. In addition, it maximizes the major strengths of FT-ICR MS, those of high mass accuracy and resolution, thereby facilitating improved peak identification via the calculation of empirical formulae.

It has also been shown that regions of unusual noise artefacts exist in the spectrum, which if ignored when using the manufacturer's software, introduce large amounts of noise into the resulting peak list. By setting exclusion regions, these peaks are removed, resulting in a spectrum with fewer noise peaks. A further important characteristic of the FT-ICR MS instrument revealed here is the non-linear quantification in SIM mode, demonstrated by comparing the intensity of peaks as the scan region is varied. This has allowed this problem to be corrected, and the spectra derived from similar instruments should be carefully checked for this undesirable characteristic.

On a practical point, the collection of spectra in narrow SIM windows produces many files, of much smaller size than comparable WSR mode spectra which can be unwieldy to access and analyse. Additionally, the SIM-stitch technique has been integrated into a tool with a graphical user interface, see Appendix A. Together with a user's manual, this software is used extensively in our lab, and beyond, to assist with organising and processing the sets of data files created. The coupling of a NanoMate to the LTQ FT enables a fully automated analysis in approximately 5.5 minutes per sample, which is conducive for high-throughput analyses.

In summary, this SIM-stitching approach [18], which is being adopted locally [16, 79] and by other laboratories [38, 80], has wide applicability to the measurement of any complex chemical mixture by FT-ICR MS. In particular, for biological samples, this new method has increased the quantity of metabolites detected, and enhanced the quality of metabolite identification [81, 79, 16, 82, 83].

3.5 Future Developments

The method presented in this chapter results in significant improvements in the quality of mass spectra obtainable using FT ICR MS. During the investigation, several areas of further development were identified:

1. It was observed that the noise level with the SIM windows varies as a function of m/z. Analysis of this variation and multiple measurements of the noise level within each SIM window would result in a slightly more accurate noise measurement, and hence SNR value for peaks.

2. As shown in Figure 3.6, the edge effect is a complex phenomenon that would benefit from more extensive characterisation to reduce the SIM window overlap where possible, thus increasing throughput.

3. The density of mass spectra of even the most complex biological samples is not consistent, with the result that some SIM windows contain fewer ions, reflected by a lower total ion current (TIC). A further optimisation of the SIM-stitching method would use *data dependent analysis*, where the width and location of the SIM windows be dynamically configured depending upon the density of peaks, and their intensity, in the mass spectrum. This would maximise the benefits of sensitivity and mass accuracy that SIM-stitching brings by ensuring that each SIM window contains the optimal number of ions.

4. Implementation benefits could be realised, in terms of the time required for the SIM-stitching algorithm to process data from large experiments. Techniques such as parallelisation using computing clusters, or on a smaller scale, the use of multi-processor graphics boards, would improve this significantly.

CHAPTER 4

NOISE FILTERING

4.1 Introduction

In Chapter 3, a new method to improve the sensitivity and mass accuracy of a mass spectrum is presented. In this chapter, the challenge of objective 2, as described in Section 1.3, is addressed — that of reliably filtering the noise present in the mass spectra to maximise the number of real metabolites detected by DI nESI FT-ICR MS, while retaining the high-throughput characteristics of this approach compared with chromatographic-based MS methods.

Sources of noise in nESI FT-ICR MS include:

1. Background chemical noise that arises as the result of compounds that interfere with the composition of the biological sample entering the instrument. During analysis of data collected for this thesis, a contaminant likely to have originated from air was found with high abundance and high variability, in the spectrum of a sample of fish liver extract. The compound, decamethylcyclopentasiloxane, is from the siloxane family and is a known atmospheric contaminant [84]. As well as such gas-phase contamination, liquid impurities can appear in the sample, e.g. as a result of sample preparation methods.

2. Electrical and thermal shot noise [85], present in any sensitive electronic equipment, increase the noise levels.

3. By far the most problematic contributor of noise in the spectrum is ionisation noise, discussed in Section 2.3, and particularly ionisation suppression and enhancement [2], which causes large variations in the observed quantity of molecules.

4. Once ionised, ions are affected by Coulombic interactions due to the large number of ions present in a small space [42, 44]. The effect of these interactions is to modify

the orbiting frequency of the ions in the ICR cell, adding frequency noise to the spectrum.

5. The DI nESI FT-ICR MS used for this thesis is prone to the occasional loss of the electrospray, most likely due to particulates blocking the spray nozzle (a 'drop-out'). When 'drop-out' occurs, the total ion current (TIC) of the signal temporarily drops significantly, often to $< 50\%$ of typical values, and the resulting spectrum shows elevated noise levels and is not representative of the sample composition [17].

6. FT artefacts such as Gibbs oscillations, while reduced by apodisation as described in Section 2.6.2, are still present.

Some of this noise can be reduced prior to the measurement, for example by optimisation of the ICR detector cell parameters [44, 86], or after the measurement, during the pre-processing stages [87]. As discussed in Section 2.6.1, the signal-to-noise ratio (SNR) of the mass spectrum can be improved by averaging multiple scans, where the theoretical SNR gain, defined in equation (2.8), is $\sqrt{n}$ (where n is the number of scans acquired). In reality, the actual SNR realised is limited by other noise effects, including instrument parameter drift and changes in the sample composition, which can contribute to an increased variability of measured ion intensities over time. Furthermore, these acquisitions are in the context of a high throughput study, and as such there is a balance in achieving high SNR, while minimising the number of scans acquired. Each scan takes approximately one second; in a typical SIM-stitch study comprising 21 SIM windows, three replicates and n scans per SIM window, each sample requires $63n$ seconds. Where a study contains many samples, it is important to reduce n as much as possible without compromising SNR excessively. Therefore, as part of the optimisation of signal quantification, the actual benefit realised is measured as the number of scans is increased when analysing a liver tissue extract.

Despite optimising the signal acquisition stage, ultimately a mass spectrum will contain both signal and noise. Profiling the complete metabolome will necessarily require the detection of low intensity metabolites, particularly as it has been shown that, at least for NMR studies, the number of unique metabolites is inversely proportional to their abundance [17]. The crucial step is in discarding noise while retaining and then measuring real peaks. For a single spectrum, this is typically achieved by setting a peak area threshold [88, 89] or a SNR threshold, and retaining only peaks that exceed this specification. SNR thresholds used include 3:1 [90, 1, 91, 92, 42] (the often quoted 'limit of detection'), 5:1 [93], 10:1 [94] (the 'limit of quantification') and higher.

However, a hard threshold technique such as this either results in many low abundance ions not being detected, or many noise peaks being falsely counted as real. In addition, it

can incorrectly remove ions with abundances that are intermittently reduced by adverse effects during ESI. This potential loss of data has been discussed recently by Wilson *et al.* [95], who developed an algorithm using liquid chromatography mass spectrometry (LC-MS) data and proposed first applying a peak intensity threshold across all mass spectra to generate one set of 'real' peaks. This matrix of peak intensities will contain multiple missing values (i.e. zero intensities) since peaks present in some spectra will not be detected in others, as they are below the detection threshold. This threshold is then removed, and all the missing values are filled in by reintegrating the areas around every 'real' peak. The problem of reliably detecting low intensity ions remains, however, since a hard threshold must initially be used. Other means of filtering noise peaks in mass spectra include by peak variance [96] or resolution [97, 98], and filtering by peak shape is used in LC-MS [99, 94]. While using additional information such as peak resolution, shape or variance will certainly provide further evidence for the existence of a 'real' peak, it has been observed that in practise, noise does not appear sufficiently distinct from low intensity signal peaks to differentiate solely on peak parameters or variance.

It is worth noting that ion capture and detection in FT-ICR MS is a discrete process, due to the manner in which ions are collected and then 'held' by the magnetic field during the detection period in the ICR detector cell. After detection, the ions are released and a newly collected ion package is transferred to the ICR detector cell, and the process repeats. It is therefore appropriate to use a 'replicate filter', where a single biological sample is measured multiple times, and only those peaks common to a minimum number of the replicates are retained; and/or a 'sample filter', where only those peaks common to a minimum number of biological samples within a group (e.g. a treatment group) are kept in the dataset. To date, these filters have been used empirically, and there is no consensus on the optimal number of replicates or samples, or the filtering parameters. An example replicate filter applied to DI FT-ICR MS data stipulates that each peak be present in at least two out of three replicate analyses [80], and Quick *et al.* first applied this filter to DI-ESI time-of-flight (TOF) MS data with the requirement that each peak be present in all 3 replicate analyses [96, 46]. Example sample filters include requiring a peak to occur in 50% [94], 60–75% [80] or 80% [88] of the total number of samples in the group.

In Chapter 3, a significant increase in detection sensitivity in DI nESI FT-ICR MS based metabolomics has been demonstrated by recording each wide-scan mass spectrum as a series of overlapping selected ion monitoring (SIM) windows. In this chapter, SIM-stitching of DI nESI FT-ICR MS data is applied to both real and simulated data to assess the performance of a three-stage filtering method, shown in Figure 4.1. This method requires multiple samples to be acquired in triplicate, and applies a SNR hard threshold followed by a replicate filter and a sample filter, to produce a filtered peak list that contains the

intensity of each peak in each sample. The results are compared with one- and two-stage filtering methods, thus showing how an optimal filtering method can be objectively selected.

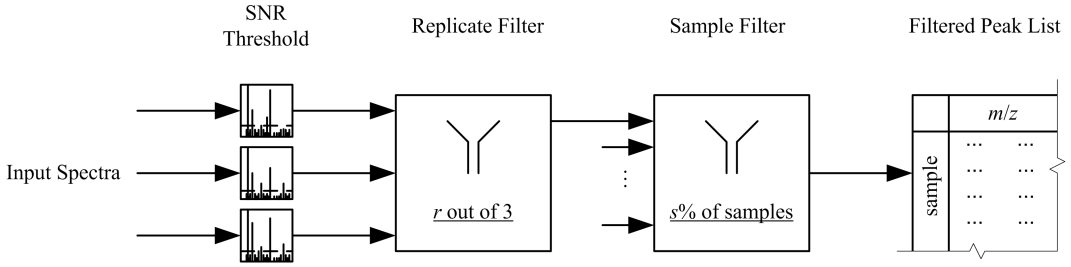

Figure 4.1: The three stage filtering schema, showing SNR threshold, replicate filter with parameter r, and sample filter with parameter s.

4.2 Optimising Number of Scans

Consider a single scan in which a peak's intensity is measured as a true signal component μ, plus a noise component normally distributed with zero mean and standard deviation σ; the peak intensity will therefore have a normal (μ, σ^2) distribution. By averaging n scans, each with true signal intensity μ and a similarly distributed noise component, the mean peak intensity has a normal $(\mu, \sigma^2/n)$ distribution [100], i.e. the standard deviation (SD) is $\sigma\sqrt{n}$, and hence the relative (to μ) standard deviation (RSD) is reduced by a factor of $\sqrt{n}$. However, as discussed previously, this benefit is limited by other sources of noise. Since a series of replicate spectra are composed of many peaks, where each peak has a RSD across the spectra, a median RSD value can be derived for the dataset. This measure has recently been proposed as a valuable measure of spectral variability in a metabolomics experiment [101]. Therefore, to determine the optimal number of scans, n, that should be acquired for each spectrum, a sample was analysed for at least 95 scans with the instrument configured in SIM scan mode and using three separate m/z range settings: 110–140 m/z, 230–260 m/z and 470–500 m/z. For each m/z range setting, the first five scans were discarded, then three sets of n (n from 3 to 30) consecutive scans were averaged, e.g. when $n = 4$, scans 1–4, 5–8 and 9–12 were averaged to produce three mean spectra. The locations of peaks common to all three spectra were then determined, and the RSD values, calculated as the standard deviation relative to the mean intensity of these peaks, were evaluated. The median RSD value was then taken as a representative measure of spectral reproducibility in each case.

The results in Figure 4.2 show a rapidly decreasing median RSD with increasing n for all mass ranges. Data was acquired three times during the same sample run, which reveals

considerable variability between repeats, and this could be caused by slow fluctuations in the electrospray process. The highest decrease in median RSD appears at around $n = 5$ scans, which shows an approximately two-fold improvement in median RSD of 5% compared to $n = 1$ scan, and so $n = 5$ is used for the rest of the modelling presented in this section. Beyond $n = 5$, the median RSD tends to decrease only marginally, potentially due to longer-term variations which counteract the anticipated gain due to signal averaging. The technical variation associated with the DI nESI FT-ICR measurement of a liver extract, ca. 5% as shown in Figure 4.2, is comparable with the technical variation observed in NMR studies [101].

4.3 Simulated Mass Spectra

Spectra simulated to quantify the performance of the filtering methods described here must be realistic in terms of the distribution of peak intensities in a spectrum, the m/z position of the peaks, the variation of peak intensities across technical replicates and the amount of noise present. First, the transient signal y, at sample point n, as acquired by FT-ICR MS at the point the ions are inside the ICR detector cell is modelled. The ion-induced amplitude model, based on the noise-free form shown in equation (2.4), is

$$y_n = \sum_{i=1}^{N_{real}} A_i' \sin\left(2\pi f_i n T + \Phi\right) \exp\left(-\lambda n T\right) + b_n, \tag{4.1}$$

where n is the sample number; N_{real} is the number of 'real' ion clouds in the cell during acquisition; A' is the observed ion intensity, including a random noise component; f is the cyclotron frequency; Φ and λ are the phase and decay constant, respectively (both assumed constant, and independent of ion species); T is the duration of the transient sampling; and b_n is a random additive noise. Frequency perturbations from Coulombic interactions are not included in this model, since the filtering method presented here focusses on noise in measured peak intensities.

The first stage in creating the model spectra, involves determining the parameters of the statistical simulation from experimental mass spectral data, including the distribution of A', f and b. These three parameters are considered in turn. The intensity of each simulated real peak is modelled as

$$A_i' = A_i c, \tag{4.2}$$

where A_i is the true intensity of peak i (i.e. determined experimentally), and c (≥ 0) is a random error associated with the intensity of that real peak. Since this is a statistical

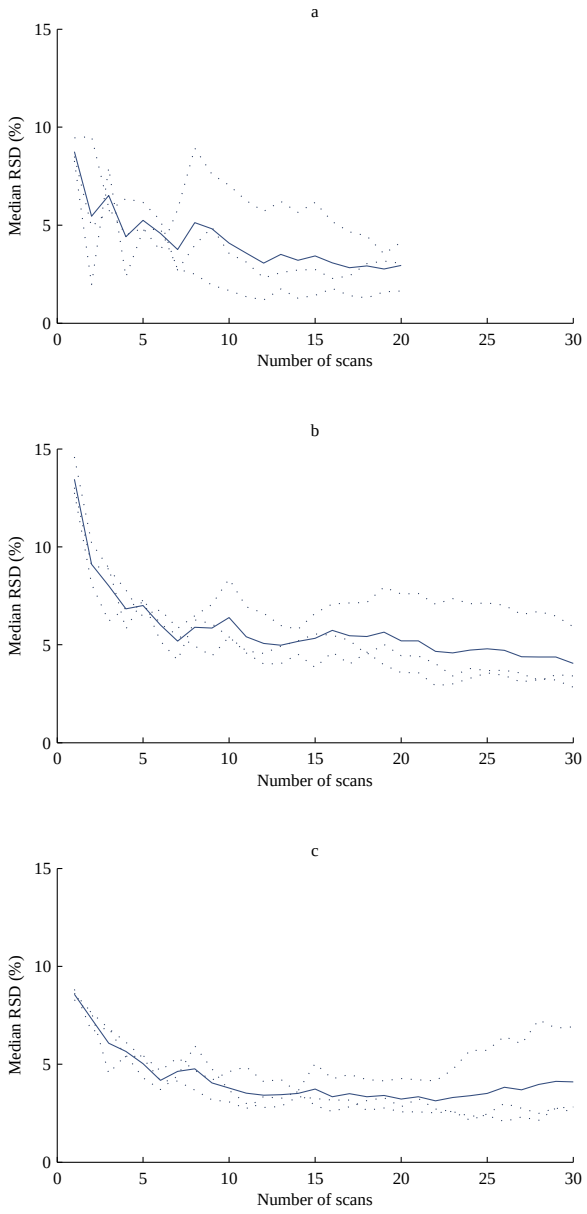

Figure 4.2: Median RSD of three spectra with mass ranges 110–140 m/z (a), 230–260 m/z (b), and 470–500 m/z (c) as the number of scans averaged to form each spectrum, n, is increased. The experiment was repeated three times each, shown as dotted lines, with the average as a solid line. The maximum number of scans in the range 110–140 m/z is limited by the second experiment repeat, which produced an insufficient number of total scans (66 obtained, 90 required).

model with parameters based on measured data, it is necessary to determine A_i for each i such that the measured and simulated spectra have matching distributions of real peak intensities. To determine the intensity distribution of real peaks (A_i), the distribution of noise is estimated and subtracted from the distribution of observed peaks, which contain both signal and noise. The distribution of noise peaks is estimated from a simulated spectrum containing only time-domain AWGN, as expected in real spectra (see Section 2.6.5). A spectrum with signal, as well as noise, will contain fewer noise peaks than the simulated noise spectrum over the same mass range due to the masking of some noise peaks by real signal. Therefore, in order to better match the number of peaks in the simulated spectrum to the measured spectrum, 6000 peaks uniformly distributed on the m/z axis are removed from the simulated noise spectrum. This number was determined empirically and provides the best match between the peak intensity distributions of the real and simulated spectra.

This process of creating noise distributions is repeated ten times. The distribution of observed peaks is acquired from ten replicate measured spectra from the same sample. The mean and range of noise peaks at a given SNR for the ten spectra is plotted in Figure 4.3, together with the relative difference. A good match is seen at low SNR, with the difference between histograms (for SNR$<$2.5) remaining within 0.3% of the number of peaks in the measured spectrum, supporting the model. Since the relative difference between the histograms is negligible below a SNR of 2.5, it is assumed that these peaks are noise in the measured spectrum. Thus, the difference between distributions for SNR$\geq$2.5 represents the population of real peaks in the measured spectrum, and an empirical cumulative distribution function (CDF) can be formed, as shown in Figure 4.4, from the difference distribution. By linearly interpolating over the empirical CDF, one can create a representative population of peak intensities using inverse transform sampling. This residual distribution is also used to estimate the number of real peaks in the model (N_{real}).

To capture the noise associated with real peaks in the model, the parameter c should be randomly distributed according to the intensity of noise seen in real measured peaks. To this end, three mass spectra of the same sample were acquired and compared. Peaks identified as common between all three spectra and with an arbitrarily high SNR of at least 6.5 in the first spectrum were extracted, to represent peaks likely to be real. For each of these peaks, the peak area, Y, relative to the mean peak area across the three replicates (Y_{mean}) was calculated. Taking the logarithm of this ratio allows a good, dense distribution to be revealed and a histogram of the resulting values is shown in Figure 4.5. The histogram is parameterised as a normal distribution, with $\mu = -0.0024$ and $\sigma = 0.0738$, and used as the random model for parameter $\log_{10}(c)$.

The model includes f, the peak frequencies, which are derived from real data to represent

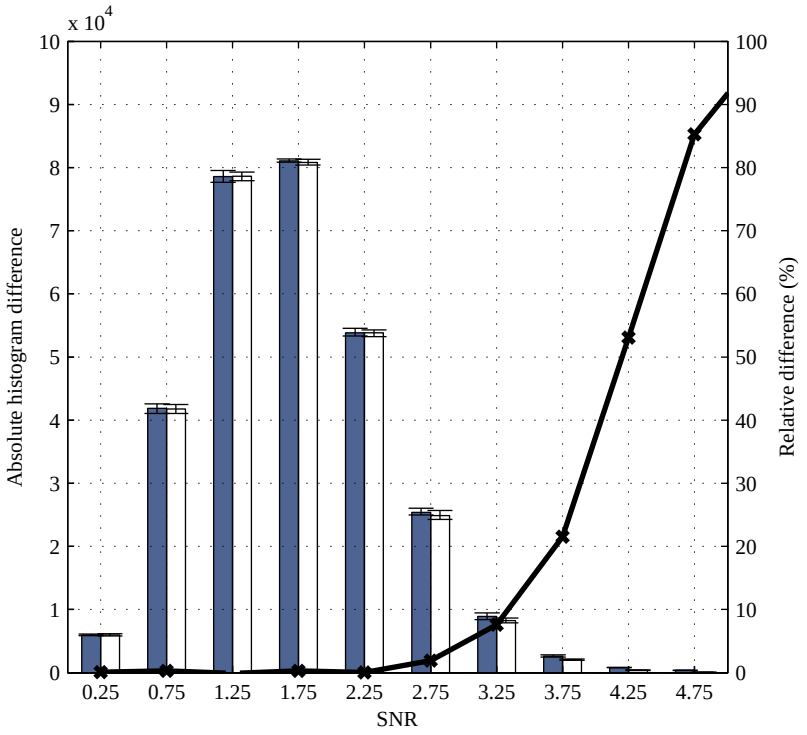

Figure 4.3: Mean ($n = 10$) histogram of peak SNR in measured spectra (shaded bars) and simulated noise spectra (clear bars) and the difference between these relative to the mean number of peaks in the measured spectra, plotted as the thick line. The histograms show very good correspondence at low SNR (< 2.5). The error bars represent the range.

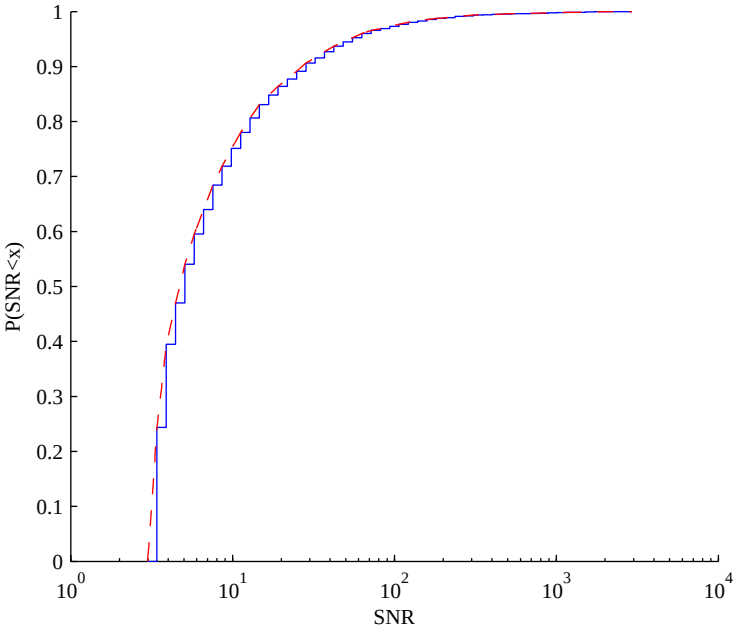

Figure 4.4: Empirical CDF of the population of real peaks (estimated from the difference between the simulated spectrum and the measured spectrum) showing the probability that the SNR is below the corresponding SNR on the x-axis. The broken line shows the linear interpolation for all SNR values.

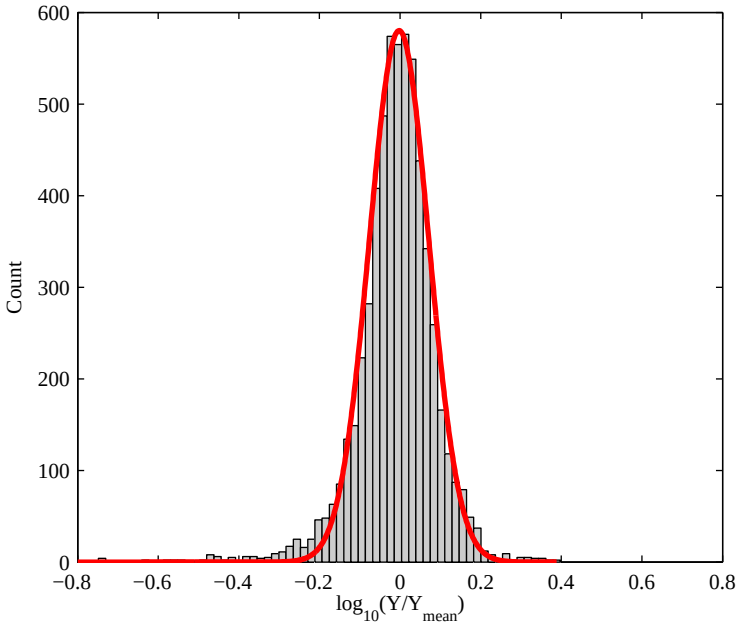

Figure 4.5: Distribution of the intensity of replicate peaks relative to the mean intensity of the peak across all replicates. A normal distribution is shown fitted to the data. Two outliers, both at -1.4, are not shown.

the varying peak densities across the spectrum and therefore any associated artefacts, such as peak overlap. The empirical CDF of peaks in a measured spectrum with a minimum SNR of 6.5 is firstly generated. The CDF is then used to determine the frequency f_i of each peak in the model by inverse transform sampling. Finally, noise b is modelled as Rayleigh distributed random noise, with Rayleigh parameter σ selected to provide an average noise level of unity in the frequency domain such that peak SNR as defined in equation (2.8) is equal to peak height.

All components of the model are now known, with phase shift ϕ and decay constant λ arbitrarily selected as 0 and 2, respectively. The number of data points M is set to the number of data points in the measured spectrum. The transient signal is thus constructed according to equation (4.1). The transient is then apodised using a Hanning function, zero-filled once and Fourier transformed using the FFT algorithm. Note that multiple scans per spectrum are not explicitly modelled, but the end effect of averaging multiple scans is included by virtue of the fact that the set of peak SNR values is measured directly from the (multiple scan) measured spectra.

