The Use of Recently Developed Mass Spectrometry-Based Proteomic Approaches for the Study of *Methylocella silvestris* BL2

Nisha Arunkumar Patel B.Sc. (Hons)

A thesis submitted for the degree of Doctor of Philosophy

University of Warwick
School of Life Sciences

September 2012

Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to my supervisor Professor Jim Scrivens and Susan Slade, for firstly selecting me for a technician role and for their continued encouragement and belief in myself, which led me to pursue my PhD studies. I am grateful for their support and expertise, which has enabled me to exceed expectations in my research and in myself.

I would like to acknowledge my collaborators Professor Colin Murrell, Dr Yin Chen and Dr Andrew Crombie for their contributions and support.

This work was financially supported by the Engineering and Physical Sciences Research Council (EPSRC) and Royal Society of Chemistry (RSC).

I am lucky to have met like-minded colleagues during my time working in the Scrivens group, the majority of which are now very special friends. I would like to thank members past and present; Sarah, Vib, Richard, Charlie, Gill, Jonathon, Fran, Elle, Matt, Eamonn, George, Baharak and Krisztina. I would like to separately thank my friend Kostas. Although I have never been his responsibility, he has always found time for me, even after leaving the group.

I also thank John and Marie Goodwin for essentially being second parents. They have made my time at Spencer Avenue a safe and enjoyable one and have understood the stresses associated with my studies.

Finally, I would like to thank my family for their love and care, and supporting my decision to further educate myself.

Declaration

I hereby declare that this thesis, submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy and entitled "The Use of Recently Developed Mass Spectrometry-Based Proteomic Approaches for the Study of *Methylocella silvestris* BL2", represents my own work and has not been previously submitted to this or any other institution for any degree, diploma or other qualification. Work undertaken by my colleagues is explicitly stated where appropriate.

Nisha A. Patel

September 2012

Summary

The study of the protein complement, termed proteomics, has advanced over the last twenty years as a consequence of developments in mass spectrometry. Currently, improvements in mass spectrometry-based approaches are targeted towards achieving information on both the identity and abundance of proteins.

Increased numbers of protein identifications are obtained by simplifying the analyte of interest. This can be achieved with the use of separation techniques, including two-dimensional liquid chromatography (2D-LC). Ion mobility coupled to mass spectrometry has recently been shown to be a useful post-ionisation separation tool for proteomic studies. The utility of these technologies for obtaining both qualitative and quantitative information is not extensively addressed in the current literature.

The use of a recently developed 2D-LC system, together with a method of ion mobility separation and a label-free quantitative approach for proteomic studies has been evaluated here for characterising the proteome of the bacterium *Methylocella silvestris*. This bacterium is the first methane-utilising bacteria also discovered to grow on substrates containing carbon-carbon bonds, and has great biotechnological potential. The metabolism of this bacterium was studied by obtaining information on its soluble proteome when grown with methane, propane, succinate, acetate, methanol, methylamine or trimethylamine.

The benefits and limitations of 2D-LC and ion mobility for profiling and label-free quantitative studies were demonstrated for simple mixtures and complex bacterial extracts. The combination of both 2D-LC and ion mobility was also achieved, resulting in wider proteome coverage when compared to the respective stand-alone approaches.

A cluster of expressed genes that were greatly up-regulated under trimethylamine growth and monomethylamine growth were proposed to be involved in the indirect pathway for trimethylamine metabolism. It was further verified that one of these genes expresses the previously unidentified trimethylamine monooxygenase. A propane assimilation route was proposed, based on information obtained on the levels of primary oxidation enzymes and downstream central metabolic pathways.

Abbreviations

1D	One-dimensional
1D RP-LC	One-dimensional reversed phase nano ultra pressure liquid chromatography
2D	Two-dimensional
2D RP-RP-LC	High pH reversed phase – Low pH reversed phase nano ultra pressure liquid chromatography

A

ACN	Acetonitrile
ADC	Analog-to-digital converter
ADH	Alcohol dehydrogenase
AMRT	Accurate mass retention time
ATP	Adenosine triphosphate
AUC	Area under the curve

B

BEH	Bridged ethyl hybrid
BLAST	Basic Local Alignment Search Tool
BSA	Bovine serum albumin

C

CDIT	Culture-derived isotope tags
cDNA	Complementary deoxyribonucleic acid
CID	Collisionally-induced dissociation
CsI	Caesium iodide
CV	Coefficient of variation

D

Da	Dalton
DAVID	Database for Annotation, Visualization and Integrated Discovery
DC	Direct current
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DiGE	Differential in-gel electrophoresis
DNA	Deoxyribonucleic acid

E

ECD	Electron capture dissociation
EM	Electron multiplier

EMC	Ethylmalonyl-CoA pathway
emPAI	Empirical protein abundance index
ENO	Enolase
ESI	Electrospray ionisation
ETD	Electron transfer dissociation
eV	Electron volt

F

FAB	Fast atom bombardment
FAD	Formate dehydrogenase
FADH	Formaldehyde dehydrogenase
FMO	Flavin-containing monooxygenase
fmol	Femtomole
FPR	False positive rate
FTICR	Fourier transform ion cyclotron resonance
FWHM	Full width at half maximum

G

g	Centrifugal acceleration
Gb	Gigabyte
GFP	[Glu^1]-Fibrinopeptide B
GMA	γ-glutamylmethylamide
GMAS	γ-glutamylmethylamide synthetase
GOGAT	Glutamate synthase
GRAVY	Grand average of hydropathicity
GS	Glutamine synthetase

H

h	Hour
H_4F	Tetrahydrofolate
H_4MPT	Tetrahydromethanopterin
hADC	Hybrid analog-to-digital converter
$HDMS^E$	Ion mobility-assisted data-independent acquisition
HILIC	Hydrophilic interaction liquid chromatography
HPLC	High performance liquid chromatography
HSS	High strength silica

I

ICAT	Isotope-coded affinity tagging
ICPL	Isotope-coded protein label
IEF	Isoelectric focussing
IMMS	Ion mobility mass spectrometry
iTRAQ	Isobaric tags for relative and absolute quantification

K

KAAS	KEGG Automatic Annotation Server
kDa	Kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
kpsi	Kilo pound per square inch
kV	Kilovolt

L

LC	Liquid chromatography
LIT	Linear ion trap
LOD	Limit of detection
LOQ	Limit of quantification
LTQ	Linear ion trap

M

M	Molar
m/s	Metres per second
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionisation
MCAT	Mass coded abundance tagging
MCP	Microchannel plate detector
MD LC	Multidimensional liquid chromatography
MDH	Methanol dehydrogenase
MeOH	Methanol
mg	Milligram
min	Minute
mL	Millilitre
mM	Millimolar
MMA	Monomethylamine
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
ms	Millisecond
MS/MS or MS^n	Tandem mass spectrometry
MS^E	Alternating low/elevated collision energy data-independent acqusition
μg	Microgram
μL	Microlitre
μm	Micrometre
μM	Micromolar
μs	Microsecond

N

NADH	Nicotinamide adenine dinucleotide
NaI	Sodium iodide
ng	Nanogram
nm	Nanometre
NMG	*N*-methylglutamate
NMGDH	*N*-methylglutamate dehydrogenase
NMGS	*N*-methylglutamate synthase
ns	Nanosecond

O

oa	Orthogonal acceleration
OD	Optical density
ORF	Open reading frame

P

PAGE	Polyacrylamide gel electrophoresis
PAI	Protein abundance index
PCR	Polymerase chain reaction
PhosB	Glycogen phosphorylase *b*
PIPES	Piperazine-N,N'-bis(2-ethanesulfonic acid)
PLGS	ProteinLynx Global Server
PMF	Peptide mass fingerprinting
pMMO	Particulate methane monooxygenase
PMO	Propane monooxygenase
ppm	Parts per million
PQQ	Pyrroloquinoline quinone
PRIDE	Proteomics identifications database
PTM	Post-translational modification

Q

QIT	Quadrupole ion trap
qPCR	Quantitative polymerase chain reaction
Q-TOF	Quadrupole time-of-flight

R

RF	Radio frequency
RNA	Ribonucleic acid
RPLC	Reversed phase liquid chromatography
RT	Retention time
RT-PCR	Reverse transciptase polymerase chain reaction
RuMP	Ribulose monophosphate pathway

S

s	Second
SAGE	Serial analysis of gene expression
SCX	Strong cation exchange
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SILAC	Stable isotope labelling in cell culture
sMMO	Soluble methane monooxygenase

T

TCA	Tricarboxylic acid cycle
TDC	Time-to-digital converter
TMA	Trimethylamine
TMAO	Trimethylamine *N*-oxide
TOF	Time-of-flight
TPR	True positive rate
TWIMS	Travelling-wave ion mobility spectrometry

U

UPLC	Ultra performance liquid chromatography
UV	Ultra-violet

V

V	Volt

X

XDIA	Extended data-independent acquisition
XIC	Extracted ion current

Standard 3 and 1 letter amino acid abbreviations are used throughout this thesis.

Amino acid	**3 Letter Code**	**1 Letter Code**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Chapter 1

Introduction

1.1 From Genome to Protein Function

After over a year of sequencing effort and a cost which accumulated to near $2 million, the bacterium *Haemophilus influenzae* was the first organism to have its genome completely sequenced (Fleischmann *et al.* 1995; Fraser *et al.* 2002). Today, it is possible to sequence and even annotate an entire bacterial genome in a day at a significantly reduced cost (Nagarajan and Pop 2010). The ease at which bacterial genomes can be sequenced has led to an extensive collection of microbial sequences. Currently, GenBank lists 1790 fully completed bacterial and archaebacterial genomes and 5230 genomes near to completion (Cruz *et al.* 2011).

1.1.1 Post-genomic technologies

The raw genomic sequence encodes a number of genes that may be transcribed into messenger RNA (mRNA) molecules, which are in turn translated to produce protein. Genome annotation is the initial step in defining the complement of proteins and functional RNAs that are available to the cell. The use of computational algorithms for identifying functional genes in genomes has been essential for predicting genes and their functions, together with the location and structure of their protein products (Gao and Chen 2010). Based on the genome annotation of the bacterium *Methylococcus capsulatus* (Bath) in 2004 (Ward *et al.* 2004), a number of previous studies on carbon assimilation, nitrogen fixation and other key metabolic enzymes were confirmed simply by the presence of the respective components in the genome. The genome can also provide a blueprint of an organism's theoretical metabolic capability. A challenge in genome annotation is that a single gene can give rise to multiple distinct proteins due to sequence polymorphisms, post-translational modifications, and in eukaryotes, alternative splicing. The annotation of the genome alone has limited value in relation to a complete understanding of gene function and cellular physiology.

The genome is a static entity which is generally unperturbed under different environmental conditions in a population of cells. Variation arises due to different patterns in gene expression. Transcriptomics is the study of mRNA transcripts present and their relative abundance. The transcriptomics revolution started when DNA (complementary DNA, genomic or plasmid libraries) microarrays were

developed. A microarray is developed by immobilising DNA to a surface. mRNA can be extracted from a cell or tissue and used as a template to create cDNA molecules which are labelled with fluorescent dye. The labelled cDNA molecules are then allowed to hybridise with the corresponding DNA molecules present on the array, those DNA molecules that hybridise are then identified by fluorescence after laser excitation (Schena *et al.* 1995). Other techniques used to study the transcriptome include quantitative polymerase chain reaction (qPCR), serial analysis of gene expression (SAGE), and more recently, RNA-seq, which involves directly sequencing cDNA generated from RNA preparations. Within the last three years, the study of bacterial transcriptomes has undergone a revolution due to the development of novel tools and increased sequencing capacity, which has unexpectedly revealed that transcription in bacteria resembles that in eukaryotes in terms of complexity (Guell *et al.* 2011).

1.1.2 Limitations of genomics and transcriptomics

Analysis of DNA and RNA can provide a primary readout of the cell and important information regarding the state of gene expression. These molecules can ultimately result in the production of proteins, which are important biological molecules in the cell that carry out the majority of essential biochemical functions. The challenge of determining gene function must therefore be complemented with approaches used to characterise the system at the gene, transcript and protein level.

Numerous studies have been carried out comparing mRNA expression with protein concentration. The vast majority of these studies concluded that the correlation was poor (Gygi *et al.* 1999b; Maier *et al.* 2009). These comparisons, however, generally suffered from a number of biological and methodological limitations. Most of these studies were limited by the fact that mRNA levels are compared to protein concentrations that were measured on a different day or in a different laboratory. To understand gene expression comprehensively; transcription, mRNA degradation, translation and protein degradation need to be studied collectively. A recent study simultaneously measured protein and mRNA turnover and levels in mouse fibroblasts (at steady-state growth)(Schwanhausser *et al.* 2011). Proteins were reported to be five times more stable (median half-life 46 h) than mRNAs (9 h). In the types of cells used in these analyses, the authors observed that

protein levels were best explained by the rate of translation, followed by transcription rates, with mRNA and protein stability being less important. This study highlighted the importance of measuring the cellular state at the protein level. The rates of protein/mRNA synthesis and degradation are of course independent of cell type.

The activation, stability and localisation of a protein can be dependent upon post-translational modifications (PTMs). The activation of a protein can also be dependent on its location. There have been almost 400 PTMs identified to date, of which phosphorylation and glycosylation have been the most widely studied. When attempting to characterise PTMs, additional complexity can arise due to the fact than an individual protein can be modified at several sites, thereby producing the possibility of many isoforms, each of which can have a distinct biological activity. Protein modification cannot be determined directly using genomic approaches (Pandey and Mann 2000).

The cell has mechanisms which regulate protein function by proteolysis, recycling and sequestration in cell compartments. These mechanisms can affect the gene products but not the genes. Protein-protein interactions and the molecular composition of cellular structures such as organelles can only be determined at the protein level.

There are medically important biofluids that simply cannot be analysed at the gene or transcript level, such as blood, plasma, urine and spinal fluid. The absence of genetic information in these samples requires exploration at the protein level.

1.1.3 The origins of proteomics

Annotation of an open reading frame (ORF) in genomic data can provide an initial, but not definitive, indication of a functional gene. The analysis of protein-based gene expression was therefore a logical step to elucidate the biological function of genes. The term "proteome" was proposed by Wilkins in 1995 and was defined as the *prot*ein complement expressed by a gen*ome*, or by a cell or tissue type (Wilkins *et al.* 1995). Analysing large numbers of proteins from a cell line or organism actually pre-dated the post-genomic era, when a graduate student published his research on the development of two-dimensional gel electrophoresis (O'Farrell 1975). In this technique, proteins are separated according to isoelectric point by isolelectric focusing in the first dimension, and according to molecular weight by sodium dodecyl sulphate

electrophoresis in the second dimension. Proteins are visualised by adding a staining solution. O'Farrell was able to provide an image of a gel showing the resolution of 1100 spots from an *Escherichia coli* protein extract. His paper was quickly recognised by molecular biologists as a true technological breakthrough. The ultimate aim of researchers was to catalogue images of 2D gels with annotated protein spots in a database. These gels were used as reference maps to be accessed across different laboratories, which was why improving the reproducibility of the technique was so important.

A bigger challenge than improving the reproducibility of this technique was identifying the large number of proteins spots that were resolved. The techniques available at the time were expensive and/or time and labour intensive in characterising so many proteins in a high-throughput manner. Techniques such as N-terminal sequencing, amino acid analysis and immunoblotting could only be used after electophoretically blotting proteins to membranes (Wilkins *et al.* 1995).

The analysis of biological molecules using mass spectrometry was revolutionised by the development of the ionisation techniques electrospray ionisation (ESI) (Fenn *et al.* 1989) and matrix-assisted laser desorption ionisation (MALDI) (Karas and Hillenkamp 1988) in the late 1980s. A new rapid mass spectrometry-based protein identification technique, peptide mass fingerprinting, was described soon after in 1993 (Pappin *et al.* 1993). This method transformed mass spectrometry into a major large-scale, functional genomics technique. The aforementioned techniques and concepts are still core elements of all modern mass spectrometry-based proteomics and are discussed further in section 1.2.

1.1.4 The state of current proteomics

Proteomics encompasses a large range of techniques that allow different aspects of the field to be studied; these include identification and quantification of proteins, characterisation of protein-protein interactions, cellular localisation, post-translational modifications, protein structure and stoichiometry/orientation of protein complexes. Techniques that have been used to study these aspects include protein microarrays (MacBeath 2002), mass spectrometry (Aebersold and Mann 2003), large-scale two-hybrid analyses (Causier 2004), and high-throughput protein production and crystallisation (Stevens *et al.* 2001). It is unsurprising that protein microarrays have

not caught up with its respective transcriptomic technology in terms of throughput. Protein microarrays are most suited for targeted hypotheses requiring the analysis of a relatively few number of proteins. They entail producing antibodies that have been designed to target a specific protein, and may require a significant amount of protein to adhere to these antibodies. Unfortunately, there is no equivalent technique to PCR to amplify, therefore analysis is limited to proteins isolated from natural sources.

There is now a large amount of data publically available for transcripts, metabolites and proteins. Users may be overwhelmed with potential targets or pathways to explore. A main concern for users of this information is the quality of the data in these databases (number of false positive results), and only recently has it been accepted that more than one peptide should be used to confidently identify a protein. The confidence of users can be increased with follow up studies, which may be costly and time consuming. The PRoteomics IDEntifications (PRIDE) database is a centralized, standards compliant, public data repository for proteomics data. It has been developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications. There are no obligations for laboratories to share their proteomic data, and generally, the primary users of this information are typically the generators of these data. A major issue that needs to be addressed with any discussion of proteomics data is how the user has reported the quality of data acquired.

The most widely used application of proteomics is the comparative analysis of biological samples that differ in specific physiological phenotypes, such as diseased and normal. Mass spectrometry-based proteomics has successfully identified previously unknown pathways with key roles in a wide range of biological systems, these have included a metabolic pathway in aggressive cancer cells and phosphoprotein networks involved in DNA-damage response (Cravatt *et al.* 2007). Developments in proteomics have been focussed on identifying more and more proteins without consideration of the confidence of assignment. It is fair to say that the identification of as many proteins as possible is still a current goal, although the development and optimisation of MS methods to make quantitative measurements, whilst maintaining and where possible improving the quality of the experimental data that is obtained is becoming more important.

1.2 Instrumentation

Mass spectrometry is a rapid and sensitive technique, which has emerged as a preferred method in characterising protein components of biological systems. Proteomics today is underpinned by the need for continuous advances in instrument and method development. The fundamentals, advantages and limitations of mass spectrometers and commonly used proteomic approaches are presented below, with the choice of experimental approach for this project discussed.

1.2.1 Mass spectrometer components

The concept of measuring mass can be a simple, yet powerful method of identifying and characterising a biological molecule, determining accurate molecular mass together with the masses of its component building blocks after fragmentation (McLafferty 2008). The use of mass spectrometry (MS) to study biological molecules is a relatively recent application in the history of the technique, which started as physicists' tool to study small molecules and then after its commercialisation used by industrial chemists to characterise complex mixtures.

A mass spectrometer measures the mass-to-charge (m/z) ratio of ionised molecules in the gas phase. A modern mass spectrometer contains three main components; an ionisation source, mass analyser(s) and a detector (Figure 1.1).

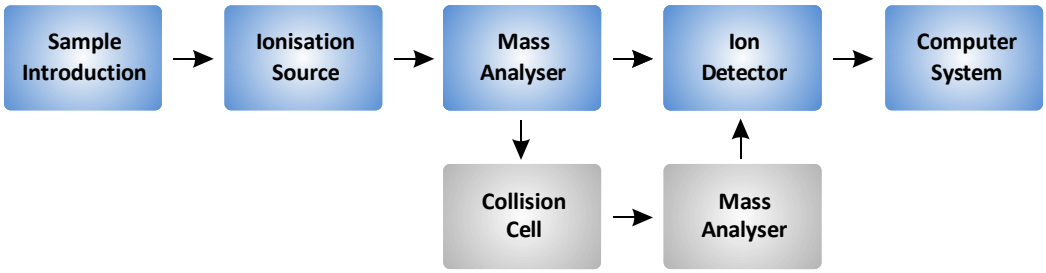

Figure 1.1. Schematic block diagram of a mass spectrometer
Elements in grey are used by tandem in space instruments.

There are two main ionisation methods which are used in mass spectrometers when utilised for proteomics applications. These are MALDI and ESI (section 1.2.2). Ions generated by these techniques are resolved by a mass analyser according to their m/z ratios. Commonly used mass analysers include quadrupoles, time-of-flight and the

Orbitrap (section 1.2.3). Ions can be fragmented by an experimentally appropriate method of dissociation and measured by a second mass analyser (in space) or by the same mass analyser (in time) (section 1.2.5). The detector registers the number of ions at each m/z value and converts it into a useable signal which can be displayed by the computer system (section 1.2.4).

1.2.2 Ionisation

Classical ionisation techniques including electron ionisation, photo ionisation and chemical ionisation require the vaporisation of the analyte into the gas phase and the ionisation of the gas phase analyte. Molecules of biological interest tend to be large and polar, and cannot be vaporised into the gas phase without extensive decomposition. Ionisation methods based on desorption, that is the direct formation or emission of ions from liquid or solid surfaces, were developed to circumvent the need for volatilization prior to ionisation. These methods included plasma desorption, laser desorption, fast atom bombardment (FAB) and thermospray ionisation. FAB combined with magnetic sector mass spectrometers was a powerful method in the early 1980s for the analysis of biomolecules up to 4000 Da, that were capable of being studied. A limitation of FAB ionisation coupled to mass spectrometers at this time was the decreasing efficiency of ionisation at increasing large M_r.

In the space of two years (1987-1988), two ionisation techniques demonstrated that biomolecules exceeding a molecular weight of 100,000 Da could be studied. These methods, matrix-assisted laser desorption ionisation (MALDI) and electrospray ionisation (ESI), revolutionised biomolecule analysis and have played a huge role in research within the life sciences. MALDI and ESI are the predominant ionisation techniques used in mass spectrometry-based proteomics approaches today. The theories and applications of these techniques are discussed below.

Matrix assisted-laser desorption ionisation (MALDI)

Karas and Hillenkamp introduced the technique in 1987 (Karas *et al.* 1987; Karas and Hillenkamp 1988), but it was Tanaka's earlier introduced but less practically applicable work on soft laser desorption (Tanaka *et al.* 1988) which led to the Nobel Prize being awarded to himself and John B. Fenn (developer of ESI).

The use of a matrix in this method serves the purpose of desorption and ionisation. The analyte to be studied is mixed with a solvent containing a matrix compound and then deposited onto a surface, the solvent evaporates allowing the co-crystallisation of the analyte and matrix. Under vacuum conditions, a laser is applied to the solid surface in pulses. One theory states that the energy supplied by the laser induces rapid heating of the matrix molecules, causing sublimation of the matrix molecules, which expand into a plume containing the intact analyte. It is not known for a particular system whether ionisation takes place in the solid phase prior to desorption or in the expanding plume in the gas phase, or if both mechanisms apply. In most cases, ions formed by this method predominantly hold one charge ($[M+H]^{1+}$).

The pulse mechanism of the laser in this approach means that ions are produced in bundles, which is well suited to analysis using a time-of-flight mass analyser (TOF). This method of ionisation has advantages for proteomics studies as it is amenable to samples containing salts, which are predominant in purification procedures. It is effective for profiling less complex samples, but it is not ideal for quantitative measurements. The technique can suffer from poor reproducibility, which can be especially true when complex samples are analysed. Non-linear relationships have been observed between relative signal intensities and concentrations (Szájli *et al.* 2008).

Electrospray ionisation (ESI)

ESI involves the application of a strong electric field, under atmospheric pressure, to a liquid containing the analyte passing through a capillary tube. This field induces a charge accumulation at the liquid surface located at the end of the capillary, which breaks into highly charged droplets containing the analyte. There are two main theories that suggest how gaseous analyte ions are formed from this highly charged droplet: (i) the charged residue theory and (ii) the ion evaporation theory, shown in Figure 1.2.

The charged residue theory proposes that the charged droplet undergoes cycles of solvent evaporation and subdivision into smaller droplets. Subdivision is thought to occur when the coulombic repulsion forces on the surface overcomes the surface tension of the droplet over the same area, this is commonly known as the Rayleigh limit. Solvent evaporation and subdivision occurs until small droplets are formed that contain one analyte molecule. As the last solvent molecules evaporates from this

ultimate droplet, the residual analyte molecule would retain some or all of the charges on the droplet to become a gas phase analyte ion (Dole *et al.* 1968). A single analyte can result in a broad range of ions which differ in the number of charges held, where $[M+nH]^{n+}$ and $n \geq 1$. An alternative explanation was presented by Iribane and Thompson, who argued that before a charged droplet became small enough to contain only one analyte molecule, the charge density on the surface would become so high that the resulting field would push or evaporate analytes out forming gaseous charged ions (ion evaporation theory).

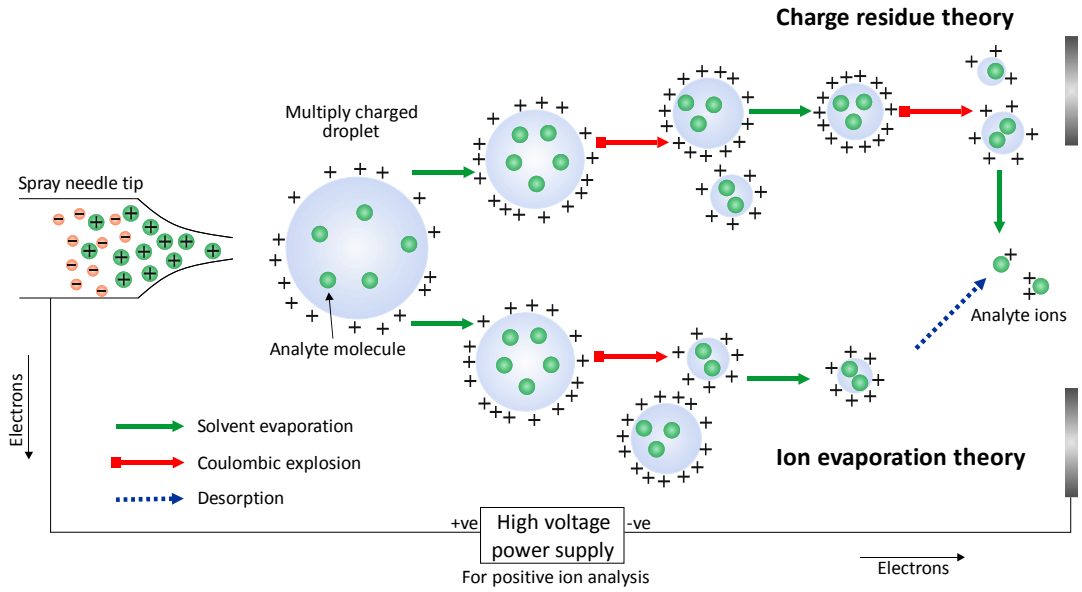

Figure 1.2. Electrospray ionisation theories

Adapted from (Nguyen and Fenn 2007)

Samples analysed by this method are required to be solubilised in an aqueous solvent. Typically, an organic modifier is added, which aids in the desolvation process and makes the analyte more amenable to protonation (addition of charge) by the act of denaturation. The solvent can additionally be acidified to promote protonation. It has been widely documented that ESI response can vary significantly among different analytes that have identical solution concentrations. In the case of peptides, the response is higher for non-polar amino acids when compared to polar ones (Cech and Enke 2001). These inherent properties of ESI are considered when it is used to make quantitative measurements in proteomics experiments. These are discussed further in section 1.4.

1.2.3 Mass analysers

The mass analyser used is central to MS instruments, so much that instruments are commonly named after the mass analysers they contain. There are a number of different types of mass analysers, which can generally be grouped into those that scan, detecting ions of each m/z in turn, and those that detect ions of all m/z simultaneously. For proteomics research, five predominant mass analysers have used; quadrupole (Q), ion trap (quadrupole ion trap, QIT; linear ion trap, LIT/LTQ), time-of-flight (TOF), Fourier-transform ion cyclotron resonance (FTICR) and the Orbitrap mass analyser. Hybrid instruments have been built to include the combination of various mass analysers; examples include Q-q-Q, Q-q-LIT, Q-TOF, TOF-TOF and LTQ-FTICR. The configurations of these various mass analysers in mass spectrometers for the purposes of proteomic experiments are discussed in section 1.3.1. The individual properties of quadrupole and time-of-flight mass analysers are discussed further, as are the Q-TOF-type instruments which were used in this research.

Quadrupole mass analyser

The quadrupole is a scanning mass analyser. It consists of four parallel conductive rods that are arranged as two pairs of electrically connected rods. A combination of direct current (DC) and alternating radio frequency (RF) potentials are applied to each pair of rods. This causes ions to traverse the centre of the quadrupole mass analyser in a complex oscillating motion. By changing the electric field that the ions are exposed to, all ions can be passed through the quadrupole or the passage restricted to selected ions of a particular m/z. If the quadrupole is operated in scanning mode, in which a particular value of DC and RF voltages are applied to the rods, then ions of a specific m/z will have a stable trajectory. In this mode, a selected ion will pass through the analyser and into the detector. The spectrum may then be scanned by varying the DC applied. The remaining ions, of different m/z value, will either collide into the rods or be lost at the walls of the analyser. The quadrupole can also act as an ion guide. In this mode, it does not discriminate ions based on their m/z value, rather, suitable RF voltages are applied to all four rods to optimise stable trajectories for all ions to pass through.

Time-of-flight mass analyser

The time-of-flight analyser measures the m/z of an ion based on the time it takes for that ion to traverse a field-free region. An electrical pulse (V) pushes a group of ions of different mass (m) with a fixed amount of kinetic energy (E_k), which provides them with different velocities (v) (Equation 1.1). This means they will traverse a field-free region of fixed distance in different times (Equation 1.2). This time is related to their mass and charge (Equation 1.3) (de Hoffmann and Stroobant 2002). Ions travel at a velocity inversely proportional to the square root of their mass. In general, the larger the ion, the longer it takes to traverse the tube.

For a fixed E_k, an ion with a mass (m) and total charge (ze), where z is the number of charges and e is the electronic charge, will have a velocity (v) determined by:

$$v = \sqrt{\frac{2E_k}{m}} = \frac{\sqrt{2zeV}}{m}$$

Equation 1.1

The time (t) an ion takes to traverse a flight tube of fixed length (d) is inversely proportional to v:

$$t = \frac{d}{v}$$

Equation 1.2

Given that v is a constant and V and d are known:

$$t = \sqrt{\frac{m}{z}} \times constant$$

Equation 1.3

For each pulse, the detector records a corresponding spectrum, called a transient. Many transients are summed to create a mass spectrum. The resolving power of a TOF mass analyser can be affected by multiple factors. These include; variations in the kinetic energies given to the ions, the length of the ion formation pulse and the length of the flight tube. Variations in kinetic energy given to ions of the same m/z will cause ions to reach the detector at slightly different times, resulting in peak broadening and reduced resolution. As mass resolution is proportional to ion flight time, higher resolving power can be achieved with longer flight tubes.

Higher mass resolution is achieved by reflectron instruments by reflecting ions back down the flight tube by the use of a series of ring electrodes, which act as a mirror, shown in Figure 1.3. The use of a reflectron effectively doubles the flight path of an ion and compensates for differences in initial kinetic energies of ion. Ions with higher kinetic energies travel further into the reflectron compared to those with less kinetic energies. The result of this is that both ions reach the detector simultaneously. The consequences of using a reflectron for improved resolution can include reduction in sensitivity and a reduction in upper mass limit of detection.

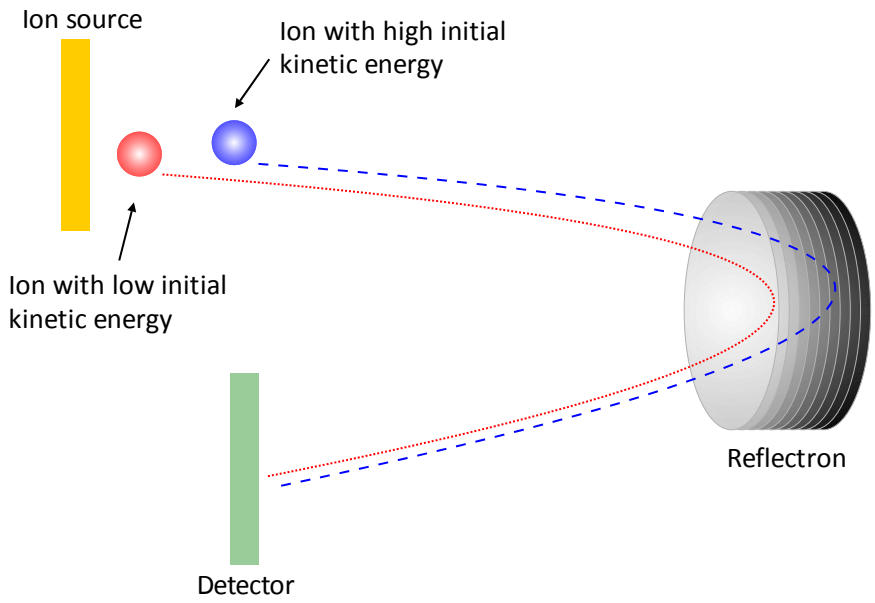

Figure 1.3. Schematic of time-of-flight (TOF)

Improved resolution can be obtained by applying a controlled delay between ion formation and acceleration. This is referred to as pulsed ion extraction or time-lag focussing. Ions that are formed with high kinetic energy move further away from the source electrode and take up less energy from the pulse applied. If correct times and voltages are applied, all ions of the same m/z will arrive at the detector simultaneously.

1.2.4 From detector to mass spectra

Ions of varying m/z that have been separated by the mass analyser are detected and transformed into a useable signal by the detector. Several types of detectors exist and the choice of detector is dependent on the instrument design and the application. For example, TOF instruments require detectors with very rapid readout and response. Detectors may be classified into two major categories; (i) those that directly measure ions that impinge upon a surface, such as photographic plates and (ii) those that are based on the amplification of an ion signal, the most popular of which are electron multipliers (EM).

The EM is the most common ion detector in use today. Figure 1.4(a) shows the schematic of a single continuous dynode, which is capable of acting solely as an EM detector. Ions striking this continuous dynode produce secondary electrons. The released electrons in turn strike the surface, emitting more electrons. This process is repeated generating an electron cascade, which effectively multiplies a single ion incident by 10^6 or more.

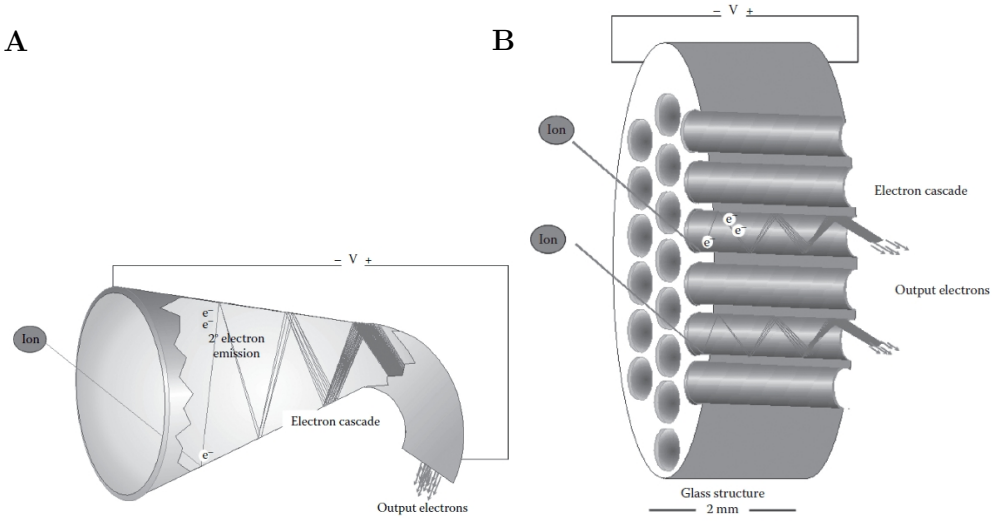

Figure 1.4. Composition of a multichannel plate detector
(a) Schematic of a single continuous dynode. (b) Multichannel plate. Taken from (Parker *et al.* 2010)

The multichannel plate shown in Figure 1.4(b) is a variation on the continuous dynode EM. In this configuration, a series of microchannels are arranged a few tens of micrometers in diameter and a few millimetres in length. Short electron pulse widths (~1 ns) can be achieved due to the length of these microchannels, which makes it an ideal detector or detector component for TOF mass spectrometers. Similar to

the operation of the continuous dynode, ions strike the surface near the entrance of the microchannel, initiating a cascade of electrons. The cloud of ions exiting the microchannel can be directed onto an anode where it can be detected and converted into a mass spectrum by two methods, namely time-to-digital or analog-to-digital conversion.

Traditionally, TOF mass spectrometers have used a detector with a time-to-digital converter (TDC). A TDC records an arrival time for each mass within a transient (ions resulting from one TOF pulse). A TDC records the arrival time for the first ion of a given mass, so it cannot tell the difference between one ion at a given mass or several ions at that given mass arriving in a transient. The result is highly sensitive single ion detection capability, however, the inability to determine how many ions are arriving at a particular time can severely limit the instrument's dynamic range.

An analog-to-digital converter (ADC) tries to capture all the ions of a given mass arriving in a single transient. In this mode, the total signal strength is recorded at very frequent intervals (as often as a billion times per second). The ADC can range between one ion of a given mass to many ions of a given mass arriving in a single transient. This results in an improved dynamic range and an ability to cope with more abundant species when compared to the TDC.

Recently, a hybrid detector, termed an hADC, has been described which offers the advantages of both ADC and TDC detectors. Peak detection is applied on signals from each TOF push, whilst maintaining the speed, high resolution, and low electronic noise signature of TDC detection with the enhanced dynamic range afforded by ADC electronics (Mazzeo *et al.* 2010).

1.2.5 Tandem mass spectrometry

Mass spectrometers containing either single (depending on the type) or multiple mass analysers can be capable of carrying out tandem mass spectrometry experiments (MS/MS or MS^n). In general, for mass spectrometers containing two mass analysers, an ion of interest is selected by the first mass analyser, dissociated into smaller fragments, and then measured using the second mass analyser.

Tandem mass spectrometry experiments are important for obtaining primary sequence information of peptide ions. A widely used technique involves selecting an

ion of a particular m/z (the precursor), allowing it to collide with an inert gas and then analysing the fragment ions formed (known as product ions). This process is known as collisionally induced dissociation (CID) and was pioneered by Jennings (Jennings 1968) and McLafferty (McLafferty and Bryce 1967). When a dissociation technique is applied to a peptide ion, fragmentation has been observed to occur at a number of different sites. Figure 1.5 shows the nomenclature for the possible types of fragment ions that can be produced from a peptide proposed by Roepstorff, Fohlmann and Beimann (Roepstorff and Fohlman 1984), (Biemann 1992).

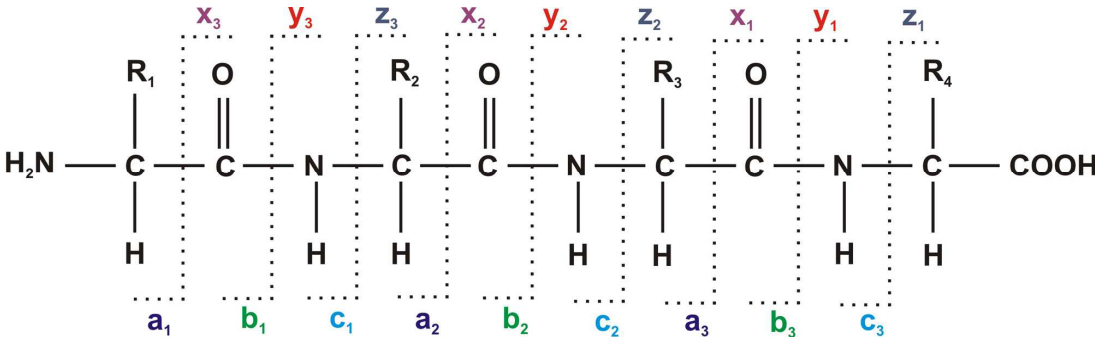

Figure 1.5. Peptide fragmentation nomenclature

Low energy CID typically produces a series of b and y type ions from the peptide fragmenting at the amide bonds of the peptide backbone. The respective ions are formed when the charge is retained on either the N- and C-terminal fragment. This process can produce a collection of consecutive fragment ions that differ in mass by a single amino acid residue, allowing one in principle to interpret the amino acid sequence of the peptide. This mode of fragmentation can be carried out on a wide range of mass spectrometers, has reproducible fragmentation patterns, and can produce a significant coverage of fragment ions from a typical tryptic peptide (approximately 20 amino acids residues or less). Some labile PTMs are prone to dissociation under CID conditions, making them difficult to analyse by this technique. Electron capture dissociation (ECD) (Zubarev *et al.* 1998) and electron transfer dissociation (ETD) (Syka *et al.* 2004) are relatively new fragmentation techniques, which dissociate peptides in a non-ergodic manner, keeping PTMs intact and inducing random breakage of the peptide backbone from the cleavage of the Cα-N bond to produce complementary c and z ions. ECD is currently limited to use on a Fourier transform ion cyclotron resonance mass spectrometer, which is not a widely accessible

instrument and requires significant technical expertise. ETD is thought to create fragment ions in a pathway analogous to ECD and is available on a variety of mass spectrometers such as quadrupole ion traps, Orbitraps, and more recently some Q-TOFs. These analysers are lower cost, lower maintenance and more widely-accessible when compared to FTICRs. For the products of ECD and ETD to be observable, the precursor ion must hold at least two, but preferably three or more, positive charges. The capture of an electron by a singly charged ion would result in a net charge of zero, which is undetectable by mass spectrometry. These techniques have been shown to produce more diagnostic fragment ions than CID when used to study large peptides and proteins, and have been shown to produce similar results for tryptic peptides. A large number of proteomics experiments have been described which combine the use of ETD and CID for more comprehensive peptide sequence analysis (Kim *et al.* 2011; Shen *et al.* 2011).

1.2.6 Data-dependent and data-independent acquisition

Mass spectrometers can be automated to carry out multiple tandem mass spectrometry experiments. This is essential for analysing samples containing thousands of peptides. It is, however, currently too technically challenging to sequentially select every single observed ion and perform MS/MS on each one. Proteomic experiments can be focussed on the identification and quantification of as many peptides/proteins as possible. This can be achieved by separating peptides by liquid chromatography and collecting product ion spectra on the eluting peptides by data-dependent acquisition (DDA), or by data-independent acquisition (DIA). The optimum choice of method depends on the complexity of the sample, the method of quantification and the type of mass spectrometer.

A DDA experiment is a serial process. The cycle starts by looking for intense peptide molecular ions (typically 3 – 10) that exceed a user-defined threshold value. Each of these peptide ions are serially isolated by the mass analyser for MS/MS acquisition during an allotted period of time, or until a certain ion current value is met. After product ion spectra are collected on each peptide, the cycle starts again by looking for the next set number of intense peptides, usually on the condition that the previous masses have been excluded so they will not be selected again. A complex biological sample typically contains a large majority of proteins that are at least two

to three orders of magnitude lower in concentration that the most abundant protein. This poses a challenge in DDA-based experiments. Typically, the number of selected precursors will increase and the MS/MS acquisition time for each molecular ion decreases with increasing sample complexity. For low abundant peptide ions, a longer amount of acquisition time needs to be spent collecting MS/MS data in order to reach a preset threshold, during which important information is lost on other peptide ions present at that time. If there are many co-eluting ions, there may be insufficient time to acquire MS/MS data for all ion species and thus some low-abundance ions may never be identified.

The major advantage of the DDA approach is the specificity of ion selection which aids primary structural information. This approach can suffer from a lack of reproducibility, low sequence coverage and may result in a large number of protein identifications based on a single peptide. Due to the nature of the DDA experiment, precursor intensities obtained using this method cannot be used to infer quantification of a given protein (Geromanos *et al.* 2009). Section 1.4 describes how some quantitative methods have been developed to take advantage of the selective nature of DDA.

Some mass spectrometers are capable of data-independent acquisition (DIA). This technique exists in different forms, but by definition all of these acquire precursor ion and product ion data irrespective of what ions are present. These methods do not involve the selection of a single precursor at any one point for subsequent MS/MS. They either fragment all or a selected fraction of precursor ions. This results in complex product ion spectra which may contain fragment ions from a number of co-eluting precursors. The major challenge in these approaches is the development of software to correlate which fragments originated from which precursor ions. DIA has recently been demonstrated using a fast scanning linear ion trap. In this approach, windows of 10 m/z are sequentially isolated for MS/MS until a desired mass range has been covered. The technique was shown to improve the signal-to-noise compared to DDA, and demonstrated the benefits of the DIA approach for quantification, however, no attempt was made to distinguish fragment ions resulting from more than one precursor (Venable *et al.* 2004). A similar approach, termed extended data-independent acquisition (XDIA), was developed using an Orbitrap mass spectrometer (Carvalho *et al.* 2010). An XDIA experiment involves the use of a high-resolution scan of precursors and a series of consecutive events of 20 m/z

isolation windows overlapping each other by 1 m/z. Fragmentation was achieved using ETD followed by CID in two synchronised scan events. Specialised software was developed to process this data for protein identification. This technique was shown to provide more detailed product ion information and hence more complete peptide identification when compared to DDA.

Section 1.4.3 describes a version of DIA termed MS^E. This method has been shown to identify more peptides when compared to the use of DDA on the same instrumentation, and quite importantly can be processed with specialised software to provide both identification and quantification information. It is the most widely used DIA method available to date and has been employed in a number of proteomics-based experiments since its introduction in 2005. The development of MS^E and its applicability to proteomic experiments is discussed in section 1.4.3.

1.2.7 Liquid chromatography

Technological advances in liquid chromatography have played a major role in mass spectrometry-based proteomics. As a standalone technique, mass spectrometry would not be able to analyse a complex mixture containing tens of thousands of components of varying concentration. Detectors have a limited dynamic range and ESI can be sensitive to ion suppression and itself has a limited dynamic range (~10^5). The relationship between LC and MS is a symbiotic one. The use of a mass spectrometer acts as a real time detection system and yields a wealth of information that is unattainable by other detection methods traditionally used with LC such as UV.

High performance liquid chromatography (HPLC) was developed in the 1960s (Giddings 1963), (Horvath and Lipsky 1966) but was not tested for the purposes of peptide separation until approximately 10 years later. This approach is the LC method of choice to be interfaced with MS. All HPLC methods have in common two phases; a stationary phase, which is designed to differentially interact with the compounds of interest and a mobile phase (in this case a liquid), which is designed to differentially elute substances of interest. The two phases are kept in contact and analytes partition differentially between the two. Physical properties of compounds such as size, binding affinities, hydrophobicity, charge and pI have been used to separate compounds of interest.

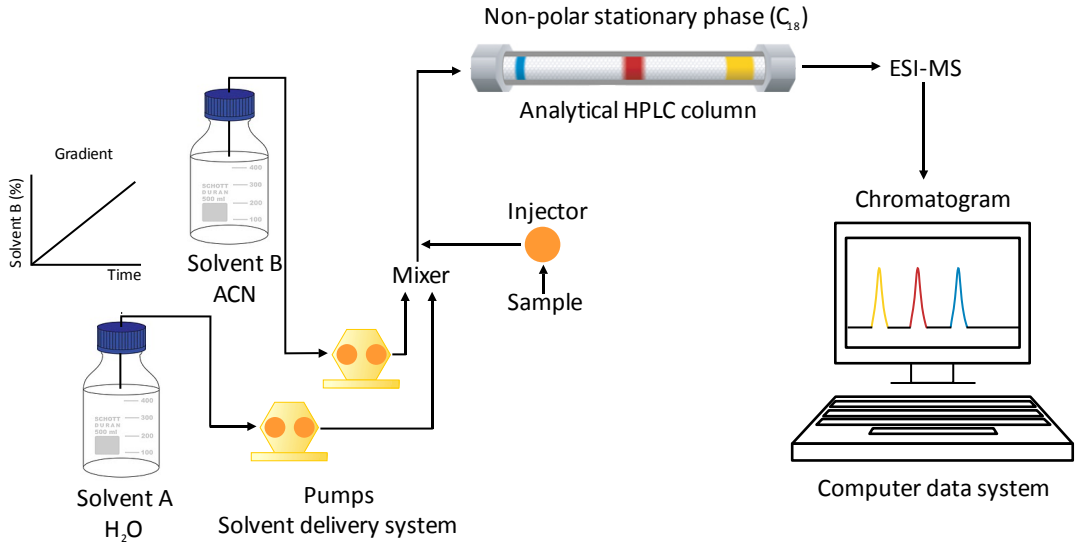

Figure 1.6. Schematic of reversed phase LC under gradient conditions
Adapted from (McDonald 2009)

Reversed phase-HPLC (RPLC) has emerged as the preferred method for peptide separation. Peptides are separated on a column containing sorbent particles bridged to non-polar functional groups (e.g. C_{18}). Organic non-polar analytes (e.g. peptides) from polar solutions (e.g. water) are able to be retained by these non-polar functional groups via non-polar attractive forces. In order to elute peptides, a mobile phase with sufficient polar character (e.g. acetonitrile or methanol) is required to disrupt the binding mechanism between the analytes and the stationary phase. RPLC is usually operated under gradient conditions to achieve optimal separation. This entails eluting peptides in a mixture of aqueous and organic solvent, whereby the concentration of organic content is increased over a period of time (gradient time). This results in hydrophilic peptides eluting at the start of the gradient, and increasingly hydrophobic peptides eluting throughout the rest of the gradient (Figure 1.6). RPLC benefits from the ability to analyse peptides dervied from experimentally challenging proteins (membrane, highly acidic/basic etc.) which can be difficult to analyse by PAGE methods. A major advantage of RPLC is that the buffers used are compatible with ESI, making it a suitable system to be coupled directly to a mass spectrometer.

Advances in both LC and MS technologies impact directly on development of mass spectrometry-based proteomics. One of the main areas that has been targeted in LC development is improvement in separation in order to reduce the complexity of a sample prior to MS analysis. The separation efficiency of HPLC under gradient conditions can be described as the peak capacity (P), which represents the maximum number of components that could be theoretically separated on a column with a

gradient time (t_g). There are three main strategies that have been employed to improve peak capacity: (i) decreasing the gradient slope (column length is fixed), (ii) increasing the column length (L) with a proportional increase in gradient time, and (iii) employing columns packed with smaller sorbent particles. Strategies involving increasing t_g and L have diminishing returns, as peak capacity does not increase linearly with these two variables. Column back-pressure and column efficiency are inversely proportional to the square of the particle diameter. This means that as the particle size decreases, the column back-pressure and efficiency increase. The reduction of particle size must result in pressures which are experimentally practical. An important advancement in HPLC, called ultra performance liquid chromatography (UPLC), successfully employed the use of smaller particles at elevated temperatures operating under an ultrahigh pressure regime (20 kpsi) (Shen *et al.* 2005), and was shown to double the number of protein identifications when compared to conventional HPLC (Motoyama *et al.* 2006). The increase of the column length while decreasing particle size using UPLC has led to improved peak capacity, resolution, sensitivity, and analysis time.

The combination of two or more LC separations, often referred to as multi-dimensional LC, can be an attractive option since it has the potential to significantly improve the separation power. This approach entails fractionation of peptides into discrete (off-line) or discontinuous (on-line) fractions, and then further analysis of each one of the fractions. When coupling together two or more LC separations, it is important that the two separation approaches are orthogonal to each other as possible, meaning that they utilise distinct approaches to differentially retain compounds. Multi-dimensional LC techniques used in proteomic applications are discussed further in Chapter 3.

1.3 Mass Spectrometry-Based Proteomics

The main steps involved in mass spectrometry-based proteomics experiments include; sample preparation, protein and/or peptide fractionation, mass spectrometry, and data processing to infer protein identification and quantification. An overview of the experimental and theoretical approaches used in a mass spectrometry-based proteomics experiment is presented in Figure 1.7.

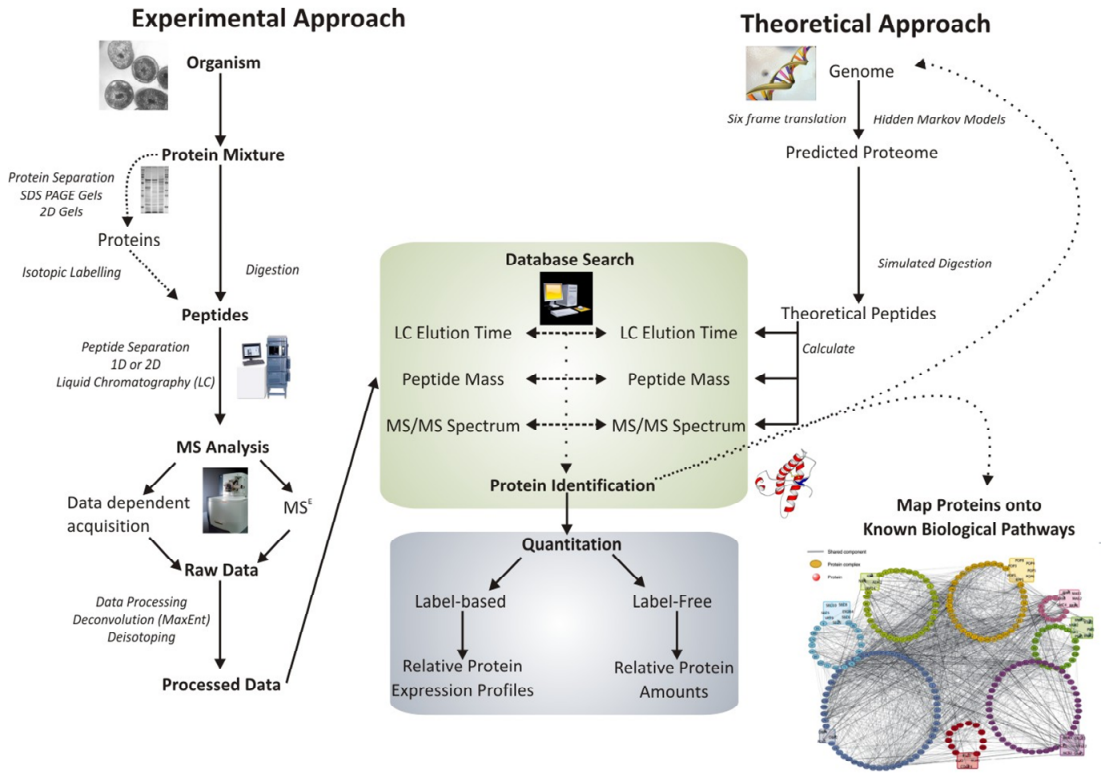

Figure 1.7. Mass spectrometry-based proteomics overview

Proteomics approaches can be generally categorised based on how a sample is treated prior to mass spectrometry analysis. Methods include: (i) separation on the peptide level (shotgun proteomics) (ii) on the protein level (top-down proteomics), (iii) or a combination of protein and peptide separation (bottom-up proteomics). In the following description of proteomic experiments, the terms profiling and qualitative describe the identification of proteins, and the terms quantitative and comparative relate to the quantification of proteins.

1.3.1 Bottom-up and shotgun approaches

Bottom-up and shotgun experiments are similar in that they both require proteins to be proteolytically digested into peptides. They both require that proteins are extracted from the organism or cell under investigation. In the bottom-up approach, proteins, not peptides, may be separated prior to proteolysis. One-dimensional and two-dimensional polyacrylamide gel electrophoresis (1D and 2D PAGE) are methods which have been used to separate and visualise proteins. Gel approaches can suffer from low resolution and only allow the analysis of the most abundant proteins. Other potential drawbacks include insufficient resolving power to fully separate all proteins within a sample and restricted throughput. This approach has a number of potential limitations including the fact that low abundance proteins are difficult to visualise and detect, and proteins with a high pI, high molecular mass and strongly hydrophobic proteins are difficult to analyse (Beranova-Giorgianni 2003). An alternative approach for protein separation is two-dimensional liquid chromatography, which separates by isoelectric point (chromatofocussing) and hydrophobicity (reversed phase). This method can have advantages over a gel-based approach since it benefits from higher resolving power and can provide the option of collecting and analysing proteins intact (as in top-down) or at the peptide-level since proteins can be collected and stored in discrete liquid fractions. The extraction of intact proteins from gels is extremely difficult, and so enzymatic digestion into peptides is the preferred approach to allow analysis using mass spectrometry. Trypsin is a popular choice of enzyme since it is cheap, robust, commercially available and has a high specificity, cleaving proteins after the basic residues arginine and lysine, unless they are adjacent to a proline. Tryptic peptides provide some advantages in positive ion mass spectrometry experiments due to the terminal basic amino acid conferring good sensitivity and predictable fragmentation.

The protein separation step is bypassed when using the shotgun approach. The whole proteome is subjected directly to proteolytic digestion producing a complex mixture containing tens of thousands of peptides. In the bottom-up approach, gel spots or liquid fractions of interest are selected for digestion which results in a relatively simple mixture of peptides. The next step in both approaches is to analyse the peptides using mass spectrometry. MALDI–TOF MS has been typically employed for simple mixtures of proteins, experimental molecular peptide ion mass information being matched with theoretical peptide masses. As sample complexity increases, the

use of peptide mass alone does not provide sufficient specificity to identify a protein with high confidence. In these cases, the samples are analysed using liquid chromatography coupled to tandem mass spectrometry (MS/MS) experiments, which generates structural information as well as intact peptide mass. The shotgun approach, which starts with a large complex peptide mixture, typically employs an additional peptide separation step before reversed phase LC-MS/MS experiments. Strong cation exchange (SCX) has been the most commonly used approach. This method separates peptides using an increasing salt gradient. Reversed phase LC is commonly employed as a secondary separation method since it is capable of desalting peptides from the SCX step.

Thousands of product ion mass spectra are generated from a proteomics experiment. These need to be analysed using bioinformatic approaches. If a database including protein sequences is available for the organism(s) under study, *in silico* digestion can be performed to generate a list of theoretical peptide masses and associated fragment ions, depending on the fragmentation mechanism and protease used. During a database search, experimental data is matched to theoretical data to infer identification of the peptide and the protein it originated from. If a database is unavailable for the organism, experimental data can be searched against a non-specific database to find closely related sequences in other organisms. Alternatively, data can be interpreted *de novo*, which involves manually assigning primary sequence from the MS/MS spectra.

