INCREASED CONFIDENCE OF METABOLITE IDENTIFICATION IN HIGH-RESOLUTION MASS SPECTRA USING PRIOR BIOLOGICAL AND CHEMICAL KNOWLEDGE-BASED APPROACHES

by

RALF JOHANNES MARIA WEBER

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

School of Biosciences
College of Life and Environmental Sciences
The University of Birmingham
March 2011

ABSTRACT

Mass spectrometry-based metabolomics aims to study endogenous, low molecular weight metabolites and can be used to examine a variety of biological systems. To substantially increase the accuracy of metabolite identification and increase coverage of the metabolome detected by high-resolution (HR) mass spectrometry I developed, optimised and/or employed several analytical and bioinformatics methods. Biological samples contain thousands of metabolites that are related through specific substrate-product transformations. This prior biological knowledge together with a mass error surface, which represents the mass accuracy of peak differences within mass spectra, were employed to significantly reduce the false positive rate of metabolite identification. To maximise the sensitivity of the Thermo LTQ FT Ultra mass spectrometer, the existing direct-infusion SIM-stitching acquisition parameters (Southam *et al.*, 2007) were re-optimised, yielding a ca. 3-fold increase in sensitivity. Finally, relative isotopic abundance measurements (RIA) using HR direct-infusion MS were characterised on the two most popular Fourier transform MS instruments (FT-ICR and Oribitrap) using the re-optimised SIM-stitching acquisition parameters. Several novel observations regarding RIA measurements were reported. Utilising these RIA characterisations within a putative metabolite identification pipeline increased the number of single true empirical formula assignments compared to using accurate mass alone. To conclude, analytical and bioinformatics methods developed in this thesis have successfully facilitated the putative identification of hundreds of metabolites in several metabolomics studies.

Dedicated to my parents, Maria and Jürgen Weber.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Prof. Mark Viant, for all his support, guidance and enthusiasm over the duration of my PhD. I offer my thanks to Drs. Stentiford, Lyons and Feist (Cefas) for the fish livers, Drs. Lodi and Tiziani for the NMR dataset, Dr. Piddock for the bacteria/media samples, Dr. Zampronio for help collecting LTQ Orbitrap Velos spectra and Dr. Khanim for a critical review of my thesis. I would additionally like to thank all members of the Viant research group and colleagues in the School of Biosciences and the Centre for Systems Biology for their help and support during my research. Also, I would like to thank my girlfriend and friends for their support and fun during my PhD, which has made the last three years so memorable. Special thanks go to my family for their patience and unconditional support which have helped me to reach the stage I am at today. Finally, I offer my thanks to Darwin Trust of Edinburgh for funding my PhD project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AGC	automatic gain control
ANOVA	analysis of variance
APCI	atmospheric pressure chemical ionisation
CI	chemical ionisation
CRM	charge residue model
DC	direct current
DFT	discrete Fourier transform
DI-MS	direct infusion-mass spectrometry
EI	electron impact
ESI	electrospray ionisation
FFT	fast Fourier transform
FID	free induction decay
FNR	false negative rate
FPR	false positive rate
FT-ICR	Fourier transform ion cyclotron resonance
FWHM	full width at half maximum
GC-MS	Gas chromatography-mass spectrometry
HMDB	Human Metabolome Database
HPLC	high-performance liquid chromatography
HR	high-resolution
Hz	Hertz
IEM	ion evaporation method
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	liquid chromatography-mass spectrometry
LTQ	linear trap quadrupole (Thermo Scientific product name)

M	neutral compounds
m/z	mass-to-charge ratio
MALDI	matrix assisted laser desorption/ionisation
MOA	mode of action
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MS^n	multiple-stage tandem mass spectrometry
MSI	Metabolomics Standard Initiative
NMR	nuclear magnetic resonance
OECD	Organisation for Economic Co-operation and Development
PCA	principal components analysis
PLS-DA	partial least squares-discriminant analysis
PLSR	partial least squares regression
ppm	parts per million
R	resolution
RF	radio-frequency
RIA	relative isotopic abundance
SIM	selected ion monitoring
SNR	signal-to-noise ratio
T	Tesla
TOF MS	time-of-flight mass spectrometry

CHAPTER ONE:

INTRODUCTION[1]

[1] Parts of this chapter, including Figure 1.2 - 1.4, are in preparation to be published in *Methods in Molecular Biology Series by Cambridge University Press* (Weber & Viant, 2010a).

1.1 Overview

Biology has become an increasingly data-rich subject. The accumulating amount of biological data harvested using high-throughput "omics" technologies (e.g. genomics, proteomics, transcriptomics and metabolomics) has necessitated the development of new and improved analytical and bioinformatics methods (Romero *et al.*, 2006). Metabolomics, one of the newest "omics" sciences, can be defined as the study of endogenous, low molecular weight molecules or metabolites and can be used to examine a variety of biological systems, from single cells to tissues, organs, or even whole organisms (Fiehn, 2002; Oliver *et al.*, 1998). Mass spectrometry (MS) has proven to be an indispensable analytical tool to measure and study metabolites (Junot *et al.*, 2010; Villas-Bôas *et al.*, 2005). Here I present several novel and improved analytical and bioinformatics methods to collect and annotate high-throughput high-resolution (HR) MS data.

This introduction is divided into eight sections describing various aspects of the field of metabolomics and the research subsequently presented within this thesis. First, Section 1.2 provides a general overview of metabolomics, related research fields and applications. Metabolomics uses non-targeted and targeted strategies to study the whole or specific parts of the metabolome (i.e. complete collection of metabolites); both approaches are described in Section 1.3. These approaches require a variety of platforms, techniques and methodologies, from which the advantages and disadvantages of the most popular analytical platforms (i.e. nuclear magnetic resonance (NMR) spectroscopy and MS) are detailed in Sections 1.4 and 1.5. In addition Section 1.6 highlights HR (direct infusion (DI)) FTMS, the focus of this thesis, as being one of the most appropriate platforms for sensitive, accurate and high-throughput analysis of metabolites. Section 1.7 describes and explains the principles of the most standard processing steps that are required to convert raw spectral data into accurate and precise HR Fourier transform (FT) mass spectra. The interpretation of these HR FT mass spectra is essential to study biological systems and to yield new biological knowledge. An overview of challenges, including examples, that are involved in the interpretation and annotation of high resolution mass spectra of biological origin are described in Section 1.8. Finally, the introduction concludes with the objectives of the research presented here.

1.2 Systems biology and Metabolomics

"The *whole* is greater than the sum of its *parts*" is one of Aristotle's (384-322 BC, Metaphysica) famous quotes. From a biology perspective it describes the principle that an entire biological system determines the behaviour of its separate components. This is now considered as the central dogma in multidimensional or systems biology (Kitano, 2002). The emergence of studying entire biological systems has been encouraged by the different "omics" tools (e.g. genomics, transcriptomics, proteomics and metabolomics) and the vast amount of associated high-throughput data (Romero *et al.*, 2006). Building comprehensive biological models/systems that take functional networks (e.g. cellular metabolism, signaling and gene expression) into account is a demanding challenge. This is mainly caused by the complexity of biological processes, which are not static and homogeneous in space, and therefore require dynamic and spatially resolved data regarding the cell types and environmental conditions.

Three different strategies are typically applied to reconstruct and study biological systems: (i) top-down (Brenner *et al.*, 2001; Bruggeman & Westerhoff, 2007), (ii) bottom-up (Brenner *et al.*, 2001; Bruggeman & Westerhoff, 2007) and (iii) middle-out (Brenner *et al.*, 2001). (i) The top-down approach identifies molecular interaction networks on the basis of correlated molecular behaviour between components of the entire system based on genome-wide "omics" data. The entire system is decomposed in a levelled manner, whereby each level or component, closer to the "bottom" of the system, provides more details until the most primitive level is reached. An iterative process is applied to measure correlations between these different levels or components, which potentially result in the discovery of new and essential links between molecular components that play a role in the behaviour of the system. This iterative process includes perturbation of the system, collection of high-throughput experimental data, analysis of experimental data, hypothesis testing and data integration. Whilst the top-down approach gives insights by decomposing the system from the top level, (ii) the bottom-up approach starts from the bottom and characterises the interactive behaviour (e.g. enzyme kinetics) of each subsystem or base element. These experimentally derived characterisations are eventually combined together to predict behaviour of the entire (sub)system. (iii) The middle-out approaches involve system reconstruction based on a level for which there is a good understanding of the experimental data and molecular processes. This level is then used to improve the understanding of the total system by linking this to knowledge and experimental data of higher and lower levels of the system.

Several studies that utilise one or a combination of the three strategies described above have been successfully carried out in the last few years (Kell, 2004), including reconstruction of metabolic networks (Förster *et al.*, 2003; Goelzer *et al.*, 2008; Herrgård *et al.*, 2008), deterministic (Srivastava *et al.*, 2002; Twycross *et al.*, 2010), stochastic (Srivastava *et al.*, 2002; Twycross *et al.*, 2010; Wilkinson, 2009) and constraint-based modeling (Mahadevan *et al.*, 2006; Sun *et al.*, 2009) and metabolic flux analysis (Kohlstedt *et al.*, 2010; Zamboni & Sauer, 2009).

Metabolomics (also referred to as metabonomics (Nicholson *et al.*, 1999)) (Fiehn, 2002), now a well-established scientific field, contributes to systems biology by providing an additional dimension of biological data. It enables the state of a biological system to be measured at a particular time within a certain environmental context which therefore reflects phenotypic changes (Brown *et al.*, 2005; Fiehn, 2002). More precisely, these metabolic snapshots represent quantitative and qualitative measurements of the intermediates and products of enzymatic reactions, i.e. metabolites, which are represented by lipids, amino acids, carbohydrates, hormones etc., providing more insights into upstream gene and protein activity (Griffin & Shockcor, 2004). Metabolites can be defined as endogenous or exogenous; where biologically catabolised metabolites in the cell or organism are defined as endogenous and the latter as metabolites that are consumed by organisms, such as food nutrients and pharmaceuticals (Dunn, 2008). The overall sizes of metabolic snapshots or metabolomes remain somewhat unknown. Several estimates have been proposed, ranging from many hundred metabolites in *yeast* (Herrgård *et al.*, 2008), several thousand in humans (Wishart *et al.*, 2009) up to a several hundred thousand in *plants* (Fiehn, 2002).

The concept of studying metabolite levels or metabolic profiles in biological systems has been around for several decades. Although the term “metabolomics” and “metabolome” were not used (Oliver *et al.*, 1998) until the late 1990s, the first studies regarding metabolic profiles were presented in the early 1970s (Horning & Horning, 1971; Pauling *et al.*, 1971). However, interest in the field has increased exponentially in the last ten years and since both terms were accepted by the biosciences community the number of publications per year, indexed by Web of Knowledge using the keywords “metabolomics or metabonomics” has increased from tens in 2002, through a few hundered in 2005, up to more than a thousand in 2010 (Figure 1.1) (Griffiths *et al.*, 2010). This increased popularity has arisen from the success of various “omics” methodologies/techniques and the realisation of the power of metabolomics to characterise phenotypes at the molecular level.

As mentioned previously, metabolomics is a key-player in systems biology. It has been applied concurrently in a growing number of fields including drug development (Wishart, 2008a); human health (Watkins & German, 2002); disease diagnosis (Ellis *et al.*, 2007; Kaddurah-Daouk *et al.*, 2008); environmental science and toxicology (Aliferis & Chrysayi-Tokousbalides, 2010; Bundy *et al.*, 2009); and nutrition and food science (Astle *et al.*, 2007; Wishart, 2008b).

In general the aim of all fields in relation to metabolomics is to characterise biological indicators (i.e. biomarkers) or profiles that are involved in general biological and pathogenic processes, or pharmacologic responses to drugs treatments. These can then be used to define molecular mechanisms and, monitor organisms or environmental systems. Numerous applications have been reported that show the benefit of metabolomics to characterise biomarkers and study molecular mechanisms in microbial (Bradley *et al.*, 2009), plant (Tarpley *et al.*, 2005), environmental (Taylor *et al.*, 2010; Taylor *et al.*, 2009) and mammalian (Lewis *et al.*, 2008) systems.

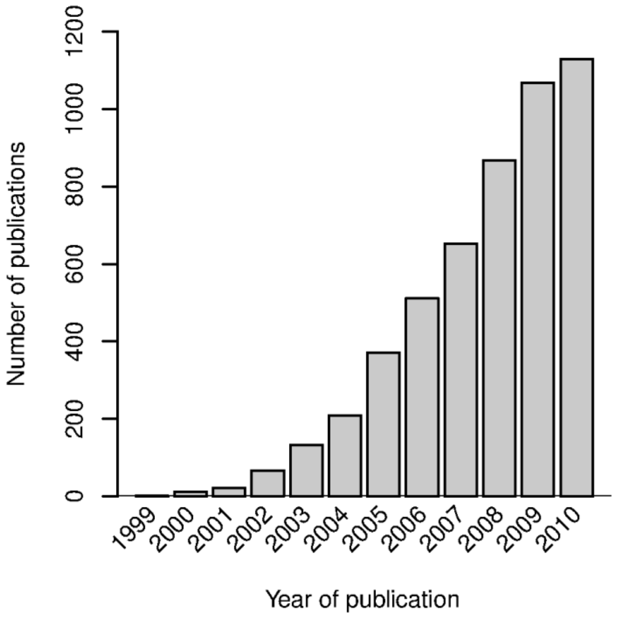


Figure 1.1 Number of publications in each year indexed by Web of Knowledge using the keywords "metabolomics or metabonomics" during the period 1999 - 2010.

1.3 Overview of metabolomics strategies

Metabolomics strategies attempt to achieve the precise and accurate quantification and identification of all metabolites, also referred to as non-target analysis (Dunn *et al.*, 2005). Unfortunately, this unbiased strategy is limited by the sensitivity and specificity of available analytical techniques and data processing methods which, so far, result in incomplete coverage of the metabolome. Furthermore, the number of available bioinformatics tools and databases for metabolite identification is small and incomplete (Aliferis & Chrysayi-Tokousbalides, 2010; Brown *et al.*, 2009), which results in measurements of signal that are unknown. The work presented in this thesis describes novel and improved methods for metabolite identification (in HR FT mass spectra) and therefore contributes to this non-targeted approach.

Nevertheless, a more global non-targeted approach has been widely used in metabolomics to overcome the challenge of annotating all observed signals, called metabolic fingerprinting (Dunn *et al.*, 2005; Ellis *et al.*, 2007). This less comprehensive but often high throughput analysis uses signal profiles of spectra to screen, examine and compare samples of different biological status or origin. These non-target signal profiles represent multivariate data (i.e. *m/z* and associated intensities) and so metabolic fingerprinting analyses typically necessitate unsupervised and supervised multivariate statistical methods (Boccard *et al.*, 2010). Principal components analysis (PCA), an unbiased multivariate statistical method that reduces the dimensionality of a data while retaining most of the variation, is often used to assess similarities and differences between MS samples (i.e. phenotypes) and determine whether samples can be grouped

(Ringner, 2008). Partial least squares discriminant analysis (PLS-DA) and Partial least squares regression (PLSR) are examples of supervised multivariate statistical methods, and used for the classification of samples and regression analysis, respectively (Boccard *et al.*, 2010). Cross-validation and permutation testing are typically included in such supervised statistical analysis (or models) to avoid over-fitting of the data. Although metabolomics is by nature multivariate, it is possible to utilise univariate analyses, such as analysis of variance (ANOVA) and t-test. A false discovery type correction is required to avoid incorrect significance level calculations for this type of multiple comparison analysis which involves large datasets (Benjamini & Hochberg, 1995).

Targeted metabolite analysis involves identification and quantification of limited preselected compounds (i.e. part of a predefined group of compounds (e.g. lipids) or specific chosen metabolic pathway) (Griffiths *et al.*, 2010). There are several platforms and techniques that suit targeted analysis, such as liquid chromatography (LC) MS and tandem MS (MS/MS) (see next section for more details). The last non-targeted approach to mention is the global measurement of extracellular metabolites (e.g. in cell culture media) called metabolic footprinting and is similar to metabolic fingerprinting (Mapelli *et al.*, 2008).

1.4 Analytical platforms to perform metabolomics

A variety of platforms using different analytical methodologies have been developed in order to cover the requirements of the different metabolomics strategies (see Section 1.3). An ideal platform to measure metabolite levels can be described as follows (Lenz & Wilson, 2007): able to perform direct sample analysis, i.e. without the need for sample preparation; high-throughput; unbiased with respect to metabolite class; both highly and equally sensitive to all compounds in the sample; robust and reproducible with a wide dynamic range; and all these criteria should be combined with enough information to allow the identification of key metabolites. Currently, none of the available platforms can supply all of these properties, which results in a trade-off between technologies and objectives (see Figure 1.2).

Currently, NMR and MS are the most common analytical techniques used to measure metabolite levels (Dunn *et al.*, 2005; Junot *et al.*, 2010; Viant *et al.*, 2003; Weljie *et al.*, 2006). Whilst NMR is known as a robust, reproducible and quantitative technique (Keun *et al.*, 2002; Lewis *et al.*, 2007; Weljie *et al.*, 2006), widely accepted as a successful platform in metabolomics, MS-based metabolomics is gaining popularity since it provides significantly greater sensitivity (Dettmer *et al.*, 2007). Nevertheless, MS does suffer from higher analytical variation, and is considered to be less quantitative in comparison to NMR.

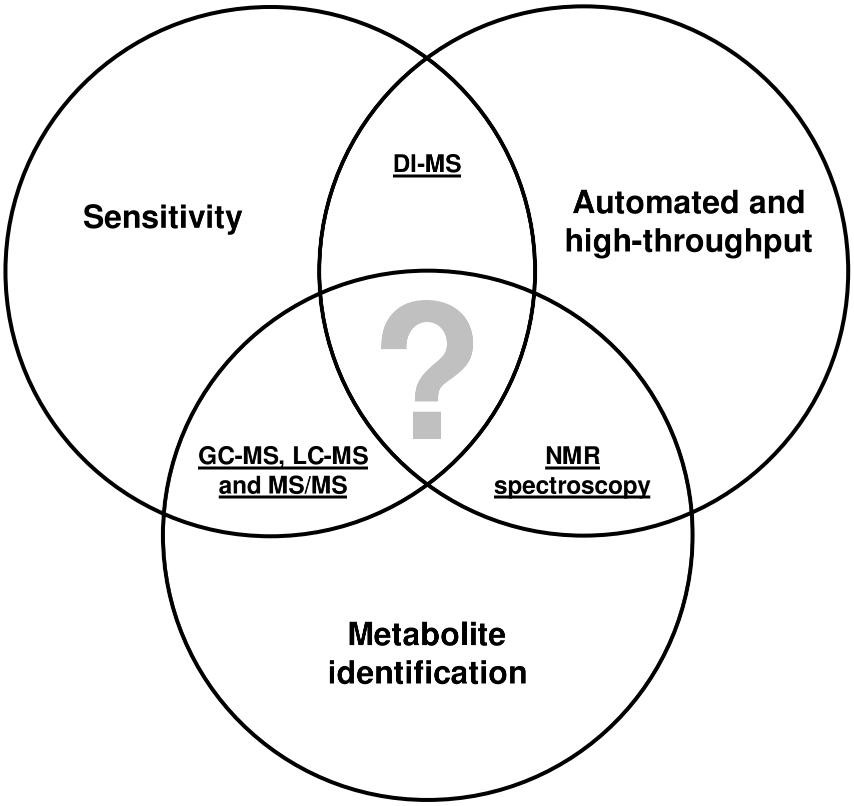


Figure 1.2 The trade-off between analytical platforms (underlined) and some of the objectives of metabolomics analysis, adapted from Dunn *et al.* (2005). DI-MS - direct infusion mass spectrometry; NMR – nuclear magnetic resonance spectroscopy; LC-MS - Liquid chromatography-mass spectrometry; GC-MS – gas- chromatography mass spectrometry; MS/MS – tandem mass spectrometry.

Briefly, NMR uses a magnetic field to spin active nuclei (e.g. ^{1}H, ^{13}C, ^{31}P and ^{15}N). Their nuclear energy states become quantised when this take place. Transitions from low to high energy states may be induced by the absorption of energy in the form of incident electromagnetic radiation. When a photon of the correct energy strikes such a nucleus, the resonance condition is met and a transition is induced. Different nuclei within the same molecule may experience different magnetic field strengths due to differences in the local electronic environment. This results in differences in their resonant frequencies and is referred to as the chemical shift (Williams & Fleming., 1995). In pulsed-Fourier transform NMR (FT-NMR) (Williams & Fleming., 1995), many frequencies are excited at once through the application of a short and intense burst of radio frequency radiation. Following this initial excitation, the nuclei slowly return to their low energy states, emitting the previously absorbed energy. This release of energy is measured as a time domain free induction decay (FID) emission signal (Williams & Fleming., 1995). This time domain signal is converted into a frequency domain NMR spectrum through Fourier transformation, which deconstructs the waveform into the individual frequencies of the nuclei in the sample. To date ^{1}H NMR is the most common NMR technique used in metabolomics. This is for two reasons: (i) ^{1}H nuclei are present in almost all metabolites making this approach non-biased and; (ii) ^{1}H nuclei have the highest relative sensitivity of all naturally occurring spin-active nuclei, leading to the highest signal intensity. A typical ^{1}H NMR spectrum presents a profile of peaks where each peak corresponds to a hydrogen nucleus in a particular chemical functional group of a particular metabolite. The intensity of each peak reflects the abundance of the type of hydrogen atom giving rise to it. As previously mentioned, a metabolite with multiple functional groups containing ^{1}H nuclei will exhibit a series of peaks in the NMR spectrum. Spectra can increase in complexity when a mixure of metabolites is measured (e.g. real biological sample) and therefore prior metabolite structural knowledge (reference databases) can help to identify signals from within the NMR spectrum that presents a mixture of metabolites.

MS is introduced in more detail below as it has been used throughout the present work. Although the field of metabolomics provides several successful approaches to analyse metabolites, it suffers from a lack of direct comparability and integration as data is measured by multiple platforms with different protocols.

1.5 Mass spectrometry

MS involves the ionisation of neutral compounds (M) from within a sample, which are then accelerated into the mass analyser, separated according to their mass to charge ratio (*m/z*) and detected to produce a mass spectrum (intensity *versus m/z*). To achieve this, MS instruments consist of three components (Figure 1.3): (i) ion source, (ii) mass analyser and (iii) detector. Additionally, a computer system is required to control the mass spectrometer and, process and store the data that is collected.

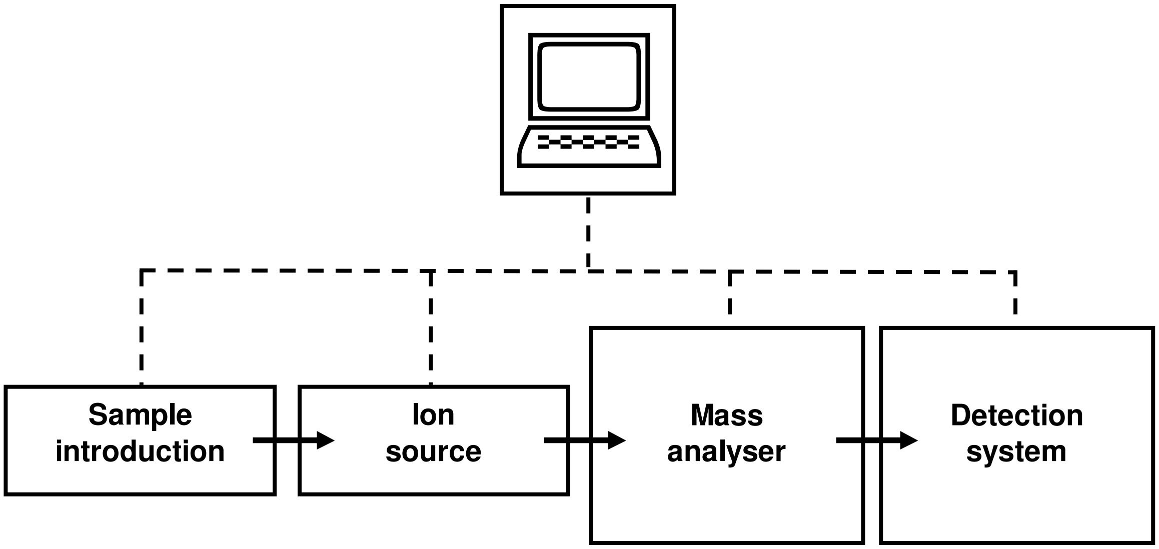


Figure 1.3 Schematic block diagram of a typical mass spectrometer operation, adapted from (Dunn, 2008**).** Samples are introduced via an inlet system to an ion source were positively or negatively charged ions are created. Ions are subsequently separated according to their *m/z* ratio in the mass analyser prior to detection either physically at a detector or as an image current of ion orbital frequencies. A computer system is used to control all instrument parameters and, process and store collected data.

Prior to ionisation, several separation techniques such as gas chromatography (GC) (Dunn, 2008), liquid chromatography (LC) (Dunn, 2008) and capillary electrophoresis (CE) (Monton & Soga, 2007) are often used to reduce ionisation suppression (Annesley, 2003) and improve quantification and identification (Dunn *et al.*, 2005; Griffiths *et al.*, 2010; Villas-Bôas *et al.*, 2005). Separation techniques transform a mixture of compounds into distinct groups that differ by chemical and physical properties. Separation is typically presented as retention time, i.e. the time before a compound elutes from the chromatograph to enter the (i) ion source.

Only compounds that are sufficiently volatile and thermally stable are suitable for GC separation (Bedair & Sumner, 2008; Dunn, 2008). Therefore sample preparation is often required, prior to GC, to make metabolites, such as sugars and amino acids, volatile and thermally stable. It is likely that this type of sample preparation introduces variability in the data. GC separates the mixture of compounds, after the solution vaporises, regarding their relative interaction with the coating of the separation column and the carrier gas (e.g. helium), respectively named stationary phase and mobile phase (Dunn, 2008). Although, GC is one of the widely accepted separation techniques in mass spectrometry, LC, including high performance (HP)LC (Wilson *et al.*, 2005), is increasingly more popular. A mixture dissolved in the mobile phase (e.g. water, acetonitrile, methanol and ethanol) enters a column under high pressure and separation of the compounds occurs based on the structure of the compounds and the specifications of the column (e.g. stationary phase, dimension size, pore size, column dimension) (Dunn, 2008; Pitt, 2009). Also, here the retention time is measured, which in this case is dependent on the pressure used, the nature of the stationary phase, solvent composition and column temperature. CE is a separation technique that is less often used in MS-based metabolomics. However, several successful applications have been presented recently (Ramautar *et al.*, 2011). This technique separates compounds based on charge and size (Ramautar *et al.*, 2011; Villas-Bôas *et al.*, 2005). The benefits of separation techniques are at the expense of longer analysis time, making this less cost effective particularly in terms of high-throughput MS. GC and LC, most commonly used in mass spectrometry-based metabolomics, are typically followed by different types of ion sources as both techniques require different environments to ionise compounds prior to detection (Dunn, 2008). GC is typically followed by electron ionisation (EI) or chemical ionisation (CI) to ionise compounds in a gas-phase environment. In short, EI uses electron beams to cause ionisation by electron ejection from the analyte or by analyte decomposition. Alternatively CI uses a reagent gas (e.g. ionised methane using EI) and high pressure conditions to transfer a proton from the reagent to the analyte. Electrospray ionisation (ESI) and atmospheric-pressure chemical ionisation (APCI) are the most common ionisation techniques applied to ionise compounds after LC analysis prior to detection. APCI is similar in principal to CI, but allows the ionisation of liquid samples as the sample is heated and vaporised before ionisation (Dunn, 2008). Finally, matrix-assisted laser desorption/ionization (MALDI) can be used with LC, and is often coupled with a time-of-flight TOF MS instruments (Macha & Limbach, 2002). MALDI is based on the bombardment of compounds with a laser light to initiation ionisation. The sample is mixed with a matrix solution and applied to a sample support/plate, and then dried to form a crystallised surface (Macha & Limbach, 2002). The matrix transforms the energy of the laser light into excitation energy for the sample, which leads to the formation of ions on the surface (Macha &

Limbach, 2002). Ionisation techniques, as mentioned above, are mainly divisible into two types: soft ionisation, which keeps the majority of compounds intact during ionisation (e.g. ESI and MALDI) and hard ionisation techniques that produce ions and/or fragments (e.g. EI, CI and APCI).

ESI without prior separation techniques, which allows high-throughput MS, has been used throughout the work presented here (Villas-Bôas *et al.*, 2005). As mentioned previously, DI ESI suffers from ionisation suppression. Ionisation suppression (i.e. compounds that compete to ionise) is highly dependent on the structures of the compounds in the sample undergoing ionisation (Annesley, 2003). For example, compounds with a greater proton affinity may cause ionisation suppression of less polar compounds, which affects the amount of ions of each type of compound that reaches the detector (Annesley, 2003). Other components, such as salts, buffers and other additives, may also influence this process (Annesley, 2003). Several analytical techniques different from separation techniques have been presented to minimise this, such as optimised sample preparation (Buhrman *et al.*, 1996) and improved ESI techniques (e.g. nano-ESI) (Annesley, 2003; Hop *et al.*, 2005). ESI uses electrical energy to transfer positive or negative ions (charged molecules) from the sample into the gas-phase prior to MS analysis. Aside from the existing ions in solution several other ion products may be formed. Several phenomena occur during ionisation including fragmentation, evaporation, collisions and ionic interactions (Bruins, 1998). Although, the exact process of ionisation is somewhat unclear it involves three main steps (Bruins, 1998). First, a spray comprising of charged droplets is formed, by creating potential difference between a needle and a capillary in a nebulising gas (e.g. nitrogen) environment. Secondly, charged droplets or ions (charge residue model or the ion evaporation method (Kebarle, 2000)) undergo evaporation which results in smaller ion clouds or separated ions. Finally, the highly charged ion clouds and single ions enter the instrument (capillary) and evaporate further where possible. As with most other ionisation techniques, this ionisation process is highly dependent on the structure of the compound(s) and the composition of the sample. Ideally, both ionisation modes (i.e. positive and negative) should be used to maximise metabolome coverage. Ions occurring within the sample or formed during ionisation comprise of the following molecule structures:

- Molecular ion – i.e. [M]±;
- (de-)protonated ion – i.e. $[M \pm H]^{\pm}$;
- Adducts – e.g. $[M + Na]^{+}$, $[M + K]^{+}$, $[M + Cl]^{-}$ and $[M + CHCOO]^{-}$;
- Fragments - the neutral structure may fragment into multiple smaller molecules which subsequently ionise in the aforementioned ways.

Once the compounds are ionised a (ii) mass analyser is used to determine the *m/z* values of all gas-phase ions. A variety of MS analysers have been employed in the field of metabolomics, including TOF MS (Kind *et al.*, 2009; Timischl *et al.*, 2008), quadrupole mass filters/ion traps (Kind *et al.*, 2009; Koulman *et al.*, 2007; March, 1997), ion cyclotron resonance (ICR) MS (Marshall *et al.*, 1998) and Orbitrap (Hu *et al.*, 2005).

To improve control and performance (e.g. sensitivity, selectively filtering ions of a particular *m/z* value or region and fragmentation) of MS metabolomics analysis several combinations of MS analysers have been employed, including triple-quadrupole MS (QqQ – two quadrupole mass analysers separated by a quadrupole collision cell (Stolker *et al.*, 2004)) and a linear ion trap with either an FT-ICR analyser or an Orbitrap analyser (Kiefer *et al.*, 2008; Taylor *et al.*, 2009).

Time-of-flight MS works on the principle that ions of different *m/z* with the same kinetic energy travel at different speeds (i.e. the lower the mass the higher the speed) (Brown & Lennon, 1995). During TOF analysis, ions are accelerated using an electric field and the time taken for them to reach a (iii) detector in a field-free region is measured, which is directly proportional to their *m/z* (Brown & Lennon, 1995).

The quadrupole mass analyzer acts as an ion filter and is similar to the quadrupole ion trap (March, 1997). The quadrupole ion trap is designed to pass the ions rather than collect the ions. The filtering of the quadrupole mass analyser involves the passage of a beam of ions through the centre of four parallel rods into a detector system (March, 1997). Radio frequency (RF) voltages are applied to the rods, causing certain ions to repel from the quadrupole and therefore only the ions that are not repelled will continue their stable trajectory to detector system.

The quadrupole ion trap can be used on the front end of HR FT mass spectrometers (e.g. FT-ICR and Orbitrap), allowing accumulation of certain ions before their passage to the detector cell (Douglas *et al.*, 2005; Hager, 2002). It uses specific RF waveforms to the electrode to create a 2D or 3D field to trap ions (Douglas *et al.*, 2005; Hager, 2002). The 3D ion trap utilises a circular electrode with two ellipsoid caps on the top and bottom to create a 3D field (Douglas *et al.*, 2005). The linear trap consists of four parallel electrodes evenly spaced to create a cylindrical ion trapping region with open ends (Hager, 2002). The small size and enclosed shape of 3D traps makes ion trapping relatively inefficient in comparison to the linear ion trap. The process of the linear ion trap involves the transmission of beams of ions through the centre of four parallel rods. Specific RF voltages are applied to the rods, whilst static DC potentials are applied to block the ends of the trap. All ions stored in the trapping field are ejected together for detection.

HR FTMS, a type of MS that uses electrostatic and/or magnetic forces to detect metabolites, is introduced in more detail below (section 1.6) as it has been used throughout the present work. HR FT mass spectrometers are usually combined with a linear ion trap to filter and accumulate ions prior to detection.

1.6 HR FTMS

HR FTMS has been accepted as the most advanced type of mass spectrometry to measure charged molecules or ions, represented by FT-ICR MS (Marshall *et al.*, 1998) and Orbitrap technologies (Hu *et al.*, 2005). HR FTMS instruments are usually equipped with an ionisation source and linear ion trap to ionise the sample and regulate the number of ions that enter the ICR cell for ion detection (Section 1.5). In addition, the linear ion trap is often used to select certain ions of interest for further and more extensive analysis (e.g. fragmentation analysis, such as tandem and multiple-stage tandem mass spectrometry (MS^n)) (Wang *et al.*, 2000). Regulating the number of ions that enter the ICR cell for ion detection is important to maximise sensitivity while maintaining high mass accuracy (Senko *et al.*, 1997; Soule *et al.*, 2010; Southam *et al.*, 2007). The linear ion trap uses a control loop, named automatic gain control (AGC), to maintain the ion injection time to collect a predefined number of ions set by the user (i.e. AGC target) for each scan.

FT-ICR MS uses a magnetic field, up to 14.5 T, and electrostatic forces to trap ions in the ICR detector cell. The trapped ions, which undergo cyclotron motion, are not detectable at this stage. The radius of the cyclotron motion is too small and the motion of the ions is not in phase. Therefore RF forces are applied to the field to increase the cyclotron radius of the ions and to let them move coherently in phase (i.e. low *m/z* ions having a higher cyclotron frequency than high *m/z* ions and ions of the same *m/z* have the same cyclotron frequency) (Hu *et al.*, 2005). The net charge of packages of ions that pass close to two electrode detector plates (opposite each other) is recorded as a function of time (i.e. free induction decay (FID)). Fourier transform is finally used to deconvolute the signals into their constituent frequencies, which can then be converted into *m/z* since ion frequency is directly proportional to *m/z*.

The Orbitrap utilises a detector cell comprising an outer “barrel-shaped” electrode and a central spindle electrode. Electrostatic forces are applied to the spindles, which traps the ions in the detector cell and causes them to orbit around and oscillate along the central spindle (Hu *et al.*, 2005). Similarly to the FT-ICR, the ion oscillation is recorded by detector plates, which is then converted to *m/z*.

The detection systems used by the FT-ICR and Orbitrap analysers allow higher mass accuracy and resolution compared to more traditional TOF MS and quadrupole mass analysers. Unlike the Orbitrap, the FT-ICR possesses a super-conducting magnet, which makes the system physically bigger and more expensive with higher running costs than the Orbitrap. However, FT-ICR MS has the highest mass accuracy (i.e. ≤1 parts per million (ppm)) and resolution (i.e. ≤ 1,000 K, full width at half maximum (FWHM)) of all commercially available mass spectrometers.

Analytical methods which improve the performance of mass spectrometers, such as sensitivity and mass accuracy, are essential to maximise the detection of the metabolome and assist in metabolite identification. Therefore, Southam *et al.* (2007) have published a spectral stitching strategy which collects a series of adjacent selected ion monitoring (SIM) windows using DI ESI FT-ICR MS, increasing the sensitivity and dynamic range of the analysis while maintaining high mass accuracy. For example, consider a low abundance ion, X. If a wide scan range (containing X) is used, only a relatively small number of X ions will be present (assuming the total number of ions is fixed by the automatic gain control (AGC), yielding a low signal-to-noise ratio (SNR). However, for a narrow scan range, X has a much higher relative abundance and this increases the SNR. Concurrently, space-charge effects are kept constant throughout, so mass accuracy is not reduced (Southam *et al.*, 2007). To yield a single mass spectrum, these SIM windows are aligned and then "stitched" together. The more recent Thermo Scientific LTQ FT Ultra has a larger trap volume in its ICR detector cell compared to previous systems, which allows a larger radius of ion cyclotron motion and considerably reduces space-charge effects meaning it is possible to increase sensitivity without compromising mass accuracy. Therefore, a re-optimisation of the SIM-stitching acquisition parameters was carried out as part of the present work (see Chapter 4).

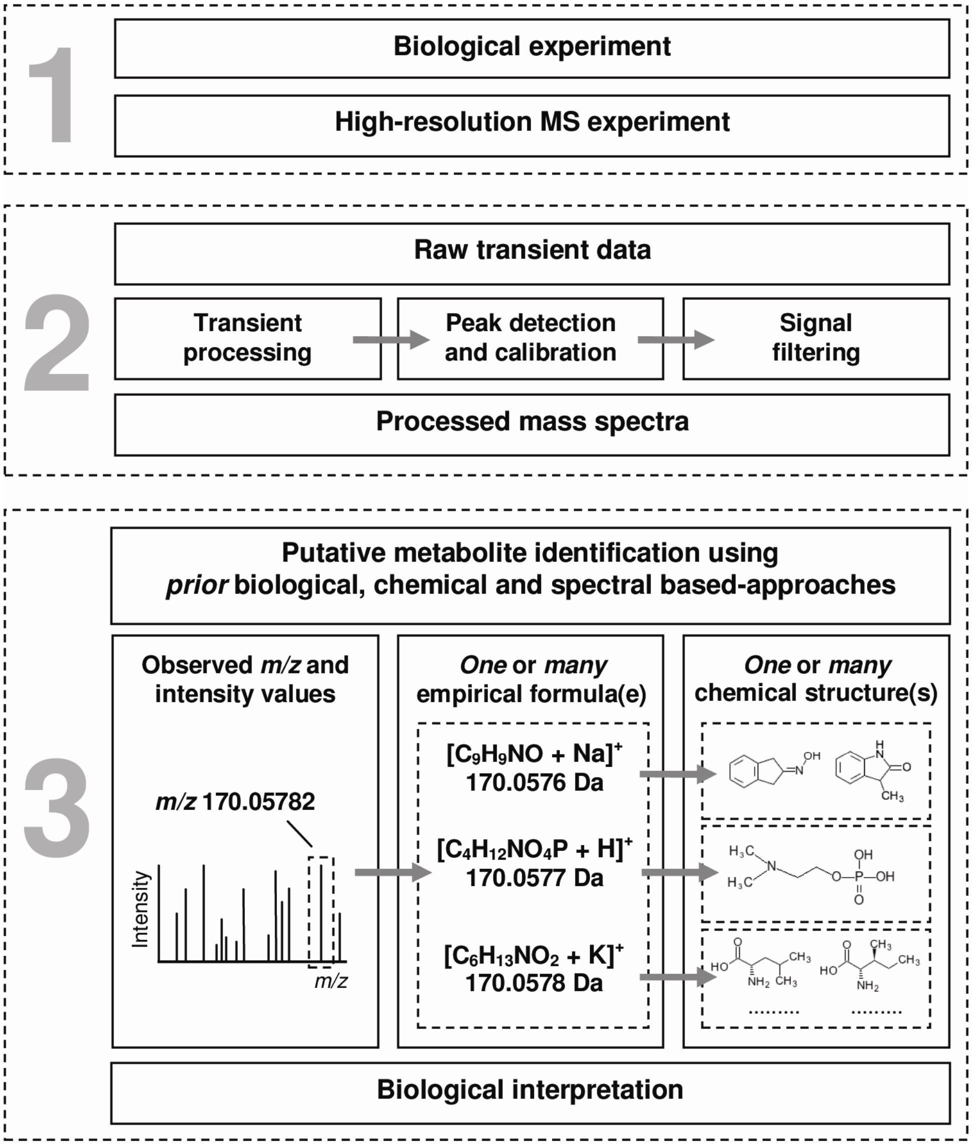

1
Biological experiment
High-resolution MS experiment
2
Raw transient data
Transient processing
Peak detection and calibration
Signal filtering
Processed mass spectra
3
Putative metabolite identification using
prior biological, chemical and spectral based-approaches
Observed *m/z* and intensity values
One or *many* empirical formula(e)
One or *many* chemical structure(s)
m/z 170.05782
Intensity
m/z
[C_9H_9NO + Na]$^+$
170.0576 Da
[$C_4H_{12}NO_4P$ + H]$^+$
170.0577 Da
[$C_6H_{13}NO_2$ + K]$^+$
170.0578 Da
OH
N
H
O
CH_3
H_3C
H_3C
OH
O—P—OH
O
HO
NH_2
CH_3
....
....
Biological interpretation

Figure 1.4 Schematic representation of the workflow for (1) collection, (2) processing and (3) identification of signals in HR FT mass spectra.

