LC-MS for the metabonomic study of human urine samples

Simon John Cubbon

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of York

University of York

Department of Chemistry

December 2007

Abstract

The field of metabonomics is beginning to grow rapidly due to the ability to analyse biofluids, providing a 'snapshot' of biological processes that have happened (cf. proteomic/transcriptomic studies, which predict what *may* happen), making it possible to profile responses over time. The work described in this thesis was motivated by the aim of profiling clinical urine samples obtained from fracture patients, with a view to identifying potential biomarkers related to failed fracture healing. This led to the need to develop and evaluate metabonomic approaches, specifically a orthogonal separation approach complementary to the commonly-used reversed phase (RP) separation methods, namely hydrophilic interaction liquid chromatography (HILIC).

Urine samples from healthy volunteers were collected and used to develop an LC-MS 'metabonomic toolbox'. This development evaluated various aspects of a metabonomic study that are commonly poorly reported within the literature: study design, sample collection storage and handling considerations, data extraction, normalisation and scaling methods, and multivariate data analysis tools.

From the literature, the commonly-used method of normalising to creatinine was found to be unsuitable due to perturbations in the urinary excretion of creatinine due to factors such as illness. Methods used to evaluate system stability were also developed and added to the 'toolbox'. HILIC was successfully used as a separation technique orthogonal to RP, producing comparable results but using different metabolites; this highlights the fact that much potential information is possibly being lost when only RP-LC-MS methods are used for analysis. The need to use both modes of ionisation polarity were also addressed for an increased coverage in biofluid metabolite profiles.

The knowledge gained in the development of the 'metabonomic toolbox' was used for the analysis of clinical urine samples. Despite the lack of properly time-setted samples and none of the recruited patients suffering delayed fracture healing, potential metabolites related to fracture healing were found. However, the samples were very different to previously-analysed samples from healthy volunteers; they showed very large amounts of protein, which had a large range of molecular weights. These were identified proteomically.

Finally, ESI-Q-o-ToF MS/MS, MALDI-ToF/ToF MS/MS and racemic amino acid analysis were used for the structural determination of a pseudomonad biosurfactant, which was identified, unexpectedly, as the cyclic lipopeptide white line inducing principle, WLIP.

Table of contents

Chapter Two – Experimental methods

Chapter Four – Clinical urine sample analysis

Chapter Five – Lipopeptide analysis

List of figures

List of tables

List of abbreviations

APCI	Atmospheric pressure chemical ionisation
ASCII	American Standard Code for Information Interchange
BuOH	Butanol
CE	Capillary electrophoresis
CI	Chemical ionisation
CID	Collision induced dissociation
CLP	Cyclic lipopeptide
cps	Counts per second
CRM	Charge residue model
Da	Daltons
DC	Direct current
DE	Delayed extraction
EI	Electron ionisation
ELISA	Enzyme-linked immunosorbent assay
ESI	Electrospray ionisation
FA	Formic acid
FAB	Fast atom bombardment
FWHM	Full width half maximum
GC	Gas chromatography
HCl	Hydrochloric acid
HILIC	Hydrophilic interaction chromatography
HMW	High molecular weight
HPLC	High performance liquid chromatography
IEM	Ion evaporation model
IT	Ion trap
KE	Kinetic energy
LC	Liquid chromatography
LV	Latent variable
LMW	Low molecular weight
m/z	Mass to charge ratio
MALDI	Matrix assisted laser desorption/ionisation
mAU	Milli absorbance units
MCP	Multi-channel plate
MeCN	Acetonitrile
MeOH	Methanol

MOWSE	Molecular weight search
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MW	Molecular weight
NMR	Nuclear magnetic resonance
OPLS-DA	Orthogonal partial least squares – discriminant analysis
PC	Principal component
PCA	Principal components analysis
PGC	Porous graphitised carbon
PLS-DA	Partial least squares – discriminant analysis
ppm	Parts per million
Q-o-ToF	Quadruploe orthogonal acceleration time of flight
RF	Radio frequency
RMM	Relative molecular mass
rpm	Revolutions per minute
SDS-PAGE	Sodium dodecyl sulfate – polyacrylamide gel electrophoresis
SIMCA	Soft independent modelling of class analogy
Th	Thompsons
TIC	Total ion count
ToF	Time of flight
t_R	Retention time
UPLC	Ultra performance liquid chromatography
UV	Ultraviolet

Acknowledgements

Firstly, I would like to give a huge thank you to Jane Thomas-Oates for letting me pursue a Ph.D. within her group. Over the past three years Jane has helped me tremendously with advice, encouragement and support as well her unhealthy dedication to ensuring that everything I touched during the preparation of this thesis turned bright **pink** (or , or **green**, or **blue**), and had a glittery shine – this is one thing that I am sure I *will* not miss.

A big thank you to all other members of the JTO group past and present, who have made it such a great place to work: Adrian, Barbara, Caroline, Carla, Dave, Emma, Ed, Jimbo, João, Karl, Kriang, NJ, Sally, Sarah, Siân, Tim, and our fellow office-mates: Chris, Deborah, Jayne, Matt and Phil. Thanks must also go to Jerry Thomas, for giving me access to '*his*' instruments in Biology and for his comedic actions over the past three years, Dave Ashford for helping me pacify the somewhat temperamental QStar, and to Julie Wilson for her help with understanding statistics! I am grateful to the EPSRC (DTA Chair) and Smith & Nephew (Steve Fenwick and Martin Todman) for providing financial support that enabled this Ph.D. work to be undertaken.

I would also like to thank all of my friends who have made the last several years in York and Leeds an absolutely amazing time; special thanks are due to the following people: Bungle and Becs, Froglet and Ellie, Blag and Beth, Woody and Ruth, Crofty, Dr. Revvitt, Jimbo, Wheway, Greg and Sally, Arran, SuePoo, G, Tom and Claire, and Nick.

Finally, I would like to thank my family and Lynne. A big thanks goes to my Mum and Dad for being amazingly good parents, to my sisters Emma, Laura and Alison, and to the two dogs in my life: Whiskey and Wilma. Thank you to Lynne, for her friendship, love and support over the last three years, especially the last two months when writing this thesis when I would sit in my own little world paying too little attention to her – here's to Central and South America!

Author's declaration

I hereby declare that the work described in this thesis is my own, except where otherwise acknowledged, and has not been submitted previously for a degree at this or any other university.

S. J. Cubbon

Chapter One

Introduction

1.1. Overview

The work described in this thesis was motivated by the original aim of profiling clinical urine samples obtained from fracture patients, in research originated in collaboration with Smith & Nephew, UK. This led to the work described in this thesis being primarily focussed upon investigating certain aspects of, and developing further methods for, liquid chromatography-mass spectrometric metabonomic studies. As the field of metabonomics is relatively young (Nicholson, 1999) compared to other better established 'omic' techniques, there was (and still is) a need to consider many different aspects of the experimental approach that are not always covered within the literature, so that a robust 'metabonomic toolbox' can be created and utilised for the study of the clinical urine samples, with the original hope of identifying candidate biomarkers for further study.

In addition to the metabonomic studies of human urine samples, the same methods as are used for metabonomic studies were exploited for the structural identification of a biosurfactant from a soil bacterial isolate.

1.2. Introduction

The human skeletal system consists of 206 bones (upon maturity) that function to support, protect and store. Whilst being strong, bones can fracture; given the nature of human activity bone fractures will and do occur. Most fractures that occur within the human body will heal successfully given some degree of medical intervention, however, there are occasions where a fracture will fail to heal successfully (delayed or non-union) and require further medical attention. The number of non-unions has been described as "alarming"; in the United States alone there were nearly 100,000 non-union treatments performed in 2003 (Jones, 2005).

Metabolic bone diseases such as Paget's[1], osteoarthritis[2] and osteoporosis[3] are a common cause of pathological fractures in the elderly. Over three million people suffer from osteoporosis in the UK, and as a result there are over 200,000 fractures annually (NHS-Direct). Studies have utilised biofluids such as serum and urine to try

---

[1] Increased and irregular bone formation leading to larger, weaker bones.
[2] The most common type of arthritis, primarily affecting joints.
[3] Bone structure becomes porous (osteoporosis literally means 'porous bones').

and find indicators of metabolic bone disease, with a view to the possibility of earlier diagnosis and treatment of these pathological bone diseases.

The overall processes behind the biochemistry of bone formation and resorption (bone turnover) are reasonably well known, but in-depth knowledge is lacking; numerous studies are currently trying to gain a further insight into the complex processes involved in bone turnover (Fisher *et al.*, 2005; Heer *et al.*, 2005; Mandelin *et al.*, 2005; Leu *et al.*, 2006; Nancollas *et al.*, 2006; Viguet-Carrin *et al.*, 2006).

Many studies have identified biomarkers in both serum and urine that have been used to monitor bone turnover (Calvo *et al.*, 1996; Eyre, 1996; Woitge *et al.*, 1998; DeLaurier *et al.*, 2004; Miller, 2005; Weisman and Matkovic, 2005). However, it has been noted that no single biomarker can provide a reliable results for monitoring bone turnover (Calvo *et al.*, 1996). As analytical techniques such as gas chromatography-mass spectrometry (GC-MS), high performance liquid chromatography-mass spectrometry (HPLC-MS) and nuclear magnetic resonance (NMR) provide increased selectivity and sensitivity over more traditional assay based techniques, they will become an increasingly important techniques for biomarker analysis and identification.

Pathological bone fractures, which are caused by metabolic bone diseases, have received most of the attention in the literature on bone turnover studies; osteoporosis being the predominant disease is therefore the most studied. Many potential serum biomarkers such as osteocalcin, alkaline phosphatase, pyridinoline, deoxypyridinoline, tartrate resistant acid phosphatase, hydroxylysine glycosides and urinary biomarkers including pyridinoline, deoxypyridinoline, osteocalcin or cross-linking telopeptides have been identified as biochemical markers that are suitable for use in the management of osteoporosis (Ebeling and Åkesson, 2001). Research has shown that recently sustained pathological fractures influence the levels of biochemical markers of bone turnover (Obrant *et al.*, 2005); given that the levels of biomarkers can be profiled for fractures that are a result of a metabolic bone disease, there is good reason to propose that the biochemical response to a non-pathological fracture may also be profiled in a similar way.

Since this study was commissioned, there have only been a handful of papers published that study non-pathological fractures and their delay or failure to heal. Of these papers none studied urine as a biofluid, all having chosen to concentrate on

serum. Zimmermann *et al.*, (Zimmermann *et al.*, 2005) concentrated on the transforming growth factor family, and in particular TGF-β1 as a marker of delayed fracture healing, whilst Henle *et al.*, (Henle *et al.*, 2005) studied matrix metalloproteinases and failed fracture healing. These studies were designed to look at specific compounds within serum; as serum is a homeostatically controlled biofluid there may be more hope for the analysis of urine, which is a biofluid that should contain biomarkers that have been removed from serum as they are no longer required, and are therefore a functional end-point.

One area of analytical science that is suitable for the analysis of biofluids, and has seen substantial growth are the related fields of metabonomics and metabolomics. These 'omic' technologies typically involve a less targeted approach to analysis, being hypothesis forming, rather than hypothesis driven. They utilise analytical methodologies such as capillary electrophoresis (Pisitkun *et al.*, 2006; Ullsten *et al.*, 2006; Monton and Soga, 2007), gas chromatography (Kopka, 2006) and high performance liquid chromatography (Plumb *et al.*, 2005; Wilson *et al.*, 2005; Lenz and Wilson, 2007), which are typically coupled to a mass spectrometer, or use NMR (Robertson *et al.*, 2000; Constantinou *et al.*, 2005; Bertram *et al.*, 2007) (or recently HPLC-NMR-MS (Bajad *et al.*, 2003)). These methods typically provide separation of the complex biofluid matrix components, and their subsequent detection (and ideally quantification).

Techniques such as NMR and HPLC-MS create vast amounts of data, which have been called mega-variate data, which would be impossible to visually interpret. Thankfully, methods such as multi-variate data analysis (MVDA) which encompass techniques such as principal component analysis (PCA) and partial least squares – discriminant analysis (PLS-DA) have been developed, which allows the comparison of large datasets with few observations (samples) and many thousands of variables. The dimensionality of datasets is reduced before any differences (such as metabolites or groups of metabolites that are related to the condition being investigated) can be pinpointed and subjected to further investigation (Dunn and Ellis, 2005; Wilson *et al.*, 2005; Chen *et al.*, 2007).

1.3. The skeletal system

The skeletal system performs a vital role within the human body; it supports muscle, which allows locomotion (mechanical role), affords protection for vital organs and stem cells (protective role) and also provides a large reserve of ions (metabolic role). The body stores a total of about 1000 g of calcium and 600 g of phosphate, 99 and 65 % of which respectively, is housed within the skeletal system (American Society for Bone and Mineral Research, 1999).

1.3.1. Bone remodelling

Bone is a living tissue and as such is in a dynamic state as it is continually being formed by osteoblasts and broken down by osteoclasts (figure 1.3.1); this is known as bone turnover. As the skeleton is in a dynamic state, it can continually adapt to changes in its physiological and mechanical environment, i.e. periods of heavy, repeated stress on the skeleton would cause more bone to be laid down, whereas long periods of inactivity can cause bone to be resorbed, weakening the skeleton.

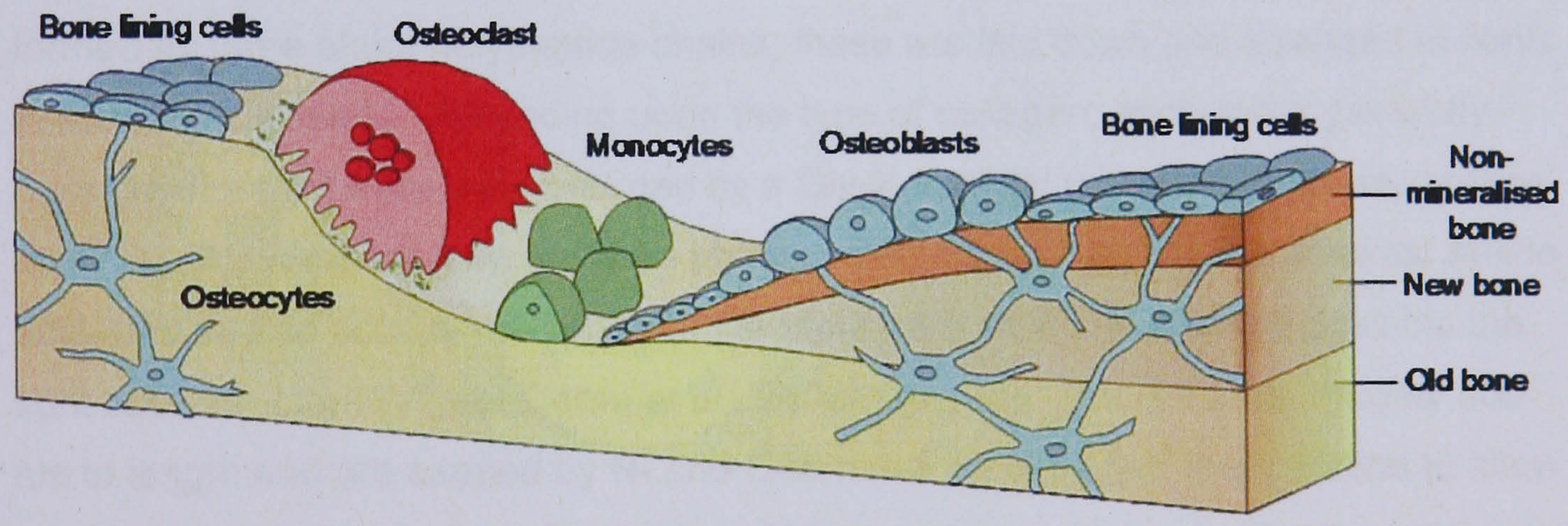


Figure 1.3.1. A schematic representation of the components involved in the bone remodelling process. Courtesy of Leah Etheridge, Dept. Biology, University of York, UK.

The process of bone turnover is highly complex and is still not fully understood. In order to maintain bone (dynamic tissue) it is removed by resorption so that new bone can be laid down in its place. Osteoclasts are giant nucleated cells that contain four to twenty nuclei; parathyroid hormone (PTH) is released by the parathyroid gland. This induces the production of vitamin D_3 [1,25$(OH)_2$ vitamin D_3], which in turn

induces the differentiation of marrow monocytes, from bone marrow, into osteoclasts. Each resorptive site only contains one or two osteoclasts that secrete a range of enzymes which function to digest the bone and release the stored calcium and other minerals into the blood stream. The secretion of tartrate resistant acid phosphatase (TRAP) lowers the pH at the resorptive site. TRAP actively digests the hydroxyapatite crystals that are linked to collagen; collagenase digests the exposed collagen fibres, leaving a site that is now ready for new bone to be laid down.

Normally, bone formation only occurs at a site that has previously undergone bone resorption and typically deposits around 0.55 μm of new bone per day (American Society for Bone and Mineral Research, 1999). Osteoblasts are formed by the proliferation of stem cells (found in bone marrow), which is induced by a drop in the serum levels of PTH, caused by an increase in the hormone oestrogen. Whereas osteoclasts are few in number at a resorptive site, osteoblasts form clusters of 100 to 400 cells and are rich in alkaline phosphatase, and secrete type I collagen that matures and allows the deposition of minerals.

Bone matrix is a two-phase system; collagen provides ductility and the ability to absorb energy whereas minerals provide rigidity. Collagen has a triple helix structure formed by three alpha polypeptide chains; these are laid down and arranged in fibrils in concentric strands. Depending upon the type of collagen, each chain generally consists of a tight triple helix afforded by a Gly-X-Y triplet (where X is usually proline and Y hydroxyproline). Glycine is an absolute requirement as it is the smallest amino acid, and so can occupy the centre of the triple helix structure, making possible the tight helix structure (Viguet-Carrin *et al.*, 2006). The collagen fibrils are roughly 300 nm in length and are capped by N- and C-terminal propeptides[1] that function to allow the association of newly synthesised procollagen chains. 95 % of collagen content in bone comprises type I collagen (and also accounts for ~ 80 % of the total protein content in bone); type III and V collagens are present at low levels and function to modulate the diameter of the type I collagen fibrils.

Collagen is initially un-mineralised. As bone matures, hydroxyapatite crystals are deposited and the bone becomes calcified, increasing the stiffness. Further stability is derived from inter and intramolecular cross-links as the N- and C-terminal

[1] The prefix pro- refers to an inactive protein or peptide. They can be activated by posttranslational modification (a chemical modification).

telopeptides[1] are oxidatively deaminated. This results in cross links being formed through condensation with a lysyl or hydroxylysyl side chain on an adjacent collagen fibril.

The three dimensional architecture of bone, its shape and geometry and the intrinsic properties of the matrix (mineral and collagen), give the skeleton the capacity to resist mechanical forces. However, the amount of force that a bone can withstand before giving way and causing a fracture is dependent upon the quality and quantity of bone tissue that is laid down.

1.3.2. Fractures

Bone is anisotropic in nature, meaning that it has different mechanical properties when it is loaded along its different axes; bones are typically strongest when weight is applied along their length. Fractures occur when too much pressure is loaded onto any one of the axes, or a combination of axes. Pathological fractures are caused by diseases such as osteoporosis[2] and Paget's[3]; bones are weaker and more likely to suffer from a fracture due to the inherent bone structure being degraded. The majority of fractures are not a result of a pathological disease; they are caused by either direct or indirect forces that result in the fracture being classed as one of three types (Wraighte and Scammell, 2006).

Transverse fractures (figure 1.3.2a) are a result of a direct force or bending and are considered to be a simple fracture as they are typically stable, which favours union of the broken bone. They may require some kind of external fixation such as a plaster cast to ensure that there is minimal movement. Oblique and spiral fractures (figures 1.3.2b and c respectively) are usually caused by compression/loading and torsional forces; these types of fractures are inherently less stable than transverse fractures and can suffer from bone shortening and/or displacement. All three types of fractures can become more complex when multiple bone fragments are present; these fractures are referred to as comminuted[4] fractures.

---

[1] The prefix telo- means 'at the end'.
[2] A metabolic bone disease where the structure of bone becomes porous (osteoporosis literally means 'porous bones'), it typically affects those over the age of 45.
[3] Paget's is a metabolic bone disease where bone cells begin to form bone in an uncontrolled and irregular pattern, causing poor bone structure, it typically affects those over the age of 45.
[4] Many broken, splintered or crushed bone fragments.

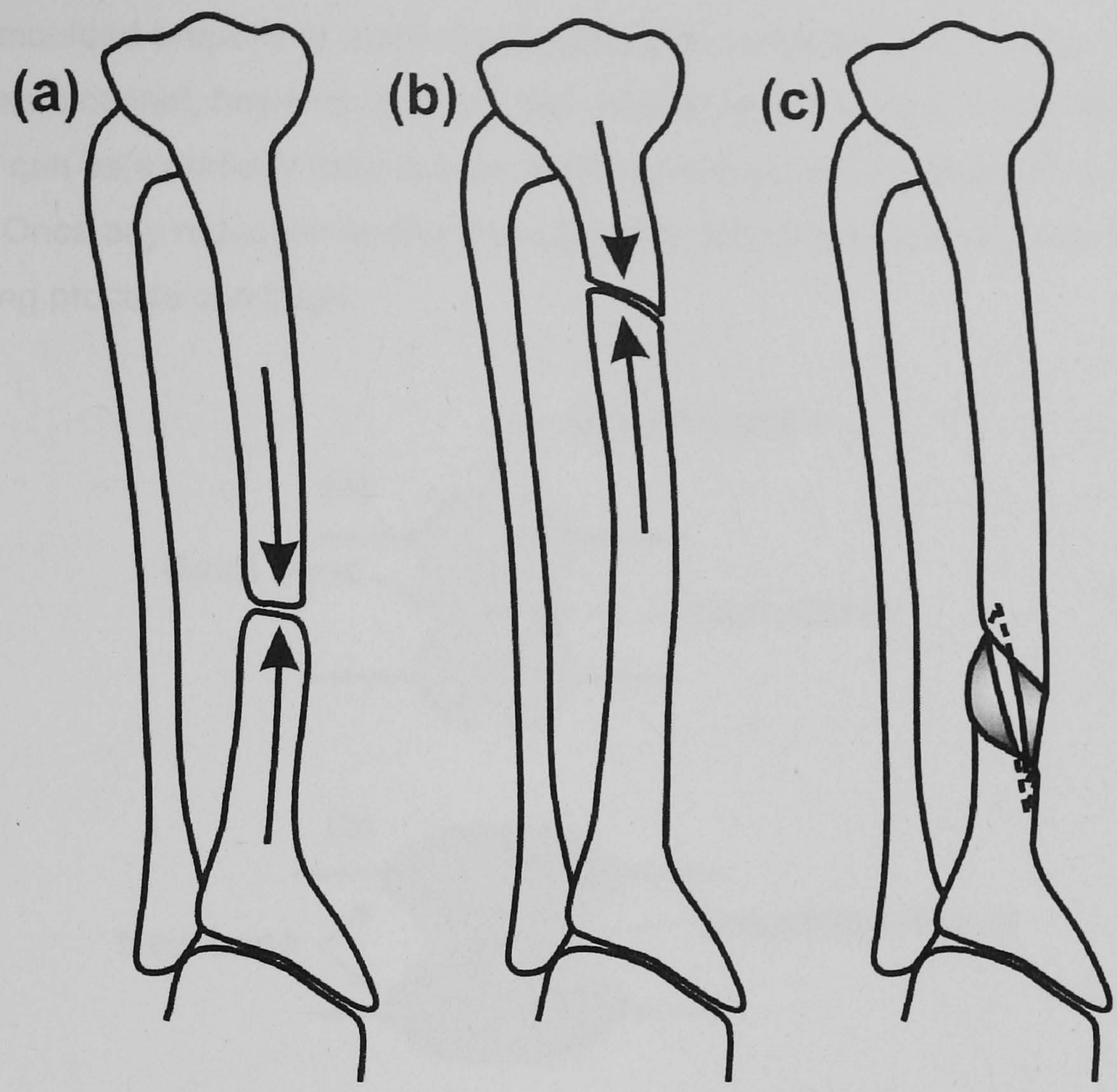


Figure 1.3.2. Diagram to illustrate the three types of non-pathological fractures: (a) transverse fracture. (b) oblique fracture, (c) spiral fracture.

1.3.3. Fracture healing

Bone has the amazing ability to successfully repair itself and heal without leaving any visible signs of scarring. There are two types of fracture healing, primary and secondary. Primary fracture healing can take place after reduction and rigid internal fixation that leaves the fracture surfaces unable to move. However, there has to be a direct bony union. More commonly, secondary fracture healing occurs. Secondary fracture healing affords some controlled movement, as there is relative stability; the fracture is usually cast in a plaster cast or is externally fixed.

Fracture treatment in the UK follows the RIP (reduction, immobilisation and physiotherapy) model (Wraighte and Scammell, 2006). Reduction is only undertaken if there is any significant displacement of the fracture and is typically used when oblique or spiral fractures occur. Manipulative reduction involves the reversal of direction that initially caused the fracture in an attempt to re-align the fractured bones. If a fracture is considered to be stable then a plaster cast can be used but

must be moulded properly to avoid any subsequent re-displacement of the bones. Plaster casts cannot, however, prevent any shortening of the bone. Less stable fractures can be externally fixed but are often accompanied by internal fixation devices. Once any reduction and/or immobilisation technique has been carried out, the healing process can begin.

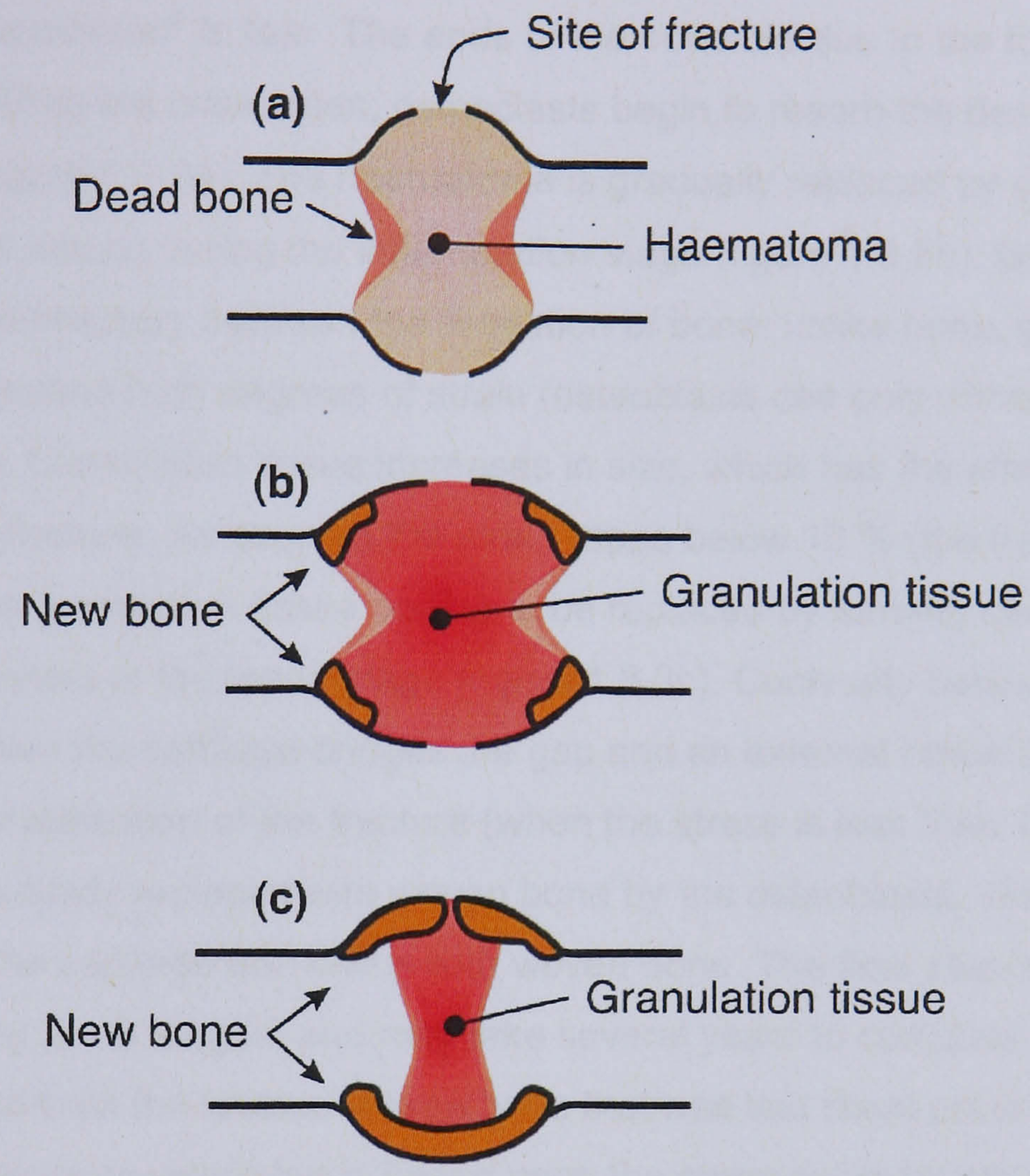


Figure 1.3.3. Diagram to represent the secondary fracture healing process: (a) formation of a haematoma at the fracture site, (b) inflammation stage, (c) final repair stage.

Primary healing requires direct bony union and absolute stability; this is only achieved through reduction and rigid internal fixation, usually with some compression. As there is no motion, there is very little (if any) strain present. The gap between the bone ends is less than 200 μm. This allows osteoclasts to "*…tunnel across the fracture line…*" (Wraighte and Scammell, 2006). Osteoblasts can follow, laying down the bone matrix to establish continuity. However, primary healing is a slow process as it is basically an extension of normal bone turnover. The bones must remain internally fixed until the healing process has finished and the remodelling is complete.

Secondary healing is the most common form of fracture healing and occurs when there is controlled movement of the fracture; the bone is relatively stable and is usually supported by an external fixation device such as a plaster cast. Secondary healing follows four phases that overlap.

The first phase involves the formation of a haematoma[1] at the site of the fracture, causing the periosteum[2] to tear. The ends of the bone die due to the formation of the haematoma, killing the osteocytes; osteoclasts begin to resorb the dead bone at the fracture site (figure 1.3.3a). The haematoma is gradually replaced by granulation tissue (fracture callus) during the inflammation stage (figure 1.3.3b). Granulation tissue is an intermediary between the formation of bone; unlike bone, granulation tissue can withstand high degrees of strain (osteoblasts can only withstand strains of less than 1 %). Granulation tissue increases in size, which has the effect of stabilising the fracture. As soon as the strain drops below 10 % (the fracture begins to stabilise) the granulation tissue begins to be replaced by forming cartilage, signalling the onset of the repair stage (figure 1.3.3c). Continuity between the bones is achieved when the cartilage bridges the gap and an external callus is formed. Upon further stabilisation of the fracture (when the stress is less than 1 %), the cartilage is gradually replaced with woven bone by the osteoblasts, which go on to replace all of the cartilage and callus with woven bone. The final step in secondary fracture healing is the longest and may take several years to complete. Lamellar bone is created from the weaker woven bone that was laid down previously. The bone shape begins to return but is based upon the stresses under which the bone has been placed during the whole repair process. Eventually, when the osteoblasts and osteoclasts begin to once more work in unison, the bone has healed and typically bears no scar tissue or evidence that a fracture has even occurred.

1.3.4. Factors that affect healing

There are many factors that can affect the rate at which a fracture heals, and also whether a fracture heals at all. Many fractures take longer than expected to heal, termed delayed union, whilst others fail to heal at all and are classed as non-union. Age has a large effect upon the rate of healing as young people, especially children, tend to heal at a much faster rate than older. Over the age of 40 to 50 there is an

---

[1] A collection of blood as the result of haemorrhage/internal bleeding.
[2] A layer of irregular tissue membrane that covers the outer surface of bone (not at joints).

increased risk of fractures due to pathological disease. These fractures are likely to take much longer to heal than the non-pathological fractures that are common to the younger age groups. Other factors such as gender may play a part, but lifestyle differences may account for the delayed or non-union of fractures. People who smoke, drink and eat excessively or poorly are at a higher risk of fractures that fail to heal successfully (Wraighte and Scammell, 2006).

1.4. Metabolism

Metabolism is the word used to describe many biological processes that are the means by which the body can synthesise many compounds and also generate energy. The synthesis of compounds is termed anabolism, and requires energy; this energy is generated by catabolism, where large molecules are broken down to generate smaller molecules, releasing energy in the process. Hormones typically control both anabolism and catabolism. The process of metabolism (both anabolic and catabolic) can be summarised as primary or secondary metabolism depending upon the function: primary metabolism (or basic metabolism) refers to any metabolic processes that are necessary to keep a cell alive (e.g. energy production, biosynthesis of molecules), whereas secondary metabolism relates to compounds that are produced/broken down, but are not important for the survival of the cell itself. However, secondary metabolism is important for an organism as a whole, as secondary metabolites are produced to enhance survival (their absence may not immediately result in death, but prolonged absence would). Molecules are typically synthesised from fats, carbohydrates and proteins (obtained from food) by enzymes. Many metabolic pathways interlink, and need to be able to detect the function or status of other pathways for efficient metabolism.

One important factor that is gaining interest in the field of metabolism (and metabonomics/metabolomics) is gut microflora (Nicholson, J *et al.*, 2002; Nicholson, Jeremy K. and Wilson, 2003; Nicholson, J *et al.*, 2004; Nicholson, J *et al.*, 2005; Gill *et al.*, 2006; Bertram *et al.*, 2007; Goodacre, 2007; Rezzi *et al.*, 2007a; Rezzi *et al.*, 2007b). Nicholson *et al.,* and Gill *et al.,* report that the human intestinal tract is home to some 10 to 100 trillion (10^{14}) microbes, which are essential for function (Nicholson, J *et al.*, 2005; Gill *et al.*, 2006). Given that the human base pair genome stands at 2.85 billion, the estimated ≥100 times greater number of genes contained within the human gut dwarfs this value. Gut microflora (microbiome) have been shown to enhance the metabolism of amino acids, glycans and xenobiotics, as well as the synthesis of vitamins (Gill *et al.*, 2006).

Whilst much more research is required to understand the effects of age, diet and pathological status upon the gut microbiome, Bertram *et al.* have shown that diet strongly affects the gut microbiome (Bertram *et al.*, 2007). Rezzi *et al.* have observed that the gut microbiome, once considered to be relatively stable, are more related to diet than once previously thought (Rezzi *et al.*, 2007a; Rezzi *et al.*, 2007b).

As the gut microbiome also varies upon demographic profile (personal conference notes), it is clear that metabolites are very varied, and largely influenced by many factors.

1.4.1. Renal handling and urinary excretion

The kidney is a key organ in the regulation of metabolic products, as it is the route by which they can be excreted. Kidneys regulate the composition of the blood by maintaining the appropriate composition of many ions, removing waste compounds such as urea, ammonia and xenobiotics, and regulating the pH. Compounds can be modified to either make them more soluble (e.g. sulfonation, alkylation, acetylation or glucuronidation where a compound is bound to glucuronic acid via a glycosidic bond), or to contain a functional marker group to allow excretion rather than retention.

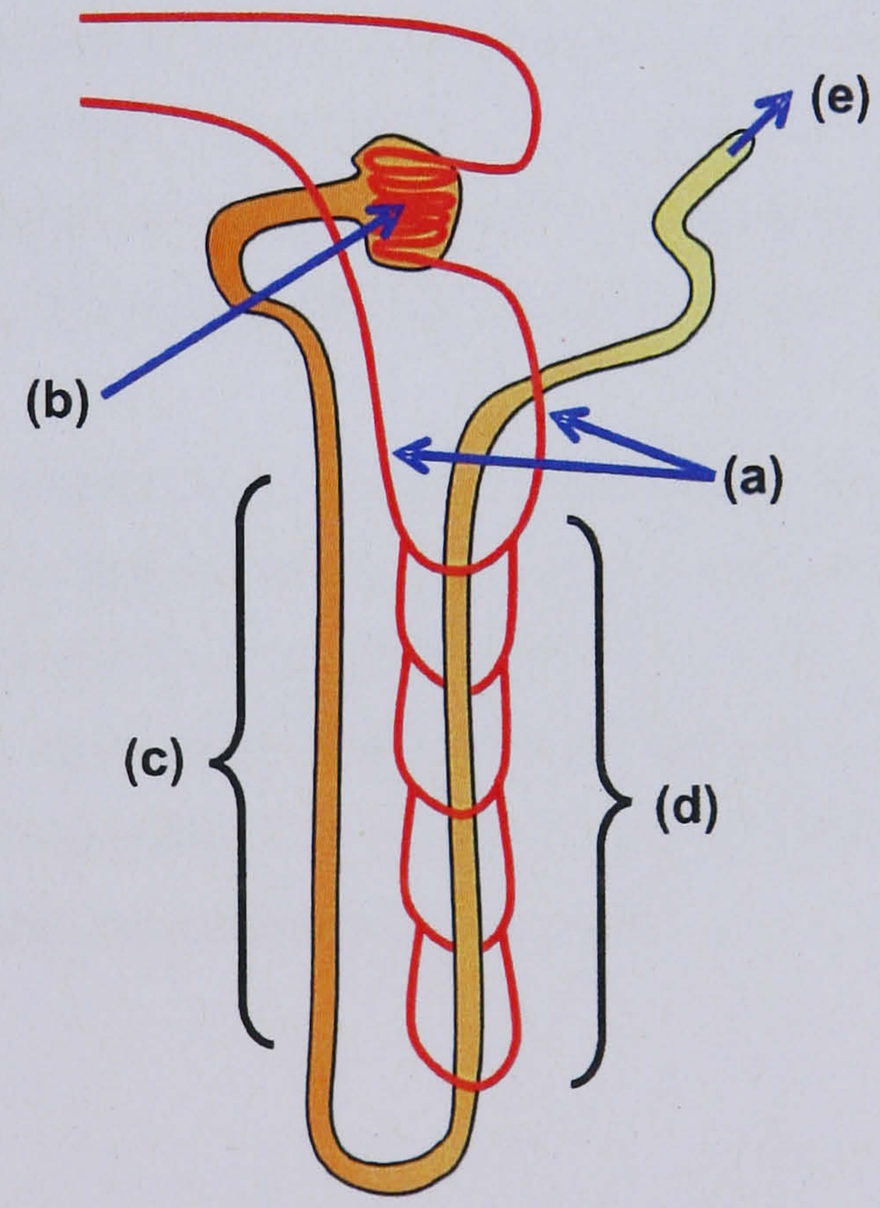


Figure 1.4.1. A schematic of a nephron from a kidney: (a) capillaries, (b) glomerulus and glomerular capillaries, (c) descending loop of Henle, (d) ascending loop of Henle, (e) distal tubule to the bladder.

In the nephron (figure 1.4.1), blood received from the heart gets filtered under pressure in the glomerulus, creating a filtrate. The glomerulus acts as a physical and

electrical charge barrier to most plasma proteins, although some do pass into the proximal tubule (figure 1.4.1c and d) dependent upon their size and concentration in plasma (Christian and Watson, 2004). The filtrate consists of water, ions such as sodium, potassium and chloride, glucose, amino acids and small proteins (less than 40 kDa (Gonzalez-Buitrago *et al.*, 2007)). The loop of Henle consists of transporters, which are selective for specific molecules. The descending loop of Henle has transporters that are specific for glucose and sodium, creating a concentration gradient that causes water to be passively reabsorbed in the ascending loop of Henle by osmosis; low molecular weight proteins can be degraded in the proximal tubule (figure 1.4.1c and d), either being reabsorbed or passed to the bladder. Other molecules can be actively or passively absorbed; the majority of glucose, amino acids and inorganic salts are reabsorbed, leaving a filtrate that consists of any excess molecules and waste products or compounds. The final filtrate passes through the distal tubule (figure 1.4.1e) to the bladder to form urine.

As urine is one of the body's waste excretion methods, urine should contain a wealth of metabolites that are related to many processes within the body, meaning that the study of urine should enable a picture of what has gone on in the body to be explored; Goodacre describes the study of such biofluids as *"...the way forward..."* (Goodacre, 2007), although this is nothing new as metabolites in urine have been used to detect diabetes, pregnancy and liver function to name but a few.

Given that the bone remodelling process causes secretion of compounds related to this process into the bloodstream, serum has been analysed for biomarkers of the remodelling process (Chapurlat *et al.*, 2000; Ebeling and Åkesson, 2001; Igarashi and Yamaguchi, 2002; Henle *et al.*, 2005; Zimmermann *et al.*, 2005). As the kidneys filter blood around 25 times a day, urine might be expected to contain excreted biomarkers relating to the bone remodelling process.

1.5. Metabonomics

The field of systems biology has given rise to an impressive array of 'omes' and 'omics', such as the genome, transcriptome, proteome and metabolome along with their studies giving genomics, transcriptomics, proteomics and metabolomics/metabonomics. The ultimate goal of these technologies has to be the complete understanding of a biological system from the genome right down to the metabolome, and also how each of these different 'omes' relates to one another given the bigger picture; this is clearly a long way from being reality, but with continual developments in both technology and its application, it may one day become feasible.

Within the scientific community there has been much debate with respect to what to name the field related to the metabolome, or indeed the metabonome. 'Metabolomics' versus 'metabonomics' may at first appear to merely be a case of semantics, but there are significant differences that emerge between the way the two terms are used in the literature, even though some prefer to use the terms interchangeably.

The term 'metabonomics' was coined by J. K. Nicholson *et al.* in 1999 (Nicholson *et al.*, 1999), and was defined as "*...the quantitative measurement of the multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification...*" with metabonomics being derived from the Greek "*meta*" and "*nomos*" meaning changes and rules/laws respectively (Lindon *et al.*, 2004). 'Metabolomics' first appeared in 2000 and was coined by O. Fiehn *et al.*, (Fiehn *et al.*, 2000) and was defined as "*...the quantitative measurement of all low molecular mass metabolites in an organism's cells at a specific time under specific environmental conditions...*". The main noticeable difference is that metabonomics is concerned with dynamic changes within a system that is a response to some sort of stimulus, whereas metabolomics looks at a 'snap-shot of the cellular metabolome' (Tang and Wang, 2006). The two terms are sometimes differentiated between based upon the analytical method utilised; metabolomics initially utilised hyphenated mass spectrometry techniques, whereas metabonomics utilised NMR based techniques, however, this is clearly an out-dated way of differentiating between metabonomics and metabolomics as is shown in the papers by Wilson *et al.*, and Lindon *et al.*, (Lindon *et al.*, 2004; Wilson *et al.*, 2005). Other arguments amount to metabolomics (and the metabolome) relating to plant and microbial systems whereas

metabonomics (and the metabonome) relate to animal models (Tang and Wang, 2006). Whichever definitions people use to decide if their study is metabolomics or metabonomics based, it is academic and arguments only detract from the science that is being carried out. Studies in the field of metabolomics and metabonomics are increasing steadily as is shown by a PubMed search[1] (figure 1.5.1). Metabolomics appears to be the favoured term used but may be caused by the fact that many people (incorrectly) choose to use the terms interchangeably; the field in general is still lagging behind that of proteomics which is still receiving much attention, although it may be beginning to reach its maturity as the 'hype' dies down (2430 PubMed citations for 'Proteomics' in 2007[1]).

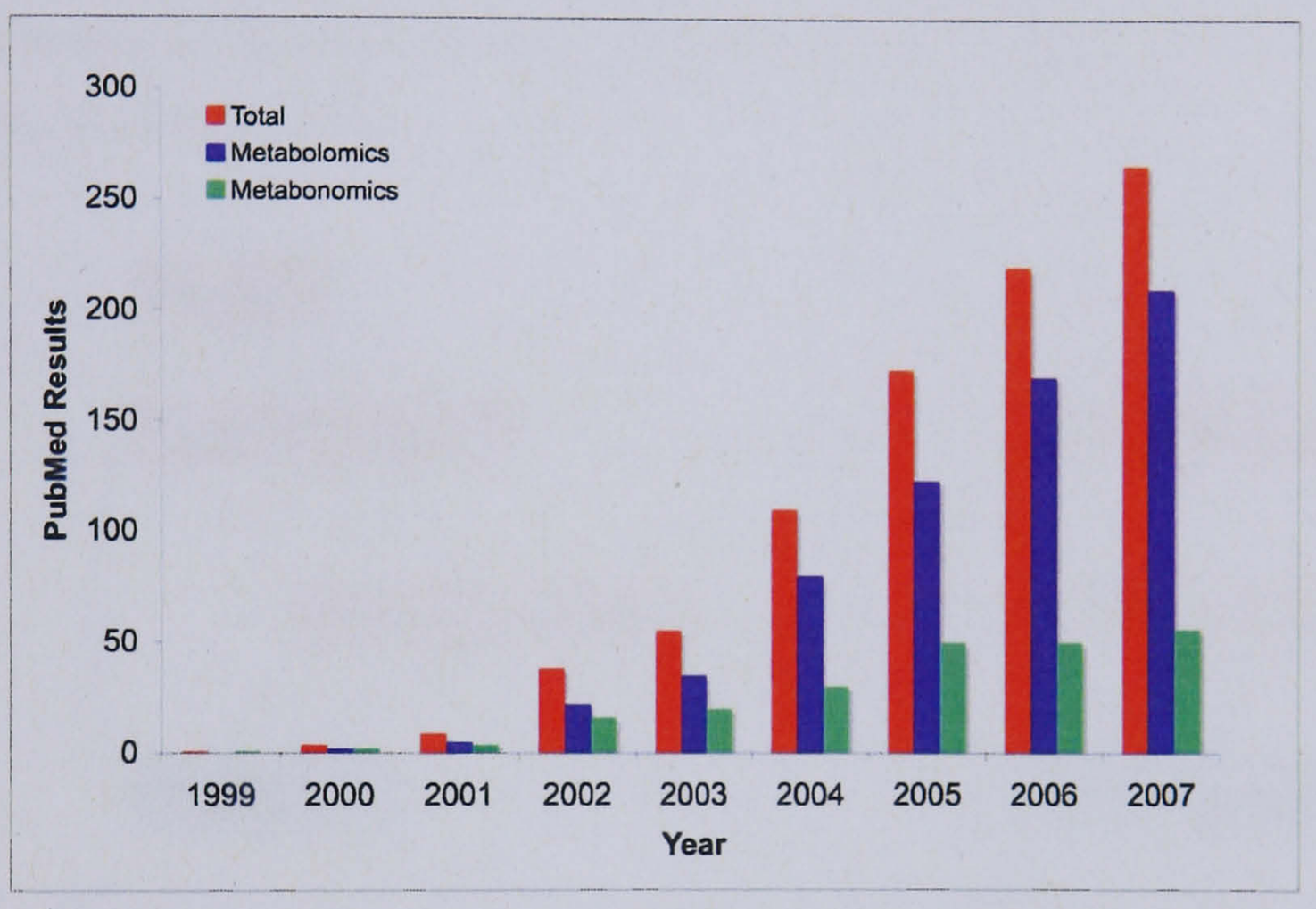


Figure 1.5.1. Graph to illustrate the increasing number of papers published citing 'metabolomics' or 'metabonomics'.

For the purpose of this study, it was chosen to use the term 'metabonomics', based upon the fact that that human clinical urine samples that were not from one given time, and the body's response to a stimulus, in this case bone fractures, were studied.

Metabonomics can provide a real biological endpoint[2]; other 'omic' techniques such as proteomics and transcriptomics are very useful, but cannot provide any biological endpoint markers that could be used to relate to a specific disease or the effects of drug metabolism. The genome, proteome and metabonome/metabolome are all

---

[1] http://www.pubmed.com accessed November 2007.
[2] Metabonomic analyses of biofluids such as urine can provide a snapshot of biological processes that *have* happened, whereas proteomic/transcriptomic studies can only predict what *may* happen.

related to one another (figure 1.5.2), and as such, environmental factors could affect each level either directly or indirectly through a knock-on effect. Through the process of homeostasis, the body automatically attempts to maintain a constant internal environment, even when disease, drugs or toxins affect concentrations and fluxes of endogenous metabolites. In order to maintain this constant internal environment, increased levels of endogenous metabolites and any exogenous metabolites are eliminated through the body's waste (urine and faeces). Many biofluids contain a wealth of information, but some such as blood may not be the best to use when real biological endpoints are sought as blood is a homeostatically controlled biofluid, the composition of which is therefore heavily regulated. There are other 'exotic biofluids' that can be sampled such as saliva, semen, bile etc., (Mukhopadhyay, 2006) but the biofluid that receives the most attention for good reason in metabolomic and metabonomic studies is urine.

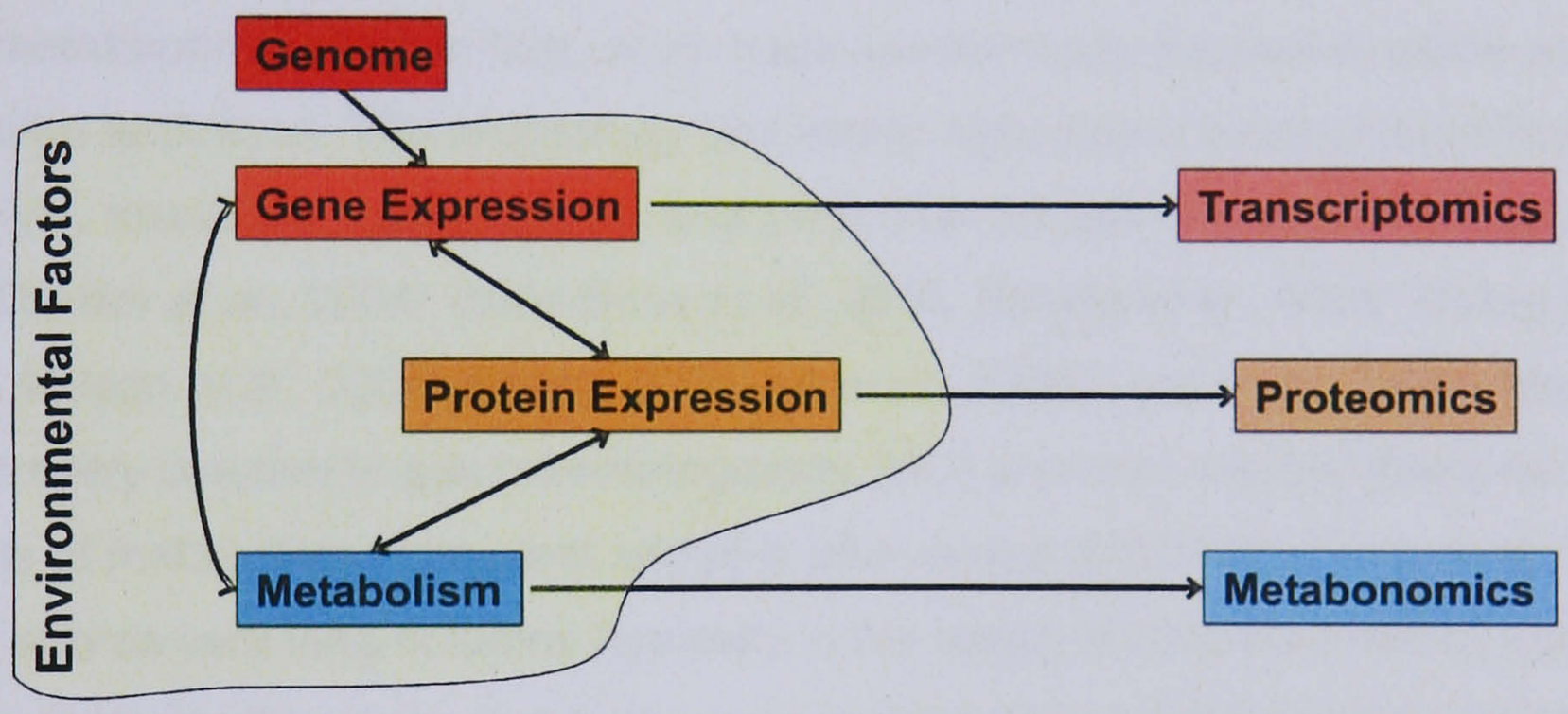


Figure 1.5.2. Diagram to represent the relationship between the genome, proteome and the metabolome, and how environmental factors affect each one.

Initial urinary metabonomic studies focussed upon the use of ^{1}H NMR as the analytical platform (Nicholson *et al.*, 1999; Lenz *et al.*, 2000; Robertson *et al.*, 2000). NMR requires very little sample preparation and is also a non-destructive technique, which can be useful if samples are precious. As urine is primarily water, special pulses have to be applied to the sample in order to suppress the overwhelming signal that would otherwise be present in the spectra, meaning increased analysis times. Even though NMR is a reasonably fast technique and shows fantastic reproducibility, when compared to MS techniques it shows significantly lower sensitivity. Given that there are estimates of 2,000 to 20,000 human metabolites[1] (compared to an

[1] The Human Metabolome Database Project has recently catalogued 2500 metabolites in the first draft of the human metabolome. (http://www.hmdb.ca, accessed November 2007).

estimated 30,000 genes and 1,000,000 proteins (Schmidt, 2004)) it would appear that profiling metabolites would be comparatively easy. However, there is a large dynamic range in concentrations of endogenous metabolites that range from femtomolar (and lower) to millimolar; coupled to the fact that the chemical complexity (lipids, sugars, amino acids, nucleotides, sulfates and many more atomic arrangements) and biological variation between samples can be large, this only serves to further complicate the analysis of biofluids. NMR can fail to detect many metabolites that are present due to '*...NMR-invisible moieties...*' (Williams *et al.*, 2005). It is not to say that NMR is a redundant technique though, as it can often provide information that other techniques cannot, as well as being a complementary technique to MS that can allow a more comprehensive picture of the components present within a biofluid, and the metabonomic question being posed to be revealed (Kenney and Shockcor, 2003; Williams *et al.*, 2005; Forshed *et al.*, 2006).

Many metabonomic studies now utilise mass spectrometry coupled to some sort of separation technique. The technology has vastly improved in terms of sensitivity, selectivity, resolution, reproducibility and price over the last decade (Plumb *et al.*, 2003; Drexler *et al.*, 2004; Villas-Boas *et al.*, 2004; Beattie *et al.*, 2005; Idborg *et al.*, 2005b; Wilson *et al.*, 2005; Kopka, 2006; Lu *et al.*, 2006; Lutz *et al.*, 2006). Mass spectrometry coupled to gas chromatography (GC) provided a sound basis for the analysis of metabolites from plant samples after derivatisation (Gullberg *et al.*, 2004) as GC affords very long columns (typically in the range of 30 to 60 meters) with very high separation efficiencies due to the large number of theoretical plates. However, complex sample matrices such as urine and blood do not contain large numbers of volatile metabolites and therefore require derivatisation. The process of derivatisation means that the chemical structure of metabolites is being altered; given that not all components are amenable to derivatisation means that a substantial amount of potentially important metabolic information may not be detected. Other separation techniques such as capillary electrophoresis (CE) afford high levels of separation efficiency as well as being able to cope with 'dirty' samples, reducing the amount of sample preparation needed[1] (Ullsten *et al.*, 2006; Monton and Soga, 2007). Coupling CE to ESI-MS has proven to be a problem; there are commercially available sources that allow the 'easy' coupling of CE to ESI-MS, but the technique suffers from poor reproducibility due to CE. The future may hold great promise for CE-MS with further development, but at present is not reliable enough for large scale, high throughput

[1] This was found to be contradictory to the results of E. Edwards from the JTO group (Edwards, 2007).

metabonomics experimentation. By far the most common separation technique used in MS metabonomics is reversed-phase liquid chromatography.

HPLC can separate a wide range of complex compounds; it is a liquid method and therefore does not require compounds to be volatile in order to be analysed. Compounds often require little or no preparation prior to their injection onto the column where they are separated based upon their partitioning behaviour between the liquid mobile phase, and the solid stationary phase. The performance of modern HPLC columns is far lower than that of GC columns, typical analytical column sizes of 4.6 × 100 mm can only resolve an upper limit of around 300 theoretical peaks (Sumner, 2006). This poor peak capacity can be increased by coupling columns together or by decreasing the particle size; this is at the expense of higher backpressures, which presents a technical challenge when designing HPLC systems. Recently, monolithic columns have begun to make their mark as they afford higher flow rates, meaning faster separations (Ishizuka *et al.*, 2002; Ikegami and Tanaka, 2004; Svec, 2004), they have also been applied to metabonomic studies (Tolstikov *et al.*, 2003; Wilson *et al.*, 2005). Correctly so, hydrophilic interaction chromatography (HILIC) is beginning to make an impact on metabonomic research (Tolstikov *et al.*, 2003; Idborg *et al.*, 2005a; Kind *et al.*, 2007), as it is selective towards polar analytes (which should be present in large amounts in aqueous biofluids). HPLC appears as though it may be superseded by ultra-performance liquid chromatography (UPLC, or small particle liquid chromatography as it should properly be called). UPLC utilises smaller particle and column sizes, typically sub 2 µm diameter particles, and a column length of 3 cm × 2 mm for traditional packed style columns (monolithic columns can also be used on UPLC systems). As the columns are much smaller than traditional HPLC columns, UPLC requires much greater backpressures of around 12,000 psi to be obtained (compared to ~ 6000 psi for traditional HPLC systems).

The result of this new technology is a vast increase in the column efficiency and the number of compounds detected. Wilson *et al.*, showed that traditional HPLC-MS allowed the detection of ~ 1,500 ions in 10 min, whereas capillary HPLC-MS could detect twice as many ions in the same time and UPLC-MS over 5,000 ions in only five minutes (Wilson *et al.*, 2005). Even with these developments in HPLC, we are still a long way from satisfying the complete picture of the metabonome as many metabolites may fail to be detected. A recent analytical development involving LC coupled to NMR and MS (LC-NMR-MS) is emerging and is a promising field for metabonomic studies (Burton *et al.*, 1997; Bajad *et al.*, 2003; Bollard *et al.*, 2005).

Whichever analytical platform is used, vast amounts of data are typically produced; once called multivariate data, megavariate data is now far too large to be analysed without the use of complex statistical methods. The overwhelming majority of metabonomic studies use a statistical method called principal components analysis (PCA). PCA is a descriptive technique and serves to show any trends or similarities inherent in the data. Another statistical method that is commonly employed is partial least squares – discriminant analysis (PLS-DA). This is a discriminative technique as it utilises *a priori* knowledge of group classification.

1.5.1. Sample collection and preparation

Biofluids are generally easy to obtain, with blood (invasive collection) and urine (non-invasive collection) the most widely analysed. A biofluid only provides a 'snapshot' of the metabolome at a specific time point, and as such is representative of the system under investigation at that time. To avoid subsequent changes (metabolic reactions, chemical modifications and microbial growth) the biofluid should ideally be flash frozen using liquid nitrogen and be stored at a suitable temperature (-20 °C, or better -80 °C). Studies (LeBeau *et al.*, 2001; Schneider *et al.*, 2002; Fura *et al.*, 2003) have shown that freezing can alter the levels of endogenous metabolites, as can repeated freeze/thaw cycles; even though freezing is detrimental to the stability of endogenous metabolites, the effects are minimal (compared to storage above freezing) and the concentrations of endogenous metabolites appears to stabilise after two weeks (Schneider *et al.*, 2002).

Samples may require preparation prior to analysis such as acid hydrolysis, derivatisation, dilution or centrifugation depending upon the analytical platform chosen. Analysis by GC requires analytes to be volatile and thermally stable. Some compounds therefore require derivatisation. Whilst the derivatisation of compounds in biofluids may allow analysis by GC and increase the detection of some compounds, the derivatisation process is not 100 % efficient; many compounds may not be completely derivatised, and some may not be derivatised at all. Non-volatile compounds that are not derivatised may not be detected using GC; therefore when analysing the metabolome of a biofluid, there are compounds that cannot be detected. GC is biased against non-volatile, high molecular weight compounds.

Most biofluids are analysed using HPLC-MS due to the ease of sample preparation and the possibility of online coupling of the HPLC to the MS using liquid introduction interfaces. Samples are typically extracted using solid phase extraction or liquid-liquid extraction when identifying specific metabolites. Typically, biofluids can be analysed directly (after only centrifugation or dilution) for unbiased analyses, avoiding the unwanted exclusion of metabolites when study of the whole metabolome is required.

1.5.2. Sample variation

Biological samples are subject to biological and analytical variation. Biofluids such as blood and urine exhibit a large diurnal variation (Ebeling and Åkesson, 2001; Wilson *et al.*, 2005); physiological factors such as state of health, age, diet, stress or diurnal cycles (Antti *et al.*, 2004) affect the composition and result in variation between samples. Analytical variation, whilst being accepted as less influential than biological variance, contributes to overall variation; sample storage, treatment, preparation and instrumental variation may all affect the data recorded.

Together, biological and analytical variance affects the ability to obtain reproducible data; controlling the variation in a sample is challenging. Analytical variation can be minimised by ensuring each sample is treated in as similar a manner as possible, but being analysed in a random order to spread any variation across the dataset. Biological variance is harder to control; inclusion/exclusion criteria can be used, as can regulating the diet and lifestyle, although this is easier said than done when using human volunteers. The inclusion of internal standards, or the pooling of samples, is another way of accounting for biological and analytical variation.

1.5.3. Standards and normalisation

The inclusion of an internal standard allows analytical variation to be accounted for, but has to be carefully considered. An ideal internal standard should not co-elute with other compounds in the biofluid being analysed, and not cause ion suppression in the ionisation process during LC-MS analysis.

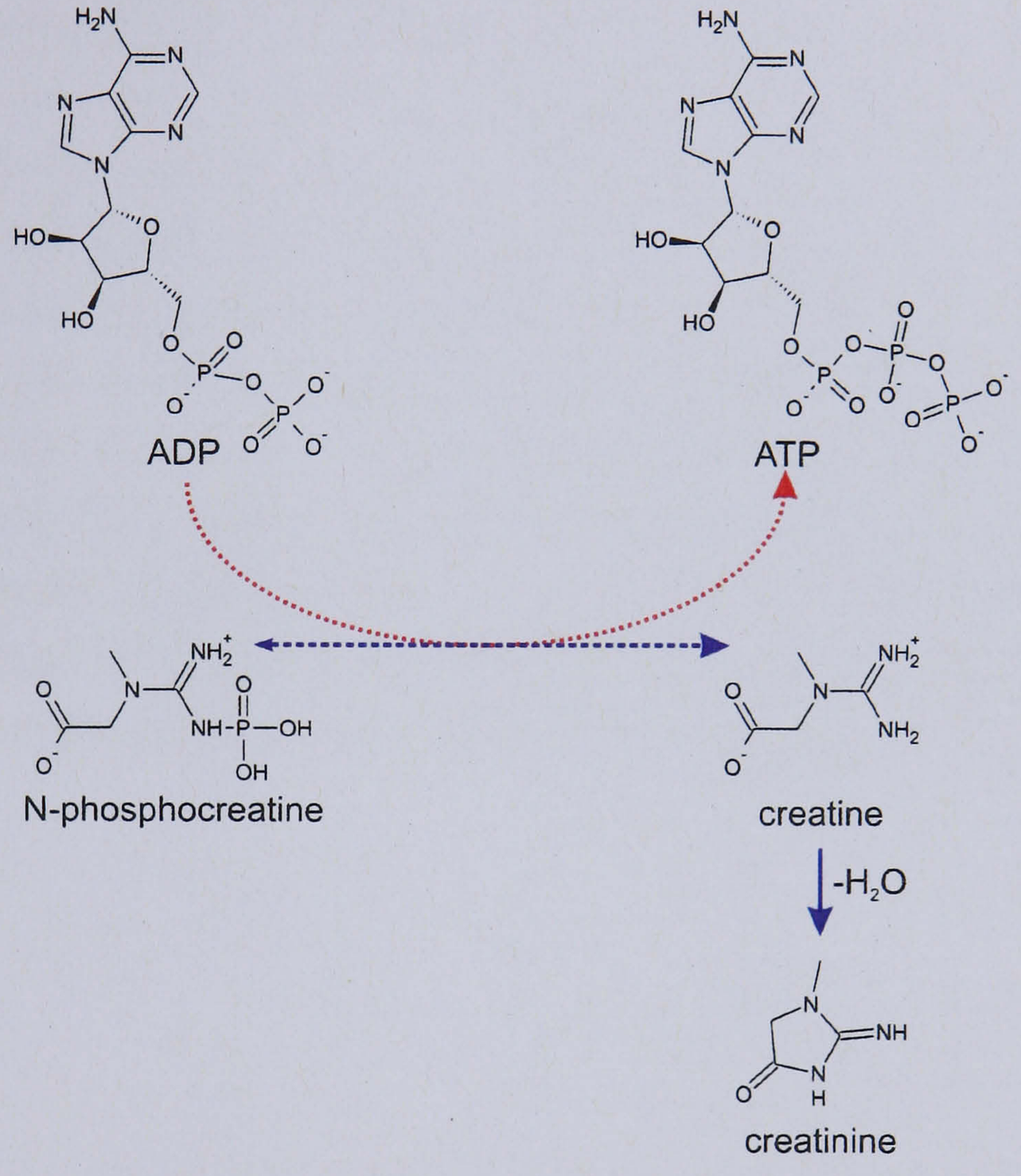


Figure 1.5.3. A schematic representing the breakdown of creatine to form creatinine. Adenosine diphosphate (ADP) is phosphorylated to adenosine triphosphate (ATP) by the removal of a phosphate group from N-phosphocreatine.

Biological variation is a much harder variable to control, especially in biofluids such as urine and serum. Creatinine is an endogenous metabolite that is excreted in urine; it is produced as a waste breakdown product during the synthesis of the body's energy source, adenosine triphosphate (ATP) (figure 1.5.3). The level of creatinine in an individual's urine is relatively stable and related to the individual's muscle mass. Creatinine has been utilised in many studies (Woitge *et al.*, 1999; Schneider *et al.*, 2002; Schoenau and Rauch, 2003; Felitsyn *et al.*, 2004; Husková *et al.*, 2004; Idborg *et al.*, 2004; Svoboda and Kasai, 2004; Obrant *et al.*, 2005) and is often used to express the concentrations of urinary metabolites (urinary metabolite : creatinine

ratio). When statistical analysis of data is undertaken, to account for concentration variance between samples, the data can be normalised to the intensity of the peak corresponding to creatinine in the MS chromatogram.

It is common in NMR metabonomic studies to normalise the data based upon the total signal intensity (Kenney and Shockcor, 2003; Antti *et al.*, 2004; Williams *et al.*, 2005). Using the total intensity normalisation method accounts for any variation in the concentration of individual urine samples, and removes the need for any internal standard to be included. More recently, LC-MS metabonomic studies have begun to use the total ion counts of individual samples to normalise the data to account for any concentration differences and analytical variation (Plumb *et al.*, 2005; Williams *et al.*, 2005). Many papers do not report how they normalise their data, but it appears that it is increasingly common to use the total ion count over creatinine (personal conference notes). Creatinine, whilst a good indicator of basal metabolism and individual muscle mass, may be perturbed by illness and other factors (Schneider *et al.*, 2002), meaning that utilising it for normalisation may not be as desirable as first thought. It is clear that the area of normalisation requires further in-depth study to determine the suitability of the different normalisation methods that are currently used.

1.5.4. Statistical analysis

Modern analytical techniques generate vast amounts of high-dimensional data (e.g. a 30 min LC-MS acquisition generates ~ 50 MB of data, so 200 samples would be 10 GB of data!), far more than can be dealt with manually. Thankfully, with the computing power now available, complex algorithms can be utilised to analyse these large, complex datasets. The field of chemometrics is defined as "*...the science of relating measurements made on a chemical system or process to the state of the system via the application of mathematical or statistical methods...*" (International Chemometrics Society). There are many different statistical methods that can be employed in order to analyse the megavariate data from metabonomic experiments. The most common tools are the descriptive principal components analysis (PCA) and the discriminative partial least squares – discriminant analysis (PLS-DA), soft independent modelling by class analogy (SIMCA) or more recently orthogonal partial least squares – discriminant analysis (OPLS-DA) (Nicholson *et al.*, 1999; Granger *et al.*, 2003; Idborg *et al.*, 2004; Wilson *et al.*, 2005). Whichever statistical approach is chosen, the user must be aware of the possible errors that can occur. There are typically two types of error that are considered, types I and II (or α and β errors respectively). A type I error is a 'false positive', this is when the hypothesis tested is rejected, although it is correct; a type II error, or a 'false negative', occurs when the hypothesis tested was not rejected when it was false.

1.5.4.1. Principal component analysis

Principal components analysis is an unsupervised technique; it does not require any prior knowledge, as it is a descriptive technique. PCA was first described in 1901 by Karl Pearson (Pearson, 1901) but was only able to be used for two to three variables due to the complex calculations involved; in 1933, Hotelling published practical computing methods, although these could not be realised until the advent of the modern computer (Hotelling, 1933). The early use of statistics could only cope with 'long and thin' data matrices, whereas today's data uses 'short and fat' data matrices as modern analytical techniques typically record many variables for few samples (or observations).

The function of PCA is to simplify, or reduce the dimensionality, of large amounts of data; data are broken down into two smaller tables, the scores and loadings. Scores

summarise the observations[1] and enable any patterns or trends inherent in the data to be visualised, whilst the loadings summarise the variables[2] and help to explain the position of the observations in the scores plot. PCA generates principal components (PCs) that are linear combinations of the original data. The first principal component accounts for the greatest amount of variation within the data; successive principal components account for the maximum variation possible that has not already been accounted for by the previous PC. Each observation is represented by a point in 'n – 1' dimensional space, where n = the number of variables (figure 1.5.4a). A line that accounts for the greatest amount of variation is fitted through the origin of the data (if the data have been mean centred) and is termed the first PC (figure 1.5.4b). The second principal component is orthogonal to the first, and accounts for the next largest amount of variation within the data (figure 1.5.4c). For a PC 1 vs. PC 2 plot, all of the points are projected onto a plane (figure 1.5.4c), this is termed the scores plot (figure 1.5.5a).

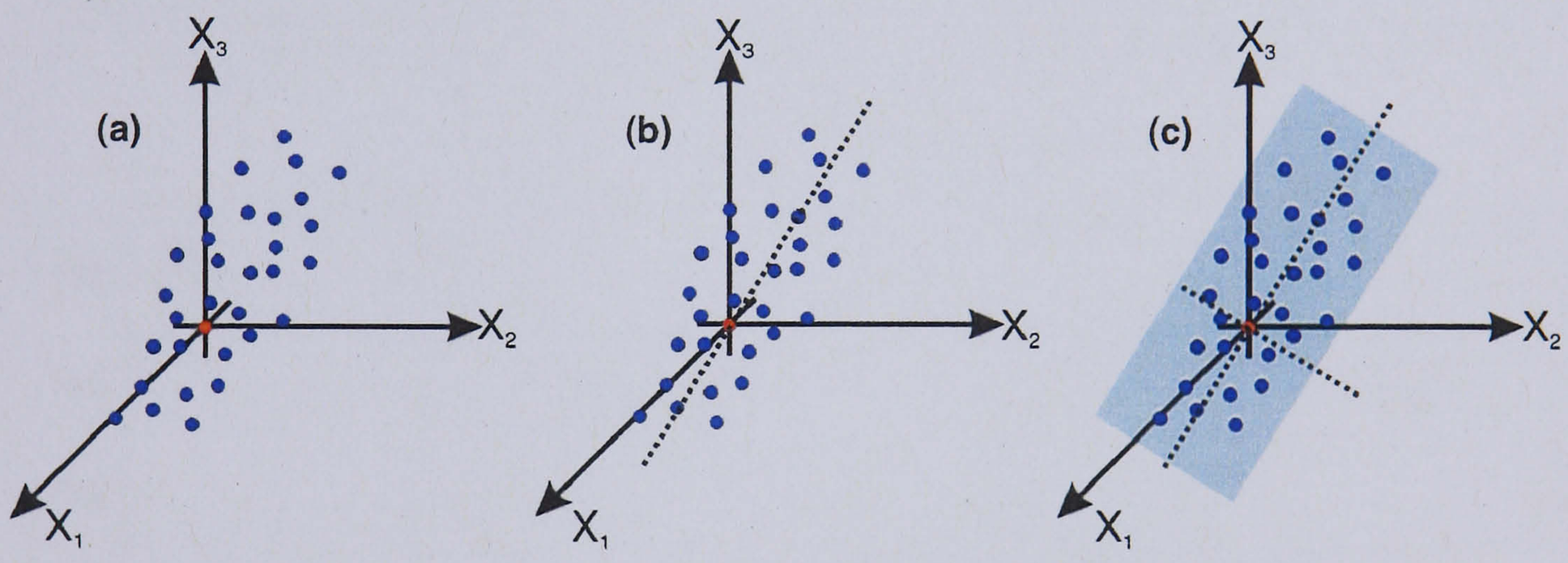


Figure 1.5.4. Schematic demonstrating how data are treated for PCA analysis, each blue dot represents an observation (an individual sample). (a) Each observation exists in n-1 dimensional space. (b) A line accounting for the greatest variation in the data is fitted, PC 1. (c) A subsequent line accounting for the next greatest variation is orthogonally placed.

Residual values based upon each point's distance in (n – 1) dimensional space from the plane (figure 1.5.4c) can help to identify outliers, as can the use of Hotelling's T^2 tool which is used to show a 95 % confidence limit within which data that fits the model well should appear, illustrated by the oval ring in the scores plot (figure 1.5.5a). Outliers can be easy to find, but can cause problems as the PCs are skewed

[1] 'Observations' relate to each sample analysed. For example: each urine sample analysed would be classed as an observation when their resulting data is analysed statistically.
[2] A 'variable' is a compound detected in one or more 'observation' (a variable would be an *m/z* value and a retention time, along with the relative intensity for each observation (or sample). This nomenclature shall be used whenever data are analysed statistically.

by the presence of an outlier and are therefore not a true representation of the data. However, to justify the removal of an outlier there should be sound scientific reasoning.