1.3.2 Top-down approach

Top-down proteomics involves the identification and characterization of intact proteins from tandem mass spectrometry experiments and has the potential to identify post-translational modifications (PTMs), which may not be readily identified in a bottom-up approach. The fragmentation of whole proteins can lead to complex spectra which can be difficult to interpret. A Fourier-transform mass spectrometer (FTMS) has been the instrument of choice for this type of experiment due to the high resolution capability and mass accuracy that it affords. Intact proteins can be fragmented in a tandem MS experiment using ECD and sequence information may be deduced from the product ion spectra. This approach is currently limited to studying proteins of molecular weight 40-50 kDa and lower, since larger proteins exhibit more

stable structures which may prevent efficient fragmentation. This is an important factor to consider in studying the proteome, as a significant component (>50%) of the proteome has a molecular weight above this limit (Armirotti and Damonte 2010).

1.3.3 Mass spectrometer configurations for proteomics

Individual properties exhibited by mass analysers are major contributors to the overall performance of the instruments. Hybrid mass spectrometers have been built which contain more than one mass analyser tailored to specific requirements. The performance of mass spectrometers as a whole may be classified in terms of mass range, analysis speed, resolution, mass accuracy, sensitivity, ion transmission, and dynamic range.

For the analysis of tryptic peptides, an ability to measure from 50 – 2000 m/z is sufficient for ions generated from peptide and fragment ions.

The speed of analysis, or scan rate, can be an important factor when considering the complexity of a sample and the requirement to interface with LC. On-line coupling of UPLC to FTICR has been technically unfeasible due to the different timescales of the experiments.

The resolving power of a mass spectrometry analyser can be characterised as the ability to separate ions of different, but defined, m/z values. A common definition of resolution is the full width at half-maximum height (FWHM). The width (Δm) is taken of a singly charged mass spectral peak at a position corresponding to half of the peak height. The resolving power is determined by the ratio m/Δm. Higher resolving power may be required to measure ions that are similar in m/z. For peptides, the likelihood that ions generated from different peptides have similar m/z, and elute from LC separation at the same time is high. The resolving power of a mass spectrometer can dictate whether one or two ions are measured in these instances. This is important in quantitative approaches, which may require the measurement of two ions that differ by 1 Da due to the incorporation of a stable isotope (section 1.4.1).

Mass accuracy is a figure which defines the uncertainty between a measured m/z value and its theoretically calculated value. It is usually reported in parts per million (ppm), which is calculated from the difference between the measured and theoretical m/z. Higher mass accuracy may required when trying to identify a

peptide purely based on its precursor ion mass and not its structural composition. Accurate mass measurements can also introduce constraints in database searches.

Sensitivity, detection limit and dynamic range are terms used to help describe the ability of a mass spectrometer to make quantitative measurements. The ability to detect the lowest abundant ion (limit of detection, LOD) is usually better than the ability to quantify the lowest abundant ion (limit of quantification, LOQ). Dynamic range refers to the ratio between the lowest and the highest abundant ions detected. Ionisation processes, mass analysers and detectors may all have a characteristic dynamic range associated with them.

A characteristic of some mass spectrometers is that some properties may be compromised in order to achieve optimum performance in other areas. This is the case with FTICR, which currently provides the highest mass accuracy and mass resolution, but has a relatively slow analysis time and limited dynamic range. This makes it suitable for top-down proteomics but limits the application of the approach in bottom-up approaches. Table 1.1 provides a summary of the most commonly used mass spectrometers and their main contribution to proteomic studies.

The Orbitrap is a relatively newly introduced mass analyser that is increasingly being used in proteomics experiments. The Oribtrap mass analyser was invented by Makarov in 1999 (Makarov 2000) and its utility for proteomics was demonstrated in 2005 by Hu *et al.* (Hu *et al.* 2005). The Oribtrap uses orbital trapping of ions in its static electrostatic fields, in which the ions orbit around a central electrode and oscillate in axial direction. Both Oribtrap and ICR instruments use a fast Fourier Transform algorithm to convert time-domain signal into a frequency which can then be converted to a m/z scale. The LTQ-Orbitrap mass spectrometer combines the advantages of high resolution and mass accuracy obtained from the Orbitrap with the speed and sensitivity of the LTQ. Overall, the LTQ-Orbitrap offers mass accuracy comparable to FTICR at a lower purchase cost.

Table 1.1. Performance characteristics of commonly used mass spectrometers for proteomics

Instrument	Mass resolution	Mass accuracy (ppm)	Sensitivity	*m/z* range	Scan rate	Dynamic range	Fragmentation (n>2, up to 13)	Ionisation source	Main applications
Quadrupole ion trap (QIT)	1000	100-1000	Picomole	50-2000; 200-4000	Moderate	1×10^3	MS^n	ESI	Protein identification of low complex samples; PTM identification
Linear ion trap (LTQ)	2000	100-500	Femtomole	50-2000; 200-4000	Fast	1×10^4	MS^n	ESI	High throughput protein identification of complex mixtures by online LC-MS^n; PTM identification
Triple quadrupole (Q-q-Q)	1000	100-1000	Attomole to Femtomole	10-4000	Moderate	6×10^6	MS/MS	ESI	Quantification of selected peptides in single or multiple reaction monitoring mode; PTM detection
Quadrupole linear ion trap (Q-q-LIT)	2000	100-500	Femtomole	5-2800	Fast	4×10^6	MS^n	ESI	Quantification of selected peptides in single or multiple reaction monitoring mode; PTM detection
Time-of-flight (TOF)	10,000-20,000	10-20; <5	Femtomole	No upper limit	Fast	1×10^4	Post-source decay	MALDI	Protein identification from in-gel digestion of gel separated proteins by peptide mass fingerprinting

Table 1.1. continued.

Time-of-flight Time-of-flight (TOF-TOF)	10,000-20,000	10-20; <5	Femtomole	No upper limit	Fast	1×10^4	MS/MS	MALDI	Protein identification from in-gel digestion of gel separated proteins by peptide mass fingerprinting or sequence tagging via CID MS/MS
Quadrupole Time-of-flight (Q-q-TOF)	10,000-20,000	10-20; <5	Femtomole	No upper limit	Moderate to fast	1×10^4	MS/MS	ESI MALDI	Protein identification from complex peptide mixtures; Intact protein analysis; PTM identificaiton
Fourier transform ion cyclotron resonance (FTICR)	50,000-750,000	<2	Femtomole	50-2000 200-4000	Slow	1×10^3	MS^n	ESI MALDI	Top-down proteomics; High mass accuracy PTM characterisation
LTQ-Orbitrap	30,000-100,000	<5	Femtomole	50-2000 200-4000	Moderate to fast	4×10^3	MS^n	ESI MALDI	Top-down proteomics; High mass accuracy PTM characterisation; Protein identification from complex peptide mixtures; Quantification

Adapted from (Han *et al.* 2008)

1.4 Quantitative Proteomics

The overall objective of quantitative proteomics is to obtain information regarding the levels of protein expression, whether that may be in one condition or between a pair of differing conditions. This was initially carried out by comparing the intensity of staining on a 1D or 2D PAGE gel (Unlu *et al.* 1997). The use of gels for quantification is not, however, ideal since they do not easily allow the detection of low abundance proteins and issues may arise with overlapping bands/spots. It was not until 1999 that a commercially available method based on LC-MS was developed. Since then, numerous LC-MS based quantitative approaches have been developed. These approaches may be placed into three main categories; discovery, directed and targeted.

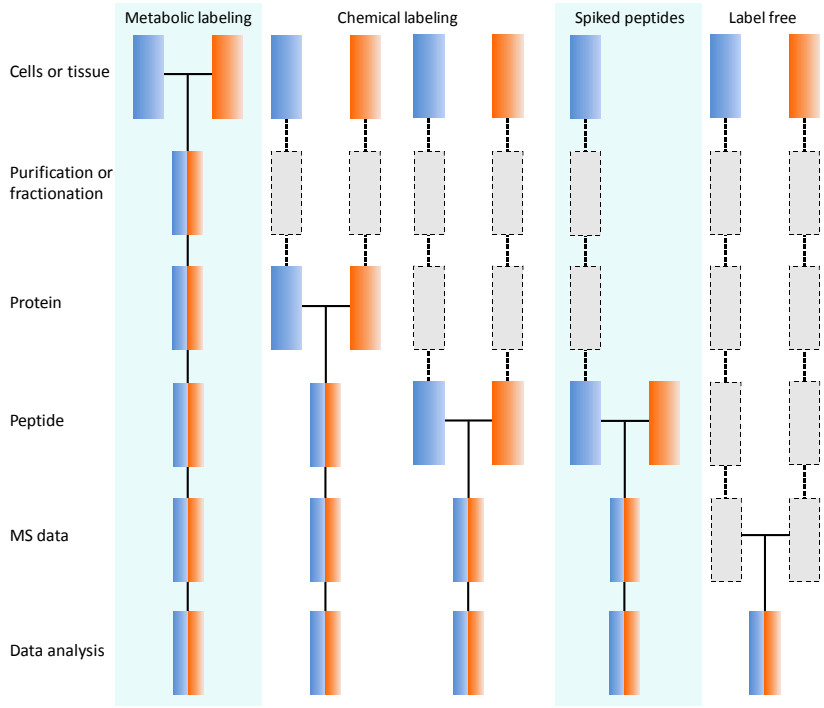

Figure 1.8. Experimental variation in common quantitative mass spectrometry workflows

Boxes in *blue* and *orange* represent two experimental conditions. *Horizontal lines* indicate when samples are combined. *Dashed lines* indicate points at which experimental variation and thus quantification errors can occur. Adapted from (Ong and Mann 2005).

Discovery-based approaches usually entail a single LC-MS/MS experiment with the aim being to identify and quantify as many proteins as possible. These experiments are normally employed for relatively uncharacterised systems. In a directed approach,

an LC-MS experiment is conducted for each sample of interest to obtain intensity information from all peptide precursor ions. The peptide intensities are compared to determine those that have been differentially expressed. A list of these differentially expressed peptides can then be included in a DDA experiment for selection and subsequent fragmentation in order to obtain sequence information. The targeted approach is not focussed on quantifying as many proteins as possible. It is a hypothesis-driven approach, which focuses on the detection and quantification of unique peptides from a selection of interesting proteins. This technique is especially useful in the validation of potential biomarkers.

ESI-MS is not inherently quantitative since proteolytic peptides may exhibit a wide range of physiochemical properties such as size, charge, hydrophobicity. This can lead to differences in mass spectrometric response. This can be overcome to an extent by the comparison of the interesting individual peptides between experiments. This can be achieved in a number of ways, shown in Figure 1.8. A major approach is based on stable isotope dilution theory, which is believed to be the most accurate way to determine relative changes in protein expression, since variation in subsequent experimental steps are minimised by combining samples at a very early stage. Label-free approaches are in theory more susceptible to experimental variation since samples are treated indendently at each stage and are only compared at the data analysis level.

1.4.1 Chemical and metabolic labelling techniques

Stable isotope labels may be chemically added to a sample of interest after extraction/purification from cells, tissue etc. (chemical labelling), or incorporated intrinsically during growth (metabolic labelling). These labelled samples are usually analysed at the peptide level, and are compared to non-labelled or differentially labelled samples in order to gain a relative measure of peptide abundance. A fundamental principle that is shared by all labelling techniques is that a labelled peptide is chemically identical to its native counterpart and therefore the two peptides should behave identically during chromatographic and/or mass spectrometric analysis. Given that MS can recognise the mass difference between a labelled and non-labelled form of a peptide, quantification is achieved by comparing the respective signal intensities. A summary of the most popular chemical and metabolic labelling

techniques is provided in Table 1.2. For each technique, it details where chemical labels are incorporated (peptide/protein), what functional group they are targeting (sulfhydral, amines, lysines, carboxyl, PTMs etc), how many different conditions may be compared and provides an indication of how widely the approach has been used. SILAC and N^{15} are the most popular metabolic labeling techniques available, and iTRAQ is currently the most widely used chemical labelling technique.

The factors that users must consider when choosing a quantitative approach is the nature of the sample, the hypothesis being evaluated, the cost, the effectiveness and the instrumentation available. Metabolic and chemical labeling techniques have been shown to be very effective, especially in the comparison of two different experimental states. These techniques, however, do require expensive isotope labels, specific software, and significant technical expertise to acquire and analyse data. Sample preparation and downstream analysis may be time consuming. This is the case for iTRAQ experiments, which can require 6 days for sample preparation and up to 60 hours instrumental time depending on the number of fractionation steps carried out (Patel *et al.* 2009). Metabolic labelling stategies cannot easily be applied to all types of samples. This can be the case for higher eukaryotic organisms, but has been achieved using immortilized cell lines. Some of these approaches can require a significant quantity of starting material, which may be challenging for the analysis of biofluids, which can be limited in supply. Some methods are specific and may not allow the quantification of some peptides. This is the case with the ICAT method, which only targets cysteine containing peptides for labelling and quantification. Another issue to be considered with all chemical labelling approaches is the possibility of incomplete labelling.

Table 1.2. Summary of metabolic and chemical labelling techniques

Technique	Chemical/ Metabolic	Target	Incorporation of label	Notes	No. of times cited	Refs.
DiGE	Chemical	Amines	Protein	**Difference gel electrophoresis**. Three protein states can be compared by labelling each sample set with one of three cyanine dyes. Protein samples are combined and separated on one 2D gel. Quantification is achieved with the use fluorescence imaging.	1009	(Unlu *et al.* 1997)
ICAT	Chemical	Cysteinyl residues	Protein	**Isotope-coded affinity tags**. Two different cell types are compared by labelling reduced proteins with either a light or heavy version of the tag. The samples are combined, enzymatically cleaved into peptides, and then only tagged peptides are purified for further analysis. Relative quantification is achieved by comparing the summation of the peak areas corresponding to co-eluting tagged peptides in LC-MS/MS. Only cysteine containing peptides can be used for quantification.	3113	(Gygi *et al.* 1999a) (Hansen *et al.* 2003) (Zhou *et al.* 2002)
ICPL	Chemical	Amines	Protein	**Isotope-coded protein label**. Three different cell states can be quantified by labelling proteins with nicotinoylation reagents (with varying stable isotopes). Labelled samples are combined and resolved on a 2D gel, then enzymatically digested and analyzed by MALDI-TOF/TOF. Heavy and light peptides ions appear as doublets/triplets in the mass spectrum. Relative quantification is achieved by calculating the ratio of the labeled peptide ion intensities.	216	(Schmidt *et al.* 2005)
iTRAQ	Chemical	Amines	Peptide	**Isobaric tags for relative and absolute quantitation.** Up to 8 proteome samples can be enzymatically cleaved into peptides, which are then labelled and pooled together. Peptides are separated by SCX and analysed by ESI-LC-MS/MS. Relative quantification is achieved from the fragmentation of each co-eluting tagged peptide and comparing the specific reporter ion intensities.	1216	(Ross *et al.* 2004)

Table 1.2. continued.

Dimethylation	Chemical	Amines	Peptide	**Stable-isotope dimethyl labeling.** Formaldehyde is used to label amine residues via reductive amination. The ionic state and the physiochemical properties of the ions are not affected like other amine-targeted techniques. Quantification is achieved by LC-MS and identification by LC-MS/MS.	137	(Hsu *et al.* 2003)
MCAT	Chemical	Lysines	Peptide	**Mass-coded abundance tagging.** Tryptic peptides from one sample state are treated with *O*-methylisourea (+42 mass units). This is combined with untreated tryptic peptides of a different sample state. This technique can be modified to improve *de novo* sequencing.	168	(Cagney and Emili 2002)
Methyl esterification	Chemical	Carboxyl	Peptide	**Methyl esterification.** Tryptic peptides from two sample states are esterified with d3- or d0-methanol. The two samples are mixed and analysed by LC-MS/MS.	201	(Goodlett *et al.* 2001)
^{15}N labelling	Metabolic	Nitrogen	Protein	**^{15}N labelling.** ^{15}N-ammonium salts can be used to incorporate ^{15}N into all amino acids in simple organisms like bacteria and yeast, enabling the quantification of all peptides. Labelled simple organisms have been fed to complex organisms such as rats, these experiments, however, incur high costs. The mass difference between labelled and non-labelled peptides is sequence dependent, which complicates spectra interpretation.	632	(Oda *et al.* 1999) (Conrads *et al.* 2001)
SILAC	Metabolic	1 – 2 essential amino acids	Protein	**Stable isotope labeling by amino acids in cell culture.** An essential amino acid containing stable isotopes is supplied in growth media and is introduced to newly synthesized polypeptides in a sequence-specific fashion	493	(Ong *et al.* 2002)
CDIT	Metabolic	1 – 2 essential amino acids	Protein	**Culture-derived isotope tags.** SILAC labelled cells are used to create internal standards for the comparative analysis between two tissue samples.	95	(Ishihama *et al.* 2005)

1.4.2 Label-free techniques

Label-free approaches do not require that the sample to be analysed is modified. The methods currently available may be divided into two distinct groups, shown in Figure 1.9:

(i) Spectral counting, which is based on counting the number of fragment ion spectra identifying peptides of a given protein

(ii) Area under the curve (AUC), extracted ion current (XIC), or signal intensity measurement of peptide precursor ions belonging to a particular protein

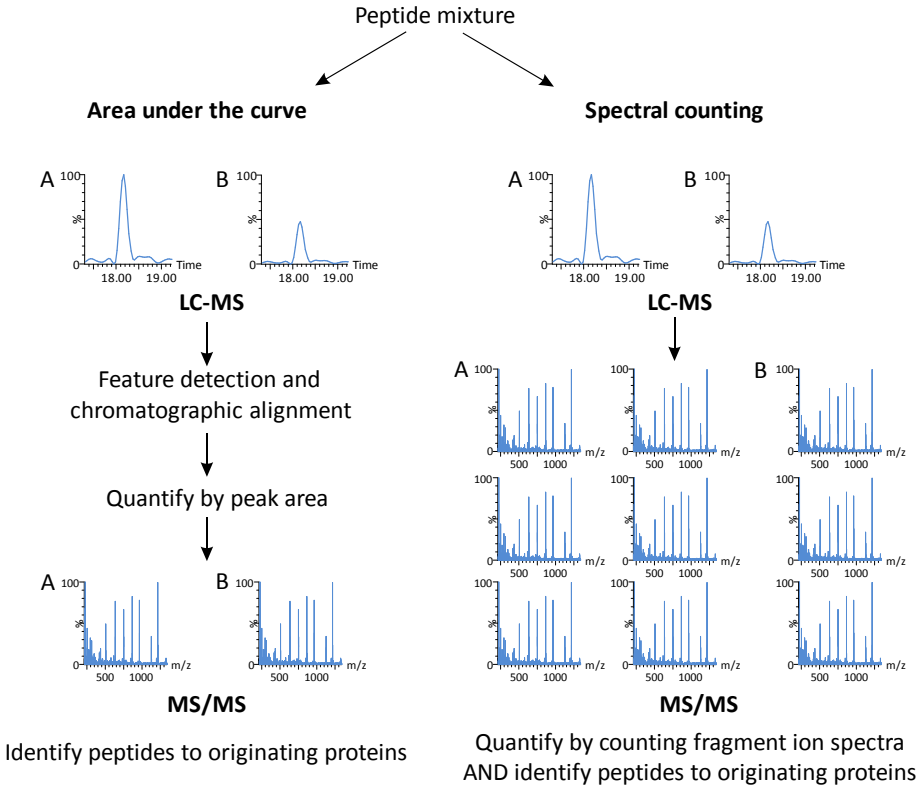

Figure 1.9. Schematic of two main label-free approaches

Theoretical samples A and B and present in a ratio of 2:1. AUC quantification is performed at the MS level. Those peptides exhibiting differential expression are identified by MS/MS either in a simultaneous or subsequent fragmentation step. Identification and quantification is achieved simultaneously at the MS/MS level.

Spectral counting

Spectral counting is based on the observation that more abundant peptides will be preferentially selected for fragmentation and will produce more abundant MS/MS spectra during a DDA experiment. An estimate of protein abundance in a sample may be calculated using the protein abundance index (PAI), which may be defined as the number of observed peptides divided by the number of observable peptides per

protein. The relationship between the number of peptides observed and the protein amount within a given sample has shown to be logarithmic, leading to the concept of an exponentially modified emPAI. This method of quantification does not need additional expensive software since it is available with publically available MASCOT database searching. The method has been shown to provide a rough estimate of protein abundance within a factor of 3 – 5 of the true value (Ong and Mann 2005).

Area under the curve (AUC) or signal intensity measurement

In these approaches, a measure of abundance is achieved by extracting an ion chromatogram for every peptide observed in an LC-MS/MS run and integrating the peak areas over the chromatographic time scale. The intensity value for each peptide in an experiment may then be compared to signals in one or more other experiments to yield relative quantification. There are three important parameters than affect the analytical accuracy of quantification using ion intensities. The first is the accuracy of mass measurement. A mass spectrometer operating at high mass accuracy would minimise the influence of interfering signals of similar but distinct masses. A second parameter is LC reproducibility. The chromatographic profile of a peptide should be similar between different experiments to aid in the identification of the correct corresponding peptide. The final parameter is the experimental balance in DDA experiments between peptide mass and fragment mass information. While the mass spectrometer is obtaining fragment mass information, peptide mass information is being lost. This can result in a loss of signal for a peptide ion, affecting its ion chromatogram. Limitations in AUC based quantification may be minimised by normalisation using spiked in calibrants or by using abundant non-changing peptides as landmarks.

AUC quantification relies heavily on the use of appropriate software. There are a number of platforms available that are commercially and freely available, and are either vendor-specific or vendor-independent. These include SIEVE (Thermo), Elucidator, Progenesis (Nonlinear), ProteinLynx Global Server (Waters) and MSQuant.

1.4.3 Data-independent acquisition (MS^E) for label-free absolute and relative quantification

A potential problem with AUC-based quantification approaches carried out when the mass spectrometer is operating in DDA mode is the balance between the use of survey (MS) and fragment scans (MS/MS) in order to obtain quantification data whilst maximising protein identification. This can be overcome, in part, by performing multiple analyses separately in MS and MS/MS modes and then by matching accurate mass and retention times (AMRT) to achieve peptide identifications (Silva *et al.* 2005). Obtaining quantification data from these label-free types of analyses relies heavily on software analysis to efficiently cluster detected peptides. Silva *et al.* describe Expression Informatics software for the analysis of AMRTs, which include peak detection, chromatographic spectral alignment, charge-state reduction and deisotoping for monitoring relative changes in protein abundances between two conditions.

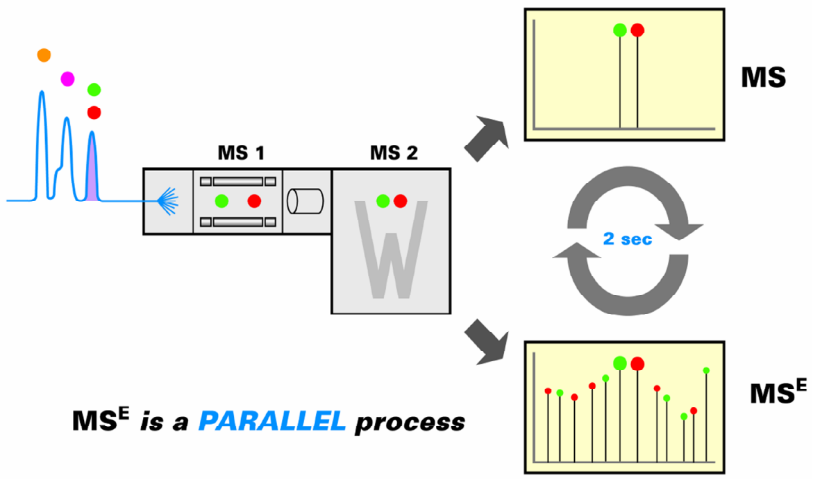

Figure 1.10. Schematic of data-independent acquisition

The method of obtaining MS and MS/MS in the same chromatographic run without selecting the selection of precursor ions has been patented and termed MS^E (Bateman *et al.* 2002). In this mode of acquisition, the quadrupole acts as an ion guide, transmitting all precursor ions in one scan and all associated product ions in alternate scans, where the collision energy is set at a low and elevated level, respectively (Figure 1.10). This methodology aims to overcome the issues associated with mass spectrometry duty cycle and the accuracy of quantification when operating in DDA mode. The duty cycle may be defined as the percentage of time a mass analyzer

spends transmitting those ions of interest during a single experimental cycle. During MS^E, data points are collected during the elution of a peptide to obtain a profile of peptide abundance. In a direct comparison of samples analysed by a Q-TOF instrument operating in either DIA (MS^E) or DDA mode, integrated intensities were compared for the same peptides measured by each method. Little correlation was observed between the two methods with the DDA method producing a greater measurement error, suggesting that the precursor intensities derived by DDA were not appropriate for use as measure of protein quantity (Geromanos *et al.* 2009).

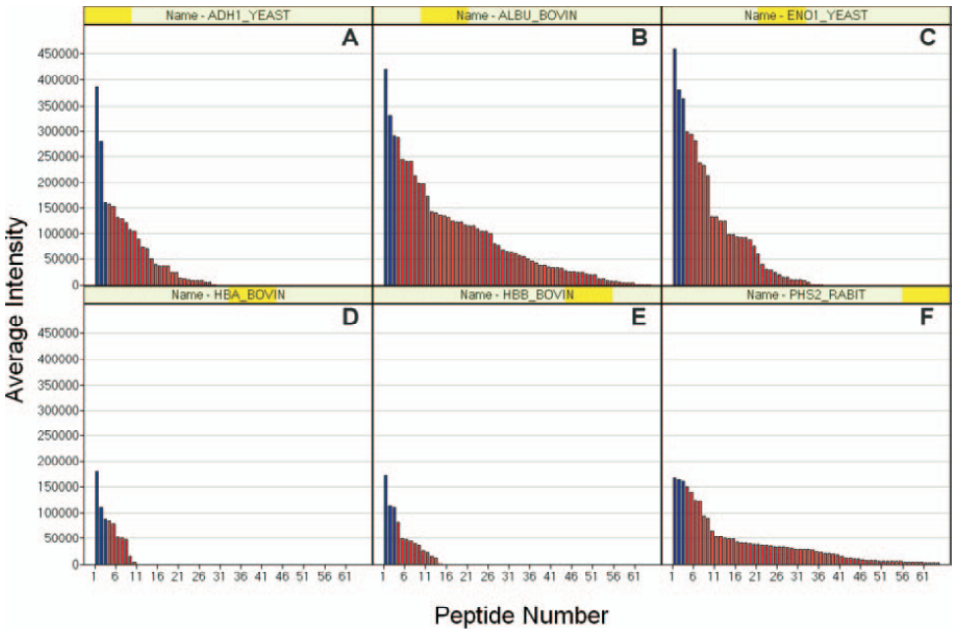

Figure 1.11. Peptide intensities measured by LC-MS^E organized by decreasing intensity for each protein

The bar plots illustrate the peptides identified to 10 pmol alcohol dehydrogenase (A), 12.5 pmol serum albumin (B), 15 pmol enolase (C), 10 pmol bovine hemoglobin α (D), 10 pmol bovine haemoglobin β (E), and 6 pmol phosphorylase B(F). Average intensities were obtained from three technical replicates. Taken from (Silva *et al.* 2006).

Relative quantification methods have dominated the quantitative proteomics field. In 2005, a method for label-free quantification that took advantage of the parallel nature of MS^E approach was described (Silva *et al.* 2006). This method was the first of its kind to provide absolute quantities of the majority of identified peptides rather than just detecting relative changes and then seeking out identity of those differentially expressed peptides. It is known that different peptides at equimolar concentrations do not necessarily have the same response when studied using MS. This is shown in

Figure 1.11, taken from the original MS^E paper which demonstrated the proof-of-principle of the technique. The authors demonstrate that the sum of the intensity of the top three most abundant peptides observed for a given protein could be related to the absolute concentration of the protein (Figure 1.12). These results demonstrated that the signal response curve measured for the three most intense tryptic peptides is similar for all proteins. Because the response curve is independent of the protein, it may be used as a universal method to obtain an absolute quantity of any other well characterised protein present in the mixture.

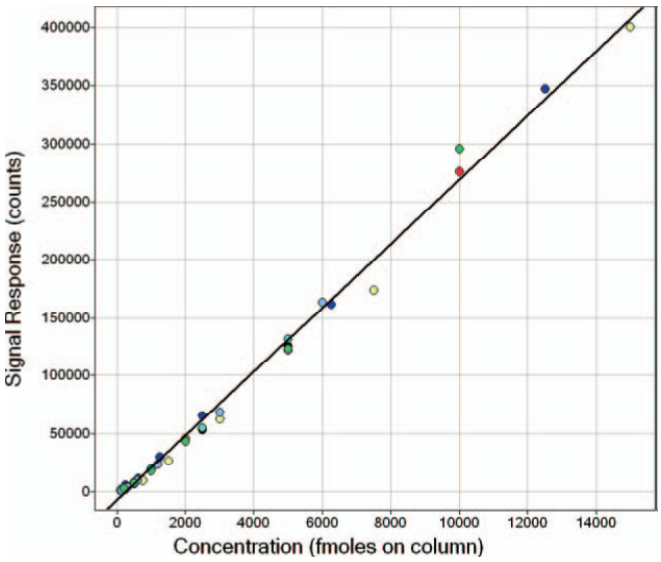

Figure 1.12. Relationship between top three peptide signal response and absolute quantity

Data are coloured to represent different proteins. Taken from (Silva *et al.* 2006).

A label-free MS^E experiment involves taking a tryptic digested sample and spiking it with a protein digest of known concentration (internal standard). The signal intensity measured from the top three most intense peptides for the internal standard is then used to create a response factor (signal intensity/fmol), which is then applied to the top three most intense peptides that are assigned to proteins in the sample. This provides an estimation of protein abundance, and is referred to as the Hi3 approach. As an example, six proteins of known concentration were mixed with a complex mixture of human serum proteins and measured using the label-free approach. The results showed that the technique was able to quantify the spiked proteins with a relative error of between -13.3 to 8.7 %.

A major disadvantage of this approach is that it is currently limited to instruments developed by Waters. Despite this, numerous peer-reviewed publications have been published since 2005 (Waters). In this research, the recently developed techniques of travelling wave ion mobility-mass spectrometry (a mass spectrometer capable of separating ions based on shape and m/z) and online two dimensional reversed phase-reversed phase (RP-RP) chromatography have been used together with the label-free MS^E technology for comparative and profiling proteomic experiments. There are only a few studies which include the combination of MS^E with 2D-LC or ion mobility, and none that provide comprehensive analytical, technical and experimental insight into the potential impact of the approach in proteomics.

1.5 Methanotrophy

Methylotrophs may be defined as microorganisms able to grow at the expense of reduced carbon compounds containing one or more carbon atoms but not containing carbon-carbon bonds (Anthony 1982). Methanotrophs, or methanotrophic bacteria, are a unique subset of the methylotrophy group that can use methane (and sometimes intermediates in the oxidation of methane) as a sole carbon and energy source (Whittenbury and Dalton 1981). The first methanotroph, *Bacillus methanicum*, was discovered in 1906 (Sohngen 1906), but it was not until much later that over 100 new methane-utilizing bacteria were characterised into 5 genera (Whittenbury *et al.* 1970). To date, methane-utilising bacteria have been classified into 14 recognized genera (Jiang *et al.* 2010). Methane oxidising bacteria have historically been grouped into type I and type II methanotrophs based upon their carbon assimilation pathways, namely ribulose monophosphate (RuMP) and the serine cycle, as well as other characteristics including G+C content, nitrogen fixation and membrane arrangements. A new group, Type X, was added to include methanotrophs that possessed enzymes of both carbon assimilation pathways, amongst other distinctive characteristics.

1.5.1 Roles and potential applications of methanotrophs

Greenhouse gases from natural sources and human activity can result in climate change in the long term by affecting the atmosphere chemically. Carbon dioxide and methane are thought to be the two main contributors to the greenhouse effect. At present, methane is less abundant, but on a molar basis it is approximately 23 times more effective as a greenhouse gas than CO_2. Unfortunately, recent reports indicate that, after a decade with little change in atmospheric methane, global atmospheric methane showed renewed growth in 2007. It has been estimated that 70% of global methane emissions are anthropogenic (Jiang *et al.* 2010).

Major anthropogenic sources of methane emissions include landfills, oil and gas recovery, coal mines and agriculture (Hanson and Hanson 1996). A number of biological systems have been developed to mitigate methane emissions in landfill, including biocovers, vented biofilters, biowindows and daily-use biotarps (Huber-Humer *et al.* 2008). There are many factors that need to be taken into consideration for these technologies to be widely used, such as the collection of gas, optimisation of

microbial growth conditions and the response of methanotrophs to different methane concentrations. The engineering and optimization of methanotrophs involved in bioprocesses for the control of landfill methane emission is therefore, of great importance. The use of methanotrophs to remove methane from coal mine atmospheres to reduce greenhouse gas emission and also prevent severe explosions has been investigated (Sly *et al.* 1993). The development of bioprocesses for the removal of coal mine gas is limited by the need for methanotrophs with high methane-oxidising capability coupled with operational difficulties. Continued progress in research on methanotrophs may enable the use of methanotrophs to alleviate coal mine gases and prevent dangerous explosions.

The unique methane-oxidising enzymes that methanotrophs possess are also capable of co-metabolising a wide range of substrates, including highly toxic and xenobiotic organic compounds. In particular, a mixed community of methanotrophs has been shown to co-metabolise methane and halogenated hydrocarbons in soil. These are major contaminants in soil and water from industrial processes.

1.5.2 Facultative methanotrophy – a controversial topic

Obligate methanotrophs can be classified by the ability to grow at the expense of methane and reduced carbon compounds containing one or more carbon atoms but containing no carbon-carbon bonds. Facultative methanotrophs are additionally capable of growing at the expense of a variety of other multi-carbon compounds.

The existence of facultative methanotrophs has been a controversial topic in the field of one-carbon metabolism, dating back almost 40 years. An early example of this controversy comes from the study of *Methylobacterium*. This genus is composed of facultative methylotrophs, and so, by definition, cannot utilise methane as a carbon and energy source. One of the species belong to this genus, *Methylobacterium organophilium*, was discovered to grow at the expense of methane and so was classified as a facultative methanotroph (Patt *et al.* 1974). The strain *M. organophilium* XX, was uniquely found to possess the methane-utilising ability, however, this was not independently verified from cultures that were stored in national repositories (Green and Bousfield 1983). In most of the cases where the existence of a facultative methanotroph was claimed, it was suggested that the

capacity for methane metabolism was lost (from the loss of a plasmid). There were also difficulties in the replication of the results by other laboratories.

In 2004, another species of the *Methylobacterium* genus was claimed to be able to use methane as the sole source of carbon and energy (Van Aken *et al.* 2004). This claim was rebuked by experts in the field for not providing enough evidence to confirm the capability of methane-utilisation (Dedysh *et al.* 2004b).

1.5.3 The *Methylocella* genus is truly facultative

In 2005, Dedysh *et al.* provided the first unequivocal proof of the existence of facultative methanotrophs belonging to the genus *Methylocella* (Dedysh *et al.* 2005). The species *Methylocella palustris* (Dedysh *et al.* 1998; Dedysh *et al.* 2000), *Methylocella silvestris* (Dunfield *et al.* 2003) and *Methylocella tundrae* (Dedysh *et al.* 2004a) were isolated from acidic peat, forest, and tundra soils, respectively. The *Methylocella* species are phylogenetically closely related to a non-methanotrophic heterotroph, *Beijerinckia indica*. The *Methylocella* species possess several features that distinguish them from other methanotrophs. Studies have shown the existence of the soluble form of the methane monooxygenase enzyme but not the particulate form (Dedysh *et al.* 2000; Theisen *et al.* 2005), the absence of which is shared only by a subset of other methanotrophs. The species also differ morphologically, lacking a well developed intracytoplasmic membrane to which particulate methane monooxygenase binds in other methanotrophs, and having a series of membrane-bound vesicles adjacent to the inner cell membrane instead. These species form a distinct taxonomic cluster of acidophilic, methanotrophic bacteria and belong to *Alphaproteobacteria* (type II methanotrophs). All three species were found to be able to utilise multi-carbon compounds acetate, pyruvate, succinate, malate and ethanol as a sole carbon and energy source. They contain enzymes for one-carbon assimilation via the serine pathway. Meticulous experiments were carried out on *M. silvestris* BL2 to verify facultative growth, which has finally dispelled any doubts that have surrounded the existence of such an organism (Dedysh *et al.* 2005). Further studies on this genus were continued using this species, since it grows most robustly.

1.5.4 Biochemistry of methanotrophs

Figure 1.13 shows the common features of methane metabolism in methanotrophs, including the central role of formaldehyde and the pathways used to assimilate carbon via intermediates. The use of enzymes known as methane monooxygenases (MMOs) to catalyse the oxidation of methane to methanol is a defining characteristic of methanotrophs. Two forms of MMOs have been found in methanotrophic bacteria. One form is a soluble MMO (sMMO), which consists of three main components; a hydroxylase ($M_r \sim$ 251 kDa), a reductase (38.6 kDa), and a coupling protein (15.5 kDa). The hydroxylase is comprised of three different subunits (α, β, and γ) of $\alpha_2\beta_2\gamma_2$ stoichiometry.

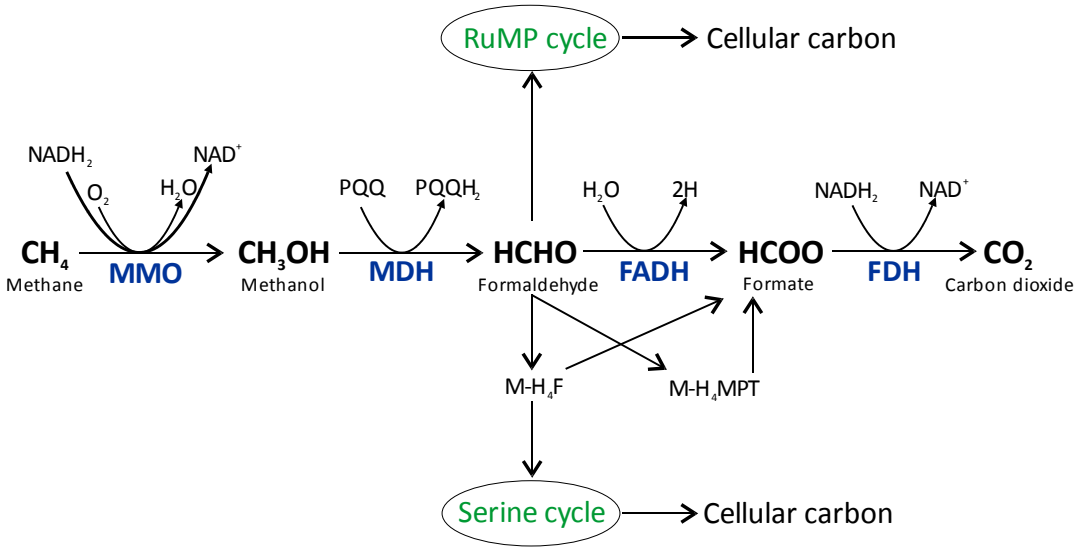

Figure 1.13. Pathways of methane oxidation and formaldehyde assimilation

MMO, methane monooxygenase; MDH, methanol dehydrogenase; FADH, formaldehyde dehydrogenase; FAD, formate dehydrogenase; H_4F, tetrahydrofolate; H_4MPT, tetrahydromethanopterin.

The majority of methanotrophs are capable of forming particulate or membrane-bound MMO (pMMO), and a minority also the sMMO. pMMO is only active in intracytoplasmic membranes when copper concentrations exceed a certain level. Below this concentration, sMMO activity is predominant in the methanotrophs that contain both enzymes.

Methanol, the product of methane oxidation via MMO, is oxidised to formaldehyde by methanol dehydrogenase. Methanol dehydrogenase is a tetramer consisting of large and small subunits in a $\alpha_2\beta_2$ configuration and requires the co-factor pyrroloquinoline quinone (PQQ). Methanol dehydrogenase and two specific forms of

cytochrome *c* are required for activity. These proteins are soluble, located in the periplasm of gram-negative methylotrophs and are present in high concentrations.

The formaldehyde produced is either assimilated into cell biomass or is further oxidised to CO_2 to generate reducing power for biosynthesis and methane hydroxylation. Methanotrophs are capable of utilising the serine and/or the ribulose monophosphate (RuMP) pathways for carbon assimilation. Some methanotrophs require the reaction of formaldehyde with co-factors tetrahydrofolate (H_4F) and tetrahydromethanpterin (H_4MPT), generating methylene-folate derivatives. These are assimilated subsequently into cellular biomass via the serine cycle (H_4F-dependent pathway) or oxidised to carbon dioxide to generate energy (H_4MPT-dependent pathway). Formaldehyde can be oxidised by formaldehyde dehydrogenase to formate, and finally to CO_2 by formate dehydrogenase.

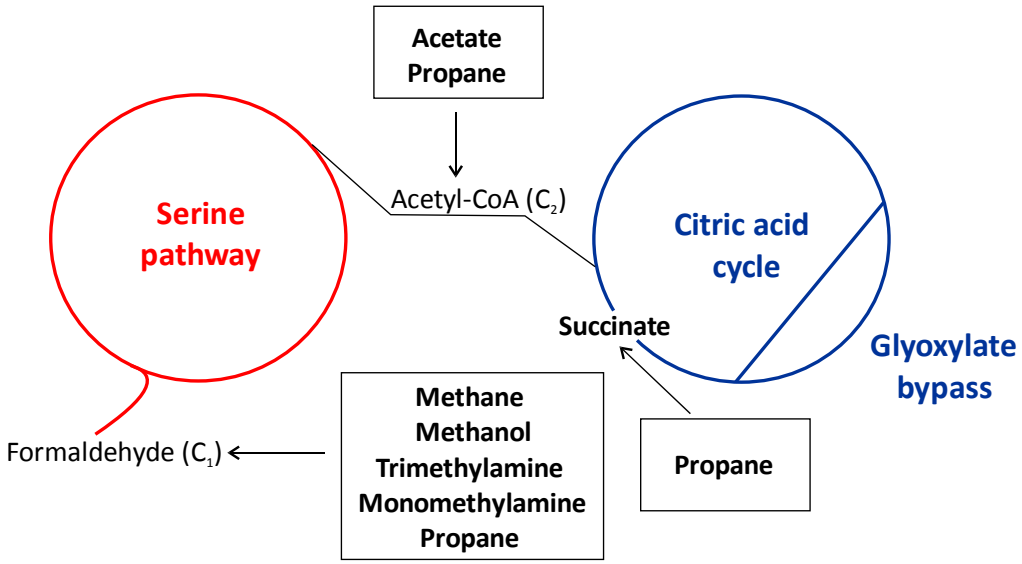

Figure 1.14. Central pathways and possible entry points from substrates used in this study

Substrates in *bold* were used as growth substrates in this study.

Figure 1.14 shows a simplistic view of the overall reactions that can occur on growth with C_1 and C_2 compounds, and highlights the growth substrates that have been used in this research. Assimilation of all carbon substrates when provided as the sole source of carbon requires their conversion to the intermediates of central metabolism containing 3 or 4 carbon atoms, which then provide precursors for biosynthesis. Methanotrophs growing on C_1 compounds must be able to create every carbon-carbon

bond. For this, *M. silvestris* uses the serine pathway which ultimately generates the C_2 compound acetyl-CoA. The citric acid cycle, or the tricarboxylic acid cycle (TCA) have two potential functions. The first is to catalyse the oxidation of acetyl-CoA, providing NADH and then ATP for biosynthesis. The second is to provide carbon precursors for biosynthesis. For this function, oxaloacetate must be replenished either by carboxylation of a C_3 compound or by the glyoxylate bypass, which was first described in 1957. This pathway, however, does not operate in some methylotrophs whilst growing on C_2 compounds that lack the key glyoxylate enzyme isocitrate lyase. After 35 years of research, the ethylmalonyl-CoA (EMC) pathway for growth on C_2 compounds has recently been elucidated in photosynthetic bacteria and has been shown to operate in methylotrophs growing on C_2 compounds, and on C_1 compounds by way of the serine cycle (Anthony 2011).

In this research, *M. silvestris* was grown with compounds containing no carbon-carbon bonds (methane, methanol, methylamine and trimethylamine) and with multi-carbon containing compounds (propane, acetate and succinate). Further background describing the utilisation of these compounds can be found in Chapters 5 and 6.

1.6 Project Aims

The work presented here combines the investigation of two novel aspects; the study of metabolic pathways in the first truly authenticated facultative methanotroph and the use of recently developed mass spectrometry-based approaches for profiling and comparative proteomics studies using the LC-MS^E technology.

A two dimensional liquid chromatography technique was recently developed which utilises two reversed phases operated using differing pH. This system may be directly coupled to a mass spectrometer. This system was made commercially available in early 2009. Ion mobility mass spectrometry-based proteomic experiments have predominantly been carried out on in-house built instruments in specialised laboratories. The Synapt HDMS G2 system, introduced in December 2009, was the first commercially-available ion mobility mass spectrometer with the ability to carry out mobility-assisted data-independent acquisition. The combination of both multidimensional LC and traveling wave ion mobility separation forms a relatively

new approach which has been made feasible due to advancements in LC column chemistry and mass spectrometry instrumentation. This research is presented in chronological order to focus on the development of new technologies and their individual impact on the label-free MS^E technology. An assessment of a number of approaches was conducted and applied to the analysis of *M. silvestris* BL2 protein extracts. The main aims of the research were:

- To evaluate the use of the recently developed technique online two-dimensional RP-RP LC for profiling and comparative proteomics in a label-free MS^E quantitative workflow (Chapter 3).

- To evaluate the use of the traveling wave ion mobility separation as an improved method for profiling and comparative proteomics in a label-free MS^E quantitative workflow (Chapter 4).

- To characterise the protein component extracted from *M. silvestris* BL2 grown at the expense of monomethylamine, trimethylamine and methanol to identify and quantify proteins involved in methylated amine metabolism (Chapter 5).

- To characterise the protein component extracted from *M. silvestris* BL2 grown at the expense of propane, succinate, acetate and methane to identify and quantify proteins involved in multi-carbon metabolism (Chapter 6).

- To quantify processes downstream of primary oxidation pathways and determine core proteins by mapping identifications to known biochemical pathways (Chapter 5 and Chapter 6).

1.7 Research and Conference Papers

The research presented in this thesis has to date resulted in two peer-reviewed publications:

Yin Chen., **Nisha A. Patel**, Andrew Crombie, James H. Scrivens and J. Colin Murrell. (2011). Bacterial flavin-containing monooxygenase is trimethylamine monooxygenase. *Proceedings of the National Academy of Sciences.* **108**, 17791-17796.

Nisha A. Patel, Andrew Crombie., Susan E. Slade, Konstantinos Thalassinos, Chris Hughes., Joanne B. Connolly, James Langridge, J. Colin Murrell and James H. Scrivens (2012). Comparison of one- and two-dimensional liquid chromatography approaches in the label-free quantitative analysis of *Methylocella silvestris*. *Journal of Proteome Research*. **11**, 4755-4763

Results of this research have been presented at the following conferences:

Konstantinos Thalassinos, **Nisha A. Patel**, Susan E. Slade, Vibhuti Patel, Andrew Crombie, J. Colin Murrell, Chris Hughes, Joanne B. Connolly, James Langridge, James H. Scrivens. Probing changes in biochemical pathways in environmentally important methanotrophs. *Proc. 57th ASMS Conf. on Mass Spectrometry and Allied topics, 2009, Philadelphia, USA*

Nisha A. Patel, Konstantinos Thalassinos, Susan E. Slade, Andrew Crombie, J. Colin Murrell, James H. Scrivens. A proteomic study into the facultative metabolism of methanotrophs. *RSC Analytical Research Forum, 2009, Kent, UK*

Nisha A. Patel, Konstantinos Thalassinos, Susan E. Slade, Andrew Crombie, J. Colin Murrell, James H. Scrivens. A proteomic study into the facultative metabolism of methanotrophs. *Proc. 18th International Mass Spectrometry Conference, 2009, Breman, Germany*

Nisha A. Patel, Konstantinos Thalassinos, Yin Chen, Susan E. Slade, Andrew Crombie, J. Colin Murrell, James H. Scrivens, Label-Free Quantitative Proteomic

Studies by 1D- and 2D-RP UPLC on an Environmentally Important Methanotroph. *LifeScience ProteoMMX Strictly quantitative 1st Conference, 2010, Chester, UK*

Nisha A. Patel, Konstantinos Thalassinos, Yin Chen, Susan E. Slade, Andrew Crombie, J. Colin Murrell, James H. Scrivens. Label-free quantitative proteomic studies on an environmentally important methanotroph. *Proc. 58th ASMS Conf. on Mass Spectrometry and Allied topics, 2010, Salt Lake City, USA*

Konstantinos Thalassinos, **Nisha A. Patel**, Susan E. Slade, James Langridge, Chris Hughes, Scott Geromonos, Marc V. Gorenstein, Dan Golick, Andrew Crombie, J. Colin Murrell, James H. Scrivens. Characterisation of a bacterial proteome by means of ion mobility mass spectrometry-based proteomics. *Proc. 58th ASMS Conf. on Mass Spectrometry and Allied topics, 2010, Salt Lake City, USA*

James H. Scrivens, **Nisha A. Patel**, Frances D.L. Kondrat, Jonathon R. Snelling, Susan E. Slade, Konstantinos Thalassinos, Charlotte A. Scarff. Shape selective studies of complex systems utilizing travelling wave ion mobility mass spectrometry. *Proc. 31st Annual meeting of the British Mass Spectrometry Society, 2010, Cardiff, UK*

Krisztina Radi, Charlotte A. Scarff, Susan E. Slade, **Nisha A. Patel**, James H. Scrivens. The development of rapid diagnostic approaches for the characterisation of hemoglobin disorders. *Proc. 59th ASMS Conf. on Mass Spectrometry and Allied topics, 2011, Denver, USA*

James H. Scrivens, Baharak Vafadar-Isfahani, **Nisha A. Patel**, Susan E. Slade, Lee Gethings, Chris Hughes, James Langridge. The use of an on-line two-dimensional (RP/RP) liquid chromatography mobility-enabled approach for the characterisation of the cellular proteomes. *Proc. 33rd Annual meeting of the British Mass Spectrometry Society, 2012, AstraZeneca, Alderley Park, UK*

Baharak Vafadar-Isfahani, **Nisha A. Patel**, Susan E. Slade, Lee A. Gethings, Chris Hughs, Jim Langridge, Andrew Crombie, J. Colin Murrell, James H. Scrivens. The use of an on-line two-dimensional (RP/RP) liquid chromatography mobility-enabled approach for the characterisation of the proteome of bacterial extracts. *Proc. 60th ASMS Conf. on Mass Spectrometry and Allied topics, 2012, Vancouver, Canada*

1.8 References

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature.* **422**, 198-207.

Anthony, C. (1982). The Biochemistry of Methylotrophs. London, Academic Press Inc.

Anthony, C. (2011). How half a century of research was required to understand bacterial growth on C1 and C2 compounds; the story of the serine cycle and the ethylmalonyl-CoA pathway. *Science Progress.* **94**, 109-137.

Armirotti, A. and Damonte, G. (2010). Achievements and perspectives of top-down proteomics. *Proteomics.* **10**, 3566-3576.

Bateman, R., Carruthers, R., Hoyes, J., Jones, C., Langridge, J., Millar, A. and Vissers, J. (2002). A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *J. Am. Soc. Mass Spectrom.* **13**, 792-803.

Beranova-Giorgianni, S. (2003). Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. *TrAC, Trends Anal. Chem.* **22**, 273-281.

Biemann, K. (1992). Mass Spectrometry of Peptides and Proteins. *Annu. Rev. Biochem.* **61**, 977-1010.

Cagney, G. and Emili, A. (2002). *De novo* peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat Biotech.* **20**, 163-170.

Carvalho, P. C., Han, X., Xu, T., Cociorva, D., Carvalho, M. d. G., Barbosa, V. C. and Yates, J. R. (2010). XDIA: improving on the label-free data-independent analysis. *Bioinformatics.* **26**, 847-848.

Causier, B. (2004). Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom. Rev.* **23**, 350-367.

Cech, N. B. and Enke, C. G. (2001). Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **20**, 362-387.

Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Paša-Tolić, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D. and Smith, R. D. (2001). Quantitative Analysis of Bacterial and Mammalian Proteomes Using a Combination of Cysteine Affinity Tags and 15N-Metabolic Labeling. *Anal. Chem.* **73**, 2132-2139.

Cravatt, B. F., Simon, G. M. and Yates III, J. R. (2007). The biological impact of mass-spectrometry-based proteomics. *Nature.* **450**, 991-1000.

Cruz, J., Liu, Y., Liang, Y., Zhou, Y., Wilson, M., Dennis, J. J., Stothard, P., Van Domselaar, G. and Wishart, D. S. (2011). BacMap: an up-to-date electronic atlas of annotated bacterial genomes. *Nucleic Acids Res.* **40**, D599-D604.

de Hoffmann, E. and Stroobant, V. (2002). Mass Spectrometry Principles and Applications. Chichester, John Wiley and Sons, Ltd.

Dedysh, S. N., Berestovskaya, Y. Y., Vasylieva, L. V., Belova, S. E., Khmelenina, V. N., Suzina, N. E., Trotsenko, Y. A., Liesack, W. and Zavarzin, G. A. (2004a). *Methylocella tundrae* sp. nov., a novel methanotrophic bacterium from acidic tundra peatlands. *Int. J. Syst. Evol. Microbiol.* **54**, 151-156.

Dedysh, S. N., Dunfield, P. F. and Trotsenko, Y. A. (2004b). Methane utilization by *Methylobacterium* species: new evidence but still no proof for an old controversy. *Int. J. Syst. Evol. Microbiol.* **54**, 1919-1920.

Dedysh, S. N., Knief, C. and Dunfield, P. F. (2005). *Methylocella* species are facultatively methanotrophic. *J. Bacteriol.* **187**, 4665-4670.

Dedysh, S. N., Liesack, W., Khmelenina, V. N., Suzina, N. E., Trotsenko, Y. A., Semrau, J. D., Bares, A. M., Panikov, N. S. and Tiedje, J. M. (2000). *Methylocella palustris* gen. nov., sp. nov., a new methane-oxidizing acidophilic bacterium from peat bogs, representing a novel subtype of serine-pathway methanotrophs. *Int. J. Syst. Evol. Microbiol.* **50**, 955-69.

Dedysh, S. N., Panikov, N. S., Liesack, W., Grosskopf, R., Zhou, J. and Tiedje, J. M. (1998). Isolation of Acidophilic Methane-Oxidizing Bacteria from Northern Peat Wetlands. *Science.* **282**, 281-284.

Dole, M., Mack, L. L., Hines, R. L., Mobley, R. C., Ferguson, L. D. and Alice, M. B. (1968). Molecular Beams of Macroions. *J. Chem. Phys.* **49**, 2240-2249.

Dunfield, P. F., Khmelenina, V. N., Suzina, N. E., Trotsenko, Y. A. and Dedysh, S. N. (2003). *Methylocella silvestris* sp. nov., a novel methanotroph isolated from an acidic forest cambisol. *Int. J. Syst. Evol. Microbiol.* **53**, 1231-1239.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. and Whitehouse, C. M. (1989). Electrospray ionization for mass-spectrometry of large biomoecules. *Science.* **246**, 64-71.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* **269**, 496-512.

Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. and Salzberg, S. L. (2002). The Value of Complete Microbial Genome Sequencing (You Get What You Pay For). *J. Bacteriol.* **184**, 6403-6405.

Gao, J. and Chen, L.-L. (2010). Theoretical methods for identifying important functional genes in bacterial genomes. *Res. Microbiol.* **161**, 1-8.

Geromanos, S. J., Vissers, J. P. C., Silva, J. C., Dorschel, C. A., Li, G.-Z., Gorenstein, M. V., Bateman, R. H. and Langridge, J. I. (2009). The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics.* **9**, 1683-1695.

Giddings, J. C. (1963). Liquid Chromatography with Operating Conditions Analogous to Those of Gas Chromatography. *Anal. Chem.* **35**, 2215-2216.

Goodlett, D. R., Keller, A., Watts, J. D., Newitt, R., Yi, E. C., Purvine, S., Eng, J. K., Haller, P. v., Aebersold, R. and Kolker, E. (2001). Differential stable isotope labeling of peptides for quantitation and *de novo* sequence derivation. *Rapid Commun. Mass Spectrom.* **15**, 1214-1221.

Green, P. N. and Bousfield, I. J. (1983). Emendation of *Methylobacterium* Patt, Cole, and Hanson 1976; *Methylobacterium rhodinum* (Heumann 1962) comb. nov. corrig.; *Methylobacterium radiotolerans* (Ito and Iizuka 1971) comb. nov. corrig.; and *Methylobacterium mesophilicum* (Austin and Goodfellow 1979) comb. nov. *Int. J. Syst. Evol. Microbiol.* **33**, 875-877.

Guell, M., Yus, E., Lluch-Senar, M. and Serrano, L. (2011). Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Micro.* **9**, 658-669.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. and Aebersold, R. (1999a). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994-999.

Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999b). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720-1730.

Han, X., Aslanian, A. and Yates Iii, J. R. (2008). Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12**, 483-490.

Hansen, K. C., Schmitt-Ulms, G., Chalkley, R. J., Hirsch, J., Baldwin, M. A. and Burlingame, A. L. (2003). Mass Spectrometric Analysis of Protein Mixtures at Low Levels Using Cleavable 13C-Isotope-coded Affinity Tag and Multidimensional Chromatography. *Mol. Cell. Proteomics.* **2**, 299-314.

Hanson, R. S. and Hanson, T. E. (1996). Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439-71.

Horvath, C. G. and Lipsky, S. R. (1966). Use of Liquid Ion Exchange Chromatography for the Separation of Organic Compounds. *Nature.* **211**, 748-749.

Hsu, J. L., Huang, S. Y., Chow, N. H. and Chen, S. H. (2003). Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* **75**, 6843-6852.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430-443.

Huber-Humer, M., Gebert, J. and Hilger, H. (2008). Biotic systems to mitigate landfill methane emissions. *Waste Manag. Res.* **26**, 33-46.

Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T. and Oda, Y. (2005). Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards. *Nat. Biotech.* **23**, 617-621.

Jennings, K. R. (1968). Collision-induced decompositions of aromatic molecular ions. *Int. J. Mass. Spectrom. Ion. Phys.* **1**, 227-235.

Jiang, H., Chen, Y., Jiang, P., Zhang, C., Smith, T. J., Murrell, J. C. and Xing, X.-H. (2010). Methanotrophs: Multifunctional bacteria with promising applications in environmental bioengineering. *Biochem. Eng. J.* **49**, 277-288.

Karas, M., Bachmann, D., Bahr, U. and Hillenkamp, F. (1987). Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *Int. J. Mass Spectrom. Ion Processes.* **78**, 53-68.

Karas, M. and Hillenkamp, F. (1988). Laser desorption ionization of proteins with the molecular masses exceeding 10000 Daltons. *Anal. Chem.* **60**, 2299-2301.

Kim, M.-S., Zhong, J., Kandasamy, K., Delanghe, B. and Pandey, A. (2011). Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides. *Proteomics.* **11**, 2568-2572.

MacBeath, G. (2002). Protein microarrays and proteomics. *Nat. Genet.* **32**, 526-532.

Maier, T., Güell, M. and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* **583**, 3966-3973.

Makarov, A. (2000). Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156-1162.

Mazzeo, J. R., Chen, W., Geromanos, S. J. and Berger, S. J. (2010). Comparison Of Detector Technologies For LC/QTOF MSE Biotherapeutic Protein Peptide Mapping Studies. *Proc. 58th ASMS Conf. on Mass Spectrometry and Allied topics*, Salt Lake City, USA.

McDonald, P. D. (2009). The Quest for Ultra Performance in Liquid Chromatography. Milford, MA, Waters Corporation.

McLafferty, F. W. (2008). Mass spectrometry across the sciences. *Proc. Natl. Acad. Sci.* **105**, 18088-18089.

McLafferty, F. W. and Bryce, T. A. (1967). Metastable-ion characteristics: characterization of isomeric molecules. *Chemical Communications (London).*

Motoyama, A., Venable, J. D., Ruse, C. I. and Yates, J. R. (2006). Automated Ultra-High-Pressure Multidimensional Protein Identification Technology (UHP-MudPIT) for Improved Peptide Identification of Proteomic Samples. *Anal. Chem.* **78**, 5109-5118.

Nagarajan, N. and Pop, M. (2010). Sequencing and Genome Assembly Using Next-Generation Technologies. *Computational Biology*, Springer. **673**: 1-17.

Nguyen, S. and Fenn, J. B. (2007). Gas-phase ions of solute species from charged droplets of solutions. *Proc. Natl. Acad. Sci.* **104**, 1111-1117.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007-4021.

Oda, Y., Huang, K., Cross, F. R., Cowburn, D. and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 6591-6.

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics.* **1**, 376-386.

Ong, S. E. and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252-262.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature.* **405**, 837-846.

Pappin, D. J. C., Hojrup, P. and Bleasby, A. J. (1993). Rapid Identification Of Proteins By Peptide-Mass Fingerprinting. *Curr. Biol.* **3**, 327-332.

Parker, C. E., Warren, M. R. and Mocanu, V. (2010). Neuroproteomics, CRC Press.

Patel, V. J., Thalassinos, K., Slade, S. E., Connolly, J. B., Crombie, A., Murrell, J. C. and Scrivens, J. H. (2009). A Comparison of Labeling and Label-

Free Mass Spectrometry-Based Proteomics Approaches. *J. Proteome Res.* **8**, 3752-3759.

Patt, T. E., Cole, G. C., Bland, J. and Hanson, R. S. (1974). Isolation and Characterization of Bacteria That Grow on Methane and Organic Compounds as Sole Sources of Carbon and Energy. *J. Bacteriol.* **120**, 955-964.

Roepstorff, P. and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biolomed. Mass Spectrom.* **11**, 601.

Ross, P. L., Huang, Y. L. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. and Pappin, D. J. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics.* **3**, 1154-1169.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science.* **270**, 467-470.

Schmidt, A., Kellermann, J. and Lottspeich, F. (2005). A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics.* **5**, 4-15.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature.* **473**, 337-342.

Shen, Y., Tolić, N., Xie, F., Zhao, R., Purvine, S. O., Schepmoes, A. A., Moore, R. J., Anderson, G. A. and Smith, R. D. (2011). Effectiveness of CID, HCD, and ETD with FT MS/MS for Degradomic-Peptidomic Analysis: Comparison of Peptide Identification Methods. *J. Proteome Res.* **10**, 3929-3943.

Shen, Y., Zhang, R., Moore, R. J., Kim, J., Metz, T. O., Hixson, K. K., Zhao, R., Livesay, E. A., Udseth, H. R. and Smith, R. D. (2005). Automated 20 kpsi RPLC-MS and MS/MS with Chromatographic Peak Capacities of 1000-1500 and Capabilities in Proteomics and Metabolomics. *Anal. Chem.* **77**, 3090-3100.

Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G.-Z., McKenna, T., Nold, M. J., Richardson, K., Young, P. and Geromanos, S. (2005). Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Anal. Chem.* **77**, 2187-2200.

Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. and Geromanos, S. J. (2006). Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel MS Acquisition. *Mol. Cell. Proteomics.* **5**, 144-156.

Sly, L. I., Bryant, L. J., Cox, J. M. and Anderson, J. M. (1993). Development of a biofilter for the removal of methane from coal mine ventilation atmospheres. *Appl. Microbiol. Biotechnol.* **39**, 400-404.

Sohngen, N. L. (1906). Uber bakterien welche methan ab kohlenstoffnahrung und energiequelle gerbrauchen (On bacteria which use methane as a carbon and energy source). *Z. Bakteriol. Parazitenk. (Infektionster).* **15**, 513–517.

Stevens, R. C., Yokoyama, S. and Wilson, I. A. (2001). Global Efforts in Structural Genomics. *Science.* **294**, 89-92.

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. and Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9528-9533.

Szájli, E., Fehér, T. and Medzihradszky, K. F. (2008). Investigating the Quantitative Nature of MALDI-TOF MS. *Mol. Cell. Proteomics.* **7**, 2410-2418.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. and Matsuo, T. (1988). Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2**, 151-153.

Theisen, A. R., Ali, M. H., Radajewski, S., Dumont, M. G., Dunfield, P. F., McDonald, I. R., Dedysh, S. N., Miguez, C. B. and Murrell, J. C. (2005). Regulation of methane oxidation in the facultative methanotroph *Methylocella silvestris* BL2. *Mol. Microbiol.* **58**, 682-692.

Unlu, M., Morgan, M. E. and Minden, J. S. (1997). Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis.* **18**, 2071-2077.

Van Aken, B., Peres, C. M., Doty, S. L., Yoon, J. M. and Schnoor, J. L. (2004). *Methylobacterium populi* sp. nov., a novel aerobic, pink-pigmented, facultatively methylotrophic, methane-utilizing bacterium isolated from poplar trees (*Populus deltoides* x nigra DN34). *Int. J. Syst. Evol. Microbiol.* **54**, 1191-1196.

Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. and Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Meth.* **1**, 39-45.

Ward, N., Larsen, Ø., Sakwa, J., Bruseth, L., Khouri, H., Durkin, A. S., Dimitrov, G., Jiang, L., Scanlan, D., Kang, K. H., Lewis, M., Nelson, K. E., Methé, B., Wu, M., Heidelberg, J. F., Paulsen, I. T., Fouts, D., Ravel, J., Tettelin, H., Ren, Q., Read, T., DeBoy, R. T., Seshadri, R., Salzberg, S. L., Jensen, H. B., Birkeland, N. K., Nelson, W. C., Dodson, R. J.,

Grindhaug, S. H., Holt, I., Eidhammer, I., Jonasen, I., Vanaken, S., Utterback, T., Feldblyum, T. V., Fraser, C. M., Lillehaug, J. R. and Eisen, J. A. (2004). Genomic Insights into Methanotrophy: The Complete Genome Sequence of *Methylococcus capsulatus* (Bath). *PLoS Biol.* **2**, e303.

Waters. "waters.com/webassets/cms/library/docs/720002330en.pdf." Retrieved 4th May 2012.

Whittenbury, R. and Dalton, H. (1981). The methylotrophic bacteria. *The prokaryotes*. M. P. Starr, H. Stolp, H. G. Truper, A. Balows and H. G. Schlegel. Berlin, Springer-Verlag: 256-261.

Whittenbury, R., Phillips, K. C. and Wilkinson, J. F. (1970). Enrichment, Isolation and Some Properties of Methane-utilizing Bacteria. *J. Gen. Microbiol.* **61**, 205-218.

Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F. and Williams, K. L. (1995). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13**, 19-50.

Zhou, H., Ranish, J. A., Watts, J. D. and Aebersold, R. (2002). Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nat Biotech.* **20**, 512-515.

Zubarev, R. A., Kelleher, N. L. and McLafferty, F. W. (1998). Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120**, 3265-3266.

Chapter 2

Materials and Methods

2.1 Sample Preparation

Bacterial protein extracts used in this research were cultivated in the laboratory of our collaborator Prof. Colin Murrell. Bacterial growth and protein extractions were performed by Andrew Crombie and Yin Chen, School of Life Sciences, University of Warwick. *E. coli* proteome digests (MassPREP Ecoli Digestion Standard) were obtained from Waters.

2.1.1 Bacterial growth and protein extraction

Small-scale cultivation (Succinate bio rep 1 and propane bio rep 1)

Cells were grown in 1 L quickfit flasks containing 200 mL medium, sealed with Subaseal (Sigma Aldrich) rubber stoppers. Succinate (5 mM final concentration) was added to the growth medium and propane (10 v/v) supplied by injection through the rubber septum. Cells were harvested during mid-late exponential phase at a density (OD 540) of 0.24 (succinate) or 0.22 (propane).

Large scale-cultivation (acetate, methane, propane bio rep 2, propane bio rep 3, succinate bio rep 2, succinate bio rep 3)

For large scale cultivation, *M. silvestris* BL2 cells were grown in 4 litre (LH Series 210, Stoke Poges, UK) or 2 litre (Fermec 300, Electrolab, Tewkesbury, UK) fermenters supplied with methane or propane (20 – 30 ml/min) and air (100 – 250 ml/min). Acetate to 5 mM was added when growth slowed, indicating depletion of substrate. Succinate in the culture medium was measured enzymatically using a Megazyme succinate kit and replenished to 5 mM when the concentration fell below 1 mM. Under these conditions cultures rapidly exhaust the relatively small amount of fixed nitrogen initially supplied in the growth medium and are capable of nitrogen fixation allowing uninterrupted growth to high cell densities. Actively-growing nitrogen-fixing cells were harvested by centrifugation, washed in growth medium, re-suspended in 0.1 M PIPES buffer (piperazine-N,N'-bis[2-ethanesulfonic acid], pH 7.0), and frozen in liquid nitrogen. The OD 540 at harvesting was 10.0 (methane), 11.9 (propane bio rep 2 and bio rep 3), 3.3 (succinate bio rep 2), 3.7 (succinate bio rep 3) and 1.1 (acetate).

Protein extraction

Frozen cells were thawed and re-suspended in PIPES buffer containing 1 mM benzamidine and broken by three passes through a French pressure cell (137 MPa) (American Instrument Co., Silver Spring, MD). Cell debris and membranes were removed by two centrifugation steps (13,000 × *g*, 30 min, 4 °C, followed by 140,000 × *g*, 90 min, 4 °C). The supernatant, containing soluble cytoplasmic proteins, was used for analysis. A protein assay was conducted on the soluble extract, using a Bio-Rad protein assay reagent according to the manufacturers protocol.

2.1.2 Enzymatic digestion

A total of 700 μg of soluble protein extract was re-suspended in 1 mL 0.1% RapiGest (Waters Corporation, Manchester, U.K.) and concentrated using a 3 kDa centrifugal filter unit (Millipore). The solution was then heated at 80 °C for 15 min. Reduction was carried out with dithiothreitol at 60 °C for 15 min at a final concentration of 9 mM. Alkylation was carried out in the dark with iodoacetamide at ambient temperature for 30 min at a final concentration of 16.7 mM, and digested with 1:50 (w/w) sequencing grade trypsin (Promega, Southampton, U.K.) at 37 °C overnight. RapiGest was hydrolysed by the addition of 2 μL 24 M formic acid, which was then incubated at 37 °C for 20 min. The resulting mixture was vortexed and centrifuged in a 0.22 μm Nylon membrane filter (Corning Incorporated, Corning, NY) to remove particulate matter.

2.1.3 Bacterial sample preparation for LC-MS^{E}

This section describes the methods for combining bacterial protein digests with an internal standard of known concentration for quantification purposes using the Hi3 method (Silva *et al.* 2006). When comparing relative quantification between two conditions, the same internal standard was used for each condition. All internal standards were purchased, pre-weighed and pre-digested from Waters. The standards did not share sequence homology with the bacterial proteins studied. All solvents were obtained from JT Baker, unless specified.

1D RP-LC analyses

Samples were diluted with glycogen phosphorylase *b* (MassPREP Phosphorylase *b* Digestion Standard) digest to give a final sample concentration of 500 - 1000 ng/μL with an internal standard concentration of 50 fmol/μL in 0.1% formic acid (aqueous). Glycogen phosphorylase *b* was diluted in 0.1 % formic acid (aqueous) to 50 fmol/μL as a quality control check and was resolved by 1D RP-LC prior to any experiments. It was also sampled after every three experiments to provide quality control check points during the course of multiple experiments.

2D RP-RP-LC analyses

Selected bacterial samples were diluted to include 23 fmol yeast alcohol dehydrogenase per 2.5 μg of protein digest in 20 mM ammonium formate. The majority of bacterial samples were diluted accordingly to allow 2 - 5 μg sample per 100 fmol glycogen phosphorylase *b* in 20 mM ammonium formate. 100 fmol rabbit glycogen phosphorylase *b* was separated by 2D RP-RP-LC in a pseudo 1D configuration (see section 2.2.3) for quality control purposes as described previously for 1D RP-LC analyses.

2.1.4 Four protein mixture for system assessment

In order to assess quantification limits, accuracy and reproducibility, measurements were conducted on a mixture of four proteins of known concentration in the absence and presence of a complex matrix.

The mixture (MassPREP Protein Digestion Standard Mix 2), contained yeast alcohol dehydrogenase, rabbit glycogen phosphorylase *b*, yeast enolase and bovine serum albumin present at a ratio of 1: 0.34: 1: 3.87 or 1: 0.41: 1.29: 5.79 when normalised to alcohol dehydrogenase, using a TDC or ADC detector, respectively. The mixture was diluted to allow the loading of each protein as presented in Table 2.1.

Table 2.1. Experiments conducted using four-protein mixture standard

Mass spectrometer	LC-MS	Detector	ADH (fmol)	Phos b (fmol)	ENO (fmol)	BSA (fmol)	Biological matrix
Synapt	1D MS^E	TDC	25.0	8.5	25.0	96.8	-
Synapt	1D MS^E	TDC	25.0	8.5	25.0	96.8	130 ng methane digest
Synapt	1D MS^E	TDC	25.0	8.5	25.0	96.8	170 ng methane digest
Synapt	1D MS^E	TDC	25.0	8.5	25.0	96.8	240 ng methane digest
Synapt G2	1D MS^E	hADC	5.0	2.1	6.5	29.0	-
Synapt G2	1D MS^E	hADC	12.5	5.1	16.1	72.4	-
Synapt G2	1D MS^E	hADC	25.0	10.25	32.3	144.8	-
Synapt G2	1D MS^E	hADC	50.0	20.5	64.5	289.5	-
Synapt G2	1D $HDMS^E$	hADC	5.0	2.1	6.5	29.0	-
Synapt G2	1D $HDMS^E$	hADC	12.5	5.1	16.1	72.4	-
Synapt G2	1D $HDMS^E$	hADC	25.0	10.25	32.3	144.8	-
Synapt G2	1D $HDMS^E$	hADC	50.0	20.5	64.5	289.5	-
Synapt G2	2D MS^E	hADC	50.0	20.5	64.5	289.5	300 ng propane digest
Synapt G2	2D $HDMS^E$	hADC	50.0	20.5	64.5	289.5	300 ng propane digest

Alcohol dehydrogenase (ADH); glycogen phosphorylase *b* (Phos b); enolase (ENO); bovine serum albumin (BSA)

2.2 Liquid Chromatography and Mass Spectrometry Configurations

2.2.1 One-dimensional reversed phase nanoLC

BEH analytical column

Nanoscale LC separations of tryptic peptides were performed using a nanoACQUITY system (Waters Corporation) with a Symmetry C_{18} trap column (180 μm × 20 mm, 5 μm) and a BEH C_{18} analytical column (75 μm × 250 mm, 1.7 μm). The composition of solvent A was 0.1 % formic acid in water, solvent B was 0.1 % formic acid in acetonitrile (pH 2.7). 1 μL injections of each sample (~500 - 1000 ng total protein and 50 fmol internal standard) were loaded onto the trap column and flushed with 99.9% A for 2 min at a flow rate of 15 μL/min. The organic solvent concentration was increased from 3% to 40% B over 90 min, followed by 2 min wash at 90% solvent B. The column was re-equilibrated with 3% B for 20 min prior to the next run. All analyses were conducted in triplicate. For the analysis of single proteins or 4-protein mixture in the absence of a biological matrix, the organic solvent concentration was increased from 3% to 40% B over 30 min.

2.2.2 High pH reversed phase-low pH reversed phase nanoLC

2 - 5 μg of digest was loaded onto the first dimension column, XBridge BEH130 C18 NanoEase Column (300 μm x 50 mm, 5 μm) using a 2D nanoACQUITY system equilibrated with solvent A (20 mM ammonium formate pH 10) at 2 μL/min. The first fraction was eluted from the first dimension column at 10.8 % B (100% acetonitrile) for 4 min at a flow rate of 2 µL/min and transferred to the second dimension trap column via a 1:10 dilution with 99.9% solvent A (0.1% formic acid in water) at 20 µL/min, over a time period of 20.5 min. Peptides were eluted from the trap column and resolved on the analytical column (BEH or HSS). Subsequent fractions were isolated by increasing the concentration of organic solvent from the first dimension and reducing the loading time from 20.5 min to 15.5 min. First dimension fractions were eluted at 13.1, 17.7 and 50.0% buffer B for three-fraction experiments, 10.8, 14.0, 16.7, 20.4, 50 and 60% buffer B for the six-fraction

experiments and 6.9, 10.4, 12.1, 13.5, 14.7, 15.9, 17.3, 18.8, 20.9, 23.9 and 65% buffer B for the eleven-fraction experiments.

2.2.3 One-dimensional nanoLC performed on two-dimensional configuration

Selected samples were resolved using the second dimensional reversed phase column only when the 2D nanoACQUITY system was configured with two reversed phase columns. During two-dimensional configurations, 1D analyses using only the separation power of the second analytical column require that samples pass through the first dimensional column. A method was developed in which only one fraction was eluted from the first dimension. The first dimensional column was equilibrated in 1% buffer B at 1 μL/min. Peptide elution onto the second dimension was achieved using an acetonitrile pulse (60% buffer B) for 4 min commencing 1 minute into the run. The eluate was transferred to the second dimension trap column via 1:20 or 1:10 dilution with 0.1% buffer B (0.1% formic acid in water) at 20 µL/min, over a time period of 20.5 min. The Xbridge column eluent buffer composition for the remainder of the analyses was held at 1% buffer B. The peptides were then eluted from the trapping column as described below.

HSS analytical column

Peptides were eluted from the Symmetry C_{18} trap column (180 μm × 20 mm, 5 μm) onto a nanoACQUITY UPLC HSS T3 C_{18} analytical column (75 μm × 150 mm, 1.8 μm). The composition of solvent A was 0.1 % formic acid in water, solvent B was 0.1 % formic acid in acetonitrile (pH 2.7). The organic solvent concentration was increased from 5% to 40% B over 90 min using a linear gradient at a constant temperature of 35 °C. The column was re-equilibrated with 3% B for 20 min prior to the next run.

2.2.4 Data-independent and mobility-assisted data-independent acquisition

Data-independent acquisition (MS^E) – Synapt and Synapt G2

Precursor ion masses and associated fragment ion spectra of the tryptic peptides were acquired using a Synapt HDMS or Synapt HDMS G2 mass spectrometer (Waters Corporation)(Pringle *et al.* 2007; Giles *et al.* 2011), using an applied capillary voltage of 3.2 - 3.5 kV. The time-of-flight analyser of the mass spectrometer was calibrated with NaI (1 mg/mL) and CsI (50 ng/mL) or by using the MS/MS spectra of the doubly charged precursor of [Glu^1]-Fibrinopeptide B (GFP) (Sigma Aldrich, St. Louis, MO). Data were lockmass-corrected using the monoisotopic mass of the doubly charged precursor of GFP which was delivered at 500 fmol/μL using the auxiliary pump at a flow rate of 500 nL/min and sampled from the reference sprayer every 60 s. Accurate mass data were collected in a data-independent mode of acquisition (MS^E), whereby the energy applied in the trap section is alternated between low (3 eV) and elevated (15 to 30 eV ramped) energy scan functions for experiments conducted using the Synapt HDMS. 6 eV (low) and 10 – 30 eV (elevated) energy settings were chosen for experiments conducted using the Synapt HDMS G2, operating in resolution mode. This method provides a near 100% duty cycle as precursors and product ion data for isotopes on all charge states are collected across the entire chromatographic peak width. The correlation of precursor ions to their corresponding MS/MS spectra is achieved by the use of a retention time alignment algorithm (Li *et al.* 2009). Spectral data were collected over a m/z range of 50 – 1950, and the acquisition rate was 0.9 s with a 0.1 s interscan delay.

Mobility assisted data-independent acquisition ($HDMS^E$) – Synapt G2

For $HDMS^E$ experiments, precursors were mobility separated in the mobility cell with a wave velocity of 600 m/s and a wave height 40 V. Precursors were subjected to alternate energy functions in the transfer section of the mobility cell (0 eV for low energy and 19 – 45 eV ramped for elevated energy scans).

The correlation of precursor ions with their corresponding MS/MS spectra was achieved by the use a retention time alignment algorithm, followed by a further correlation process that is based on the physicochemical characteristics of peptides when they undergo collision induced fragmentation (Li *et al.* 2009). Mobility resolved

precursor and product ion spectra were obtained by acquiring 200 spectra (each of 64 μs duration) during low and elevated collision energy regimes. Precursor peptides that co-elute in the absence of mobility separation may be separated according to mass, charge and shape (rotationally averaged cross-section). This additional degree of separation is designed to improve the quality of the product ion spectra obtained. In conventional MS^E experiments accurate mass precursor and product ion spectra, together with LC retention time, are collected but here, in addition, mobility arrival times are also measured.

2.2.5 Overview of experiments

Table 2.2 outlines the experiments carried out to characterise the soluble protein content of *M. silvestris* grown at the expense of various growth conditions. An *E. coli* extract was analysed using selected LC-MS configurations.

Table 2.2. Experiments conducted with bacterial extracts

Numbers in brackets denote the number of first dimensional fractions employed.

Sample	LC	Column	MS	Mode of operation	Internal standard
Trimethylamine	1D RP-LC	BEH	Synapt	MS^E	Phos b
Methylamine	1D RP-LC	BEH	Synapt	MS^E	Phos b
Methanol	1D RP-LC	BEH	Synapt	MS^E	Phos b
Propane – bio rep 1	1D RP-LC	BEH	Synapt	MS^E	Phos b
	2D RP-RP-LC (6)	BEH	Synapt	MS^E	ADH
Propane – bio rep 2	1D RP-LC	BEH	Synapt	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	$HDMS^E$	Phos b
	2D RP-RP-LC (6)	BEH	Synapt	MS^E	Phos b
Propane – bio rep 3	1D RP-LC	HSS	Synapt G2	MS^E	Phos b
	1D RP-LC	HSS	Synapt G2	$HDMS^E$	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	MS^E	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	$HDMS^E$	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	MS^E	4-protein mix
	2D RP-RP-LC (3)	HSS	Synapt G2	$HDMS^E$	4-protein mix
Succinate – bio rep 1	1D RP-LC	BEH	Synapt	MS^E	Phos b
	2D RP-RP-LC (6)	BEH	Synapt	MS^E	ADH
Succinate – bio rep 2	1D RP-LC	BEH	Synapt	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	$HDMS^E$	Phos b
Succinate – bio rep 3	1D RP-LC	HSS	Synapt G2	MS^E	Phos b
	1D RP-LC	HSS	Synapt G2	$HDMS^E$	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	MS^E	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	$HDMS^E$	Phos b

Table 2.2 continued.

Sample	LC	Column	MS	Mode of operation	Internal standard
Methane	1D RP-LC	BEH	Synapt	MS^E	Phos b
	1D RP-LC	BEH	Synapt	MS^E	4-protein mix
	2D RP-RP-LC (6)	BEH	Synapt	MS^E	Phos b
	2D RP-RP-LC (11)	BEH	Synapt	MS^E	Phos b
Acetate	1D RP-LC	HSS	Synapt G2	MS^E	Phos b
	1D RP-LC	HSS	Synapt G2	$HDMS^E$	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	MS^E	Phos b
	2D RP-RP-LC (3)	HSS	Synapt G2	$HDMS^E$	Phos b
E. coli	1D RP-LC	BEH	Synapt	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	MS^E	Phos b
	1D RP-LC	BEH	Synapt G2	$HDMS^E$	Phos b

2.3 Data Analysis

Optimal processing settings were chosen for data according to the amount of sample loaded, the LC separation, the mode of operation and the version of software employed. All data was processing using the most up to date database available. Further details may be found in Table 2.3.

2.3.1 Data processing and database searching

Each LC-MS^E raw data file was processed using ProteinLynx Global Server (PLGS) v.2.4 or v.2.5 in order to generate centroided, charge-state reduced and deisotoped peptide masses with associated fragment ions. For $HDMS^E$ raw data files, processing was carried out using a workstation equipped with a Graphics Processor Unit. This is required to compute intensive calculations carried out during the processing of these large data files. Ion detection, clustering, and normalisation of data has been described in detail elsewhere (Silva *et al.* 2005). In brief, all fragment ions within a retention time window associated to one tenth of the chromatographic peak width of a precursor ion are time-aligned or assigned to the precursor. The resulting precursor-product ion list is then queried against a database utilising an iterative three-step process. The first step considers data that matches certain criteria, namely those that are correctly cleaved, with mass tolerances better or equal to 10 ppm for precursor

ions and 20 ppm or better for product ions and those with a fixed modification. As a consequence of these database search tolerances, each submitted precursor can provide multiple tentative peptide identifications. Overall, the strategy of the algorithm aims to result in a single precursor assigned to one peptide. All other lower ranking tentative peptide identifications annotated with each securely identified precursor are not considered. The product ions used for the validation of each high ranking precursor are then removed from the precursor-product list of other co-eluting precursors, thereby eliminating them for consideration when identifying any coincidently detected precursors. These tentative peptides are then ranked and scored based on how well they conform to 14 predetermined models of specific, physicochemical attributes. These include (i) the presence of certain amino acids at of near the N-terminus, affecting the ratio of the total *y* ion intensity to the total *b* ion intensity, (ii) the presence of product ions corresponding to loss of H_2O and NH_3 from peptides containing specific amino acids, (iii) complementary C or N terminal product ions, (iv) how well the tentative amino acid sequence supports the observed peptide charge states, and (v) how well the experimental retention time matches the theoretical retention time. During this second step, precursor and product ions that have not yet been assigned are queried against a subset database of identified proteins from the first step. This includes missed cleavages, in-source fragments, neutral losses, and variable modifications. During this final step, remaining unidentified ions are considered against the complete database for additional protein identifications, which includes peptide identifications without fragment ion information (Li *et al.* 2009). An overview of this process is shown in Figure 2.1.

Processed data files from 2D analyses were merged prior to database searching. All data were searched against a *M. silvestris* database (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi). A randomised version of the query database was generated by randomizing the protein sequences, while holding the peptide amino acid composition and the number of tryptic peptides constant. A random decoy database was generated whereby each protein in the query database had been randomised once. The decoy database was then merged with the original database along with sequences for the internal standards used and common contaminants. The false positive rate is estimated by the number of randomised entries identified (false positive rate, FPR) divided by the number of correct identifications (true positive rate, TPR) expressed as a percentage. A fixed

modification of carbamidomethyl-C was specified. Variable modifications considered included acetyl N-terminus, deamidation N, deamidation Q and oxidation M. Automatic settings for mass accuracy were used and two missed tryptic cleavages were allowed. The protein identification criteria included the number of fragment ions per peptide detected, product ion matches per protein and peptides per protein. Minimum values of these criteria are outlined in Table 2.3.

Table 2.3. Data processing and database search parameters

Sample loading (μg)	LC	MS	Mode of operation	Software	Fragment ions/ peptide	Fragment ions/ protein	Peptides /protein	Low intensity threshold	High intensity threshold	Intensity threshold
0.05 – 0.5	1D	Synapt	MS^E	v.2.4	3	7	1	250	100	1500
0.05 – 0.5	1D	Synapt	MS^E	v.2.5	3	7	1	100	30	750
2.5 – 5.0	2D	Synapt	MS^E	v.2.4	3	7	1	250	75	1000
0.05 – 0.1	1D	G2	MS^E	v.2.5	3	7	1	80	20	500
0.05 – 0.1	1D	G2	$HDMS^E$	v.2.5	2	5	1	80	20	500
0.5 – 2.0	1D/2D	G2	MS^E	v.2.5	3	7	1	100	30	750
0.5 – 2.0	1D/2D	G2	$HDMS^E$	v.2.5	2	5	1	100	30	750

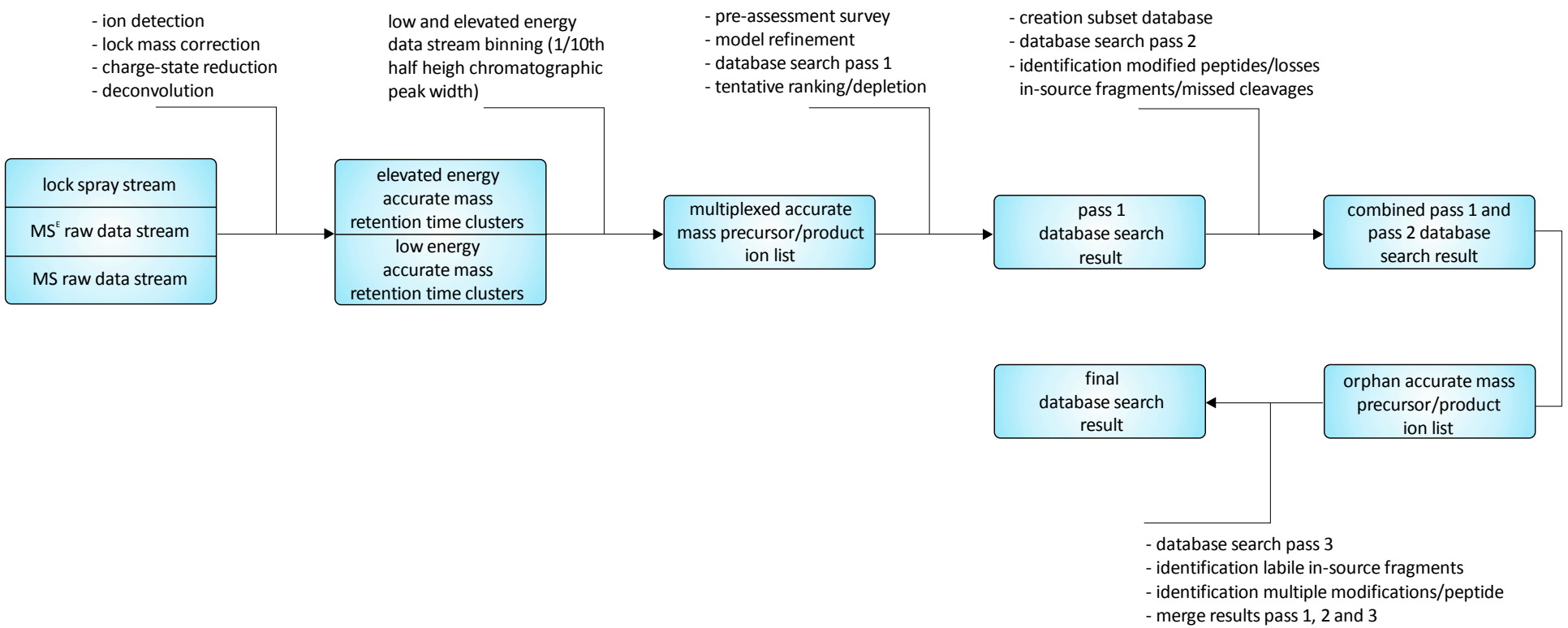

Figure 2.1. Workflow ion accounting database search algorithm illustrating data processing and the iterative search process

Adapted from (Li *et al.* 2009).

2.3.2 Filtering and normalisation

Protein and peptide identifications were filtered to include proteins that had been identified in two or more technical replicates. By using replication as a filter, the false positive rate is minimised since false positive identifications are assumed to have a random nature and as such do not tend to replicate across injections. The use of accurate mass information for the precursor and associated fragment ions for alignment purposes together with the high number of matched peptides per protein achieved with DIA methods compared to a typical DDA experiment, results in a high level of confidence >95%. Since this search algorithm is not probability based, no selection of individual MS/MS spectra is undertaken.

Protein abundances estimated using the Hi3 method were reported by the algorithm as weight (ng) and molar quantity (fmol) for each protein. Protein abundances were converted to a percentage of total detected protein by weight or moles accordingly. For example, in the discussion of subunit stoichiometry the abundance of proteins was best described in molar quantities. Excel (Microsoft Corp., Redmond, WA) was used to filter results obtained from technical analyses and to compute average values for the data obtained (protein scores, sequence coverage, matched peptides, protein quantities etc.).

2.3.3 Relative expression analyses

The regulation of proteins observed under different growth conditions was monitored using the Expression tool in PLGS. The data were auto-normalised, which allowed for small differences in loading amounts. For example, in an experiment which entailed the measurement of a mixture containing 25 fmol ADH and 50 fmol ADH, a ratio of 1:1 is reported when the data are auto-normalised. When normalisation of this data was employed a ratio of 1:2 was calculated. A more consistent representation of regulated proteins is obtained when comparing the levels between equal loading amounts, therefore data was auto-normalised.

Log(e) ratios, scores, standard deviations and probabilities were reported for proteins shared between two different conditions. The probability value is assigned from 0 to 1; those values that are 0.05 or lower and 0.95 and higher, representing regulation likelihood greater than 95%, have been assumed to be statistically significant. Output was filtered to exclude ratios with a score less than 200 - 250 and

to include proteins that were present in at least two technical replicates. A cut-off score was determined by assessing the number of random entries that had entered after filtering and by assessing the standard deviation. Significant biological regulation was determined at a 30% fold change, which is approximately -0.3 and 0.3 on the natural log scale and is estimated to be 2-3 times the estimated error of the intensity measurement. Those identifications with relative expression values less than or equal to -0.3 were identified as down-regulated and those with values of more than, or equal to, 0.3 were assigned as up-regulated.

2.3.4 Bioinformatic tools

Venn diagrams were produced using VennMaster-0.37.3. The entire theoretical proteome of *M. silvestris* and selected experimentally observed proteins were subjected to cellular localisation prediction using PSORTb v3.0.2 (Yu *et al.* 2010). This program seeks structural motifs, signal peptides, experimentally observed data etc. to predict protein localisation (cytoplasmic, inner membrane, periplasmic, outer membrane or extracellular) for gram-negative bacteria. The tool TOPCONS was used to predict transmembrane helices in selected proteins (Bernsel *et al.* 2009).

Data analysis of large protein lists is a very important downstream task to probe their biological significance. Data analysis of such highly complex and large volume data sets is a challenging task, which requires support from bioinformatics software packages. Selected lists of protein accession numbers were submitted to the online database DAVID (the database for annotation, visualization and integrated discovery) (Huang *et al.* 2008) which is able to extract biological features/significance associated with large gene lists. This database combines a number of tools that carry out functional classification and pathway mapping. Proteins were mapped to pathways present in the Kyoto Encyclopedia of Genes and Genomes database (KEGG), which combined pathways that have been experimentally verified in a large number of organisms.

2.4 Validation

RT-PCR was carried out targeting Msil_3604, to confirm the presence of its transcript in methylamine-grown cells. The forward primer and reverse primer used, shown in Table 2.4, produced a PCR product of 268 base pairs at an annealing temperature of 55 °C.

Table 2.4. Primers designed to target Msil_3604

Forward Primer	Reverse Primer
CGCCAGCCAATGCTATAAAT	ATTGTCCTGCCAGAAAATGC

2.5 References

Bernsel, A., Viklund, H., Hennerdal, A. and Elofsson, A. (2009). TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **37**, W465-W468.

Giles, K., Williams, J. P. and Campuzano, I. (2011). Enhancements in travelling wave ion mobility resolution. *Rapid Commun. Mass Spectrom.* **25**, 1559-1566.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols.* **4**, 44-57.

Li, G. Z., Vissers, J. P. C., Silva, J. C., Golick, D., Gorenstein, M. V. and Geromanos, S. J. (2009). Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics.* **9**, 1696-1719.

Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T. and Scrivens, J. H. (2007). An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass spectrom.* **261**, 1-12.

Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G.-Z., McKenna, T., Nold, M. J., Richardson, K., Young, P. and Geromanos, S. (2005). Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Anal. Chem.* **77**, 2187-2200.

Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. and Geromanos, S. J. (2006). Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel MS Acquisition. *Mol. Cell. Proteomics.* **5**, 144-156.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S. b., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J. and Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics.* **26**, 1608-1615.

Chapter 3

Comparison of Liquid Chromatography Approaches

3.1 Introduction

3.1.1 Challenges in proteomic studies

The proteomic component of biological samples is typically made up of a highly complex mixture of proteins. Serum proteomes can contain thousands, or even tens of thousands of proteins with concentrations that may span twelve orders of magnitude (Anderson and Anderson 2002). The number of peptides created from a simple prokaryotic organism forms a challenging analytical problem; if 30 - 40 peptides are created from each protein (~2,000), the resulting tryptic digest may contain more than 80,000 peptides. Regardless of the type of proteomics experiment carried out (bottom-up, top-down etc.), separation is typically required before mass analysis. One-dimensional high performance liquid chromatography advances have been made by laboratories packing their own columns. The Smith group were able to identify 12,000 peptides relating to 2,000 *Shewanella oneidensis* proteins (~40% of predicted) in a 12 h run on a 1D-LC set up coupled to a linear ion trap (Shen *et al.* 2005).

The coupling of two or more LC approaches may afford improved separation, which can lead to the identification of more peptides. Ideally, both these approaches should operate at high peak capacity, have good separation power, be highly reproducible and be compatible with MS. Additional considerations for choosing a separation strategy are associated with the type and number of samples. The time it takes to prepare samples, carry out the separation and undertake mass analysis may be a concern when dealing with large numbers of samples. The generation of sufficient quantity of sample can be challenging. A more recent requirement of separation technologies is how well they can be adapted for quantitative proteomic studies.

Here, 1D RP-LC and 2D RP-RP-LC separation techniques were used for label-free quantitative analysis using the MS^E approach. These methods were assessed using simple mixtures and applied to characterise protein extracts from *M. silvestris* grown on substrates methane, succinate and propane.

3.1.2 Multi-dimensional liquid chromatography for peptide separations

Multi-dimensional liquid chromatography (MD LC), which can be described at the combination of two or more forms of LC, is a technique that can afford higher resolving power than uni-dimensional LC.

MD LC can be performed either online or offline. Offline MD-LC can be more flexible because it can allow the use of incompatible buffers in the various LC dimensions. First dimensional fractions may be collected in discrete fractions, which are then snap frozen and lyophilised before being reconstituted in a compatible buffer for the second dimension. For offline systems, the potential for sample losses is higher and a larger initial protein loading is required in order to minimize losses and optimize the number of proteins identified. Online MD LC faces a greater challenge in coupling together two dimensions where incompatible buffers may be used. Various approaches have been developed, including the use of storage loops, trap columns, parallel analytical columns in the second dimension or a stop-flow approach for collecting first dimension fractions online (Malerod *et al.* 2010). Although these systems are challenging to operate, they allow automation, with minimal sample loss and contamination, high-throughput operation, and the analyses of small quantities of sample.

A variety of first dimension technologies have been used in MD LC approaches for proteomics, including reversed phase (RP) (Gilar *et al.* 2005b), strong cation-exchange chromatography (SCX) (Link *et al.* 1999), size exclusion chromatography (SEC) (Gilar *et al.* 2005a), hydrophilic interaction liquid chromatography (HILIC) (Boersema *et al.* 2008) and isoelectric focussing (IEF) techniques (O'Farrell 1975) (Slebos *et al.* 2008). The method most commonly used for peptide separation typically includes SCX in the first dimension and RP in the second dimension. This approach separates peptides according to solute charge and hydrophobicity, respectively (Malerod *et al.* 2010).

Alternative combinations of separation technologies have been evaluated and compared to SCX-RP, showing similar or improved peptide separation and identification (Dowell *et al.* 2008). Two groups have independently shown that a combination of high-pH RP and low-pH RP can be an effective way of separating peptides and provides results that are at least comparable to SCX-RP (Gilar *et al.* 2005b; Gilar *et al.* 2005a; Toll *et al.* 2005; Gilar *et al.* 2009), and in some cases

significantly better (Zhou *et al.* 2010). Although both high-pH RP and low-pH RP separate peptides based on their hydrophobicity, the two separations have been shown to be significantly orthogonal. Each dimension exhibits altered selectivity because of the different charges of the amino acid side chains of the peptides at different values of pH. A recently developed online high pH RP-low pH RP UPLC system (2D RP-RP-LC) couples the two dimensions together via a dilution step. In this system, a fraction is eluted from the first dimensional column containing a percentage of organic at high pH. This fraction cannot be directly analysed by the second dimension due to the presence of organic solvent. Instead, the fraction is diluted with a highly aqueous solvent prior to separation on the second column. Details of this configuration are shown in Figure 3.1.

MD LC-based studies have been shown to provide better proteome coverage by identifying more peptides. Increasingly information regarding the concentration of these peptides and the proteins from which they originated from is desirable. A number of quantitative proteomics techniques are available and the choice of which should be based on the composition of the biological sample and the hypothesis being tested (Domon and Aebersold 2010). A subset of these techniques can be incorporated into shotgun strategies, which can include the use of MD LC systems. The chemical labelling approach iTRAQ (isobaric tags for relative and absolute quantification) (Ross *et al.* 2004) typically involves SCX-RP separation of labelled peptides prior to data-dependent MS acquisition. This is because a relatively simple mixture of peptides in unit time is optimal for MS operated using data-dependent acquisition. More recently, a complex whole lysate sample analysed using label-free methods included a SCX-RP step prior to data-independent MS acquisition. This provided qualitative information about membrane proteins without the need for enrichment (Yang *et al.* 2011). There is, however, a need to assess the suitability of label-free methods for quantitative and comparative proteomics studies where MD LC systems have been employed.

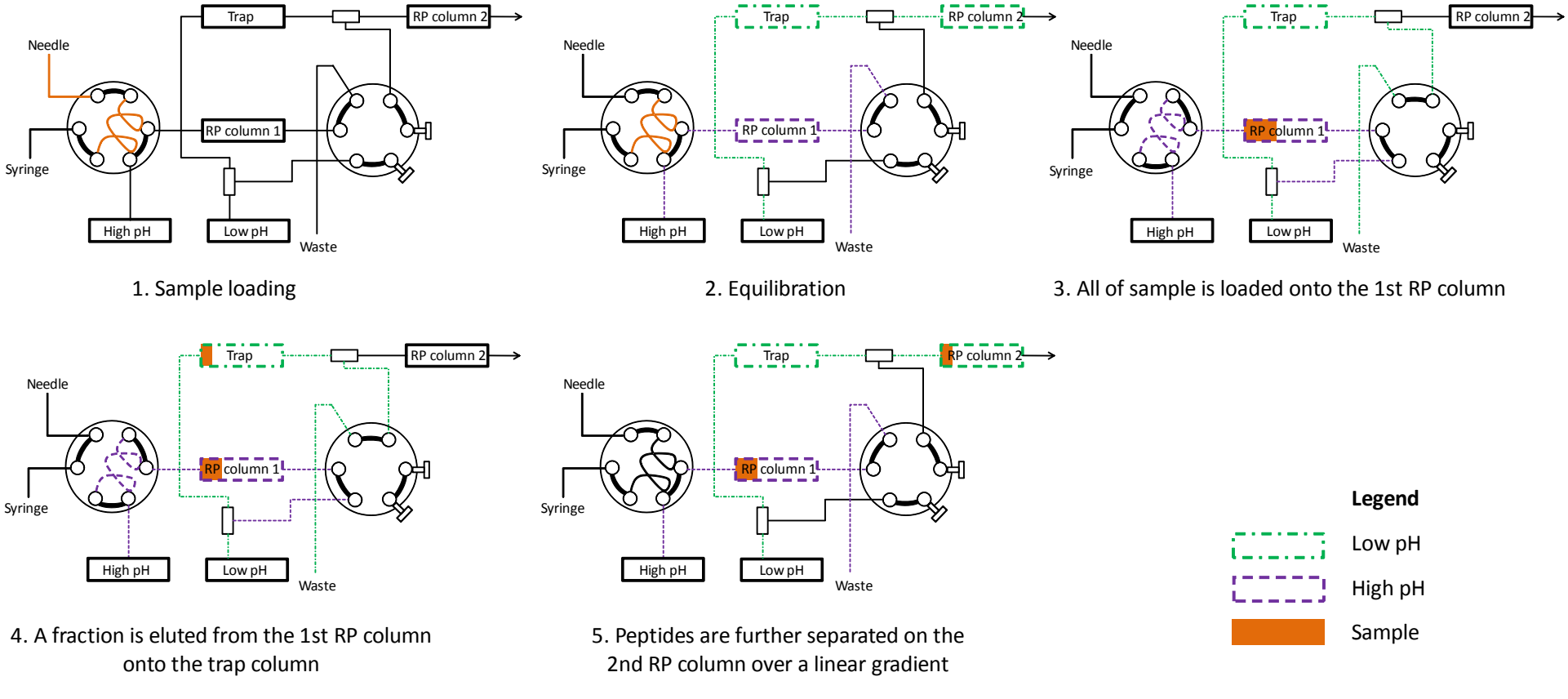

Figure 3.1. Schematic of an online 2D RP-RP-LC system

3.2 Aims

Improvements in LC separation techniques are crucial in advancing profiling and comparative proteomic studies. The recently developed high-pH RP low-pH RP system operates in an online manner, reducing the time it takes to collect fractions and individually transfer each one for further LC separation. Increasing selected variables in LC (i.e. gradient time, column length etc.) can improve peak capacity but may eventually have diminishing returns in terms of the number of proteins identified and the time required. In the case of 2D-LC, there is an additional variable. The user has a choice of how many fractions are eluted from the first LC dimension to use for further separation by the second dimension.

The aims of this work were:

- To assess the quantitative accuracy of the various experimental approaches using simple protein mixtures

- To evaluate 1D RP-LC MS^E and 2D RP-RP-LC MS^E methods for profiling and comparative experiments using a label-free quantitative approach

- To investigate the effect of varying the number of first dimensional fractions used for 2D RP-RP-LC MS^E

3.3 Methods

Bacterial protein preparation, LC-MS^E settings and data processing were detailed in Chapter 2. This section provides an overview of the experiments conducted to achieve the specific aims of this work.

3.3.1 System assessment

The accuracy and reproducibility of the quantification method was assessed by analysing a mixture of four pre-digested proteins (MassPREP Protein Digestion Standard Mix 2). The mixture containing yeast alcohol dehydrogenase, rabbit glycogen phosphorylase *b*, yeast enolase and bovine serum albumin are present at a ratio of 1: 0.34: 1: 3.87 or 1: 0.41: 1.29: 5.79 when normalised to alcohol dehydrogenase when quantified with a TDC or ADC detector, respectively. This mixture was resolved by 1D RP-LC and analysed using the conditions outlined in Table 3.1. The ability of the approach to quantify these proteins against a complex background was investigated in experiments 1 – 4. The 4-protein mixture was maintained at the stated concentrations and the amount of biological matrix was varied. All experiments were conducted in triplicate.

Table 3.1. Experiments conducted with standard proteins using 1D RP-LC

Experiment	Mass spectrometer	Mode of operation	Detector	Mixture (fmol)				Biological matrix
				ADH	Phos b	ENO	BSA	
1	Synapt	MS^E	TDC	25.0	8.5	25.0	96.8	-
2	Synapt	MS^E	TDC	25.0	8.5	25.0	96.8	130 ng methane digest
3	Synapt	MS^E	TDC	25.0	8.5	25.0	96.8	170 ng methane digest
4	Synapt	MS^E	TDC	25.0	8.5	25.0	96.8	240 ng methane digest

3.3.2 Comparison of LC techniques for profiling and comparative proteomics

Experiments were conducted on protein extracts from propane-, succinate- and methane-grown cells to evaluate the 1D RP-LC MS^E and 2D RP-RP-LC MS^E approches. The experiments that were used in these comparisons are detailed in

Table 3.2. Information regarding the growth substrate supplied, LC configuration, number of first dimensional fractions employed and the mass spectrometer are provided. LC-MS^E settings, data processing and database searching details were outlined in Chapter 2.

Table 3.2. Experiments conducted on methane, succinate and propane samples to compare 1D RP-LC MS^E and 2D RP-RP-LC MS^E

Sample	LC	MS	Mode of operation	No. of 1st dimensional fractions
Propane – bio rep 1	1D RP-LC	Synapt HDMS	MS^E	-
	2D RP-RP-LC	Synapt HDMS	MS^E	6
Succinate – bio rep 1	1D RP-LC	Synapt HDMS	MS^E	-
	2D RP-RP-LC	Synapt HDMS	MS^E	6
Methane	1D RP-LC	Synapt HDMS	MS^E	-
	2D RP-RP-LC	Synapt HDMS	MS^E	6
	2D RP-RP-LC	Synapt HDMS	MS^E	11

3.4 Results and Discussion

3.4.1 Overview

Two types of LC configurations were employed, prior to label-free data-independent MS analysis, to provide qualitative and quantitative information regarding the proteome of *M. silvestris* grown utilising different substrates. The number of peptides and proteins observed from each growth condition and experimental approach are outlined in Table 3.3.

Table 3.3. Proteome and protein coverage observed using 1D and 2D approaches in the study of *M. silvestris* grown with succinate, propane or methane

	Succinate		Propane		Methane	
	1D	2D*	1D	2D*	1D	2D*
Total number of protein identifications	327	918	248	656	178	603
Number of peptide identifications	2649	8610	2149	5453	1194	5998
Matched peptides per protein	11	19	11	17	12	22
Sequence coverage (%)	26	35	27	32	33	36
Protein false discovery rate (%)	0.0	0.5	0.4	0.6	0.6	0.2

*6 first dimensional fractions

Protein and peptide identifications were carried out using minimum fragment ion criteria with an additional repeatability filter as described in Chapter 2. The number of proteins identified by the 1D approach was 327, 248 and 178 for succinate-, propane- and methane-grown extracts, respectively, with an approximate three-fold increase in the number of identifications observed for each condition when the 2D approach was employed. A similar trend was observed when comparing the number of peptide identifications between the 1D and 2D approach. In addition to improved proteome coverage, the quality of individual protein coverage in terms of matched peptides and sequence coverage was also higher in the 2D method. Protein false discovery rates were estimated by the use of randomised decoy sequences during database searching; this rate was relatively low regardless of the condition or technique (<0.6%).

3.4.2 Peptide detections and separations

In a 1D RP-LC experiment, peptides are resolved by increasing the organic content over time to elute increasingly more hydrophobic peptides. In a 6 fraction 2D experiment, peptides are eluted with fixed amounts of acetonitrile per fraction at 10.8, 14.0, 16.7, 20.4, 50 and 60% of the total composition. Figure 3.2 shows the average isoelectric point and GRAVY score, (a measure of a hydrophobicity), of peptides in each of these fractions. A larger GRAVY score indicates more hydrophobic properties. As would be expected for a reversed phase experiment, peptides on average are more hydrophobic in high organic content fractions, shown in Figure 3.2(b). This is also apparent when visually assessing the separation space used in the 2nd RP dimension in a 2D experiment (Figure 3.5), which shows a shift in the earliest peptide retention time (RT) when organic content is increased.

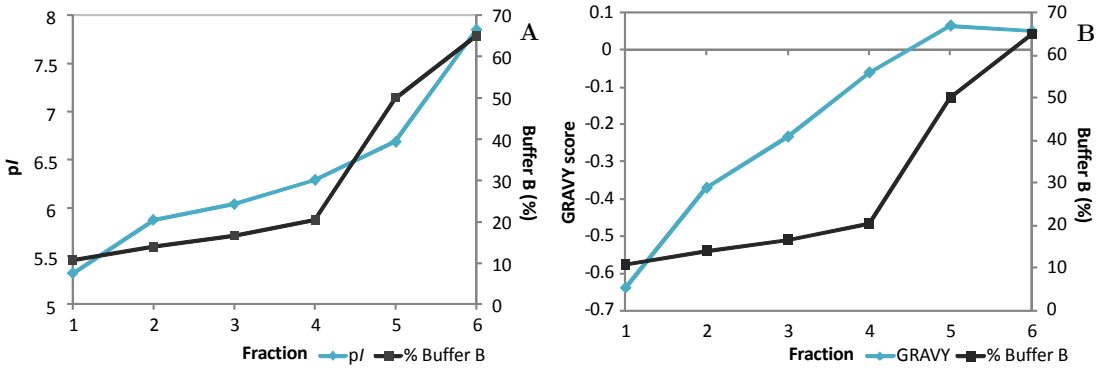

Figure 3.2. Physiochemical properties of peptides from six fractions in a high-pH low-pH reversed phase experiment

Average pI (a) and GRAVY score (b) of peptides from individual fractions

A high pH RP-low pH RP separation is possible since peptides exhibit different properties at extreme pHs. Peptides are charged molecules comprised of ionisable basic and acidic functional groups, which means the mobile phase pH will have a pronounced effect of their retention behaviour. It has been shown that acidic peptides ($pI < 5.5$) are more strongly retained at pH 2.6 (pH of the second dimension), compared to basic peptides ($pI > 7.5$), which are more strongly retained under pH 10 conditions. This was also observed in experiments here, shown in Figure 3.2(a).

It has been shown that non-repeatable identifications, corresponding to a peptide/protein that has only been identified in one technical replicate, are often false

positive identifications (Gilar *et al.* 2009). Here, a comparison was made between peptide identifications made using the 1D and 2D (6 fractions) approaches after a replication filter was applied (Figure 3.3). As expected, a significant increase in peptide identifications was observed using the 2D approach; with up to a three-fold increase for propane and succinate and five-fold for the methane condition.

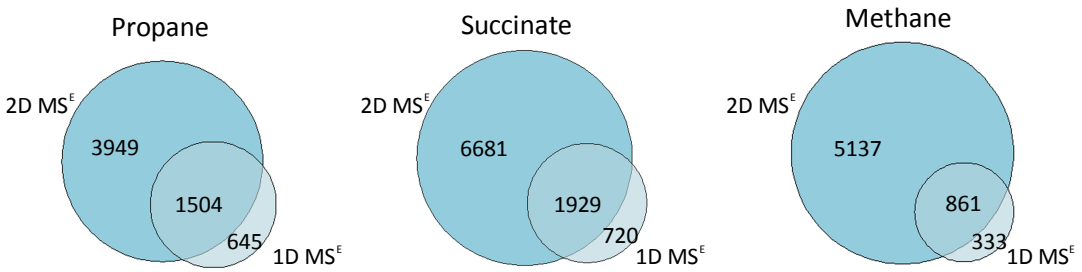

Figure 3.3. Comparison of peptide identifications between 1D RP-LC MS^E and 2D RP-RP-LC MS^E

Peptides identified using the 1D approach should also be present when using the 2D approach. In these experiments, 72% of peptides identified by the 1D approach were also identified in the 2D approach.

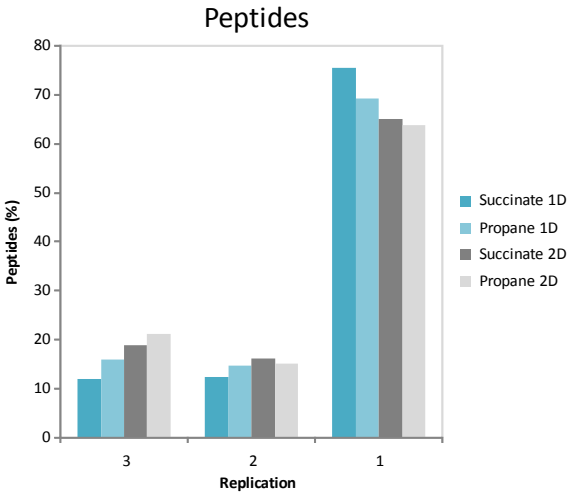

Figure 3.4. Peptide replication rate in technical replicates

Figure 3.4 highlights the importance of carrying out technical replicate analyses. The proportion of peptides that are replicated in one, two or three technical analyses are shown for selected substrates analysed using the 1D and 2D methods. Approximately 70% of peptides are only ever identified in one technical replicate using the 1D approach. This is slightly improved with the 2D approach with 65% of peptides only seen in one replicate.