1.7 Data handling of HR FT mass spectrometry data

Raw data recorded using HR FTMS consists of multiple pre-processed files and associated transient (i.e. time-domain) files. Each pre-processed file, usually produced by a commercial mass spectrometer, corresponds to a specific sample and includes detailed information regarding the MS experiment that has been performed (e.g. acquisition parameters, external calibration and mass spectral data). Several commercial, e.g. Xcalibur (Thermo Scientific), and open-source software packages, e.g. MZmine 2 (Pluskal *et al.*, 2010), XCMS (Smith *et al.*, 2006) and Midas (Freitas *et al.*, 2003), are available to the user to process the pre-processed files or to convert them into more general formats, e.g. mzXML (Lin *et al.*, 2005) or a peak list (i.e. *m/z* values and associated intensities). Unfortunately, the majority of software packages do not allow the user to process transient data, which limits the flexibility and control over spectral processing and potentially results in less accurate and sensitive mass spectra. Although the available protocols to process transient data can vary, the main steps are normally indistinguishable, and are as follows: transient averaging; Fourier transform; peak detection and calibration (see Figure 1.4 and Figure 1.5).

The summation of transients (see Figure 1.5a) in the time domain provides a gain in the SNR. Since the noise in a transient signal is normally distributed, a reduction of √N times the noise level is achieved, where N is the number of averaged transients. Next, apodisation is required to reduce the discontinuities (known as Gibb's oscillations) at the beginning and end of a frequency-domain peak. This is a result of the abrupt start and end of the transient signal. Applying the Hanning window function reduces Gibb's oscillations at the expense of overall intensity and resolution of the signal (see

Figure 1.5b and Figure 1.5d). Although, the Hanning window function is the most commonly used one in HR FTMS, several others exist and the trade-offs between them have been reported previously (Brenna & Creasy, 1989).

Prior to Fourier transform zero-filling can be applied to increase the total number of data points in the processed spectrum, i.e. $2^{m}N$ where m = number of times zero-filling is applied and N = the original number of data points in the transient (see Figure 1.5c and d) (Comisarow & Melka, 1979). It results in a beneficial smoothing of the signal, hence improving the ability to locate peaks. Finally, the time domain data is transformed into frequency domain data using a discrete fast Fourier transform (FFT) (see Figure 1.5d).

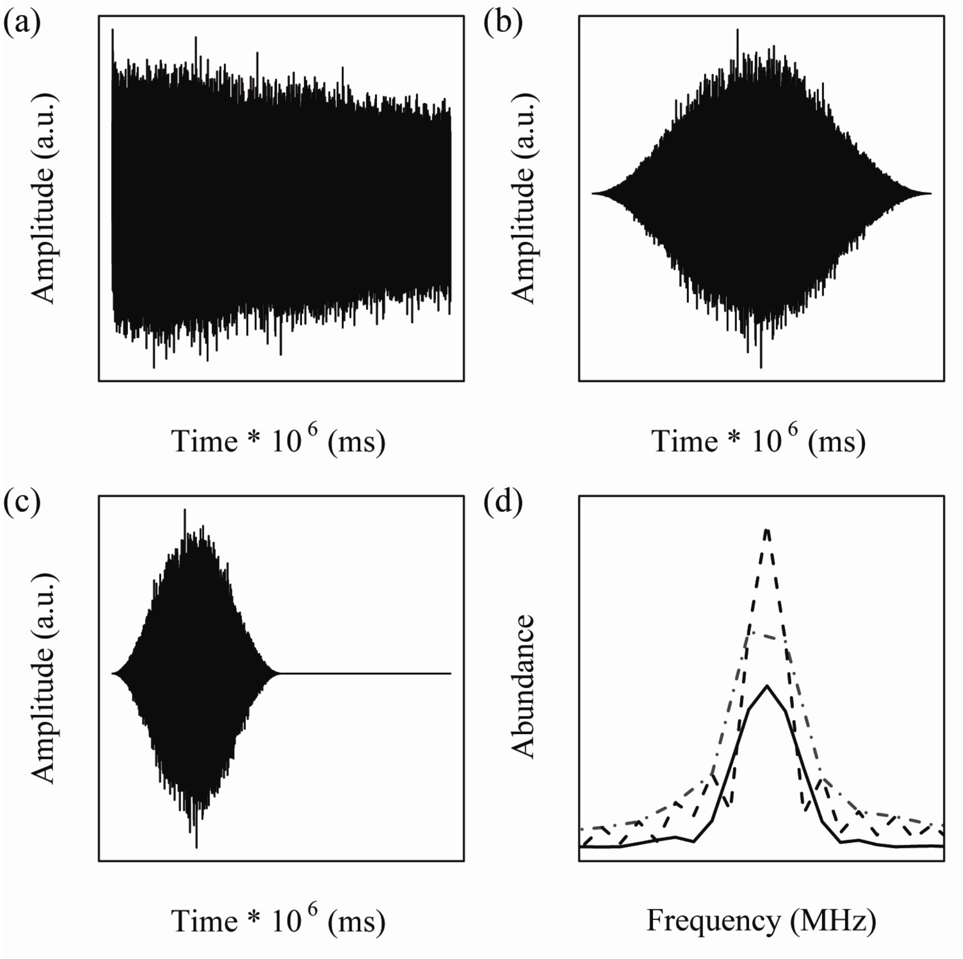


Figure 1.5 Representation of several processing steps for a particular SIM scan (*m/z* 280-380): (a) transient summation, **(b)** Hanning apodization, and **(c)** zero-filling. Panel **(d)** represents the effect of several processing steps on a single peak after Fourier transform, specifically: dotdash (- — - —): without applying Hanning function or zero-filling; dash (— — —): after zero filling only, and solid (———): after applying Hanning function and zero filling.

High mass accuracy and precision are essential for metabolite identification and, therefore, calibration is an important step for achieving the lowest possible mass error. External and internal calibration are achieved by mapping frequency domain peaks (i.e. in Hz) to *m/z* domain, using a calibration equation with (i) pre-defined calibration parameters (derived from instrument calibration) or (ii) a list of defined calibrants with known exact masses that occur in the mass spectrum, respectively. A variety of calibration equations that are available (e.g. $m/z = A/f + B/(f^2)$, A = ion charge density and trap geometry, B = magnetic field strength, f = frequency) have been extensively discussed by Zhang *et al.* (2005). Peak-detection, or peak-picking, required for calibration can be divided into two steps. Firstly, an accurate estimation of noise in the spectrum is required to avoid false peak detection, and secondly either peak height or area is measured (both providing information on relative quantification) to distinguish real peaks from the estimated noise level. HR FT mass spectra include many different types of noise, including technical and biological. A signal-free region is required from which the noise level can be measured (Limbach *et al.*, 1993). Peaks are usually allocated using constraints, shape criteria or fitting-models (e.g. local maxima, centroid and interpolation).

Despite the processing described above, a proportion of the “peak list” generated will still include noise. For the work presented here, a three-stage signal filter was used to discriminate authentic metabolite signals from this noise (Payne *et al.*, 2009). This approach requires that each sample is analysed in triplicate. Then the first stage of filtering comprises of a hard SNR threshold (typically >3.5) for peak picking, followed by a “replicate filter” and finally a “sample filter”. The replicate filter requires that a signal deemed “real” is present in at least 2 out of the 3 replicate measurements. The sample filter then requires that this signal is present in at least 50% typically of all samples measured in a given study. In addition, a fourth “filter” can be applied, whereby peaks observed in the solvent blank spectra are also removed from the dataset.

Prior to statistical analysis, as introduced in Section 1.3, the intensities of the mass spectral data are normalised, e.g. central tendency, linear regression, and locally weighted regression, to remove systematic *bias* while preserving biological information, and also pre-treated to deal with missing values (Callister *et al.*, 2006).

1.8 Automated metabolite identification of HR FT mass spectra

Processed HR FT mass spectra, as described above, yield complicated profiles typically comprised of thousands of peaks arising from: molecular ions (i.e. $[M - e]^+$); (de-)protonated ions (i.e. $[M{\pm}H]^{\pm}$); adducts (e.g. $[M + Na]^+$); naturally occurring isotopes (e.g. $^{13}C^{12}C_{n-1}$) and fragments of metabolites (e.g. from in-source fragmentation). Overall, the identification of signals can be divided into non-structural (i.e. identification of an empirical (or chemical) formula, $C_cH_hN_nO_oP_pS_s$, Figure 1.4) and structural assignments (i.e. leading to a unique metabolite and name, Figure 1.4). Compounded by the finite mass accuracy of HR FTMS, a single peak is typically assigned multiple empirical formulae (particularly for ions >300 *m/z*), which in turn, can represent numerous naturally-occurring structural isomers. Without additional information, such as that provided by chromatographic retention time and/or by fragmentation patterns from tandem mass spectrometry and MS^n, the definitive identification of metabolic peaks is impossible (Sumner *et al.*, 2007). Accordingly, it is well-established that a high number of false positive peak assignments occur during metabolite identification (Matsuda *et al.*, 2009). Furthermore, although there has been considerable progress in the development and availability of bioinformatic tools and databases (e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2008), Human Metabolome Database (HMDB) (Wishart *et al.*, 2009), KNApSAcK (Shinbo *et al.*, 2006), BioCyc (Caspi *et al.*, 2010)) for visualisation and identification (Aliferis & Chrysayi-Tokousbalides, 2010; Brown *et al.*, 2009), they remain incomplete and therefore unidentified signals in mass spectra are commonplace.

To link mass spectral data to metabolic content, spectral peaks must first be assigned to single, correct empirical formulae using the so-called Diophantine equations (See Equation 1) (Junot *et al.*, 2010). Although, HR FTMS can measure highly mass accurate spectra, only in certain cases can a single empirical formula be assigned correctly to a low mass value (typically <200 *m/z*), as shown in Figure 1.6. Increasing the mass accuracy increases the likelihood of assigning the correct empirical formula, however, definitive assignment using mass measurements alone is practically unachievable due to technical limitations of available mass spectrometers. Nevertheless, internal calibration can improve the mass accuracy significantly, for example, up to a root-mean-square mass error of 0.18 ppm (Southam *et al.*, 2007).

$$\text{Mass}^{\text{positive ion}}\ (m/z) \pm \text{mass error boundary}\ (m/z) = [12.0000{*}^{12}\text{C}_a + 1.0078{*}^{1}\text{H}_b + 14.0031{*}^{14}\text{N}_c + 15.9949{*}^{16}\text{O}_d + 30.9738{*}^{31}\text{P}_e + 31.9721{*}^{32}\text{S}_f + {}^{23}\text{Na}{*}22.9898] - \text{electron}{*}0.0005$$

Equation 1 A typical Diophantine equation to calculate empirical formulae (containing carbon, hydrogen, nitrogen, oxygen, potassium, sulfur and sodium) for a sodium adduct $[M + Na]^+$ in a particular mass range.

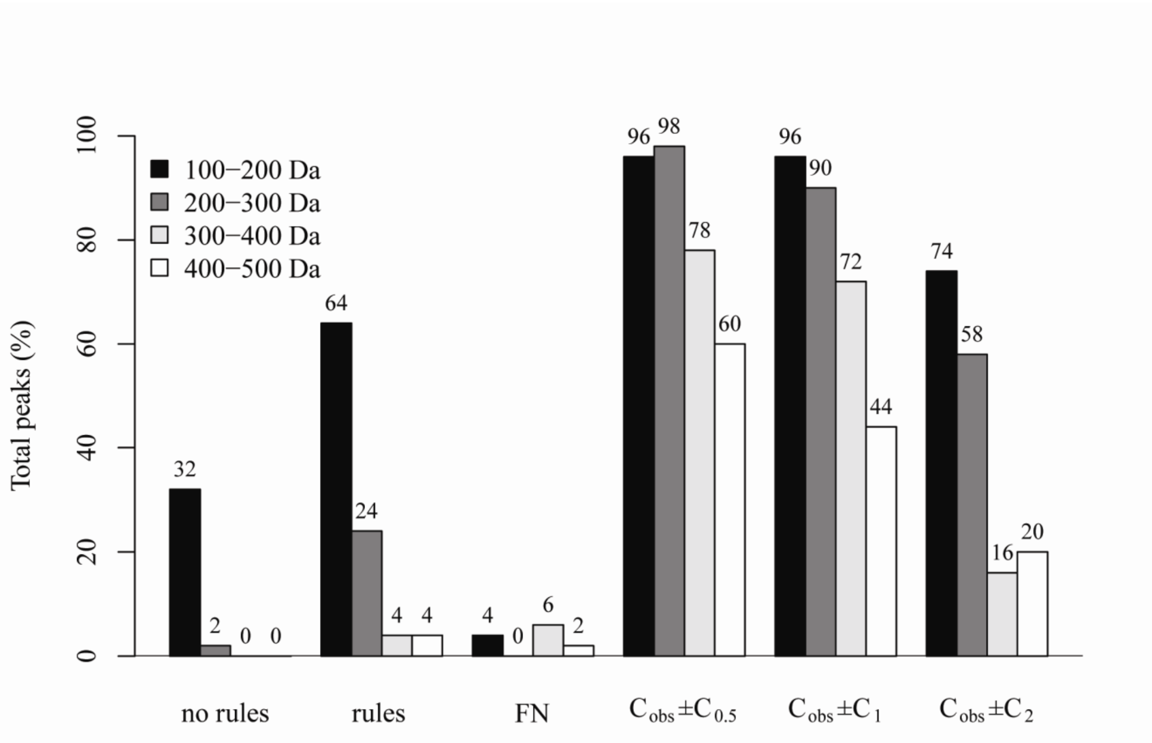


Figure 1.6 Percentages of theoretical peaks with a single correct empirical formula assignment without (“no rules”) and with (“rules”) the use of heuristic rules and carbon isotopic information ($C_{observed}$ ± C_n), for three different carbon error ranges. The third group shows the false negative (“FN”) rate after using heuristic rules. The data presented is based on four theoretical peak lists, each peak list represents a different mass range and comprises of 50 adduct/ion masses selected randomly from the modified KEGG COMPOUND database (Kanehisa *et al.*, 2008).

To reduce the false positive assignment of incorrect empirical formulae and to increase the number of “true positive” correct assignments, several criteria can be utilised. First, restricting the elements and elemental ranges (e.g. $C_{1\text{-to-}34}$) included in the calculation minimises the number of unrealistic empirical formulae assignments whilst saving computational time. For most biological studies focusing on polar metabolites, formulae comprised of ^{12}C,^{1}H,^{14}N,^{16}O,^{31}P and ^{32}S will satisfy the assignment of most natural compounds. The maximum absolute number of each element in the formula can be calculated by dividing the observed mass by the elemental mass (e.g. 170.05782/12.0 for carbon). However, Kind and Fiehn (2007) have shown the appropriacy of even stricter limits.

Observed peaks correspond to charged molecular ions and adducts of neutral metabolites requiring addition of several atoms and molecules, predominantly $[M + H]^+$, $[M + {}^{23}Na]^+$ and $[M + {}^{39}K]^+$ for positive ion mode and $[M – H]^-$, $[M + {}^{35}Cl]^-$ and $[M + {}^{37}Cl]^-$ for negative ion mode, depending on the composition of the biological matrix. Further constraints are typically used to evaluate the empirical formulae generated, such as the nitrogen rule and valence considerations. The heuristic rules, again developed by Kind and Fiehn (2007), set elemental limits which assess the likelihood of each elemental composition and clearly increase the number of single correct assignments, particularly for low masses (<300 *m/z*) as shown in Figure 1.6. However, a small number of false negative assignments are observed as a result of the filters being set too strictly (see Figure 1.6). Even with strict constraints the number of single correct assignments is relatively small for higher mass ranges (>300 *m/z*).

Finally, for further improvement in the accuracy of putative metabolite identification, relative isotopic abundance (RIA) measurements can be used, i.e. originating in the fact that $^{13}C^{12}C_{a}{-}_{1}{}^{1}H_{b}{}^{14}N_{c}{}^{16}O_{d}{}^{31}P_{e}{}^{32}S_{f}$ has a natural abundance of 1.1% relative to the parent peak $^{12}C_{a}{}^{1}H_{b}{}^{14}N_{c}{}^{16}O_{d}{}^{31}P_{e}{}^{32}S_{f}$. The ratio of these peak intensities provides an estimate of the number of carbon atoms in the formula, which can potentially be used to further filter the assignments of empirical formulae from mass measurements alone. HR FT mass spectrometers are of particular value here as they have the ability to distinguish closely spaced peaks (isotopic distributions), for example, $[^{12}C_{a}{}^{1}H_{b}{}^{14}N_{c}{}^{16}O_{d}{}^{31}P_{e}{}^{32}S_{f}+{}^{41}K]^{+}$ and $[^{12}C_{a}{}^{1}H_{b}{}^{14}N_{c}{}^{16}O_{d}{}^{31}P_{e}{}^{34}S{}^{32}S_{f-1}+{}^{39}K]^{+}$ which differ by only 0.0022 Da. Characterising the accuracy and precision of RIA FT measurements in HR mass spectra is important before such a method can be applied. Furthermore, high sensitivity is important for accurate RIA measurements for two reasons: (i) isotope peaks may not be detected in mass spectra if the instrument's sensitivity is too low; (ii) the accuracy and precision of RIA measurements is highly dependent on the measured signal intensity (Xu *et al.*, 2010). When an RIA prediction accuracy of ±1 carbon atoms (for a given empirical formula) is applied for the ranges 100-200 and 200-300 *m/z*, more than 90% of the peaks are assigned a single correct empirical formula, highlighting the value of RIA measurements for improving empirical formulae assignments.

Alternatively, a probabilistic model using statistics, a posterior probability distribution over the set of empirical formulae assignments, and a predefined list of biochemical transformations to produce a list of the most likely empirical formulae assignments, has been investigated by Rogers *et al.* (2009). These predefined biochemical reactions include, for example, $\pm H_2O$ for condensation/dehydration reactions and $\pm HO_3P$ for phosphorylation/dephosphorylation reactions. Further improvement was achieved by the inclusion of isotopic information. The success of this particular approach highlights the value of prior biochemical knowledge of endogenous reactions for improving the confidence of empirical formulae identification.

Assigning empirical formulae or mass measurements to putative structural identities or metabolite names is the next goal. The latter is generally achieved *via* automated searches against a compound database (what I term a "single-peak search"), such as KEGG (Kanehisa *et al.*, 2008). The mass accuracy of the MS analysis, search threshold (i.e. mass tolerance) and quality (density, completeness, correctness and specificity) of the compound database used all play an important role in the success of automated database searches. The greater the density of the database, the more metabolite entries for a given mass range, i.e. search window, and the more likely the search will result in a high false positive rate as one empirical formula typically occurs as numerous different structural isomers (Matsuda *et al.*, 2009). Incomplete databases will obviously result in a higher false positive rate, while inaccuracies result in redundant assignments. A lack of organism specificity in the database will also cause a high false positive rate, as some metabolites will only be found in certain organisms and tissues (Weber & Viant, 2010b). Although some organism-specific databases exist, such as the Human Metabolome DataBase (HMDB) (Wishart *et al.*, 2009) for human and EcoCyc (Keseler *et al.*, 2009) for *Escherichia coli*, they are currently limited in number (Aliferis & Chrysayi-Tokousbalides, 2010; Brown *et al.*, 2009). Advancements of analytical techniques and careful annotation of the metabolite signals have the potential to improve the situation in the near future.

In addition to the “single-peak search” approach described above, several more sophisticated strategies that integrate prior biological knowledge to improve the accuracy of metabolite identification have been published (Breitling, Pitt & Barrett, 2006; Breitling, Ritchie, Goodenowe, Stewart & Barrett, 2006; Brown *et al.*, 2009; Gipson *et al.*, 2008; Jourdan *et al.*, 2008; Rogers *et al.*, 2009). Chapter 3 includes a more extended introduction followed by the novel work that was carried out here to contribute to more accurate, automated (putative) metabolite identification in MS-based metabolomics. It is important to emphasise that the final list of putatively identified empirical formulae or structural identities are, if required, further analysed using additional analytical techniques such as MS^n fragmentation to fulfil the requirements proposed by the Metabolome Standard Initiative (MSI) (Sumner *et al.*, 2007).

1.9 Research Objectives

The aim of this work is to develop novel methods and improve existing methods to measure and subsequently annotate HR MS data, and to apply these methods to real-world metabolomics studies. The specific research objectives are therefore:

1. Review, investigate and develop methods that integrate prior biological and spectral knowledge to reduce incorrect assignments in automated metabolite identification (Chapter 3);
2. Re-optimise acquisition parameters for the direct infusion SIM-stitching method (developed originally for a Thermo Scientific LTQ FT (Southam *et al.*, 2007)) on a Thermo Scientific LTQ FT Ultra to maximise sensitivity and mass accuracy. This will therefore allow a more complete and precise identification of metabolites (Chapter 4);
3. Characterise the accuracy and precision of RIA measurements in HR MS, conducted by two leading FTMS instruments (i.e. Thermo Scientific LTQ FT Ultra and Thermo Scientific LTQ Orbitrap) (Chapter 5);
4. Investigate to what extent RIA in HR MS spectra contribute to decreased incorrect assignments in automated metabolite identification (Chapter 5);
5. Review and investigate several applications in MS-based metabolomics that make use of methods developed in this work (Chapter 3).

CHAPTER TWO:

MATERIALS AND METHODS

This section will detail general methods that were used in more than one experimental chapter. Methods that were unique to a single experimental section of this thesis will be detailed in the chapter where they were used.

2.1 Preparation biological samples and chemical standards[2]

Biological samples, (Human K562 myeloid leukaemia cells (Tiziani *et al.*, 2009) and freshwater flea (*Daphnia magna*) (Taylor *et al.*, 2009), Chapter 3; liver samples from a marine flatfish (the Dab, *Limanda limanda*) (Southam *et al.*, 2008), Chapters 4 and 5) were homogenised in methanol 8 µL/mg wet tissue mass (all solvents HPLC grade, Fisher Scientific, Loughborough, UK) and water (2.5 µL/mg) (Wu *et al.*, 2008). Homogenisation was carried out for 20 sec at 64,000 rpm using a Precellys-24 bead-based homogeniser (Fisher Scientific). Homogenates were transferred into 1.8 mL glass vials using glass Pasteur pipettes and chloroform (8 µL/mg added with a Hamilton syringe, Fisher Scientific) and water (4 µL/mg) were added (Wu *et al.*, 2008). The samples were vortexed for 30 sec, allowed to stand on ice for 10 min and then centrifuged at 2500-*g*, (4 °C). The polar and non-polar phases were carefully transferred into fresh 1.8 mL glass vials using a Hamilton syringe. Polar extracts were dried using a centrifugal evaporator (Thermo Scientific Savant, Holbrook, NY) and stored at -80°C until analysis. Non-polar extracts were stored at -80°C, and these were not used in this thesis. Prior to analysis, polar extracts were resuspended in 3 times their original volume

[2] Experimental work, regarding biological samples, described in this section was performed by Drs. S. Tiziani, A. Southam and N. Taylor.

of 80:20 v/v methanol:water with either 0.25% formic acid for positive ionisation mode (Fisher Scientific), or 20 mM ammonium acetate for negative ionisation mode (Fisher Scientific).

A solution of 0.0005% PEG200 and 0.0005% PEG600 (Sigma-Aldrich, UK) was prepared in 80:20 v/v methanol:water with 0.25% formic acid (all HPLC grade, Fisher Scientific, UK). Frozen liver samples from a marine flatfish (Southam *et al.*, 2008) were extracted and prepared as described above and aliquots equivalent to 1.5 mg of biomass were dried. These extracts were each resuspended in 75 µl of the polyethylene glycol (PEG) solution, and samples were pooled. This solution is referred to as BioPEG and used for the re-optimisation of the SIM-stitching acquisition parameters (Chapter 4) and the characterisation of HR FTMS isotopic abundance measurements (Chapter 5).

2.2 Direct infusion mass spectrometry

Prior to analysis, samples were centrifuged (5,000-g) for 10 minutes, and the supernatant (>5 µL) was loaded into a standard 96-well sample plate (Abgene, Epsom, UK). The sample plate was sealed with foil using a heat-sealing device (Thermo-sealer, Abgene). Samples were analysed using a hybrid 7T FT-ICR mass spectrometer (LTQ FT or LTQ FT Ultra, Thermo Electron Corp., Bremen, Germany) equipped with a chip-based direct infusion nanoelectrospray ionisation assembly (TriVersa, Advion Biosciences, Ithaca, NY). During sample analysis, the 96-well sample plate sits on a cooler plate within the NanoMate/TriVersa, which retains the samples at 10°C. The sample is picked up from the 96-well sample plate using a new disposable plastic tip each time, placed against a new electrospray nozzle and an electric current is applied to the chip, which initiates nanoelectrospray of the sample. Nanoelectrospray conditions comprised a 200 nL/min flow rate, 0.3 psi backing pressure and ±1.7 kV electrospray voltage (positive or negative ion mode), controlled by ChipSoft software (version 8.1.0, Advion Biosciences). For the LTQ FT each sample was analysed in triplicate using the SIM-stitching FT-ICR method from *m/z* 70–500 (SIM windows of 30 *m/z* overlapping by 10 Da), as reported previously (Southam *et al.* 2007). Parameters included an AGC target of 1×10^5 and the mass resolution (defined for an ion at *m/z* 400) was fixed at 100 K throughout, and data was recorded for 5 min 45 sec per replicate analysis. All parameters mentioned above were re-optimised for the LTQ FT Ultra as part of this thesis, and detailed in Chapter 4.

2.3 Data processing

Data were obtained either as processed mass spectra with associated peak lists (Xcalibur, version 2.0, Thermo Electron), or as transient files (i.e. scans recorded in the time domain), which were processed using the SIM-stitching algorithm (Southam *et al.*, 2007), including averaging of transients, Hanning apodisation, zero filling once, and application of a fast Fourier transformation (see Chapter 1). Next, all adjacent SIM windows (see Section 1.2) were stitched together using the same SIM-stitching algorithm (Southam *et al.*, 2007). All peaks with a SNR below 3.5 were rejected, and then each mass spectrum was internally calibrated using two-parameter equation (i.e. m/z = A/f + B/(f^2), A = ion charge density and trap geometry, B = magnetic field strength, f = frequency) and a pre-defined calibrant list of known metabolites. Calibrated mass spectra were processed using a three-stage signal filtering algorithm (Payne *et al.*, 2009): (i) signals in the mass spectra were selected as peaks if above a SNR of 3.5:1; (i) peaks were only retained if present in at least two out of three replicate measurements (1.5 ppm spread), and peaks in the extract blank were removed from the biological mass spectra unless twice as intense as in the biological samples (1.5 ppm spread); and (iii) peaks were only retained if present in at least 50% of all biological samples (2.0 ppm spread).

Finally, the signal intensity matrix after three-stage signal filtering was further processed by the addition of missing values (Sangster *et al.*, 2007), normalised using the probabilistic quotient method (Dieterle *et al.*, 2006), and generalised log transformed (Parsons *et al.*, 2007). Although, the final steps to process the signal intensity matrix are not essential for the work presented in this thesis, they are part of the automated pipeline used to process mass spectral datasets as described previously by Taylor *et al.* (2009).

2.4 Metabolite identification

Putative identification of metabolites is an important part of the presented work in this thesis. The next two sections describe the general approaches that were used to putative identify metabolites in HR mass spectra (see Chapter 3, 5 and 6). Both approaches are extensively validated and integrated into MI-Pack as part of Chapter 3.

2.4.1 Putative identification of empirical formulae

Between zero and many potential empirical formulae ($C_CH_HN_NO_OP_PS_S$) were calculated for each of the experimentally detected peaks in the data matrix described above. This was achieved using a custom-written elemental composition calculator (in Python and SQLite, as part of MI-Pack, see Chapter 3), in which the elements were restricted (Kind

& Fiehn, 2007), see each experimental section of chapter for more details on restrictions. In addition, since the observed peaks correspond to charged molecular ions or adducts of the neutral metabolites, the masses of the several common positive or negative ion adducts were effectively added to the elemental composition calculator. In practice, the calculation was conducted for each positive ion peak, allowing for $[M + H]^+$, $[M + Na]^+$ and $[M + {}^{39}K]^+$, or each negative ion peak, allowing for $[M - H]^-$, $[M + {}^{35}Cl]^-$, $[M + {}^{37}Cl]^-$. Only those empirical formulae with an absolute mass error of 1.0 ppm (see each experimental section of each chapter for more details regarding mass accuracy) were recorded, and then all potential formulae were filtered using four heuristic rules reported by Kind and Fiehn (2007). Specifically: (i) restricted number of atoms per element, (i) Lewis and Senior rules, (i) H/C ratio, and (iv) elemental ratio of N, O, P, and S *versus* C. Note that these rules were applied to the neutral metabolites following subtraction of the adduct, e.g. applied to $[C_5H_9NO_2]$ not to $[C_5H_9NO_2 + Na]^+$. In some cases, the ^{13}C isotope of $C_CH_HN_NO_OP_PS_S$ was identified, allowing the approximate determination of the number of carbon atoms in the molecule based upon the intensity ratio of the ^{12}C and ^{13}C peaks, which is more precisely characterised for HR FTMS in Chapter 5.

2.4.2 Putative identification of metabolite names

The putative metabolite identity of each of the observed peaks was determined based upon accurate mass measurements, which was achieved using custom-written script (in Python and SQLite, as part of MI-Pack, see Chapter 3) and the KEGG database (Kanehisa *et al.*, 2008). First the KEGG LIGAND database was downloaded and the exact mass of each entry was re-calculated based on the associated empirical formula;

this served to increase the mass accuracy to 6 decimal places. Then the exact mass of each entry (neutral compound) was modified to allow for the formation of several ions listed above. For example, the database record for glucose (KEGG ID C00031, *m/z* 180.063390) was extended by different adduct masses such as *m/z* 203.052611 for [glucose + Na]$^+$, *m/z* 215.032792 for [glucose + ^{35}Cl]$^-$, etc. Next the *m/z* values of all the experimentally observed peaks were compared to the modified KEGG data and matches with an absolute mass error of 1.0 ppm were recorded (see each experimental section of each chapter for more details regarding mass accuracy). See Chapter 3 for a detailed report and discussion on novel methods to improve putative identification of metabolites.

2.5 Hardware and Software

2.5.1 High performance computing cluster architecture

Two high performance computing cluster architectures, named BlueBEAR (http://www.bear.bham.ac.uk/) and Xenia (http://www.biochemistry.bham.ac.uk/hpcc/), were used to run computationally expensive jobs (e.g. calculating empirical formulae). Less extensive jobs were performed on a dual-core processor desktop system. Specifications of each system are detailed in the next three sections.

2.5.1.1 BlueBEAR cluster

- 384 twin dual-core processors (4 cores/node) 64-bit 2.6 GHz AMD Opteron 2218
- Total of 1536 cores (8 and 16 (16 cores) GB memory)

- 4 quad-processor dual-core (8 cores/node) 64-bit 2.6 GHz AMD Opteron 8218 (32 GB of memory)
- 150 TB storage (IBM's - general parallel file system (GPFS))
- Operating software: Scientific Linux 5.2.

2.5.1.2 Xenia cluster

- 32 twin dual-core processors (4 cores/node) 64bit 2.4 GHz AMD64
- Total of 128 cores (8 GB memory)
- 10 TB of RAID 5 storage
- Operating software: Red Hat Enterprise Linux 4 (RHEL4)

2.5.1.3 Dual-core processor desktop system

- Dell Dimension E520
- Intel® Core™ 2 Duo processors (E6300) @ 1.86 GHz
- 2 GB of memory
- 160 GB hard disk
- Operating software: Windows XP (http://www.microsoft.com/windows/) and Ubuntu (http://www.ubuntu.com)

2.5.2 Application software and programming languages

The following application software and programming languages were used to perform data analysis and programming throughout the presented work:

- Matlab (The Mathworks, http://www.mathworks.co.uk)
    - Bioinformatics Toolbox (The Mathworks, http://www.mathworks.co.uk)

  - SIM-stitching algorithm (Southam *et al.*, 2007)
  - Statistics Toolbox (The Mathworks, http://www.mathworks.co.uk)
- Microsoft Office 2003 (http://office.microsoft.com)
- Molecular Weight Calculator (http://ncrr.pnl.gov/software/)
- Mozilla Firefox web browser (http://www.mozilla.com)
- MySQL (http://www.mysql.com)
- Python programming language (http://www.python.org)
  - Rpy (http://rpy.sourceforge.net)
  - Parallel Python (http://www.parallelpython.com)
  - Numpy (http://numpy.scipy.org)
- Putty (http://www.chiark.greenend.org.uk/~sgtatham/putty/)
- R programming language (http://www.r-project.org)
- SQLite (http://www.sqlite.org)
- SQLite Manager (http://code.google.com/p/sqlite-manager/)
- WinScp (http://winscp.net/)
- Xcalibur (Thermo Electron)
- Chenomx NMR Suite (version 5.0; Chenomx Inc., Canada)

CHAPTER THREE:

MI-PACK: INCREASED CONFIDENCE OF METABOLITE IDENTIFICATION IN MASS SPECTRA BY INTEGRATING ACCURATE MASSES AND METABOLIC PATHWAYS[3]

[3] Parts of this Chapter have been published in *Chemometrics and Intelligent Laboratory* Systems (Weber & Viant, 2010b). Applications of putative identification of metabolites in biological HR FT-ICR mass spectra presented in this Chapter have been published in *Metabolomics* (Taylor *et al.,* 2009*) and Toxicological Sciences* (Taylor *et al.,* 2010).

3.1 Introduction

3.1.1 Putative identification of metabolites in biological HR FT-ICR mass spectra

HR FTMS is a leading technique for measuring metabolites in biological samples (Dunn, 2008; Giavalisco *et al.*, 2008; Han *et al.*, 2008; Taylor *et al.*, 2009; Zhang *et al.*, 2009). However, automatic identification of the hundreds or potentially thousands of metabolites detected remains one of the greatest challenges in metabolomics, and is critical for generating new knowledge of biological systems (Breitling *et al.*, 2008). While tandem mass spectrometry (MS/MS) and MS^n can identify the structures of selected metabolites, it is a less feasible approach for the automated assignment of many thousands of peaks, particularly if high sample throughput is required. Ideally, information intrinsic to a high resolution wide-scan mass spectrum can at least provide *putative* metabolite identifications (Sumner *et al.*, 2007). Subsequently a limited number of putatively identified metabolites that are found to be of particular importance to the biological phenomena being investigated can be analysed by MS/MS to confirm their identity. Putative metabolite identification can in principle be achieved using accurate mass measurements, typically by searching against a database on a peak-by-peak basis (here termed "single-peak search") (Kind & Fiehn, 2006; Smith *et al.*, 2006). However a single accurate mass measurement can be assigned to one or more empirical formula(e) of the form $[C_AH_BN_CO_DP_ES_F\text{+adduct}]^{\pm}$. Furthermore each formula can occur in different chemical structures, each representing a unique metabolite; e.g. $C_6H_{12}O_6$ at 180.06339 Da corresponds to several carbohydrate metabolites. This complexity, together with the technology-limited mass accuracy of any mass spectrometer, the occurrence of thousands of metabolites in nature and the somewhat low number of available species- and metabolite-specific databases (e.g. KNApSAcK and HMDB) (Brown *et al.*, 2009; Draper *et al.*, 2009; Shinbo *et al.*, 2006; Wishart *et al.*, 2009), has the potential to induce a high false positive rate (FPR) of metabolite identification, in particular for a single-peak search.

Kind and Fiehn (2007) published a set of 'golden rules' (implemented here) to remove empirical formulae that were incorrectly assigned to an accurate mass, thus helping to reduce this error rate. A second strategy to increase the accuracy of metabolic identification is to use prior biological knowledge of the interconnectivity of metabolites into metabolic networks. Previously, prior knowledge has been implemented into the annotation of a high resolution mass spectrum using mass differences between pairs of peaks (Breitling, Pitt & Barrett, 2006; Breitling, Ritchie, Goodenowe, Stewart & Barrett, 2006; Jourdan *et al.*, 2008). For example, Gipson *et al.* (2008) used a clustering method and biological database to assign a peak-pair to a substrate-product biochemical reaction. Although this approach is dependent upon well-defined peak shapes from liquid chromatography (LC) data, and hence not generalisable to all high resolution mass spectra (in particular to direct infusion MS), it demonstrates how prior knowledge of metabolic pathways can increase the confidence of metabolite identification. Subsequently, Rogers *et al.* (2009) reported an elegant probabilistic method that compared observed mass differences of peak pairs to a list of common biochemical reactions in order to assign the most likely empirical formulae to the peaks. These predefined biochemical reactions included, for example, ±H_2O for condensation/dehydration reactions and ±HO_3P for phosphorylation/dephosphorylation reactions. The success of this particular approach demonstrates the value of prior knowledge of types of biochemical reactions for improving the confidence of identifying empirical formulae. Brown *et al.* (2009) also employed mass differences to increase confidence in metabolite identification, specifically utilising differences in adduct and fragment patterns, rather than bio-transformations. Some of the identification strategies described above depend on the reliability and completeness of existing metabolic networks, e.g. KEGG. Therefore recent progress in the reconstruction of metabolic networks is highly relevant to these strategies to improve the accurate representation of biochemical, metabolic and signaling networks, ultimately improving metabolite annotation (Herrgård *et al.*, 2008).

Error rates for metabolite identification are inherently dependent upon both the actual technology-limited mass accuracy of the MS and the mass error range used in the identification algorithm. MS techniques with high mass accuracy (e.g. FTMS with 0.18 ppm root-mean-square mass error (Southam *et al.*, 2007)) are ideal. However, if the selected mass error range for the algorithm is too small relative to this instrumental error, some metabolites that are present will not be identified; i.e. a high false negative rate (FNR). Conversely, if the mass error range is set too large, there will be many incorrect assignments; i.e. a high FPR. Therefore to maximise the accuracy of metabolite identification it is important to determine the instrumental error over the entire mass spectrum and use this to define the mass error range for the identification algorithm. While it is routine to calculate mass errors of individual metabolites in a spectrum by comparing the theoretical and measured mass/charge (*m/z*) values, the errors associated with the *mass differences* of peak pairs have not previously been reported. These errors should be characterised experimentally to maximally exploit peak-pair differences for metabolite identification.