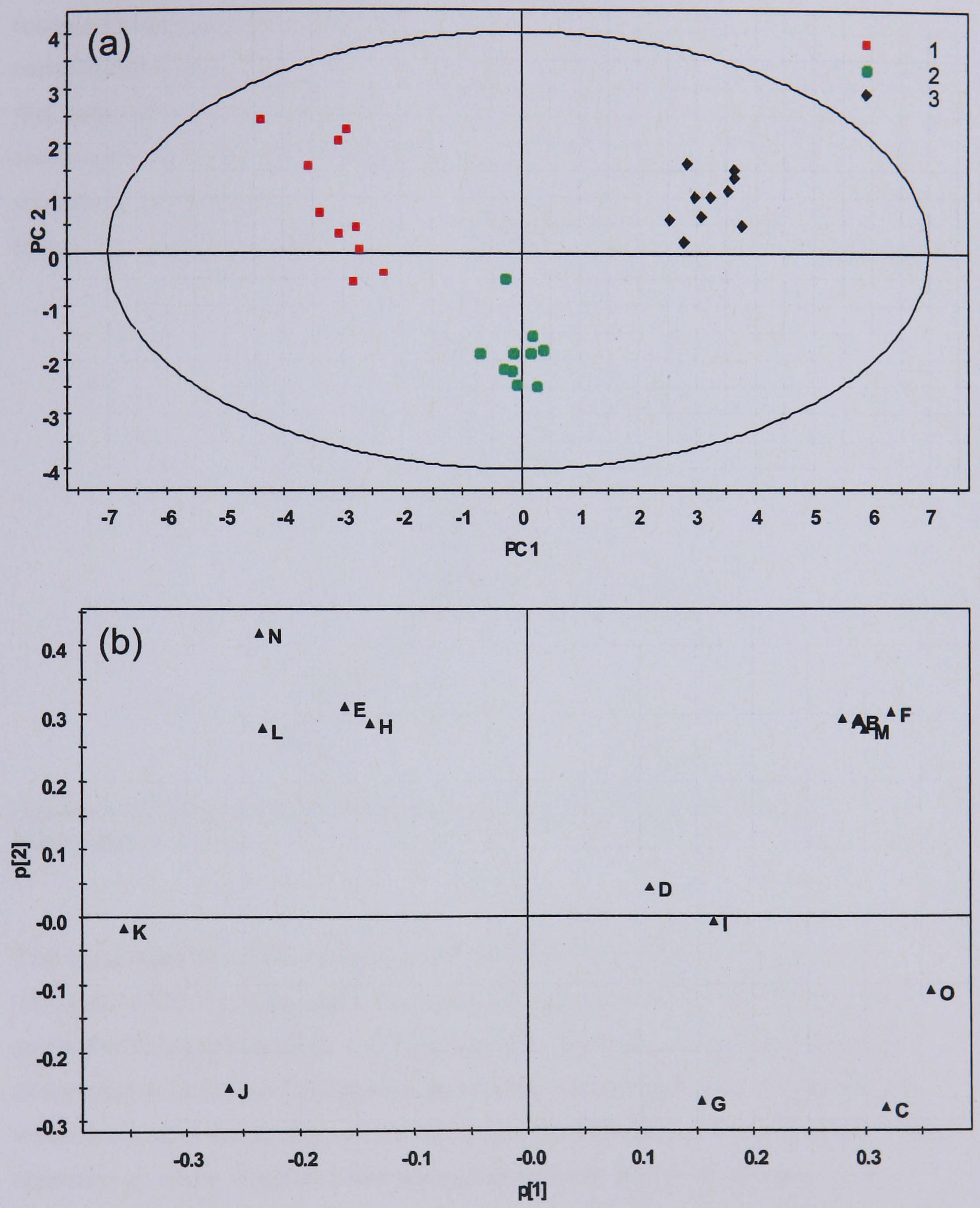


Figure 1.5.5. (a) Scores plot with three groups of data, clustering separately. (b) Loadings plot showing which factors are responsible for the clustering, the positions of which relate to the clustering observed in the corresponding PCA score plot (a).

Loadings determine how observations positions in the scores plot relate to the variables that were used. Loadings plots allow the user to interpret the scores plot, as the loadings relate to each variable. Loadings are calculated by relating the angles between each PC and the variable axis. The cosine of each angle is taken with respect to each axis and a loadings vector is generated (figure 1.5.6). If a particular variable has a strong influence then the cosine is close to 1 as the cosine of 0 is 1 (the angle of the component to the variable is close to 0 degrees); conversely, if a component has very little influence then the loading will be close to zero as the variable is nearly orthogonal (cosine of 90 degrees is 0). The loadings scale runs from -1 to +1 giving negative loadings, which correspond to a cosine of 180 degrees.

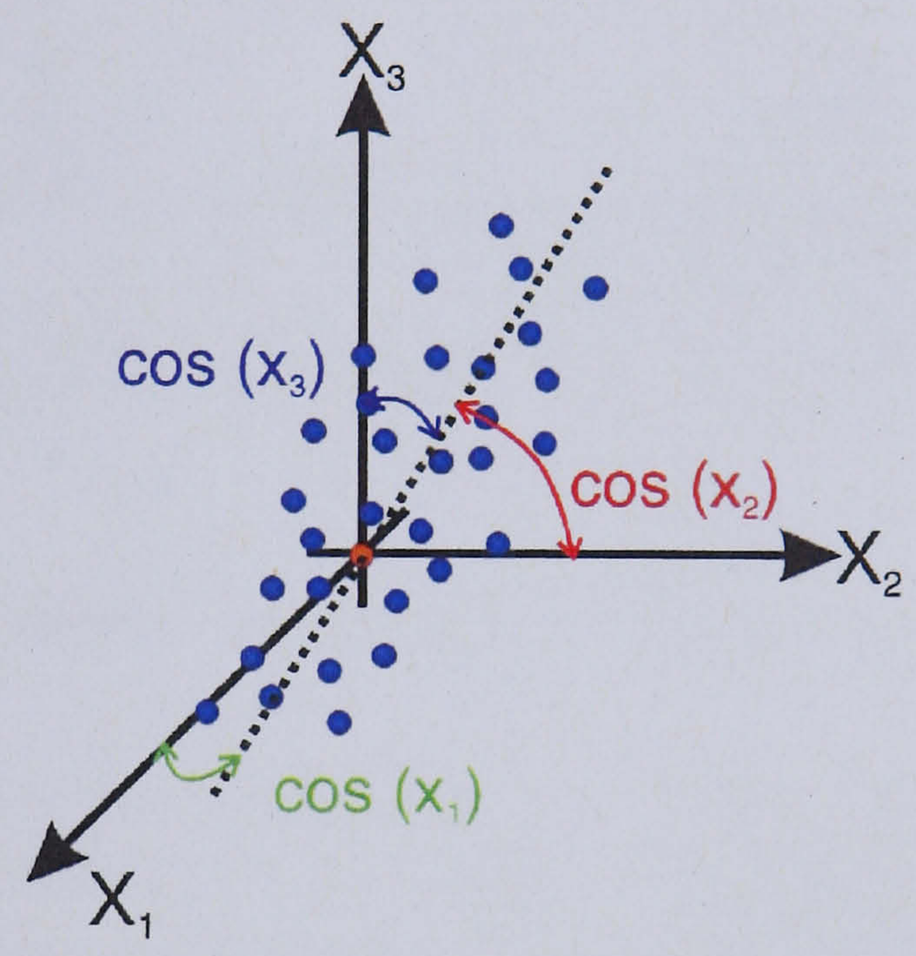


Figure 1.5.6. Diagram illustrating how loadings are determined from the observations.

The 'goodness' of a PCA model can be determined by utilising internal cross-validation (CV). It is also useful to ensure that the model is not being over-fitted, given that there are up to (n – 1) PCs available. Internal CV works by removing observations (and their corresponding variable values) and using the developed model to predict the position of the removed observations in the scores plot. This is repeated as many times as there are removed observations. If the new model component that is generated enhances the predictive power, then that component is retained. There is a trade off between the fit of a model and its predictive ability; the R^2 value is the 'goodness' of a fit, whilst the Q^2 value the 'goodness' of prediction. The optimal number of PCs can be determined by the relationship between the R^2 and Q^2 values. If, as shown in figure 1.5.7, the R^2 and Q^2 values are close then that

PC can be considered to exhibit a good fit. When the number of PCs increases, so does the complexity of the model. When the Q^2 value tails off, the model is beginning to become over-fitted and too complex. The ellipse in figure 1.5.7 highlights the optimal number of components; any further PCs would be over-fitted, as the Q^2 values decrease rapidly.

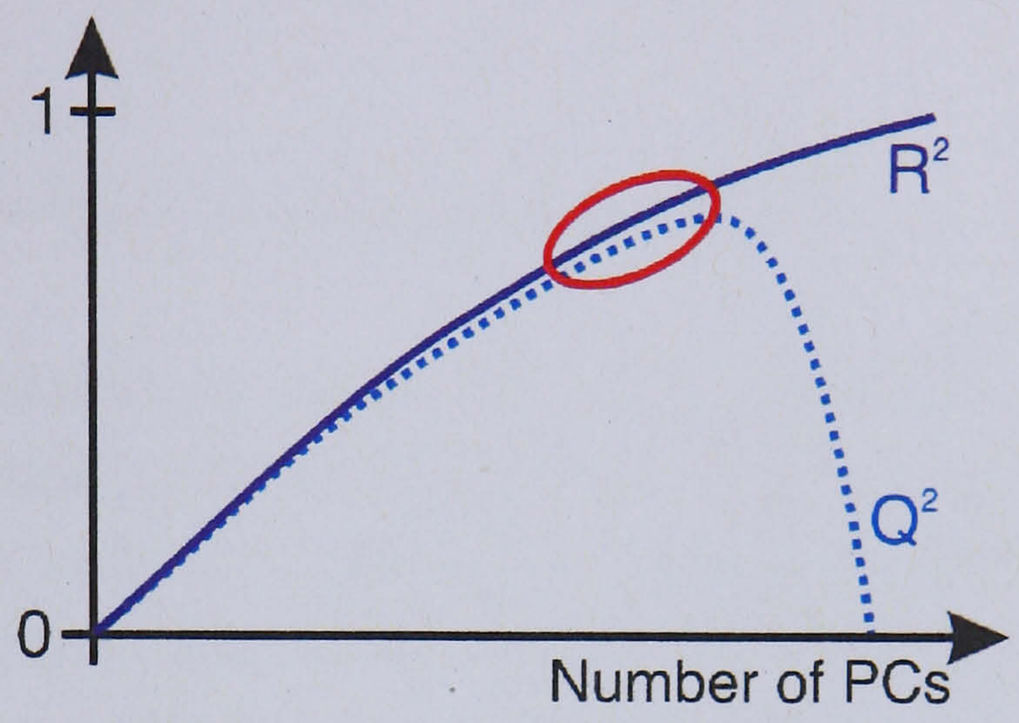


Figure 1.5.7. Diagram to show how the fit of a model and its predictive ability must be controlled with respect to the number of PCs used.

1.5.4.2. Partial least squares – discriminant analysis

Partial least squares – discriminant analysis (PLS-DA) is a regression analysis, and as such is a discriminative function which seeks to separate two or more groups based upon *a priori* knowledge (Wold *et al.*, 2001). As with PCA, PLS-DA represents each observation with a point in multi-dimensional space. PLS-DA seeks to obtain the greatest separation, which is obtained using the prior class knowledge. However, it can only be used to separate groups when there is a difference, which is why it is useful to analyse data with PCA first.

In contrast to PCA, where all of the data are extracted into a single X matrix (the explanatory variables), PLS-DA allows the comparison of two blocks of data, X and Y, where the Y matrix contains 'dependent' data, indicating class belonging. Latent variables (LVs, equivalent to PCs) are generated using both the X and Y matrices. PLS-DA seeks to maximise any inherent differences between the pre-defined classes in the Y matrix. Using this class knowledge, it is possible for PLS-DA to predict the class of one of the pre-defined classes in the Y matrix. Internal CV is once more essential in ensuring that any developed model is not being over-fitted.

Using PLS-DA to classify unknowns to groups requires that the model is trained on a representative dataset; the developed model should subsequently be tested (external CV) with data (an external test set) that were held back and not used to develop the discriminative model, and the developed model judged upon its classification success. The variables that are most influential and cause the separation of the groups can be found by using the loadings plots or by analysing the coefficients.

1.5.4.3. Soft independent modelling by class analogy

Soft independent modelling by class analogy (SIMCA) is useful when there are too many independent classes for PLS-DA (PLS-DA falls down when there are too many groups), but lacks any information on why the groups are different. SIMCA generates a localised PCA model for each defined group and is able to work with overlapping groups. Again, as with PLS-DA, there has to be a training set and a test set if a useful model is to be developed. The 'localised' PCA models use residuals to determine which observations belong to which group. It is these residual values that are used to predict classes for the new (and test) data.

1.6. High performance liquid chromatography

The development of high performance liquid chromatography dates back to 1906 when Mikhail Tswett used glass columns of 50-500 cm in length, and 1-5 cm in diameter, to separate chlorophyll solutions in carbon disulfide (Tswett, 1906). Modern HPLC is a technique used to separate compounds within a solution according to a specific property, such as hydrophobicity/hydrophilicity. The compounds within a solution are separated according to the basis of their partitioning behaviour between a stationary phase and a mobile phase. The principle of partitioning can be shown as the equilibrium of a compound (C) between the mobile and stationary phases:

$$\text{C (Mobile Phase)} \leftrightarrows \text{C (Stationary Phase)} \qquad \textbf{Equation 1.6.1}$$

The equilibrium shown above (equation 1.6.1) can be described by a partition coefficient, which is defined as:

$$k_D = \frac{[\text{C (Stationery Phase)}]}{[\text{C (Mobile Phase)}]} \qquad \textbf{Equation 1.6.2}$$

A solution containing a mixture of compounds to be separated is injected into a flow of liquid (mobile phase) that is pumped through a column containing a solid medium (stationary phase). When compounds come into contact with the stationary phase, they distribute between the stationary phase and mobile phase (equation 1.6.1). The mobile phase is continually passing thorough the column, causing compounds to distribute between the mobile and stationary phase many different times. Compounds with a higher k_D (equation 1.6.2) spend a larger amount of time in the stationary phase, therefore taking longer to elute from the column than compounds with lower k_D values; this is the basis upon which compounds are separated.

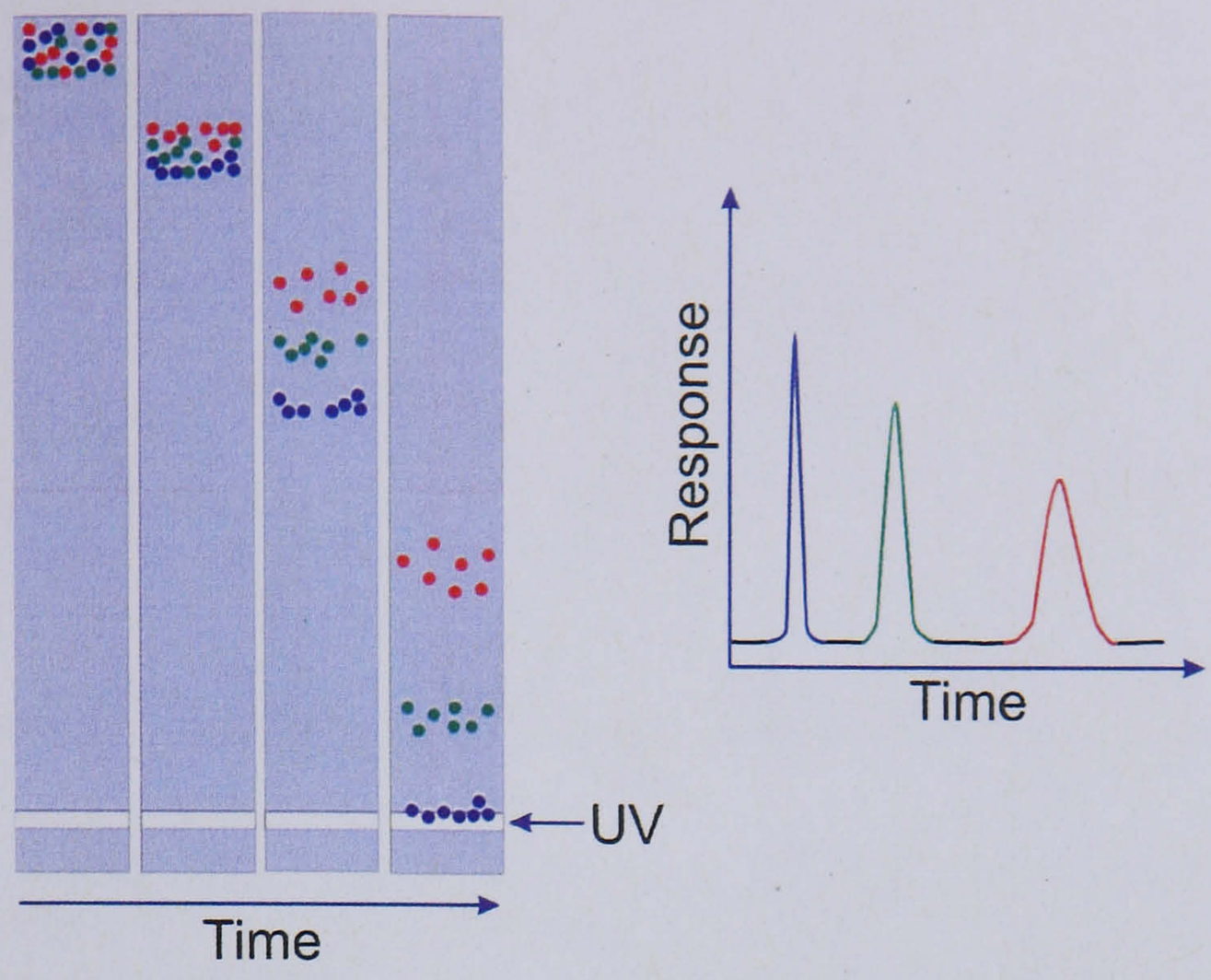


Figure 1.6.1. A schematic diagram illustrating how three different components migrate through a column and separate over time, as each different compound has a different partition coefficient. As each compound elutes from the column, they are detected using UV, with the area of each peak being proportional to the concentration.

Figure 1.6.1 shows the separation of three compounds over time. The compounds each migrate through the column at different rates, each having different partition coefficients (k_D), before being eluted from the column and detected (here using UV). The resulting separation of the analytes produces a chromatogram containing peaks with areas that are related to the concentration of each of the separated compounds.

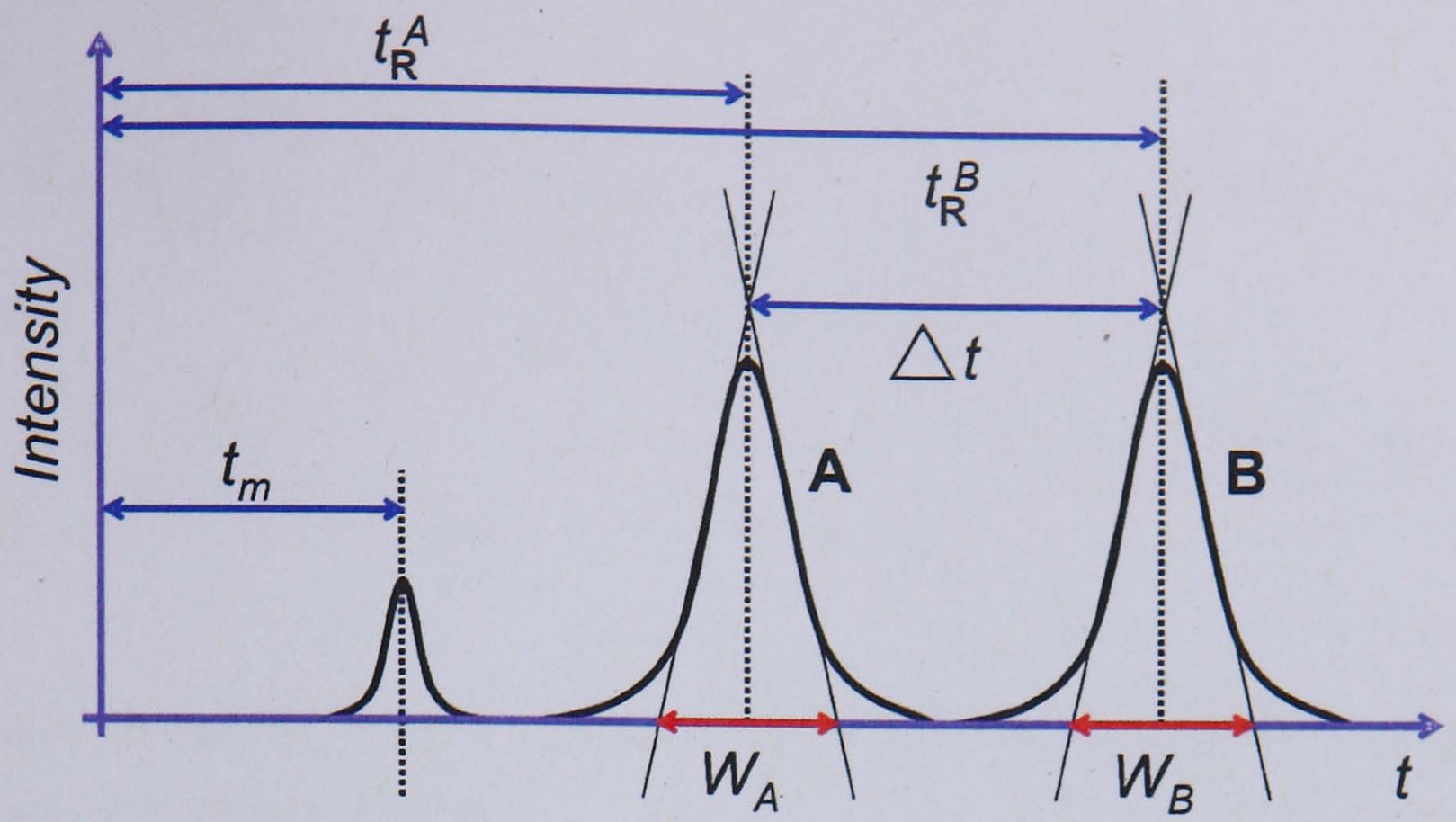


Figure 1.6.2. A schematic illustrating the chromatogram of a mixture of two compounds, A and B. t_m = unretained compound (dead volume); t_R^A = retention time of compound A, t_R^B = retention time of compound B; W_A = baseline width of peak A and W_B = baseline width of peak B, calculated by the intersection of an extension of their point of inflection (shown); $\Delta t = t_R^B - t_R^A$.

The capacity factor of a compound (k') can be used to describe the migration rate of a compound on a column, and is calculated according to equation 1.6.3 (using the terms shown in figure 1.6.2), which uses a compound's retention time (t_R) and also the time that the mobile phase takes to pass through the column (t_m).

$$k'_A = \frac{t_R{}^A - t_m}{t_m}$$ **Equation 1.6.3**

The capacity factor for two compounds can be used to describe the selectivity factor (α), a measure of peak separation:

$$\alpha = \frac{k'_A}{k'_B}$$ **Equation 1.6.4**

The efficiency and resolution that can be obtained using a column is of importance, and can be calculated using the peak widths, retention times and column length (figure 1.6.2). The efficiency of a column is measured by the number of theoretical plates, N, and is calculated using equation 1.6.5 (where HM = half maximum):

$$N = 16\left(\frac{t_R{}^A}{w_A}\right) = 5.545\left(\frac{t_R{}^A}{w_A(HM)}\right)$$ **Equation 1.6.5**

$$R = \frac{2\Delta t}{w_A + w_B}$$ **Equation 1.6.6**

Resolution, R, can be calculated by the ability of a column to separate two peaks (equation 1.6.6 and figure 1.6.2).

The ability of a column to separate two compounds is dependent upon any 'band broadening' that occurs during separation, and leads to a reduced efficiency as fewer compounds can be separated. Plate height, H, can be used to describe the performance of a column independent of the columns length:

$$H = \frac{L}{N}$$ **Equation 1.6.7**

A column can be considered to have a number of 'theoretical plates' where the separation of a compound can be described as a series of independent, consecutive equilibration events occurring between the stationary phase and mobile phase. Each plate is the distance required for an 'equilibration event' to occur; therefore the greater the number of plates, the more efficient the column. Plate height is dependent upon the linear flow rate, *u*, of the mobile phase, and can be explained using the van Deemter equation:

$$H = A + \frac{B}{u} + Cu$$ **Equation 1.6.8**

Where 'A' relates to eddy diffusion, 'B' to longitudinal diffusion and 'C' to mass transport; these terms are described in further detail below:

- Eddy diffusion (A) can be summarised as the different path lengths that compounds can take in a column. Whilst each molecule of a compound may start at the same position, they take different paths through the column as the mobile phase flows around the particles; this causes band broadening, but is independent of the velocity of the mobile phase.

- Longitudinal diffusion (B) creates band broadening, as molecules of a compound move from an area of high concentration (the centre of an analyte band) to an area of low concentration (the edges of the band). The extent of the diffusional band broadening depends upon the amount of time each compound spends on the column, thus band broadening is inversely proportional to the flow rate of the mobile phase.

- Mass transport (C) is the main factor that contributes to band broadening, particularly for silica particle based chromatography. During separation, molecules are partitioned between the stationary and mobile phase, depending upon their interaction with the stationary phase; this is much faster than the diffusion of a molecule into and out of a pore within a silica particle where there is no mobile phase flow. Due to the difference in partitioning and diffusion, band broadening occurs. The effects of mass transport are proportional to the velocity of the mobile phase, meaning that lower flow rates lead to better equilibration between partitioning and diffusion, leading to less band broadening.

Various aspects of an LC system can be altered to increase the resolution and efficiency. The mobile phase composition can be altered to change the capacity factors of compounds. Gradient mobile phases change the composition of solvents over time, enabling a decrease in the capacity factor of compounds that are strongly retained, thus decreasing analysis time. Reducing particle sizes increases the surface area of the stationary phase, which leads to an increase in resolution and separation efficiency; although this is at the expense of backpressure, as increased pressure is required to force the mobile phase through a more tightly packed stationary phase. To decrease eddy diffusion, the internal diameter of a column can be reduced, giving better peak shapes. However, less analyte can be loaded onto a column with a small internal diameter compared to columns with a larger internal diameter.

The most commonly used separation mode is reversed phase (RP) chromatography, where components are separated based upon their hydrophobicity with the most hydrophobic compounds eluting last. The majority of RP columns utilise derivatised silica particles, typically (although not exclusively) with n-octyl and octadecyl functional groups, where a predominantly aqueous based mobile phase is used.

1.6.1. Monolithic columns

An important development in RP column technology has been the introduction of monolithic columns (Svec and Frechet, 1992). The word monolith derives from the Latin *monolithus*, meaning single stone; the stationary phase in a monolithic column differs greatly from a traditional packed silica-based column as they consist of a single, porous piece of polymerised material as the stationary phase. The porous nature of a monolithic column means that the mobile phase can pass through with less hindrance than a traditional packed column; this leads to a substantial decrease in back pressure, meaning that higher flow rates can be used (and therefore faster separation achieved). As the mobile phase flows through a highly porous structure, the mass transfer process is no longer limited by diffusion into pores on silica particles, but now occurs primarily by convection.

The use of monolithic columns has allowed a reduction in analysis time and an increase in sensitivity and stability to be obtained for proteomic experiments (Premstaller *et al.*, 2001; Wienkoop *et al.*, 2004; Chen *et al.*, 2005; Rodrigues, 2005; Ault, 2007; Sumpton, 2007). However, despite the use of monolithic columns for proteomic studies, their uptake has been much slower for metabonomic studies where only a handful of studies have discussed or utilised this type of stationary phase (Pham-Tuan *et al.*, 2003; Tolstikov *et al.*, 2003; Dunn and Ellis, 2005; Wilson *et al.*, 2005; Cubbon *et al.*, 2007). Due to the success of monolithic columns for proteomic research within the JTO group (Robinson and MacDonell, 2004; Rodrigues, 2005; Ault, 2007; Sumpton, 2007), a monolithic C_{18} column was used for all RP separations presented within this thesis.

1.6.2. Hydrophilic interaction chromatography

The use of RP within the field of metabonomics is widespread, and may be caused by scientists using technologies that have been commonplace. However, for a field such as metabonomics, this could be a critical mistake given that many biofluids are predominantly aqueous and are thus likely to contain a wealth of polar content (that is poorly retained when using RP stationary phases, as discussed in chapter 3.2).

The lack of 'normal phase' separation use for metabonomic studies can be accounted for by its poor reproducibility and separation efficiencies (hence the popularity of RP stationary phases), but more importantly its incompatibility with MS due to the solvents typically employed (e.g. hexane). The void in polar analyte separation has been filled by the recent introduction of hydrophilic interaction liquid chromatography (HILIC). The principle of HILIC was first described in 1952 by Samuelson and Sjöström for the separation of monosaccharides using Amberlite IRA-400[1] as a stationary phase (Samuelson and Sjöström, 1952), but the acronym and the functional stationary phase used today was first suggested by Alpert *et al.* (Alpert, 1990).

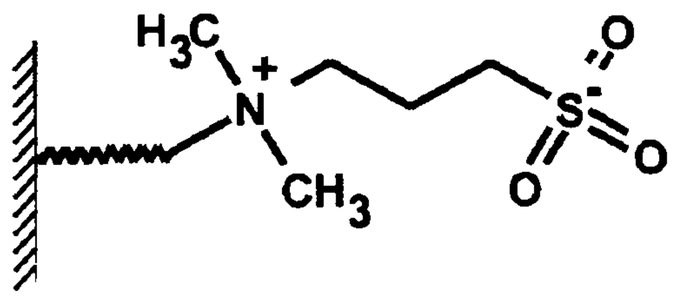


Figure 1.6.3. The zwitterionic bonded stationary phase for HILIC.

Figure 1.6.3 shows the bonded stationary phase for HILIC, a zwitterionic functional group with a net charge of zero; there are two functional ion exchange groups that are pH independent[2]. HILIC is thought to work by the hydrophilic partitioning of compounds into the water rich stationary phase, and also weak electrostatic interactions (Hemström and Irgum, 2006). HILIC is an orthogonal separation method to RP (Wang *et al.*, 2005), and as such the solvent strength is opposite to that of RP. The use of HILIC for metabonomic studies is not common at the present time, but a handful of studies (Idborg *et al.*, 2005; Kind *et al.*, 2007; Mawhinney *et al.*, 2007) have shown that this technique shows great promise, and was therefore the focus of research within this thesis (chapter three).

---

[1] Amberlite IRA400 is an anionic exchange resin.
[2] Counter ions come from buffer contained within the mobile phase. The stationary phase remains hydrated.

1.7. Mass spectrometry

1.7.1. History

Mass Spectrometry (MS) began with experiments in 1899 by J.J. Thompson (Thompson, 1899); by 1913 he had described how rays of positive electricity could be applied to chemical analyses. Coming up to its centenary, MS has seen exponential growth in both its use and development; MS is widely used to aid the identification and quantification of unknown substances and to probe their physical and chemical properties, all with a high degree of sensitivity and selectivity.

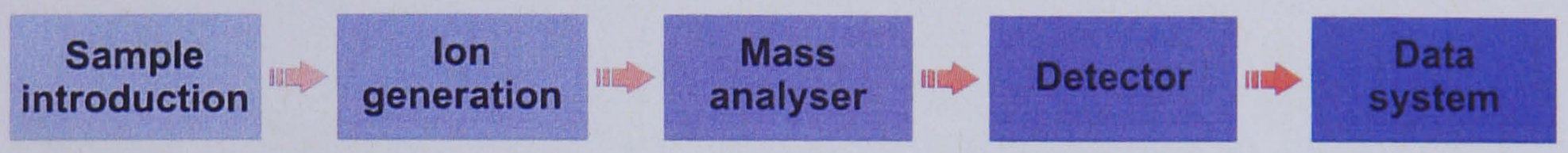


Figure 1.7.1. A schematic highlighting the components of an MS system.

The general principles of MS are shown in figure 1.7.1. Sample is introduced and transferred into the gas phase, creating positive or negatively charged gas phase ions. Ions are then separated according to their mass to charge ratio (*m/z*), and then detected, with ions' abundance being measured. The ionisation process can be at atmospheric pressure or under vacuum, while the mass analyser is under vacuum, as the gas phase ions require a mean free path so that they can traverse space. MS is a destructive analytical technique. However, utilising the latest techniques allows nanograms or less of an analyte to be used for analysis.

1.7.2. Ionisation methods

1.7.2.1. Electron ionisation

Electron impact was first used by Dempster *et al.* in 1921 to study the isotopes of lithium and magnesium (Dempster, 1921). This method was subsequently improved by Bleakney (Bleakney, 1929) and then Nier (Nier, 1947), and is now referred to as electron ionisation (EI).

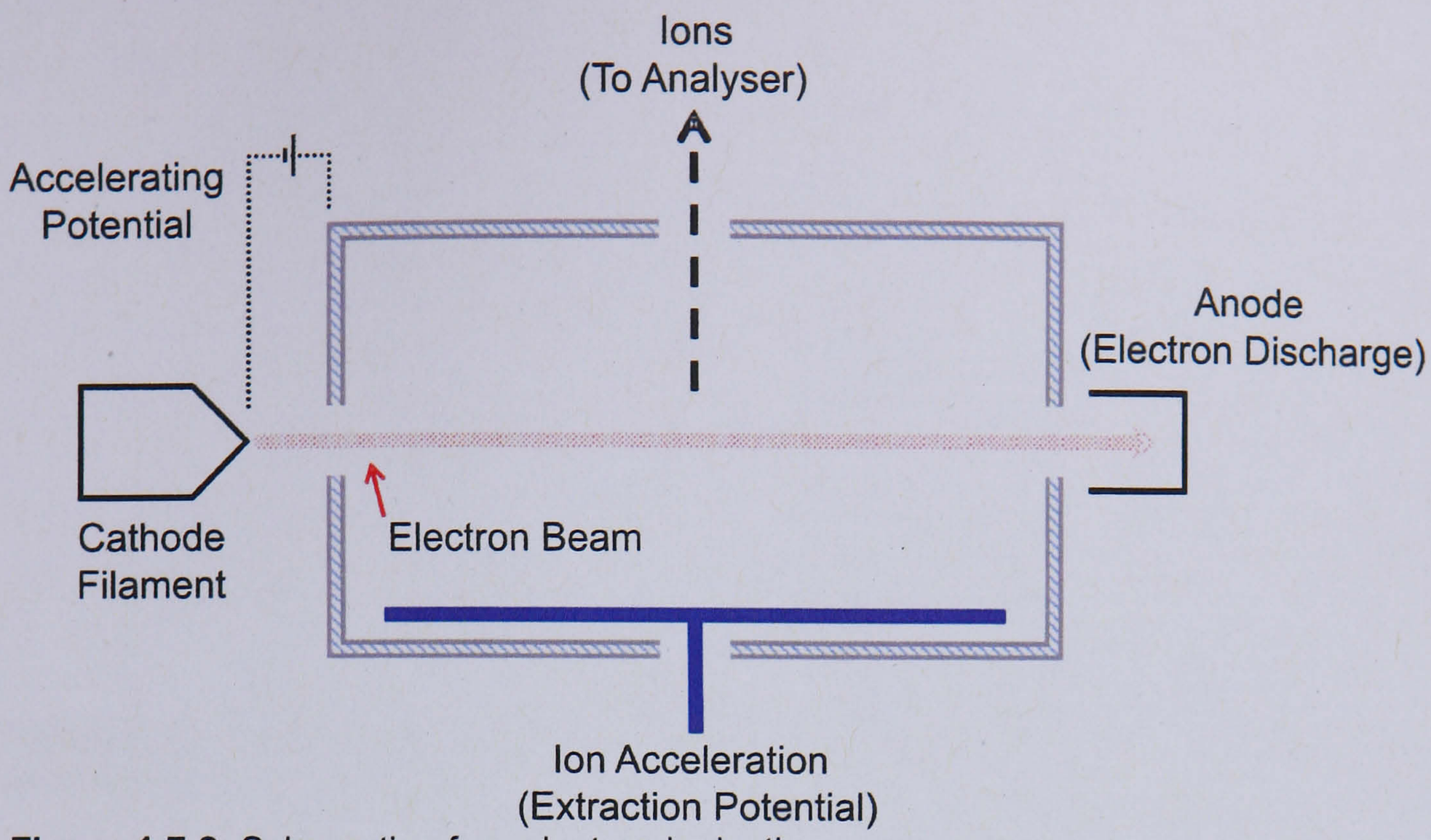


Figure 1.7.2. Schematic of an electron ionisation source.

Figure 1.7.2 shows a schematic of an EI source. The sample is introduced as a vapour into the source under vacuum; this is to ensure that unwanted ion-molecule collisions do not occur as well as maintaining a mean free path. Heating a cathode element causes the emission of electrons, which are accelerated into the source by an accelerating potential of ca. 100 V. The electrons then traverse the source.

The interaction of the electron beam and the gaseous sample may cause the ejection of an electron from the sample to create a positively charged radical ion (equation 1.7.1). This is achieved by the electron beam being associated with a particular wavelength, λ, where h = Planck's constant, m = mass of electron and υ = velocity (equation 1.7.2).

$$M + e^- \rightarrow M^{+\bullet} + 2e^- \qquad \textbf{Equation 1.7.1.}$$

$$\lambda = \frac{h}{m\upsilon} \qquad \textbf{Equation 1.7.2.}$$

The typical kinetic energy of 70 eV used in the EI source corresponds to a wavelength of 140 pm. This is sufficient to cause excitations within the gaseous sample that can cause the expulsion of an electron.

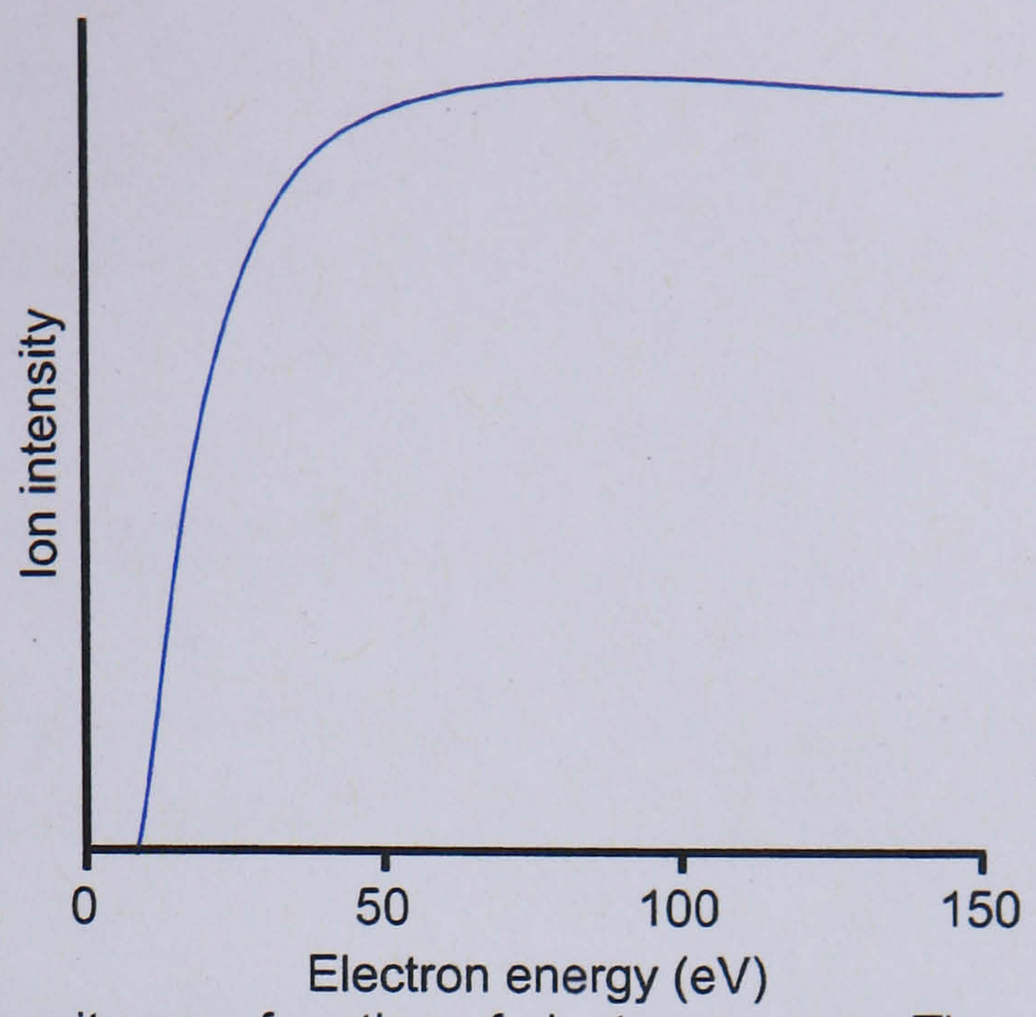


Figure 1.7.3. Ion intensity as a function of electron energy. The maximum ion intensity is less than 70 eV.

The typical energy of 70 eV is used, as this is in the plateau region, away from the sharp drop in ion intensity at lower electron energies (figure 1.7.3). This allows a reproducible spectrum to be obtained, which is helpful for the comparison of spectra from different instruments, making the use of spectral libraries for the identification of unknowns feasible. Most molecules only require around 10 eV of energy to become ionised, but the number of ions generated is insufficient to gain any structural information. Using 70 eV produces ions with excess energy; this excess energy causes the fragmentation of the ion, which is useful, as structural information can be determined from the fragments obtained. One drawback of this fragmentation caused by the excess energy is that the molecular ion is not always observed.

1.7.2.2. Chemical ionisation

Chemical Ionisation (CI) was developed by Munson and Field in 1966 and is a softer ionisation technique than EI, as ions are generated with little excess energy (Munson and Field, 1966). CI therefore produces an easily recognisable MH^+ peak. Ions are produced by the collision of a primary ion (formed in the source) with a molecule introduced into the source, the source being at a suitable pressure that allows collisions between the primary ions and gaseous sample to occur.

A reagent gas, such as methane, is introduced into the source and is ionised by EI to create the primary ions; these primary ions collide with the gaseous sample

introduced into the source. A plasma is subsequently created where many different reactions can occur, causing both positive and negative ions to be formed by proton transfer, hydride abstraction, adduct formation or charge transfer.

1.7.2.2.1. Methane as a reagent gas

The reagent gas is ionised by means of a reaction with an electron beam, as in an EI source, which creates a radical cation that can react further (equation 1.7.3):

$$CH_4 + e^- \rightarrow CH_4^{+\bullet} + 2e^-$$
$$CH_4^{+\bullet} \rightarrow CH_3^+ + H^\bullet$$
$$CH_4^{+\bullet} \rightarrow CH_2^{+\bullet} + H_2$$

Equation 1.7.3.

As the concentration of methane within the source is higher than that of the analyte, the most likely collision is the methane radical cation with methane (equation 1.7.4):

$$CH_4^{+\bullet} + CH_4 \rightarrow CH_5^+ + CH_3^\bullet$$

Equation 1.7.4.

Collisions of methane with $CH_4^{+\bullet}$ or its breakdown product occur (equation 1.7.5):

$$CH_3^+ + CH_4 \rightarrow C_2H_5^+ + H_2$$
$$CH_2^{+\bullet} + CH_4 \rightarrow C_2H_3^+ + H_2 + H^\bullet$$
$$C_2H_3^+ + CH_4 \rightarrow C_3H_5^+ + H_2$$

Equation 1.7.5.

The reaction of CH_5^+ with the analyte molecule, M, ionises M by proton transfer (equation 1.7.6):

$$CH_5^+ + M \rightarrow MH^+ + CH_4$$

Equation 1.7.6.

1.7.2.2.2. Proton Transfer

Proton transfer is the most common ionisation reaction that occurs within the plasma and is effectively an acid/base reaction. The reagent gas forms a Brönsted acid, eg. $CH_4^{+\bullet}$, which donates a proton to the Brönsted base (the analyte molecule, M). The energetics of the acid/base reaction can be controlled by the use of different reagent gases such as ammonia and isobutane.

1.7.2.2.3. Adduct formation

Given the wide range of reactions that can occur in the plasma, adducts can be formed by a third body collision. A protonated molecule, $[M+H]^+$, can form an adduct with another analyte molecule, M (or indeed a different analyte molecule, N), or fragment ions, F^+; these reactions are useful in aiding the confirmation of the protonated molecule peak, or for assessing the purity of a compound by assessing the different molecular peaks in a spectrum (equation 1.7.7).

$$MH^+ + M \rightarrow (2M + H)^+$$
$$F^+ + M \rightarrow (F + M)^+$$
$$MH^+ + N \rightarrow (N + MH)^+$$

Equation 1.7.7.

1.7.2.2.4. Charge transfer

As with EI, radical cations can be obtained upon the use of rare gases. Xenon gas can be ionised to produce a radical cation that subsequently reacts with the analyte molecule, M, to transfer a charge. Xenon radicals transfer less energy to the analyte molecule compared to methane gas, meaning less fragmentation of the molecule occurs (equation 1.7.8).

$$Xe + e^- \rightarrow Xe^{+\bullet} + 2e^-$$
$$Xe^{+\bullet} + M \rightarrow M^{+\bullet} + Xe$$

Equation 1.7.8.

1.7.2.3. Matrix assisted laser desorption ionisation

Matrix Assisted Laser Desorption Ionisation (MALDI) is an ionisation technique that utilises short, intense laser pulses to produce gas phase ions from a mixture of an analyte with a matrix. Two groups developed different methods of laser desorption ionisation. Tanaka *et al.* used a fine cobalt powder in a glycerol matrix with the analyte added to it; a nitrogen laser at 337 nm was used to irradiate the mixture, producing intact ions that are generally singly charged species, the cobalt powder was used to reflect the laser irradiation onto the analyte (Tanaka *et al.*, 1988). Karas *et al.* developed a similar method, where the laser irradiation used does not directly cause the ionisation of the analyte (Karas *et al.*, 1987). They added picomolar amounts of an analyte to a solution containing an excess of an organic matrix that absorbs strongly at the wavelength of the laser. The mixed solution was spotted onto a target plate and allowed to dry, forming crystals. Laser radiation was directed onto the spot, where the organic matrix absorbs the laser energy. Tanaka was controversially awarded the Nobel Prize in 2002 for his work, but it is the technique described by Karas *et al.* that is generally used today.

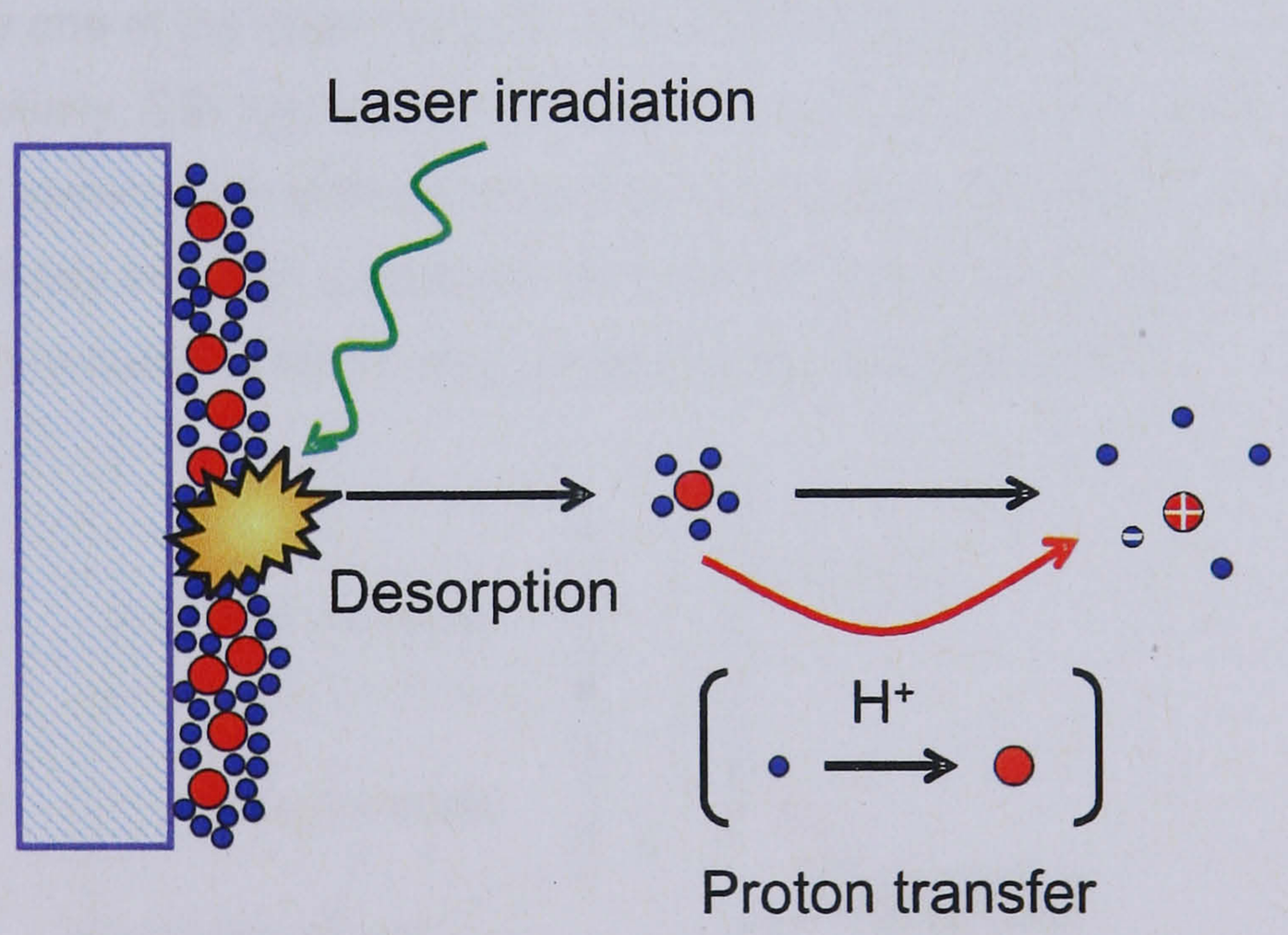


Figure 1.7.4. MALDI ionisation. The formation of gas phase ions from analyte dissolved in a liquid matrix by laser irradiation.

The mechanisms of ion formation by MALDI are not fully understood, but it is generally accepted that upon laser pulsing (10^6 to 10^{10} $W\,cm^{-2}$ of energy deposited) the matrix absorbs the laser radiation, causing rapid heating and subsequent expansion of the matrix and analyte into the gas phase in a plume. The matrix molecules are electronically excited, and can either transfer a proton to or from the

analyte molecule, yielding $[M+H]^+$ or $[M-H]^-$. Figure 1.7.4 shows the desorption of the analyte surrounded by matrix, but in addition other processes also take place, such as the ejection of single molecule aggregates and clusters of the matrix (which are evident in the resulting spectrum).

MALDI is a soft ionisation technique that produces almost exclusively protonated (cationised) molecules. It is also a very sensitive technique, as millimolar (and lower) concentrations are typically spotted onto a target for analysis. MALDI allows the study of large, involatile and thermally labile species and is commonly used for the analysis of biological molecules. Its pulsed nature makes it a suitable ionisation device for time of flight mass analysis (section 1.7.3.7).

1.7.2.4. Electrospray ionisation

Developed by Fenn *et al.* in 1988 (Fenn *et al.*, 1989), electrospray ionisation (ESI) was based on a concept proposed in 1968 by Dole *et al.* (Dole *et al.*, 1968). ESI has since become one of the most versatile and widely used methods for the ionisation of an analyte. Initially, ESI was used for the analysis of proteins, but as interest increased, its applications widened to synthetic polymers and small molecules. Its broad applicability and high sensitivity were further enhanced by the ease with which it could be hyphenated to HPLC and capillary electrophoresis (CE).

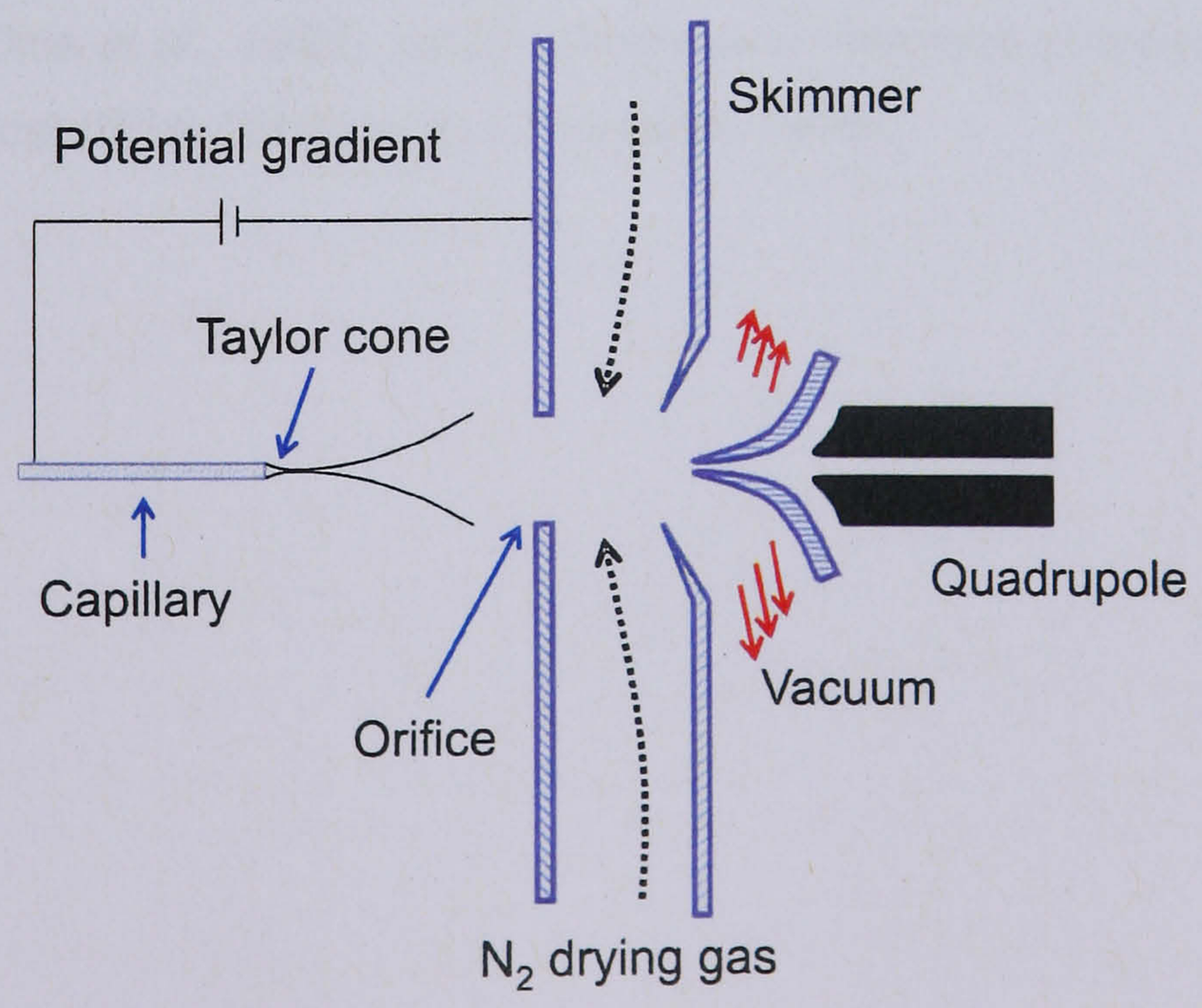


Figure 1.7.5. Schematic of an electrospray ionisation source.

The analyte is dissolved in a suitable solvent, placed in a syringe, and pumped into the capillary by the use of a syringe driver. Alternatively, sample can be received from the output from HPLC or CE. A high voltage is applied to the capillary tip, creating a potential gradient between the capillary end and the orifice (figure 1.7.5). Given that the ESI source is at atmospheric pressure and the mass analyser is under vacuum, a series of lenses and skimmers exist, that act as a physical barrier; roughing pumps and turbo pumps are used to create a vacuum in the analyser.

The voltage (0.5-5 kV) is required to induce charge accumulation at the end of the capillary, meaning that the liquid eluting from the capillary is highly charged. The potential has the effect of inducing a Taylor cone; this occurs at a specific voltage, the onset voltage, where surface tension is overcome causing the release of a plume of small charged droplets. At any voltage lower than the onset voltage, droplets are not emitted, as the charge is not sufficient to break the surface tension. Depending on the flow rate, sometimes a nebulising gas is applied around the capillary, which aids in the formation of the Taylor cone and the emission of droplets. The fine plume of charged droplets can be subjected to a flow of heated nitrogen gas, the drying gas, which causes the solvent to evaporate; the drying gas is not always heated and is sometimes not used at all, this is dependent on the flow rate and the source used.

The formation of gas phase ions is proposed to occur by one of two mechanisms. Dole, who proposed the concept of an ESI source proposed the charge residue model (CRM) (Dole *et al.*, 1968), whilst Iribarne and Thomson proposed an ion evaporation model (IEM) (Iribarne and Thomson, 1976).

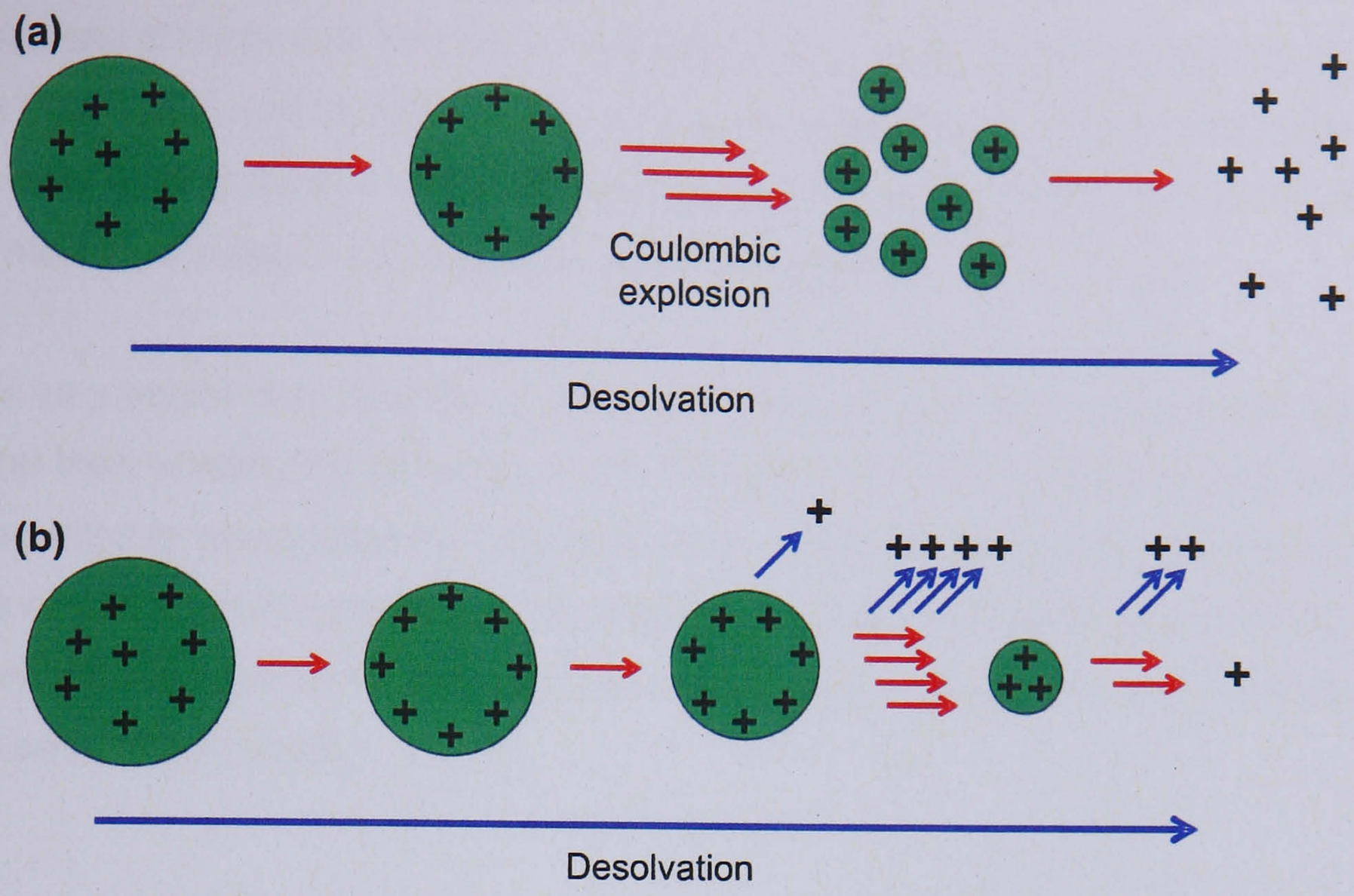


Figure 1.7.6. (a) The charge residue model, as proposed by Dole, (b) the ion evaporation model, as proposed by Iribarne and Thompson.

$$q^2 = 8\pi^2 \varepsilon_0 \gamma D^3$$ **Equation 1.7.8.**

The Rayleigh limit (equation 1.7.8, where q = charge, ε_0 = permittivity of environment, γ = surface tension, D = diameter of a droplet) occurs by the desolvation from the drying gas, causing the evaporation of the solvent, which decreases the volume of the droplet, causing charge to develop at the surface of the droplet. The CRM (figure 1.7.6a) suggests that at the Rayleigh limit, the charges on the droplet coulombically repel each other and subsequently overcome the surface tension, causing a coulombic explosion. This explosion releases smaller droplets, which subsequently undergo the same process until single ions are produced.

The IEM (figure 1.7.6b) involves desolvation by the drying gas; as the droplet shrinks due to the solvent evaporating, charge accumulates on the surface causing coulombic repulsion to occur. At the Rayleigh limit, instead of an explosion, ions are directly released from the surface of the droplet.

If an analyte has ionisable sites, then it is usually protonated (deprotonated in negative mode). However, if the analyte lacks any ionisable sites it can still be ionised through the adduction of anions/cations such as sodium, potassium,

ammonium, chloride and acetate to name but a few. Large molecules tend to have more than one ionisable site, which can lead to multiply charged ions, and given that in a mass spectrometer, the mass to charge ratio is measured, the analysis of very high molecular weight compounds using ESI is possible.

ESI is very sensitive to flow. At very low flow rates (nL min^{-1}) the sensitivity is much greater than at high flow rates (mL min^{-1}). This has led to the development of various ESI sources to incorporate the vast diversity in obtainable flow rates. The various types of ESI source available are important as the flow rates used may be varied depending upon the amount of available sample, and whether the source is being coupled to HPLC or CE.

1.7.2.4.1. Nanospray

Wilm and Mann worked on the concept of the miniaturisation of the ESI source, initially developing a micro-electrospray source, which was later to be renamed the nano-electrospray source (Wilm and Mann, 1994). The setup is similar to that shown in figure 1.7.5, but there is reduced drying gas flow and no need for a nebulising gas. Nanospray uses a capillary with an internal diameter of 1-2 µm, compared to that of 20-100 µm for ESI. Only 0.2 to 2 µL of sample solution is loaded onto the capillary, which can provide around 30 minutes of analysis time whilst only consuming ~1 µL of sample. The capillary is positioned around 1-5 mm directly in front of the orifice, but slightly off-axis, with a potential being applied to the capillary. The potential alone is sufficient to induce the formation of a Taylor cone with a flow rate in the region of 20 nL min^{-1} (ESI originally used 1-10 µL s^{-1}).

Nanospray produces droplets that are less than 200 nm in diameter, compared to around 1 µm for droplets produced by ESI. The small and monodisperse droplets have a high surface to volume and charge to volume ratio, meaning that the generation of gas phase ions is almost instantaneous. Given the size of the droplets, Dole's charge residue process is the most plausible, as only one molecule per droplet is possible, meaning that only one fission step can occur (compared to multiple in ESI). Nanospray is up to 500 times more efficient at ion generation than ESI (Juraschek *et al.*, 1999), and generally has a better signal to noise ratio and a reduced amount of clustering due to the direct emission of ions, which creates a higher tolerance to salts.

1.7.2.4.2. High flow rate ESI

The ESI source used for this study was the Applied Biosystems TurboIonSpray (figure 1.7.7). Ionisation efficiency is increased at high flow rates (5-1000 µL min^{-1}) with a solvent composition that can vary from 100 % aqueous to 100 % organic. The main function is to allow the direct coupling to HPLC using analytical columns without the need for reduced flow or splitting of the flow post HPLC.

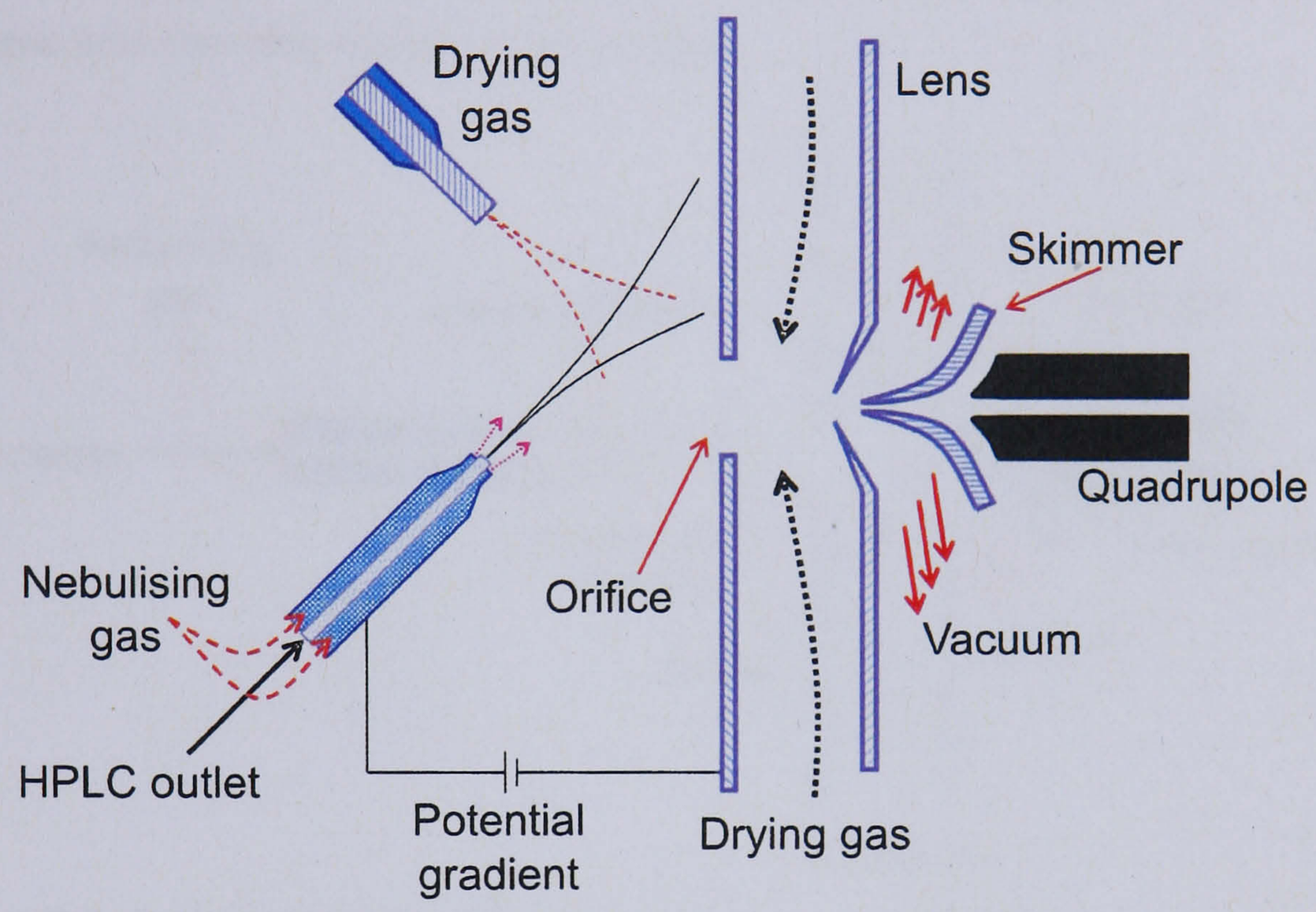


Figure 1.7.7. The Applied Biosystems TurboIonSpray source (used for ESI in this study).

The ionisation efficiency for high flow rates is increased by the addition of a second drying gas (figure 1.7.7). The capillary is set at 45° to the orifice and is offset so that the Taylor cone is not directed directly into the orifice. Perpendicular to the Taylor cone is the additional drying gas, which aids in the desolvation of the spray from the capillary. The additional drying gas can be heated to a temperature of 500 °C.

1.7.2.5. Atmospheric pressure chemical ionisation

Atmospheric Pressure Chemical Ionisation (APCI) is a soft ionisation technique that is appropriate for coupling to HPLC; it is used for the analysis of compounds that are not ionised efficiently by other methods. Unlike CI, which uses an electron beam, APCI uses a corona discharge to produce primary ions from the nebulising gas, which reacts with solvent molecules, forming reagent ions (secondary ions). Ionisation is achieved by proton transfer or abstraction from the secondary ions, and gas phase ions can also adduct to the analyte.

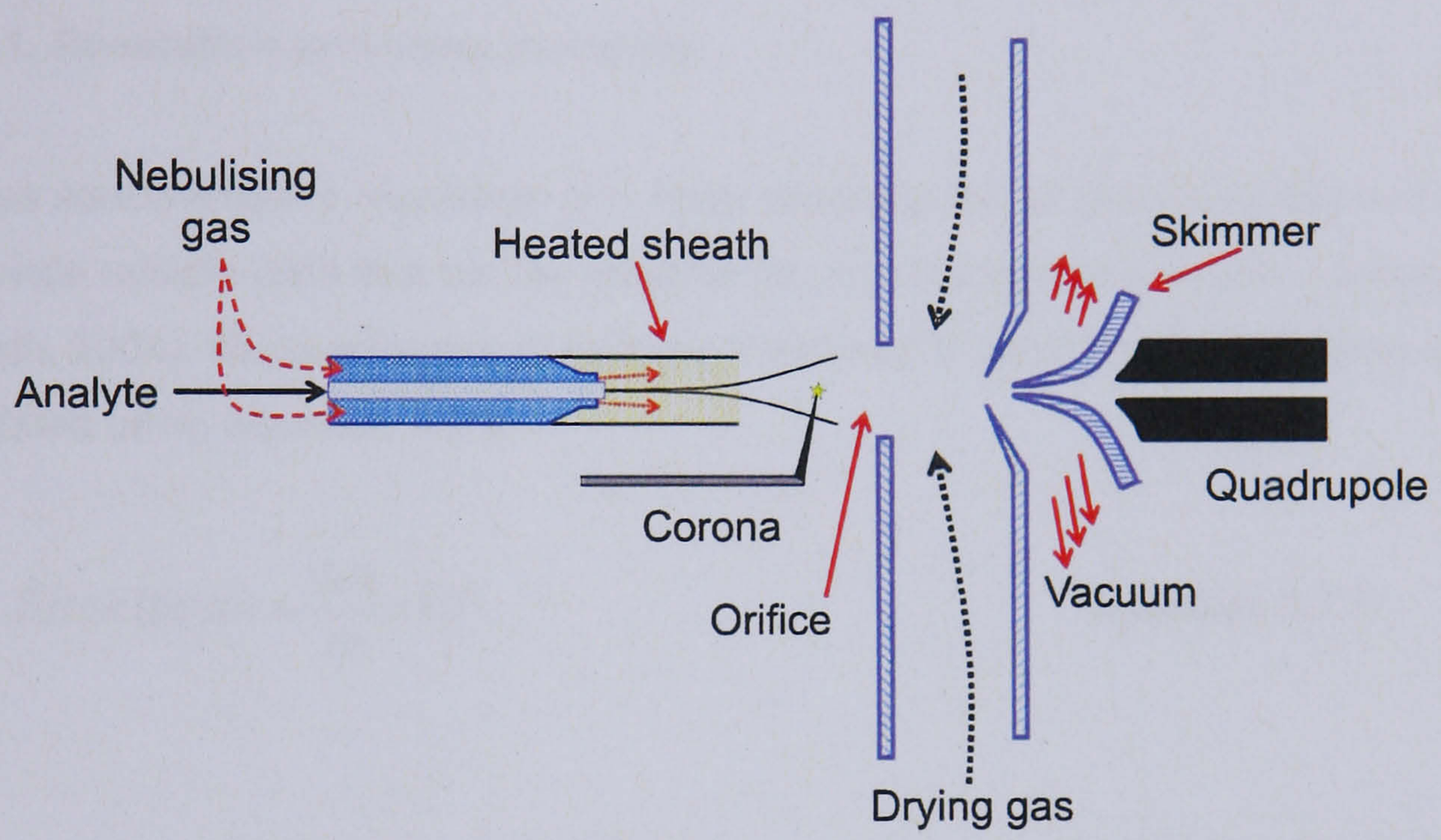


Figure 1.7.8. Schematic of the APCI source.

The main features of an APCI source are shown in figure 1.7.8. The analyte, dissolved in a suitable solvent, is passed through a capillary and nebulised by nitrogen gas at high pressure. The sheath around the capillary is heated, causing the nebulised solvent to vapourise. A corona discharge needle produces electrons, when the vapourised sample comes into contact with the corona discharge needle, ionisation occurs.

1.7.3. Mass analysers

In the mass analyser, ions are separated according to their mass to charge (*m/z*) ratio. Magnetic sector instruments were the first type of mass spectrometers (MS) used. Along with quadrupole MSs, these are scanning analysers, meaning that ions of different *m/z* ratios are transmitted over a period of time. Later inventions such as time of flight (ToF) separate ions in time, and ion traps (IT) in space.