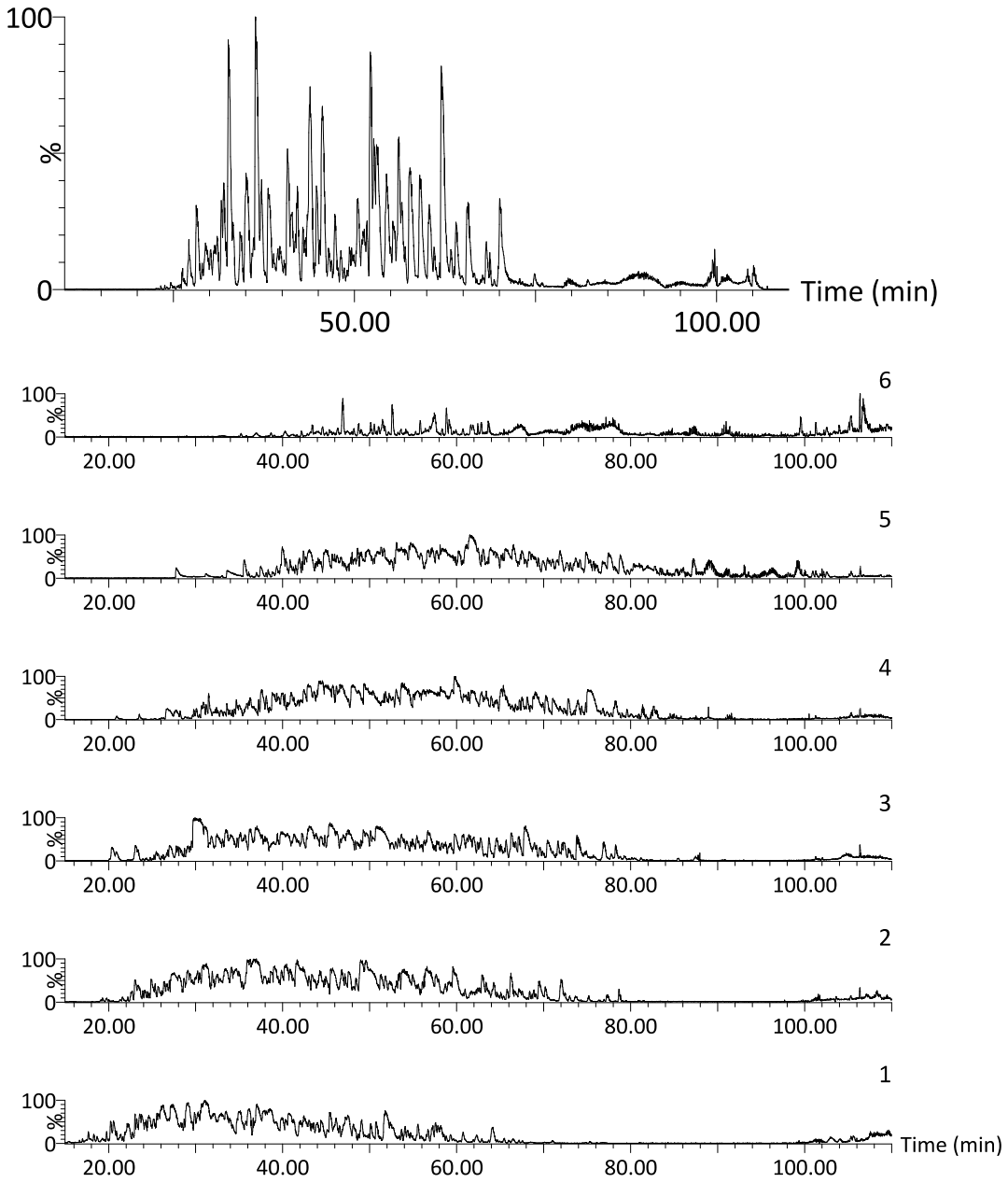

Figure 3.5. 1D and 2D chromatograms of complex bacterial extracts

Upper shows chromatogram of 1D RP-LC experiment. *Lower* shows chromatograms from a 6 fraction method for a 2D RP-RP-LC experiment, labelled 1-6 from lowest→highest organic content.

In the 2D approach, the amount of sample injected was approximately five-fold greater than the required amount for 1D. This ensures that fractions from the first dimension contain an appropriate amount of sample that can then be resolved by the second dimension separation without overloading the column. Multi-dimensional chromatographic approaches, in part, are capable of identifying lower abundant proteins due to the fact that total sample loadings are higher. As a consequence, high

abundance peptides may be over loaded. This was shown by peak broadening observed in certain selected fractions (Figure 3.5). Mass error distributions for precursors and fragment ions were plotted for each approach in the analysis of succinate-grown cells to show possible adverse effects due to overloading (Figure 3.6). The accuracy of mass measurement does not appear to be perturbed by the increased loading amounts required for 2D analysis.

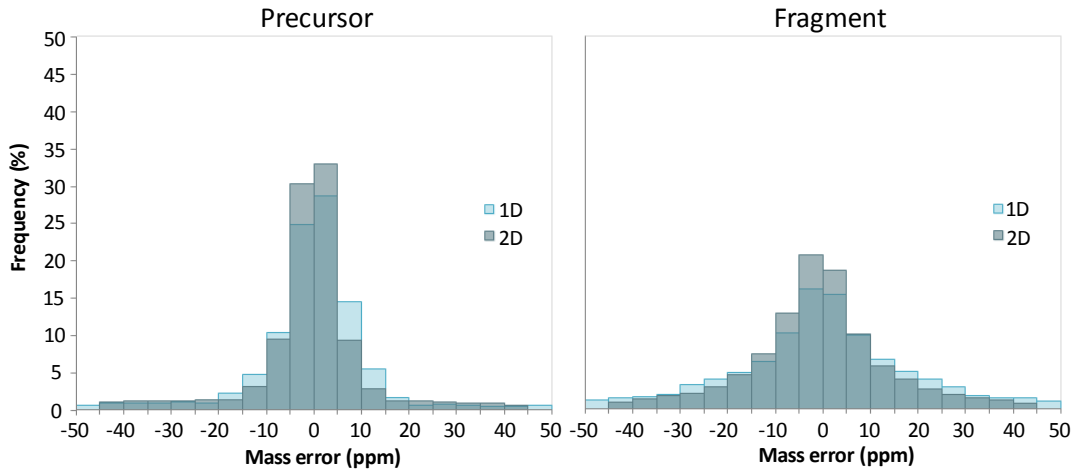

Figure 3.6. Precursor and fragment mass accuracy

In addition to accurate mass, highly reproducible peptide retention times are essential for performing comparative analyses between two data sets, which require that features are aligned according to their accurate mass and retention time. Peptide retention times for these data were shown to be highly reproducible, with 90% of peptides shown to have a coefficient of variation less than 0.1 for both 1D-LC and 2D-LC separations across technical replicates (Figure 3.7).

It has been shown that retention times of peptides commonly identified in 1D and 2D experiments have a strong correlation (in an experimental set up where a high pH RP separation was carried out offline and individual fractions were analysed at low pH RP) (Gilar *et al.* 2009). A good correlation suggests that peptides are confidently identified and few false positive assignments are made. Figure 3.8 shows the correlation observed between retention times of peptides (from at least 2 technical replicates) identified in succinate samples identified by means of both the 1D and 2D approach.

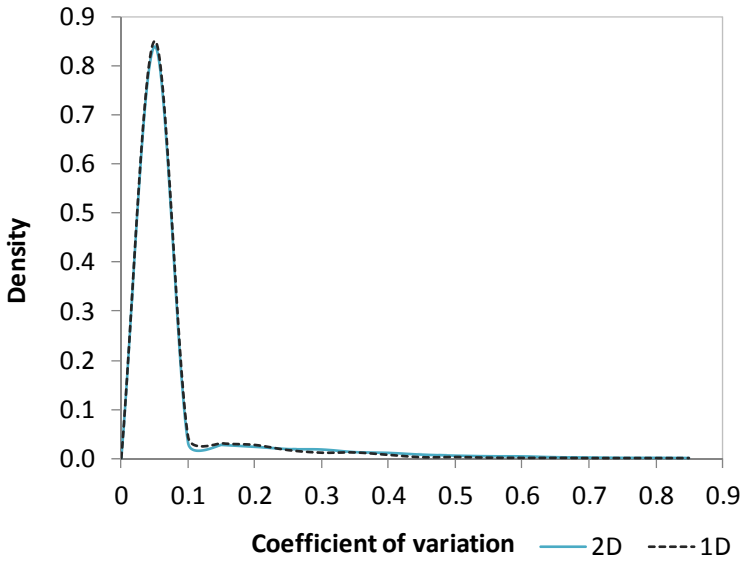

Figure 3.7. Retention time variability for 1D and 2D approaches

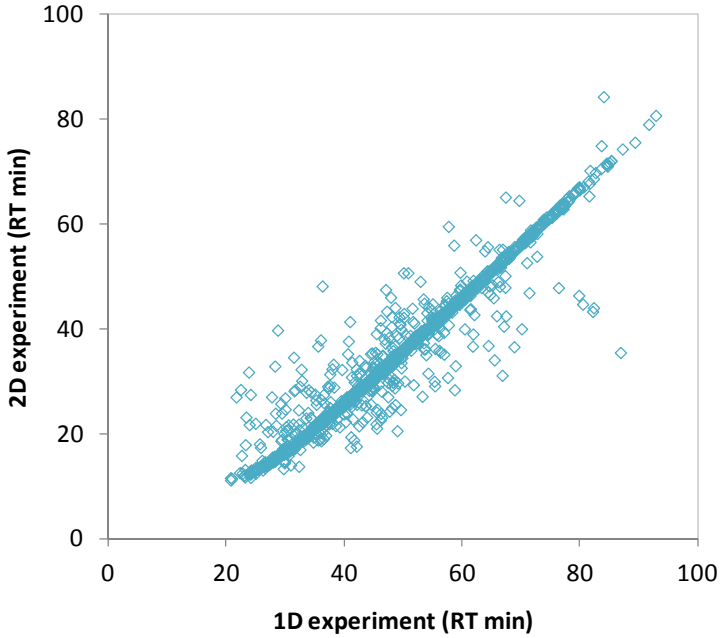

Figure 3.8. Correlation between experimental RTs of peptides identified in 1D and 2D experiments

The majority of peptides have similar retention times, which provides a strong indication that those peptides are confidently assigned. Despite the fact that these peptides have been identified in at least two replicates and by two different techniques, there are still a number of peptide RTs that do not agree. This could correspond to false positive assignments. Peptide retention prediction algorithms exist within database search strategies to improve the confidence of peptide assignment. Where a 1D and 2D experiment has been conducted on the same sample,

it could be suggested that correlation of RT can be applied to improve the confidence of peptide assignment.

Figure 3.5 shows the chromatograms obtained for a 1D RP-LC separation and for each of the six RP-LC separations in a 2D experiment. This provides an indication of the separation space or capacity that is being practically used. For a single RP-LC run, 1D or 2D, the majority of peptides elute over a 60 min period. In a multi-dimensional experiment, it is desirable to have an even distribution of peaks between fractions. Data obtained from 2D experiments are usually processed for each individual fraction, which are then merged prior to database searching. Merging processed data prior to database searching integrates signals for a peptide that may occur in more than one fraction, which is desirable for a more accurate measurement of peptide concentration.

Table 3.4. Detections made succinate-grown extracts analysed by 1D RP-LC MS^E and 2D RP-RP-LC MS^E (6 fractions) for one technical replicate

LC	Total		Peptides	
	Number of detected ions	Intensity ($\times 10^8$)	Number of detection ions	Proportion of detected ion intensity
1D RP-LC	35,982	6.00	3858	40%
2D RP-RP-LC (Total)	70,237	8.68	18,049	54%
Fraction 1	10,444	1.24	2,130	48%
Fraction 2	13,485	1.64	3,481	61%
Fraction 3	15,141	1.88	3,988	54%
Fraction 4	14,291	1.66	4,275	61%
Fraction 5	18,239	2.03	5,182	61%
Fraction 6	3,768	0.24	913	61%

The number of ions detected in 1D and 2D experiments are shown in Table 3.4. Fractions were processed and searched against a database individually to provide an indication of the first dimensional separation capability. The Table provides the number of detections, which refers to deconvoluted charge-state reduced ions, together with their summed intensity. After database searching, a proportion of these detections are assigned as peptides. This is shown in the Table as the proportion identified divided by the total detected ion intensity. For the 2D experiment, the number of ions detected is shown for the overall experiment and for each individual fraction to provide an indication of the distribution of peptides. The number of

detected ions was not significantly enriched in any particular fraction. These ranged from 14.8 – 25% of the total number of detected ions in each of the first five fractions. The final fraction contained a higher organic content to serve the purpose of a wash step and therefore a lower number of detected ions was expected. From these data, one can assess the extent of peptide overlap between fractions by summing the peptides detected from individual fractions and comparing this to the merged peptide identifications. Here, the sum of peptide identifications from individually database searched fractions was 19,969. When the data was merged prior to database searching, 18,049 peptides were assigned, indicating that only 1,920 peptides were observed in more than one fraction.

Dynamic range considerations in complex protein mixtures can limit protein identifications in a number of ways. These include ion suppression of low abundance peptides, co-eluting peptides from LC separation and the duty cycle, and the detection dynamic range and resolving power of the mass spectrometer. The data-independent mode of acquisition utilised in these experiments has the advantage of a high duty cycle. An increase in peptide and protein identifications was obtained using the 2D approach, compared to the 1D method. This may be explained as being due to the simplification in the complex peptide mixture entering the mass spectrometer. With the gradients and samples analysed here, a significantly larger number of peptides elute per second during a 1D RP-LC gradient compared with those observed when a second dimension gradient has been used in a 2D experiment. The precursor intensities recorded during an analytical gradient that were attributed to protein identifications were compared using succinate-grown extracts analysed by 1D and 2D. 40% of the total recorded intensity was assigned to peptides using the 1D approach and 54% in the 2D approach. A larger proportion of peptide masses are assigned to proteins by the 2D approach, however, in both approaches, there are a number of peptide masses that are not assigned. It is evident that issues such as ion suppression and co-eluting peptides are reduced with the adoption of a fractionation-based strategy.

Several factors may affect the ability to assign ion detections to peptides. One such problem can occur at the sample preparation/storage stage. Protein extracts are treated with protease inhibitors to stop ongoing digestion of peptides prior to the addition of trypsin. If proteolytic enzyme activity continues to take place, or protein degradation occurs, this may lead to the formation of non-tryptic peptides. These

cannot easily be assigned during the database search process. The quality of the fragmentation data can affect how well a peptide is assigned. This is dependent on the abundance of the precursor ion, charge state, residue content and the sequence of the peptide. A factor more specific to this study may affect the ability to assign ion detections. This is associated with the data-independent nature of the MS acquisition. The raw data are complex and require alignment of fragment ions to their precursor ions. After the data is processed, a fragment ion spectrum for a given precursor ion typically contains fragments formed by the peptide and, to a lesser extent, fragments from other peptides eluting at a similar time. This is particularly a problem for co-eluting peptides that exhibit the same retention profile. This results in a complex fragment ion spectrum that may contain a mixture of ions from more than peptide. This further complicates the search algorithm. The database search stage may also limit the number of peptide assignments, if not set-up correctly by the user. If the amino acid sequence for a protein in the database is incorrect, even by a single amino acid, then the resulting peptide will not be matched correctly by the search algorithm. Post-translational modifications that are not defined by the user during the database search process can also hinder the matching process.

Fragmentation data that were not assigned to peptides in the data sets collected here were interrogated further. The manual interpretation of all unassigned fragment ion spectra, however, is an unfeasible task. A small number of abundant ions, present in some substrates, were not assigned to a protein. The product ion spectra of some of these peptides contained a sufficient number of fragment ions to enable manual interpretation. Figure 3.9 shows the fragment ion spectra from a detected ion that was not assigned to a peptide during the database search. Selected b and y ions have been annotated. The precursor ion of 1387.82 $(M+H)^+$ with RT 61 min, was detected in all technical replicates for the succinate- and propane-extracts. The fragment ion spectrum for this precursor was manually interpreted and assigned the sequence VPSTI/LTI/LAPYVVK. An unambiguous portion of this sequence ("APYVVK") was searched against the *M. silvestris* database, and matched to a single tryptic peptide with the sequence VGSTLTIAPYVVK, originating from the protein Msil_0471 PQQ dependent dehydrogenase methanol ethanol family. Interestingly, the ion 1347.80 $(M+H)^+$, RT 56 min, relating to the peptide VGSTLTIAPYVVK was also present in the search results and was assigned correctly. This is also shown in Figure 3.9. The manually sequenced peptide with a proposed

glycine→proline substitution was searched against a non-redundant database using BLAST (Basic Local Alignment Search Tool), to search for any known methanol dehydrogenases with this sequence. No known sequences for methanol dehydrogenase contained this substitution. There are also currently no known post-translational modifications that would account for a +40 difference on a glycine residue.

It had previously been suggested that elevated pH may induce deamidation of asparagines and glutamic acids, converting them into the corresponding aspartic and glutatmic acids (Gilar *et al.* 2005b). This was not observed in the analyses performed here, with an average of 0.21 and 0.16 deamidation modifications per peptide observed for the 1D and 2D analyses.

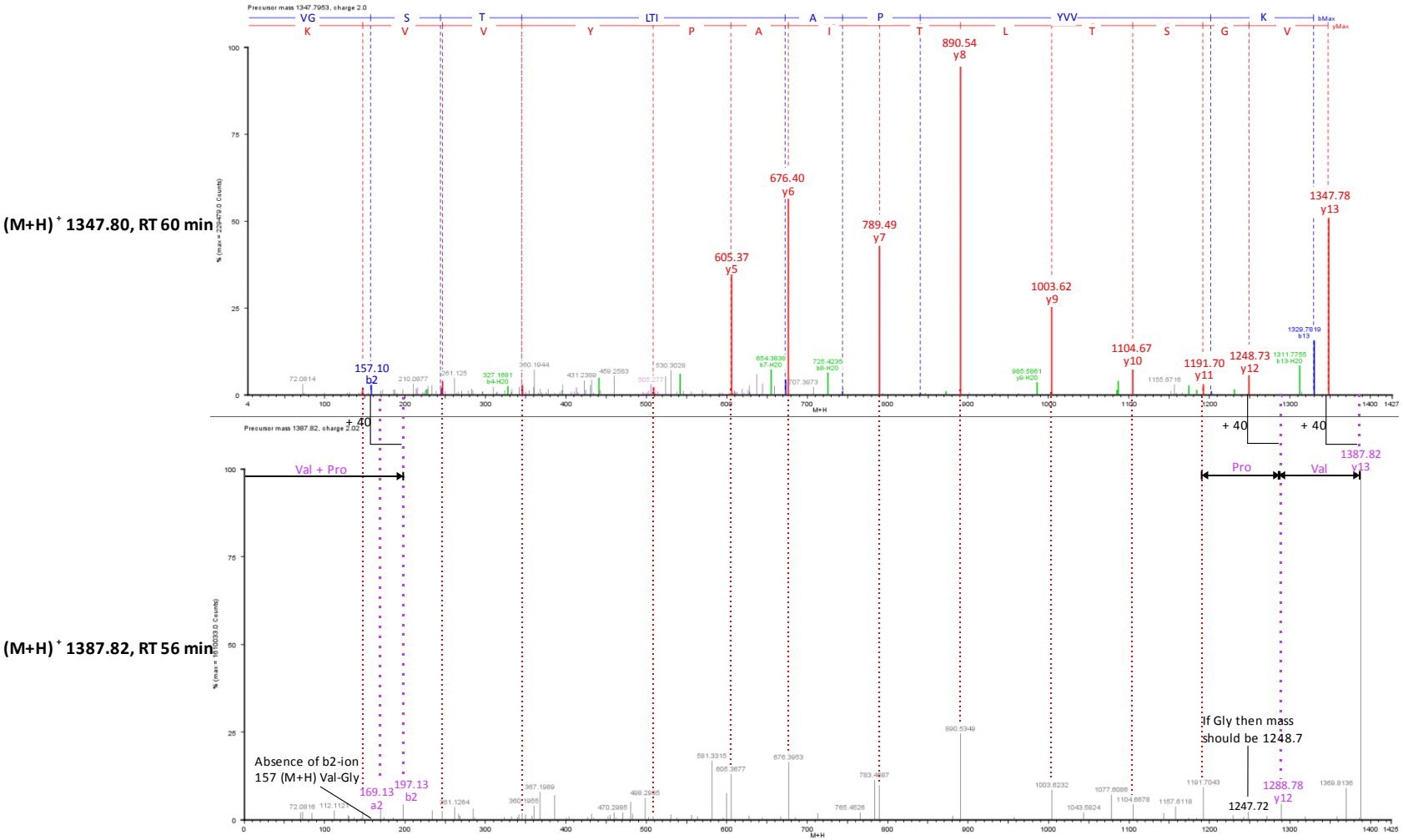

Figure 3.9. Database matched and unassigned annotated fragment ion spectra

3.4.3 Protein identifications

The number of proteins identified using the 1D approach was 327, 248 and 178 for succinate-, propane- and methane-grown extracts, respectively, with an approximate three-fold increase in identifications observed for each propane and succinate condition and a five-fold increase for the methane condition when the 2D approach was employed. Fewer protein identifications were observed in the methane-grown protein sample. This could be attributed to a small number of abundant proteins present in the sample. The label-free method estimated that, in the 1D analyses, five proteins constituted 50% of the total protein abundance detected in the methane-grown sample. The five most abundant proteins constituted 33 and 27% for the propane- and succinate-grown samples.

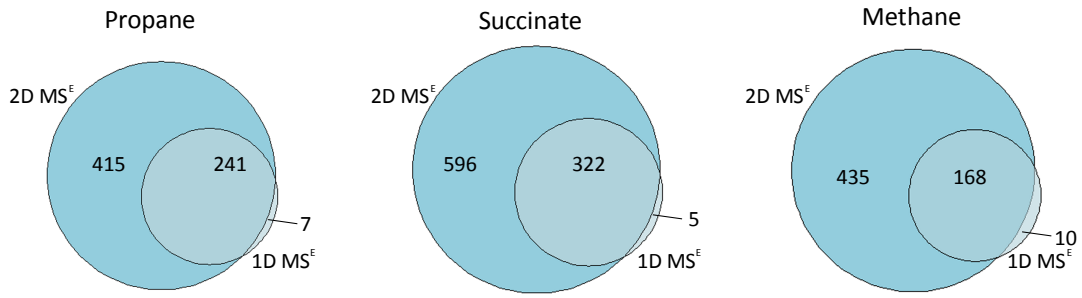

Figure 3.10. Proteins identified in two or more technical replicates

97% of proteins identified by the 1D approach were also identified in the corresponding 2D experiment (Figure 3.10). This is similar to observations elsewhere (Gilar *et al.* 2009) (Taylor *et al.* 2009). The majority of the proteins that were not identified using the 2D method were assigned low scores by the search algorithm or were random entries.

Figure 3.11 shows the proportion of proteins found in one, two and three technical replicates. The replication rate of proteins is greater than those of peptides. This is similar to observations made elsewhere (Nagaraj *et al.* 2012). Approximately 60% of identifications were made in two or more replicate analyses for the 1D approach, improved to 70% in the 2D approach.

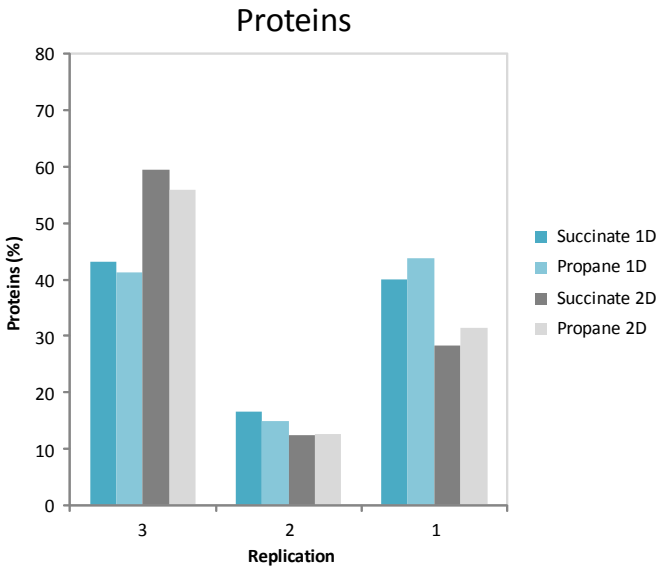

Figure 3.11. Protein replication rate in three technical replicates

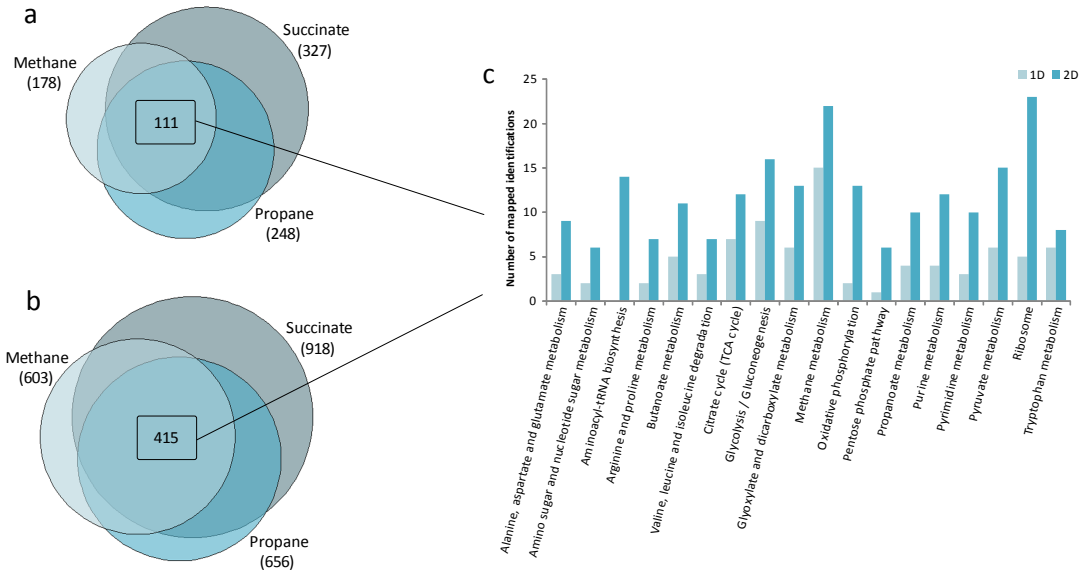

Figure 3.12. Overlap of proteins identified in ≥2 replicates between all growth conditions

Data shown for (a) 1D and (b) 2D methods. (c) A subset of pathways populated by proteins common to propane, succinate and methane-grown cells identified by either the 1D or 2D approach.

Figures 3.12(a-b) illustrate the overlap of proteins confidently identified across all three studied substrates for both the 1D and 2D approaches (418 and 1043 non-redundant identifications). 111 and 415 proteins were found to be common between succinate, propane and methane for the 1D and 2D approach, respectively, which is

27% and 40% of the total number of non-redundant proteins. The vast majority of proteins shared by all conditions from the 1D approach were also found using the 2D approach. It may be suggested that proteins present in all growth conditions are part of a core set of proteins required by the bacteria, irrespective of their environment. These proteins were populated onto pathways present in the KEGG database, a subset of which is represented in Figure 3.12(c). The majority of mapped proteins were assigned metabolic functions with a more comprehensive coverage of populated pathways observed when the 2D experimental data were used. This is an indication that the additional proteins identified using the 2D approach do have relevant metabolic functions.

3.4.4 Absolute and relative quantitative measurements

A pre-digested mixture of four proteins; including glycogen phosphorylase *b*, enolase, bovine serum albumin and alcohol dehydrogenase, was spiked into various quantities of digested methane extracts prior to separation by 1D RP-LC. The mixture was diluted with the peptide sample to ensure that peptide corresponding to 25 fmol alcohol dehydrogenase were loaded onto the column for each injection. Using this protein as an internal standard, other proteins, according to the manufacturers specifications, should be present at 8.5, 25, and 96.75 fmol for glycogen phosphorylase *b*, enolase and BSA, respectively. Figure 3.13 shows the expected quantities for the four proteins determined by the manufacturer along with the estimated quantities of these proteins that were spiked into 130, 170 and 240 ng of soluble protein extract from methane-grown bacteria. These proteins were accurately measured using the 1D approach (average error of 5%) and, within the experimental dynamic range, background complexity in the sample has not influenced the quantification data.

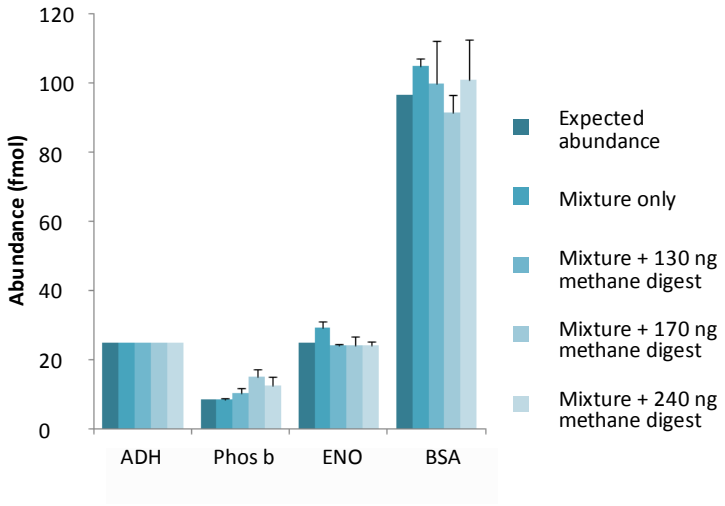

Figure 3.13. Estimated quantities of standard proteins spiked into various loadings of digested methane-grown extract

Alcohol dehydrogenase was used as an internal standard (Phos b – glycogen phosphorylase b, ADH – alcohol dehydrogenase, ENO – enolase, BSA – bovine serum albumin).

The reproducibility of protein quantification between technical replicates of the two approaches was assessed by calculating coefficient of variation (CV). The variation of quantification of proteins across technical replicates was calculated for 1D- and 2D-analysis on propane-grown extracts, shown in Figure 3.14. 90% of quantified proteins display CV values of 0.5 or lower for both approaches, with the 1D approach having a lower distribution. The distributions of precursor error do not appear to be significantly different between the two approaches. This highlights the potential utility of 2D RP-RP-LC for reproducible label-free quantitative analysis.

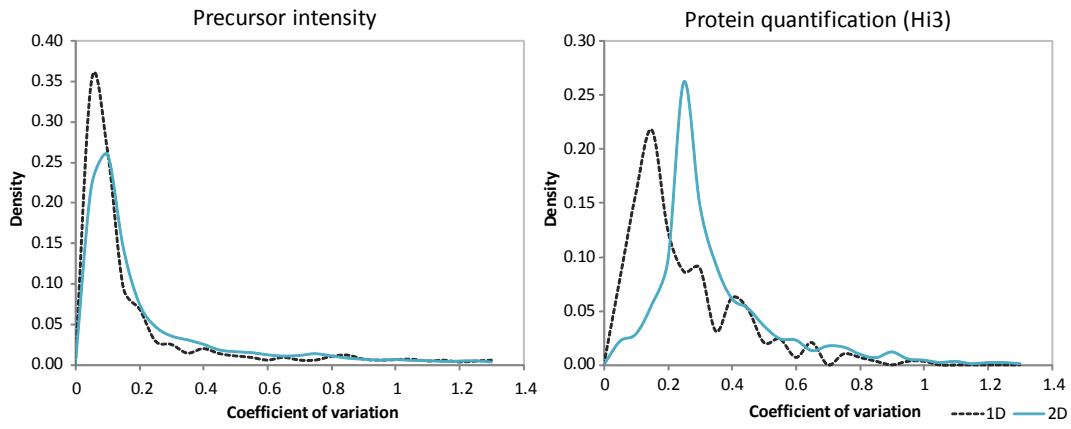

Figure 3.14. Precursor intensity and protein quantification technical variation

Figure 3.15 shows a plot of abundance of the 656 proteins identified using the 2D approach from propane-grown extracts. This highlights identifications confirmed by both 1D and 2D approaches and those unique to each technique. It is evident that lower abundance proteins are predominantly identified using the 2D approach whereas the 1D technique is limited to identification of more abundant proteins. The quantitative dynamic range observed using the 1D approach was approximately three orders of magnitude between the most and least abundant detected protein in the analysis of propane and succinate extracts. This is improved to almost four orders with the use of 2D RP-RP-LC.

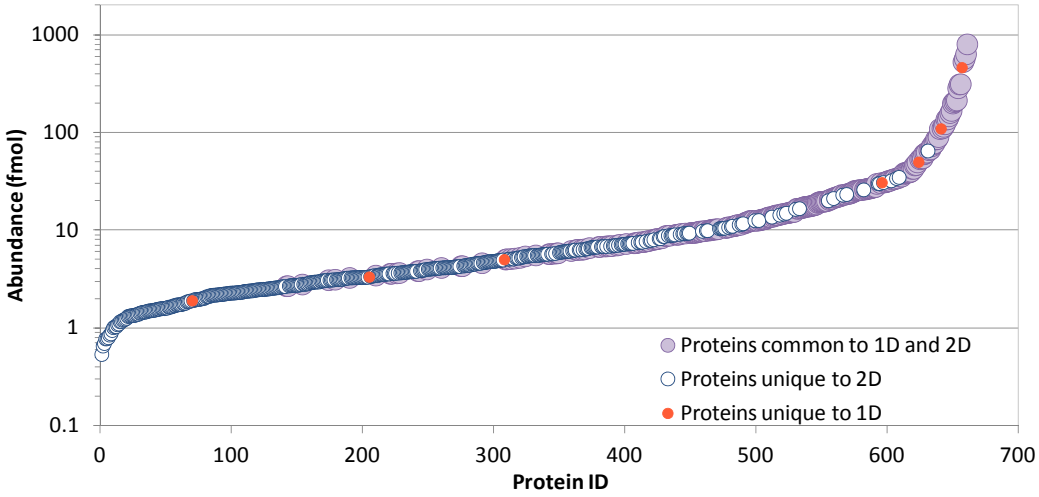

Figure 3.15. A plot of protein abundances from propane-grown cells determined using the 2D approach

Figure 3.16 shows the relative protein expression for the regulated proteins (common to propane and succinate substrates). Relative abundances were plotted as the natural log of the ratio as reported from the Waters Expression software, with standard deviation values calculated from the technical replicates. 28 (106) proteins were indicated to be down-regulated in the succinate sample compared to propane, and 25 (87) were indicated to be up-regulated. The remaining 97 (314) showed no distinct differential expression for either the 1D and 2D approach, respectively. All 150 ratios obtained using the 1D approach were also obtained using the 2D approach. Of the 150 ratios measured from the 1D approach, 77% were in agreement with those values obtained using 2D in terms of direction of regulation (i.e. up- or down-regulated).

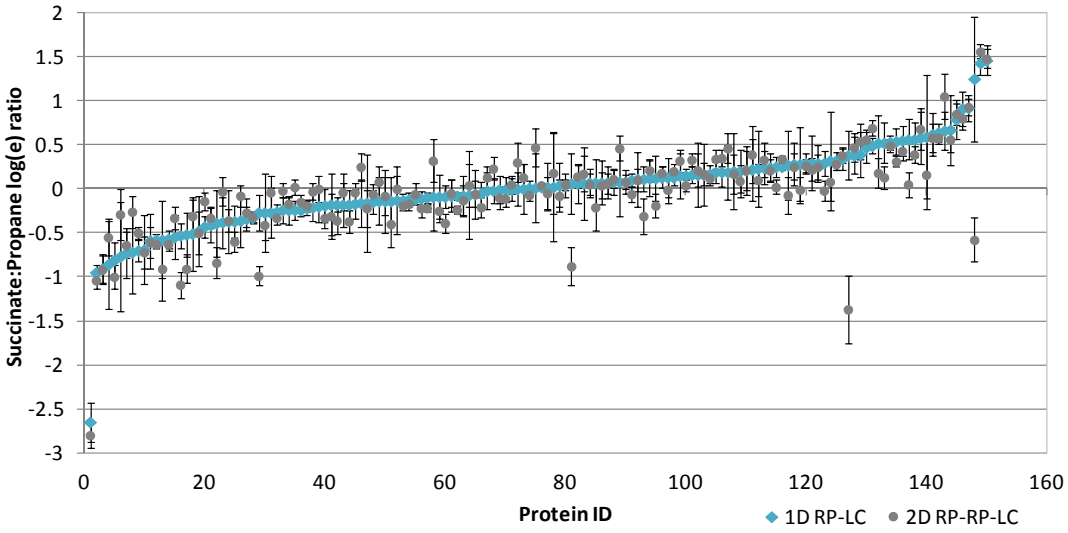

Figure 3.16. Relative changes between succinate:propane expressed as log(e) ratio for 1D and 2D approaches

A scatter plot comparing the regulation values calculated from the two methods showed a reasonable correlation, with a R^2 value of 0.66, and three distinct outliers. When these outliers were removed, the correlation improved significantly, with a R^2 value of 0.82 obtained, shown in Figure 3.17. The outliers may be largely attributed to results obtained from an anomalous technical replicate. An example is shown in Figure 3.18, which shows the normalised average peptide intensities used for relative measurements between succinate and propane using both the 1D and 2D approaches.

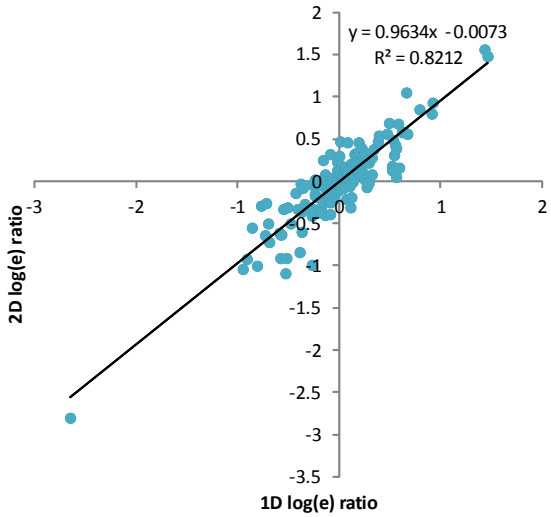

Figure 3.17. Comparison of methods for estimating relative changes in protein expression

The protein YP 002360972 ribosomal protein L19 was shown to be up-regulated in succinate using the 1D approach (log(e) ratio 1.25) and down-regulated in succinate using the 2D approach (log(e) ratio -0.58). This disagreement can be explained by the peptide MNIIAELEAEQAAK, which was not measured effectively in the succinate condition by the 1D approach. This is shown in Figure 3.18.

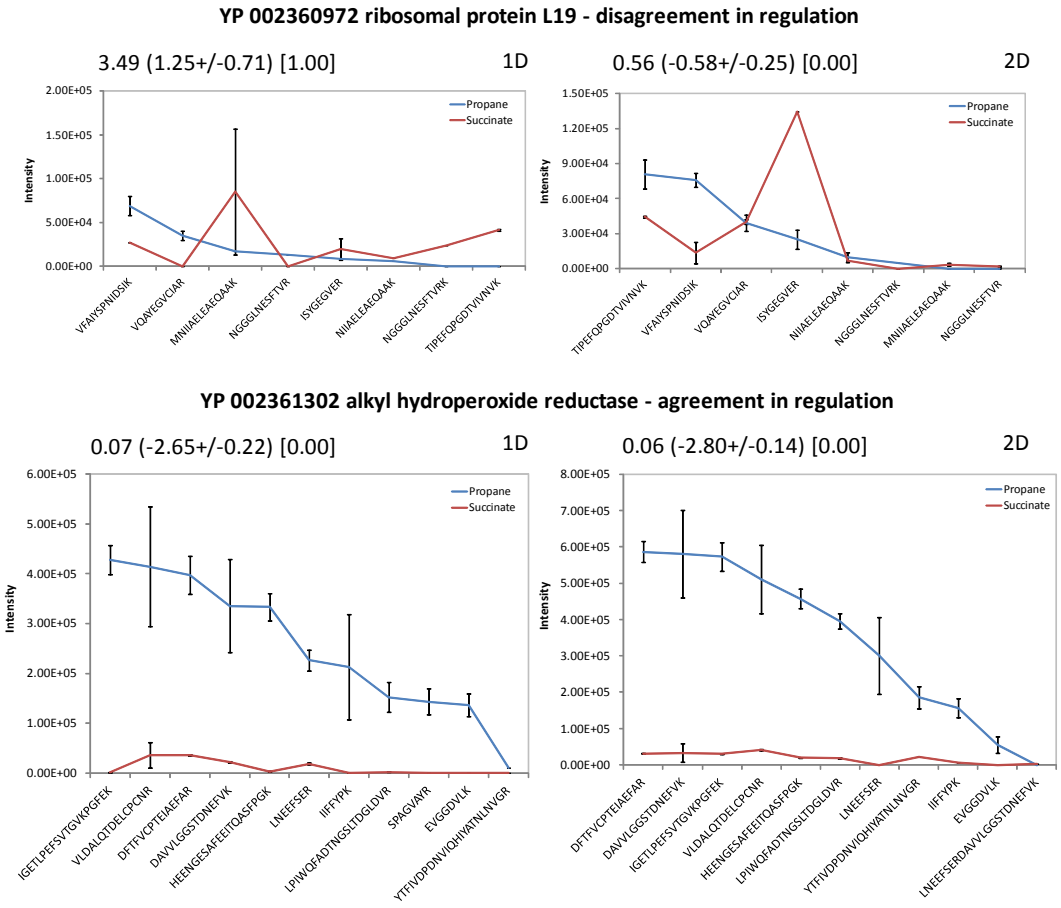

Figure 3.18. Peptide intensity profiles for differentially expressed proteins

Differentially expressed proteins are expressed by $a(b$ +/- $c)$ $[d]$, where a = succinate:propane, b = log(e) succinate:propane, c = log(e) standard deviation and d = probability.

Disagreements due to an incorrect measurement in one technical replicate were not exclusive to any approach, so one cannot speculate whether the 1D or the 2D approach is more reliable. Figure 3.18 shows an example of a protein with similar estimated regulation. The protein YP 002361302 alkyl hydroperoxide reductase was shown to be down-regulated in succinate by the 1D approach (log(e) ratio -2.65) and 2D approach (log(e) ratio -2.80). It is interesting to note that the relative peptide intensity values are not the same for both growth conditions, or even for both methods for these two proteins. For the purposes of making relative expression

measurements, the total intensity for all detected peptides are taken into account. It is therefore possible that the order of peptide intensities should not be an issue. The relative quantification of a protein should not rely on a single peptide since their relative intensity compared to the rest of the protein may not the same in different mixtures.

The estimated absolute amount (Hi3) reported for the ribosomal protein was used to calculate a relative expression value between succinate and propane. This was then used to study the effect of the imprecise measurement of a single peptide, which resulted in conflicting measurements using the Expression method. For the 1D approach, this protein is estimated to be up-regulated to a lesser extent in succinate (log(e) ratio 0.48) compared to the value estimated using the Expression method (log(e) ratio 1.25) and un-regulated in succinate by the 2D approach (log(e) ratio -0.17), which provides results comparable to the Expression method (log(e) ratio -0.58). This shows that the anomalous measurement of a single peptide can affect the estimation of quantification using both the Hi3 and Expression methods. The use of probability and standard deviations are commonplace for assessing the quality of regulation values, and are typically used to filter out values of low confidence. The reporting of absolute quantification data could also benefit from additional information to provide an indication on reproducibility.

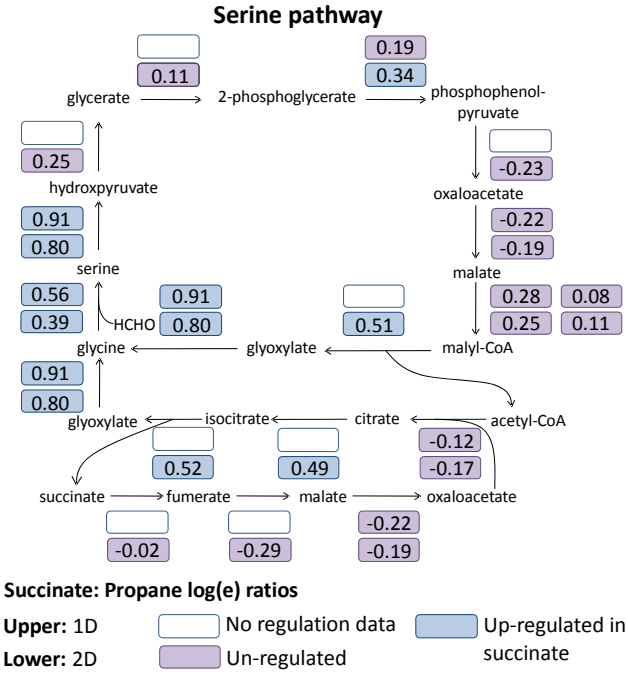

Figure 3.19. The serine and glyoxylate pathway

Annotated with Expression log(e) ratio succinate:propane for the enzymes involved ascertained using 1D and 2D methods.

Figure 3.19 illustrates the regulation of the enzymes involved in the serine and glyoxylate pathway between succinate and propane conditions calculated using either the 1D or 2D approach. There is excellent agreement in the direction and values of regulation observed when comparing 1D- and 2D-obtained values for the enzymes in the pathway. This also shows that the 2D method was able to provide more relative values, as in the 1D approach some of these enzymes were only identified in one substrate or not identified at all.

3.4.5 Effect of increasing the number of first dimensional fractions in the study of methane-grown extracts

The results initially obtained from 2D analyses were obtained from a 2D approach which included six fractionation steps. Additional experiments were carried out on the methane-grown extract investigating the effect of increasing the number of discontinuous fractions from six to eleven. The experimental considerations relating to all three types of LC separation and the results obtained for each are outlined in Table 3.5.

Table 3.5. A comparison of the experimental requirements for each of the approaches in the study of methane extracts

	1D RP-LC	2D RP-RP-LC 6 fractions	2D RP-RP-LC 11 fractions
Protein loading (for one technical replicate) (μg)	0.5	2.5	5
Instrument time (LC and MS acquisition) (h)	120 min × 3 replicates 6	140 min × 6 fractions × 3 replicates 42	140 min × 11 fractions × 3 replicates 77
Data file size (Gb)	6 Gb × 3 replicates 18	7 Gb × 6 fractions × 3 replicates 126	7 Gb × 11 fractions × 3 replicates 231
Dynamic range	10^3 - 10^4	10^3 - 10^4	10^4
Number of proteins identified (≥2 replicates)	178	603	777
Average number of matched peptides per protein	12	22	25
Average sequence coverage (%)	33	36	42

The amount of sample required for each LC method takes into account that the final analytical column has an optimised loading of approximately 0.5 - 1 μg of protein digest during a gradient. Increased sample loading can be used if more separation steps are employed and would improve the dynamic range of the experiment. The values outlined in Table 3.5 relate to the loadings associated with one technical replicate; if three replicates were employed, in order to obtain more confident results, then the minimum amount of material needed for an 11 fraction method would be 15 μg. Sample restriction can be an important consideration when performing proteomics experiments, since it can be a challenge to generate suitable amounts from some biological systems. If the amount of sample is restricted, technical replicates and optimisation may not be achievable for some 2D experiments.

A significant increase in confident protein identifications was made when comparing the 1D method (178) with the 2D method using six fractions (603). The extension from six to eleven fractions affords an additional 174 protein identifications. Of the 603 identifications made using the six-fraction method, 51 of these were not confirmed by the eleven-fraction method (Figure 3.20). The identified peptides relating to these proteins were investigated further with respect to properties such as retention time and hydrophobicity. No obvious relationship was observed which

would explain the absence of these peptides in the eleven-fraction method. It may be suggested that these differences could result from problems with peptides eluting from the trapping column during the dilution step.

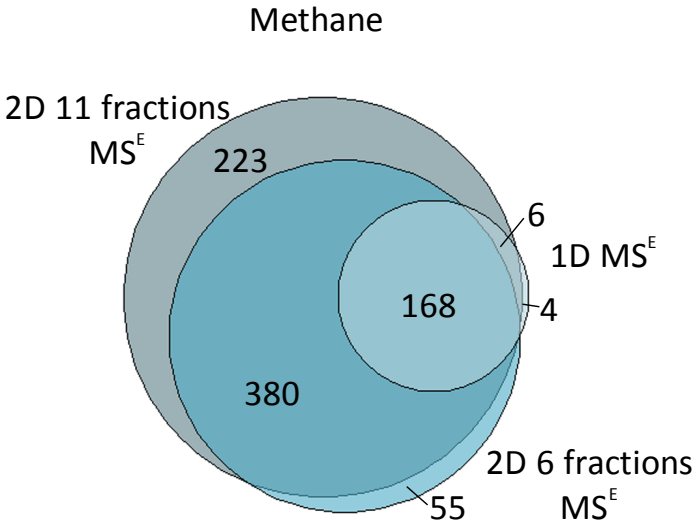

Figure 3.20. Protein identifications in methane-grown extracts analysed by various approaches

The average number of peptides matched per protein assignment was 12 for the 1D approach, which almost doubled to 22 in the 2D approach with 6 fractions, and further improved to 25 when 11 fractions were employed. The average sequence coverage obtained using 1D RP-LC was 33%. The 2D approach with 6 fractions produced on average sequence coverage of 36% and the 11 fractions experiment an average of 42%. These observations suggest that the extra proteins identified through 2D separation contain proteins identified by the 1D approach as well as additional proteins. The increased number of peptides and higher sequence coverage confers additional confidence in the identifications obtained.

Experimental and instrumental times are relevant considerations when selecting a proteomics strategy, and can play a key factor for high-throughput studies. Including the time required for the analysis of three technical replicates, the one-dimensional approach required 6 h LC and MS time for one sample, which increased to 77 h for the 11 fractions method, using the columns and gradient conditions used in these experiments. Increasing amounts of information obtained by multi-dimensional techniques can concurrently be associated with increasing data file sizes. The complexity of the sample, the mass spectrometer scan rate, they type of data acquisition and the period of data acquisition are factors that effect the size of data

files. For the samples and LC-MS settings used here, 231 Gb of data was collected for three technical replicates using the 11 fraction method. This can place significant pressure on data storage and retrieval requirements.

3.5 Conclusions

An ideal proteomics experiment would provide comprehensive proteome coverage and reproducible quantification information with minimal experimental time. Pre-fractionation is one method of reducing sample complexity prior to MS analysis; typically allowing larger sample loading which in turn may lead to a higher number of protein identifications. A larger number of protein identifications, higher sequence coverage and the identification of lower abundant proteins were achieved when 2D experiments were employed. These experiments included six fractionation steps. Results showed that increasing the experimental time significantly to include more fractions than six has a diminishing return in terms of output. The 1D approach is advantageous in terms of sample requirement, instrumental time requirements and downstream processing time. When optimising a 2D experiment, if sample availability were low, a combination of 1D and 2D methods would be ideal, with the 1D experiment providing quantitative information, which can then be used to optimise sample loading in 2D experiments. Good agreement was observed between 1D and 2D experiments in terms of identifications and relative and absolute quantification for each condition. This suggests that label-free quantitative analysis does not suffer when utilising a multi-dimensional separation approach.

Label-free quantification within an online 2D RP-RP-LC approach has been assessed and compared to that obtained from 1D RP-LC experiments. Good qualitative and quantitative agreement between data sets was obtained. The 2D RP-RP-LC label-free quantitative approach also has significant potential in wider qualitative and quantitative proteomic studies where simplification of the complexity of the sample leads to improved dynamic range and data quality.

3.6 References

Anderson, N. L. and Anderson, N. G. (2002). The Human Plasma Proteome. *Mol. Cell. Proteomics.* **1**, 845-867.

Boersema, P., Mohammed, S. and Heck, A. (2008). Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal. Bioanal. Chem.* **391**, 151-159.

Domon, B. and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotech.* **28**, 710-721.

Dowell, J. A., Frost, D. C., Zhang, J. and Li, L. (2008). Comparison of Two-Dimensional Fractionation Techniques for Shotgun Proteomics. *Anal. Chem.* **80**, 6715-6723.

Gilar, M., Olivova, P., Chakraborty, A. B., Jaworski, A., Geromanos, S. J. and Gebler, J. C. (2009). Comparison of 1-D and 2-D LC MS/MS methods for proteomic analysis of human serum. *Electrophoresis.* **30**, 1157-1167.

Gilar, M., Olivova, P., Daly, A. E. and Gebler, J. C. (2005a). Orthogonality of Separation in Two-Dimensional Liquid Chromatography. *Anal. Chem.* **77**, 6426-6434.

Gilar, M., Olivova, P., Daly, A. E. and Gebler, J. C. (2005b). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J. Sep. Sci.* **28**, 1694-1703.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. and Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotech.* **17**, 676-682.

Malerod, H., Lundanes, E. and Greibrokk, T. (2010). Recent advances in on-line multidimensional liquid chromatography. *Anal. Methods.*110-122.

Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O. and Mann, M. (2012). System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics.* **11**.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007-4021.

Ross, P. L., Huang, Y. L. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. and

Pappin, D. J. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics.* **3**, 1154-1169.

Shen, Y., Zhang, R., Moore, R. J., Kim, J., Metz, T. O., Hixson, K. K., Zhao, R., Livesay, E. A., Udseth, H. R. and Smith, R. D. (2005). Automated 20 kpsi RPLC-MS and MS/MS with Chromatographic Peak Capacities of 1000-1500 and Capabilities in Proteomics and Metabolomics. *Anal. Chem.* **77**, 3090-3100.

Slebos, R. J. C., Brock, J. W. C., Winters, N. F., Stuart, S. R., Martinez, M. A., Li, M., Chambers, M. C., Zimmerman, L. J., Ham, A. J., Tabb, D. L. and Liebler, D. C. (2008). Evaluation of Strong Cation Exchange versus Isoelectric Focusing of Peptides for Multidimensional Liquid Chromatography-Tandem Mass Spectrometry. *J. Proteome Res.* **7**, 5286-5294.

Taylor, P., Nielsen, P. A., Trelle, M. B., Hørning, O. B., Andersen, M. B., Vorm, O., Moran, M. F. and Kislinger, T. (2009). Automated 2D Peptide Separation on a 1D Nano-LC-MS System. *J. Proteome Res.* **8**, 1610-1616.

Toll, H., Oberacher, H., Swart, R. and Huber, C. G. (2005). Separation, detection, and identification of peptides by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry at high and low pH. *J. Chromatogr. A.* **1079**, 274-286.

Yang, X., Levin, Y., Rahmoune, H., Ma, D., Schöffmann, S., Umrania, Y., Guest, P. C. and Bahn, S. (2011). Comprehensive two-dimensional liquid chromatography mass spectrometric profiling of the rat hippocampal proteome. *Proteomics.* **11**, 501-505.

Zhou, F., Cardoza, J. D., Ficarro, S. B., Adelmant, G. O., Lazaro, J.-B. and Marto, J. A. (2010). Online Nanoflow RP-RP-MS Reveals Dynamics of Multicomponent Ku Complex in Response to DNA Damage. *J. Proteome Res.* **9**, 6242-6255.

Chapter 4

Ion Mobility-Assisted Proteomic Approaches

4.1 Introduction

4.1.1 Ion mobility mass spectrometry for proteomic applications

Ion mobility spectrometry (IMS) coupled to mass spectrometry (MS) has been shown to provide a powerful tool for structural studies of proteins and protein complexes and a rapid separation technique. Advantages of mobility separation in mass spectrometry include reduction in spectral congestion and the ability to separate isobaric species by shape. As a separation tool, ion mobility mass spectrometry (IMMS) has been used to aid in the analysis of components within complex mixtures (Hilton *et al.* 2008).

Ion mobility separation offers significant potential advantages in the field of proteomics. Sample preparation methods such as depletion of abundant proteins or LC at the peptide or protein level increases the time scale of proteomics experiments, but are essential to provide a more comprehensive coverage of the proteome and achieve quantification at the lower limits of detection. IMMS experiments involving the analysis of peptide mixtures or complex proteome samples, to date, have predominantly been carried out on instruments developed by individual research laboratories. A number of studies utilising these home-built technologies have demonstrated the power of the technique by analysing samples by direct infusion with no prior sample depletion or LC separation (Valentine *et al.* 1998; Guevremont *et al.* 2000; Hoaglund-Hyzer and Clemmer 2001; Kurulugama *et al.* 2008; Valentine *et al.* 2009; Ibrahim *et al.* 2011). Whilst this process greatly reduces instrumental time, which is ideal for high throughput experiments, the individual proteome coverage is typically compromised and as a stand-alone technique it is only suited to the characterisation of abundant proteins. The millisecond timescale of mobility separation, however, can be incorporated into an LC-MS experiment timescale to increase information content without increasing the timescale of the experiment.

The ability to identify and quantify peptides and proteins from tissues and biological fluids is essential for discovery of potential biomarkers. It is, therefore, of little surprise that the majority of IMMS-based proteomic experiments have been carried out on complex eukaryotic samples. These include blood plasma (Venne *et al.* 2005; Valentine *et al.* 2006; Liu *et al.* 2007; Belov *et al.* 2008), urine (Moon *et al.* 2003), and lower order organisms (Taraszka *et al.* 2005a; Taraszka *et al.* 2005b). LC

was used in an off-line manner in conjunction with IMMS in these experiments to provide an improved coverage of the proteome (Liu *et al.* 2007). High-throughput proteomic studies have also been achieved with the use of short reversed-phase capillary LC separations coupled on-line to an IMS-MS platform. This demonstrated the benefits of using mobility to separate low abundance species from a complex background matrix (Baker *et al.* 2010).

4.1.2 Travelling wave ion mobility mass spectrometry for proteomic applications

A travelling-wave (T-wave) ion mobility separation approach has been incorporated into commercial quadrupole time-of-flight (TOF) instrumentation. A schematic is shown in Figure 4.1 (Giles *et al.* 2004; Pringle *et al.* 2007). The travelling-wave mobility cell comprises a series of electrodes arranged orthogonally to the ion transmission axis. Opposite phases of an RF voltage are applied to adjacent electrodes, creating pulses of T-waves along which the ions are propelled through a background gas. The drag due to the presence of gas causes ions to periodically slip over the waves, with ions of higher mobility slipping over less often than those of lower mobility and so exiting the device first (Riba-Garcia *et al.* 2008).