Here I present the Metabolite Identification Package (MI-Pack) that incorporates several modules to support metabolite identification in mass spectra. The core identification algorithm is based on an approach of mapping an experimentally-derived empirical formula difference of a peak-pair to a known empirical formula difference between substrate-product pairs derived from KEGG (Kanehisa *et al.*, 2008), which I term transformation mapping (TM). This uses a similar concept to that reported by Rogers *et al.* (2009), except instead of using a predefined list of common biochemical reactions to identify empirical formulae, here our algorithm exploits prior knowledge of metabolite interconnectivity from the KEGG database to improve the accuracy of identifying metabolite names. This is achieved according to the assumptions that metabolites within a particular pathway will all ionise and that they are all present at detectable concentrations. As an additional extension to the Gipson *et al.* (2008) and Rogers *et al.* (2009) studies, and to further increase the accuracy of metabolite identification, I have developed a novel semi-automated method to calculate the instrumental errors associated with peak differences. This generates an error surface that is a function of both the mass difference and the average *m/z* of a peak-pair. The TM algorithm and mass error surface are extensively validated and compared against the most widely used method for the identification of metabolite names, the single-peak search (for a range of ppm mass measurement errors). This employed four types of dataset: (i) simulated dataset comprising a list of metabolite masses extracted from one KEGG pathway to demonstrate the TM algorithm; (ii) experimental mass spectra of human cancer cells to derive a FPR (i.e. for peaks incorrectly identified as non-human metabolites in a human sample); (iii) simulated dataset comprising of randomly generated peak lists to calculate a second FPR (i.e. for random peaks incorrectly identified as metabolites); and (iv) experimental NMR spectra of the same human cancer cells as above, to determine the FNR for metabolite identification, i.e. known metabolites that were not identified. Metabolite Identification Package (MI-Pack) and details of implementation and use are freely available from http://www.biosciences-labs.bham.ac.uk/viant/mipack/.

3.1.2 Applications

The research presented in this Chapter has focused on the development of algorithms to assist the experimentalist in automatically assigning putative metabolite identities. During the last three years, to demonstrate the value of these algorithms in real-world examples, I have contributed to several applied research projects. Here I present several applications of automated putative identification of metabolites for two different organisms (*Daphnia magna* and *Salmonella enterica* serovar typhimurium) as part of three different studies. Two of these three studies have been published (Taylor *et al.*, 2010; Taylor *et al.*, 2009). Several novel and non-novel identification approaches were applied, including empirical formulae assignments, “Single-Peak Search” (SPS) and “Transformation Mapping” (TM) all part of MI-Pack. In addition, for one of the three studies the re-optimised SIM-stitching acquisition parameters (see Chapter 4) were used to collect the MS dataset. I focus upon the results of the automated putative metabolite identification, not on the biological questions being addressed in the research. However, for completeness, a brief introduction to all three studies is provided below.

3.1.2.1 Metabolic profiling of *Daphnia magna* using FT-ICR mass spectra for toxicity screening (Taylor *et al.*, 2009)

Currently there is a surge of interest in exploiting toxicogenomics to screen the toxicity of chemicals, enabling rapid and accurate categorisation into classes of defined mode-of-action (MOA), and prioritizing chemicals for further testing. Direct infusion FT-ICR MS-based metabolomics can provide a sensitive and unbiased analysis of metabolites and therefore has considerable potential for chemical screening. The water flea, *Daphnia magna*, is an Organisation for Economic Co-operation and Development (OECD) test species and is utilised internationally for toxicity testing. However, no metabolomics studies of this species have been reported. Here we optimised and evaluated the effectiveness of FT-ICR MS metabolomics for toxicity testing in *D. magna*. It was confirmed that high-quality mass spectra can be recorded from as few as 30 neonates (<24 h old; 224 μg dry mass) or a single adult daphnid (301 μg dry mass). An OECD 24 h acute toxicity test was conducted with neonates at copper concentrations of 0, 5, 10, 25 and 50 μg l^{-1}. A total of 5447 unique peaks were detected reproducibly, of which 4768 were assigned at least one empirical formula and 1017 were putatively identified based upon accurate mass measurements (using algorithms developed in Chapter 3). Significant copper-induced changes to the daphnid metabolome, consistent with the documented MOA of copper, were detected thereby validating the approach. In addition, *N*-acetylspermidine was putatively identified as a novel biomarker of copper toxicity. Collectively, these results highlight the excellent sensitivity, reproducibility and mass accuracy of FT-ICR MS, and provide strong evidence for its applicability to high-throughput screening of chemical toxicity in *D. magna*.

3.1.2.2 Discriminating between different acute chemical toxicities via changes in the daphnid metabolome (Taylor *et al.*, 2010)

Previous work (see Section 3.1.2.1) has shown that mass spectra can be recorded from whole-organism homogenate or haemolymph of single adult *Daphnia magna*. Here we develop multivariate models and discover perturbations to specific metabolic pathways that can discriminate between the acute toxicities of four chemicals to *D. magna* using FT-ICR MS metabolomics. We focus on model toxicants (cadmium, fenvalerate, dinitrophenol, and propranolol) with different MOAs. First, we confirmed that a toxicant-induced metabolic effect could be determined for each chemical in both the haemolymph and the whole-organism metabolome, with between 9 and 660 mass spectral peaks changing intensities significantly, dependent upon toxicant and sample type. Subsequently, supervised multivariate models were built that discriminated significantly all four acute metabolic toxicities, yielding mean classification error rates (across all classes) of 3.9 and 6.9% for whole-organism homogenates and haemolymph, respectively. Following extensive peak annotation (using algorithms developed in Chapter 3), we discovered toxicant-specific perturbations to putatively identified metabolic pathways, including propranolol-induced disruption of fatty acid metabolism and eicosanoid biosynthesis and fenvalerate-induced disruption of amino sugar metabolism. We conclude that the metabolic profiles of whole daphnid homogenates are more discriminatory for toxicant action than haemolymph. Furthermore, our findings highlight the capability of metabolomics to discover early-event metabolic responses that can discriminate between the acute toxicities of chemicals.

3.1.2.3 Metabolomics to discover and identify molecules exported by the AcrAB-TolC multi-drug resistant efflux pump in *S. typhimurium*

Efflux pumps are attractive drug targets to reduce unwanted export of antibiotics from cells. However, the understanding of the natural roles of efflux pumps (AcrAB-TolC of *S. typhimurium* in this particular study) and their contribution to host-pathogen interactions is incomplete. An understanding of how the efflux pump affects pathogenicity is fundamental to the development of new drug treatments whether the host is plant, animal or human. The AcrAB-TolC pump forms a tri-partite complex; (i) AcrA is a periplasmic adaptor protein that links AcrB and TolC; (ii) AcrB is an integral cytoplasmic membrane protein; (iii) TolC is an outer membrane protein. The presence and absence of the pump AcrAB-TolC (or parts of the pump) and the effect on the resistance of the *Salmonella* to several compounds have been widely investigated (Karatzas *et al.*, 2007; Ricci & Piddock, 2009a; Ricci & Piddock, 2009b). The overall aim of this preliminary metabolomics study is to identify the natural substrate molecules exported by AcrAB-TolC thereby providing new data on the molecular basis of pathogenicity of *Salmonella enterica*. This was achieved using a MS metabolic footprinting approach, the re-optimised SIM-stitching method in Chapter 4, and the algorithms developed in this Chapter collectively enabling the putative identification of these exported compounds into the media.

3.2 Material and methods

3.2.1 Experimental and Simulated datasets[4]

3.2.1.1 Experimental NMR and HR FT-ICR mass spectra of cell extracts

Human K562 myeloid leukaemia cells (n=18 biological samples) were cultured, the intracellular metabolites extracted, and one-dimensional ^{1}H NMR spectra measured and processed, all as reported previously (Tiziani *et al.*, 2009). Metabolites were identified using the Chenomx NMR Suite (version 5.0; Chenomx Inc., Canada). Direct infusion FT-ICR mass spectra of the identical cancer cell extracts were measured and processed as described in Sections 2.1 - 2.3 (LTQ FT, see Table 4.2). Each mass spectrum comprised of ca. 4350 *m/z* values and intensities, and the final peak list after noise removal contained 3925 entries.

3.2.1.2 Experimental HR FT-ICR mass spectra of *Daphnia magna* and *Salmonella typhimurium*

The culturing of *Daphnia magna* as well as the biomass optimisation study and toxicity testing (to copper, cadmium, dinitrophenol (DNP), fenvalerate and propranolol) where performed as described by Taylor *et al.* (2009, 2010).

[4] Experimental work described in this section was performed by Drs. S. Tiziani, J. Kirwan, M. Webber and N. Taylor.

S. typhimurium wild type (L354, n=6) and mutant strains (L644, L742 and L976, each n=6) were cultured as described before (Webber *et al.*, 2009). Once cultured, the *S. typhimurium* cells were removed and 0.5 ml media was flash frozen in liquid nitrogen and diluted 1:2 v/v with methanol and 0.25% formic acid, further diluted 1 in 5 with 80:20 v/v methanol:water and 0.25% formic acid to make a final concentration of 74:16 v/v methanol:water.

Analyses were conducted using a hybrid 7-Tesla linear ion trap FT-ICR mass spectrometer (LTQ FT or LTQ FT Ultra, Thermo Scientific, Germany) equipped with a Triversa chip-based nanoelectrospray ion source (Advion Biosciences, NY, USA). Nanoelectrospray conditions (controlled by ChipSoft software version 8.1.0, Advion Biosciences) comprised 0.5 psi backing pressure, +1.7 kV electrospray voltage and ca. 200 nL/min sample flow rate. A mass resolution of 100,000 and maximum linear ion trap fill-time of 1s were fixed throughout the analyses (see Table 3.1, Section 2.2 and Table 4.2)).

Mass spectra were processed using the SIM-stitching algorithm (see Southam *et al.,* 2007 and Section 2.3) and the three-stage signal filtering algorithm detailed in Payne *et al.,* 2009 and Section 2.3.

Table 3.1 Overview of experimental details for DI FT-ICR MS metabolomics studies. D1, D2, S1 correspond to the studies summarised in Sections 3.1.2.1, 3.1.2.2 and 3.1.2.3, respectively.

Study	**Classes**	**No. of samples**	**Total no. of samples**	***m/z* range**	**Ion mode**
D1	(1) Control exposure copper - 0 μg l^{-1}	6	30$^{\#}$	70-500*	+ / -
	(2) Exposure copper - 5,10,25 and 50 μg l^{-1}	(4×6) 24			
D2	(1a) Control - whole organism	10	100$^{\#}$	70-500*	-
	(1b) Exposure 4 chemicals - whole organism	(4×10) 40			
	(2a) Control - haemolymph	10			
	(2b) Exposure 4 chemicals - haemolymph	(4×10) 40			
S1	(1) Control - L354	6	24$^{\#}$	70-590**	+
	(2) modified strains - L644, L742 and L976	3×6 (18)			

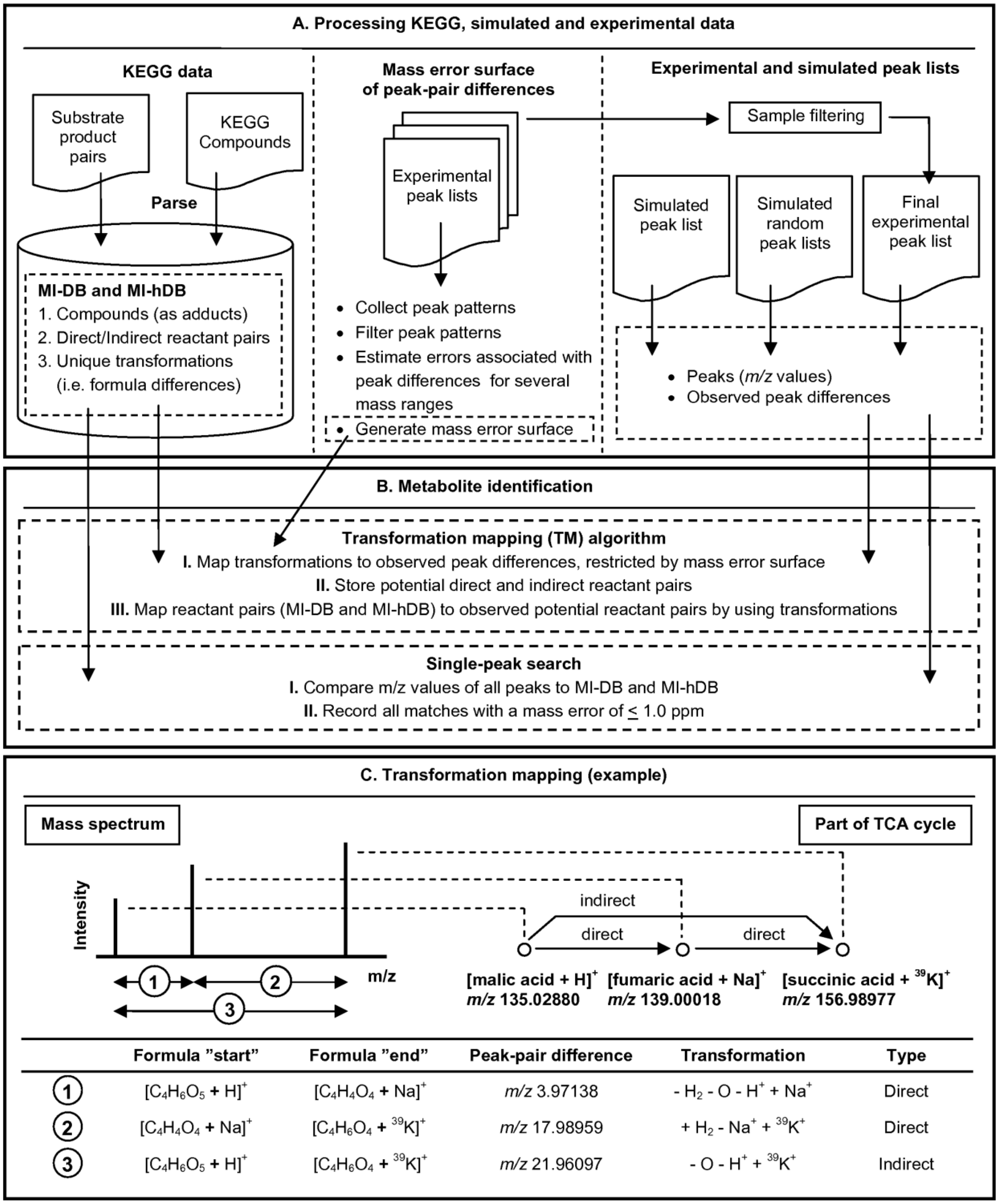
$^{\#}$measured in triplicate, *SIM-stitching parameters Thermo FT LTQ (table 4.2), **SIM-stitching parameters Thermo FT LTQ Ultra (table 4.2)

3.2.1.3 Simulated mass spectrum metabolic cycle

The KEGG identifiers of 16 small-molecule metabolites (M) occurring in the KEGG-representation of the tricarboxylic acid (TCA) cycle were extracted and then each was randomly assigned to one of three potential adduct forms ($[M + H]^+$, $[M + Na]^+$ or $[M + {}^{39}K]^+$). The accurate masses of these 16 adducts were then obtained from the modified database, described below, to generate the simulated peak list.

3.2.1.4 Simulated random mass spectra

Mass spectra comprising of randomly located peaks, although with realistic *m/z* values (i.e. clustered around integral *m/z*) were simulated as reported previously (Payne *et al.*, 2009). To achieve this, the simulation used prior knowledge of the peak distribution of the cancer cell mass spectra recorded above. In total 51 peak lists were generated, each containing ca. 4000 *m/z* values and intensities (see supplementary information by Payne *et al.* (2009)).


A. Processing KEGG, simulated and experimental data
KEGG data
Substrate product pairs
KEGG Compounds
Parse
MI-DB and MI-hDB
1. Compounds (as adducts)
2. Direct/Indirect reactant pairs
3. Unique transformations (i.e. formula differences)
Mass error surface of peak-pair differences
Experimental peak lists
Collect peak patterns
Filter peak patterns
Estimate errors associated with peak differences for several mass ranges
Generate mass error surface
Experimental and simulated peak lists
Sample filtering
Simulated peak list
Simulated random peak lists
Final experimental peak list
Peaks (*m/z* values)
Observed peak differences
B. Metabolite identification
Transformation mapping (TM) algorithm
I. Map transformations to observed peak differences, restricted by mass error surface
II. Store potential direct and indirect reactant pairs
III. Map reactant pairs (MI-DB and MI-hDB) to observed potential reactant pairs by using transformations
Single-peak search
I. Compare m/z values of all peaks to MI-DB and MI-hDB
II. Record all matches with a mass error of ≤ 1.0 ppm
C. Transformation mapping (example)
Mass spectrum
Part of TCA cycle
Intensity
m/z
1
2
3
indirect
direct
direct
[malic acid + H]$^+$ *m/z* 135.02880
[fumaric acid + Na]$^+$ *m/z* 139.00018
[succinic acid + ^{39}K]$^+$ *m/z* 156.98977

	Formula "start"	Formula "end"	Peak-pair difference	Transformation	Type
1	$[C_4H_6O_5 + H]^+$	$[C_4H_4O_4 + Na]^+$	*m/z* 3.97138	$- H_2 - O - H^+ + Na^+$	Direct
2	$[C_4H_4O_4 + Na]^+$	$[C_4H_6O_4 + {}^{39}K]^+$	*m/z* 17.98959	$+ H_2 - Na^+ + {}^{39}K^+$	Direct
3	$[C_4H_6O_5 + H]^+$	$[C_4H_6O_4 + {}^{39}K]^+$	*m/z* 21.96097	$- O - H^+ + {}^{39}K^+$	Indirect

Figure 3.1 Schematic representation of (a) the overall processing workflow for KEGG and mass spectral data, and **(b)** the metabolite identification workflow for the single-peak and TM approaches. **(c)** Example of the TM algorithm for three enzymatically-related metabolites in the TCA cycle, showing the peaks in the mass spectrum, the transformations between peak pairs (1, 2, 3), and their mapping to direct and indirect transformations stored in MI-DB.

3.2.2 Methods

3.2.2.1 Metabolite identification database (MI-DB and MI-hDB)

The KEGG COMPOUND database (Kanehisa *et al.*, 2008), containing ca. 15,500 metabolites, was downloaded (March 2009) and the exact mass of each compound re-calculated using the associated empirical formula (i.e. by multiplying the number of each element by the exact mass of that element, for all elements in the formula) (Wieser & Berglund, 2009); this increased the mass accuracy to 6 decimal places. The exact mass of each neutral compound was modified to account for the formation of three potential adduct forms, $[M + H]^+$, $[M + Na]^+$ and $[M + {}^{39}K]^+$ (Draper *et al.*, 2009); e.g. glycine (KEGG identifier C00037, *m/z* 75.03203) was modified to *m/z* 76.03931 for $[\text{glycine+H}]^+$ and *m/z* 98.02125 for $[\text{glycine+Na}]^+$ etc. All modified compound records were stored in a new database (MI-DB) (Figure 3.1a). Next the KEGG REACTION database (Kanehisa *et al.*, 2008), containing ca. 11,340 substrate-product pairs (here termed "direct reactant pairs"), was downloaded (March 2009) and parsed to the MI-DB. Subsequently, all direct reactant pairs were joined by an SQL statement to obtain "indirect reactant pairs" (i.e. two metabolites that are not directly linked by an enzyme, but that share a common third metabolite to which they are both enzymatically connected; Figure 3.1c). The empirical formula difference and mass difference of each direct and indirect reactant pair was added to each entry; e.g. for the substrate-product pair $[\text{oxaloacetic acid+H}]^+$ and $[\text{phosphoenolpyruvic acid+Na}]^+$, the empirical formula difference and mass difference were represented by $[-C+H+O+P+H^+-Na^+]$ and *m/z* 35.97650. A transformation reference list, comprising the information for all direct and indirect reactant pairs (mass differences and associated transformations), was inserted into the MI-DB (Figure 3.1a). Finally, to generate an equivalent human-specific database (MI-hDB), a reference list of human compound identifiers was obtained from the "human" KEGG pathways, i.e. using human classified compounds and pairs. All parsing-scripts were written in Python (including SQLite, http://www.python.org) and form modules of MI-Pack.

3.2.2.2 Assignment of peak patterns

This section describes the method used to locate specific peak patterns in experimental mass spectra in order to measure particular peak differences. The patterns sought included: (i) ^{13}C–^{12}C patterns (*m/z* difference of 1.00336), (ii) various patterns between adducts of metabolites (e.g. $[M + {}^{39}K]^+ - [M + Na]^+$, *m/z* 15.97394, $[M + Na]^+ - [M + H]^+$, *m/z* 21.98195, and $[M + {}^{39}K]^+ - [M + H]^+$, *m/z* 37.95588), and (iii) phosphorylation reactions ($[M + PO_3H + adduct]^+ - [M + adduct]^+$, *m/z* 79.96633). These peak differences were subsequently used to estimate the mass error associated with peak differences over the mass range *m/z* 70 – 500 (i.e. mass error surface).

The assignment of peak patterns comprises 5 steps and is illustrated here for adduct patterns.

1. For each of the 18 experimental peak lists (of cancer cell extracts), seven de-adducted lists were calculated by subtracting the accurate masses of seven potential adducts; i.e. subtracting an electron (e), proton (H^+), or various metal ions, for the adduct forms $[M - e]^+$, $[M + H]^+$, $[M + Na]^+$, $[M + {}^{39}K]^+$, $[M + {}^{41}K]^+$, $[M + 2Na - H]^+$, and $[M + 2{}^{39}K - H]^+$.

2. All de-adducted lists, for each experimental peak list, were concatenated and sorted from low to high mass. This resulted in all the adduct forms of any one metabolite becoming centered on the neutral mass of M.
3. Adduct patterns were identified from each final modified peak list by locating these "clusters" of peaks that lay within an *m/z* error range of 1.5 ppm (>3 times the largest experimental mass error reported for our FT-ICR method (Southam *et al.,* 2007)) based on the highest original mass starting with the highest *m/z* value. For each identified adduct pattern the original measured masses and the names of the adduct ions were recorded.
4. Peak pairs (i.e. part of a larger peak pattern or sometimes the entire peak pattern) were located within defined mass range windows, starting at 70 Da. The mass range of each window corresponded to the sum of the mass difference being searched for plus 10 Da. The next mass range window started at the highest peak assigned in the previous window.
5. Finally, the experimental mass differences of the various peak pairs used (e.g. $[M + Na]^+$ - $[M + H]^+$, $[M + {}^{39}K]^+$ - $[M + H]^+$ and $[M + {}^{39}K]^+$ - $[M + Na]^+$), were calculated for each mass range window that was defined.

The same method was applied to select ^{12}C-^{13}C patterns and phosphorylation reactions, except that the first step of the method was slightly different. An original peak list and a "subtracted" peak list (subtracting *m/z* 1.00336 for ^{12}C-^{13}C patterns or *m/z* 79.96633 for phosphorylation reactions) were concatenated to make up 18 final modified peak lists for each of both types of peak patterns.

3.2.2.3 Rules for confidence in peak-pattern assignment

Several rules were applied to reject coincidental matches and increase confidence in peak identification (i.e. to reject a peak-pair that has the same mass difference as a ^{12}C-^{13}C pair, but which does not in fact arise from those isotopes).

1. For adduct patterns, only patterns comprising of ≥3 adducts were retained (except for patterns at low mass, from *m/z* 70-150).
2. Illogical adduct patterns were rejected, when the pattern included a $[M + {}^{41}K]^+$ peak without the higher abundance $[M + {}^{39}K]^+$ peak.
3. For carbon isotopes (^{12}C, ^{13}C), only peak pairs with an allowed $^{12}C/^{13}C$ intensity ratio were retained (see Section 3.2.2.4).
4. Only peak patterns were retained when their *m/z* values could be assigned to at least one identical neutral empirical formula within an accuracy of 1.5 ppm. To achieve this, a custom-written elemental composition calculator (including rules by Kind and Fiehn, 2007) was used (part of MI-Pack).

Only adduct patterns comprising of ≥3 adducts were retained as described above (section 3.2.2.2). Therefore using seven adduct forms increased the number of retained adduct patterns. At the same time it increased the confidence of selecting adduct patterns as it lowered the number of wrong (random) pattern assignments. Consequently it resulted in a more robust mass error surface.

The 3 adduct forms ($[M + H]^+$, $[M + {}^{39}K]^+$ and $[M + Na]^+$) used for construction of the MI-DB were chosen as they are the most common ion forms in direct infusion-based MS. This has been confirmed by several colleagues in our field (e.g. (Brown *et al.*, 2009; Draper *et al.*, 2009; Taylor *et al.*, 2009). Table 8.1 represents this finding as most of the

NMR-detected metabolites are identified in the mass spectra by one or more of the 3 main adducts. Other adducts forms, e.g. $[M + 2Na - H]^+$, and $[M + 2^{39}K - H]^+$, were not used to minimise false assignments and hence the FPR .

3.2.2.4 Confidence in the assignment of carbon isotope patterns

Several studies have shown that ^{12}C–^{13}C isotope patterns in mass spectra are important to increase confidence in metabolite identification (Giavalisco *et al.*, 2008; Kind & Fiehn, 2007). ^{12}C–^{13}C isotopes show a well defined intensity pattern (e.g. ^{13}C has a relative abundance of 1.1% compared to ^{12}C) which was used in this study to increase the confidence in the assignments of carbon isotope patterns. This peak pair is one of the peak patterns that was used for estimating the mass error for peak differences (see above).

Our approach used the following strategy, effectively extracting prior knowledge of metabolite empirical formulae from KEGG. First, all unique formulae were obtained from the MI-DB for the mass range 0 - 500 Da and were divided in multiple 10 Da windows. For each 10 Da window the minimum and maximum number of carbon atoms from known metabolites were derived by simply inspecting the empirical formulae in that mass range. Both values were then multiplied by 1.1 % to calculate the estimated minimum and maximum peak abundance arising from the one ^{13}C-containing metabolite (relative to the all ^{12}C metabolite) over the defined mass range. From those estimated one ^{13}C-containing peak abundances, all ^{12}C abundances were calculated and plotted as a function of the average mass range (Figure 3.2). Linear regression was applied to retrieve the estimated linear equation to calculate the minimum and maximum ^{12}C *versus* ^{13}C abundances for a particular mass. The application of these criteria is shown in Table 3.2.

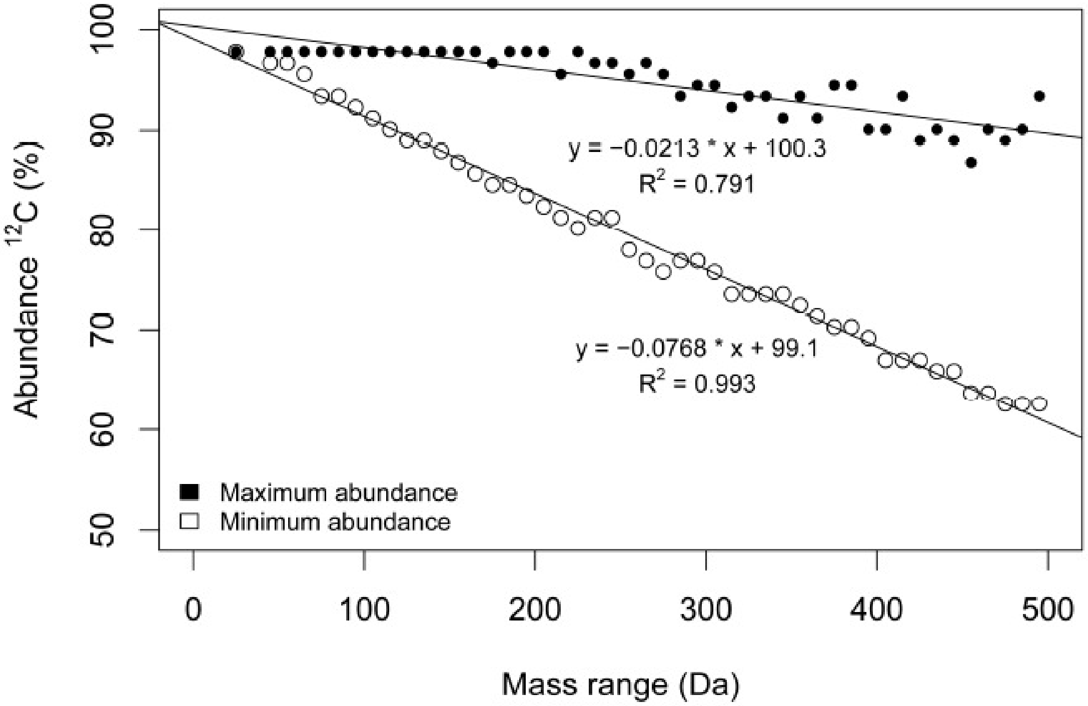


Figure 3.2 Minimum and maximum abundance of ^{12}C *versus* ^{13}C based on all formulae (0-500 Da) in the KEGG database.

Table 3.2 Examples of experimental ^{12}C-^{13}C patterns in mass spectra of human cancer cell extracts. The experimental ^{12}C to ^{13}C abundance was compared to the theoretical minimum and maximum abundances described above, and classified as a real (true) or non-real (false) isotope pattern.

peak^{12}C (*m/z*)	**peak^{13}C (*m/z*)**	**Intensity peak^{12}C**	**Intensity peak^{13}C**	**Minimum abundance (theoretical)[a]**	**Maximum. abundance (theoretical)[b]**	**Abundance (experimental)[c]**	**Criteria[d]**
229.15618	230.15960	66384.00	11228.04	81.48	95.46	85.53	true
292.95931	293.96299	11013.45	10993.69	76.58	94.10	50.04	false
330.27653	331.28004	32288.16	7087.93	73.71	93.31	82.00	true
360.33058	361.33411	11272.27	6821.34	71.40	92.66	62.30	false
388.28221	389.28554	35293.71	9383.07	69.26	92.07	79.00	true
443.27295	444.27603	2007.07	5415.45	65.03	90.90	27.04	false
444.28180	445.28507	199948.60	44524.89	64.96	90.88	81.79	true

[a]Minimum theoretical abundance = -0.0768 * *m/z* value peak^{12}C + 99.1, [b]Maximum theoretical abundance = -0.0213 * *m/z* value peak^{12}C + 100.3, [c]Experimental abundance = Intensity peak^{12}C / (Intensity peak^{12}C + Intensity peak^{13}C), [d]Experimental abundance > Minimum theoretical abundance and Experimental abundance < Maximum theoretical abundance

3.2.2.5 Mass error surface for peak differences

A mass error surface representing the error associated with the differences between peak pairs was generated as a function of both the mass difference and the average *m/z* of a peak-pair (Figure 3.1). Our approach was based upon identifying particular peak pairs in the experimental mass spectra with confidence (see Section 3.2.2.3 and 3.2.2.4) from which errors were calculated by comparing the theoretical and measured mass differences. Peak pairs comprised of (i) ^{13}C-^{12}C patterns (difference of *m/z* 1.00336), (ii) various patterns between adducts of metabolites ([M + ^{39}K]$^+$ - [M + Na]$^+$, *m/z* 15.97394; [M + Na]$^+$ - [M + H]$^+$, *m/z* 21.98195; and [M + ^{39}K]$^+$ - [M + H]$^+$, *m/z* 37.95588), and (iii) phosphorylation reactions ([M + PO_3H + adduct]$^+$ - [M + adduct]$^+$, *m/z* 79.96633). The mass error tolerance used to initially locate the peaks was 1.5 ppm, which is >3 times larger than the maximum mass error reported for our FT-ICR approach (Southam *et al.*, 2007). The peak pairs were then grouped into defined mass ranges in terms of their average *m/z* values (see Section 3.2.2.3). For each mass range and type of peak-pair, the multiple observations were visualised using a boxplot, any outliers were deleted, and the final mass error estimated as twice the standard deviation of a fitted normal distribution. Then the mass error surface was derived by interpolating between these errors, average *m/z* values and theoretical peak-pair differences. All processing, analyses and visualisations were executed using custom-written Python code (including SQLite, http://www.python.org) supported by the R scripting language (http://www.r-project.org/).

3.2.2.6 Metabolite identification algorithms

Two identification algorithms were employed (summarised in Figure 3.1b). The single-peak search simply compared the *m/z* values of all experimentally observed peaks, one at a time, against the MI-DB, and recorded all matches within a defined mass error (default value of ≤1.0 ppm, although errors ranging from ≤0.2 to ≤1.0 ppm were investigated). The TM approach calculated the *m/z* differences of all experimentally observed peak pairs and then compared these to the known mass differences in the KEGG-derived transformation reference list (see above). A match was made if the error between experimental and theoretical *m/z* differences was less than the threshold defined by the mass error surface. Each match, comprising of the experimentally measured peak-pair, mass difference and associated empirical formula difference (hereafter termed a transformation; Figure 3.1c), was recorded as a potential direct or indirect reactant pair. The matches were then mapped by SQL joining statements to the direct and indirect reactant pairs in MI-DB based on the transformation. Although the joining statement is mainly based on transformations, multiple starting points in the form of KEGG compound identifiers (i.e. a peak located within 1.0 ppm of a known metabolite in KEGG) are required to initiate the TM algorithm. This strategy avoids mapping the wrong reactant pair in the MI-DB as several reactant pairs in this database can be separated by the same transformation. It is also important to note that this approach requires high accuracy of the *m/z* values of these single "starting point" peaks. The TM algorithm then operated under one of three mapping constraints to improve the flexibility and robustness of metabolite identification (Figure 3.1c): map experimental observations to direct transformations only (default setting); to direct-*or*-indirect transformations (more flexible, recording a match even when a shared metabolite is not detected; e.g. even if fumaric acid is not detected, both succinic acid and malic acid are still identified, see Figure 3.1c); or to direct-and-indirect transformations (more robust, recording a match only when two "downstream" metabolites are present; e.g. if malic acid, fumaric acid and succinic acid occur in the mass spectrum, malic acid is identified as it forms a direct reactant pair with fumaric acid and indirect pair with succinic acid). These examples are given for the simplified case of only 3 metabolites, which is obviously not the case for real experimental data. All metabolite identification scripts have been written in Python (including SQLite) (http://www.python.org) and form modules of MI-Pack.

3.2.3 Putative metabolite identification of *Daphnia magna* and *Salmonella typhimurium* HR FT-ICR mass spectra

Peaks were assigned an empirical formula(e) using MI-Pack software (see table 3.2), based upon their accurate mass measurement and, where available, $^{12}C/^{13}C$ intensity ratios were used to estimate the number of carbon atoms. In many cases, each empirical formula was putatively assigned a metabolite name using the “single-peak search” or the “transformation mapping” approach (direct transformations only, see table 3.2).

Table 3.3 Overview details used for putative metabolite identification of for DI FT-ICR mass spectra. D1, D2 and S1 correspond to the studies summarised in Sections 3.1.2.1, 3.1.2.2 and 3.1.2.3, respectively.

Study	**No. of peaklists**	**Strategy for putative identification**	**Adduct types included**	**Mass error tolerance (ppm)**	**Atom ranges allowed***
D1	1	Empirical formula calculator Single-peak search (SPS)	$[M + H]^+$, $[M + Na]^+$, $[M + {}^{39}K]^+$, $[M + {}^{41}K]^+$, $[M + 2Na - H]^+$, $[M + 2{}^{39}K - H]^+$ $[M + e]^-$, $[M - H]^-$, $[M + {}^{35}Cl]^-$, $[M + {}^{37}Cl]^-$, $[M + Na - 2H]^-$, $[M + {}^{39}K - 2H]^-$, $[M + \text{acetate}]^-$	<0.75	$^{12}C_{0-34}$ H_{0-72} N_{0-15} O_{0-19} P_{0-4} $^{32}S_{0-3}$
D2	8**	Empirical formula calculator Transformation mapping (TM)	$[M - H]^-$, $[M + {}^{35}Cl]^-$, $[M + {}^{37}Cl]^-$	≤1.0	$^{12}C_{0-34}$ $^{13}C_{0-1}$ H_{0-72} N_{0-15} O_{0-19} P_{0-4} $^{32}S_{0-3}$ $^{34}S_{0-1}$ $Na^{35}Cl_{0-4}$ $Na^{37}Cl_{0-4}$
S1	1	Empirical formula calculator Transformation mapping (TM)	$[M + H]^+$, $[M + Na]^+$, $[M + {}^{39}K]^+$, $[M + {}^{41}K]^+$	≤1.0	$^{12}C_{0-34}$ $^{13}C_{0-1}$ H_{0-72} N_{0-15} O_{0-19} P_{0-4} $^{32}S_{0-3}$ $^{34}S_{0-1}$

*empirical formula assignments only, **one peaklist for each of the four chemicals (whole organism and haemolymph)

3.3 Results and Discussion

3.3.1 Demonstration of TM algorithm using simulated data

The value of the TM algorithm was demonstrated by comparing it against a single-peak search using a simulated peak list of 16 TCA cycle metabolites. All 16 metabolites were identified by both identification strategies (Table 8.1). However the single-peak search, using the MI-DB, falsely identified 35 metabolite names (of a total of 51 matches, giving a FPR of 69%). This high error rate is caused by many metabolites having the same exact mass and empirical formulae; e.g. nine metabolites have an exact mass of *m/z* 192.02700, all of which were assigned by the single-peak search even though only two of these metabolites (citric and isocitric acid) are part of the TCA cycle. In comparison, the TM algorithm falsely identified only 17 metabolites, reducing the number of false positive assignments by more than 50%. This highlights the substantial improvement offered by the TM algorithm for metabolite identification, gained by using prior knowledge of enzymatic relations between reactant pairs. Most reactant pairs comprised of a substrate and product with different empirical formulae (Figure 3.1c). Some reactions however involve only a structural rearrangement with no net loss or gain of atoms. These result in a false positive identification even for the TM algorithm; e.g. even though such structural rearrangements do not explicitly occur in the TCA cycle, the TM algorithm falsely identified maleic acid (not part of TCA cycle) as it is a structural isomer and product of correctly identified fumaric acid (part of TCA cycle). This resulted in one false positive identification of *m/z* 139.00018 ($[M + Na]^+$) for the TM algorithm yet, for the single-peak search, two falsely identified metabolites (two isomers for this *m/z* value) were identified, again highlighting the improvement offered by the TM approach.

Three mapping constraints can be used in the TM approach. Following the validation of the 'direct transformation' (default) constraint, see above, the other two constraints were applied to the TCA cycle both successfully identifying all 16 metabolites. Compared to the default constraint, the direct-or-indirect transformation constraint resulted in only one additional false identification, which is again ca. 50% lower than the number of false identifications using a traditional single-peak search. The more robust constraint, direct-and-indirect transformations, decreased the number of false identifications by three. Although this constraint increases the confidence in those metabolites that are identified, it has the potential to not identify a metabolite that is present in the mass spectrum (see Section 3.3.4). Overall, all three constraints significantly decreased the false positive identification error rate compared to a single-peak search.

3.3.2 Mass error surface for peak-pair differences

A novel semi-automated approach to estimate mass errors associated with peak differences in experimental mass spectra was developed (Figure 3.3). The resulting mass error surfaces, represented as both absolute and relative errors, are shown in Table 3.4. Errors were calculated for peak differences in the range *m/z* 0-80 only. This range was selected for two reasons. First, peak differences >80 *m/z* are not particularly abundant and are not so well distributed throughout the mass spectra. Second, as the empirical cumulative distribution of all substrate-product mass differences (selected from MI-DB) showed that ca. 80% of these are below 80 Da (Figure 3.4a), thus representing the most important range. In principle, the mass error surface could be extrapolated beyond this range using mass differences such as mono- to disaccharides, fatty acids to glycerides and lipids. A wider mass range of the mass spectrum is required as the abundance of these mass differences is particular low in the mass range *m/z* 70-500. However, exploiting other types of mass differences could be of high interest for lipid analysis where particular mass differences occur in high numbers in particular ranges of the mass spectrum. Nevertheless, here I focused on actual, more reliable empirical peak differences. The initial location of peaks assumed a maximum mass error of 1.5 ppm. If this value was unrealistically small, compared to the experimental error, the mass error surface would have underestimated the errors. However, Figure 3.4b clearly shows that all the experimental observations lie well within this error range. Having confirmed that I have not underestimated the errors, the considerable improvement (i.e. smaller mass error tolerances) from using empirically-derived mass errors of peak-pair differences compared to using a fixed maximum mass error on both peaks (Jourdan *et al.*, 2008) can be seen in Table 8.2; e.g. for a peak-pair spacing of 2.01565 *m/z* (i.e. H_2) in the range *m/z* 300.00000 to 302.01565, the mass error surface has a ca. 3-fold smaller error than a fixed 1.0 ppm error approach. It is also worth noting that the relative mass errors associated with peak differences can be dramatically larger than one might expect in high mass accuracy FT-ICR MS. For example, ^{13}C-^{12}C peak-pair differences near *m/z* 400 have an error of >300 ppm, reinforcing the importance of using an empirically derived error surface to maximise the accuracy of metabolite identification in the TM algorithm.