4.3.1 Assessing the Noise Filter

The spectrum model described above generates a mass spectrum with realistic peak intensities and frequencies. It includes 4441 signal peaks, derived from measured spectra, and random noise. By generating spectra using this model, the performance of the filtering strategy can be assessed. A set of 30 random spectra is created, as described above, labelled $1a$, $1b$, $1c$, $2a$, $2b$, $2c$, ..., $10a$, $10b$, $10c$, where (a, b, c) indicate the replicate spectra and (1...10) the repeat experiments. After applying a filtering configuration, the resulting peak list is compared with the 'true' signal frequencies, f. Those peaks common to f to within 2.5 ppm are considered correctly identified, while the remainder are considered noise. The 2.5 ppm error window allows for error in the frequency estimation caused by noise-induced peak shape distortion while avoiding detection of neighbouring peaks. As stated previously, one measure of filter performance is the number of real peaks identified. This is added to the empirical probability that a detected peak is real, defined as shown in equation (4.3).

$$\mathrm{P(detected\ and\ real)} = \frac{\text{number of detected peaks which are real}}{\text{number of detected peaks}} \tag{4.3}$$

4.4 Three Stage Noise Filtering

4.4.1 Method

Given a spectrum, containing signal and noise, the aim of any noise filter is to remove noise while retaining signal. Therefore, to assess the performance of a noise filter it is necessary to know which peaks are due to real signal, and which are due to noise. However, given a spectrum acquired from a complex tissue extract, one cannot easily classify low intensity peaks as signal or noise. To address this problem, a statistical model is developed, as described in Section 4.3, that allows the realistic simulation of mass spectra, including signal and noise. The simulated FT-ICR mass spectra are generated in the time domain with signal and noise features distributed according to observations from measured data, and provide a means of quantifying the performance of the three-stage noise filtering method. The simulated spectra include random noise, both as additive white noise induced by thermal effects, and as random fluctuations in the intensities of 'real' signal peaks across replicate spectra caused by variations in the ESI process and ion detection.

Multiple simulated spectra are created, representing replicate spectra of multiple biological samples (e.g. of the same phenotype and treatment group). These are then used as the input to three different configurations of the noise filter, which, with reference to Figure

4.1 are:

1. A SNR threshold filter applied to a single replicate spectrum, such that only peaks with a SNR above a threshold are considered real;
2. A SNR threshold filter, followed by a 'replicate' filter, in which real peaks must be present in at least r out of three replicates;
3. A SNR threshold filter, followed by both a 'replicate' filter and a 'sample' filter. In the sample filter, each peak must be present in at least $s\%$ of the number of biological samples.

By comparing the output (filtered peak list) with the signals used to generate the simulated spectra, the performance of each filter can be quantified.

4.4.2 Results and Discussion

SNR Threshold Noise Filtering

This strategy attempts to reduce noise from the spectrum by applying a hard threshold to the SNR of each peak such that only peaks with a SNR above a threshold are retained. For this experiment, only the first replicate of each sample is used $(1a, ..., 10a)$, giving a total of 10 spectra. Each spectrum is filtered as described above. The results for typical SNR thresholds of 3.0, 5.0, 8.0 and 10.0, for ten repeat experiments, are shown in Table 4.1. It is apparent that one can either detect many real peaks (at low SNR threshold) or have high confidence that a detected peak is real (at high SNR threshold), but not both. Since it is necessary to achieve both of these in metabolomics studies, SNR threshold noise filtering alone is not recommended.

SNR threshold	N_{real}	N_{noise}	$P_{d,r}$
3.0	3427	10987	0.238
5.0	2159	50	0.977
8.0	1298	14	0.989
10.0	1044	8	0.992

Table 4.1: Metrics for simulated spectra, filtered by SNR threshold only: mean number of real peaks, mean number of noise peaks and mean probability that a detected peak is real, based on 10 repeats.

SNR threshold	Replicate filter	N_{real}	N_{noise}	$P_{d,r}$
3.0	2 out of 3	3413	807	0.809
3.0	3 out of 3	2332	29	0.988
5.0	2 out of 3	2023	29	0.986
5.0	3 out of 3	1573	10	0.994

Table 4.2: Metrics for simulated spectra, filtered by SNR threshold and replicate filtered: mean number of real peaks, mean number of noise peaks and mean probability that a detected peak is real, based on 10 repeats.

SNR threshold	Replicate filter	Sample filter	N_{real}	N_{noise}	$P_{d,r}$
2.6	2 out of 3	40%	3913	150	0.963
2.6	2 out of 3	60%	3646	50	0.986
2.8	2 out of 3	30%	3956	253	0.940
2.8	2 out of 3	50%	3753	62	0.984
3.0	2 out of 3	30%	3929	267	0.959
3.0	2 out of 3	50%	3659	56	0.985
3.5	2 out of 3	20%	3882	156	0.961
3.5	2 out of 3	30%	3705	96	0.975
4.0	2 out of 3	10%	3825	196	0.951
5.0	2 out of 3	10%	2857	84	0.971

Table 4.3: Metrics for simulated spectra and three stage filtering: number of real peaks, number of noise peaks and probability that a detected peak is real, based on a single repeat.

SNR Threshold and Replicate Noise Filtering

In this strategy, firstly SNR threshold filtering is applied as described above, then all remaining peaks are replicate-filtered with $r = 2$ (such that peaks must be present in at least two out of three replicate spectra) and also $r = 3$ (peaks must be present in all three spectra). The experiment is repeated over all ten samples. The results for typical SNR threshold and replicate filtering with $r = 2$ and also $r = 3$ are shown in Table 4.2. Since the replicate filter removes a large amount of the noise, it is now feasible to use a low SNR threshold, e.g. with SNR=3.0, 80.9% of the peaks detected by the two-stage filter are real, compared to only 23.8% using a SNR threshold alone, with both methods detecting just over 3400 real peaks. However, there are still many noise peaks also being detected even when a $r = 2$ replicate filter is applied. This is because the low SNR threshold still allows many noise peaks to pass (Table 4.1), and consequently even with replicate filtering many of these remain. While applying a three out of three replicate filter mitigates these effects, many real peaks are also lost, and therefore this is not a good option. The third filtering method aims to address this shortfall.

Three-Stage Noise Filtering

Here, results are presented using the three-stage filter (see Figure 4.1). During experimentation, it was observed that occasionally, individual peak intensities 'drop out' below the noise level, apparently disappearing. Therefore, an $r = 2$ out of 3 replicate filter is selected as optimal, which reduces the loss of real peaks. The two-stage filter is applied first, with a varying SNR threshold and $r = 2$, resulting in ten peak lists. The final stage acts across these samples to retain peaks present in at least $s\%$ of the samples, using $s = 10, 20, \ldots, 100\%$. The experiment uses all of the available simulated spectra $(1a, 1b, 1c, ..., 10a, 10b, 10c)$, where (a, b, c) indicate the replicate spectra, and $(1 \ldots 10)$, the samples. An important assumption here is that the variation observed between replicate samples also captures the variation between samples. While this is not expected to be true in a real experiment since the concentration of metabolites will vary between biological samples, no methods have been reported to model the general variation between samples. Therefore, to maintain generality, the approximation is made that, as is often the case, samples within one biological group will all be similar, e.g. the same phenotype. Additionally, the assumption is that variations in metabolite populations will not be extreme, i.e. all metabolites will be present across all samples, albeit in varying concentrations.

The results of the experiment with selected values of the SNR threshold, r and s, shown in Table 4.3 reveal that no single solution can definitively be described as optimal. In general, the compromise available is between high likelihood that a detected peak is real,

which is achievable with high SNR threshold and/or high r and s; and a high number of correctly detected peaks but also incorrectly identified noise, which is achievable with low SNR threshold and/or low r and s.

It should be noted that approximations and assumptions are applied to the model and therefore, emphasis should be placed on the relative performance of different filter configurations instead of absolute values such as number of peaks detected and likelihood that a detected peak is real. To this extent, several configurations of the filter are compared, which demonstrate the trade-off between these two metrics, as shown in Table 4.3. It is apparent that selecting the optimal configuration may depend upon the data being filtered. If the samples are likely to be very similar, then a higher s is acceptable since it is anticipated that a high number of peaks will be common across samples. In this case, a lower SNR threshold of 2.6–3.0 may be used, yielding approximately 3800–3900 real peaks with a likelihood of 0.940–0.986 that a detected peak is real. If more variability is observed between samples, reducing s to 10% requires the SNR threshold to be raised to 4.0 to yield a comparable number of real peaks (3825) and probability of a detected peak being real (0.951).

To help select the optimal set of parameters for an experiment, all results with $r = 2$ are presented as contour plots, as shown in Figure 4.6. The first plot shows that the number of real peaks detected is optimal with low SNR threshold and low sample filter, s, as represented by the lighter regions. The second plot shows the probability that a detected peak is real, and is optimal with high SNR and high s. Thus, by considering a third metric of the product of these first two metrics, the overall relative performance of the filter can be visualised, as shown in the third plot. Using this plot, optimal parameter sets are identified that encompass the lightest region. These are listed in Table 4.3, and comprise of SNR=2.6, $r = 2$, $s = 40 - 60\%$; SNR=2.8, $r = 2$, $s = 30 - 50\%$; SNR=3.0, $r = 2$, $s = 30 - 50\%$; SNR=3.5, $r = 2$, $s = 20 - 30\%$; and SNR=4.0, $r = 2$, $s = 10\%$. The optimal setting of SNR=5.0 is also shown in Table 4.3 for comparison with one- and two-stage filtering, which are shown in Table 4.1 and Table 4.2, respectively. Three-stage filtering has been successfully used with settings that favour higher probability of a peak being real, i.e. SNR=3.5, $r = 2$ and $s = 50\%$, as recently reported for an FT-ICR MS-based metabolomics study of an aquatic invertebrate [16]. The effect of the filtering stages in this study are reflected by these typical total numbers of positive and negative ion peaks at each stage: 13979 peaks after the SNR threshold filter, 8532 peaks after replicate filtering and 5447 peaks after sample filtering. Of these 5447 peaks, more than one thousand metabolites were putatively identified [16]. Overall, this three-stage filter shows significant improvement in either the number of real peaks detected or the probability that a detected peak is real, or both.

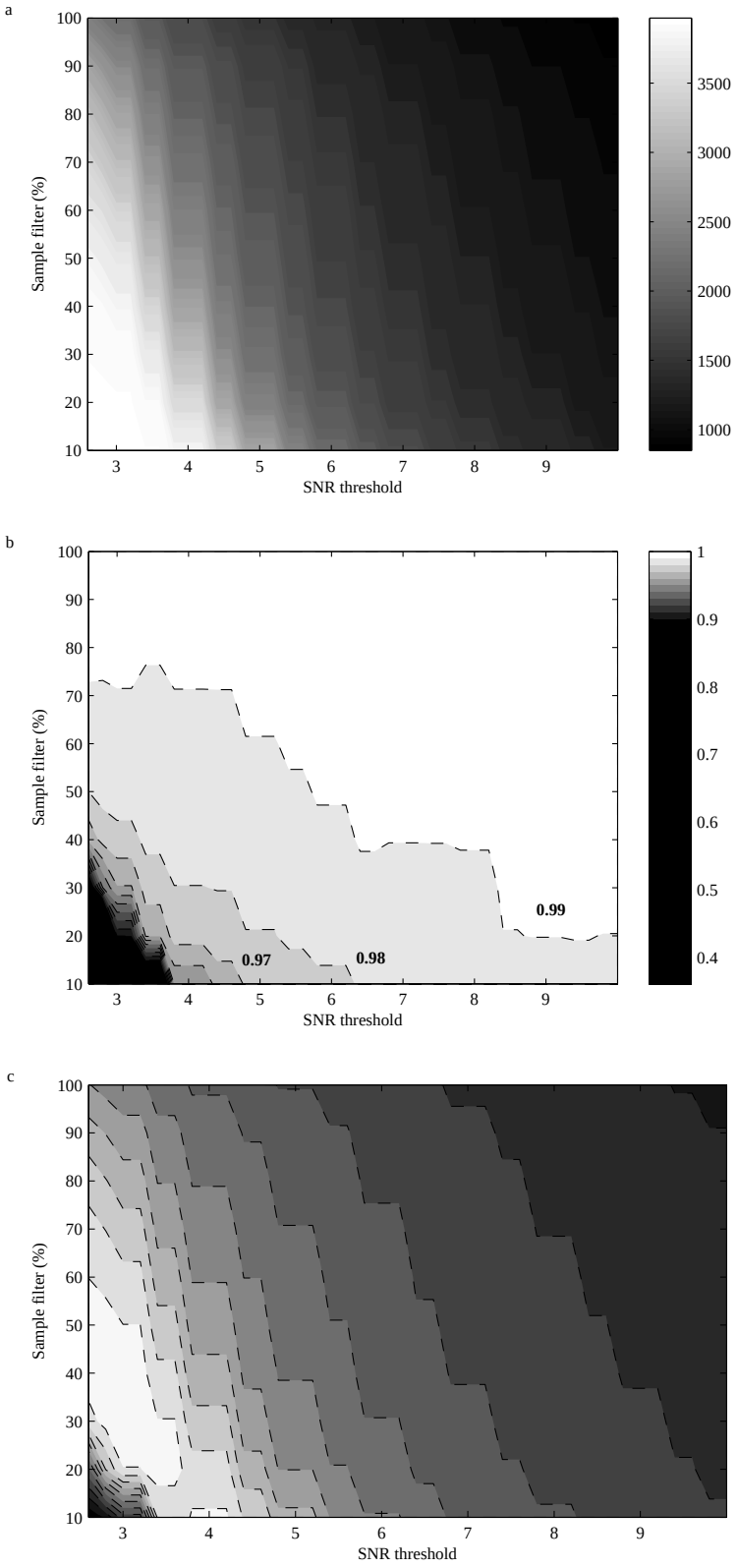

Figure 4.6: Contour plots showing the quality of the simulated spectrum, after three-stage filtering has been applied for varying SNR threshold and sample filter, and fixed replicate filter, and quantified in terms of: (a) number of real peaks correctly identified, (b) probability that a detected peak was real, and (c) the product of (a) and (b).

4.5 Conclusions

The second objective of this work was to minimise the noise present in FT-ICR mass spectra. It has been shown that the number of scans acquired for each spectrum is an important issue to be addressed, and that at least five scans are required for a median RSD of approximately 5%. Having addressed this issue, a three-stage noise filtering method was described, and through the use of a simulated spectrum, with known 'real' and 'noise' peaks, significant benefit was shown to the quality of the data, compared to single- and two-stage filtering methods. The parameters of the filter should be chosen carefully to maximise the metrics of interest, and optimal values for these parameters have been presented. While the results presented here may vary as a result of the features of the particular empirical data, experiments from a different type of biological sample (spectra of human cell line extracts, with over a 2-fold increase in peak density) have yielded similar results. When devising a data filtering strategy in FT-ICR MS experiments, the parameters should be selected using a scheme such as this as a guide. Such an approach, while applied here to DI nESI FT-ICR MS data, is likely to be relevant to any high resolution MS data where there are significant numbers of low intensity peaks that would be lost if a simple hard threshold was applied.

Thee methods developed in this chapter have been integrated into the SIMStitch tool. The components of the tool are shown in Figure 4.7. The SIM-stitch method described in Chapter 3, and the noise filtering methods presented in this chapter, can be readily used by biological scientists. The graphical user interface shown in Figure 4.8 assists with managing the input, intermediate, and output files necessary for an experiment.

4.6 Future Developments

Imposing hard constraints on any classification rarely represents the 'real world'; instead, a level of belief is associated with each classification — in this case 'noise' or 'real' peak. Even when an expert mass spectroscopist is asked to indicate real peaks, they will make a judgment based on how 'likely' they believe it is that a certain peak is real, given the information available. Considering the presence of the peak in other samples has been shown to improve the quality of the judgment. Therefore, a future development of this filtering process would be to quantify the belief in the peak classification so that down-stream operations are provided with a list of peaks, and an associated likelihood that each is real or noise. One approach that would benefit from this is a Bayesian model, as described by Rogers *et al.* [102], where the likelihood that a peak is real could be included as an additional term in the model. Also, this information could be included

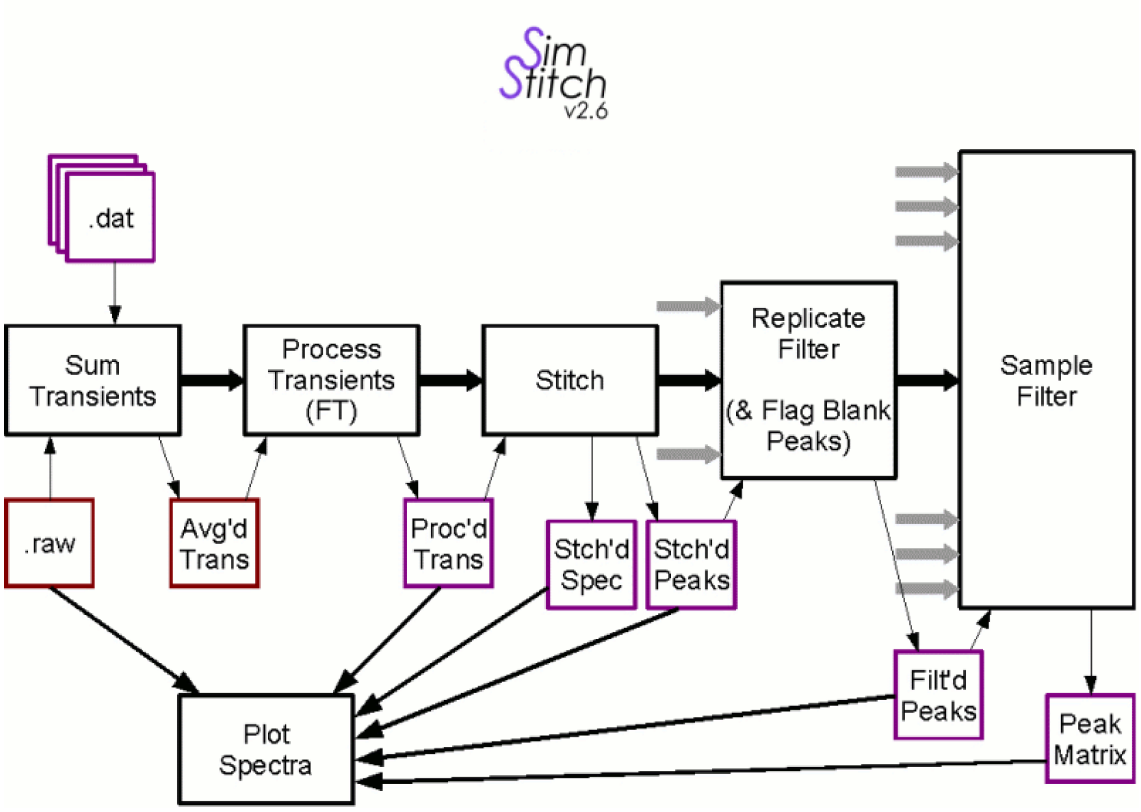

Figure 4.7: The schema of SIM-stitch. The red boxes represent essential files, while the purple boxes represent temporary and output files. The '.dat' files are the transient files generated by the instrument. The '.raw' files are also generated by the instrument and contain essential information regarding the transient files, and instrument configuration. The stages in processing are shown, from summation (averaging) of the transient files, Fourier transformation, SIM-stitching and finally the three-stage filter. A separate module allows the spectral results at most stages of the processing to be viewed.

as an additional continuous metric to the constraints based profiling method, presented in Chapters 5 and 6, to further enhance the quality of the search for sample metabolic content.

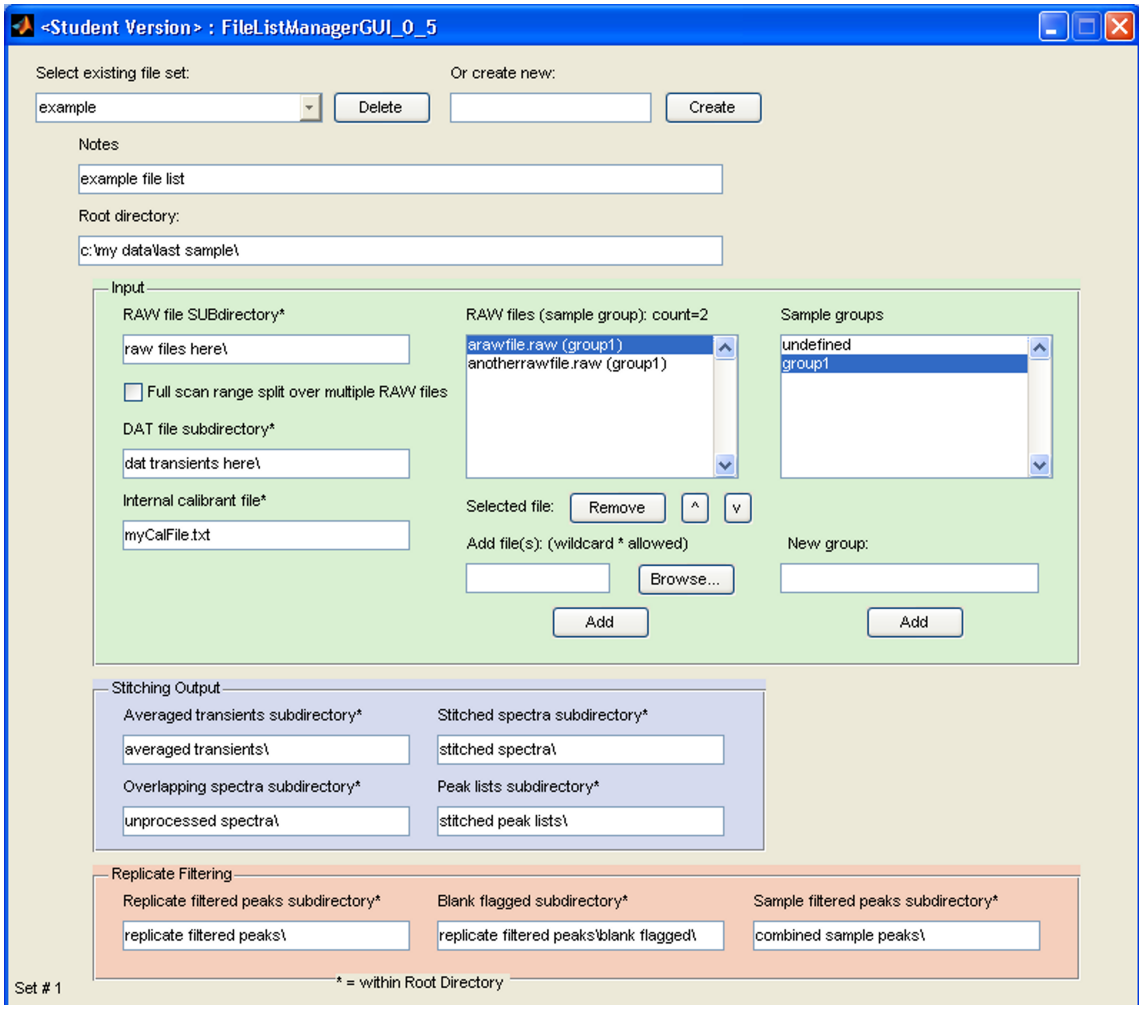

Figure 4.8: The SIM-stitch graphical user interface.

CHAPTER 5

INTERPRETING THE MASS SPECTRUM

5.1 Profiling the Metabolome

Chapters 3 and 4 have focussed on the process of accurately and sensitively extracting the molecular weight and intensity information of ions present in a mass spectrum. As discussed in Chapter 1, the ultimate goal of mass spectrometry is to establish the content of a sample in terms of the compounds present, commonly referred to as *profiling*. Therefore, in this and the subsequent chapter, an approach to making this final step in identifying compounds is presented. This chapter introduces metabolic profiling from mass spectra, in terms of the available information present in the spectrum, how it can be used to identify metabolites, and the issues that arise during the process. The chapter includes a review of current available methods for metabolic profiling, and finally discusses a novel strategy that can improve on existing methods. This new approach is discussed further in Chapter 6.

As has been shown in preceding chapters, FT-ICR MS provides a breakdown of sample-derived ions by their m/z and intensity. Assuming that the ion charge, z, can be identified, the ion mass can be determined. From this mass measurement, it is trivial to establish one or more possible empirical formulae that match the observed mass. For example, an observed mass of 342.29648 Da in a biological sample could be the naturally-occurring molecule with empirical formula $C_{12}H_{22}O_{11}$. However, it quickly becomes clear when one searches for this formula in a compound database, such as PubChem [4], which returns 506 hits, that $C_{12}H_{22}O_{11}$ corresponds to several different compounds including sucrose, lactose and maltose. Called *isomers*, such compounds are commonplace in nature and have the same elemental composition but different structures. Two examples of an isomer are shown in Figure 5.1.

Consequently, an empirical formula alone is typically insufficient to identify compounds

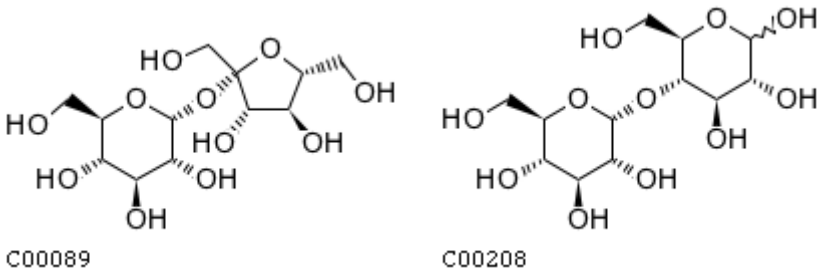

Figure 5.1: Structural diagram of isomers sucrose (ID C00089) and maltose (ID C00208), as present in the KEGG database [3].

and so other 'axes of information', in addition to molecular mass, are required to fully profile the metabolome by FT-ICR MS [23]. The types of information have been classified into four groups, as shown in Figure 5.2. The use of accurate mass, structural information and biological *a priori* are discussed in the subsections below. Two additional analysis tools, van Krevelen diagrams and Kendrick mass defect analysis are also considered.