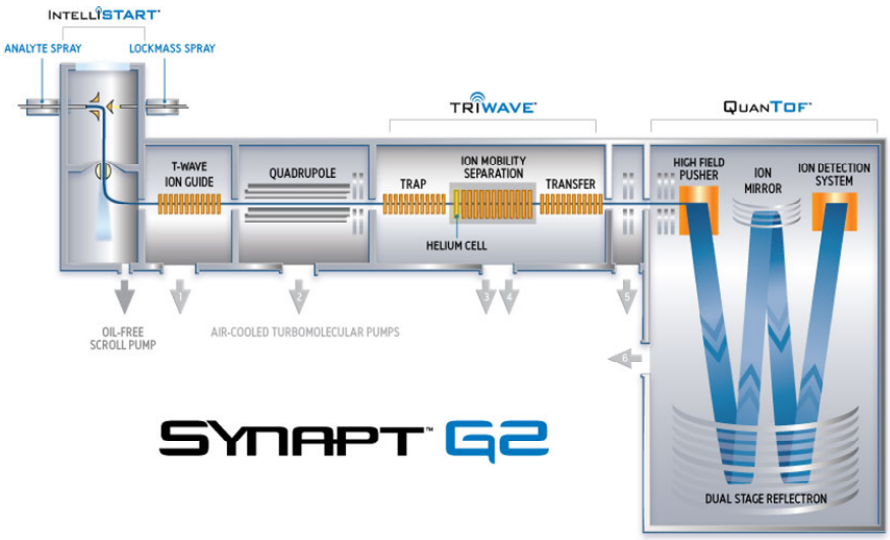

Figure 4.1. Hybrid Q-TOF mass spectrometer with travelling wave ion mobility capability

Limited dynamic range, when operated in mobility mode, restricted the utility of the original instrumentation for LC-MS proteomic studies. A second generation of this instrument combined ion mobility separation with data-independent acquisition, termed HDMS^E (Giles *et al.* 2011; Rodríguez-Suárez *et al.* 2011). In this instrument, the use of a novel ion detection system has significantly increased the dynamic range obtainable in mobility mode and enabled proteomic experiments to be undertaken. The DIA experiment entailed the transmission of all precursor peptides through the quadrupole analyser. The peptides were then subjected to low and elevated energy conditions, within the collision cell, on alternative scans (Silva *et al.* 2005). When mobility separation was enabled, precursors were separated, according to mobility, prior to fragmentation, providing additional experimental information which may be used to align product and precursor ions. This is shown in Figure 4.2. The ability to obtain precursor and product ion information with no precursor ion selection incorporating mobility separation in a single experiment has previously been achieved with the use of split-field drift tube (Koeniger *et al.* 2005).

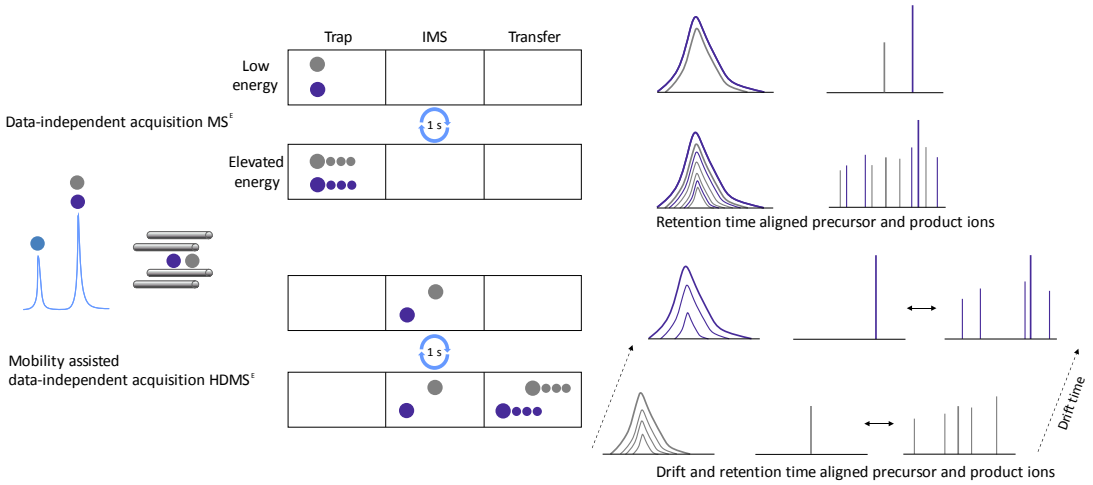

Figure 4.2. Schematic of mass spectrometer arrangement operating in MS^E and HDMS^E

The use of travelling wave ion mobility mass spectrometry (TWIMS) in a structural-based proteomic study has demonstrated the ability to distinguish between phosphorylated and non-phosphorylated peptides (Thalassinos *et al.* 2008). To date, there are few studies that provide a direct comparison of experiments conducted with and without mobility separation on the same instrument for proteomic studies. Of the comparative studies that have been carried out, most have entailed the interfacing

of a mobility device to the mass spectrometer (Venne *et al.* 2005; Saba *et al.* 2009). The Synapt G2 instrument has the ability to switch between experiments conducted with and without mobility separation without physically changing the configuration.

4.2 Aims

The widespread use of ion mobility for proteomic applications is now an achievable aspiration, made with the availability of the first commercial mass spectrometer containing the specific requirements. Currently, ion mobility separations alone do not rival the resolving power of LC as a stand-alone technique. The post-ionisation separation technique, however, fits perfectly into the timescale of an LC-MS^E experiment, increases the peak capacity of the system and has been shown to provide improved peptide detections. An ideal outcome from employing an ion mobility separation step would be to improve dynamic range and increase protein identifications without the need for time consuming multi-dimensional LC. A combination of multi-dimensional LC and ion mobility forms an exciting prospect for mass spectrometry-based proteomics to accomplish one of its major goals, which is to achieve greater coverage of the proteome.

The aims of this work were:

- To assess the effective dynamic range of selected LC-MS configurations using simple standard protein mixtures

- To investigate the linearity exhibited by certain peptide properties for quantitative purposes and compare to the Hi3 peptide approach

- To evaluate 1D RP-LC MS^E, 1D RP-LC $HDMS^E$, 2D RP-RP-LC MS^E and 2D RP-RP-LC $HDMS^E$ methods for profiling and comparative proteomic experiments using a label-free quantitative approach

4.3 Methods

Bacterial protein preparation, LC-MS^E settings and data processing were as detailed in Chapter 2. This section provides a specific overview of the experiments conducted to achieve the specific aims of this work.

4.3.1 System assessment and alternative quantitative methods

The accuracy and reproducibility of the quantification method applied within various LC-MS configurations was assessed by analysing a mixture of four pre-digested proteins (MassPREP Protein Digestion Standard Mix 2). The mixture containing yeast alcohol dehydrogenase, rabbit glycogen phosphorylase *b*, yeast enolase and bovine serum albumin are present at a ratio of 1: 0.34: 1: 3.87 or 1: 0.41: 1.29: 5.79, normalised to alcohol dehydrogenase when quantified using a mass spectrometer operating with a TDC or ADC detector, respectively. This mixture was resolved by 1D RP-LC and 2D RP-RP-LC using the conditions outlined in Table 4.1. All 1D RP-LC experiments were conducted in triplicate.

Table 4.1. Experiments conducted on Synapt G2 with standard proteins

Exp.	LC	Mass spectrometer	Mode of operation	Detector	Mixture (fmol)				Biological matrix
					ADH	Phos b	ENO	BSA	
1	1D	Synapt G2	MS^E	hADC	5.0	2.1	6.5	29.0	-
2	1D	Synapt G2	MS^E	hADC	12.5	5.1	16.1	72.4	-
3	1D	Synapt G2	MS^E	hADC	25.0	10.25	32.3	144.8	-
4	1D	Synapt G2	MS^E	hADC	50.0	20.5	64.5	289.5	-
5	1D	Synapt G2	$HDMS^E$	hADC	5.0	2.1	6.5	29.0	-
6	1D	Synapt G2	$HDMS^E$	hADC	12.5	5.1	16.1	72.4	-
7	1D	Synapt G2	$HDMS^E$	hADC	25.0	10.25	32.3	144.8	-
8	1D	Synapt G2	$HDMS^E$	hADC	50.0	20.5	64.5	289.5	-
9	2D	Synapt G2	MS^E	hADC	50.0	20.5	64.5	289.5	300 ng propane digest
10	2D	Synapt G2	$HDMS^E$	hADC	50.0	20.5	64.5	289.5	300 ng propane digest

Experiments 1 – 4 and 5 – 8 involved the dilution of the 4-protein mixture to four different concentrations. These four dilutions were analysed without addition of a complex background by means of 1D RP-LC MS^E (1 – 4) and 1D RP-LC $HDMS^E$ (5 – 8). The dynamic range, accuracy and reproducibility obtained were compared between these two approaches.

Experiments 9 and 10 involved addition of the 4-protein mixture to a propane-extract digest prior to analysis using 2D RP-RP-LC MS^E and 2D RP-RP-LC $HDMS^E$.

Experiments 1 – 4 were used to investigate the peptide features (including top three peptide intensities) suitable for quantitative measurements. The following six equations were applied to calculate protein abundance based on response factor or linear regression.

The Hi3 label-free method used here for quantifying bacterial proteins is based on Equation 4.1. In the case of the four protein mixture, the known concentration of alcohol dehydrogenase is used to quantify the other proteins in the mixture. The intensity of the top three most abundant peptides for alcohol dehydrogenase are summed and divided by the molar quantity to produce a response factor (intensity/mol). This response factor is then multiplied by the summed intensity of the top three most abundant peptides of the other proteins to provide an estimate of cocnentration. In essence, a single protein has been used to create a one-point calibration.

$$\sum \textit{top three abundant peptide intensities}$$

Equation 4.1

Alternatively, some AUC-based label-free quantitative approaches consider the sum of all peptide intensities for a given protein, Equation 4.2.

$$\sum \textit{all observed peptide intensities}$$

Equation 4.2

Higher molecular weight proteins provide a larger number of observable peptides after tryptic digestion. The sum of all observed peptide intensities can, therefore, be biased towards larger proteins. A modified version of this equation has been termed

intensity based absolute quantification (iBAQ) (Schwanhausser *et al.* 2011), Equation 4.3, which attempts to overcome the MW bias by dividing the sum of all peptide intensities by the number of theoretically observable peptides.

$$\frac{\sum all\ observed\ peptide\ intensities}{no.\ of\ theoretically\ observable\ peptides}$$

Equation 4.3

The iBAQ approach uses a commercially available standard mixture of ~40 proteins present at a wide dynamic range (~10^5). This is then mixed with the sample of interest. The summed intensities of the standard proteins are plotted against theoretical abundance, and the gradient and intercept of a linear regression fit is used to calculate protein abundances from the sample. Quantification is, therefore, based on multiple calibration points rather than just one. This concept was adapted here to utilise the sixteen concentrations of proteins from the four dilutions. These ranged from 2 – 290 fmol. They were then applied as an external rather than internal calibrant. Equation 4.4 is an adapted form of Equation 4.3, including the molecular weight of the protein as an additional variable.

$$\frac{\sum all\ observed\ peptide\ intensities}{no.\ of\ theoretically\ observable\ peptides} \times MW$$

Equation 4.4

In theory, the molecular weight of a protein should not have a direct effect on the quantification calculated using summed intensities of the top three most abundant peptides, except for those proteins which have three or less observable peptides. Equations 4.3 and 4.4 were adapted to include the top three most abundant peptides rather than the sum of all peptide intensities (Equations 4.5 and 4.6).

$$\frac{\sum top\ three\ abundant\ peptide\ intensities}{no.\ of\ theoretically\ observable\ peptides}$$

Equation 4.5

$$\frac{\sum top\ three\ abundant\ peptide\ intensities}{no.\ of\ theoretically\ observable\ peptides} \times MW$$

Equation 4.6

Equations 4.1 – 4.6 were applied to standard proteins at sixteen varying concentrations. For each of these equations, a single-point calibration was created using a response factor for alcohol dehydrogenase, and a multiple-point calibration was created using all proteins.

4.3.2 Evaluation of mobility-assisted MS^E for profiling and comparative proteomics

Ion mobility-MS^E experiments ($HDMS^E$) have been directly compared with conventional MS^E experiments carried out using the same instrument without physically changing the configuration. Table 4.2 provides details on the growth conditions studied, the LC configuration employed and mass acquisition parameters used for the comparisons.

Table 4.2. Experiments conducted on bacterial extracts for evaluation of mobility-assisted MS^E

Sample	LC mode (column)	No. of 1st dimensional fractions	MS	Mode of operation
Propane – bio rep 2	1D RP-LC (BEH)	-	Synapt G2	MS^E
	1D RP-LC (BEH)	-	Synapt G2	$HDMS^E$
Succinate – bio rep 2	1D RP-LC (BEH)	-	Synapt G2	MS^E
	1D RP-LC (BEH)	-	Synapt G2	$HDMS^E$
Propane – bio rep 3	1D RP-LC (HSS)	-	Synapt G2	MS^E
	1D RP-LC (HSS)	-	Synapt G2	$HDMS^E$
	2D RP-RP-LC (HSS)	3	Synapt G2	MS^E
	2D RP-RP-LC (HSS)	3	Synapt G2	$HDMS^E$
Succinate – bio rep 3	1D RP-LC (HSS)	-	Synapt G2	MS^E
	1D RP-LC (HSS)	-	Synapt G2	$HDMS^E$
	2D RP-RP-LC (HSS)	3	Synapt G2	MS^E
	2D RP-RP-LC (HSS)	3	Synapt G2	$HDMS^E$
Acetate	1D RP-LC (HSS)	-	Synapt G2	MS^E
	1D RP-LC (HSS)	-	Synapt G2	$HDMS^E$
	2D RP-RP-LC (HSS)	3	Synapt G2	MS^E
	2D RP-RP-LC (HSS)	3	Synapt G2	$HDMS^E$

Where possible, data sets were compared in which similar experimental conditions were used, for instance, where the same reversed phase analytical column was employed (BEH or HSS). The data sets from which results were extracted and which are discussed are shown in the text or in Table and Figure captions.

4.4 Results and Discussion

4.4.1 Overview

Ion mobility-MS^E experiments ($HDMS^E$) have been directly compared with conventional MS^E experiments using the same instrument without physically changing the configuration. Experimental conditions such as increased protein loading, segmented or longer LC gradients, were not optimised specifically for individual experiments to enable a non-biased comparison between the two approaches. The focus is on the quality of data obtained (Sections 4.4.2 and 4.4.3) and the ability to reproducibly and accurately quantify protein concentrations (Section 4.4.4). Tryptic digests of bacterial proteomes extracted from various growth conditions were analysed using various LC-MS combinations of 1D RP-LC, 2D RP-RP-LC, MS^E and $HDMS^E$. The number of peptides and proteins observed from each growth condition and experimental approach are outlined in Table 4.3.

The $HDMS^E$ data were searched with less stringent criteria based on experimental observation of higher quality fragmentation spectra (discussed later). A problem that may be associated with the lowering of threshold settings can be an increase in false positive identifications. The false positive rate measured here was low for all experimental approaches ($< 2\%$). This demonstrates that the addition of ion mobility separation results in data sets of similar confidence.

The number of protein and peptide identifications increased 2-fold and 1.7-fold from 1D RP-LC MS^E to 1D RP-LC $HDMS^E$. The improvement in protein identifications was lower when comparing 1D RP-LC MS^E to 2D RP-RP-LC MS^E (1.6-fold) but the identification of peptides increased 2-fold. When comparing the system with the lowest peak capacity (1D RP-LC MS^E) to the highest (2D RP-RP-LC $HDMS^E$), a 2.8-fold and 3.3-fold increase in protein and peptide identifications was observed.

Table 4.3. Proteins and peptides observed using various approaches in the study of *M. silvestris* grown with succinate, propane or acetate

	1D RP-LC MS^E	1D RP-LC $HDMS^E$	2D RP-RP-LC MS^E	2D RP-RP-LC $HDMS^E$
Propane – bio rep 3				
Protein identifications	236	584	404	809
Peptide identifications	2386	4595	5404	9350
False discovery rate	0.8 %	0.9 %	0.7 %	1.9 %
Succinate – bio rep 3				
Protein identifications	357	753	544	941
Peptide identifications	2795	4928	5477	8575
False discovery rate	0.8 %	0.9 %	1.3 %	1.1 %
Acetate				
Protein identifications	303	509	450	750
Peptide identifications	2142	3421	4114	6712
False discovery rate	0.7 %	0.8 %	0.4 %	0.7 %

4.4.2 Peptide detections and separation

There are a number of fractionation methods that have been commonly used to reduce sample complexity. These include depletion of abundant proteins, enrichment of a subset of proteins and MD LC. These methods result in an increase of peak capacity, which enables the detection of more components. Improvements in mass measurement accuracy have also been demonstrated to be beneficial for proteomics experiments. The excellent mass resolution and mass accuracy afforded by Orbitrap-based instrumentation has been shown to improve overall specificity of proteomics experiments, since more restrictive search tolerances can be applied. A study on the impact of resolution parameters on tandem MS experiments in an Orbitrap instrument showed limited improvement in peptide/protein identifications above a mass resolution of 30,000 FWHM (Kim *et al.* 2010). The Synapt G2 Q-Tof-type instrument used in the experiments here was operated at a relatively high mass resolution (18,000 FWHM), comparable to the optimal mass resolution parameters suggested by the systematic study using an Orbitrap instrument (15,000 FWHM).

Figure 4.3(a) shows precursor and product ion mass measurement error distributions. These illustrate the stability of mass measurement with and without the application of a mobility separation step. Approximately 90% of the precursor

and 70% of the product ions display a mass error of 0 ± 5 ppm. This mass resolution and mass accuracy was maintained when the mass spectrometer was operated in HDMS^E mode.

The use of ion mobility as a post-ionization separation tool allows the measurement of drift time to be recorded for precursor ions. Figure 4.3(b) shows a plot of mass-to-charge against drift time for precursor ions.

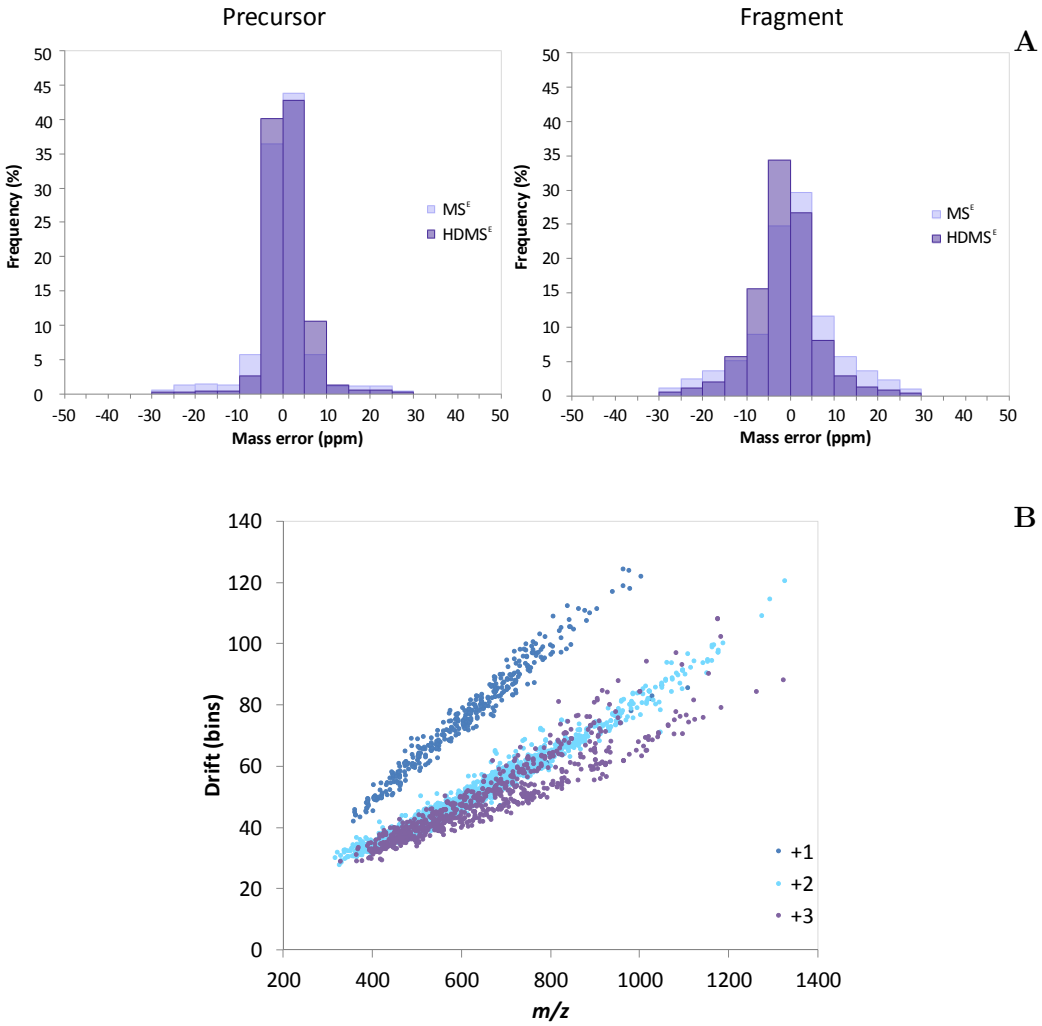

Figure 4.3. Precursor/fragment mass accuracy and precursor drift times
Data shown for succinate biological replicate 2 analysed by 1D RP-LC MS^E and HDMS^E

The +1 and +2 families lie on discrete trend lines while the +3 ions appear to adopt varying conformations for similar m/z. Highly charged ions have been shown to adopt extended conformations in order to minimize coloumb repulsions between charge sites (Valentine *et al.* 2001).

Variance of precursor drift times across technical replicates and different growth conditions was compared. Drift time measurements were shown to be highly reproducible, with 98% of ions with a CV less than 1% across technical replicates and 95% of ions from various growth conditions with a CV less than 1%. This demonstrates the potential of using drift time as a constraint to improve confidence of peptide identification thus allowing a reduction in false discovery rates. The use of ion mobility drift time data to improve peptide identification has been suggested (Shah *et al.* 2010; Valentine *et al.* 2011).

Figure 4.4 shows the analytical space occupied in the m/z, retention time and drift time dimensions during a 1D RP-LC HDMS^{E} experiment. Since the number of elements required to form natural peptides is limited, peptides generally have confined characteristics. This has been observed for peptides separated under reversed phase conditions, where the majority of peptides tend to elute between 25 – 50 min during the gradient conditions used here (0 – 90 mins). This is also seen during separation using ion mobility. Figure 4.4(a) shows the arrival time distributions (analogous to an LC chromatogram) of ions present in an MS scan, the mass spectrum of which is shown in Figure 4.4(b). The separation of peptides by ion mobility is dependent on charge and mass, and like LC separation, the separation space is not fully utilised due to the limited potential composition of peptides. The mass spectrum relating to a selected arrival time distribution is shown for the time point of 30 min during an LC run. Here one would expect a high rate of peptides eluting per second. As expected, baseline separation is not achieved at this level of complexity. Figure 4.4(c) shows an extracted arrival time distribution for a doubly charged peptide, showing a resolution of 40 FWHM.

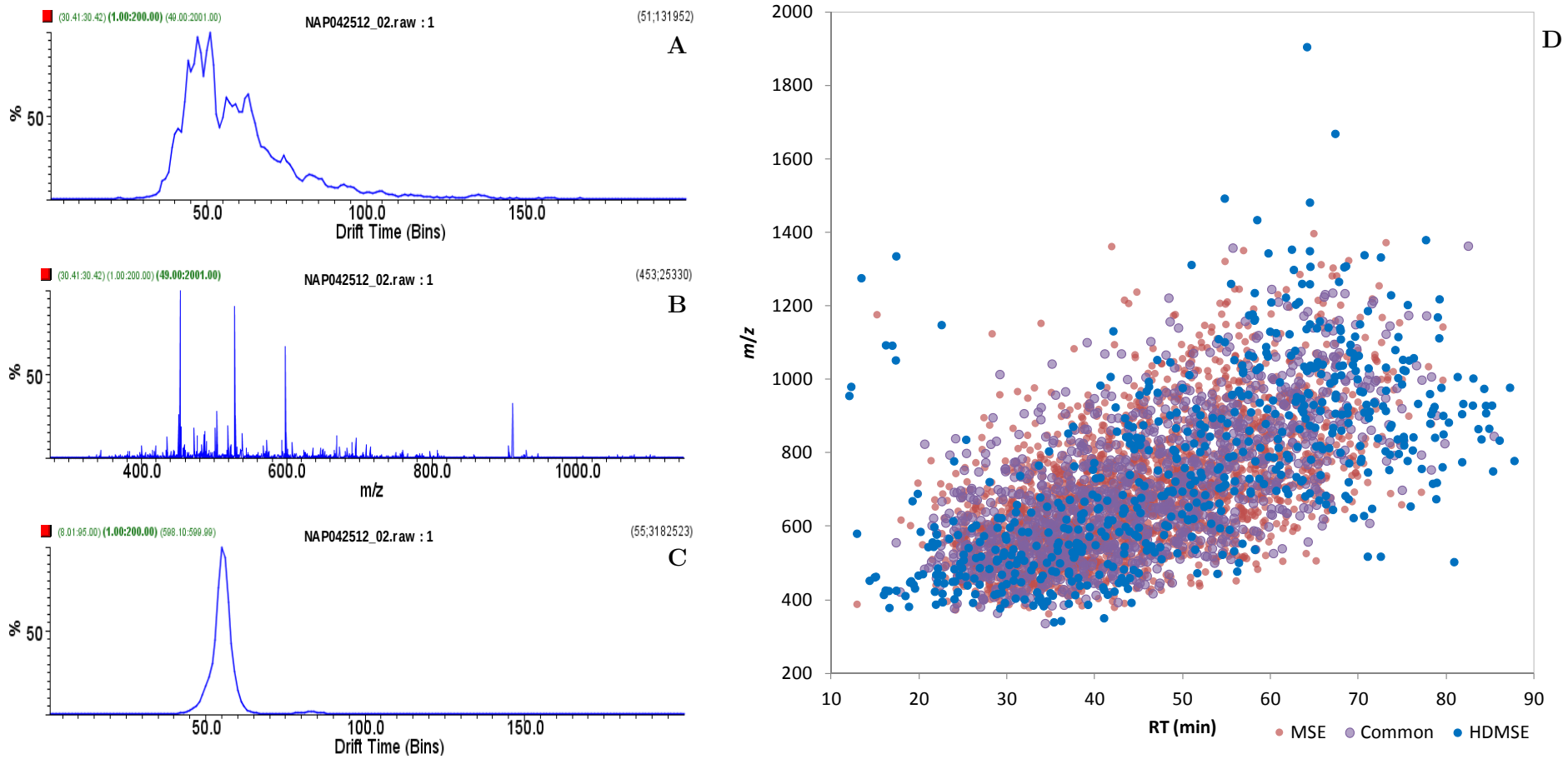

Figure 4.4. Analytical space occupied during a 1D RP-LC HDMS^{E} experiment

Data shown for succinate bio rep 3 (a) Arrival time distribution of peptides at RT 30 min. (b) Mass spectrum at RT 30 min. (c) Extracted arrival time distribution of doubly charged ion. (d) Plot of m/z versus RT for peptides identified from 1D RP-LC MS^{E} and HDMS^{E} experiments.

Figure 4.4(d) shows a plot of a peptide retention time, measured m/z, and detection during a 1D RP-LC MS^E experiment, a 1D RP-LC $HDMS^E$ experiment, or both. It might be expected that additional peptides identified when ion mobility separation was applied may be higher during the portion of the LC gradient where most peptides elute. This does not seem to be the case here. Additional mobility-identified peptides were evenly distributed throughout the whole gradient. There do appear to be additional identifications in the analytical space not occupied by peptides identified during 1D RP-LC MS^E, notably at the start of the gradient (10 – 20 min) and the end (70 – 90 min). Species eluting during 70 – 90 min are the result of high organic content wash, and tend to comprise non-peptide components. It is possible that the ion mobility step results in the separation of non-peptide species, increasing the chances of identification of peptide ions.

A significant benefit of the use of ion mobility in mass spectrometry experiments is the separation of ions with similar m/z but different motilities. This may enable information to be extracted from the individual components. This would be beneficial in proteomics experiments, when two ions of similar m/z elute together. This possibility was investigated in data obtained for succinate analysed using 1D RP-LC MS^E and 1D RP-LC $HDMS^E$. Considering ions with precursor ions masses within ± 0.5 m/z and LC retention times ± 0.5 min, 33 pairs of ions, measured using the $HDMS^E$ approach, had similar accurate mass and retention times. These related to peptides of different compositions. A number of ion pairs with similar masses and retention times were also present in the MS^E dataset. This does not support the hypothesis that ion mobility predominantly separated isobaric species.

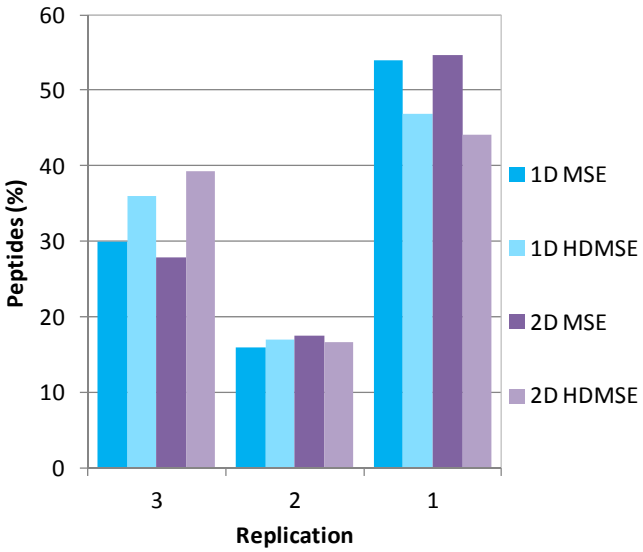

Figure 4.5. Peptide replication rate in technical replicates for various experimental approaches

Data shown for the analysis of succinate biological replicate 3.

Figure 4.5 shows peptides observed in one, two and three technical replicates. For the succinate growth condition using the 1D or 2D MS^E approach, 54% of peptides were observed in one injection, 17% in two and 29% in all three injections. When mobility separation was employed 45% of peptides were identified in one injection and the fraction of peptides identified in all three injections increased to 38% with 17% replicated in two injections. It is likely that this is due to the fact that fewer, non-specific, low intensity ions, which tend to be sampled in one replicate, are present in the mobility-assisted data acquisition. Considering only peptides that were observed in at least two replicates, an average 1.7-fold increase in peptide identification was observed using the mobility-assisted approach.

Reduction in spectral complexity and improvement in signal-to-noise are two features associated with IMMS-based proteomics studies. Data-independent acquisition requires that product ion spectra are extracted from higher energy data acquisitions. The algorithm utilises accurate mass and retention time, amongst other properties, to time align fragment ions with their originating precursor ions. The resulting fragment ion spectrum for a peptide contains ions originating from the precursor, unidentified ions and fragment ions that have been assigned to other peptides. Further analysis of selected product ion spectra of peptides fragmented in normal MS^E mode and $HDMS^E$

mode show that there is a significant (up to five-fold) decrease in the level of chemical noise.

Figure 4.6 shows extracted product ion spectra from the peptide MGATLEALAK m/z 1004.54, which originated from the protein acetyl CoA acetyltransferase. This was observed with and without mobility separation, displaying either the fragment ions that were assigned to the peptide or ions that were not assigned to any peptide. This example shows a typical reduction in noise observed in mobility spectra. The fragmentation pattern is, however, similar between the two approaches. This reduction in complexity observed in HDMS^{E} mode together with improved acquisition specificity allowed the use of less stringent fragmentation criteria during database searching.

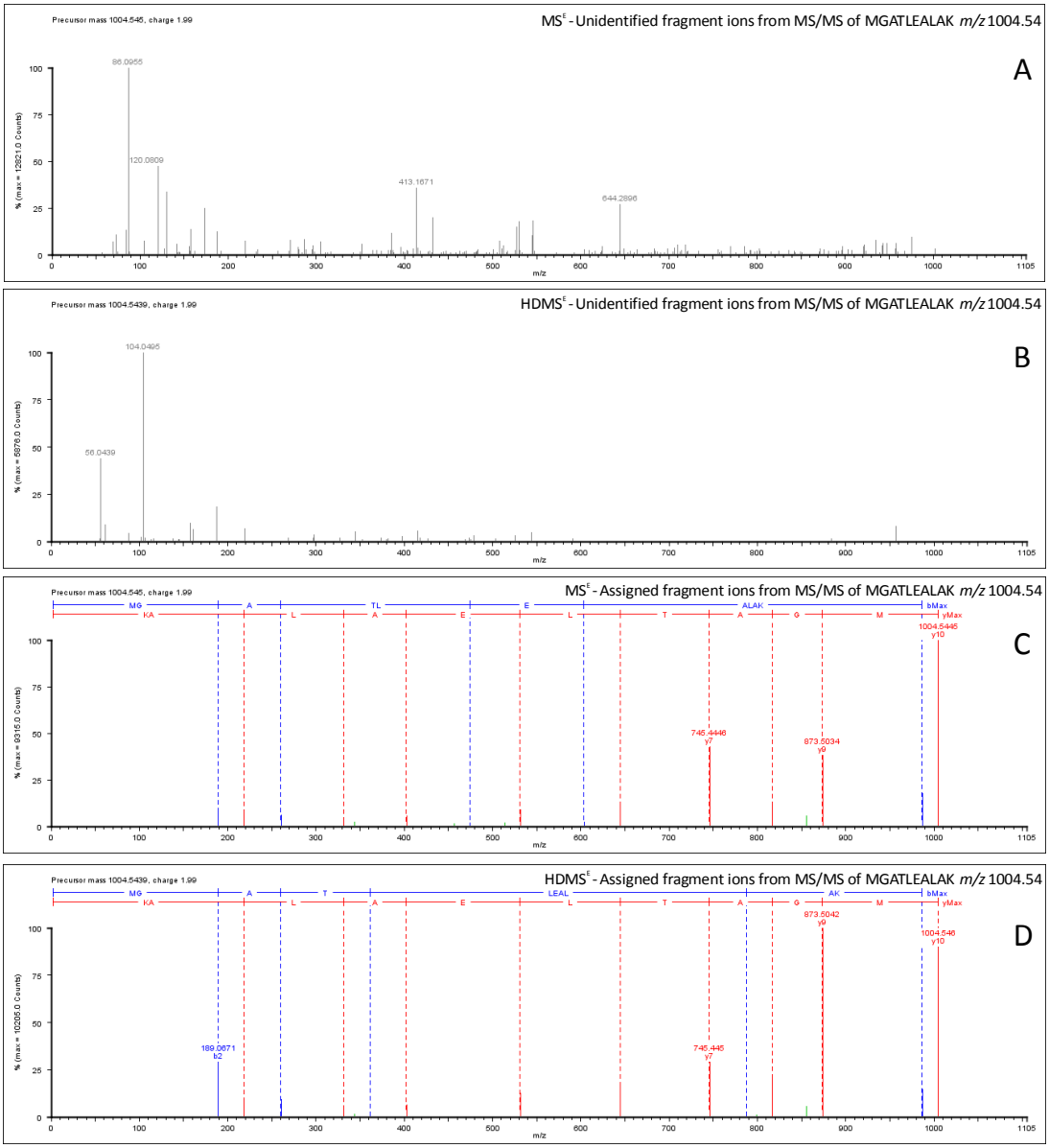

Figure 4.6. Assigned fragment ion spectra acquired in MS^E and HDMS^E modes

Unassigned fragment ions for peptide MGATLEALAK *m/z* 1004.54 acquired in (a) MS^E and (b) HDMS^E mdoes. Fragment ions assigned peptide from (c) MS^E and (d) HDMS^E data.

The quality of fragment ion spectra was assessed by considering the number of matched and consecutive fragment ions for peptides identified using both MS^E and $HDMS^E$ approaches. Figure 4.7 shows a plot of the number of matched or consecutive matched fragment ions assigned to a peptide using the 1D RP-LC MS^E or 2D RP-RP LC $HDMS^E$ approach.

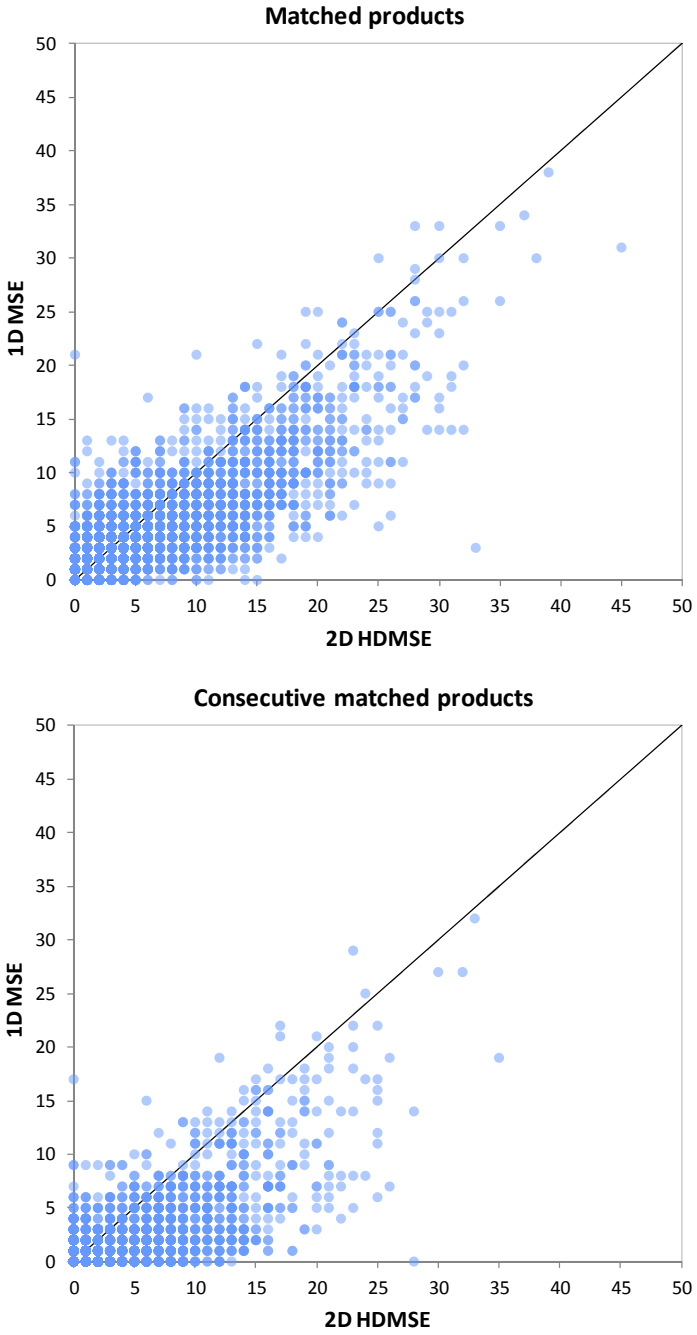

Figure 4.7. Matched products and consecutive matched fragment ions per peptide

For peptides with a greater number of fragment ions identified by either the 1D MS^E or 2D $HDMS^E$ approach, the data point is closest to the y- and x-axis, respectively. Data points that lie on y = x indicates instances that two methods did not differ. More peptides were assigned with additional matched and consecutive matched ions by the $HDMS^E$ experiment, indicating the improved quality of product ion spectra obtained.

Raw data were processed to produce deconvoluted and deisotoped ion masses. For a single experiment, Table 4.4 contains information regarding the number of ions detected after processing for each of the various experimental approaches, and the summed intensities of the ions. After processing, a proportion of these masses were assigned as peptides during database searching. The number of peptides assigned is shown in Table 4.4, together with the proportion they constitute of the total recorded intensity. The number of detections was shown to reduce by half when employing an ion mobility separation step in both 1D-LC and 2D-LC configurations. This does not equate to a reduction in the number of assigned peptides, which do not significantly differ with the addition of ion mobility. These data represent a single data set. The experiments were conducted in triplicate to filter out identifications only found in a single experiment. It has been shown that more peptides tend to replicate in a single experiment in the MS^E experiments when compared to $HDMS^E$ experiments. The number of peptide detections, after applying a replication filter, increased when ion mobility was employed (Table 4.1).

Table 4.4. Deconvoluted ion masses and peptides

	Deconvoluted ion masses		Peptide detections	
Technique	**Number**	**Intensity (× 10^8)**	**Number**	**Proportion of total intensity**
1D MS^E	95,354	2.42	6498	40.9%
1D $HDMS^E$	47,655	0.99	7775	58.6%
2D MS^E	309,763	11.7	13,773	40.3%
2D $HDMS^E$	116,710	3.12	13,031	56.4%

Data shown for succinate bio rep 3 from one technical analysis.

These data show a larger proportion of the total detected intensities are attributed to peptides in the mobility experiments. This is in agreement with the observation that fragment ion spectral quality is improved with the inclusion of ion mobility

separation. The total intensity recorded provides an indication of the amount of sample that has been loaded. Here, the total recorded intensity is approximately three times larger in the 2D approaches when compared to the respective 1D approaches. This is to be expected, since the 2D approach used here employed three first dimensional fractions, which led to three second dimensional separations. For quantitative experiments, it is important to load similar amounts of sample for comparative purposes. In methods such as iTRAQ, this is ensured by carrying out a BCA assay for estimating protein quantification. Protein estimation methods are particularly inaccurate for complex mixtures of proteins. These data show that MS can be used to estimate sample loading, which can be helpful in comparative studies. It is apparent from the total detected intensities that ion mobility experiments suffer from loss of ions or limited detection capabilities. This is discussed further in section 4.4.4.

4.4.3 Protein identifications

Chapter 3 describes how proteins identified using a 1D approach form largely a subset of proteins identified using the 2D approach. The overlap of protein identifications between experimental approaches that employ varying first dimensional chemistries have been shown to exhibit less commonality. A comparison of identifications obtained using SCX-RP and electrostatic repulsion-hydrophilic interaction chromatography (ERLIC)-RP revealed a 52% overlap (Hao *et al.* 2010). The comparison of proteins identified by the approaches with and without the use of ion mobility separation has been less well documented in the literature.

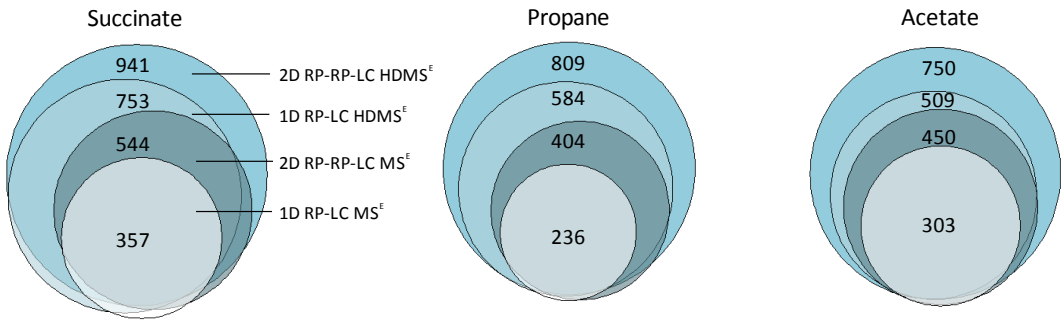

Figure 4.8. Overlap of protein identifications using various experimental approaches
Data shown for biological replicates 3, proteins observed in ≥2 technical replicates

Figure 4.8 shows the overlap of protein identifications from the four experimental approaches used here, for succinate-, propane- and acetate-growth conditions. As with the comparison of 1D LC and 2D LC approaches, these results show an excellent overlap of protein identifications between all four techniques for all studied substrates. The proteins identified using the 2D mobility approach almost completely encompass the identifications from the other three approaches. On average, 7% of protein identifications were not confirmed by the 2D mobility approach. When comparing pair-wise datasets between two approaches, there is also a good degree of overlap. From 1D MS^E to 1D $HDMS^E$, only 3% were not identified by the latter method, from 1D $HDMS^E$ to 2D MS^E, 12% were not identified and from 2D MS^E to 2D $HDMS^E$ only 6% were not confirmed (values from analysis of succinate).

Obtaining this extent of commonality between four proteomic methodologies is not common. These data suggest that protein identifications from all four approaches are highly confident. This shows great promise for the use of ion mobility in future proteomic studies.

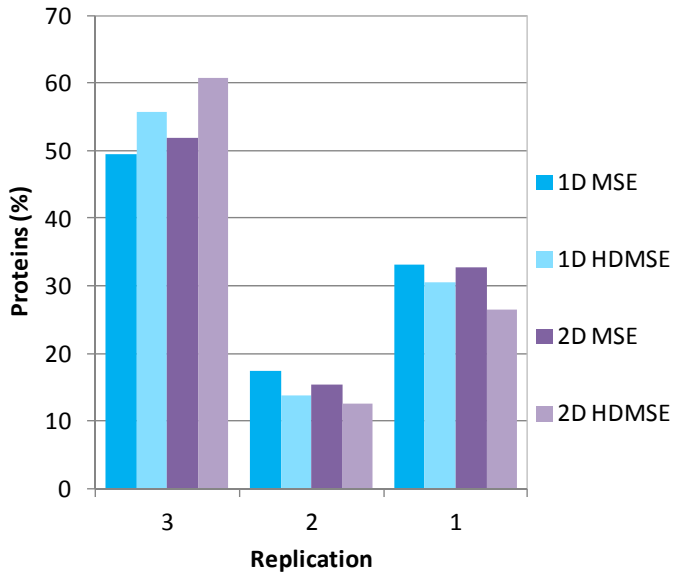

Figure 4.9. Protein replication rate in technical replicates

Data shown for succinate biological replicate 3.

For the succinate-growth condition, using the 1D/2D MS^E approach, approximately 33% of proteins were observed in one injection, 16% in two and 50% in all three injections. These values were improved with the mobility-assisted experiment where 27% of proteins were identified in one injection, shown in Figure 4.9. The majority of proteins (59%) were identified in all three injections with 14% replicated in at least

two injections. The improved reproducibility is in agreement with that observed for peptide replication. Proteins were more frequently observed than peptides, as has been previously observed (Nagaraj *et al.* 2012).

Figure 4.10 illustrates the functions of proteins identified by MS^E and $HDMS^E$ experiments. Higher protein coverage onto known biological pathways was observed for mobility-assisted compared to normal MS^E experiments, giving additional confidence that the proteins identified through the mobility-assisted experiment were of biological relevance.

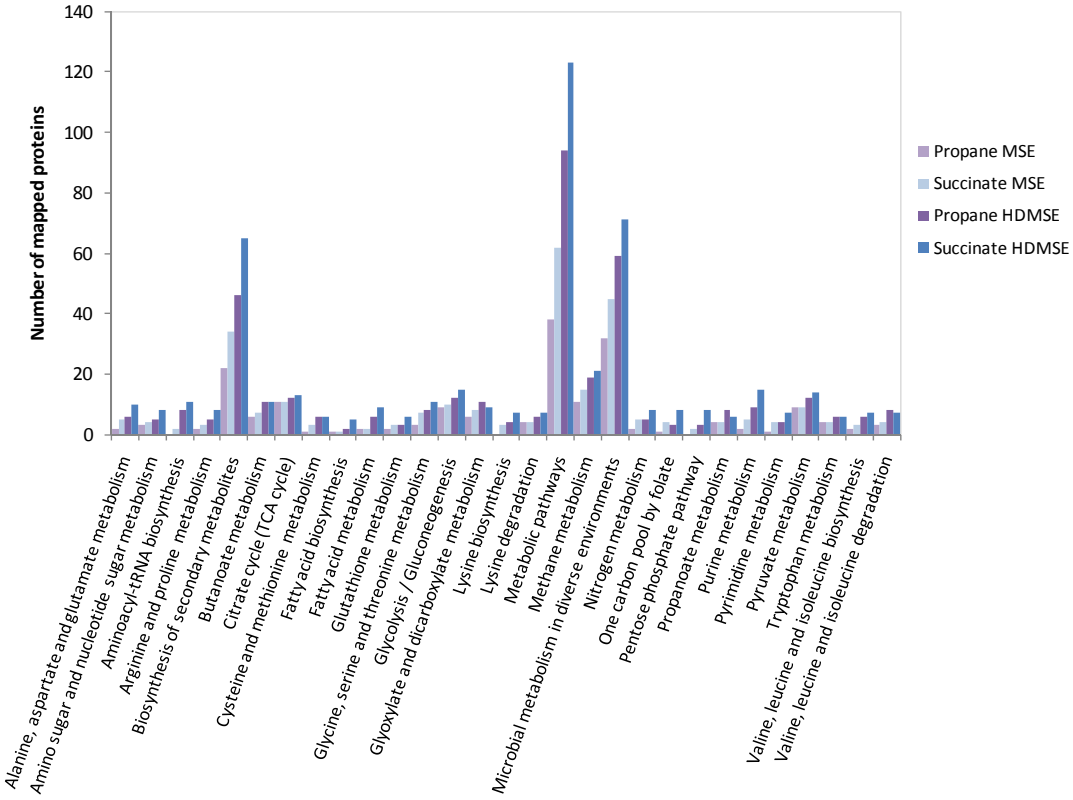

Figure 4.10. Pathways populated by proteins identified in succinate and propane grown cells

Data shown for analysis of biological replicates 2 separated by 1D RP-LC.

4.4.4 Absolute quantitative measurements

In addition to protein identification, an ability to achieve accurate and reproducible quantification measurements of the identified proteins forms an essential component of a mass spectrometry-based proteomics approach. Quantification in MS^E and $HDMS^E$ experiments was performed by employing the Hi3 label-free system (Silva *et al.* 2006). The accuracy and reproducibility of absolute and relative quantification measurements was assessed in Chapter 3 in the evaluation of the characteristics of a 2D-LC system. Measurements made for complex bacterial protein mixtures and a simple protein mixture have been assessed to validate the use of mobility-assisted experiments in an MS^E approach.

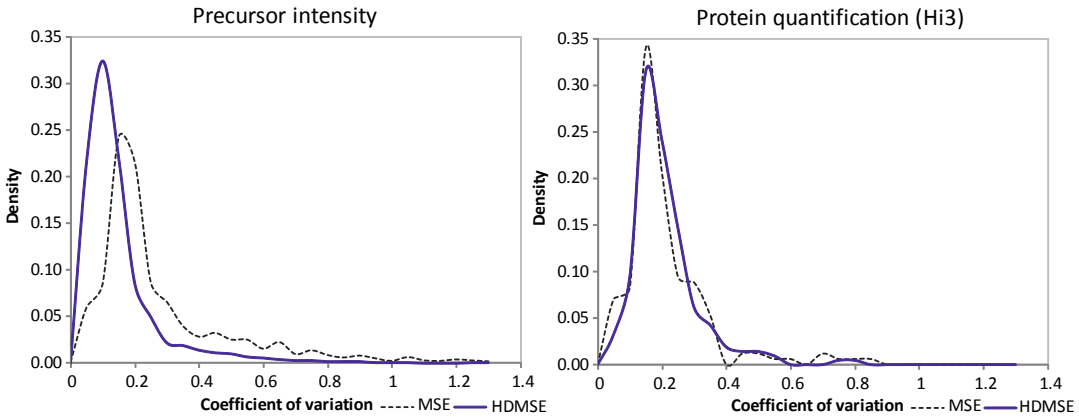

Figure 4.11. Precursor intensity and protein quantification technical variation
Data shown for succinate biological replicate 2 analysed by 1D RP-LC MS^E and 1D RP-LC $HDMS^E$

A key factor required for accurate quantification is a high reproducibility of abundance measurements. Reproducibility of protein quantification across technical replicates of the two methodologies was assessed using their coefficient of variation (CV). Figure 4.11 shows the variation associated with protein quantification (Hi3 method) and precursor intensity measurements from MS^E and $HDMS^E$ experiments on succinate-grown extracts. Average CVs were 0.18 and 0.13, respectively for precursor intensities. CVs were slightly higher for protein concentration; 0.23 for MS^E and 0.20 for $HDMS^E$. Variation in protein concentration from replicate injections was not found to be significantly concentration-dependent over the normal measured concentration range. Peptides reported at 2 fmol were found to show little difference to peptides reported at 100 fmol.

For bacterial extracts, column loading was maintained between MS^E and $HDMS^E$ experiments (~500 ng), enabling direct comparison of protein quantities estimated using the Hi3 approach. Figure 4.12(a) illustrates that protein concentration measurements were in good agreement between MS^E and $HDMS^E$ approaches (when both 1D- and 2D-LC are employed) for proteins reported to be in the ~3 – 400 fmol concentration range. For proteins with concentrations estimated to be between ~400 – 8000 fmol agreement between the two approaches was less clear. This could be attributed to the fact that mobility separated ions reach the detector in relatively sharp packets, associated with separation in the mobility cell, which could result in detector saturation.

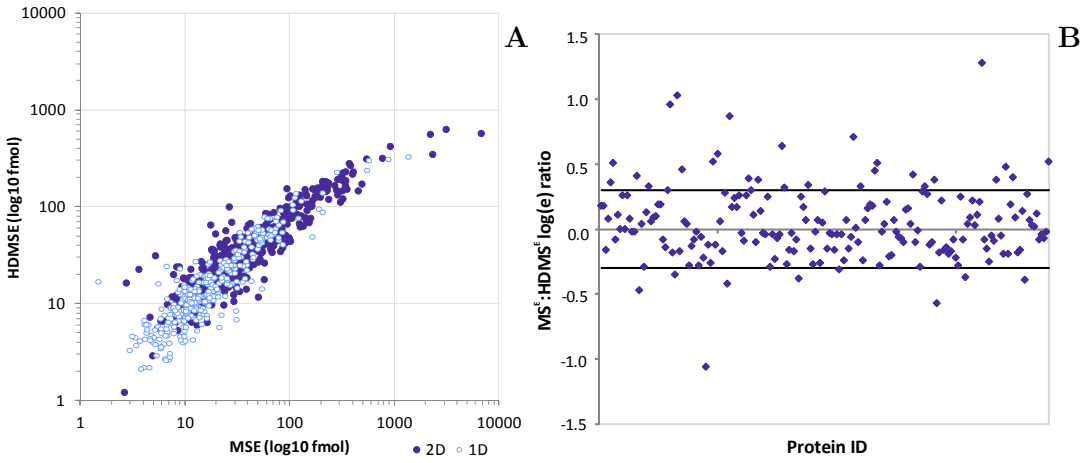

Figure 4.12. Comparison of MS^E and $HDMS^E$ methods for protein quantification
(a) Estimation of protein abundance 1D RP-LC MS^E *versus* 1D RP-LC $HDMS^E$ and 2D RP-RP-LC MS^E *versus* 2D RP-RP-LC $HDMS^E$. (b) Relative quantification performed between 1D RP-LC MS^E and 1D RP-LC $HDMS^E$.

Relative changes in protein quantification between different growth conditions can be evaluated using Expression software, which takes into account all assigned measured peptide intensities rather than just the top three most intense peptides. This software tool was utilized to measure changes in protein concentrations in the same growth condition to compare values obtained from MS^E and $HDMS^E$ experiments. Figure 4.12(b) shows the relative concentration values of proteins quantified in succinate extracts using $HDMS^E$ and MS^E. A log(e) ratio of approximately 0 ± 0.3 indicates minor experimental difference between the values obtained from the two approaches. This represents the majority of the data within three times the expected signal-to-

noise of non-regulated proteins. The vast majority (140 out of 198) lie within these values indicating no experimentally significant differences were observed.

Table 4.5. Estimated quantities of a simple protein mixture at varying concentrations

(- not observed/applicable, * observed with no quantification)

Protein	LC	Matrix	Theoretical (fmol)	MS^E (fmol)	Error	$HDMS^E$ (fmol)	Error
Phos b	1D	-	**2.1**	-	-	-	-
ADH	1D	-	**5.0**	5.0	-	-	-
BSA	1D	-	**29.0**	32.0	10.3 %	*	-
ENO	1D	-	**6.5**	6.5	0 %	-	-
Phos b	1D	-	**5.1**	4.7	-7.8 %	-	-
ADH	1D	-	**12.5**	12.5	-	12.5	-
BSA	1D	-	**72.4**	74.1	2.3 %	74.4	2.8 %
ENO	1D	-	**16.1**	13.8	-14.3 %	-	-
Phos b	1D	-	**10.3**	8.6	-16.5 %	8.1	-21.4 %
ADH	1D	-	**25.0**	25.0	-	25.0	-
BSA	1D	-	**144.8**	137.7	-4.9 %	155.8	7.6 %
ENO	1D	-	**32.3**	24.1	-25.4 %	23.4	-27.6 %
Phos b	1D	-	**20.5**	16.2	-21.0 %	20.3	-1.0 %
ADH	1D	-	**50.0**	50.0	-	50.0	-
BSA	1D	-	**289.5**	242.0	-16.4 %	155.2	-46.4 %
ENO	1D	-	**64.5**	52.3	-18.9 %	56.6	-12.2 %
Phos b	2D	300 ng Propane	**20.5**	24.1	17.6	25.7	25.4
ADH	2D	300 ng Propane	**50.0**	50.0	-	50.0	-
BSA	2D	300 ng Propane	**289.5**	316.2	9.2	292.3	1.0
ENO	2D	300 ng Propane	**64.5**	46.7	-27.6	51.8	-19.7

In order to evaluate the suitability of mobility-assisted approaches to measure accurate protein concentration levels, a simple mixture of four known digested proteins was prepared and analysed using the four LC-MS configurations. The theoretical and experimentally derived quantities are outlined in Table 4.5. For levels that have been shown to be within the concentration range where MS^E and $HDMS^E$ approaches can both be used, agreement with expected values and between approaches was excellent. The majority of the estimated values were slightly

underestimated and fell within a percentage error range of -27 to 10%. The dynamic range of the HDMS^E approach was, as has already been observed, slightly lower than that of the MS^E experiment and so, for significantly higher concentrations, the values are in less good agreement. This is observed for BSA, which, when present at 290 fmol, has an estimated value given by the 1D RP-LC HDMS^E approach of 155 fmol (error of -46.4 %). This was investigated further by examination of the MS response exhibited by tryptic peptides from BSA.

Figure 4.13 shows the integrated MS response for the ion 722.42 m/z, which relates to the doubly charged peptide YICDNQDTISSK + carbamidomethyl from the protein BSA. The ESI response was shown to be different, but linear, at increasing concentrations of 29, 72, 145 and 290 fmol (1: 2.5: 5: 10) between 1D MS^E and 1D HDMS^E modes.