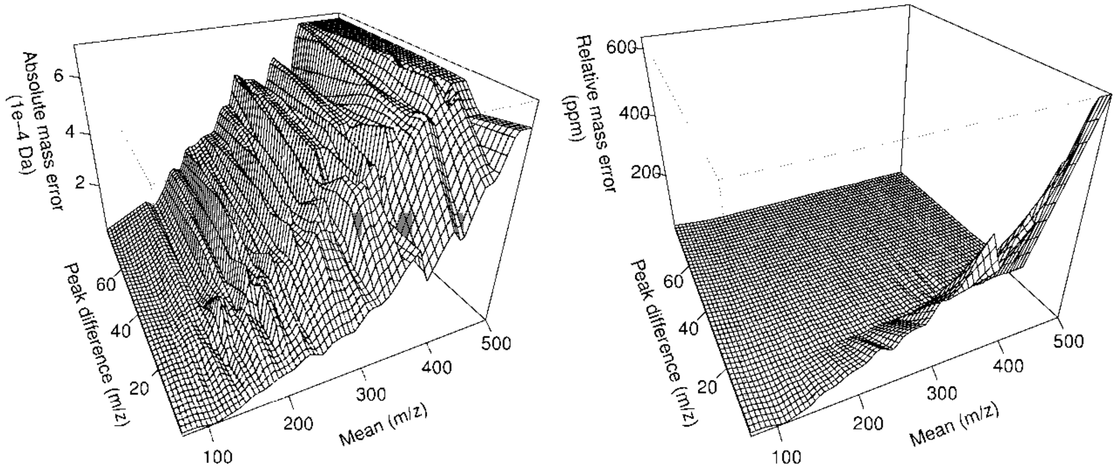


Figure 3.3 Mass error surface derived from experimental mass spectral data (human cancer cell extracts), which represents (a) absolute mass error (Da) and (b) relative mass error (ppm) associated with the mass difference between a pair of peaks. This error is a function of both the mass difference of the peak-pair as well as the mean of the *m/z* values of multiple peaks present within a sliding window between 70 and 500 Da.

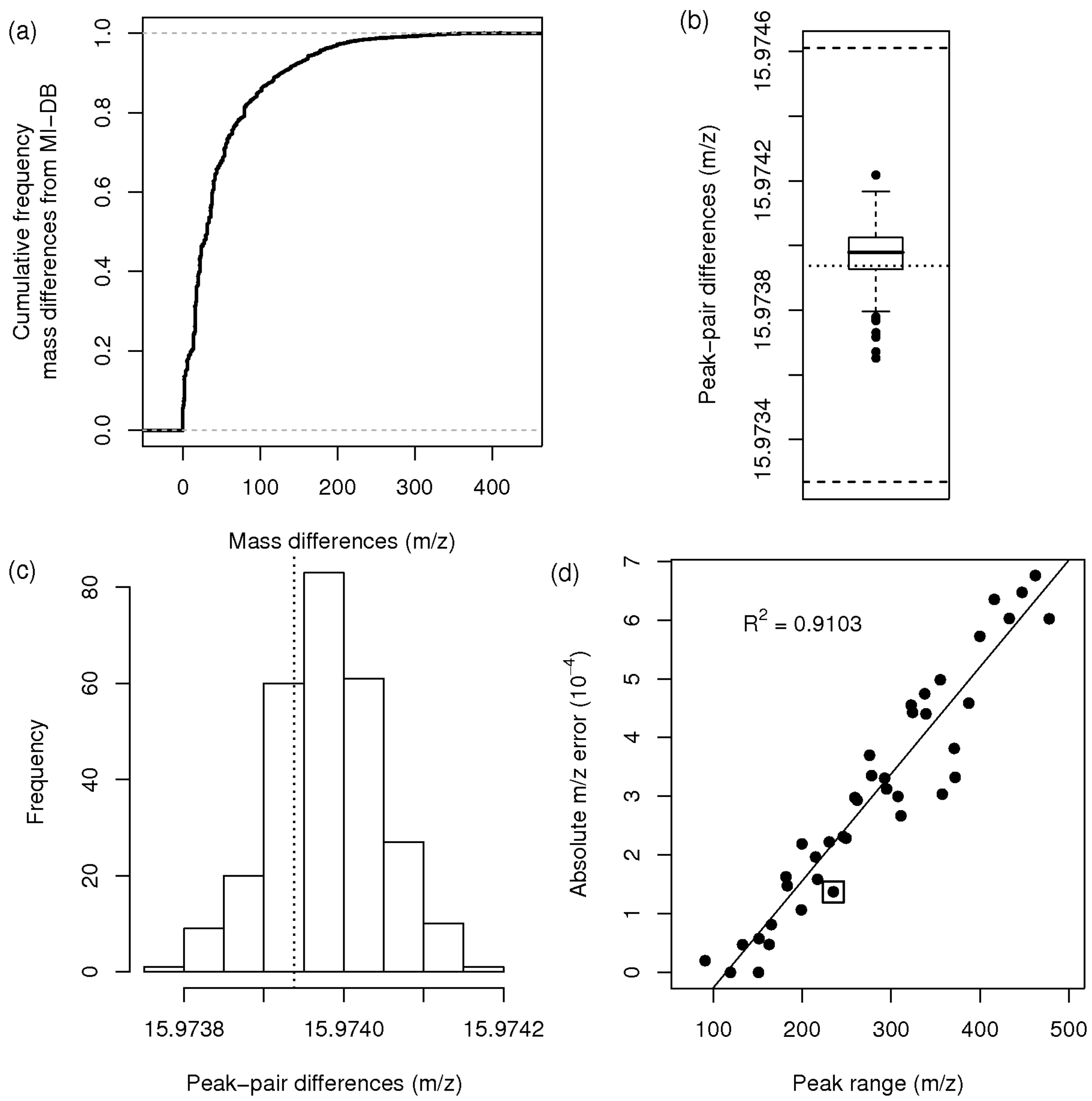


Figure 3.4 (a) Empirical cumulative distribution of all modified reactant-pair mass differences confirming that ca. 80% of substrate-product reactions involve a mass change of <80 Da. (b) Boxplot containing 280 $[M + Na]^+$ - $[M + {}^{39}K]^+$ peak-pair differences occurring in the range *m/z* 217-248; eight measurements were rejected as outliers. The horizontal dotted line illustrates the theoretical *m/z* difference of 15.97394, and the two horizontal dashed lines illustrate the fixed 1.5 ppm error boundaries. (c) Same data as in b (with outliers removed) represented as a histogram. The absolute mass error was estimated as twice the standard deviation of a fitted normal distribution (here, *m/z* ±0.00014). (d) All estimated absolute mass errors for the $[M + Na]^+$ - $[M + {}^{39}K]^+$ peak-pair differences. The error estimated from c is marked by a box.

3.3.3 Assessment of false positive rates for metabolite identification

Incorrect (i.e. false positive) metabolite identification represents a significantly underestimated problem in MS based metabolomics. Therefore, to examine and compare the robustness of the single-peak search (initially with default mass error ≤1.0 ppm) and TM identification strategies, two types of FPRs were determined: (i) the incorrect identification of non-human metabolites within experimental mass spectra of human cancer cell extracts, by comparing identification results between the MI-DB (i.e. for all species) and the human-specific MI-hDB; and (ii) the incorrect identification of random peaks as metabolites in simulated mass spectra. Table 3.4 summarises the numbers of observed peaks (in the mass spectra), de-adducted *m/z* values and named KEGG metabolites that have been putatively identified in these datasets, using both identification strategies. FPRs for incorrect assignment of non-human metabolites were calculated from the ratio of non-human assignments to the total number of assignments (e.g. for a single-peak search, 2116 metabolites - 678 / 2116 = 68.0% error rate; Figure 3.5a). This calculation provides a lower limit on the FPR since there may also be false assignments of some human metabolites. The default TM algorithm is >4 times more effective at not incorrectly identifying non-human metabolites compared to a single-peak search, with a FPR of only 15.9%. This considerable difference in error rates arises in part because the human metabolites in a human cancer cell are strongly related to each other *via* direct and indirect enzymatic transformations; i.e. the metabolites are linked in non-random networks. The two other mapping constraints in the TM algorithm had a logical but minor affect on the FPRs (Figure 3.5a). Considering the 3925 peaks in the cancer cell mass spectra, the single-peak search identified (including incorrect assignments) ca. 20%

while the TM algorithm identified ca. 11%, based on the MI-hDB. This relatively low number of assignments is caused by several factors, including that only three adducts ($[M + H]^+$, $[M + Na]^+$, $[M + {}^{39}K]^+$) were considered yet several others can form in the ion source (Draper *et al.*, 2009), that in-source fragmentation was not considered, and that the KEGG database is known to be incomplete. Even these low percentages however equate to >250 metabolites (or de-adducted m/z values) putatively identified in the mass spectra, using the TM algorithm, which is >5 times greater than for a typical NMR metabolomics study (Southam *et al.*, 2008; Tiziani *et al.*, 2009).

Table 3.4 Summary of the total number (3925) of observed peaks, "de-adducted" *m/z* values and named metabolites that were assigned in the final peak list of human cancer cells and in simulated random mass spectra (average of 51 random peak lists), using a single-peak search or TM algorithm with three different constraints. Two databases were employed, one containing metabolites from all species (MI-DB) and the other being human-specific (MI-hDB).

	Observed peaks		**"De-adducted" *m/z* values**		**Metabolite names**	
	DB	**hDB**	**DB**	**hDB**	**DB**	**hDB**
Human cancer cell extract peak list						
Single-peak search	795	437	646	305	2116	678
TM (Direct)	411	370	289	247	629	529
TM (Direct-or-indirect)	445	396	319	271	729	612
TM (Direct-and-indirect)	367	325	252	214	528	451
Simulated random peak lists						
Single-peak search	179	58	174	53	408	106
TM (Direct)	21	18	17	15	35	29
TM (Direct-or-indirect)	37	32	33	27	75	59
TM (Direct-and-indirect)	10	9	7	7	15	13

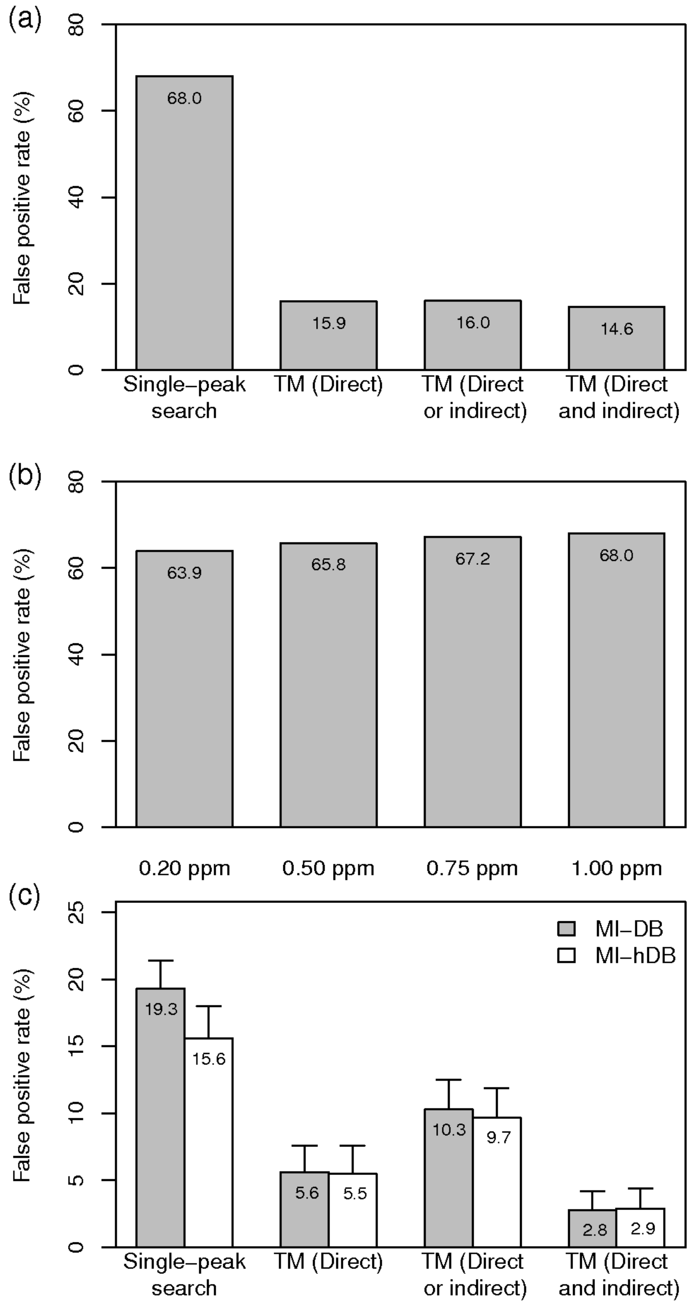


Figure 3.5 (a) Error rates for the false positive identification of non-human metabolites within human cancer cell mass spectra, for different identification algorithms. **(b)** Error rates for the false positive identification of non-human metabolites within human cancer cell mass spectra, for single-peak search using different ppm errors. **(c)** Error rates for the incorrect assignment of metabolites within simulated random mass spectra, for different identification algorithms and for the MI-DB and human-specific MI-hDB.

The calculations above for the single-peak search were based on matches with a mass error ≤1.0 ppm, which is a commonly used error boundary for metabolite identification in high resolution mass spectra (Kind & Fiehn, 2006; Kind & Fiehn, 2007). However, to provide a fair comparison between the single-peak search (with fixed mass error) and TM approach (with mass error surface), the single-peak search was repeated over a series of mass measurement errors. As anticipated, a smaller ppm error decreases the number of human and non-human named metabolite assignments from 2116 (for 1.0 ppm mass error) to 1629 (for 0.5 ppm), which increases the likelihood of missing correct assignments (i.e. increases the FNR). Against this, the FPR does decrease as a function of decreasing ppm mass error (Figure 3.5b), but only very slightly. This latter result in particular highlights the significant benefit of the TM approach (Figure 3.5a) *versus* a single-peak search, even when an unrealistically low 0.2 ppm mass error is employed for the single-peak approach.

Incorrectly assigning metabolites to random peaks and noise, which are inherent to mass spectra even when using strict noise filters during spectral processing (Payne *et al.*, 2009), also reduces the accuracy of metabolite identification. FPRs for incorrectly assigning random peaks were calculated by referencing the number of incorrect assignments (simulated random dataset) to the total number of assignments (cancer cell dataset); e.g. for a single-peak search using the MI-DB, 408 incorrect metabolite assignments / 2116 = 19.3% error rate (Figure 3.5c). This FPR was >3 times lower for the default TM algorithm, at only 5.6%. With the more robust mapping constraint this error decreases to 2.8%. If the human-specific MI-hDB is employed, the error rate decreases somewhat for the single-peak search only, but remains ca. 3 times larger than for the TM algorithm (Figure 3.5c).

Having considered two error rates associated with false metabolite identifications, I have clearly documented the significant advantage of the TM algorithm compared to a single-peak search. The enzymatic relationships between metabolites play an essential role in this improved performance, confirming that inclusion of prior biological knowledge provides one route to more robust metabolite identification. Furthermore, the source of prior knowledge also impacts on the error rates; e.g. for the default TM algorithm a total of 629 and 529 metabolites are putatively identified from the MI-DB and MI-hDB respectively. This difference almost certainly reflects a greater number of false positive assignments when using the considerably larger database. Therefore, a further strategy for improving the accuracy of metabolite identification is to use species-specific databases, though these are still somewhat scarce (Brown *et al.*, 2009).

3.3.4 Assessment of false negative rates for metabolite identification

As shown above, the TM algorithm identifies fewer metabolites in the mass spectra of human cancer cell extracts than a single-peak search (see Table 3.4). The data demonstrate that this results, at least in part, from the lower FPR of the TM algorithm. However the mapping constraints imposed by TM, that identified metabolites must be directly or indirectly enzymatically linked to other identified metabolites, might result in some metabolites within the cancer cells remaining unidentified (i.e. effectively a false negative assignment). Arguably, this could partly contribute to the reduction in the number of identified metabolites using the TM approach. To assess the FNR, I first confirmed the assignments of some metabolites within the same cancer cell extracts using a complementary analytical method, NMR spectroscopy. A total of 39 cellular metabolites were identified in the NMR spectra, of which 26 were subsequently confirmed to be present in the positive ion FT-ICR mass spectra using manual methods (using a 1.0 ppm mass error range and searching for only the $[M + H]^+$, $[M + Na]^+$ or $[M + {}^{39}K]^+$ adducts). The remaining 13 metabolites were not detected as ions of any of these three adduct forms, preferring to form negative ions only. The FNRs of a single-peak search and TM algorithm were evaluated by their effectiveness at identifying these 26 confirmed metabolites in the cancer cell mass spectra (see Table 3.5).

Table 3.5 Summary of the number of metabolites identified (and percentage of total), out of a maximum of 26 compounds confirmed by NMR spectroscopy, using a single-peak search and TM algorithm.

	No. of metabolites identified		
	with unique name	**with >1 name[a]**	**Not identified[b]**
Single-peak search	3 (11.5%)	23 (88.5%)	0 (0.0%)
TM (Direct)	8 (30.8%)	17 (65.4%)	1 (3.8%)
TM (Direct-or-indirect)	5 (19.2%)	21 (80.8%)	0 (0.0%)
TM (Direct-and-indirect)	6 (23.1%)	16 (61.5%)	4 (15.4%)

[a]i.e. includes false positive assignment, [b]i.e. represents a false negative assignment

The single-peak search identified all 26 metabolites, whilst the default TM algorithm successfully identified 25 of these compounds (yielding a FNR of 3.8%). This failure to identify citric acid, due to the mapping constraints of the default TM algorithm (direct transformations only), did not occur when a direct-or-indirect transformation constraint was used and all 26 metabolites were identified. Conversely, the more robust direct-and-indirect constraint increased the FNR to 15.4%. Considering those metabolites that were identified unambiguously, both the single-peak search and default TM algorithm identified glutathione, pantothenic acid and taurine, yet the TM algorithm additionally identified creatine, phosphoethanolamine, glycine, methionine and succinic acid as unique hits. This >2-fold increase in unambiguous identification further highlights the superior performance of the TM approach *versus* the traditional method. The remaining metabolite identifications were ambiguous in that the NMR-confirmed metabolites occurred at identical *m/z* values to other compounds. However, even for this case the TM algorithm yielded fewer false positive assignments than the single-peak search; e.g. the single-peak search resulted in 4 false positive and 1 true assignment of tyrosine, whereas the default TM algorithm yielded only 1 false positive and 1 true assignment. Considering all metabolites assigned to more than one name, the TM algorithm is >3 times more effective at decreasing the FPR. Consistent with our results above (default TM algorithm generated 50 named metabolites, and the single-peak search generated 151 names; see Table 8.3).

As anticipated I have demonstrated that the constraints imposed by the TM algorithm, *via* the inclusion of prior biological knowledge, causes a non-zero FNR for metabolite identification (i.e. if a metabolite is present in a sample that is not directly or indirectly enzymatically related to another identified metabolite, it will not itself be identified). However, this FNR of <5% is trivial relative to the advantage of the TM approach for drastically reducing the FPRs.

3.3.5 Applications – Putative identification of metabolites in biological HR FT-ICR mass spectra

Here three putative identification strategies were employed to three different studies. In the first strategy, empirical formulae were assigned to the peaks using an elemental composition calculator with appropriate constraints. In addition, in cases where carbon isotope patterns were detected, the ^{12}C-^{13}C ratio was used to calculate the approximate number of carbon atoms in the metabolite. The second strategy utilised the high mass accuracy of the FT-ICR measurements to match the observed *m/z* values of the peaks to putative metabolite identities in a modified KEGG LIGAND database. Finally, for the third strategy the TM algorithm was applied.

The number of matches (empirical formula(e) and metabolite names(s)) per peak ranged from zero to many and are show in Table 8.4 and 8.5. A summary of the total number of peaks observed, and the number of empirical formula(e) and metabolite name assignments are summarised in Table 3.6. When positive ion mode was used to record spectra (SIM-stitching acquisition parameters for the LTQ FT in this case), significantly less peaks were observed (see Table 3.6, study D1). This may reflect that the majority of the metabolites (i.e. chemical structures) in this particular sample type are more likely to form negative adducts during ionisation. Although the majority of peaks (73%-91%) in all three studies were assigned at least one empirical formula (including false positive assignments), several remained unassigned. This is likely to arise due to the presence of isotopes or alternative adducts which were not included in the assignment algorithms. A few hundred peaks were assigned to one or many metabolite names using TM or SPS.

The incompleteness of the KEGG database, used for both algorithms, is the underlying reason for the low proportion of putative identified peaks. Putative identification algorithms, such as TM, are highly dependent on the completeness of existing metabolic networks such as KEGG. Therefore, reconstruction of existing and novel metabolic networks is required to ultimately improve annotation of mass spectra. Furthermore, the number of assignments using SPS was significantly higher compared to TM. Dataset D1 and S1 have approximately the same number of peaks but different number of assignments (see Table 3.6). However, both MS datasets represent a different set of metabolites (D1: metabolome *versus* S1: AcrAB-TolC exported compounds into the media), which complicates the comparison. As shown in Chapter 3, ~68% of the SPS assignments are false-positives in comparison to ~16% for TM.

Several empirically-calculated numbers of carbon atoms derived from the ^{12}C-^{13}C intensity ratio are consistent with the actual empirical formula assignments (see Table 8.4 and Table 8.5). The incorporation of RIA characterisation into MI-Pack is essential future work as explained in Chapter 7. Note that none of the metabolite assignments reported in this Chapter can be regarded as definitive since they depend only on accurate mass and therefore do not fulfil the Metabolomics Standards Initiative criteria for metabolite identification (Sumner *et al.*, 2007).

Table 3.6 Summary of the total number of peaks in positive and/or negative ion FT-ICR MS datasets of *D. magna* (D1 and D2) and *S. typhimurium* (S1), and the number of those peaks that could be assigned empirical formula(e) and putative metabolite identities. D1, D2 and S1 correspond to the studies summarised in Sections 3.1.2.1, 3.1.2.2 and 3.1.2.3, respectively.

Number of peaks	**Positive ion mode**		**Negative ion mode**	
Study	**D1[a]**	**S1**	**D1[a]**	**D2[b]**
Total (after signal filtering)[c]	1848	2051	3599	5627
Assigned to ≥1 empirical formulae	1478 (80%)	1487 (73%)	3290 (91%)	5007 (89%)
Assigned to unique empirical formula	675 (37%)	393 (19%)	231 (6%)	469 (8%)
Assigned to ≥1 KEGG LIGAND ID	283[SPS] (15%)	115[TM] (6%)	734[SPS] (20%)	158[TM] (3%)

[a] Following removal of assignments in which the match was a non-endogenous metabolite such as a drug, plasticiser or pesticide (Taylor *et al.*, 2009), [b] Average number over 8 peaklists, [c] Payne *et al.* (2009), [SPS] Single-peak search, [TM] Transformation Mapping.

In addition to the general discussions above regarding putative identification of the three MS datasets, it is worth mentioning several other successful findings that were previously published in *Metabolomics* (Taylor *et al.*, 2009) and *Toxicological Sciences* (Taylor *et al.*, 2010).

Automated putative metabolite identification of thousands of peaks in the MS dataset used for investigating copper toxicity in *D. magna* has resulted in a list of putative metabolites that are potentially involved in oxidative stress (Taylor *et al.*, 2009). These were selected based on top weighted peaks calculated using a PLS-DA model. Although, the list included mainly amino acids, one particular metabolite named *N*-acetylspermidine was putatively identified as a novel marker for copper toxicity. Based on this particular metabolite and the available related literature the following hypothesis was constructed by Taylor *et al.* (2009): "We hypothesise that the known induction of reactive oxygen species (ROS) due to copper exposure also increases the activity of spermidine/spermine *N*-acytyltransferase (SSAT) in *D. magna*, resulting in the observed increase in *N*-acytylspermidine". Although the focus of this study was the optimisation and evaluation of the effectiveness of DI FT-ICR MS metabolomics for toxicity testing in *D. magna*, the ability to determine and putatively identify metabolic changes has a huge potential in the future of ecotoxicogenomics research.

A subsequent investigation into whether different acute chemical toxicities can be discriminated via changes in the *Daphnid* metabolome also included putative identification of metabolites (Taylor *et al.*, 2010). In this case eight large MS datasets were annotated and investigated (see Table 8.4). Based upon the putative annotation and significant fold changes of peak intensities between different sample classes (controls and toxicants), several observations were reported. For example, a unique increasing trend of multiple fatty acid and oxylipid metabolites was induced by propranolol (Taylor *et al.*, 2010). This observation was then more extensively investigated (using KEGG pathways and other putatively identified metabolites) and linked to the disruption of the eicosanoid biosyntheses pathway (KEGG pathway ko00590). This example of metabolic biomarker discovery again clearly shows the benefit of putative identification of MS datasets.

Finally, the overall aim of the third application was to identify natural substrate molecules exported by AcrAB-TolC to provide new data on the molecular basis of pathogenicity of *S. typhimurium*. First, the 1478 peaks were putatively assigned to one or more empirical formulae and 675 of these assignments were assigned to a single empirical formula. Applying the TM algorithm to the 2051 peaks resulted in 115 peaks being assigned to one or many metabolite name(s). Although further analysis is required, this dataset of putatively identified peaks has the potential to guide the study to identify natural substrate molecules exported by AcrAB-TolC and construct novel hypothesis, as shown for the two *D. magna* studies above.

3.4 Conclusions

Metabolite identification is of central importance to all metabolomics studies as it provides the route to new knowledge. Here I have documented extremely high FPRs of metabolite identification (up to 68%) for the traditional method of automated database searching accurate masses on a single-peak-by-peak basis. I therefore confirm the recent conclusion of Draper *et al.* (2009) that the annotation of a large proportion of peaks in high resolution mass spectra of metabolite mixtures is not possible by this simple approach. Biological samples are not, however, composed of random mixtures of metabolites, but instead comprise of thousands of compounds that are related through specific chemical transformations that ultimately form large metabolic networks. Here I have demonstrated how prior biological knowledge of transformations between substrate-product pairs, extracted from the KEGG database, can significantly decrease FPRs of metabolite identification while maintaining minimal FNRs compared to the single-peak search. Integral to this improved accuracy of metabolite identification is a novel method to determine the mass errors associated with peak-pair differences, which generates a mass error surface. Overall, our transformation mapping approach with mass error surface (using the default constraint of mapping to direct transformations only) is >4 times more effective at not incorrectly identifying non-human metabolites in a human sample (error rate of 15.9%), and >3 times more effective at not incorrectly identifying random peaks as metabolites (error rate of 5.6%). Furthermore, by comparing the FT-ICR MS data against NMR measurements of the same biological samples, I have confirmed a low FNR for metabolite identification of only 3.8%. Our findings also confirm that error rates can be reduced by using a species-specific database, and demonstrate that the FPR and FNR are dependent on the accuracy of the database(s) employed. Since all metabolite databases and metabolic reconstructions are currently incomplete, in particular for lipid metabolism (Nookaew *et al.*, 2008), further improvements of these resources are urgently required. The methods reported here could, following minor alterations (e.g. data compatibility and integration of other data sources), utilise reconstructions of metabolic networks and therefore further contribute to the field of systems biology. Furthermore, although this study has focused on positive ion mass spectra, the methods and principles are equally applicable to negative ion data by considering the different types of ions formed (e.g. $[M - H]^-$ and $[M + {}^{35}Cl]^-$ instead of $[M + H]^+$ and $[M + Na]^+$). Overall, I conclude that inclusion of prior biological knowledge in the form of known metabolic pathways provides one route to more accurate putative metabolite identification in MS based metabolomics. The next step towards more robust metabolite identification is to include further experimental measurements, such as retention time data from LC-MS studies.

In addition, I have demonstrated the application of several algorithms and analytical methods (see Chapter 4) for putative identification of metabolites. This has been achieved for two different sample types (i.e. *D. magna* and *S. typhimurium*), using two different metabolomics strategies, fingerprinting and footprinting. The MS datasets recorded in positive ion mode resulted in significantly less peaks in comparison to the datasets recorded in negative ion mode. This suggests that certain chemical structures are more likely to form negative ions. Between 6% and 37% of the peaks in the different datasets were assigned to a single empirical formula and the percentage of peaks assigned to one or many empirical formulae varied from 73% up to 91%. The highest percentage of single empirical formulae assignments was found for mass spectra recorded in positive ion mode. However, even when appropriate constraints were applied to filter down the number of empirical formula(e) assignment(s) many false positives were found (Kind & Fiehn, 2007). A small percentage of the peaks of each dataset were assigned to one or many metabolite names (SPS ≤20% and TM ≤6%). Incompleteness of the database used for identification is likely the main reason for this observation, and therefore reconstruction of metabolic networks, as already mentioned, is highly important to improve the accuracy and completeness of metabolite identification. Nevertheless, the incomplete putative identification of MS data has still shown to be successful for guiding ecotoxicogenomics research as shown in the two studies by Taylor *et al.,* (2009, 2010).

CHAPTER FOUR:

RE-OPTIMISATION OF SIM-STITCHING PARAMETERS FOR THE LTQ FT ULTRA[5]

[5] Parts of this chapter, including Figure 4.2, 4.3, 4.5, 4.8 and table 4.2, have been published in *Analytical Chemistry* (Weber *et al.,* 2011).

4.1 Introduction

HR FTMS, including FT-ICR MS (Marshall *et al.*, 1998) and Orbitrap technologies (Hu *et al.*, 2005), has become a leading analytical platform in metabolomics (Junot *et al.*, 2010). Analytical methods that maximise the sensitivity and mass accuracy are essential to enable the detection of as large a proportion of the metabolome as possible and to assist in metabolite identification. To achieve this goal we previously developed the SIM-stitching method for direct infusion HR FTMS, which is based upon the collection and subsequent 'stitching' together of multiple adjacent SIM scans (Southam *et al.*, 2007).

The upgrade of our LTQ FT to an LTQ FT Ultra, with a larger trap volume in its ICR detector cell, provided an opportunity to re-optimise the SIM-stitching acquisition parameters and further improve the quality of metabolomics analysis. The more advanced ICR detector cell, of the LTQ FT Ultra, allows a larger radius of ion cyclotron motion, therefore considerably more charges can be trapped without an increase in space-charge effects. This results in higher sensitivity without compromising mass accuracy.

The objectives to achieve this goal are therefore to carry out the analytical tests developed by (Southam *et al.*, 2007) for wide-scan, high dynamic range DI FT-ICR MS analysis of complex biological mixtures of low molecular weight metabolites. The optimisation process was applied to a chemically defined mixture of PEG, as well as to real biological samples. The chemically defined mixture enabled assessment of the mass accuracy of the Thermo Scientific LTQ FT Ultra as a function of the number of ions transferred into the ICR detector cell, and was also used to assess the mass accuracy of the SIM-stitching method. Liver extracts were used to determine the ideal width of each SIM window, and to assess the improvement in dynamic range by comparing the total number of peaks detected using the former SIM-stitching approach and the re-optimised SIM stitching approach. In addition, the intensity profiles across SIM-stitched mass spectra were investigated as done previously by (Payne *et al.*, 2009).

4.2 Material and methods

4.2.1 Preparation of standards and biological samples for HR FTMS

PEG standards, fish liver extract samples, BioPEG (i.e. polar metabolites from a fish liver extract resuspended in a 80:20 v/v methanol:water solution with 0.25% formic acid, 0.0005% PEG200 and 0.0005% PEG600 (Sigma-Aldrich, UK)) were prepared as described in more detail in Section 2.1.

4.2.2 HR FTMS

Analyses were conducted using a hybrid 7-Tesla linear ion trap FT-ICR mass spectrometer (LTQ FT Ultra, Thermo Scientific, Germany) equipped with a Triversa chip-based nanoelectrospray ion source (Advion Biosciences, NY, USA). Nanoelectrospray conditions (controlled by ChipSoft software version 8.1.0, Advion Biosciences) comprised 0.5 psi backing pressure, +1.7 kV electrospray voltage and ca. 200 nL/min sample flow rate. All analyses were recorded in positive ion mode. A mass resolution of 100,000 and maximum linear ion trap fill-time of 1 sec were fixed throughout the analyses (see Section 2.2 for more details).

4.2.3 Optimisation of number of ions transferred to ICR cell

To optimise the AGC target, BioPEG or PEG alone was directly infused into the mass spectrometer and a mass range of *m/z* 300-400 was acquired using the wide SIM mode at R=100,000. The AGC target was varied from 1×10^5, 2.5×10^5, 5×10^5, 7.5×10^5, 1×10^6, 2.5×10^6, 5×10^6 to 1×10^7. In addition, the scan mode "narrow SIM" (Thermo Scientific terminology) was investigated using the standard fixed window width of 30 Da for a mass range of *m/z* 340-370. Analyses were conducted in triplicate (n=3 mass spectra of prepared PEG standards).

4.2.4 Optimisation of SIM window size

Several "wide scan" (Thermo Scientific terminology) SIM window sizes centered on *m/z* 200 were investigated (i.e. 200, 150, 100, 75, 50, 40, 30 Da) as shown in Figure 4.1. The number of transients recorded for AGC targets of 5×10^5 and 1×10^6 was kept in proportion to the size of the window (i.e. ca. 15 transients for 100 Da window and ca. 30 transients for 200 Da window), so that the overall acquisition time is constant and therefore comparable. The average peak count was calculated over several shared *m/z* regions (*m/z* 185-215, 185-190, 190-195, 195-205, 205-210 and 210-215), as shown in Figure 4.1. Analyses were conducted in triplicate (n=3 mass spectra of fish liver extract samples).

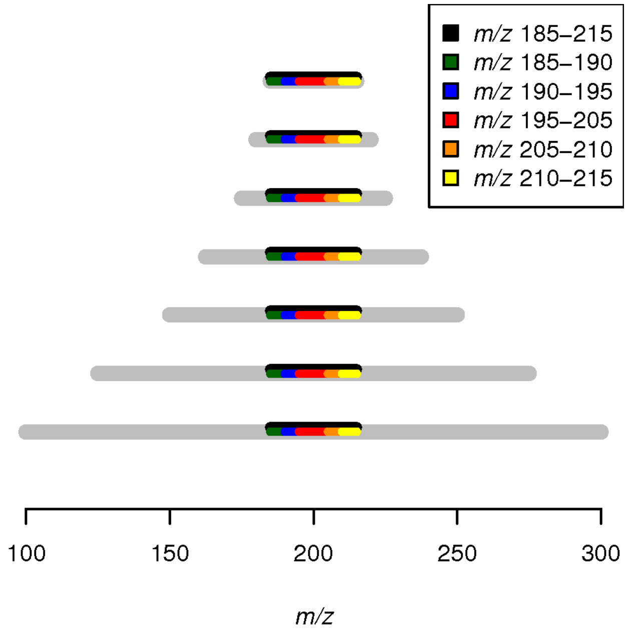


Figure 4.1 Experimental design for the study that re-optimised the SIM window size by counting the number of peaks within several shared *m/z* ranges (six different colours) for several different window sizes (grey, *m/z* 30, 40, 50, 75, 100, 150 and 200).

4.2.5 Ion intensity across the SIM window

To evaluate signal loss at the ends of each SIM window (termed 'edge-effects' (Payne *et al.*, 2009; Soule *et al.*, 2010; Southam *et al.*, 2007), several wide SIM scans of *m/z* 100 (each shifted by 5 *m/z*, Figure 4.2) were recorded using an AGC target of 1×10^6. The intensities of several carbon isotope-pairs were measured at different positions throughout the SIM scan to determine the extent of the edge-effects (Figure 4.2).

4.2.6 Quantification of dynamic range in biological mass spectra

The following window widths using scan mode "wide scan" and an AGC target of 1×10^6 were compared to quantify the overall sensitivity (i.e. dynamic range) and mass accuracy in biological mass spectra: *m/z* 75, *m/z* 125 and *m/z* 150. Overlapping regions for each of these window sizes were altered (see Table 4.1) to keep the total *m/z* range of the final SIM-stitched mass spectra approximately the same. The overall acquisition time was kept to 5.25 min which is identical to the protocol for the LTQ FT (Southam *et al.*, 2007). Analyses were conducted in triplicate (n=3 mass spectra of prepared BioPEG sample). Peaks were counted in the mass range of *m/z* 80 – 485 after signal filtering was applied (Payne *et al.*, 2009), i.e. a SNR threshold of $\geq$7 and only peaks that occur in at least 2-out-of-3 replicate mass spectra were maintained, both used previously for mass spectra collected on the LTQ FT (Southam *et al.*, 2007)).

Table 4.1 Parameters used for the comparison of observed number of peaks for the original (Southam *et al.*, 2007) and the re-optimised SIM-stitching protocols. *removed from each end of the SIM window.

Size of SIM window (*m/z*)	**Overlap (*m/z*)**	**Total range (*m/z*)**	**Number of SIM Windows**	**AGC target**	**Mass spectrometer**
30	10 (5*)	70-500	21	1×10^5	LTQ FT
75	40 (15*)	70-495	11	1×10^6	LTQ FT Ultra
100	45 (15*)	70-500	7	1×10^6	LTQ FT Ultra
150	57 (15*)	70-499	4	1×10^6	LTQ FT Ultra

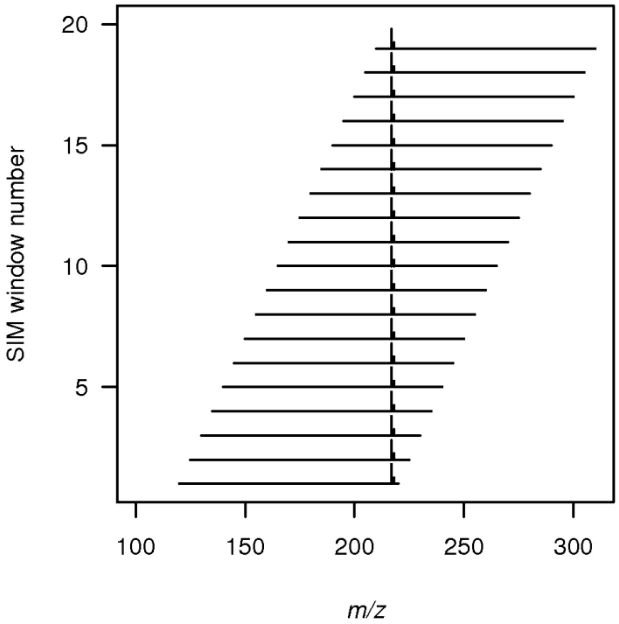


Figure 4.2 Schematic representation of the experimental design to characterise 'edge effects' for a *m/z* 100 wide SIM scan. Several PEG isotope pairs were measured across the window, in this case the figure shows one particular PEG isotope pair as an example (*m/z* 217.10465 and *m/z* 218.10800). This enabled the intensities of these peaks to be measured as a function of their position within the SIM scan.

4.2.7 Data processing

Mass spectra collected for re-optimisation were processed using the SIM-stitching algorithm (Southam *et al.*, 2007) (see Chapter 2 for more details).

4.2.8 Calibration of mass spectra

Spectra collected for optimisation of the number of ions transferred to the ICR cell (section 4.2.3) and the quantification of dynamic range in biological mass spectra (section 4.2.6) were internally calibrated, either by a single calibration using all identified PEG peaks or by multiple calibration using calibrants chosen at random. For multiple calibration (Section 4.2.3), calibrants (18 PEG peaks in total, *m/z* 300-400) occurring within spectra (i.e., PEG peaks ($[(C_2H_4O)_n+H_2O+H]^+$, $[(C_2H_4O)_n+H_2O+Na]^+$, $[(C_2H_4O)_n+H_2O+K]^+$ and carbon isotopes of all three) within ≤1.5 ppm mass error tolerance were randomly divided into two equal subsets. One subset was used for calibration (ca. 9 PEG peaks) and the other for mass error measurements (ca. 9 PEG peaks). To retrieve reliable mass error measurements this process was repeated 100 times. Other mass spectra were non-random internally calibrated using the calibrants mentioned above (Sections 4.2.4 - 4.2.6).