1.7.3.1. Resolution and mass accuracy

A mass spectrometer's resolution and mass accuracy are of great importance if it is to provide reliable data that can be used for the identification of unknown compounds (Balogh, 2004). Mass accuracy is typically measured in parts per million (ppm) and is calculated using equation 1.7.9:

$$\text{Error (ppm)} = \frac{\delta m}{m} x 10^6 \qquad \textbf{Equation 1.7.9.}$$

There are two methods for determining the resolution of an MS. Initially, resolution was defined as the ability of an MS to resolve two peaks of similar mass. With that definition, two peaks were usually considered resolved if the valley between the peaks is equal to or less than 10 % of the lowest intensity peak, as shown in figure 1.7.9a. Figure 1.7.9b shows the full width half maximum (FWHM) method, which is now the most commonly chosen method of reporting a peak's resolution.

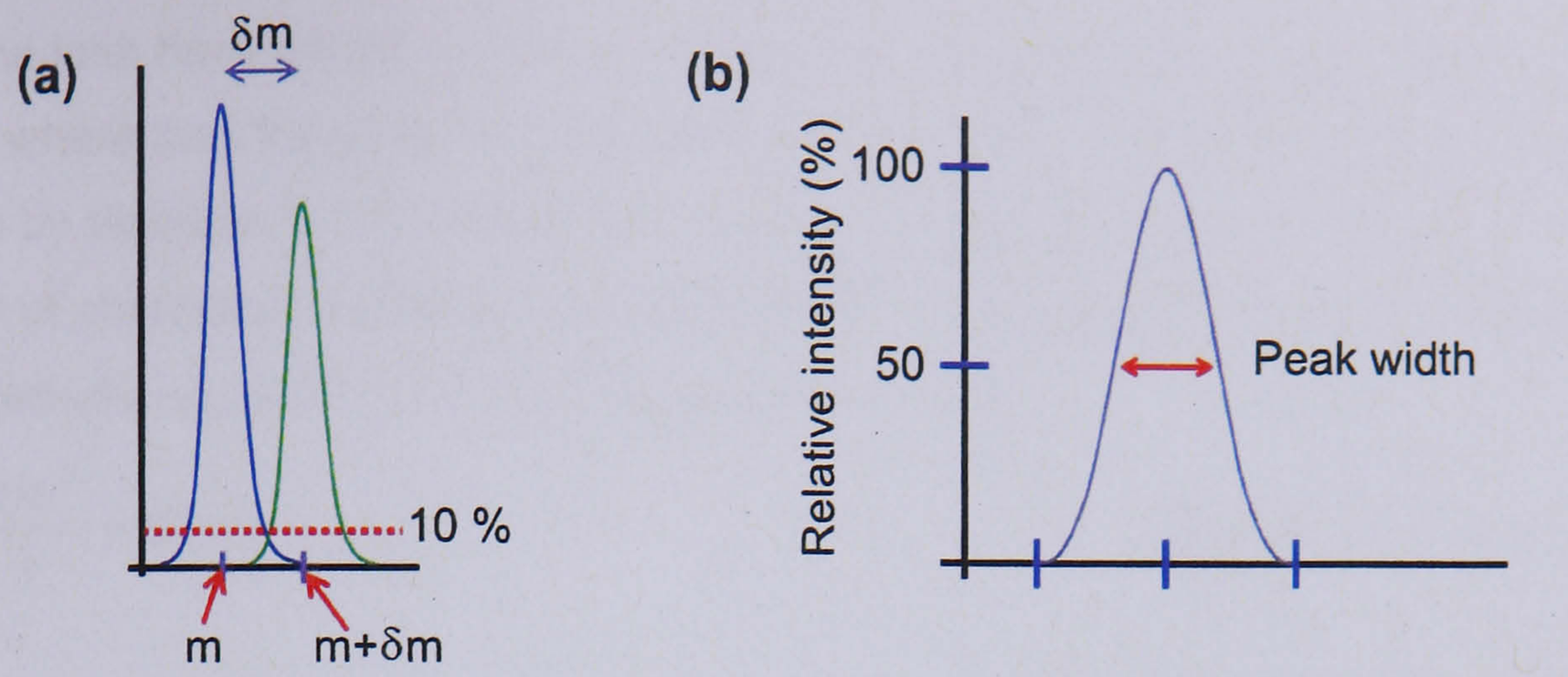


Figure 1.7.9. (a) Two peaks are considered resolved, as the valley is at 10 % of the intensity of the weaker peak. The resolution, R, is given by the mass, m (where z = 1), divided by the change in mass, δm (R = m/δm); (b) The full width half maximum (FWHM) definition takes the peak width at half maximum, or 50 % of the relative intensity.

1.7.3.2. Magnetic sector analysers

Magnetic sector instruments were the first type of commercially available mass spectrometer. Ions are generated (usually by EI) and subsequently accelerated from the source. The ions gain kinetic energy through the acceleration, which is given by equation 1.7.10 (where z = number of charges, e = charge of an electron, V_{acc} = accelerating potential, m = mass and v = velocity of an ion.)

$$zeV_{acc} = \frac{mv^2}{2}$$ **Equation 1.7.10.**

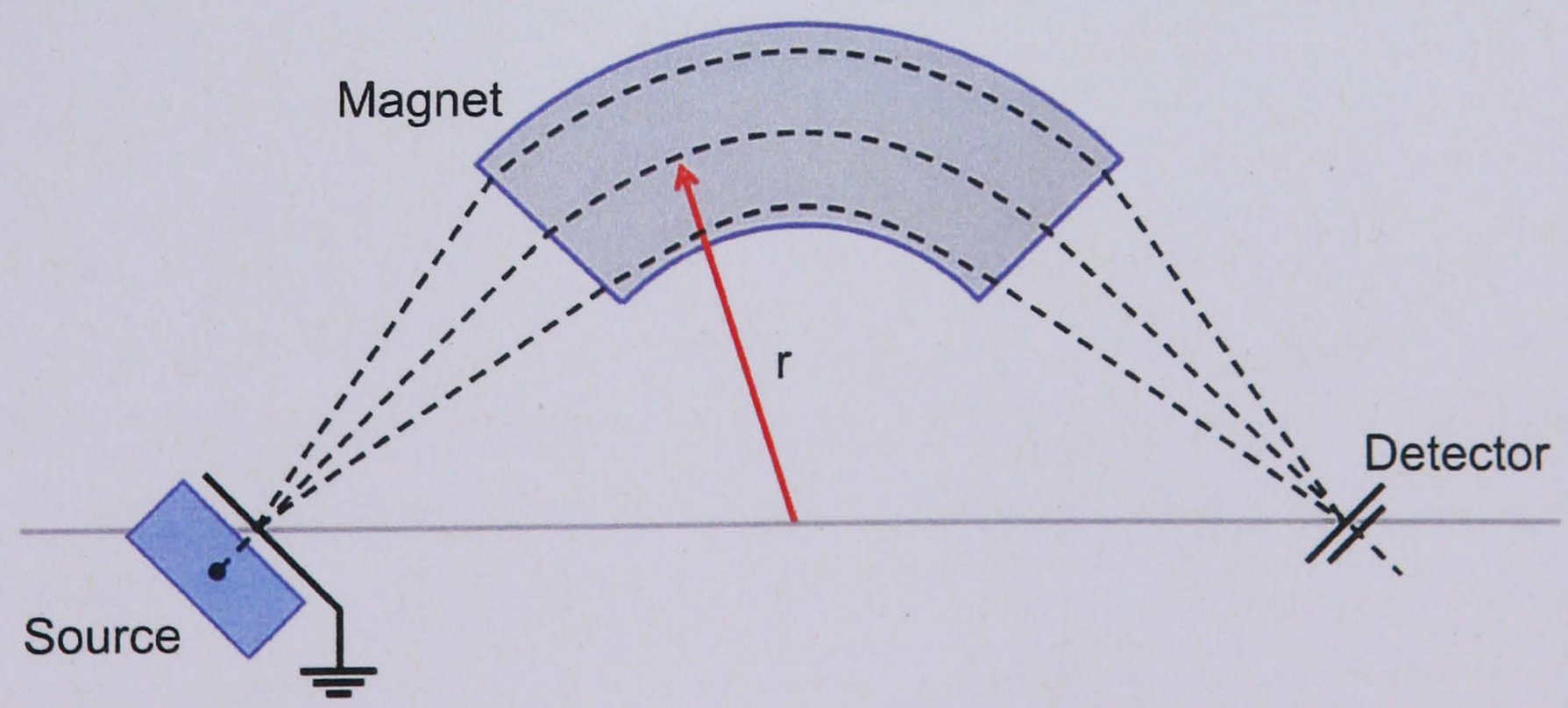


Figure 1.7.10. A magnetic sector instrument. The magnetic field has a strength of B, given in Tesla; the ions traverse the field through a radius, r.

Once the ions have left the source, they enter the mass analyser region (figure 1.7.10) where they traverse the magnetic field. The motion of the ion through the field is given by equation 1.7.11 (where B = magnet strength, in Tesla, r = radius, z = number of charges, e = charge of an electron, υ = velocity of an ion), which, when combined with equation 1.7.10, forms equation 1.7.12:

$$\frac{m\upsilon^2}{r} = Bze\upsilon$$ **Equation 1.7.11.**

$$\frac{m}{z} = \frac{B^2 r^2 e}{2V}$$ **Equation 1.7.12.**

Equation 1.7.12 shows that by varying the values of B or V, different *m/z* values can traverse the field to reach the detector. Scanning is carried out as an exponential function beginning with the high masses. Magnetic sector instruments have a constant resolution over all masses, which means at lower masses, there is a low δm (greater separation) compared to a high δm for high masses. If the scanning were to be constant and not exponential, then not all of the ions at lower masses would be transmitted; hence the exponential scanning allows the detection of all of the ions at lower masses. The exponential scan mode function is given by equation 1.7.13 (where m = mass at time t, m_0 = starting mass at t = 0) :

$$m = m_o e^{-kt}$$ **Equation 1.7.13.**

1.7.3.3. Quadrupole analyser

In 1953, Paul and Steinwedel developed the principle of the quadrupole which was subsequently developed and refined and is now an integral part of many modern mass spectrometers (Paul and Steinwedel, 1953).

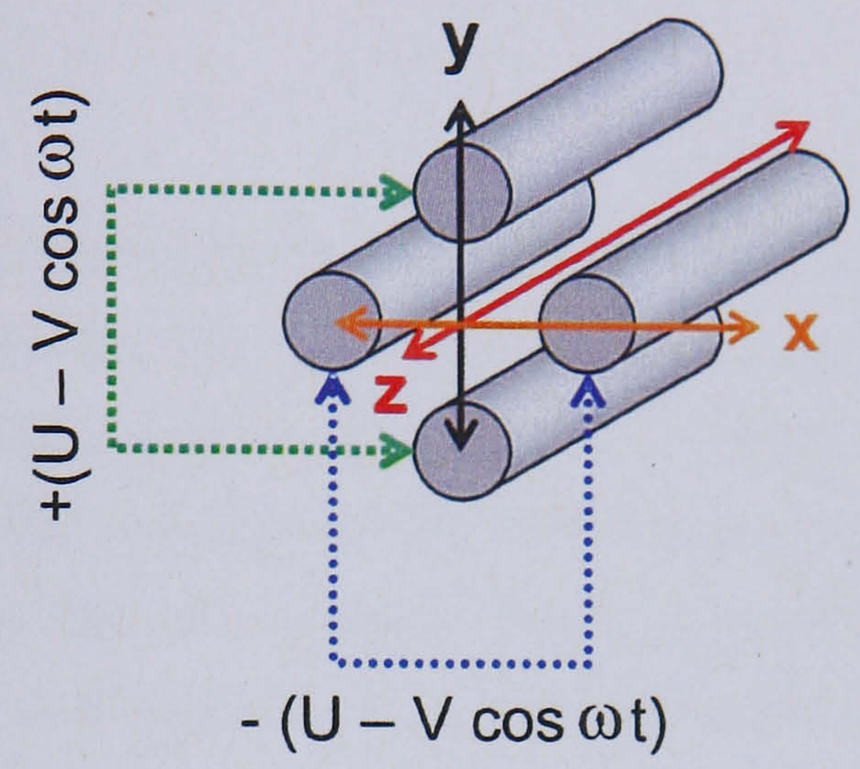


Figure 1.7.11. Schematic of a quadrupole. U = DC potential, V = amplitude, ω = angular frequency (2πv, where v = frequency).

Quadrupoles consist of 4 parallel rods (figure 1.7.11) that have a circular or hyperbolic cross section. Opposite rods are electronically connected to one another; one pair is connected to a positive potential + *(U – Vcosωt)*, and the other pair to a negative potential – (*U – Vcosωt*). *U* is a direct current (DC) component which is variable; an alternating radio frequency (RF) with amplitude *V*, and frequency ω, is applied. The RF potentials that are applied to the two sets of rods are 180° out of phase. Ions enter the quadrupole at a constant velocity where they are subjected to the alternating RF field, which is superimposed onto a constant field, the positive and negative DC potentials. Ions oscillate through the quadrupole along the z axis; depending on the ion's mass, the potentials *U* and *V* and the RF frequency applied. Ions of a certain *m/z* ratio have a stable trajectory and therefore pass through the quadrupole. Ions which do not have a stable trajectory are not transmitted as they are attracted towards the rods where they collide, losing their charge. Altering the DC and RF potentials, but maintaining a constant DC/RF ratio, allows the successive transmission of ions with different *m/z* ratios.

E. Mathieu developed equations in 1866 that described vibrations across a stretched membrane; the equations were found to be applicable to the motion of an ion and are therefore used to determine the parameters required to allow an ion of a particular

m/z to be transmitted through the quadrupole with a stable trajectory (equation 1.7.14, where z = charge on ion, e = charge of electron, m = mass, ω = angular frequency, r_0 = radius from z-axis to quadrupole (diameter is $2r_0$)).

$$a_u = \frac{8zeU}{m\omega^2 r_0^2} \quad \text{and} \quad q_u = \frac{4zeV}{m\omega^2 r_0^2} \qquad \textbf{Equation 1.7.14.}$$

From equation 1.7.14, a mass stability diagram (figure 1.7.12) can be created. The areas where an ion is stable in the x and y dimensions according to the solutions to the Mathieu parameters are circled. The areas of ion stability can be used to provide mass discrimination; not all of the areas are commonly used in mass spectrometry due to the high potentials that are required. Therefore most mass spectrometers utilise the area that lies on the q_u axis, as the potentials are suitable for use within a quadrupole.

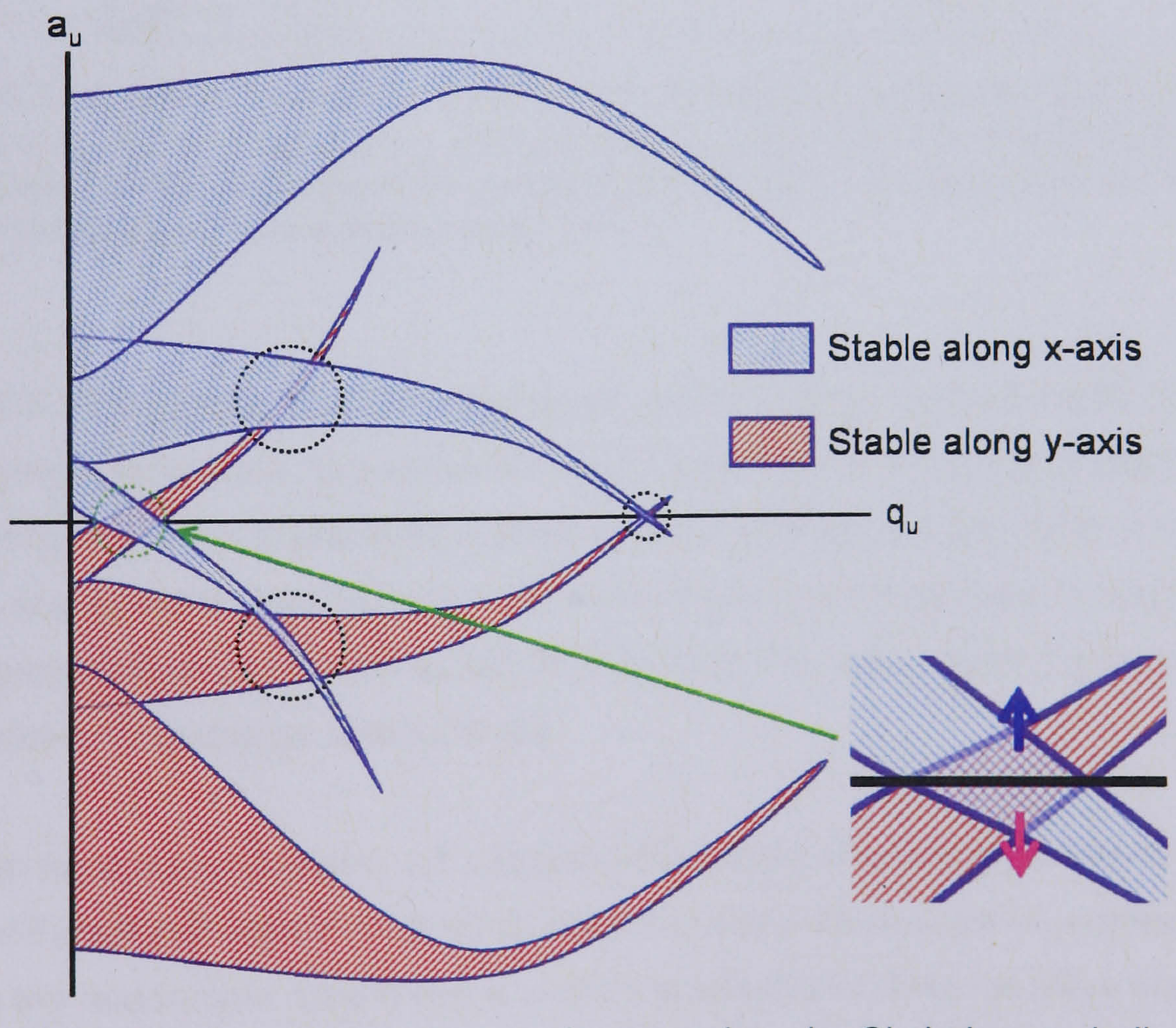


Figure 1.7.12. Mathieu stability diagram for a quadrupole. Circled areas indicate where ions are stable in a quadrupolar field. Inset is an expansion of the area that is utilised in most mass spectrometers. Above the q_u-axis corresponds to +U, and below to –U.

Providing that an ion's trajectory does not exceed the fixed r_0 value (where it would collide with a quadrupole and discharge) its trajectory is stable along the length of the quadrupole (z-axis) and is transmitted; when an ion moves away from the z axis of the quadrupole the potential field that the ion experiences is greater, causing the ion to be focussed back to the centre along the z-axis.

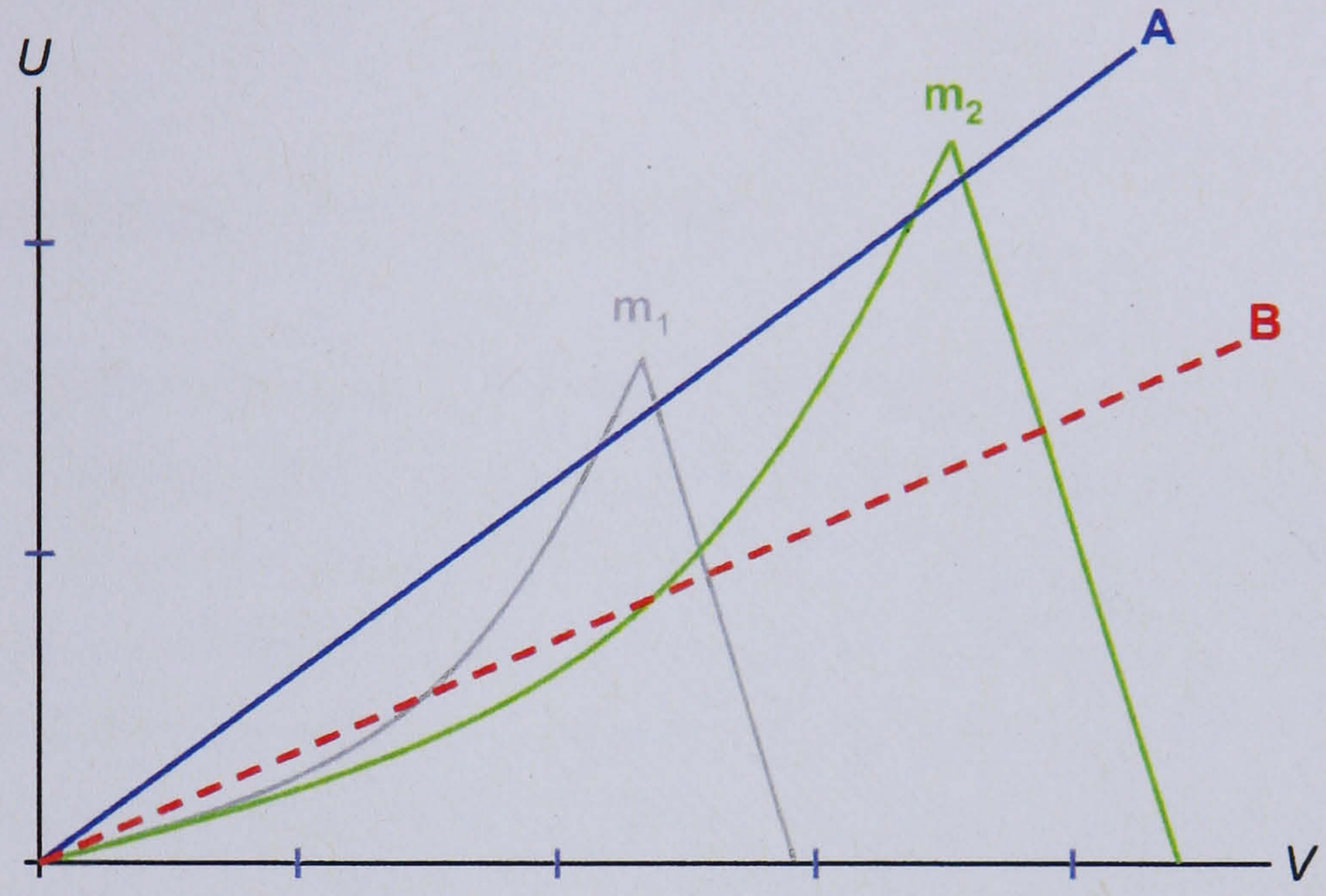


Figure 1.7.13. Stability areas for positive ions of different masses (m_1 and m_2) as a function of a_u and q_u. Ions have a stable trajectory if the scan line (A and B) passes through the stability area; mass resolution is achieved by operating the scan line close to the apices of the stability areas, line A.

Figure 1.7.13 illustrates how ions of different *m/z* values can be transmitted through a quadrupole, giving mass discrimination. Mass separation is achieved by altering the *U* and *V* values whilst maintaining a constant ratio between the two (*U/V*). Ions that have a_u and q_u values that fall within the stable region are transmitted through the quadrupole; ions whose a_u and q_u values lie outside the stable region have unstable trajectories and discharge against a rod.

The scan lines A and B (figure 1.7.13) both pass though the stability areas for two ions of different *m/z* values, m_1 and m_2, meaning that both ions are transmitted through the quadrupole. Line B has a shallow gradient and does not allow mass discrimination as both ions of m_1 and m_2 have stable trajectories and are detected together and therefore not resolved. Line A has a steep gradient and passes through the apices of the stability areas for the two ions; these ions are resolved and discrimination between the two masses is now possible.

A quadrupole's resolution is controlled by altering the DC component, *U*. A high *U/V* ratio has a higher resolution but reduced transmission; initial ion velocities (amount of kinetic energy) and the position of entry into the quadrupole means that operating close to the boundaries of stability are impractical, hence why quadrupoles usually only operate at unit resolution.

1.7.3.4. Quadrupoles as a collision cell – MS/MS

Soft ionisation techniques such as ESI, MALDI and APCI only produce mainly protonated molecules with little, if any, fragmentation. For structural information to be obtained, a molecule must be fragmented to yield structurally informative fragment ions. Molecules can be fragmented by their collision with an inert gas. Jennings and McLafferty described the concept of colliding ions with a collision gas in the 1960s (Jennings, 1968; McLafferty, 1968); originally called collision activated dissociation (CAD), it is now commonly referred to as collision induced dissociation (CID). If an ion collides with an inert gas then part of its kinetic energy is converted into internal energy. This internal energy, if sufficient, induces fragmentation.

If a quadrupole is operated in RF only mode ($U = 0$), provided that V is still within the area of stability, all ions have a stable trajectory and pass through the quadrupole; they are thus focussed along the z-axis. Ions that enter an RF only quadrupole can fragment through metastable dissociation, or by their collision with an inert gas present at a low pressure, CID; when ions fragment within an RF only quadrupole it is referred to as a collision cell.

The coupling of two mass analysers, such as quadrupoles, separated by a collision cell, allows the analysis of individual components from a mixture. The coupling of mass analysers creates a tandem mass spectrometer that is tandem in space; tandem in time instruments are described later (section 1.7.3.7). Tandem mass spectrometry enables a series of experiments that allows a wide range of information about an ion to be obtained (figure 1.7.14).

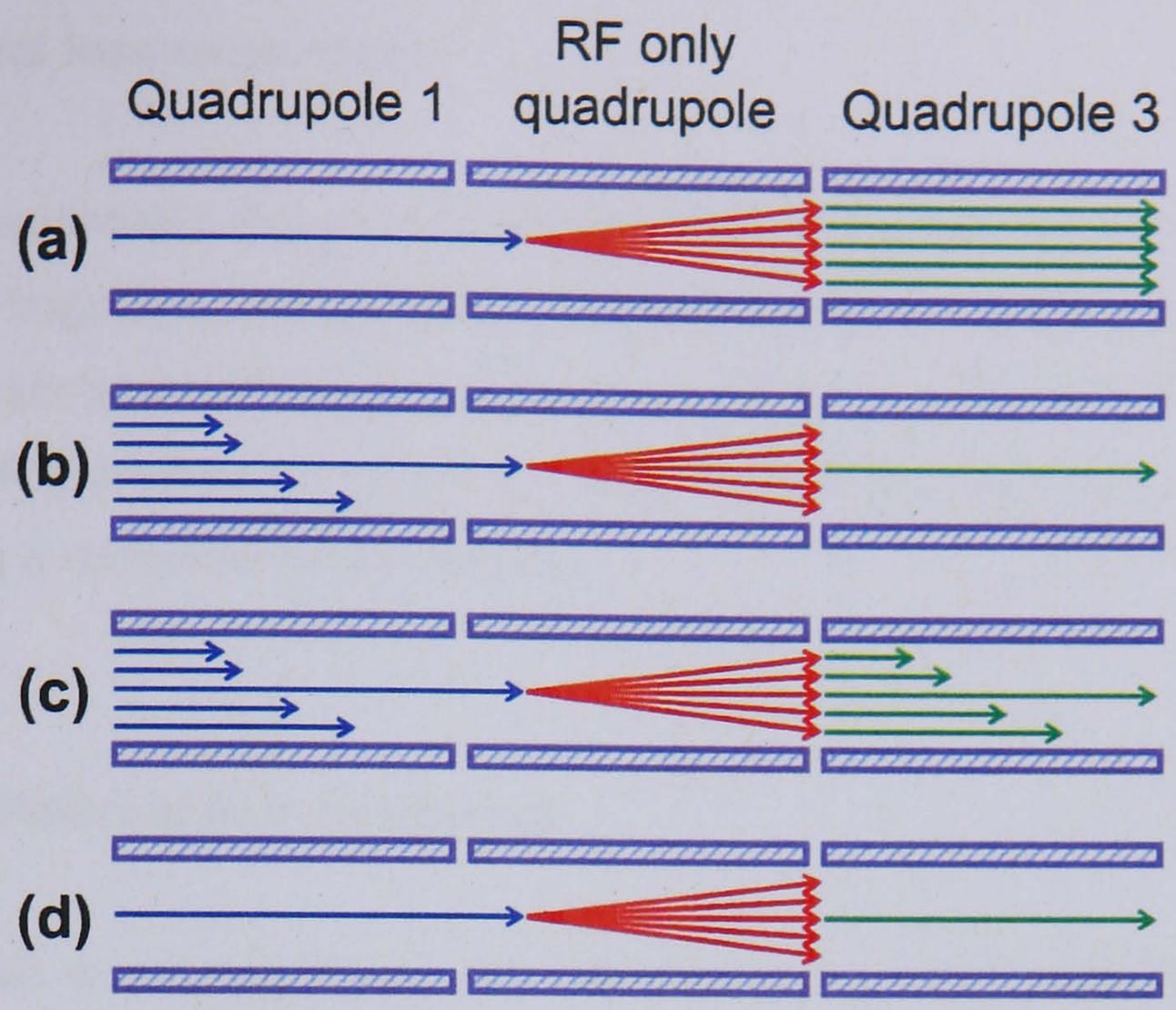


Figure 1.7.14. Summary of the four tandem mass spectrometry experiments. (a) product ion experiment. (b) precursor ion experiment. (c) neutral loss experiment. (d) selected reaction monitoring.

1.7.3.4.1. Product ion experiment

A product ion experiment (figure 1.7.14a) is where a precursor ion is selected in the first mass analyser and transmitted to the collision cell, where it fragments by metastable dissociation or by CID. The fragments are analysed in the second mass analyser, yielding the product ion spectrum.

1.7.3.4.2. Precursor ion experiment

The precursor ion experiment (figure 1.7.14b) is conceptually the opposite of a product ion scan. The second mass analyser is set to only allow ions of a specific *m/z* ratio to be transmitted. Ions are scanned through the first mass analyser and are fragmented in the collision cell; the detector gives a response if an ion transmitted through mass analyser one fragments to the product ion *m/z* value set in mass analyser two.

1.7.3.4.3. Neutral loss experiment

A neutral loss experiment (figure 1.7.14c) identifies a precursor ion that loses a specific neutral fragment, such as water. As with the precursor ion experiment, the first mass analyser sequentially allows ions through to the collision cell to undergo fragmentation. Any product ion that has a mass corresponding to the constant mass difference gives a response at the detector.

1.7.3.4.4. Selected reaction monitoring

Selected reaction monitoring (figure 1.7.14d) records a specific precursor ion giving a specific fragment ion. The first mass analyser transmits specific precursor ions that are subsequently fragmented in the collision cell. The second mass analyser is set to transmit specific fragments; a signal is only given if the selected precursor ion yields the correct fragment ion.

1.7.3.5. Ion traps

The quadrupole Ion Trap (IT) was described by Paul *et al.* in 1953, and initially adopted by physicists to investigate the properties of trapped ions (Paul and Steinwedel, 1953). The initial problems of poor resolution, mass range and the ability to study (trap) only one *m/z* value at a time due to 'mass selective stability' were eventually overcome by the work by Stafford's group (Stafford, 2002). Stafford developed the 'mass selective instability mode', meaning that ions could be trapped and sequentially ejected from the trap and detected; the addition of helium gas to the trap was found to vastly improve the peak shape and resolution.

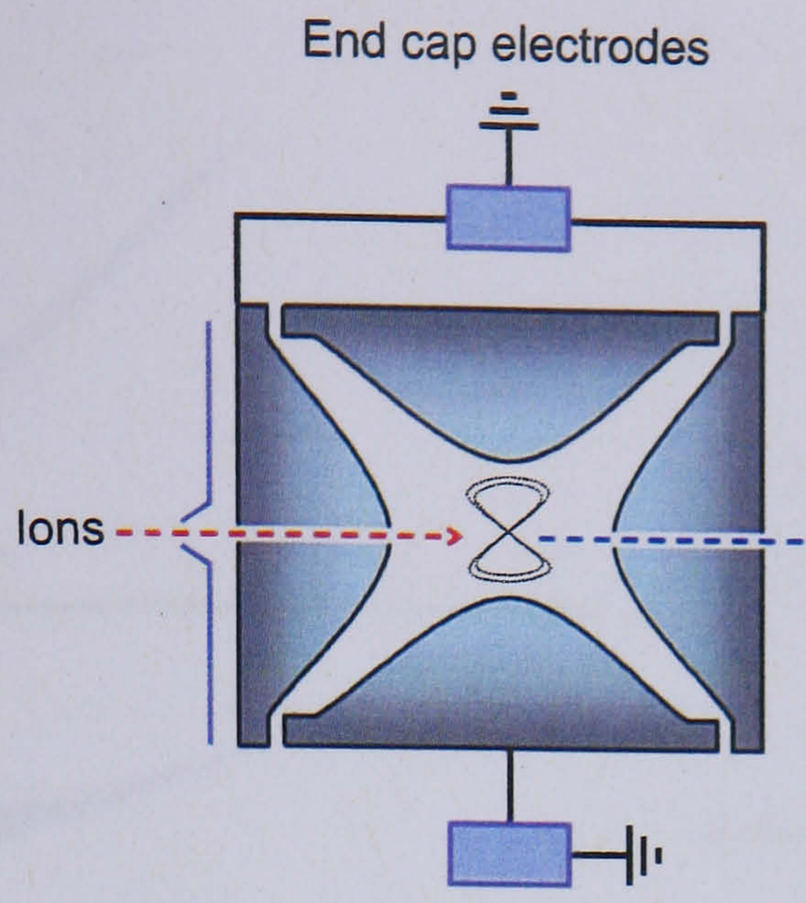


Figure 1.7.15. A schematic of an ion trap. The ring electrode utilises a fundamental RF whereas the end caps utilise variable RF for the excitation or ejection of ions. Typical ion trap geometries are less than 1 cm^3.

The ion trap (figure 1.7.15) consists of a ring electrode and two end caps which are all hyperbolically shaped. Ions are received from a multipole (which has focussed the ions) where they enter the trap through the end cap (when a repelling potential on the gate lens is dropped) and are subjected to a multipole field. As with quadrupoles, ions are only stable within a certain field that is determined by solutions to the Mathieu equations. Since the ion trap is cylindrical with respect to the ring electrode, the x and y coordinates can be reduced to r^2 ($x^2 + y^2 = r_0^2$).

The stability of an ion can be expressed using only the *z* and *r* coordinates. The application of DC (*U*) and RF (*V*) is required and is applied in the Mathieu parameters (equation 1.7.15, where $\omega = 2\pi f$).

$$q_z = \frac{4eV}{mr_0^2\omega^2} \quad \text{and} \quad a_z = \frac{-8eU}{mr_0^2\omega^2}$$ **Equation 1.7.15.**

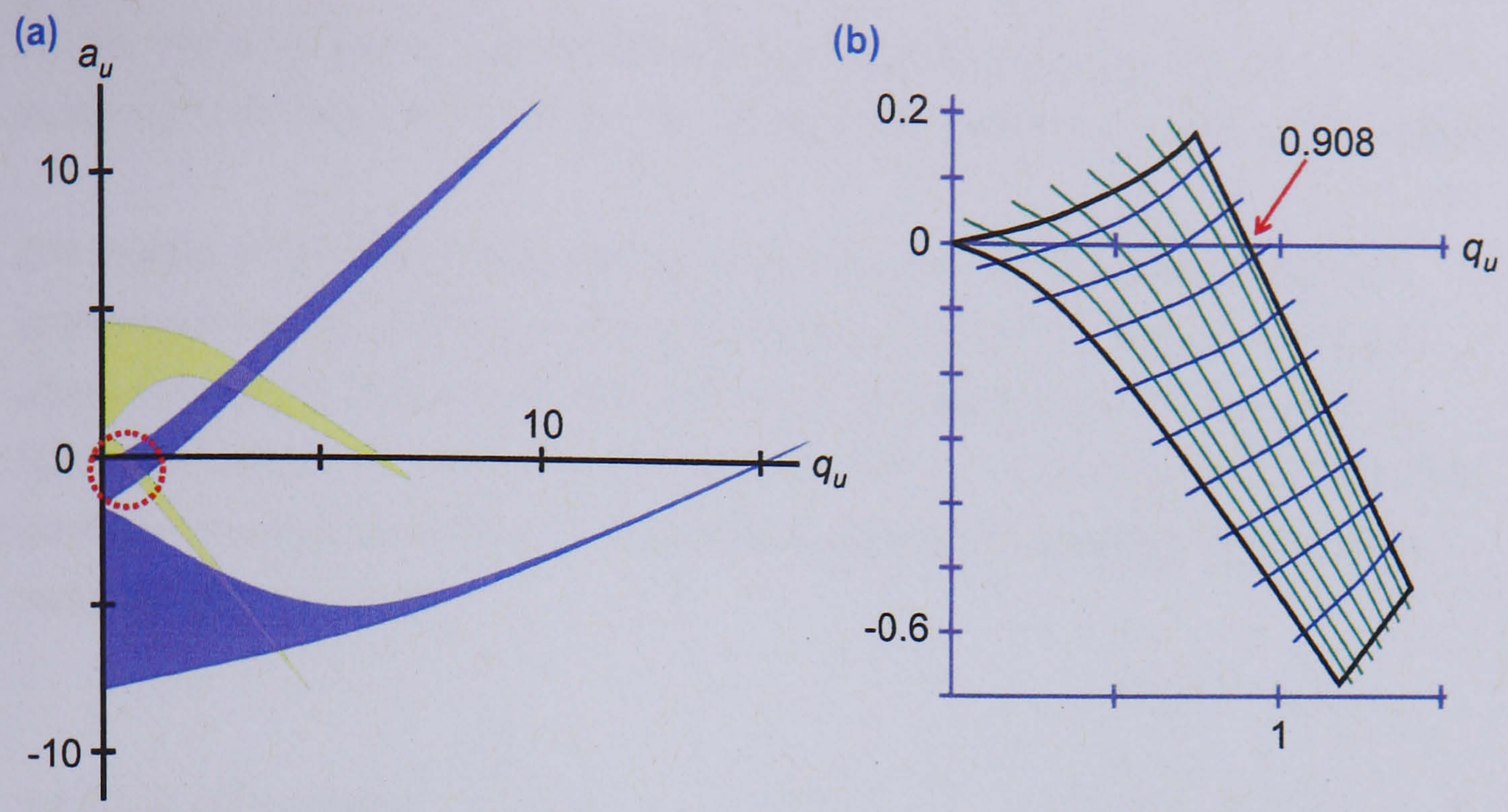


Figure 1.7.16. (a) The stability diagram for ions in an ion trap. Blue areas indicate stability on the r-axis (between the ring electrode) and the yellow areas indicate stability on the z-axis (between the end caps). The area of stability circled is the one generally utilised and is expanded and shown in (b).

For an ion to be stable in an ion trap, it must be stable on both the r and z axis. The three areas of this stability are shown in figure 1.7.16a; the area circled (and shown in figure 1.7.16b) is the only one that can be practically utilised, as the parameters that would be required to use the other areas cannot be used as arcing between the ring and end cap electrodes would occur. The mass selective instability mode allows the maximum number of ions possible to be trapped; this is operated with no DC potential, meaning that $U = 0$ and therefore $a_z = 0$. As $a_z = 0$, the ions in the trap effectively 'sit' on the q_z axis with ions of high mass at low q_z values and ions of low mass at high q_u values.

1.7.3.5.1. Injection and trapping of ions

The gate lens at the entrance of the ion trap is pulsed from positive to negative (in positive mode) to allow a packet of ions from a quadrupole to enter the trap. The ions are subjected to an RF field that is applied to the ring electrode; the RF frequency is held constant but its amplitude (V) can be varied (discussed later). The RF field increases linearly from the centre of the trap, which causes the ions to be focussed into the centre of the trap, where they oscillate in a 'figure of eight' pattern (figure

1.7.15). Helium is an inert gas that is introduced at low pressures into the trap to aid focussing of the ions, as it reduces their excess kinetic energy by 'collisional cooling'.

The number of ions that enters the trap has to be carefully controlled to avoid a phenomenon called the 'space charge effect'. Ions need to be at a concentration where ions do not 'see' one another, effectively existing in infinite space without interaction with another ion. Too many ions result in the ions being able to 'see' one another, causing a distortion of the quadrupole field and consequent loss of mass accuracy and resolution.

1.7.3.5.2. Ion ejection

A simple MS experiment requires the ejection and detection of the ions that are stored in the trap. Figure 1.7.16b shows a q_z value of 0.908, this is called the q_{eject} and the q value at which ions stored in the trap become unstable in the axial direction (between the end caps) but remain stable in the radial direction (between the ring electrode). The ions are ejected from the trap when they reach this value by ramping the RF amplitude (V), but due to the geometry of the trap, only 50 % of the ions reach the detector as the ions pass towards both end caps, only one of which leads to the detector in the most commonly used IT design.

In order to eject the ions, the RF amplitude (V) is sequentially ramped causing the ions to become unstable and be ejected in order of increasing *m/z* value. There is a problem that arises which means there is a limit to the potential that can be applied to the trap before arcing between the electrodes occurs; ions of a high mass are therefore not excited enough to become unstable and therefore remain in the trap (figure 1.7.17a,i,ii,iii).

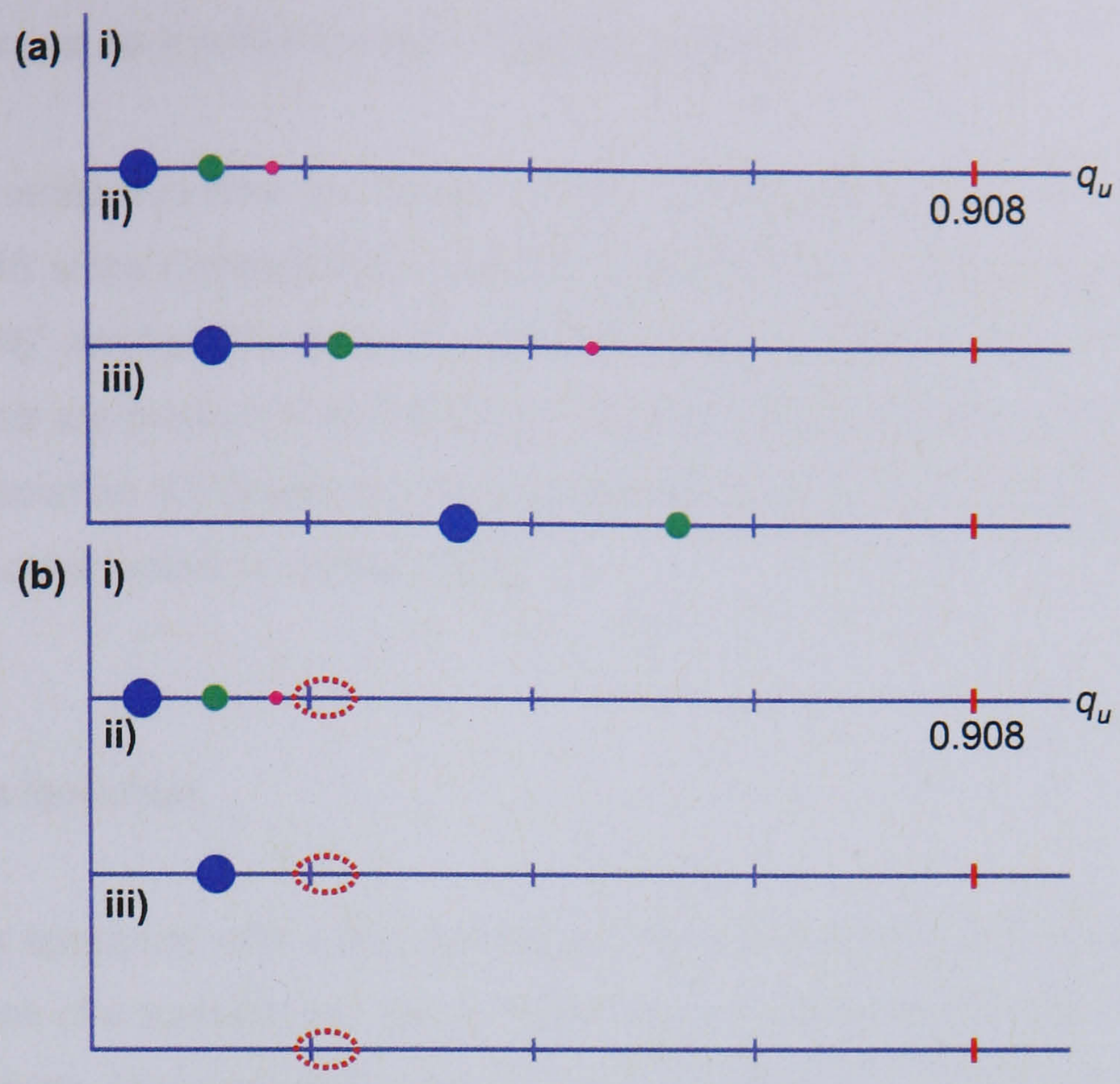


Figure 1.7.17. An expansion of the q_z axis to illustrate the problem of high mass ions remaining in the trap at the maximum RF amplitude. (a) i) The ions are collisionally cooled and focused and are at a low q_z value. ii) The RF amplitude has been increased, the low mass ions (pink) begin to move along the q_z axis as they become excited and unstable. iii) The amplitude has been increased to its maximum, the low mass ions have been ejected whereas the high mass ions remain in the trap. (b) A secular frequency creates an area of instability at lower q_z values (red elipse). i) The ions are cooled and focussed in the centre of the trap. ii) An increase in amplitude has ejected the ions of a low mass and intermediate mass (green) through the area of instability. iii) At maximum amplitude, all of the ions have been ejected from the trap by the creation of the area of instability at a q_z value sufficient to allow the high mass ions to be ejected.

Trapped ions of a given *m/z* oscillate at a specific frequency, called the secular frequency. When ions are excited by a secular frequency, they gain energy, becoming unstable and move towards the end caps. Figure 1.7.17b,i,ii,iii, shows how the application of a secular frequency creates a 'hole' where ions can be ejected from the trap and detected at a q_z values lower than 0.908; the position of the 'hole' depends upon the amplitude of the frequency that is applied. This technique is called resonance ejection and is used to extend the mass range of the ion trap.

1.7.3.6. Ion traps as a collision cell - MS/MS and MS^n

Ion traps are tandem in time and allow not only for MS/MS experimentation, but for successive MS steps creating MS^n capacity, thus allowing in-depth fragmentation of an analyte. MS^n involves the selection of a particular *m/z* value, so that that only ions of selected *m/z* are present within the trap. The ions are then fragmented by collision induced dissociation (CID) and the fragments ejected (or specific fragments isolated in the trap and subjected to further CID).

1.7.3.6.1. Ion isolation

To produce a spectrum with a fragmentation that is specific for ions of a particular *m/z* value, ions of a specific *m/z* value must first be isolated in the trap. Once an ion has been chosen, for example the ion at *m/z* 573 in figure 1.7.18a, ions of other *m/z* values in the trap are ejected by the application of a multi frequency resonance RF across the end caps (except at the frequency that corresponds to the selected ion). The isolated ion is then isolated in the trap.

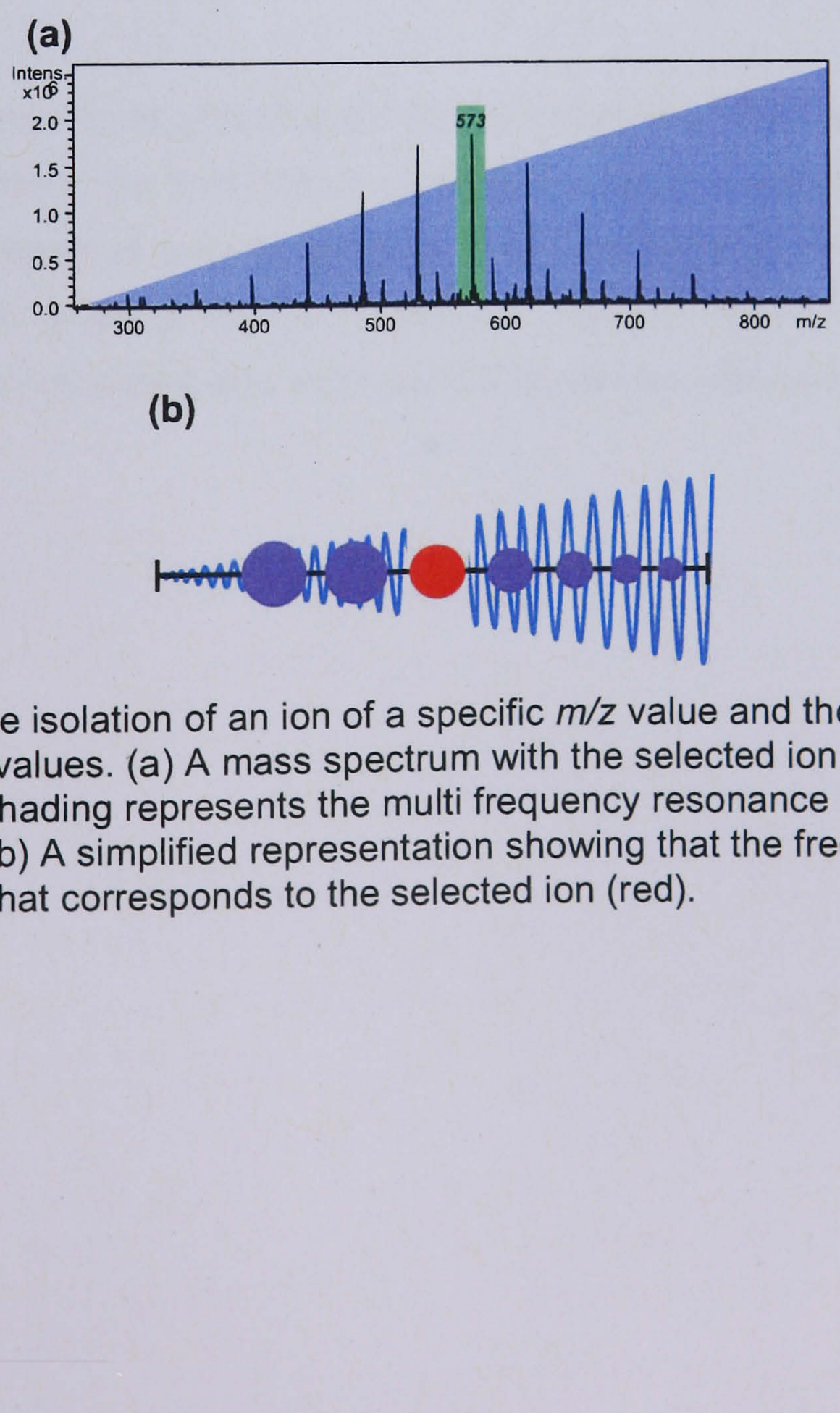


Figure 1.7.18. The isolation of an ion of a specific *m/z* value and the expulsion of ions of other *m/z* values. (a) A mass spectrum with the selected ion highlighted in green. The blue shading represents the multi frequency resonance RF that is applied to the end caps. (b) A simplified representation showing that the frequency is not applied to the q_u that corresponds to the selected ion (red).

1.7.3.6.2. Collision induced dissociation

Once ions of a particular *m/z* have been isolated in the trap, an RF waveform that is specific to the isolated ions is applied to the end caps (lower RF than causes ejection to occur); this is sometimes referred to as a 'tickle voltage'. The ions begin to oscillate in the trap and gain kinetic energy, the ions then collide with the helium gas causing some of the kinetic energy to be converted into internal energy, inducing fragmentation by CID. After CID has occurred, the precursor ion and its fragments are focussed into the centre of the trap before they are sequentially ejected and detected.

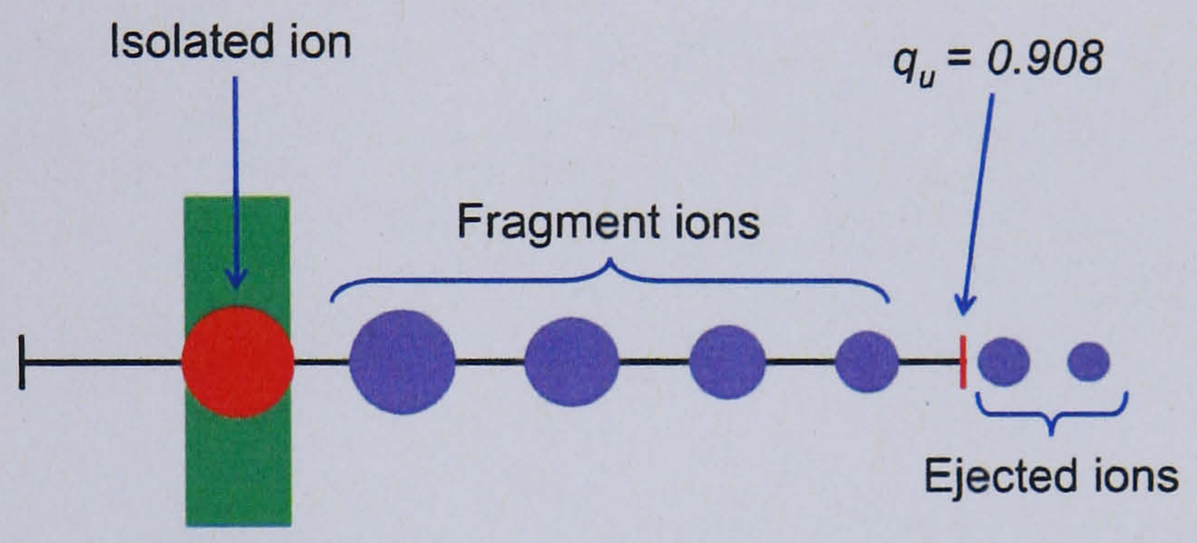


Figure 1.7.19. Low mass cut off. Fragment ions of 25 % or less of the isolated ions *m/z* value are ejected and not detected.

One limitation caused by fragmenting an ion is that not all of the fragment ions generated are stable in the q_u direction, causing them to be expelled from the trap during the CID process. In order to effectively fragment an ion in a trap it has to be set at a q_u of ~0.25, at which the 'low mass cut off' typically results in ions of less than 25 % of the initial parent ions *m/z* value to be ejected without detection (figure 1.7.19).

1.7.3.7. Time of flight analyser

Initially, Time of Flight (ToF) mass spectrometers were linear and first described by Stephens in 1946 (Stephens, 1946). The resolution and accuracy of the first instruments was low, but the advent of delayed extraction and reflectron increased both the resolution and mass accuracy greatly.

1.7.3.7.1. Linear ToF

A linear ToF is essentially a long tube under vacuum (figure 1.7.20). Ions are accelerated from the source by an electric field into a field free region; the ions traverse the field free region and are detected. A pulsed ion source such as MALDI is well suited to use with a ToF analyser as it produces discrete packets of ions; the starting time of the ions acceleration into the field free region can easily be recorded.

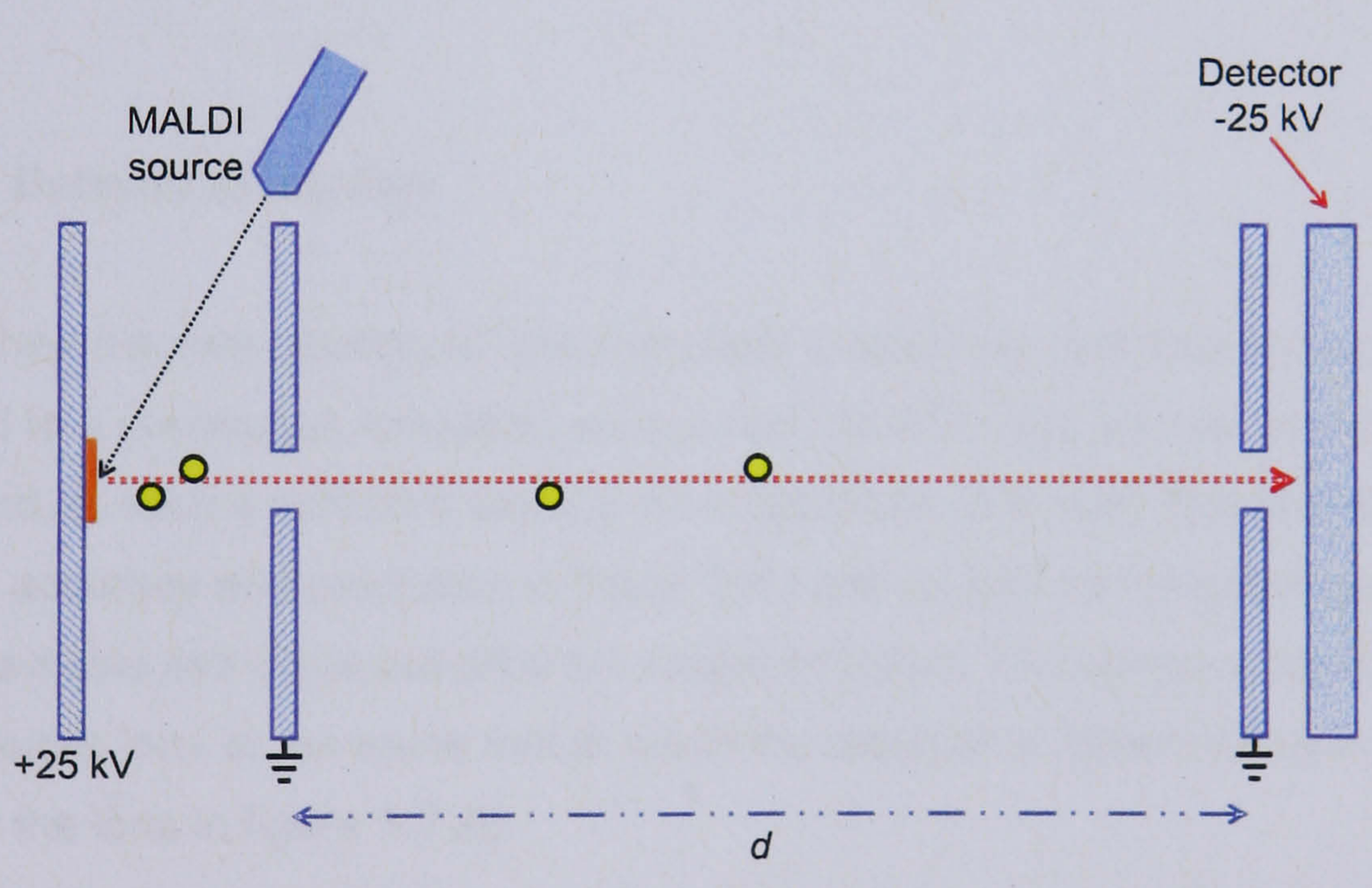


Figure 1.7.20. A schematic of a linear ToF. Ions are generated and accelerated into a field free region, d, where the time taken for ions to traverse this region is recorded and their mass determined.

All ions of same charge gain the same kinetic energy. The velocity depends upon the mass, equation 1.7.16. Ions of a different mass therefore separate in space and time in a ToF analyser. If an ion's 'time of flight' can be recorded, then its mass can be determined. An ion leaves the source with a kinetic energy (E_k) described by equation 1.7.16 (where m = mass, q = total charge (q = ze), v = velocity, V = applied potential):

$$E_k = qV = \frac{mv^2}{2}$$ **Equation 1.7.16.**

$$t = \frac{d}{v}$$ **Equation 1.7.17.**

$$t^2 = \frac{m}{z}\left(\frac{d^2}{2V_s e}\right)$$ **Equation 1.7.18.**

Equation 1.7.18 is a combination of equations 1.7.16 and 1.7.17, and shows that measuring the time that it takes a particular ion to traverse the field free region allows its *m/z* value to be determined. The time taken to record a full spectrum, even at low resolution, was enough to ensure ToFs found wide application in research and were therefore developed to increase the resolution and accuracy that could be obtained.

1.7.3.7.2. Delayed extraction

MALDI forms discrete packets of ions that allow a definitive start time to be recorded compared to a continuous ionisation source such as ESI; ions are continuously formed and as such a definitive starting time cannot be recorded. The problems of low mass accuracy and resolution in linear ToFs are caused by the generation of ions of the same *m/z* value but different kinetic energies. The spread of kinetic energy causes ions of the same *m/z* to reach the detector at different times, as shown by the ions in figure 1.7.20.

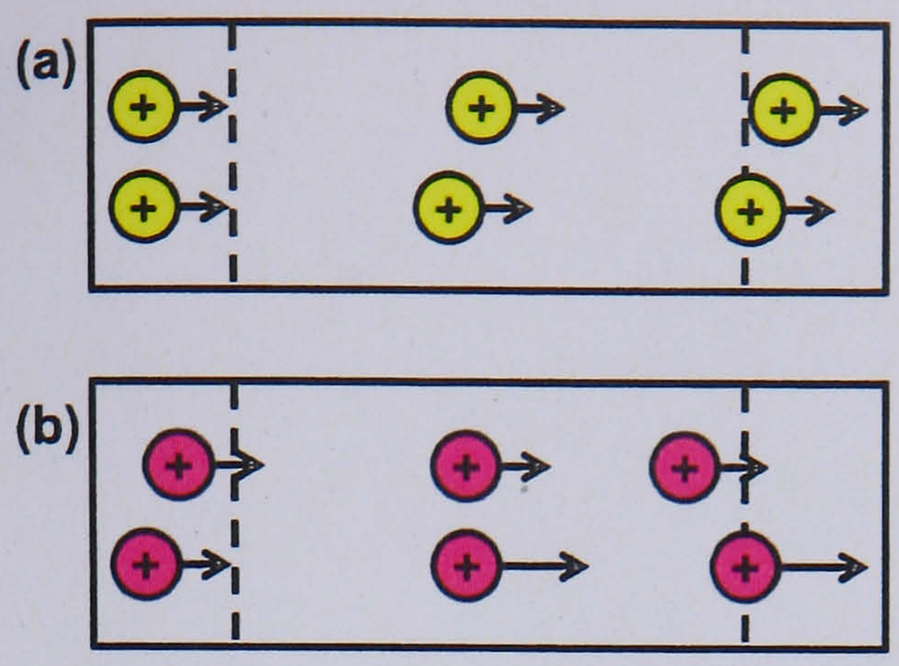


Figure 1.7.21. Factors that reduce mass accuracy and resolution in ToFs: (a) temporal distribution, ions are formed at different times, (b) spatial distribution, ions are formed at different locations, (c) kinetic energy distribution, ions gain different amounts of kinetic energy.

Temporal distribution (figure 1.7.21a) involves the formation of ions at different times during the ablation process in MALDI; the ions arrive at the detector at slightly different times because of their different time of formation. The formation of ions at different locations either within the matrix or plume gives rise to a spatial distribution (figure 1.7.21b); ions that remain in the accelerating field (between the grids) for longer periods of time gain more kinetic energy than ions that spend less time in the accelerating field. The range of kinetic energies that is imparted to the ions (figure 1.7.21c) causes their arrival at the detector at different times; a combination of all three distributions causes a decrease in mass accuracy and resolution.

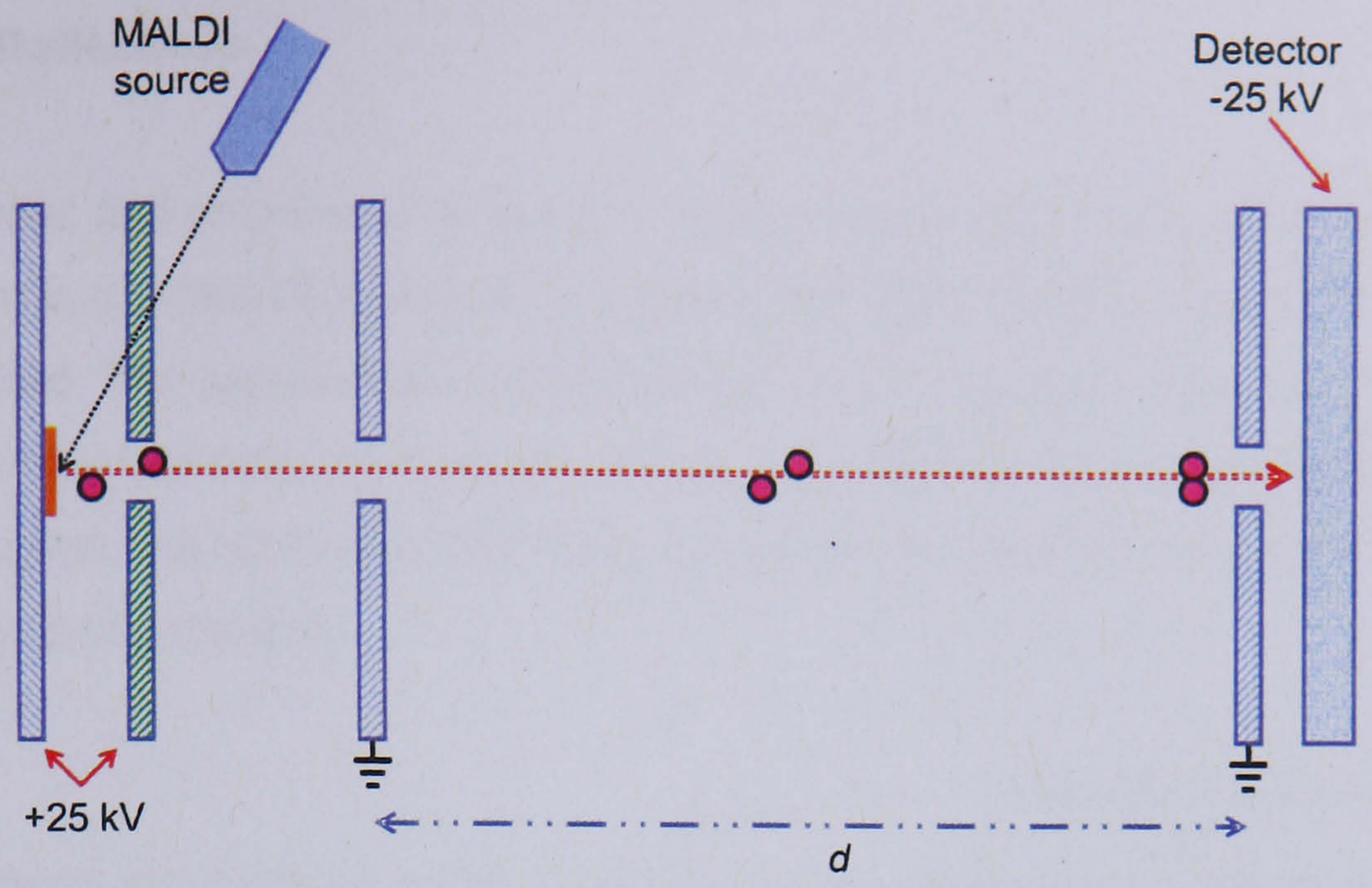


Figure 1.7.22. A linear ToF with a DE source. The two acceleration grids are initially at ground, the ions expand into a field free region before the acceleration potential is applied to the ions.

Delayed extraction (DE) was developed to try and correct for the kinetic energy spread and temporal distribution of ions with the same *m/z* value. Linear ToFs were modified (figure 1.7.22) to include a DE source. The ions initially expand into a field free region (the grids are temporarily held at earth); ions of greater kinetic energy move further towards the detector than those with less. After a short delay, the extraction potential is applied (to two grids); ions with a lower kinetic energy remain closer to the grids and are subsequently accelerated more than the ions with higher kinetic energy that are further away from the grids which are accelerated to a lesser degree. The DE of ions has the effect of correcting for the initial kinetic energy spread by focussing ions of the same *m/z* onto the detector, increasing the resolution and mass accuracy.

1.7.3.7.3. Reflectrons

Mamyrin *et al.* first proposed the use of an electrostatic reflector or reflectron in 1973. A reflectron is a series of grids and ring electrodes that create an ion mirror or retarding field. The purpose of a reflectron is to correct for any initial energy dispersion that ions may have; as the energy dispersion is corrected for and the flight path is doubled, the resolution and mass accuracy are improved but at the cost of mass range and sensitivity.

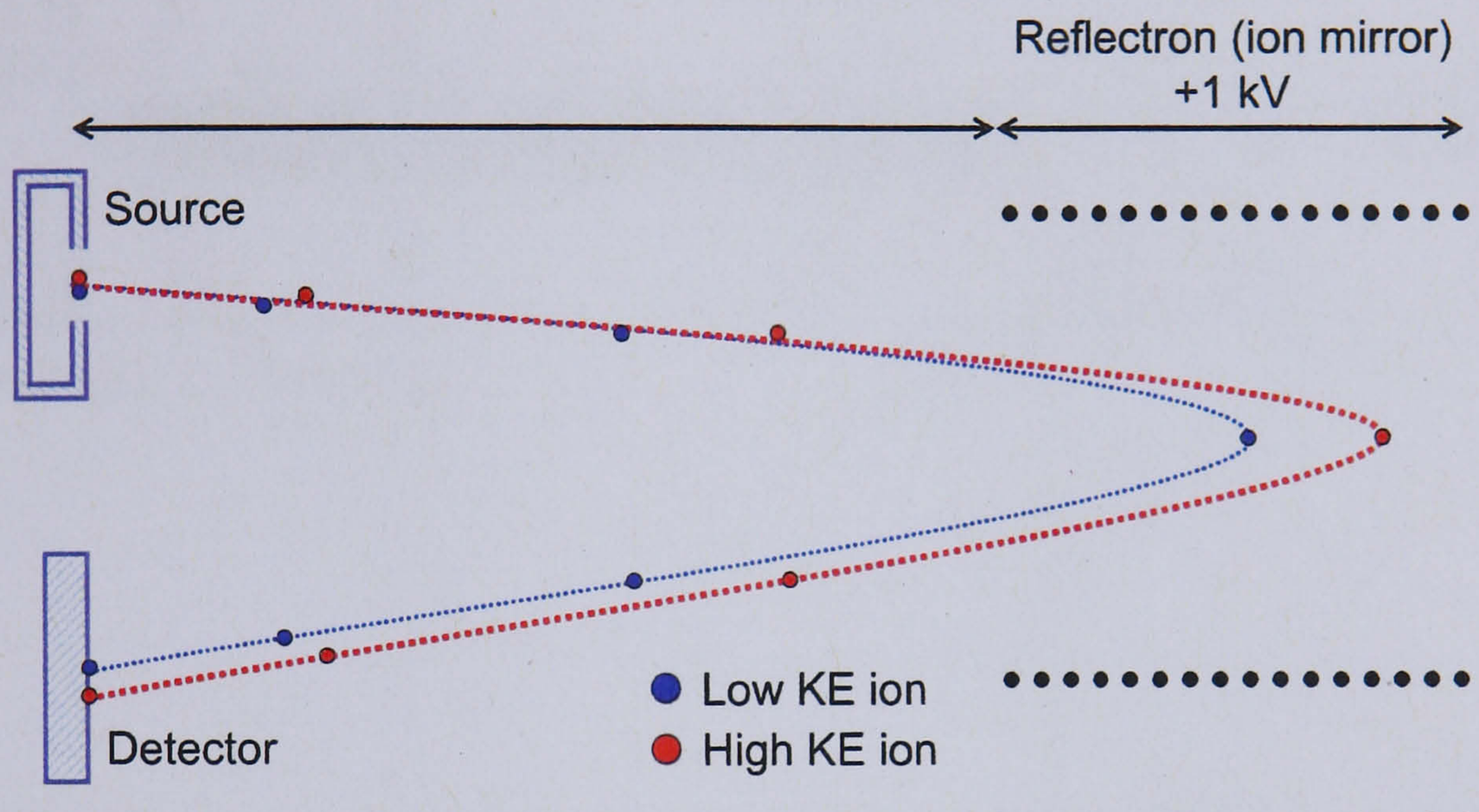


Figure 1.7.23. A schematic of a ToF analyser with a reflectron. The length of the ion path is doubled due to the ion's paths being inverted. KE = kinetic energy.

Ions with high kinetic energy penetrate the reflectron more than an ion with low kinetic energy (figure 1.7.23). The ions do not gain or lose any kinetic energy as a result of the reflectron, it merely corrects for the difference in energy. Ions of high kinetic energy spend less time in the field free region and more in the reflectron, conversely low kinetic energy ions spend more time in the field free region and less in the reflectron; the result is that ions of the same *m/z* but different kinetic energies arrive at the detector at the same time.

The combination of delayed extraction and reflectrons in some ToFs has been used in order to gain greater resolution. The reflectron can only account for the kinetic energy spread of ions; DE can correct for spatial and temporal distribution, causing ions affected by these factors to be focussed, allowing an increase in the resolution as the ions reach the detector with less spread in time.

1.7.3.7.4. Orthogonal acceleration ToF (Applied Biosystems QStar Pulsar *i*)

ToFs require discrete packets of ions, so a continuous ionisation source such as ESI requires the continuous ion beam to be pulsed to create discrete packets of ions. One way to create a pulsed ion beam is to effectively store the ions before creating a pulse that allows the stored ions to enter the ToF, or alternatively to pulse the ion beam into the ToF which is orthogonal to the ion beam (figure 1.7.24).