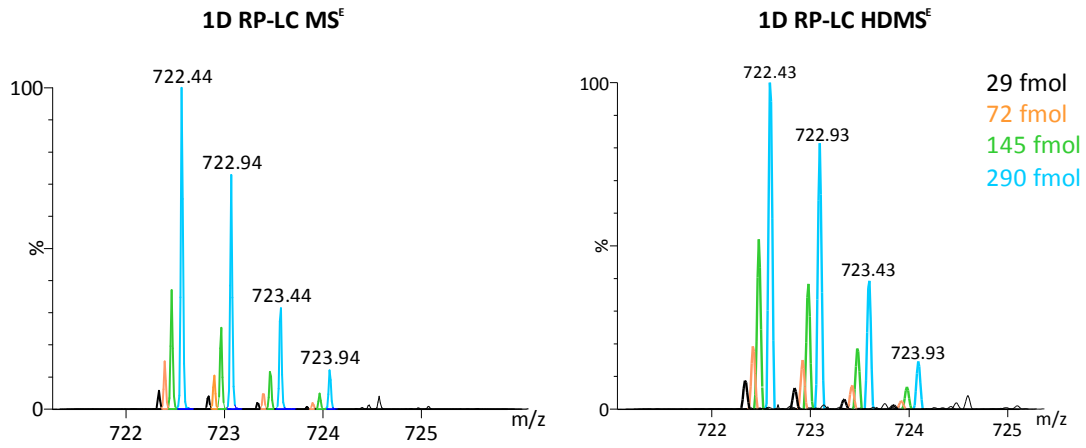

Figure 4.13. MS response of BSA peptide at varying concentrations

The absolute intensities recorded during mobility and non-mobility approaches are not comparable due to the decreased signal observed using the mobility approach, and therefore comparisons were made using relative peptide intensities. These are shown for peptides identified from BSA at concentrations 72 and 290 fmol (Figure 4.14). When the top three most abundant peptides are considered, all three peptides appear to be more abundant at 290 fmol when using the MS^E approach compared to that using the HDMS^E method. This is not observed at the 72 fmol concentration.

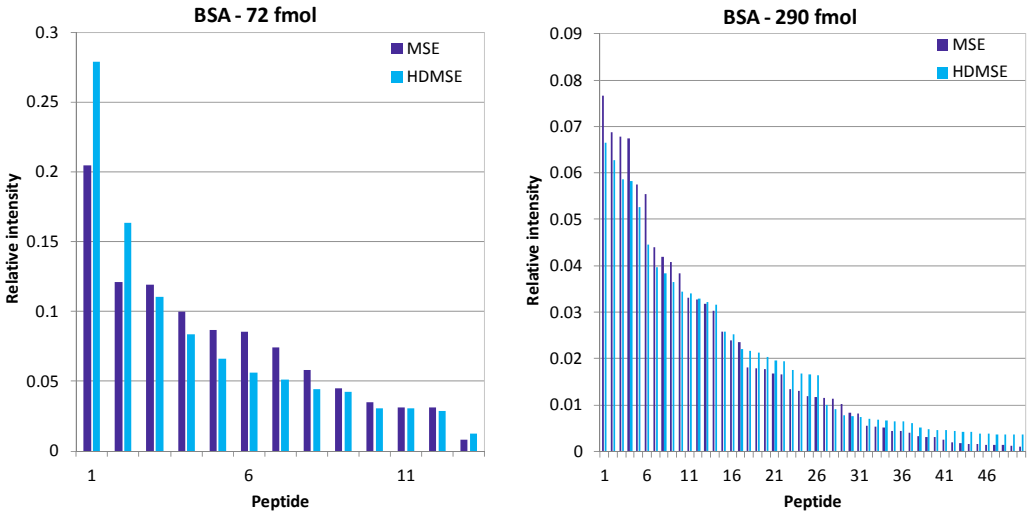

Figure 4.14. Relative peptide intensities measured by MSE and HDMSE organised by decreasing intensity for BSA

The 2D HDMS^E approach did not appear to be limited for quantifying higher abundant proteins against a complex background, with all four proteins at varying concentrations being quantified to within -27.6 to 25.5% of their theoretical value. This could be explained by considering the increased complexity associated with uni-dimensional LC separation. More peptides would be predicted to elute per second during a 1D RP-LC experiment when compared to 2D RP-RP-LC. The number and concentrations of ions can have a marked effect on the ionisation efficiency and the detection. The reduced number of components reaching the detector during 2D separation, which would then be concentrated during mobility separation, would be expected to be within the limits of the detection system.

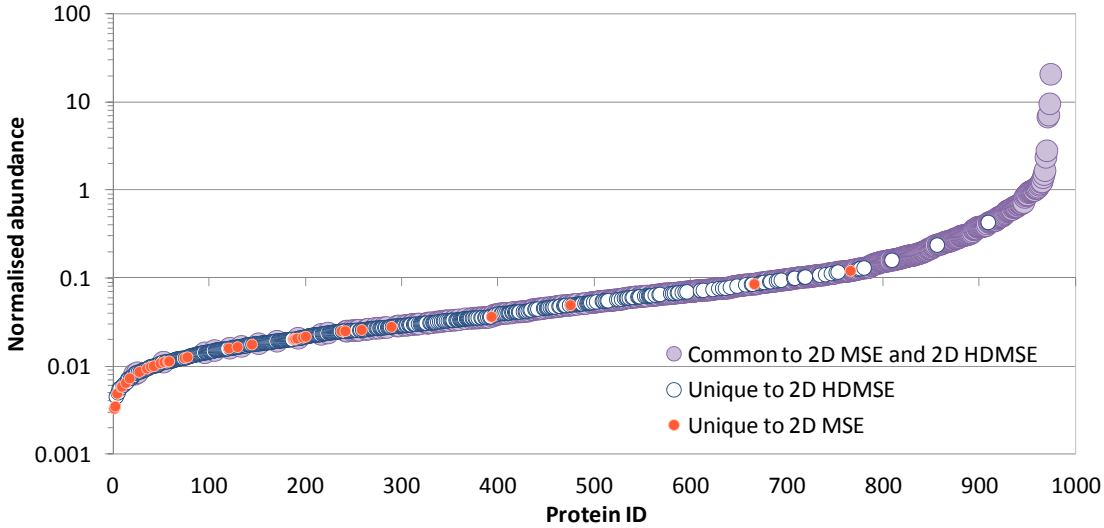

Figure 4.15. Abundances for proteins quantified using 2D RP-RP-LC with MS^E and $HDMS^E$

Data shown for succinate bio rep 3.

Work in Chapter 3 indicated that additional proteins identified using the 2D approach were predominantly lower in abundance. Figure 4.15 shows a plot of protein identifications ordered in ascending abundance for proteins identified using the 2D RP-RP-LC MS^E and $HDMS^E$ approaches. The proteins uniquely identified by the mobility approach vary in concentration and are not predominantly low abundant proteins. This in contrast to observations between 1D RP-LC MS^E and 2D RP-RP-LC MS^E approaches. Ion mobility may reduce complexity by separating out existing components over a wide range of concentrations, which increases specificity. Here, protein loading was not increased in 2D mobility experiments, which may limit the detection of additional low abundance components. The few proteins that were not identified by the mobility experiment but only by the normal MS^E approach were lower in abundance. It is apparent that the mobility approach can have limitations for the detection of low abundant proteins. This may be overcome by increasing the sample loading, not carried out here in order to ensure that a non-biased comparison could be undertaken.

4.4.5 Relative quantitative measurements

Figure 4.16 shows a comparison of MS^E and $HDMS^E$ approaches for relative quantification measurements. A scatter plot comparing the regulation values

measured from the 1D MS^E and 1D HDMS^E methods shows a reasonable correlation, with an R-squared value of 0.76, shown in Figure 4.16(a).

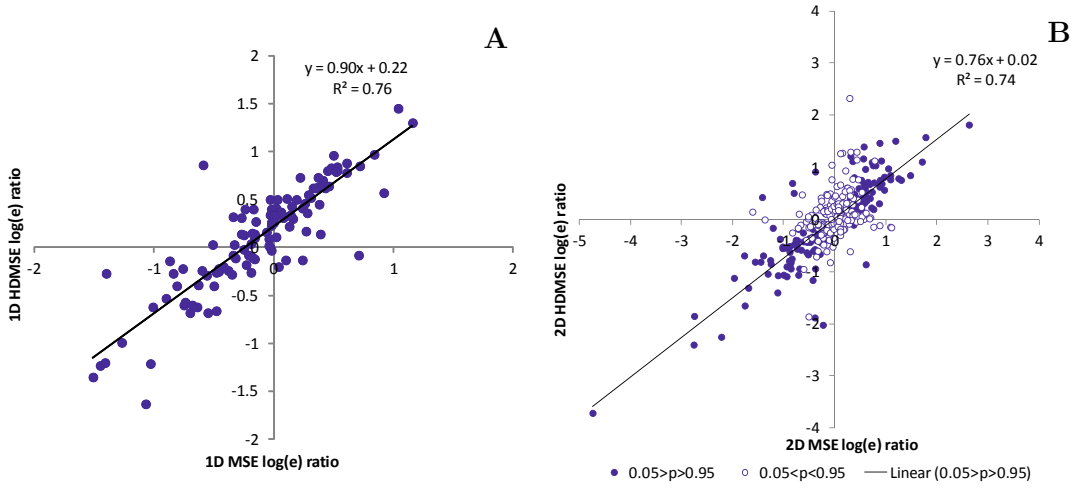

Figure 4.16. Comparison of methods for estimating relative changes in protein expression

Data shown for succinate:propane analysed by (a) 1D RP-LC MS^E/HDMS^E and (b) 2D RP-RP-LC MS^E/HDMS^E

306 relative measurements were shared by the 2D HDMS^E and 2D MS^E approaches, shown in Figure 4.16(b). 139 out of 306 pairs of measurements included confident *p*-values from both methods. The majority of regulation values that were reported as significant ($0.05>p>0.95$) were in agreement between the two approaches (112 out of 139). There was less agreement between relative values that were not within the confidence values ($0.05<p<0.95$). When considering just the confident values, a reasonable correlation was observed (R-squared 0.74).

Figure 4.17 shows the citric acid cycle populated with proteins that displayed differential expression when the Expression tool was used to compare protein concentration differences between succinate- and propane-grown cells. This was measured using all four LC-MS approaches. These values provide two examples that illustrate improved proteome and pathway coverage when mobility-assisted approaches were employed; (1) relative expression of the enzyme aconitate hydratase was exclusive to the 2D RP-RP-LC HDMS^E approach and (2) regulation of the beta subunit of succinate dehydrogenase was only obtained using the 1D RP-LC HDMS^E and 2D RP-RP-LC HDMS^E approaches.

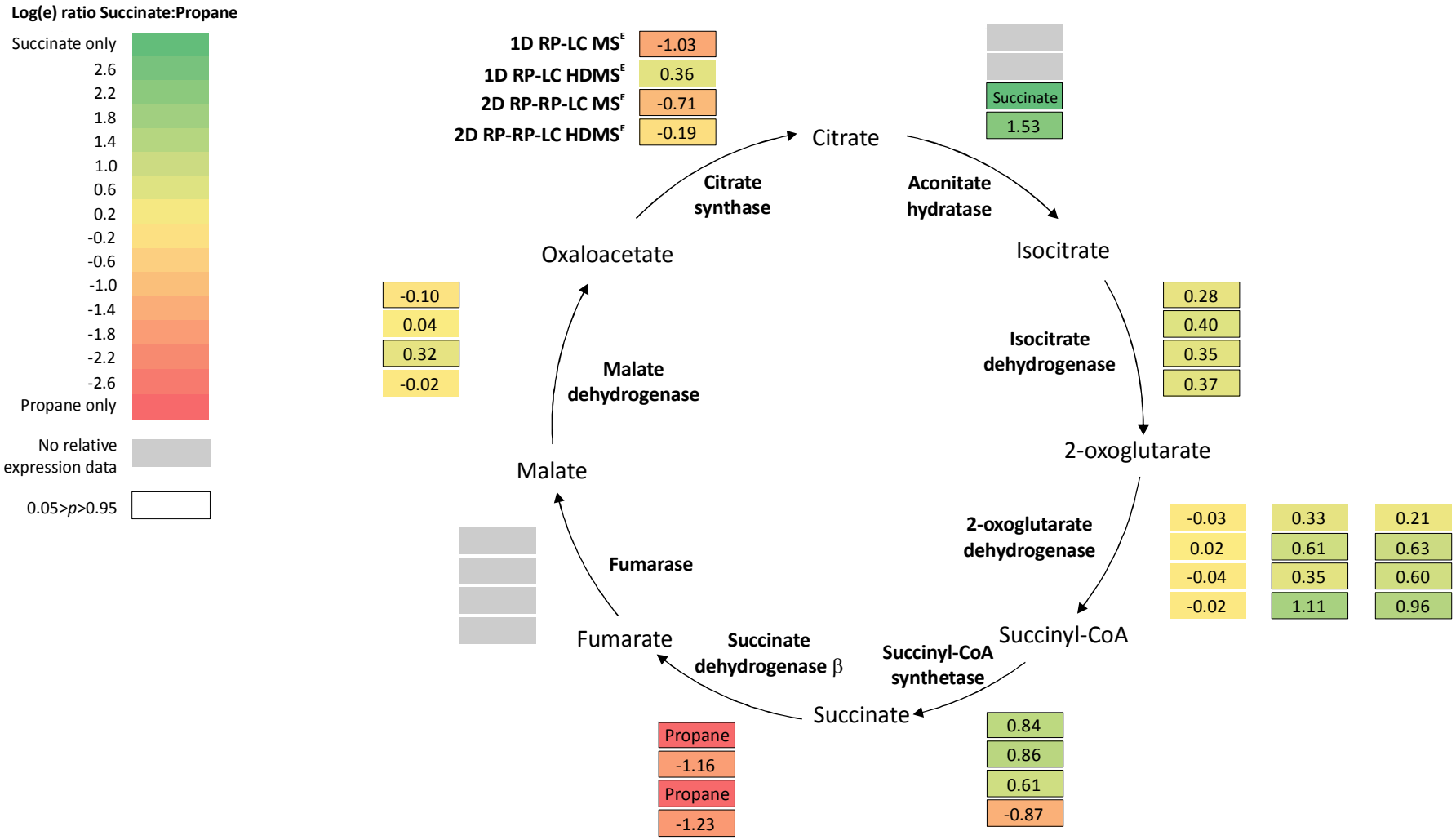

Figure 4.17. Relative expression values of citric acid cycle enzymes

The enzyme, fumarase (Msil_0238), was detected for both succinate- and propane-growth conditions using the 2D RP-RP-LC HDMS^E approach. This protein did not meet the criteria set for the Expression software, and therefore, no relative quantification information could be obtained. Relative expression is shown for the three subunits of the enzyme 2-oxoglutarate dehydrogenase. Most of the techniques were able to confidently show that two out of three subunits were up-regulated in the succinate-growth condition. When considering significant regulation values, all relative measurements were in excellent agreement between all four analytical techniques, with only one exception (succinyl-CoA synthetase). These results demonstrate the potential for further use of both ion mobility and 2D RP-RP-LC in label-free MS^E approaches in order to obtain information regarding the regulation of protein expression.

4.4.6 Alternative label-free quantitative approaches

The use of ESI-MS response for estimating the abundance of proteins forms the basis of the label-free approach used in all studies carried out in this work. The Hi3 technique, which involves summing the top three most abundant peptides is justified by the fact that all proteins, when subjected to tryptic digestion, exhibit similar relative peptide intensity profiles, and exhibit linear properties at increasing concentration. Figure 4.18(a) shows the relationship between theoretical protein concentration and the measured top three peptide intensities for standard proteins ranging from 5 - 290 fmol with an excellent R-squared value (0.98). These values were obtained from four separate experiments carried out without replicate analyses. Figure 4.18(b) shows the standard deviation for the coefficient of determination (R-squared) experimentally derived for the top three peptide intensities and other peptide properties from replicate analyses.

The utilisation of summed peptide intensities has been described for techniques such as iBAQ, for which data are recorded on a mass spectrometer operating in DDA mode. The use of summed peptide intensities was investigated here for data acquired by MS^E, which is a DIA method. The relationships between summed peptide intensities and protein concentration with and without the inclusion of parameters accounting for MW bias, were very similar (R-squared 0.84 – 0.87). It is perhaps unsurprising that the equation that divides the top three peptide intensities

by the number of observable peptides has an adverse affect on the relationship with concentration, since the top three peptide intensities are largely independent of molecular weight.

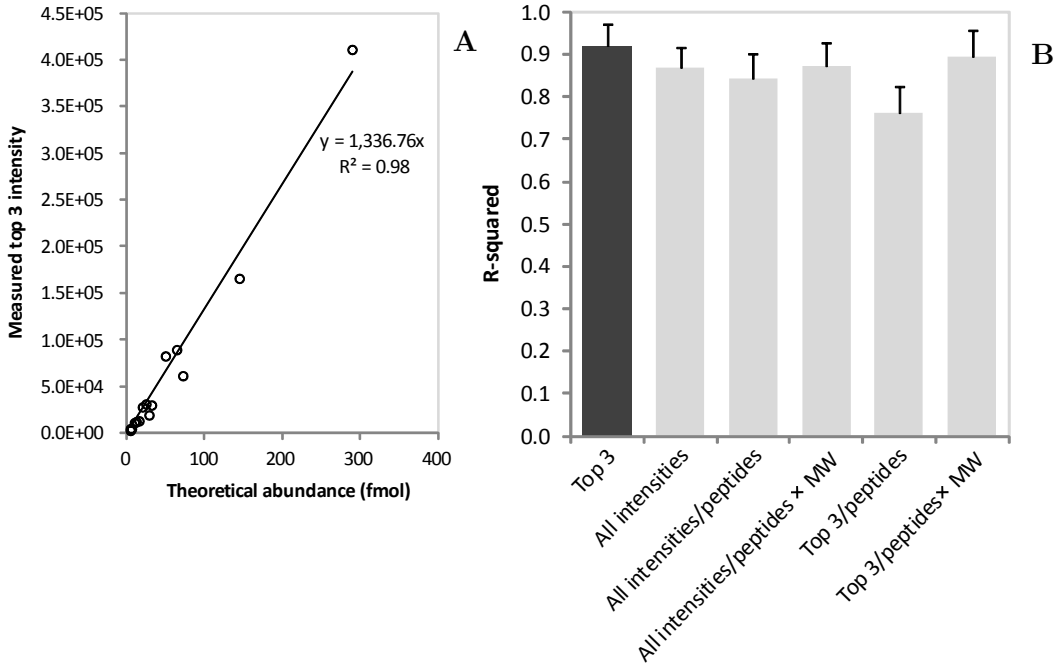

Figure 4.18. Relationship between theoretical abundance and various peptide properties

(a) Linear fit of theoretical abundance against sum of three most intense peptides for four-protein mixture present at 5 – 290 fmol. (b) R-squared values for various peptide properties against theoretical abundance. Error bars indicate variation over three technical analyses.

The iBAQ method has recently been described, whereby peptide intensities are log-transformed and plotted against known log-transformed absolute molar amounts of proteins spiked into a sample. Linear regression was then used to fit the sum of all intensities to absolute standard protein amounts. The slope and intercept from this calibration curve was then used to convert intensities to molar amounts for all identified proteins (Schwanhausser *et al.* 2011). This was applied to the detections made for the 4-protein mixture dilutions, which were analysed in the absence of sample, and shown in Figure 4.19(a). The slope and intercept from this calibration curve was used to convert intensities recorded from the analysis of methane-grown extracts that were analysed separately. Figure 4.19(b) shows the estimation of methane-grown protein abundances obtained using this method and a comparison to the top three peptide response factor method.

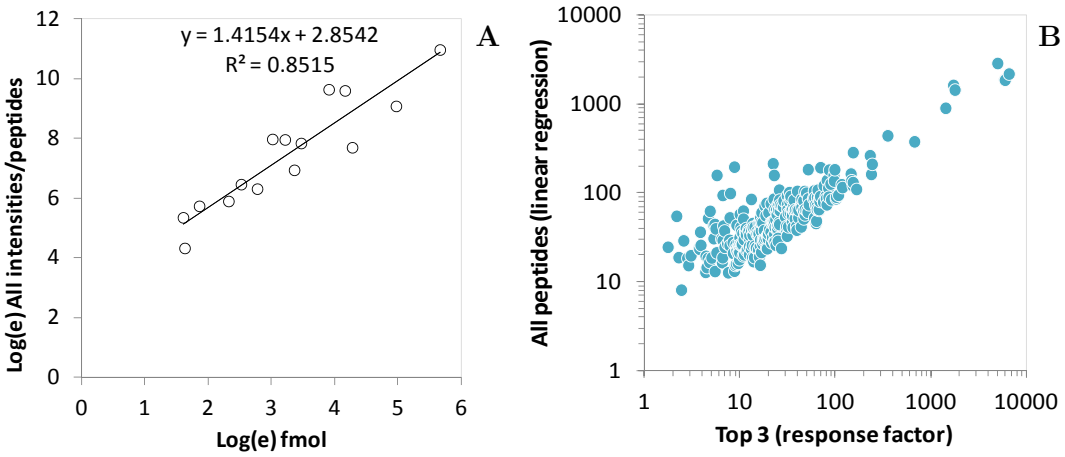

Figure 4.19. Absolute quantification based on sum of all peptide intensities

(a) Plot of theoretical abundance against all peptide intensities divided by theoretical observable peptides. (b) Estimation of protein abundance from methane-grown cells by the top three approach using a response factor and the "iBAQ" like approach using the equation from Figure 4.19(a).

The majority of estimated abundances were in good agreement between the two quantification methods. The larger differences were observed between the less abundant proteins. General agreement between these two approaches does pose an important question, regarding the development of new label-free approaches. A new technique, in general, should offer improvement in one or more aspects, for instance, accuracy, reproducibility or cost. Utilising all peptide intensities, rather than the top three peptides does not appear to offer any significant advantages. If a calibration curve was made using a commercially available set of proteins, the cost of this would be significantly higher than the use of a single point calibration (£452.20 for Universal Proteome Standard 2, Sigma compared to £68 for glycogen phosphorylase *b*).

The accuracy of the quantitative approaches investigated here could only be verified by the application of other techniques, such as western blotting or targeted proteomic approaches. Figure 4.20 shows an estimation of protein abundance for selected proteins after these approaches were re-applied to the standards that were used to create a response factor or a calibration curve. In general, the response factor methods were more accurate than those used using a calibration curve. The accuracy is reasonable for linear regression methods when calibration points from separate LC-MS experiments were utilised. An internal standard offers advanatges over external calibrants since it reflects the response of the mass spectrometer at the same time as sample analysis.

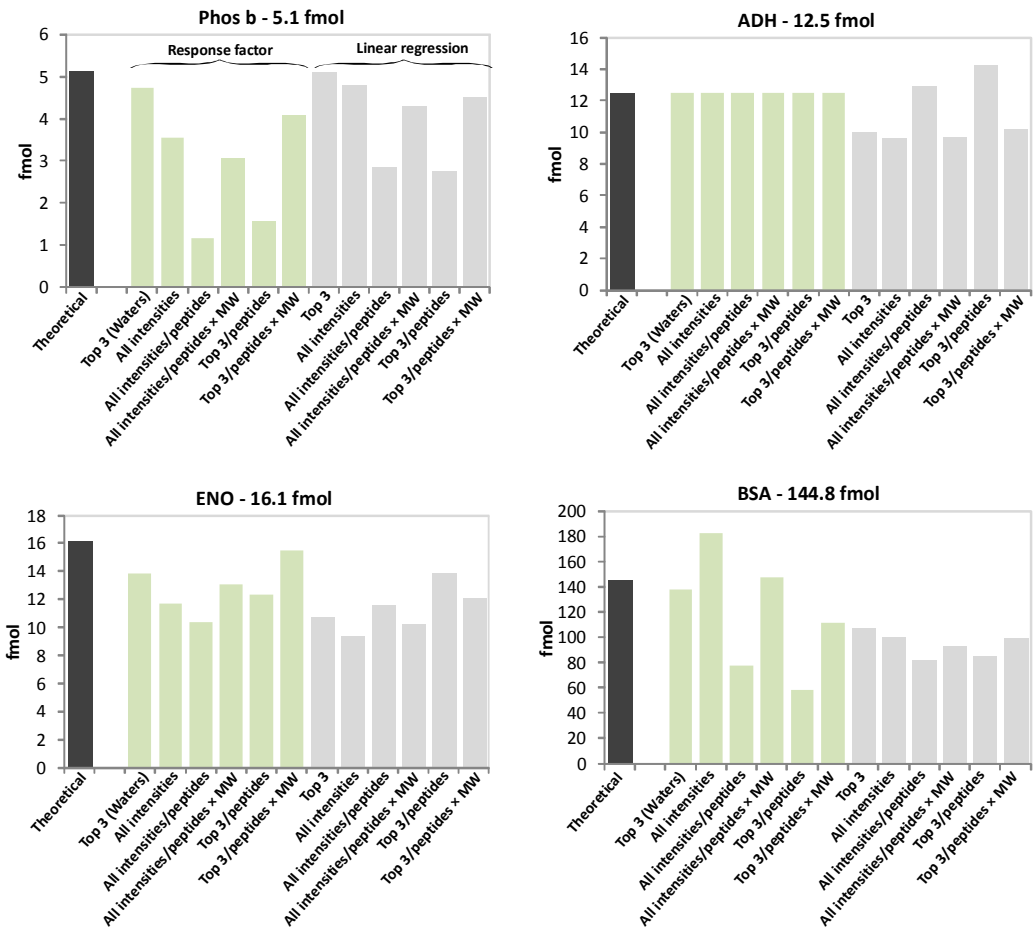

Figure 4.20. Various quantification methods applied to standard proteins

4.4.7 Comparison of 2D RP-RP-LC HDMS^E to current mass spectrometry-based proteomic approaches

Large scale proteomic studies have largely been undetaken using the model organism Yeast. A chronological review of the study of this organism shows the improvement in mass spectrometry-based approaches, which started with identifying hundreds (Shevchenko *et al.* 1996) but, more recently, thousands of proteins (de Godoy *et al.* 2008). These studies were carried out with an aim of identifying as many proteins as possible, and as such, involve time consuming protein and/or peptide fractionation steps and the combination of two or more experimental approaches. A summary of the study carried out by de Godoy *et al.* may be found in Table 4.6. A recent study was carried by Nagaraj *et al.* in 2012, with the aim of providing a comprehensive coverage of the Yeast proteome, and an alternative goal of reducing sample requirement and experimental time. The major technical aspects of these two studies

on Yeast may be compared with those in a 2D RP-RP-LC HDMS^E experiment. This is not an entirely fair comparison since the theoretical proteome size of Yeast is approximately four-times larger than the soluble proteome of *M. silvestris*. The comparison does however allow important practical experimental comparisons to be made. Both the Yeast experiments were conducted using Orbitrap-based mass spectrometers, which are becoming widely adopted by proteomics-focussed laboratories. In these studies, the Orbitraps were operated at a mass resolution approximately 3-fold higher than that used in the Synapt G2 experiments. The de Godoy method combined three experimental approaches, the first requiring a large amount of sample for protein fractionation (100 μg per lane for 1D SDS-PAGE) and the others requiring a large amount of sample for offline IEF peptide fractionation (75 μg). This study combined the total number of protein identifications from three approaches. In comparison to the online approach taken here for *M. silvestris*, the de Godoy approach took a significantly longer amount of experimental time (up to 34-fold more time for a single replicate), requiring larger amounts of sample for MS analysis (up to 250-fold more for a single replicate), ultimately achieving a similar quantitative dynamic range and proteome coverage.

The study of Nagaraj *et al.* was a 1D RP-LC approach, which does not involve additional protein/peptide fractionation prior to MS analysis. As such, this online method is comparable to the online 2D RP-RP-LC approach taken here. The main technological advancement which enabled an excellent coverage of the Yeast proteome using the Nagaraj approach was the use of longer LC-MS columns (50 cm compared to 15 cm for the other approaches). Similar experimental times (5 – 7 hr) and sample requirements (2 – 4 μg) were required for the 1D RP-LC approach with long columns and the 2D RP-RP-LC approach used here. Both of these approaches achieved a similar coverage of the theoretical proteome.

An additional consideration for the selection of a suitable proteomic approach can be availability of equipment. The 1D RP-LC approach used for the study of Yeast appears an attractive option, which does not require the expertise required for the operation of online 2D systems or the labour required for offline 2D systems. The long column used in this particular experiment, however, was packed in-house and was not commercially available.

	Methylocella silvestris	Yeast (de Godoy *et al.* 2008)	Yeast (Nagaraj *et al.* 2012)
Proteome type	Soluble	Entire	Entire
Theoretical number of ORFs	~1,700	~6,600	~6,600
Experimental description	A single 2D method with no offline fractionation steps in order to simultaneously identify and quantify proteins.	Three methods all including an initial offline fractionation of peptides/proteins. Identification results are combined for all approaches.	A single uni-dimensional LC approach utilising ultra performance LC with longer columns.
Peptide/protein fractionation	Online 2D RP-RP-LC (3 fractions) for peptides	1. Protein fractionation (5 fractions) SDS-PAGE (15 slices) 2. Peptide IEF (24 fractions) 3. Peptide IEF (24 fractions) All followed by 1D RP-LC	1D RP-LC for peptides
LC-MS column	75 μm × 150 mm, 1.8 μm Column commercially available	75 μm × 150 mm, 3 μm Column packed in-house	75 μm × 500 mm, 1.8 μm Column packed in-house
Mass spectrometer	Synapt G2, HDMS^E	LTQ-Orbitrap, DDA or mass range selected DDA	Q-Exactive (Quadrupole-Orbitrap), DDA
Experimental time (fractionation + MS) (1 rep)	7 hr	1. ~150 hr (DDA) 2. ~80 hr (DDA) 3. ~240 hr (mass selected DDA)	5 hr
Mass resolution	18,000	60,000 at m/z 400	50,000 at m/z 200
Sample requirement for LC-MS analysis	2 μg	1. 500 μg 2. 75 μg 3. 75 μg	4 μg
Quantification **Dynamic range**	Label-free 10^3 – 10^4	SILAC $>10^4$	SILAC (spiked-in standard) 10^5 (based on peptide intensities)
Proteome coverage	67%	67%	58%

Table 4.6. Comparison of 2D RP-RP-LC HDMS^E to current mass spectrometry-based approaches

4.5 Conclusions

Ion mobility separation has been shown to be achievable within the timescale of LC and mass analysis. It has been utilised here as an experimental component of a label-free proteomic study using data-independent acquisition. Obtaining added mobility data does not add to the timescale of the experiment, but the information associated with this additional measurement does significantly increase the size of data files (by a factor of two) and requires powerful computation hardware for effective and timely processing.

A non-biased comparison of MS^E and mobility-assisted MS^E methods was carried out for the analysis of bacterial extracts obtained using different growth conditions and simple standard protein mixtures. A reduction in spectral congestion in extracted product ion spectra was observed when using the mobility-assisted approach. This resulted in higher confidence and better quality assignments from database searching. An increased number of peptide and protein identifications were achieved using the mobility-assisted approach and these were more reproducible as a result of the reduction in observation of random low intensity ions, which tend not to be replicated.

For ion mobility to be established within a proteomics workflow in the future, the ability to reproducibly and accurately quantify peptide intensity is of great importance. Good agreement with the conventional MS^E method was obtained. A small reduction in effective experimental dynamic range was observed with the 1D RP-LC $HDMS^E$ experiment due to detector saturation under the stringent requirements of the mobility-assisted data acquisition. This was improved when 2D RP-RP-LC $HDMS^E$ experiments were employed. The method combines the benefits of both 2D-LC and ion mobility and was shown to be comparable to results obtained using current approaches that aimed to provide comprehensive proteome coverage. Similar proteome coverage and dynamic ranges were obtained, with experimental advantages that include less sample requirement, reduced experimental time and easier availability.

Alternative label-free approaches were also investigated and compared to the Hi3 approach. Further work is required to investigate the potential benefit of using an approach which utilised the sum of all peptides and multiple calibration points.

4.6 References

Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F., Prior, D. C., Ibrahim, Y. M., LaMarche, B. L., Mayampurath, A. M., Schepmoes, A. A., Hopkins, D. F., Tang, K., Smith, R. D. and Belov, M. E. (2010). An LC-IMS-MS Platform Providing Increased Dynamic Range for High-Throughput Proteomic Studies. *J. Proteome Res.* **9**, 997-1006.

Belov, M. E., Clowers, B. H., Prior, D. C., Danielson III, W. F., Liyu, A. V., Petritis, B. O. and Smith, R. D. (2008). Dynamically Multiplexed Ion Mobility Time-of-Flight Mass Spectrometry. *Anal. Chem.* **80**, 5873-5883.

de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C. and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature.* **455**, 1251-1254.

Giles, K., Pringle, S. D., Worthington, K. R., Little, D., Wildgoose, J. L. and Bateman, R. H. (2004). Applications of a travelling wave-based radio-frequency only stacked ring ion guide. *Rapid Commun. Mass Spectrom.* **18**, 2401-2414.

Giles, K., Williams, J. P. and Campuzano, I. (2011). Enhancements in travelling wave ion mobility resolution. *Rapid Commun. Mass Spectrom.* **25**, 1559-1566.

Guevremont, R., Barnett, D. A., Purves, R. W. and Vandermey, J. (2000). Analysis of a Tryptic Digest of Pig Hemoglobin Using ESI-FAIMS-MS. *Anal. Chem.* **72**, 4577-4584.

Hao, P., Guo, T., Li, X., Adav, S. S., Yang, J., Wei, M. and Sze, S. K. (2010). Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun Proteomics: Comprehensive Profiling of Rat Kidney Proteome. *J. Proteome Res.* **9**, 3520-3526.

Hilton, G. R., Jackson, A. T., Thalassinos, K. and Scrivens, J. H. (2008). Structural Analysis of Synthetic Polymer Mixtures Using Ion Mobility and Tandem Mass Spectrometry. *Anal. Chem.* **80**, 9720-9725.

Hoaglund-Hyzer, C. S. and Clemmer, D. E. (2001). Ion Trap/Ion Mobility/Quadrupole/Time-of-Flight Mass Spectrometry for Peptide Mixture Analysis. *Anal. Chem.* **73**, 177-184.

Ibrahim, Y. M., Prior, D. C., Baker, E. S., Smith, R. D. and Belov, M. E. (2011). Characterization of an ion mobility-multiplexed collision-induced dissociation-

tandem time-of-flight mass spectrometry approach. *Int. J. Mass Spectrom.* **293**, 34-44.

Kim, M.-S., Kandasamy, K., Chaerkady, R. and Pandey, A. (2010). Assessment of Resolution Parameters for CID-Based Shotgun Proteomic Experiments on the LTQ-Orbitrap Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **21**, 1606-1611.

Koeniger, S. L., Valentine, S. J., Myung, S., Plasencia, M., Lee, Y. J. and Clemmer, D. E. (2005). Development of Field Modulation in a Split-Field Drift Tube for High-Throughput Multidimensional Separations. *J. Proteome Res.* **4**, 25-35.

Kurulugama, R. T., Valentine, S. J., Sowell, R. A. and Clemmer, D. E. (2008). Development of a high-throughput IMS-IMS-MS approach for analyzing mixtures of biomolecules. *J. Proteomics.* **71**, 318-331.

Liu, X., Valentine, S. J., Plasencia, M. D., Trimpin, S., Naylor, S. and Clemmer, D. E. (2007). Mapping the Human Plasma Proteome by SCX-LC-IMS-MS. *J. Am. Soc. Mass Spectrom.* **18**, 1249-1264.

Moon, M. H., Myung, S., Plasencia, M., Hilderbrand, A. E. and Clemmer, D. E. (2003). Nanoflow LC/Ion Mobility/CID/TOF for Proteomics: Analysis of a Human Urinary Proteome. *J. Proteome Res.* **2**, 589-597.

Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O. and Mann, M. (2012). System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics.* **11**.

Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T. and Scrivens, J. H. (2007). An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass spectrom.* **261**, 1-12.

Riba-Garcia, I., Giles, K., Bateman, R. H. and Gaskell, S. J. (2008). Evidence for structural variants of a- and b-type peptide fragment ions using combined ion Mobility/Mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**, 609-613.

Rodríguez-Suárez, E., Hughes, C., Gethings, L., Giles, K., Wildgoose, J., Stapels, M., Fadgen, K. E., Geromanos, S. J., Vissers, J. P. C., Elortza, F. and Langridge, J. I. (2011). An Ion Mobility Assisted Data Independent LC-MS Strategy for the Analysis of Complex Biological Samples. *Curr. Anal. Chem.* **In press**.

Saba, J., Bonneil, E., Pomiès, C., Eng, K. and Thibault, P. (2009). Enhanced Sensitivity in Proteomics Experiments Using FAIMS Coupled with a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *J. Proteome Res.* **8**, 3355-3366.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature.* **473**, 337-342.

Shah, A. R., Agarwal, K., Baker, E. S., Singhal, M., Mayampurath, A. M., Ibrahim, Y. M., Kangas, L. J., Monroe, M. E., Zhao, R., Belov, M. E., Anderson, G. A. and Smith, R. D. (2010). Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bioinformatics.* **26**, 1601-1607.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H. and Mann, M. (1996). Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci.* **93**, 14440-14445.

Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G.-Z., McKenna, T., Nold, M. J., Richardson, K., Young, P. and Geromanos, S. (2005). Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Anal. Chem.* **77**, 2187-2200.

Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. and Geromanos, S. J. (2006). Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel MS Acquisition. *Mol. Cell. Proteomics.* **5**, 144-156.

Taraszka, J. A., Gao, X., Valentine, S. J., Sowell, R. A., Koeniger, S. L., Miller, D. F., Kaufman, T. C. and Clemmer, D. E. (2005a). Proteome Profiling for Assessing Diversity: Analysis of Individual Heads of *Drosophila melanogaste*r Using LC-Ion Mobility-MS. *J. Proteome Res.* **4**, 1238-1247.

Taraszka, J. A., Kurulugama, R., Sowell, R. A., Valentine, S. J., Koeniger, S. L., Arnold, R. J., Miller, D. F., Kaufman, T. C. and Clemmer, D. E. (2005b). Mapping the proteome of *Drosophila melanogaster*: Analysis of embryos and adult heads by LC-IMS-MS methods. *J. Proteome Res.* **4**, 1223-1237.

Thalassinos, K., Grabenauer, M., Slade, S. E., Hilton, G. R., Bowers, M. T. and Scrivens, J. H. (2008). Characterization of Phosphorylated Peptides Using Traveling Wave-Based and Drift Cell Ion Mobility Mass Spectrometry. *Anal. Chem.* **81**, 248-254.

Valentine, S. J., Counterman, A. E., Hoaglund, C. S., Reilly, J. P. and Clemmer, D. E. (1998). Gas-phase separations of protease digests. *J. Am. Soc. Mass Spectrom.* **9**, 1213-1216.

Valentine, S. J., Ewing, M. A., Dilger, J. M., Glover, M. S., Geromanos, S., Hughes, C. and Clemmer, D. E. (2011). Using Ion Mobility Data to Improve Peptide Identification: Intrinsic Amino Acid Size Parameters. *J. Proteome Res.* **10**, 2318–2329.

Valentine, S. J., Kulchania, M., Barnes, C. A. S. and Clemmer, D. E. (2001). Multidimensional separations of complex peptide mixtures: a combined high-performance liquid chromatography/ion mobility/time-of-flight mass spectrometry approach. *Int. J. Mass Spectrom.* **212**, 97-109.

Valentine, S. J., Kurulugama, R. T., Bohrer, B. C., Merenbloom, S. I., Sowell, R. A., Mechref, Y. and Clemmer, D. E. (2009). Developing IMS-IMS-MS for rapid characterization of abundant proteins in human plasma. *Int. J. Mass Spectrom.* **283**, 149-160.

Valentine, S. J., Plasencia, M. D., Liu, X. Y., Krishnan, M., Naylor, S., Udseth, H. R., Smith, R. D. and Clemmer, D. E. (2006). Toward plasma proteome profiling with ion mobility-mass spectrometry. *J. Proteome Res.* **5**, 2977-2984.

Venne, K., Bonneil, E., Eng, K. and Thibault, P. (2005). Improvement in Peptide Detection for Proteomics Analyses Using NanoLC-MS and High-Field Asymmetry Waveform Ion Mobility Mass Spectrometry. *Anal. Chem.* **77**, 2176-2186.

Chapter 5

Proteomic Studies on Methylated Amine Metabolism

5.1 Introduction

Methylated amine compounds, including monomethylamine (MMA) and trimethylamine (TMA), are ubiquitous in the environment. The putrefaction of food products (Kamiya and Ose 1984) and degradation of nitrogen-containing pesticides (Bhadbhade *et al.* 2002) can release methylated amines. In the marine environment, methylated amines are released from the degradation of quaternary amines, which marine organisms use for regulating osmotic-related stress (Welsh 2000). The microbial oxidation of methylated amines is a key part of the environmental cycling of carbon and nitrogen, as well as an essential step in preventing formation of methane, the second most important greenhouse gas.

Methylated amines, including monomethylamine (CH_3NH_2), dimethylamine $(CH_3)_2NH$ and trimethylamine $(CH_3)_3N$, are substrates that can be utilised as a sole source of carbon and nitrogen by methylotrophic bacteria. Methylotrophs have evolved to accomplish their metabolic goals through a combination of pathways, for instance, the assimilation of formaldehyde can occur via the serine cycle or the ribulose monophosphate cycle. Similarly, methylated amine metabolism pathways vary between methylotrophs.

5.1.1 Trimethylamine metabolism

There are two possible pathways which have been proposed for the oxidation of trimethylamine (Figure 5.1). The first of these routes involves oxidative *N*-demethylation of TMA by the enzyme TMA dehydrogenase, yielding dimethylamine and formaldehyde directly (Colby and Zatman 1973). The existence of the enzyme TMA monooxygenase was proposed by Large *et al.* to explain the ability of *Aminobacter aminovorans* to grow on trimethylamine since there was no evidence for a TMA dehydrogenase (Large *et al.* 1972; Boulton *et al.* 1974). In this pathway, termed the indirect pathway, TMA is converted to trimethylamine *N*-oxide (TMAO), which is then ultimately converted to formaldehyde (assimilated as carbon) and ammonium (assimilated as nitrogen) via the intermediate MMA. Even though the activity of trimethylamine monooxygenase has been known for over four decades, the encoding gene has never been identified. There is no indication from the *M. silvestris*

genome of the existence of this enzyme. There is also the absence of a coded protein with significant homology to known trimethylamine dehydrogenases.

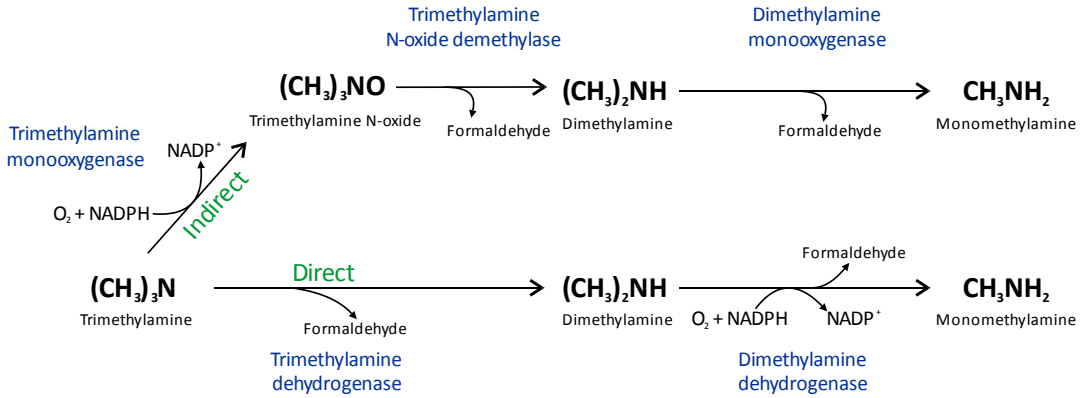

Figure 5.1. Direct and indirect metabolism of trimethylamine

5.1.2 Monomethylamine metabolism

The majority of studies carried out on MMA oxidation in bacteria were undertaken in the 1960s. Two major routes for methylamine oxidation were proposed. One of these was the direct oxidation of MMA, carried out by methylamine dehydrogenase in proteobacteria (Eady and Large 1968) and by methylamine oxidase in Gram-positive bacteria (Zhang *et al.* 1993). The second route entails the conversion of methylamine to methylated amino acids *N*-methylglutamate (NMG) (Shaw *et al.* 1966), *N*-methylalanine (Lin and Wagner 1975) or γ-glutamylmethylamide (GMA) (Kung and Wagner 1969), followed by oxidation to formaldehyde and the respective amino acids. It is believed that the pathways involving NMG and GMA are separate, although recent data suggests that they may be in the same pathway (Chen *et al.* 2010).

The genetics of the enzymes involved in the direct pathway have been well characterised. Only recently have the genes involved in a pathway including NMG and GMA as intermediates been characterised in *Methyloversatilis universalis* FAM5 (Latypova *et al.* 2010) and *Methylocella silvestris* BL2 (Chen *et al.* 2010). In this study the proteins induced by growth on MMA in *M. silvestris* BL2 from soluble and membrane fractions were analysed by means of 1D gels and MALDI-MS.

5.2 Aims and Methods

The genes relating to the indirect pathway for monomethylamine metabolism have only recently been characterised in *M. silvestris*, and the genes relating to indirect metabolism of trimethylamine have never been identified in any organism. The main aims of this work were to identify and quantify the enzymes involved in the respective metabolism pathways and to determine relative expression levels downstream of the initial metabolic processes.

Soluble extracts obtained from *M. silvestris* grown on trimethylamine, monomethylamine or methanol + ammonium were tryptically digested, spiked with an internal standard and subjected to 1D RP-LC MS^E analysis. Relative expression analysis was performed between TMA:MMA, MMA:MeOH and TMA:MeOH. RT-PCR was carried out to validate the expression of selected transcripts. Further details are outlined in Chapter 2.

This work was combined with genetic, biochemical and bioinformatic studies carried out by Yin Chen and Andrew Crombie and published:

Chen, Y., Patel, N. A., Crombie, A., Scrivens, J. H. and Murrell, J. C. (2011). Bacterial flavin-containing monooxygenase is trimethylamine monooxygenase. *Proceedings of the National Academy of Sciences.* **108**, 17791-17796.

5.3 Results and Discussion

5.3.1 Proteome coverage and data assessment

Using the criteria outlined in Chapter 2, 309, 265 and 116 proteins were confidently identified from methanol (MeOH), monomethylamine (MMA) and trimethylamine (TMA) samples, respectively. Figure 5.2 illustrates a 60% overlap obtained between these data sets, resulting in the identification of 378 non-redundant proteins. This significant overlap confirmed the commonality of methylotrophic metabolism downstream of primary oxidation reactions. On average, proteins were identified by matching 11 peptides with a sequence coverage of 32%. The quality of data is presented for selected proteins listed in Tables 5.1 and 5.2. A low false positive rate was calculated for all datasets (<1%).

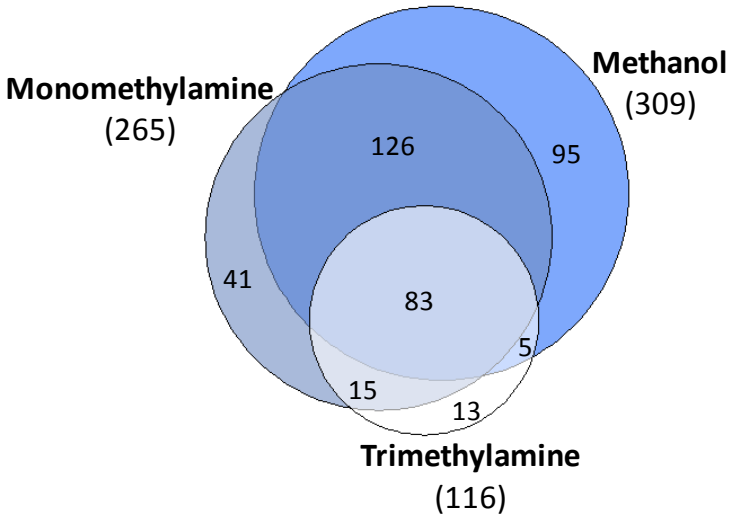

Figure 5.2. Proteins identified from trimethylamine, monomethylamine and methanol extracts

A core set of 83 proteins were identified in all three substrates. 76 out of 83 of the commonly identified proteins were categorised using the DAVID tool (Huang *et al.* 2008; Huang *et al.* 2009). Figure 5.3 shows the biological processes, molecular functions and pathways that were core to all three studied substrates. These 83 proteins made up 22% of the non-redundant protein set in terms of identifications, but contributed to the majority of the detected quantity in all three substrates (MeOH 63%, MMA 55%, TMA 80%). The functional analysis tool revealed that these proteins were involved in essential biological processes including protein folding,

energy and acetyl-CoA processes, oxidation reduction reactions and generation of precursor metabolites.

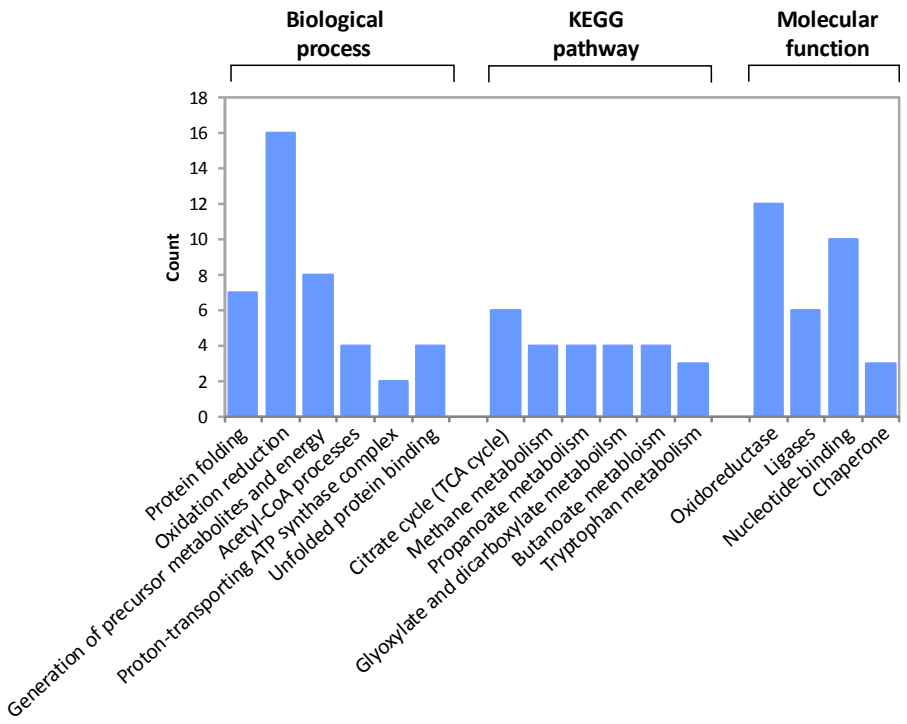

Figure 5.3. Functional analysis of proteins common to TMA-, MMA- and MeOH-grown extracts

Figure 5.4 shows the predicted primary locations of the theoretically expressed proteins from the genome, and from the 378 non-redundant proteins observed in MeOH, MMA and TMA extracts. From the genome, approximately a third of proteins were assigned as Unknown. Of this subsection, 56% have Hypothetical or Unknown descriptions. A smaller proportion of observed proteins (23%) were assigned with unknown locations, however, the majority of these proteins were assigned with homology-based descriptions (80%), and as such were more likely to be true proteins. The number of hypothetical proteins that remain in a proteomics data set can sometimes be used as a marker of quality, analogous to the use of a randomised database to generate a false positive rate. *M. silvestris* is a relatively less well characterised organism, and has not been updated with experimental verification on a regular basis compared to more widely-studied organisms. The presence of proteins with unknown descriptions and/or locations cannot therefore be explained as false positive identifications.

When assessing the proteome coverage, there are a number of factors that need to be considered. Functional and sequence redundancy is present in the genome and a number of ORFs can encode proteins with the same function. In some cases only one of these is actually expressed. There are some methanotrophs that have been found to encode both RuMP and serine cycle enzymes, but have only been shown to utilise one of these pathways. Expression of proteins is an energetically costly activity, and therefore, it is beneficial for bacteria to express certain proteins only under conditions where those proteins are needed for say growth or survival. It is difficult to assess the true coverage of the theoretical proteome without taking into account these factors. It is likely that cytoplasmic and periplasmic proteins are most likely to be present in soluble extracts, which in this case corresponds to 1685 proteins from the theoretically expressed genome. It is clear that cytoplasmic and periplasmic proteins are enriched when the soluble protein extract protocol is followed.

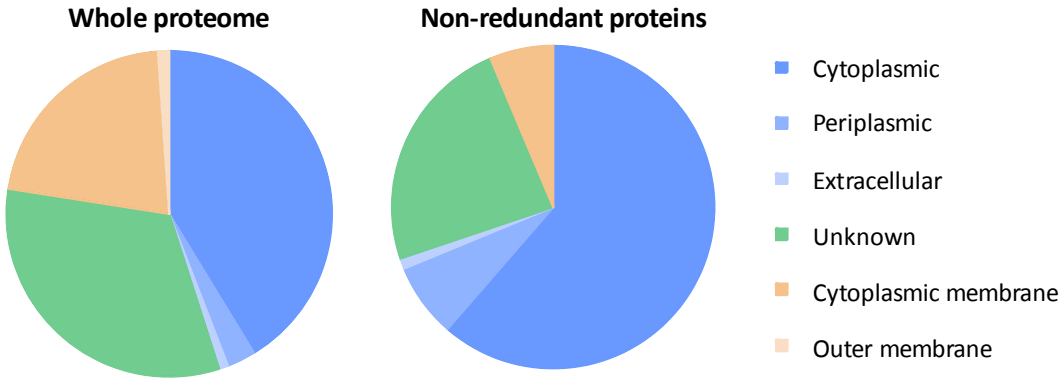

Figure 5.4. Cellular localisation of (i) predicted proteome (ii) non-redundant identifications

A small proportion of expressed proteins (6%) were predicted to be primarily located in the cytoplasmic membrane. Further interrogation of these proteins revealed that most of them were known membrane proteins including transporters, unknown/hypothetical membrane proteins and succinate dehydrogenase. Succinate dehydrogenase is a heterotetramer, which then trimerizes to form a functional complex. The complex comprises two hydrophilic subunits, a flavoprotein (alpha) and iron-sulfur protein (beta) subunit, and two hydrophobic membrane anchor subunits, (delta and gamma). Only the alpha subunit was identified from this complex. This could be explained by the fact that it was positioned in the cytoplasm and was furthest away from the cytoplasmic membrane within the complex.

The number of proteins identified in the TMA soluble extract (116) was notably lower when compared to those identified in MMA and MeOH extracts (~300). The composition of a mixture of proteins can have a significant effect on the number of proteins that can be identified using an LC-MS approach. The analysis of human plasma by 1D RP-LC MS^E typically results in the identification of 40 - 90 proteins. After depletion of the 14 most abundant proteins, which comprise 90% of the total amount of protein content by weight, the number of identifications made from LC-MS analyses increases significantly (~3-fold) (Slade 2012).

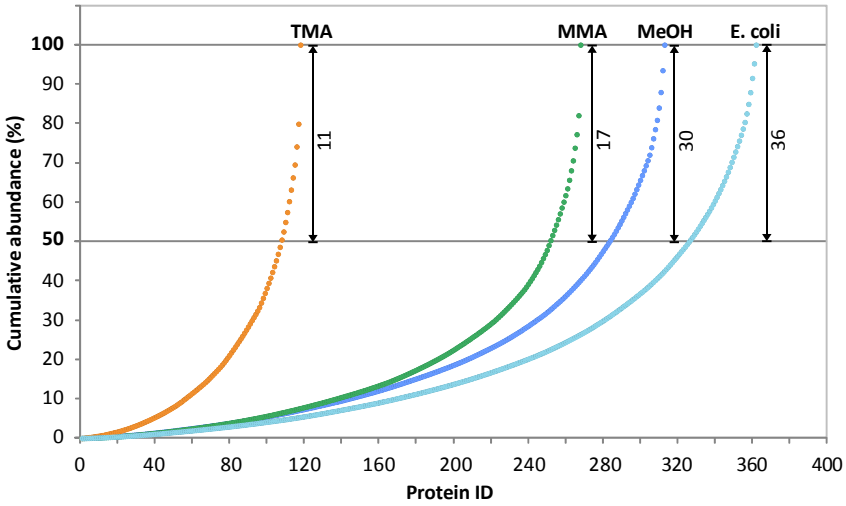

Figure 5.5. Normalised cumulative abundances for proteins in MeOH, MMA, TMA and *E. coli* extracts

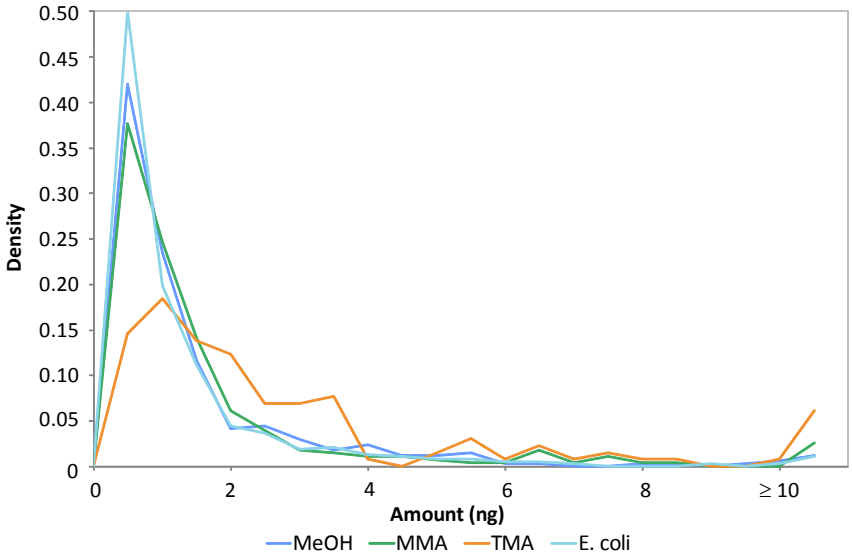

Figure 5.6. Protein abundance distributions for proteins inMeOH, MMA, TMA and *E. coli* extracts

Figure 5.5 shows the cumulative protein abundances which were obtained from the 1D RP-LC MS^E analysis of TMA, MMA and MeOH grown-extracts compared to the *E. coli* proteome. The plot provides an indication of the number of proteins that constitute 50% of the total amount of protein detected. Fewer proteins were identified when a small proportion of proteins are present at higher concentrations. This can be clearly seen in the TMA extract, where 11 proteins account for 50% of the protein content compared to 36 proteins in the *E. coli* extract. The density plot in Figure 5.6 shows the distribution of protein abundances for these studied substrates. This also indicates a larger proportion of proteins are present at more than 10 ng and fewer proteins less than 2 ng for TMA proteins when compared to other conditions.