4.3 Results and Discussion

4.3.1 Optimisation of the number of ions transferred to the ICR cell

The more advanced ICR detector cell, of the LTQ FT Ultra, allows a larger radius of ion cyclotron motion, therefore considerably more charges can be trapped without an increase in space-charge effects. This results in higher sensitivity without compromising mass accuracy. In theory, the AGC target (i.e. the number of charges to be transferred from the front-stage ion trap to the ICR detector cell) would be set extremely low to achieve near perfect mass accuracy. However, lowering the AGC target reduces sensitivity and so a compromise must be reached. Therefore, the number of charges transferred from the LTQ to the ICR detector cell was optimised to maximise sensitivity whilst minimising the mass error. AGC target values of 1×10^5 up to 1×10^6 yielded similar mass errors (root-mean-squared (RMS) error of 0.12-0.18 ppm and maximum absolute error of 0.27-0.61 ppm, at R=100,000; Figure 4.3). As the AGC target increased beyond 1×10^6 a trend towards higher RMS errors emerged, likely caused by increased space-charge effects. Therefore an AGC target of 1×10^6 was selected corresponding to an RMS mass error of 0.16 ppm and maximum absolute mass error of 0.29 ppm. This maximum mass error was considerably better than the 0.48 ppm error recorded previously on the LTQ FT, while the RMS mass error was similar (Southam *et al.*, 2007). Scan mode "narrow SIM" was also investigated, however, the fixed window size of 30 Da for this scan mode resulted in a low number of observed calibrants for random internal calibration (e.g. an increase in AGC target ($\geq 1\times10^5$) resulted in the disappearance of calibrants at the edges of the SIM window). Internal calibration is essential for measuring the decrease of mass accuracy caused by space-charge effects and for that reason scan mode “narrow SIM” was not investigated further in this study. The primary advantage of this newly optimised method is that 10 times as many charges enter the ICR detector cell thereby allowing increased detection sensitivity.

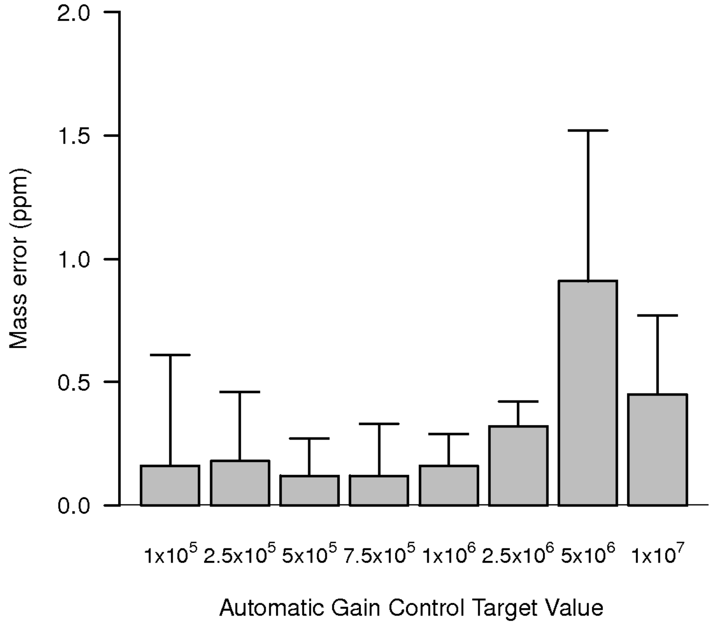


Figure 4.3 Effect of the number of ions transferred to the ICR detector cell on the absolute mass accuracy of 9 non-calibrated PEG compound *m/z* measurements, after (n=100) random internal calibrations, in a *m/z* 100 wide SIM window (*m/z* 300-400). A total of 18 PEG compounds (i.e. 9 of them used for calibration *versus* 9 of them used to measure mass accuracy) were used, including $[(C_2H_4O)_n + H_2O + H]^+$, $[(C_2H_4O)_n + H_2O + Na]^+$, $[(C_2H_4O)_n + H_2O + {}^{39}K]^+$, $[(C_2H_4O)_n + H_2O + {}^{41}K]^+$ and carbon isotopes of all four. Solid bars denote the average root-mean-squared (RMS) mass error, and the error bars represent the average maximum absolute mass error.

4.3.2 Optimisation of SIM window size

The next optimisation step included the maximisation of the number of ions detected with the optimised AGC target of 1×10^6 and also with an additional AGC target of 5×10^5. Fish liver extract samples were analysed using seven increased SIM window sizes, each sharing a common *m/z* 30 region centered on the theoretical *m/z* value of 200 (Figure 4.2). Although both AGC targets have a similar trend in the average number of observed peaks, the highest number of average peaks was observed for an AGC target of 1×10^6, which confirms the findings in Section 4.3.1 (see Figure 4.4). As shown by Southam *et al.* (2007) the number of peaks observed increases for SIM windows with a narrower mass range. A similar observation was also observed for the average peak counts in our study, but only for the shared regions *m/z* 195-205 in windows size up to 100 Da (red) and in particular for an AGC target of 5×10^5 (e.g. 150 peaks for 30 Da window *versus* 86 peaks for a 100 Da). Surprisingly the average number of observed peaks for this particular *m/z* region increases again for window sizes ≥ *m/z* 150 (yellow). This trend was also found for the regions *m/z* 205 - 210 and *m/z* 210 – 215, except that the increase in the average number of peaks started at a window size of 100 Da. Furthermore, the average number of peaks for several shared regions for a window size 30 Da (i.e. *m/z* 185-190, *m/z* 205-210 and *m/z* 210-215) and 40 Da (i.e. *m/z* 185-190 and *m/z* 210-215, particular for 1×10^6) were proportionally smaller in comparison to the other SIM window sizes. This is a result of the decrease in signal at the lower- and upper-end of the window. This also explains the difference in average peak count over the total shared region (black). Edge effects were characterised up to 15 Da at each end of the window (see Section 4.3.3). Therefore SIM window sizes *m/z* 30, *m/z* 40 and *m/z* 50 were excluded from further investigation as more than 50% of the window is discarded to avoid edge effects. Overall, both 5×10^5 and 1×10^6 AGC targets demonstrated a similar profile as mentioned before except for the region *m/z* 195-205 (red). For an AGC target of 5×10^5 there was a decrease in the number of peaks for the first 3 window sizes (i.e. *m/z* 30, 40, 50), which was not the case for an AGC target of 1×10^6. The observation of similar peak counts for relatively small window sizes (i.e. *m/z* 10 up to 30) was also mentioned by Southam *et al.* (2007). By considering that, (i) ~200 ions of the same metabolite are required to generate a detectable signal in the ICR cell (Marshall *et al.*, 1998), (ii) a total of only 1×10^5 ions are transferred to the ICR cell as determined by the AGC, and (iii) the narrower the SIM window the fewer different metabolites are transferred to the cell and, therefore, the greater the likelihood that a low-abundance metabolite will be present in sufficient ion number to be detected. It appears, however, that the reduction in acquisition time for windows narrower than *m/z* 30 counteracts any sensitivity benefit that would be gained by a further reduction of window size (i.e. there is insufficient signal averaging to increase the SNR to detect more peaks). In conclusion, window sizes *m/z* 75, *m/z* 100 and *m/z* 150 were selected for further investigation.

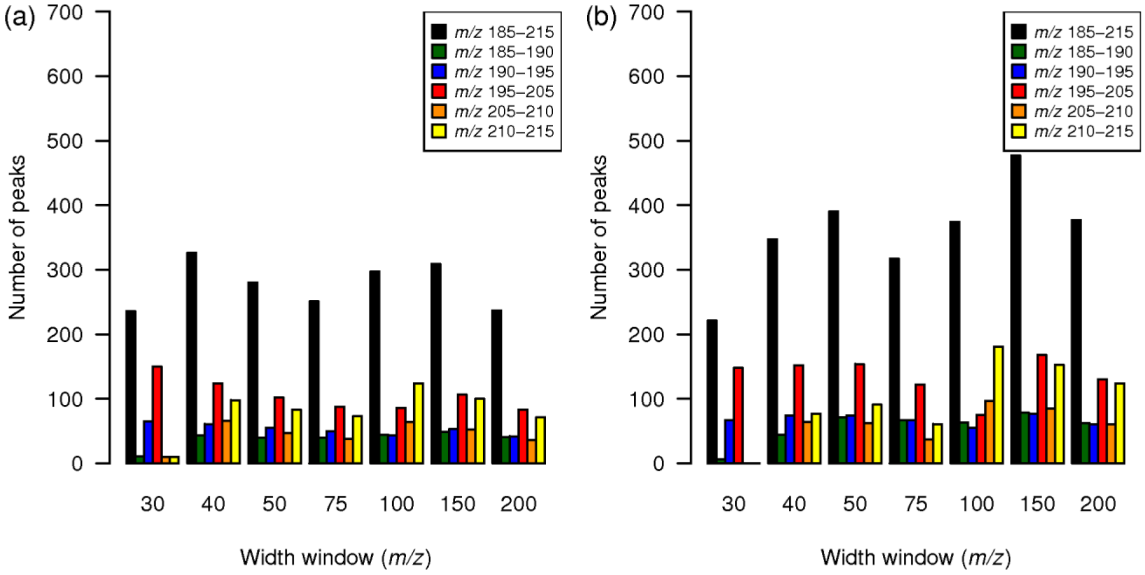


Figure 4.4 Effect of window size on the average number of peaks in several shared *m/z* ranges detected in BioPEG for two different AGC targets: **(a)** 5×10^5 and **(b)** 1×10^6.

4.3.3 Ion Intensity across the SIM window

The loss of signal intensity at the ends of a SIM window (‘edge effect’) was characterised by measuring the intensities of several PEG isotope pairs at different locations across an *m/z* 100 SIM window (Figure 4.2). The empirically-derived carbon differences of these PEG carbon isotope pairs (theoretical number of carbons *versus* experimental derived carbons, see Chapter 5) increases or decreases dramatically in the lower and upper 15 Da regions of each scan (Figure 4.5) in comparison to the empirically-derived carbon differences derived from the centre of the SIM window. Finally, these regions were chosen to be excluded during spectral processing so that they did not contribute to the SIM-stitched mass spectrum, which resulted in an overlapping region of *m/z* 30 for each adjacent window.

The intensity across a 100 Da SIM scan was also investigated for several groups of peaks (see Figure 4.6) Similar profiles were measured (e.g. measured signal intensity increases as the peak moves across the SIM window) as reported previously by Payne *et al.* (2009) (see Figure 4.6a). However, several additional profiles were observed. For example, signal intensities that were constant for the first half of the window, increased in the second half of the window (see Figure 4.6d). The profiles also show the “edge effect” reported above, which is presented by the disappearance or decrease of signal intensity in the lower and upper ends of the window. Next, two different SNR thresholds (i.e. ≥3.5 and ≥35) were used to examine if these signal intensity profiles are affected by noise. Although, profiles based on peaks selected with a less strict SNR filter were somehow noisier, the overall profiles of both SNR thresholds are similar (see Figure 4.6). The trend of the gradients of all signal intensity profiles is shown in Figure 4.7 and is considerably different from the increasing gradient shown previously (Payne *et al.*, 2009). Therefore the increasing linear function that was parameterised by Payne *et al.* (2009) does not fulfil the correction of signal intensities measured by the re-optimised SIM-stitching protocol. Further investigation would be needed to parameterise this new signal intensity trend.

The quality of SIM-stitched mass spectra is partly dependent on the performance of waveform ion-isolation method used in the linear ion trap to select ions in a certain *m/z* range (i.e. SIM window) prior to detection in the ICR cell. Several waveform ion-isolation methods for the linear ion trap have been reported previously with mass-selective instability filtering the most commonly applied for linear ion traps as it is here (Song *et al.*, 2009; Soni & Cooks, 1994). This type of ion filtering is typically done by adjusting the RF and DC components of the linear ion trap to cause excitation of a range of ions. Unwanted ions are then subsequently ejected from the ion trap leaving the trapped ions of interested behind. It is not uncommon that ions close to the edges of the RF/DC stability field are discriminated or completely ejected due the field imperfection, which causes the “edge effects” as shown in Figures 4.4 and 4.5 (Song *et al.*, 2009). Additionally, it is not unlikely that the type of waveform ion-isolation method in some manner has an effect on the overall signal intensity profile of the SIM window as shown in Figures 4.6 and 4.7.

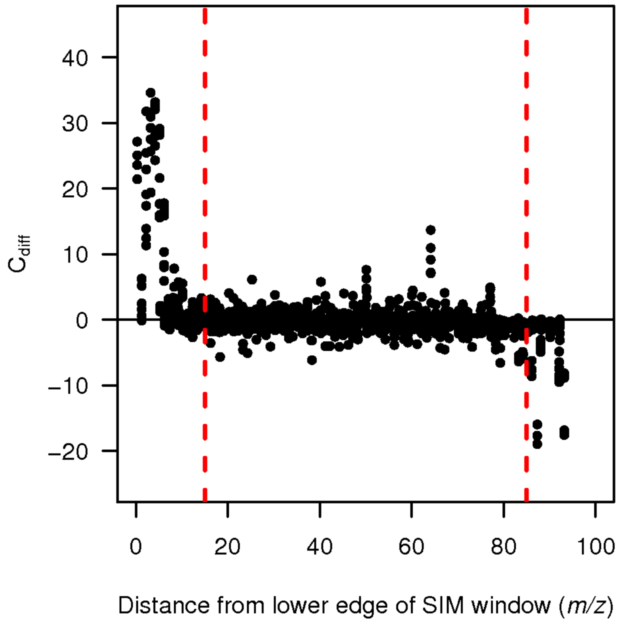


Figure 4.5 C_{diff} (see Chapter 5) for several PEG isotope pairs in the range *m/z* 70-590 measured across multiple SIM window *m/z* ranges (measured as shown in Figure 4.2), showing larger and therefore less accurate C_{diff} values for isotope pairs measured close to the lower and upper edges of the SIM window. The vertical dashed lines, *m/z* 15 from each end, indicate the recommended cutoff for *m/z* 100 wide SIM windows.

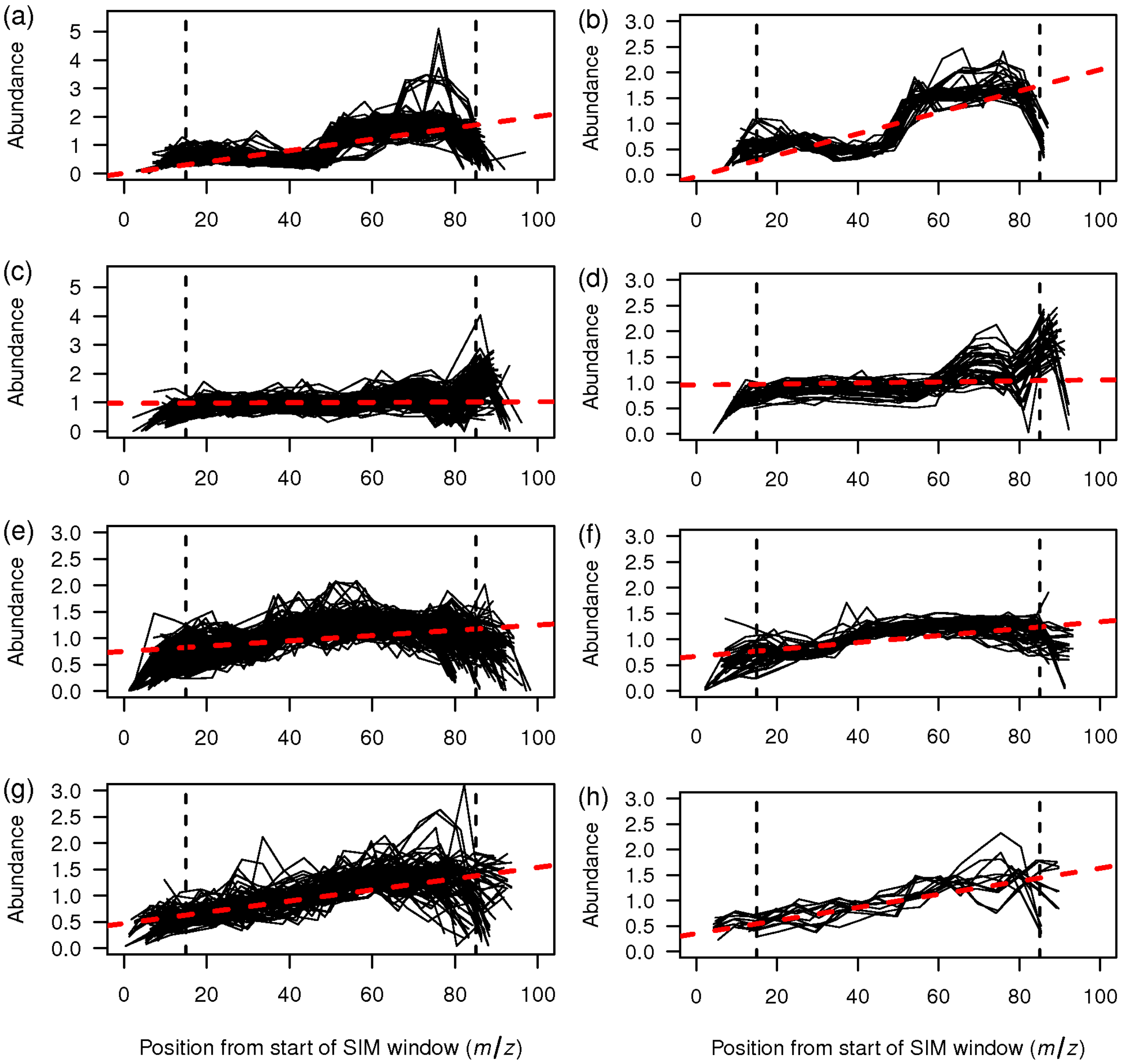


Figure 4.6 The abundance of several signals relative to their mean values as measured across multiple SIM windows (100 Da) with two different SNR thresholds (four figures on the left side: ≥3.5 and four figures on the right side: ≥35.0). (a, b) peaks in the range *m/z* 185-195 (c, d) peaks in the range *m/z* 295-305 (e, f) peaks in the range *m/z* 435-445 (g, h) peaks in the range *m/z* 515-525.

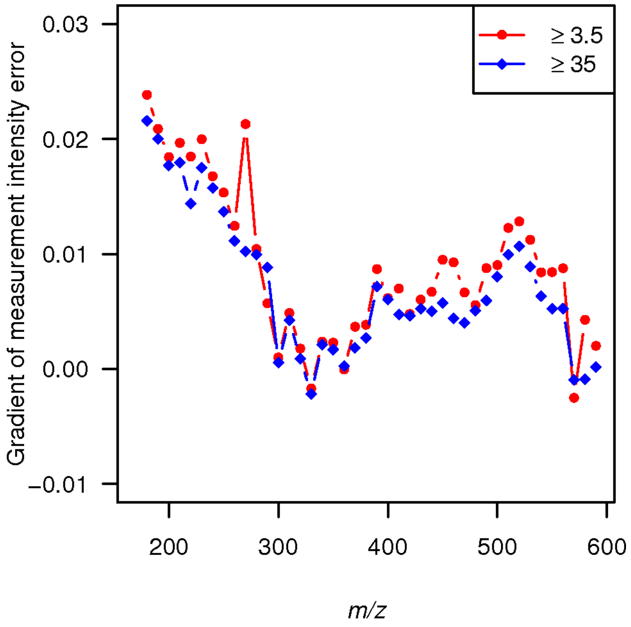


Figure 4.7 The gradient of the measurement intensity error for a selection of peaks for two different signal-to-noise ratio thresholds, ≥3.5 and ≥35, located between 70 and 600 *m/z*.

4.3.4 Quantification of dynamic range in biological mass spectra

The final step was to optimise the SIM window width to maximise the number of peaks detected (using AGC target of 1×10^6 and removing 15 *m/z* from each end of each SIM window). After signal filtering (Payne *et al.*, 2009), the LTQ FT Ultra detected ca. 3 times more peaks than was achieved previously (Southam *et al.*, 2007) with relatively little dependency on the SIM window width (Figure 4.8). A SIM window width of *m/z* 100 was selected, with the final optimised method using 7 SIM windows overlapping by *m/z* 30, to cover *m/z* 70-590. This region contains the highest density of polar metabolites. See Table 4.2 for summary of optimal acquisition parameters. Given that the final method comprised of fewer SIM windows than before, the total acquisition time was much reduced (now only 2.25 min/sample), which along with improved sensitivity and mass accuracy, is advantageous for high-throughput metabolomics.

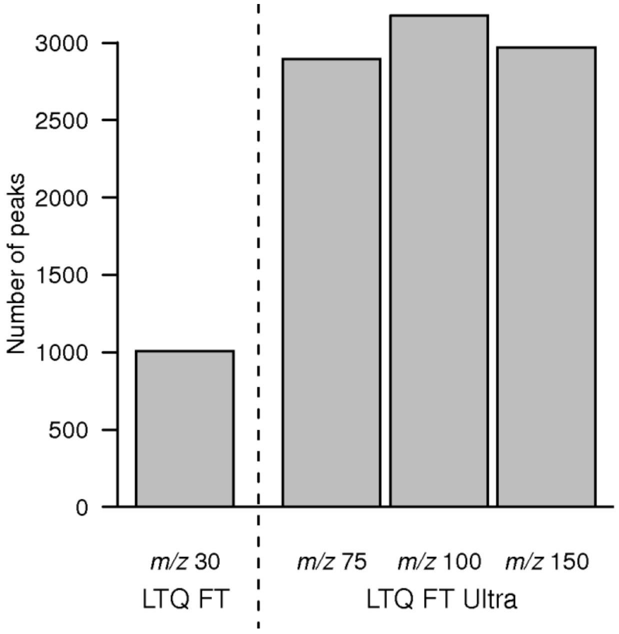


Figure 4.8 Number of peaks detected in positive ion direct infusion FT-ICR SIM-stitched mass spectra of liver extracts following 2-out-of-3 replicate filtering (Payne *et al.*, 2009). This highlights the ca. 3-fold increase in sensitivity using the re-optimised SIM-stitching method on an LTQ FT Ultra mass spectrometer compared to the original method on an LTQ FT (Southam *et al.*, 2007). Re-optimisation shows relatively little dependency on the SIM window width, with *m/z* 100 selected for the final method.

Table 4.2 Parameters for direct infusion SIM-stitching method implemented originally on a Thermo LTQ FT mass spectrometer (Southam *et al.*, 2007) and re-optimised here on a Thermo Scientific LTQ FT Ultra.

Parameter	**LTQ FT**	**LTQ FT Ultra**
AGC target	1×10^5	1×10^6
SIM scan range	*m/z* 30	*m/z* 100*
Overlap of SIM scans	*m/z* 10 (*m/z* 5 removed from each end)	*m/z* 30 (*m/z* 15 removed from each end)
Time for each SIM scan (no. of transients)	15 sec (10)	15 sec (10)
Total range	*m/z* 70-500	*m/z* 70-590
Total number of overlapping SIM scans	21	7
Total acquisition time per sample **	5 min 45 sec	2 min 15 sec

*Scan mode: Wide SIM. **Including a 30 s start delay of dummy scans.

4.4 Conclusions

Direct infusion SIM-stitching was re-optimised for the LTQ FT Ultra, resulting in a ca. 3-fold increase in sensitivity, shorter analysis time (2.25 min *versus* 5.75 min), and ultra-high mass accuracy (rms error 0.16 ppm and maximum absolute error of 0.29 ppm) compared to our previous method implemented on a LTQ FT (rms error 0.18 ppm and maximum absolute error of 0.48 ppm). Intensity across a wide scan SIM window and a total SIM-stitched mass spectrum were investigated. The intensities of ions were severely reduced up to 15 Da at each end of the wide scan SIM window. To compensate for this "edge effect" adjacent windows were overlapped by a region of 30 Da, and 15 Da was discarded from each end. This results in seven adjacent windows representing a mass spectrum from *m/z* 70 up to *m/z* 590. The overall trend of the gradients of all signal intensity profiles of several peaks across a SIM-stitched mass spectrum was considerably different from the increasing gradient shown previously (Payne *et al.*, 2009). Further investigation is required to parameterise this new signal intensity trend.

CHAPTER FIVE:

CHARACTERISATION OF ISOTOPIC ABUNDANCE MEASUREMENTS IN HIGH-RESOLUTION FT-ICR AND FT-ORBITRAP MASS SPECTRA FOR IMPROVED CONFIDENCE OF METABOLITE IDENTIFICATION[6]

[6] The contents of this Chapter, including the figures, have been published in *Analytical Chemistry* (Weber *et al.,* 2011).

5.1 Introduction

The high mass accuracy and resolution of HR FT MS instruments, together with strategies to improve their dynamic range, e.g. SIM-stitching (Southam *et al.,* 2007; Chapter 4), provide the critical combination of specifications for DIMS to investigate the metabolome in detail (Soule *et al.*, 2010). However, automated identification of the thousands of signals arguably remains the greatest challenge in metabolomics. This includes assigning one (or more) empirical formula(e) to each peak in the mass spectrum, assigning metabolite name(s) to these empirical formula(e) and, for definitive identification, further analytical measurements such as, liquid chromatography and/or fragmentation (MS^n) (Brown *et al.*, 2011; Wolf *et al.*, 2010)(Brown *et al.*, 2011; Heinonen *et al.*, 2008; Wolf *et al.*, 2010).

Here I focus on the primary step in metabolite identification, namely assigning empirical formula(e) to accurate mass measurements. High mass accuracy greatly facilitates this process by allowing a small mass error tolerance during searches, thus reducing the number of potential empirical formula(e) found. Heuristic rules and restrictions can be applied to remove incorrect elemental compositions (e.g. the number and type of atoms, and atom ratios), ultimately yielding the most likely empirical formula(e) (Kind & Fiehn, 2007). However, the number of potential empirical formulae per peak increases significantly with *m/z*; e.g. whereas *m/z* = 148.00389 can be assigned to a single formula, *m/z* = 450.01864 has 47 possible assignments (for ≤1 ppm mass error tolerance and $[C_{0\text{-}34}H_{0\text{-}72}N_{0\text{-}15}O_{0\text{-}19}P_{0\text{-}7}S_{0\text{-}8}\text{+H or Na or K}]^{+}$). Consequently, further strategies must be employed to reduce false-positive assignments and improve confidence in metabolite identification. Relative isotopic abundance (RIA) measurements (from isotope-pairs such as $^{12}C_n$ and $^{13}C^{12}C_{n-1}$, or $^{32}S_n$ and $^{34}S^{32}S_{n-1}$) can in principle be highly informative for assigning empirical formula(e) by providing an estimate of the numbers of atoms present. Kind & Fiehn (2006) and others reported a considerable benefit of incorporating RIA into empirical formula(e) determination (Kaufmann, 2010; Miura *et al.*, 2010); e.g. for a peak at *m/z* = 500.00000 (assuming 3 ppm mass accuracy), 33 potential empirical formulae can be assigned, yet on adding RIA measurements (assuming 2% accuracy) this decreases to only 3 possible assignments. This demonstrates the value of RIA measurements in cases where the intensity measurements are accurate, but how appropriate is this assumption for HR FTMS instruments used in metabolomics?

It has been demonstrated that FT-ICR MS RIA measurements of a PEG solution are relatively poor above *m/z* 500, with absolute errors in the number of carbons ranging from 0 to 96 (Stenson *et al.*, 2003). Using the same type of instrument, a precision of ±1.6 carbons was reported for peaks <500 *m/z* with a SNR ≥25 for a natural organic mixture but without defined chemical standards, but this became considerably poorer for peaks with SNR <25 (Koch *et al.*, 2007). Both studies reported that digital resolution and the *bias* of closely spaced cyclotron frequencies may lead to imprecise RIA measurements. Evaluation of RIA accuracy on an LTQ Orbitrap (*m/z* >600) showed that intensity error measurements increased when resolution was increased, with 3% accuracy for a resolution of R=7,500 and increasing to 10% for R=100,000 (Xu *et al.*, 2010). It has also been reported that absolute mean RIA errors for an LTQ Orbitrap are 16% for positive ion mode and 12% for negative ion mode (Xu *et al.*, 2010). In general, these RIA errors are considerably greater than those used by Kind & Fiehn (2006), drawing into question the value of isotopic intensity measurements in HR FTMS metabolomics.

Furthermore, none of the FT-ICR MS studies used the ideal sample type for characterising RIA, i.e. a biological extract spiked with defined chemical standards, leaving the benefit of RIA measurements for peak annotation in metabolomics uncertain. This is reiterated by Junot *et al.* (2010), who stated "…there is very limited information available in the literature about the ability of mass spectrometers to achieve accurate measurements of relative isotopic ion abundances".

Here I seek to determine the accuracies and precisions of RIA measurements on two leading HR FTMS platforms to definitively evaluate the extent to which such measurements can improve automated metabolite identification. To achieve this I had two specific objectives. Initially, I sought to characterise the accuracies of RIA measurements on the FT-ICR MS (including at ultra-high resolution: R=100,000-750,000) and Orbitrap instruments (R=100,000), using the re-optimised stitching parameters (Chapter 4) and biological samples spiked with PEG standards. The second objective was to assess the efficacy of RIA measurements in the metabolomics workflow by calculating the extent to which these measurements reduce the number of empirical formulae in automated metabolite identification.

5.2 Material and methods

5.2.1 Preparation of standards and biological samples for HR FTMS.

A BioPEG solution, polar metabolites from a fish liver extract resuspended in a 80:20 v/v methanol:water solution with 0.25% formic acid, 0.0005% PEG200 and 0.0005% PEG600 (Sigma-Aldrich, UK), was prepared as described in more detail in Section 2.1.

5.2.2 HR FTMS

Analyses were conducted using either a hybrid 7-Tesla linear ion trap FT-ICR mass spectrometer (LTQ FT Ultra, Thermo Scientific, Germany) or a hybrid LTQ Orbitrap Velos mass spectrometer (Thermo Scientific) both equipped with a Triversa chip-based nanoelectrospray ion source (Advion Biosciences, NY, USA). Nanoelectrospray conditions (controlled by ChipSoft software version 8.1.0, Advion Biosciences) comprised 0.5 psi backing pressure, +1.7 kV electrospray voltage and ca. 200 nL/min sample flow rate (see Section 2.2 and Table 4.2 for more details).

To determine RIA accuracies for the LTQ FT Ultra (n=100 mass spectra, at R=100,000) and LTQ Orbitrap Velos (n=30 mass spectra), using the optimised acquisition parameters (see Chapter 4), BioPEG was analysed at R=100,000. In addition the effect of resolution on RIA error was assessed by analysing BioPEG in positive ion mode on the LTQ FT Ultra, at R=100,000, 200,000, 400,000 and 750,000 (n=3 mass spectra each).

5.2.3 Data processing

Transient MS data from the LTQ FT Ultra were processed using the SIM-stitching algorithm (Southam *et al.* (2007)). Spectra were internally calibrated (i.e. PEG peaks ($[(C_2H_4O)_n+H_2O+H]^+$, $[(C_2H_4O)_n+H_2O+Na]^+$, $[(C_2H_4O)_n+H_2O+K]^+$ and carbon isotopes of all three) within ≤1.5 ppm mass error tolerance). Data collected on the Orbitrap were processed as externally calibrated raw files (using Xcalibur, Thermo Scientific) as the transient data was not available. To facilitate comparison of the LTQ FT Ultra and

Orbitrap data, the LTQ FT Ultra data was additionally processed as externally calibrated raw files in Xcalibur.

$$C_{diff} = \frac{100 \times \text{relative abundance } {}^{13}\text{C}{}^{12}\text{C}_{\text{n-1}} \text{ PEG peak}}{1.10 \times \text{relative abundance } {}^{12}\text{C}_{\text{n}} \text{ PEG peak}} - \textit{theoretical no. of carbons}\,\text{PEG}_{\text{parent}} \quad (1)$$

RIA accuracy calculations in FT mass spectra were based mainly on intensities ratios of carbon isotope-pairs of known PEG peaks. Therefore, all observed peaks in each mass spectrum were compared to a list of theoretical masses for parent (${}^{12}C_n$) and isotopic (${}^{13}C{}^{12}C_{n-1}$) PEG standards. Only isotope-pairs for which both peaks have a mass error of <1.0 ppm (for internally calibrated spectra) or <1.5 ppm (externally calibrated spectra) were used. The intensities of these carbon isotope-pairs were used to estimate the theoretical number of carbon atoms present. C_{diff} was calculated by subtracting the theoretical number of carbons from the empirically-calculated value (1). The same procedure was repeated for oxygen and potassium RIA measurements assuming ${}^{18}O{}^{16}O_{n-1}$: 0.20% natural abundance and $[M + {}^{39}K]^+$-$[M + {}^{41}K]^+$: 6.73%. Presentation of RIA measurements in this format makes the interpretation informative and easy to implement, in particular for defining ranges to identify empirical formula(e) with confidence. In addition to C_{diff}, RIA errors as a percentage were calculated for the Orbitrap RIA measurements to allow comparison to previous studies (Xu *et al.*, 2010).

Empirical formula(e) were calculated for each parent PEG peak (from the isotope-pairs) that appeared in >50% of the BioPEG LTQ FT Ultra mass spectra. The type and number of atoms were restricted as follows: $[C_{0\text{-}34}H_{0\text{-}72}N_{0\text{-}15}O_{0\text{-}19}P_{0\text{-}7}S_{0\text{-}8}$+H or Na or K$]^+$ with a mass error tolerance of ≤1 ppm. Neutral empirical formula(e) ($C_6H_{12}O_6$ and not

$[C_6H_{12}O_6+H]^+$) were then filtered using heuristic rules (Kind & Fiehn (2007)). Formula(e) calculations, data mining and statistics were performed using the R-scripting language and Python, including MI-Pack (see Section 2.4.1).

5.3 Results and Discussion

5.3.1 Accuracy of relative isotopic abundance measurements by FT-ICR MS

The matrices of biological samples used in metabolomics studies can complicate HR FTMS measurements, e.g. by introducing ion suppression and/or enhancement (Annesley, 2003). Here, to maximise applicability to real samples, the characterisation of RIAs was performed on known PEG peaks spiked into a biological matrix. Furthermore, the PEG standards were diluted to cover the intensity distribution of the biological metabolites. Specifically, as shown in Figure 5.1, the two SNR distributions are sufficiently similar which makes this study relevant to biological mass spectra collected using Thermo Fisher FT instruments (in particular the LTQ FT Ultra). Carbon-13 is the most widely detected natural isotope in mass spectra of complex metabolite mixtures and therefore the focus here. The PEG-spiked biological extracts were analysed repeatedly (n=100 mass spectra) over a wide mass range (*m/z* 70-590), generating ca. 2700 $^{12}C_n$ - $^{13}C^{12}C_{n-1}$ PEG peak-pair intensity measurements, which were used to calculate the number of carbons in the standards. The calculated C_{diff} values are strongly dependent on the SNR of $^{13}C^{12}C_{n-1}$-containing PEG peaks (Figure 5.2). For example, low intensity peaks (SNR of $^{13}C^{12}C_{n-1}$ $\geq$3.5 and <50) exhibit a standard deviation (SD) of the error in the number of carbons of 2.38; mid intensity peaks (SNR $\geq$50 and <250) have a SD=1.52 carbons; whilst high intensity peaks (SNR $\geq$250 and <500) exhibit a SD=0.68 carbons. Interestingly, the mean of the error in the number of carbons for each of these three groups of peak intensities is -0.41, -0.15 and -0.52, respectively. These negative values imply that most of the empirically-calculated numbers of carbons are lower than the theoretical values, an effect reported previously for FT-ICR MS (Figure 5.3) (Koch *et al.*, 2007). This offset is observed even for measurements in which the $^{12}C_n$ and $^{13}C^{12}C_{n-1}$ peak intensities are high and very highly correlated (R^2≈1.0; Figure 5.3c). In addition to ^{13}C measurements, both ^{18}O and ^{41}K isotopes were detected for several PEG standards, the latter as $[M + {}^{41}K]^+$ adducts (Figure 5.3b and d). Note that these findings are not dependent on the data pre-processing methods (i.e. our custom-written SIM-stitching algorithms) as similar results were produced using Xcalibur 2.1 (see Figure 5.4 and Figure 5.5). The development of methods for pre-processing mass spectral data (including peak integration) is important to improve the accuracy of RIA measurements, as presented by Xu *et al.* (2010) To facilitate comparisons between datasets the data obtained by pre-processing using the SIM-stitching algorithm versus Xcalibur (version 2.1) were compared (Figure 5.4). In total, approximately 30% more carbon isotope pairs (≥3.5 SNR for $^{13}C^{12}C_{n-1}$ peak) were measured using the SIM-stitching algorithm (n=40 mass spectra). For carbon isotope pairs observed using both pre-processing methods and using peak height as a measure of quantity, the following means and standard deviations of the carbon errors were calculated: SIM-stitching -0.54±1.54 carbons and Xcalibur -1.19±1.51 carbons. Based upon these preliminary results the SIM-stitching algorithm appears to outperform Xcalibur by detecting more isotope pairs and with greater accuracy. Peak height versus integrated peak area (SIM-stitching algorithm only) was also compared and no significant differences were observed (Figure 5.5), except for one outlier that was caused by two overlapping peaks increasing the area of the ^{13}C-containing peak (Figure 5.6).

Overall, our initial conclusion is that RIA measurements can provide useful information on the numbers of atoms for the more intense peaks, which again highlights the importance of using analytical methods that maximise signal intensity such as the SIM-stitching approach.

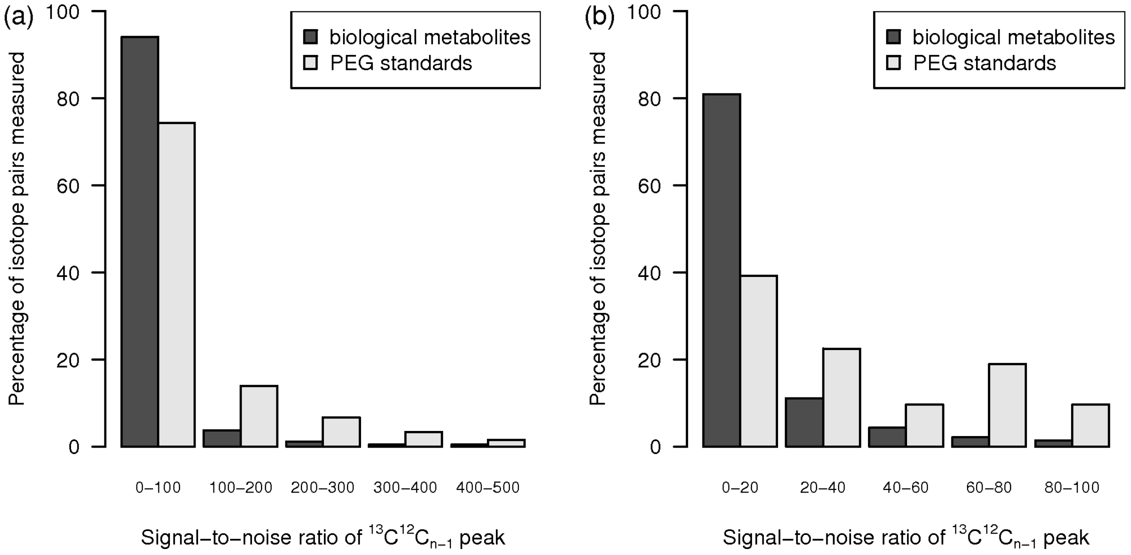


Figure 5.1 Distribution of signal-to-noise ratios for several PEG and biological metabolite $^{13}C^{12}C_{n-1}$ measured signals in the range *m/z* 70-590 (n=100 mass spectra of BioPEG). (**a**) SNR of $^{13}C^{12}C_{n-1}$ peak ≤500. (**b**) SNR of $^{13}C^{12}C_{n-1}$ peak ≤100.

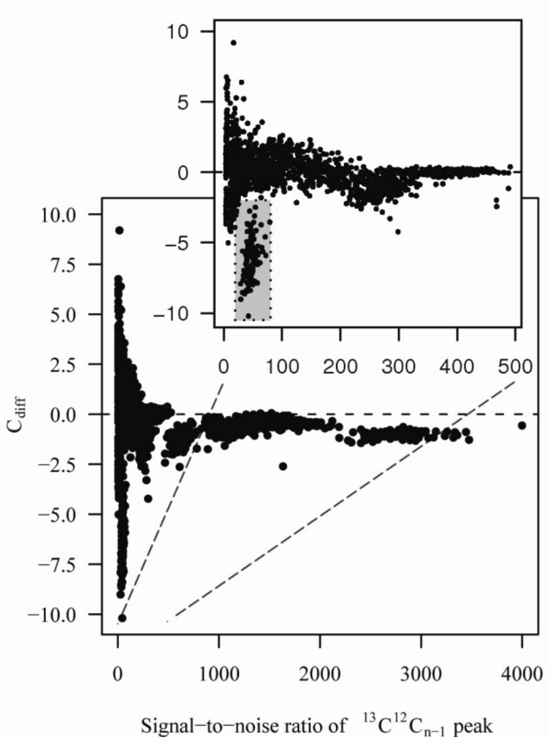


Figure 5.2 Characterisation of the accuracy of RIA measurements on the LTQ FT Ultra. C_{diff} values (empirically-calculated number of carbons minus actual number) of several PEG standards, measured in a biological sample matrix, are plotted against the signal-to-noise ratios of the $^{13}C^{12}C_{n-1}$-containing PEG peaks. Inset shows a subset of the data, highlighting the particularly large RIA errors at low signal-to-noise. One group of PEG isotope-pair measurements (*m/z* = 503.30621, $[(C_2H_4O)_{11}+H_2O+H]^+$) was visually classified as an outlier (gray box). All spectral processing was conducted using the SIM-stitching algorithm (Southam *et al.*, 2007).