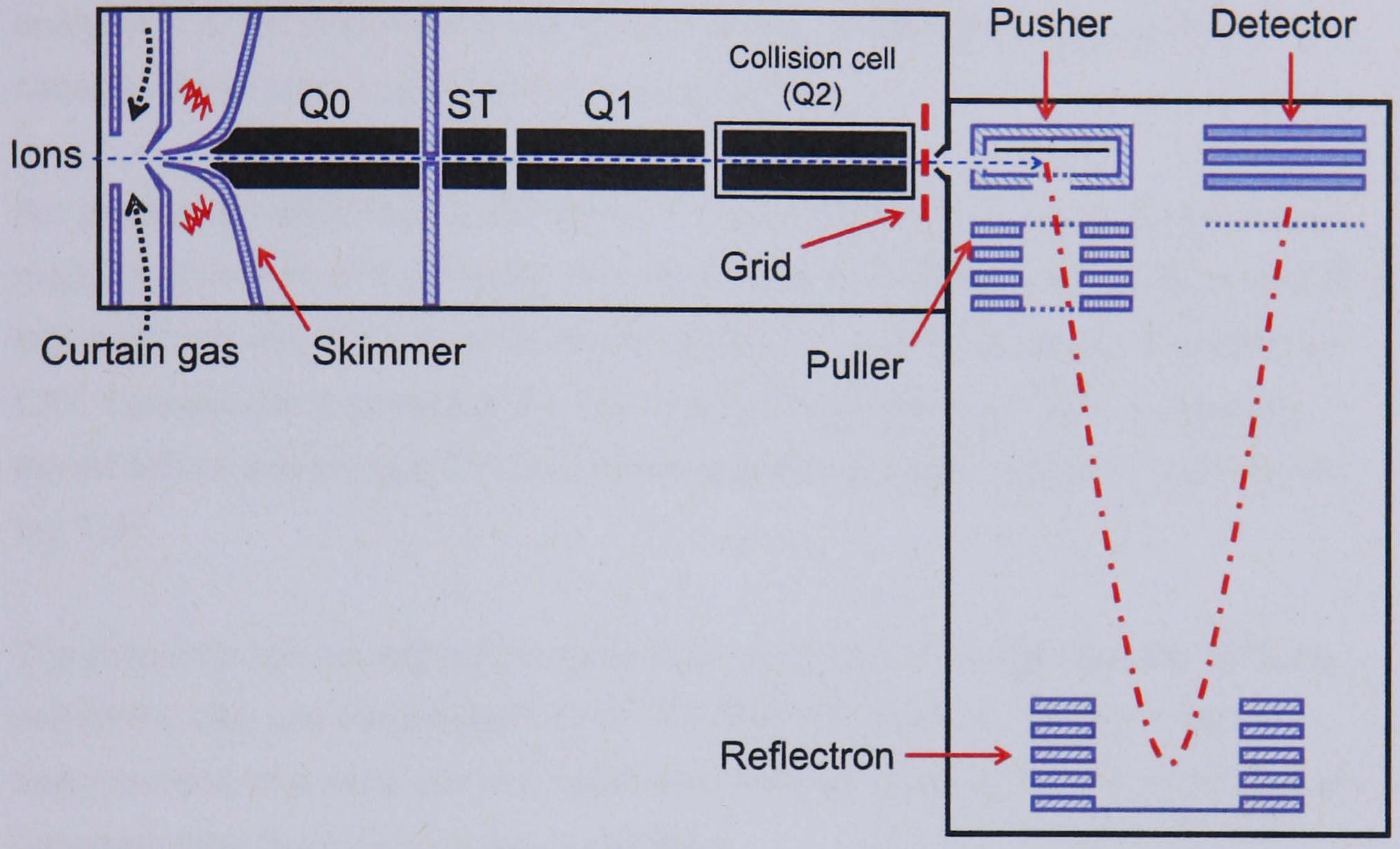


Figure 1.7.24. A schematic of the Applied Biosystems QStar Pulsar *i*. ST = Stubbies, a small quadrupole.

Ions enter the orthogonal acceleration ToF (oa-ToF) where they are focussed by an RF only quadrupole, Q0 which transmits all ions (figure 1.7.24); the ions are then transferred to the high vacuum area of the MS. The voltage that is applied to Q0 and ST is held at a constant fraction of the RF voltage that is applied to Q1, the mass filter quadrupole which has no DC potential applied for MS. The collision cell contains a collision gas at low pressures; the cell is operated in RF mode, which is stepped over several ranges to transmit a wide range of m/z values to the ToF. The ions enter the ToF where they are orthogonally 'pushed' by an accelerating voltage, creating a discrete packet of ions, which allows a starting time for the time of flight to be recorded. The ions enter the field free region where they are separated according to

their *m/z* value before being detected; ions with a large *m/z* have a longer flight time than ions with a low *m/z* value.

1.7.3.7.4.1. MS/MS

oa-ToF is tandem in space as it has two mass analysers that are separated by a collision cell. The first mass analyser is a quadrupole that operates in RF only mode for MS, but uses a DC potential to allow mass discrimination. The second mass analyser is a ToF and is not a scanning analyser, which means that oa-ToF is only capable of precursor and product ion experiments.

For product ion experiments, Q1 (figure 1.7.24) is operated in mass discrimination mode to allow ions of a particular *m/z* value to be transmitted to the collision cell. The ions enter the collision cell where they collide with a collision gas and fragment by CID; if a potential is applied to the exit lens of the collision cell then the ions are stored before being passed to the pusher where they are orthogonally pushed into the ToF.

The precursor ion experiment involves Q1 scanning ions through into the collision cell where they are fragmented using CID. The ToF is set to only give a signal if fragment ions of a particular *m/z* value are detected. If the correct fragment ions are detected, then the precursor ion is identified.

1.7.3.8. MALDI-ToF/ToF (Applied Biosystems 4700 proteomics analyser)

The Applied Biosystems 4700 proteomics analyser is a MALDI linear tandem ToF mass spectrometer, which can perform high energy CID MS/MS experiments. The MALDI source utilises a Nd:YAG solid state laser, with a repetition rate of 200 Hz, and is equipped with delayed extraction optics. For MS, ions are extracted with a high potential (ca. 20 kV) and pass through the MS1 region, the collision cell and the MS2 region (figure 1.7.25); the ions' time of flight are calculated when they reach either the linear or reflectron detectors, depending upon which mode is chosen.

For tandem MS experiments, the ions are subjected to DE and then accelerated using a high potential into MS1, where the timed ion selector is used to select a precursor ion. The timed ion selector is a dual stage electrode, where the first

electrode is switched on, retarding all ions; as the time of flight that corresponds to the precursor ion being selected occurs, the voltage is dropped allowing the transmission of the precursor ions before the second electrode is turned on to repel any subsequent ion transmission. The retarding lens decelerates the transmitted precursor ion to an energy of ~1 keV where it enters the collision cell, which now contains a collision gas. After the precursor ion has been subjected to CID, the fragment ions are re-accelerated in the second source; pulsed ion acceleration provides a start point for ToF measurement of the fragment ions.

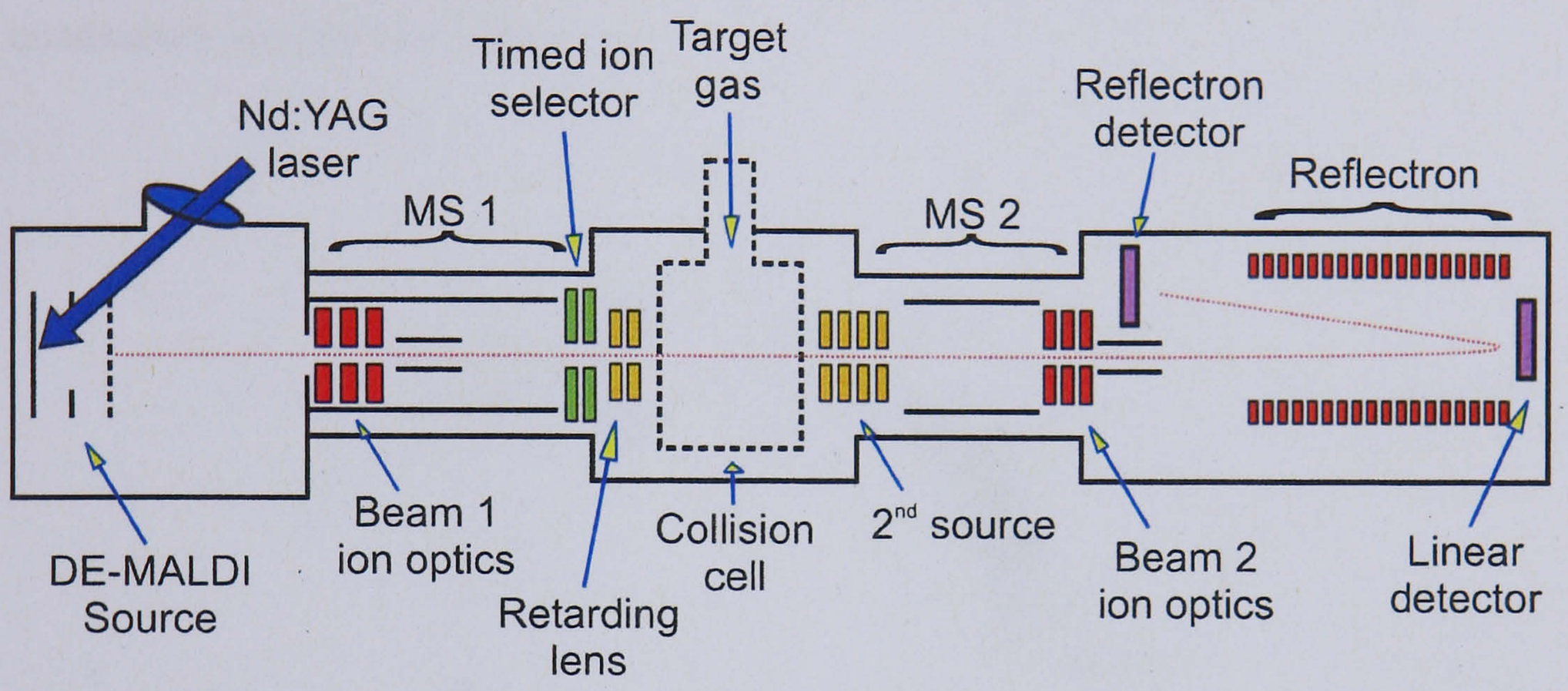


Figure 1.7.25. A schematic of an Applied Biosystems 4700 proteomics analyser with ToF/ToF optics.

1.7.3.9. Detectors

Given that ions can be separated according to their *m/z* value by mass analysers, they need to be detected for any structural information to be obtained. Initially, ions used to be detected directly by a photographic plate or a Faraday cup. Ions of the same *m/z* would reach the plate at the same place (the intensity of the spot would be proportional to the intensity); a scale could be applied that would determine the *m/z* of the spots. Faraday cages cause an ion to discharge upon a collision; the current of the discharge is amplified and then detected. Modern mass spectrometers use electron multipliers, photomultipliers or array detectors such as the microchannel plate.

1.7.3.9.1. Electron multiplier

Positive or negative ions are attracted towards a conversion dynode where they impact causing the emission of secondary particles (figure 1.7.26). The secondary particles (electrons) are accelerated into a horn shaped device where they collide with a cathode, releasing more electrons which go on to collide again and again with further cathodes, causing a cascade of electrons. The cascade of electrons travels towards ground where their current can be measured (it is proportional to intensity). Typical amplification can reach 10^7; electron multipliers are commonly used for quadrupole and ion trap instruments.

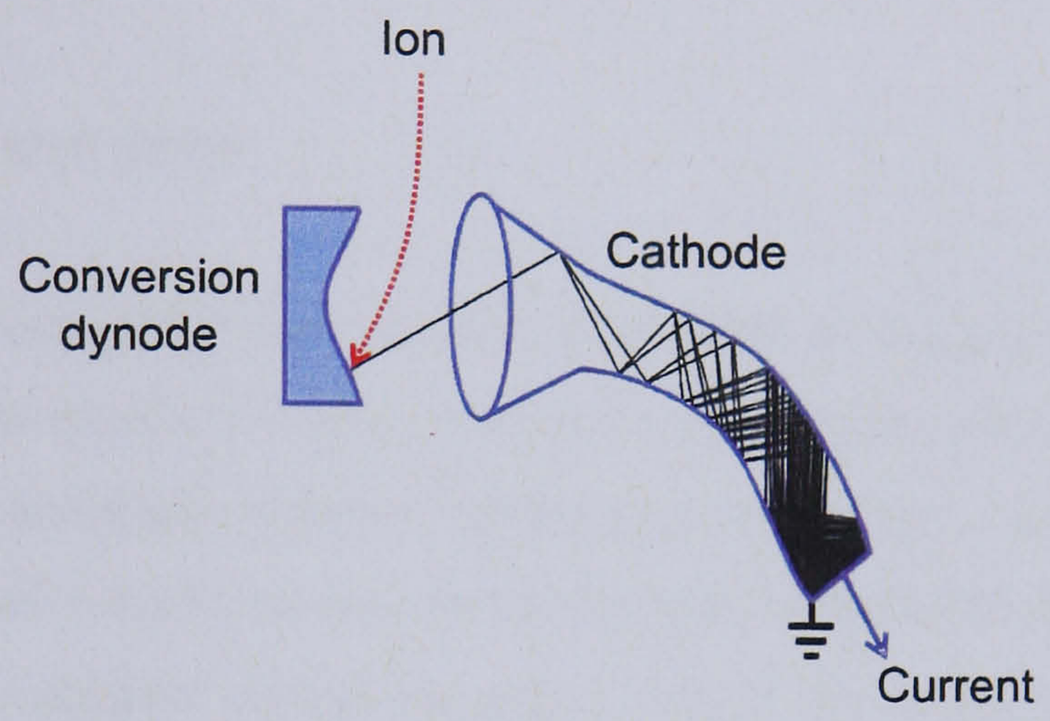


Figure 1.7.26. A schematic of an electron multiplier.

1.7.3.9.2. Photomultiplier

Ions are converted into electrons by their collision with a conversion dynode, as in an electron multiplier. The secondary particles are accelerated towards a phosphorescent screen, from which photons are emitted on impact of the electrons (figure 1.7.27). The photons are detected by a photomultiplier. The lifetime of a photomultiplier is longer than that of an electron multiplier as it is sealed in glass, preventing contamination; however the amplification is less, at 10^4 to 10^5.

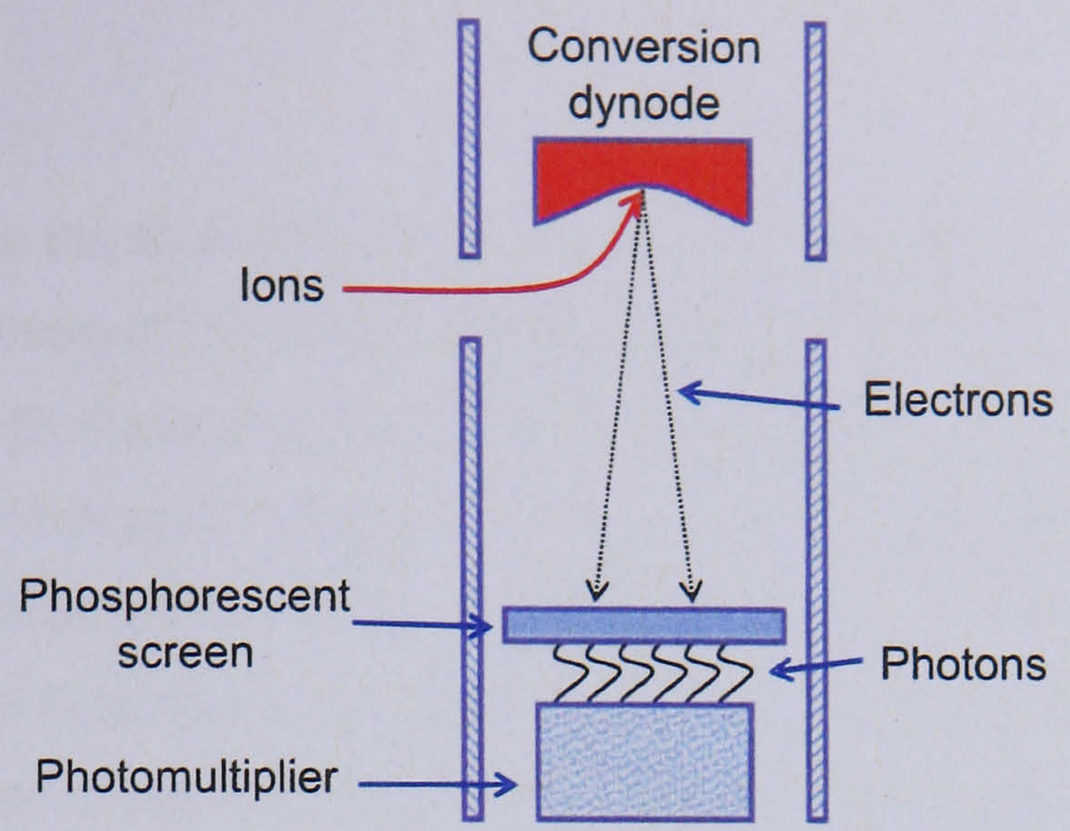


Figure 1.7.27. A schematic of a photomultiplier.

1.7.3.9.3. Microchannel plate

The microchannel Plate (MCP) is a plate that consists of many small (4-25 μm diameter) glass tubes (figure 1.7.28a), that have a geometry (figure 1.7.28b) which allows a cascade of electrons to occur (analogous to electron multipliers). Both faces of the disk are coated in metal so that each channel is electronically connected, with a potential difference applied across the plate. Ions of the same *m/z* value can arrive at a different location on the plate but can still all be detected at the same time with a typical amplification of 10^6. The connection of plates in series (figure 1.7.28 c-d) can increase the amplification up to 10^8. MCPs are commonly used in ToFs as they allow the simultaneous detection of ions over space.

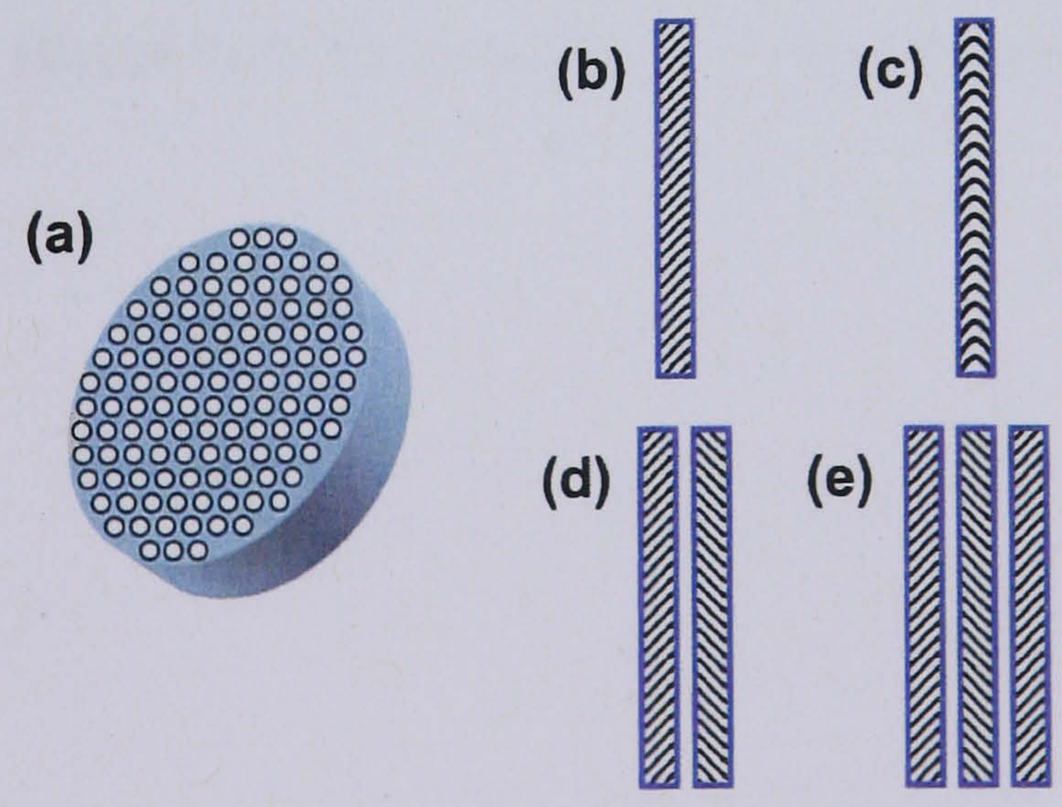


Figure 1.7.28. (a) The microchannel plate detector. (b) Continuous dynode electron multiplier geometry. (c-d) The connection of MCPs together to increase the amplification of ion current.

1.8. Aims

This thesis describes the application of LC-MS for the analysis of human urine samples. The objectives of the research described in this thesis were to critically assess current LC-MS metabonomic methodologies alongside a new separation method, hydrophilic interaction chromatography (HILIC), in an attempt to increase the coverage of metabolites within human urine samples. A further objective was to analyse an extract of *Pseudomonas chlororaphis* PCL 1391 using the same analytical methods as a metabonomic study, but in a completely different manner.

Chapter Three thus discusses and evaluates the various analytical platforms available for metabonomic studies, comparing and contrasting different separation methods, ionisation methods and detection methods. It describes considerations involved in sample collection, storage and manipulation, as well as subsequent extraction of raw LC-MS data and the related statistical analysis of these data and their assessment and development. The main focus of the work described in Chapter Three is the development of HILIC as a complementary orthogonal separation method to reversed phase for an increased coverage of urinary compounds, and a proposal of the required elements of a robust LC-MS 'metabonomic toolbox'.

Chapter Four uses the methods discussed and developed in Chapter Three for the analysis of clinical urine samples obtained from patients who had suffered a fracture.

Chapter Five discusses the structural determination of a *Pseudomonad* biosurfactant using ESI-MS(MS), MALDI-ToF/ToF and racemic amino acid analysis.

Chapter Two

Experimental methods

2.1. Urine collection

2.1.1. Samples collected from volunteers within the Department of Chemistry, University of York, UK

An e-mail was sent to all academic, administrative and Ph.D. student members of the Department of Chemistry, University of York, UK, asking for volunteers to donate two urine samples. All potential volunteers were made aware of the purpose for collecting their urine samples (for the development of a 'metabonomic toolbox'), and were guaranteed anonymity. They were informed that they would have to collect the first void of the day, and any subsequent void after 15:00 on the same day, with a preference for 'mid-stream' urine. Volunteers anonymously collected sealed bags containing two randomly labelled sample tubes (Bibby Sterilin), further instructions and two further sealable bags in which to place their filled sample tubes. The only information that was requested was gender, age (age groups were acceptable) and if they were a smoker; no restrictions were placed upon volunteers diet or lifestyle. Volunteers were asked to place their filled sample tubes into one of three boxes spread across the department within two hours of donation; all samples were transferred to -80 °C storage within two hours of donation, indicated by the majority of samples still being warm upon collection of the boxes. A total of 62 samples were donated, 39 from males and 23 from females. All data relating to the samples is presented in appendix A.

2.1.2. Clinical urine sample collection

Full ethical approval was obtained for the collection of urine samples from patients suffering bone fractures. A registrar orthopaedic surgeon collected clinical urine samples from patients suffering a fracture who were admitted to York District Hospital NHS Trust's accident and emergency department between October 2004 and February 2005. The inclusion criteria was that patients were between the ages of 18-45 to reduce any chance of pathological fractures or incomplete skeletal development; initially only long bone fractures were considered, but due to the lack of long bone fractures, the study was opened up to wrist and ankle fractures. Further exclusions were patients who had suffered multiple injuries, had malignancy, head injuries, spine/foot/hand fractures, pregnant or nursing mothers and any unconscious patients. A total of 61 patients were deemed suitable for inclusion into the study (45

males and 16 females); of these, 11 declined consent stating either lack of interest or a needle phobia[1], with a further two patients withdrawing consent at a later date, and one being transferred into the care of another health trust. This left a total of 48 patients (36 males and 12 females, with an age range of 19 to 47 years old; average age = 29.5, standard deviation = 8.2), who each donated between one and four urine samples, ranging from a period of t = 0 (time of fracture) to 133 days (19 weeks) after the initial fracture (average = 6 weeks).

A total of 12 different fracture types were included in the study, with the largest sample cohort being ankle fractures (51 urine samples). All obtained data relating to the clinical urine samples are presented within appendix B.

2.2. Sample storage

2.2.1. Samples collected from volunteers within the Department of Chemistry, University of York, UK

Upon collection, donated urine samples were stored at -80 ºC for a period of four weeks before being subject to further manipulations (section 2.3), and stored at -80 ºC after these manipulations.

2.2.2. Clinical urine samples

Clinical urine samples were stored at -80 ºC at Smith & Nephew, York Science Park, UK. Samples were aliquotted (section 2.3) and stored at -80 ºC until transport to the Department of Chemistry on dry ice for analysis.

2.3. Sample manipulations

All urine samples were defrosted at room temperature before being aliquotted into microcentrifuge vials (Sarstedt) and re-frozen at -80 ºC prior to any further sample preparation and analysis. The samples collected from within the department were

[1] Serum samples were also collected for a parallel study by Smith & Nephew into serum markers related to fracture repair; any patients with needle phobia were excluded from the trial.

used to create a pooled reference sample by transferring 1 mL from each sample into a beaker before mixing and aliquotting this pooled reference sample.

2.3.1. Samples separated using RP

For analysis using RP-LC-MS, all urine samples were defrosted at room temperature before being centrifuged at 10,186 x g for 8 min (GenFuge 24D, Progen). The supernatant was collected and passed through a 0.45 µm PVDF syringe filter (VWR) into sample vials fitted with 250 µL deactivated glass inserts (Agilent Technologies), ready for analysis.

2.3.2. Samples separated using HILIC

For analysis using HILIC-MS, all urine samples were defrosted at room temperature before being mixed in a 1:1 ratio with MeCN (Fisher Scientific). The samples were then centrifuged at 10,186 x g for 8 min (GenFuge 24D, Progen). The supernatant was collected and passed through a 0.45 µm PVDF syringe filter (VWR) into sample vials fitted with 250 µL deactivated glass inserts (Agilent Technologies), ready for analysis.

2.3.3. Sample re-analysis using RP (clinical samples)

To precipitate and remove the protein present in the clinical urine samples prior to RP-LC-MS analysis, the samples were treated as described in section 2.3.2.

2.4. HPLC separations

2.4.1. RP separation

Urine samples were separated on a 100 x 4.6 mm Chromolith RP18e column (Merck), along with a guard column (5 x 4.6 mm), on an an Agilent 1100 LC (Agilent Technologies). Mobile phase A was 0.1% (v/v) formic acid (Fisher Scientific), while mobile phase B was MeCN (Fisher Scientific) modified by the addition of 0.1% (v/v) formic acid (Fisher scientific). The gradient started with 5% mobile phase B,

increasing to 20% B at 9 min, and then to 95% B at 21 min. The mobile phase was held isocratic for 3 min before returning to the starting conditions within 3 min (total run time was 30 min). The injection volume was 20 µL, and the column was eluted at a flow rate of 600 µL min^{-1}.

2.4.2. HILIC separations

Urine samples were separated using a 3.5 µm, 100 x 4.6 mm ZIC-HILIC column (SeQuant), along with a guard column (20 x 2.1 mm), on an an Agilent 1100 LC (Agilent Technologies). Three gradients were developed, with gradient three (section 2.4.2.3) being used for all HILIC separations in this thesis.

2.4.2.1. Gradient 1

Mobile phase A consisted of 0.1 % (v/v) formic acid (pH 4, Fisher Scientific), while mobile phase B consisted of MeCN (Fisher Scientific) modified by the addition of 0.1 % (v/v) formic acid (Fisher Scientific). The gradient started with 5 % mobile phase A increasing linearly to 20 % over a period of 9 min, with a further increase to 95 % over 12 min before being held isocratic for 3 min before returning to the starting conditions within 3 min. The mobile phase was kept at 5 % A for the remaining time to allow equilibration (total run time was 30 min). The injection volume was 20 µL, and the column eluted at a flow rate of 600 µL min^{-1}.

2.4.2.2. Gradient 2

Mobile phase A consisted of 0.1 % (v/v) formic acid (pH 4, Fisher Scientific), while mobile phase B consisted of MeCN (Fisher Scientific) modified by the addition of 0.1 % (v/v) formic acid (Fisher Scientific). The gradient started with 5 % mobile phase A increasing linearly to 95 % over a period of 15 min. The mobile phase was held isocratic for 4 min before returning to the starting conditions within 30 s. The mobile phase was kept at 5 % A for the remaining time to allow equilibration (total run time was 30 min). The injection volume was 20 µL, and the column eluted at a flow rate of 600 µL min^{-1}.

2.4.2.3. Gradient 3

Gradient 3 used the same parameters as shown in section 2.4.2.2 with the only difference being the addition of 5 mM ammonium acetate (Fisher Scientific) to mobile phase A.

2.5. LC-MS(MS) analysis

2.5.1. ESI parameters

A TurbolonSpray source (Applied Biosystems) was used for all ESI analyses. The LC outlet from an Agilent 1100 series HPLC was directly coupled with no splitting. The capillary voltage was held at ±2500 V depending upon the ionisation mode; N_2 nebulising gas, 3.3 L min^{-1}; and N_2 drying gas, 6.0 L min^{-1} at 300 °C.

2.5.2. APCI parameters

The LC outlet from an Agilent 1100 series HPLC was directly coupled to an APCI source (Applied Biosystems) with no splitting. The nebulising current was held at ±2 (arbitrary units) depending upon the ionisation mode; N_2 nebulising gas, 3.8 L min^{-1}; and N_2 drying gas, 1.5 L min^{-1} at 425 °C.

2.5.3. MS parameters

ESI and APCI Q-o-ToF MS experiments were performed using an Applied Biosystems QStar pulsar *I* quadrupole orthogonal time of flight tandem MS. The MS was operated in full scan mode with *m/z* range of 40-1000 using the following parameters (depending upon the ionisation polarity being used): focussing potential = ±145 V; declustering potential = ±45 V; declustering potential 2 = ±15 V; quadrupole 2 gas pressure = 2 (arbitrary units). For positive ionisation mode the following additional parameters were used: mirror = +985 V; liner = -400 V; plate = +330 V; grid = -400 V; offset = -11.3 V, and for negative ionisation mode: mirror = -985 V; liner = +400 V; plate = -330 V; grid = +410 V; offset = -25.4 V. Data were recorded

using the Analyst QS v1.1 software (Applied Biosystems).

2.5.4. MS/MS parameters

All CID tandem MS experiments used the same conditions as those in section 2.5.3 (except the quadrupole 2 gas which was set at 5 (arbitrary units)). In addition to these conditions, the independent data acquisition setting was used with the following settings: the four most intense peaks were selected for CID at collision energies of 20 and 25 (arbitrary units), with dynamic exclusion set to 120 s to prevent the re-analysis of precursor ions already selected for CID. An 'include list' was used to include any ions that should be automatically selected for CID should they appear in an MS survey 'scan'.

2.6. Data extraction and normalisation

2.6.1. Data extraction

Raw LC-MS data were exported using the metabolomics export script (Applied Biosystems). Peaks files were created prior to generating a 3D data matrix of *m/z* and t_R versus intensity for each sample analysed. The following settings were used: t_R tolerance, 0.5 min; LC peak width (min/max), 0.1/10 min; intensity threshold, 10 counts s^{-1} for positive ionisation mode and 1 count s^{-1} for negative ionisation mode; mass accuracy, 200 ppm; maximum peak number, 5000. The data were exported as a text file (ASCII format), ready for import into Excel (Microsoft) for further data manipulation and the addition of sample information.

2.6.2. Normalisation

Extracted data were imported into Excel (Microsoft) and were either normalised to creatinine intensity or total ion count, depending upon the experiment. Normalisation to creatinine was performed by finding the most intense creatinine value, and dividing all other observations (samples) creatinine intensity by this value. The resulting scale factor for each observation was then used to multiply each variable within that observation.

Normalisation to total ion count was performed by summing the intensities of all variables for each observation (sample), with the resulting value for each observation being divided by the largest total ion count value. The resulting scale factor for each observation was then used to multiply each variable within that observation.

All resulting data were saved as a text file in the ASCII format.

2.7. Statistical analysis

All statistical analyses used the SIMCA-P+ statistical software versions 11 and 11.5 (Umetrics).

2.7.1. Data import

All datasets were imported as text files into SIMCA-P+, where they were transposed so that each column represented an observation. Any information such as sample name, sample data etc. were assigned one of the following formats: primary observation ID; secondary observation ID; X variable (all of the variable intensities), and Y variable (discriminatory variables). Once all formats were set, the resulting imported data were ready for statistical analysis.

2.7.2. Principal component analysis

For principle component analysis (PCA), all Y variables were excluded from the data and the resulting X matrix scaled using either mean centering, pareto scaling or unit variance. The number of principal components (PCs) developed was determined by R^2 and Q^2 values; these values relate to the explained variation and give an indication of the fit of the model and its predictive ability. Internal cross-validation (CV) was used to determine an optimal balance between fit and predictive ability and to determine the number of components used for each model. Data points in the resulting scores plot can be coloured according to any secondary observation IDs set (section 2.7.1).

2.7.3. Partial least squares

For partial least squares (PLS) analysis ~$^2/_3$ of any dataset were used for PLS model development (with the remaining ~$^1/_3$ being held back to form an external test set, section 2.7.4), with a Y variable being included to indicate class belonging. The X matrix was scaled using either mean centering, pareto scaling or unit variance. The number of latent variables (LVs) developed was determined using the R^2 and Q^2 values, as described in section 2.7.2. Variable importance for projection (VIP) scores were used to identify and assess unimportant variables that did not add any predictive ability to the developed model. Any unimportant variables were removed and the model rebuilt with the process repeated. Once a satisfactory PLS model has been built, the resulting scores plot can be coloured according to any secondary observation IDs or Y variables set (section 2.7.1).

2.7.4. External classification

To determine the 'true' predictive ability of any developed PLS model, the remaining ~$^1/_3$ of a dataset were imported as a secondary dataset and manipulated according to section 2.7.1, before being selected as a prediction dataset. The resulting predicted Y variable values were used to indicate the external classification rate based upon their closeness to their actual Y variable values.

2.8. Proteomics

2.8.1. Bradford assay

Each of the clinical urine samples were diluted by adding 5 µL to 995 µL of HPLC grade water (Fisher Scientific). Eight diluted bovine serum albumin standards (0, 50, 125, 250, 500, 750, 1000 and 1500 µg/mL) were created. 10 µL of each of the standards and diluted clinical urine samples were pipetted into 96-well microtitre plates (Corning Inc.). 200 µL of Coomassie brilliant blue dye (Sigma) was added to each of the standards or samples. A photometer microplate reader (Dionex Technologies) set to measure absorbance at 570 nm was used to determine protein concentration.

2.8.2. 1-D gel electrophoresis

For the separation of urinary proteins using 1-D SDS-PAGE analysis, 18 µL of diluted clinical urine samples (5 µL in 1000 µL water) and MW marker standards (Invitrogen) were added to 7.5 µL of NuPAGE buffer (Invitrogen) and 4.5 µL of 2-mercaptoethanol (Invitrogen), giving a total volume of 30 µL. These solutions were incubated at 75 °C for 10 mins before 20 µL of each sample were loaded onto a 1-D NuPAGE 4-12 % Bis-Tris 1 mm 10-well gel (Invitrogen). The gel was run at a constant voltage of 200 V for ca. 50 min. Once completed, the gel was washed in water for 20 min before being stained overnight using Coomassie brilliant blue dye (Sigma).

2.8.3. In-gel tryptic digestion

Protein bands were excised from the stained gel, chopped into smaller segments and placed into 0.5 mL microcentrifuge vials (Sarstedt). Each of the excised gel pieces were washed twice using 20 mM ammonium bicarbonate (Sigma) in MeCN (Fisher Scientific) for 20 min. The gel pieces were then washed for 5 min using MeCN, before being dried for 20 min in a SpeedVac (Savant) and then incubated at 65 °C for 1 h in 10 mM dithioerythritol (Sigma) in 100 mM ammonium bicarbonate. The gel pieces were then subsequently washed with 100 mM and then 25 mM ammonium bicarbonate for 15 min before being washed with MeCN for 5 min. The gel pieces were dried in a SpeedVac for 20 min before being digested overnight by incubating the gel pieces at 37 °C with 10 µL of 0.02 µg/µL trypsin (porcine trypsin, Promega) in 20 mM ammonium bicarbonate. The supernatant from each microcentrifuge vial was then extracted using C_{18} ZipTips (Millipore) before 0.5 µL was spotted onto MALDI plates, ready for analysis.

2.8.4. MALDI-ToF/ToF analysis

The 4700 proteomics analyzer (Applied Biosystems) was used to analyse the digested proteins from the clinical urine samples. To each of the spotted digested samples, 0.5 µL of α-cyano-4-hydroxycinnamic acid (Sigma) in 0.1 % TFA (Sigma)

were spotted and allowed to air dry. First, MS spectra were recorded for each sample spot to identify the ten most intense peaks for subsequent MS/MS analysis, using 1500 laser shots and accumulating the resulting data. MS/MS data were acquired using the default 1 kV MS/MS method, with a maximum of 2000 laser shots being allowed for each spectrum; air was used as the collision gas.

2.8.5. Protein identification by database searching

Data obtained from MS/MS analyses were submitted for database searching using an in-house MASCOT server (Matrix Science). GPS Explorer v3.6. (Applied Biosystems) was used to submit the data for database searching using the following parameters: MS/MS ion search; trypsin enzyme; monoisotopic mass values; unrestricted protein mass; 200 ppm mass tolerance; ±0.1 Da fragment mass tolerance. All searches were performed against the NCBInr protein sequence database (06 July 2007 build).

2.9. Lipopeptide analysis

2.9.1. Sample information

A HPLC fraction obtained from an ethyl acetate extract of *Pseudomonas chlororaphis* PCL 1391 spent growth medium was provided by the Department of Biology, University of Leiden, the Netherlands. The fraction was reconstituted in 300 µL of MeOH (Fisher Scientific) before being diluted to 25 % using MeOH modified by the addition of 0.1 % (v/v) formic acid (Fisher Scientific) for further analysis.

2.9.2. ESI-MS(MS) analysis

The same parameters were used as shown in section 2.5.3 and 2.5.4.

2.9.3. MALDI-ToF/ToF analysis

The same parameters were used as shown in section 2.8.4.

2.9.4. Chemical methods

To cleave the ester bond to create a linear lipopeptide, 10 µL of diluted sample was added to a 1:1 mixture of 35 % ammonia solution (Sigma) and MeOH, and left overnight at room temperature. The sample was reduced to dryness using a SpeedVac and reconstituted in MeOH with 0.1 % (v/v) formic acid.

The above procedure was also undertaken using a 1:1 mixture of 35 % ammonia solution and BuOH (Fisher Scientific).

2.9.5. Racemic amino acid analysis

Racemic amino acid analysis was performed with the assistance of Dr. Kirsty Penkman (Department of Chemistry, University of York). 10 µL of stock lipopeptide was added to 200 µL of 7 M HCl (Fisher Scientific), the vial flushed with N_2 gas, and then placed in an oven (Binder Ovens) for 24 h at 110 °C, to hydrolise the peptide bonds, releasing free amino acids into solution. The hydrolysed sample was reduced to dryness in a SpeedVac before being rehydrated in 40 µL 0.01 M HCl and 1.5 mM sodium azide (Sigma), containing the non-protein amino acid L-*homo*-Arg at a concentration of 0.01 mM.

The rehydrated solution was analysed using RP-HPLC (C_{18} HyperSil BDS column, 5 x 250 mm, Agilent 1100 series LC, Agilent Technologies) where 2 µl of sample was injected and mixed online with 2.2 µl of derivitising reagent (260 mM *N*-Iso-L-butyryl L-cysteine (Sigma), and 170 mM *o*-phthaldialdehyde (Sigma) in 1M potassium borate buffer (Sigma), adjusted to pH 10.4 with potassium hydroxide pellets). Mobile phase A consisted of 23 mM sodium acetate tri-hydrate, 1.5 mM sodium azide, 1.3 µM EDTA, adjusted to pH 6.00 with 10% acetic acid and sodium hydroxide (all Sigma), mobile phase C was methanol and mobile phase D was MeCN. Initially 95% A and 5% C was used at a flow rate of 0.56 ml min^{-1}, changing to 50% C and 2% D after 95 min. Fluorescence detection used a Xenon-arc flash lamp at a frequency of 55Hz, with excitation wavelength of 230nm and emission wavelength of 445nm.

Chapter Three

Development of a 'metabonomic toolbox'

The work presented in this section formed part of the following publication:

"Hydrophilic Interaction Chromatography for Mass Spectrometric Metabonomic Studies of Urine"

Simon Cubbon, Timothy Bradbury, Julie Wilson, and Jane Thomas-Oates

Analytical Chemistry, Volume 79, Number 23, pages 8911 – 8918.

3.1. Introduction

One of the earliest metabolite profiling experiments was performed by Pauling *et al.* in 1971 (Pauling *et al.*, 1971); they quantitatively studied around 250 substances in a sample of breath and 280 substances in a sample of urine vapour using gas-liquid chromatography. The field now known as metabonomics, defined by Nicholson *et al.* in 1999 (Nicholson *et al.*, 1999), is in its infancy today, and is very rapidly changing/developing direction. Many overlapping fields of medicine/biology, analytical science and statisticians are all bringing many different ideas to metabonomic studies; the literature is currently very patchy as some details are just not reported, so it cannot be found if the details were, or were not considered, or how they were undertaken. These problems are further confounded, as some considerations that are important to analytical scientists may be very different from those that are important to medics (e.g. replicates, appropriate controls, sample storage and analytical conditions). The metabolomics standards initiative (MSI) steering group (Sansone *et al.*, 2007) seeks to standardise most aspects of metabolomic (metabonomic) experiments. However, as the MSI has only just published some draft guidelines (Metabolomics volume 3, 2007), the existing literature studies are likely to be non-standard.

It is therefore of great importance that all aspects of a metabonomic study are carefully considered and controlled to stand any chance of obtaining results that are both robust and informative. The main goal of a metabonomic study is to obtain as much information as possible. This may encompass the use of several different analytical platforms, as well as considering and employing methods specific to each analytical platform. The only fixed aspect of the study which is the subject of this PhD. was the use of HPLC-MS as the analytical platform, hence the need to develop LC methods that allow for the separation, and subsequent detection, of as broad a range of compounds within human urine samples as is possible.

3.1.1. Aims

The purpose of the work described in this chapter was to develop a robust LC-MS platform for the analysis of human urine samples, as well as the full consideration and development of all components of every step of a metabonomic study, as shown in figure 3.1.1.

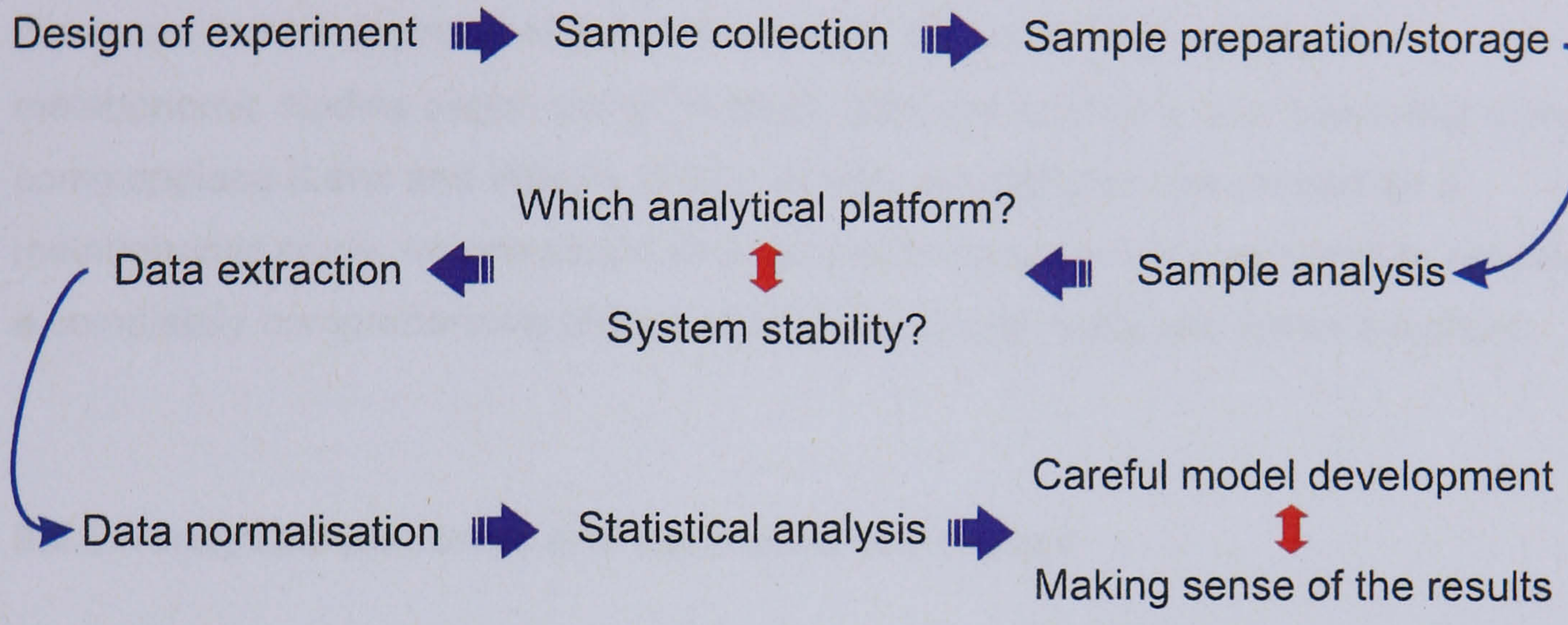


Figure 3.1.1. Schematic representation of the steps involved in a metabonomic study.

As the analytical platform was fixed (LC-MS was chosen as the platform for this study), the methods that were available at the start of this work, along with the problems associated with LC-MS are first highlighted. This is followed by the results of my in-depth study of all of the aspects involved in a metabonomic study (figure 3.1.1).

Considering all of the points of a metabonomic study should allow the development of a robust LC-MS method (or a metabonomic toolbox) that can be effectively used within a broader experimental protocol. To increase the coverage of the metabolite content within human urine, a hydrophilic interaction liquid chromatography method (HILIC) was developed and compared to the performance of a traditional reversed phase (RP) approach.

3.2. Analytical Platform Considerations

3.2.1. Introduction

The analytical platforms used for this study were an Agilent 1100 series HPLC coupled to an Applied Biosystems QStar *Pulsar i* ESI-Q-o-ToF, however, other analytical platforms are available and should be compared to the platforms used here to allow the relative advantages and disadvantages for each to be assessed, as these can have a dramatic effect on the results obtained. The majority of metabonomic studies began using ^{1}H-NMR, although LC-MS is now becoming more commonplace (Lenz and Wilson, 2007). Whichever platforms are utilised for a metabonomic study, no one single analytical technique can be considered to provide a completely comprehensive picture of the compounds contained within a biofluid.

3.2.2. Analytical platforms and separation techniques

Early metabonomic studies used ^{1}H-NMR as their analytical platform. ^{1}H-NMR is a very reproducible technique that suffers less analytical bias than other platforms due to its universal detection of compounds, providing that they contain a proton. As LC-MS systems increased in reliability, and the field of metabonomics evolved, the increased sensitivity (superior to that of NMR) and stability afforded by the latest LC-MS platforms has led to its increased usage within the field.

There are many different MS platforms available; the Q-o-ToF instrument used in this study has the capability to provide tandem MS data with resolutions approaching 10,000 with a mass accuracy typically better than 20 ppm. However, the data collection rate is rather slow (typically 1 s per 'scan') compared to other MS techniques. Triple quadrupoles (QqQ) can also perform tandem MS (along with other MS experiments) but have lower levels of resolution and mass accuracy compared to ToFs. Ion traps (IT) can perform successive CID experiments, however, they typically have to perform many CID steps to gain similar data to those in a single step on a Q-o-ToF. IT's resolution and mass accuracy are poor compared to a Q-o-ToF, which is not desirable for accurate mass measurements – something that can be important for metabonomic studies. A recent 'enhancement' of the IT is the orbitrap; it combines a linear IT with a subsequent IT where ions are maintained within the trap; their mass being recorded by the cyclic motion of ions passing between the ends of the trap.

This technique affords very high resolution, mass accuracy, data acquisition rates and in-depth structural characterisation properties comparable with a traditional IT.

Whichever MS detection technique is used, some form of front-end separation is required for complex biofluid matrices.

Gas chromatography (GC) coupled to MS was the traditional approach to metabonomic studies due to its high resolving power and reproducibility. Since as much information as possible is desired for metabonomic studies, GC as a separation method is largely unsuitable for biofluid analysis (Kopka, 2006). This is due to the fact that compounds need to be volatile in order to be separated and therefore analysed. As urine predominantly contains compounds of a polar nature, the majority of these would fail to be detected using GC as a separation technique. Derivatisation can be carried out in order to increase the number of volatile components, but can be a lengthly process that is not 100 % efficient and also changes the chemical structure of compounds being derivatised. Given the nature of metabonomic studies, high throughput experimentation is often required which cannot be obtained using GC due to the lengthly analysis times.

One separation method which is gradually gaining more attention for metabonomic studies is capillary electrophoresis (CE) (Wang and Liao, 2004; Iadarola *et al.*, 2005; Pisitkun *et al.*, 2006; Ullsten *et al.*, 2006). CE has the ability to separate components, in urine for example, using an aqueous medium with only small injection volumes (in the nL range, compared to μL range used for GC and LC). Despite the apparent benefits that CE should offer as a separation technique, it has failed to make any appreciable mark in metabonomic research. Only a handful of papers have utilised CE (some coupled to MS) successfully (Ullsten *et al.*, 2006; Monton and Soga, 2007; Soria *et al.*, 2007), and even then the systems were targeted and therefore not comprehensive. The reasons behind the apparent failure of CE-MS to make headway in metabonomic research may be due to the poor reproducibility of migration times, as well as the problems associated with joining the capillary to an ES source successfully. Work previously carried out within the JTO and associated groups (Emma Edwards, Ed Bergström, Cristina Soria and Julie Wilson) has highlighted CE's current unsuitability for metabonomic studies. Undertaking statistical analysis using data generated by CE proved to be near impossible due to large deviations in migration time, baseline shifts, capillary degradation and instability of the capillary interface. As the technology advances, CE may become a viable

separation technique, but for now appears to be inappropriate for the separation of biofluids in large-scale metabonomic experiments.

Separation using HPLC is now the most common method currently employed for the separation of components in biofluids prior to analysis by MS (Bajad and Shulaev, 2007; Chen *et al.*, 2007; Hodson *et al.*, 2007; Lenz and Wilson, 2007; Wagner *et al.*, 2007). Using such a separation technique can reduce (but not remove) the effects of matrix suppression (Taylor, 2005; Chambers *et al.*, 2007) when coupled to an ESI/APCI interface. Matrix suppression can reduce the number of compounds that can be detected within a sample, thus reducing the coverage of compounds present and increasing the selectivity of LC-MS as a metabonomic platform. Despite LC being the most common separation method used for MS studies, many fail to appreciate the importance of choosing a column, as this dictates the bias towards particular classes of compounds that can be retained and therefore detected. The overwhelming majority of metabonomic studies utilise a reversed phase approach; this discriminates against polar compounds, which are likely to be the main components of biofluids, due to their aqueous nature. Some studies have utilised hydrophilic interaction LC (HILIC) (Idborg *et al.*, 2005; Hemström and Irgum, 2006; Mawhinney *et al.*, 2007) as a complementary separation method. The development of a HILIC separation approach is assessed in section 3.6.

Whilst reasonable analysis times can be obtained using HPLC[1], development of this method called ultra-performance LC (UPLC), which should perhaps more correctly be called small particle LC, is gaining attention within the literature (Wilson *et al.*, 2005; Crockford *et al.*, 2006; Nordstrom *et al.*, 2006; Bruce *et al.*, 2007; Lenz *et al.*, 2007; Rainville *et al.*, 2007). UPLC uses sub 2 µm particles, which enable far superior separation and resolution than particle sizes typically used in HPLC (> 2 µm) (Churchwell *et al.*, 2005; Plumb *et al.*, 2005; Wilson *et al.*, 2005). HPLC has a maximum pumping pressure of less than 400 bar, whereas UPLC typically uses pressures that can exceed 700 bar (Churchwell *et al.*, 2005), which is required when using sub 2 µm particles. At such high pressures, run times can be reduced to a fraction of those used in HPLC separations, whilst maintaining a greater resolving

[1] Monolithic columns (chapter one) have allowed reductions in analysis time and increases in sensitivity and stability to be obtained for proteomic experiments (Premstaller *et al.*, 2001; Wienkoop *et al.*, 2004; Chen *et al.*, 2005; Ault, 2007; Sumpton 2007); their uptake in metabonomics seems much slower.

power and increasing the possibility of allowing the detection of more compounds (Churchwell *et al.*, 2005; Plumb *et al.*, 2005; Wilson *et al.*, 2005).

Using UPLC separation methods does discriminate against certain MS platforms due to the peak widths that can be obtained (less than 1 s at half-height). For example, the Q-o-ToF used in this study typically has an acquisition speed of 1 s, meaning that some peaks may go undetected. Lenz *et al.* utilised an orthogonal ToF MS with UPLC separation, obtaining peak widths of ~3 s at half-height (Lenz *et al.*, 2007); this would still be too short for tandem MS experiments where the total acquisition times can approach 10 s (again for the Q-o-ToF used in this study).

3.2.3. Ionisation methods

The output of CE and LC is in liquid form, and MS requires gas phase ions; as such, there are different sources that can generate gas phase ions from liquid. The most common and universal interface is electrospray ionisation (ESI) and its related techniques (micro ESI and nano ESI), the use of which is dependent upon the flow rate from CE/LC. The largest problem with ESI is the fact that it suffers from matrix effects (Taylor, 2005); this is where co-eluting compounds compete for the 'ion-stream', generally resulting in a decrease in ionisation efficiency for some compounds (or the enhancement of other compounds). Whilst the exact mechanism of matrix effects are unknown, it is postulated that polar compounds fail to reach the surface of the charged droplets formed, therefore not being transferred into the gas phase (Bonfiglio *et al.*, 1999; King *et al.*, 2000). Given that most biofluids contain a high proportion of polar content, then the use of ESI as an ionisation method seems perhaps not ideal. APCI is a complementary method of ionisation that is analogous to ESI.

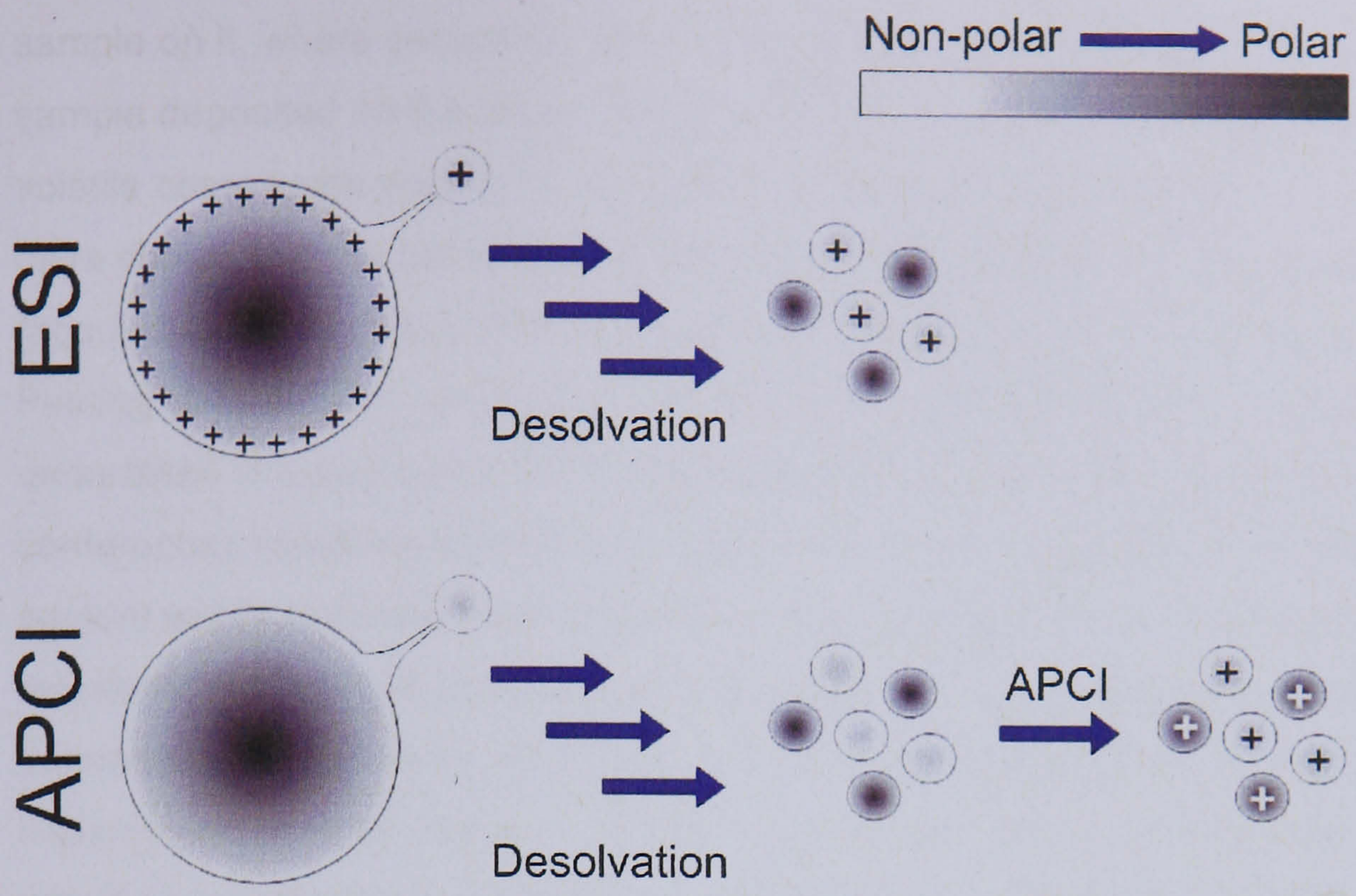


Figure 3.2.1. A comparison of the different methods of ionisation for the two complementary ionisation methods: ESI and APCI.

Figure 3.2.1 compares the two mechanisms of ionisation in ESI and APCI. ESI creates many small, charged droplets. The polar more contents tend towards the centre of a droplet, away from the charged surface where the more hydrophobic content resides, therefore having a greatly reduced chance of being transferred into the gas phase as an ion. Conversely, APCI generates neutrals first, meaning that a mixture of polar and non-polar neutrals is created. After the generation of neutrals, molecules are charged (creating gas phase ions) by the production of electrons at a corona discharge needle, which in turn allows proton transfer (or adduction) to occur. This ionisation method does not discriminate against polar compounds to the extent that ESI does, therefore allowing greater ionisation efficiency for polar compounds.

A recent advance by Shimadzu is the development of a dual APCI/ESI source which allows both methods of ionisation to occur at the same time, should increase the overall ionisation efficiency over a broad range of compounds, from highly polar to non-polar, thus making it potentially highly suitable for LC-MS metabonomic studies (Shimadzu).

One final ionisation source that has recently been reported for use in metabonomic studies is desorption-ESI (DESI) (Pan *et al.*, 2007). DESI-MS works by 'firing' a stream of gas phase ions (generated by ESI of a suitable solvent) at a target with a

sample on it, where secondary ions are formed by collisions of primary ions with the sample deposited on the target. Whilst this is a crude method where many non-volatile compounds may fail to be ionised (personal conference notes), Pan *et al.* have shown that the technique can provide a rapid analysis of many samples (acquisition times of less than 1 min per sample). This work parallels the work of Pauling *et al.* in 1971, which was one of the first metabolite profiling experiments undertaken (Pauling *et al.*, 1971). Despite this ionisation method being 'snubbed' at conferences (personal conference notes) for its poor coverage of the metabolite content within biofluids, therefore technically not being a method suitable for metabonomics by definition, this simple approach may actually yield relevant biomarkers. Research by Willis *et al.* showed that dogs could be trained to identify patients with bladder cancer "...*on the basis of urine odour more successfully than would be expected by chance alone*..." (Willis *et al.*, 2004). This suggested that volatile compounds related to tumours were present within the urine from cancer patients. There are many other stories of animals being able to detect illnesses much before any medical symptoms can be detected (BBC, 2007). As much of the research within the field of metabonomics concerns illnesses, as is so often the case with nature, maybe science can learn a lesson that sometimes simple methods may produce the best results?

3.2.3.1. APCI validation

In order to make a comparison of APCI and ESI sources for analyses of human urine samples, the only APCI source available for the Applied Biosystems QStar Pulsar *i* Q-o-ToF was used to analyse the urine samples obtained from volunteers from the Department of Chemistry, University of York. The urine samples were analysed using all of the considerations laid out in this chapter. From the first instance the APCI source proved unsuitable for metabonomic experiments, as any of the settings with respect to the positioning of the corona and nebulising source could not be maintained throughout data acquisition, due to the poor design of the source.

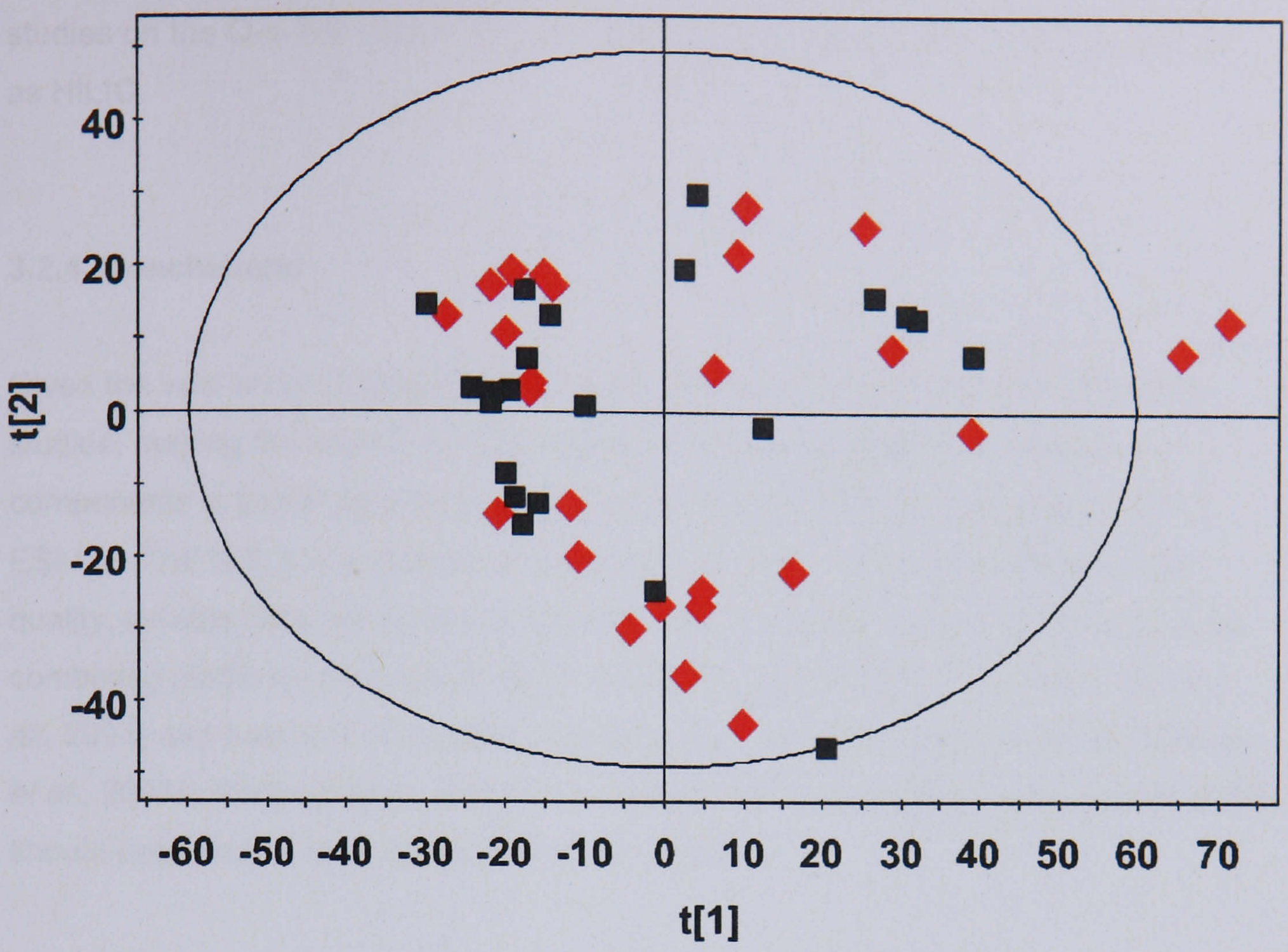


Figure 3.2.2. PLS scores plot for a gender response variable from positive mode RP-LC-APCI-MS data. ■ = samples from male volunteers, ◆ = samples from female volunteers.

Urine samples from volunteers within the Department of Chemistry were first analysed using RP-LC-APCI-MS. The PLS scores plot from analysing the positive mode RP-LC-APCI-MS data shows no clearly definable clusters based upon gender (figure 3.2.2). Clustering according to gender should have been observed, as was seen for PLS analysis of the same samples using RP-LC-ESI-MS (section 3.6), but proved impossible to obtain. This failure to obtain suitable data could be caused by the fact that the APCI source used was unable to cope with long data acquisition times (two to three days), as the corona discharge needle quickly became excessively dirty with deposits being formed, greatly reducing the ionisation efficiency over time; this was further highlighted when the urine samples were analysed using negative mode RP-LC-ESI-MS, as the corona again formed heavy deposits, completely reducing the ionisation efficiency, causing no data to be obtained for these analyses. The inability of the APCI source to produce robust data, along with the fact that any settings were hard to maintain with respect to the positioning of the corona and nebulising source made its further use impractical. Had a better APCI source been available, then APCI may have been appropriate for metabonomic

studies on the Q-o-ToF, especially in conjunction with a separation technique such as HILIC.

3.2.4. Conclusions

Given the vast array of different analytical techniques available for metabonomic studies, making the correct choice to obtain a 'global fingerprint' of biofluid components is easier said than done. Despite this study being restricted to HPLC-ESI-Q-o-ToF MS, it is a platform that has the potential to provide sufficient, high quality, reliable data. As no single technique can provide a comprehensive analysis, combining platforms such as LC-NMR-(ESI/APCI)MS (Burton *et al.*, 1997; Bajad *et al.*, 2003) and fusion of complementary data (Kenney and Shockcor, 2003; Forshed *et al.*, 2007a; Forshed *et al.*, 2007b; Lenz *et al.*, 2007; Zhengzheng and Daniel, 2007) should become the rule, rather than the exception.

3.3 Sample collection and analysis

3.3.1 Introduction

The design and implementation of a metabonomic study has to be very carefully considered if a successful output stands any chance of being achieved; that is achieving the original goals, or answering a hypothesis generated by analysing the data obtained. From the initial design of a study, to collecting the samples, and then their subsequent processing prior to analysis requires many different steps and challenges, all of which can have a considerable effect upon the end result.

The main goal of a metabonomic study should be to produce a robust and comprehensive fingerprint of the biofluid chosen (Wilson *et al.*, 2005; Lenz and Wilson, 2007; Sangster, T P *et al.*, 2007), however, this is much easier said than done. Ideally, a broad range of analytical platforms (and techniques specific to each platform) should be encompassed to allow this idea of a 'global' fingerprint to, at the very least, be considered. In reality, many studies, including this one, are limited by the available analytical platforms, amount of biofluid available, money and most important of all, time.

3.3.2 Aims

The aim of this section is to compare some of the methods currently used within the literature, and to highlight some of the many aspects of sample collection and analysis, which are all too often omitted from metabonomic studies. The data presented within this section was obtained by the analysis of urine collected from fit and healthy volunteers from the Department of Chemistry, University of York (unless otherwise stated). This section comprises initial considerations for sample collection, sample storage and pre-treatment, sample treatment prior to analysis, sample stability during analysis, repeat/blank injections, random analysis and importantly, system stability.

3.3.3 Results and discussion

To begin a metabonomic study, it needs to be decided what question is trying to be answered; it is of no use collecting a biofluid from exclusively healthy donors/volunteers/animals if biomarkers related to kidney disease are sought for example (Dihazi and Muller, 2007). A carefully designed plan for collecting the desired biofluid should first be constructed. For the initial development work undertaken for this thesis, a broad cross section of urine samples that would ideally not be perturbed by influences of illness were sought. This was to both aid the development of a robust LC-MS system, highlighting some of the pitfalls that can be encountered with metabonomic studies, and to develop complementary LC techniques to attempt to increase the coverage of the urinary fingerprint by LC-ESI-MS methods (see section 3.6 – HILIC development).

Members of the Department of Chemistry, University of York, UK, were contacted by e-mail requesting their help by donating two urine samples to aid methodology development. It was stressed that anonymity would be maintained throughout. Any volunteers were able to collect a sealed pack containing two randomly numbered sterile 25 mL sample tubes, two sealable bags and an instruction sheet informing them once more of anonymity and how to collect/deposit the donated urine samples. Volunteers were asked to provide the two mid-stream urine samples from the same day, the first being the first void of the day, and the second being any void after 15:00 (but before 18:00).

The only information requested from the volunteers was gender, time of collection, age and whether they were a smoker or not. All volunteers were advised that if they felt uncomfortable with providing age then they could provide an age range, similarly, they did not have to declare being a smoker if they did not wish to do so. After each sample was donated, the volunteers were asked to deposit them into one of three large, sealed, red boxes across the department as soon after donation as possible (confirmed by all samples still being warm upon collection). Any deposited samples were recorded and immediately stored at -80 °C until all samples were collected (appendix A contains all recorded data from the collection of these samples).

The careful consideration of what samples were required, and their prompt collection and storage, is the initial priority towards obtaining good data.

3.3.3.1 Sample storage, stability and preparation

Many metabonomic studies require time setted data, and as such, many samples are required to be stored for some time before all samples can be analysed. Some studies have added preservatives prior to storage, such as sodium azide[1], to inhibit bacterial growth (Saude and Sykes, 2007); immediate storage at -80 °C should be sufficient to inhibit bacterial growth (Lauridsen *et al.*, 2007). The addition of preservatives prior to storage may have adverse effects upon the composition of the samples, and was therefore not used with samples collected for this study.

Studies into the effects of urine storage at various temperatures have all come to the same general conclusions (LeBeau *et al.*, 2001; Schneider *et al.*, 2002; Fura *et al.*, 2003; Gika *et al.*, 2007). Storing urine at room temperature without first filtering caused the concentration of some compounds such as benzoate, lactate and creatine to fluctuate, meaning that the urine samples were subject to degradation. Filtering the urine samples prior to storage at room temperature diminished the effects of degradation of benzoate and lactate, but failed to have any appreciable effect upon stopping the concentration of creatine from fluctuating (Saude and Sykes, 2007). This suggests that filtering samples removes possible causes of degradation (such as bacteria) and can increase the stability of urine stored at room temperature. Whilst filtration may exclude some compounds from the samples (e.g. large proteins), filtering has a two-fold benefit, also removing sediment.

Despite the centrifugation of urine samples collected from volunteers within York Chemistry Department at 10,186 g for 8 min, it was noticed that particulate matter still remained in some of the urine samples; injecting these samples onto an LC column would have quickly degraded its performance, as interparticulate spaces (or the guard column) could easily have become blocked, therefore increasing back pressure. Because of this, all samples were filtered through 0.45 μm PVDF syringe filters, removing any particulate matter and also helping to increase the stability of the samples at room temperature prior to analysis. Whilst the work by Saude and Sykes showed that for the short term (i.e. less than 8 hours), samples should be reasonably stable at room temperature and are therefore fine to be racked for analysis, it would be more suitable to store the samples in a temperature controlled rack to try and

[1] Sodium azide (NaN_3) is a biocide that inhibits bacterial growth of gram-negative bacteria.

diminish any degradation of compounds as much as possible (as was done by Gika *et al.*).

A study comparing endogenous urinary metabolites stored at room temperature and at -80 °C reported that the concentration of all metabolites studied altered significantly at room temperature, but remained reasonably stable over a four week period of storage at -80 °C (Saude and Sykes, 2007). Gika *et al.* reported that over a period of four weeks, storage at either -20 or -80 °C did not highlight any appreciable differences when the data collected from the LC-MS analysis of urine samples were compared by PCA. They do however, correctly point out that this is a 'blunt analysis tool', and that some metabolites which do not have large influences upon the developed PCA model (heavily dependent upon the scaling method utilised) could in fact be subject to degradation, and not be highlighted by a change in the observed clustering shown by the PCA scores plot (Gika *et al.*, 2007).

As a result of earlier studies (LeBeau *et al.*, 2001; Schneider *et al.*, 2002; Fura *et al.*, 2003), it was decided that any samples collected for this study should be stored at -80 °C and allowed a period of at least four weeks for any degradation of urinary components to become consistent across the cohort, and only centrifuge and filter prior to analysis. More recent research specifically tailored for metabonomic experiments confirmed that the original choice to just store samples at -80 °C, and to centrifuge and filter prior to analysis were optimal (Gika *et al.*, 2007; Saude and Sykes, 2007).

After an initial four week storage period, aliquotting the urine samples collected from the Department of Chemistry into smaller portions was required to prepare aliquots for subsequent analysis and so minimise the number of freeze/thaw cycles required. Whilst Saude and Sykes recommend that the number of freeze/thaw cycles are kept to a minimum, Pisitkun *et al.* showed that up to four freeze/thaw cycles had little effect upon the composition of urine, and Gika *et al.* (again from PCA results) showed that up to nine freeze/thaw cycles did not effect clustering (Pisitkun *et al.*, 2006; Gika *et al.*, 2007; Saude and Sykes, 2007). Despite the apparent stability of urine to freeze/thaw cycles, stored urine samples should be treated identically, and all possible manipulations should be undertaken at the first freeze/thaw cycle when samples are aliquotted; this should hopefully decrease the chances of any unnecessary degradation of urinary compounds.

3.3.3.2 System Stability

Metabonomic studies were initially carried using NMR analytical platforms (Nicholson *et al.*, 1999; Lenz *et al.*, 2000; Robertson *et al.*, 2000). NMRs are renowned for their reproducibility, not only from day to day, but also from laboratory to laboratory. These high levels of reproducibility are very desirable for metabonomic experiments. However, despite poorer levels of reproducibility, LC-MS metabonomic studies are becoming more common due to its higher sensitivity which should be more compatible for the large dynamic range and chemical complexity seen in a urine matrix (Want, E J *et al.*, 2005; Wilson *et al.*, 2005; Want, E. J. *et al.*, 2007). This is not to say that NMR is a redundant technique though; not everything can be detected by LC-MS (especially when ESI is utilised). NMR should be seen as a complementary method of analysis for metabonomic studies. However, it still remains that many changes within biofluids may be below the LOD obtainable from NMR studies.