5.3.2 Highly up-regulated cluster during MMA and TMA growth are *N*-methylglutamate pathway enzymes

All the enzymes which have been proposed to be involved in the *N*-methylglutamate pathway (Msil_2632 to 2639) for monomethylamine utilisation were identified and quantified in MMA- and TMA-extracts, but not in methanol-extracts, when the label-free 1D RP-LC MS^E approach was used. Five components, within this cluster, have been previously identified from *M. silvestris* MMA-grown soluble or membrane fractions separated by 1D SDS-PAGE and analysed by MALDI-MS. It was shown that transcripts relating to eight of these proteins were absent in methanol-grown cells (Chen *et al.* 2010). Table 5.1 shows the descriptions for all eight proteins based on BLASTP similarity and gene assignment based on similarity to the experimentally characterised *N*-methylglutamate genes in *Methyloversatilis universalis* FAM5. The three main enzymes (*genes*) involved in this pathway are *N*-methylglutamate synthtase (*mgsABC*), γ-glutamylmethylamide synthetase (*gmaS*) and *N*-methylglutamate dehydrogenase (*mgdABCD*). It was experimentally verified that γ-glutamylmethylamide synthetase is expressed by Msil_2635, and it is likely that Msil_2632 to Msil_2634 and Msil_2636 to Msil_2639 are *N*-methylglutamate synthase and *N*-methylglutamate dehydrogenase, respectively (Chen *et al.* 2010).

Table 5.1 compares the results obtained from 1D RP-LC MS^E to the published 1D SDS-PAGE analysis in terms of number of peptides matched and sequence

coverage, with additional quantitative information for MS^E identified proteins. On average, proteins were identified with 2.7-fold additional peptides with a 1.7-fold increase in protein coverage. It is perhaps surprising that Msil_2634 was not identified from the 1D SDS-PAGE analyses since it was shown by the 1D RP-LC analyses to be the most abundant protein in this cluster, constituting 3.17% and 1.41% of total detected protein in MMA and TMA extracts, respectively. The proteins from Msil_2632 – 2639 were highly expressed in methylamine-grown cells, resulting in 9.5% and 6.5% of the total detected protein by weight in MMA and TMA extracts, respectively.

Figure 5.8(a-b) shows the abundance of the subunits relating to the enzymes *N*-methylglutamate synthase, γ-glutamylmethylamide synthetase and *N*-methylglutamate dehydrogenase estimated using the Hi3 label-free quantification approach, and where they are proposed to be placed in the pathway. Since this approach provided an estimation of molar quantity, subunit stoichiometry could be inferred to some extent. The majority of structural studies carried out on methylamine-related enzymes were carried out in the 60s and 70s, typically by means of native gels or sedimentary methods, which provided an estimate for the molecular weight of whole protein complexes and individual subunits. The data obtained from the label-free experiments have been compared to these early studies to provide additional insight into some overlooked enzymes.

γ-Glutamylmethylamide synthetase (Msil_2635)

γ-Glutamylmethylamide synthetase is proposed to be the initial enzyme in the pathway, acting on monomethylamine and glutamate to generate γ-glutamylmethylamide. Msil_2635 encodes a subunit of 48.4 kDa which was present at 240 fmol in MMA-extracts, 10-fold higher than the subunits relating to downstream enzyme *N*-methylglutatamte dehydrogenase (present at ~20 fmol) and similar to those subunits representing *N*-methylglutamate synthase (150 – 350 fmol). Previous studies have shown that in *Methylovorus mays*, this enzyme has a native molecular weight between 410 – 470 kDa (estimated by means of gel filtration) with 51 kDa subunits (estimated by means of SDS-PAGE) (Yamamoto *et al.* 2007). Kimura *et al.* also demonstrated this enzyme in *Methylophaga* sp. AA-30 had a native molecular weight of 440 kDa and existed as a homo-octomer (Kimura *et al.* 1992).

N-Methylglutamate synthase (Msil_2632, Msil_2633 and Msil_2634)

The substrate of *N*-methylglutamate synthase is a contentious issue in the study of methylotrophy. It was first proposed to act directly on monomethylamine and glutamate to form *N*-methylglutamate (Shaw *et al.* 1966). In the proposed pathway for *M. silvestris* this enzyme was thought to use γ-glutamylmethylamide, not monomethylamine, as a substrate. Recent studies have shown the production of formaldehyde and ammonium when γ-glutamylmethylamide was incubated with glutamate and cell-free extracts of *M. silvestris,* indicating that γ-glutamylmethylamide may be transformed to *N*-methylglutamate, which can be further oxidised by *N*-methylglutamate dehydrogenase to yield formaldehyde.

The enzyme *N*-methylglutamate synthase comprises about 5% of the soluble protein in *Pseudomonas* MA (ATCC 23819). In *M. silvestris*, it was calculated to be 5.4% and 2.8% of the detected amount in MMA- and TMA- extracts, respectively. This enzyme was originally identified in *Pseudomonas* MA, now commonly known as *Aminobacter aminovorans*, by Pollock and Hersh in 1971. By carrying out sedimentation experiments they estimated the native enzyme to have a molecular weight of 350,000, which was composed of approximately 12 subunits (30,000 to 35,000 molecular weight)(Pollock and Hersh 1971). Since these early studies, no experiments have been carried out to determine the structural composition of this enzyme. Although the genetics of this enzyme was recently characterised in *M. universalis* FAM5, the sequence for all three subunits was incomplete and the spectral counting method used in these studies only allowed differentially expressed proteins to be analysed. Using the predicted molecular weights from the sequences of Msil_2632 to Msil_2634 (32.0, 24.7, and 46.9 kDa) and the abundance of proteins estimated using the label-free approach used here, a complex containing 9 subunits with stoichiometry of 2: 3.0 (±0.30): 4.1 (±0.28) is implied, which amounts to an estimated molecular weight of 325,753. *Aminobacter aminovorans* does not have a fully annotated genome so the stoichiometry estimated here cannot be applied and compared to the estimation of 350 kDa. Purification of this enzyme would aid in the determination of the number and type of subunits and cofactors and could resolve the issues associated with substrate specificity.

ORF	Molecular mass (kDa)	Top BLASTP match (identity [%])	Gene assignment[a]	1D SDS-PAGE gel + MALDI-MS[b]		1D RP-LC + ESI-MS[E]			
				No. of peptides detected	Sequence coverage (%)	No. of peptides detected	Sequence coverage (%)	Protein abundance (fmol)	Protein abundance (%)[c]
Msil2632	32.0	Glutamine amidotransferase, class II (*Azohizobium caulinodans* ORS 571 [67]) Glutamine amidotransferase, class II (*Xanthobacter autotrophicus* Py2 [65])	*mgsA*	5	21	20	62	160	1.05
Msil2633	24.7	Glutamate synthase domain 3-like protein (*Methylobacterium populi* BJ001 [72]) Glutamate synthase family protein (*Xanthobacter autotrophicus* Py2 [66])	*mgsB*	4	24	15	60	241	1.22
Msil2634	46.9	Putative glutamate synthase GltB2 subunit (*Methylobacillus flagellatus* KT) [87]) Glutamate synthase domain 2-like protein (*Azohizobium caulinodans* ORS 571 [86])	*mgsC*	-	-	36	60	330	3.17
Msil2635	48.4	Putative type III glutamine synthetase (*Rhodopseudomonas palustris* HaA2 [69]) γ-Glutamylmethylamide synthetase (*Methylovorus mays* [41])	*gmaS*	7	19	26	54	240	2.37
Msil2636	44.8	Putative sarcosine oxidase β subunit (*Azohizobium caulinodans* ORS 571 [75]) Sarcosine oxidase β subunit (*Corynebacterium* sp. P-1 [49])	*mgdA*	6	18	14	24	24	0.22
Msil2637	10.3	Putative sarcosine oxidase, δ subunit (*Bradyrhizobium* sp. BTAi1 [55]) Sarcosine oxidase δ subunit (*Corynebacterium* sp. P-1 [37])	*mgdB*	-	-	6	62	26	0.05
Msil2638	104.3	Putative sarcosine oxidase α subunit (*Methylobacterium populi* BJ001 [55]) Sarcosine oxidase α subunit (*Corynebacterium* sp. P-1 [36])	*mgdC*	7	10	31	39	61	1.31
Msil2639	21.8	Putative sarcosine oxidase γ subunit (*Rhodopseudomonas palustris* HaA2 [37]) Sarcosine oxidase γ subunit (*Corynebacterium* sp. P-1 [31])	*mgdD*	-	-	4	27	30	0.14

Table 5.1. Mass spectrometry and BLAST search results of proteins induced by monomethylamine compared to published data

[a] Based on similarity to *Methyloversatilis universalis* (Latypova *et al.* 2010); [b] Average from soluble and membrane fractions (Chen *et al.* 2010); [c] Proportion of total protein detected

N-Methylglutamate dehydrogenase (Msil_2636 – Msil_2639)

The enzyme *N*-methylglutamate dehydrogenase uses *N*-methylgultamate as a substrate to generate a molecule of glutamate and formaldehyde. This enzyme has been characterised in a variety of bacteria and has shown to be composed of different subunits, require various cofactors and be active in soluble and membrane fractions. Bamforth and Large described the partial purification of this enzyme from *Pseudomonas aminovorans*, demonstrating the enzyme's resemblance to integral membrane properties like those exhibited by succinate dehydrogenase (Bamforth and Large 1977b). An improved method of purification showed that *N*-methylglutamate dehydrogenase gave one band on an SDS-PAGE corresponding to a molecular weight of 130,000. Binding studies with detergent also suggested it existed as a homotetramer (Bamforth and Large 1977a). In the bacterium AT2, *N*-methylglutamate dehydrogenase was shown to exhibit more extrinsic membrane properties rather than intrinsic, with a native molecular weight of 407 kDa consisting of three to four subunits with molecular weight of 108 kDa (Boulton *et al.* 1980).

In *M. silvestris*, this enzyme does not appear to exist as a homotetramer as it has a predicted subunit composition of 44.8, 10.3, 104.3 and 21.8 kDa, which were estimated to be present at 24, 26, 61 and 30 fmol, respectively. This enzyme is similar to that recently described in *M. universalis* FAM5, which has a predicted subunit composition of 44, 100, 12, and 10 kDa (Latypova *et al.* 2010). Unlike *M. universalis* FAM5, however, *N*-methylglutamate dehydrogenase has been shown to be rather abundant in the membrane fraction of *M. silvestris*, as suggested from intensity staining on a 1D SDS-PAGE (Chen *et al.* 2010). These results show that this enzyme is also abundant in the soluble fraction. A possible explanation for the enzyme's multiple locations in *M. silvestris* is that it may contain both integral membrane components and cytoplasmic subunits. Bioinformatic analysis predicts a trans-membrane region in only one of the four subunits. Figure 5.7 shows the consensus of four prediction tools for the location of the helix in Msil_2636. This subunit was also present at the lowest abundance in comparison to the others, which may be a reflection of its position in the membrane and its difficulty to solubilise during the extraction process.

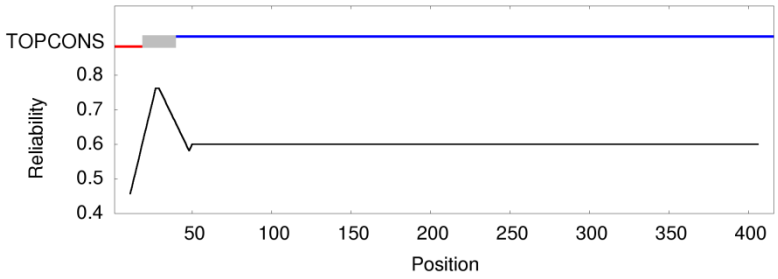

Figure 5.7. Trans-membrane helix prediction in Msil_2636

Red line indicates inside the membrane, *blue* line indicates outside the membrane and grey box is TM-helix with topology in→out.

Molar quantities estimated for each of the expressed genes in the cluster Msil_2632 – Msil_2639, range from 24 to 330 fmol. Proteins encoding *N*-methylglutamate synthase (160 – 330 fmol) were found to be much more abundant than proteins encoding the enzymes down-stream (24 – 60 fmol) and similar to those up-stream (240 fmol). Taking into account estimated stoichiometry based on label-free quantification and previous structural studies, it could be estimated that γ-glutamylmethylamide synthetase, *N*-methylglutamate synthase and *N*-methylglutamate dehydrogenase complexes were present between 30 – 80 fmol.

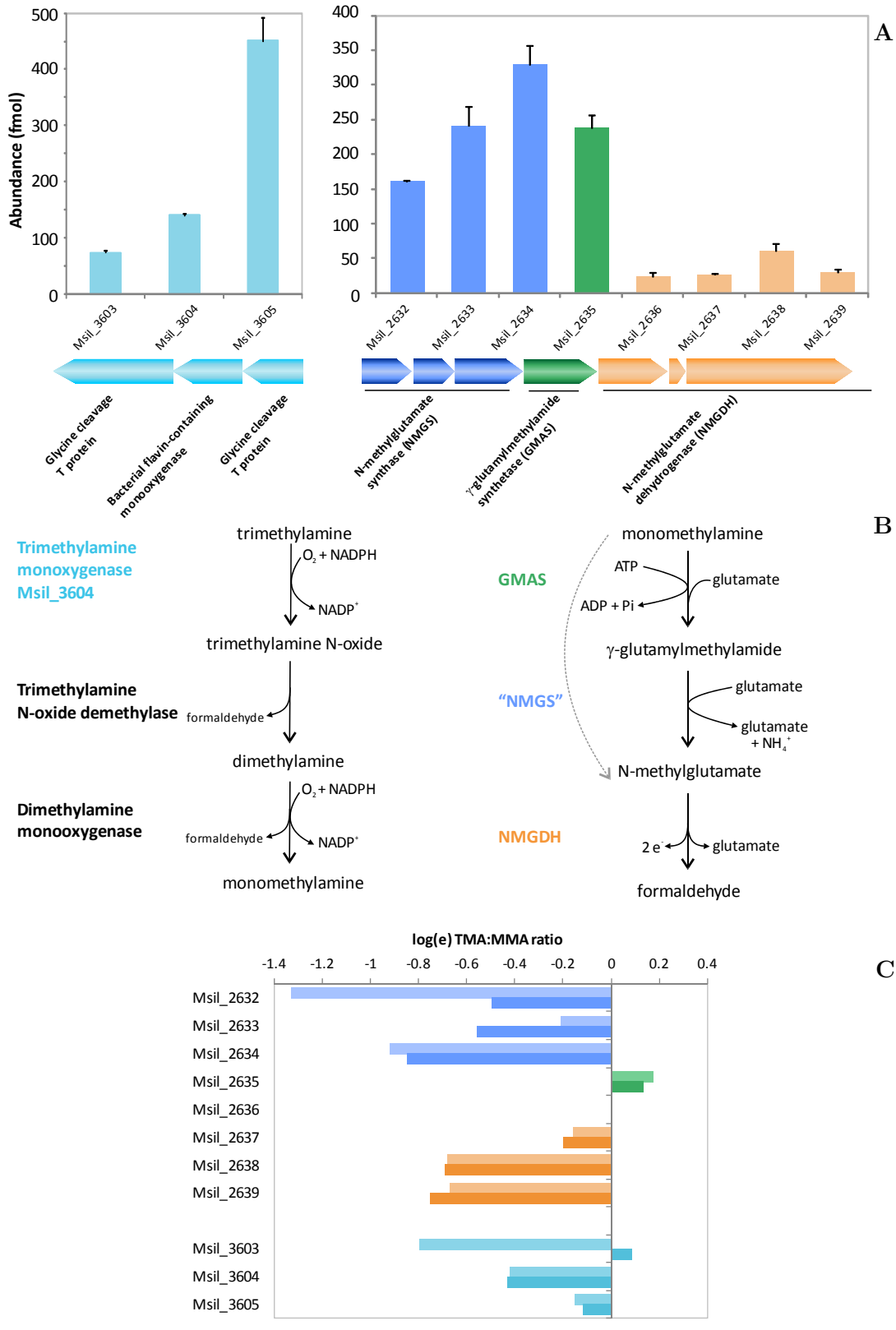

Figure 5.8. MMA- and TMA-proposed pathways in *M. silvestris*

(a) TMA-related enzyme levels are shown from analysis of TMA extracts and MMA-related enzyme levels are shown from analysis of MMA extracts by 1D RP-LC MS^E. (b) Proposed pathways. (c) Relative expression determined using all peptide intensities (light) and top three peptide intensities (dark). Negative ratios indicate up-regulation in monomethylamine growth and positive ratios indicate up-regulation in TMA growth

5.3.3 Highly up-regulated cluster during MMA and TMA growth encodes the enzyme trimethylamine monooxygenase

In addition to identifying the key enzymes involved in monomethylamine oxidation, a cluster of three genes, Msil_3603, Msil_3604 and Msil_3605, were found to be highly abundant in MMA- and TMA-grown extracts, constituting 8.5% and 7.5% of the total detected protein by weight, respectively. These proteins were not detected in MeOH-grown extracts. Figure 5.8(a) shows the abundances for each of these proteins measured from TMA-extracts using the 1D RP-LC MS^E approach. The relative levels for these three proteins were similar in the MMA-extracts. The descriptions given for Msil_3603 (82.4 kDa) and Msil_3605 (42.0 kDa) based on BLASTP similarity was Glycine cleavage T protein and Msil_3604 (52.0 kDa) was flavin-containing monooxygenase. It may be suggested that these enzymes were involved in an indirect pathway for trimethylamine oxidation.

Flavin-containing monooxygenase is trimethylamine monooxygenase (Msil_3604)

Flavin-containing monooxygenases (FMOs) are known to use a reduced form of an organic cofactor to activate molecular oxygen. According to the classification system proposed by van Berkel *et al.*, flavin-containing monooxygenases are a part of the subclass B of flavoprotein monooxygenases. Features of this subclass include the ability to bind the cofactor flavin adenine dinucleotide (FAD), the dependence on NADPH as coenzyme and being encoded by a single gene (van Berkel *et al.* 2006). FMOs have been widely characterised in various eukaryotic systems, including mammals, yeast, plants and fungi. In the human body, FMOs play an important role in the detoxification of drugs and other xenobiotics, including TMA, found in the diet of mammals. An FMO recently identified in the bacterium *Methylophaga* sp. Strain SK1 (Choi *et al.* 2003), had an estimated molecular weight of 105 kDa and consisted of two identical 54 kDa subunits. This enzyme was also shown to be able to utilise TMA as a substrate. The presence of FMOs have been observed in many bacterial genomes, however, the physiological roles of bacterial FMOs are unknown.

The label-free quantitative mass spectrometry approach employed here demonstrated that Msil_3604 flavin-containing monooxygenase was highly expressed

in *M. silvestris* grown with MMA and TMA, and was undetectable during methanol growth. This, along with the fact that eukaryotic FMOs have been shown to oxidise TMA to trimethylamine *N*-oxide, suggested that bacterial FMO may be involved in the indirect TMA metabolic pathway (trimethylamine monooxygenase, trimethylamine *N*-oxide demethylase and dimethylamine monooxygenase), which was first proposed by Large *et al.* in the 1970s. Thus, it may be suggested that Msil_3603 – Msil_3605 encoded for these enzymes.

The function of Msil_3604 was identified by cloning the gene from *M. silvestris*, overexpressing it in *Escherichia coli* and performing steady-state kinetic assays with a variety of substrates. Of the substrates tested, the enzyme had the highest affinity for TMA. A mutant of Msil_3604 was also constructed, which could no longer grow on TMA. Primers were designed, targeting the Msil_3604 gene, to confirm the presence of its transcript in MMA and MeOH-grown cells, shown in Figure 5.9.

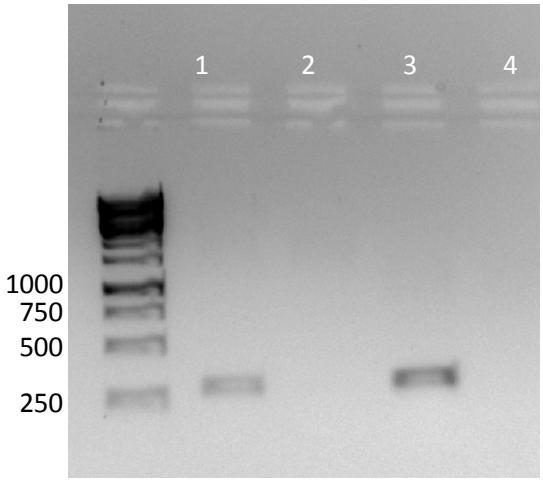

Figure 5.9. Agarose gel of RT-PCR products targeting Msil_3604
(1) Monomethylamine cDNA; (2) Monomethylamine mRNA; (3) Genomic DNA – positive control; (4) Water – negative control

A transcript relating to Msil_3604 was also identified in methanol-grown cells, which could not be supported by the proteomic data. It is possible that Msil_3604 may be transcribed under methanol-growth but has been translationally repressed, or that the protein is expressed at a low abundance and was undetectable using the proteomics approach employed in these experiments. Data obtained from proteomic, genetic and bioinformatics experiments suggests that bacterial FMO encodes trimethylamine

monooxygenase (Chen *et al.* 2011). Figure 5.8(b) shows the proposed pathway for trimethylamine metabolism in *M. silvestris*, which involves the release of two formaldehyde molecules from the enzymes trimethylamine *N*-oxide demethylase and dimethylamine monooxygenase. Monomethylamine is then proposed to be metabolised by the *N*-methylglutamate pathway, releasing an additional molecule of formaldehyde and ammonium.

Glycine cleavage T proteins may be trimethylamine *N*-oxide demethylase and dimethylamine monooxygenase (Msil_3603 and Msil_3605)

Sequence analysis of Msil_3603 and Msil_3605 revealed the presence of tetrahydrofolate-binding domains (involved in C_1 transfer), which suggested they may encode the enzymes trimethylamine *N*-oxide demethylase and dimethylamine monooxygenase, each of which may be proposed to release a molecule of formaldehyde. Further genetic or biochemical studies which would infer the exact identities of these genes have not been conducted on these enzymes, and few structural-based studies of the bacterial versions of these enzymes have been carried out. The protein expressed by Msil_3605 was estimated to be ~450 fmol, significantly higher than Msil_3603, which was present at ~60 fmol. Inducible activities of trimethylamine *N*-oxide demethylase and dimethylamine monooxygenase were detected in cell-free extracts of *M. silvestris* when it was grown on methylated amines, at 71.2 ± 16.3 and 21.2 ± 0.7 $nmol \cdot min^{-1} \cdot mg^{-1}$, respectively (Chen *et al.* 2011). The increased activity observed for trimethylamine *N*-oxide demethylase compared to that of dimethylamine monooxygenase may indicate that Msil_3605 is trimethylamine *N*-oxide demethylase. Purification and characterisation of these two genes would be required to verify their identities.

MMA- and TMA-enzymes are up-regulated during growth on MMA

Figure 5.8(c) shows the relative levels of MMA- and TMA-enzymes found to be expressed under different growth conditions. TMA-specific enzymes were up-regulated during MMA growth, as well as a number of MMA-specific enzymes. The fact that TMA enzymes were expressed at all under MMA-growth may be due to a reaction product, such as MMA, inducing upstream TMA enzymes. This is similar to

observations which have been made elsewhere; trimethylamine *N*-oxide aldolase (demethylase) activity was detected in *P. aminovorans* grown on monomethylamine and trimethylamine (Boulton and Large 1979). *N*-methylglutamate pathway enzymes may be up-regulated during MMA growth due to a larger external supply of MMA, which may induce expression of these enzymes.

In addition to the estimation of relative expression levels made using the Expression software, Figure 5.8(c) shows the relative expression levels calculated using absolute values estimated by the Hi3 approach. The values are generally in good agreement, with the exception of Msil_3603, which was estimated to be up-regulated in MMA by the Expression method and was not significantly differentially expressed when the Hi3 method was employed. Although the Expression approach uses more rigorous and statistical methods to obtain relative quantification, the use of top three peptide intensities to provide a rough estimate for relative quantification may be a promising, inexpensive, alternative for users without access to specialised commercial software.

5.3.4 Methylotrophic and central metabolic pathways

The proteins with the ten highest levels of measured concentration are largely common for each growth condition (Table 5.2). They may be categorised into known methylotrophic metabolic functions, or core processes such as protein folding (chaperonins) and translation (translation elongation factor Tu).

The major methanol dehydrogenase subunit, Msil_0471 PQQ dependent dehydrogenase methanol ethanol family which is the enzyme responsible for oxidising primary (and sometimes secondary) alcohols such as methanol, was the most abundant protein detected for all substrates. This protein was present at a relative concentration of 6.6% - 19.3% in the substrates studied here. This is comparable to the estimation that this enzyme usually constitutes between 5% and 15% of the soluble protein, determined from the specific activity of methanol dehydrogenase from different bacterial crude extracts and pure enzymes (Anthony 1982). The small subunit of methanol dehydrogenase, Msil_0474 methanol dehydrogenase beta subunit, was also identified in all substrates at a concentration approximately 2.5-fold lower, along with the associated cytochrome *c* (Msil_0473). The regulation of sMMO and the factors affecting their expression has been the target of several studies over the

last three decades. Soluble methane monooxygenase was not detected in TMA, MMA or methanol-grown cells, using the 1D RP-LC MS^E approach. Other studies have also shown that sMMO expression is not observed in methanotrophs growing on methanol, however, details are unclear (Theisen 2006). Further study of sMMO regulation may provide insights into its activation and/or repression under these growth conditions.

An interesting observation concerns the abundance of Msil_1388 expressed during growth on TMA. This protein was given the description of a hypothetical protein and was the fifth most abundant protein during TMA growth. This protein was present at a concentration approximately ten-fold lower during MeOH and MMA growth. Bioinformatic analysis revealed no predicted conserved domains, which may have been used to infer function or location of this protein. A homology-based search revealed that this protein was similar to other conserved hypothetical proteins which were predicted to be exported. A spectral counting-based proteomic study on the methylotroph *Methylobacillus flagellatus* showed that a hypothetical protein was the second most abundant protein observed under growth on methanol and methylamine. This protein did not share any sequence homology with the hypothetical protein expressed in *M. silvestris*.

Table 5.2. Abundant proteins in methanol, methylamine and trimethylamine extracts

Proteins identified in 2 or more replicates with average values. Abundant proteins are sorted in descending order according to % amount of total detected protein by weight.

	ORF	MW (kDa)	Description	Amount (%)	Amount (ng)	Amount (fmol)	Peptides	
1.	Msil_0471	68.5	YP 002360806 PQQ dependent dehydrogenase methanol ethanol family	6.6	27.3	398	44	Methanol
2.	Msil_0795	57.6	YP 002361127 chaperonin GroEL	5.6	23.2	402	41	
3.	Msil_1714	42.6	YP 002362023 aminotransferase class V	4.0	16.5	388	23	
4.	Msil_2390	18.6	YP 002362682 formaldehyde activating enzyme	3.3	13.1	705	12	
5.	Msil_2996	40.2	YP 002363269 acetyl CoA acetyltransferase	2.4	9.9	246	29	
6.	Msil_2342	55.7	YP 002362634 Aldehyde Dehydrogenase	2.3	9.7	175	17	
7.	Msil_0582	43.1	YP 002360916 translation elongation factor Tu	2.3	9.5	220	26	
8.	Msil_2955	67.8	YP 002363228 chaperone protein DnaK	1.9	8.0	118	40	
9.	Msil_1712	59.9	YP 002362021 Formate tetrahydrofolate ligase	1.4	5.9	99	25	
10.	Msil_1226	70.9	YP 002361557 acetate CoA ligase	1.3	5.4	76	23	
1.	Msil_0471	68.5	YP 002360806 PQQ dependent dehydrogenase methanol ethanol family	17.6	86.1	1257	65	Monomethylamine
2.	Msil_3605	41.9	YP 002363852 glycine cleavage T protein aminomethyl transferase	4.7	23.2	553	28	
3.	Msil_0795	57.6	YP 002361127 chaperonin GroEL	3.4	16.4	285	43	
4.	Msil_2634	46.9	YP 002362919 ferredoxin dependent glutamate synthase	3.2	15.5	330	36	
5.	Msil_3604	51.7	YP 002363851 Flavin containing monooxygenase	2.5	12.1	235	24	
6.	Msil_2635	48.4	YP 002362920 glutamine synthetase type III	2.4	11.5	238	26	
7.	Msil_2342	55.7	YP 002362634 Aldehyde Dehydrogenase	2.1	10.4	188	22	
8.	Msil_2996	40.2	YP 002363269 acetyl CoA acetyltransferase	1.7	8.3	207	27	
9.	Msil_2955	67.8	YP 002363228 chaperone protein DnaK	1.6	7.9	116	44	
10.	Msil_2390	18.6	YP 002362682 formaldehyde activating enzyme	1.5	7.2	388	13	
1.	Msil_0471	68.8	YP 002360806 PQQ dependent dehydrogenase methanol ethanol family	19.3	82.9	1205	51	Trimethylamine
2.	Msil_2390	18.6	YP 002362682 formaldehyde activating enzyme	5.6	23.8	1284	13	
3.	Msil_3605	42.0	YP 002363852 glycine cleavage T protein aminomethyl transferase	4.4	18.9	450	19	
4.	Msil_0795	57.6	YP 002361127 chaperonin GroEL	3.2	13.5	235	33	
5.	Msil_1388	22.4	YP 002361715 hypothetical protein	3.0	12.7	570	15	
6.	Msil_2635	48.8	YP 002362920 glutamine synthetase type III	2.8	12.1	249	25	
7.	Msil_3881	53.6	YP 002364125 Aldehyde Dehydrogenase	2.7	11.4	212	26	
8.	Msil_2996	40.4	YP 002363269 acetyl CoA acetyltransferase	2.4	10.4	257	23	
9.	Msil_2342	56.0	YP 002362634 Aldehyde Dehydrogenase	2.3	10.0	178	22	
10.	Msil_1714	42.7	YP 002362023 aminotransferase class V	2.0	8.5	199	16	

Formaldehyde may be either assimilated via the serine pathway or oxidised to carbon dioxide. Formaldehyde can be directly oxidised to formate via formaldehyde dehydrogenase or via the H_4MPT-linked pathway, however, the genome lacks coding for formaldehyde dehydrogenase. Formaldehyde activating enzyme is the first enzyme involved in the H_4MPT-linked formaldehyde oxidation path, and has been shown to be highly expressed in methylotrophs (Vorholt *et al.* 1999). This was also observed in *M. silvestris* present as one of the top ten most abundant proteins in all growth conditions. Figure 5.10 shows the relative expression levels of proteins involved in methylotrophic and central metabolic pathways. The majority of enzymes involved in H_4MPT-linked formaldehyde oxidation were identified in the studied substrates, and do not significantly differ in expression during MeOH, TMA and MMA growth.

The genome encodes several proteins which relate to the enzyme formate dehydrogenase. This uses formate as a substrate to generate NADH and CO_2. The genes Msil_3654 to Msil_3658, and Msil_3325 to Msil_3328 encode up to four subunits (α,β,γ,δ or α,β,γ) of formate dehydrogenase and an accessory protein, and the gene Msil_2953 has been given the description formate dehydrogenase. Methylotrophs such as *Mycobacterim vaccae* (Karzanov et al. 1991) and *Pseudomonas* sp. 101 (Karzanov *et al.* 1989) are known to possess and utilise two forms of formate dehydrogenases. These differ in cellular location, molecular mass and specificity towards electron acceptors. Four types of formate dehydrogenase have been described in bacteria (Chistoserdova *et al.* 2007). These include a soluble NAD^+-linked enzyme and a membrane-bound one (Anthony 1982). The subunits encoded by Msil_3325 to Msil_3328, which encode an enzyme which includes three subunits, have been predicted to contain more transmembrane helices than the equivalent subunits from Msil_3654 to Msil_3658. It is likely that the cluster encoding three subunits relates to the membrane form of the enzyme, similar to that characterised in *E. coli* (Graham and Boxer 1981). Two of the four subunits relating to the genes that most likely encode a soluble form of the enzyme were identified in these studies, along with Msil_2953. A proteomic study of the membrane fraction would be required to confirm whether both forms of formate dehydrogenase were used in this bacterium.

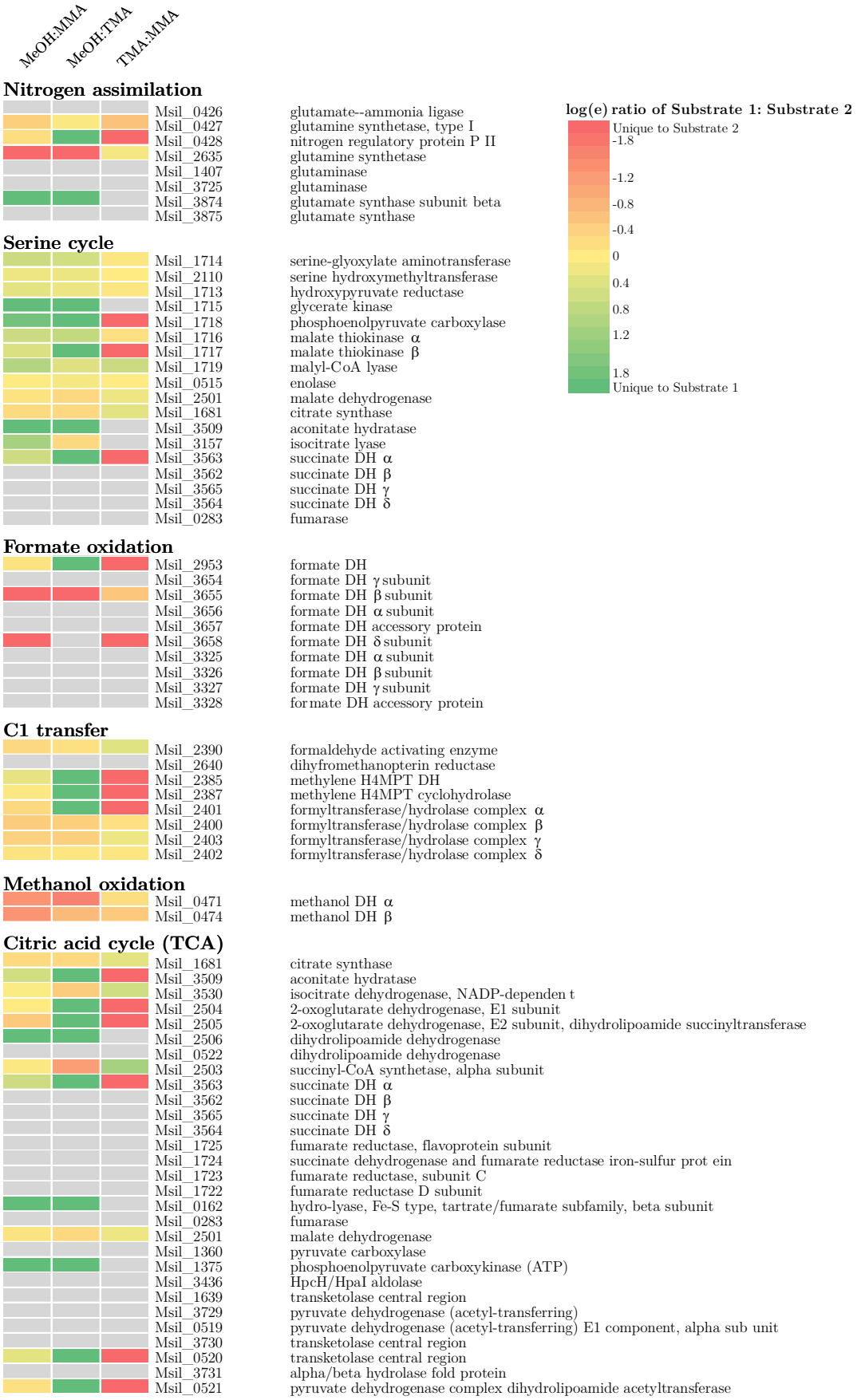

Figure 5.10. Relative levels of proteins in TMA, MMA and MeOH extracts

Proteins that were not identified in either substrate are denoted by *grey* boxes. *Proposed GMAS.

From the genome it may be proposed that nitrogen, in the form of ammonium, is assimilated through glutamine synthetase (GS) and glutamate synthase (GOGAT) (GS-GOGAT pathway), since genes encoding glutamate dehydrogenase or alanine dehydrogenase were not present. The genome encodes a glutamine synthetase, type III (Msil_2635) a glutamine synthetase, type II (Msil_0426) and a glutamine synthetase, type I (Msil_0427). Sequence analysis has shown that Msil_2635 lacks a conserved ammonia-binding domain site and Msil_0426 was not detected in these analyses, suggesting that Msil_0427 was the correct glutamine synthetase. Msil_2635 has also proposed to be γ-glutamylmethylamide synthetase, which is part of the MMA oxidation pathway. Two ORFs are encoded for glutamate synthase, Msil_3874 and Msil_3875. The data presented here shows that glutamine synthetase (Msil_0427) and a nitrogen regulatory protein (Msil_0428) are slightly up-regulated in MMA compared to MeOH and TMA, however, glutamate synthase was only detected in methanol-grown extracts. Since ammonium was supplied to methanol-grown cells, it would be expected that similar levels of GS and GOGAT would be present in MMA and TMA cells through the generation of ammonium in the *N*-methylglutamate pathway.

The majority of proteins known to be involved in the serine pathway have been identified. Two enzymes that were not identified were fumarase and succinate dehydrogenase, the latter of which is thought to be localised in the membrane. The up- or down-regulation of serine pathway enzymes may provide an indication of the cells requirements for assimilation under the different growth conditions. Some enzymes are shared between the serine cycle and the citric acid cycle, and as such the characterisation of serine pathway-exclusive enzymes is necessary to provide a more accurate perspective on the regulation of the pathway. The enzymes that are most likely to be exclusive to the serine pathway are serine-glyoxylate aminotransferase (Msil_1714), hydroxypyruvate reductase (Msil_1713) and serine hydroxymethyltransferase (Msil_2110). These enzymes were not significantly differentially expressed in the studied growth conditions. This would be expected, since downstream reactions would be expected to be the same after the primary oxidation of methanol, trimethylamine and monomethylamine into formaldehyde.

5.4 Conclusions

A relatively simple mass spectrometry-based approach has been used with no pre-fractionation of samples, to provide qualitative and quantitative information on the soluble protein content of *M. silvestris* grown at the expense of TMA, MMA or MeOH. A good overlap of identifications was achieved between these studied growth conditions. Cytoplasmic and periplasmic proteins were enriched and core proteins were involved in essential functions. Quantitative data implied that the distribution of protein abundances expressed during TMA growth was notably different in comparison to the other studied growth substrates.

All enzymes proposed to be involved in the *N*-methylglutamate pathway for MMA metabolism were confidently identified in MMA- and TMA-grown extracts, but not in methanol. Label-free quantification was used to provide an estimate for the concentration of these enzymes as a total of the detected protein content, and was used to suggest subunit stoichiometry.

A cluster of expressed genes that were significantly up-regulated under TMA and MMA growth were proposed to be involved in an indirect pathway for TMA metabolism. It was verified that one of these genes expressed the previously unidentified trimethylamine monooxygenase. Relative quantification showed that MMA- and TMA-specific enzymes were up-regulated under MMA growth compared to TMA.

The biotechnologically important enzyme, soluble methane monooxygenase, was not detected under TMA, MMA or MeOH growth conditions. Good coverage was achieved for central pathways such as formaldehyde oxidation and the serine pathway, which were shown to be expressed at similar levels across all substrates. A comprehensive study would provide a more detailed assessment of the regulation of these pathways under the various growth conditions.

These studies revealed some potentially interesting enzymes, in a unique bacterium, which have the potential to be studied further. Purification and biochemical studies would be necessary to confirm the identities of dimethylamine monooxygenase and trimethylamine *N*-oxide demethylase. The structural composition and substrate specificity of some *N*-methylglutamate pathway enzymes would also warrant further investigation.

5.5 References

Anthony, C. (1982). The Biochemistry of Methylotrophs. London, Academic Press Inc.

Bamforth, C. W. and Large, P. J. (1977a). The molecular size of *N*-methylglutamate dehydrogenase of *Pseudomonas aminovorans*. *Biochem J.* **167**, 509-512.

Bamforth, C. W. and Large, P. J. (1977b). Solubilization, partial purification and properties of *N*-methylglutamate dehydrogenase from *Pseudomonas aminovorans*. *Biochem J.* **161**, 357-370.

Bhadbhade, B. J., Sarnaik, S. S. and Kanekar, P. P. (2002). Biomineralization of an organophosphorus pesticide, Monocrotophos, by soil bacteria. *J. Appl. Microbiol.* **93**, 224-234.

Boulton, C. A., Crabbe, M. J. and Large, P. J. (1974). Microbial oxidation of amines. Partial purification of a trimethylamine mono-oxygenase from *Pseudomonas aminovorans* and its role in growth on trimethylamine. *Biochem. J.* **140**, 253-263.

Boulton, C. A., Haywood, G. W. and Large, P. J. (1980). *N*-Methylglutamate Dehydrogenase, a Flavohaemoprotein Purified from a New Pink Trimethylamine-utilizing Bacterium. *J. Gen. Microbiol.* **117**, 293-304.

Boulton, C. A. and Large, P. J. (1979). Inactivation of trimethylamine *N*-oxide aldolase (demethylase) during preparation of bacterial extracts. *FEMS Microbiol. Lett.* **5**, 159-162.

Chen, Y., Patel, N. A., Crombie, A., Scrivens, J. H. and Murrell, J. C. (2011). Bacterial flavin-containing monooxygenase is trimethylamine monooxygenase. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17791-17796.

Chen, Y., Scanlan, J., Song, L. J., Crombie, A., Rahman, M. T., Schafer, H. and Murrell, J. C. (2010). gamma-Glutamylmethylamide Is an Essential Intermediate in the Metabolism of Methylamine by *Methylocella silvestris*. *Appl. Environ. Microbiol.* **76**, 4530-4537.

Chistoserdova, L., Crowther, G. J., Vorholt, J. A., Skovran, E., Portais, J.-C. and Lidstrom, M. E. (2007). Identification of a Fourth Formate Dehydrogenase in *Methylobacterium extorquens* AM1 and Confirmation of the Essential Role of Formate Oxidation in Methylotrophy. *J. Bacteriol.* **189**, 9076-9081.

Choi, H. S., Kim, J. K., Cho, E. H., Kim, Y. C., Kim, J. I. and Kim, S. W. (2003). A novel flavin-containing monooxygenase from *Methylophaga* sp. strain

SK1 and its indigo synthesis in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **306**, 930-936.

Colby, J. and Zatman, L. J. (1973). Trimethylamine metabolism in obligate and facultative methylotrophs. *Biochem. J.* **132**, 101–112.

Eady, R. R. and Large, P. J. (1968). Purification and properties of an amine dehydrogenase from *Pseudomonas* AM1 and its role in growth on methylamine. *Biochem. J.* **106**, 245-255.

Graham, A. and Boxer, D. H. (1981). The organization of formate dehydrogenase in the cytoplasmic membrane of *Escherichia coli*. *Biochem J.* **195**, 627-637.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols.* **4**, 44-57.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1-13.

Kamiya, A. and Ose, Y. (1984). Study of odorous compounds produced by putrefaction of foods. *J. Chromatogr.* **292**, 383-391.

Karzanov, V. V., Bogatsky, Y. A., Tishkov, V. I. and Egorov, A. M. (1989). Evidence for the presence of a new NAD+-dependent formate dehydrogenase in *Pseudomonas* sp. 101 cells grown on a molybdenum-containing medium. *FEMS Microbiol. Lett.* **60**, 197-200.

Karzanov, V. V., Correa, C. M., Bogatsky, Y. G. and Netrusov, A. I. (1991). Alternative NAD+-dependent formate dehydrogenases in the facultative methylotroph *Mycobacterium vaccae* 10. *FEMS Microbiol. Lett.* **81**, 95-99.

Kimura, T., I, S., Hanai, K. and Tonomura, Y. (1992). Purification and Characterization of γ-Glutamylmethylamide Synthetase from *Methylophaga* sp. AA-30. *Biosci. Biotechnol. Biochem.* **56**, 708-711.

Kung, H.-F. and Wagner, C. (1969). Gamma-glutamylmethylamide. A new intermediate in the metabolism of methylamine. *J. Biol. Chem.* **244**, 4136-4140.

Large, P. J., Boulton, C. A. and Crabbe, M. J. (1972). The reduced nicotinamide-adenine dinucleotide phosphate- and oxygen-dependent *N*-oxygenation of trimethylamine by *Pseudomonas aminovorans*. *Biochem J.* **128**, 137P–138P.

Latypova, E., Yang, S., Wang, Y. S., Wang, T. S., Chavkin, T. A., Hackett, M., Schafer, H. and Kalyuzhnaya, M. G. (2010). Genetics of the glutamate-mediated methylamine utilization pathway in the facultative

methylotrophic beta-proteobacterium *Methyloversatilis universalis* FAM5. *Mol. Microbiol.* **75**, 426-439.

Lin, M. C. and Wagner, C. (1975). Purification and characterization of *N*-methylalanine dehydrogenase. *J. Biol. Chem.* **250**, 3746-3751.

Pollock, R. J. and Hersh, L. B. (1971). *N*-methylglutamate synthetase purification and properties of the enzyme. *J. Biol. Chem.* **246**, 4737-4743.

Shaw, W. V., Tsai, L. and Stadtman, E. R. (1966). The Enzymatic Synthesis of *N*-Methylglutamic Acid. *J. Biol. Chem.* **241**, 935-945.

Slade, S. E. (2012). 1D RP-LC MSE analysis on depleted and undepleted plasma samples. *Personal communication.*

Theisen, A. R. (2006). Regulation of Methane Oxidation in the Facultative Methanotroph *Methylocella silvestris* BL2. *PhD.* Department of Biological Sciences, University of Warwick

van Berkel, W. J. H., Kamerbeek, N. M. and Fraaije, M. W. (2006). Flavoprotein monooxygenases, a diverse class of oxidative biocatalysts. *J. Biotechnol.* **124**, 670-689.

Vorholt, J. A., Chistoserdova, L., Stolyar, S. M., Thauer, R. K. and Lidstrom, M. E. (1999). Distribution of Tetrahydromethanopterin-Dependent Enzymes in Methylotrophic Bacteria and Phylogeny of Methenyl Tetrahydromethanopterin Cyclohydrolases. *J. Bacteriol.* **181**, 5750-5757.

Welsh, D. T. (2000). Ecological significance of compatible solute accumulation by micro-organisms: from single cells to global climate. *FEMS Microbiol. Rev.* **24**, 263-290.

Yamamoto, S., Wakayama, M. and Tachiki, T. (2007). Characterization of Theanine-Forming Enzyme from *Methylovorus mays* No. 9 in Respect to Utilization of Theanine Production. *Biosci. Biotechnol. Biochem.* **71**, 545-552.

Zhang, X., Fuller, J. H. and McIntire, W. S. (1993). Cloning, sequencing, expression, and regulation of the structural gene for the copper/topa quinone-containing methylamine oxidase from *Arthrobacter* strain P1, a gram-positive facultative methylotroph. *J. Bacteriol.* **175**, 5617-5627.

Chapter 6

Proteomic Studies on Facultative Methanotrophy

6.1 Introduction

6.1.1 Methane and acetate utilisation

The study of methane metabolism regulation is of great interest due to the atmospheric impact of this global warming gas and also due to the biotechnological potential of monooxygenases. Investigating the effect of multi-carbon substrates on facultative methanotrophic activity can improve our understanding of methane cycling in the environment. The addition of acetate to a methane-grown culture of *Methylocella* has been shown to initially inhibit methane oxidation, but to have a long term stimulatory effect (Dedysh *et al.* 2005). The regulation of methane oxidation in *M. silvestris* BL2 has been investigated during growth on methane and acetate (Theisen *et al.* 2005). Acetate is a relevant model multi-carbon substrate since it is a major product of fermentation in flooded soils. Conclusive evidence for the absence of pMMO in *M. silvestris* BL2 was provided, in accordance with its absence in the genome and that fact that *M. silvestris* does not have an extensive intracytoplasmic membrane system to which pMMO would be bound in other methanotrophs. *In vitro* culture studies (Theisen *et al.* 2005) and DNA stable-isotope probing carried out using environmental samples (Rahman *et al.* 2011) demonstrated that transcription of the *mmo* operon could be repressed by acetate. The exact regulatory mechanism operating the switch from methane to acetate utilisation remains unclear.

Proteomic studies have been carried out on the soluble content of *M. silvestris* when grown with methane or acetate (Patel *et al.* 2009). This study identified 423 non-redundant proteins using a 1D RP-LC MS^E approach, and achieved relative quantification information by means of a label-free method (Hi3) and a labeling method (iTRAQ). The main focus of this study was the comparison of mass spectrometry-based approaches. The *M. silvestris* database at the time of this study was incomplete, which posed a challenge for the bioinformatic analysis.

The route by which C_2 substrates are assimilated is not only of interest for substrates such as acetate, but also for numerous substrates that share the initial conversion of acetyl-CoA as an entry point for central carbon metabolism. Many organisms use the glyoxylate cycle which was identified more than 50 years ago. This operates in conjunction with the TCA cycle. The pathway requires two key enzymes; isocitrate lyase and malate synthase. Some organisms lack isocitrate lyase and an operating glyoxylate cycle, but may retain the ability to grow on acetate. This

alternative pathway has long remained a mystery, until the discovery of the ethylmalonyl-CoA (EMC) pathway in *Rhodobacter sphaeoirdes* (Alber *et al.* 2006). This pathway has also been shown to be connected to the serine pathway in the facultative methylotroph *Methylobacterium extorquens* AM1 (Peyraud *et al.* 2009).

6.1.2 Bacterial oxidation of propane

Organisms capable of utilising C_2 – C_4 gaseous hydrocarbons have been investigated for their potential use in petroleum exploration and as sources for single cell protein to be used as animal feeds. Microbes with an ability to grow on short *n*-alkanes (other than methane) have recently been investigated for their potential use in the biotechnological production of epoxides, synthesis of chiral epoxyalkanes and as catalysts in bioremediation systems. It has been proposed that these organisms may have mitigated harmful effects during recent catastrophic events. The Deepwater Horizon event was the largest offshore oil spill in history to date. The spillage occurred at great depth, with the potential release of methane, ethane, propane and butane to the atmosphere if they were to reach shallow water. The majority of microbial respiration in deep water plumes contributed to the utilisation of propane and ethane. It was proposed that propane and ethane promoted rapid hydrocarbon respiration and primed bacterial populations for the degradation of other hydrocarbons (Valentine *et al.* 2010). Contrary to belief, the delivery of hydrocarbons into the environment is not entirely associated with industrial activities. A wide variety of saturated and unsaturated hydrocarbon gases are released from processes such as organic matter degradation by methanogenesis or cellulose fermentation (Shennan 2006), as well as by release from geological deposits (Etiope and Ciccioli 2009).

Propane oxidation is predominantly confined to the complex of Gram-positive bacteria *Corynebacterium-Mycobacteria-Nocardia-Rhodococcus* (the CMNR group). Even though Gram-negative propane-utilisers have been reported, it is the CMNR group that are frequently found in enrichment cultures from soil and freshwaters samples. The oxidation of propane is rather more complex when compared to that of methane. Oxidation of propane can occur by action of a monooxygenase on the terminal carbon atom, sub-terminal carbon atom, or a combination of both, yielding

propan-1-ol, propan-2-ol, and 1,2-propandiol, respectively. The roles of terminal and sub-terminal oxidation for propane metabolism have not been fully resolved.

Detailed studies on a pathway for propane utilisation have been made using the bacterium *Rhodococcus rhodochorus* PNKb1 (Woods and Murrell 1989). This bacterium was found to be capable of growing on intermediates of both the terminal and sub-terminal pathways, suggesting that the initial monooxygenase is indiscriminate towards initial oxygen insertion. SDS-PAGE of cell-free extracts of *R. rhodochorus* PNKb1 grown on intermediates of propane metabolism showed the presence of three proteins with molecular weights of 69, 59 and 57 kDa. These were proposed to comprise propane monooxygenase. Further studies indicated the requirement of a primary and secondary alcohol dehydrogenase for propan-1-ol and propan-2-ol, respectively. A soluble NAD^+-dependent secondary alcohol dehydrogenase induced during propane growth was purified and shown to consist of two subunits with a total molecular weight of 86 kDa. A primary alcohol dehydrogenase was not purified, but inferred from the isolation of mutants unable to grow on propan-1-ol. Experimental evidence suggested that this bacterium utilised the terminal pathway, as elevated levels of propanal dehydrogenase (propanal → propanoate) were present. The enzyme propionyl-CoA synthetase was also elevated. This carries out a condensation reaction which produces propionyl-CoA.

Figure 6.1 summarises the previously identified propane assimilation pathways. These pathways result in the conversion of propane to acetyl-CoA, formaldehyde, pyruvate or succinate, which then serve as entry points into the central carbon metabolism.

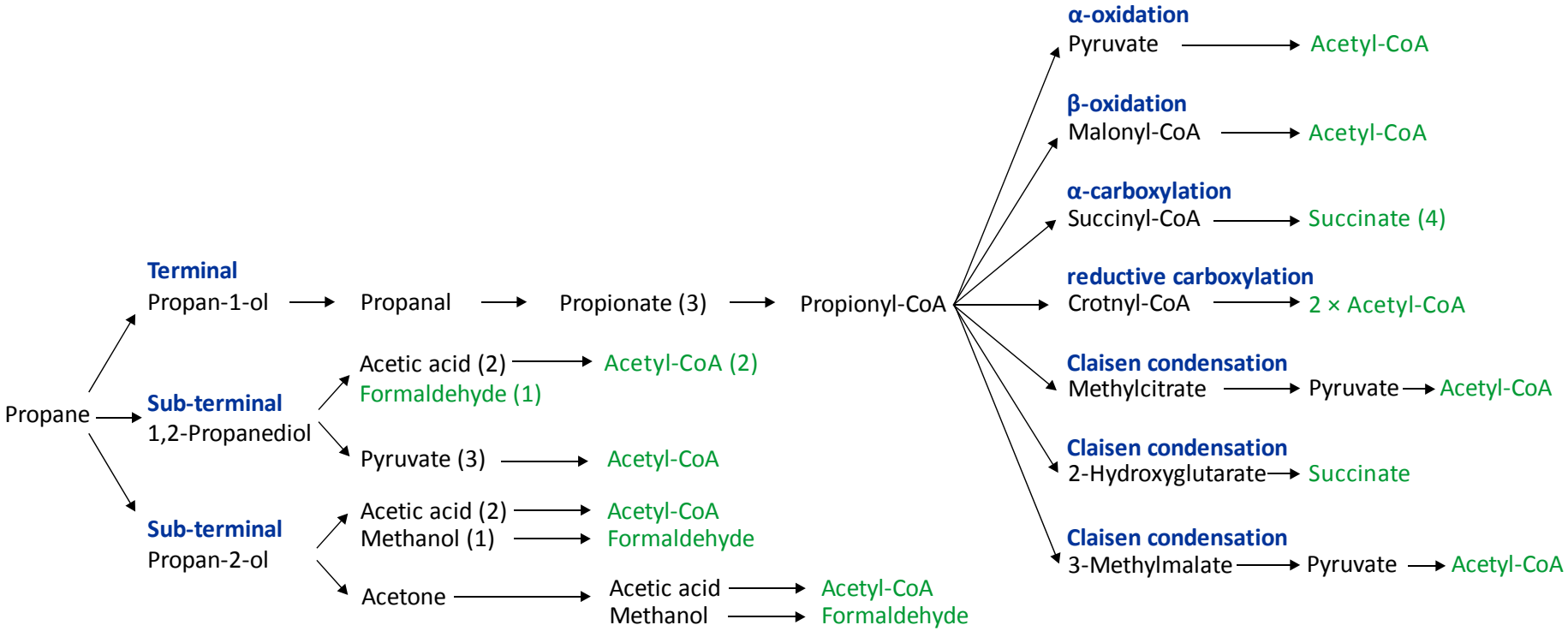

Figure 6.1. Possible propane assimilation pathways

Green text indicates entry point to central metabolism, *blue text* indicates pathway name/type of reaction, *black-solid* arrows indicate one or more enzyme steps, *brackets* denote number of carbon atoms.

6.2 Aims

This study of the proteome of the methanotroph *M. silvestris* BL2 provided an opportunity to understand metabolic pathways. The utility of multi-carbon compounds as a growth substrate has been well documented in a subset of methylotrophs, but only recently in the methanotrophic genus *Methylocella.* *M. silvestris* BL2 was the first bacterium which had been shown to grow at the expense of the multi-carbon compounds propane and acetate, as well as methane, and remains the only example of an organism able to use either of these hydrocarbons as a sole source of carbon and energy.

The aims of this Chapter were:

- To employ mass spectrometry-based proteomic approaches to identify and quantify proteins from soluble protein extracts obtained from *M. silvestris* grown at the expense of propane, succinate, acetate and methane
- To interrogate these results to characterise propane and acetate assimilation pathways, and to compare them to control substrates succinate and methane
- To conduct all analyses using technical replicates and for selected substrates using technical and biological replicates, in order to assess the variation associated with the results obtained

6.3 Methods

Bacterial protein preparation, LC-MS^E settings and data processing were detailed in Chapter 2. Where results are discussed, the analytical techniques used are referred to in the text or figure/table legends. Figure 6.2 provides an overview of the methods that were used for obtaining biological replicates for the analysis of propane- and succinate-growth conditions.