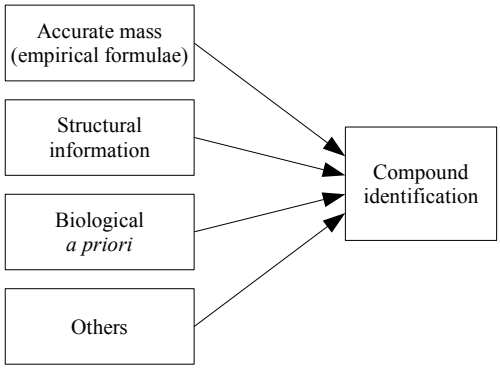

Figure 5.2: Information sources of compound identification using MS.

5.1.1 Accurate Mass

The first axis of information, used in this work, is the m/z and intensity of ions in the spectrum and the subsequent interpretation of this as *empirical formula* and abundance of a compound. For example, given an observed mass m, the relationship to the empirical formula expressed as $C_cH_hN_nO_oP_pS_s$ is shown in equation (5.1).

$$m = 12.000000c + 1.007825h + 14.003074n + 15.994915o + 30.973763p + 31.972072s - m_e \tag{5.1}$$

The coefficient of each variable is the exact mass of each element (in Daltons), and the

subtraction of the electron mass m_e implies a positive-ion mode mass spectrum. Equation (5.1) can be solved as a combinatorial problem, to yield one or more possible empirical formulae.

The mass accuracy and resolution of the DI nESI FT-ICR instrument are very high and because of this, in certain cases, ions at low mass can be measured with such precision that, by application of equation (5.1), a single empirical formula can be assigned. Such an assignment is not conclusive, for example, if there may be additional elements present that are not considered. However, if the elements likely to be present in the sample are known, an empirical formula can be assigned with high confidence.

Beyond ca. 138 Da and even with instrument mass accuracy as low as 0.5 ppm, it is insufficient to assign empirical molecular formulae based solely on accurate mass, since multiple formulae exist with matching mass due to the higher number of possible element permutations [103]. In this case, additional information found within the spectrum should be used. As described in Section 2.3, the ESI process results in ions that typically belong to one of four classes: molecular ions, (de-)protonated ions, adducts or fragments. Each of these forms is a modified molecule, with the original, neutral 'parent' molecule being the unknown compound. Additionally, any ion may be present in naturally-occurring *isotope* forms, which contain varying numbers of neutrons but identical numbers of protons and electrons. Each of these types of ion can be used to help identify a parent molecule: the more ions related to a particular molecule, the stronger the case for that molecule existing in the sample. A formalisation of this is presented by Rogers *et al.* [102]. As discussed, ion fragments are not included in this work due to the use of a 'gentle' nESI ion source. Thus, aside from the ionised form of the metabolite $[M]^+$, the features that are considered to be present in the spectrum are adduct peaks, including for simplicity, protonated and de-protonated forms; and isotope peaks.

Adducts

Adducts represent forms of the molecule, where not only an electron is lost (or gained), but where additional elements have been lost or gained. For example, a common feature is the protonated form of the molecule, i.e. $[M + H]^+$, where a hydrogen atom less its electron (i.e. a proton) have become joined with the metabolite. Other forms include $[M + K]^+$ and $[M + Na]^+$ — where components from the sample matrix react with the metabolites [38]. Together with their isotope forms, these adducts account for over 40% of the spectral peaks [38]. The presence and abundance of each adduct form is, at present, mostly unpredictable when using direct injection MS. This is evident from the wide variety of adduct forms described in published material [89, 104, 38]. Recent work

has begun to identify some trends for certain sample types [105, 106], and, in general, at least the protonated $[M + H]^+$, sodiated $[M + Na]^+$ and potassiated $[M + K]^+$ adduct forms are expected to be present within the mass spectra of natural organic material [16]. Despite this uncertainty, the presence alone of one or more adducts aids molecular formula assignment by allowing patterns to be discovered. For example, if a peak is found corresponding to $[M + H]^+$, and another corresponding to $[M + K]^+$, and a further corresponding to $[M + Na]^+$, this is strong evidence that these are the correct assignments over any other unrelated empirical formulae that could be assigned to these three peaks.

Metal ions arising from salts, such as sodium or potassium, form particularly useful adducts. This is because most metabolites in natural organic matter will not contain these metals bound covalently. Therefore, formulae containing metals can be readily identified as adducts in a spectrum, for example $[M + K]^+$ and $[M + Na]^+$.

Isotopes

The second type of feature found in a mass spectrum is isotopes peaks. These arise as a result of naturally-occurring forms of the elements with different numbers of neutrons [23]. Most elements present in organic matter have one or more isotopes. The expected naturally-occurring abundance of each isotope form is well known. Therefore, for a particular metabolite, one can predict the combinations of element isotopes that will be seen for a certain molecular formula. The set of combinations results in an 'isotope pattern', unique to each molecular formula. As well as the presence of isotope peaks indicating the *presence* (or absence) of a particular element being in the molecular formula, the intensity of the peaks relative to each other reflects the *quantity* of those elements in the formula.

For example, an isotope occurring with relatively high abundance is carbon-13, which has an additional neutron compared to the monoisotopic form carbon-12. As carbon-13 peaks are readily located in mass spectra, many published profiling experiments consider *only* carbon-13 isotopes when searching for isotopes — three examples are readily located [90, 107, 103]. Carbon-12 has a relative natural abundance of 0.989, compared to 0.011 for carbon-13. Therefore, an abundance of $\frac{0.011}{0.989} = 0.0111$ for the carbon-13 peak relative to the carbon-12 peak would indicate the presence of a single carbon atom in the molecule. By extension, the number of carbon atoms can be deduced from peak abundance. However, other isotopes are also present, and while generally at low abundance even when compared to the carbon-13 isotope, they are predicted to be observable in the high dynamic range mass spectra produced using FT-ICR MS, and therefore should be used. For example, as shown in Table 5.1, the compound proline with empirical formula $C_5H_9NO_2$ and nominal mass 116 Da has a total of 8 isotope peaks that are theoretically measurable

in a mass spectrum with dynamic range 1×10^4, assuming the monoisotopic peak is the most abundant, and so has relative abundance of 100%. Generally, only the carbon-13 peaks (rows in Table 5.1 with relative abundance 5.56% and 0.12%) would be considered in a profiling effort, even though several other isotopes may be present, as indicated in the Table 5.1.

Formula	Relative Abundance
$^{1}H_9\ ^{12}C_5\ ^{14}N\ ^{16}O_2$ -e	1.0000
$^{1}H_9\ ^{12}C_5\ \underline{^{15}N}\ ^{16}O_2$ -e	0.0037
$^{1}H_9\ ^{12}C_4\ \underline{^{13}C}\ ^{14}N\ ^{16}O_2$ -e	0.0556
$^{1}H_9\ ^{12}C_5\ ^{14}N\ ^{16}O\ \underline{^{17}O}$ -e	0.0008
$^{1}H_8\ \underline{^{2}H}\ ^{12}C_5\ ^{14}N\ ^{16}O_2$ -e	0.0014
$^{1}H_9\ ^{12}C_4\ \underline{^{13}C}\ \underline{^{15}N}\ ^{16}O_2$ -e	0.0002
$^{1}H_9\ ^{12}C_5\ ^{14}N\ ^{16}O\ \underline{^{18}O}$ -e	0.0040
$^{1}H_9\ ^{12}C_3\ \underline{^{13}C_2}\ ^{14}N\ ^{16}O_2$ -e	0.0012
$^{1}H_9\ ^{12}C_4\ \underline{^{13}C}\ ^{14}N\ ^{16}O\ \underline{^{18}O}$ -e	0.0002

Table 5.1: Proline and all naturally-occurring isotopes with relative abundance over 0.01%. Isotopic elements are underlined.

5.1.2 Structural Information

Information about the structure of the compound allows potential formulae to be filtered to those with a matching structure. Structural information can be acquired by several means, the most common of which are chromatography and fragmentation [23].

Chromatography involves coupling a column to the mass spectrometer, through which the sample passes before being ionised and detected. The column contains either a liquid solvent for liquid chromatography (LC), or a gas for gas chromatography (GC), and separates the compounds along the additional axis of *elution time*, which is the time taken for the compound to pass ('elute') through the column. The recorded elution time depends upon their structural characteristics [26] and consequently compounds will appear in spectra within different time frames. This additional information about each compound can be used to mediate between candidate compounds where m/z alone is insufficient.

In fragmentation, ions are forcibly separated into their constituent parts, thus providing evidence about the structure of the parent, or 'precursor' ion [1], from which the fragment came. Fragmentation can occur in a single stage or in multiple stages, referred to as MS/MS and MS^n, respectively, and is achieved using a variety of methods including collision-induced dissociation (CID) in which the ions are collided with neutral molecules.

In this work, neither chromatography nor fragmentation are used, because both methods increase acquisition time significantly and are consequently unsuitable when using DI-ESI FT-ICR, in order to meet the high throughput requirements of this study outlined in Section 1.3. While some fragmentation of fragile molecules can naturally occur during the ionisation process [38, 35], the nESI source used in this work is sufficiently 'soft' that very little fragmentation occurs [23].

5.1.3 Biological Information

A further way in which compounds can be more certainly identified using DI FT-ICR MS is by combining mass measurements with biological information. This includes information that describes how compounds interact within biological organisms. It has been shown that by using such knowledge, certain compounds can be anticipated to be present in the sample, based upon the presence of other compounds [108]. Such knowledge sources are typically compound databases such as KEGG [3] and the PubChem database [4]. Currently, such databases are relatively general, containing many compounds that would not be expected to be present in all species, however as they become more complete and specific, more accurate interpretation of mass spectra can be afforded. As well as being general, such databases also suffer from containing errors and being incomplete, with the result that reliable identification using such sources becomes difficult. This issue is compounded by the presence of biological and non-biological contaminants in the sample that would have to be considered separately, since they would be present in such databases.

5.1.4 van Krevelen Diagrams and Kendrick Mass Defect Analysis

van Krevelen diagrams are a means of classifying empirical molecular formulae by their elemental composition [109]. In particular, by plotting the oxygen/carbon element count ratio against hydrogen/carbon element count ratio, the compound class of a molecular formula of interest can be estimated based upon its spatial position. Such a diagram is useful as a visual aid when describing the composition of a biological sample, however there is little evidence to suggest that *all* compounds with the prescribed area belong to the compound groups, or to what class those outside any group belong. Furthermore, the boundaries of the classes are somewhat inconsistent [110, 109]. Therefore, it is difficult to use van Krevelen diagrams as a means of quantifying the relative likelihood of multiple possible molecular formulae, and so they are not pursued as a means of identifying molecular formulae in this study.

Kendrick mass defect (KMD) analysis uses the offset of exact molecular mass from nominal mass to identify ions belonging to a homologous group. Ions belonging to the same group,

for example differing by CH_2, have the same Kendrick mass defect, which is calculated as shown in equation (5.2), in which m is the peak mass in Daltons [109].

$$\text{KMD} = \text{observed nominal mass} - 14/14.01565m \tag{5.2}$$

Thus, the KMD can be used to arbitrate between two possible molecular formulae for a peak, since the one that belongs to a group present in the sample can be preferentially chosen. Such an approach is particularly useful where large, homologous groups of compounds can be expected to exist, for example in crude oil [111]. This study is focussed on general natural organic matter for which Kendrick mass defect analysis is of limited use [34].

5.2 Search Strategy

Presented so far in this chapter, is the information that is available when searching, within a mass spectrum, for the metabolic content of the sample. The way the information is used and how the solution is found are key to obtaining a good answer to the problem in terms of the sample content. In this section, current strategies in solving the problem are explored, and key shortfalls that should be addressed are identified. The following section presents a new approach with the aim of tackling these shortfalls.

5.2.1 Current Solutions

One technique that captures the philosophical approach adopted by most laboratories was recently published by Fiehn and Kind in 2007 [104]. Labelled by the authors as the 'golden rules', they describe how seven rules and checks are used step by step to generate, prune and rank potential molecular formulae corresponding to peaks in a mass spectrum. Peaks are analysed individually, and each peak is presented in a tabular format. Within the table, each row represents a unique empirical formula generated using the constraints on formulae to limit the starting list, and the columns contain the results of the rules, i.e. pass or fail. Each rule effectively applies a constraint to the solution space in order to derive a smaller set of solutions. The remaining solutions are ranked according to a score indicating the match between observed and calculated isotope patterns. The final stage is to search for all alternatives in the PubChem database and to select the highest ranking of these as the solution. Fiehn and Kind demonstrate the approach on an FT-ICR mass spectrometer using two substances with nominal mass 765 Da and 854 Da. In both cases,

the algorithm resulted in several alternative compounds, from which the correct one was selected after looking-up each empirical formula in compound databases.

Such methods are suitable for simple spectra, where there are relatively few peaks that can be analysed individually. However, as spectra are acquired at increasing mass resolution, and many thousands of peaks are detected, this process becomes infeasible. Furthermore, as more peaks are detected, the number of peaks that have conflicting assignments increases. For example, a peak at a particular location could represent the isotope of one compound, or the adduct of a different compound, or the fragment of yet another compound. The methods described above do not allow the 'best' assignment to be made. Rather, the peak will be assigned to the first compound that is considered, and the peak will then no longer be considered as a potential peak for any other compound. It seems pertinent to determine which compound the peak is *most likely* to belong to.

A second but equally important shortfall seen in such search implementations to date, is that with the exception of the isotopic pattern score, the rules themselves are all 'hard', in the sense that they define rigid boundaries for particular characteristics. For example, defining an acceptable range for the ratio of hydrogen to carbon atoms in a molecular formula is a good approach where only the most common compounds are expected to be present in the sample, but fails to capture outliers. For example, the ratio of hydrogen to carbon elements in a formula ('H/C ratio') is one indicator that the formula represents a real compound [104]. Fiehn and Kind use the H/C ratio as one indicator to decide if a potential formulae is 'real' or not, i.e. whether or not it exists in a compound database. In order to do this, they count the number of compounds in the database as a function of the H/C ratio. They then define boundaries of 0.8–2.8 for the ratio, within which a molecular formula is classed as having an acceptable H/C ratio, and beyond which the formula is rejected.

However, the H/C ratio of known compounds is not discrete, for example many more exist with ratio 0.8–1.0 than 2.6–2.8 [104]. In Fiehn and Kind's approach, a molecular formula with H/C ratio of 0.9 is as likely to represent a compound in the sample as one with H/C ratio of 2.7, all else being equal, and so although the first molecular formula is more likely to be a real compound than the second, this is not considered when arriving at the solution. This is a shortfall of their approach, which is similar to other element count rules frequently found in profiling literature [110, 112, 113].

Fienn and Kind have shown that 0.6% of a test set of formulae taken from the PubChem database did not pass all the tests, and therefore valid compounds would have been excluded from the solution space. Considering that PubChem confirms that there are at least several million unique compounds, this equates to a large number of compounds missed

from the solution space. The boundaries of the tests could be altered, however as they are extended they become less effective at reducing the search space and useless if 100% of compounds are to be included, at which point the ranking by isotope pattern is the sole filtering mechanism.

Particularly key is the *exact mass*, which is often filtered according to a hard threshold that defines an allowable range of values. For example, a metabolite is typically considered as 'allowed' if the m/z distance is in the order of < 1 ppm from where the true mass is expected to lie [113]. The measured error is dependent upon the instrument and acquisition parameters, and since FT-ICR MS is effectively providing a continuous measurement of mass, the error measured is also continuous in nature. Applying a hard threshold does not capture this information, resulting in the loss of potentially useful information. For example, a peak that is present with 1.1 ppm m/z error will be discarded from the set of possible solutions, even though it is almost as likely as a peak with 1.0 ppm m/z error. Consider also that a peak assignment with a m/z error of 0.8 ppm is less likely to be correct than a peak assignment with 0.4 ppm error, due to the normally-distributed nature of the mass measurement error [18]. Hence there is also a loss of fidelity information that would indicate how much confidence could be placed in an assignment. Therefore, such a hard threshold should not be used unless it can be relaxed to include the vast majority of expected errors, while allowing the 'nearness' of the measurement to ideal to be captured, such that alternative solutions to the measurement can be ranked. Since the information contained in the observation is useful and important, it should be included in the metabolite assignment decision process.

The example above is one example of additional information that could be used in the identification of metabolites from mass spectra, and this idea is developed in the rest of this chapter.

5.2.2 Model-Based Approaches

Approaches based on models use knowledge about the mass spectrum, to determine the 'best' solution of the identification of peaks. The benefits of such approaches over those described in the previous sub-section are that the knowledge about the spectrum can be more systematically and rigorously combined. Ultimately, any models used should allow *all* available knowledge about the spectrum and measurement system to be included and represented *completely*, for example mass error should be represented as a continuous parameter.

The following are four of the main criteria that should be included in a model:

- The mass measurement error;
- The isotope pattern accuracy;
- The presence of adduct peaks;
- The likelihood that the molecular formula exists in any biological sample.

Use of *mass measurement error* should consider the continuous nature of the measurement when determining the 'quality' of the measurement. Instead of the current approach of considering hard boundaries, as described in Section 5.2.1, a wider window should be allowed to exist, and the proximity of the measurement to 'ideal' within that window used as an indication of the quality of the match between molecular formula and measured m/z.

The *isotope pattern accuracy* should continue to be used in a way that assigns a value to the match between theoretical isotope pattern and the pattern observed [104]. The number of isotopes considered in the pattern should be extended beyond low numbers of carbon-13 isotopes to include all isotopic elements that could potentially be observed in the mass spectrum. Given the high dynamic range of ESI FT-ICR MS (see Section 3.1), there should exist many isotope peaks, as discussed in Section 5.1.1, which should be used wherever possible.

As many *adduct peaks* as possible should be considered in the search for the metabolome, and more importantly, the number of adducts present should be used as a criteria to establish the quality of the molecular formula assignment. As described in Section 5.1.1, the more adducts present, the higher the chance of a correct molecular formula assignment. Where possible, evidence to suggest the likelihood of certain adduct patterns being present should be taken into account when calculating the overall figure of merit.

The *likelihood of the molecular formula existing in the sample* should be calculated in a manner similar to that adopted by Fiehn and Kind, and discussed in Section 5.2.1, but with the key difference that instead of using a set of rules to decide whether or not a molecular formula exists, certain characteristics of the formula are used to quantify the likelihood that the formula exists. As has already been noted, a small portion of molecular formulae existing in compound databases do not meet Fiehn and Kind's rules, and relaxing the rules would lead to a vast increase in the number of potential formulae, including many that would not exist naturally.

These four criteria should be combined in a manner that reflects their value to the mass spectroscopist as indicators of a compound existing in a sample. Given the incomplete and noisy information available from mass spectrometry, as discussed in Chapter 2, the aim is to find the solution that *best fits* the observations. It should also be possible to include

additional information in the search as it becomes available. For example, should the behaviour of adducts become predictable, for example by means of a continuous probability distribution dependent upon parent formulae, the system should permit such information to be represented accurately. The remainder of this section describes an existing Bayesian approach, and a new constraints-based approach, to solving this problem.

5.2.3 Probabilistic Methods

One model-based approach to solving the problem of extracting molecular formulae from mass spectra is the use of Bayesian probability to identify the optimal solution, as reported by Rogers *et al.* [102]. In their work, Rogers *et al.* use statistical models to determine the probability of a peak being assigned to a compound, given the observed peaks. The model includes the mass measurement error and a set of potential biochemical transformations that suggests related compounds. A Gibbs sampling scheme is used to determine the posterior probability of the assignment of peaks to formulae. The model is also extended to include isotope information, and allows additional information to be included in the form of additional terms in the model.

An advantage of the probabilistic method is the probabilities of the assignments quantify the probability that the assignments are correct. This allows *relative* comparisons to be made between assignments and possible alternatives in a way that is easily understood. A further benefit of using Bayesian probability as described by Rogers *et al.* is the efficient estimation of posterior probabilities using a Gibbs sampling scheme. In their work, they calculate the posterior probabilities of 379 potential formulae being assigned to 446 peaks in approximately four minutes.

A disadvantage of this approach is that the information that is included in the model must be described as a probability distribution. It may not always be possible to quantify accurately and precisely the information, for example if there is insufficient experimental data to calculate a distribution. Furthermore, identified compounds must exist within a compound database, limiting the opportunity of discovering new compounds.

5.2.4 Constraints Optimisation Methods

A novel approach to the assignment of molecular formulae to peaks is by modelling the problem as a constraints optimisation problem (COP). The model describes the information that can be found in the spectrum, such as isotopes, as a set of constraints. These constraints are 'soft' since instead of defining a criteria that *must* be met, they define a value that is a function of the degree to which the constraint is met. For example, a higher

value would be placed on an assignment where an isotope peak was also observed, than if no isotope was observed. By combining the soft constraints into a single value, the goal of the COP search is to maximise this overall value.

This method is similar to the probabilistic method described in Section 5.2.3, in that the each molecular formula assignment in the solution has an associated value quantifying the degree of confidence in that assignment. Also, both methods use statistical uncertainty where it is available, for example in measured mass error.

The constraints and probabilistic methods differ in that the COP method does not require information to be described as a statistical distribution. This is due to the soft constraints, which can contain any function to place value on an assignment, and may be based on qualitative descriptions as will be shown in the following chapter. The benefit of such flexibility is that the qualitative experience of the expert mass spectroscopist can be much more easily captured where statistical data is insufficient. For example, there is little information regarding the occurrence and abundance of adducts that appear in spectra, however experiments may demonstrate an increased abundance of some adduct types over others, and this can be incorporated in a constraints model.

This method therefore requires a value to be placed on information, and more accurate results will be obtained when these values more closely represent the real world. It may be undesirable to place value on information for which good statistical models are available. For example, the classification of peaks as real or noise, as described in Section 4.6, could be developed to yield a belief value for a peak being real. Assuming that the probability distribution of the classification is known, it might be preferable to use Bayesian methods [102] when including this information in the model.

A difficulty with the constraints approach is the efficient implementation of the search. Whereas Bayesian models can be sampled to estimate the solution, as shown by Rogers *et al.* [102], no such method is obvious for the constraints search.

5.3 Challenges

The constraints solution strategy outlined above presents several challenges that will be addressed. The first is that a continuous search space contains infinite solutions. To overcome this issue, it will be shown in Chapter 6 that the search space can be segmented by discretising the individual variables. This improves the feasibility of the search while still avoiding the problem of hard thresholds imposed using current solutions. The quantity and location of values in each variable domain should be selected carefully so as to maximise capture of the continuous nature of the variables.

The second challenge is how to assign relative weighting to the variables so that they can be combined to find the optimal solution. Again, in Chapter 6 it will be shown how Utility Theory [114] can be used to achieve this.

The third major challenge is how to efficiently search for the solution given the vast search space that will be created - even with discrete variables. However, in this work the emphasis is on demonstrating the validity of the approach, rather than dealing with practical issues, which can be solved in time.

5.4 Conclusions

As outlined in Chapter 1, 'metabolic profiling' implies interest in all metabolites within a certain chemical group related to the sample preparation techniques used. Of interest are both the *identity* and *concentration* of the metabolic composition in a sample, and the biologist is likely to be interested in both. While signal intensity does not directly represent metabolite concentration, and is a largely unsolved problem, the focus will be on the identification of metabolites, since this is by far the biggest challenge, not least because many metabolites are as yet unknown. The problem is made more complex by the necessary presence in the spectrum of adducts $[M+A]^{\pm}$, where M is the neutral molecule and A represents one or more elements that become attached to the neutral molecule during ionisation. Without such element exchanges, many neutral molecules would not be observed as ions. Furthermore, molecules may fragment, both in solution and in the gas-phase. As discussed in Section 5.1.1, naturally occurring molecular isotopes are present, where one or more of the constituent atoms are present with an unusual number of neutrons. Together with contamination within the sample from the atmosphere and other noise sources, these issues create a highly complex spectrum that it is not easy to interpret. As has been shown, solutions to identifying metabolites from a complex spectrum are evolving rapidly, but one common drawback is the limited use of the data available. This limitation is imposed by the use of hard thresholds when filtering solutions. The opportunity in this challenge is therefore to make use of *all* available data when making decisions about the sample composition, specifically by capturing the continuous nature of features. Chapter 6 describes a computational approach, using constraint satisfaction methods, to achieve better metabolic profiling.

CHAPTER 6

PROFILING THE METABOLOME AS A CONSTRAINTS OPTIMISATION PROBLEM

6.1 Introduction

Chapter 5 describes how profiling the metabolome identifies and quantifies the metabolites in a biological sample, from the mass spectrum. As described in Section 5.2.1, current methods search for certain patterns within the mass spectrum, and apply a series of tests to filter out unlikely solutions. Section 5.2.4 describes a new approach that, by calculating an overall figure of merit for solutions, allows the optimal solution to be found. This approach encompasses more outliers in the set of possible solutions, and includes multiple criteria in the determination of the best solution. Several challenges are described in Section 5.3 that arise from this method.

The method adopted here to profile the metabolome is through the use of constraints satisfaction methods. A constraints satisfaction problem (CSP) is one that is described in terms of the goal and the rules of the problem, rather than specifying the method that should be used to find the solution [114]. Therefore, the focus is on an accurate and complete description of the problem, rather than getting overly involved in mechanisms to solve the problem. Providing the CSP is accurately and well described, existing search methods can very efficiently find the solution, given only minimal essential problem specification [114]. An example is the n-queens chess problem, in which n queen pieces must be placed on a chessboard of size n by n so that none of the pieces are under attack [115]. The problem is described as a CSP by n, the attack pattern of the queen and the criterion that no queen must be under attack. In order to find a solution, standard search methods can be applied to the CSP. Using constraints methods, both discrete and non-discrete constraints can be included [115], and by using Utility Theory [114], a meaningful, systematic and qualitative weighting can be applied to the find the optimal solution. Additionally, a scalable method

will be demonstrated that can accommodate future constraints as the knowledge becomes available.