As LC-MS systems exhibit less reproducibility than NMR systems, the resulting data produced should be carefully scrutinised. There are many methods that can be used to increase and monitor the stability of an LC-MS system.

LC-MS systems generally require some time to allow the whole setup (both the LC and MS side) to equilibrate. It is good practice to allow an LC column to equilibrate by first running a gradient (no sample injection) to condition the column. This initial gradient also allows the ESI chamber to heat up to the selected desolvation temperature (300 °C for all experiments described within this chapter), and electronics/optics to stabilise.

During the aliquotting of samples collected from within the department, ca. 100 μL from each sample was held back to create a 'pooled' sample that was representative of all the samples to be analysed. In order to evaluate system stability over a whole run, pooled samples were randomly included throughout any run with at least three pooled samples being run back to back at the beginning of the run, before the analysis of individual samples.

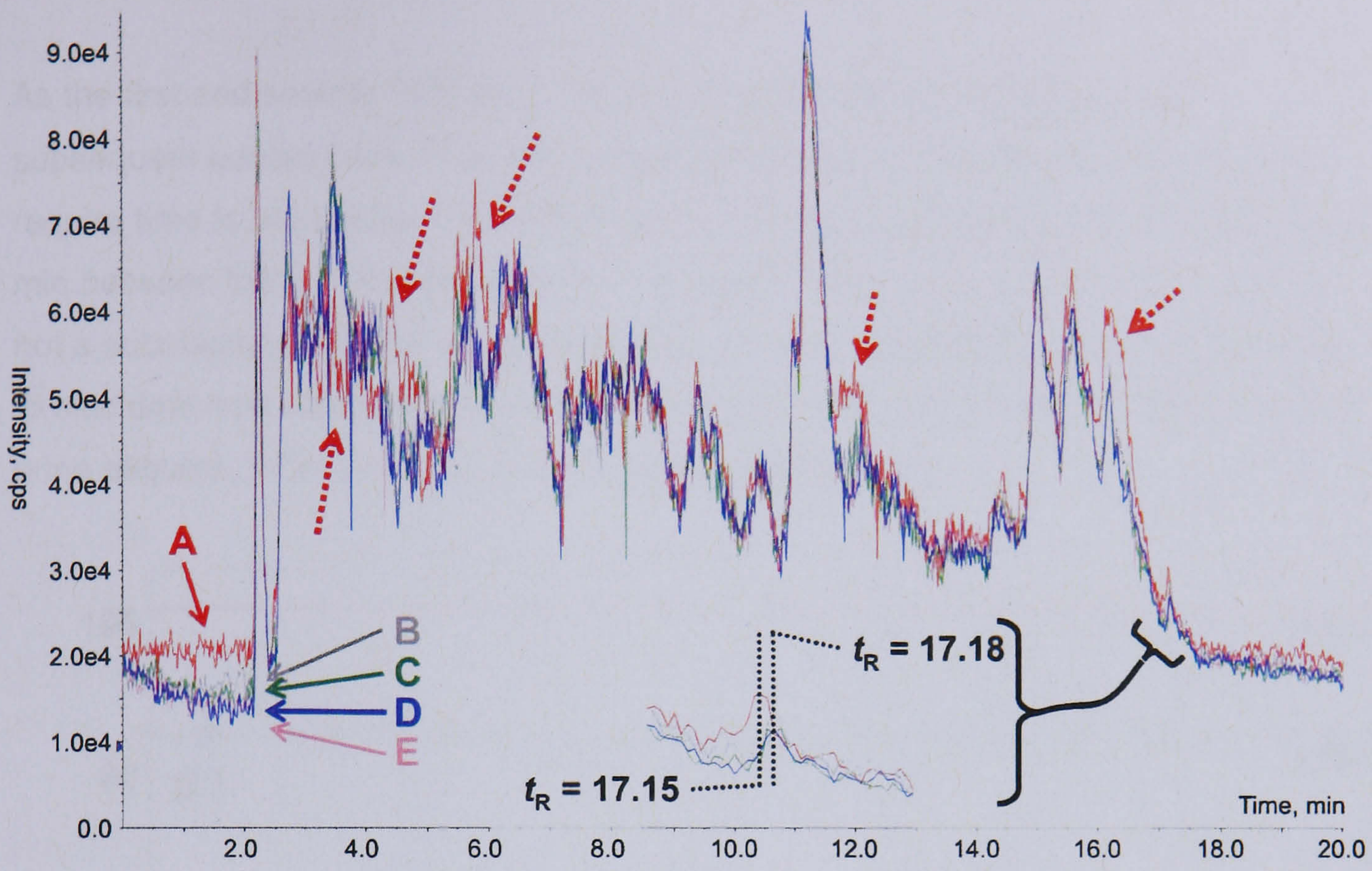


Figure 3.3.1. Five positive mode RP-LC-MS TICs from aliquots of pooled urine. The first three samples (A – C) were from run back to back at the beginning of data acquisition, with the remaining samples (D – E) being analysed throughout the run. Hashed red arrows indicate deviation from sample A. Inset shows a magnified portion of the TICs, highlighting a minor deviation in retention time.

The data presented within figure 3.3.1 show five TIC traces from five replicate injections from pooled urine aliquots of the collected urine set from the Department of Chemistry. The first three traces were from back to back injections after the first initial conditioning gradient. The TIC traces generally appear to exhibit the same trend throughout the whole 30 min acquisition (only 20 min shown as this was the information rich section of TICs). The most noticeable deviations are for TIC A, the first sample analysed, and are highlighted by the hashed red arrows; all samples show some minor deviation in intensity. Inset into figure 3.3.1 is a magnified section of a low intensity peak at ca. 17.2 min. The first TIC (A) gave a retention time for this peak at 17.15 min, with the second subsequent TIC (B) at 17.18 min; despite the peak from TIC B having a retention time equal to all subsequent TICs, the intensity is slightly higher (and higher still for TIC A). The last sample from the back to back pooled urine injections (C) show a retention time of 17.18 min, which is consistent with the two further sample TICs (D & E) from pooled samples which were analysed

after individual samples had been analysed; the intensity of the three peaks (C to E) is also consistent.

As the first and second TICs from the pooled urine aliquots differ from any subsequent pooled urine TICs, this suggests that the LC-MS system does indeed require time to equilibrate. The difference in retention time shown from 17.15 to 17.18 min between the first and third (and subsequent) TICs correspond to <2 s, which is not a substantial deviation in retention time. To investigate these effects further, and to elucidate how different any eluting compounds were from TIC to TIC of the pooled urine aliquots, PCA was used to analyse the resulting data.

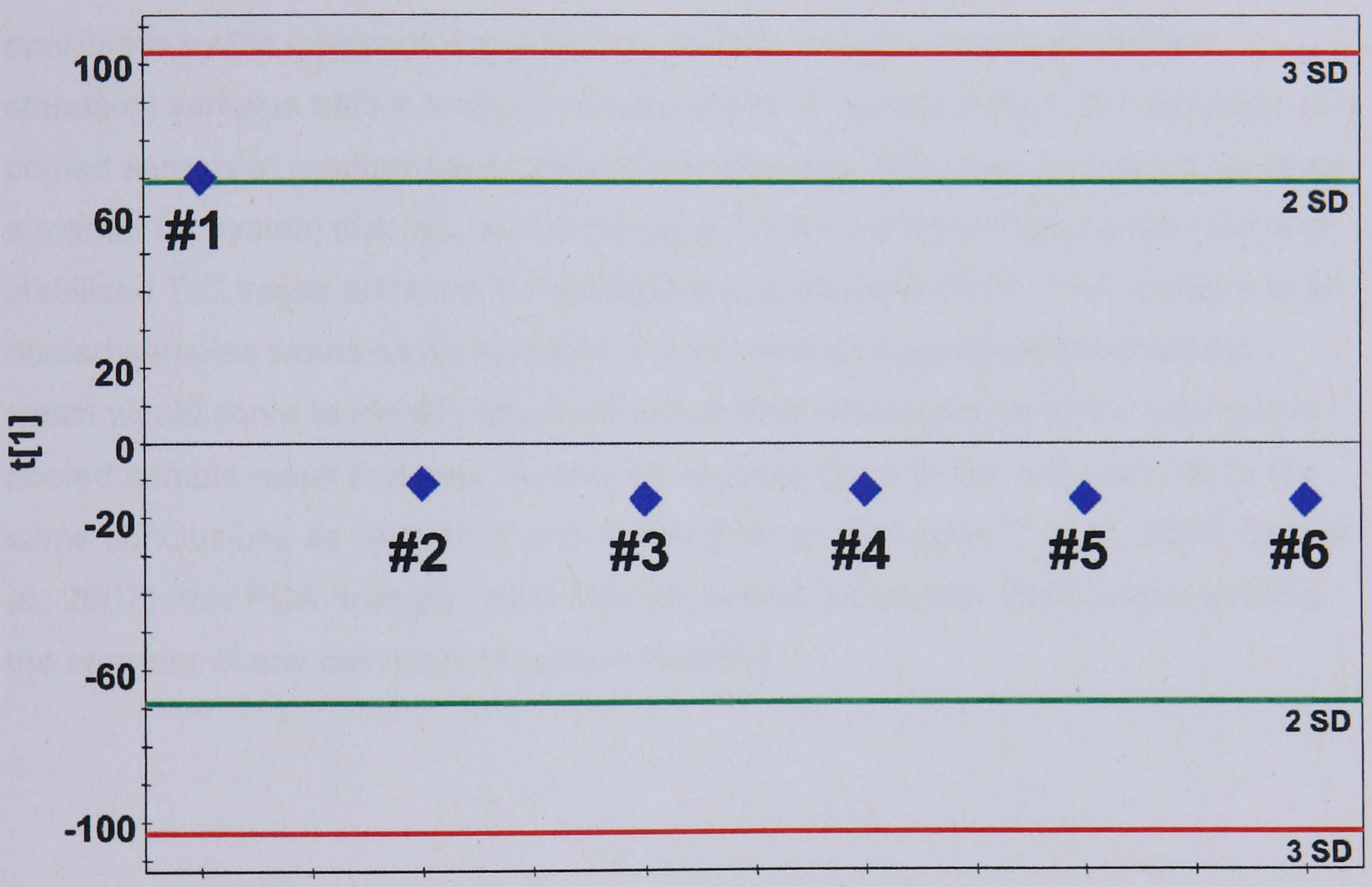


Figure 3.3.2. PCA scores plot using one principal component (y axis) comparing data from six LC-MS analyses of pooled urine aliquots.

Figure 3.3.2 shows the resulting scores plot from the PCA analysis of six replicate LC-ESI-MS runs of pooled urine aliquots, with only one principal component being developed for the model (t[1] on the y axis). Point number one corresponds to the first injection, and resides just outside the green line, meaning that for this data point, it is over two standard deviations outside the average of the analysed data. All subsequent samples data (points two to six) are well within the two standard deviation lines and illustrate how each sample analysis was, statistically at least, equivalent to one another.

From the TIC traces (figure 3.3.1) and the PCA analysis of resulting LC-MS data from the analysed pooled urine samples, it is clear that the LC-MS system requires one to two samples to be injected and separated to enable the system to equilibrate. The reason why the LC-MS system does not produce reproducible results from the offset may lie with the chromatography itself. The LC column may require certain binding sites to be masked, or may just require conditioning to the type of compounds that are being analysed. The MS optics and electronics may heat up at the start of data acquisition, and therefore require a short period to reach equilibrium.

The above data highlight the fact that LC-MS systems should be allowed to equilibrate by the injection of at least two pooled (or some other well defined standard) samples before analysis of samples from the test cohort. The inclusion of a pooled sample at random points throughout data acquisition can also act to serve as a marker for system stability, as the resulting TIC traces should be representative of stabilised TIC traces obtained at the beginning of data collection. PCA analysis of all pooled samples would easily highlight any anomalous pooled sample analyses, which would serve to identify any test sample data acquired around the anomalous pooled sample result that may need to be re-acquired, with this work coming to the same conclusions as Gika *et al.* and Sangster *et al.* (Sangster, T *et al.*, 2006; Gika *et al.*, 2007) (this PCA analysis could also be carried out on-line, immediately warning the operator of any deviation of system stability).

One further way to evaluate the ongoing stability and repeatability of an LC-MS system is to analyse a sample from the test cohort twice, with a number of other samples analysed inbetween each analysis. This method serves not only to evaluate the stability of the system, but to also evaluate the reproducibility of sample preparation methods, as any errors should become evident when the resulting data were processed.

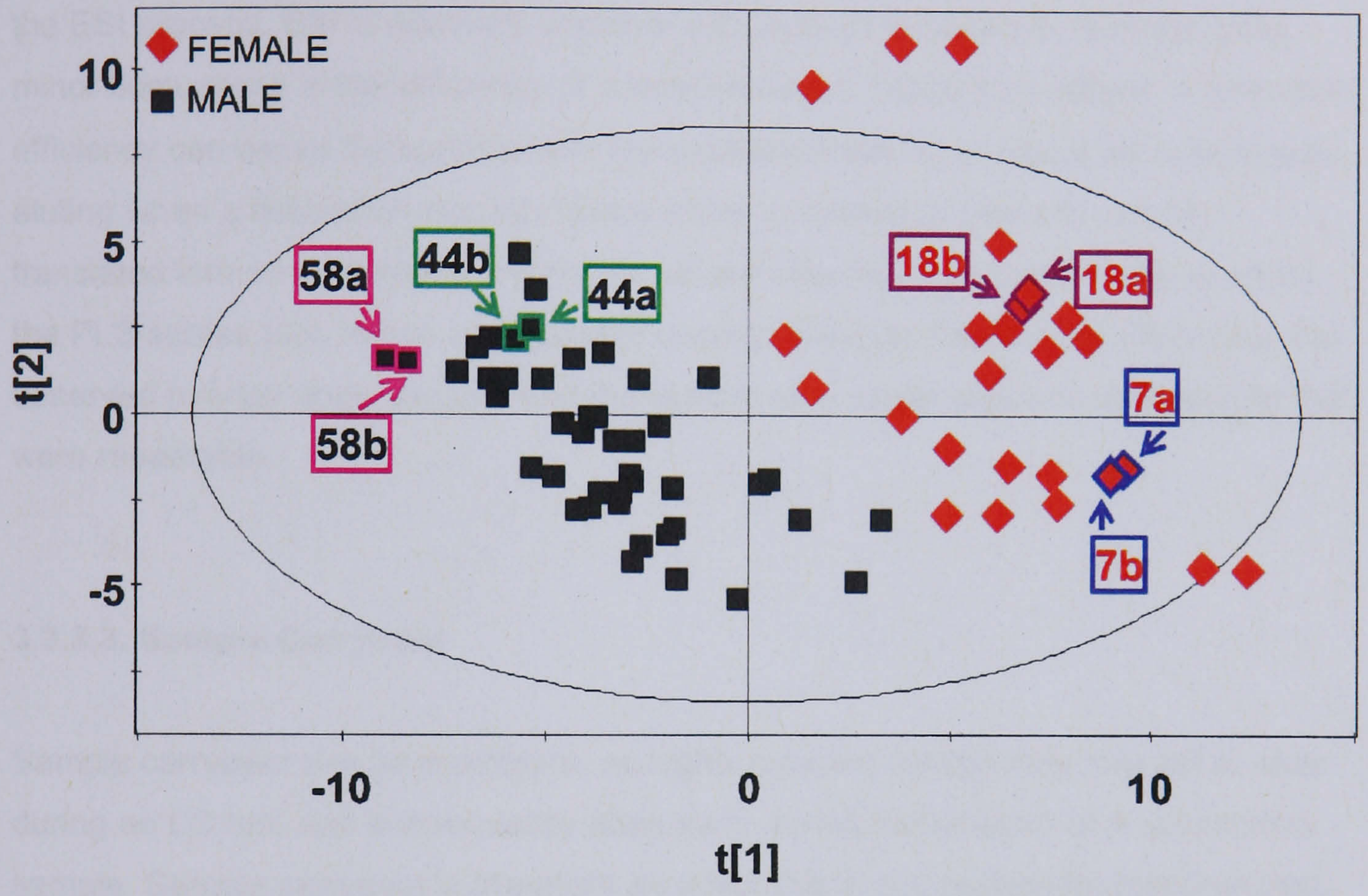


Figure 3.3.3. PLS scores plot of positive mode RP-LC-MS data separated according to gender. The model utilised two latent variables and was optimised to methodology outlined in section 3.5.

The data presented in figure 3.3.3 shows a PLS scores plot for the data from samples collected within the Department of Chemistry, separated according to gender using two latent variables. There is clear separation between samples donated by males and females, as was expected. The four points (all female) that lie outside the 95 % confidence margin (represented by the ellipse) were only found to be marginal outliers (see section 3.5). Four pairs of points correspond to the analysis of duplicate aliquots of four different samples, randomised throughout analysis. Samples 18a and 18b were analysed with ten other samples inbetween, which corresponds to 6.3 h between each sample being analysed; 7a and 7b had 11 samples analysed inbetween, corresponding to 7.0 h between each sample; 58a and 58b had 9 samples run inbetween, meaning 5.7 h inbetween each sample, and

samples 44a and 44b had 13 samples analysed inbetween, meaning 8.2 h lapsed between the replicate sample analysis.

Each of the four pairs of points show the duplicate analyses residing within the same area on the PLS scores plot. Obviously, if the LC-MS system were highly reproducible, one would expect the two points to overlap exactly. The reason that each of the pairs of points are slightly offset from one another, can be explained by the ESI process. ESI is relatively unstable with respect to absolute reproducibility; minor fluctuations in the efficiency of ionisation occur. These fluctuations in ionisation efficiency can cause fluctuations with the recorded intensity of signal for compounds eluting when a fluctuation occurs. These minor variations in intensity can be translated into minor deviations from the values seen for a replicate analysis within the PLS scores plot, hence an imperfect overlap. Despite these minor variations, the observed overlap does suggest that the system as a whole was providing results that were repeatable.

3.3.3.3. Sample Carryover

Sample carryover can be a problem, as highly retained compounds may fail to elute during an LC run, and subsequently elute early during the analysis of a successive sample. Sample carryover is therefore an effect that is not desired for metabonomic studies as it could give rise to spurious results. To try and combat this problem, all LC gradients used for analysis within this project had extended wash cycles at the end of each gradient (see section 3.6 for more details).

To evaluate the possibility of sample carryover, blank runs were randomly carried out during LC-MS data acquisition; this involved a gradient being run without the injection of any urine samples; only a 1:1 mixture of MeCN:H_2O was injected. Any sample carryover from the previous sample injection should be visible in the resulting TIC trace of the blank run.

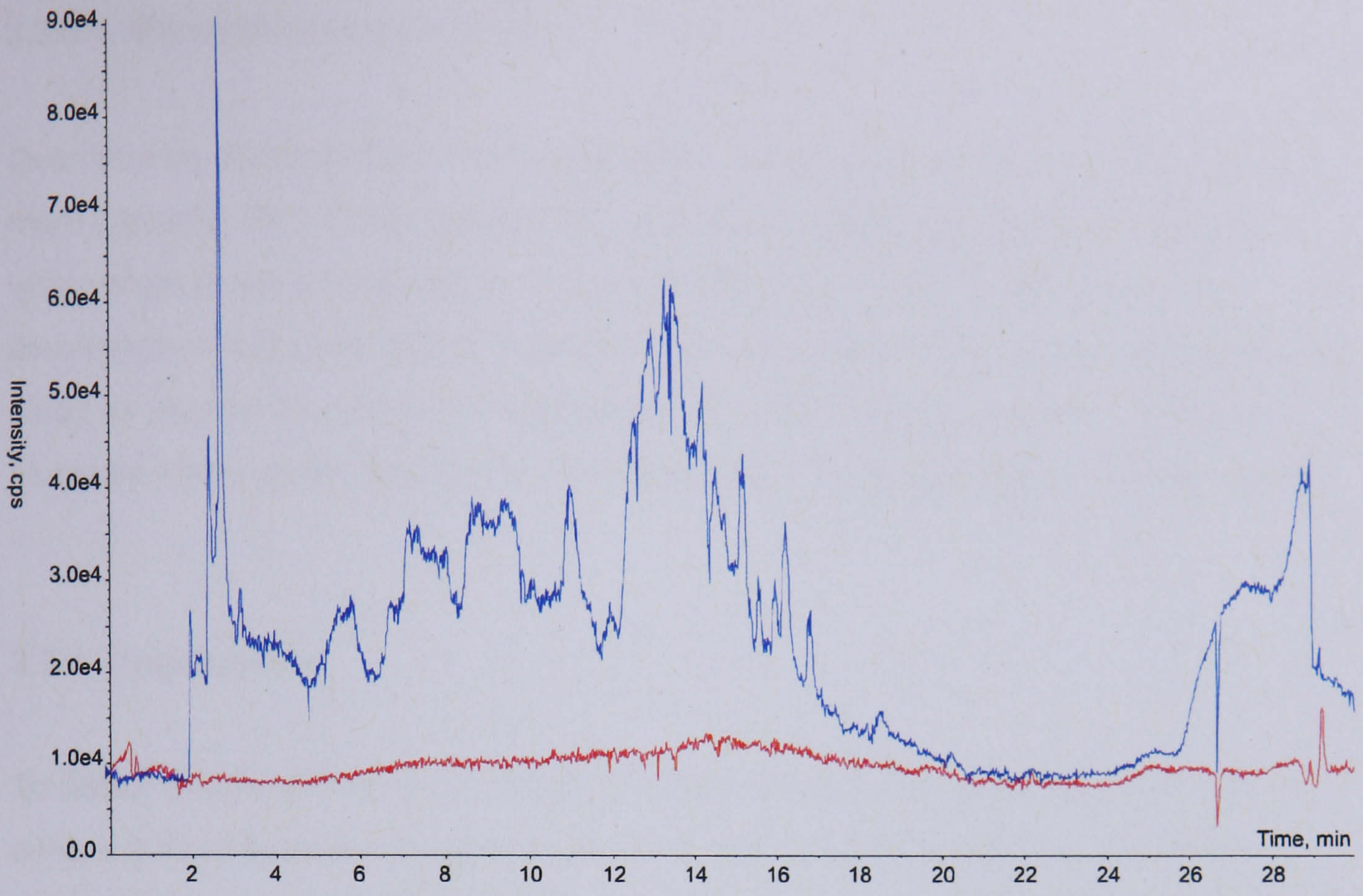


Figure 3.3.4. Two TIC traces from a positive RP-LC-MS analysis. The blue trace corresponds to a sample run before a blank gradient with no sample injection was recorded (red trace).

Figure 3.3.4 shows comparison of a sample TIC and a subsequent blank injection of solvent; the red trace corresponds to the blank while the TIC from the sample that was recorded immediately before it, is shown by the blue trace. The baseline from the blank gradient does not show any appreciable deviation from the 1000 cps intensity value seen for both the sample (first two minutes, corresponding to the dead volume) and the blank. There is a sharp dip in intensity for both traces evident at ca. 27 min, which may be caused by a persistent contaminant eluting from the column which causes suppression of ionisation (matrix effects); this could also be true for the only minor peak shown at ca. 29 min. The peak present in the blank TIC corresponds to a mass of 84.95 Da, and was also present in every TIC trace for all other blanks, pooled samples and samples, suggesting that it is a persistent contaminant in solvent. Clearly, the adopted cleaning strategy was sufficient to avoid sample carryover.

3.3.3.4. Random sample analysis

Over time as many samples are analysed the levels of contaminants may alter, or more typically, the LC-MS system may develop a systematic drift. If all samples within a group were analysed consecutively, then any trends could falsely aid discrimination between different groups. Therefore, when performing a metabonomic study (or indeed any study) involving many samples, which is aimed at trying to elucidate some underlying factors, samples should be analysed in a random order.

3.3.4. Conclusions

To obtain results that enable robust statistical models to be generated, one has to ensure that all sources of possible error, or unwanted perturbations to data are controlled as much as is possible to ensure that the phrase 'rubbish in = rubbish out' does not become a reality.

All studies should begin with careful planning and preparation before samples are collected. The work presented within this section shows that a carefully constructed, logical approach to sample collection and data acquisition is required. From the literature, it is recommended that samples should be frozen at -80 °C as soon as possible, and be left for a period of at least one month before being aliquotted to allow the degradation of any compounds to be consistent across the sample cohort; the data presented herein suggest that this was suitable. Freeze/thaw cycles should be kept to a minimum, meaning that any sample manipulation should be undertaken during aliquotting to ensure that all samples are treated equally and subjected to the same minimal levels of freeze/thaw cycles and manipulations.

As the goal of a metabonomic study is to obtain as representative a picture of the compounds present within the chosen biofluid as possible, any sample pre-treatment which could exclude compounds should be avoided unless absolutely necessary; this includes common clean-up steps such as SPE. Prior to analysis, samples should be centrifuged and filtered (minimal loss of compounds) to remove any particulate matter.

Representative pooled samples should be created during sample aliquotting. Pooled samples should be analysed first (minimum of three) to allow the LC-MS system to

come to equilibrium. Throughout data acquisition, further random analysis of pooled samples, and also replicate analysis of some samples should be carried out make it possible to check whether the LC-MS system as a whole remains stable. The inclusion of blank runs where no sample is analysed should be undertaken to ensure that sample carryover does not occur. The random analysis of samples is an absolute must, especially given that scientists are creatures of habit and like things to be ordered! Randomisation of samples across the whole data acquisition helps to avoid any time related shifts in intensity of contaminants and to minimise the effects of any instrument drift.

Overall, from sample collection, storage and pre-treatment, to its analysis, there are many different steps that are required in order to generate robust, reproducible (within run), and hopefully, meaningful results from an LC-MS analysis.

3.4 Data extraction

3.4.1 Introduction

Once the experimental data has been collected, one crucial step remains before the data can be statistically analysed: data extraction. This involves converting individual LC-MS data files into a single file that contains details of *m/z* and t_R along with intensity for each component detected across all of the sample cohort. Whilst this step may appear to be trivial, it is of great importance that raw data (here, LC-MS data) be carefully extracted with much consideration for the many parameters that can be applied. It is pointless ensuring that raw LC-MS data was rigorously collected so that it was the best available, if poorly chosen parameters or data extraction algorithms are chosen.

Unfortunately, there are relatively few data extraction algorithms available, as LC-MS vendors try to 'force' the user to utilise their software. Some freely available data extraction software programs, such as XCMS (Smith *et al.*, 2006), are available. However, non-proprietary extraction programs still require raw data to be extracted into a user-friendly format such as NetCDF, mzXML or mzData. Raw data files obtained from an Applied Biosystems QStar, used for obtaining all data presented in chapters three and four, are stored in a complex binary *wiff* format file. ProteomeCommons.org are currently offering a $1000 bounty for someone to create a *wiff* reader, thus highlighting the problem of non-standardised raw data output files. As the raw data generated for chapters three and four would have required converting from *wiff* into another file format before being extracted for statistical analysis, possibly losing data, it was decided to use the Applied Biosystems proprietary software, Metabolomics Export Script v1.0.0.3.

Problems that are associated with LC-MS raw data extraction (and that/those from other separation/detection methods) are typically peak picking errors that are caused by peak shapes, shifts of retention times, mass and signal intensity, all of which can cause errors when peak picking parameters are poorly selected (Sangster *et al.*, 2007).

The Metabolomics Standards Initiative (Fiehn *et al.*, 2007; Sansone *et al.*, 2007) aims to address some of the above issues relating to data extraction, along with the more important problem of standardising the raw mass spectrometric data reporting file

format by publishing a series of papers on standard reporting features (Metabolomics volume 3, 2007).

3.4.2 Aims

The aims of this section were to explore problems that are encountered when using data extraction software, along with investigating some possible solutions and areas of new research, which should hopefully aid a more accurate representation of extracted raw LC-MS data.

3.4.3 Results

The metabolomics export script works by generating *peaks* files (containing information about each detected ion) from each sample's raw LC-mass spectrometric data file (*wiff* format). The generated peaks files are then converted into aligned *peaks* files based upon many parameters that are chosen. The most critical of parameters relate to the retention time tolerance, mass accuracy and LC peak window size.

The retention time tolerance relates to how much retention time a particular *m/z* value can differ by and still be considered as deriving from the same peak. Studying TICs highlighted that the LC-MS system used in this study performs well, as there are not any notable deviations in retention time (this can be seen in figure 3.3.1 (section 3.3), where TICs from pooled samples are overlaid to provide an idea of system stability). A retention time tolerance setting of 0.5 min (giving ± 0.25 min) was found to be more than sufficient, and is similar to that reported within the literature from groups who have also used the metabolomics export script (Gika *et al.*, 2007; Sangster *et al.*, 2007).

Working alongside the retention time tolerance to determine what forms a peak, the mass tolerance setting is another important setting. As all of the datasets within chapters three and four average ca. 100 samples, the recording of all the data for each ionisation mode or separation method took around three days (each LC-MS run averages around 35 mins when the conditioning gradient and syringe cleaning step are included). During the three days of data acquisition, the mass spectrometer

cannot be expected to maintain a high level of mass accuracy (i.e. less than 10 ppm for a well calibrated QStar ToF MS). Without accounting for any drift in mass accuracy, extracting the raw data could force some genuine peaks to go undetected as their mass 'drifts' past the tolerance set within the metabolomics export script. When accounted for, any drift within mass accuracy was not considered to be a problem. This was because there are known compounds (e.g. creatinine, hippurate) within urine which can be used to re-calibrate the recorded LC-MS data. A mass accuracy setting of 500 ppm was used within the metabolomics export script. This is equal to ± 0.08 Da at *m/z* 150, which should be sufficient to cover any drift encountered for a ToF MS.

The last critical parameter, the LC peak window, is designed to remove any noise from each data file. Originally, this parameter was set to values of 0.1 and 3 min for the minimum and maximum values respectively. This meant that any peaks narrower than 0.1 min or wider than 3 min would not classify as a peak and therefore not be extracted. Upon studying the resulting extracted data matrix from the analysis of urine samples collected from within the Department of Chemistry, University of York, some abnormalities within the data were noticed.

/Sample Name	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
/Yvar	1	1	2	1	2	2	1	2	1
/Variablelist									
114.06_2.80	531.82	591.60	0.00	94.85	0.00	817.46	983.54	545.83	967.21
114.08_3.20	0.00	0.00	232.85	0.00	641.14	0.00	0.00	0.00	0.00
114.05_3.55	114.69	140.83	123.88	956.34	920.74	401.18	821.17	677.70	549.12

Figure 3.4.1. An excerpt of the extracted data matrix where each column represents a sample, and each row a concatenated *m/z* and t_R value (other than headers).

The two highlighted cells within figure 3.4.1 are the only peak areas extracted for that specific variable (114.08_3.20); all other samples apparently yielded no peak at all for this variable. Examining the variable in the row above (114.06_2.80) shows that two corresponding cells are empty, meaning that no peaks have been extracted for this variable (114.06_2.80) for those particular samples (S_3 and S_5 from the '/Sample Name row'). As the missing peak values for the first variable (114.06_2.80) correspond to the two values for the second variable (shaded cells, 114.08_3.20), this highlighted the fact that there must be some problem with the parameters used

for extracting the data. Upon analysis of the raw data, it was noticed that there were slightly broader peaks (not returning to baseline) for the above extracted values at *m/z* 114. Increasing the LC peak window extraction parameters to 0.1 and 20 min for minimum and maximum values respectively caused the above effect to disappear.

Because of the above problem, considering how extraction algorithms may treat peaks highlights some potentially important issues.

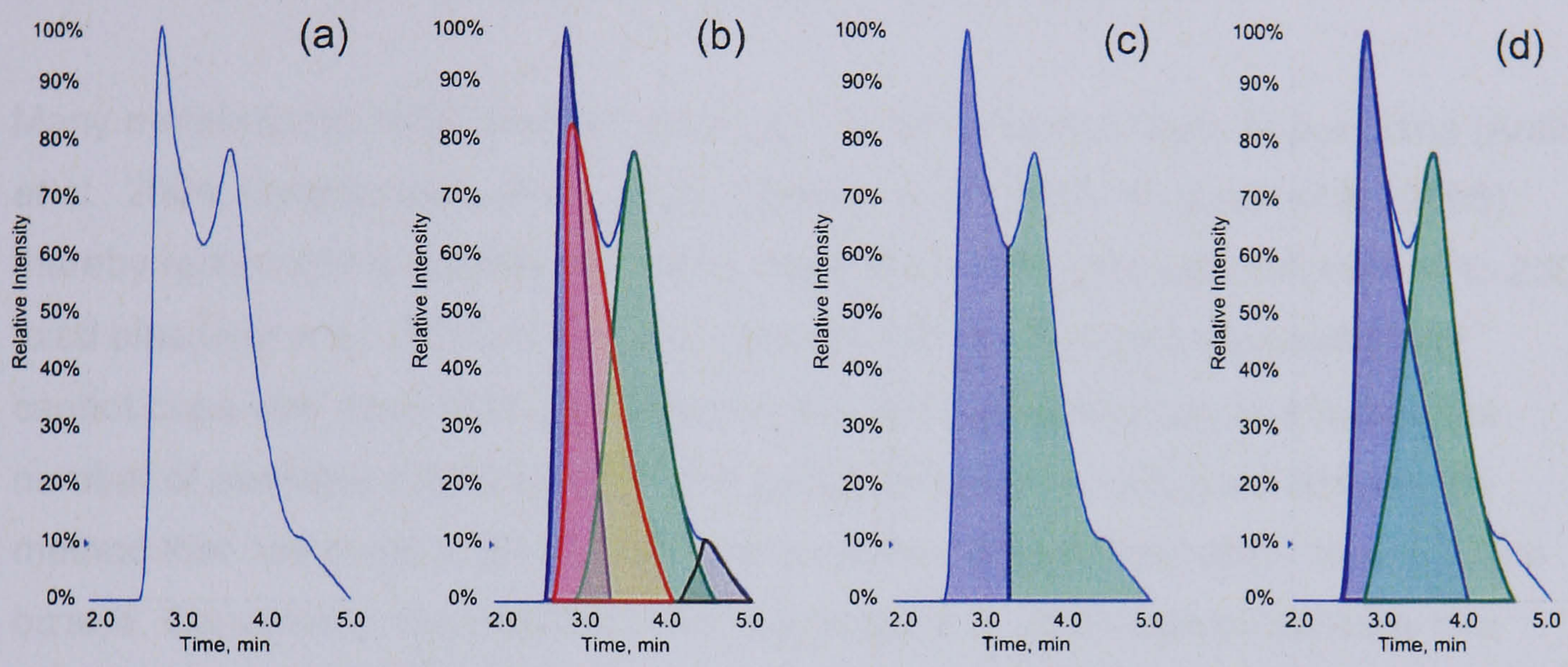


Figure 3.4.2. An extracted ion chromatogram of the peak cluster at *m/z* 114, highlighted within the previous figure. (a) XIC of *m/z* 114 from positive mode RP-LC-MS (b) Possible assignment of 4 peaks within the XIC (c) Peak area assignment not based on Gaussian peak distribution (d) Peak area assignment based upon Gaussian peak distribution.

The XIC of mass *m/z* 114 (figure 3.4.2a) does not correspond to a single, Gaussian shaped peak. There are two well definable peaks present, with a small hump on the tail of the second peak, which could correspond to another, lower intensity peak. Figure 3.4.2b illustrates how there could be up to four peaks contributing to the XIC trace shown. The most intense peak (blue) has a slight shoulder, which could correspond to another peak (red), or could just be caused by peak tailing. Extraction algorithms may not be able to distinguish between peaks of close *m/z* values and retention times (without some form of dynamic algorithm that can utilise peak shape), even when given suitable parameter values. When peaks are extracted within the metabolomics export script, it is hard to know how they are being treated; does the extraction algorithm just drop a line to the baseline when a valley is found (figure 3.4.2c), or does the algorithm try and force Gaussian style peaks (figure 3.4.2d)? Analysing the raw data does not give any clues as to which method the metabolomics export script uses, as the resulting extracted data matrix was scaled

by an unknown factor, meaning that peak areas from raw data do not directly correlate with their extracted peak areas.

One further problem that may not be addressed by many extraction algorithms for LC-MS data, is baseline shifting. Where does, and should, a baseline be when extracting data? This baseline problem has been identified and addressed for NMR data (Chang *et al.*, 2007), but does not appear to have been addressed within the literature for LC-MS studies, something which requires attention.

Many metabolomic NMR studies choose to bin their spectra into 0.04 ppm bins (Antti *et al.*, 2004; Constantinou *et al.*, 2005; Williams *et al.*, 2005; Kochhar *et al.*, 2006), thereby reducing the relative resolution. Worryingly, 0.04 ppm bins correspond to 250 fixed bins over a 0 – 10 ppm range to coincide with the fact that Microsoft Excel cannot cope with more than 256 columns; the fact that something as trivial as the number of available columns within one particular computer program dictates the method that has come to be accepted for a whole metabonomic study seems a little bizarre. Conversely, the metabolomics export script in effect utilises movable bins with no fixed *m/z* width (other than that constrained by carefully chosen parameters). As fixed bins have the potential to split peaks across two bins, novel methods such as using wavelet transforms (Davis *et al.*, 2007) which are an adaptive binning process are now being reported for use with NMR data. Many of these new applications for NMR (adaptive binning, baseline correction) should hopefully cross the boundary and find a much-needed application in LC-MS data extraction and processing.

3.4.4 Conclusions

Despite the limited choice of data extraction software available, the metabolomics export script that was used for extracting all of the data within chapters three and four currently offers the best chance of obtaining useful LC-MS metabonomic data, provided that: the parameters used are suitable for the data being extracted and that once extracted, the data is carefully scrutinised for any errors, such as those pointed out within this section. Sangster *et al.* has published work which also highlights some of the problems that have been found within this section when using the metabolomics data export script, and has shown that an improvement to the original algorithm provided better statistical results over the previous algorithm (Sangster *et*

al., 2007). Despite some improvements to data extraction algorithms, it is clear that this vital step within the field of metabonomic studies (be it NMR or MS research) requires much more attention than it is currently receiving. As the field of metabonomics expands, and a greater cross section of scientists begin to co-operate on projects (e.g. chemists, biologists and importantly, statisticians and computer scientists) it is hoped that greater emphasis is placed on correctly extracting raw data.

3.5. Multivariate data analysis

3.5.1. Introduction

Metabonomic studies which involve a reasonable sized study of ca. 100 samples generate vast amounts of high dimensionality data; this is further increased when studies utilise multiple separation methods, e.g. reversed phase and hydrophilic interaction chromatography, different ionisation methods (ESI, APCI, positive and negative ionisation), or indeed different detection platforms (MS and/or NMR). Given that resulting datasets contain large amounts of data, which may or may not be linked, it is impossible to manually interpret these vast sets of data. Whilst it is recommend that data are at the very least 'visually checked' initially, the subsequent analysis has to involve some kind of multivariate data analysis.

As modern computing power and storage capacity is cheap and easily obtainable, the field of metabonomics greatly utilises chemometrics to try and elucidate any useful information from the mountains of data that can all too easily be generated. There are many different statistical approaches that can be employed in order to analyse what has been described as 'megavariate' rather than multivariate data (Griffin and Bollard, 2004) from metabonomic experiments. The overwhelming majority of studies use descriptive and discriminative statistics, these predominantly being principal components analysis (PCA) and partial least squares (PLS) respectively (Lu *et al.*, 2006; Lutz *et al.*, 2006; Ullsten *et al.*, 2006; Gu *et al.*, 2007; Hodson *et al.*, 2007; Katajamaa and Oresic, 2007; Kell, 2007; Lenz and Wilson, 2007; Pizzolato *et al.*, 2007; Sanchez-Ponce and Guengerich, 2007; Trygg *et al.*, 2007; Zhengzheng *et al.*, 2007). There are many other statistical methods that are available (genetic algorithms and ANOVA to name but two); whichever is chosen, the user must be acutely aware that statistics can (and does) show you what you want to see.

Many metabonomic studies use both PCA and PLS without much consideration for both how to use the statistical methods, and what the resulting outputs mean. Kell summarises this rather eloquently: "*...the literature is full of complete rubbish resulting from a combination of over-optimism in the face of ostensibly positive findings, statistical ignorance and the fear of journals to scrutinise data too carefully lest they find something unpleasant...*" (Kell, 2007).

3.5.1.1. Aims

Obviously, there is no substitute for 'good' data to begin with, and it is with this in mind that great care has been taken with statistical analyses by investigating the many parameters (most of which are rarely, if ever, reported within the literature) and the effects that they have upon the resulting data. This work aimed to define optimal statistical settings using urine samples collected from healthy volunteers within the Department of Chemistry, University of York, UK. A thorough analysis of the collected data is reported in this chapter (chapter 3.6.) using the developed optimised methods presented herein.

3.5.1.2. Results

All data presented within this section (unless otherwise stated) was generated by the analysis of urine samples collected from healthy volunteers within the Department of Chemistry, University of York.

3.5.1.2.1. Initial data analysis: principal components analysis

After 'looking' at the raw data to ensure that nothing appears to be out of place (e.g. abnormal TICs or UV chromatograms), the initial steps should involve a global view of the data. That is, the data is analysed using an unbiased method. PCA is such a method as it is unsupervised; it does not require *a priori* knowledge of class belonging. The dataset(s) as a whole are considered and represented within a matrix, 'X'. PCA groups observations that contain similar variables. The greatest variation is accounted for within the first principal component (PC), with subsequent PCs accounting for the remaining variation within the data (each successive PC accounts for less variation).

3.5.1.2.2. Outlier detection

PCA is very useful for finding observations from the dataset which do not fit well with the bulk of the data. Observations that fall outside the 95 % confidence limit (represented by an ellipse on the scores plot, figure 3.5.1) on the scores plot, e.g.

point A (figure 3.5.1), are classed as outliers. Outlying points are caused by abnormalities within the data, such that the outlying observation contains variables that are not present, or are of a different magnitude to those in the bulk of the data. Outliers cannot just be removed from the dataset; there has to be valid justification for doing so first. If a point falls just outside the 95 % confidence limit, it may not be a strong outlier. Studying the DModX[1] plot (figure 3.5.2) aids in determining whether an observation is a weak or strong outlier.

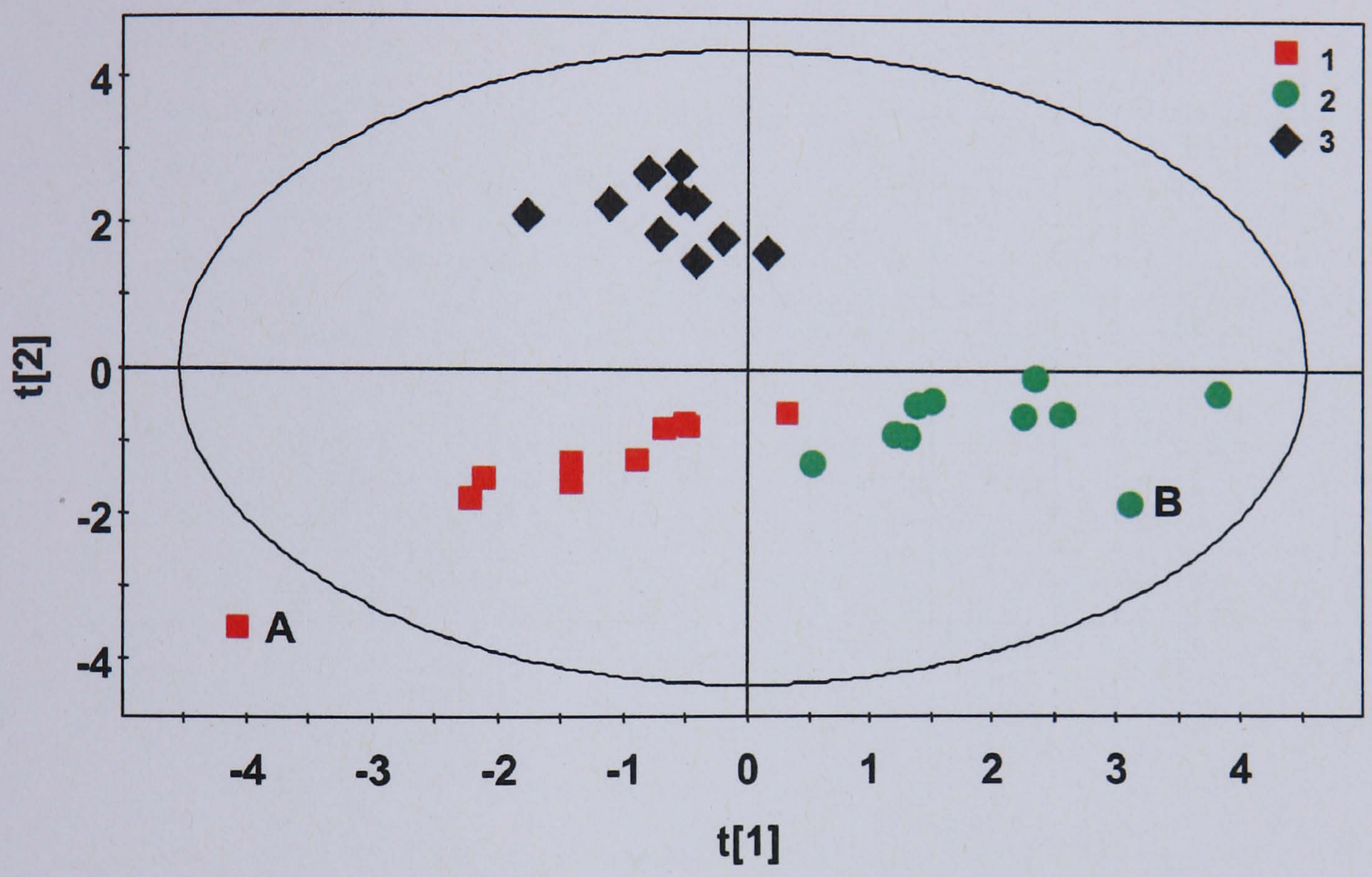
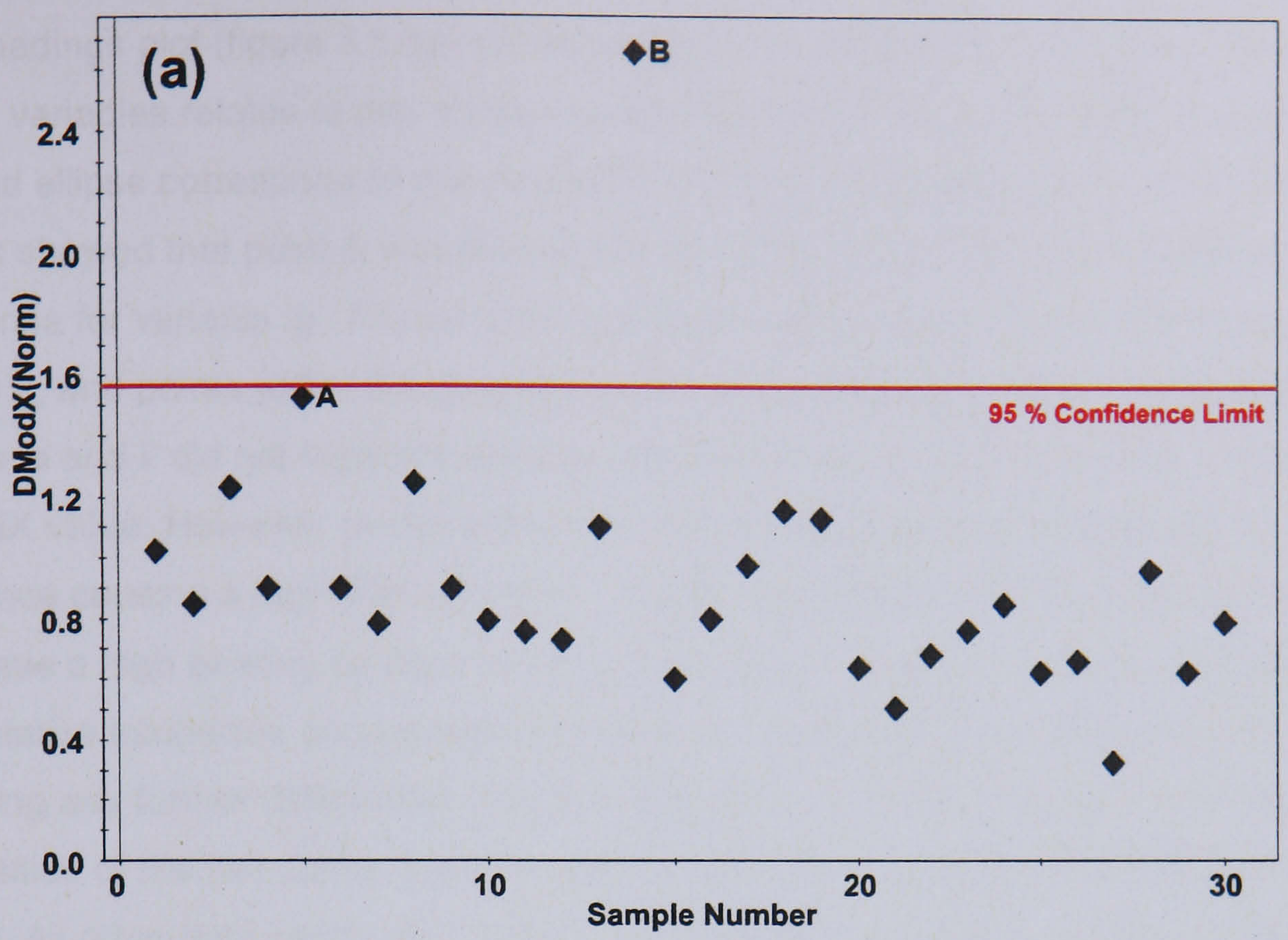


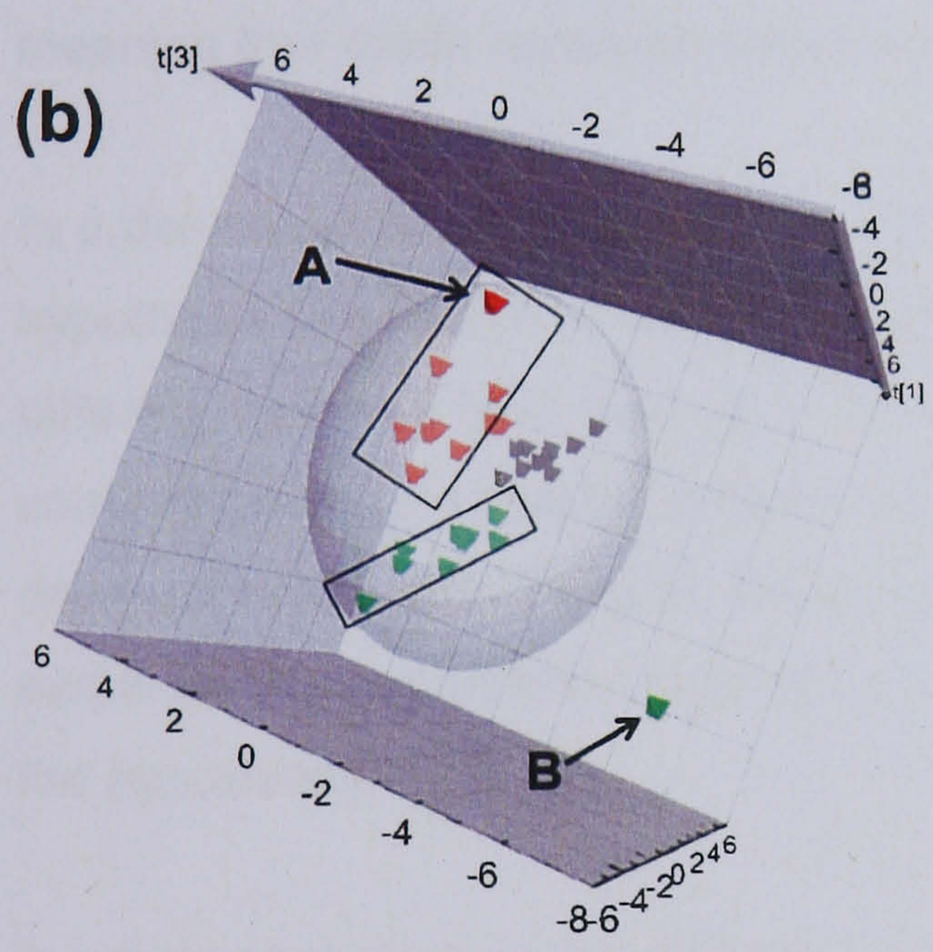
Figure 3.5.1. A PCA scores plot for three different classes of hypothetical data: 1, 2 and 3. The ellipse represents the 95 % confidence limit, any data outside this can be classed as an outlier.

Point A is an outlier according to the scores plot (figure 3.5.1), but according to the DModX plot (figure 3.5.2a) it lies in plane with the majority of the data. Whilst point B is not classed as an outlier on the scores plot, it is 'out of plane' with the majority of the data, as it has a DModX value which is higher than the critical value (95 %). Figure 3.5.2b and c help to illustrate how point A is in keeping with the data, whereas point B is not. Point A is still an outlier, but is on the same plane as the rest of group 1 (red data points). Point B now exists outside the 95 % confidence limits (grey sphere) and does not lie on the same plane as the rest of group 2 (green data points). Justification for removing outliers can only be complete when the loadings

[1] DModX is the distance each observation is from the model plane.

plot (and the raw data) has been examined to elucidate what is causing the observations to be classed as an outlier.





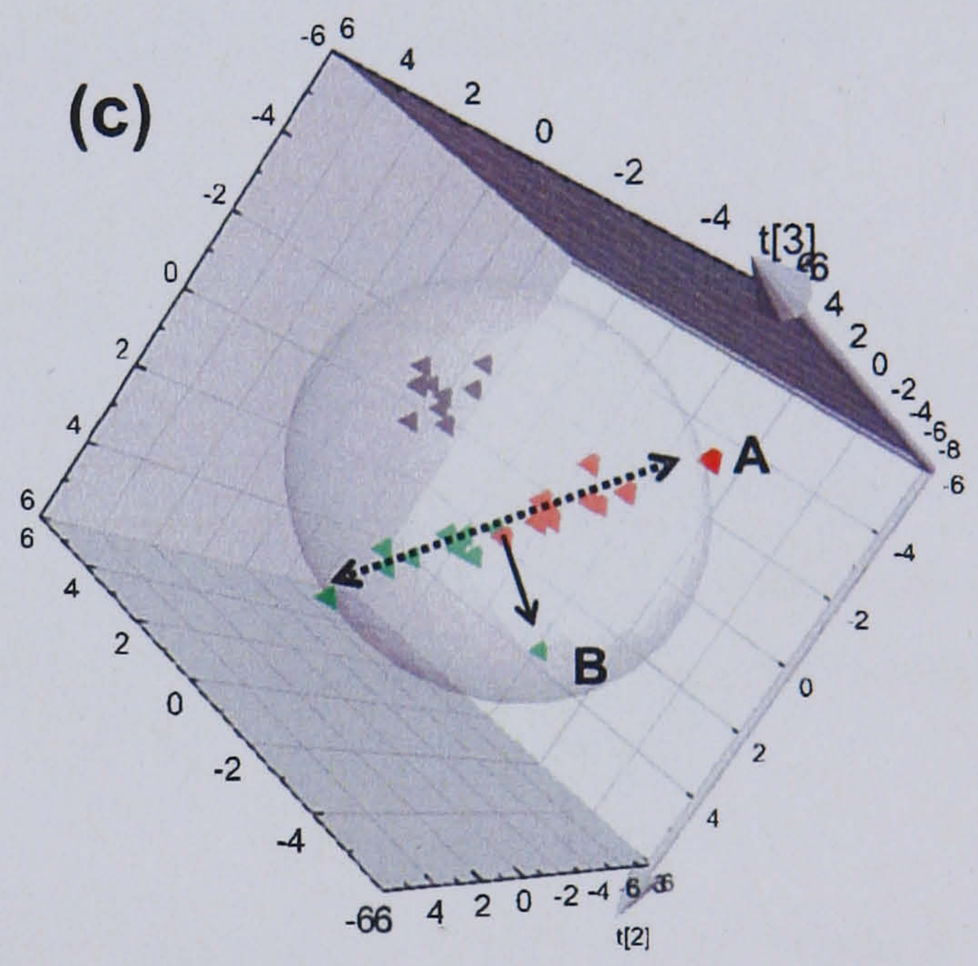
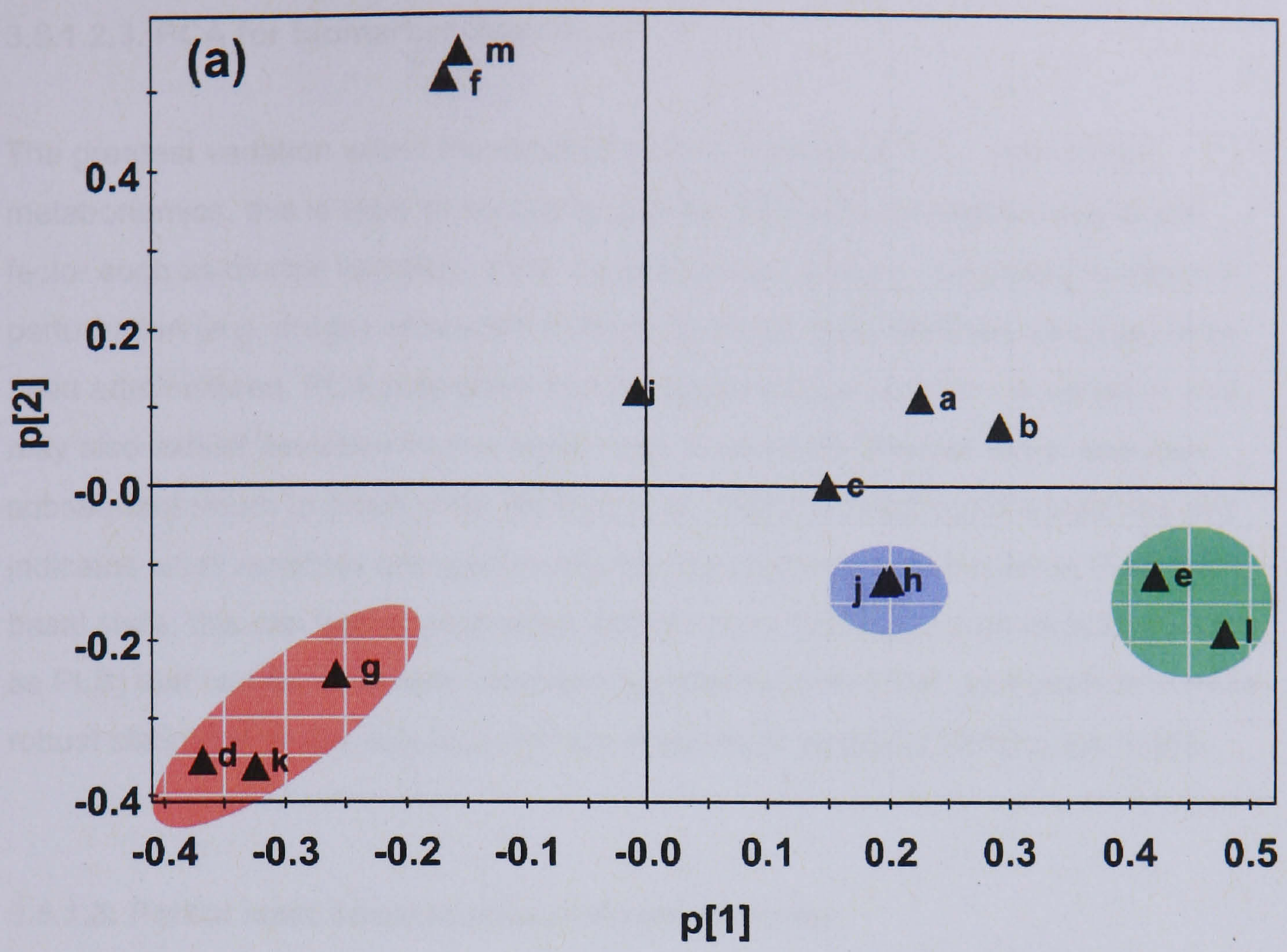


Figure 3.5.2. (a) DModX plot highlighting two points, A and B. Point B has a DModX value outside the 95 % confidence limit. (b) 3D scores plot highlighting how point A is not significantly 'different' from the rest of group 1 (red points), whereas point B has a large distance from group 2 (green points), hence the high DModX value. (c) 3D scores plot from a different angle, illustrating how point A lies 'in plane' with group 1, and how point B lies outside the plane of group 2.

The loadings plot (figure 3.5.3a) corresponds to the scores plot and shows how each of the variables relates to any clustering exhibited within the scores plot. Points within the red ellipse correspond to the clustering of group 1; analysing each of the three points showed that point A was classed as an outlier due to having an elevated response for variable '**g**'. Points within the green ellipse cause the clustering for group 2, and points within the grey ellipse the clustering for group 3. Studying the points '**e** and **l**' did not highlight any data that could cause point B to have a high DModX value. However, points within the blue ellipse only gave a response for point B, hence causing a high DModX value. The remaining points on the loadings plot do not have a high bearing on the clustering as they are close to the origin, meaning that the relative intensities across each of the three classes must be fairly even, not allowing any further differences between classes to be found. Figure 3.5.3b plots the intensities of the two points '**j** and **h**' which cause point B to have a high DModX value. All other data points do not have any response for the two variables '**j** and **h**', meaning that these variables are exclusively causing the perturbation of point B.

In order to justify removing either point A or point B, the question "what is the hypothesis in question or being generated?" needs to be asked. Point B, having different variables from the bulk of the data may derive from a sample that is contaminated, or it may be a perturbation that is relevant to the hypothesis being drawn. Point A, containing the same variables as the bulk of the dataset, may be an outlier as the perturbation could be caused by something which is not consistent with the hypothesis.

It is only once all avenues of such investigation have been followed that an outlying observation can be removed from the dataset, providing that the reasons for doing so are justifiable.



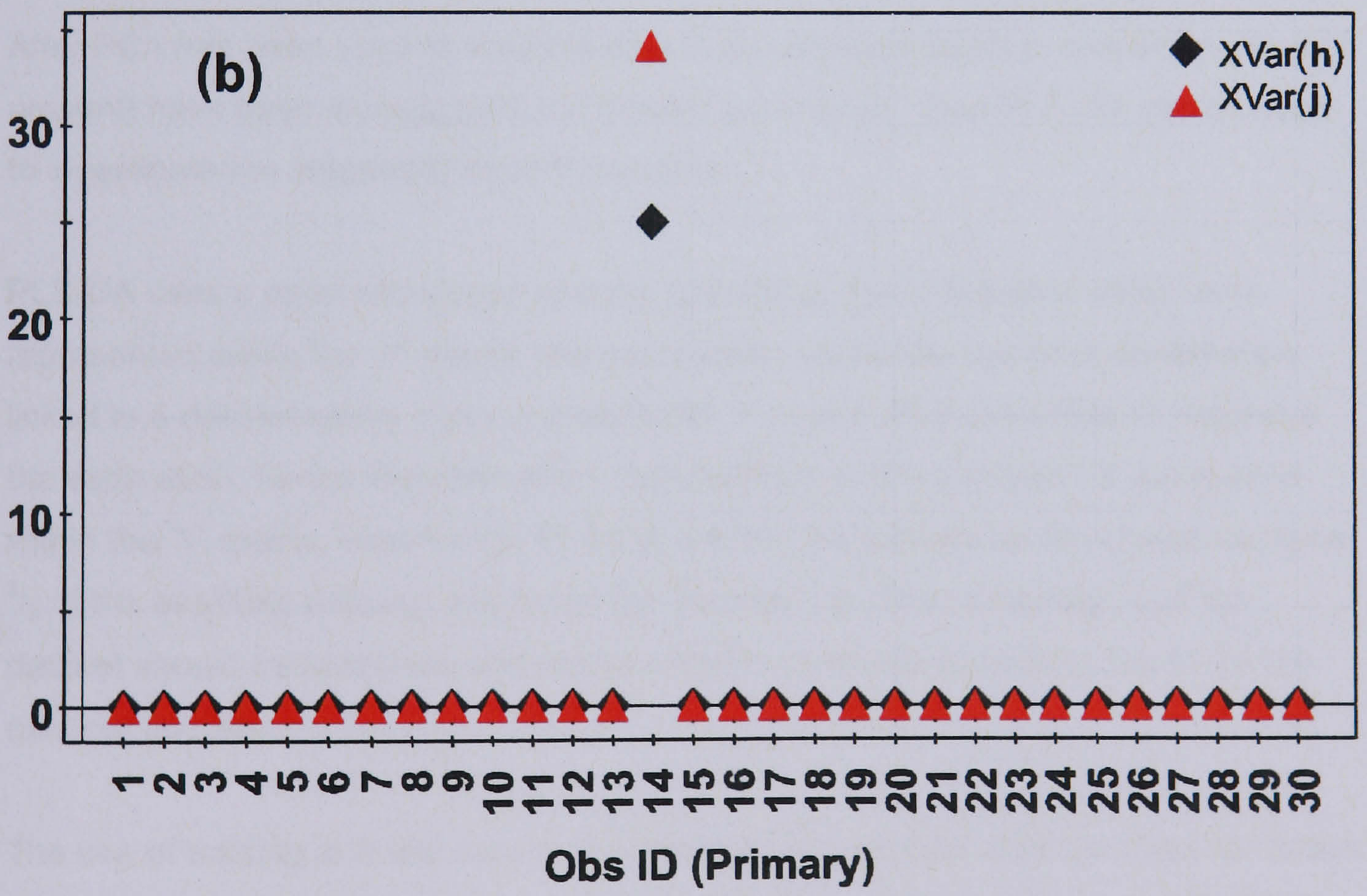


Figure 3.5.3. (a) Loadings plot explaining the clustering shown in the scores plot. (b) A plot of the intensities for two variables, **h** and **j**. These two variables cause the high DModX value for point B.

3.5.1.2.3. PCA for biomarker detection?

The greatest variation within the dataset is accounted for in PC 1. For urinary metabonomics, this is likely to be due to gender difference or another significant factor such as diurnal variation, if the samples have not been subjected to external perturbation (e.g. drugs). However, if drugs (or some other external stimulus) have been administered, PCA may show this as the greatest source of the variation and may also exhibit deviation from a basal state towards an affected state, and then subsequent return to basal state (Bollard *et al.*, 2005). Examining the loadings plot indicates what variables are responsible for any separation, or deviation from the basal state; this can lead to biomarker identification. Usually, a bias technique (such as PLS) that can be externally validated is preferred over PCA, as it leads to a more robust statistical model with less chance of spurious variables forming the model.

3.5.1.3. Partial least squares (discriminant analysis)

After PCA has been used to view the data in an unbiased fashion, and any outliers (if present) have been investigated and treated accordingly, then PLS-DA can be used to investigate the dataset(s) more thoroughly.

PLS-DA uses *a priori* knowledge of class belonging, that is the data which were represented within the 'X' matrix (the explanatory variables) has each observation linked to a discriminatory class in a separate 'Y' matrix. PLS-DA seeks to maximise the separation, hence 'discrimination', between two or more groups as designated within the 'Y' matrix. Importantly, PLS-DA models should only be developed using ca. $^2/_3$ of the available dataset; this forms the 'training' set. The remaining $^1/_3$ of the dataset should be held back and not be used in model development; this forms the external test set.

The use of training and test sets is of paramount importance. Internal cross-validation (typically venetian blind) can give models which fit the data exceedingly well; given that there may be many thousands of variables with which a model can be built, it is unsurprising to note that such models are easily obtained and should be treated with scepticism. It is only through proper model development and external validation using test sets that contain data that were not used to form the discriminative model, that potential biomarkers can confidently be obtained.

3.5.1.3.1. PLS model development – problems and solutions

In order to highlight a problem with discriminative statistics, which is one of the points which Kell makes (Kell, 2007), data from positive mode ESI-RP-LC-MS study of the urine from healthy male and female volunteers within the Department of Chemistry were analysed using PLS. The data were randomly assigned to one of six classes, meaning that the 'Y' matrix did not bear any resemblance to any 'real' possible groupings within the 'X' matrix (such as gender, diurnal variation, age etc.)

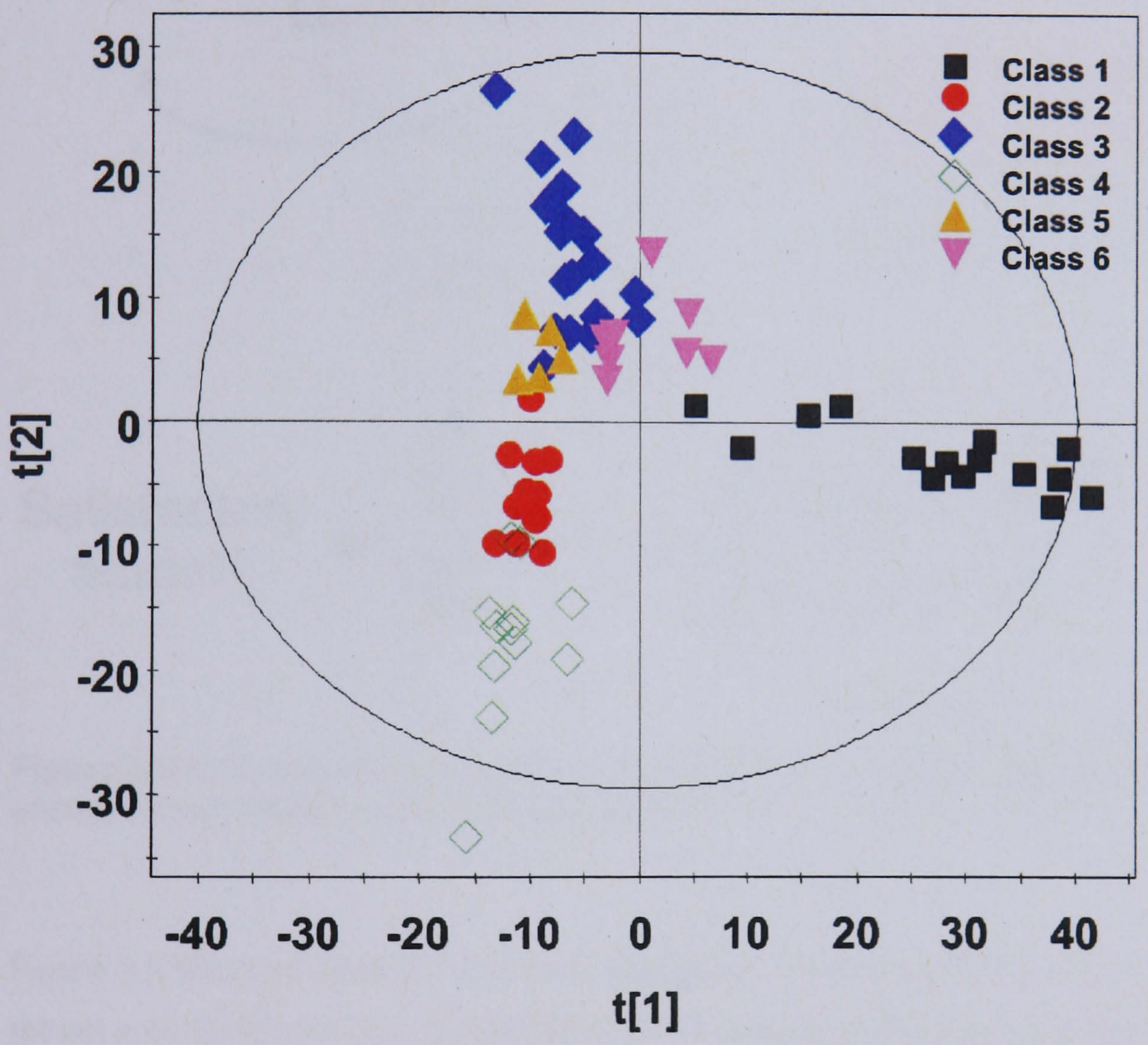


Figure 3.5.4. PLS scores plot illustrating how seemingly good models can be created given enough variables, despite there being no basis for the grouping shown.

Figure 3.5.4 shows the resulting PLS scores plot after PLS model development; internal CV values were satisfactory. Both LVs are required to separate all of the seven groups, but it does highlight the fact that a seemingly good PLS model can be built, even when there should be no relationship between the 'X' and 'Y' matrices. Given that PLS models can be developed and effectively made to show the user what they were asking, it is of paramount importance that any developed model is treated with a great deal of scepticism. The PLS scores plot in figure 3.5.4 is

overfitted as it was developed using all of the available variables from the dataset (many thousands). The variables that direct the separation shown can be listed in order of importance to the model, this is termed the Variable Importance for Projection (VIP).