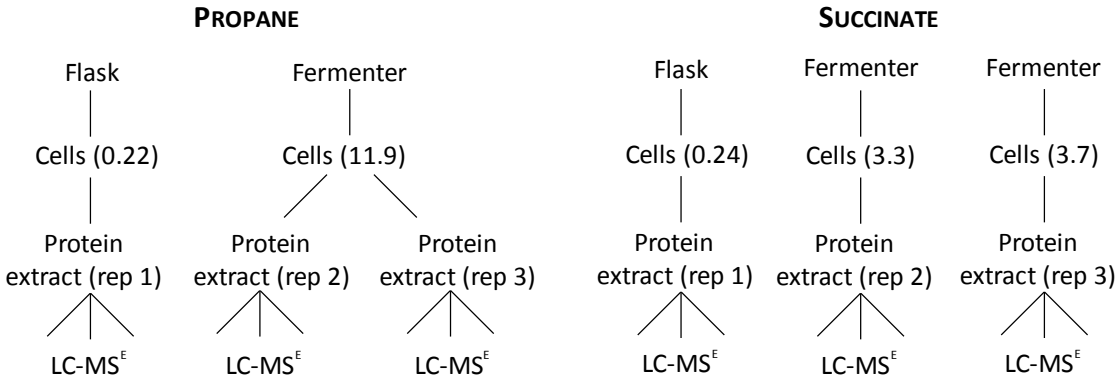

Figure 6.2. Overview of bacterial extract preparations

Number in brackets denotes OD 540 at time of harvesting.

6.4 Results and Discussion

6.4.1 Biological replication

The ability to make confident, biologically relevant conclusions requires that similar observations can be made between the analysis of various sample preparations (biological replication) as well as for multiple analyses of the same sample (technical replication). In terms of proteomic studies, these analyses should ideally provide similar information in terms of protein identifications and concentrations. The ability to carry out both biological and technical replicates is a demanding process, requiring increased experimental time for sample preparation, data acquisition and data analysis. Analysis of biological replicates is not yet required when reporting proteomics results for publication, but is clearly desirable.

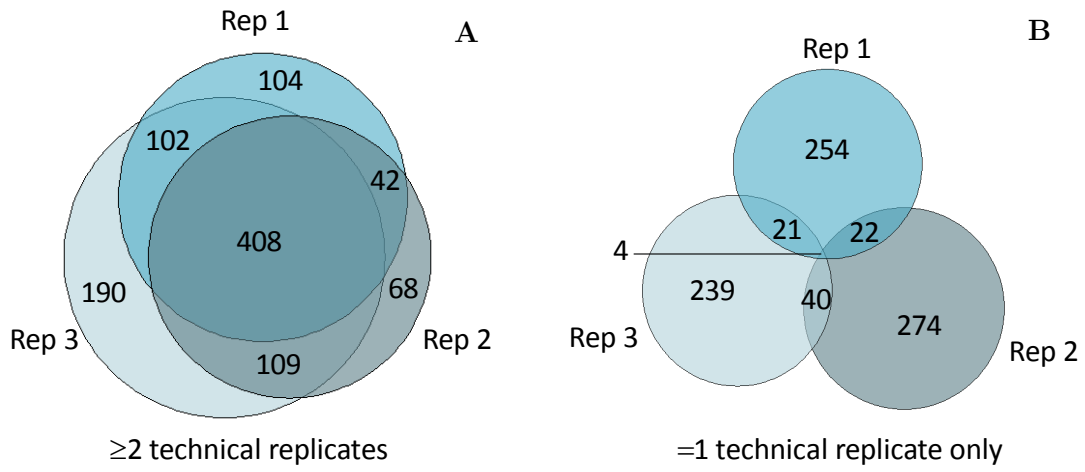

Figure 6.3. Protein identifications from three biological replicates of propane-grown extracts

(a) Protein identifications observed in 2 or more technical replicates. (b) Discarded protein identifications observed in a single technical replicate only. Datasets were obtained by the following means:

Rep 1 – 2D RP-RP-LC (6 fractions) Synapt MSE
Rep 2 – 2D RP-RP-LC (6 fractions) Synapt MSE
Rep 3 – 2D RP-RP-LC (3 fractions) Synapt G2 HDMSE

Soluble proteins extracts were obtained from different cell cultures or by obtaining multiple extracts from the same culture (Figure 6.2). The results discussed in Chapters 3 and 4 have addressed the variation that may occur between different LC-MS approaches and the technical variation that exists within each approach. The methods employed here for the production of protein extracts allowed the assessment of variation that may occur between small and large scale growth (i.e. bio rep 1 vs bio rep 2 and bio rep 1 vs bio rep 3 for succinate and propane), and variation of protein extraction within the same population of cells (bio rep 2 vs bio rep 3 for propane).

Figure 6.3(a) shows the overlap of protein identifications made between three biological replicates for the propane-growth condition when analysed by various LC-MS methods. The LC-MS approaches that provided the most protein identifications were used for these comparisons in order to reflect the optimal overlap between biological replicates. 1023 non-redundant proteins were observed when three datasets were combined, with a 65% overlap (≥2 biological replicates) and 408 proteins common to all biological replicates. These data sets represent proteins that were observed in two or more technical replicates. This is achieved by discarding these proteins that were only identified in a single replicate. Single replicating identifications have been assumed to be random in nature and are considered to be false positives. Figure 6.3(b) shows the overlap of discarded identifications from each

biological dataset. A poor correlation was observed, indicating that single replicating identifications, when three technical replicates were employed, were not replicated across the nine technical replicates from the analysis of three similar samples. This provides confidence that the majority of those identifications that were not included in the individual datasets were indeed false positives. There were 87 single replicating proteins that were identified in two or three technical replicates when considering nine technical replicates. Whether these identifications should be considered as true positives forms an interesting topic for discussion.

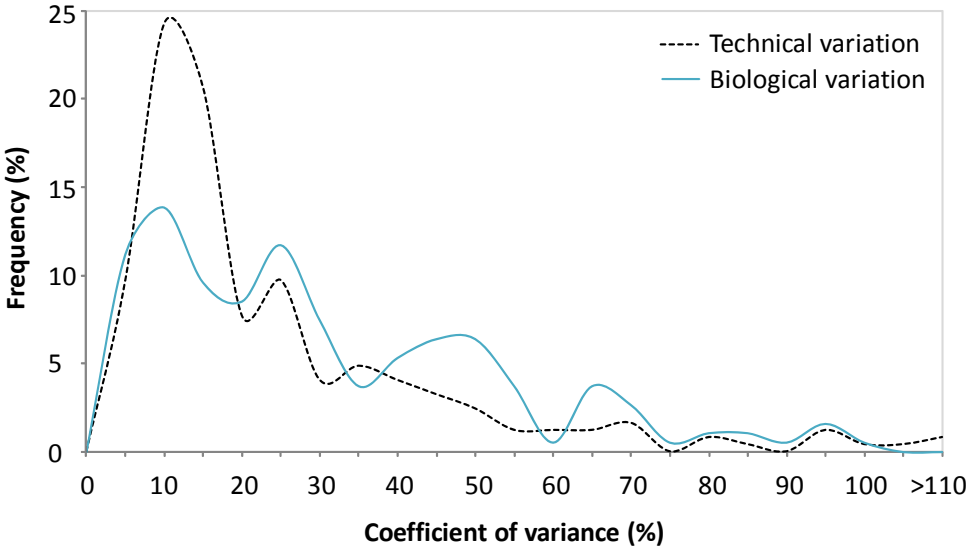

Figure 6.4. Technical and biological variation for protein quantification in propane-extracts

Since biological replicates 2 and 3, for the propane condition were obtained from the same culture of cells, one would expect a significant similarity between these samples. In terms of protein identifications, this does appear to be the case, with 82% of proteins from biological replicate 3 shared by replicate 2, compared to 77% of proteins from biological replicate 1 shared by replicate 2. The comparison of protein quantification between different biological replicates has been limited to the LC-MS technique used. Figure 6.4 shows the CVs of protein quantification estimated for proteins common to biological replicates 1 and 2, which were grown using small- or large-scale cultures, but analysed under similar LC-MS conditions. The technical variation associated with the analysis of one biological condition is also shown for comparison. This direct comparison illustrated that the technical variation is significantly smaller than the biological variation, with approximately 70% of proteins

displaying a CV of 50% or lower when considering only the biological variation, compared to 90% of proteins displaying a CV of 50% or lower for the technical variation.

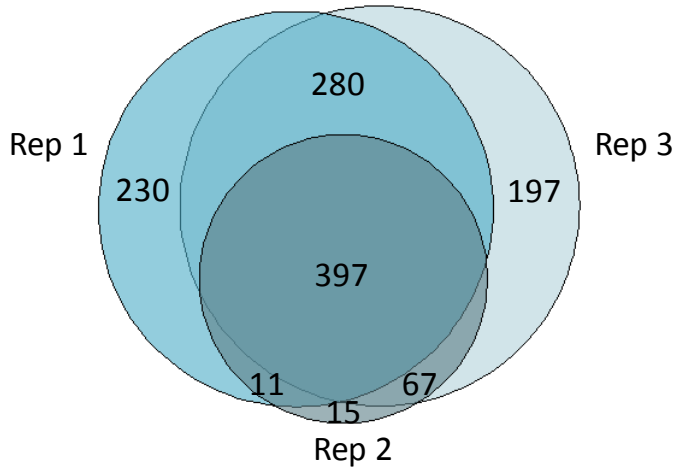

Figure 6.5. Protein identifications from three biological replicates of succinate-grown extracts

Datasets were obtained by the following means:
Rep 1 – 2D RP-RP-LC (6 fractions) Synapt MSE
Rep 2 – 1D RP-LC Synapt G2 HDMSE
Rep 3 – 2D RP-RP-LC (3 fractions) Synapt G2 HDMSE

Figure 6.5 shows the overlap of protein identifications made between three biological replicates for the succinate-growth condition when analysed by various LC-MS methods. 1197 non-redundant proteins were observed when three data sets were combined, with a 63% overlap (≥2 biological replicates) and 397 proteins common to all biological replicates. Fewer identifications were made from the second biological replicate, but almost all of these were included in the other two samples. Biological replicates 2 and 3 were obtained under similar conditions, i.e. grown in fermenters and harvested at similar densities. This is reflected by the observation that a greater number of proteins were shared between replicates 2 and 3 (464) when compared to those shared between replicates 2 and 1 (408).

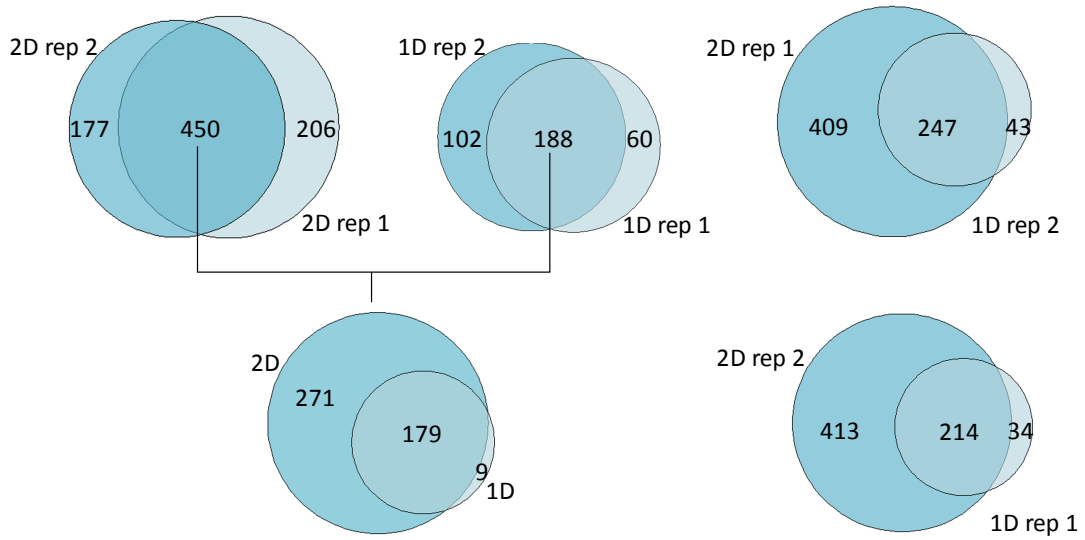

Figure 6.6. Protein identifications in two biological replicates of propane-grown extracts analysed by 1D and 2D LC-MS^{E} methods

Proteins that seem to be unique to a particular biological replicate can partially be explained by the dynamic range of the mass spectrometry approach used. This can be more easily explained by a comparison of the results obtained from the two cultures of *M. silvestris* grown with propane and analysed by 1D and 2D approaches. These are shown in Figure 6.6. Protein identifications were compared between different biological samples for both LC approaches, with 54% of non-redundant proteins being common to both biological replicates by either method. The overlap in biological replication is, as expected, lower than that observed for technical replication and was similar to observations made elsewhere (Nagaraj and Mann 2011). The proteins that were seemingly unique to one biological replicate in the comparison of 1D data sets were compared for the 2D data sets, for example biological replicate 1 analysed by 1D vs. biological replicate 2 analysed by 2D. From these results it can be shown that a proportion of the proteins that were found to be uniquely identified from the 1D approach were also identified in the second sample using the 2D approach. This may be attributed to the fact that proteins present in the different samples were not detected due the limit of quantification of the 1D method.

6.4.2 Proteome coverage and assessment

Confident identification results were combined from all biological replicates and LC-MS experiments conducted for the analysis of succinate-, propane-, acetate- and methane-grown extracts to provide comprehensive data sets. 1230, 1018, 796 and 829 non-redundant proteins were compiled for succinate-, propane-, acetate- and methane-grown extracts, respectively. Figure 6.7 illustrates the overlap of these identifications. 525 proteins comprised a core set, which represented 35% of total identifications that were observed across all growth conditions.

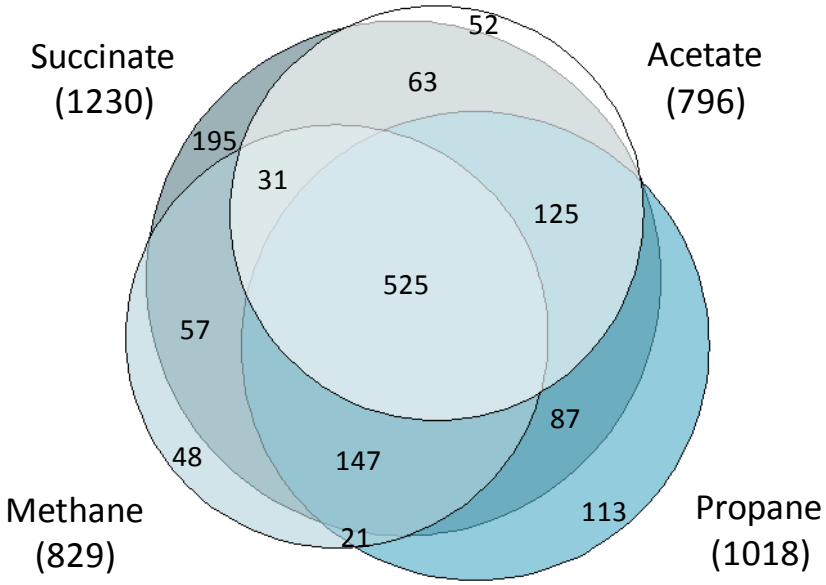

Figure 6.7. Protein identifications combined from all LC-MS experimental approaches

A total of 1465 proteins were identified considering results combined from all growth substrates, experimental conditions and biological replicates. Only 28% of these proteins were uniquely attributed to a single growth condition. Assuming the expression of all 1685 proteins predicted to be in the soluble fraction (cytoplasmic and periplasmic), an 85% coverage of the theoretical proteome was obtained. Figure 6.8 shows an enrichment of soluble proteins seen previously from the LC-MS analysis of proteins expressed during methanol and methylated amine growth.

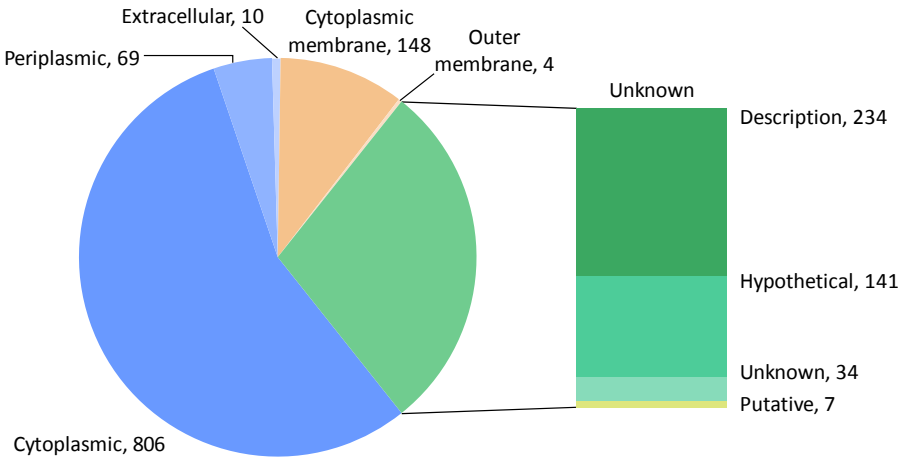

Figure 6.8. Cellular localisation prediction of detected proteome

A small proportion of membrane proteins were also identified. These included large protein complex assemblies, which tend to compose of subunits that are integrated or associated with the membrane as well as subunits that are predominantly located in the cytoplasm. The latter were classed by the algorithm as membrane proteins due to their association with integral membrane proteins. Of the four outer membrane proteins identified, two were hypothetical. Proteins with unknown locations were experimentally under-represented. 234 proteins, whose locations were unable to be predicted using the algorithm, had homology-based descriptions. Approximately 300 detected proteins overall were given hypothetical, unknown or putative descriptions, of which half were assigned locations by the algorithm. Interestingly, 91 of these proteins were identified in all four growth conditions, adding further confidence that these are true positives and that these are essential proteins that are required by the bacterium. Further investigation of hypothetical proteins would be required to deduce their functions.

Figure 6.9 shows the functional categorisation of proteins that were expressed under all growth conditions. Pathways involving the biosynthesis of cellular components (amino acids, fatty acids etc.) and metabolic processes from the KEGG database were populated. Molecular functions including oxidoreductases and nucleotide binding were present, amongst other housekeeping functions such as transcription, translation, and ribosomal structure.

Figure 6.9. Functional analysis of proteins common to succinate-, propane-, acetate- and methane-grown extracts

6.4.3 C_1 and C_2 assimilation pathways

Figure 6.10 shows the abundance of proteins expressed from the *mmo* cluster during growth on succinate, propane, acetate and methane. Molar quantities are shown to provide stoichiometry information and the values were normalised to allow for different loading amounts.

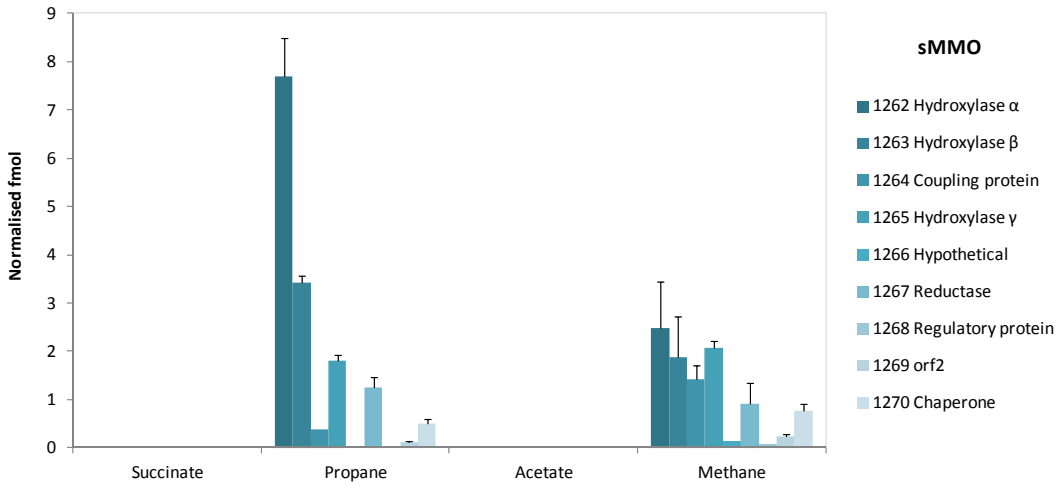

Figure 6.10. Quantification of soluble methane monooxygenase components

Data is shown for:
Succinate (bio rep 1) – 2D RP-RP-LC MS^E (6 fractions) Synapt
Propane (bio rep 3) – 2D RP-RP-LC MS^E (3 fractions) Synapt G2
Acetate – 2D RP-RP-LC MS^E (3 fractions) Synapt G2
Methane – 2D RP-RP-LC MS^E (6 fractions) (Synapt)

All proteins expressed from the *mmo* cluster (Msil_1262 – 1270) were identified in the methane-growth condition. They consisted of 8.2% of the total detected amount by weight (as estimated from 2D RP-RP-LC 6 and 11 fraction analysis). This included the identification of a hypothetical protein (Msil_1266), "Orf2" (Msil_1269) and a regulatory protein (Msil_1268), which were not identified in the previous study by Patel *et al.* (Patel et al. 2009). Similar open reading frames to the hypothetical "Orf2" are present at analogous positions in both *M. trichosporium* OB3b and *M. capsulatus* (Bath). This suggested that Orf2 may be a *bona fide* gene and could have a role in transcription of the *mmo* operon. The expression of "Orf2" under methane and propane growth conditions supported the hypothesis that this gene product is existent. Unlike other methanotroph *mmo* clusters such as those of *M. capsulatas* (Bath), the gene *mmoR,* which expresses the σ^{54}-transcriptional regulatory protein

(Msil_1268), is not transcribed independently of the structural genes. This suggests that a novel control mechanism may occur in *M. silvestris*. Transcriptional regulatory proteins are known to exist at relatively low abundance in the cell, so it is perhaps unsurprising that Msil_1268 was only identified when a more extensive fractionation strategy was used. This regulatory protein was exclusively identified during methane and propane growth by the 2D-MS^E and 2D-$HDMS^E$ approach, respectively. These data show the advantage that fractionation strategies provide for the quantification of lower abundance proteins.

The proteomics study carried out by Patel *et al.* demonstrated that structural subunits of sMMO were elevated 60-fold under methane growth when compared to acetate growth. No proteins in the Msil_1262 – 1270 cluster were identified using the conventional 1D RP-LC MS^E approach or by using the extensive fractionation strategies for studying acetate growth. This is in agreement with transcriptional analysis, which showed an absence of sMMO during acetate growth (Theisen *et al.* 2005).

Data obtained from the proteomic analyses conducted here demonstrated that sMMO is largely absent during growth on succinate (one subunit present at 0.009% in bio rep 3) and abundant during growth on propane (consisting of 5 – 15.9% of the total detected amount from the three biological replicates tested). The possibility that propane may induce the expression of sMMO is an exciting prospect, which may be of biotechnological importance. Whether succinate acts as a repressor, or propane an inducer, are also factors that may warrant further investigation.

The requirement for isocitrate lyase during growth on C_1 and C_2 compounds has been demonstrated using a *M. silvestris* deletion strain (Crombie and Murrell 2011). Deletion of the gene relating to the enzyme had little effect on growth on succinate, but abolished growth on methanol and severely restricted growth on acetate. Proteins involved in the recently discovered EMC pathway for acetate metabolism were searched against the *M. silvestris* database. These revealed a set of homologous proteins. The identities and locations of these genes, however, suggests that the EMC pathway does not operate in *M. silvestris*. Overall, it is likely that acetate is assimilated via the glyoxyate cycle.

The current version of the *M.silvestris* database was used to convert protein identification results published from the Patel study to enable comparison with the data acquired in this work. Relative quantification using Expression software was not

achievable with the data resulting from the analysis of methane and acetate growth conditions here, due to the different experimental conditions which were employed. An approximate measure of relative abundance was obtained by using the absolute quantity values by the Hi3 method, which were previously shown to be in fairly good agreement with relative levels calculated using the Expression software (section 5.3.3)

Figure 6.11 shows the levels of glyoxylate enzymes measured between different growth conditions, including those values obtained from the Patel study. Considering that relative quantification was performed using either the Expression software or by taking the estimated absolute abundances, and that these studies were carried out two years apart by slightly different methods on different samples, the results from the 2D RP-RP-LC MS^E approach for methane versus acetate are in good agreement with the values from the 1D RP-LC MS^E approach employed in the Patel study. The key enzyme, isocitrate lyase, previously unidentified during growth on methane, was identified during growth on all studied substrates, predominantly using the 2D-LC approaches. This enzyme was found to be up-regulated during acetate-growth in comparison to that under succinate-, methane- and propane-growth. Enhanced operation of the glyoxylate cycle during growth on acetate may be due to the increased availability of acetyl-CoA, which is generated from acetate. Acetyl-CoA is also generated by the serine cycle, during the assimilation of methane by *M. silvestris*. Operation of the serine cycle, in the absence of the EMC pathway, requires the activity of isocitrate lyase, together with enzymes of the TCA cycle. The fact that the glyoxylate-specific enzyme isocitrate lyase and the majority of other glyoxylate enzymes are up-regulated in acetate when compared to methane may suggest that the bacterium favours growth on acetate over methane. This is also reflected in the comparatively slow growth on methane compared to acetate. Growth on succinate is feasible since it serves to replenish 2-oxaloacetate, (which is withdrawn from the cycle to provide carbon precursors for biosynthesis) without the requirement of glyoxylate enzymes. There appears to be a minor role for glyoxylate enzymes in succinate when compared to acetate, shown by the up-regulation of isocitrate lyase during acetate growth, and no consistent variations in the expression of other glyoxylate enzymes. These other enzymes, however, have roles in the citric acid cycle and the serine pathway, and so expression levels may not be indicative of a single pathway. Expression of glyoxylate enzymes and other pathways utilised during propane-growth are discussed in section 6.4.4.

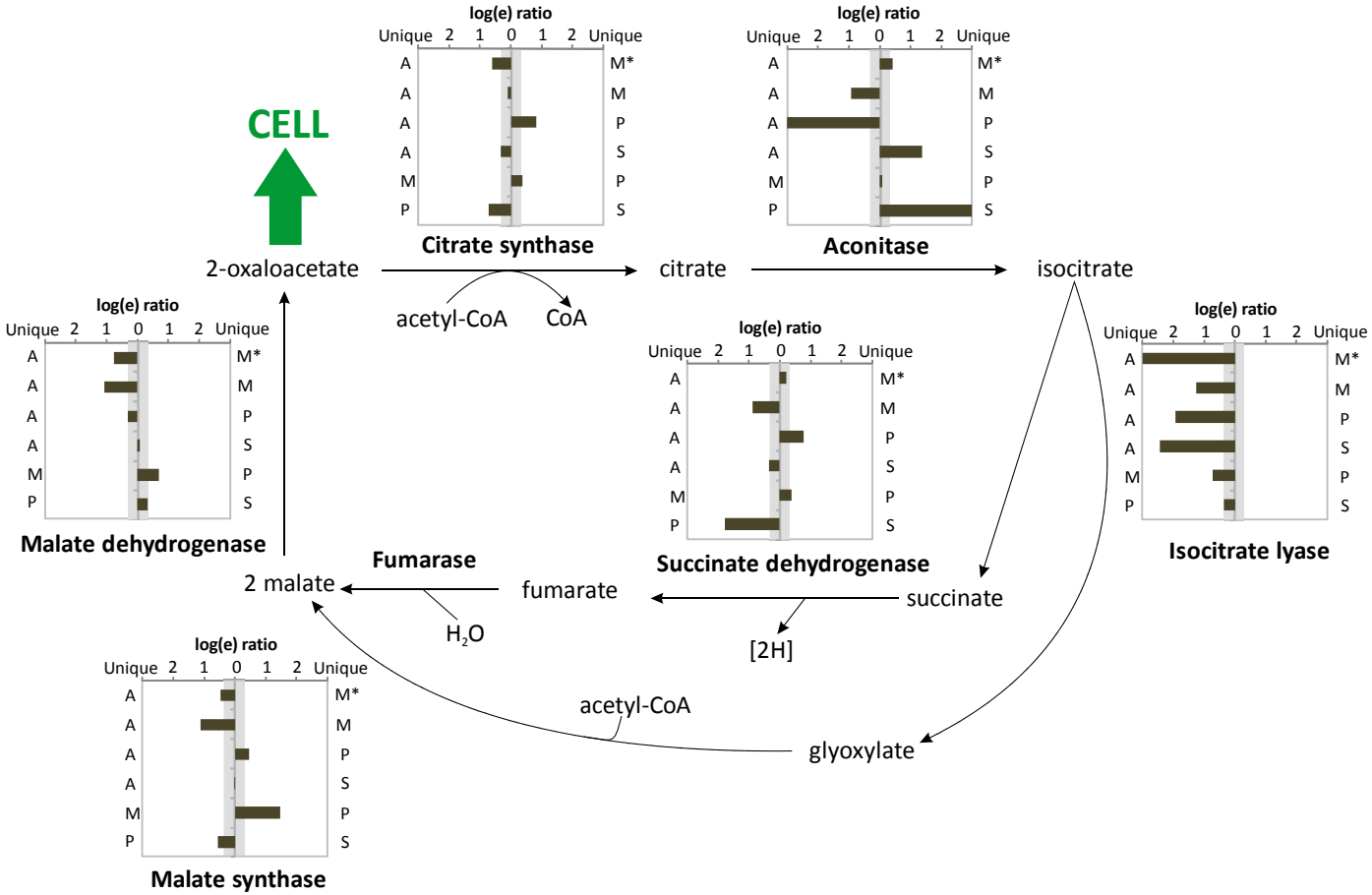

Figure 6.11. Expression of glyoxylate enzymes

*Methane versus acetate relative expression values from (Patel *et al.* 2009). Log(e) values of -0.3 to 0.3 fall within *grey-shaded boxes*. Acetate (A), Methane (M), Succinate (S) and Propane (P). Data shown for: A:M 2D-LC HDMSE and 2D-LC MSE; P (bio rep 3):A 2D-LC HDMSE; S (bio rep 3):P (bio rep 3) 2D-LC HDMSE; P (bio rep 2):M 2D-LC MSE; S (bio rep 3):A 2D-LC HDMSE

6.4.4 Propane assimilation and central metabolic pathways

Structural genes relating to the initial enzymes involved in propane oxidation (hydroxylase, reductase and coupling protein) were predicted from the genome of *M. silvestris* based on homology (Msil_1646 – 1651). The proteins relating to these genes were highly expressed upon growth with propane when compared to levels observed from methane-, acetate- and succinate-growth. Msil_1646, a regulatory protein, was not identified in any propane-grown extracts. From the analyses of three propane-grown biological replicates, propane monooxygenase (PMO) was estimated to be present at 4.7 – 13.5 % of the soluble protein content. Figure 6.12(a) shows the normalised levels of the components expressed from Msil_1647 – 1651. The relative levels of PMO components were similar to those expressed by equivalent components in the *mmo* cluster.

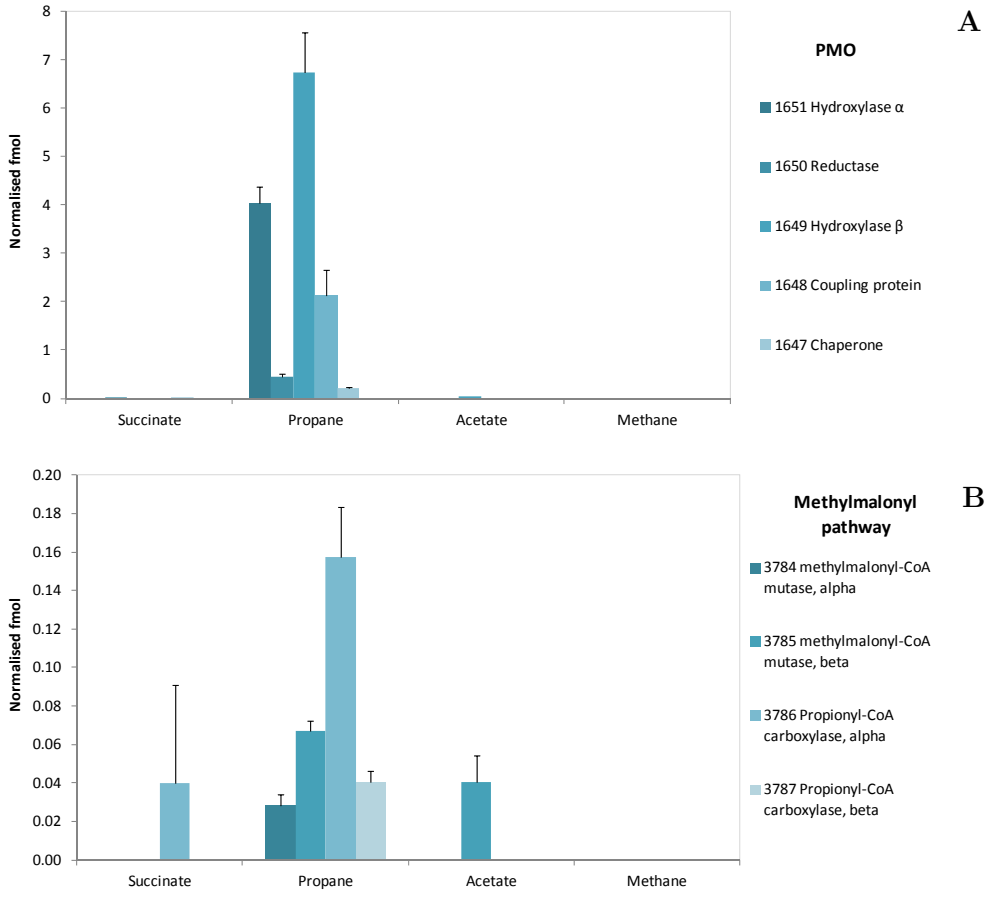

Figure 6.12. Protein quantities relating to propane assimilation

Absolute quantification of (a) propane monooxygenase components and (b) methylmalonyl pathway enzymes. Results shown from data sets used in Figure 6.10.

Additional proteins, detected from propane samples, which may be involved in the propan-1-ol metabolism (terminal oxidation pathway) included propionyl-CoA carboxylase (Msil_3786 – 3787 alpha and beta), methylmalonly-CoA mutase (Msil_3784 – 3785 alpha and beta) and methylmalonyl CoA epimerase (Msil_2934). These form part of the methylmalonyl-CoA pathway (α-carboxylation) of propionate assimilation which results in four-carbon succinate (Wegener *et al.* 1968). The majority of these proteins were not detectable in succinate-, methane- and acetate-grown samples (Msil_3786 and Msil_3785 were detected at lower levels during succinate-growth and acetate-growth, respectively), suggesting that propane may be, in part, assimilated by this route. Concentrations of these proteins are shown in Figure 6.12(b). These proteins were estimated to constitute 0.5 – 1.2% of the soluble content during growth on propane. Msil_3433 and Msil_3434 are also described by the genome as propionyl-CoA carboxylase and methylmalonyl-CoA mutase. These proteins were not detected during propane-growth, suggesting that these may be redundant copies or that they may be expressed under different conditions.

The products formaldehyde (one-carbon) and acetyl-CoA (two-carbon) are assimilated via the serine and glyoxylate cycles, respectively. Possible alternative routes for propane assimilation, via terminal or sub-terminal pathways, may result in formaldehyde or acetyl-CoA. This would require the induction of the serine cycle or the glyoxylate cycle for assimilation. A comparison of these enzyme levels during propane- and succinate-growth did not reveal any consistent pattern of up- or down-regulation, which may suggest little preference for these alternative carbon assimilation pathways. Further biochemical and genetic methods would be required to test this hypothesis. Formaldehyde is a known product of methane oxidation, so it may be expected that serine cycle enzymes would be up-regulated in methane-grown cells when compared to succinate-grown cells. This was experimentally observed. It may be suggested that formaldehyde and serine cycle enzymes would be up-regulated in methane when compared to propane, if propane was not assimilated via formaldehyde. This was observed, with one of three possible serine-pathway exclusive enzymes (Msil_1713 hydroxypyruvate reductase) up-regulated in methane along with the majority of other serine pathway enzymes. These relative measurements were obtained with confident regulation values ($0.05>p>0.95$) and are shown in Table 6.1.

Table 6.1. Proteins involved in central metabolism

[a]*Substrate1: Substrate2* ratio is expressed as log(e). Positive values indicate up-regulation in *Substrate 1* and negative values indicate up-regulation in *Substrate 2*. Where proteins were uniquely identified in a particular substrate, the name of the substrate is stated instead of the ratio.

[b]*p* values as reported by Expression software. Statistically regulated proteins fall within 0.05>p>0.95.

Description	Gene No. Msil_	[a]Propane: Acetate	[b]*p* value	Succinate: Propane	*p* value	Propane: Methane	*p* value	Succinate: Acetate	*p* value
C_1 transfer									
formaldehyde activating enzyme	2390	0.77	1	0.11	0.85	2.85	1	-0.05	0.16
methylene H4MPT DH	2385	0.10	0.66	0.25	0.91	Methane		0.07	0.68
methylene H4MPT cyclohydrolase	2387	-0.21	0.05	0.60	1	-0.12	0.29	0.34	1
formyltransferase/hydrolase complex α	2401	-0.20	0.14	0.22	0.95	0.03	0.57	0.37	0.97
formyltransferase/hydrolase complex β	2400	0.04	0.58	0.19	0.98	0.2	0.97	0.06	0.65
formyltransferase/hydrolase complex γ	2403	0.14	0.7	-0.05	0.35	0.09	0.73	-0.22	0.06
formyltransferase/hydrolase complex δ	2402	-0.78	0	0.18	0.82	-		-0.61	0
Formate oxidation									
formate DH	2953	Propane		0.11	0.85	-0.53	0	Succinate	
formate DH β subunit	3655	Acetate		-		Methane		Acetate	
formate DH α subunit	3656	Acetate		-		Methane		Acetate	
Citric acid cycle									
citrate synthase	1681	0.46	0.94	-0.19	0	0.37	1	-0.29	0
aconitate hydratase	3509	-0.39	0	1.53	1	0.07	0.64	1.09	1
isocitrate dehydrogenase	3530	-0.19	0	0.37	1	0.02	0.61	-0.09	0.07
2-oxoglutarate dehydrogenase	2504, 2505, 2506	0.47, 0.18, -1.04	1, 0.96, 0	- 0.02, 1.11, 0.96	0.4, 1, 1	0.95, 0.69, 0.68	1, 1, 1	0.22, 0.27, -0.1	1, 0.99, 0.23
succinyl-CoA synthetase	2503	1.87	1	-0.87	0	1.3	1	0.53	1
succinate DH β	3562	1.04	1	-1.23	0	Propane		-0.20	0.13
malate dehydrogenase	2501	0.35	1	-0.02	0.4	0.71	1	0.20	1

Table 6.1. continued.

[c]Likely to be exclusive to the serine cycle

Description	Gene No. Msil_	Propane: Acetate	*p* value	Succinate: Propane	*p* value	Propane: Methane	*p* value	Succinate: Acetate	*p* value
Serine cycle									
serine-glyoxylate aminotransferase	[c]1714	-0.41	0	0.52	1	0.15	0.23	0.45	1
serine hydroxymethyltransferase	[c]2110	-0.21	0	0.31	1	-0.24	0	0.09	0.96
hydroxypyruvate reductase	[c]1713	-0.11	0.18	1.81	1	-1.70	0	0.94	1
glycerate kinase	1715	-1.20	0	1.53	1	-1.60	0	0.30	1
phosphoenolpyruvate carboxylase	1718	-0.18	0.17	0.85	1	Methane		0.83	1
malate thiokinase α	1716	0.71	1	0.44	1	-1.01	0	1.04	1
malate thiokinase β	1717	0.14	0.76	0.30	1	-0.95	0	0.66	1
malyl-CoA lyase	1719	-0.55	0	1.46	1	-1.35	0	1.12	1
enolase	0515	-0.16	0.01	0.08	0.87	-0.13	0.02	-0.04	0.19
malate dehydrogenase	2501	0.35	1	-0.02	0.4	0.71	1	0.20	1
citrate synthase	1681	0.46	0.94	-0.19	0	0.37	1	-0.29	0
aconitate hydratase	3509	-0.39	0	1.53	1	0.07	0.64	1.09	1
succinate DH β	3562	1.04	1	-1.23	0	Propane		-0.20	0.13
Pyruvate (C_3 metabolism)									
pyruvate carboxylase	1360	0.94	1	-0.91	0	0.90	1	0.08	0.81
phosphoenolpyruvate carboxykinase	1375	-0.69	0	0.53	1	Propane		-0.35	0

Data shown for:

Propane (bio rep 3):Acetate 2D RP-RP-LC HDMS[E]
Succinate (bio rep 3):Propane (bio rep 3) 2D RP-RP-LC HDMS[E]
Propane (bio rep 2):Methane 2D RP-RP-LC MS[E]
Succinate (bio rep 3):Acetate 2D RP-RP-LC HDMS[E]

The majority of serine pathway enzymes were found to be up-regulated during succinate-growth when compared to propane- and acetate-growth. This suggested that there may be a role for these enzymes during succinate-growth. There was no consistent pattern of regulation for serine pathway enzymes between propane- and acetate-growth.

Propane assimilation may occur via the metabolite pyruvate, which proceeds via its conversion into acetyl-CoA by pyruvate dehydrogenase. It may be expected if this were the case that this enzyme would be up-regulated in propane when compared to methane. This protein was not found to be significantly up-regulated, with a log(e) ratio of 0.28 (propane:methane).

C_1 transfer proteins are involved in oxidising formaldehyde (into formate) or assimilating formaldehyde (via the serine pathway). It be might expected that these proteins would be up-regulated during growth on methane. C_1 transfer proteins were not differentially expressed between any conditions. The presence of C_1 transfer proteins during acetate-, succinate- and propane-growth suggests that these proteins are constitutively expressed. Unlike C_1 transfer proteins, formate oxidation enzymes were shown to be up-regulated in or unique to the methane-growth condition. This is in agreement with the generation of formaldehyde/formate during methane oxidisation. These results suggest that formate, and not formaldehyde assimilation/oxidation is a control point in this pathway. Formate has been shown to be the main branch point for methylotrophic metabolism in facultative methylotroph *Methylobacterium extorquens* AM1 (Crowther *et al.* 2008).

The majority of enzymes involved in the citric acid cycle were up-regulated during propane-growth when compared to acetate- and methane-growth. This may suggest a preference for growth on propane. There was no consistent pattern observed for the levels of these enzymes between succinate- and propane-growth. This may be indicative of propane assimilation via succinate, a citric acid cycle intermediate.

Table 6.2. Number of differentially expressed proteins and fold-changes

Substrate 1 /Substrate 2	Up-regulated Substrate 1/ (%)	Unregulated/ (%)	Up-regulated Substrate 2/ (%)	Fold-change	
				Range	Median
Propane/Acetate	217 (40)	170 (31)	154 (29)	1.3 - 60	1.9
Succinate/Propane	185 (28)	289 (43)	195 (29)	1.3 - 81	1.9
Propane/Methane	108 (25)	228 (52)	101 (23)	1.3 - 45	1.8
Succinate/Acetate	211 (32)	298 (46)	137 (22)	1.3 - 17	1.7

Same datasets used for Table 6.1.

Table 6.2 provides details of the numbers of proteins that were up-regulated for a particular growth condition and those which showed no distinct differential expression. For a single comparison, approximately half the proteins were not differentially expressed with the other half up-regulated for a particular growth condition. Of the proteins that were differentially expressed, they ranged from a fold-change of 1.3 up to 81. The majority of proteins did not exhibit extreme fold-changes, as shown by the median fold-changes of 1.7 – 1.9 observed. The extent of differential regulation within a particular system can be an important factor when designing an approach to measure relative quantification levels. If relative errors of 25% were present for a protein present in two samples with a fold-change of 2, this would result in a measured fold-change of 1.2, which would not normally be considered to be a differentially regulated protein.

6.5 Conclusions

The proteomic study of *M. silvestris* BL2 grown on succinate, propane, acetate or methane has provided information regarding the utilisation of C_1 and C_2 compounds. These data are in good agreement with previous biochemical and transcriptional studies. In comparison to a previously published related proteomic study, higher proteome coverage and similar quantitative relative measurements were achieved for the analysis of the acetate- and methane-growth conditions. Data obtained from 2D-LC and $HDMS^E$ experimental approaches confirmed the expression of a key glyoxylate enzyme, isocitrate lyase, during all studies growth conditions. A relatively low abundance transcriptional regulator for the *mmo* operon was exclusively identified during growth on methane. Soluble methane monooxygenase components were also identified during growth on propane. This is an exciting discovery that may be of significant biotechnological importance.

Biological replicates were analysed for the succinate- and propane-growth conditions. Considering the slight variations in bacterial growth conditions, good agreement was observed between identifications as well as a reasonable correlation in the protein abundance measurements.

Data sets from various LC-MS analyses and biological replicate analyses were combined for each of the four studied growth conditions. A good overlap of identifications was obtained between the various growth conditions, comprising a core set of proteins that may be involved in essential functions. These core proteins may contain a subset of proteins that do not differ in their expression when subjected to a wide range of environmental stimuli. Further investigation of these proteins may provide a list of abundant non-changing peptides that could be used for normalisation in label-free studies. 1465 non-redundant proteins were identified across all four studied growth conditions, which corresponds to approximately 85% of the theoretical soluble proteome.

Propane monooxygenase components were up-regulated during growth on propane when compared to all other growth conditions. Expression of additional enzymes suggests that assimilation proceeds, in part, via propan-1-ol and the methylmalonyl-CoA pathway. A differential regulation of downstream pathways supported the hypothesis that the preferred propane assimilation occured via the

molecule succinate. These results do not, however, conclusively dismiss the possibility that propane could also be assimilated via acetyl-CoA, pyruvate or formaldehyde.

6.6 References

Alber, B. E., Spanheimer, R., Ebenau-Jehle, C. and Fuchs, G. (2006). Study of an alternate glyoxylate cycle for acetate assimilation by *Rhodobacter sphaeroides*. *Mol. Microbiol.* **61**, 297-309.

Crombie, A. and Murrell, J. C. (2011). Development of a system for genetic manipulation of the facultative methanotroph *Methylocella silvestris* BL2. *Method. Enzymol.* **495**, 119-33.

Crowther, G. J., Kosály, G. and Lidstrom, M. E. (2008). Formate as the Main Branch Point for Methylotrophic Metabolism in *Methylobacterium extorquens* AM1. *J. Bacteriol.* **190**, 5057-5062.

Dedysh, S. N., Knief, C. and Dunfield, P. F. (2005). *Methylocella* species are facultatively methanotrophic. *J. Bacteriol.* **187**, 4665-4670.

Etiope, G. and Ciccioli, P. (2009). Earth's Degassing: A Missing Ethane and Propane Source. *Science.* **323**, 478.

Nagaraj, N. and Mann, M. (2011). Quantitative Analysis of the Intra- and Inter-Individual Variability of the Normal Urinary Proteome. *J. Proteome Res.* **10**, 637-645.

Patel, V. J., Thalassinos, K., Slade, S. E., Connolly, J. B., Crombie, A., Murrell, J. C. and Scrivens, J. H. (2009). A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches. *J. Proteome Res.* **8**, 3752-3759.

Peyraud, R., Kiefer, P., Christen, P., Massou, S., Portais, J.-C. and Vorholt, J. A. (2009). Demonstration of the ethylmalonyl-CoA pathway by using 13C metabolomics. *Proc Natl Acad Sci U S A.* **106**, 4846-4851.

Rahman, M. T., Crombie, A., Moussard, H., Chen, Y. and Murrell, J. C. (2011). Acetate Repression of Methane Oxidation by Supplemental *Methylocella silvestris* in a Peat Soil Microcosm. *Appl. Environ. Microbiol.* **77**, 4234-4236.

Shennan, J. L. (2006). Utilisation of C2–C4 gaseous hydrocarbons and isoprene by microorganisms. *J. Chem. Technol. Biotechnol.* **81**, 237-256.

Theisen, A. R., Ali, M. H., Radajewski, S., Dumont, M. G., Dunfield, P. F., McDonald, I. R., Dedysh, S. N., Miguez, C. B. and Murrell, J. C. (2005). Regulation of methane oxidation in the facultative methanotroph *Methylocella silvestris* BL2. *Mol. Microbiol.* **58**, 682-692.

Valentine, D. L., Kessler, J. D., Redmond, M. C., Mendes, S. D., Heintz, M. B., Farwell, C., Hu, L., Kinnaman, F. S., Yvon-Lewis, S., Du, M., Chan, E. W., Tigreros, F. G. and Villanueva, C. J. (2010). Propane Respiration Jump-Starts Microbial Response to a Deep Oil Spill. *Science.* **330**, 208-211.

Wegener, W. S., Reeves, H. C., Rabin, R. and Ajl, S. J. (1968). Alternate pathways of metabolism of short-chain fatty acids. *Bacteriol. Rev.* **32**, 1-26.

Woods, N. R. and Murrell, J. C. (1989). The Metabolism of Propane in *Rhodococcus rhodochrous* PNKb1. *J. Gen. Microbiol.* **135**, 2335-2344.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

7.1.1 Label-free MS^E approaches for comparative and profiling proteomic studies

Since the first application of MS^E technology for proteomic studies (Silva *et al.* 2006), there have been 111 peer-reviewed publications which have involved the use of MS^E to provide information on protein identification and quantification. The use of data-independent acquisition for proteomic applications is becoming recognised as a valuable method which can in some circumstances offer advantages over traditional data-dependent methods. The first discussion group dedicated to data-independent approaches at ASMS Conference on Mass Spectrometry and Allied Topics 2012 was another example of the increasing popularity of the approach. The development of a targeted data-independent approach termed SWATH (AB Sciex) provides an indication of demand for data-independent approaches (Gillet *et al.* 2012).

This research has demonstrated recently developed LC and MS technologies that have contributed to improved label-free MS^E proteomic approaches. The use of separation technologies such as 2D-LC and ion mobility is expected to offer advantages including increased protein identification. The use of 2D RP-RP-LC experiments was shown to provide improved dynamic range and data quality, together with quantitation that was comparable to 1D RP-LC. An approach which included ion mobility-assisted MS^E coupled with 1D RP-LC separation was shown to provide a higher number of protein identifications with a significantly reduced experimental time but at the expense of a reduced effective dynamic range, when compared to an approach using 2D RP-RP-LC and MS^E. The effective dynamic range was improved by using a combination of 2D RP-RP-LC and ion mobility-assisted MS^E. This method was also shown to provide similar proteome coverage when compared to other comprehensive mass spectrometry-based proteomic approaches.

7.1.2 Metabolic insights into facultative methanotrophy

This research revealed some potentially interesting enzymes in a unique bacterium, which have the potential to be studied further. The gene expressing the enzyme trimethylamine monooxygenase was identified for the first time, and was found to be highly abundant in the genomes of bacteria in the surface ocean. Previously identified

N-methylglutamate pathway enzymes for monomethylamine metabolism were characterised further. The structural composition and substrate specificity of some these enzymes would benefit from further investigation.

Methylocella silvestris was the first bacterium discovered with an ability to grow at the expense of both propane and methane. Proteomic studies of the bacterium grown with propane, succinate, acetate or methane suggested that propane, in part, is assimilated via the methylmalonyl pathway. The expression of methane monooxygenase during growth on propane is an exciting discovery, which may be of biotechnological importance.

7.2 Future Directions

The continuing development of analytical and computational approaches will improve label-free MS^E quantification of proteomes. Recent developments in computational approaches have included a version of Progenesis LC-MS (Nonlinear dynamics), which supports Thermo, Bruker, Waters, Agilent and AB Sciex data formats. This program enables the comparison of different experimental conditions using a high number of replicates, and has been shown to provide the most accurate ratios compared to other available bioinformatic platforms (Sprenger *et al.* 2012). To combine and compare results from different experiments, Progenesis LC-MS aligns them to compensate for between-run variation in the LC separation technique. This is shown in Figure 7.1. The peak picking algorithm can handle complex samples and can differentiate between overlapping peptide ions. An important and perhaps overlooked feature of this program is the ability to select peptides for quantification. This can be a problem with current PLGS software, which can include peptides for quantification which may be shared between two or more proteins.

Despite impressive developments in LC-MS based proteomics techniques, identification and quantification variation is still a concern. Thalassinos *et al.* recently described the use of a data-independent fragment ion repository for protein identification, quantification and validation (Thalassinos *et al.* 2012). Information is stored in a relational database and may be used to create peptide- and protein-specific fragment ion maps that can be queried in a targeted fashion against raw or processed ion detections. Individual or group queries were shown to provide information

regarding pathway and molecular machinery analysis of the protein complement from two closely related mammalian species. It was also demonstrated that LC-MS performance metrics could be extracted and utilized to ensure optimum performance of the analytical workflows employed.

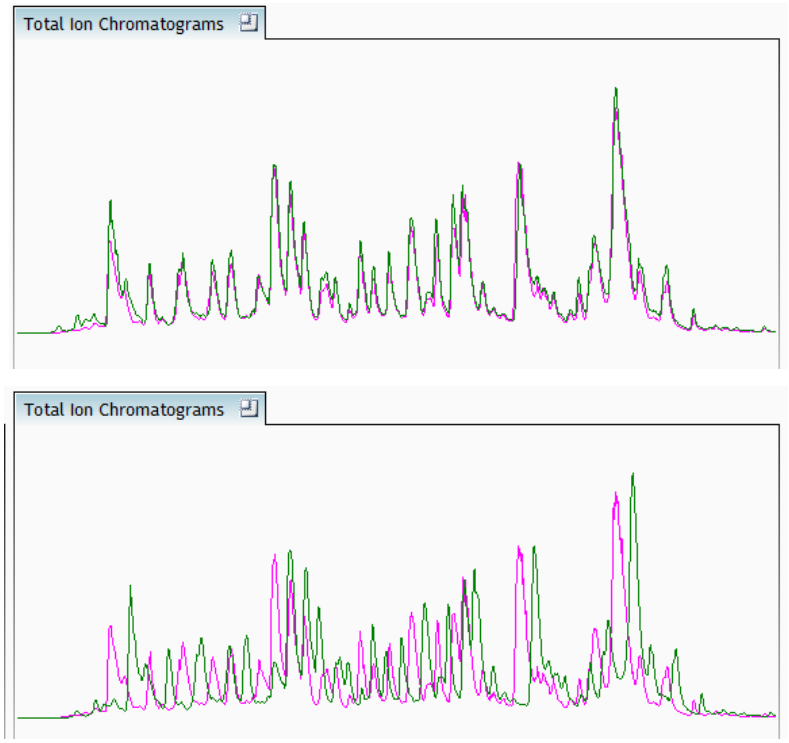

Figure 7.1. Chromatographic alignment by Progenesis LC-MS (Nonlinear dynamics)

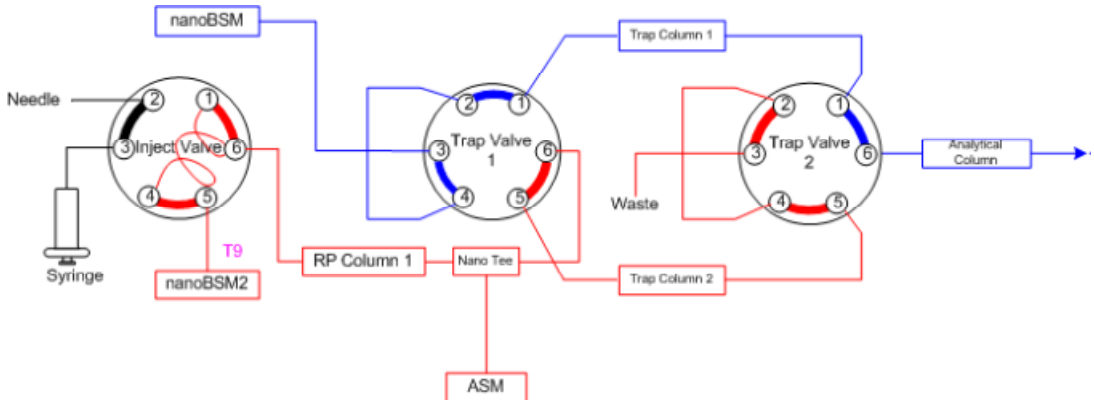

Figure 7.2. High throughput online 2D-LC system

Taken from (Stapels *et al.* 2012)

Figure 7.2 shows a schematic of a prototype online 2D-LC system which employs the use of two trapping columns. This essentially reduces the time it takes to load sample per fraction and may greatly reduce experimental time (Stapels *et al.* 2012). The experimental time for a six fraction experiment could be reduced by 90 min if each loading step was 15 min. The availability of high-throughput online 2D-LC may provide an attractive option when developing strategies which involve the analysis of a large number of samples.

A number of important subfields of proteomics lie outside the scope of this research. These include targeted quantitative proteomics and the identification of PTMs. Targeted approaches are considered the most accurate method for obtaining quantification information, but can be low-throughout and relatively expensive when compared to discovery approaches. A reduction in the cost of peptide synthesis may contribute to the increased use of targeted approaches to validate quantification information obtained using less accurate discovery-based approaches. The significant challenge posed by relating proteomes and physiological states could be overcome by integrating information obtained from various -omic technolgies and by continuing improvements in data-handling and instrumentation.

7.3 References

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R. and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics.* **11**, 17.

Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V., Li, G. Z., Richardson, K., Wall, D. and Geromanos, S. J. (2006). Simultaneous qualitative and quantitative analysis of the *Escherichia coli* Proteome - A sweet tale. *Mol. Cell. Proteomics.* **5**, 589-607.

Sprenger, R. R., Casanovas, A., Sidoli, S., Schwämmle, V., Ejsing, C. S. and Jensen, O. N. (2012). Label-free absolute quantification of proteomes: evaluation and comparison of bioinformatics platforms and strategies. *Proceedings of the 60th ASMS Conference on Mass Spectrometry and Allied Topics*, Vancouver, Canada.

Stapels, M., Thompson, J. W., Moseley, M. A. and Langridge, J. (2012). Increased Throughput for 2D LC in the Analysis of Human Placental Samples. *Proceedings of the 60th ASMS Conference on Mass Spectrometry and Allied Topics* Vancouver, Canada.

Thalassinos, K., Vissers, J. P. C., Tenzer, S., Levin, Y., Thompson, J. W., Daniel, D., Mann, D., DeLong, M. R., Moseley, M. A., America, A. H., Ottens, A. K., Cavey, G. S., Efstathiou, G., Scrivens, J. H., Langridge, J. I. and Geromanos, S. J. (2012). Design and Application of a Data-Independent Precursor and Product Ion Repository. *J. Am. Soc. Mass Spectrom.* In press.