6.1.1 Constraints Satisfaction

A constraints satisfaction problem S, can be formally represented as [115]:

$$S = \langle X, D, C \rangle \,, \tag{6.1}$$

where X is a set of variables that describe the problem, D is the domain of each variable in X and C is a set of constraints that apply to the variables X and describe acceptable assignments to each variable. The solutions of S are the solutions to the problem. Solving a CSP is not dissimilar to solving any search problem, where the assignment of one or more variables is sought, given a domain and some conditions that limit the values that can be assigned to each variable.

The significant value added by CSP solvers is in limiting the search space through the use of *constraint propagation* [114], which aims to ensure that, at every stage of the search, the remaining un-searched solutions are viable. This is achieved by applying inference [114] to restrict the domain of variables, by using the constraints and the domain of variables to limit the domain of other variables in the problem. For example, suppose variable x has domain $d_x = \{1, 2, 3, 4, 5\}$; variable y as domain $d_y = \{3, 4, 5, 6\}$; and a system constraint c is $x > y$. Before finding the solution to the problem, constraint propagation enables the domain of x and y to be reduced by inferring the additional constraint $x > 3$. This removes the impossible solutions to x, and the domain of x become $d_x = \{4, 5\}$. Constraint propagation is key to solving a CSP efficiently, since there are typically many variables with large domains and consequently a vast space of potential solutions.

6.1.2 Constraints Optimisation

A constraints optimisation problem (COP) extends the concept of a CSP by finding the optimal solution, as described by some optimisation function f_{opt} of the variables, and is defined as [115]:

$$S_o = \langle X, D, C, f_{opt} \rangle \,. \tag{6.2}$$

The optimisation function maps each solution (in terms of the variable assignments) to a numerical value [115]. In the example in Section 6.1.1 above, there are eight solutions for

(x, y), namely $4, 3$; $(4, 4)$; $(4, 5)$; $(4, 6)$; $(5, 3)$; $(5, 4)$; $(5, 5)$; and $(5, 6)$. If, for example, some of these solutions are more preferable than others, the problem can be extended to a COP by adding an *optimisation function*, for example $f_{\text{opt}} = x + y$. The optimal solution is defined as the one that minimises (or maximises) f_{opt}. Calculating f_{opt} for each solution yields the optimal, single solution that minimises f_{opt}, $x = 4$ and $y = 3$.

It will be shown how constraints optimisation can be applied, together with a suitable optimisation function that places suitable value on a metabolic profiling solution, to a mass spectrum to identify and quantify the metabolites within the mass spectrum.

6.2 From Sample to Metabolite ID: a Model

In order to describe the problem of profiling the metabolome from the mass spectrum as a COP, the first stage is to abstract the main stages in measurement of the molecules. This will allow the variables, domains and constraints to be distilled to form a suitable constraints optimisation problem.

The stages in measurement, from molecules in the sample through to a list of metabolites, as described in Chapters 2 and 5, is represented as a flow of data, as shown in Figure 6.1, and in which the observed peaks are derived from a series of processes that are applied to the sample metabolites.

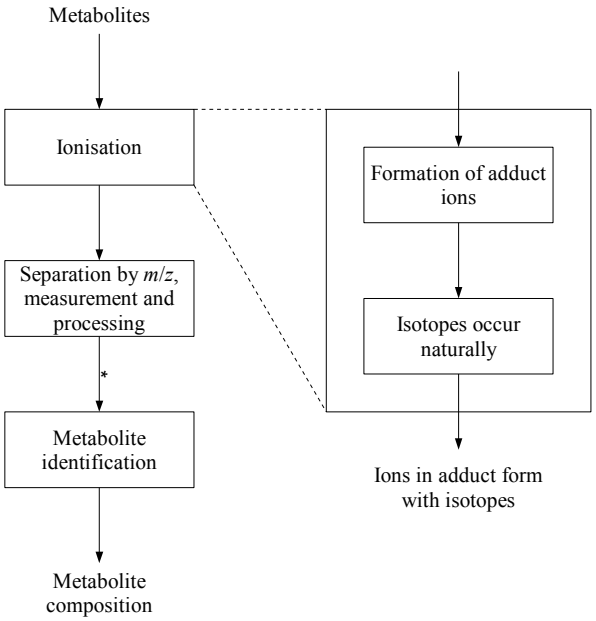

Figure 6.1: Representation of the system as a data flow diagram. The asterisk (*) indicates the stage at which the peak list is generated.

By considering each process as a transformation of the input, the goal is to estimate each

transformation together with the unknown input to the system, i.e. the metabolites. The measurements that are available are the peaks from the mass spectrometry, representing the output of the system. If the system input and transformations can be determined, such that the system output matches the observed output, then the complete system is a good match to the observations. The individual stages captured by the system are:

1. Neutral molecules, including metabolites, exist in the sample. The sample is prepared in a form suitable for mass spectrometry [18].

2. The ionisation process creates charged ions consisting of neutral molecules combined with an adduct, or neutral molecules with a proton loss or gain (see Section 5.1.1). The naturally-occurring isotope forms of elements result in multiple ions being present, representing the different isotope forms of the ion (see Section 5.1.1). In Figure 6.1, the ionisation stage is represented as two separate processes that create adduct ions followed by isotopes; as described in Chapter 5, this is a simplification, since isotopes are present from the start, but the end result is the same.

3. Ions are separated by m/z, and are detected by the instrument to produce a mass spectrum, as described in Section 2.6, which is then processed and filtered to produce a peak list as described in Chapters 2 to 4.

4. Metabolites are identified, using the peak list, to produce the metabolite composition of the sample. This is the stage required to be solved.

The first transformation shown in Figure 6.1 is the formation of adduct ions during the ionisation stage, as described in Section 5.1.1. Adduct ions form either from the combination of molecules in the sample with adducts, which are subsequently ionised during ionisation, or they are formed during the ionisation process itself. Beginning at the input to the system, let the set of molecular formulae in the sample (including metabolites) be defined as $\mathbf{mf} = \{mf_1, mf_2, \ldots\}$. Also consider a given set of potential adducts, $\mathbf{adduct} = \{adduct_1, adduct_2, \ldots\}$. Then it follows that the set of adducted ions that *potentially* are present, **adion**, could be expressed as

$$\mathbf{adion} = \begin{bmatrix} mf_1 + adduct_1 \\ mf_1 + adduct_2 \\ \vdots \\ mf_2 + adduct_1 \\ mf_2 + adduct_2 \\ \vdots \end{bmatrix},$$

where $\mathbf{adion} = \{adion_1, adion_2, \ldots\}$.

Secondly, the isotope forms of each adducted ion are present, as described in Section 5.1.1. This gives rise to isotopic, adducted, ion forms *isoion*, expressed as

$$\mathbf{isoion} = \begin{bmatrix} \text{isotope\_pattern}(adion_1) \\ \text{isotope\_pattern}(adion_2) \\ \vdots \end{bmatrix},$$

where $\mathbf{isoion} = \{isoion_1, isoion_2, \ldots\}$, and in which the function 'isotope_pattern' returns the one or more isotope peaks present for a given monoisotopic molecular formula. Since the presence of these isotopic ions is being measured using mass spectrometry, as described in Section 5.1.1, one expects to observe a series of peaks, **mz**, that may be expressed as

$$\mathbf{mz} = \text{mass}(\mathbf{isoion}),$$

where $\mathbf{mz} = \{mz_1, mz_2, \ldots\}$, and the function 'mass' returns the mass of each isotopic molecular formula in **isoion**.

The relationships between **mf**, **adion**, **isoion** and **mz** can be represented graphically, as shown in Figure 6.2. The nodes, from top row to bottom row, represent the sets of potential sample molecular formulae (mf), adducted ions ($adion$) and isotopic forms of adducted ions ($isoion$) together with the observed peak $m/z(mz)$. The vertices indicate a transformation, from one form to another. The graph describes the complete set of transformations that may occur, for instance in the example in Figure 6.2, mf_1 may be expressed as up to six peaks.

6.3 The Constraints Satisfaction Problem

6.3.1 Viewpoint and Variables

There may be multiple valid perspectives from which a problem can be viewed, and these are known as *viewpoints* [115]. The viewpoint adopted will result in a specific set of variables, domains and constraints that describe all valid solutions to the problem; different viewpoints describe the same problem and so yield the same solution, but use differing variables and therefore domains and constraints [115]. Some viewpoints can be expressed and solved more efficiently, depending on the objects and functions available in the solver, and so a viewpoint is selected with the aim of finding the optimal solution to the COP as quickly as possible.

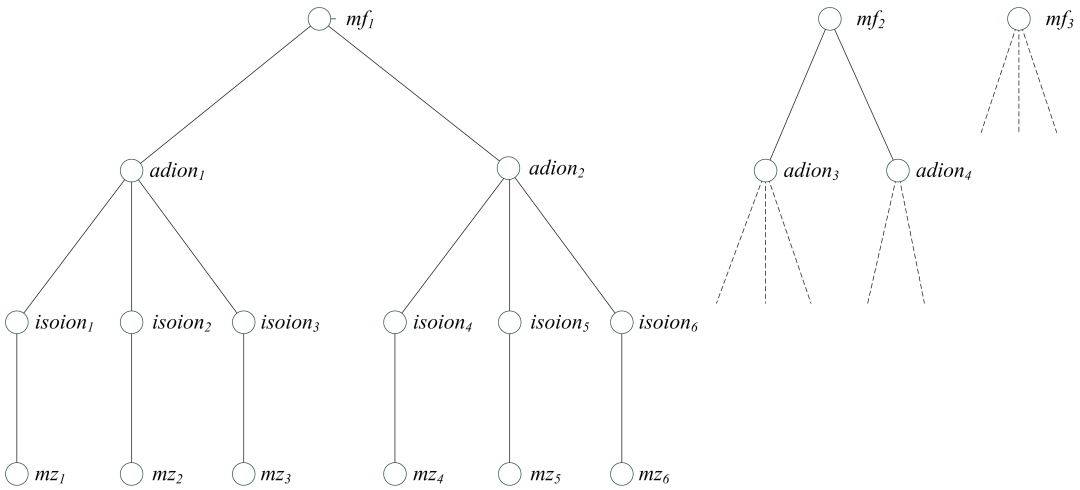

Figure 6.2: Representation of mass spectrometry measurement as a series of transformations between components: molecular formulae, mf; monoisotopic ions plus adducts, $adion$; isotopic ions plus adducts, $isoion$; and mass spectral peaks, mz.

The viewpoint selected here is to implement the system structure as described in Section 6.2, and use the binary matrix variables **A** and **F** to indicate the presence (binary '1') or absence (binary '0') of each transformation. **A** represents the transformations from **mf** to **adion**, and **F** the transformations from **adion** to **isoion**. Hence, **A** and **F** are the same size as **adion** and **isoion**, respectively. If **A** and **F** are determined, the complete solution to the COP is known; no other variables are required. This can be understood by observing that, given **A** and **F**, each mz can be related to one mf, and from this, the quantity of each mf in the solution can be determined. The allowable assignments for **A** and **F** that represent valid solutions will be implemented by a set of hard constraints as described in Section 6.1.1. The optimisation function will be specified in terms of the system variables, as described in Section 6.4.1.

The solution will indicate which molecular formulae mf satisfy the hard constraints of the system and also result in the highest optimal value for the optimisation function. The abundance for each mf can be calculated from the system as the sum of the set of mz intensities that contribute to the mf, and depends upon **A** and **F**.

6.3.2 Constraints

As described above, the system connectivity will be represented as hard constraints that must be satisfied by any solution. In addition, other constraints should be included.

The first hard constraint imposed on the system is that there must exist exactly one isotopic ion per peak — in other words, peaks are due to a single ion formula only.

Thus, the assumption being made is that instrument resolution is sufficient to resolve ions with different molecular formulae. This is a reasonable approximation, since the instrument has extremely high resolution, as described in Chapter 2. In theory, this constraint could be relaxed, or removed, to represent the finite resolution of the instrument, and so consideration is lent to the possibility that peaks are overlapping and are not fully resolved in the spectrum. This would, in the current implementation, have an excessively negative impact on the time needed to find a solution. This first hard constraint can be expressed, for each mz, as

$$\sum_i f_i = 1,$$

where i are the indices of F that are linked to each mz, and f_i is the i^{th} element of $\mathbf{F}$.

The second hard constraint is that each monoisotopic ion plus adduct, *adion*, present in the solution must be linked to exactly one neutral molecular formula, mf. This means that detected ions are the adducted form of a single molecular formula only. This assumption reduces the search space considerably by requiring that each mz be the product of a single mf only, meaning that molecular adducts are distinct. For example, a peak cannot represent the identical molecular formulae $[M_1+H]^+$ and $[M_2]^+$ simultaneously; it is either from M_1 or M_2, but not both. Again, this constraint could be relaxed at the expense of search time. The constraint can be expressed, for each *adion*, as

$$\sum_i a_i = 1,$$

where i are the indices of $\mathbf{A}$ that are linked to each valid ion plus adduct, *adion*, and a_i is the i^{th} element of $\mathbf{A}$.

6.3.3 Solution Search

Before the system can be solved, the graph that describes all possible scenarios is created, including all nodes and vertices. Firstly, the set of isotopic ions, *isoion*, is generated for each observed peak. This is achieved by describing a CSP in which the variable is the count of each element and the constraint is that the total mass is within a 0.75 ppm window of the peak m/z. This window should capture the maximum expected error of the instrument, however a conservative approximation is necessary here to achieve feasible search execution times. Within this CSP, an additional hard constraint is the element base set from which the molecular formulae are constructed: here, this is limited to the elements C, H, N, O, P, S, Na and K, these being the elements typically present in natural organic matter and the salts that are added to the sample in order to encourage ionisation [16].

Next, the monoisotopic ions plus adducts, *adion*, are generated by simplifying the isotopic ion into its monoisotopic form. Finally, the potential adducts are removed, where possible, from each *adion* to yield the set of neutral molecular formulae, mf.

An example is shown in Figure 6.3, in which a peak at 230.0892 m/z is observed in the mass spectrum, resulting from the presence of sodiated tryptophan. Tryptophan, with molecular formula $H_{12}C_{11}N_2O_2$, is shown as the highest node in the graph. In the first stage, the molecule is shown to combine with a sodium atom to form the sodiated form of the compound, variable $adion_1$. One of the many isotopic forms of this sodiated molecule, with three carbon-13 atoms, is shown as variable $isoion_1$, from which the mass 230.0892 m/z is derived, as shown for variable mz_1.

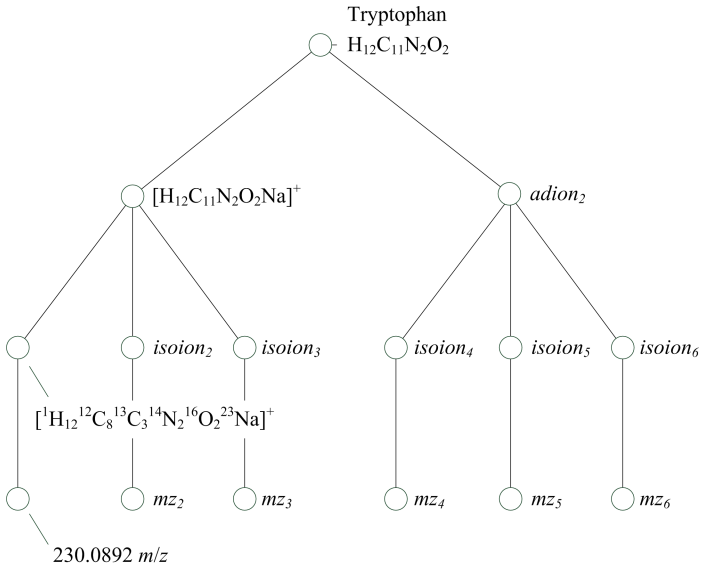

Figure 6.3: Example of representing the presence of tryptophan in a compound via a sodium adduct transformation, showing in full the presence of the [M+3] carbon-13 isotope and resulting in a peak occurring at 230.0892 m/z.

Having established the graph with all potential transformations, the objective is to solve the system in terms of **F** and **A**, and hence determine which **mf** were present in the sample, given the observations **mz**, and certain *a priori*.

The assumption when solving the system is that the system is correctly described both in terms of the variables and the relationships between those variables. Once **F** and **A** are found, the solution is indicated by the mf that are active, i.e. those that are connected in the system, via transformations, to an observed mz.

The form of the graph in Figure 6.3 can be inverted to obtain a more natural visualisation of the system, where the 'input' is **mz** at the top, and the output is **mf** at the bottom.

This is shown for the example in Figure 6.4, in which a mass spectrum of the unknown sample has been obtained where only $mz_{1...4}$ are present.

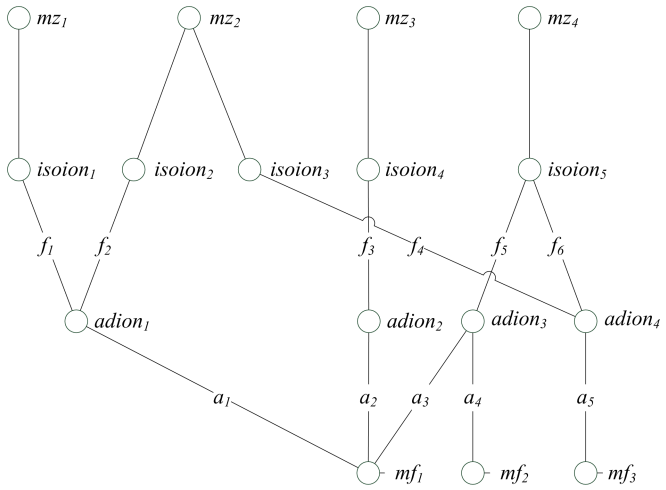

Figure 6.4: An example system graph, showing four observed peaks, and three potential molecular formulae.

From inspection of the graph in Figure 6.4, it would be a good hypothesis that the sample contains mf_1, but not mf_2 or mf_3, since mf_1 is linked to each mz, meaning that mz_1 and mz_2 are isotopes of one adduct form, and mz_3 and mz_4 are two additional adduct forms of the molecular formula mf_1. This is strongly suggestive of the presence of mf_1 as a compound in the sample. If one wished to solve this computationally, the problem could be expressed as a CSP, as described above. In this case, with reference to Figure 6.4, the constraints to ensure that the transformation variables **F** and **A** are fully connected are:

$$\begin{aligned} a_1 &= f_1 \vee f_2 \\ a_2 &= f_3 \\ a_3 &= f_5 \\ a_4 &= f_6 \\ a_5 &= f_7 \vee f_4. \end{aligned}$$

The constraints that ensure that each peak mz is represented in the solution are:

$$\begin{aligned} f_1 = mz_1 &= 1 \\ f_2 \vee f_4 = mz_2 &= 1 \\ f_3 = mz_3 &= 1 \\ f_5 \vee f_6 \vee f_7 = mz_4 &= 1. \end{aligned}$$

There are several solutions to these constraints, since the observed peaks could be result of mf_1 alone, or a combination of mf_1 with mf_2 and/or mf_3. An optimisation function

(see Section 6.4), together with a COP search is therefore required, to find the most likely solution. In addition, other factors should be considered, such as whether or not mf_1 is a 'possible' molecular formula.

6.4 The Constraints Optimisation Problem

In this section, it is shown how the model of the system presented in the previous section can be expressed as a COP, which can then be solved. There are three facets to this COP. The first is the set of variables are required to describe the problem. The second is the hard constraints that describe all *valid* solutions, and which should implement the structure and connectivity of the system based on the graph constructed, for example as shown in Figure 6.4. The third facet is the optimisation function that places value on each solution and allows the optimal solution to be determined.

6.4.1 Optimisation Function

As stated previously, one would assume that in the example in Figure 6.4, the optimal solution would be $\mathbf{mf} = (1, 0, 0)$, since one would expect that the presence of multiple adducts suggests the presence of the ion mf_1 over mf_2 and mf_3, which only appear to be present with a single adduct ion each. Logically, it is tempting to define the optimal solution to the COP as the solution with the smallest number of mf's, i.e. the solution where the smallest number of molecular formulae can explain the observed peaks. However, this approach would introduce a bias against the inclusion of additional m, even if such an additional mf is highly likely to exist in the sample. For example, if mf_2 was present with a good isotope pattern but a single adduct, it would still be expected that mf_2 was present in the sample, in addition to mf_1.

The proposal for the optimisation function is to instead determine a value v for each mf that captures 'how well' that mf itself is explained by the observed spectrum. For example in Figure 6.4, $\mathbf{mf} = (1, 0)$ might have value vector $\mathbf{V} = (50, -)$, where each element in V is the value of the corresponding mf, and a value of '$-$' signifies the absence of the corresponding molecular formula. Compare this with the case where $\mathbf{mf} = (1, 1)$ in which, say, $V = (20, 60)$, where the value of mf_1 is lower due to fewer adducts present in the solution of mf_1. The overall mean value is higher in the second case, therefore the optimisation function proposed is:

$$f_{opt}(\mathbf{F}, \mathbf{A}) = \frac{\sum \mathbf{V}}{n_{mf}},$$

where n_{mf} is the number of valid molecular formulae in the solution, $\mathbf{F}$ and $\mathbf{A}$ are the

transformation vectors, and $\mathbf{V}$ is the vector of values for each mf.

By taking the *mean* value over all 'valid' elements in V as the overall solution value, one calculates that the first and alternate solutions have $f_{opt} = 50$ and $f_{opt} = 40$, respectively. The first solution is therefore preferable. If, however, a good isotope pattern was measured for mf_2, the value vector for the second solution could become $V = (20, 90)$, giving $f_{opt} = 55$ and indicating that the second solution is preferable to the first.

The challenge is to derive a function that places a *realistic* and *quantitative* value on each solution, and this will necessarily include the aspects of the system that have been identified as contributing evidence for the existence of a certain molecular formula. The system aspects, identified in Section 5.2.2, that may provide evidence for a molecular formulae include:

1. Mass measurement error, as defined by the difference in the observed peak m/z and the *expected* m/z of the $isoion$ assigned to the peak;
2. The presence of multiple adducts;
3. The presence and intensity accuracy of an isotope pattern;
4. The likelihood of the molecular formula existing in the sample by virtue of its formula characteristics.

The four aspects above are important in placing confidence on the identification of ion molecular formulae to peaks in a spectrum, and must therefore be captured within the optimisation function. In order to achieve this, the aspects must be expressed quantitively and they should be combined in such a way as to yield a single numerical value indicating the overall value of the solution.

The following section describes how utility theory can be used to quantify and combine the value of each aspect.

6.5 Optimisation - Metrics and Utility Theory

In this section, the four metrics listed in Section 5.2.2 that will contribute to the optimisation function are considered in turn. The aim is to specify the metrics in such a way that they can be combined. However, since the metrics are orthogonal, they cannot be combined directly, so utility theory is used to design the overall value of the solution, as a function of each metric.

To avoid excessive complication, the metrics will be defined on a 'per molecular formula' basis, i.e. each molecular formula will be assigned utility, in terms of each metric, and the

overall value of the solution will be calculated as an average over all mf, as described in Section 6.5.5.

6.5.1 Mass Measurement Error

The mass measurement error (MME) is typically measured in parts per million (ppm), and is defined as shown in equation (6.3) [23]. It is the measurement error that is observed for an ion population hypothesised to exist, under the assumption that the peak being measured is caused by that ion population.

$$\text{MME} = \frac{\text{measurement error}}{\text{accurate mass}} \times 1e6 \text{ ppm} \tag{6.3}$$

As discussed in Section 6.2, each isotopic ion formula, $isoion$, is derived from a single peak (mz) in the measured spectrum, and consequently there is a MME associated with each $isoion$ in the solution. A single value is required to quantify the overall MME for a particular mf, and if equal weighting is placed on all $isoion$'s for a particular mf, the mean is adopted in this work as representative of the overall MME. Therefore, the mean MME for each mf, over the $isoion$'s linked to that mf, is mme_{mf}:

$mme_{mf} = \Sigma mme_{isoion}/n_{isoion}$,

where mme_{isoion} is the mass measurement error across the related set of $isoion$'s and n_{isoion} is the number of $isoion$'s related to mf.

It is generally the case that the mass measurement error probability density function (PDF) is normally distributed, and this may be attributed to the central limit theorem [32], which states that the sum of independent random variables has a normal distribution. The theorem can be extended to this case, in which many noise sources, such as thermal noise and electrical noise, can be expected to be random and independent, as described in Section 2.6.1. In order to remove the dependence of the MME on the instrument or sample, the MME is expressed as a standard error, se, defined as shown in equation (6.4) [49]. A random variable standardised in this way has zero mean and unity variance. Intuitively, a variable expressed as a standard error describes the number of standard deviations the variable is from its expected value, and therefore in the case of MME, indicates a measure of the likelihood that the relationship between the hypothesised measurement, mz, and isotopic ion formulae, $isoion$, is correct.

$$se = \frac{x - \mu}{\sigma} \tag{6.4}$$

In equation (6.4): x is the normally-distributed dependent variable, in this case MME; and μ and σ are the mean and standard deviation of x, respectively.

6.5.2 Isotope Pattern Accuracy

As described in Section 5.1.1, while it is expected that each molecule in the sample is present in every isotope form possible, the number of isotopes observed in a mass spectrum will depend on the concentration of the compound in the sample and the dynamic range of the measurement. The peaks expected to be seen in a spectrum for a compound are known collectively as the 'isotope pattern' for that compound.