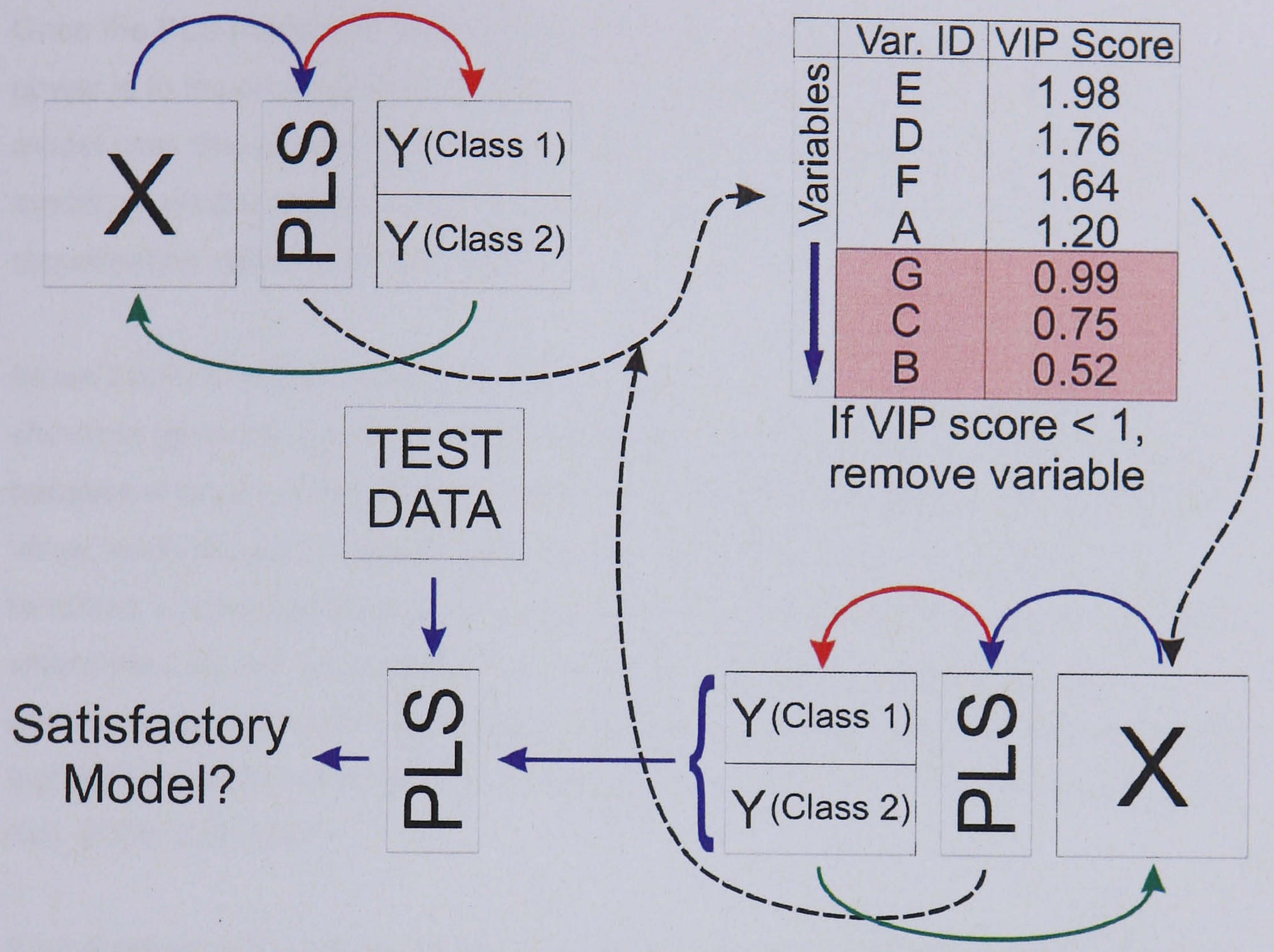


Figure 3.5.5. Schematic representing our method of PLS model development to ensure robust, reliable models are developed.

Figure 3.5.5 summarises the methods that should be followed to develop a more robust and significant PLS model. After the first development of a PLS model, the VIP list gives all of the variables with their respective score. Any variable with a VIP score of less than one should be removed from the model, as these are not considered statistically significant. A subsequent PLS model is then developed using only the remaining variables with VIP scores greater than one. The whole process is repeated until a suitable number of variables remains and the internal CV is satisfactory[1]; the number of remaining variables should be decided using 'common sense', as if too many remain, then the developed model still has a chance of overfitting (I typically aim to use less than 10 variables). As the PLS model's

[1] Defined by R^2 and Q^2 values in SIMCA. R^2 is the 'goodness' of fit, or the explained variance and Q^2 is the fraction of total variation of the X matrix that can be predicted by a component (as estimated by internal CV).

complexity is reduced by reducing the number of variables, it is also worthy to note that excessive latent variables should not be used in a model. The general rule is the fewer the better (as for PCA, the most variation should be contained within the first few components of the model).

Once the PLS model has been deemed satisfactory, the true test of its predictive power is to import a secondary dataset, the test set (figure 3.5.5). The developed model uses the allowed variables from the test set (as dictated by the variables in the model) to predict class belonging. If the model is robust, then a high external classification rate should be obtained.

When the PLS model reduction scheme (figure 3.5.5) is followed using the data shown to generate figure 3.5.4, a meaningful model cannot be developed. This is because a large number of variables were removed which, whilst having a low VIP value, were actually all being used to predict class belonging; once these were removed, a working model is no longer achieved. We are unaware of any literature where the authors have reported use of such a scheme of discriminant model development within their experimental sections, other than a few papers published to highlight the problems of poor use of statistics (Handl *et al.*, 2005; Broadhurst and Kell, 2006; Kell, 2007).

It is therefore of the utmost importance that for any discriminative study, a rigorous approach such as that highlighted above, is followed to ensure that any developed models genuinely relate to the hypothesis, rather than to 'statistical junk'.

3.5.2. Data normalisation and scaling

3.5.2.1. Introduction

Metabonomic studies that use carefully controlled animal subjects should theoretically have fewer issues with large biological variation between urine samples (than that from using human subjects). Conversely, the majority of studies where human volunteers are involved do not have such a luxury, as the volunteers generally cannot be controlled to the same degree as animal subjects. This 'large biological variation' between subjects requires to be corrected for if statistical analyses are to be less biased towards the largest biological variation; generally related to differences in the concentration of each individual urine sample.

Normalisation is another area that is usually vaguely described within the literature, if at all; data is either not normalised, scaled to creatinine or most recently to the total ion count (TIC). Of the literature which does report its normalisation methods, the most common is to use scaling to creatinine (Woitge *et al.*, 1999; Schoenau and Rauch, 2003; Husková *et al.*, 2004; Idborg *et al.*, 2004; Svoboda and Kasai, 2004; Obrant *et al.*, 2005). Normalisation to TIC stems from NMR studies where normalising data to the total signal intensity is common (Kenney and Shockcor, 2003; Antti *et al.*, 2004; Williams *et al.*, 2005). LC-MS metabonomic studies have begun to adopt this method of normalisation (Plumb *et al.*, 2003; Williams *et al.*, 2005) as a means of accounting for concentration differences as well as some forms of analytical variation.

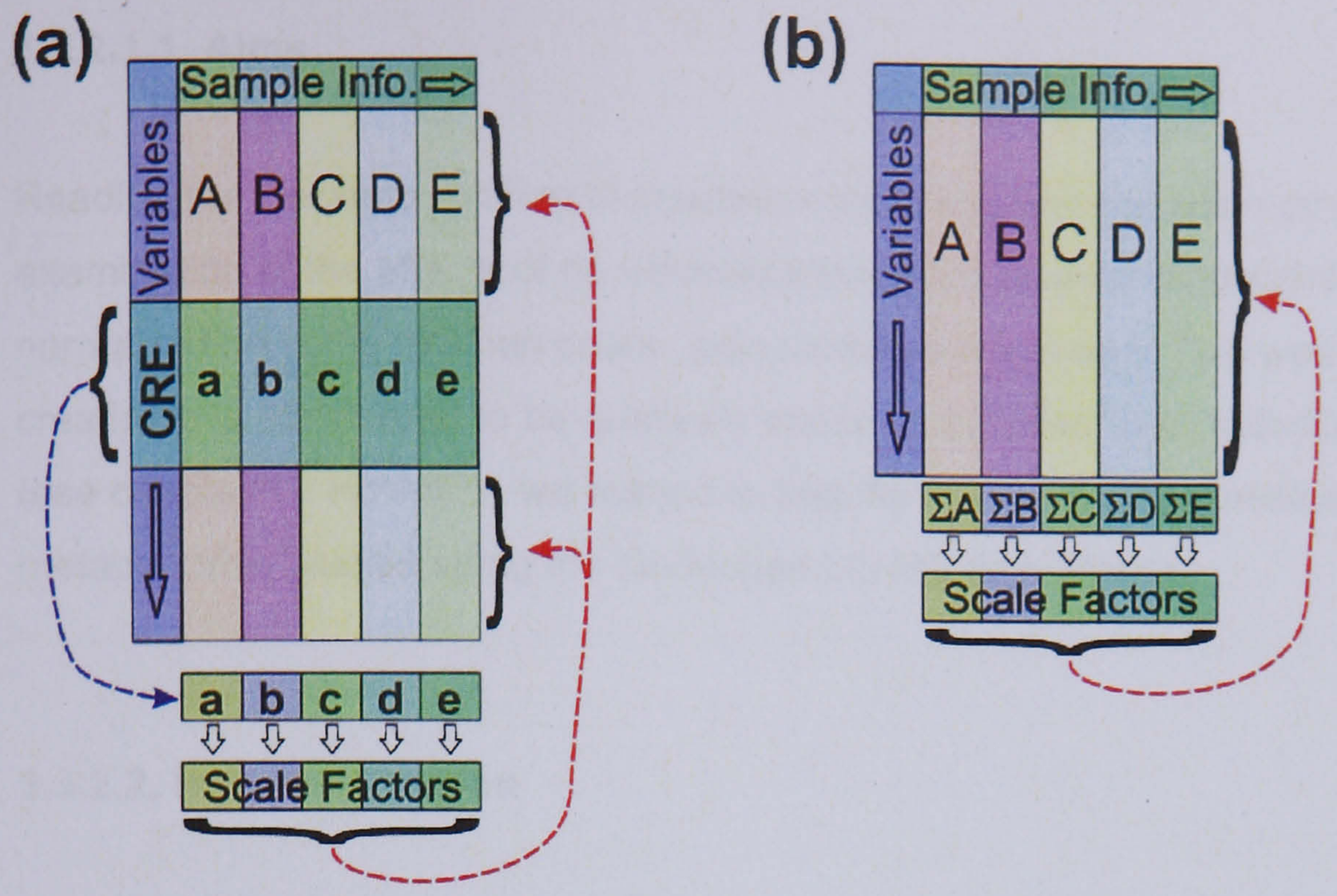


Figure 3.5.6. (a) Schematic representing how normalisation to creatinine works. (b) Schematic representation of normalisation to total ion count.

Figure 3.5.6 shows schematics of two normalisation techniques: the more common normalisation to creatinine (figure 3.5.6a), and the newer normalisation to total ion count (figure 3.5.6b). Normalisation to creatinine involves each sample's creatinine intensity being extracted (*m/z* 114.07 = $[M + H]^+$ and *m/z* 112.05 = $[M - H]^-$ with t_R for RP = 2.7 min and HILIC = 8.2 min), represented by each observations column letter in lower case with a prime (e.g. A is a'), and used to create a scale factor by scaling each observation's creatinine value to the same value. The scale factor is then used to scale each variable within the original 'X' matrix (e.g. a' scales column A variables, b' column B variables etc.). Normalisation by total ion count is undertaken using a similar method to that of creatinine normalisation. The sum of each observation's intensities (represented by ΣA, ΣB etc. in figure 3.5.6b) is used to create a scale factor, by dividing each TIC value by the largest value obtained from the sum of each observation. As with creatinine normalisation, the generated scale factors are then used to scale each variable (by the generated scale factor) within the 'X' matrix.

For both normalisation methods, the scale factors should be visually checked to first ensure that the values are roughly consistent with one another, and that there are no inconsistent values. If the scale factors are consistent, then the X matrix data can be normalised by the relevant scale factor.

3.5.2.1.1. Aims

Reading the literature relating to creatinine synthesis and excretion prompted the examination of the effects of no normalisation, normalisation to creatinine and normalisation using total ion count upon model development. This was because creatinine is considered to be relatively stable and related to an individual's mass (see chapter 1). However, we wished to test the relevance of creatinine scaling for metabonomic studies using the developed LC-MS approaches.

3.5.2.2. No normalisation

Leaving the data 'as is' and not using normalisation means that any clustering evident upon PCA, would most likely be linked to the concentration of each individual sample (dependent upon any subsequent scaling methods used, see chapter 3.5.2.4). Whilst this may be desired for some studies, it is evident that the majority of metabonomic studies require some kind of normalisation to account for any biological (and/or analytical) variation, so that any differences in the concentrations of individual urine samples are not the overriding factor directing clustering.

One method that has been used in the literature, and is technically 'no normalisation', is 24 h urine collection. The collection of each void of urine over a 24 h period should remove any effects caused by diurnal variation, and also minimise the effects of varying concentrations. Whilst 24 h urine collection has been described as the 'gold standard', it suffers many disadvantages (Heavner *et al.*, 2006). Generally, 24 h collection is completely unfeasible for a metabonomic study as it would involve volunteers/'patients' completing a full 24 h collection, where the sample integrity could be compromised by bacterial infection and incomplete sample collection for example (Woitge *et al.*, 1999; Heavner *et al.*, 2006). However, most obvious is the issue of storing many volunteers/'patients' 24 h urine samples (given that the average human produces 1-2 L of urine daily), especially for time-set metabonomic studies.

3.5.2.3. Creatinine normalisation

A thorough analysis of literature relating to normalisation by creatinine suggested that normalisation using creatinine may not have been as robust as once thought. Whilst creatinine excretion is usually consistent, some of the literature suggests that the levels of excreted creatinine are easily altered by internal and external factors such as therapeutic interventions, disease and growth (Boeniger *et al.*, 1993; Schoenau and Rauch, 2003). These findings were consistent with the worry that levels of creatinine may be easily perturbed, hence the need to further research the literature on creatinine normalisation and alternative normalisation methods.

If creatinine levels in urine are susceptible to large fluctuations caused by disease for example, then using the concentration of creatinine to normalise data would seem rather hazardous. Heavner *et al.* suggests that whilst using creatinine to normalise data appears "*...to be a valid and effective method...*", its use would not improve correlation to exposure to dose, and could in fact make it worse (Heavner *et al.*, 2006). This is backed up by research undertaken by Antti *et al.*. Antti dosed Sprague-Dawley rats with differing amounts of hydrazine[1] to aid the design of experiment and data screening for adverse drug effects (Antti *et al.*, 2004). Upon analysis of ^{1}H NMR data, the resonance at δ 3.92 increased in positive correlation with increasing hydrazine dose. This resonance corresponded to that of creatinine, showing that creatinine levels were perturbed by induced disease. This effect may not have been noticed had the data been normalised to creatinine. Heavner *et al.* also state that creatinine exhibits variability due to gender, age, muscularity, physical activity, diet, disease state, pregnancy and creatinine intake (also noted by Scheonau and Rauch (Schoenau and Rauch, 2003)).

Further to creatinine being perturbed by many factors, some literature suggests that creatinine is not as stable as once thought. Schneider *et al.* describe how after excretion, creatinine is influenced by variations in both pH and temperature (Schneider *et al.*, 2002); the levels of creatinine decreased by around 20 – 25 % within the first 12 days of urine storage, stabilising after this initial period. This has also been noted by Saude and Sykes (Saude and Sykes, 2007) who also suggested that many urinary metabolites' concentrations fluctuate within the first 7 – 14 days of storage at either – 20 or – 80 °C, again stabilising after this period.

---

[1] N_2H_4: a hepatotoxin which induces steatosis, lipid retention.

With the literature proving that using creatinine to normalise urinary metabonomic data is not ideal, it is clear that alternative methods should be utilised. A set of experiments to compare no normalisation, normalisation to creatinine, or to total ion count were designed. These results are discussed in section 3.5.2.5.

3.5.2.4. Scaling techniques

In addition to normalisation, some kind of scaling should also be applied post normalisation. Scaling is typically used to account for the large range in the concentrations between metabolites excreted in urine. Metabolites with a large concentration are not necessarily of greater importance than metabolites with a significantly lower concentration; without some kind of scaling, the dominant features (metabolites (variables) with large concentrations) would dominate any statistical model developed.

A literature search highlights that scaling is another area of metabonomic studies that are scantily or not reported. Only van den Berg *et al.* (van den Berg *et al.*, 2006) has conducted a study into different scaling methods for metabonomic / genomic data analysed by PCA. The general conclusions from van den Berg *et al.* can be applied to PLS, but this author suggests it is an area that requires more consideration from both statisticians and biologists working together.

We chose to evaluate three scaling techniques (those being the only methods available in SIMCA P+ v11.5, Umetrics, Sweden): mean centring (ctr), pareto (par) and unit variance (UV).

$$\text{ctr} = \left(x_{ij} - \bar{x}_{ij}\right) \qquad \textbf{Equation 3.5.1}$$

$$\text{par} = \frac{\left(x_{ij} - \bar{x}_{ij}\right)}{\sqrt{SD}} \qquad \textbf{Equation 3.5.2}$$

$$\text{UV} = \frac{\left(x_{ij} - \bar{x}_{ij}\right)}{SD} \qquad \textbf{Equation 3.5.3}$$

Equation 3.5.1 represents how mean centring works. Each variable, x_{ij}, has the mean, $\bar{x}_{ij}$, for that variable subtracted from itself. This has the effect of converting the mean of all of the data to zero, centring the data around the origin in PCA and PLS analyses. Mean centring is not a true scaling technique, because the variation is only between the samples themselves with no variables being scaled. However, mean centring forms part of both pareto and unit variance scaling, as the data is first centered.

Pareto and unit variance (equations 3.5.2 and 3.5.3 respectively) both utilise mean centring and also the standard deviation of each variable. Standard deviation is a measure of the spread of the data. As pareto uses the square root of the standard deviation, it reduces the effect of large intensities more than small, therefore making intense variables less dominating. Unit variance scales each variable to have a standard deviation of one (dividing each sample's response for a variable by the standard deviation for that variable), meaning that any differences are based upon correlations within the data, as all variables are now equally important.

Pareto, and especially unit variance, requires normalised data to have been carefully extracted (from the raw LC-MS data). If any 'noise' has been extracted then these scaling methods would enhance this.

3.5.2.5. Results

In order to evaluate each normalisation and scaling technique (no normalisation, normalisation to creatinine, normalisation using total ion count, mean centring, pareto and unit variance), we used datasets generated from urine samples collected from volunteers within the York Chemistry Department, who considered themselves to be generally fit and free from disease. As mentioned earlier, creatinine has been shown to be easily perturbed by illness or other states; it is with this in mind that we chose people who should have been free from anything that would have perturbed the basal levels of creatinine excretion, hence allowing a comparison using normalisation to creatinine.

For both positive and negative modes of ESI, and also for two complementary separation methods (RP and HILIC, the development of which is described later in chapter 3.5.6), discriminative PLS models were developed and optimised according to the methodology outlined earlier (figure 3.5.5). The optimised models were then evaluated with ~ $^1/_3$ of the initial data which were held back to form an external test set.

PCA was used to first analyse each dataset to detect any outliers; no data points were considered to be outliers based upon the DModX values of any points outside the 95 % confidence limit, thus all data were retained for PLS analysis. Table 3.5.1 shows all of the methods used to develop models for datasets obtained using separation techniques, ionisation modes, normalisation methods and scaling methods. For this study, gender was used as the discriminatory variable within the Y matrix. Gender was chosen as it should be the largest discriminatory variable amongst healthy volunteers, and therefore allow the generation of robust PLS models.

Table 3.5.1. All methods used to develop PLS models (totalling 36) to allow the comparison of different normalisation and scaling techniques.

Separation Method	Ionisation Mode	Normalisation Technique	Scaling Method
Reversed Phase	Positive	None	ctr
			par
			UV
		Creatinine	ctr
			par
			UV
		Total Ion Count	ctr
			par
			UV
	Negative	None	ctr
			par
			UV
		Creatinine	ctr
			par
			UV
		Total Ion Count	ctr
			par
			UV
HILIC	Positive	None	ctr
			par
			UV
		Creatinine	ctr
			par
			UV
		Total Ion Count	ctr
			par
			UV
	Negative	None	ctr
			par
			UV
		Creatinine	ctr
			par
			UV
		Total Ion Count	ctr
			par
			UV

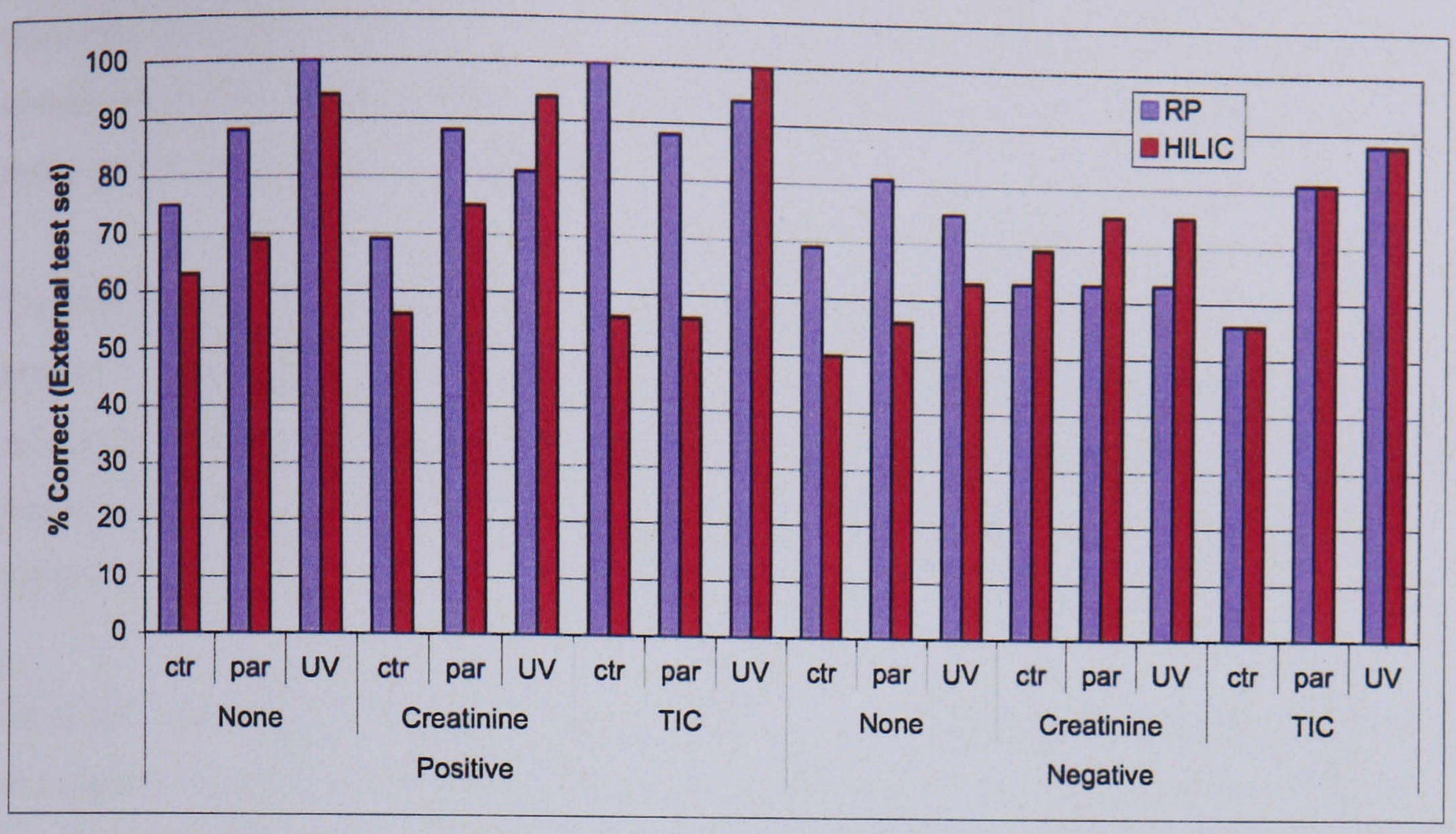


Figure 3.5.7. Graphical representation of external classification results for all 36 developed PLS models (as shown in table 3.5.1).

The urine samples from the Department of Chemistry were analysed using both positive and negative ionisation mode RP- and HILIC-ESI-MS (discussed later within the RP and HILIC development section, chapter 3.5.6). After each dataset (±RP and ±HILIC) was extracted into an X matrix, they were left as is, normalised by creatinine or normalised to total ion count before being imported into SIMCA. Each dataset was then scaled by one of three methods (ctr, pareto or UV) to generate a discriminative PLS model based upon gender using the earlier developed methods (figure 3.5.5). Once satisfactory PLS models had been developed, external test sets were imported to evaluate the external classification rates, and therefore the accuracy of the developed models.

Comparing the external classification results as percentages across all 36 differently developed models shows that each different method used gave a different external classification result (figure 3.5.7). Analysing the data for each different scaling method for all modes of normalisation shows the general trend of ctr < par < UV with respect to the observed classification results. This may be caused by the fact that UV gives each variable an equal weighting, meaning that the models have in effect access to lower intensity metabolites, whereas par and ctr do not. The three normalisation methods (none, creatinine and TIC) all show comparable external classification results for each of the three scaling methods, with normalisation to TIC performing the best when combined with UV scaling. Evaluating positive and

negative ionisation modes across all normalisation, scaling and separation methods shows us that positive ionisation mode generally allowed the development of more robust PLS models for discrimination by gender.

The results presented here are in agreement with that of van den Berg *et al.* who stated "*...autoscaling and range scaling seem to perform better than other methods...*" (where autoscaling is analogous to UV). I also agree that "*...data pre-treatment is often overlooked or is applied in an ad hoc way...*" (van den Berg *et al.*, 2006).

As-such, care should be taken when applying scaling methods; the results obtained should be carefully scrutinised to ensure that variables with the highest VIP values are determined and identified so that they can be linked to the question being asked (this is discussed later in much greater detail within the RP and HILIC method development section, chapter 3.5.6).

Analysis of the normalisation results for both modes of ionisation shows that no normalisation led to classification rates that were broadly comparable to normalisation to creatinine or TIC. This shows that for discrimination by gender, concentration differences do not play a major role. Providing that the volunteers were not ill or diseased, then the results for normalisation by creatinine were very comparable to TIC normalisation. This was very promising as it showed that if, for this example, normalisation to TIC was comparable to normalisation to creatinine, TIC is therefore a suitable normalisation method that can be used for normalising urinary metabonomic data. Normalisation to TIC should remove any problems that are associated with normalising to creatinine (perturbations due to illness, disease etc.), which could affect studies such as the clinical study (chapter four) which was the ultimate overall aim of the work described in this thesis.

3.5.2.6. Conclusions

It has been shown that careful consideration has to be given to how data are manipulated prior to statistical analysis. Given that many papers confirmed our suspicion that normalisation to creatinine should no longer be considered the 'gold standard', we have also shown that when data are normalised using TIC, it is highly comparable to normalising to creatinine (at least when there are expected to be no

perturbations to the creatinine levels). Our results showed that scaling by UV was most favourable for metabonomic studies, in agreement with research by van den Berg *et al.*.

Overall, for either positive or negative ESI, we suggest that normalisation using TIC followed by scaling using unit variance should be employed for urinary metabonomic studies. However, other scaling methods should not be discounted and should always, at the very least, be considered during statistical analysis.

3.5.3. Data fusion

3.5.3.1. Introduction

The goal of a true metabonomic study should be to encompass a representative 'fingerprint' which contains the largest amount of information that is available. Studies may collect NMR data, along with LC/GC/CE-MS data. Mass spectrometric data should at the very least contain data from positive and negative ionisation modes, but ideally data from more separation methods (e.g. RP and HILIC for LC) or different ionisation methods (e.g. ESI and APCI).

Typically, each data cohort (i.e. RP+, RP-, HILIC+, HILIC- for ESI-LC-MS) is statistically analysed as a separate entity. Whilst this is a necessary step, the data should also be analysed as a whole. Data fusion has been described in a handful of papers (Idborg *et al.*, 2005; Forshed *et al.*, 2007a; Forshed *et al.*, 2007b) but has yet to find a place as a mainstream data analysis method within metabonomic studies.

3.5.3.2. Aims

The aim of this 'data fusion' study was to examine how fusing together four different LC-MS datasets (RP+, RP-, HILIC+ and HILIC-) would effect PLS model development. Each individual data cohort, once optimised, produces a list of VIP values for the variables forming the discriminative model. Developing a 'fused' model involving several datasets should produce robust statistical models, as all variables across all datasets can now contribute to forming a single PLS model.

3.5.3.3. Results

Four sets of data were obtained by analysing the urine from healthy volunteers from the department, by RP and HILIC LC-MS in both positive and negative ESI modes. Each dataset was individually exported into Excel as a text file (in ASCII format) (Microsoft Excel 2004 for Mac) and normalised to TIC as described (chapter 3.5.2). Figure 3.5.8 illustrates how the individual datasets were concatenated together. All of the observations were ordered and aligned so that each column corresponds to data from the same urine sample. The variables for each dataset were labelled to identify

which dataset they belonged to (i.e. all variables from RP+ were subsequently labelled with '_RPpos'). The datasets were then concatenated, one below the other, to form one single X matrix of the data.

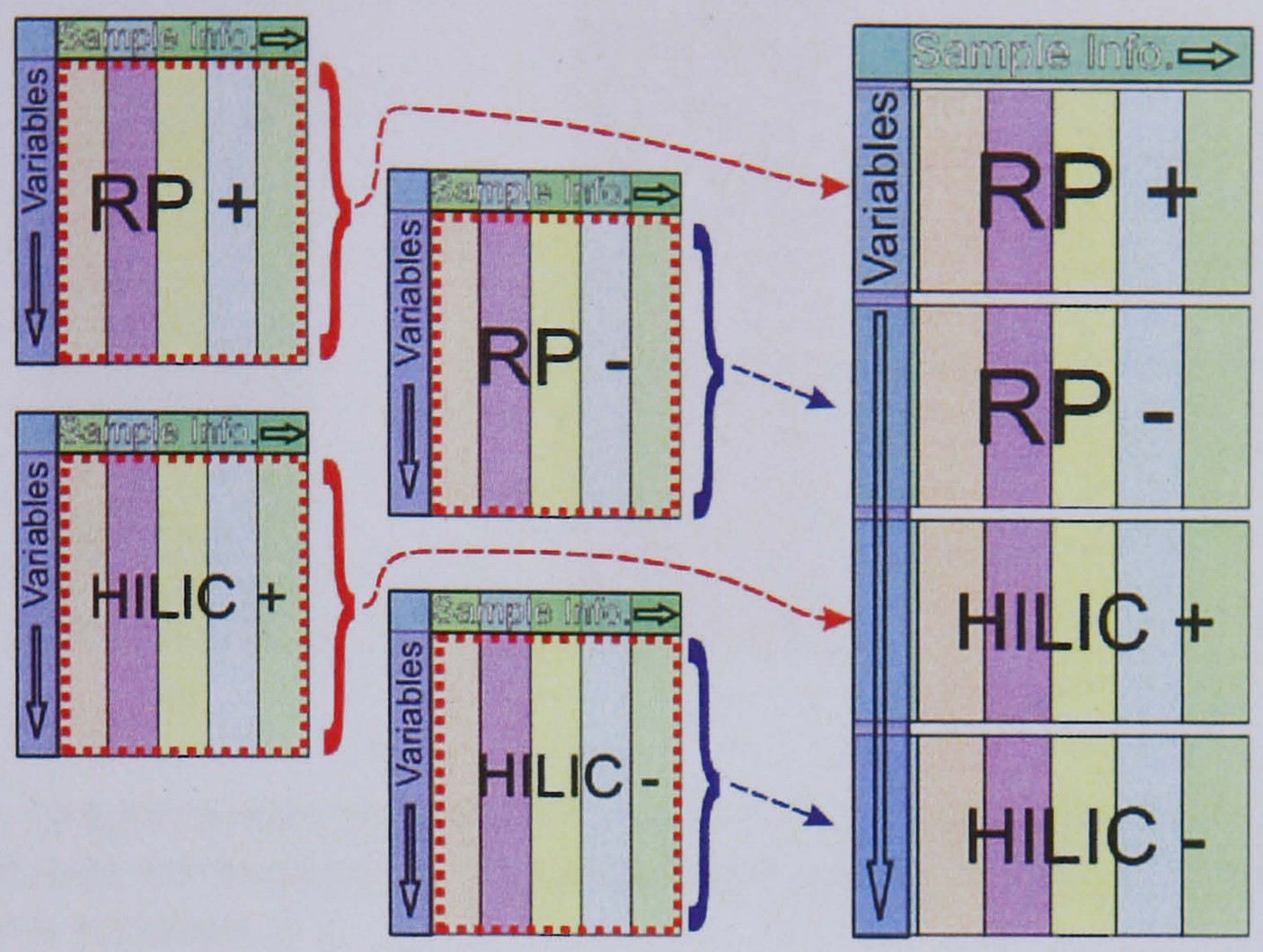


Figure 3.5.8. Representation of data fusion of four different LC-MS datasets.

Before PLS analysis, the data were analysed by PCA to determine if there were any anomalous data points. As the data appeared to be satisfactory, response variables based upon gender discrimination were assigned to the Y matrix. As the PLS model contained four times the usual number of variables, it was crucial to ensure that the model was optimised according to the methodology previously laid out (figure 3.5.5).

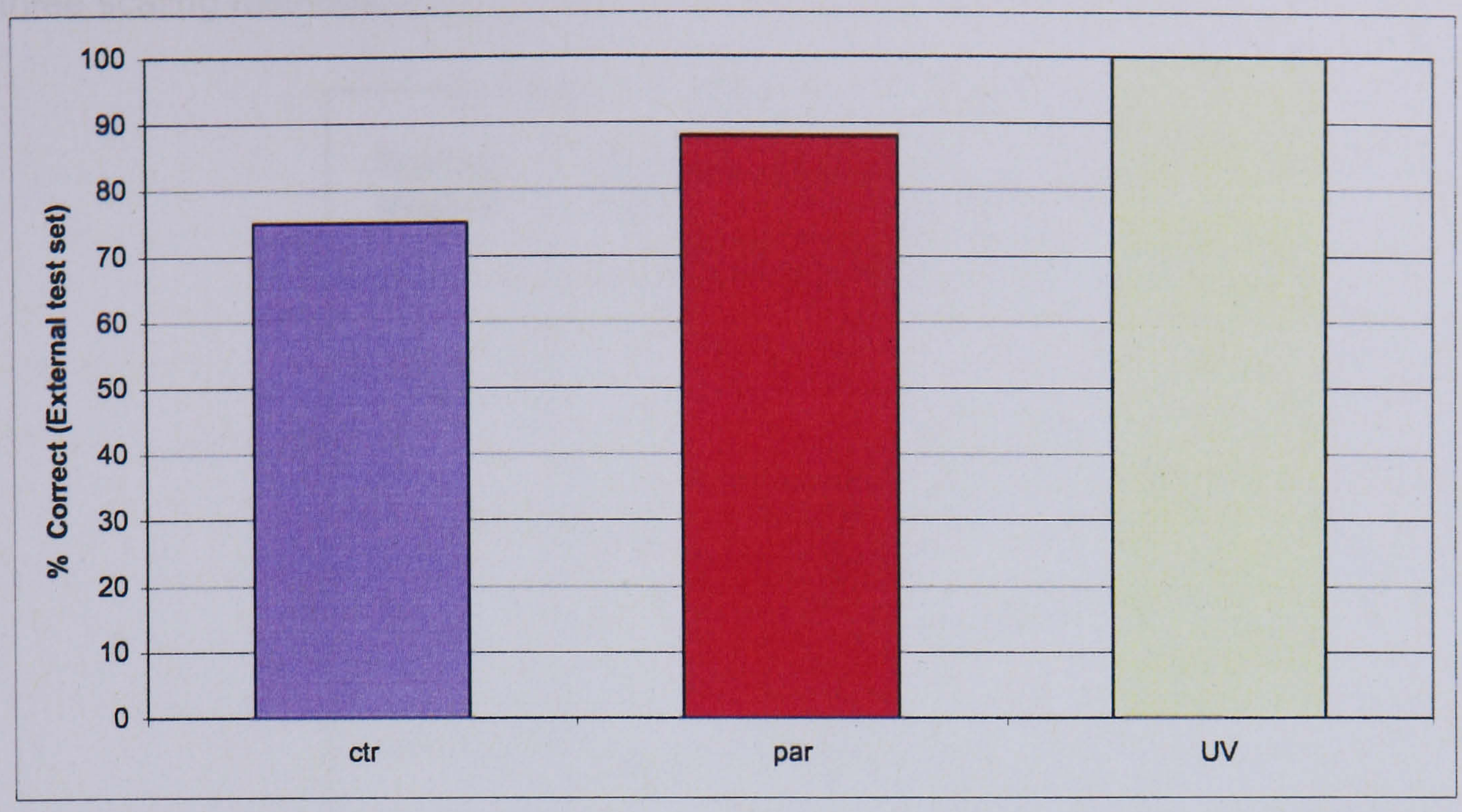


Figure 3.5.9. Graphical representation of external test set classification results for concatenated data normalised by TIC, using three scaling methods: mean centering, pareto and unit variance.

The concatenated dataset was imported into SIMCA for analysis by PLS using three scaling methods (mean centring, pareto and UV). When satisfactory PLS models had been developed, external test sets were imported to evaluate the predictive ability of each of the three developed PLS models. What we see is that the external test set classification results (figure 3.5.9) from the optimised PLS models show that scaling to unit variance once again gave the highest classification rate, followed by pareto and then mean centring. UV scaling was expected to have the highest classification rate as each variable is given the same statistical significance, meaning that at the very least, the model should perform as well as the best single PLS model developed for the non-concatenated model. As the concatenated model contains information from two different separation methods and polarities, such an unbiased scaling method is required; each different analytical method may produce variables of varying intensities, which are not comparable between analytical methods.

Table 3.5.2. Comparison of the top five variables forming each PLS model for all three scaling methods.

Scaling Method	Rank (VIP)	Separation Method	Polarity
ctr	1	RP	+
	2	RP	-
	3	RP	-
	4	HILIC	-
	5	RP	+
par	1	RP	+
	2	RP	+
	3	RP	-
	4	RP	-
	5	RP	-
UV	1	HILIC	+
	2	HILIC	-
	3	HILIC	+
	4	RP	-
	5	RP	-

Comparing the top five variables as determined from their VIP values for each scaling method, showed that a mixture of variables from across each of the four datasets formed the model (table 3.5.2). For scaling by mean centring and pareto, separation by reversed phase gave the top five variables (other than position four for mean centring which used a HILIC variable), with the positive ionisation mode providing the most important variable. This is interesting as it suggests that separation by reversed phase may give rise to variables that are more intense, and therefore more discriminating than HILIC variables – something which pareto scaling would enhance. Scaling by unit variance gave five completely different variables. The first three variables were from HILIC separation, with the last two from reversed phase separation. Once again, the top variable was from positive ionisation mode.

A more in-depth discussion of these data fusion results, along with each variable's *m/z* and t_R values, is presented in chapter 3.5.6 where the development of RP and HILIC LC-MS methods are discussed, along with a thorough analysis of the data used to develop these models.

3.5.3.4. Conclusions

Whilst much information can be obtained through the statistical analysis of individual datasets, the concatenation (or fusion) of multiple datasets can enhance information obtainable from metabonomic studies. Even though the data presented here were purely LC-MS data, this additional method of data analysis is not exclusive to this analytical platform. Forshed *et al.* has evaluated different techniques for the fusion of ^{1}H-NMR data with LC-MS data (Forshed *et. al.,* 2007a).

Despite the example for data fusion using gender as a discriminatory factor which was easily modelled, we predict that data fusion may be of a greater use for more complex models where changes may be more subtle, or linked to a series of different compounds which are only detectable across different analytical platforms. Fusing the data and applying discriminative statistics may enable compounds which, in a single dataset model, would not form a robust model, to be combined with compounds from other datasets with similar poor classification results to form a single, more robust model.

We feel that as the field of metabonomics progresses, data fusion should become a more common feature within the published literature.

3.6. Development of reversed phase and hydrophilic interaction liquid chromatographic methodologies for mass spectrometric metabonomic studies of urine

3.6.1. Introduction

Metabonomics comprises a suite of 'omic' technologies and has seen substantial growth in recent years, perhaps due to its success within the pharmaceutical industry, where metabonomics is now used for the identification of potential markers of disease, efficacy and toxicity (Nicholson *et al.*, 2002; Drexler *et al.*, 2004; Lindon *et al.*, 2004; Walgren and Thompson, 2004; Wilson *et al.*, 2005; Robertson *et al.*, 2007).

It cannot be stressed enough, that a 'true' metabonomic study should involve a comprehensive analysis with no pre-selection of analytes, in order to obtain as much information as possible. ^{1}H-NMR spectroscopy is a non-discriminatory analytical technique (provided that there is a proton!) and provides information about all of the metabolites above the limit of detection (LOD) within a biological sample. However, sensitivity is a problem in NMR, and metabolites present in low concentrations may not be detected. Mass spectrometry, on the other hand, has lower LODs (orders of magnitude) but is a more selective technique. When used in conjunction with HPLC, which is required to provide separation of the components within a chosen biofluid and to reduce selectivity, mass spectrometry allows detection and quantification of low-level metabolites (Dettmer *et al.*, 2007).

A comprehensive LC-MS study should utilise both positive and negative ionisation modes with a chromatographic method that allows the retention and separation of as many components as possible. However, many LC-MS metabonomic studies only use reversed phase chromatography, which instantly discriminates against highly polar analytes. Despite the fact that only non-polar and mildly polar analytes are retained, RP-LC-MS is still the most widely used metabonomic MS platform (Plumb *et al.*, 2005; Williams *et al.*, 2005; Wilson *et al.*, 2005; Lu *et al.*, 2006; Lutz *et al.*, 2006; Sumner, 2006; Tang and Wang, 2006; Hodson *et al.*, 2007; Robertson *et al.*, 2007). As urine is predominantly aqueous, a significant proportion of the content is likely to be highly polar, and would typically be unretained using a traditional RP approach, and thus not contribute to the data obtained.

Hydrophilic interaction chromatography (HILIC) is analogous to normal phase chromatography in that it utilises a polar stationary phase, allowing the retention of polar analytes (Hemström and Irgum, 2006). However, unlike normal phase, HILIC allows the use of aqueous solvents, making this separation technique compatible with ESI-MS. In direct contrast to RP-LC, gradient elution HILIC begins with a low polarity organic solvent and elutes polar analytes by increasing the polar aqueous content. Compounds are retained by partitioning into a water rich layer which is partially immobilised on the stationary phase. MS compatible buffers are typically used to reduce any undesirable electrostatic interactions between the analytes and stationary phase (Hemström and Irgum, 2006).

It is because of HILIC's compatibility with MS and ability to retain polar compounds, therefore possibly increasing the coverage of urinary compounds, that it was chosen to assess its suitability for application in LC-MS based metabonomic studies.

3.6.2. Aims

The aim of this work was to analyse urine collected from fit and healthy members of the Department of Chemistry, University of York, comparing a traditional RP-LC-MS approach with HILIC-LC-MS, using both positive and negative electrospray ionisation modes. Both RP and HILIC gradients were optimised to obtain a good separation of compounds over the analysis time. The resulting data were analysed by PCA and PLS to visualise the information-rich data, using the techniques developed and described earlier in this chapter. PLS discriminative models were developed and classified for gender, time of collection and age in order to compare the performance of HILIC-LC-MS with the traditional RP-LC-MS approach. Further to the analysis of each individual dataset (RP and HILIC in both ionisation modes), data fusion was performed to evaluate its potential for generating robust models based upon a more complete dataset.

3.6.3. Results and discussion

3.6.3.1. RP and HILIC gradient optimisation

So that a direct comparison could be made between reversed phase and hydrophilic interaction liquid chromatography separation methods, each column's dimensions were identical at 4.6 x 100 mm, and each gradient's total run time was set to 30 min. This allowed comparisons to be made under optimal developed conditions for each column. As RP-LC is the most commonly utilised separation method for LC-MS metabonomic studies, there are many gradients described in the literature that have been optimised. A stepwise gradient from 0 % MeCN, increasing to 20 % then 95 % before returning to starting conditions appears to be one of the most common gradients described in the literature (Granger *et al.*, 2003; Plumb *et al.*, 2005; Gika *et al.*, 2007); this gradient scheme was modified slightly to avoid a completely aqueous mobile phase, thus increasing ionisation efficiency and reducing viscosity (therefore reducing the back pressure), and is presented in table 3.6.1.

Table 3.6.1. Gradient profile for RP-LC-MS metabonomic studies. Solvent A = H_2O with 0.1 % v/v formic acid; B = MeCN with 0.1 % v/v formic acid. The additional steps from 30-35 min correspond to column washing and re-equilibration steps.

Time min^{-1}	% A	% B
0	95	5
9	80	20
21	5	95
24	5	95
27	95	5
30	95	5
31	0	100
32	0	100
34	95	5
35	95	5

30 min total run times were chosen to allow for as much separation as possible whilst allowing for a reasonable sample throughput. Obviously, non-academic institutions

would require much shorter analysis times, as very high throughput experimentation is demanded; in industry, time is money, whereas in academia, time is the students.

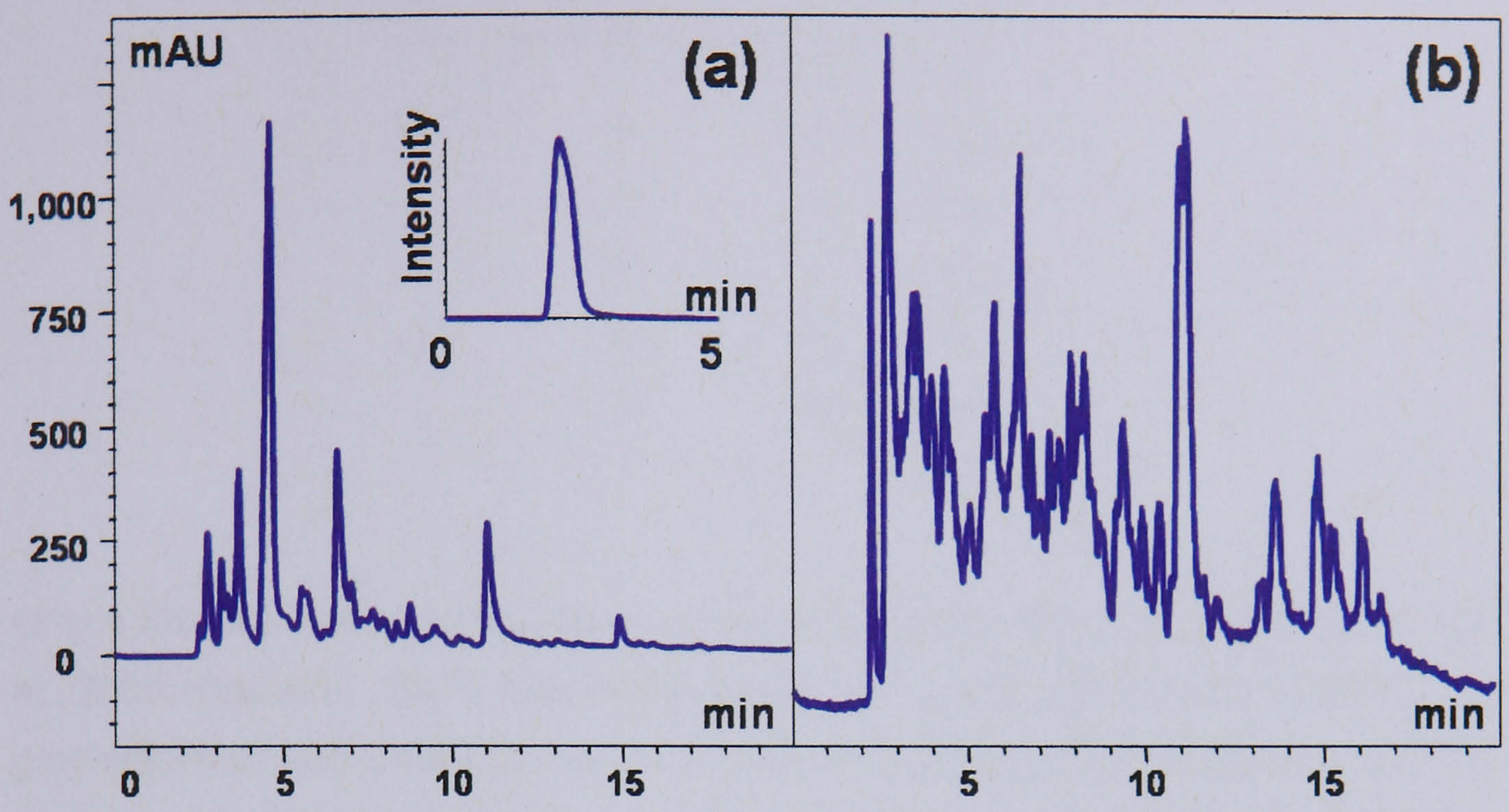


Figure 3.6.1. (a) Typical UV_{254} chromatogram obtained using RP separation of a urine sample. Inset shows an XIC for creatinine (*m/z* 114.07 = $[M + H]^+$ with a retention time of 2.1 min. (b) Positive mode TIC of the same urine sample, normalised to the most intense peak.

A typical RP-LC UV chromatogram is shown in figure 3.6.1a for the separation of a urine sample. Despite many urinary components not having a chromophore, meaning that UV chromatograms do not contain many peaks, there are some intense peaks that are Gaussian shaped present within the chromatogram. The corresponding TIC (figure 3.6.1b) is much more information rich, showing many more peaks than the UV chromatogram. Changing the gradient profile did not show any appreciable change to either the UV chromatogram or the TIC, suggesting that the gradient was suitable for the separation of urinary components. Both the UV chromatogram and TIC return to the baseline at the end of each run (only 20 min of gradient shown to highlight the information rich sections). The high organic content wash at the end of each run (table 3.6.1) was applied to remove any highly hydrophobic components, avoiding late eluting compounds being present in subsequent analyses.

Table 3.6.2. Gradient profile for HILIC-LC-MS metabonomic studies. Solvent A = MeCN with 0.1 % v/v formic acid; B = H_2O with 0.1 % v/v formic acid.

Time min^{-1}	% A	% B
0	95	5
9	80	20
21	5	95
24	5	95
27	95	5
30	95	5

Only a handful of papers have been published regarding the use of HILIC (Idborg *et al.*, 2005; Kind *et al.*, 2007; Mawhinney *et al.*, 2007), and most of these utilise a gradient which was similar to that chosen for RP separation; this is because HILIC is in effect reversed-reversed phase, as the solvents used are the same, but the other way around. The first HILIC gradient used H_2O and MeCN both modified by the addition of 0.1 % v/v formic acid (table 3.6.2). Whilst the separation achieved appeared to be similar to that obtained by RP-LC (figure 3.6.1a), repeat injections showed that retention times, peak shapes and intensities were not reproducible (not shown). As HILIC columns require longer equilibration times than RP-LC (Hemström and Irgum, 2006), a longer equilibration time was afforded at the end of the run (19.5 to 30 min) at the expense of a shorter gradient elution profile (table 3.6.3).

Table 3.6.3. Gradient profile for HILIC-LC-MS metabonomic studies utilising a longer equilibration period (19.5 to 30 min). Solvent A = MeCN with 0.1 % v/v formic acid; B = H_2O with 0.1 % v/v formic acid.

Time min^{-1}	% A	% B
0	95	5
15	5	95
19	5	95
19.5	95	5
30	95	5

The resulting UV chromatogram (figure 3.6.2a) and the repeat analysis of the same urine sample (figure 3.6.2b) show that comparable peaks were obtained. However, as for the first HILIC gradient (table 3.6.2), retention times, peak shapes and intensity were not repeatable. The peak labelled '1' exhibited different intensities and retention times ($t_R \cong 3.0$ and 2.4 min for the first and second analyses respectively). The poorly resolved peaks highlighted by arrow '2' show no comparable peak shapes or retention times, and the most intense peak in the UV chromatogram highlighted by arrow '3', despite being of comparable intensities in the two chromatograms, show a shift in retention time of ca. 1 min.

These deviations in retention time, peak shapes and intensity were clearly unacceptable. As HILIC columns work using a zwitterionic stationary phase, and urine is a salty matrix, the reproducibility problems could have been caused by a build up of salt on the column, which was effecting how compounds were being retained. It was decided to add a buffer to the aqueous solvent; the addition of this buffer aids in stabilising the pH of the mobile phase and maintaining a constant salt content, buffering the zwitterionic stationary phase and avoiding any unwanted ionic interactions.

The addition of 5 mM ammonium acetate to the aqueous phase of the gradient shown in table 3.6.3 yielded very reproducible results, as the retention times, peak shapes and intensities of all the peaks within the UV chromatograms were identical (figure 3.6.2c and d, cf. panels a and b).

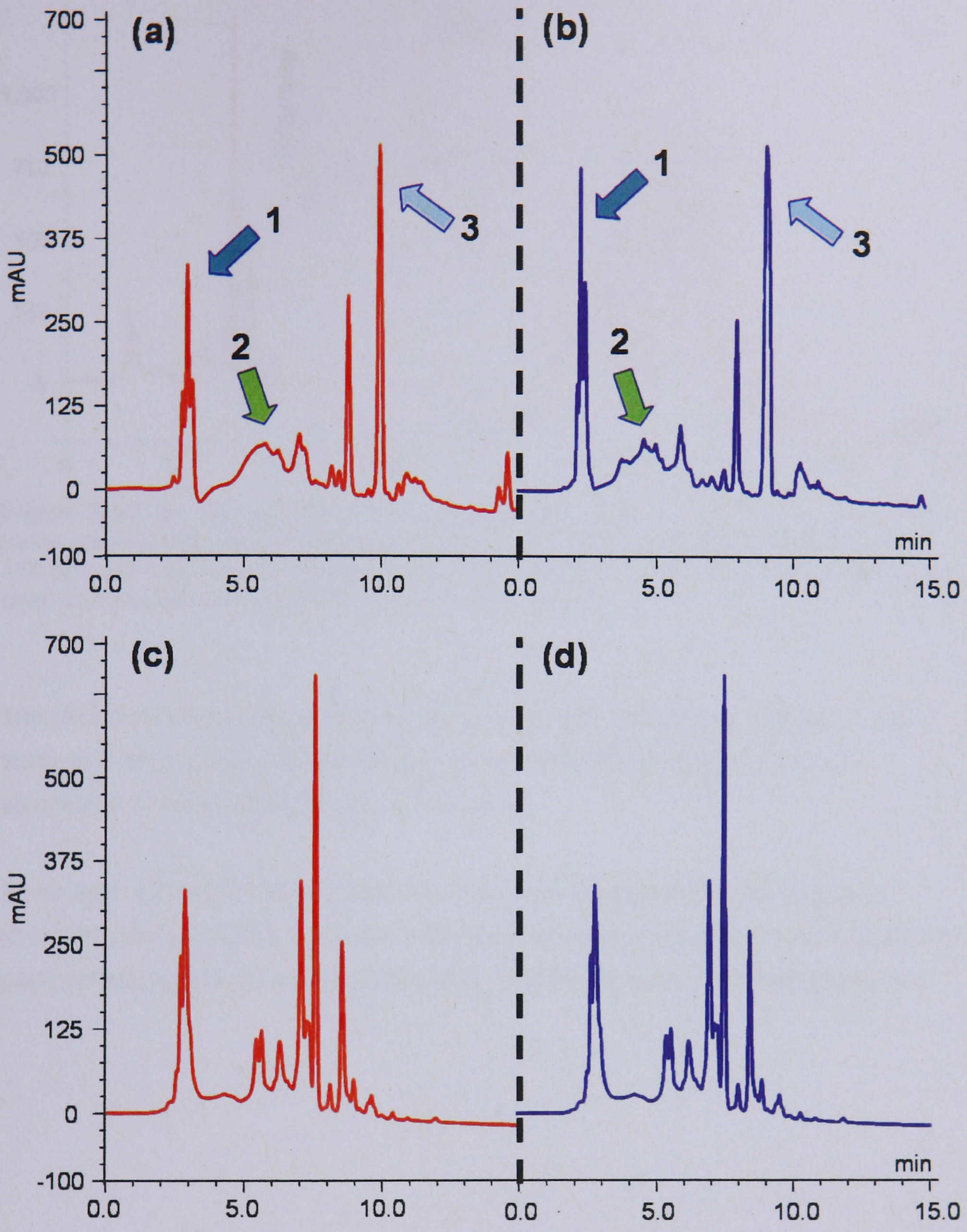


Figure 3.6.2. (a - b) first and second UV_{254} chromatograms of the same urine sample using the gradient described in table 3.6.3, where the arrows refer to specific peaks (see comments within main text). (c – d) two replicate injections of the same urine sample using same gradient, but with the addition of 5 mM ammonium acetate to the aqueous phase.

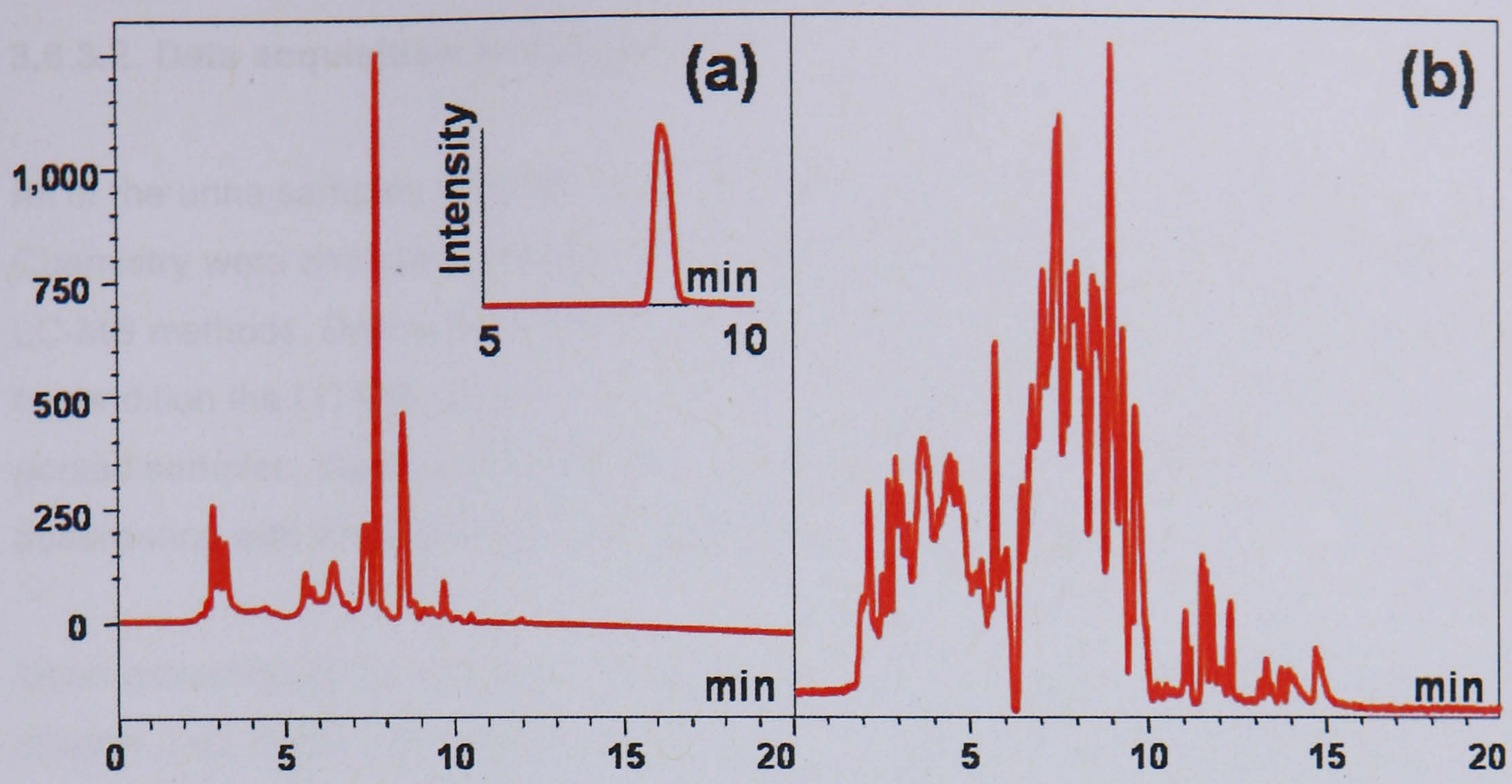


Figure 3.6.3. (a) Typical UV_{254} chromatogram obtained using HILIC separation of the same urine sample as used for figure 3.6.1. Inset shows an XIC for creatinine (*m/z* 114.07 = $[M + H]^+$ with a retention time of 8.3 min. (b) Positive mode TIC of the same urine sample, normalised to the most intense peak.

The UV chromatograms and TICs shown in figure 3.6.3 possess fewer peaks than those of a RP analysis (cf. figure 3.6.1), however, there still appears to be an abundance of information present within the TIC.

As suitable RP and HILIC gradients of comparable length were developed and shown to produce robust, reproducible results, the next steps allowing comparison of each column's ability to produce meaningful statistical results could be carried out.

3.6.3.2. Data acquisition and extraction

All of the urine samples obtained from the volunteers from the Department of Chemistry were analysed positive and negative mode ESI and both RP and HILIC LC-MS methods. Before any samples were analysed, pooled urine samples were run to condition the LC-MS system. Throughout each of the four data acquisitions, pooled samples, blank runs and replicate sample analyses were carried out in accordance with the methodologies laid out earlier in this chapter.

Upon extraction of the resulting data using the metabolomics export script (see chapter 3.4), it was noticed that for both separation methods, fewer peaks were extracted from negative ionisation mode compared to positive. This was not a surprise, as for positive ionisation mode, more compounds are generally ionised (general observation); however, negative ionisation mode tends to generate much clearer spectra than positive with fewer background peaks being observed. It was a surprise to find that more variables were generated from RP-LC-MS than HILIC-LC-MS data. This could be due to the fact that HILIC only retains mildly polar and polar compounds (of which it was hypothesised that there should be a substantial in urine). RP columns can retain mildly polar and hydrophobic compounds, but typically generate multiple peaks from the same compounds (due to salts); this fact may account for the increase in extracted data when using RP separation techniques.

For both RP and HILIC separation, the most intense peak observed within TICs generally corresponded to creatinine. The intensity of the creatinine peak for each sample was plotted from the extracted data; this is presented in figure 3.6.4.

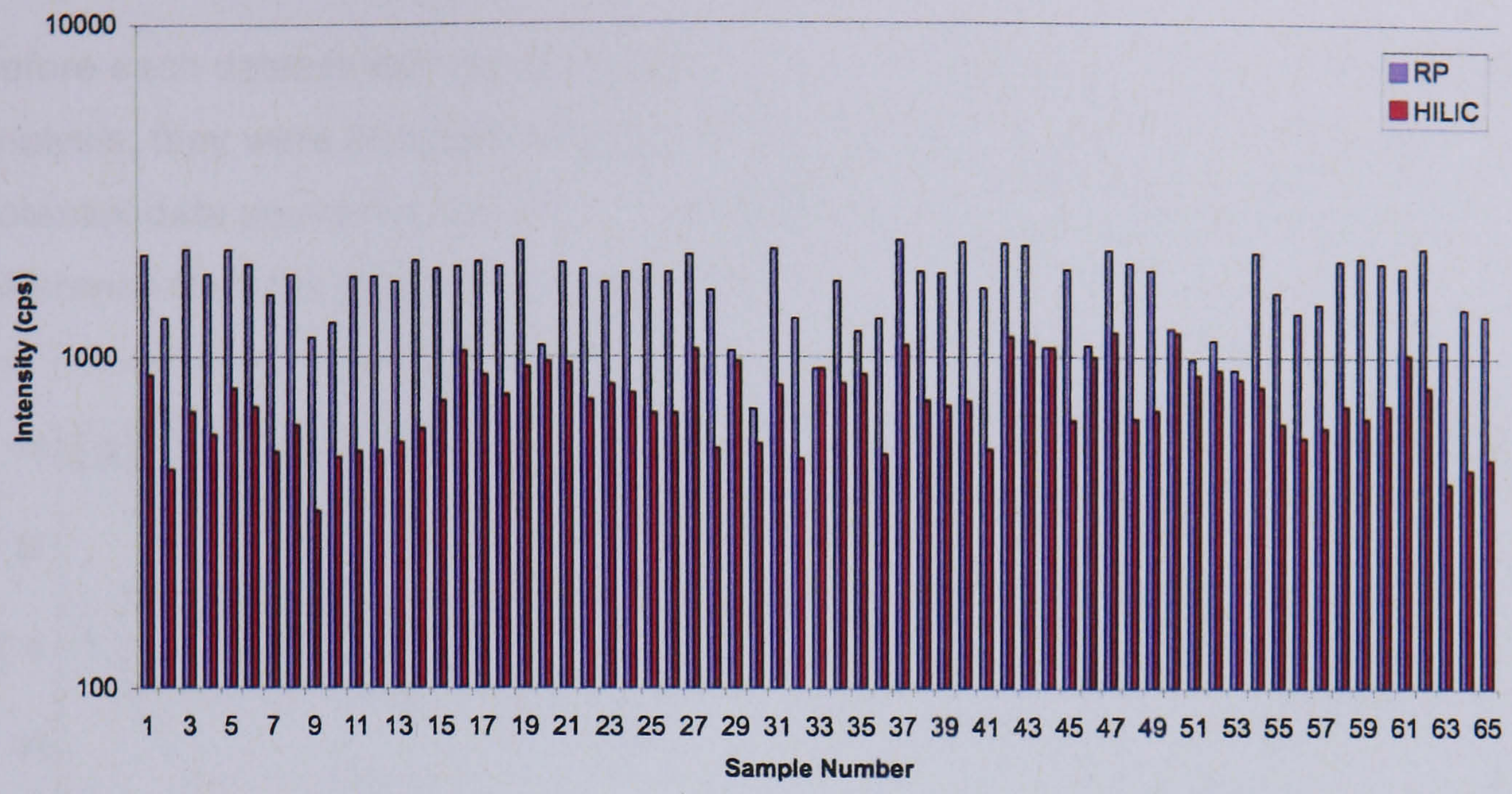


Figure 3.6.4. Graphical representation of the extracted creatinine peak intensity from positive mode data for both RP and HILIC separation.

From the data presented in figure 3.6.4, it is clear that the intensity of the creatinine peaks are generally higher using RP separation than HILIC; this was expected as all samples analysed by HILIC-LC-MS were diluted 50:50 with MeCN. It is evident that, on the whole, the extracted intensity of creatinine using both separation techniques generally follow the same trend. There are 10 samples where the difference in intensity does not follow the same trend as the majority of the data, with two samples (33 and 44) having the same intensity. The creatinine levels from HILIC-LC-MS are consistent across the majority of samples; it is the RP-LC-MS creatinine levels that appear to be lower than expected for the samples that do not follow the trend. As creatinine elutes from the RP column very close to the void, this apparent decrease in expected creatinine intensity could be due to ion suppression caused by particularly salty samples or co-eluting compounds which utilise the majority of the ion stream; this is consistent with the fact that creatinine is well retained using HILIC and does not exhibit any noticeable deviation from the expected intensity.

Despite there being a few minor discrepancies between the extracted intensities for creatinine, it is evident that both RP and HILIC-LC-MS can produce comparable results in terms of detection and ionisation, based upon creatinine.

3.6.3.3. PCA

Before each dataset was randomly split into training and test sets for discriminative analysis, they were analysed using PCA in an unbiased manner to identify any potential data points outside the 95 % confidence limits, which also had a large difference from the model (DModX in SIMCA).

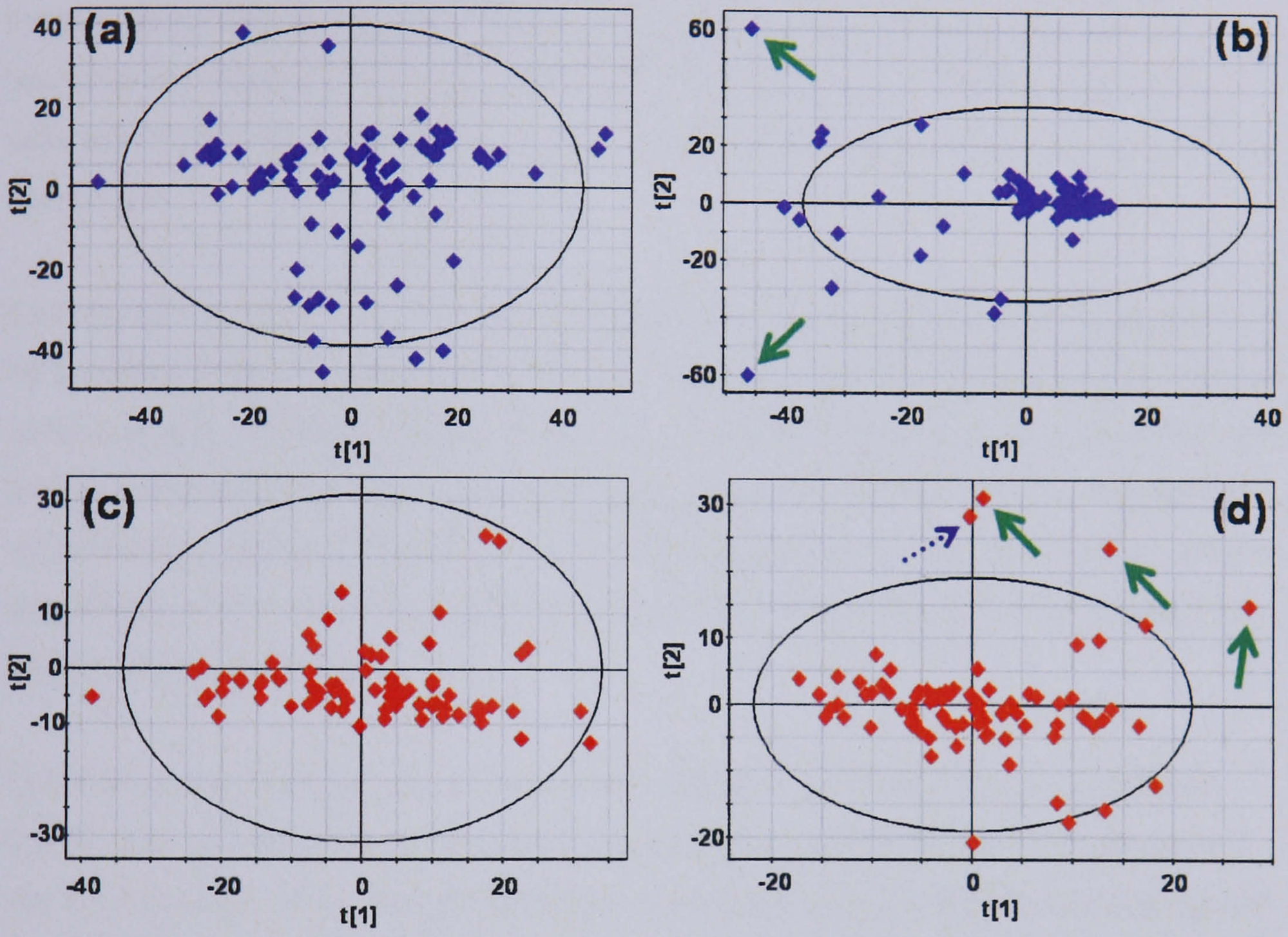


Figure 3.6.5. PCA scores plots of the first two principal components for (a) positive mode RP-LC-MS analysis, (b) negative mode RP-LC-MS analysis, (c) positive mode HILIC-LC-MS analysis and (d) negative mode HILIC-LC-MS analysis.

The four PCA scores plots presented in figure 3.6.5 show the first two principal components for each separation and ionisation method. Figure 3.6.5a shows positive mode RP-LC-MS data, with only a few data points outside of the 95 % confidence limit (shown by the ellipse), whereas for negative ionisation mode RP-LC-MS analysis (figure 3.6.5b) there are two data points (indicated by arrows) that are more substantial outliers than the other points just outside the 95 % confidence limit. For positive mode analysis by HILIC-LC-MS (figure 3.6.5c), the vast majority of the data points are within the confidence limit, with only two points outside this limit; conversely, negative mode HILIC-LC-MS analysis (figure 3.6.5d) shows many more

points outside the confidence limit, with the most extreme of these indicated by arrows.

Comparing positive ionisation mode to negative for both separation methods shows that negative mode data appear to have more 'outliers' than positive. The fact that there are more moderate outliers for negative mode than positive, may be accounted for by the fact that positive mode generates many more variables than negative. As models using negative mode data have fewer variables to form a model, any spurious variables will have a greater effect upon the model, therefore generating data points that appear to be outliers in the first two principal components of a PCA scores plot.

Examining the distance from model (DModX) plots for each PCA model showed that for positive ionisation mode for both separation methods, the few data points that are outside of the confidence limits did not have a DModX value that was larger than the critical tolerance value of 0.05 (corresponding to 95 % confidence). For this reason, all of the data points from positive mode data (figure 3.6.5 a and c) were not classed as outliers, and were subsequently retained to form datasets for further analysis by discriminative statistics.

The data points indicated by arrows in the negative ionisation scores plots (figure 3.6.5 b and d) are reasonable outliers according to the confidence limits shown on the scores plots. When the DModX plots were examined, the data points highlighted by solid green arrows had DModX values which were higher than the critical value, meaning that these data points were not only outside the 95 % confidence limits, but were also not on the same plane as the bulk of the remaining data. The data point indicated by a hashed blue arrow (figure 3.6.5d) did not have a DModX value which was above the critical value meaning that this data point was only an outlier based upon its value for the second principal component. Studying the loadings plots for both negative ionisation mode scores plots failed to highlight any particular variables responsible for the outliers indicated, thus the outlying points must be caused by a combination of many variables, and are therefore not specific outliers.

As the first few principal components failed to show any clustering based upon gender, time of collection, age or smoker status, and the outlying data points were not caused by any particular variables, all of the data points were retained and not discarded. However, the data points were recorded as being outliers and were to be

checked in any subsequent discriminative analysis to see if they once more were classed as outliers[1].

As none of the marked outliers were removed from either the RP-LC-MS or HILIC-LC-MS negative ionisation mode datasets, each of the four datasets were subsequently randomly split into training and test sets. For the subsequent PLS analyses the data were assigned the discriminatory variables gender, time of collection and age, and roughly one-third of the data were held back to form a test set. The test sets contained equal numbers of observations for gender and time of collection, with eight observations for each discriminatory class. For discrimination by age, there were fewer observations for the age groups 31-40 and 41-61 as the bulk of the samples donated were from the 21-30 age group (not surprising within a University environment). The test set therefore consisted of nine, four and three randomly chosen observations for the age groups 21-30, 31-40 and 41-61 respectively. The training and test sets were then normalised according to the methods laid out in section 3.5 before PLS analysis.