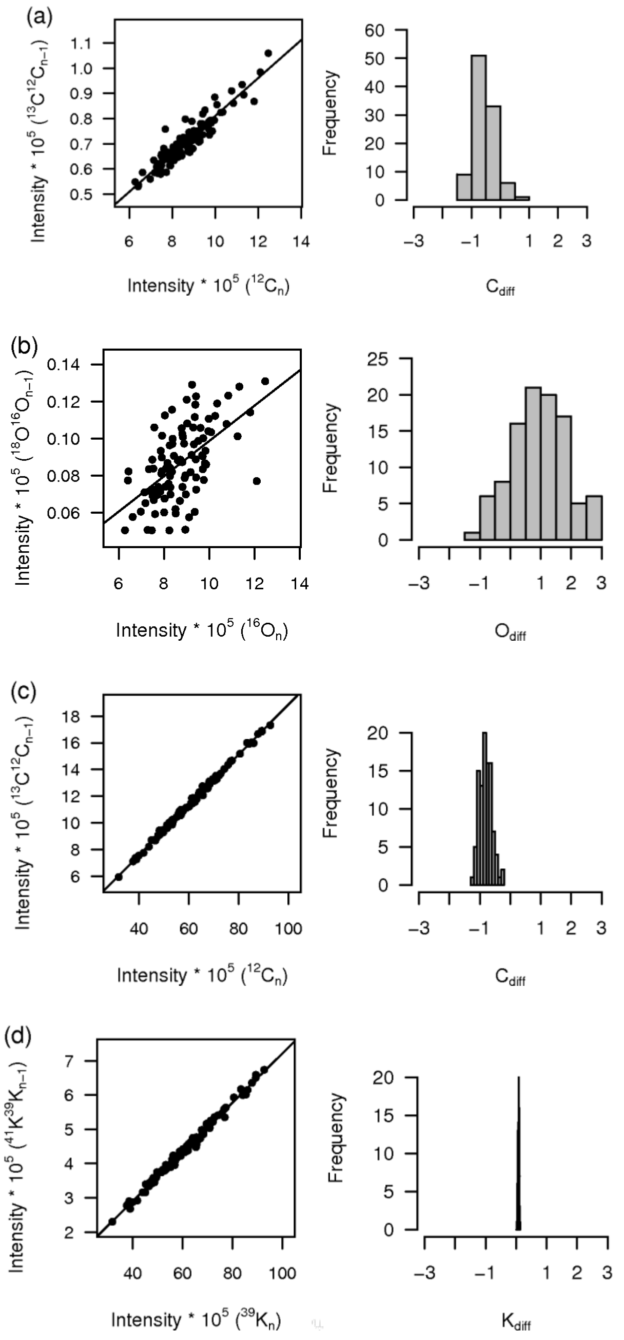


Figure 5.3 Intensity correlations between isotope pairs and associated C_{diff} distributions for two (parent) PEG peaks. (**a**) *m/z* 195.12270; $[^{12}C_8H_{18}O_5 + H]^+$ and $[^{13}C^{12}C_7H_{18}O_5 + H]^+$; R2=0.89; (**b**) *m/z* 195.12270; $[C_8H_{18}{}^{16}O_5 + H]^+$ and $[C_8H_{18}{}^{18}O^{16}O_4 + H]^+$; R2=0.39; (**c**) *m/z* 453.20966; $[^{12}C_{18}H_{38}O_{10} + K]^+$ and $[^{13}C^{12}C_{17}H_{38}O_{10} + K]^+$; R2=1.00; (**d**) *m/z* 453.20966; $[C_{18}H_{38}O_{10} + {}^{39}K]^+$ and $[C_{18}H_{38}O_{10} + {}^{41}K]^+$; R2=0.99.

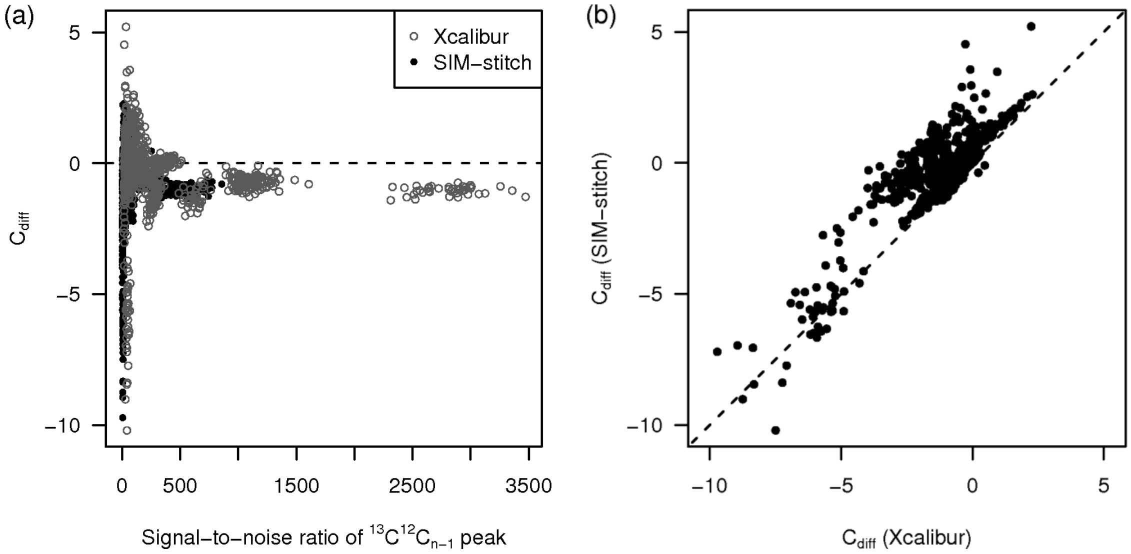


Figure 5.4 Comparison of the accuracy of isotope intensity measurements on a LTQ FT Ultra, using two types of data pre-processing (Xcalibur and the SIM-stitching algorithm (Southam *et al.*, 2007)). (a) C_{diff} values for several PEG isotope pairs in the range *m/z* 70-590 *versus* the signal-to-noise ratio of the $^{13}C^{12}C_{n-1}$ peak. **(b)** Bivariate plot of C_{diff} values indicating the similarity of the two pre-processing methods.

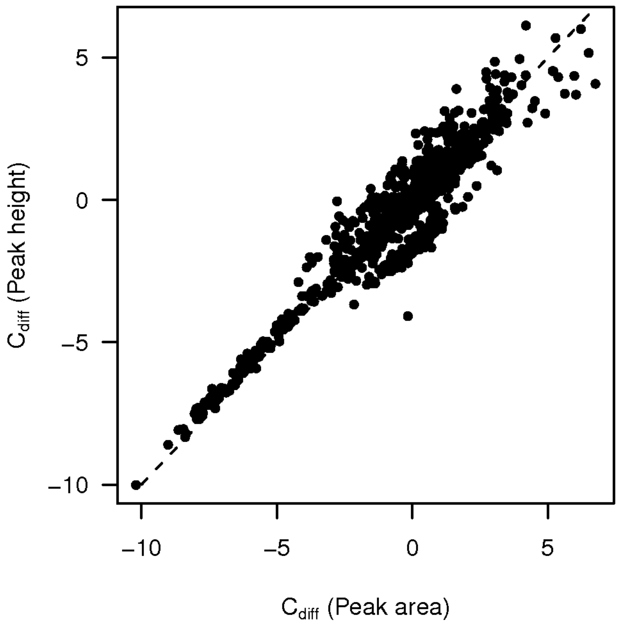


Figure 5.5 Bivariate plot of C_{diff} values of several PEG isotope pairs in the range *m/z* 70-590 for two different measures of signal quantity (i.e. peak height and peak area). Analyses were performed using an LTQ FT Ultra and all spectral processing was conducted using the SIM-stitching algorithm (Southam *et al.*, 2007).

Next, the effect of increasing mass resolution on RIA errors was evaluated, with the expectation that higher resolution decreases the likelihood of overlapping peaks and therefore improves the accuracy of RIA measurements (Soule *et al.*, 2010). Several resolution settings (100,000 to 750,000) were investigated on the LTQ FT Ultra, resulting in a stepwise decrease in peak widths (Figure 5.6). The increased resolution was, however, offset by longer signal acquisition times of ions in the ICR (doubling the resolution requires double the acquisition time of each SIM scan), making ultra-high resolution analyses less attractive for high-throughput metabolomics. Furthermore, contrary to expectations, increased resolution led to significantly greater errors between empirically-calculated and theoretical numbers of carbons (Figure 5.7a). This shift towards larger negative errors as resolution increases, is consistent with that reported previously for the LTQ Orbitrap.(Xu *et al.*, 2010) I rationalise that this is mainly caused by the reduction in signal intensities in the ultra-high resolution spectra. Said reduction is at least partially caused by the signal deterioration that occurs when the acquisition time is increased, as longer acquisition times necessarily record a greater amount of noise, overall reducing the SNR. Arguably, the acquisition of additional SIM scans would offset this deterioration of signal intensity, but the longer acquisition times would be less compatible with high-throughput FTMS analysis. This reduction in signal intensities is clearly evident from the decrease in the number of $^{12}C_n$ - $^{13}C^{12}C_{n-1}$ PEG isotope-pairs detected as resolution increases (Figure 5.7b), with ca. 50% fewer isotope-pairs at R=750,000 *versus* R=100,000. The reduction in SNR will clearly have an adverse effect on the RIA errors, based upon our earlier findings (Figure 5.2). Several other decreasing and/or increasing trends regarding ultra-high resolution and SNR were observed (Figure

5.8) that may relate to isotopic beat patterns (Easterling *et al.*, 1999; Erve *et al.*, 2009) but further investigation would be needed to clarify these trends. From these data, it can be concluded that a mass resolution of 100,000 is most appropriate for FT-ICR MS metabolomics based upon the need for high-throughput and accurate RIA measurements to assist metabolite identification.

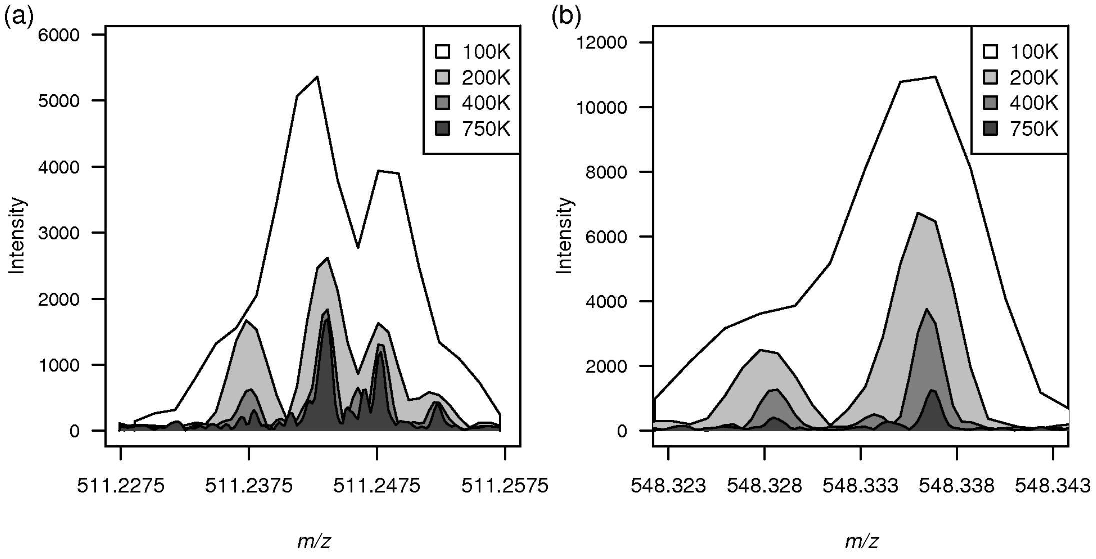


Figure 5.6 Effect of resolution on the peak widths of several PEG and unidentified biological metabolite signals measured using a Thermo Scientific LTQ FT Ultra. Resolution was varied from 100,000 up to 750,000, significantly reducing the peak widths but also decreasing the signal intensities.

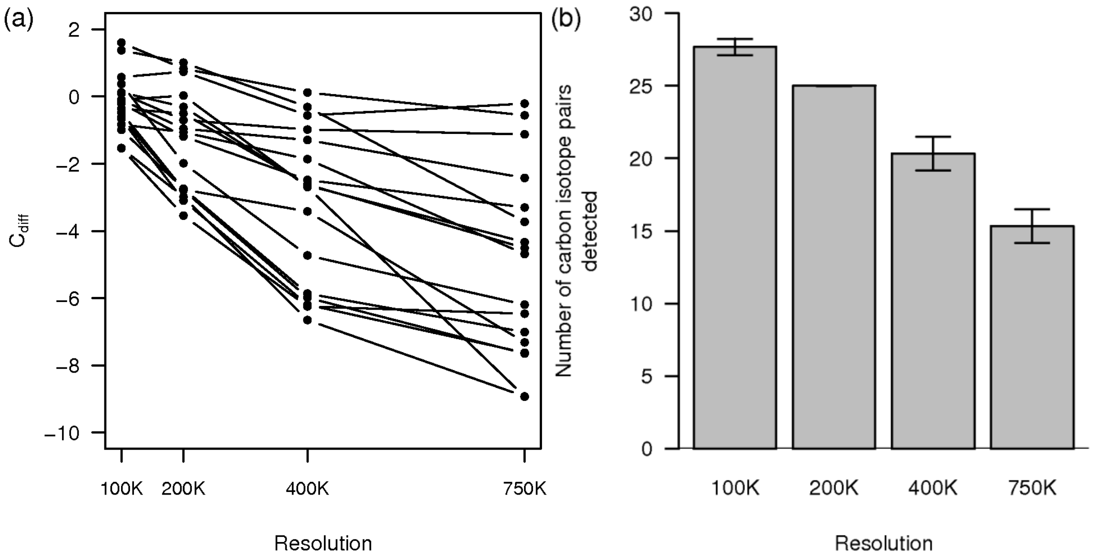


Figure 5.7 (a) C_{diff} for 15 PEG standards, plotted against four LTQ FT Ultra resolution settings from 100,000 to 750,000 (each n=3 BioPEG replicate spectra). **(b)** The average number of $^{12}C_n$ - $^{13}C^{12}C_{n-1}$ PEG isotope-pairs detected as a function of resolution between *m/z* 70-590 (error bars represent SD). Collectively these data highlight the loss of sensitivity and the significant reduction in the accuracy of RIA measurements as mass resolution is increased.

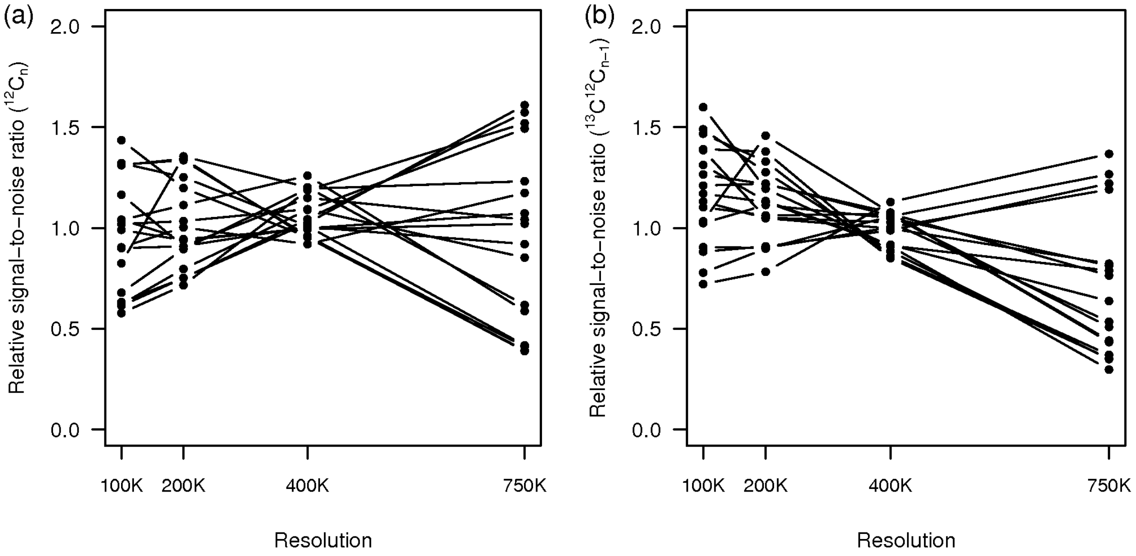


Figure 5.8 **(a)** The average signal-to-noise ratio ($^{12}C_n$) of several PEG peaks relative to their mean values as measured across four different resolutions (n=3 mass spectra of BioPEG). **(b)** The average signal-to-noise ratio ($^{13}C^{12}C_{n-1}$) of several PEG peaks relative to their mean values as measured across four different resolutions (n=3 mass spectra of BioPEG).

5.3.2 Comparison of accuracies and precisions of isotopic abundance measurements by the Orbitrap and FT-ICR MS

A similar study investigating RIA accuracy was also conducted on the Orbitrap. Identical samples, SIM-stitching acquisition parameters (LTQ FT Ultra settings, Table 4.1) and data-processing (Xcalibur 2.1) were used to enable direct comparison to the FT-ICR MS dataset. Qualitatively, the RIA accuracies of the Orbitrap and FT-ICR MS followed a similar trend, with C_{diff} being strongly dependent on the SNR of the $^{13}C^{12}C_{n-1}$-containing PEG peaks and with low intensity peaks giving the largest errors (Figure 5.9a). On closer inspection, however, peaks with a SNR ≥3.5 and ≤15 had substantially larger C_{diff} values for the Orbitrap compared to the FT-ICR. To further highlight this finding, >50% of all C_{diff} values for the Orbitrap were >3 carbons, compared to only 6.5% for the FT-ICR MS (Figure 5.9b and Figure 5.10). At higher SNR, both instruments showed diminishing C_{diff} values. Unlike the FT-ICR MS, for which C_{diff} values asymptotically approaches a small negative value (i.e. a mean±SD offset of -0.57±0.98 carbons for PEG peaks with SNR ≥50 and ≤500, when calculated using Xcalibur 2.1), the C_{diff} values for the Orbitrap averaged -0.31±0.67. A similar negative offset in the C_{diff} values was reported for the Orbitrap by Xu *et al.* (2010) and Bruker Daltonics FT-ICR MS by Koch *et al.* (2007), suggesting this small but consistent error is inherent to HR FTMS (occurring on instruments from two different manufacturers).

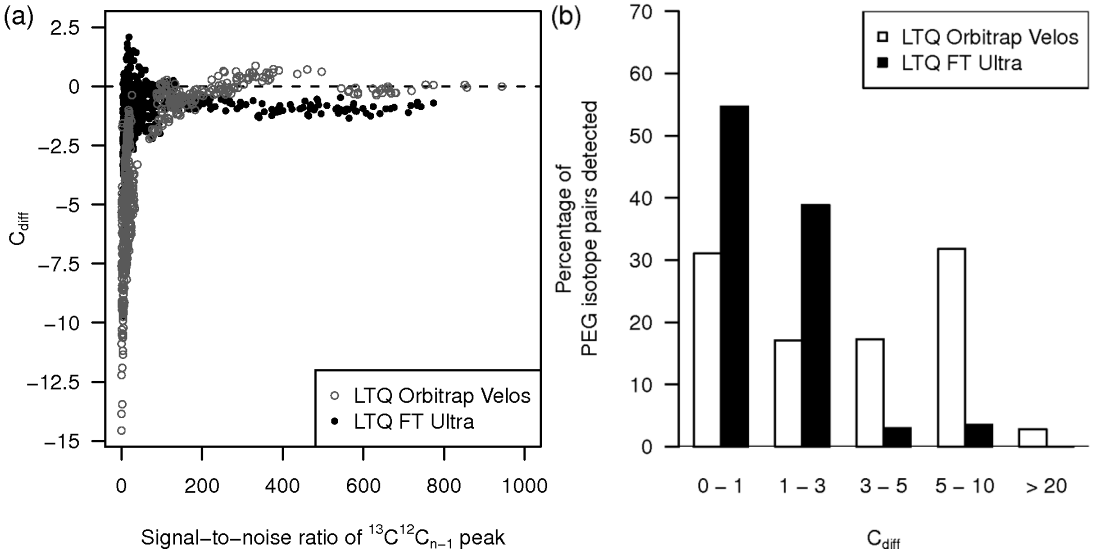


Figure 5.9 Comparison of the RIA accuracies, i.e. C_{diff} values, for the LTQ FT Ultra and LTQ Orbitrap Velos. Only measurements of $^{13}C^{12}C_{n-1}$-containing PEG peaks that are common to both datasets were used. **(a)** C_{diff} values of several PEG standards (measured in a biological matrix) plotted against the signal-to-noise ratios of the $^{13}C^{12}C_{n-1}$-containing PEG peaks. **(b)** Distribution of C_{diff} values of several PEG isotope-pairs in the range *m/z* 70-590 (SNR ≥3.5). All spectral processing was conducted using Xcalibur software.

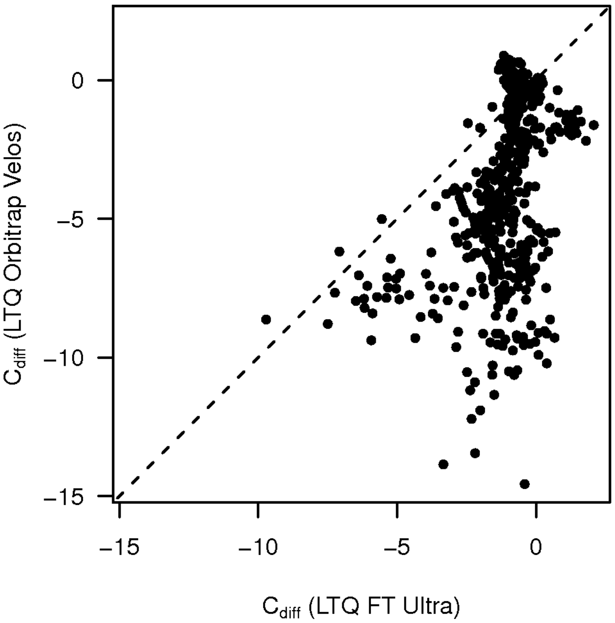


Figure 5.10 Bivariate plot of C_{diff} values facilitating the comparison of two datasets measured with two different FT instruments (LTQ FT Ultra *versus* LTQ Orbitrap Velos). All spectral processing was conducted using Xcalibur software.

Overall these results suggest that RIA measurements on the Orbitrap are less accurate than that for the FT-ICR MS. This could have occurred because the SIM-stitching acquisition parameters were optimised for the FT-ICR MS, and then implemented on the Orbitrap; e.g. the AGC target of 1×10^6 is ten times higher than the Orbitrap's factory recommendation. However, a further surprising observation is that the sensitivity of the Orbitrap appeared lower than for the FT-ICR MS, even when using the same AGC target, SIM window width, resolution and scan time. Specifically, the Orbitrap detected reproducibly only 756±18 signals (SNR $\geq$3.5, across n=30 mass spectra) compared to 1099±59 by the FT-ICR MS. The difference in instrument design (e.g. C-trap for the Orbitrap versus no C-trap for the LTQ FT Ultra) and/or tuning method is/are possible reason(s) for this difference in sensitivity. Although reducing the Orbitrap's AGC target may improve the RIA accuracy and precision, it would also further decrease the sensitivity (Figure 5.11). Since high sensitivity is a critical requirement of a metabolomics analysis, the performance of the Orbitrap at lower AGC was not evaluated. Collection of these two comparable datasets by the same analyst, using the same samples and instrument parameters, suggests some benefits of the FT-ICR MS over the Orbitrap for metabolomics.

The comparison of our Orbitrap measurements with previous studies is complicated by this being the first application of SIM-stitching on this spectrometer (Xu *et al.*, 2010). Also, authors report RIA accuracies in different formats. However, here I demonstrate that RIA measurements using an LTQ Orbitrap Velos with SIM-stitching are particularly accurate, with observed absolute RIA errors of <20% accuracy for ${}^{13}C^{12}C_{n-1}$-containing peaks with intensities between 1×10^5-1×10^6 compared to errors up to 60% reported previously for an Orbitrap XL (Figure 5.12) (Xu *et al.*, 2010). This greater RIA accuracy may arise from a combination of, the difference between Orbitrap instruments used (i.e. Orbitrap Velos versus Orbitrap XL), use of nanoelectrospray (Schmidt *et al.*, 2003) (*versus* electrospray), and acquisition *via* higher-sensitivity SIM-stitching (*versus* single full scan) (Xu *et al.*, 2010).

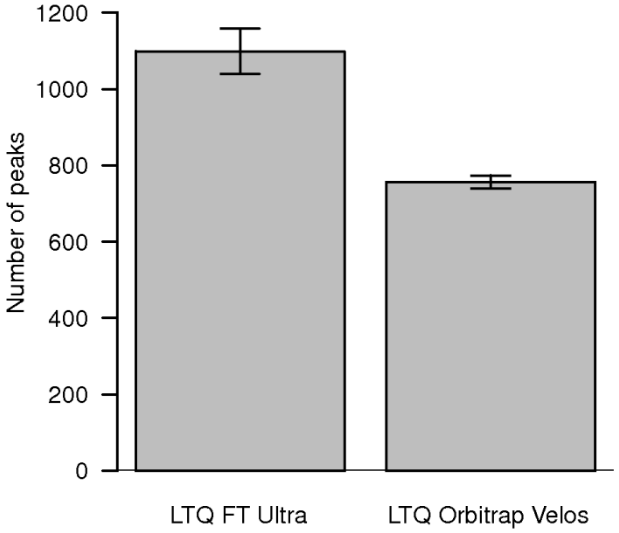


Figure 5.11 Average peak count ± SD (*m/z* 70 – 590 and signal-to-noise ratio ≥3.5) for mass spectra of liver extracts (n = 30 replicates) using two different HR FTMS instruments. To enable direct comparison both datasets were processed using Xcalibur.

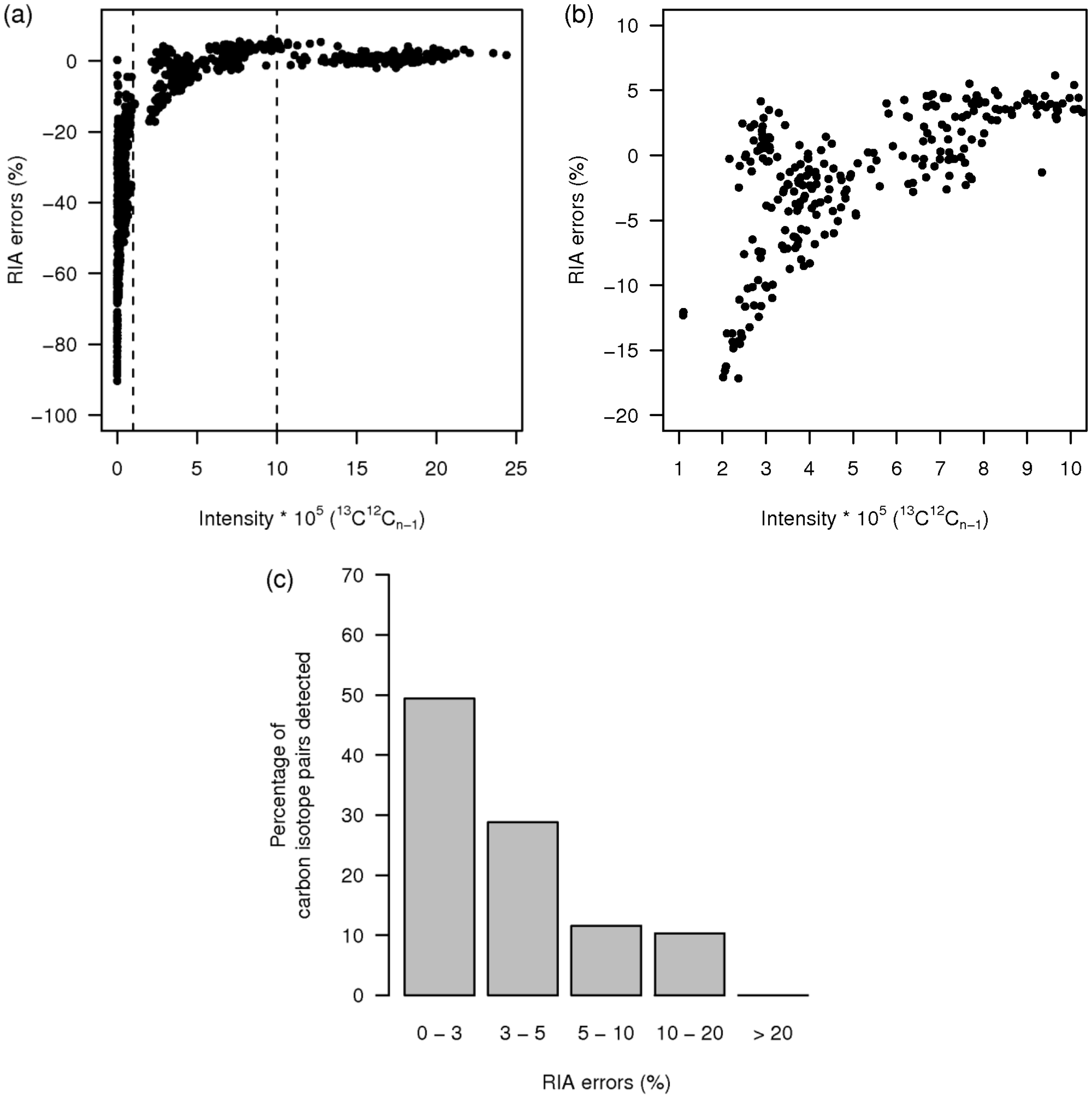


Figure 5.12 RIA measurements using a Thermo Scientific LTQ Orbitrap Velos and the SIM-stitching method (Southam *et al.*, 2007). **(a)** RIA errors (in form of %) plotted against the peak intensities (of $^{13}C^{12}C_{n\text{-}1}$ containing ion) of several PEG isotope pairs. **(b)** Expanded view of the previous figure, with RIA errors plotted against selected peak intensities ($^{13}C^{12}C_{n\text{-}1}$; intensities of 1×$10^5$ up to 1×10^6) of several PEG isotope pairs. **(c)** Distribution of RIA errors ($^{13}C^{12}C_{n\text{-}1}$; intensities of 1×$10^5$ up to 1×10^6) for several PEG isotope pairs. The dataset was collected using the LTQ Orbitrap Velos and processed using Xcalibur.

5.3.3 Increased confidence of metabolite identification using isotopic abundance measurements

As demonstrated, the RIA measurement accuracy on both HR FTMS platforms is limited, particularly for peaks with low SNR. This raises the question as to whether the information gained about the number of carbons, oxygens, potassiums, etc. can reduce the number of empirical formula(e) assigned to a peak within an automated metabolite identification pipeline. Here I evaluate this, and report an approach to minimise the false-positive rate of empirical formula(e) assignments. First all 27 PEG standards were located for which both the ${}^{12}C_n$ and ${}^{13}C{}^{12}C_{n-1}$-containing isotopes were detected in ≥50% of the 100 FT-ICR mass spectra of BioPEG (Payne *et al.*, 2009). Then the the list of all possible empirical formulae for each of these 27 measured *m/z* values (for ${}^{12}C_n$-containing PEG) were calculated with a mass error tolerance of 1 ppm. Only 4 of 27 cases (14.8%) yielded a single empirical formula, with the remaining 85.2% of cases yielding from 2-26 possible formulae (Figure 5.13 and Figure 5.14). This again highlights the challenge of metabolite identification using accurate mass measurements alone.

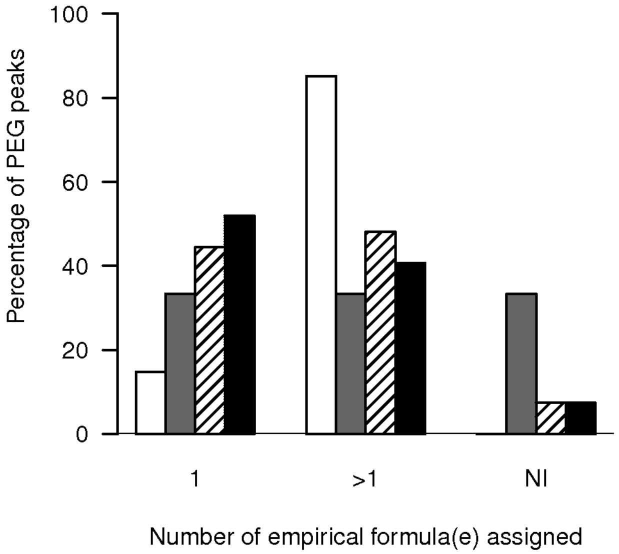


Figure 5.13 Summary of the number of empirical formula(e) assigned (1: single correct assignment, >1: more than one assignment including correct assignment, NI: not identified or many assignments without correct assignment) to each of the 27 PEG standards using different identification approaches (white bar: without RIA filtering; gray bar: with RIA filtering but without offset correction; diagonal stripes: with RIA filtering and offset correction (carbon only); black bar: with RIA filtering and offset correction (carbon (and oxygen and/or potassium if available)). A bin size of 500 SNR was used to apply the offset correction.

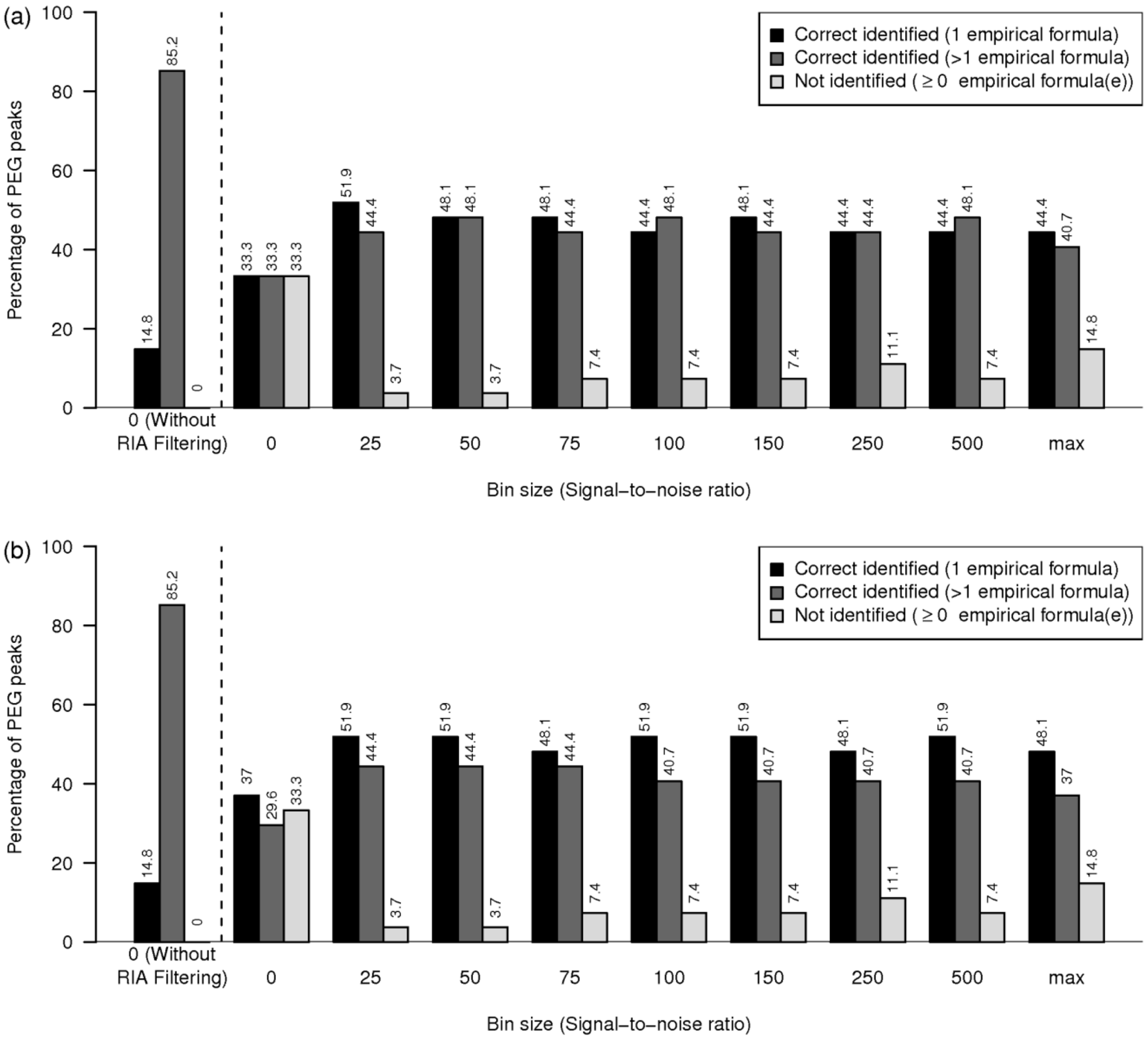


Figure 5.14 Summary of the number of empirical formula(e) assigned to each of 27 PEG standards using different identification approaches for several offset corrections (bin sizes). As shown in Figure 5.3, RIA values that are measured precisely may still result in relatively poor RIA accuracy, which is represented by a negative offset (Figure 5.2). The offset was corrected using the mean of all C_{diff} values in a SNR bin of 500 (i.e. SNR ranges ≥3.5 to <500, ≥500 to <1000, ≥1000 to <1500). This SNR bin size was chosen for three reasons: (i) It covers the majority of signal intensity measurements in biological mass spectra (see Figure 5.1); (ii) smaller bin sizes are too sensitive to local corrections instead of an overall correction of the offset (and therefore the use of a larger bin size makes it more applicable to other MS datasets); and (iii) the overall error rates for the different bin sizes are similar except for the bin size 'max'. Analyses were conducted using an LTQ FT Ultra and all spectral processing was conducted using the SIM-stitching algorithm (Southam *et al.*, 2007). **(a)** Identification approach: carbon isotope only. **(b)** Identification approach: carbon isotope (and oxygen and/or potassium isotopes if available).

To assess the value of the empirical RIA characterisations for eliminating incorrect formulae, three types of filtering were implemented. First, using the relative isotope measurements from the FT-ICR mass spectra, the number of carbons present in each of the 27 PEG standards was estimated. Then, for each PEG peak, the theoretically calculated empirical formulae were discarded if the theoretical number of carbons fell outside of this empirical range (defined by mean±3xSD). Application of the empirical RIA characterisations greatly improved the metabolite identification, with 33.3% of PEG standards definitively assigned to a single (and correct) empirical formula, which represents more than a doubling of correct identifications (Figure 5.13). However, 33.3% of the PEG peaks remained assigned to >1 empirical formula, and a further 33.3% were now incorrectly assigned (either to an incorrect empirical formula or to no formula at all; i.e. false-negatives). The relatively high FNR is particularly concerning, but can be explained by the small negative offset in C_{diff} (Figure 5.2 and Figure 5.9). If the empirically-calculated number of carbons (mean±3xSD) does not include a correction for this offset, the actual number of carbons in the correct empirical formula can lie outside this range. Therefore, our second filtering strategy additionally included an offset correction (i.e. the empirical range for the number of carbons defined as mean±3xSD-offset). Calculating the offset was the next challenge, since it showed dependency upon the SNR (Figure 5.2 and Figure 5.9). Following a thorough analysis (Figure 5.14 and Figure 5.15), I implemented a method whereby different offset corrections (i.e. mean C_{diff} values) were used for different SNRs (of the $^{13}C^{12}C_{n-1}$-containing peaks), specifically for each of the ranges ≥3.5 to <500, ≥500 to <1000, ≥1000 to <1500. This second filtering strategy not only successfully reduced the occurrence of false-negative assignments to 2 (one of these peaks corresponds to the outlier in Figure 5.2), but also further increased the number of PEG standards definitively assigned to a single correct empirical formula to 44.4% (Figure 5.13).