In the system presented here, each mf is associated with one or more *adion*, and each *adion* is associated with one or more *isoion*. The *isoion* form the observed isotope pattern, and by comparing this observed isotope pattern with the expected isotope pattern, the objective is to obtain a metric describing the quality of the match of these two patterns. In order to quantify the accuracy of the match between the measured and expected isotope patterns, isotopes will be considered in *pairs*. This allows a simple expression for the accuracy of the measured isotope ratios compared to the expected isotope ratios, and by considering the mean value over multiple pairs, an overall measure for isotope pattern match can be calculated. This approach yields a figure that is independent of the number of isotopes, but a function of each isotope ratio observed.

The accuracy of the isotope ratio, $isoacc$, for a single pair of isotopes is described in terms of the measured ratio (mr) and the expected ratio (er) for the pair:

$isoacc = (1 - |\varphi|) \times 100\%,$

where $\varphi = \frac{mr-er}{er}$ for $mr \leq 2er$, otherwise $\varphi = 1$ and where $mr, er > 0$.

This yields a measure of the accuracy of the isotope ratio for a single pair of isotopes, where $isoacc = 100\%$ represents a perfect match, and $isoacc = 0\%$ a poor match. A threshold of $me = 2er$ defines the 'worst case' accuracy, beyond which $isoacc$ is 0% to indicate an accuracy beyond interest. The isotope pattern accuracy function is illustrated in Figure 6.5 below.

Having defined an accuracy measurement for a single pair of isotopes, the measurement is extended to the case where more than two isotopes are present. This is achieved by calculating the *mean* accuracy value over each isotope ratio in the pattern. For example, consider three isotopes that are observed for an *adion* as shown in Figure 6.6.

The accuracy for pair (a, b) is:

$isoacc_{(a,b)} = \left(1 - \left|\frac{y_a/y_b - n_a/n_b}{n_a/n_b}\right|\right) \times 100\%,$

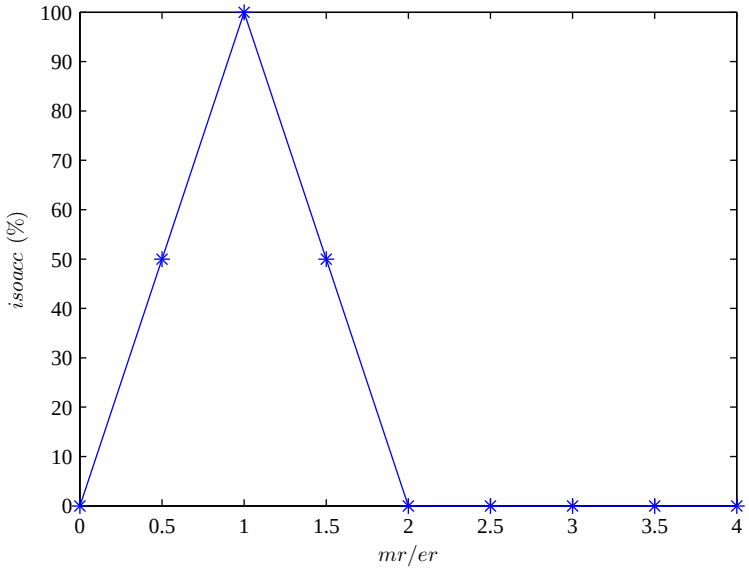

Figure 6.5: Isotope pattern accuracy, $isoacc$, for a peak pair as a function of measured ratio (mr) and expected ratio (er).

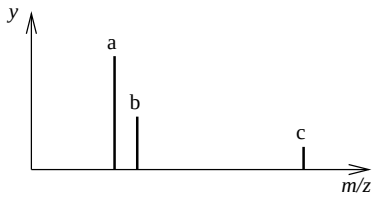

Figure 6.6: Example isotope pattern consisting of three peaks

where n_a and n_b are the natural abundances of the *isoion* associated with peaks a and b, respectively, and y_a and y_b are the measured abundances of peaks a and b, respectively. A similar result is obtained for pairs (a, c) and (b, c). The accuracy for this *adion*, $isoacc_{adion}$, is then taken as the mean over all linked *isoion*, i.e.

$isoacc_{adion} = \frac{isoacc_{a,b}+isoacc_{a,c}+isoaccb,c}{3}$.

The overall accuracy for this molecular formula, $isoacc_{mf}$ is then the mean accuracy over all linked *adion*, i.e. :

$isoacc_{mf} = \frac{isoacc_{adion1}+isoacc_{adion2}+...}{n_{adion}}$,

where n_{adion} is the number of *adion* linked to neutral molecular formula mf. In the case that no multi-peak isotope pattern is observed, $isoacc_{mf} = 0$.

6.5.3 Presence of Adducts

As discussed in Section 5.1.1, it is currently not possible to accurately predict which adducts will be present in a mass spectrum obtained using FT-ICR MS. However, there are a set of 'common' adducts that are present in positive mode FT-ICR MS experiments and that appear to form more readily than others: $[M]^+$, $[M+H]^+$, $[M+K]^+$ and $[M+Na]^+$ [38]. These correspond to the ionised, protonated, potassiated and sodiated forms of the molecules, respectively. The abundance of each of these forms depends upon the availability of the adducts within the sample. For example, a 'salty' sample containing a large amount of potassium and sodium would be expected to contain a large number of potassiated and sodiated adducts in the mass spectrum, compared to the ionised and protonated forms.

In order to capture the approximate expected relative quantities of each adduct, three groups are defined to describe combinations of adducts that appear in the spectrum. The groups are labelled 'likely', 'somewhat likely' and 'unlikely'. The description of these groups, and the adducts belonging to each, is based upon observations made during an environmental toxicity study [16]. The groups are therefore applicable to this work, which is based upon similar samples and sample preparation methods. While the group descriptions are vague and qualitative, it is the *relative* likelihood that is being captured. When the formation of adducts is better understood and can be modelled, a more directly quantitative approach can be adopted.

By observing the prevalence of adduct patterns found in research by Taylor *et al.* [16], the combinations of the four main adducts are classified as follows:

The adduct group for molecular formula mf, $adductgroup_{mf}$, is assigned by looking up

Group	Qualitative Likelihood of Being Observed	Adduct pattern
A	Likely	$[M+H]^+$, or the set $\{[M+H]^+,[M+Na]^+,[M+K]^+\}$
B	Somewhat Likely	$[M+Na]^+$, or $[M+K]^+$, or $[M]^+$, or the set $\{[M+H]^+,[M+Na]^+\}$
C	Unlikely	*all other combinations*

Table 6.1: Adduct pattern classification.

in Table 6.1 the adducts for the set of *adion* that relate to the mf.

6.5.4 Likelihood of Molecular Formula

The fourth and final optimisation criteria considered is the likelihood that the molecular formula mf exists in the sample. This should be determined without making any assumptions about the sample, such as its composition. As outlined in Section 5.2.1, one commonly adopted approach is to extract, from the existing body of knowledge, certain 'rules' about the composition of molecular formulae from databases. As discussed in Section 5.2.2, an improved method is to determine the trends in molecular formulae composition, using large databases of compounds. For example, instead of stating that $x\%$ of molecular formulae have a ratio of hydrogen to carbon atoms of between a and b, the distribution of the logarithm of the hydrogen to carbon ratio is calculated, which is seen to be approximately normal as shown in Figure 6.7.

This clearly shows the continuous nature of the feature, and suggests that one could assign a continuous value to the 'likelihood' of a molecular formula existing, based on that feature alone. As well as yielding accurate likelihoods for those molecular formulae that are known to exist, this allows extrapolation of the characteristics to as yet unknown compounds. In this way, the problem of assigning 'unusual' formulae as impossible is avoided, while also offering the benefit of capturing how 'typical' a formula is, compared to the database. Furthermore, it is discovered that Fiehn's other rules can also be described as continuous, approximately normal distributions. By assuming independence between these n features, one can calculate an overall likelihood measure based upon the distance in n-dimensional space of a single formula from the 'ideal' formula. The Mahalanobis distance, d^2_{mf}, is used to achieve this as shown in equation (6.5) [116], in which se_g is the standard error of feature g, calculated using equation (6.4).

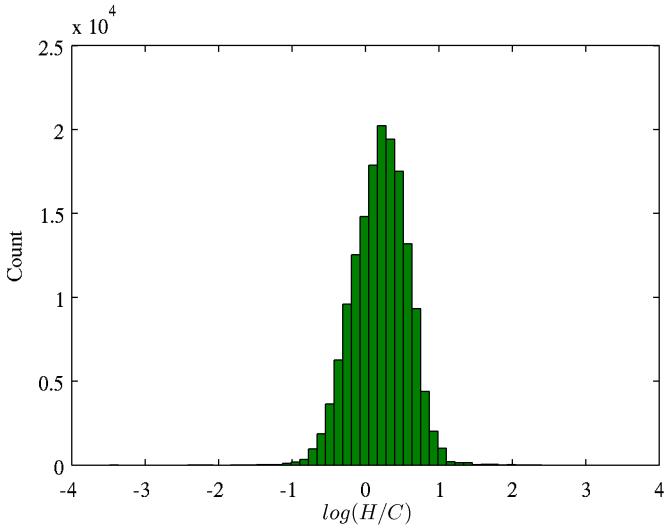

Figure 6.7: Ratio of hydrogen to carbon atom count for appropriate compounds with unique molecular formulae in index range 00000001 to 10000001 in PubChem database [4], giving a total of 156,032 unique molecular formulae included.

$$d_{mf}^2 = se_{H/C}^2 + se_{N/C}^2 + \dots \tag{6.5}$$

The Mahalanobis distance requires the independent variables to be normally distributed such that the standard error values are comparable, and so the molecular formula features that yield such a distribution are measured from the PubChem database [4], and are shown in Table 6.2. The process to select the molecular formulae is as follows. The unique molecular formulae for the first 10,000,000 compounds are extracted from the database, and then filtered based on the following criteria: they must contain at least one hydrogen, nitrogen, oxygen, phosphorus or sulfur atom; they must contain at least one carbon atom; they must not contain any Silicon atoms.

These features are plotted below for appropriate compounds with unique molecular formulae in index range 00000001 to 10000001 in the PubChem database [4].

Therefore, the Mahalanobis distance d_{mf} of a molecular formula from the 'typical' molecular formula is calculated as shown in equation (6.6), in which μ and σ are the mean and standard deviation of the $\log(feature)$ distribution, assumed to be normal.

$$d_{mf} = \left(\frac{\log(feature) - \mu}{\sigma}\right)^2 + \dots \tag{6.6}$$

In order to validate this measure as representative of the likelihood of the molecular formulae, a set of formulae from the PubChem database was extracted and d_{mf} calculated

log(Feature)	Condition	Formula Count	μ	σ
$\log(H/C)$	$H > 0$	156,032	0.301996	0.374198
$\log(N/C)$	$N > 0$	131,014	-1.803085	0.853849
$\log(O/C)$	$O > 0$	141,439	-1.388731	0.854676
$\log(P/C)$	$P > 0$	16,944	-2.083101	0.708180
$\log(S/C)$	$S > 0$	72,728	-2.206890	0.771069
$\log(N)$	$N > 0$	131,014	0.784755	0.623614
$\log(O)$	$O > 0$	141,439	1.241222	0.737060
$\log(P/m)$	$P > 0$	16,944	-5.570163	0.400399
$\log(S/m)$	$S > 0$	72,728	-5.419571	0.581384

Table 6.2: Summary of molecular formula features.

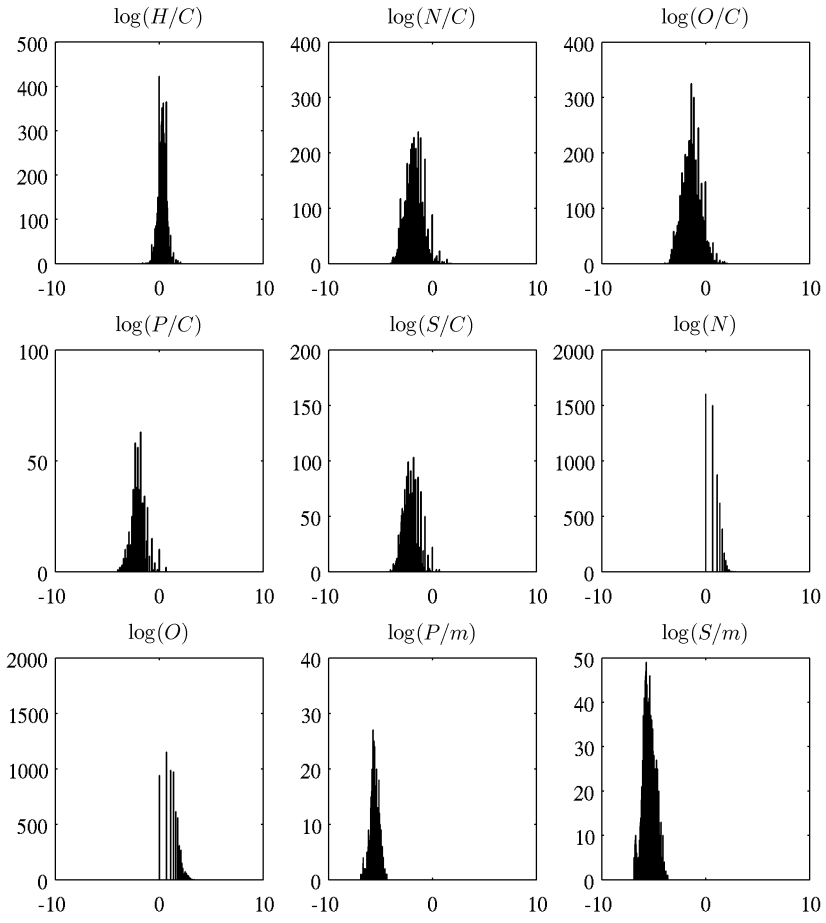

Figure 6.8: Feature distributions of all unique molecular formulae in index range 00000001 to 10000001 in the PubChem database [4].

for each using equation (6.5), and the results are shown in Figure 6.9. The feature d_{mf} is approximately normally distributed and therefore indicates that the measure is capturing the desired likelihood of the formula existing.

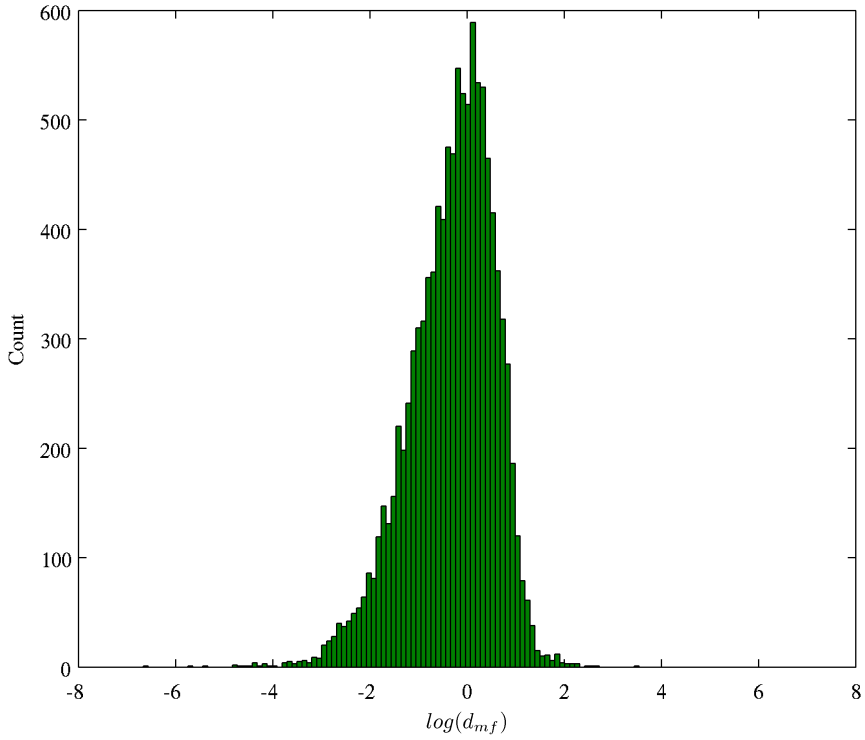

Figure 6.9: Mahalanobis distance for appropriate compounds with unique molecular formulae in index range 10000001 to 10025001 in PubChem database [4], giving a total of 11,467 unique molecular formulae included.

6.5.5 Placing Value on a Molecular Formula Assignment using Utility Theory

Described in Sections 6.5.1 to 6.5.4 are four metrics that can be used to place a confidence value on the assignment of a particular molecular formula for a measured mass. With the exception of the presence of an adduct pattern, the metrics are quantitative, but cannot be used together since they are not normalised and represent orthogonal aspects of an assignment. Ultimately, a single value is required to represent the value on each molecular formula assignment, and so a method is required to combine both quantitative metrics and the grouped adduct pattern metric, and scale them such that a meaningful comparison can be made between two values for two different assignments. Utility theory was chosen to achieve this, and place a single value on a molecular formula assignment, given the outcome of the four metrics described above.

The fundamental premise of utility theory is that different metrics are assigned a numerical value, as a function of the metric value or group, that credits that metric with a 'utility', as follows:

1. U_{me}, the utility of the mass measurement error;
2. U_{ir}, the utility of the isotope ratio pattern;
3. U_{ap}, the utility of the adduct pattern;
4. U_{ml}, the utility of the molecular 'likelihood'.

The utility should represent, on a common scale, the desirability of the metric, and has roots in economic decision-making, where un-quantifiable parameters were assigned a utility based upon the experience and intuition of the manager making the decision [117]. Utility theory is thus able to capture the *preference* of the expert in making decisions where no mathematical model exists to precisely calculate the preferred solution. There is a formalised approach to deriving the utility functions [118] that also establishes the relative contribution of the different metrics to the overall utility.

In the interests of creating a working solution as a basis for further refinement and optimisation, a set of utility functions are proposed that capture the most significant differences between the solution metrics. Following the procedure outlined by Moore and Thomas [118], and applying knowledge and experience of interpreting mass spectra gleened from experts in the field, the graphs shown in Figure 6.10 illustrate the utility functions, which map a metric such as the MME to a utility value.

The following four general statements summarise the criteria that were used to determine the utility functions and their relative value:

1. An accurate isotope ratio pattern is highly valuable;
2. The MME is only valuable if it is very low;
3. The adduct pattern and MME are less valuable than isotope ratio pattern and molecular likelihood;
4. A high molecular likelihood is more valuable than either MME or adduct pattern.

Using these individual metric utilities, the utility for each molecular formula, U_m, is simply the sum of all utilities:

$$U_m = U_{ir} + U_{me} + U_{ml} + U_{ap},$$

where U_{ir} is the utility of the isotope ratio pattern, U_{me} is the utility of the MME, U_{ml} is the utility of the molecular likelihood and U_{ap} is the utility of the adduct pattern.

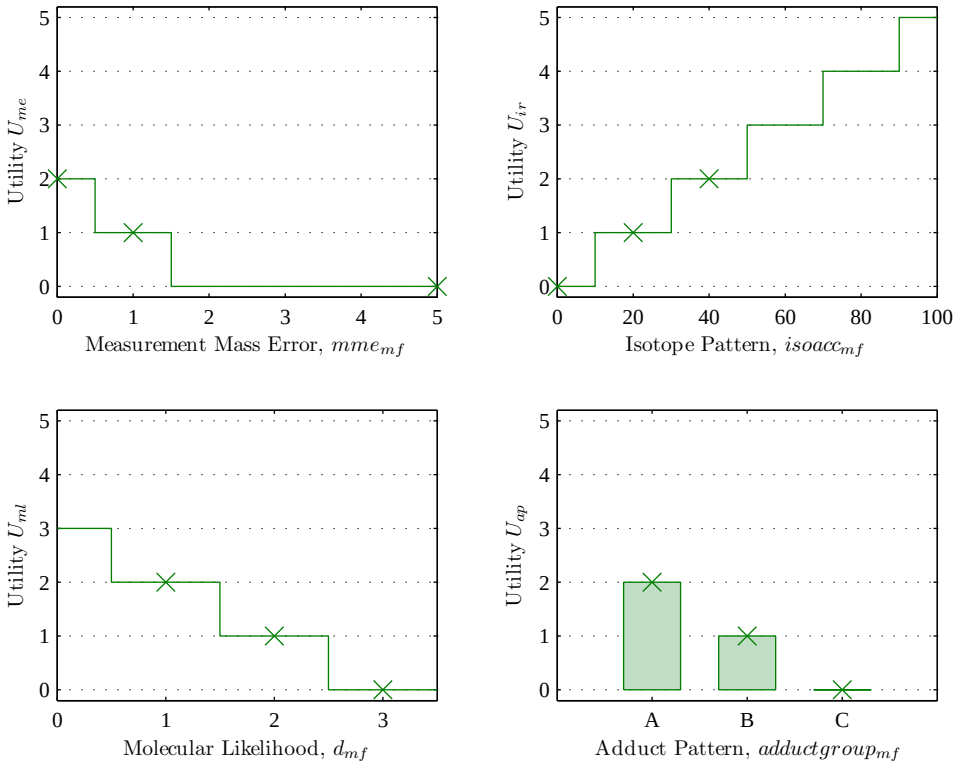

Figure 6.10: Utility graphs for measurement mass error; isotope pattern; molecular likelihood and adduct pattern.

The overall utility of the solution, U, is taken as the mean of the n individual molecular formula utilities:

$U = \overline{U_m} = \frac{\sum U_m}{n}$.

Intuitively, by setting U to be the optimisation function of the COP, the solution will identify the molecular formulae with the highest average utility, or 'quality'. In the case where a solution could be explained by multiple molecular formula assignments, additional molecular formulae will only be considered part of the solution if they increase the average utility: i.e. if they are 'better' than the worst 50% of the assignments.

To facilitate efficient optimisation branching as discussed in Section 6.6.1, two auxiliary variables are required. Auxiliary variables are not essential to solve the COP, but help implement the problem, for example by describing some intermediary quantity that enables constraints to be expressed more succinctly [115]. The auxiliary variables are M_{count} and U_{sum}, which represent the number of valid molecular formulae and the sum of the utility for all valid molecular formula, respectively. The calculation of these variables is:

$M_{count} = \sum_{\forall i} (mf_i = 1)$, and

$U_{sum} = \sum_{\forall m} U_m$.

The overall utility for a solution, the real variable U, can then be calculated as shown in equation (6.7) below. Since U is a positive real number representing the continuous value of mean utility, is is not readily represented as a finite domain integer variable. Therefore, it will not be stored as a variable directly, and instead M_{count} and U_{sum} will be used to form the optimisation function.

$$U = \frac{U_{sum}}{M_{count}} \tag{6.7}$$

6.6 Solution Search

This section describes the methods used to solve the problem that have been defined in the previous section in terms of a COP. Firstly, the search method adopted is detailed, which is followed by some of the key optimisation methods adopted.

6.6.1 Branch and Bound Search

Having established the variables, domains and constraints that will model the COP, it remains to describe the search method that will find the optimal solution. There are

several search methods, arguably the most common of which is the depth-first search, in which variables are assigned values starting from the first variable and moving to the last. At each stage of a partial solution, i.e. where some, but not all, of the variables have been assigned a value, a check is made that the domain of the unassigned variables is still consistent with the constraints [115]. In the event that a 'dead end' is reached, i.e. a variable assignment precludes any of the subsequent variables being assigned a value that satisfies the constraints, the algorithm un-assigns the most recent assignments until a potential solution exists again. This process is known as 'backtracking' [119].

The branch and bound search is an extension to the depth-first search method. In branch and bound, solution subsets are excluded from the search by estimating the upper and lower bounds of the optimisation quantity. For example, a COP consists of system variable

$X = (x_1, x_2)$,

and the constraint is

$(x_1 + 2x_2) \geq 2$.

The steps in the branch and bound search can be represented as a constraints graph, shown in Figure 6.11.

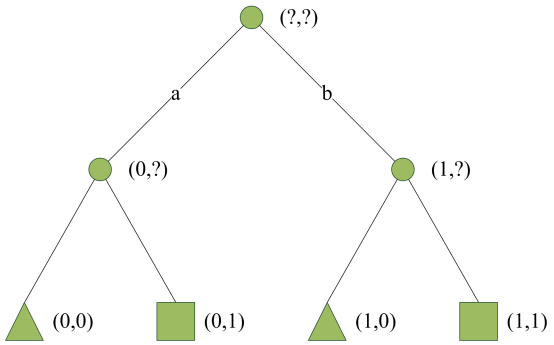

Figure 6.11: Representing a COP search as a constraints graph, for a simple example consisting of variable $X = (x_1, x_2)$. The numbers in parentheses indicate the assignments at each node to the two variables. Nodes where the variables are not yet fully assigned are shown as circular; 'failed' nodes in which the full or partial solution does not satisfy the constraints are triangular; and complete, valid solutions are represented by squares.

Initially X is unassigned. The first branch a assigns $x_1 = 0$. Subsequent branching attempts to assign $x_2 = 0$ and then $x_2 = 1$. Only when the constraint is satisfied is a solution found. The right-hand branch b assigns $x_1 = 1$ and subsequent assignments follow the same pattern as the left-hand branch. At each partial solution state, the constraints will be re-propagated, i.e. checks will be made to ensure that the domains of each variable

is consistent with the constraints, given the domains of all other variables. Where variable domains have changed, and the domains of other variables is altered as a consequence, new variable constraints are created to improve the efficiency of the search. This process is termed 'constraint propagation' [120], and is important to reduce the solution space as far as possible.

The order of choosing the variables, the 'branching order', can have a significant impact on the efficiency of the search. For example, if the branching order was instead $\{x_2, x_1\}$, the search could be illustrated by the constraints graph in Figure 6.12.