[1] Once discriminatory models were developed, the outliers indicated within figure 3.6.5 were not classed as outliers within any developed model. Further to this, the most important variables were checked to see if there was any correlation with the loadings plots from the PCA analysis; there were no correlations evident, suggesting that retaining these 'outliers' was not detrimental to subsequent model development.

3.6.3.4. PLS analysis

Each of the datasets were analysed by PCA, based on consideration of the results of which it was decided that all the data should remain without any data points being discarded. The next step to compare how HILIC separation compared to a traditional RP approach was to analyse each dataset using discriminative statistics. All of the data presented within this section were normalised to TIC and scaled using unit variance. However, the most important variables forming the models developed using no normalisation, normalisation to creatinine, as well as normalisation to TIC and all three scaling methods (mean centring, pareto and unit variance) as presented in section 3.5 are reported too for comparison.

Development of all PLS models was undertaken using the scheme laid out in section 3.5. The explained variation (in terms of R^2 and Q^2 within SIMCA P+) gave an indication of the fit of the model and its predictive ability (using internal venetian blind CV). The internal CV was used to determine the number of components in the developed models. VIP scores were used to ascertain the discriminating power for each variable, and also to identify and remove any unimportant variables that did not add any predictive power to a model. After the removal of unimportant variables, the models were re-developed and the process repeated until a satisfactory model was developed (typically less than 10 variables were used to form a PLS model).

3.6.3.4.1. Positive ionisation mode data analysis

For both separation methods, PLS plots for positive ionisation mode datasets afforded a clear separation in terms of discrimination by gender using LC-MS (figure 3.6.6). It is clear that the discriminative power of the model using data from HILIC separation (figure 3.6.6b) is comparable to, if not better than, that produced from a more traditional RP separation approach (figure 3.6.6a).

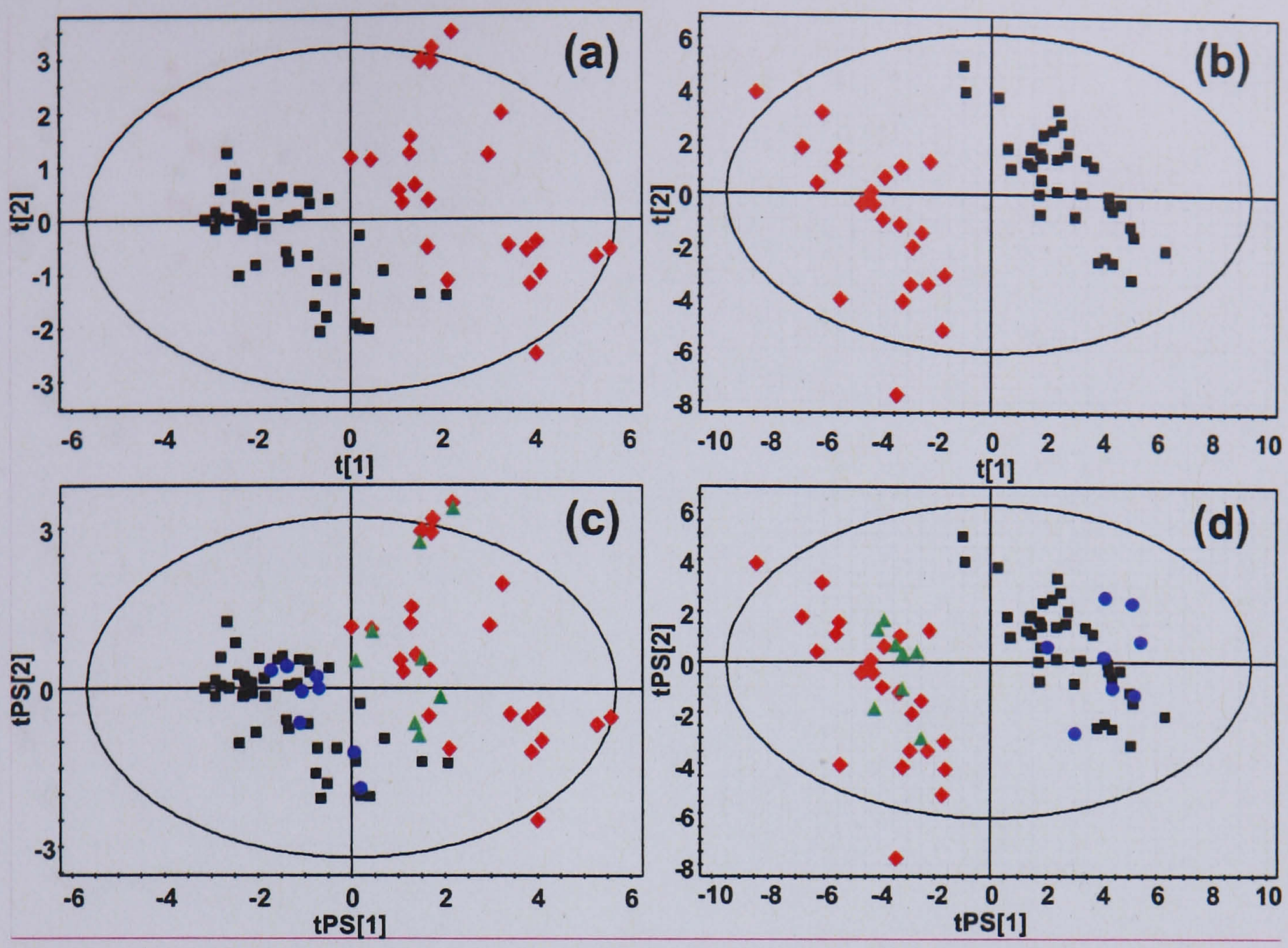


Figure 3.6.6. PLS scores plots for a gender response variable (■ = male training set, ♦ = female training set, ● = male external test set, ▲ = female external test set) analysed in positive mode ESI-MS: (a) reversed phase data training set data. (b) HILIC data training set data. (c) reversed phase data with external test set overlaid. (d) HILIC data with external test set overlaid.

To assess the predictive power of each developed PLS model, the external test sets were imported. The classification rates from the independent test sets were 94 and 100 % for RP and HILIC datasets respectively (the lower classification rate for RP-LC-MS corresponds to one fewer external test set samples being correctly classified). Figures 3.6.6c and d show the PLS scores plots with the external test set data overlaid to highlight the predictive power of the developed models. Despite the scores plots being presented using two latent variables, both developed models for discrimination by gender only utilised one latent variable for prediction, with no gain in classification rates being observed by using two latent variables.

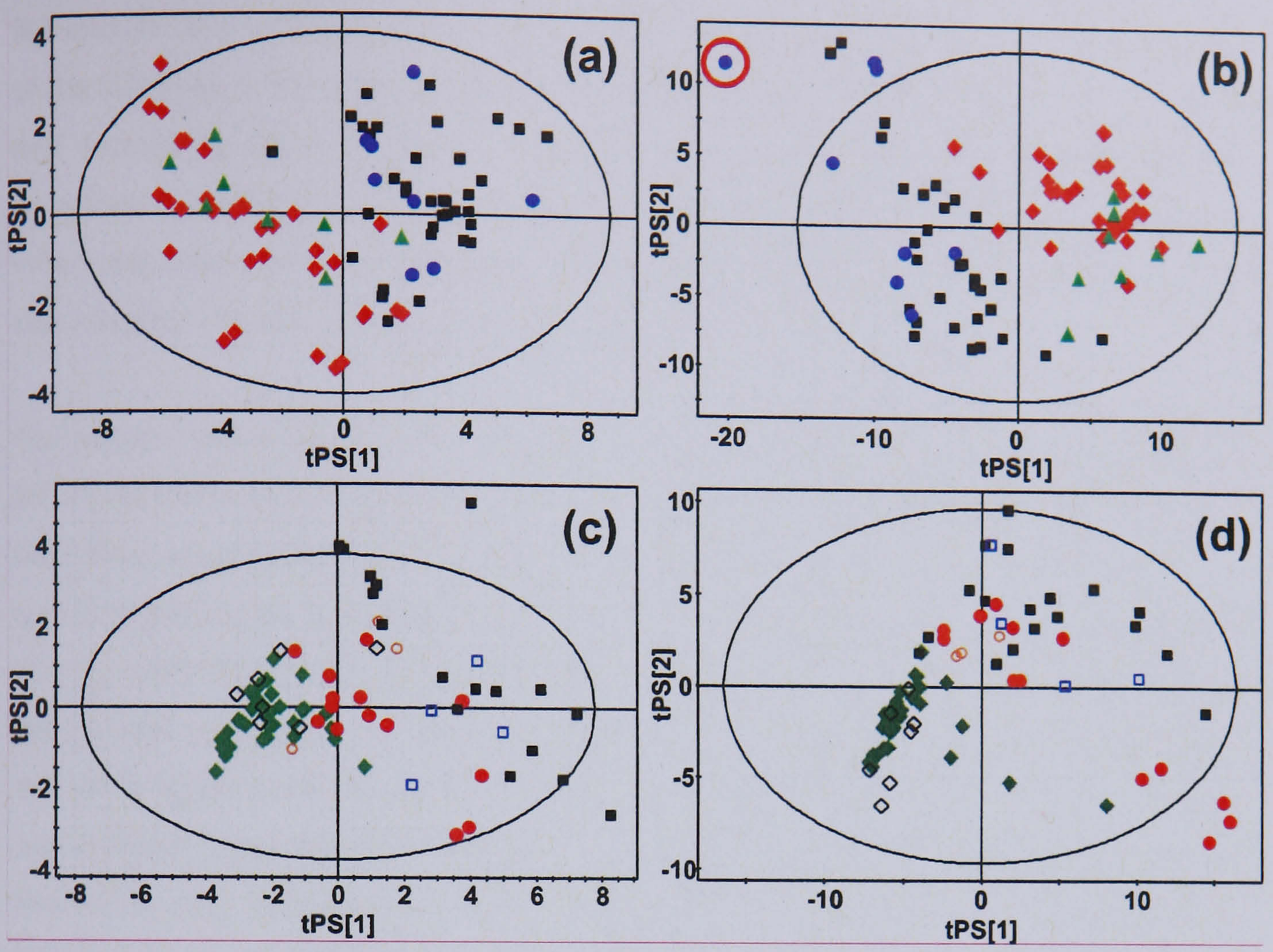


Figure 3.6.7. PLS scores plots for the response variables time of collection (a,b) where (♦ = AM training set, ■ = PM training set, ▲ = AM external test set, ● = PM external test set), and age (c,d) where (♦ = 21-30 age group training set, ● = 31-40 age group training set, ■ = 41-61 age group training set, ◊ = 21-30 age group external test set, ○ = 31-40 external age group test set, □ = 41-61 external age group test set) analysed in positive mode ESI-MS with both the training and test set data shown. (a) PLS model for discrimination by time of collection using RP-LC-MS data, (b) PLS model for discrimination by time of collection using HILIC-LC-MS data, (c) PLS model for discrimination by age using RP-LC-MS data, (d) PLS model for discrimination by age using HILIC-LC-MS data.

For further comparison of the two separation techniques using positive ionisation mode LC-MS, PLS was used to analyse data based on the time of sample collection (AM vs. PM) and age. Figure 3.6.7 shows scores plots with time of collection and age as discriminatory factors. Both RP-MS (figure 3.6.7a) and HILIC-MS (figure 3.6.7b) data exhibited good clustering for discrimination by time of collection (AM = first void of the day, PM = any subsequent void after 15:00 h), although, as might be expected, the scores plots do show some overlap. This overlap is probably caused by the fact that some volunteers may not have donated the first void of the day, or may have donated earlier than the 15:00 h collection time for the PM sample. Also, discrimination between AM/PM samples was not expected to give clear clustering, as some urinary profiles may not substantially differ over the course of the collection period. Despite the overlap, classification rates of the external test sets were 94 % for

RP and 87 % for HILIC. HILIC's lower classification rate was due to the fact that one of the samples in the external test set (a PM donation, highlighted by a red circle) lies well outside the 95 % confidence limit (figure3.6.7b) and was subsequently classed as a false positive. The PLS model for RP separation gave the highest classification when one latent variable was used, whereas the HILIC separation model gave higher classification results with two latent variables.

For discrimination by age, the donors were split into three arbitrary classes: ages 21-30, 31-40 and 41-61. A general trend with increasing age could be seen along the first latent variable for both RP and HILIC datasets, although the groups merge at age boundaries, as expected. This trend along the first principal component explains why for both RP and HILIC data models, only one latent variable was required to give the highest classification rates. The samples corresponding to the 21-30 age group are more tightly clustered those of the 31-40 and 41-61 age groups, which are less well defined. Classification results for the independent test sets were the poorest in this study, with classification rates of 71 and 86 % for RP (figure 3.6.7c) and HILIC (figure 3.6.7d) respectively.

PLS models for both RP and HILIC data were able to predict samples from the younger age groups (21-30) reasonably well, but the models were unable to accurately predict samples from the older groups (31-40 and 41-61). Samples from the external test sets are overlaid onto the RP (Fig. 3.6.7c) and HILIC (Fig. 3.6.7d) scores plots and illustrate the poor prediction of class for the older age groups.

Table 3.6.4. Comparison of the top five variables for each developed model using gender as a discriminatory factor for positive ionisation mode LC-ECI-MS data, where GD = gender for which that variable was more discriminatory for.

Ionisation Method	Normalisation Method	Scaling Method	Rank (VIP)	Separation Method					
				RP			HILIC		
				GD	*m/z*	t_R	*GD*	*m/z*	t_R
Positive	None	ctr	1	⇑M	**310.22**	15.83	⇑M	**126.06**	3.67
			2	⇑F	**229.16**	3.88	⇑M	**327.11**	3.97
			3	⇑M	**286.19**	15.18	⇑F	**432.14**	4.80
			4	⇑F	**290.16**	8.32	⇑M	**371.07**	4.15
			5	⇑M	**265.13**	11.22	⇑M	**126.07**	5.28
		par	1	⇑M	**310.22**	15.83	⇑F	**126.06**	3.67
			2	⇑F	**114.06**	2.63	⇑M	**126.07**	5.28
			3	⇑M	**100.09**	7.57	⇑M	**162.03**	9.23
			4	⇑F	**290.16**	8.32	⇑M	**327.11**	3.97
			5	⇑M	**286.19**	15.18	⇑F	**432.14**	4.80
		UV	1	⇑F	**497.19**	15.47	⇑M	**185.02**	7.12
			2	⇑F	**815.25**	11.37	⇑F	**192.11**	7.70
			3	⇑F	**83.05**	9.23	⇑M	**428.01**	2.63
			4	⇑M	**182.07**	5.73	⇑F	**325.99**	7.65
			5	⇑M	**263.20**	6.67	⇑F	**206.04**	2.55
	Creatinine	ctr	1	⇑M	**310.22**	15.83	⇑M	**327.11**	3.97
			2	⇑M	**114.07**	3.52	⇑F	**432.14**	4.80
			3	⇑M	**286.19**	15.18	⇑M	**476.20**	5.22
			4	⇑M	**100.09**	7.57	⇑M	**371.07**	4.15
			5	⇑M	**302.24**	16.12	⇑M	**415.12**	4.83
		par	1	⇑M	**310.22**	15.83	⇑F	**166.99**	2.98
			2	⇑M	**286.19**	15.18	⇑F	**177.04**	3.50
			3	⇑F	**290.16**	8.32	⇑M	**529.14**	4.48
			4	⇑M	**100.09**	7.57	⇑M	**371.07**	4.15
			5	⇑M	**202.11**	5.75	⇑F	**401.95**	7.03
		UV	1	⇑F	**497.19**	15.47	⇑F	**182.05**	2.48
			2	⇑F	**815.25**	11.37	⇑F	**231.10**	5.33
			3	⇑F	**90.05**	2.70	⇑F	**206.04**	2.55
			4	⇑F	**492.22**	15.50	⇑F	**158.02**	7.93
			5	⇑F	**290.16**	8.32	⇑F	**105.02**	2.73
	TIC	ctr	1	⇑M	**114.07**	3.52	⇑F	**126.06**	3.67
			2	⇑M	**310.22**	15.83	⇑M	**529.14**	4.48
			3	⇑M	**100.09**	7.57	⇑F	**166.99**	2.98
			4	⇑F	**229.16**	3.88	⇑M	**126.07**	5.28
			5	⇑M	**286.19**	15.18	⇑F	**265.08**	2.92
		par	1	⇑M	**310.22**	15.83	⇑F	**166.99**	2.98
			2	⇑M	**100.09**	7.57	⇑F	**177.04**	3.50
			3	⇑M	**286.19**	15.18	⇑F	**401.95**	7.03
			4	⇑F	**290.16**	8.32	⇑F	**265.08**	2.92
			5	⇑F	**229.16**	3.88	⇑M	**126.06**	3.67
		UV	1	⇑F	**171.10**	6.92	⇑F	**206.04**	2.55
			2	⇑F	**497.19**	15.47	⇑F	**192.11**	7.70
			3	⇑F	**217.13**	6.92	⇑M	**87.02**	3.42
			4	⇑M	**161.09**	10.23	⇑F	**325.99**	7.65
			5	⇑F	**815.25**	11.37	⇑F	**182.05**	2.48

It is evident that for positive ionisation mode, both separation methods allow the development of models that have comparable external classification results. Table 3.6.4 shows the five most important variables (as determined by their VIP values) for the developed PLS models using gender as the discriminatory factor, for all normalisation and scaling methods. The shaded cells within table 3.6.4 highlight variables that are duplicated in other developed models for that particular separation method. Whilst there are some unshaded cells, meaning unique variables were used in that particular model, there is a large number of shaded cells for both RP and HILIC separation method models, suggesting that these variables are important for discrimination by gender. As there are only two groups to discriminate against, the 'GD' column shows which gender each variable had an increased response for.

The CID tandem MS analysis of the most important variables determined by PLS models for discrimination by gender, time of collection and age for positive (and negative) ionisation mode are presented in section 3.6.4.

3.6.3.4.2. Negative ionisation mode data analysis

Negative mode ESI-MS was also investigated for each separation method to determine whether further information could be obtained over that produced using positive mode, in order to provide as comprehensive an MS fingerprint as possible. Again PLS was carried out on the data sets with response variables assigned to gender, time of collection and age.

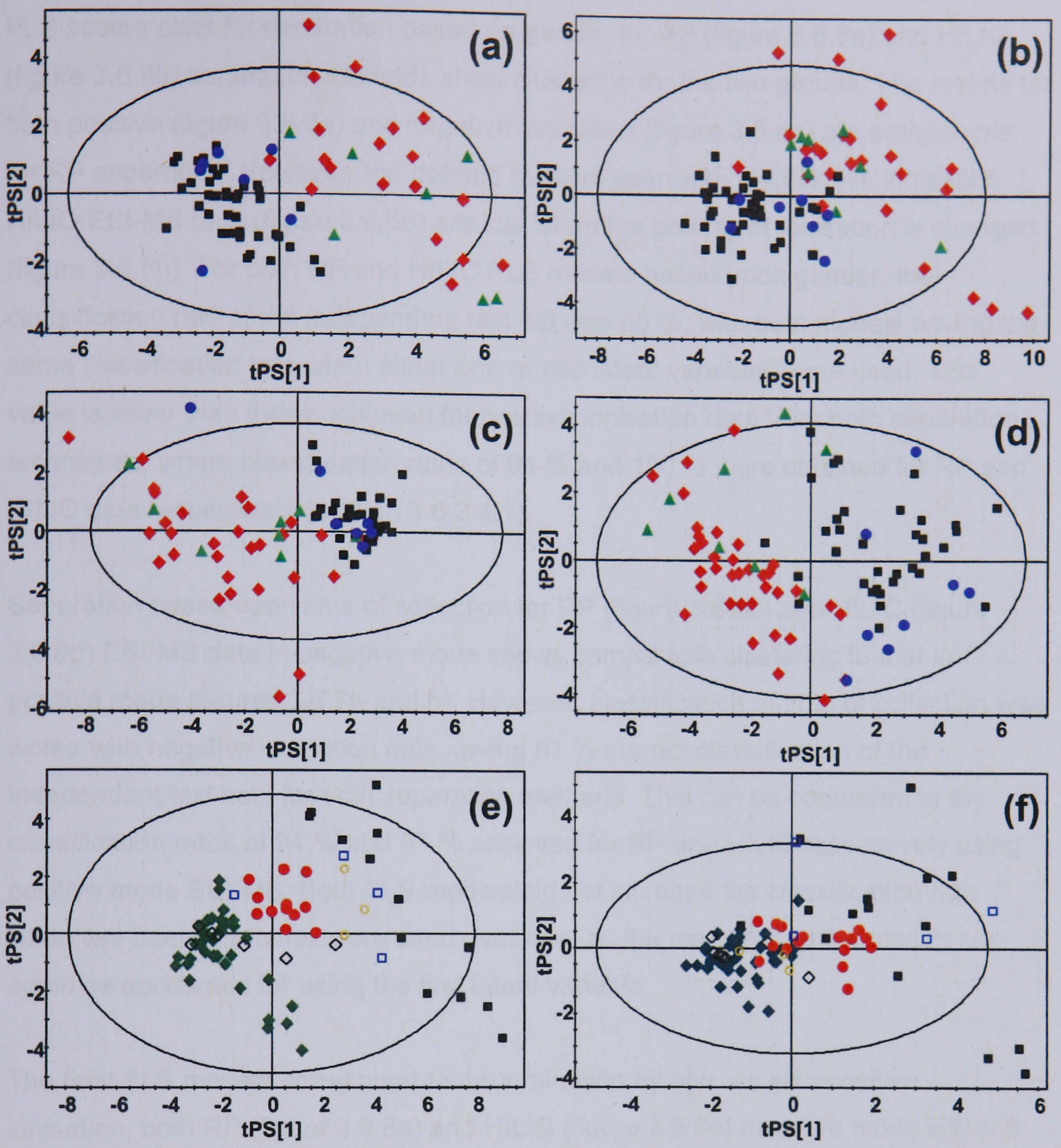


Figure 3.6.8. PLS scores plots for negative mode ESI-MS with the response variables: gender (a,b) where (■ = male training set, ♦ = female training set, ● = male external test set, ▲ = female external test set), time of collection (c,d) where (♦ = AM training set, ■ = PM training set, ▲ = AM external test set, ● = PM external test set), and age (e,f) where (♦ = 21-30 age group training set, ● = 31-40 age group training set, ■ = 41-61 age group training set, ◊ = 21-30 age group external test set, ○ = 31-40 external age group test set, □ = 41-61 external age group test set). (a) PLS model for discrimination by gender using RP-LC-MS data, (b) PLS model for discrimination by gender using HILIC-LC-MS data, (c) PLS model for discrimination by time of collection using RP-LC-MS data, (d) PLS model for discrimination by time of collection using HILIC-LC-MS data, (e) PLS model for discrimination by age using RP-LC-MS data, (f) PLS model for discrimination by age using HILIC-LC-MS data.

PLS scores plots for separation based on gender for RP (figure 3.6.8a) and HILIC (figure 3.6.8b) separation methods show clustering for the two groups. The results for both positive (figure 3.6.6a) and negative ionisation (figure 3.6.8a) are comparable for RP separation. However, the defined clusters seen with the positive ionisation HILIC-ESI-MS data (figure 3.6.6b) are lost when the polarity of ionisation is changed (figure 3.6.8b). For both RP and HILIC PLS models based upon gender, the classification rate of the independent test set was 88 %, with both models having the same classification rate when either one or two latent variables were used. This value is lower than those achieved for positive ionisation data from both separation techniques, where classification rates of 94 % and 100 % were obtained for RP and HILIC data respectively (section 3.6.3.4.1).

Separation based upon time of collection for RP (figure 3.6.8c) and HILIC (figure 3.6.8d) ESI-MS data in negative mode shows comparable clustering to that in positive mode (figures 3.6.7a and b). However, classification of time of collection was worse with negative ionisation data, giving 81 % correct classification of the independent test sets for both separation methods. This can be compared to the classification rates of 94 % and 87 % achieved for RP and HILIC respectively using positive mode ESI-MS. Both PLS models did not increase the classification rate when two latent variables were used over one, as the majority of the variation can again be accounted for using the first latent variable.

The final PLS models correspond to discrimination by age. As with positive ionisation, both RP (figure 3.6.8e) and HILIC (figure 3.6.8e) negative mode ESI-MS data show overlap of the age groups, with the general trend of increasing age along the first latent variable. With negative ionisation data, the age group 31-40 appears to form a tighter cluster than was observed for positive ESI-MS data (figure 3.6.7c and d), although the oldest age group (41-61) again shows poor clustering. Even with the tighter clustering of the second age group (31-40), classification of the external test set for both separation methods using negative ESI-MS data was poor at 64 %. For RP separation, the use of two latent variables increased the classification rates, but for HILIC separation only one latent variable was required to gain the maximum classification rate.

Table 3.6.5. Comparison of the top five variables for each developed model using gender as a discriminatory factor for negative ionisation mode LC-ECI-MS data, where GD = gender for which that variable was more discriminatory for.

Ionisation Method	Normalisation Method	Scaling Method	Rank (VIP)	Separation Method					
				RP			HILIC		
				GD	*m/z*	t_R	*GD*	*m/z*	t_R
Negative	None	ctr	1	⇑F	**186.99**	13.60	⇑F	**186.98**	2.92
			2	⇑M	**211.99**	11.53	⇑M	**211.98**	5.80
			3	⇑F	**191.00**	4.48	⇑M	**367.12**	2.13
			4	⇑F	**263.07**	11.33	⇑M	**263.08**	4.57
			5	⇑F	**178.03**	11.35	⇑F	**96.95**	10.73
		par	1	⇑F	**191.00**	4.48	⇑F	**186.98**	2.92
			2	⇑M	**211.99**	11.53	⇑M	**211.98**	5.80
			3	⇑F	**263.07**	11.33	⇑F	**96.95**	10.73
			4	⇑M	**331.14**	17.02	⇑M	**367.12**	2.13
			5	⇑M	**465.21**	17.57	⇑F	**107.04**	2.97
		UV	1	⇑M	**158.08**	7.82	⇑F	**165.03**	2.93
			2	⇑F	**101.03**	3.48	⇑M	**263.08**	2.87
			3	⇑F	**217.04**	3.40	⇑M	**88.03**	9.35
			4	⇑F	**145.01**	3.47	⇑F	**424.98**	5.80
			5	⇑F	**304.89**	2.28	⇑F	**74.02**	5.40
	Creatinine	ctr	1	⇑F	**178.03**	11.35	⇑F	**186.98**	2.92
			2	⇑F	**263.07**	11.33	⇑M	**211.98**	5.80
			3	⇑M	**211.98**	11.53	⇑F	**161.97**	3.02
			4	⇑M	**191.00**	3.93	⇑M	**367.12**	2.13
			5	⇑F	**191.00**	4.48	⇑M	**263.08**	4.57
		par	1	⇑F	**178.03**	11.35	⇑F	**186.98**	2.92
			2	⇑F	**263.07**	11.33	⇑M	**211.98**	5.80
			3	⇑F	**191.00**	4.48	⇑M	**367.12**	2.13
			4	⇑M	**211.98**	11.53	⇑F	**107.04**	2.97
			5	⇑M	**191.00**	3.93	⇑M	**263.08**	4.57
		UV	1	⇑F	**101.03**	3.48	⇑F	**74.02**	5.40
			2	⇑F	**145.01**	3.47	⇑M	**145.01**	6.75
			3	⇑F	**78.96**	4.33	⇑F	**101.02**	6.75
			4	⇑F	**495.19**	15.45	⇑M	**88.03**	9.35
			5	⇑M	**158.08**	7.82	⇑M	**151.01**	6.08
	TIC	ctr	1	⇑F	**263.07**	11.33	⇑F	**186.98**	2.92
			2	⇑F	**186.99**	13.60	⇑M	**367.12**	2.13
			3	⇑M	**191.00**	4.48	⇑F	**107.04**	2.97
			4	⇑M	**211.98**	11.53	⇑M	**263.08**	4.57
			5	⇑M	**191.00**	3.93	⇑F	**191.00**	7.18
		par	1	⇑F	**263.07**	11.33	⇑F	**186.98**	2.92
			2	⇑F	**186.99**	13.60	⇑F	**107.04**	2.97
			3	⇑M	**191.00**	4.48	⇑M	**367.12**	2.13
			4	⇑M	**465.21**	17.57	⇑M	**172.98**	2.97
			5	⇑F	**283.06**	13.30	⇑M	**167.00**	7.58
		UV	1	⇑F	**101.03**	3.48	⇑F	**184.09**	2.50
			2	⇑F	**145.01**	3.47	⇑F	**186.98**	2.92
			3	⇑M	**541.22**	15.23	⇑F	**74.02**	5.40
			4	⇑F	**495.19**	15.45	⇑F	**101.02**	6.75
			5	⇑F	**260.99**	10.28	⇑F	**107.04**	2.97

As for positive mode of ionisation, the top five most important variables for the developed PLS models using gender as a discriminatory variable, again using all normalisation and scaling methods, shows that there are relatively few unshaded cells (table 3.6.5). The gender for which each variable gave the most discrimination for are shown in the 'GD' column. The CID tandem MS analysis of the variables generated by negative ionisation mode data for discrimination by gender, time of collection and age is presented in section 3.6.4.

3.6.3.4.3. Summary of discriminative analysis of data from positive and negative modes of ionisation

Table 3.6.6. Comparison of external test set classification results for reversed phase and HILIC separation technique data from positive and negative mode ESI-MS studies. The value indicates the percentage of correct classification results.

Ionisation Mode	Y variable	Separation Method	
		RP	HILIC
Positive	Gender	94	100
	Time of collection	94	87
	Age	71	86
Negative	Gender	88	88
	Time of collection	81	81
	Age	64	64

Table 3.6.6 summarises the classification results for the independent test set. The two different chromatographic column chemistries allow comparable classification rates, showing that HILIC is a suitable separation technique to be employed for the analysis of human urine in metabonomic studies. It is clear that, for gender, diurnal variation and age, the classification rates obtained using positive mode ESI-MS are higher than for negative mode, suggesting that positive ionisation data generate more robust models. However, when a metabonomic study is undertaken, a comprehensive picture of the components present in the sample should be sought and both positive and negative ionisation considered.

Comparing tables 3.6.4 and 3.6.5 for the variables generated for discrimination by gender shows that it is clear that the variables used to construct the models from the positive and negative ionisation mode data are all totally different, as both the *m/z* and retention times (t_R) are different. This shows that different compounds are contributing to the developed models, highlighting the fact that both modes of ionisation are important in increasing the urinary metabolome coverage. Further to this, comparing the variables for RP-LC-MS and HILIC-LC-MS also reveals that different compounds are contributing to the developed PLS models. This is entirely consistent with the initial expectation that different compounds would be retained on the different columns, and again reinforcing the need to utilise complementary separation methods when as much information as possible is sought from urinary samples.

3.6.3.5. Data fusion

To evaluate how fusing the four datasets (±RP and ±HILIC) together alters the predictive ability of a PLS model for each of the three Y variables (gender, time of collection and age), the datasets were concatenated together as outlined in section 3.5. As for each individual dataset, the fused dataset was first analysed using PCA to identify any outliers. As for the single datasets, there was no clustering within the first few LVs related to gender, time of collection or age. The majority of the data points were within the 95 % confidence limit; only three data points were outside the confidence limit, but did not have a DModX value above the critical value (95 %) and were therefore retained for subsequent analysis.

The fused dataset was split into a training and test set. As the training set contained over 15,000 variables, optimising the PLS models based upon discrimination by gender, time of collection and age was of paramount importance to avoid any overfitting.

The PLS scores plot based upon separation according to gender shows good clustering (figure 3.6.9a); this is reflected in the external classification rate of 100 %. The external test set data points are plotted onto the scores plot, which shows two latent variables. To obtain the 100 % classification rate, two latent variables were required.

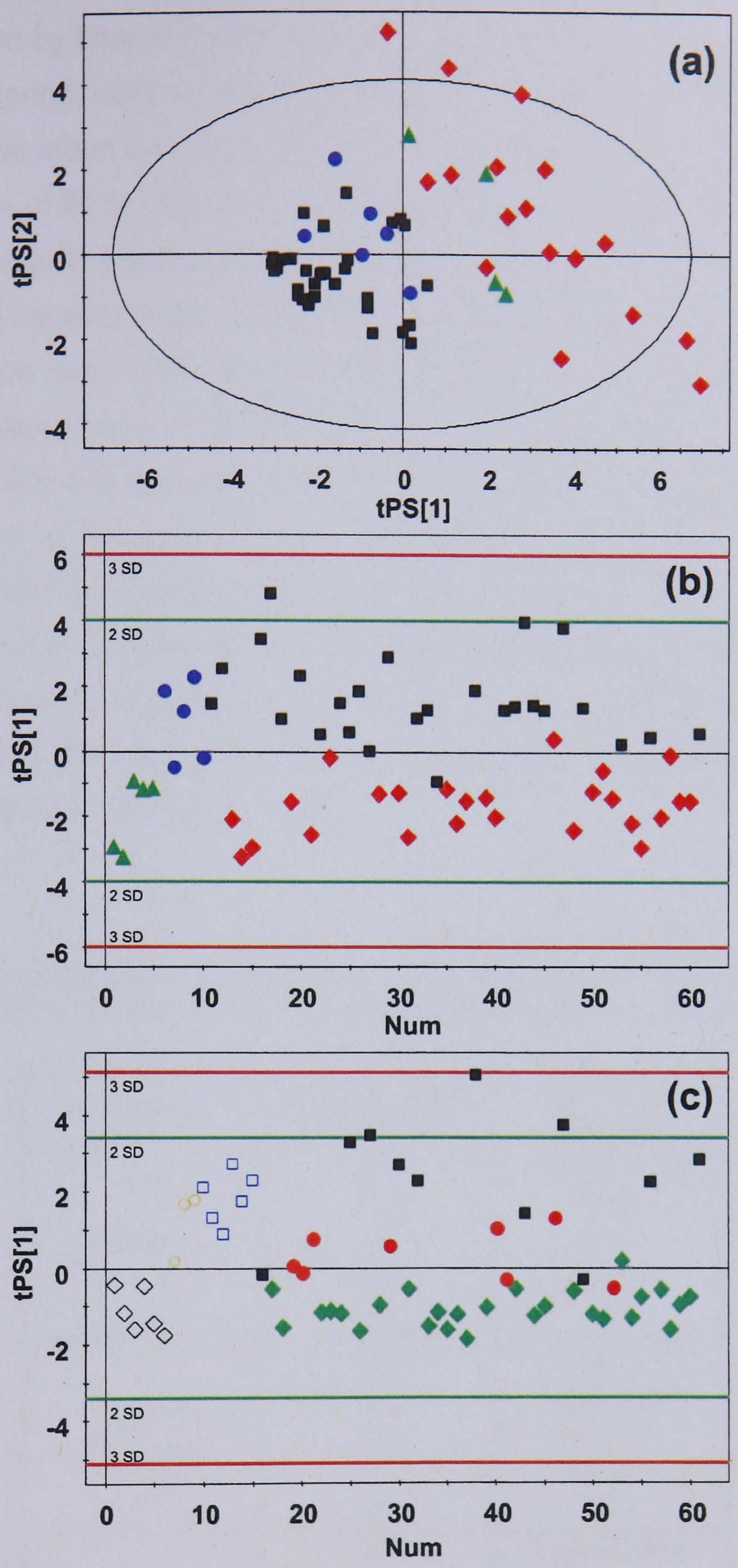


Figure 3.6.9. PLS scores plots for the response variables: gender (a) where (■ = male training set, ♦ = female training set, ● = male external test set, ▲ = female external test set), time of collection (b) where (♦ = AM training set, ■ = PM training set, ▲ = AM external test set, ● = PM external test set) and age (c) where (♦ = 21-30 age group training set, ● = 31-40 age group training set, ■ = 41-61 age group training set, ◊ = 21-30 age group external test set, ○ = 31-40 external age group test set, □ = 41-61 external age group test set).

For discrimination by time of collection, the developed PLS model again showed some good clustering, but with overlap between the two groups as expected (figure 3.6.9b). Using two latent variables did not increase the observed external classification rate of 90 % over using one latent variable. Similarly for discrimination by age, there is significant overlap between the three age groups (figure 3.6.9c). The 21-30 age group exhibits a tight cluster between zero and minus two on the y-axis, with the 31-40 age group also showing little spread across the one latent variable. The 41-61 age group has a large spread from zero to five on the y-axis, showing much more variation between the samples. The external test set data (overlaid) shows that the 21-30 age group can be well predicted, but with increasing age the level of confidence in predictive ability is much lower as there is so much overlap between the 31-40 and 41-61 age groups; this is reflected in the external classification rate of 87 % (100 % for the 21-30 age group). As the 41-61 age group has poor clustering, only one latent variable was required, as using two did not increase the external classification rate.

Table 3.6.7. Comparison of the external test set classification results for concatenated data and each of the four individual data sets (± RP and ± HILIC).

Y Variable	Classification Rate (%)	Separation Method
Gender	100	Data Fusion
	100	HILIC +
	94	RP +
	88	HILIC -
	88	RP -
Time of Collection	94	RP +
	90	Data Fusion
	87	HILIC +
	81	HILIC -
	81	RP -
Age	87	Data Fusion
	86	HILIC +
	71	RP +
	64	HILIC -
	64	RP -

Table 3.6.7 compares the highest external classification results for each of the single datasets with the concatenated dataset. It should be expected that as the data fusion models use all of the available variables from each of the four single datasets, they should perform as well as, or better than, the best performing individual dataset. As

some of the original datasets had some samples missing due to insufficient urine donated, there were fewer samples available to form the aligned concatenated dataset, resulting in slightly smaller test sets, and therefore different external classification results. For both discrimination by gender and age, data fusion generated results that were equal to, or slightly better than, the best performing individual model. For discrimination by time of collection, the data fusion model performed as well as the positive ionisation mode RP dataset with only one sample being incorrectly classified; the decrease in classification rate was due to the differently sized external test set.

Table 3.6.8. Comparison of the top five variables for each developed model highlighting which separation and ionisation mode generated each of the variables.

Y Variable	Rank (VIP)	Separation Method	Polarity	*m/z*	t_R	*Y Variable Discriminative for*
Gender	1	HILIC	+	**428.01**	2.63	⇑F
	2	HILIC	-	**184.09**	2.50	⇑M
	3	HILIC	+	**223.04**	3.00	⇑F
	4	RP	-	**101.03**	3.48	⇑F
	5	RP	-	**145.01**	3.47	⇑F
Time of Collection	1	RP	+	**121.06**	3.95	⇑PM
	2	HILIC	+	**190.11**	12.33	⇑PM
	3	RP	+	**114.90**	2.45	⇑AM (PM = 0)
	4	RP	+	**152.01**	2.80	⇑PM
	5	RP	+	**283.16**	5.57	⇑AM
Age	1	HILIC	+	**84.93**	14.52	(31-40/41-61) > (21-30)
	2	RP	-	**231.01**	4.12	(41-61) > (21-30/31-40)
	3	RP	+	**433.23**	6.42	(41-61) > (21-30/31-40)
	4	RP	+	**176.09**	6.67	⇑ (31-40/41-61)
	5	RP	+	**217.13**	6.92	⇑(41-61)

Each of the three developed PLS models all used a mixture of variables deriving from different separation and ionisation method data (table 3.6.8). The fact that each model utilises variables from each of the four single datasets, highlights the fact that by not attempting to increase the coverage of the components within a biofluid, there is the potential to miss many variables that may be important to a model's development, and therefore the question being asked.

3.6.4. Variable analysis using CID tandem MS

Identifying the ions giving rise to the variables generated by each of the developed models in section 3.6 was carried out using CID tandem MS. A selection of four urine samples was chosen for analysis based upon differences in gender, time of collection and age. Each of the four samples were analysed using the two modes of ionisation and both separation methods, RP and HILIC. Each of the most important variables for each model were added to an 'include list' within Analyst QS's independent data acquisition (IDA) setting, meaning that when an ion with a retention time and mass that corresponds to one in the 'include list' is detected, it is isolated and subjected to CID product ion analysis. A total of the four most intense peaks could be analysed using two different collision energies from any one 'survey scan'.

Upon analysis of the resulting data for each of the samples from each separation and ionisation mode, it became clear that many of the masses added to the 'include list' did not fragment (or were not isolated at intensities sufficient for CID product ion analysis), leaving just the precursor ion present in the spectrum. Unfortunately, this is a problem with the QStar's Analyst QS software, where the most intense peaks are first selected for CID tandem MS analysis at the selected collision energies, with an MS 'survey scan' being run in between each tandem MS step; this can result in some ions not being selected for CID, despite being present on the 'include list'.

The product ions that were generated on CID tandem MS analysis of the precursor ions from the variable lists, and the *m/z* values of ions identified as variables and corresponding to metabolites within the Human Metabolite Data Base (HMDB, (Wishart *et al.*, 2007))[1] or Metlin[2] database, are presented in table 3.6.9:

[1] http://www.hmdb.ca (accessed November 2007).
[2] http://metlin.scripps.edu (accessed November 2007).

Table 3.6.9. A table showing precursor ions that were isolated and subjected to CID tandem MS analysis, and *m/z* values of ions that were identified as variables and correspond to metabolites in either HMDB or Metlin. (*n/d* = not detected in CID analysis. *n/r* = not recorded, meaning no fragment ions were observed).

Separation Method	Statistical Model	Ionisation Polarity	tR	Precursor Ion Mass (*m/z*)	Product Ions (*m/z*)					Product Ion Spectrum in Appendix C
RP	Gender	Positive	15.47	497.2	321.1	303.2	186.1			1.0
		Negative	3.48	101.0	*n/r*					-
			3.47	145.0	*n/r*					-
			15.23	541.2	429.1	145.1				2.0
			15.45	495.2	357.0	267.0	260.1	181.1		3.0
	Time of Collection	Positive	*n/d*	*n/d*	*n/d*					-
		Negative	2.95	173.0	*n/r*					-
			2.95	111.0	*n/r*					-
			3.75	135.0	*n/r*					-
	Age	Positive	8.90	197.1	*n/r*					-
		Negative	8.90	196.1	*n/r*					-
			11.82	180.1	*n/r*					-
			8.92	391.1	291.1	195.1	129.0	97.0	69.0	4.0
HILIC	Gender	Positive	2.55	206.0	*n/r*					-
		Negative	5.40	74.0	*n/r*					-
	Time of Collection	Positive	9.37	290.8	*n/r*					-
			9.37	206.8	*n/r*					-
		Negative	*n/d*	*n/d*	*n/d*					-
	Age	Positive	3.33	197.0	*n/d*					-
			7.50	204.0	*n/r*					-
		Negative	5.38	275.0	*n/r*					-

Appendix B contains the CID tandem MS spectra of any precursor ions which produced fragment ions upon CID.

As the QStar Q-o-ToF MS can only provide accurate mass measurements to 2 d.p. at best for a well calibrated machine[1], it is not possible to confidently assign atomic compositions to precursor (and fragment) ion masses that do not correspond to a metabolite within one of the databases; this is often the case with metabonomic studies where, without the ability to obtain accurate masses (to 4 d.p. with mass accuracy below 1 ppm) or have access to comprehensive databases or search engines (such as those used for proteomic studies), the identification of metabolites remains the hardest goal to achieve.

The unidentified precursor ion at *m/z* 497 (appendix B1) is accompanied by a peak 2 Th higher, suggesting the presence of chlorine in the compound. The loss of 18 Th from one of the fragment ions at *m/z* 321 is consistent with the subsequent loss of water from this fragment ion.

[1] Using internal standards.

The precursor ion at *m/z* 101 did not produce any structurally diagnostic fragment ions at either collision energy used. However, upon searching the databases, a matching mass was found with 2-oxobutyric acid. This metabolite is involved in the metabolism of the amino acids Gly, Ser and Thr, and has previously been detected in human urine as an endogenous metabolite (Liebich *et al.*, 1981). Another metabolite that failed to produce any fragment ions, but has a mass that corresponded to a metabolite found upon a database search was the precursor ion at *m/z* 145. This mass could correspond to the metabolite oxoglutaric acid. Oxoglutaric acid is involved in the Krebs cycle and also in amino acid metabolism and is an endogenous metabolite found in urine (Lee *et al.*, 1998).

The two ions at *m/z* 541 and 495 (appendix B2 and 3) both produced fragment ions. However, their masses did not correspond to any metabolites contained within either of the databases, and the lack of more accurate mass data means that it is not possible to assign atomic compositions to the fragment ions, making it very difficult for a structure to be postulated.

Guneral and Bachmann have previously detected metabolites in human urine which gave deprotonated molecules at *m/z* 173 and 135 (Guneral and Bachmann, 1994). These corresponded to *cis*-aconitic acid, related to the Krebs cycle, and threonic acid, a by-product of the oxidation of ascorbic acid. The ion at *m/z* 111 could correspond to uracil, a compound that has many uses within the body. Uracil is found in RNA, can react to form nucleosides, and has previously been detected in urine as an endogenous metabolite (Jiang *et al.*, 2001; Hofmann *et al.*, 2003).

A metabonomic study by Williams *et al.* of development and ageing of rats proposed two possible atomic compositions of $C_7H_{17}O_4S$ and $C_4H_{13}N_4O_5$ (RMM = 197) for an ion they detected in negative ionisation mode that was discriminatory of age (Williams *et al.*, 2005). The ion also detected in this study at *m/z* 196 (table 3.6.9) in negative ionisation mode (corresponding to a mass of 197 Da) as an important variable for discrimination by age, could correspond to the same component that Williams *et al.* detected.

The remaining precursor ion at *m/z* 391 (appendix B4), which produced fragment ions, did not correspond to any metabolites within any of the databases; without more accurate mass data, no atomic structure can be postulated.

The precursor ion of *m/z* 206 could correspond to an ion of the same *m/z* value detected by Lenz *et al.*, Plumb *et al.*, and Hodson *et al.* (Lenz *et al.*, 2004; Plumb *et al.*, 2005; Hodson *et al.*, 2007). Research by Plumb *et al.* and Hodson *et al.* showed that an ion of *m/z* 206 detected in positive ionisation mode contributed significantly to clustering based upon discrimination according to gender (Plumb *et al.*, 2005; Hodson *et al.*, 2007). The ion observed in this study at *m/z* 206 was detected in the positive ionisation mode, and was an important variable for discrimination according to gender, in agreement with the results of Plumb *et al.* and Hodson *et al.* Lenz *et al.* determined that the ion of *m/z* 206 corresponded to the metabolite 4,8-dihydroxyquinoline-2-carboxylic acid, which is part of the tryptophan catabolism pathway (Lenz *et al.*, 2004).

The only remaining precursor ion that corresponded to a metabolite within either of the databases gave a deprotonated molecule at *m/z* 74, which is postulated to correspond to the amino acid Gly, an endogenous metabolite in human urine (Bales *et al.*, 1984).

3.6.4.1. Discussion

Despite the lack of more accurate mass data which may have led to an increased confidence in the postulated metabolites identified within this study, the postulated metabolites that were identified as discriminatory for gender, time of collection and age have all previously been identified as endogenous metabolites in urine. The ion at *m/z* 206 that was highly discriminatory for gender, which has previously been identified as a metabolite that can be used to discriminate based upon gender in two studies (Plumb *et al.*, 2005; Hodson *et al.*, 2007), suggests that the models developed were robust and consistent with previously published metabonomic research.

3.7. Discussion

The work presented in Chapter Three covers many topics that are involved in an LC-MS metabonomic study. The various platforms available for a metabonomic study were discussed and evaluated for the contributions that each method can provide. It is clear that whilst the majority of studies choose to focus on one particular method, be it LC-MS (Wilson *et al.*, 2005; Sumner, 2006; Chen *et al.*, 2007) or NMR (Lenz *et al.*, 2000; Constantinou *et al.*, 2005; Bertram *et al.*, 2007), the future of metabonomic studies lies with the use of both NMR- and MS-based analyses where all of the data generated are compared together to provide the most comprehensive analysis possible (Forshed *et al.*, 2007a; Forshed *et al.*, 2007b).

As metabonomic studies tend towards a 'comprehensive' fingerprint of the biofluid chosen for analysis (Lenz and Wilson, 2007), the consideration of how samples are collected, stored and manipulated is of paramount importance. However, this is an area that is often overlooked or poorly considered within the literature. The results presented within section 3.3 are in good agreement with similar studies described in the literature (LeBeau *et al.*, 2001; Schneider *et al.*, 2002; Fura *et al.*, 2003; Gika *et al.*, 2007; Saude and Sykes, 2007). Samples should be stored, preferably at -80 °C, as soon as possible after collection and allowed to remain frozen for a period of at least one week to allow the degradation of compounds to be consistent across the sample cohort (Saude and Sykes, 2007), with any sample manipulation kept to an absolute minimum. Samples should just be centrifuged and filtered prior to LC-MS analysis to maximise the metabolite content of urine samples (avoiding the inevitable loss of analytes using extraction methods such as solid phase extraction) (Gika *et al.*, 2007).

As LC-MS systems are renowned for poorer reproducibility than NMR, the system as a whole should be allowed to equilibrate by the analysis of a minimum of three reference samples (e.g. aliquots of a urine pool), and continually monitored by the inclusion of reference samples throughout the run; this was also proposed in very recently published research by Gika *et al.* (Gika *et al.*, 2007).

Despite being a vitally important step, the approaches to extraction of raw data into a usable format for subsequent statistical analysis has received comparatively less attention than the rest of the field of metabonomics. Research by Sangster *et al.* showed that problems with extraction algorithms led to problems with the resulting

extracted data (Sangster *et al.*, 2007); this was also clear in this study in the results presented in section 3.4.

The statistical analysis of metabonomic data (section 3.5) is one of the most important steps in a metabonomic study, after producing 'good' data in the first instance. Statistical tools such as PCA and PLS have made the analysis of large, complex datasets easier, and available to all. Because of the ease of use of statistics, spurious results can be easily generated where naivety leads to the generation of a hypothesis being generated that is proved using poor, undeveloped statistical models. At the very least, discriminative models built using PLS should be developed using only ~$^2/_3$ of the data, and subsequently refined by the removal of unimportant variables. Once a statistical model has been refined and developed, the remaining ~$^1/_3$ of the withheld data should be analysed using the developed model to evaluate the 'true' predictive ability of the developed model. However, there are many other considerations that need to be evaluated when statistics are involved, ranging from scaling methods to normalisation.

Many metabonomic studies have used creatinine to provide some degree of normalisation of differences in the concentration of urine samples prior to statistical analyses (Woitge *et al.*, 1999; Schoenau and Rauch, 2003; Husková *et al.*, 2004; Idborg *et al.*, 2004; Svoboda and Kasai, 2004; Obrant *et al.*, 2005). However, normalisation using total ion count (originating from NMR studies (Kenney and Shockcor, 2003; Antti *et al.*, 2004; Williams *et al.*, 2005)) appears also to be becoming a more accepted method of normalisation in LC-MS studies (Plumb *et al.*, 2003; Williams *et al.*, 2005). Reading the literature, it became very evident that very little consideration has been given to the consequences of using creatinine for normalisation, as studies into the effects of internal and external factors showed that excreted levels of creatinine are easily perturbed (Boeniger *et al.*, 1993; Schoenau and Rauch, 2003; Antti *et al.*, 2004; Heavner *et al.*, 2006). This was evidence that the use of creatinine to provide normalisation for metabonomic studies may be an outdated way of doing so, and is an area that definitely requires further research.

The main aim of Chapter Three was to evaluate a HILIC separation method for the retention of more polar compounds from urine samples. HILIC is an orthogonal separation method to RP, and was shown to provide an increased coverage of different metabolites from urine by the development of robust statistical models that used completely different variables than models developed using RP-LC-MS data.

This highlighted the fact that much information has potentially been missed when only RP as a separation method has been employed for metabonomic studies. The use of HILIC is quite rightly beginning to find a place in metabonomic studies (Idborg *et al.*, 2005; Kind *et al.*, 2007; Mawhinney *et al.*, 2007) as the field demands an increasing amount of information from biofluids.

Despite the lack of more accurate mass measurements, variables being identified and the poor success rate of CID tandem MS analysis, there were a number of masses that were identified from both database searches and previous metabonomic studies into urine samples. Each of the putative assignments correlated well with components previously identified in human urine, with two variables being detected in this metabonomic study as being important for discrimination by age and gender, that have also been identified in other studies.

3.8. Conclusions

The work presented in this chapter has highlighted the fact that metabonomics is still an immature field, with much potential for important scientific discoveries to be made. However, due to the immaturity of the field, there is still much that needs to be researched and standardised if LC-MS-based metabonomics is to stand a chance of becoming a mature, robust and commonly utilised 'omic' approach.

It has been shown that metabonomic experiments are typically poorly reported within the literature, with many vitally important experimental aspects either not reported or not carried out at all. The recommendations from this chapter are that metabonomic studies are first carefully thought through before sample collection begins. When a study has been developed, samples should be carefully collected, stored appropriately for a pre-determined period of time (at least one week to allow the degradation of compounds to be consistent across all samples), and subjected to minimal freeze/thaw cycles and manipulations.

Samples should only be centrifuged and filtered prior to analysis using LC-MS. For LC separation, orthogonal separation methods (RP and HILIC) should be utilised to increase the coverage of the urinary metabolome and thus the amount of information obtained. LC-MS systems should be allowed to equilibrate by the initial analysis of at least three reference samples (e.g. aliquots of a urine pool), and subsequently monitored by the inclusion of reference samples throughout the randomised urine samples.

Data should be not be normalised according to creatinine concentrations, due to the perturbation to concentrations caused by internal and external factors (such as therapeutic interactions, disease and growth), but instead to total ion count. The resulting normalised data should be carefully analysed using both PCA and PLS statistical methods. The data should first be analysed using PCA to view the maximum variation in the dataset, and to explore any outliers that may be present. Subsequent discriminative analysis by PLS should be undertaken on $\sim^2/_3$ of the dataset, with any unimportant variables being removed from the model. The optimised PLS model should then be evaluated using the remaining $\sim^1/_3$ of the dataset that was held back and not used to build the model; this determines the model's 'true' predictive ability.

As the field develops, a more comprehensive 'global' analysis of biofluids should be sought, where all of the data generated are integrated together and analysed as a whole, becoming the norm, rather than the exception.

The conclusions from the work presented in this chapter were used for the subsequent LC-MS metabonomic analysis of clinical urine samples from patients who had suffered a fracture.

3.8.1. Retrospective view

From the work undertaken within this chapter, and with the knowledge gained, it has become clear that obtaining larger sample cohorts would be a benefit allowing a greater number of samples to be held back to form the external test set. Holding back roughly $^1/_3$ of a small data set means that the data available to build a robust model may not be sufficient, let alone large enough to allow the external test set to thoroughly test any developed model.

Obtaining more information from the volunteers would have allowed further investigation into the potential information contained within this biofluid:

- Dietary intake prior to donation
- Height / weight
- Physical activity levels
- Ethnic origin
- Drug intake

The use of NMR for the analysis of the samples would have allowed a further orthogonal detection approach, and would have allowed an interesting comparison of any results obtained to be undertaken.

Collaborating with statisticians to analyse any resulting data more thoroughly would be of great benefit, as a more critical approach to the analysis of metabonomic data could then be sought.

Chapter Four

Clinical urine sample analysis:
Bone fracture profiling

4.1. Introduction

The skeletal system within the human body performs a vital role, as it supports muscle, protects vital organs and stem cells, and is a vast reserve of ions. Despite their strength, given the very nature of human activity, bones do fail and fractures occur. As bones are continuously being renewed, they exhibit an amazing ability to repair themselves and regain their original strength, usually without scarring. In spite of the body's ability to repair itself after a fracture, there are times when a fracture takes much longer to heal than normal (delayed fracture), or may not heal at all (non-union).

Fractures that are delayed or go to non-union (failed fracture healing) require further medical intervention, putting patients through further stress and increasing the time until their fracture has successfully healed. Whilst much research has been undertaken studying biofluids for biomarkers that relate to pathological fractures such as osteoporosis (Calvo *et al.*, 1996; Woitge *et al.*, 1998; Chapurlat *et al.*, 2000; Ebeling and Åkesson, 2001; Srivastava *et al.*, 2002; Garnero and Delmas, 2003), there have been very few studies published regarding non-pathological fractures (Severns *et al.*, 2003; Henle *et al.*, 2005; Zimmermann *et al.*, 2005). Of the research undertaken to attempt to elucidate biomarkers of bone resorption/formation, the overwhelming majority use serum samples and try to identify biomarkers (Chapurlat *et al.*, 2000; Yu-Yahiro *et al.*, 2001; Srivastava *et al.*, 2002, Henle *et al.*, 2005; Li *et al.*, 2005; Asaba *et al.*, 2006); few papers describe studies that have used urine as a biofluid, but those that have, are targeted studies of known breakdown products of large proteins e.g. telopeptides and collagen cross-links (Chapurlat *et al.*, 2000; Yu-Yahiro *et al.*, 2001; Qvist *et al.*, 2002; Srivastava *et al.*, 2002; Garnero and Delmas, 2003; Lamers *et al.*, 2005).

To my knowledge, no research has been undertaken using a metabonomic approach for the analysis of clinical urine samples from patients whom have sustained a non-pathological fracture. In collaboration with Smith & Nephew (Research Centre, York Science Park, York, UK) and York NHS Hospital Trust (York, UK), urine samples were collected from patients who had suffered a fracture, with a view to analysing the samples to hopefully identify biomarkers that relate to failed fracture healing. The ultimate aim was to identify candidate biomarkers that could be tested as early urinary predictors of fracture healing in order to investigate the possibility of early identification of patients whose fracture will result in delayed/non-union.

To avoid any chance of pathological fracture patients being included in the study, or those with incomplete skeletal development, all recruited patients were between the ages of 18-45. Initially, only long bone fractures were considered, but due the slow recruitment of patients suffering from long bone fractures, wrist and ankle fractures were also included in the study as it progressed. Further exclusions were any patients who had suffered multiple injuries, had malignancy, head injuries, spine/foot/hand fractures, pregnant or nursing mothers and any unconscious patients.

During the recruitment period (October 2004 to February 2005) a total of 61 patients were deemed suitable for inclusion into the study (45 males and 16 females); of these, 11 declined consent stating either lack of interest or a needle phobia[1], with a further two patients withdrawing consent at a later date, and one being transferred into the care of another health trust. This left a total of 48 patients (36 males and 12 females, with an age range of 19 to 47 years old; average age = 29.5, standard deviation = 8.2), who each donated between one and four urine samples, ranging from a period of $t = 0$ (time of fracture) to 133 days (19 weeks) after the initial fracture (average = 6 weeks). All of the samples were labelled randomly, shuffled, transported from York Hospital to Smith & Nephew, where they were stored in a semi-organised, partially catalogued state at -80 °C until the recruitment period finished.

[1] Serum samples were also collected for a parallel study by Smith & Nephew into serum markers related to fracture repair; any patients with needle phobia were excluded from the trial.

Table 4.1. Summary of fracture types, breakdown by gender and number of samples obtained.

Fracture Type	No. of Patients	Gender		Total No. of Urine Samples
		Male	Female	
Ankle	21	15	6	51
Wrist	10	6	4	24
Fibula	2	2	0	5
Radius	2	2	0	5
Tibia/Fibula	3	2	1	4
Tibial Plateau	1	1	0	4
Clavicle	3	3	0	3
Radial Head	2	2	0	3
Ulna	1	1	0	3
Radius/Ulna	1	1	0	2
Gr. Tuberosity of Humerus[1]	1	0	1	2
Pilon	1	1	0	1
Total	**48**	**36**	**12**	**107**

The largest sample cohorts obtained were not from long bone fractures, but from ankle and wrist fractures (table 4.1), with a total of 51 and 24 samples obtained respectively. After all of the samples were collected and (de)coded, they were each defrosted at room temperature before being aliquotted into 0.5 mL microcentrifuge vials and then refrozen to await analysis by ESI-LC-MS. Some samples were of sufficient volume to allow up to 13 aliquots to be produced, whereas some only afforded three aliquots (average number of aliquots from each of 107 samples = 8). All obtained data relating to the clinical urine samples are presented within appendix C.

[1] Greater tuberosity of humerus refers to the fracture of the head of a humerus.

4.2. Aims

The aim of this work was to utilise all of the developed 'metabonomic tools' described in chapter three to comprehensively analyse the obtained clinical urine samples by both positive and negative ionisation modes and reversed phase and hydrophilic interaction LC-ESI-MS. The resulting data were analysed using PCA and PLS to attempt to elucidate potential biomarkers that could be putatively linked to failed fracture healing[1], hopefully generating candidates for further research into the field, with an envisaged end goal of developing early tests for possible failed fracture healing, thus allowing earlier intervention and shorter healing times.

[1] Unfortunately, post-analysis of the clinical urine samples, it was found out that none of the recruited patients fractures went to non-union, or suffered delayed healing. Despite this, the resulting data obtained was analysed to attempt to elucidate potential biomarkers that could putatively be related to the fracture healing process instead.

4.3. Results from RP-LC-ESI-MS analysis of clinical samples

Prior to analysis by positive and negative RP- and HILIC-LC-ESI-MS, the aliquotted clinical urine samples were removed from storage at Smith & Nephew (-80 °C) and placed in a container with dry ice, for transport to the Departments of Chemistry or Biology. As there were limited aliquots available from the clinical samples, pooled urine samples were created from the urine samples collected from volunteers from within the Department of Chemistry. These 'pool' samples were used for system equilibration and ongoing system stability monitoring. All optimised methods from the 'metabonomics toolbox' (chapter three) were employed for all analyses described within this section. For the RP-LC-MS analysis of clinical urine samples, each sample was defrosted at room temperature, vortex mixed, centrifuged and then filtered to remove any remaining sediment before being transferred into LC autosampler sample vials, ready for analysis.

4.3.1. Positive mode RP-LC-ESI-MS analysis of clinical urine samples

The LC-MS system was equilibrated by the injection and separation of three identical pooled urine aliquots. The TICs shown in figure 4.3.1 all follow the same trend, except the first TIC (blue line) that shows some deviations in peak intensity (indicated by arrows) from subsequent pooled sample TICs. The observed deviation for the first pooled sample was expected from previous results (chapter 3.3) as the LC-MS system equilibrates; as subsequent TICs all followed the same trend with no deviation in intensity or retention time, the system was considered equilibrated and thus ready for the analysis of clinical urine samples.

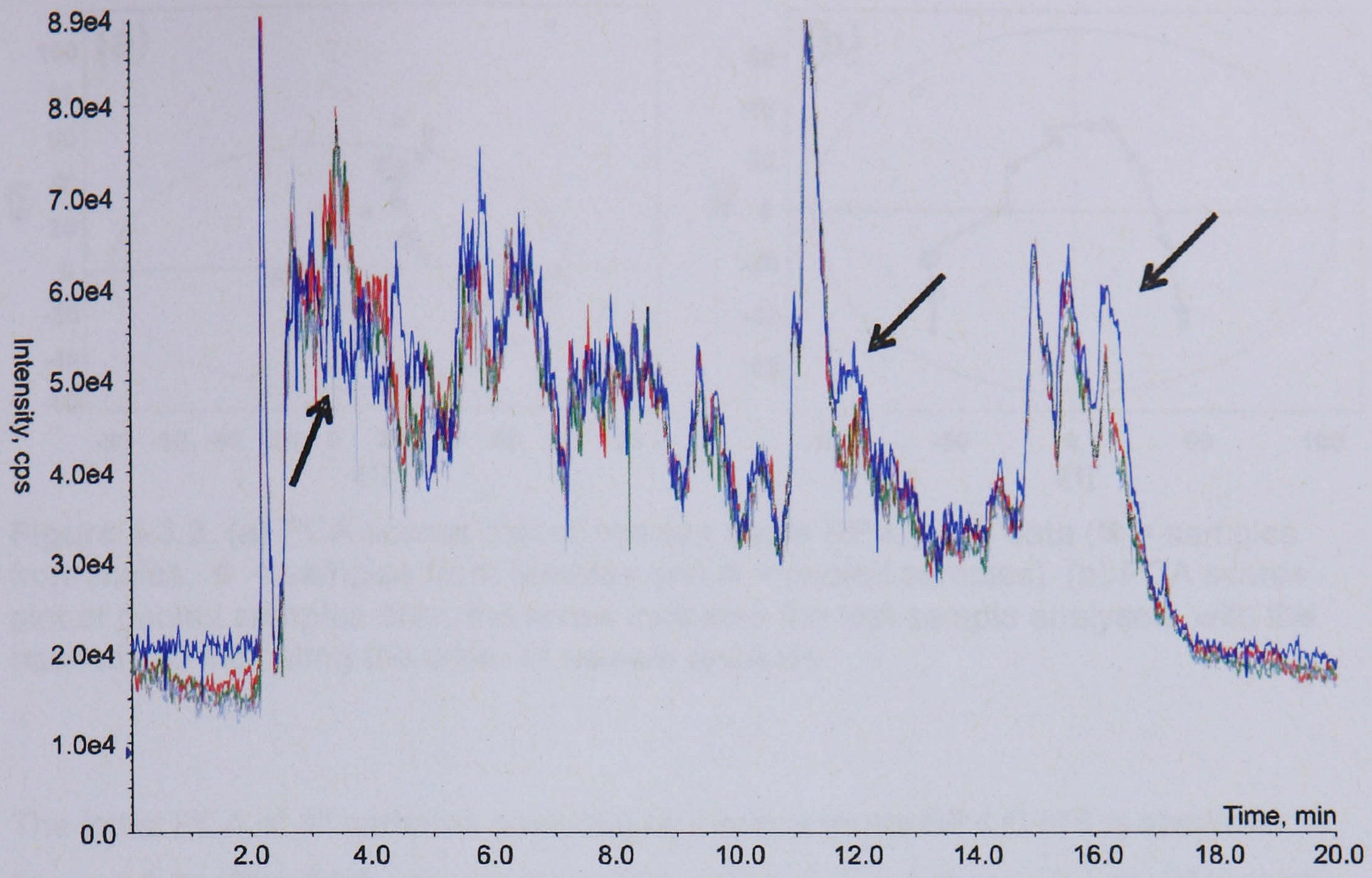


Figure 4.3.1. Five overlaid TICs from positive mode RP-LC-MS analysis of five pooled urine samples; three back-to-back analyses carried out at the very start of the analysis to monitor system equilibration, and two subsequent analyses for ongoing system stability monitoring. The arrows indicate areas where the first pooled samples TIC (blue line) deviate from subsequent TICs.

So that samples suffered minimal possible degradation, only 10-14 samples were defrosted and prepared for analysis at any one time. Subsequent sample batches were prepared and loaded into the autosampler as the analysis proceeded. Pooled samples and random repeats of clinical samples were included throughout data collection, meaning a total of 130 samples were analysed (3.5 days of analyses).

After the analysis of all samples, the raw data were extracted using the metabolomics export script (Applied Biosystems) to form a matrix for import into Excel (Microsoft Excel for Mac 2004), where data relating to each sample were added into a spreadsheet, before being subsequently imported into SIMCA P+ v11.5 (Umetrics, Sweden) for statistical analysis.

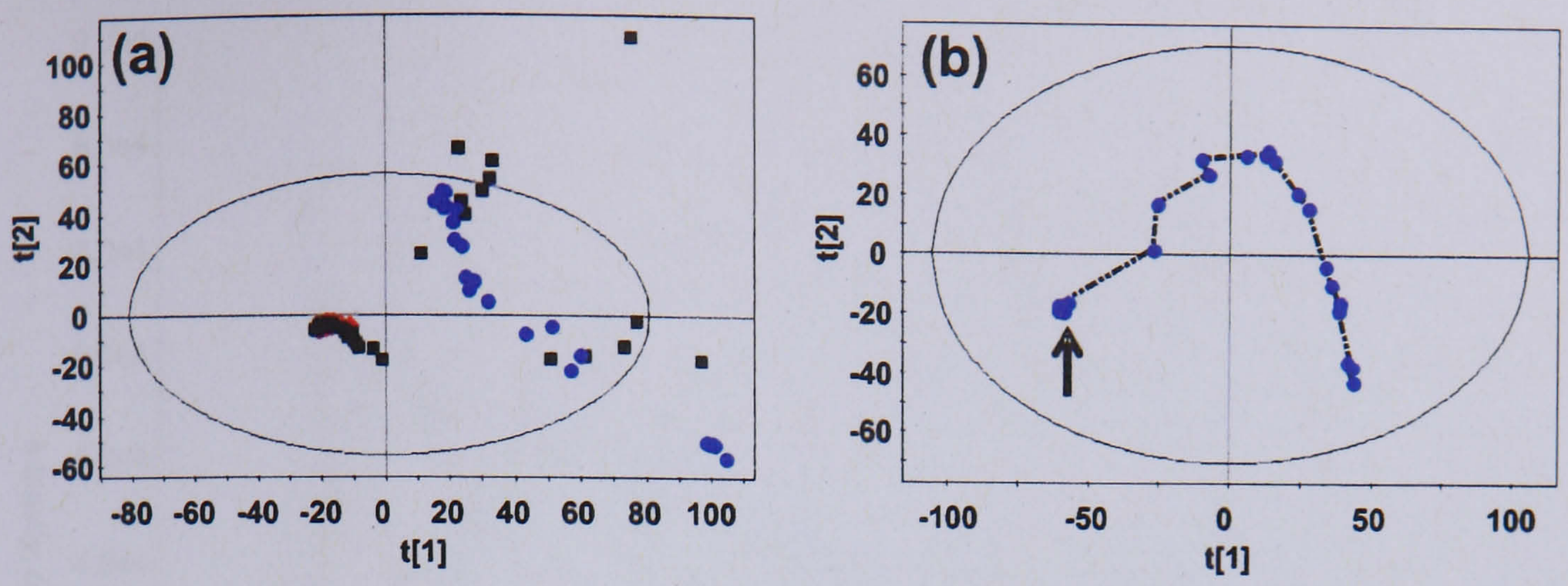


Figure 4.3.2. (a) PCA scores plot of positive mode RP-LC-MS data (■ = samples from males, ◆ = samples from females and ● = pooled samples). (b) PCA scores plot of pooled samples only, the arrow indicates the first sample analysed, with the hashed line indicating the order of sample analysis.

The initial PCA of all samples analysed by positive mode RP-LC-MS is shown in figure 4.3.2a. The PCA scores plot shows a tight cluster between 0 and -20 on both axes (PC 1 and PC 2), with all female data points within this cluster (red diamonds, masked by male data points), and the majority of the male data points (black squares). The remaining male data points are either outside the 95 % confidence limit (shown by an oval in the scores plot) or are far away from the bulk of the data in the tight cluster. The pooled samples (blue dots) appear to cluster apart from the samples analysed before any of the clinical samples (grouped at ~100 on PC 1 and ~50 on PC 2). This trend was unexpected, as the pooled samples should reside within the same area upon a PCA scores plot, given that they are the same sample just analysed a number of times.

PCA analysis of the data from just the pooled samples displayed a 'U' shape in the resulting scores plot (figure 4.3.2b) when the points were joined up in order of analysis (shown by a hashed line). The arrow indicates the three initial equilibration injections, which all form a tight cluster, with subsequent pooled samples moving away from this point. Visual inspection of the TICs of the initial injections did not highlight any substantial differences (figure 4.3.1), which is why the initial samples form a tight cluster. However, TICs of later injections of the pooled samples illustrate the reason for the 'U' trend observed in the scores plot.

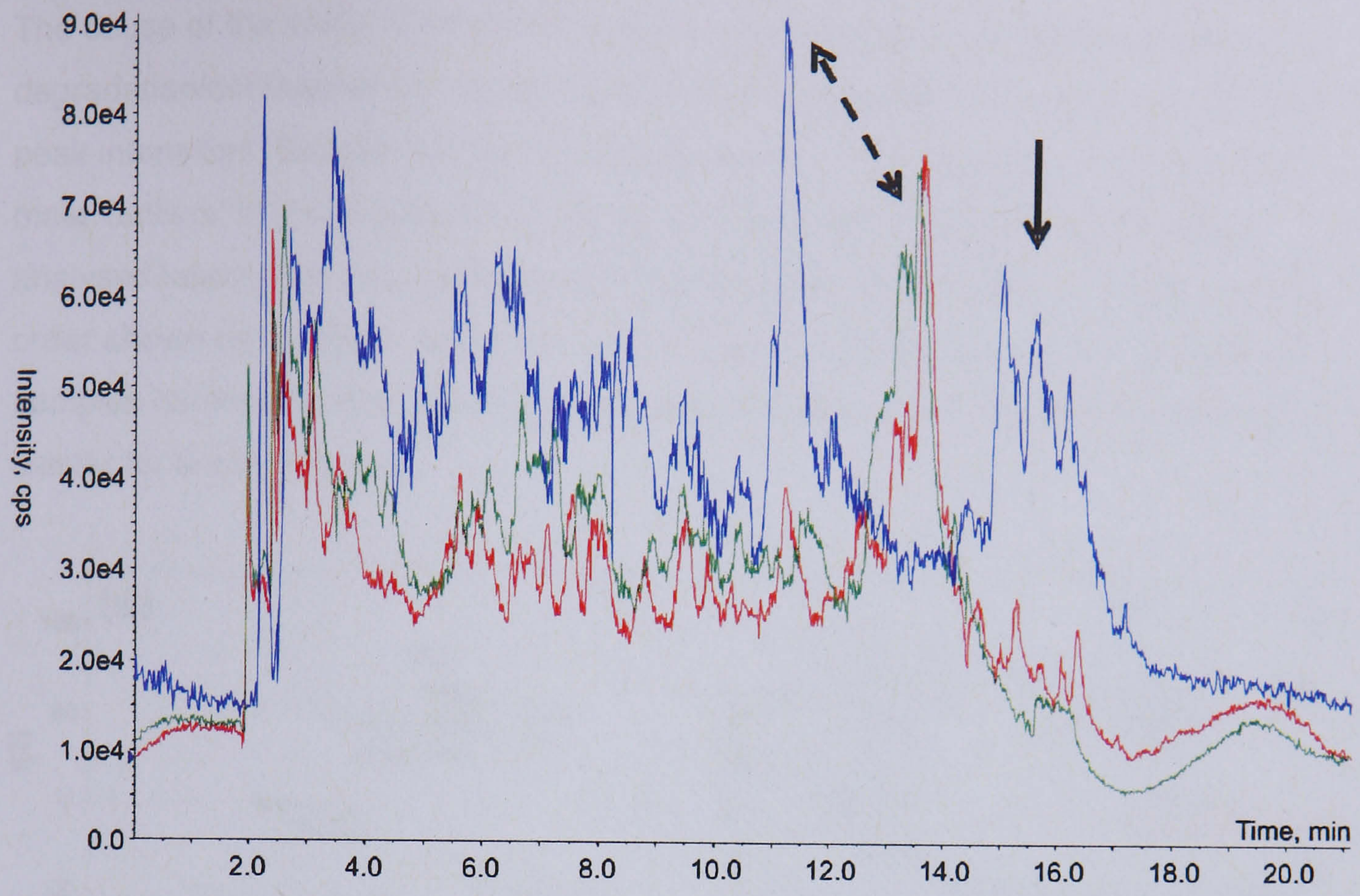


Figure 4.3.3. Three TICs from positive mode RP-LC-MS analysis of pooled urine samples. The blue TIC was from the initial analysis, with the green and red TICs obtained towards the end of sample analysis. The hashed double-headed arrow shows a large shift in the retention time of the most intense peak, and the solid arrow shows the severe reduction in intensity for an intense set of peaks.