In addition to ^{13}C isotopic abundances, the FT-ICR mass spectra of BioPEG comprised of further isotopes, including ^{18}O and ^{41}K. The third filtering strategy additionally included the refinement of empirical formulae using these isotopes in the same manner as for ^{13}C, except no offset was applied due to a lack of appropriate measurements to characterise the possible effect. This final automated filtering algorithm for eliminating incorrect empirical formula(e) further increased the number of PEG peaks assigned to one correct empirical formula to 51.9% (Figure 5.13). In summary, incorporation of RIA characterisation for ^{13}C, ^{18}O and ^{41}K resulted in a ca. 3.5-fold reduction in the number of incorrect empirical formulae being assigned and a ca. 3-fold increase in the number of compounds that can be assigned a single correct empirical formula, compared to an unfiltered approach. Importantly, this equates to assigning the correct formula to >50% of the PEG peaks measured from *m/z* 70-590, which is particularly encouraging when extrapolated to assigning metabolites in a biological sample.

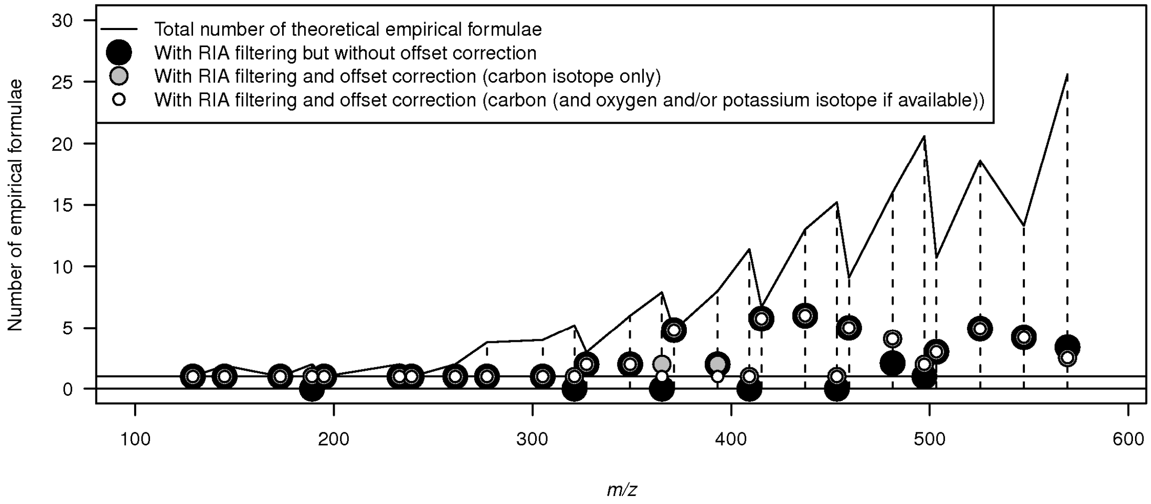


Figure 5.15 Summary of the number of empirical formula(e) assigned to each of 27 PEG standards using several different filtering approaches. A SNR bin size of 500 was used to apply the offset correction (see Figure S13). Analyses were conducted using an LTQ FT Ultra and all spectral processing was conducted using the SIM-stitching algorithm (Southam *et al.*, 2007).

5.4 Conclusions

Using the re-optimised DI SIM-stitching method for the LTQ FT Ultra (Chapter 4), RIA measurements were characterised on the LTQ FT Ultra and LTQ Orbitrap, revealing that RIA accuracies decrease with decreasing SNR. Additionally, a negative offset was discovered between the number of carbons calculated from the RIA measurements and that predicted theoretically. Increasing resolution on the LTQ FT Ultra reduced the sensitivity and consequently decreased the quality of RIA measurements. While the LTQ Orbitrap and LTQ FT Ultra showed similar RIA accuracies and precisions when isotope-pairs had mid-to-high SNR (e.g. >15), the LTQ Orbitrap performed less well for low SNR peaks. Additionally, the LTQ FT Ultra showed a 40% higher sensitivity than the LTQ Orbitrap when using SIM-stitching. However, our Orbitrap data achieved a ca. 3-fold improvement in absolute RIA accuracy compared to a previous study (Xu *et al.*, 2010), which may originate from the increased sensitivity of SIM-stitching. Finally, when RIA characterisations for ^{13}C, ^{18}O and ^{41}K were integrated into an automated peak identification pipeline, we demonstrated an approximately 3-fold increase in the number of correct single empirical formula assignments. Overall, we therefore conclude that RIA measurements on HR FTMS instruments can significantly improve the putative identification of metabolites in metabolomics applications.

CHAPTER SIX:

CONCLUSIONS

Putative metabolite identification can in principle be achieved using accurate mass measurements using a single-peak search. However a single accurate mass measurement can be assigned to one or more empirical formula(e). Furthermore each formula can correspond to different chemical structures ultimately leading to very high FPRs of identification. This complexity has stimulated the development of novel and improved bioinformatics and analytical methods to measure and subsequently annotate HR MS data, and to apply these methods to real-world metabolomics studies. At the start of this thesis five objectives were stated in order to achieve the overall aim (see Section 1.8). How well these objectives were addressed and what further work is required to fulfil these objectives is discussed here.

Chapter 3 describes the TM algorithm (part of MI-Pack) which uses metabolite interconnectivity from the KEGG database to putatively identify metabolites by name. The principle behind the algorithm is based on mapping an experimentally-derived empirical formula difference for a pair of peaks to a known empirical formula difference between substrate-product pairs derived from KEGG. I also developed a novel semi-automated method to calculate a mass error surface associated with experimental peak-pair differences. Compared to the traditional identification method of database searching accurate masses on a single-peak-by-peak basis, the TM algorithm together with the mass error surface reduces the FPR of identification by >4-fold, while maintaining a minimal FNR.

The findings also confirmed that error rates can be reduced by using species-specific databases, and demonstrate that the FPR and FNR are dependent on the accuracy of the database(s) employed. Since all metabolite databases and metabolic reconstructions are currently incomplete, further improvements of these resources are urgently required. The methods reported here could, following minor alterations (e.g. data compatibility and integration of other data sources), utilise reconstructions of metabolic networks and therefore further contribute to the field of systems biology. The next step towards more robust metabolite identification is to include further experimental measurements, such as retention time data from LC-MS studies and MS/MS. Nevertheless, the mass error surface, putative identification of metabolite names, and calculation of false positive and FNRs collectively advance and improve the traditional single-peak search. I conclude that prior biological knowledge in the form of metabolic pathways provides one route to more accurate metabolite identification.

Second, to enable high metabolome coverage and accurate metabolite identification, MS methods must have high sensitivity and mass accuracy. The higher the mass accuracy the smaller the number of empirical formula(e) assignments for a single peak. Therefore optimal acquisition parameters are required to mimimise space-charge effects that occur in the detector cell of HR FT-ICR mass spectrometers. The LTQ FT Ultra has a larger trap volume in its ICR detector cell than the earlier LTQ FT model, allowing a larger radius of ion cyclotron motion and considerably reduced space-charge effects, and hence increased sensitivity (i.e. allowing more ions into detector cell) without compromising mass accuracy. Therefore, re-optimisation of SIM-stitching was conducted on the LTQ FT Ultra (see Chapter 4), which achieved a ca. 3-fold sensitivity increase compared to the original method while maintaining a root-mean-squared mass error of 0.16 ppm. Although the re-optimised acquisition parameters allows an improvement to both sensitivity and mass accuracy in mass spectra, the LTQ FT Ultra does not fully resolve metabolome coverage and the issue of multiple false positive assignments to a single mass measurement. Additionally, the results of the re-optimisation study have highlighted a systematic error in the intensity measurements whereby the intensity of a peak is dependent upon its location within the SIM range being measured, resulting in inaccurate quantification. The overall trend observed was considerably different from the increasing trend shown previously (Payne *et al.*, 2009). Further investigation is required to parameterise this new signal intensity trend to avoid inaccurate quantification.

Currently there is limited information available on the accuracy and precision of relative isotopic abundance (RIA) measurements using high-resolution direct-infusion mass spectrometry (HR DIMS), and it is unclear if this information can benefit automated peak annotation in metabolomics. Therefore, as part of this thesis I have characterised the accuracy and precision of RIA measurements on the Thermo Scientific LTQ FT Ultra and LTQ Orbitrap Velos mass spectrometers as part of Chapter 5. I showed that the quality as measured by the accuracy and precision of empirically-derived carbons of RIA measurements is highly dependent on SNR, with RIA accuracy increasing with higher SNR. Furthermore, a negative offset between the theoretical and empirically-calculated numbers of carbon atoms was observed for both mass spectrometers. Increasing the resolution of the LTQ FT Ultra lowered both the sensitivity and the quality of RIA measurements. Overall, although the errors in the empirically-calculated number of carbons can be large, I demonstrated that RIA measurements do improve automated putative peak annotation, increasing the number of single empirical formula assignments by >3-fold compared to using accurate mass alone. Further work is now required to incorporate these RIA characterisations into MI-Pack, to allow exclusion of false positive formulae assignments.

None of the metabolite assignments can be regarded as definitive since they depend only on accurate mass and therefore do not fulfil the Metabolomics Standards Initiative criteria for metabolite identification (Sumner *et al.*, 2007). Additional analytical techniques are required to meet the requirements for definitive identification, such as LC-MS and MS/MS. Although the studies in this thesis have focused on positive ion mass spectra which are the most common ion forms, the methods and principles are equally applicable to negative ion data as shown in Section 3.3.5. However, the formation of positive and negative ions is highly depended on the sample matrix (i.e. type of sample) and modifiers added. Therefore, more investigation is needed to further understand the complexity of HR FT mass spectra and ultimately improve metabolite identification.

To conclude, all the bioinformatics and analytical methods developed and improved as part of this thesis have significantly increased the accuracy of putative metabolite identification in HR FT mass spectra and have proved to be a successful guidance to define future targeted analysis as shown in Section 3.3.5.

CHAPTER SEVEN:

REFERENCES

[Aliferis & Chrysayi-Tokousbalides, 2010] Aliferis, K. & Chrysayi-Tokousbalides, M. (2010), 'Metabolomics in pesticide research and development: review and future perspectives', *Metabolomics* pp. 1–19. 10.1007/s11306-010-0231-x. http://dx.doi.org/10.1007/s11306-010-0231-x

[Annesley, 2003] Annesley, T. M. (2003), 'Ion suppression in mass spectrometry', *Clin Chem* **49**(7), 1041–1044. http://www.clinchem.org/cgi/content/abstract/49/7/1041

[Astle *et al.*, 2007] Astle, J., Ferguson, J. T., German, J. B., Harrigan, G. G., Kelleher, N. L., Kodadek, T., Parks, B. A., Roth, M. J., Singletary, K. W., Wenger, C. D. & Mahady, G. B. (2007), 'Characterization of proteomic and metabolomic responses to dietary factors and supplements', *The Journal of Nutrition* **137**(12), 2787–2793. http://jn.nutrition.org/content/137/12/2787.abstract

[Bedair & Sumner, 2008] Bedair, M. & Sumner, L. W. (2008), 'Current and emerging mass-spectrometry technologies for metabolomics', *TrAC Trends in Analytical Chemistry* **27**(3), 238 – 250. Metabolomics. http://www.sciencedirect.com/science/article/B6V5H-4RRFN5H-1/2/14f1666c6b7fa428f5912ce143828a61

[Benjamini & Hochberg, 1995] Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300. http://dx.doi.org/10.2307/2346101

[Boccard *et al.*, 2010] Boccard, J., Veuthey, J.-L. & Rudaz, S. (2010), 'Knowledge discovery in metabolomics: an overview of MS data handling.', *J Sep Sci* **33**(3), 290–304. http://dx.doi.org/10.1002/jssc.200900609

[Bradley *et al.*, 2009] Bradley, P. H., Brauer, M. J., Rabinowitz, J. D. & Troyanskaya, O. G. (2009), 'Coordinated concentration changes of transcripts and metabolites in

Saccharomyces cerevisiae', *PLoS Comput Biol* **5**(1), e1000270. http://dx.doi.org/10.1371%2Fjournal.pcbi.1000270

[Breitling, Pitt & Barrett, 2006] Breitling, R., Pitt, A. R. & Barrett, M. P. (2006), 'Precision mapping of the metabolome.', *Trends Biotechnol* **24**(12), 543–548. http://dx.doi.org/10.1016/j.tibtech.2006.10.006

[Breitling, Ritchie, Goodenowe, Stewart & Barrett, 2006] Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. & Barrett, M. (2006), '*Ab initio* prediction of metabolic networks using Fourier transform mass spectrometry data', *Metabolomics* **2**(3), 155–164. http://dx.doi.org/10.1007/s11306-006-0029-z

[Breitling *et al.*, 2008]Breitling, R., Vitkup, D. & Barrett, M. P. (2008), 'New surveyor tools for charting microbial metabolic maps.', *Nat Rev Microbiol* **6**(2), 156–161. http://dx.doi.org/10.1038/nrmicro1797

[Brenna & Creasy, 1989] Brenna, J. & Creasy, W. R. (1989), 'Experimental evaluation of apodization functions for quantitative Fourier transform mass spectrometry', *International Journal of Mass Spectrometry and Ion Processes* **90**(2), 151–166. http://www.sciencedirect.com/science/article/B6TG6-44FMWB1-SP/2/a90b606df004810feedd718998cca719

[Brenner *et al.*, 2001] Brenner, S., Noble, D., Sejnowski, T., Fields, R., Laughlin, S., Berridge, M., Segel, L., Prank, K. & Dolmetsch, R. (2001), 'Understanding complex systems: top-down, bottom-up or middle-out?', *Novartis Foundation Symposium: Complexity in Biological Information Processing* pp. 150–159.

[Brown *et al.*, 2009] Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R. & Kell, D. B. (2009), 'Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics.', *Analyst* **134**(7), 1322–1332. http://dx.doi.org/10.1039/b901179j

[Brown *et al.*, 2011] Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L. & Dunn, W. B. (2011), 'Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets.', *Bioinformatics* . http://dx.doi.org/10.1093/bioinformatics/btr079

[Brown & Lennon, 1995] Brown, R. S. & Lennon, J. J. (1995), 'Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer', *Analytical Chemistry* **67**(13), 1998–2003. http://pubs.acs.org/doi/abs/10.1021/ac00109a015

[Brown *et al.*, 2005] Brown, S. C., Kruppa, G. & Dasseux, J.-L. (2005), 'Metabolomics applications of FT-ICR mass spectrometry.', *Mass Spectrom Rev* **24**(2), 223–231. http://dx.doi.org/10.1002/mas.20011

[Bruggeman & Westerhoff, 2007] Bruggeman, F. J. & Westerhoff, H. V. (2007), 'The nature of systems biology', *Trends in Microbiology* **15**(1), 45 – 50. http://www.sciencedirect.com/science/article/B6TD0-4MCW9NF-2/2/8e422827cb3d55c4de99444138524ca5

[Bruins, 1998] Bruins, A. P. (1998), 'Mechanistic aspects of electrospray ionization', *Journal of Chromatography A* **794**(1-2), 345 – 357. http://www.sciencedirect.com/science/article/B6TG8-3VWG2GJ-34/2/c05985c66707be2cbb7e7366ec61783d

[Buhrman *et al.*, 1996] Buhrman, D. L., Price, P. I. & Rudewicz, P. J. (1996), 'Quantitation of SR 27417 in human plasma using electrospray liquid chromatography-tandem mass spectrometry: A study of ion suppression', *Journal of the American Society for Mass Spectrometry* **7**(11), 1099 – 1105. http://www.sciencedirect.com/science/article/B6TH2-3VV60TT-2/2/6346ef2c20487c6c7bc442a58e2d8e92

[Bundy *et al.*, 2009] Bundy, J., Davey, M. & Viant, M. (2009), ‘Environmental metabolomics: a critical review and future perspectives’, *Metabolomics* **5**, 3–21. 10.1007/s11306-008-0152-0. http://dx.doi.org/10.1007/s11306-008-0152-0

[Callister *et al.*, 2006] Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W.-J., Webb-Robertson, B.-J. M., Smith, R. D. & Lipton, M. S. (2006), ‘Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics.’, *J Proteome Res* **5**(2), 277–286. http://dx.doi.org/10.1021/pr050300l

[Caspi *et al.*, 2010] Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P. & Karp, P. D. (2010), ‘The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.’, *Nucleic Acids Res* **38**(Database issue), D473–D479. http://dx.doi.org/10.1093/nar/gkp875

[Comisarow & Melka, 1979] Comisarow, M. B. & Melka, J. D. (1979), ‘Error estimates for finite zero-filling in Fourier transform spectrometry’, *Analytical Chemistry* **51**(13), 2198–2203. http://pubs.acs.org/doi/abs/10.1021/ac50049a032

[Dettmer *et al.*, 2007] Dettmer, K., Aronov, P. A. & Hammock, B. D. (2007), ‘Mass spectrometry-based metabolomics.’, *Mass Spectrom Rev* **26**(1), 51–78. http://dx.doi.org/10.1002/mas.20108

[Dieterle *et al.*, 2006] Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. (2006), ‘Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in ^{1}H NMR metabonomics.’, *Anal Chem* **78**(13), 4281–4290. http://dx.doi.org/10.1021/ac051632c

[Douglas *et al.*, 2005] Douglas, D. J., Frank, A. J. & Mao, D. M. (2005), ‘Linear ion traps in mass spectrometry’, *Mass Spectrometry Reviews* **24**(1), 1–29.

[Draper *et al.*, 2009] Draper, J., Enot, D., Parker, D., Beckmann, M., Snowdon, S., Lin, W. & Zubair, H. (2009), ‘Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive *m/z* annotation tool utilising predicted ionisation behaviour ’rules”, *BMC Bioinformatics* **10**(1), 227. http://dx.doi.org/10.1186/1471-2105-10-227

[Dunn, 2008] Dunn, W. B. (2008), ‘Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes.’, *Phys Biol* **5**(1), 11001. http://dx.doi.org/10.1088/1478-3975/5/1/011001

[Dunn *et al.*, 2005] Dunn, W. B., Bailey, N. J. C. & Johnson, H. E. (2005), ‘Measuring the metabolome: current analytical technologies.’, *Analyst* **130**(5), 606–625. http://dx.doi.org/10.1039/b418288j

[Easterling *et al.*, 1999] Easterling, M. L., Amster, I. J., van Rooij, G. J. & Heeren, R. M. A. (1999), ‘Isotope beating effects in the analysis of polymer distributions by Fourier transform mass spectrometry’, *Journal of the American Society for Mass Spectrometry* **10**(11), 1074 – 1082. http://www.sciencedirect.com/science/article/B6TH2-3XNJTJF-3/2/a85cddf9423ea482f0cd447e770e1061

[Ellis *et al.*, 2007] Ellis, D. I., Dunn, W. B., Griffin, J. L., Allwood, J. W. & Goodacre, R. (2007), ‘Metabolic fingerprinting as a diagnostic tool.’, *Pharmacogenomics* **8**(9), 1243–1266. http://dx.doi.org/10.2217/14622416.8.9.1243

[Erve *et al.*, 2009] Erve, J. C. L., Gu, M., Wang, Y., DeMaio, W. & Talaat, R. E. (2009), ‘Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination.’, *J Am Soc Mass Spectrom* **20**(11), 2058–2069. http://dx.doi.org/10.1016/j.jasms.2009.07.014

[Fiehn, 2002] Fiehn, O. (2002), ‘Metabolomics--the link between genotypes and phenotypes’, *Plant Molecular Biology* **48**, 155–171. 10.1023/A:1013713905833. http://dx.doi.org/10.1023/A:1013713905833

[Freitas *et al.*, 2003] Freitas, M. A., King, E. & Shi, S. D.-H. (2003), 'Tool command language automation of the modular ion cyclotron data acquisition system (MIDAS) for data-dependent tandem Fourier transform ion cyclotron resonance mass spectrometry', *Rapid Communications in Mass Spectrometry* **17**(4), 363–370. http://dx.doi.org/10.1002/rcm.922

[Förster *et al.*, 2003] Förster, J., Famili, I., Fu, P., Palsson, B. . & Nielsen, J. (2003), 'Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network.', *Genome Res* **13**(2), 244–253. http://dx.doi.org/10.1101/gr.234503

[Giavalisco *et al.*, 2008] Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A. C., Catchpole, G. & Willmitzer, L. (2008), 'High-resolution direct infusion-based mass spectrometry in combination with whole (13)C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas.', *Anal Chem* **80**(24), 9417–9425. http://dx.doi.org/10.1021/ac8014627

[Gipson *et al.*, 2008] Gipson, G., Tatsuoka, K., Sokhansanj, B., Ball, R. & Connor, S. (2008), 'Assignment of MS-based metabolomic datasets via compound interaction pair mapping', *Metabolomics* **4**(1), 94–103. http://dx.doi.org/10.1007/s11306-007-0096-9

[Goelzer *et al.*, 2008] Goelzer, A., Brikci, F. B., Martin-Verstraete, I., Noirot, P., Bessières, P., Aymerich, S. & Fromion, V. (2008), 'Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*.', *BMC Syst Biol* **2**, 20. http://dx.doi.org/10.1186/1752-0509-2-20

[Griffin & Shockcor, 2004] Griffin, J. L. & Shockcor, J. P. (2004), 'Metabolic profiles of cancer cells.', *Nat Rev Cancer* **4**(7), 551–561. http://dx.doi.org/10.1038/nrc1390

[Griffiths *et al.*, 2010] Griffiths, W. J., Koal, T., Wang, Y., Kohl, M., Enot, D. P. & Deigner, H.-P. (2010), 'Targeted metabolomics for biomarker discovery.', *Angew Chem Int Ed Engl* **49**(32), 5426–5445. http://dx.doi.org/10.1002/anie.200905579

[Hager, 2002] Hager, J. W. (2002), 'A new linear ion trap mass spectrometer', *Rapid Communications in Mass Spectrometry* **16**(6), 512–526. http://dx.doi.org/10.1002/rcm.607

[Han *et al.*, 2008] Han, J., Danell, R., Patel, J., Gumerov, D., Scarlett, C., Speir, J., Parker, C., Rusyn, I., Zeisel, S. & Borchers, C. (2008), 'Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry', *Metabolomics* **4**(2), 128–140. http://dx.doi.org/10.1007/s11306-008-0104-8

[Heinonen *et al.*, 2008] Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R. A. & Rousu, J. (2008), 'FiD: a software for *ab initio* structural identification of product ions from tandem mass spectrometric data.', *Rapid Commun Mass Spectrom* **22**(19), 3043–3052. http://dx.doi.org/10.1002/rcm.3701

[Herrgård *et al.*, 2008] Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Novère, N. L., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasi?, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttilä, M., Klipp, E., Palsson, B. ., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J. & Kell, D. B. (2008), 'A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology.', *Nat Biotechnol* **26**(10), 1155–1160. http://dx.doi.org/10.1038/nbt1492

[Hop *et al.*, 2005] Hop, C. E. C. A., Chen, Y. & Yu, L. J. (2005), 'Uniformity of ionization response of structurally diverse analytes using a chip-based nanoelectrospray ionization source.', *Rapid Commun Mass Spectrom* **19**(21), 3139–3142. http://dx.doi.org/10.1002/rcm.2182

[Horning & Horning, 1971] Horning, E. C. & Horning, M. G. (1971), 'Metabolic profiles: Gas-phase methods for analysis of metabolites', *Clin Chem* **17**(8), 802–809. http://www.clinchem.org/cgi/content/abstract/17/8/802

[Hu *et al.*, 2005] Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. & Cooks, R. G. (2005), 'The Orbitrap: a new mass spectrometer.', *J Mass Spectrom* **40**(4), 430–443. http://dx.doi.org/10.1002/jms.856

[Jourdan *et al.*, 2008] Jourdan, F., Breitling, R., Barrett, M. P. & Gilbert, D. (2008), 'MetaNetter: inference and visualization of high-resolution metabolomic networks.', *Bioinformatics* **24**(1), 143–145. http://dx.doi.org/10.1093/bioinformatics/btm536

[Junot *et al.*, 2010] Junot, C., Madalinski, G., Tabet, J.-C. & Ezan, E. (2010), 'Fourier transform mass spectrometry for metabolome analysis.', *Analyst* **135**(9), 2203–2219. http://dx.doi.org/10.1039/c0an00021c

[Kaddurah-Daouk *et al.*, 2008] Kaddurah-Daouk, R., Kristal, B. S. & Weinshilboum, R. M. (2008), 'Metabolomics: a global biochemical approach to drug response and disease.', *Annu Rev Pharmacol Toxicol* **48**, 653–683. http://dx.doi.org/10.1146/annurev.pharmtox.48.113006.094715

[Kanehisa *et al.*, 2008] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y. (2008), 'KEGG for linking genomes to life and the environment.', *Nucleic Acids Res* **36**, D480–D484. http://dx.doi.org/10.1093/nar/gkm882

[Karatzas *et al.*, 2007] Karatzas, K. A. G., Webber, M. A., Jorgensen, F., Woodward, M. J., Piddock, L. J. V. & Humphrey, T. J. (2007), 'Prolonged treatment of *Salmonella enterica* serovar Typhimurium with commercial disinfectants selects for multiple antibiotic resistance, increased efflux and reduced invasiveness.', *J Antimicrob Chemother* **60**(5), 947–955. http://dx.doi.org/10.1093/jac/dkm314

[Kaufmann, 2010] Kaufmann, A. (2010), 'Strategy for the elucidation of elemental compositions of trace analytes based on a mass resolution of 100,000 full width at half maximum.', *Rapid Commun Mass Spectrom* **24**(14), 2035–2045. http://dx.doi.org/10.1002/rcm.4612

[Kebarle, 2000] Kebarle, P. (2000), ‘A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry.’, *J Mass Spectrom* **35**(7), 804–817. http://dx.doi.org/3.0.CO;2-Q

[Kell, 2004] Kell, D. B. (2004), ‘Metabolomics and systems biology: making sense of the soup.’, *Curr Opin Microbiol* **7**(3), 296–307. http://dx.doi.org/10.1016/j.mib.2004.04.012

[Keseler *et al.*, 2009] Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G. & Karp, P. D. (2009), ‘EcoCyc: a comprehensive view of escherichia coli biology.’, *Nucleic Acids Res* **37**(Database issue), D464–D470. http://dx.doi.org/10.1093/nar/gkn751

[Keun *et al.*, 2002] Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E., Lindon, J. C. & Nicholson, J. K. (2002), ‘Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis’, *Chemical Research in Toxicology* **15**(11), 1380–1386. http://pubs.acs.org/doi/abs/10.1021/tx0255774

[Kiefer *et al.*, 2008] Kiefer, P., Portais, J. C. & Vorholt, J. A. (2008), ‘Quantitative metabolome analysis using liquid chromatography-high-resolution mass spectrometry’, *Analytical Biochemistry* **382**(2), 94–100.

[Kind & Fiehn, 2006] Kind, T. & Fiehn, O. (2006), ‘Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm.’, *BMC Bioinformatics* **7**, 234. http://dx.doi.org/10.1186/1471-2105-7-234

[Kind & Fiehn, 2007] Kind, T. & Fiehn, O. (2007), ‘Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry.’, *BMC Bioinformatics* **8**, 105. http://dx.doi.org/10.1186/1471-2105-8-105

[Kind *et al.*, 2009] Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S. & Fiehn, O. (2009), 'FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry.', *Anal Chem* **81**(24), 10038–10048. http://dx.doi.org/10.1021/ac9019522

[Kitano, 2002] Kitano, H. (2002), 'Systems biology: a brief overview.', *Science* **295**(5560), 1662–1664. http://dx.doi.org/10.1126/science.1069492

[Koch *et al.*, 2007] Koch, B. P., Dittmar, T., Witt, M. & Kattner, G. (2007), 'Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter.', *Anal Chem* **79**(4), 1758–1763. http://dx.doi.org/10.1021/ac061949s

[Kohlstedt *et al.*, 2010] Kohlstedt, M., Becker, J. & Wittmann, C. (2010), 'Metabolic fluxes and beyond-systems biology understanding and engineering of microbial metabolism.', *Appl Microbiol Biotechnol* **88**(5), 1065–1075. http://dx.doi.org/10.1007/s00253-010-2854-2

[Koulman *et al.*, 2007] Koulman, A., Tapper, B. A., Fraser, K., Cao, M., Lane, G. A. & Rasmussen, S. (2007), 'High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics', *Rapid Communications in Mass Spectrometry* **21**(3), 421–428. http://dx.doi.org/10.1002/rcm.2854

[Lenz & Wilson, 2007] Lenz, E. M. & Wilson, I. D. (2007), 'Analytical strategies in metabonomics.', *J Proteome Res* **6**(2), 443–458. http://dx.doi.org/10.1021/pr0605217

[Lewis *et al.*, 2008] Lewis, G. D., Wei, R., Liu, E., Yang, E., Shi, X., Martinovic, M., Farrell, L., Asnani, A., Cyrille, M., Ramanathan, A., Shaham, O., Berriz, G., Lowry, P. A., Palacios, I. F., TaÅŸan, M., Roth, F. P., Min, J., Baumgartner, C., Keshishian, H., Addona, T., Mootha, V. K., Rosenzweig, A., Carr, S. A., Fifer, M. A., Sabatine, M. S. & Gerszten, R. E. (2008), 'Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury', *The Journal of Clinical Investigation* **118**(10), 3503–3512. http://www.jci.org/articles/view/35111

[Lewis *et al.*, 2007] Lewis, I. A., Schommer, S. C., Hodis, B., Robb, K. A., Tonelli, M., Westler, W. M., Sussman, M. R. & Markley, J. L. (2007), 'Method for determining molar concentrations of metabolites in complex solutions from two-dimensional ^{1}H-^{13}C NMR spectra.', *Anal Chem* **79**(24), 9385–9390. http://dx.doi.org/10.1021/ac071583z

[Limbach *et al.*, 1993]Limbach, P. A., Grosshans, P. B. & Marshall, A. G. (1993), 'Experimental determination of the number of trapped ions, detection limit, and dynamic range in Fourier transform ion cyclotron resonance mass spectrometry', *Analytical Chemistry* **65**(2), 135–140. http://dx.doi.org/10.1021/ac00050a008

[Lin *et al.*, 2005] Lin, S. M., Zhu, L., Winter, A. Q., Sasinowski, M. & Kibbe, W. A. (2005), 'What is mzXML good for?', *Expert Rev Proteomics* **2**(6), 839–845. http://dx.doi.org/10.1586/14789450.2.6.839

[Macha & Limbach, 2002] Macha, S. F. & Limbach, P. A. (2002), 'Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry of polymers', *Current Opinion in Solid State and Materials Science* **6**(3), 213 – 220. http://www.sciencedirect.com/science/article/B6VS5-46DM1GM-B/2/035504c840e8e518ff67ff3136465381

[Mahadevan *et al.*, 2006] Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nuñez, A., Coppi, M. V., Palsson, B. O., Schilling, C. H. & Lovley, D. R. (2006), 'Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling.', *Appl Environ Microbiol* **72**(2), 1558–1568. http://dx.doi.org/10.1128/AEM.72.2.1558-1568.2006

[Mapelli *et al.*, 2008] Mapelli, V., Olsson, L. & Nielsen, J. (2008), 'Metabolic footprinting in microbiology: methods and applications in functional genomics and biotechnology.', *Trends Biotechnol* **26**(9), 490–497. http://dx.doi.org/10.1016/j.tibtech.2008.05.008

[March, 1997] March, R. E. (1997), 'An introduction to quadrupole ion trap mass spectrometry', *Journal of Mass Spectrometry* **32**(4), 351–369. http://dx.doi.org/10.1002/(SICI)1096-9888(199704)32:4<351::AID-JMS512>3.0.CO;2-Y

[Marshall *et al.*, 1998]Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. (1998), 'Fourier transform ion cyclotron resonance mass spectrometry: a primer.', *Mass Spectrom Rev* **17**(1), 1–35. http://dx.doi.org/3.0.CO;2-K

[Matsuda *et al.*, 2009] Matsuda, F., Shinbo, Y., Oikawa, A., Hirai, M. Y., Fiehn, O., Kanaya, S. & Saito, K. (2009), 'Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches.', *PLoS One* **4**(10), e7490. http://dx.doi.org/10.1371/journal.pone.0007490

[Miura *et al.*, 2010] Miura, D., Tsuji, Y., Takahashi, K., Wariishi, H. & Saito, K. (2010), 'A strategy for the determination of the elemental composition by Fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios.', *Anal Chem* **82**(13), 5887–5891. http://dx.doi.org/10.1021/ac902931x

[Monton & Soga, 2007] Monton, M. R. N. & Soga, T. (2007), 'Metabolome analysis by capillary electrophoresis-mass spectrometry', *Journal of Chromatography A* **1168**(1-2), 237 – 246. Editors' Choice I. http://www.sciencedirect.com/science/article/B6TG8-4N49VMW-D/2/012151403e4b813854241f8eed79d2c8

[Nicholson *et al.*, 1999] Nicholson, J. K., Lindon, J. C. & Holmes, E. (1999), ''metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.', *Xenobiotica* **29**(11), 1181–1189. http://dx.doi.org/10.1080/004982599238047

[Nookaew *et al.*, 2008] Nookaew, I., Jewett, M. C., Meechai, A., Thammarongtham, C., Laoteng, K., Cheevadhanarak, S., Nielsen, J. & Bhumiratana, S. (2008), 'The genome-scale metabolic model *ilN800* of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism.', *BMC Syst Biol* **2**, 71. http://dx.doi.org/10.1186/1752-0509-2-71

[Oliver *et al.*, 1998] Oliver, S. G., Winson, M. K., Kell, D. B. & Baganz, F. (1998), 'Systematic functional analysis of the yeast genome', *Trends in Biotechnology* **16**(9), 373 – 378. http://www.sciencedirect.com/science/article/B6TCW-3TVXPC7-3/2/00d1b9dece47c68d632ecf6c5ecedc70

[Parsons *et al.*, 2007] Parsons, H. M., Ludwig, C., Günther, U. L. & Viant, M. R. (2007), 'Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation.', *BMC Bioinformatics* **8**, 234. http://dx.doi.org/10.1186/1471-2105-8-234

[Pauling *et al.*, 1971] Pauling, L., Robinson, A. B., Teranishi, R. & Cary, P. (1971), 'Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography', *Proceedings of the National Academy of Sciences* **68**(10), 2374–2376. http://www.pnas.org/content/68/10/2374.abstract

[Payne *et al.*, 2009] Payne, T. G., Southam, A. D., Arvanitis, T. N. & Viant, M. R. (2009), 'A signal filtering method for improved quantification and noise discrimination in fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data.', *J Am Soc Mass Spectrom* **20**(6), 1087–1095. http://dx.doi.org/10.1016/j.jasms.2009.02.001

[Pitt, 2009] Pitt, J. (2009), 'Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry.', *Clin Biochem Rev* **30**(1), 19–34–. http://ukpmc.ac.uk/abstract/MED/19224008

[Pluskal *et al.*, 2010] Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. (2010), 'MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.', *BMC Bioinformatics* **11**, 395. http://dx.doi.org/10.1186/1471-2105-11-395

[Ramautar *et al.*, 2011] Ramautar, R., Mayboroda, O. A., Somsen, G. W. & de Jong, G. J. (2011), 'CE-MS for metabolomics: Developments and applications in the period 2008-2010.', *Electrophoresis* **32**(1), 52–65. http://dx.doi.org/10.1002/elps.201000378

[Ricci & Piddock, 2009*a*] Ricci, V. & Piddock, L. J. V. (2009*a*), 'Ciprofloxacin selects for multidrug resistance in *Salmonella enterica* serovar Typhimurium mediated by at least two different pathways.', *J Antimicrob Chemother* **63**(5), 909–916. http://dx.doi.org/10.1093/jac/dkp054

[Ricci & Piddock, 2009*b*] Ricci, V. & Piddock, L. J. V. (2009*b*), 'Only for substrate antibiotics are a functional AcrAB-TolC efflux pump and RamA required to select multidrug-resistant *Salmonella typhimurium*.', *J Antimicrob Chemother* **64**(3), 654–657. http://dx.doi.org/10.1093/jac/dkp234

[Ringner, 2008] Ringner, M. (2008), 'What is principal component analysis?', *Nat Biotech* **26**(3), 303–304. http://dx.doi.org/10.1038/nbt0308-303

[Rogers *et al.*, 2009] Rogers, S., Scheltema, R. A., Girolami, M. & Breitling, R. (2009), 'Probabilistic assignment of formulas to mass peaks in metabolomics experiments.', *Bioinformatics* **25**(4), 512–518. http://dx.doi.org/10.1093/bioinformatics/btn642

[Romero *et al.*, 2006] Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J. P., Friel, L. A., Erez, O., Mazaki-Tovi, S., Than, N. G., Hassan, S. & Tromp, G. (2006), 'The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome.', *BJOG* **113 Suppl 3**, 118–135. http://dx.doi.org/10.1111/j.1471-0528.2006.01150.x

[Sangster *et al.*, 2007] Sangster, T. P., Wingate, J. E., Burton, L., Teichert, F. & Wilson, I. D. (2007), 'Investigation of analytical variation in metabonomic analysis using liquid chromatography/mass spectrometry.', *Rapid Commun Mass Spectrom* **21**(18), 2965–2970. http://dx.doi.org/10.1002/rcm.3164

[Schmidt *et al.*, 2003] Schmidt, A., Karas, M. & Dülcks, T. (2003), 'Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI?', *Journal of the American Society for Mass Spectrometry* **14**(5), 492 – 500. http://www.sciencedirect.com/science/article/B6TH2-489B3S6-1/2/40e2600cfc9482f53dc7881652350448

[Senko *et al.*, 1997] Senko, M. W., Hendrickson, C. L., Emmett, M. R., Shi, S. D.-H. & Marshall, A. G. (1997), 'External accumulation of ions for enhanced electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry', *Journal of the American Society for Mass Spectrometry* **8**(9), 970 – 976. http://www.sciencedirect.com/science/article/B6TH2-3ST27XK-4/2/8b72dae5fec3e198c2055b1b80145c94

[Shinbo *et al.*, 2006] Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D. & Kanaya, S. (2006), Knapsack: A comprehensive species-metabolite relationship database, *in* K. Saito, R. A. Dixon & L. Willmitzer, eds, 'Plant Metabolomics', Vol. 57 of *Biotechnology in Agriculture and Forestry*, Springer Berlin Heidelberg, pp. 165–181. 10.1007/3-540-29782-0_13. http://dx.doi.org/10.1007/3-540-29782-0_13

[Smith *et al.*, 2006] Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. (2006), 'XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.', *Anal Chem* **78**(3), 779–787. http://dx.doi.org/10.1021/ac051437y

[Song *et al.*, 2009] Song, Q., Smith, S. A., Gao, L., Xu, W., Volny, M., Ouyang, Z. & Cooks, R. G. (2009), 'Mass selection of ions from beams using waveform isolation in radiofrequency quadrupoles', *Analytical Chemistry* **81**(5), 1833–1840. PMID: 19178148. http://pubs.acs.org/doi/abs/10.1021/ac802213p

[Soni & Cooks, 1994] Soni, M. H. & Cooks, R. G. (1994), 'Selective injection and isolation of ions in quadrupole ion trap mass spectrometry using notched waveforms created using the inverse Fourier transform', *Analytical Chemistry* **66**(15), 2488–2496. http://pubs.acs.org/doi/abs/10.1021/ac00087a013

[Soule *et al.*, 2010] Soule, M. C. K., Longnecker, K., Giovannoni, S. J. & Kujawinski, E. B. (2010), 'Impact of instrument and experiment parameters on reproducibility of ultrahigh resolution ESI FT-ICR mass spectra of natural organic matter', *Organic Geochemistry* **41**(8), 725–733. http://dx.doi.org/10.1016/j.orggeochem.2010.05.017

[Southam *et al.*, 2008]Southam, A. D., Easton, J. M., Stentiford, G. D., Ludwig, C., Arvanitis, T. N. & Viant, M. R. (2008), 'Metabolic changes in flatfish hepatic tumours revealed by NMR-based metabolomics and metabolic correlation networks', *Journal of Proteome Research* **7**(12), 5277–5285. http://pubs.acs.org/doi/abs/10.1021/pr800353t

[Southam *et al.*, 2007]Southam, A. D., Payne, T. G., Cooper, H. J., Arvanitis, T. N. & Viant, M. R. (2007), '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.', *Anal Chem* **79**(12), 4595–4602. http://dx.doi.org/10.1021/ac062446p

[Srivastava *et al.*, 2002] Srivastava, R., You, L., Summers, J. & YIN, J. (2002), 'Stochastic vs. deterministic modeling of intracellular viral kinetics', *Journal of Theoretical Biology* **218**(3), 309 – 321. http://www.sciencedirect.com/science/article/B6WMD-46YBBVF-G/2/b15fc67b5f30555e51c51019c97e92d5

[Stenson *et al.*, 2003] Stenson, A. C., Marshall, A. G. & Cooper, W. T. (2003), 'Exact masses and chemical formulas of individual suwannee river fulvic acids from ultrahigh resolution electrospray ionization fourier transform ion cyclotron resonance mass spectra', *Analytical Chemistry* **75**(6), 1275–1284. http://pubs.acs.org/doi/abs/10.1021/ac026106p

[Stolker *et al.*, 2004] Stolker, A. A. M., Niesing, W., Hogendoorn, E. A., Versteegh, J. F. M., Fuchs, R. & Brinkman, U. A. T. (2004), 'Liquid chromatography with triple-quadrupole or quadrupole-time of flight mass spectrometry for screening and confirmation of residues of pharmaceuticals in water', *Analytical and Bioanalytical Chemistry* **378**(4), 955–963.