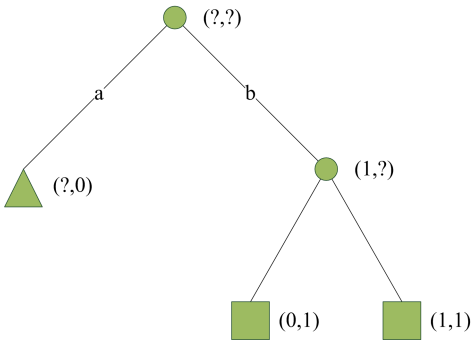

Figure 6.12: The search assignments when the order of the search variables is changed.

At branch a, the value 0 is assigned to x_2 and as the constraint is propagated, it is found that the constraint cannot be met regardless of the value of x_1. Compared to the search in Figure 6.11, this search contains fewer nodes, and so highlights the importance of selecting a suitable order of branching. The variable order finally chosen, for search of the solution to the profiling problem addressed here, is shown in Table 6.3. The value order (e.g. minimum to maximum, or *vice-versa*) for assigning variable values is also shown. These were chosen after experimentation of the search method as the variable order is changed. It is important to note that the solution of the search is not affected by this choice, since the COP being described is unchanged.

Variable	Order of value assignment
U	max. to min.
U_{sum}	max. to min.
M_{count}	min. to max.
mf	1 then 0
A	1 then 0
F	1 then 0

Table 6.3: Variable search order and order of value assignment.

The target mean utility U is selected first. Since this is an integer variable representing

the positive real value U_{sum}/M_{count}, the integer value of U represents multiple value assignments of U_{sum} and M_{count}. For example, the result of $U = U_{sum}/M_{count} = 10/2$ and $11/2$ both yield the integer result $U = 5$. In order to address this issue, the resolution of U is increased by instead branching on $U \times 1000$, defined as $U_{1000} = U_{sum} \times 1000/n$. Next, the **mf** variables are assigned, corresponding to the presence of molecular formulae. Having selected **M**, the **adion** variables are selected (**A**) followed by the **isoion** (**F**).

6.6.2 Partitioning

Since this is a combinatorial search problem, a method of reducing search time significantly is to split or partition the problem into completely independent sub-problems. The solutions to these sub-problems can then be directly combined to yield the overall solution. The advantage of this approach is that partitioning enforces independence between the variables that form each sub-problem, and consequently the search space is reduced considerably.

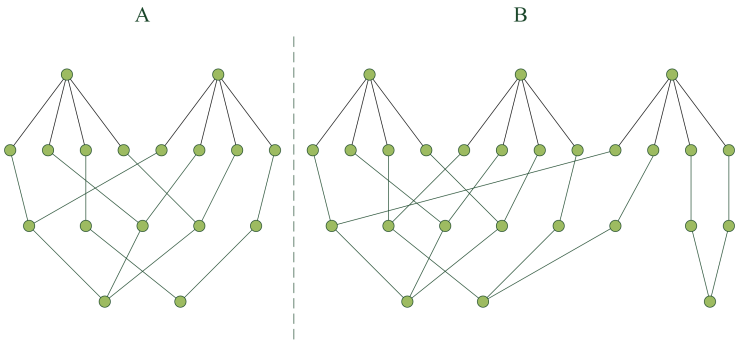

Figure 6.13: Example search space split into two independent partitions A and B.

In the example shown in Figure 6.13, partition A is found to be independent of B, and the problem is solved as two sub-problems. An added advantage of this approach is that each partition can be solved as a separate process in a multi-processor environment, and with unlimited processors the solution time is limited to the maximum search time across all partitions.

6.6.3 Redundant Constraints

As their name implies, redundant constraints do not affect the solution space of the COP, but allow implicit constraints to be explicitly stated. If selected carefully, such constraints can significantly reduce the search space by allowing inference by propagation that would otherwise not occur. In this case, a redundant constraint imposed on the solution is that

the sum of the utility, U_{sum} must be less than the maximum utility possible, given the number of active molecular formulae, M_{count}. This is determined by summing the top M_{count} molecular formulae with the highest utility, i.e. :

$$U_{sum} \leq \sum_{i=1}^{M_{count}} U'_i,$$

where U' is the utility of the molecular formula, U_m, ordered by decreasing magnitude.

6.6.4 Dominance

Dominance is the concept that partial solutions of a subset of the variables can be removed from the search space where an alternative solution of the same subset of variables is guaranteed to improve, or at least not make worse, the overall value of the solution. There are two occasions where this can be used to reduce the search space in this COP.

The first application of dominance rules applies for molecular formulae, mf, connected to the same ion plus adduct, *adion*, and removes unnecessary mf from the solution. For example, consider the partial system shown in Figure 6.14.

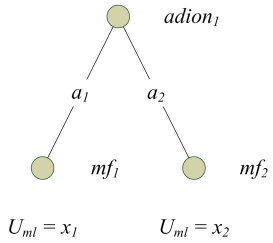

Figure 6.14: A sub-graph showing two alternative potential solutions in mf_1 and mf_2: dominance rules can eliminate one of these solutions.

Since both mf_1 and mf_2 are connected *only* to $adion_1$, the optimal solution may include valid mf_2 but will never include valid mf_1 since $(U_{ml})_2 > (U_{ml})_1$.

The second application of dominance rules also removes unnecessary mf and is applicable in the case that a mf is associated with a single m/z. Any other mf that are similarly associated with the same m/z (i.e. single isotope, single adduct) are 'parallel' options in which the selection of either branch is independent of any other variable in the system and thus branches with a lower U_m can be removed from the search space. For example, in Figure 6.15, three parallel options exist in which the two outer options, mf_1 and mf_3 can be removed from the search since they return a lower utility than mf_2.

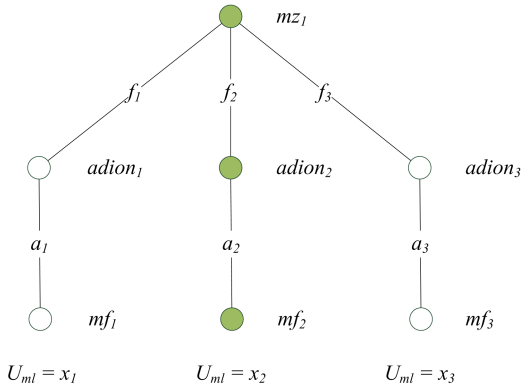

Figure 6.15: A sub-graph showing three alternative parallel solutions in mf_1, mf_2 and mf_3, with molecular likelihood utility x_1, x_2 and x_3, respectively. In the case that $x_1, x_3 < x_2$, dominance rules can again be used to eliminate the two solutions containing mf_1 and mf_3 from the search.

6.7 Results and Discussion

The COP detailed above was implemented in C++ using Gecode libraries [121]. An object-oriented approach was adopted to generate and store the elements, formulae and COP system.

Validation of the system was carried out in two stages. Firstly, simulated data was used to verify the performance of the search in case of noise-free and noisy data. Secondly, data from non-biological samples of known composition was used to investigate performance under realistic conditions. Thirdly, data from real biological samples was used to validate the technique by comparison with existing methods. Finally, further validation of the effect of noise on the method is performed.

6.7.1 Simulated Data

The objective of the COP solver was to elucidate the neutral molecules present in the sample, by interpreting the spectral artefacts such as adducts and isotopes, while being as robust to noise as possible. Therefore, the first stage is to assess the performance of the system in the ideal case, using simulated spectra without noise. This allows the validation of the utility functions that have been described above. Secondly, noise is introduced, both on the original peaks and as a source of new, purely noise peaks.

The simulated data is created from a pure hypothetical sample consisting of the common metabolites and selected adducts, as shown in Table 6.4.

Metabolite	Molecular Formula	Ion Forms Present	Total Relative Abundance (%)
Threonine	$C_4H_9NO_3$	$[M+H]^+$	105.8
Proline	$C_5H_9NO_2$	$[M+H]^+$	101.4
Glutamic acid	$C_5H_9NO_4$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	98.8
Methionine	$C_5H_{11}NO_2S$	$[M+H]^+$, $[M+Na]^+$	95.4
Histidine	$C_6H_9N_3O_2$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	89.0
Leucine, Isoleucine	$C_6H_{13}NO_2$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	82.9
Lysine	$C_6H_{14}N_2O_2$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	77.7
Arganine	$C_6H_{14}N_4O_2$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	72.6
Phenylalanine	$C_9H_{11}NO_2$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	68.5
Tyrosine	$C_9H_{11}NO_3$	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$	63.0
Tryptophan	$C_{11}H_{12}N_2O_2$	$[M+H]^+$, $[M+Na]^+$	57.2
Ophthalmic acid	$C_{11}H_{19}N_3O_6$	$[M+H]^+$	52.3

Table 6.4: Simulated spectrum composition. The total abundance was split approximately equally between the adducts.

The peaks generated include isotopic forms of the elements carbon, hydrogen, nitrogen, oxygen and sulphur, however peaks are limited to those theoretically observable given an instrument dynamic range of 1×10^4, i.e. with relative abundance $\geq 0.01\%$. The 250 peaks generated include adduct and isotope peaks and are plotted as a mass spectrum in Figure 6.16.

The COP system was used to extract molecular formulae from the peak list. It was quickly found that the solution search is very intensive and in order to obtain practical run times, it was necessary to impose limitations on the number of adducts and isotopes that are represented in the system. Therefore, only the adduct compounds $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$ were included, together with the isotope forms that include only carbon-13 or potassium-41, or both.

Using the COP system described above, a total of 94 unique molecular formulae were identified as being present in the sample. The molecular formulae of all 12 simulated compounds were identified, as shown in Table 6.5. In each case, all adducts present in the peak list were correctly identified, as well as the carbon-13 and potassium-41 isotopes. The difference in measured abundance can be explained by the missing contribution from the hydrogen, nitrogen, oxygen and sulphur isotope peaks to the total abundance. Instead, these peaks have been assigned to the 82 additional identified molecular formulae that were not present in the simulated sample, but have been identified as molecular formulae, as shown in Figure 6.17. These additional molecular formulae, however, are identified

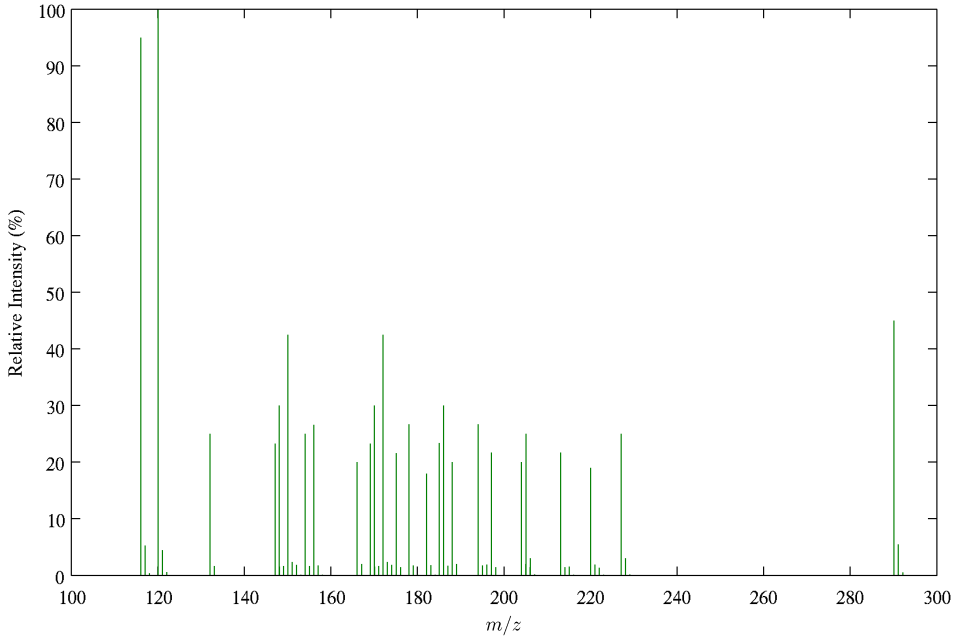

Figure 6.16: Simulated spectrum (noise-free), including isotopes.

by the system as quite distinct from the sample compounds, as shown by comparing the abundance and overall utility of the formulae. Figure 6.17 shows that the 12 correctly assigned compound formulae are identified with significantly higher abundance than all others, and generally with higher utility. This demonstrates that the system discriminates components of the solution that are noise.

In each case, the utility for mass error and isotope ratio are assigned maximum values of 2 and 5, respectively (see Figure 6.10. Thus for this simple example, the adduct patterns and molecular formula likelihood are the factors that affect the overall utility. In the case of methionine and tryptophan, the adduct pattern consisting of the protonated and sodiated forms of the molecule, i.e. $[M+H]^+$ and $[M+Na^+]$, results in lower utility. Histidine, phenylalanine, arganine, tyrosine and ophthalmic acid have molecular formulae with characteristics that result in a reduced utility of 2. Tryptophan ($H_{12}C_{11}N_2O_2$) has a reduced molecular formula utility of 1, reflecting the non-centric composition as described by the database profiling in Section 6.5.5.

The above has shown that the COP solver is robust to partial *a priori*, i.e. when isotopes are present in the spectrum that the search does not include. Shown in Table 6.5 are the results obtained for the simulated spectrum that included the following noise components:

1. Random m/z measurement noise in the form of zero mean Gaussian additive noise with standard deviation $\sigma = 0.165$, as observed in Section 3.3.1. This effectively

Metabolite	Measured Abundance	Abundance Error (%)	Utility U_m (In Noise)
Threonine	104.5	-1.2	12 (12)
Proline	100.4	-1	12 (11)
Glutamic acid	97.4	-1.4	12 (11)
Methionine	89.8	-5.8	11 (10)
Histidine	87.5	-1.6	11 (10)
Leucine, Isoleucine	82.1	-1	12 (11)
Lysine	76.6	-1.4	12 (11)
Arganine	71.1	-2	11 (10)
Phenylalanine	67.9	-0.9	11 (10)
Tyrosine	62.3	-1.2	11 (11)
Tryptophan	56.5	-1.3	9 (8)
Ophthalmic acid	50.8	-2.8	11 (10)

Table 6.5: Simulated spectrum profiling results.

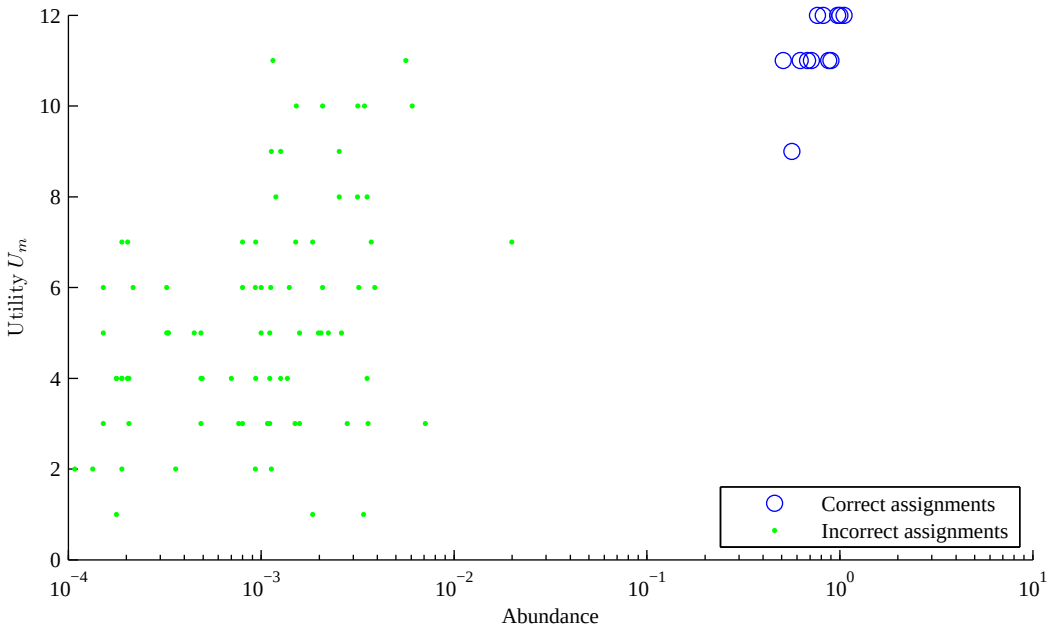

Figure 6.17: Utility and abundance of all molecular formulae identified as present in the simulated spectrum, showing the effect of including only carbon-13 and potassium-41 isotopes in the solution search.

adds an m/z measurement error to each peak.

2. Additional noise peaks uniformly distributed by m/z with intensity at the minimum level of detection as determined from the instrument dynamic range of 1×10^4. The number of noise peaks is calculated from the results in Chapter 4, where a recommended filter setting of $\text{SNR}_{thresh} = 3.5$, $r = 2$ and $s = 30\%$ yields a predicted real and noise peak count of 3705 and 96, respectively. Applying the same ratio of noise to real peaks (2.6%) results in a prediction of 7 noise peaks in this case.

It can be seen in Table 6.5 that when noise is introduced into the system, the utility of each assignment tends to decrease slightly while the abundance remains unaffected, as shown by the utility values in parentheses in the table. The effect on utility can be attributed to the reduced measured mass accuracy that has been applied, and the abundance shows that the same set of peaks have been identified for each compound.

The assignments made to the seven additional noise peaks are shown in Table 6.6. In three cases, no assignment could be made, and while potential molecular formulae were found in each case, the hard constraints on the system eliminated the assignments during the 'pruning' process, and effectively identified the peaks as random noise and not attributable to an allowable molecular formula. In the case of the remaining four peaks, the combination of isotope ratio pattern, adduct pattern, molecular likelihood and measurement error results in a very low utility and abundance: such peaks are easily identifiable as noise.

Noise Peak m/z	Peak Assignment	Molecular Formula	Abundance	Utility U_m
119.366429	-			
136.794785	-			
173.594094	$[^1H_{80}\,^{12}C^{13}C_2\,^{23}Na^{32}S]^+$	$H_{80}C_3S$	0.0001	1
193.597232	$[^1H_{74}\,^{12}C_3\,^{13}C\,^{14}N_5]^+$	$H_{73}C_4N_5$	0.0001	4
193.719173	-			
235.556069	$[^1H_{71}\,^{12}C_{10}\,^{14}N_2\,^{16}O]^+$	$H_{70}C_{10}N_2O$	0.0001	4
260.329670	$[^1H_{45}\,^{12}C_7\,^{13}C_3\,^{14}N_2\,^{16}O_2\,^{32}S]^+$	$H_{44}C_{10}N_2O_2S$	0.0001	6

Table 6.6: Simulated spectrum noise peak assignments

These results demonstrate that at the anticipated levels of noise, no significant degradation of the results is observed where good adduct and isotope patterns are present.

6.7.2 Measured PEG Data

In this section, the aim is to validate the performance of the COP solver using real, measured data.

Polyethylene glycol (PEG) is a common polymer that is often used in validation experiments since it provides a mass spectrum with a distinctive peak pattern. Being a polymer, PEG is constructed of a series of repeating structures, and is available in different lengths, according to structural equation (6.8).

$$\mathrm{HO{-}CH_2{-}(CH_2{-}O{-}CH_2{-})_n{-}CH_2{-}OH}, \tag{6.8}$$

where n indicates the number of repetitions of the central structure.

Data was collected* in triplicate for a sample containing only PEG compounds PEG-0 to PEG-11, corresponding to $n = 0$ to 11 in equation (6.8). A 'blank' sample containing only the solvent and prepared in the same way as the PEG sample was also analysed in triplicate. Data was acquired in SIM windows of width 30 m/z over the range 70–540 m/z and combined using SIM-stitch, as described in Chapter 3. During this stage, peaks were required to have a SNR$\geq$ 3.5. PEG peaks with SNR$\geq$ 10 and present in the form of $[M+H]^+$, $[M+Na]^+$, $[M+{}^{39}K]^+$ or $[M+{}^{41}K]^+$ ions were used for internal calibration. In total, 61 peaks over the whole m/z range were used for internal calibration.

Following SIM-stitching, each set of three replicate spectra were combined using an $r = 2$ out of 3 filter, as described in Chapter 4. In order to remove background peaks occurring as a result of the sample preparation method, peaks were removed from the PEG spectra where a peak in close proximity also appeared in the blank spectrum. In order to reduce the possibility of sample carry-over causing peaks to falsely appear as background peaks, peaks were only removed from the PEG spectra where the corresponding intensity in the blank spectrum was at least 33% of that for the PEG sample.

Spectral peaks analysed were limited to the range 70–300 m/z so that the COP could be solved in reasonable time. From a total of 372 peaks after removing noise regions, 196 unique molecular formulae were identified. The PEG molecular formulae located are shown in Table 6.7. Reassuringly, PEG polymers PEG-0 to PEG-4, present within the m/z range, are detected. However PEG-5, while detected as a potential molecular formula with utility $U_m = 9$, is assigned zero abundance, i.e. is not included in the solution. This is the result of the falsely detected molecular formula $H_{20}C_{11}N_7O_2$, present as a protonated ion at mass 283.175124 Da having the same utility as PEG-5 of $U_m = 9$, present as a protonated ion at mass 283.175131 Da. A close inspection reveals that PEG-5 has slightly preferable molecular formula likelihood and isotope ratio patterns, however the quantisation effect of the discrete utility effects masks this difference. Additionally, were the mass spectrum to be extended beyond 300 m/z, an overall higher utility could be expected for PEG-5 as the $[M+Na]^+$ and $[M+K]^+$ adducts would be within the m/z range of the spectrum.

*Thanks to Ralf Weber for this data

PEG-n	Formula	Nominal	Number of Adduct Peaks				Abun-	Utility
		Mass (Da)	+H	+Na	$+^{39}$K	$+^{41}$K	dance	U_m
PEG-0	$H_6C_2O_2$	62			1		0.0003	4
PEG-1	$H_{10}C_4O_3$	106		1	1 (1)	1	0.0102	9
PEG-2	$H_{14}C_6O_4$	150	1 (1)	1 (1)	1 (1)	1 (1)	0.5566	11
PEG-3	$H_{18}C_8O_5$	194	1 (1)	1 (1)	1 (1)	1 (1)	1.0117	12
PEG-4	$H_{22}C_{10}O_6$	238	1 (1)	1 (1)	1 (2)	1 (1)	1.5985	12
PEG-5	$H_{26}C_{12}O_7$	282	* (*)					

Table 6.7: Identification of PEG compounds in control sample. The table shows the number of peaks for each different adduct type observed. For example, values in the column '+H' indicate the number of $[M+H]^+$ ions found. The number in parentheses indicates the number of isotope peaks, with (1) representing a single carbon-13 isotope and (2) a double carbon-13 isotope. An asterisk (*) indicates that the peak, while present, has been assigned to a different molecular formula.

The use of the mass spectrometer to generate the data from which the above results are drawn introduces a significant level of complexity to the spectrum. This includes noise sources previously discussed, such as chemical noise from the sample preparation, which although are minimised by acquiring and filtering peaks also present in a 'blank' spectrum, are still present as additional compounds in the spectrum. There is also a degree of variability in the measurement of the PEG compounds that the filtering process aims to reduce but that will also cause some 'real' peaks to be incorrectly filtered out. This is apparent from the low abundance found for PEG-1. However, despite these confounding factors, four of the nine identified molecular formulae with utility measure $U_m > 8$ are PEG peaks, including three of the four with $U_m > 10$. This demonstrates that the system is capable of achieving accurate results in terms of molecular formula and corresponding utility.

The results are expressed graphically in terms of U_m and abundance in Figure 6.18. The presence of large numbers of other molecular formulae in the figure emphasises the importance of the utility measure in placing a confidence score on the many assignments that are unavoidable in such a complex spectrum containing hundreds of peaks.

6.7.3 Measured Biological Data

Data collected using DI nESI FT-ICR MS analysis of cancer cell line samples [32] was used to validate the COP system as an approach for identifying molecular formulae of

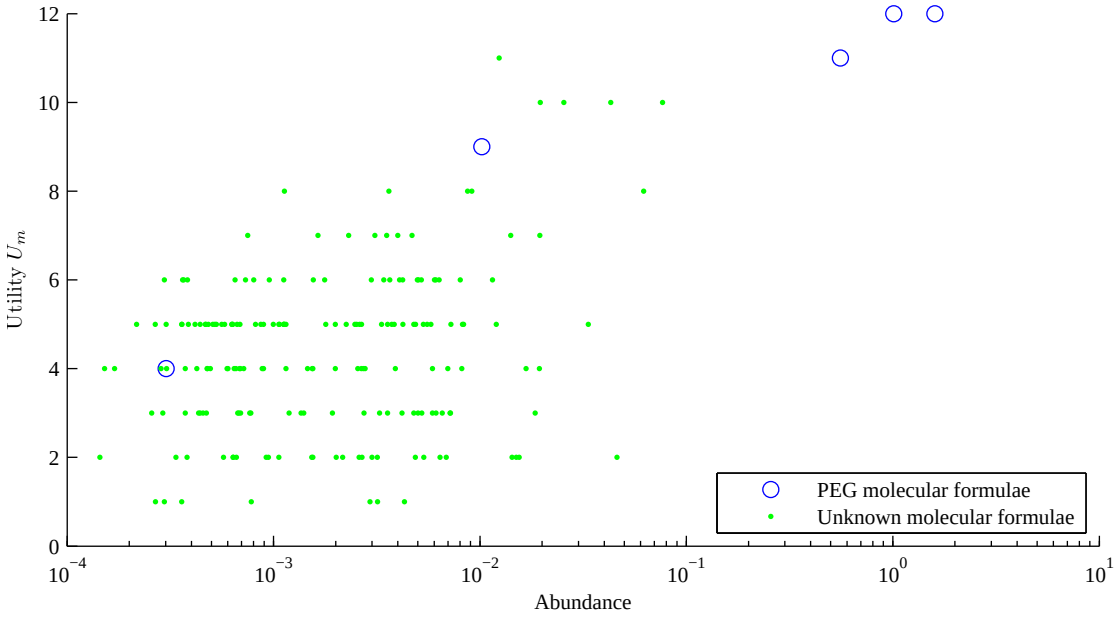

Figure 6.18: Utility and abundance of all molecular formulae identified as present in the spectrum of PEG sample.

complex biological samples. These samples form part of a study into the metabolic effect of treatment drugs on acute myeloid leukaemia cell lines. The significant challenge with analysing biological samples such as this arises from the complex, uncertain and noisy nature of the data, resulting in a lack of a 'golden' solution for comparison. The closest to such a standard is partially achieved by using NMR spectroscopy data. NMR is renowned for being a highly reproducible and quantifiable instrument capable of reliably detecting the presence of metabolites in a sample [17, 122]. As such, 18 biological samples of human K562 actue myeloid leukaemia cell extract were analysed using ^{1}H NMR as previously reported [32]. Five of the cell extract samples were also analysed more recently by our lab[†] using positive-mode FT-ICR MS, SIM-stitching [18] and filtering [21]. In the filtering stage, the following settings were used: $n = 3$ (triplicate) samples; $r = 2$ (out of three) replicate filtering; $s = 50\%$ sample filtering and peaks also present in the blank sample were removed providing their intensity in the blank was no greater than 33.3% of the sample intensity. A total of 1354 peaks remained, with 610 existing with mass $< 300m/z$. The sample was recently analysed in a similar fashion by Weber *et al.* during the comparison of a novel metabolite identification method [108], and in which 1947 peaks with mass < 300 Da were used. The data used here is based on a more recent FT-ICR MS measurement and data analysis, and a more conservative filtering method, although the cell extract sample is the same and thus valid comparisons can be made.

[†]Thanks to Ralf Weber for this data

Application of the constraints-based profiling method resulted in 96 peaks being removed, as no molecular formula could be assigned, leaving 514 peaks from which 378 molecular formulae were found, together with the utility measure and abundance of each. The distribution of the utility and abundance of the formulae identified is shown in Figure 6.19.

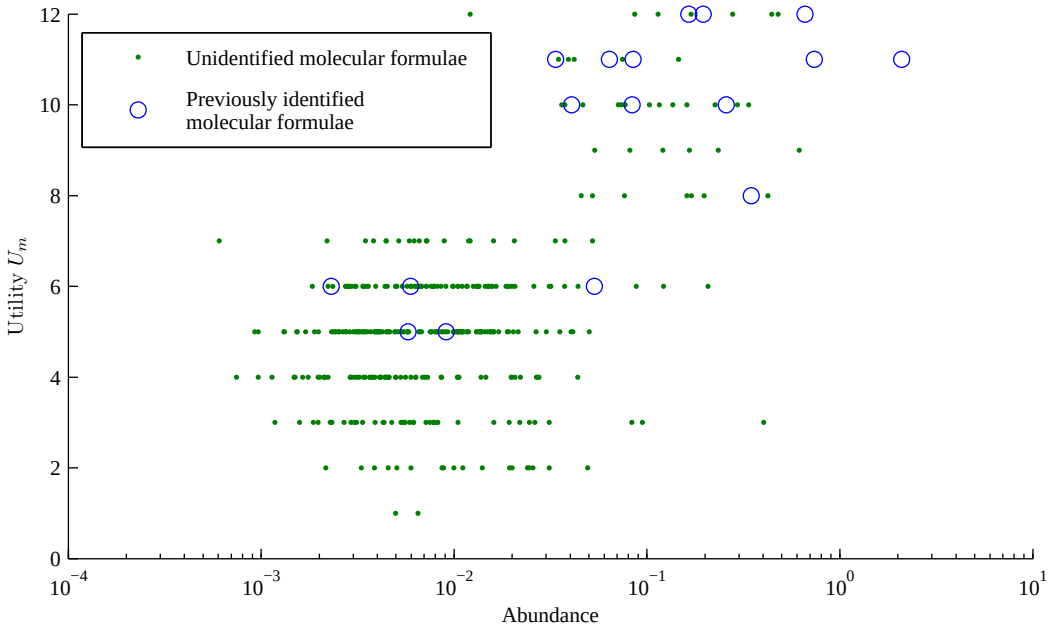

Figure 6.19: Utility and abundance of all molecular formulae identified as present in the biological sample. Molecular formulae previously identified in the sample are distinguished from those that are unknown.

The empirical formulae of 'known' compounds that have previously been found in similar data by Weber *et al.* [108] are indicated in the figure, and are listed in Table 6.8. Note that this work is concerned with the molecular formula identification: the compound identification is speculative and cannot be confirmed using DI FT-ICR MS alone, as discussed in Section 5.1.

Potential Compound Identification	Empirical Formula	Nominal Mass (Da)	Adduct (^{13}C Isotope) Peaks				Total Abundance	Utility U_m
			$[M+H]^+$	$[M+Na]^+$	$[M+^{39}K]^+$	$[M+^{41}K]^+$		
Glycine	$C_2H_5NO_2$	75	1				0.002	6
Alanine	$C_3H_7NO_2$	89	1 (1)				0.034	11
Lactate	$C_3H_6O_3$	90						
Proline	$C_5H_9NO_2$	115	1	1	1 (1)	1	0.737	11
Valine	$C_5H_{11}NO_2$	117	1	1	* (1)	1	0.196	12
Succinate	$C_4H_6O_4$	118						
Taurine	$C_2H_7NO_3S$	125		1	1 (1)	*	0.347	8
Creatine	$C_4H_9N_3O_2$	131	1 (1)	1 (1)	1 (1)	*	2.095	11
Leucine	$C_6H_{13}NO_2$	131	1	1	* (1)	1	0.165	12
Asparagine	$C_4H_8N_2O_3$	132		1	1	1	0.084	10
Aspartate	$C_4H_7NO_4$	133			1		0.006	6
Ethanolamine phosphate	$C_2H_8NO_4P$	141			*			
Glutamine	$C_5H_{10}N_2O_3$	146		*	1		0.006	5
Glutamate	$C_5H_9NO_4$	147	1 (1)	1 (1)	1 (1)	*	0.660	12
Methionine	$C_5H_{11}NO_2S$	149	1	1	1		0.053	6
Phenylalanine	$C_9H_{11}NO_2$	165	1 (1)	*	*		0.064	11
Glucose	$C_6H_{12}O_6$	180		* (*)	1 (1)	1	0.258	10
Tyrosine	$C_9H_{11}NO_3$	181	1 (1)	1	1		0.041	10
Citrate	$C_6H_8O_7$	192	*	1	*		0.009	5
Pantothenate	$C_9H_{17}NO_5$	219	1	1	1	1	0.085	11

Table 6.8: Compounds with mass < 300 Da likely to be present in cancer cell-line experiment data. An asterisk (*) indicates that the peak, while present, has been assigned to a different molecular formula.

Of the top 20 molecular formulae potentially present in the sample, as identified by Weber *et al.*, 17 were found. The three remaining formulae relate to the compounds lactate, succinate and ethanolamine phosphate. A manual check confirmed that of these, neither lactate nor succinate were present in the spectrum in $[M+H]^+$, $[M+Na]^+$ or $[M+K]^+$ forms. However, ethanolamine phosphate (empirical formula $C_2H_8NO_4P$ and mass 141.01910 Da) was found to be present in $[M+^{39}K]^+$ adduct form at mass 179.98227 Da. The reason that this adduct was not included in the solution is that the molecular formula $[C_4H_4N_4S]$, with mass 179.98226 Da has higher molecular likelihood, expressed as a greater U_{ml}, and so the peak was attributed instead to this unknown compound. If another adduct or isotope of ethanolamine phosphate were found, the overall utility associated with the compound would be increased and the peak would be correctly assigned.

Interestingly, the results reveal that isotopes are readily identified, however occasionally, the more abundant monoisotopic forms are not assigned to the solution. For example, in the case of valine, the $[M+^{39}K]^+$ adduct is identified only in the carbon-13 isotope form. The carbon-12 form is present, however the peak, located at 156.04214 m/z is instead assigned to the ion $[^1H_{11}\ ^{12}C_3\ ^{14}N\ ^{16}O_4\ ^{31}P]^+$, even though this ion has a mass error of 3.75 standard errors (equivalent to 0.62 ppm), compared with a mass error of 0.065 standard errors (0.01 ppm) for the true assignment. This assignment can be explained by observing that the unknown molecular formula $H_{10}C_3NO_4P$ has also been found in $[M+Na]^+$ and $[M+K]^+$ adduct forms, and with the additional $[M+H]^+$ form, the adduct pattern is in utility group A as apposed to group B, and therefore the overall utility is increased. Were the peak assigned instead to the molecular formula of valine, no improvement in the utility for that molecular formula which is already at maximum utility of 12, would be made, and consequently the overall mean utility U would decrease. These results demonstrate that the system works as designed, however as more analyses are made, the utility functions will likely require optimisation. For example in this case, since a good carbon-13 and potassium-41 isotope pattern is observed, the monoisotopic peak becomes unimportant in confirming the presence of valine and so it is assigned to a different formula that appears to deliver higher value. This could be resolved by introducing an increase in the overall utility as the number of isotope peaks associated with a compound is increased. This would result in an 'isotope-greedy' solution that would tend more to place peaks in isotope patterns rather than considering just the mean isotope pattern accuracy, as described in Section 6.5.2.

An additional observation is that certain adducts are not assigned to known molecular formulae, instead being attributed to unknown formulae. For example, while the $[M+Na]^+$ adduct ion for Glutamine, mass 169.05833 Da, is present in the spectrum as a peak at mass 169.05834 m/z, it is not assigned to Glutamine in the solution. Instead, the

peak is assigned to $[^{1}H_{10}\ ^{12}C\ ^{13}C\ ^{14}N_5\ ^{16}O_2\ ^{32}S]^+$, identified as the carbon-13 isotope to the $[M+H]^+$ adduct of $H_9C_2N_5O_2S$, an unknown compound not present in the KEGG database. This compound has an overall utility of 5 — the same as Glutamine. Were the peak assigned to Glutamine, the utility for the Glutamine molecular formula assignment would decrease to 4, since the adduct pattern would be group C instead of the more favourable B. This would have the effect of reducing the overall mean utility U — a consequence of the adduct pattern groupings and utility assignments, which in this case undesirably place higher value on a single $[M+K]^+$ peak being observed over both $[M+K]^+$ and $[M+Na]^+$ peaks. As new information regarding the formation of adducts becomes available, it is expected that such shortcomings would be removed and the solution found by the COP system match more closely the expert eye of a mass spectroscopist.

Overall, the results obtained by the COP solver are an excellent match to those expected to appear in the spectrum. Whereas typically, a 'target' list of metabolites is searched for within the spectrum, or specific biomarkers of interest are identified, the COP method returns a list of all molecular formulae found within the spectrum together with a score for each. As Figure 6.19 and Table 6.8 show, the scores for the 17 formulae previously identified in the sample range from 5 to 12. The majority of formulae have high utility (≥ 10), however several are assigned low utility values. This is largely because in each of these cases of low utility, the isotope pattern is not observed. The peak list used for this experiment was heavily filtered and consequently it is expected that many of the low intensity isotope peaks will not be present in the final spectrum. Indeed, only 610 peaks with mass below 300 m/z are used, compared to 1947 in the dataset used by Weber *et al.* in which 20 unique molecular formulae with mass $< 300 m/z$ are found. The lack of isotope pattern has a significant effect on the overall utility of the molecular formulae, largely due to the decision taken that a molecular formula with no isotope pattern present in the spectrum has zero isotope ratio pattern utility, i.e. $U_{ir} = 0$. It is therefore expected that a more complete mass spectrum would increase the overall utility for these compounds known to exist in the sample, reflecting the general truth that the more information provided to the system, the more accurate the solution.

6.7.4 Robustness to Noise

As previously discussed, the mass spectrum inevitably contains noise as well as signal. In order to understand the effect of this noise on this novel approach in falsely identifying metabolite empirical formulae in mass spectra, it is beneficial to observe the outcome when the algorithm is executed using a simulated spectrum consisting solely of noise, with no consistent information pertaining to metabolites. This information will give an indication of the false positive rate (FPR) of this method.

The spectrum simulated was designed to match the biological spectrum used in Section 6.7.3 in terms of the number of peaks (610) and the overall m/z range (70–300 m/z). The distribution of intensities was designed to match as closely as possible a realistic spectrum with dynamic range $1e6$, and intensity values were calculated using the formula

$$Y = 10^Z,$$

where Z is a set of uniformly distributed random real numbers with range $[-4, 2]$.

After pruning, there remained only 391 'valid' peaks in the spectrum and 768 potential molecular formulae. In this case, the partitioning process was particularly effective in simplifying the subsequent data mining since many of these formulae are, as expected, unrelated and consequently a large number of partitions (385 in total) were created. The execution time after this stage was negligible, a result of partitions falling into the category of being small, or containing straight-forward solutions that require minimal arbitration.

The search identified 389 unique molecular formulae. The most interesting result from the random data is the distribution of the utility values of the solution as shown in Figure 6.20. Such a trend is perhaps surprising, in that there are very few molecular formulae with low utility such as 1, until one recalls that the algorithm is designed to *maximise* the overall mean utility of the solution. To this extent, it could be expected that in a given mass spectrum, there will be large numbers of molecular formulae with low utility corresponding to random noise peaks as shown in Figure 6.20. The distribution of utility values for molecular formulae in a real biological sample is therefore expected to follow a similar pattern. Figure 6.21 shows the molecular formula utility assignments for the biological sample results described in Section 6.7.3. As expected, a large number of utility values exist in the range 1–7, which are expected to correspond to the random noise component of the spectrum. However, the distribution in Figure 6.21 is slightly offset to the right when compared to Figure 6.20. One should be careful to avoid over-analysing these observations since they are based on a single experimental run, however the shift of the distribution could be attributable to the presence of 'real' molecular formulae assignments which would be expected to be associated with a higher utility than the noise.

The observation of the second feature in Figure 6.21 from $U_m = 10$ is strongly suggestive of a significant number of molecular formulae assignments that are made with confidence well above that expected from a random spectrum. Within the random spectrum, none of the 389 molecular formulae had a non-zero U_{ir}, compared with 52 of the 378 (13.5%) in the biological spectrum, and it is this contribution to the overall utility that is necessary for utility values greater than 7.

With this in mind, and while comparing Figures 6.21 and 6.20, it seems reasonable to suppose that those assignments with utility above 7 are highly likely to be from non-random

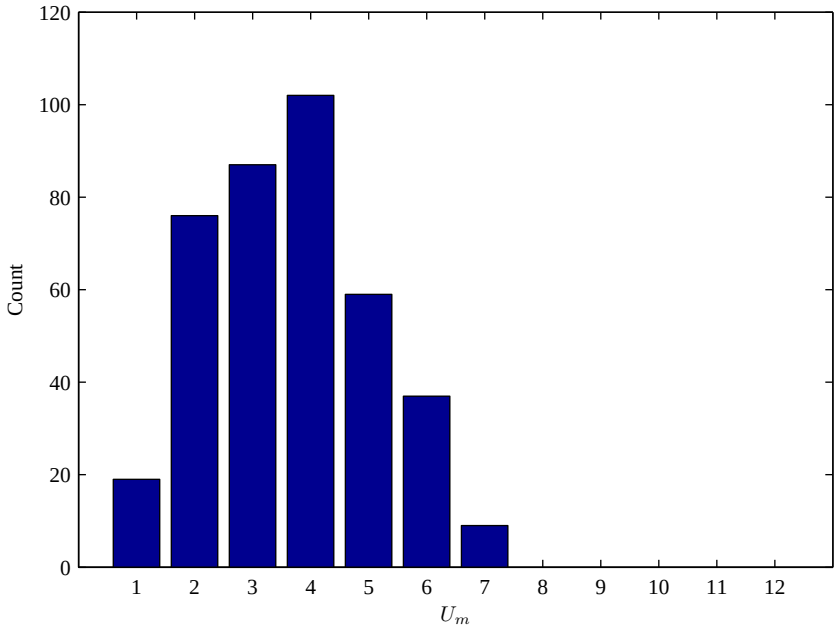

Figure 6.20: Histogram of utility values U_m for molecular formulae assignments, random data.

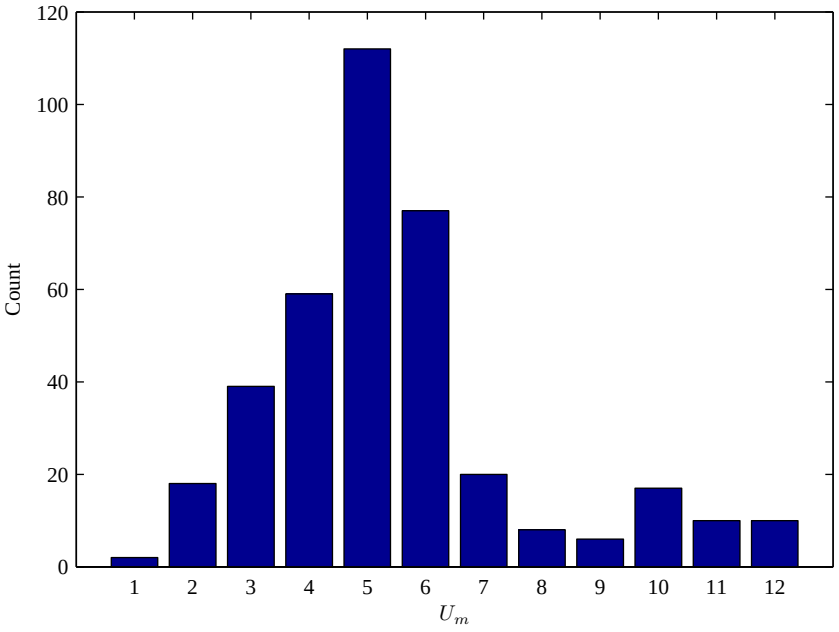

Figure 6.21: Histogram of utility values U_m for molecular formulae assignments, biological data.

features. Furthermore, the algorithm has found the most appropriate combination of peaks, that together yield the molecular formula assignment, assuming the utility functions are optimal. In the case of biological data, a total of 51 assignments meet this criteria.

This is an encouraging result: 12 of the 17 compounds (71%) known to be in the biological sample and identified by the algorithm are present with utility > 7. The low utility of the other 7 compounds (where no isotope pattern is found) is due to missing isotope peaks that are below the level of noise and hence discarded at the filtering stage. The filtering used in this example is relatively conservative in order to reduce the number of peaks input to the algorithm and hence maintain a feasible execution time. Further optimisation of the algorithm and implementation would help to increase those compounds with observed isotopes, and consequently increase the number of assignments with utility above 7.

6.8 Conclusions

This chapter has shown how constraints optimisation and a COP solver can be used to find the best solution to a mass spectrum in terms of molecular formulae present in the sample. The metrics of mass measurement error, isotope pattern accuracy, and the likelihood that molecular formulae exist in the sample are considered as continuous variables, although for practical reasons are discretised in implementation. Together with the adduct pattern, these four metrics are combined using utility theory into an overall 'utility' value for each molecular formula, that is a meaningful quantification of how well the formula is represented in the spectrum. The optimisation function maximises the average utility for all molecular formulae within the spectrum, and so determines the optimal solution to the spectrum.

In the case of a simulated spectrum, the COP solver correctly identified all the molecular formulae in the sample, and in the presence of noise proved robust. The utility measure for each molecular formula enabled the 'true' molecular formulae to be easily distinguished from the incorrect molecular formulae, assigned to noise peaks, or those not recognised by the solver.

For a real sample containing PEG compounds, the more complex spectrum resulted in many additional molecular formulae being identified, unrelated to PEG. Despite this, 5 of the 6 PEG compounds present in the spectrum were identified with high (> 8) utility in 4 of these cases. Both the PEG identified with low utility PEG-0, and the unidentified PEG compound PEG-5, are caused by missing peaks located beyond the m/z limits of the spectra acquired.

When applied to a real complex biological sample, the COP solver identified 17 of the 18 compounds known to be present in the spectrum, a true positive rate (TPR) of 94% for this subset of compounds. 12 of these compounds were identified with high utility (> 7), confirming the relevance of the utility to the quality of the molecular formula assignment.

A spectrum containing purely random noise reveals no molecular formulae assignments with utility above 7, and a pattern of low utility values that corresponds with that obtained during analysis of the biological sample. This shows that molecular formulae with high utility values are likely to be the result of real compounds present in the sample.

This work has shown how a COP solver together with a set of metrics can elucidate the molecular formulae of compounds present in an ultra high mass spectrum. As more knowledge is gained concerning the expression of compounds by FT-ICR MS, the concept of using continuous measures of metrics such as mass measurement error combined with a search method such as the COP solver promises to deliver consistent and quantitative profiling results.

6.9 Future Developments

The ultimate goal of metabolic profiling is to establish compounds rather than molecular formulae, and so an important next stage would be to include more biological information in the COP solver by means of a compound database. This would allow the output to be compounds rather than molecular formulae, although the work so far completed is still applicable since each compound is associated with a molecular formula.

The necessary discretisation of the metrics in this work does limit the performance of the COP solver in terms of the solution quality. Ideally, finer or even no discretisation would be necessary - this would ensure that the best solution is chosen where multiple options exist with similar overall utilities.

Refinement of the utility functions for the metrics would improve the solution. For example, the three adduct groups A, B and C do not always accurately reflect the likelihood of the adducts being present in the spectrum. The advantage of the COP solver however, is that such functions can easily be updated to reflect current knowledge, and the solution found will reflect the optimal solution, given this *a priori*.

The use of constraints methods for metabolic profiling has been applied here to positive mode MS. Applying the methods to negative mode MS is an extension that would yield the benefits of constraints methods, shown in this chapter, to both modes. Indeed, if spectra are acquired in both positive and negative mode, the constraints model could be expanded to include both spectra simultaneously. This would potentially further increase

the quality of the profiling, by providing yet more evidence to support each metabolite assignment.

Finally, the method described in this chapter could be further enhanced by providing alternative solutions. While a single solution offers the best optimisation of the constraints, it is heavily dependent upon accurately capturing the value of each metric during the application of utility theory. To help mitigate this, alternative, 'next best' solutions could be presented together with an overall utility for each solution. One way that this could be implemented is by iteratively re-running the COP search with the most recent optimal solution removed from the search space.

CHAPTER 7

CONCLUSIONS

The aim of this thesis was to address the challenges and shortfalls in current methods for the analysis of mass spectrometry-based metabolomics data. It has been shown how rigorous, methodical and novel methods have been applied to the problem in order to successfully achieve this goal.

Profiling the metabolome, in order to identify the metabolic profile of a sample, is a complex process. It consists of locating features within a dense, incomplete and noisy mass spectrum, and attributing those features to known metabolites. It is imperative that the mass spectra are of the highest quality before accurate profiling results can be obtained. FT-ICR MS produces high quality spectra, but with limitations in terms of the noise present in the spectra, and the quantity of ions that can be analysed simultaneously. Arguably, the most important features of the spectra are mass accuracy, sensitivity and noise. Therefore, the first two objectives of this thesis are to optimise the mass spectra in terms of these features, and within the limitations of the instrumentation. This is achieved through the development of novel approaches to the processing of the instrument's data. The first, SIM-stitching, increases the sensitivity and mass accuracy of the spectrum by intelligently 'stitching' together many smaller spectra into a single, wide spectrum. This is done in a manner that makes optimal use of internal calibrants, known substances within the sample, without compromising throughput. By allowing an optimal number of ions into the detection cell, mass accuracy is maintained, while at the same time increasing sensitivity over five-fold. The result is that more metabolites can be detected, and compounds can be detected more reliably. As a result of developing methods to meet this objective, unexpected noise was discovered in the mass spectra. The noise appears in the form of artefacts that occurred at consistent locations within the spectrum, and cannot be definitively attributed to a source. These regions were removed during the SIM-stitching process, an essential stage in avoiding spurious results and many false signal features that would otherwise occur.

Having optimised the sensitivity and mass accuracy of the spectrum, the second objective was to reduce the noise present in the spectrum. This was achieved through the use of a three-stage filter. In this approach, and in contrast to methods currently used, spectra from multiple samples are used to greatly improve the quantity and reliability of the features present in the mass spectrum. The filter was applied to both simulated and real, biological data. It is shown to yield significant benefits to the quality of the resultant spectrum. Additionally, the optimal number of scans necessary for reproducible results was found. Without knowing this, it is likely that analyses will either acquire insufficient scans for a consistent results, or will be longer than necessary. The first scenario will lead to significant degradation of the quality of the quantification data, while the latter results in unnecessary additional costs, which can be significant when using FT-ICR MS technology.

The work flows and tools that have been developed to meet the first two objectives are being used as a standard within Dr Mark Viant's laboratory at the University of Birmingham [81, 79, 16, 82, 83], and the national NERC Metabolomics Facility, also at the University of Birmingham. Both the SIM-stitch process and the three-stage filter have been integrated into a single tool, with a graphical user interface and user's manual, to assist scientists in using the methods developed. The tools have been made freely available, and have been requested by several key international laboratories. The SIM-stitch [18] and three-stage filtering methods [21] have, up to 22 April 2011, attracted 27 and 4 citations from non-involved authors, respectively.

The final objective of this thesis was to profile the metabolome, using the optimised mass spectra generated in the previous stages. A completely novel application of constraints optimisation methods and utility theory are combined to achieve this. FT-ICR mass spectra contain many derivatives of the compounds in the sample, including adducts, isotopes and fragments. As a result, the spectra are highly complex and densely populated. Historically, the interpretation of spectra was a lengthy manual process. Many computational methods currently used rely significantly on user intervention for decision making, and arbitration between multiple possible solutions. They are also often based upon a limited subset of spectral information. In this work, constraints optimisation methods have been developed that determine the combination of metabolites that best explain the observed spectrum. Four types of information are included in the solution search: adduct peaks, isotope peaks, mass accuracy and molecular likelihood. Using constraints optimisation enables easy addition of further information to the model, which will improve the quality of the result. For the first time, utility theory is used to combine these different types of information into a single score. This enables the experience of a mass spectroscopist to be captured and applied in a systematic manner. By applying utility theory and constraints

optimisation methods in this way, the optimal set of metabolites is found, based upon the mass spectrum. This approach is applied to both simulated and real data, and succeeds in both cases to correctly and robustly identify the molecular formulae of the compounds in the sample.