[Sumner *et al.*, 2007] Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O., Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A., Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J. & Viant, M. (2007), 'Proposed minimum reporting standards for chemical analysis', *Metabolomics* **3**(3), 211–221. http://dx.doi.org/10.1007/s11306-007-0082-2

[Sun *et al.*, 2009] Sun, J., Sayyar, B., Butler, J. E., Pharkya, P., Fahland, T. R., Famili, I., Schilling, C. H., Lovley, D. R. & Mahadevan, R. (2009), 'Genome-scale constraint-based modeling of *Geobacter metallireducens*.', *BMC Syst Biol* **3**, 15. http://dx.doi.org/10.1186/1752-0509-3-15

[Tarpley *et al.*, 2005] Tarpley, L., Duran, A., Kebrom, T. & Sumner, L. (2005), 'Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period', *BMC Plant Biology* **5**(1), 8. http://www.biomedcentral.com/1471-2229/5/8

[Taylor *et al.*, 2010] Taylor, N. S., Weber, R. J. M., White, T. A. & Viant, M. R. (2010), 'Discriminating between different acute chemical toxicities via changes in the daphnid metabolome.', *Toxicol Sci* **118**(1), 307–317. http://dx.doi.org/10.1093/toxsci/kfq247

[Taylor *et al.*, 2009] Taylor, N., Weber, R., Southam, A., Payne, T., Hrydziuszko, O., Arvanitis, T. & Viant, M. (2009), 'A new approach to toxicity testing in *Daphnia magna*: application of high throughput ft-icr mass spectrometry metabolomics', *Metabolomics* **5**(1), 44–58. http://dx.doi.org/10.1007/s11306-008-0133-3

[Timischl *et al.*, 2008]Timischl, B., Dettmer, K., Kaspar, H., Thieme, M. & Oefner, P. J. (2008), 'Development of a quantitative, validated capillary electrophoresis-time of flight-mass spectrometry method with integrated high-confidence analyte identification for metabolomics.', *Electrophoresis* **29**(10), 2203–2214. http://dx.doi.org/10.1002/elps.200700517

[Tiziani *et al.*, 2009] Tiziani, S., Lodi, A., Khanim, F. L., Viant, M. R., Bunce, C. M. & Gunther, U. L. (2009), 'Metabolomic profiling of drug responses in acute myeloid leukaemia cell lines.', *PLoS One* **4**(1), e4251. http://dx.doi.org/10.1371/journal.pone.0004251

[Twycross *et al.*, 2010] Twycross, J., Band, L., Bennett, M., King, J. & Krasnogor, N. (2010), 'Stochastic and deterministic multiscale models for systems biology: an auxin-transport case study', *BMC Systems Biology* **4**(1), 34. http://www.biomedcentral.com/1752-0509/4/34

[Viant *et al.*, 2003] Viant, M. R., Rosenblum, E. S. & Tjeerdema, R. S. (2003), 'NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health', *Environmental Science & Technology* **37**(21), 4982–4989. PMID: 14620827. http://pubs.acs.org/doi/abs/10.1021/es034281x

[Villas-Bôas *et al.*, 2005] Villas-Bôas, S. G., Mas, S., Akesson, M., Smedsgaard, J. & Nielsen, J. (2005), 'Mass spectrometry in metabolome analysis.', *Mass Spectrom Rev* **24**(5), 613–646. http://dx.doi.org/10.1002/mas.20032

[Wang *et al.*, 2000] Wang, Y., Shi, S. D.-H., Hendrickson, C. L. & Marshall, A. G. (2000), '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* **198**(1-2), 113 – 120. http://www.sciencedirect.com/science/article/B6VND-3YXB2MV-C/2/dcca1e813198b8639534a9c577ed9973

[Watkins & German, 2002] Watkins, S. M. & German, J. B. (2002), 'Toward the implementation of metabolomic assessments of human health and nutrition', *Current Opinion in Biotechnology* **13**(5), 512 – 516. http://www.sciencedirect.com/science/article/B6VRV-4790D52-N/2/01963fa26d2f2f8db6afea9ad6eba61d

[Webber *et al.*, 2009] Webber, M. A., Bailey, A. M., Blair, J. M. A., Morgan, E., Stevens, M. P., Hinton, J. C. D., Ivens, A., Wain, J. & Piddock, L. J. V. (2009), 'The global consequence of disruption of the AcrAB-TolC efflux pump in *Salmonella enterica* includes reduced expression of spi-1 and other attributes required to infect the host.', *J Bacteriol* **191**(13), 4276–4285. http://dx.doi.org/10.1128/JB.00363-09

[Weber *et al.*, 2011] Weber, R. J. M., Southam, A. D., Sommer, U. & Viant, M. R. (2011), 'Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification', *Analytical Chemistry* . http://pubs.acs.org/doi/abs/10.1021/ac2001803

[Weber & Viant, 2010*a*] Weber, R. J. M. & Viant, M. R. (2010*a*), *Methodologies for Metabolomics: Experimental Strategies and Techniques*, Cambridge University Press (in preparation).

[Weber & Viant, 2010*b*] Weber, R. J. M. & Viant, M. R. (2010*b*), 'MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways', *Chemometrics and Intelligent Laboratory Systems* **104**(1), 75–82. http://www.science.cjb.net/science/article/B6TFP-4YY8MTJ-3/2/0e51a548fea1418bbcb8364142c63ef5

[Weljie *et al.*, 2006] Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. (2006), 'Targeted profiling: quantitative analysis of ^{1}H nmr metabolomics data', *Analytical Chemistry* **78**(13), 4430–4442. PMID: 16808451. http://pubs.acs.org/doi/abs/10.1021/ac060209g

[Wieser & Berglund, 2009] Wieser, M. E. & Berglund, M. (2009), 'Atomic weights of the elements 2007', *Pure Appl. Chem.* **81**(11), 2131–2156.

[Wilkinson, 2009] Wilkinson, D. J. (2009), 'Stochastic modelling for quantitative description of heterogeneous biological systems.', *Nat Rev Genet* **10**(2), 122–133. http://dx.doi.org/10.1038/nrg2509

[Williams & Fleming., 1995] Williams, D. H. & Fleming., I. (1995), *Spectroscopic methods in organic chemistry*, McGraw-Hill.

[Wilson *et al.*, 2005] Wilson, I. D., Plumb, R., Granger, J., Major, H., Williams, R. & Lenz, E. M. (2005), 'HPLC-MS-based methods for the study of metabonomics', *Journal of Chromatography B* **817**(1), 67 – 76. Coupled-column Systems in the Biosciences. http://www.sciencedirect.com/science/article/B6X0P-4DD8BRP-1/2/485b5809f822eb2c17097db771d6c472

[Wishart, 2008*a*] Wishart, D. S. (2008*a*), 'Applications of metabolomics in drug discovery and development', *Drugs in R&D* **9**(5), –. http://adisonline.com/drugsrd/Fulltext/2008/09050/Applications_of_Metabolomics_in_Drug_Discovery_and.2.aspx

[Wishart, 2008*b*] Wishart, D. S. (2008*b*), 'Metabolomics: applications to food science and nutrition research', *Trends in Food Science & Technology* **19**(9), 482 – 493. http://www.sciencedirect.com/science/article/B6VHY-4S32NKG-1/2/9aac6fbd23bfaa20a62f22f1cc2ed911

[Wishart *et al.*, 2009] Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., Souza, A. D., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J. & Forsythe, I. (2009), 'Hmdb: a knowledgebase for the human metabolome.', *Nucleic Acids Res* **37**, D603–D610. http://dx.doi.org/10.1093/nar/gkn810

[Wolf *et al.*, 2010] Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. (2010), 'In silico fragmentation for computer assisted identification of metabolite mass spectra.', *BMC Bioinformatics* **11**, 148. http://dx.doi.org/10.1186/1471-2105-11-148

[Wu *et al.*, 2008] Wu, H., Southam, A. D., Hines, A. & Viant, M. R. (2008), 'High-throughput tissue extraction protocol for NMR- and MS-based metabolomics.', *Anal Biochem* **372**(2), 204–212. http://dx.doi.org/10.1016/j.ab.2007.10.002

[Xu *et al.*, 2010] Xu, Y., Heilier, J.-F., Madalinski, G., Genin, E., Ezan, E., Tabet, J.-C. & Junot, C. (2010), 'Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-orbitrap mass spectrometer for further metabolomics database building.', *Anal Chem* **82**(13), 5490–5501. http://dx.doi.org/10.1021/ac100271j

[Zamboni & Sauer, 2009] Zamboni, N. & Sauer, U. (2009), 'Novel biological insights through metabolomics and ^{13}C-flux analysis.', *Curr Opin Microbiol* **12**(5), 553–558. http://dx.doi.org/10.1016/j.mib.2009.08.003

[Zhang *et al.*, 2005] Zhang, L.-K., Rempel, D., Pramanik, B. N. & Gross, M. L. (2005), 'Accurate mass measurements by Fourier transform mass spectrometry', *Mass Spectrometry Reviews* **24**(2), 286–309. http://dx.doi.org/10.1002/mas.20013

[Zhang *et al.*, 2009] Zhang, N. R., Yu, S., Tiller, P., Yeh, S., Mahan, E. & Emary, W. B. (2009), 'Quantitation of small molecules using high-resolution accurate mass spectrometers - a different approach for analysis of biological samples.', *Rapid Commun Mass Spectrom* **23**(7), 1085–1094. http://dx.doi.org/10.1002/rcm.3975

CHAPTER EIGHT:

APPENDICES

Table 8.1 Theoretical *m/z* value, type of adduct form, metabolite identity and number of false positive assignments for the different identification strategies of all metabolites that are part of the simulated mass spectrum (TCA cycle). A zero ppm error was used for the identification approaches as the list comprises of theoretical masses, which are identical to the mass values in the MI-DB.

No.	**Theoretical *m/z* value**	**Adduct form**	**Metabolite name; (KEGG identifier)**	**No. of false positive assignments (named metabolites)**			
				Single-peak search	**TM (direct)**	**TM (direct-or-indirect)**	**TM (direct-and-indirect)**
1	126.97920	$[M + {}^{39}K]^+$	pyruvic acid; (C00022)	2	1	1	1
2	135.02880	$[M + H]^+$	malic acid; (C00149)	3	1	1	1
3	139.00018	$[M + Na]^+$	fumaric acid; (C00122)	2	1	1	1
4	147.02880	$[M + H]^+$	oxoglutaric acid; (C00026)	4	2	2	2
5	154.99510	$[M + Na]^+$	oxalacetic acid; (C00036)	2	1	1	1
6	156.98977	$[M + {}^{39}K]^+$	succinic acid; (C00042)	2	1	1	1
7	168.98965	$[M + H]^+$	phosphoenolpyruvic acid; (C00074)	1	1	1	1
8	197.00566	$[M + Na]^+$	cis-aconitic acid; (C00417)	2	1	1	0
9	213.00058	$[M + Na]^+$	oxalosuccinic acid; (C05379)	0	0	0	0
10	215.01623	$[M + Na]^+$	isocitric acid; (C00311)	7	3	3	2
11	230.99016	$[M + {}^{39}K]^+$	citric acid; (C00158)	7	3	3	2
12	426.05225	$[M + H]^+$	thiamine diphosphate; (C00068)	0	0	0	0
13	492.06041	$[M + Na]^+$	2-(alpha-hydroxyethyl)thiamine diphosphate; (C05125)	0	0	0	0
14	550.06589	$[M + Na]^+$	3-carboxy-1-hydroxypropyl-thpp; (C05381)	0	0	0	0
15	832.11501	$[M + Na]^+$	acetyl-coA; (C00024)	0	0	0	0
16	906.09442	$[M + {}^{39}K]^+$	succinyl-coA; (C00091)	3	2	3	2
			Total no. of false positive assignments:	**35**	**17**	**18**	**14**

Table 8.2 Illustrative examples of the sizes of errors associated with peak differences for a "fixed error" *versus* mass error surface approach. Comparison of the two error handling approaches clearly shows the improvement of the mass error surface, which measures an ca. 3-fold smaller mass error for relatively small peak differences in high mass ranges. Such peak differences cover most of the enzymatic reactions in the KEGG database as shown in Figure 3.4a.

Theoretical mass spectral data			**Fixed error approach**				**From mass error surface**			**Fold improvement of mass error surface**
Peak A (*m/z*)	**Peak B (*m/z*)**	**Peak difference**	**1.0 ppm error peak A (*m/z*)**	**1.0 ppm error peak B (*m/z*)**	**Total error (*m/z*)**	**Relative error (ppm)[a]**	**Absolute error (*m/z*)**	**Relative error (ppm)**	**Mean relative error (peak A and peak B)[b]**	
200.00000	201.00336	1.00336 (^{13}C–^{12}C)	± 0.00020	± 0.00020	± 0.00040	± 399.66	± 0.00009	± 87.36	± 0.22	4.6
300.00000	301.00336	1.00336 (^{13}C–^{12}C)	± 0.00030	± 0.00030	± 0.00060	± 598.99	± 0.00020	± 195.46	± 0.33	3.1
400.00000	401.00336	1.00336 (^{13}C–^{12}C)	± 0.00040	± 0.00040	± 0.00080	± 798.32	± 0.00033	± 327.76	± 0.41	2.4
200.00000	202.01565	2.01565 (H_2)	± 0.00020	± 0.00020	± 0.00040	± 199.45	± 0.00009	± 44.24	± 0.22	4.5
300.00000	302.01565	2.01565 (H_2)	± 0.00030	± 0.00030	± 0.00060	± 298.67	± 0.00021	± 102.42	± 0.35	2.9
400.00000	402.01565	2.01565 (H_2)	± 0.00040	± 0.00040	± 0.00080	± 397.89	± 0.00035	± 172.38	± 0.44	2.3
200.00000	215.02348	15.02348 (CH_3)	± 0.00020	± 0.00022	± 0.00042	± 27.62	± 0.00020	± 13.34	± 0.48	2.1
300.00000	315.02348	15.02348 (CH_3)	± 0.00030	± 0.00032	± 0.00062	± 40.94	± 0.00029	± 19.45	± 0.47	2.1
400.00000	415.02348	15.02348 (CH_3)	± 0.00040	± 0.00042	± 0.00082	± 54.25	± 0.00059	± 39.10	± 0.72	1.4
200.00000	243.98983	43.98983 (CO_2)	± 0.00020	± 0.00024	± 0.00044	± 10.09	± 0.00017	± 3.83	± 0.38	2.6
300.00000	343.98983	43.98983 (CO_2)	± 0.00030	± 0.00034	± 0.00064	± 14.64	± 0.00045	± 10.16	± 0.70	1.4
400.00000	443.98983	43.98983 (CO_2)	± 0.00040	± 0.00044	± 0.00084	± 19.19	± 0.00063	± 14.35	± 0.75	1.3

[a]Relative error calculated as total error / peak difference (ppm), [b]Relative error calculated as absolute error / ((peak A + peak B) / 2) * 0.000001 (ppm)

Table 8.3 Name of metabolite identified by NMR spectroscopy, corresponding observation in FT-ICR mass spectrum, and number of false positive assignments for the different identification strategies. The numbers of true positive assignments for the different identification strategies are as follows: 26 for single-peak search, 25 for TM (direct), 26 for TM (direct-or-indirect), and 22 for TM (direct-and-indirect). Entries marked by a box correspond to false negative assignments and were not identified by one of the selected identification approaches. The NMR and FT-ICR MS data were collected from identical biological samples, human cell extracts.

No.	Name of metabolite detected by NMR (KEGG identifier)	Measured *m/z* values (including adduct form)	No. of false positive assignments (named metabolites)			
			Single-peak search	TM (direct)	TM (direct-or-indirect)	TM (direct-and-indirect)
1	glycine; (C00037)	76.03931 $[M + H]^+$, 98.02126 $[M + Na]^+$, 113.99519 $[M + {}^{39}K]^+$	2	0	1	0
2	beta-alanine; (C00099)	112.03690 $[M + Na]^+$	7	3	4	3
3	alanine; (C00041)	112.03690 $[M + Na]^+$	7	3	4	3
4	sarcosine; (C00213)	112.03690 $[M + Na]^+$	7	3	4	3
5	lactic acid; (C00186)	113.02092 $[M + Na]^+$	7	3	4	$\boxed{2}$
6	proline; (C00148)	116.07061 $[M + H]^+$, 138.05258 $[M + Na]^+$, 154.02647 $[M + {}^{39}K]^+$	2	1	1	1
7	valine; (C00183)	118.08627 $[M + H]^+$, 140.06823 $[M + Na]^+$, 156.04216 $[M + {}^{39}K]^+$	9	2	2	2
8	creatine; (C00300)	132.07675 $[M + H]^+$, 154.05868 $[M + Na]^+$, 170.03259 $[M + {}^{39}K]^+$	1	0	0	0
9	isoleucine; (C00407)	132.10190 $[M + H]^+$, 154.08384 $[M + Na]^+$, 170.05775 $[M + {}^{39}K]^+$	10	3	3	3
10	leucine; (C00123)	132.10190 $[M + H]^+$, 154.08384 $[M + Na]^+$, 170.05775 $[M + {}^{39}K]^+$	10	3	3	3
11	asparagine; (C00152)	133.06077 $[M + H]^+$, 155.04274 $[M + Na]^+$, 171.01663 $[M + {}^{39}K]^+$	5	2	2	2
12	phosphoethanolamine; (C00346)	142.0264 $[M + H]^+$, 164.00834 $[M + Na]^+$, 179.98227 $[M + {}^{39}K]^+$	1	0	1	0
13	glutamine; (C00064)	147.07644 $[M + H]^+$, 169.05833 $[M + Na]^+$, 185.03230 $[M + {}^{39}K]^+$	4	2	2	1
14	taurine; (C00245)	148.00391 $[M + Na]^+$, 163.97784 $[M + {}^{39}K]^+$	0	0	0	$\boxed{0}$
15	glutamic acid; (C00025)	148.06046 $[M + H]^+$, 170.04234 $[M + Na]^+$, 186.01629 $[M + {}^{39}K]^+$	9	6	6	5
16	methionine; (C00073)	150.05835 $[M + H]^+$, 172.04025 $[M + Na]^+$, 188.01437 $[M + {}^{39}K]^+$	3	0	0	0
17	succinic acid; (C00042)	156.98978 $[M + {}^{39}K]^+$	2	0	1	$\boxed{0}$
18	phenylalanine; (C00079)	166.08627 $[M + H]^+$, 188.06819 $[M + Na]^+$, 204.04215 $[M + {}^{39}K]^+$	3	1	1	1
19	aspartic acid; (C00049)	172.00064 $[M + {}^{39}K]^+$	2	1	1	1
20	myo-inositol; (C00137)	181.07069 $[M + H]^+$, 203.05262 $[M + Na]^+$, 219.02655 $[M + {}^{39}K]^+$	41	12	12	10
21	tyrosine; (C00082)	182.08120 $[M + H]^+$, 204.06313 $[M + Na]^+$, 220.03707 $[M + {}^{39}K]^+$	4	1	1	1
22	citric acid; (C00158)	215.01626 $[M + Na]^+$, 230.99009 $[M + {}^{39}K]^+$	8	$\boxed{1}$	4	$\boxed{1}$
23	pantothenic acid; (C00864)	220.11796 $[M + H]^+$, 242.09992 $[M + Na]^+$, 258.07387 $([M + {}^{39}K]^+$	0	0	0	0
24	glutathione; (C00051)	308.09109 $[M + H]^+$, 330.07283 $[M + Na]^+$, 346.04693 $[M + {}^{39}K]^+$	0	0	0	0
25	AMP; (C00020)	348.07043 $[M + H]^+$, 370.05256 $[M + Na]^+$, 386.02622 $[M + {}^{39}K]^+$	4	2	2	1
26	ADP; (C00008)	450.01898 $[M + Na]^+$, 465.99278 $[M + {}^{39}K]^+$	3	1	1	1
		Total no. of false positive assignments:	**151**	**50**	**60**	**44**
		Total no. of true positive assignments:	**26**	**25**	**26**	**22**

Table 8.4 Putative metabolite identification of a subset of peaks detected in the FT-ICR mass spectra of *D. magna* (whole organism) following exposure to cadmium.

a) Average intensity across all samples.

b) Not listed explicitly for cases where >5 empirical formulae can be identified.

c) Calculated for the specified ion form of the empirical formula.

d) Error between the observed and theoretical masses, presented as parts per million of the theoretical mass.

e) Number of carbon atoms derived from the ^{12}C - ^{13}C isotope intensity pattern (when observed).

f) Derived from the transformation mapping approach for metabolite identification, described in Chapter 3. Not listed explicitly for cases where >5 putative metabolite names can be identified.

Observed		Identification					
m/z	Average intensity[a]	Empirical formula[b]	Ion form	Theoretical mass (Da)[c]	Mass error (ppm)[d]	No. C atoms[e]	Putative metabolite name(s)[f]
78.95889	4476.5	0					0
78.95905	5658.3	0					0
79.00352	14514.2	0					0
79.00368	22502.1	0					0
79.95720	4769.6	0					0
79.95736	7571.7	0					0
80.97472	18333.3	0					0
80.97488	28180.9	CH2O2	[M+Cl]-	80.97488	-0.02		["Formate"]
81.96996	4794.6	0					0
81.97012	4426.7	0					0
82.97176	8445.7	0					0
82.97193	10467.9	CH2O2	[M+(37Cl)]-	82.97193	-0.02		["Formate"]
87.00859	13015.4	0					0
87.00876	18824.1	C3H4O3	[M-H]-	87.00877	-0.10		["3-Oxopropanoate", "Pyruvate"]
88.04023	9801.5	0					0
88.04040	14666.4	C3H7NO2	[M-H]-	88.04040	-0.03		["D-Alanine", "L-Alanine", "beta-Alanine"]
89.99585	4490.6	0					0
90.02758	5428.6	0					0
91.00350	52625.5	0					0
91.00368	79520.5	C2H4O4	[M-H]-	91.00368	-0.04		0
91.99893	13767.0	0					0
92.01132	32417.4	0					0
92.01150	59291.4	0					0
92.92802	14770.2	[NaCl]	[M+Cl]-	92.92802	-0.05		0

94.92507	9152.8	[Na(37Cl)]	[M+Cl]-	94.92507	-0.05		0
94.92507	9152.8	[NaCl]	[M+(37Cl)]-	94.92507	-0.05		0
96.95985	4869.2	0					0
96.96010	25420.0	0					0
96.96962	6185.3	0					0
97.02950	5416.1	C5H6O2	[M-H]-	97.02950	-0.04		0
103.04007	5820.8	C4H8O3	[M-H]-	103.04007	0.01		["(R)-3-Hydroxybutanoate"]
105.02005	2878.6	0					0
113.02441	4001.2	C5H6O3	[M-H]-	113.02442	-0.08		["2-Hydroxy-2,4-pentadienoate"]
114.05604	6671.7	C5H9NO2	[M-H]-	114.05605	-0.11		["D-Proline", "L-Proline"]
126.90498	108366.6	0					0
141.01656	3800533.1	C2H2N6O2	[M-H]-	141.01665	-0.62		0
....	....	....	....	....	....	....	....
203.03260	52924.7	C5H12O6	[M+Cl]-	203.03279	-0.94		0
203.08148	7497.4	C5H13N6OP	[M-H]-	203.08157	-0.45		0
203.08261	109406.4	C11H12N2O2	[M-H]-	203.08260	0.04	11.6	["L-Tryptophan"]
203.10373	12721.9	C8H16N2O4	[M-H]-	203.10373	-0.01		0
203.88987	8614.5	CHNO3S[Na(37Cl)]	[M+(37Cl)]-	203.88984	0.14		0
203.99077	28690.4	C2H5N5O3[NaCl]	[M-H]-	203.99059	0.90	4.9	0
203.99077	28690.4	C3H7N5S[Na(37Cl)]	[M-H]-	203.99061	0.77	4.9	0
203.99077	28690.4	C3(13C)H6N4[NaCl]	[M+Cl]-	203.99063	0.71	4.9	0
203.99077	28690.4	C5H7N3O2S2	[M-H]-	203.99069	0.37	4.9	0
204.02642	21499.8	C6(13C)H8N2O3	[M+Cl]-	204.02625	0.84	5.8	0
204.02642	21499.8	C7(13C)H10N2S	[M+(37Cl)]-	204.02628	0.70	5.8	0
204.08595	14007.4	C10(13C)H12N2O2	[M-H]-	204.08596	-0.03	11.6	0
205.01755	30016.4	C10H6N2O	[M+Cl]-	205.01741	0.66		0
205.01755	30016.4	C5H10N2O3[Na(37Cl)]	[M-H]-	205.01754	0.05		0
205.01755	30016.4	C7H10O5S	[M-H]-	205.01762	-0.34		0
205.02730	10309.6	C7H11O5P	[M-H]-	205.02714	0.80		0
205.02730	10309.6	C8H10O4	[M+Cl]-	205.02731	-0.06		["2-Hydroxy-6-oxo-octa-2,4-dienoate"]
205.02730	10309.6	C9H12OS	[M+(37Cl)]-	205.02734	-0.19		0
205.02730	10309.6	C2H8N8P2	[M-H]-	205.02744	-0.69		0

205.03539	37605.4	C11H8N2	[M+(37Cl)]-	205.03520	0.93		0
205.03539	37605.4	C7H10O7	[M-H]-	205.03538	0.06		0
205.06372	13727.6	C8H15O4P	[M-H]-	205.06352	0.97		0
205.06372	13727.6	C9H14O3	[M+Cl]-	205.06370	0.11		0
205.06372	13727.6	C10H16S	[M+(37Cl)]-	205.06372	-0.02		0
205.12341	50459.2	C13H18O2	[M-H]-	205.12340	0.03		0
205.88692	12351.0	C4HNOS4	[M-H]-	205.88683	0.46		0
206.09535	26045.1	C8H18NO3P	[M-H]-	206.09516	0.94		0
206.09535	26045.1	C9H17NO2	[M+Cl]-	206.09533	0.09		0
206.98029	13236.7	C4H10N2P2[Na(37Cl)]	[M-H]-	206.98032	-0.15		0
206.98029	13236.7	C4H2N6O[NaCl]	[M-H]-	206.98036	-0.32		0
206.98029	13236.7	C6H10O2P2S	[M-H]-	206.98040	-0.54		0
206.98029	13236.7	C7H4N4S2	[M-H]-	206.98046	-0.84		0
207.06847	8305.1	0					0
207.06966	71835.5	C6H16N2O2[Na(37Cl)]	[M-H]-	207.06957	0.41		0
207.06966	71835.5	C8H16O4S	[M-H]-	207.06966	0.02		0
207.07935	34825.8	C8H17O4P	[M-H]-	207.07917	0.86		0
207.07935	34825.8	C9H16O3	[M+Cl]-	207.07935	0.02		0
207.07935	34825.8	C10H18S	[M+(37Cl)]-	207.07937	-0.11		0
207.11572	10130.7	C9H21O3P	[M-H]-	207.11556	0.79		0
207.11572	10130.7	C10H20O2	[M+Cl]-	207.11573	-0.06		0
208.06492	8893.4	C5H15N3O2[Na(37Cl)]	[M-H]-	208.06482	0.46		0
....	....	....	....	....	....	....	....
472.33727	1350.8	C13H43N15O2S	[M-H]-	472.33721	0.13		0
472.33727	1350.8	C18H45N11[NaCl]	[M-H]-	472.33729	-0.03		0
472.33727	1350.8	C27H49NO3	[M+(37Cl)]-	472.33770	-0.90		0
472.88357	3614.4	310					0
472.97068	3983.5	237					0
473.07632	14536.5	120					0
473.15234	5836.5	58					0
473.26095	42823.1	14					0
473.27433	9933.3	14					0
473.32876	4172.4	C30H42N4O	[M-H]-	473.32858	0.37		0

473.32876	4172.4	C21H48N4O5	[M+(37Cl)]-	473.32892	-0.34		0
473.32876	4172.4	C27H52[Na(37Cl)]	[M+(37Cl)]-	473.32902	-0.56		0
473.32876	4172.4	C22H48N6OP2	[M-H]-	473.32921	-0.95		0
473.32876	4172.4	C16H45N12P	[M+(37Cl)]-	473.32923	-0.99		0
473.34974	3691.8	C27H46N4O3	[M-H]-	473.34971	0.05		0
473.34974	3691.8	C28H53OP	[M+(37Cl)]-	473.34985	-0.24		0
474.26445	36644.2	26				19.2	0
474.93752	3229.4	276					0
475.15507	4170.3	49					0
475.17687	19218.5	43					0
475.26800	7730.6	29				19.2	0
475.34479	6881.3	C27H54[Na(37Cl)]	[M+(37Cl)]-	475.34467	0.24		0
475.34479	6881.3	C22H50N6OP2	[M-H]-	475.34486	-0.14		0
475.34479	6881.3	C16H47N12P	[M+(37Cl)]-	475.34488	-0.18		0
475.34479	6881.3	C23H49N6P	[M+Cl]-	475.34503	-0.51		0
475.38115	5478.1	C23H54N6P2	[M-H]-	475.38124	-0.20		0
476.05960	6550.0	118					0
476.20260	68225.2	30					0
476.23144	15318.9	25					0
476.28045	19209.2	10				24.3	0
477.02955	12147.5	172					0
477.28395	5143.8	19				24.3	0
477.36060	16916.7	C27H56[Na(37Cl)]	[M+(37Cl)]-	477.36032	0.58	28.9	0
477.36060	16916.7	C22H52N6OP2	[M-H]-	477.36051	0.19	28.9	0
477.36060	16916.7	C23H51N6P	[M+Cl]-	477.36068	-0.18	28.9	0
477.37365	6975.7	C33H50O2	[M-H]-	477.37380	-0.32		0
478.36403	5369.0	6				28.9	0
478.87518	2273.3	334					0
478.90111	5562.0	310					0
478.98844	30489.8	241					0
479.07170	10143.9	125					0
479.25016	4486.3	23					0
....	....	....	....	....	....	....	....

Table 8.5 Putative metabolite identification of a subset of peaks detected in the FT-ICR mass spectra of *S. typhimurium*.

a) Average intensity across all samples (L354, L644, L742 and L976).

b) Not listed explicitly for cases where >5 empirical formulae can be identified.

c) Calculated for the specified ion form of the empirical formula.

d) Error between the observed and theoretical masses, presented as parts per million of the theoretical mass.

e) Number of carbon atoms derived from the ^{12}C - ^{13}C isotope intensity pattern (when observed).

f) Derived from the transformation mapping approach for metabolite identification, described in Chapter 3. Not listed explicitly for cases where >5 putative metabolite names can be identified.

Observed		Identification					
m/z	Average intensity[a]	Empirical formula[b]	Ion form	Theoretical mass (Da)[c]	Mass error (ppm)[d]	No. C atoms[e]	Putative metabolite name(s)[f]
70.03424	144391.7	0					0
74.46765	3419.6	0					0
75.02905	14362.1	0					0
77.04181	30287.8	0					0
80.94781	5593.9	0					0
84.08074	12127.4	C5H9N	[M+H]+	84.08078	-0.42		0
86.09639	113681.8	C5H11N	[M+H]+	86.09643	-0.41		0
87.09975	6418.7	0					0
90.05492	19299.3	C3H7NO2	[M+H]+	90.05496	-0.39		["Alanine", "D-Alanine", "L-Alanine", "Sarcosine", "beta-Alanine"]
90.97662	13604	C2H3PS	[M+H]+	90.97659	0.37		0
96.92175	6450.6	0					0
103.05420	8031.2	C8H6	[M+H]+	103.05423	-0.26		0
104.05282	7749.6	C4H9NS	[M+H]+	104.05285	-0.26		0
104.10696	9998.4	C5H13NO	[M+H]+	104.10699	-0.29		0
106.95055	5747.9	0					0
110.07124	25849	C5H7N3	[M+H]+	110.07127	-0.30		0
112.03687	15525.4	C3H7NO2	[M+Na]+	112.03690	-0.27		["Alanine", "D-Alanine", "L-Alanine", "Sarcosine", "beta-Alanine"]
115.08656	4458.3	C5H10N2O	[M+H]+	115.08659	-0.26		0
116.07057	31834.2	C5H9NO2	[M+H]+	116.07061	-0.30		["D-Proline", "L-Proline"]

118.08623	239122.6	C5H11NO2	[M+H]+	118.08626	-0.22		["4-Methylaminobutyrate", "5-Aminopentanoate", "Betaine", "L-Valine"]
119.08958	14448.3	0					0
120.08075	132096	C8H9N	[M+H]+	120.08078	-0.21		0
121.08410	9904.7	0					0
125.07092	4086.4	C6H8N2O	[M+H]+	125.07094	-0.16		["Methylimidazole acetaldehyde"]
127.08657	8979.4	C6H10N2O	[M+H]+	127.08659	-0.15		0
....	....	....	....	....	....	....	....
227.07910	405435.4	C11H12N2O2	[M+Na]+	227.07910	0.00	11.0	["L-Tryptophan"]
227.07910	405435.4	C6H15N2O5P	[M+H]+	227.07914	-0.16	11.0	0
227.07910	405435.4	C8H16N2O3	[M+K]+	227.07925	-0.67	11.0	["N2-Acetyl-L-lysine", "N6-Acetyl-L-lysine"]
227.10264	53878.2	C10H14N2O4	[M+H]+	227.10263	0.02		["Porphobilinogen"]
227.11387	21346.5	C9H14N4O3	[M+H]+	227.11387	0.01		["Carnosine"]
227.13901	7451.9	C11H18N2O3	[M+H]+	227.13902	-0.04		0
227.15836	23216	0					0
228.08246	49211	C10(13C)H12N2O2	[M+Na]+	228.08245	0.02	11.0	0
228.08246	49211	C5(13C)H15N2O5P	[M+H]+	228.08249	-0.14	11.0	0
228.08246	49211	C7(13C)H16N2O3	[M+K]+	228.08261	-0.65	11.0	0
228.08246	49211	C5H13N3O7	[M+H]+	228.08263	-0.74	11.0	0
228.13427	30526.8	C10H17N3O3	[M+H]+	228.13427	0.01		0
228.14550	18004.3	C9H17N5O2	[M+H]+	228.14550	-0.01		0
228.17066	44547.1	C11H21N3O2	[M+H]+	228.17065	0.03		0
229.10053	20976.4	C10H16N2O2S	[M+H]+	229.10053	0.02		0
229.11828	77295.7	C10H16N2O4	[M+H]+	229.11828	-0.02		0
229.12952	26841.5	C9H16N4O3	[M+H]+	229.12952	0.01		["Deoxyguanidinoproclavaminic acid"]
229.15467	667562.1	C11H20N2O3	[M+H]+	229.15467	0.00		0
230.11354	94717.8	C9H15N3O4	[M+H]+	230.11353	0.03		0
230.14993	25478.5	C10H19N3O3	[M+H]+	230.14992	0.05		0
230.15803	81882.2	0					0

231.09755	28639.3	C9H14N2O5	[M+H]+	231.09755	0.00		0
231.17031	164313.7	C11H22N2O3	[M+H]+	231.17032	-0.04		0
231.97239	26158	C5H2N7P	[M+(41K)]+	231.97221	0.79		0
231.97239	26158	C6H3N5S2	[M+Na]+	231.97221	0.77		0
231.97239	26158	C3H7N5OS2	[M+K]+	231.97236	0.11		0
231.97239	26158	C6HNO9	[M+H]+	231.97241	-0.09		0
231.97239	26158	C3H5N5O3S	[M+(41K)]+	231.97259	-0.86		0
232.89664	20406.1	C8H2P2S	[M+(41K)]+	232.89652	0.50		0
233.09545	11889.7	C9H16N2O3S	[M+H]+	233.09544	0.04		0
233.12606	149312.3	C11H18N2O2	[M+Na]+	233.12605	0.05		0
234.02442	58651.9	C4H13N5S2	[M+K]+	234.02440	0.09		0
234.02442	58651.9	C7H7NO8	[M+H]+	234.02445	-0.11		0
234.02442	58651.9	C4H11N5O2S	[M+(41K)]+	234.02462	-0.87		0
234.08492	81315.1	C9H13N3O3	[M+Na]+	234.08491	0.03		0
234.08492	81315.1	C6H17N3O4	[M+K]+	234.08507	-0.63		0
234.12941	18498	0					0
234.14485	23358.2	C9H19N3O4	[M+H]+	234.14483	0.07		0
….	….	….	….	….	….	….	….
454.30263	32516.5	C11H39N15O2	[M+(41K)]+	454.30239	0.53		0
454.30263	32516.5	C22H39N5O5	[M+H]+	454.30240	0.51		0
454.30263	32516.5	C17H36N13P	[M+H]+	454.30270	-0.16		0
454.30263	32516.5	C19H42N7O2P	[M+Na]+	454.30298	-0.78		0
454.30263	32516.5	C15H39N11O3S	[M+H]+	454.30308	-1.00		0
454.70246	36518.3	C6HO2P3S8	[M+H]+	454.70280	-0.74		0
454.70246	36518.3	C8H2P2S8	[M+K]+	454.70291	-0.99		0
455.17515	22053.1	25					0
455.21397	183145.8	17					0
456.10192	19678	37					0

456.21711	35482.1	16					0
456.23565	107249	12					0
456.28190	75989.9	9					0
456.31823	36907.5	C22H41N5O5	[M+H]+	456.31805	0.40		0
456.31823	36907.5	C17H38N13P	[M+H]+	456.31835	-0.27		0
456.31823	36907.5	C19H44N7O2P	[M+Na]+	456.31863	-0.88		0
456.73748	53718.7	7					0
457.15434	22764.1	30					0
457.19065	29777.2	22					0
457.26589	54206.4	9					0
457.27709	26286.6	9					0
457.28503	18456.7	C19H43N6OPS	[M+Na]+	457.28489	0.31		0
457.28503	18456.7	C30H36N2O2	[M+H]+	457.28495	0.17		0
457.28503	18456.7	C15H40N10O2S2	[M+H]+	457.28499	0.09		0
457.28503	18456.7	C24H49OPS	[M+(41K)]+	457.28545	-0.92		0
457.28503	18456.7	C15H32N14O3	[M+H]+	457.28546	-0.93		0
458.19872	30754.4	18					0
458.22477	73274.4	16					0
458.26111	38462.8	10					0
458.29749	24530.3	7					0
459.18513	35199.1	24					0
459.19642	36871.3	26					0
459.24514	37481	14					0
460.27673	51593.5	9					0
460.31307	19610.1	C31H42NP	[M+H]+	460.31276	0.67		0
460.31307	19610.1	C21H41N5O6	[M+H]+	460.31296	0.24		0
460.31307	19610.1	C18H45N9S	[M+(41K)]+	460.31314	-0.15		0
460.31307	19610.1	C16H38N13OP	[M+H]+	460.31327	-0.43		0
460.31307	19610.1	C23H48N3P3	[M+H]+	460.31339	-0.69		0
461.22449	25653.8	17					0
....	....	....	....	....	....	....	....