The application of constraints optimisation to metabolic profiling has been shown to offer benefits to the metabolomics community over some current techniques. One notable benefit is that the method is autonomous, requiring no user intervention beyond expressing preference for the utility of spectral features. By so doing, utility theory allows the mass spectroscopist to apply suitable weighting to each type of information in a methodical and consistent way. A further benefit is the easy integration into the constraints tool of additional information about the mass spectra. As more is understood about the mass spectrometer, and particularly electrospray ionisation, more patterns may become apparent in the spectra. Due to these benefits, the profiling part of the thesis is currently in preparation for publication under the working title 'Improved Interpretation of FT-ICR Mass Spectra using Utility Theory and Constraints Optimisation'.

Finally, this thesis has demonstrated a successful interdisciplinary collaboration between electronic, electrical and computer engineering, and the discipline of biological sciences. Measurements from biological organisms are no different from those obtained in other real-world scenarios, and so engineering principles should be applied. This thesis has adopted this attitude when processing FT-ICR mass spectra, to the benefit of many biological scientists.

7.1 Future Developments

Due to the availability of data, there are some consistencies in the type of data used during this work. For example, all data analysed is *positive mode* FT-ICR MS. Therefore, the SIM-stitch, filtering and profiling methods have all been developed based upon these conditions. However, there are benefits in using alternative configurations of the instrument, which can yield additional information. For example, negative mode FT-ICR MS results in a different set of adducts from positive mode, and can therefore allow compounds to be detected that do not readily ionise using positive mode. A future development of these methods could cater for the needs of negative mode ionisation, with relatively little modification. Additional value would be gained by including both positive and negative mode spectra in the profiling stage. In this case, the presence in both positive and negative mode spectra of peaks related to a compound increases the likelihood of that compound being present in the sample. This could be captured through the use of utility theory. A further example of a different instrumentation application is the use of chromatography

MS, which provides a second axis of information, and is therefore of benefit during the profiling stage.

During development of the SIM-stitch algorithm, certain characteristics of the SIM windows were discovered to be inconsistent. The process would benefit from a rigorous analysis of SIM windows, and allow any subtle variations to be characterised and corrected. One enhancement to the algorithm would be a data dependent variation of the SIM window position and width. A key factor is the number of ions in each SIM window, which should ideally be constant. Allowing the SIM windows to be moved would maximise the benefit obtained by SIM-stitching, regardless of the distribution and intensity of peaks in the spectrum. Currently, the SIM-stitch process is computationally intensive, and in large experiments may become a bottleneck in the processing. This is mainly due to the large number of Fourier transform operations required. The users would therefore benefit from optimisation of this process.

Classifying peaks in the spectrum as signal or noise is a difficult task. While a significantly improved three-stage filter has been proposed, common to this and other published methods is the binary classification of peaks as either signal or noise. Since in reality it is impossible to classify peaks with 100% accuracy, significant benefit would be gained by instead quantifying the likelihood that each peak is signal or noise. This better captures the reality of the noise filtering process, and the additional information could prove a useful input to the profiling stage. The measure of belief that a peak is signal would form part of the decision process, by being an additional constraint in the search.

Profiling the metabolome using constraints optimisation methods is a new area that offers many opportunities for future development. It has been shown how constraints can be used to identify metabolic formulae present in the sample. Metabolic profiling aims to identify compounds, i.e. molecular formulae and structures. Therefore, the next stage in developing the constraints methods would be to extend the search to compounds. This could be achieved by including biological information in the COP, such as a compound database. Since several compounds often share the same molecular formula, methods would need to be developed to infer structural information. The use of fragmentation or chromatography would potentially meet such a need. One current area of research interest is the potential relationship between the adduct patterns observed for a compound, and the structure of that compound. As any such relationships are established between adducts and compound, they can be used as further constraints in the profiling method presented here. In terms of implementation, the constraints methods presented here leave space for optimisation. The high computational cost of searching such a large space is reflected in the long program run times, which has made several approximations necessary, for example a limited number of elements. If the search strategy and implementation is

further optimised, these approximations can be relaxed. The end effect will be a higher quality result.

REFERENCES

[1] Marshall A, Hendrickson C, Jackson G. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: A Primer. Mass Spectrometry Reviews. 1998;17:1–35.

[2] Cech NB, Enke CG. Practical Implications of Some Recent Studies in Electrospray Ionization Fundamentals. Mass Spectrometry Reviews. 2001;20:362–387.

[3] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28:27–30.

[4] Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry. 2008;4:217–214.

[5] King KM, Rubin G. A history of diabetes: from antiquity to discovering insulin. British Journal of Nursing. 2003;12:1091–1095.

[6] Villas-Bôas SG, Mas S, Åkesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. Mass Spectrometry Reviews. 2005;24:613–646.

[7] Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. Trends in Biotechnology. 1998;16:373–378.

[8] Schmidt CW. Metabolomics: what's happening downstream of DNA. Environmental Health Perspectives. 2004;112:A410–A415.

[9] Dunn W, Ellis D. Metabolomics: Current Analytical Platforms and Methodologies. Trends in Analytical Chemistry. 2005;24:285–294.

[10] Roessner U, Bowne J. What is metabolomics all about? BioTechniques. 2009;46:363–365.

[11] Blow N. Biochemistry's new look. Nature. 2008;455:697–700.

[12] Fiehn O. Metabolomics the link between genotypes and phenotypes. Plant Molecular Biology. 2002;48:155–171.

[13] Adams A. Metabolomics: Small-Molecule 'Omics. The Scientist. 2003;17:38–40.

[14] Nicholson JK, Wilson ID. Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. Nature Reviews Drug Discovery. 2003;2:668–676.

[15] van Ravenzwaay B, Coelho-Palermo Cunha G, Leibold E, Looser R, Mellert W, Prokoudine A, et al. The use of metabolomics for the discovery of new biomarkers of effect. Toxicology Letters. 2007;172:21–28.

[16] Taylor NS, Weber RJM, Southam AD, Payne TG, Hrydziuszko O, Arvanitis TN, et al. A New Approach to Toxicity Testing in *Daphnia Magna*: Application of High Throughput FT-ICR Mass Spectrometry Metabolomics. Metabolomics. 2009;5:1573–3882.

[17] Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, Westler WM, et al. Method for Determining Molar Concentrations of Metabolites in Complex Solutions from Two-Dimensional ^{1}H-^{13}C NMR Spectra. Analytical Chemistry. 2007;79:9385–9390.

[18] Southam AD, Payne TG, Cooper HJ, Arvanitis TN, Viant MR. Dynamic Range and Mass Accuracy of Wide-Scan Direct Infusion Nanoelectrospray Fourier Transform Ion Cyclotron Resonance Mass Spectrometry-Based Metabolomics Increased by the Spectral Stitching Method. Analytical Chemistry. 2007;79:4595–4602.

[19] Payne TG, Southam AD, Viant MR, Cooper HJ, Arvanitis TN. The Rich Tapestry of Metabolomics: Stitching Spectra. In: Metabolomics Society's 3rd Annual International Conference, Manchester; 2007. p. 125–126.

[20] Payne TG, Southam AD, Viant MR, Cooper HJ, Arvanitis TN. An Integrated Calibration and Stitching Algorithm for Optimising Mass Accuracy and Dynamic Range in FT-ICR Mass Spectrometry Based Metabolomics. In: Metabomeeting 3; 2006. p. 21.

[21] Payne TG, Southam AD, Arvanitis TN, Viant MR. A Signal Filtering Method for Improved Quantification and Noise Discrimination in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry-Based Metabolomics Data. Journal of the American Society for Mass Spectrometry. 2009;20:1087–1095.

[22] Ledford EB, Dhaderi S, White RL, Spencer RB, Kulkarni PS, Wilkins CL, et al. Exact Mass Measurements by Fourier Transform Mass Spectrometry. Analytical Chemistry. 1980;52:463–468.

[23] Gross JH. Mass Spectrometry: A Textbook. Springer-Verlag; 2004.

[24] Griffiths WJ, Karu K, Hornshaw M, Woffendin G, Wang Y. Metabolomics and metabolite profiling: past heroes and future developments. European Journal of Mass Spectrometry. 2007;13:45–50.

[25] Stenson AC, Landing WM, Marshall AG, Cooper WT. Ionization and Fragmentation of Humic Substances in Electrospray Ionization Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry. Analytical Chemistry. 2002;74:4397–4409.

[26] McMurry J. Organic Chemistry — 3rd ed. Brooks/Cole Publishing Company; 1992.

[27] Baldeschwieler JD. Ion Cyclotron Resonance Spectroscopy. Science. 1968;159:263–273.

[28] Comisarow MB, Marshall AG. Fourier Transform Ion Cyclotron Resonance Spectroscopy. Chemical Physics Letters. 1974;25:282–283.

[29] Comisarow MB, Marshall AG. Frequency-Sweep Fourier Transform Ion Cyclotron Resonance Spectroscopy. Chemical Physics Letters. 1974;26:489–490.

[30] Marshall AG, Hendrickson CL. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. International Journal of Mass Spectrometry. 2002;215:59–75.

[31] Wang TC, Ricca TL, Marshall AG. Extension of Dynamic Range in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry via Stored Waveform Inverse Fourier Transform Excitation. Analytical Chemistry. 1986;58:2935–2938.

[32] Tiziani S, Lodi A, Khanim FL, Viant MR, Bunce CM, Günther UL. Metabolic Profiling of Drug Responses in Acute Myeloid Leukaemia Cell Lines. PLoS ONE. 2009;4:e4251.

[33] Bruins AP. Mechanistic Aspects of Electrospray Ionisation. Journal of Chromatography A. 1998;794:345–357.

[34] Kujawinski EB. Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (ESI FT-ICR MS): Characterization of Complex Environmental Mixtures. Environmental Forensics. 2002;3:207–216.

[35] Reemtsma T. Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: Status and needs. Journal of Chromatography A. 2009;1216:36873701.

[36] John B. Fenn - Autobiography [online]; 2010 [cited 2010 Feb 6]. Available from: http://nobelprize.org/nobel_prizes/chemistry/laureates/2002/fenn.html.

[37] Kebarle P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. Journal of Mass Spectrometry. 2000;35:804–817.

[38] Draper J, Enot DP, Parker D, Beckmann M, Snowdon S, Lin W, et al. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an inter-

active m/z annotation tool utilising predicted ionisation behaviour 'rules'. BMC Bioinformatics. 2009;10.

[39] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology. 2004;22:245–252.

[40] Hop CECA, Chen Y, Yu LJ. Uniformity of ionization response of structurally diverse analytes using a chip-based nanoelectrospray ionization source. Rapid Communications in Mass Spectrometry. 2005;19:3139–3142.

[41] Wang Y, Shi SDH, Hendrickson CL, Marshall AG. Mass-selective ion accumulation and fragmentation in a linear octopole ion trap external to a Fourier transform ion cyclotron resonance mass spectrometer. International Journal of Mass Spectrometry. 2000;198:113–120.

[42] Masselon C, Tolmachev AV, Anderson GA, Harkewicz R, Smith RD. Mass Measurement Errors Caused by Local Frequency Perturbations in FTICR Mass Spectrometry. Journal of the American Society for Mass Spectrometry. 2002;13:99–106.

[43] D Keith Williams J, Muddiman DC. Parts-Per-Billion Mass Measurement Accuracy Achieved through the Combination of Multiple Linear Regression and Automatic Gain Control in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer. Analytical Chemistry. 2007;79:5058–5063.

[44] Gordon EF, Muddiman DC. Impact of Ion Cloud Densities on the Measurement of Relative Ion Abundances in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: Experimental Observations of Coulombically Induced Cyclotron Radius Perturbations and Ion Cloud Dephasing Rates. Journal of Mass Spectrometry. 2001;36:195–203.

[45] Limbach PA, Grosshans PB, Marshall AG. Experimental Determination of the Number of Trapped Ions, Detection Limit, and Dynamic Range in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Analytical Chemistry. 1993;65:135–140.

[46] Dunn WB, Overy S, Quick WP. Evaluation of Automated Electrospray-TOF Mass Spectrometry for Metabolomic Fingerprinting of the Plant Metabolome. Metabolomics. 2005;1:137–148.

[47] Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2:171–196.

[48] Marshall AG. Theoretical Signal-to-Noise Ratio and Mass Resolution in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Analytical Chemistry. 1979;51:1710–1714.

[49] Kreyszig E. Advanced Engineering Mathematics, sixth edition. Wiley; 1988.

[50] Goto Y. Highly Accurate Frequency Interpolation of Apodized FFT Magnitude-Mode Spectra. Applied Spectroscopy. 1998;52:134–138.

[51] Burton RD, Matuszaka KP, Watsona CH, Eylera JR. Exact Mass Measurements Using a 7 Tesla Fourier transform ion cyclotron resonance mass spectrometer in a good laboratory practices-regulated environment. Journal of the American Society for Mass Spectrometry. 1999;10:1291–1297.

[52] Papoulis A. Probability, random variables and stochastic processes. 2nd ed. McGraw-Hill; 1984.

[53] Comisarow MB, Melka JD. Error Estimates for Finite Zero-Filling in Fourier Transform Spectrometry. Analytical Chemistry. 1979;51:2198–2207.

[54] Keefe CD, Comisarow MB. Exact Interpolation of Apodized, Magnitude-Mode Fourier Transform Spectra. Applied Spectroscopy. 1989;43:605–607.

[55] Umesh S, Tufts DW. Estimation of parameters of exponentially damped sinusoids using fast maximum likelihood estimation with application to NMR spectroscopy data. IEEE Transactions on Signal Processing. 1996;44:2245–2259.

[56] Reynolds G. Magnetic Resonance Spectroscopy Methods for Paediatric Brain Tumour Classification: PhD Thesis. School of Electronic, Electrical and Computer Engineering, The University of Birmingham, UK; 2008.

[57] Reynolds G, Wilson M, Peet A, Arvanitis TN. An algorithm for the automated quantitation of metabolites in *in Vitro* NMR Signals. Magnetic Resonance in Medicine. 2006;56:1211–1219.

[58] Keefe CD, Comisarow MB. A Family of Highly Accurate Interpolation Functions for Magnitude-Mode Fourier Transform Spectroscopy. Applied Spectroscopy. 1990;44:600–613.

[59] Liang Z, Marshall AG. Precise relative ion abundances from FT ICR magnitude-mode mass spectra. Analytical Chemistry. 1990;62:70–75.

[60] Goodner KL, Milgram E, Williams KR, Watson CH, Eyler JR. Quantitation of Ion Abundances in FT ICR MS. Journal of the American Society for Mass Spectrometry. 1998;9:1204–1212.

[61] Williams CP, Marshall AG. Hartley/Hilbert Transform Spectroscopy: Absorption-Mode Resolution with Magnitude-Mode Precision. Analytical Chemistry. 1992;64:916–923.

[62] Gudbjartsson H, Patz S. The Rician Distribution of Noisy MRI Data. Magnetic Resonance in Medicine. 1995;34:910–914.

[63] Price P. Standard Definitions of Terms Relating to Mass Spectrometry: a Report from the Committee on Measurements and Standards of the American Society for Mass Spectrometry. Journal of the American Society for Mass Spectrometry. 1991;2:336–348.

[64] Ledford EB, Rempel DL, Gross ML. Space charge effects in Fourier transform mass spectrometry — mass calibration. Analytical Chemistry. 1985;56:2744–2748.

[65] Zhang L, Rempel D, Pramanik BN, Gross ML. Accurate mass measurements by Fourier Transform mass spectrometry. Mass Spectrometry Reviews. 2005;24:286–309.

[66] Muddiman DC, Oberg AL. Statistical Evaluation of Internal and External Mass Calibration Laws Utilized in Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Analytical Chemistry. 2005;77:2406–2414.

[67] Wong RL, Amster IJ. Sub Part-Per-Million Mass Accuracy by Using Stepwise-External Calibartion in Fourier Transform Ion Cyclotron Mass Spectrometry. American Society for Mass Spectrometry. 2006;17:1681–1691.

[68] Scheltema RA, Kamleh A, Wildridge D, Ebikeme C, Watson DG, Barrett MP, et al. Increasing the mass accuracy of high-resolution LC-MS data using background ions a case study on the LTQ-Orbitrap. Proteomics. 2008;8:4647–4656.

[69] D Keith Williams J, Hawkridge AM, Muddiman DC. Sub Parts-Per-Million Mass Measurement Accuracy of Intact Proteins and Product Ions Achieved Using a Dual Electrospray Ionization Quadrupole Fourier Transform Ion Cyclotron Resonance Mass Spectrometer. Journal of the American Society for Mass Spectrometry. 2007;18:1–7.

[70] Tolmachev AV, Monroe ME, Jaitly N, Petyuk VA, Adkins JN, Smith RD. Mass Measurement Accuracy in Analyses of Highly Complex Mixtures Based Upon Multidimensional Recalibration. Analytical Chemistry. 2006;78:8374–8385.

[71] Yanofsky CM, Bell AW, Lesimple S, Morales F, Lam TT, Blakney GT, et al. Multicomponent Internal Recalibration of an LC-FTICR-MS Analysis Employing a Par-

tially Characterized Complex Peptide Mixture: Systematic and Random Errors. Analytical Chemistry. 2005;77:7246–7254.

[72] Finnigan^{TM} LTQ FT^{TM} Hardware Manual 1153760 Revision B; 2004.

[73] Finnigan^{TM} Xcalibur® Getting Productive: Qualitative Analysis XCALI-97101 Revision A; 2005.

[74] Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated Approach for Quantitative Analysis of Complex Peptide Mixtures from Tandem Mass Spectra. Nature Methods. 2004;1:39–45.

[75] Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP. *Ab initio* predicition of metabolic networks using Fourier transform mass spectrometry data. Metabolomics. 2006;2:155–164.

[76] Yeh C, Röbel A. Adaptive Noise Level Estimation. In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06), Montreal, Canada, September 18–20; 2006. p. 145–148.

[77] Gal O. A Collection of Fitting Functions [online]; 2004 [cited 2010 Sep 9]. Available from: http://www.mathworks.com/matlabcentral/fileexchange/4222-a-collection-of-fitting-functions.

[78] Personal communication with Roy Goodacre, Manchester; 2009.

[79] Hrydziuszko O, Silva M, Perera MTPR, Richards D, Murphy N, Mirza D, et al. Application of Metabolomics to Investigate the Process of Human Orthotopic Liver Transplantation: A Proof-of-Principle Study. OMICS — A Journal of Integrative Biology. 2010;14:143–150.

[80] Han J, Danell RM, Patel JR, Gumerov DR, Scarlett CO, Speir JP, et al. Towards High-Throughput Metabolomics Using Ultrahigh-Field Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Metabolomics. 2008;4:128–140.

[81] Hines A, Staff FJ, Widdows J, Compton RM, Falciani F, Viant MR. Discovery of Metabolic Signatures for Predicting Whole Organism Toxicology. Toxicological Sciences. 2010;115:369–378.

[82] Viant MR. Recent developments in environmental metabolomics. Molecular Biosystems. 2008;4:980–986.

[83] Wu H, Southam AD, Hines A, Viant MR. High-throughput tissue extraction protocol for NMR and MS-based metabolomics. Analytical Biochemistry. 2008;372:204–212.

[84] Schlosser A, Volkmer-Engert R. Volatile polydimethylcyclosiloxanes in the ambient aboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. Journal of Mass Spectrometry. 2003;38:523–525.

[85] Kaur P, O'Conner PB. Algorithms for Automatic Interpretation of High Resolution Mass Spectra. Journal of the American Society for Mass Spectrometry. 2006;17:459–468.

[86] Uechi GT, Dunbar RC. Space Charge Effects on Relative Peak Heights in Fourier Transform-Ion Cyclotron Resonance Spectra. Journal of the American Society for Mass Spectrometry. 1992;3:734–741.

[87] Frahm JL, Velez CMC, Muddiman DC. Understanding the Influence of Post-Excite Radius and Axial Confinement on Quantitative Proteomic Measurements Using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Rapid Communications in Mass Spectrometry. 2007;21:1196–1204.

[88] Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. Analytical Chemistry. 2006;78:567–574.

[89] Tong H, Bell D, Tabei K, Siegel MM. Automated Data Massaging, Interpretation, and E-Mailing Modules for High Throughput Open Access Mass Spectrometry. American Society for Mass Spectrometry. 1999;10:1174–1187.

[90] Stentiford G, Viant M, Ward D, Johnson P, Martin A, Wenbin W, et al. Liver Tumors in Wild Flatfish: A Histopathological, Proteomic, and Metabolomic Study. OMICS. 2005;9:281–299.

[91] Senko MW, Beu SC, McLafferty FW. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. Journal of the American Society for Mass Spectrometry. 1995;6:229–233.

[92] Gordon EF, Muddiman DC. Quantification of Singly Charged Biomolecules by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Utilizing an Internal Standard. Rapid Communications in Mass Spectrometry. 1999;13:164–171.

[93] Hughey CA, Galasso SA, Zumberge JE. Detailed Compositional Comparison of Acidic NSO Compounds in Biodegraded Reservoir and Surface Crude Oils by Negative Ion Electrospray Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Fuel. 2007;86:758–768.

[94] Smith C, Want E, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry. 2006;78:779–787.

[95] Sangster TP, Wingate JE, Burton L, Teichert F, Wilson ID. Investigation of Analytical Variation in Metabonomic Analysis Using Liquid Chromatography/Mass Spectrometry. Rapid Communications in Mass Spectrometry. 2007;21:2965–2970.

[96] Overy SA, Walker HJ, Malone S, Howard TP, Baxter CJ, Sweetlove LJ, et al. Application of Metabolite Profiling to the Identification of Traits in a Population of Tomato Introgression Lines. Journal of Experimental Botany. 2005;56:287–296.

[97] van der Burgt YEM, Taban IM, Konijnenburg M, Biskup M, Duursma MC, Heeren RMA, et al. Parallel Processing of Large Datasets from NanoLC-FTICR-MS Measurements. Journal of the American Society for Mass Spectrometry. 2007;18:152–161.

[98] Katajamaa M, Orešic M. Processing Methods for Differential Analysis of LC/MS Profile Data. BMC Bioinformatics. 2005;6.

[99] Andreev VP, Rejtar T, Chen HS, Moskovets EV, Ivanov AR, Karger BL. A Universal Denoising and Peak Picking Algorithm for LC-MS Based on Matched Filtration in the Chromatographic Time Domain. Analytical Chemistry. 2003;75:6314–6326.

[100] Pitman J. Probability. Springer-Verlag; 1993.

[101] Parsons HM, Ekman DR, Collette TW, Viant MR. Spectral Relative Standard Deviation: a Practical Benchmark in Metabolomics. The Analyst. 2009;134:478–485.

[102] Rogers S, Scheltema RA, Girolami M, Breitling R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics. 2009;25:512–518.

[103] Kind T, Fiehn O. Metabolomics database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics. 2006;7.

[104] Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics. 2007;8.

[105] Schug K, McNair HM. Adduct formation in electrospray ionization. Part 1: Common acidic pharmaceuticals. Journal of Separation Science. 2002;25:760–766.

[106] Schug K, McNair HM. Adduct formation in electrospray ionization mass spectrometry II. Benzoic acid derivatives. Journal of Chromatography A. 2003;985:531–539.

[107] Brown SC, Kruppa G, Dasseux JL. Metabolomics Applications of FT-ICR Mass Spectrometry. Mass Spectrometry Reviews. 2005;24:223–231.

[108] Weber RJM, Viant MR. MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. Chemometrics and Intelligent Laboratory Systems. 2010;104:75–82.

[109] Sleighter RL HP. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. Journal of Mass Spectrometry. 2007;42:559–574.

[110] Kujawinski EB, Behn MD. Automated Analysis of Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra of Natural Organic Matter. Analytical Chemistry. 2006;78:4363–4373.

[111] Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG. Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. Analytical Chemistry. 2001;73:4676–4681.

[112] Hertkorn N, Frommberger M, Witt M, Koch BP, Schmitt-Kopplin P, Perdue EM. Natural Organic Matter and the Event Horizon of Mass Spectrometry. Analytical Chemistry. 2008;80:8908–8919.

[113] Koch B, Dittmar T, Witt M, Kattner G. Fundamentals of Molecular Formula Assignment to Ultrahigh Resolution Mass Data of Natural Organic Matter. Analytical Chemistry. 2007;79:1758–1763.

[114] Russell S, Norvig P. Artificial Intelligence: A Modern Approach. Prentice Hall; 1995.

[115] Smith BM. Modelling. In: Rossi F, van Beek P, Walsh T, editors. Handbook of Constraint Programming. Elsevier; 2006. p. 377–406.

[116] McLachlan GJ. Mahalanobis distance. Resonance. 1999;4:20–26.

[117] von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton University Press; 1953.

[118] Moore PG, Thomas H. The Anatomy of Decisions. Penguin Business; 1988.

[119] Golomb SW, Baumert LD. Backtrack Programming. Journal of the Association of Computing Machinery. 1965;12:516–524.

[120] Schulte C, Carlsson M. Finite Domain Constraint Programming Systems. In: Rossi F, van Beek P, Walsh T, editors. Modelling. Elsevier; 2006. p. 495–526.

[121] Schulte C, Lagerkvist M, Tack G. Generic Constraint Development Environment [online]; 2010 [cited 2010 June 8]. Available from: http://www.gecode.org.

[122] Robertson DG. Metabonomics in Toxicology: A Review. Toxicological Sciences. 2005;2:809–822.