Figure 4.3.3 shows superimposed TICs from three injections of the pooled sample. The blue TIC is from one of the initial injections run prior to the analysis of clinical samples, and the green and red TICs correspond to analyses towards the end of data acquisition (~3.5 days) to highlight the large shifts in retention time and peak intensity. The green and red TICs show a significant difference from the original TIC, with substantial deviation (~3 min) in retention time of the most intense peak indicated by the hashed double-headed arrow, and the disappearance of an intense set of peaks highlighted by the solid arrow. The large deviations in retention time and intensity occurred gradually, with a systematic drift in both the retention time and intensity of the peaks over the 3.5 day acquisition period; this means that some peaks that were present in the extracted data for the initial pooled sample injections would fail to be present in subsequent injections (an increasing number of peaks were 'lost' as time progressed). The systematic deviation in retention time and peak intensity explains the 'U' shaped trend present in the PCA scores plot for the analyses of the pooled sample (figure 4.3.2b).

The cause of the shifts in retention time and intensity was consistent with column degradation/contamination, which would cause large shifts in the retention time and peak intensities. Despite the trend observed for the pooled sample analyses, the male 'outliers' in the scores plot in figure 4.3.2a do not correspond to samples analysed later in the run, nor do they follow any time related trend (sample analysis order shown by numbers along the x-axis in figure 4.3.4b); the random analysis of samples controls for this, meaning that these samples did not form part of the tight cluster for another reason.

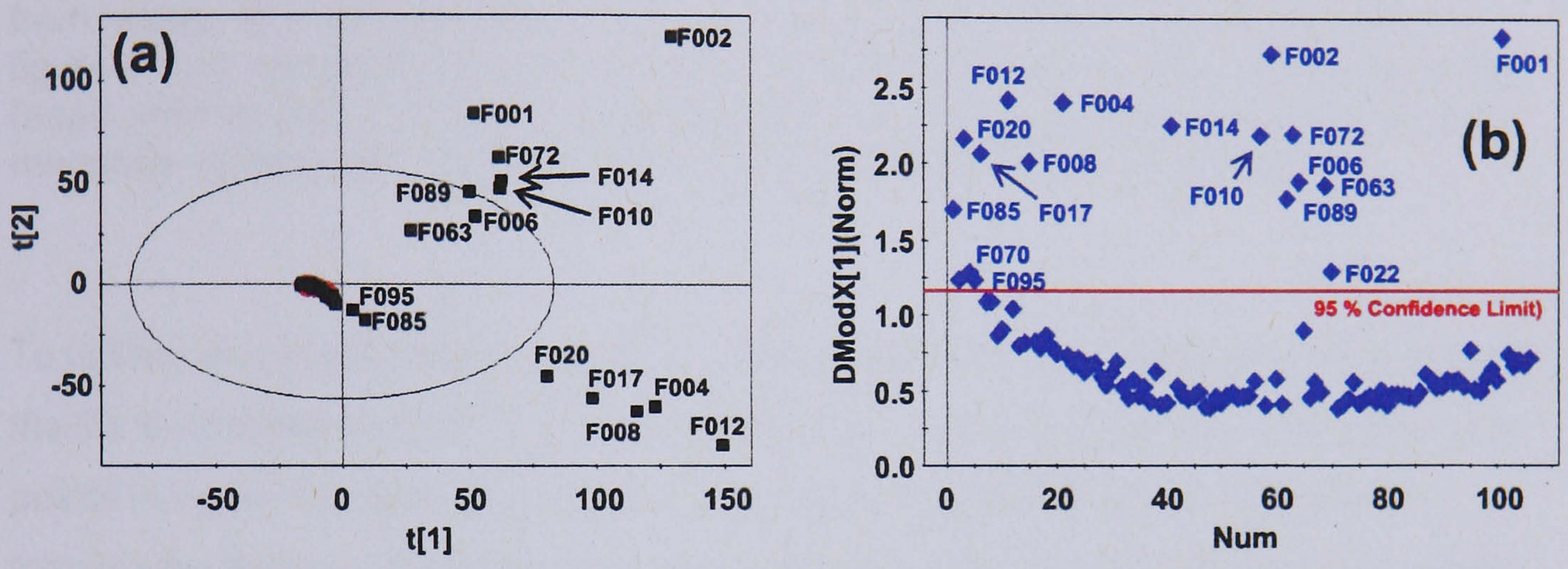


Figure 4.3.4. (a) PCA scores plot of positive mode RP-LC-MS data (■ = samples from males, ◆ = samples from females), with all samples outside of the tight cluster labelled (see appendix B). (b) Corresponding DModX plot of the data, where data points above the red line (95 % confidence limit) have large distances from the model.

PCA of the clinical samples (without the pooled sample data) results in more data points lying outside the 95 % confidence limit (figure 4.3.4a), compared to PCA of all analyses (figure 4.3.2a); again, there is a tight cluster containing the bulk of the data. When the data presented in figure 4.3.4a were viewed in 3D (not shown), the tight cluster containing the bulk of the data points formed along the same plane, with 13 data points lying outside of (or close to) the 95 % confidence limit on a different plane to the bulk of the data. This is highlighted by the many data points above the 95 % confidence limit (red line) in the corresponding DModX plot (figure 4.3.4b), where the majority of the labelled data points correspond to those labelled in the scores plot (figure 4.3.4a); the few additional data points above the confidence limit correspond to a few samples that were not outliers in the PCA scores plot, but were slightly 'out of plane' from the bulk of the data.

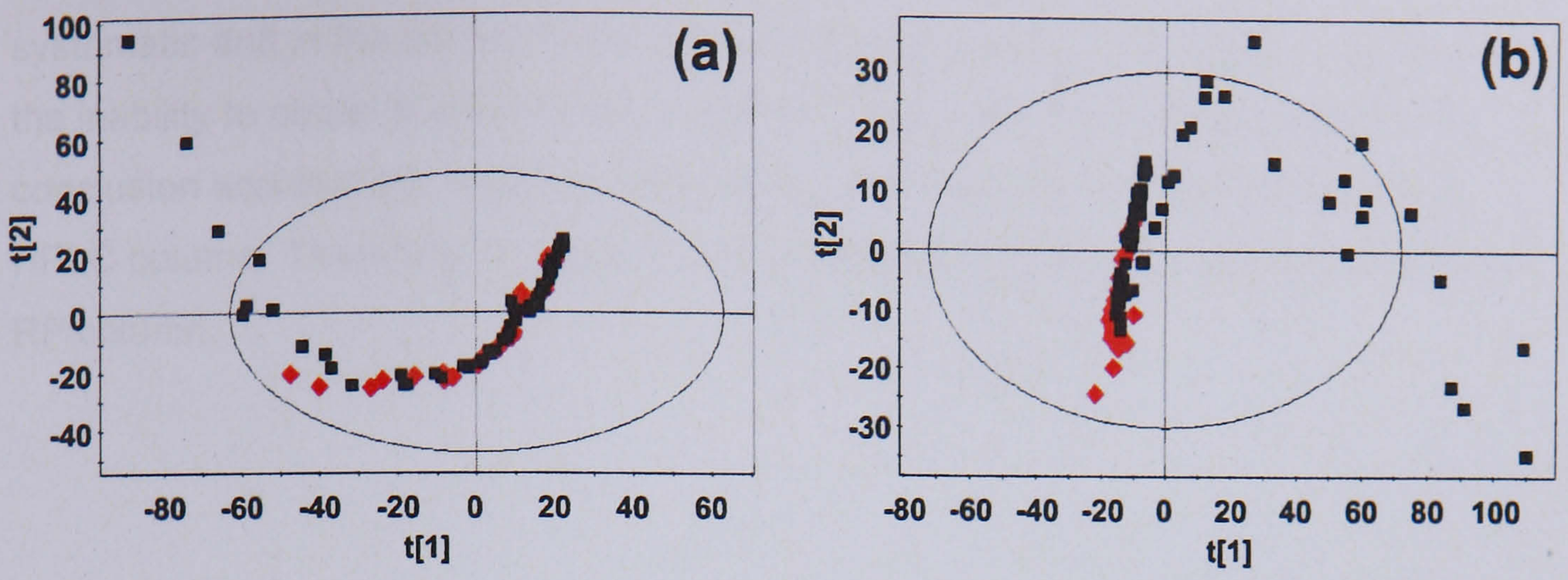


Figure 4.3.5. (a) PCA scores plot of positive mode RP-LC-MS data (■ = samples from males, ◆ = samples from females), with all data points that were labelled in figure 4.3.4a removed. (b) Resulting PLS analysis of positive mode RP-LC-MS data (again with all data points that were labelled in figure 4.3.4a removed) for a gender response variable (■ = samples from males, ◆ = samples from females).

To further investigate these trends, the data points with large DModX values (above the 95 % confidence limit, figure 4.3.4b) that also corresponded to the labelled data points in figure 4.3.4a were removed; thus PCA was performed on the data sets forming the tight cluster. The resulting PCA scores plot (figure 4.3.5a) shows a well defined trend within the data; many data points cluster together, with some data points tailing off at values lower than -10 along PC 1. Nothing in the information obtained on the patients could be used to explain the effects observed in the first two, and any subsequent, PCs. The clustering observed even with the 'outlying' samples removed, shows a systematic drift of the data. This was further shown when PLS analysis of the whole data set (excluding pooled samples) failed to discriminate by gender (figure 4.3.5b); this should have been easily obtained as was shown in chapter 3.6.

As for PCA, there is a cluster containing the bulk of the data (both male and female data points), with other data points moving away from the cluster and spreading out. Surprisingly, even when all variables were used to build the PLS model, there was no apparent clustering based upon discrimination by gender. The data points from 30 to 110 along the first latent variable (LV) correspond to the 13 data points that had large DModX values and were close to, or outside of, the 95 % confidence limit in the original PCA scores plot (figure 4.3.4).

Given that the data obtained were unsuitable for subsequent analysis due to the unacceptable drifts in both retention times and peak intensity, shown by the

systematic drift in the pooled urine sample injections upon PCA (figure 4.3.2b), and the inability to obtain clustering according to gender using PLS. The unfortunate conclusion was that the analysis was flawed, probably due to degradation of the HPLC column. Therefore, It was decided to analyse the clinical samples using a new RP column.

4.3.2. Negative mode RP-LC-ESI-MS analysis of clinical samples

The new column (and LC-MS system) was equilibrated and tested by the analysis of five aliquots of the pooled urine sample before the analysis of any clinical urine samples. As for positive mode RP-LC-MS analysis, the first aliquot analysed exhibited a TIC that was marginally different from the four subsequent near-identical TICs (not shown).

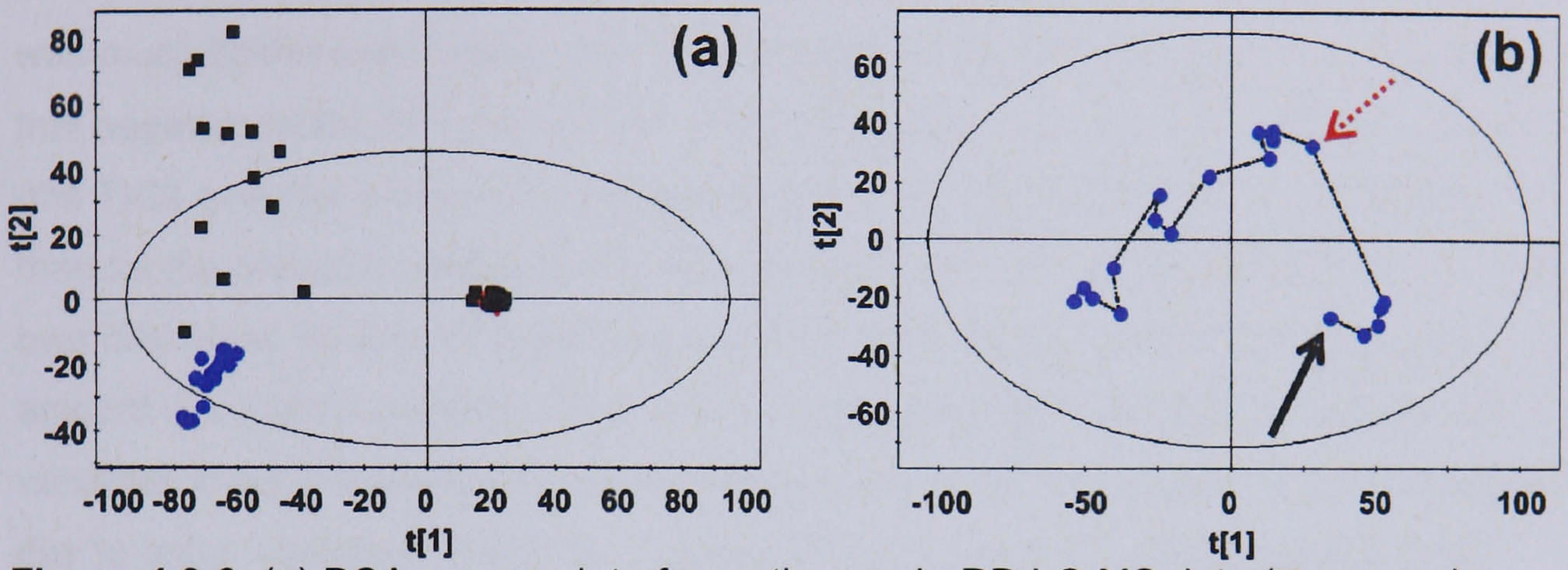


Figure 4.3.6. (a) PCA scores plot of negative mode RP-LC-MS data (■ = samples from males, ◆ = samples from females and ● = pooled samples). (b) PCA scores plot of pooled samples only, the solid black arrow indicates the first sample analysed and the hashed red arrow the sixth, with the hashed line indicating subsequent sample analysis.

The PCA scores plot for the analysis of extracted negative mode RP-LC-MS data is shown in figure 4.3.6a, and includes both the clinical and pooled sample data. The pooled sample data points cluster in the lower left hand side of the scores plot, with far less variation between samples than for the positive mode RP-LC-MS data (figure 4.3.2a); this suggests that whilst there is some variation between repeat injections of the pooled sample, on the whole, the system was more stable than observed in the positive mode RP-LC-MS analysis. The clinical data show two distinct clusters, a small tight cluster (centered around 20 and 0 on PCs 1 and 2 respectively) that contains the bulk of the data points (all female samples and the majority of the male samples), with the remaining samples forming a loose cluster between -40 and -80 on PC 1.

Analysing the data points from the pooled urine sample injections by PCA shows a trend within the data (figure 4.3.6b); the first few analyses (before the analysis of any clinical samples, with the first sample indicated by the solid black arrow) were similar, due to their close proximity to one another in the scores plot. The subsequent pooled

urine sample data points (indicated by a hashed red arrow) show an instant shift in position from the initial data points; after the initial shift, a more stepwise trend is observed. The PCA of positive mode RP-LC-MS data from pooled samples (figure 4.3.2b) showed a more stepwise pattern from the beginning, than that observed for negative mode RP-LC-MS data.

Despite the trend observed upon PCA of the pooled sample data, the initial cluster observed in the PCA scores plot of all data analysed by negative mode RP-LC-MS was much tighter than that observed for positive mode RP-LC-MS data, suggesting that negative mode RP-LC-MS data analysis produced slightly more reliable results (the TICs over the whole run exhibited less shifting in retention time and intensity than for the previous analysis). Analysing the pooled sample aliquots by PCA on their own only gives an idea of any trends, as PCA by definition seeks to find the most amount of variation between data points. Given that there are many thousands of variables in each dataset, the pooled sample injections' data points would not overlay due to minor variations within the dataset; PCA maximises these variations, thus 'increasing' the observed variation when pooled sample data are analysed using PCA.

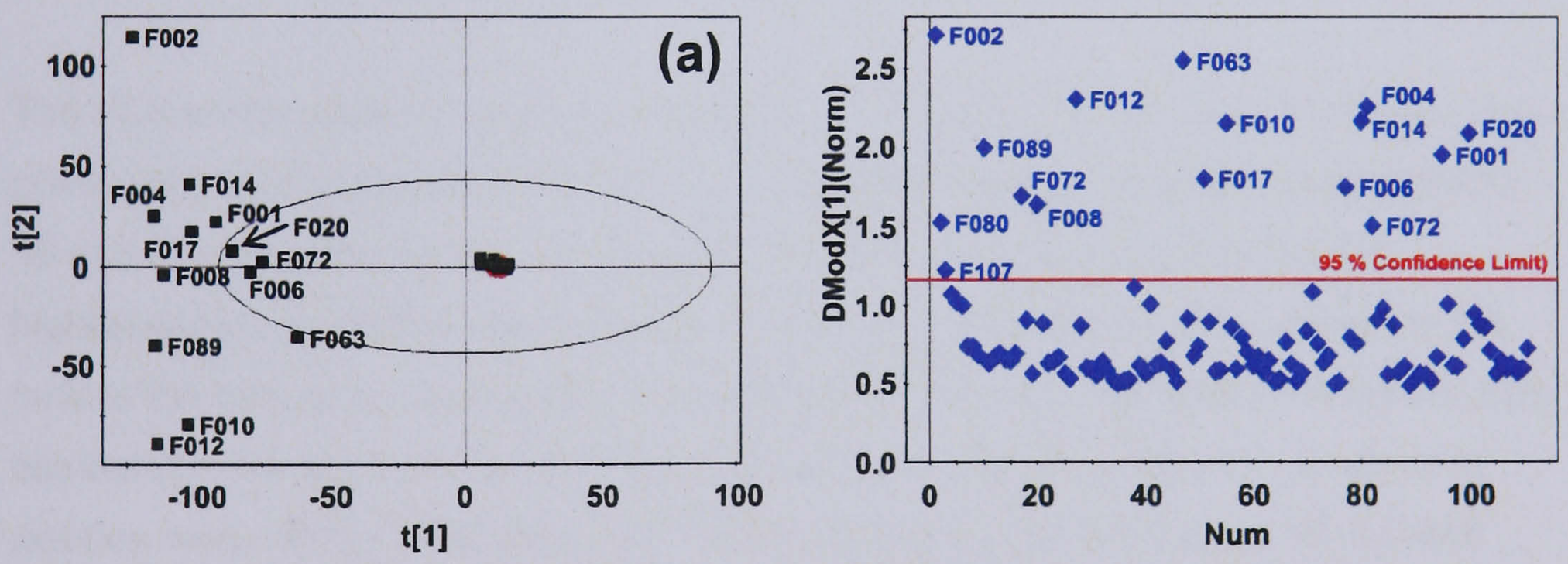


Figure 4.3.7. (a) PCA scores plot of negative mode RP-LC-MS data (■ = samples from males, ◆ = samples from females), with all samples outside of the tight cluster labelled (see appendix B). (b) Corresponding DModX plot of the data, where data points above the red line (95 % confidence limit) have large distances from the model.

When only the clinical sample datasets were analysed using PCA, the resulting scores plot (figure 4.3.7a) shows a tight cluster and a loose cluster, with samples spread over a broad range of values along the second PC. The majority of the labelled samples are outside, or very close to, the 95 % confidence limit. The labelled

data points all correspond to those labelled in the PCA from positive mode RP-LC-MS data (figure 4.3.4a). The corresponding DModX plot (figure 4.3.7b) shows many samples above the confidence limit; these samples correspond to those labelled in the corresponding PCA scores plot, again showing that these data points are 'different' to those in the tight cluster.

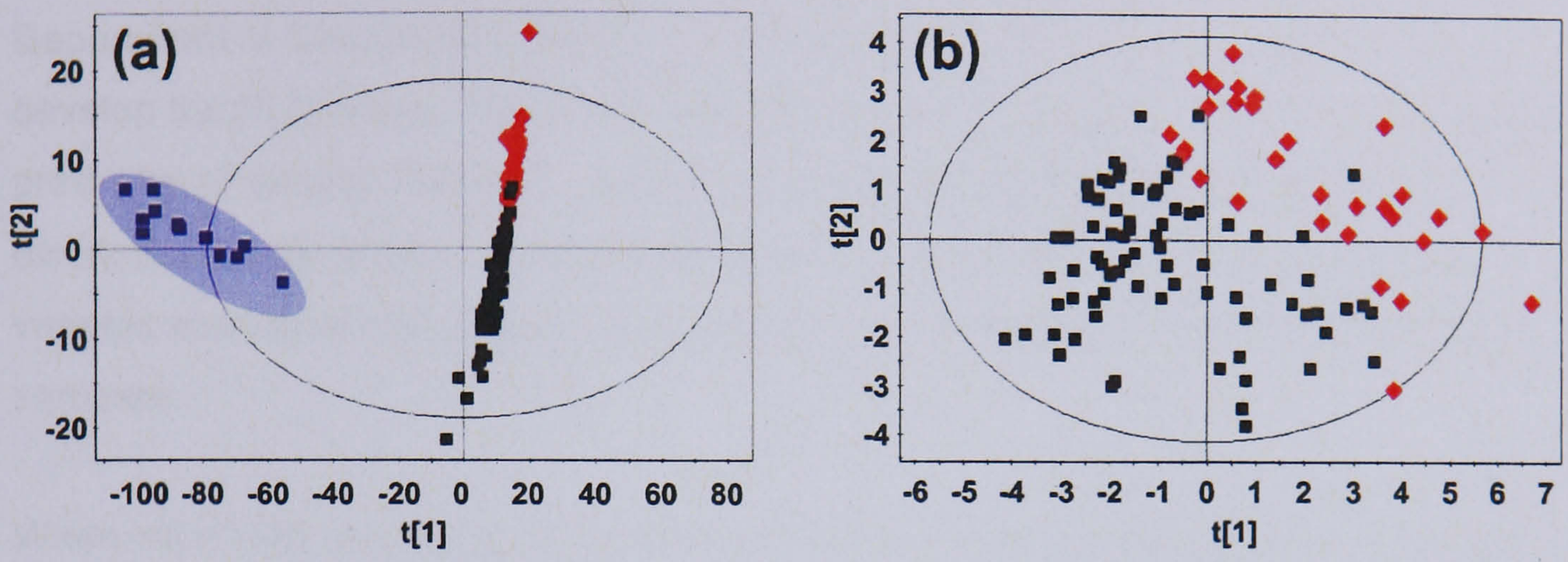


Figure 4.3.8. (a) PLS scores plot of negative mode RP-LC-MS data for a gender response variable (■ = samples from males, ◆ = samples from females), with all clinical data points included. The shaded ellipse contains all 12 samples that are 'different' from the bulk of the data. (b) Resulting PLS analysis of negative mode RP-LC-MS data (with all data from the shaded area in (a) removed) for a gender response variable (■ = samples from males, ◆ = samples from females).

The PLS scores plot for negative mode RP-LC-MS data (two thirds of all clinical data points, but not the pooled urine injections included) shows one long cluster between -10 and 20 along the first LV, and a large cluster below -50 along the first LV, highlighted by a shaded ellipse (figure 4.3.8a). The long cluster corresponds to the bulk of the data and contains all of the data points that also fell within the tight cluster observed in the PCA scores plot (figure 4.3.7a). In contrast to the PLS analysis of positive mode RP-LC-MS data, the PLS scores plot of negative mode RP-LC-MS data shows some degree of clustering based upon gender. The data points within the shaded ellipse all correspond to the labelled 'outliers' within the corresponding PCA scores and DModX plots (figure 4.3.7a and b).

Figure 4.3.8b shows the resulting PLS scores plot for a gender response variable when the initial model (figure 4.3.8a) was optimised by the removal of unimportant variables; there are two overlapping clusters that relate to samples obtained from male and female patients (in contrast to the PLS analysis of positive mode RP-LC-MS data, where no developed model could discriminate by gender). The data points from the shaded ellipse (figure 4.3.8a) no longer cluster apart from the bulk of the

data, but now form part of the cluster containing data points corresponding to male patients.

The external classification using the third of the dataset that was held back gave an external classification rate of 42 %, which is very low compared to 88 % obtained using negative mode RP-LC-MS analysis of urine samples collected from within the Department of Chemistry (chapter 3.6). This suggests that the variables used to develop the PLS model, yielding the separation shown in figure 4.3.8b, are not overly predictive of gender. However, despite the poor external classification rate for discrimination by gender, the fact that clustering based upon a gender response variable was observed showed some promise for the further analysis of the clinical samples.

When HILIC-MS was used for the analysis of the clinical samples (section 4.6), the possible cause of the issues observed during the positive and negative mode RP-LC-MS analyses of the clinical samples became evident, and was thus investigated further.

4.4. Proteomic analysis of clinical urine samples

Upon preparation of the clinical samples for analysis by HILIC-ESI-MS, they were diluted by addition of an equal volume of MeCN for injection onto the HILIC column. This had the effect of producing large amounts of precipitate in the majority of samples, something that was not observed with any of the pooled samples, or indeed any other urine samples analysed during the course of my PhD. Given that the RP gradient used for the analysis of the clinical samples utilised MeCN to elute hydrophobic compounds, the clinical samples, once injected onto the column, would have precipitated in the same manner on the column. This is proposed to be the cause of the systematic degradation of column performance seen as drifts in retention time and peak intensity over time observed on analysis of the clinical, but not the healthy volunteer samples (chapter three); the effects were most likely less evident for the negative mode RP-LC-MS analysis as this was carried out on a new column, which had therefore not suffered as much on-column precipitation as the column used for positive mode RP-LC-MS analysis. The level of precipitation caused by the addition of MeCN explained why, for the clinical samples, filtering was much harder (sometimes impossible) due to the increased viscosity; differences in surface tension were also observed between the clinical and pooled samples when transferring filtered urine into sample vials (clinical samples had a higher surface tension, clinging to the vials much more than the pooled urine sample).

To test if the precipitate was protein, 10 µL of each of a random cross section of clinical samples were added to 200 µL of Coomassie solution (see 4.4.1); an immediate change from dull red to bright blue was observed, suggesting protein was present. To further test whether the clinical urine samples contained protein, 1D Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis (SDS-PAGE) was run, with only 5 µL each from two samples loaded onto the gel.

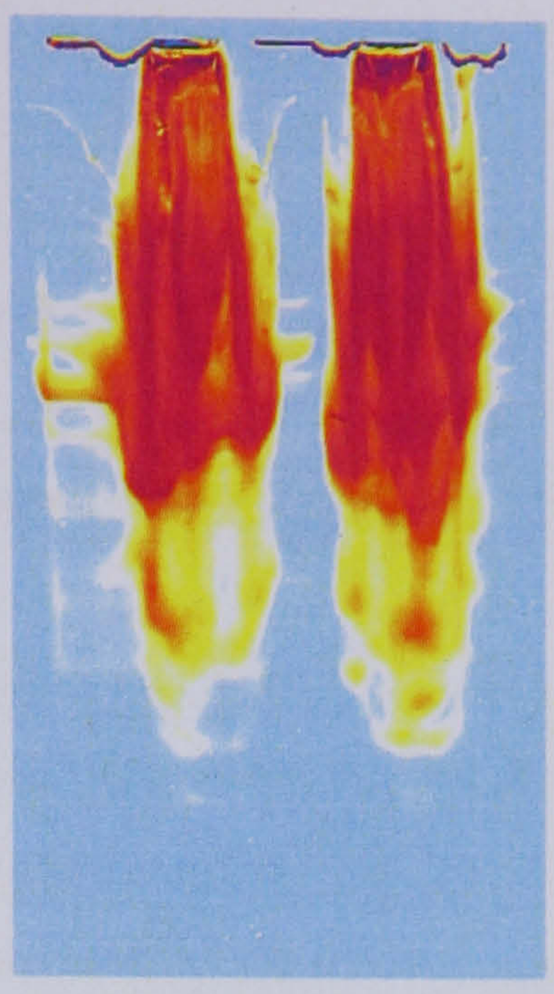


Figure 4.4.1. Coomassie stained 1D SDS-PAGE of two clinical urine samples (colour altered for clarity, red = higher concentration of protein).

Following Coomassie brilliant blue staining, the 1D SDS-PAGE of the two clinical urine samples shows a heavily overloaded gel; positively staining material in both lanes has spread outside of their lane due overloading (figure 4.4.1). As bands relating to different proteins could not be resolved, the samples required dilution before any protein identification could take place. To allow dilution of the clinical urine samples to the correct concentration for SDS-PAGE analysis, each of the clinical samples was analysed using a Bradford assay to determine the concentration of protein present.

4.4.1. Bradford assay of clinical urine samples

The Bradford assay was first described by M Bradford in 1976 (Bradford, 1976), and is a semi-quantitative colourimetric protein assay. The principle behind a Bradford assay is the colour shift of Coomassie (a dye) from a dull red at 465 nm, to a bright blue at 595 nm upon binding of protein. Coomassie binds to protein through different interactions such as van der Waals forces and ionic interactions; hydrophobic aromatic amino acids such as Phe, Try and Pro also aid in binding, as does hydrophilic Arg (Compton and Jones, 1985).

As the binding of Coomassie to protein occurs in a stoichiometric manner, the increase in absorbance at 595 nm is proportional to the concentration of protein within a sample. However, time related shifts in intensity occur, so the assay must be

completed in a timely fashion to avoid inducing error into the results obtained. Bradford assays are only linear over a small range of protein concentrations; typically 125-1500 µg/mL dependent on the protein being bound (Zor and Selinger, 1996). A dilution series of a random selection of clinical urine samples showed that 5 µL in 1000 µL H_2O was sufficient for the majority of samples to fall into the linear range, thus highlighting how concentrated the protein was in some of the clinical samples.

Two 10 µL portions of each diluted clinical sample were loaded into wells on a 96-well plate (providing a duplicate), along with duplicates of eight protein standards ranging from 0-1500 µg/mL, meaning a total of four 96-well plates were analysed (each 96-well plate therefore containing a full set of standards and a subset of clinical urine samples). 200 µL of Coomassie was loaded into each cell containing 10 µL of sample or standard, and then agitated for 30 s before being analysed by UV-Vis absorbance at 595 nm.

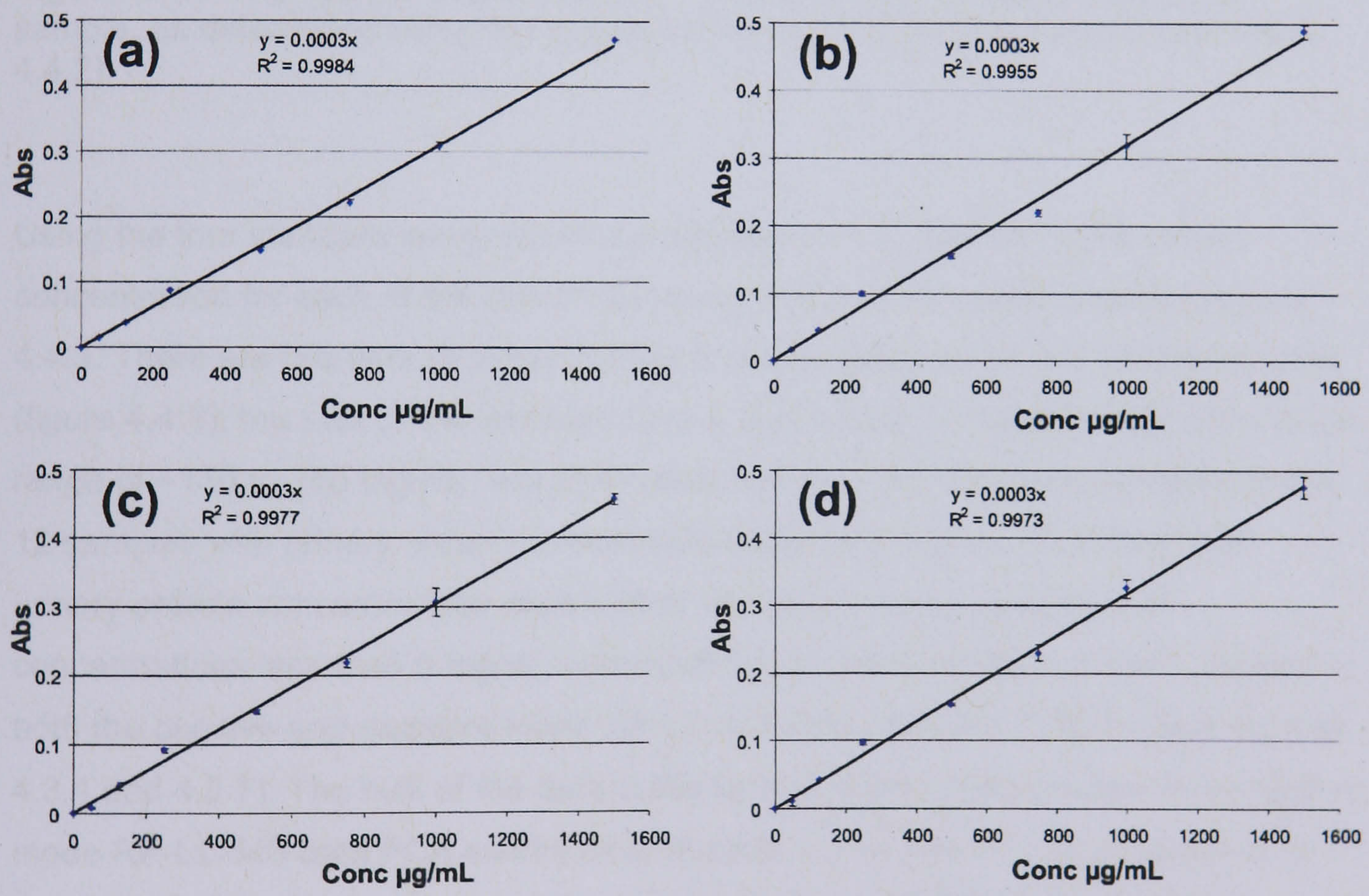


Figure 4.4.2. Standard curves for all four 96-well plate standard protein samples.

Figure 4.4.2 shows four standard curves used for the determination of the concentrations for each clinical sample present on each 96-well plate. Each of the standard curves goes through the origin and yields the same y = mx equation (where

m = 0.0003), as well as having R^2 values greater than 0.99, indicating an acceptable 'goodness of fit'.

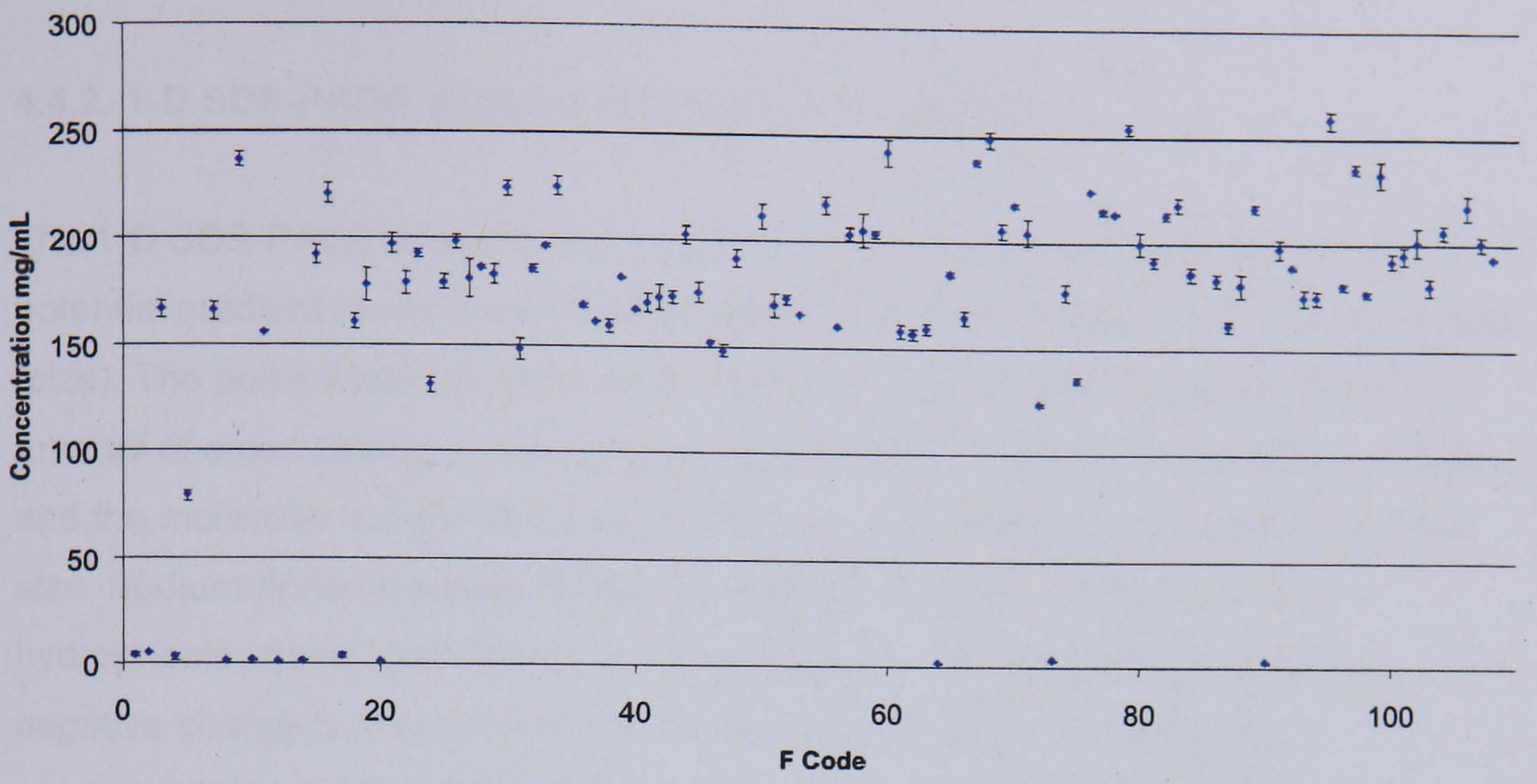


Figure 4.4.3. A graph plotting the protein concentrations of each clinical urine sample, as determined using the equation from each of the standard curves (figure 4.4.2).

Using the four standard curve equations (figure 4.4.2) to determine the protein concentration for each of the clinical samples, produced the graph shown in figure 4.4.3. There are two tiers of protein concentrations observed for the clinical samples (figure 4.4.3); the bulk of the samples have a total urinary protein concentration in the range of ~150 to 250 mg/mL, which corresponds to a ~40 fold increase between the 12 samples with urinary protein concentrations below 5 mg/mL. When the total urinary protein concentrations were further analysed, the 12 samples with concentrations less than 5 mg/mL corresponded to the samples that were 'outliers' in both the positive and negative mode RP-LC-MS data PCA and DModX plots (figures 4.3.4 and 4.3.7). The bulk of the data in the tight clusters on the positive and negative mode RP-LC-MS data PCA scores plots (figures 4.3.4a and 4.3.7a) correspond to the clinical samples with protein concentrations above 75 mg/mL (figure 4.4.3).

Values for normal urinary protein excretion rates reported in the literature vary greatly. The most common value found was ~150 mg/day total urinary protein excretion (Pisitkun *et al.*, 2006; Tyan *et al.*, 2006; Gonzalez-Buitrago *et al.*, 2007). This clearly means that the level of protein excretion observed for the majority of the clinical samples was far from normal. In order to determine if there were one or more

types of protein present within the clinical samples, 1D SDS-PAGE analysis of a cross section of samples was performed.

4.4.2. 1-D SDS-PAGE analysis of clinical urine samples

The 1-D SDS-PAGE analysis of proteins works by using polyacrylamide gel and a potential gradient to separate proteins based upon their charge and molecular weight (size). The polyacrylamide gel is cross-linked, with pore sizes determined by the amount of cross-linking; larger proteins take longer to migrate than smaller proteins, and the molecular weight (MW) range that can be separated is determined by pore size. Sodium dodecyl sulfate (SDS) denatures proteins by interacting with hydrophobic chains, and applies an overall negative charge to the protein; the negative charge is proportional to the mass of the protein, which is why polyacrylamide gel is used to 'sieve' the proteins, giving separation; further denaturing is achieved by the addition of 2-mercaptoethanol and heating to near boiling, this reduces disulfide linkages and has the effect of creating a linear shaped protein.

After protein denaturing, the sample is loaded into a cell at the top of the polyacrylamide gel (submerged in an SDS running buffer), and then subjected to a voltage. The proteins migrate towards the anode at rates proportional to their charge and mass, with the largest proteins migrating the least due to their resistance in migrating through the gel structure.

After separation is achieved, the gel is stained using Coomassie brilliant blue (Chrambach *et al.*, 1967), which visualises bands of protein within the gel. MW markers are typically run in one of the lanes to allow the MW of protein bands observed to be estimated. A flow chart of the SDS-PAGE protocol used to analyse the clinical urine samples is summarised in figure 4.4.4:

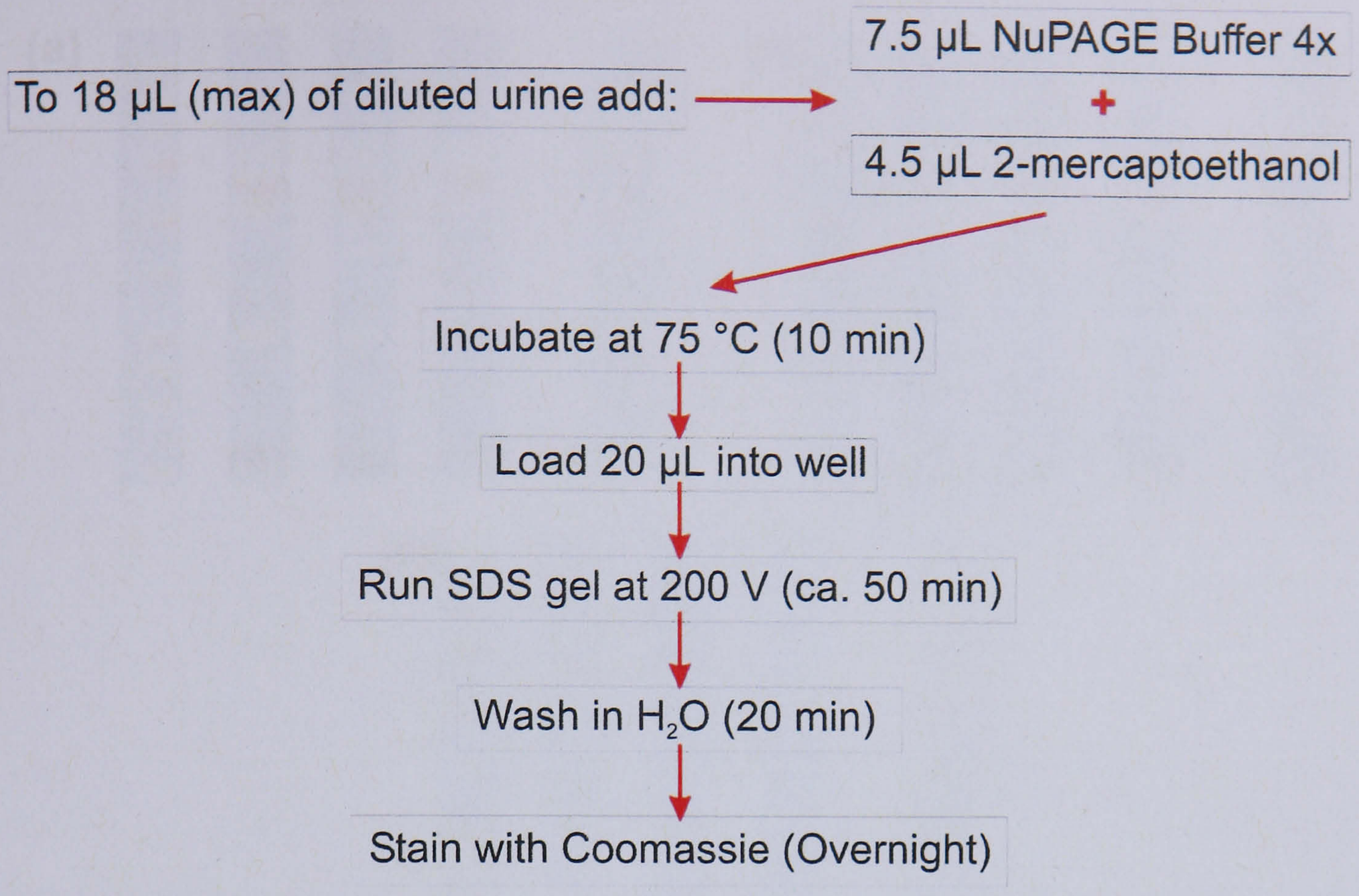


Figure 4.4.4. SDS-PAGE protocol for analysis of protein.

To obtain a representative picture of the protein content over all the clinical samples (to see if different proteins were present at different concentrations), three samples were chosen from three different protein concentration ranges for SDS-PAGE analysis. The most concentrated samples were F67, F94 and F78 (248, 258 and 253 mg/mL), with F11, F5 and F74 having median protein concentrations (156, 79 and 135 mg/mL) and F2, F89 ad F17 for the lowest concentration samples (5, 2 and 4 mg/mL).

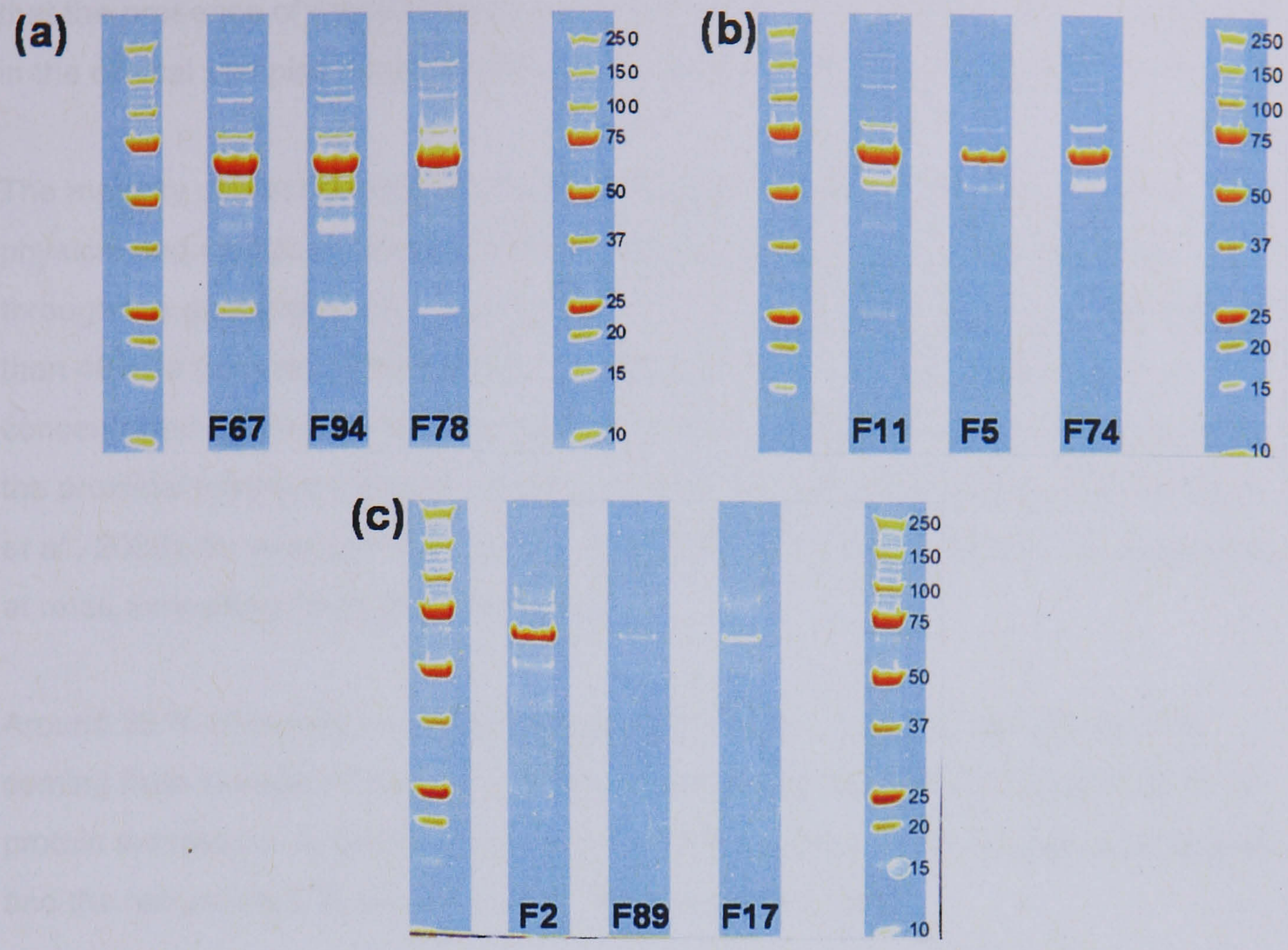


Figure 4.4.5. SDS-PAGE analysis of six randomly chosen clinical urine samples from three different concentration levels. The two outside lanes on each gel contained MW markers, the weight of which (kDa) is labelled in the right hand lane on each gel. Colour altered for clarity. (a) Highest protein concentrations of 248, 258 ad 253 mg/mL for samples F67, F94 and F78. (b) Median protein concentrations of 156, 79 and 135 mg/mL for samples F11, F5 and F74. (c) Lowest protein concentrations of 5, 2 and 4 mg/mL for samples F2, F89 and F17.

Each of the three clinical urine samples from the three different levels of protein concentrations were analysed on separate gels, with MW markers used in each of the outside lanes to aid in estimating the MWs of the different protein bands that were present (figure 4.4.5). The most concentrated urine samples (figure 4.4.5a) show that different proteins were present in these clinical urine samples, with a MW spread from ~25 to 250 kDa suggesting the presence of some large proteins. The median (~125 mg/mL) and low concentration (<5 mg/mL) cohorts (figure 4.4.5b and c) show the same bands are present, but at lower intensities. The most intense band across all samples is at ~70 kDa.

Pisitkun *et al.* reported "*…a myriad of proteins and peptides can be detected in normal urine…*" (Pisitkun *et al.*, 2006); they found the presence of more than 1000 different protein gene products and many more peptides in human urine, suggesting

that the presence of many proteins is normal. However, the concentrations observed in the clinical samples would be considered nephrotic (Dihazi and Muller, 2007).

The majority of plasma protein should be retained, as the glomerulus acts as a physical and electrical charge barrier to most proteins. Any proteins that pass through the glomerulus into the proximal tubule are generally small proteins, less than 40 kDa (Gonzalez-Buitrago *et al.*, 2007), or are those which are very concentrated within plasma. Despite this, lots of proteins and peptides that pass into the proximal tubule are scavenged/proteolysed, and therefore reabsorbed (Pisitkun *et al.*, 2006); for example, the most abundant plasma protein, albumin, is reabsorbed at rates exceeding 99 % (Sarti *et al.*, 2001).

Around 30 % of urinary proteins originate from plasma, with the remaining 70 % coming from kidneys (Tyan *et al.*, 2006; Dihazi and Muller, 2007). Roughly 49 % of protein excreted in urine is in the form of soluble proteins, 48 % consist of sediments, and the remaining 3 % are exosomes[1] (Pisitkun *et al.*, 2006).

The next step in the proteomic analysis of the clinical urine samples was to analyse the bands observed in the stained polyacrylamide gels, with the intention of identifying the proteins present.

[1] Sub 80 nm vesicles.

4.4.3. Protein identification by MALDI-ToF/ToF analysis

Proteomic identification is usually carried out by one of two different approaches, top-down and bottom-up. Top-down proteomics involves the analysis of intact proteins using high accuracy mass measurements and subsequent CID or other fragmentation method and tandem MS analyses. Bottom-up proteomics (used in this study) first involves the digestion (either separated or as a mixture) of proteins into peptide fragments (typically of 5-30 AA residues), usually using trypsin that cleaves on the C-terminal side of Lys and Arg (as long as they are not followed by Pro). The resulting peptide fragments are then analysed by MS and CID tandem MS. A summary of the bottom-up proteomics approach is shown in figure 4.4.6:

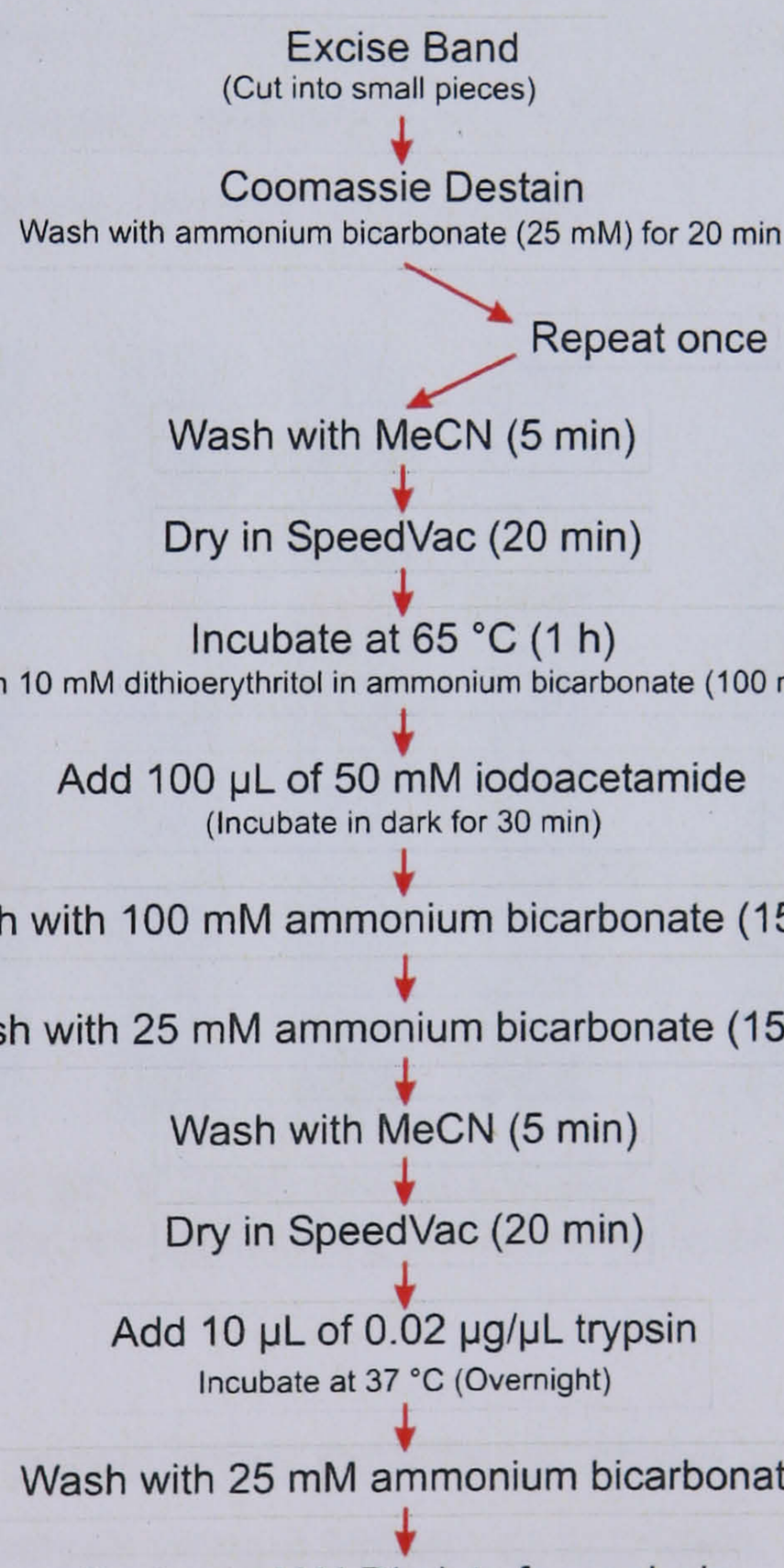


Figure 4.4.6. Flow chart of the bottom-up proteomic analysis of proteins.

Whichever proteomic approach is used, a database search is undertaken to compare the exported MS and/or MS/MS data with data expected from tryptic peptides of any of the protein sequences held in a database. A common search engine called MASCOT, used in this study[1], was developed by Perkins *et al.* in 1999 (Perkins *et al.*, 1999) as an update of the original MOlecular Weight SEarch (MOWSE) search engine and scoring algorithm developed in the early 1990s by Pappin *et al.* (Pappin *et al.*, 1993). The MOWSE search engine functioned by assigning a statistical weight to each peptide matched, which was based upon the frequency of other peptide masses in a protein with a certain MW range. MOWSE now forms part of MASCOT and is still the basis used to assign statistical confidence in any peptides matched. MASCOT uses a mathematical approach to compare theoretical fragment ions to the fragment ions recorded; confirmation of expected protein identification is obtained by linking the protein identified from a database to the mass observed in the original gel, as well as using the significance provided by the MOWSE score returned[2] and linking the role of the protein back to the biological system.

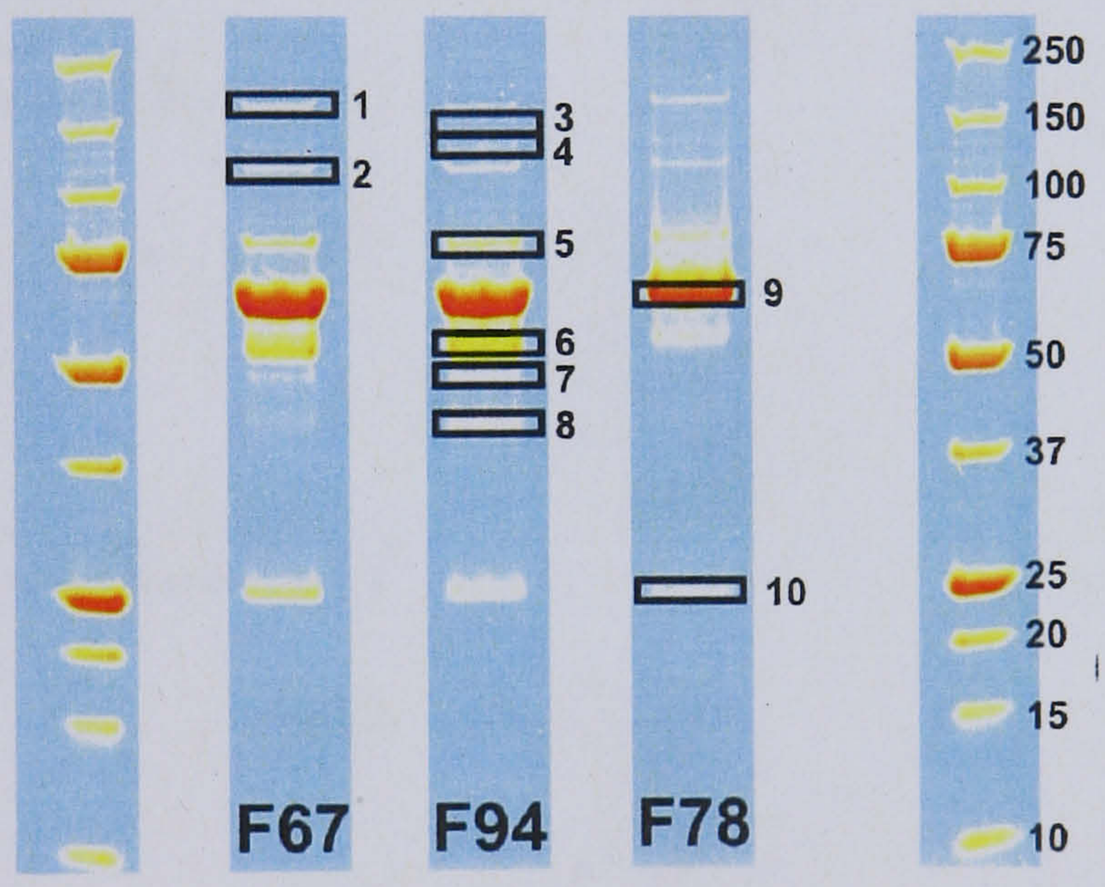


Figure 4.4.7. SDS-PAGE gel of three clinical samples F67, F94 and F78. Ten bands were excised for proteomic analysis using a bottom-up approach.

Ten of the most intense protein bands from the gel shown in figure 4.4.7 were excised for proteomic analysis using a bottom-up approach. Before each digested sample was spotted onto a MALDI plate for analysis, they were extracted using C_{18} ZipTips to remove any excess salts, with the aim of improving the S/N ratio.

---

[1] The database used for all searches was the NCBInr database, updated on 06/07/2007.
[2] An event is considered significant if it occurs at random with a frequency of less than 5 %, giving $p < 0.05$.

0.5 µL of each extracted sample was spotted onto a MALDI plate and covered with 0.5 µL of a solution of ∝-cyano-4-hydroxycinnamic acid and allowed to air dry. The resulting MS and MS/MS data were imported into GPS Explorer v3.6 (Applied Biosystems), which uses the MASCOT search engine to attempt to match any of the peptides present within each sample with those generated *in silico* from proteins in the NCBInr database.

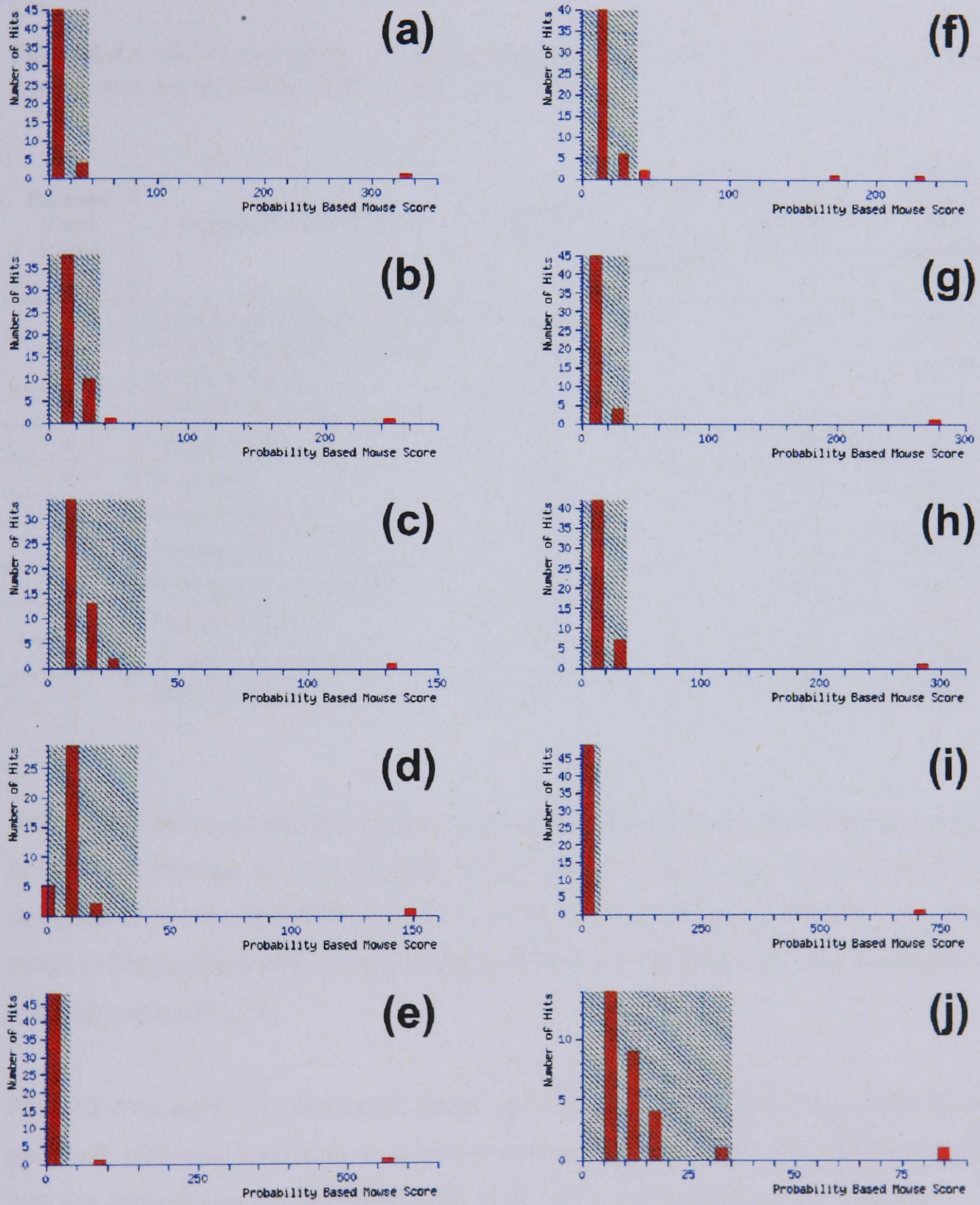


Figure 4.4.8. Probability based MOWSE score plots for each of the ten excised protein bands. (a) band 1, (b) band 2, (c) band 3, (d) band 4, (e) band 5, (f) band 6, (g) band 7, (h) band 8, (i) band 9 and (j) band 10.

A probability based MOWSE score plot shows any hits below the assigned confidence limit ($p < 0.05$) in a hashed green area, any other hits (red bars) above this green area are considered significant; the higher the MOWSE score the better. Each of the probability based MOWSE score plots shown in figure 4.4.8 shows hits above the significance level, meaning that convincing protein identifications were obtained from each of the ten excised bands MS and MS/MS data.

Table 4.4.1. Summary of each protein identified by MASCOT from the ten excised bands, along with the estimated mass from the gel.

Excised Band Label	Protein ID from MASCOT	MOWSE Score	Number of peptides Identified	Estimated mass from gel (kDa)	Mass of protein from MASCOT (Da)
1	Alpha-2-macroglobulin precursor	332	6	~170	164600
2	Chain B, human complement component C3b	246	5	~115	114238
3	Complement factor H	132	6	~150	143710
4	Ceruloplasmin	149	5	~120	116197
5	Transferrin	567	12	~80	79280
6	Alpha-1-antitrypsin	229	7	~50	46848
	Immunoglobulin G	165	3	~50	52687
7	Fibrinogen gamma chain	276	6	~48	47971
8	Preprohaptoglobin	285	6	~40	38940
9	Human serum albumin A	704	16	~70	67988
10	Apolipoprotein A-I	85	3	~25	28061

Table 4.4.1 summarises the results obtained from the bottom-up proteomic analysis of the most intense bands excised from the SDS-PAGE analysis of the clinical urine samples. From the MOWSE scores, and the closeness of each identified protein's mass to those observed in the gel, the protein identifications can be considered as confident assignments.

Alpha-2-macroglobulin precursor (band 1) is a binding host for foreign peptides and particles, and also functions as a barrier against pathogens (Borth, 1992), meaning that the protein subunit identified (165 kDa) forms part of the body's immune response. Band 2 corresponds to chain B – human complement component C3b, which is a single chain glycoprotein that is expressed in a large number of cells, and also has as immunological role. C3b is involved in clearing pathogens from an

organism by aiding cell lysis (Hamer *et al.*, 1998), and is mainly synthesised in the liver. The presence of C3b in human urine has been noted before by Pascual *et al.*, where it was present in low concentrations, showing that it was released by glomerular podocytes[1] (Pascual *et al.*, 1994). C3b is linked to both factor H and immunoglobulin, which were also detected in the clinical urine samples and correspond to excised bands 3 and 6 respectively.

Complement H (band 3) is a crucial fluid phase regulator that interacts with C3b, effectively deactivating C3b's function (Jokiranta *et al.*, 1999), thus regulating the body's immune response. The protein identified in band 6 corresponded to Immunoglobulin G, again an immune system protein. Immunoglobulin G provides the majority of antibody-based immunity against any invading pathogens.

Cerruloplasmin (band 4), transferrin (band 5) and antitrypsin (band 6) are all glycoproteins. Cerruloplasmin is synthesised in the liver and functions to transport ~90 % of the copper within plasma; similarly, transferrin is involved in the delivery of iron ions within plasma. Antitrypsin is a serine protease inhibitor that plays a key role in controlling the coagulation of blood, and is also related to controlling inflammation in the body. Linked to antitrypsin, is fibrinogen (band 7) that is also a blood clotting agent, which is produced within the liver.

Preprohaptoglobin (band 8) functions to provide a physiological defence against haemoglobin-induced toxicity (Ngai *et al.*, 2007). The by-product of preprohaptoglobin is haptoglobin, which binds to free haemoglobin to stop glomerular filtration of haemoglobin, preventing oxidative injury to the kidneys.

The presence of human serum albumin (HSA) was expected once protein was confirmed within the clinical samples, as it is the most abundant protein present in serum at a concentration of ~40 mg/mL HSA's function is to maintain osmotic pressure, and also the proper distribution of body fluids within serum; HSA also transports fatty acids, hormones and other physiologically important compounds around the body. In each of the three gels (figure 4.4.5), the band at ~70 kDa was the most intense, suggesting that HSA was the most abundant protein within all of the clinical urine samples.

---

[1] Glomerular epithelial cells.

The final protein identified was apolipoprotein A-I (band 10), which functions to extract cholesterol from body tissues and subsequently transport it to the liver where it can either be excreted or recycled (Lahoz *et al.*, 2003).

4.4.4. Discussion

The results presented within this section (4.4) were certainly a surprise, given all previously analysed urine samples did not show any precipitation upon the addition of MeCN prior to analysis by HILIC-ESI-MS. Research into urinary protein is not new, and is typically related to some physiological process such as glomerular diseases where increased levels of protein are excreted into urine (Sarti *et al.*, 2001; Christian and Watson, 2004; Pisitkun *et al.*, 2006; Barratt and Topham, 2007; Dihazi and Muller, 2007; Gonzalez-Buitrago *et al.*, 2007; Ngai *et al.*, 2007). The proteins identified within the clinical urine samples have all previously been reported in the literature, suggesting that the observation of these proteins is nothing new. However, what was in contrast to the literature was the sheer amount of protein detected within the clinical urine samples. Despite the semi-quantitative nature of Bradford assays, the levels of protein detected here (average = 166 mg/mL) were significantly higher than the average value excreted daily (despite the various different values reported in the literature) of ~150 mg/day (Pisitkun *et al.*, 2006; Gonzalez-Buitrago *et al.*, 2007). Given that a normal person produces 1-2 L of urine a day, the average protein concentration of 166 mg/mL found in the clinical samples would correspond to a daily total excretion of over 150 g/day, 1000 times the reported average daily excretion. However, many of the clinical samples were very concentrated from the observed colour (dark yellow), meaning that the patients may not have produced as much urine on the day of collection.

Other than benign causes of increased levels of protein in urine, such as fever/post exercise and also just having an upright posture (meaning an increased amount of protein present in urine towards the end of the day) (Newman *et al.*, 2000; Christian and Watson, 2004), a few papers have reported increases in urinary protein excretion caused by trauma or stress (Yu *et al.*, 1983; DeGaudio *et al.*, 1999; Sarti *et al.*, 2001). DeGaudio *et al.* reported an increase in capillary permeability, and therefore an increase in protein excretion with trauma, that was also associated with a systematic inflammatory response (as found in this study by the presence of immune system proteins); a range of different traumas were included in the

DeGaudio study, bone fractures being one of them (DeGaudio *et al.*, 1999). Yu *et al.* reported that after a severe burn, kidney function was altered, causing impaired glomerular filtration and also the concentrating ability of the distal renal tubule, causing an imbalance of H_2O/salt and an increase in urinary protein excretion (Yu *et al.*, 1983). Research by Sarti *et al.* concluded that there was a direct correlation between surgical stress score[1] and capillary permeability; they state that the rate of albumin loss can reach 5 % per hour in healthy adults, but can increase to 300 % per hour in adult septic shock (Sarti *et al.*, 2001).

Even though stress can cause an increase in the excretion of protein in urine, the levels observed in the clinical samples still appears to be very high. The majority of the clinical samples with normal levels of urinary protein (the 'outliers' labelled in figures 4.3.4a and 4.3.7a) were obtained at time = 0, therefore possibly being collected before protein passed into the bladder in large amounts. The largest period between collected urine samples was 18 weeks for patients number 16 (sample F36 at t = 18 weeks) and 19 (sample F43 at t = 18 weeks) who both suffered ankle fractures (see appendix B for full details). This should have meant that their fracture was well into the reparative phase and strong enough for load bearing, meaning little stress should have been present. However, the two samples from 18 weeks after the initial fracture had urinary protein concentrations of 169 and 173 mg/mL (F36 and F43 respectively), which is still high, suggesting that stress/trauma or some other physiological factor may also have been present.

Clearly, no firm conclusions as to the reason for the high levels of protein observed can be drawn from the information available for each sample (appendix B), but it is evident that this is an area of research that requires much more consideration due to the lack of consistent literature.

[1] The Oxford surgical stress score is used to relate stress on a scale of 1 to 14 (where 1 = the least amount of stress, and 14 = the most amount of stress).

4.5. Re-analysis of clinical samples using RP-LC-ESI-MS

The demonstration of protein in the majority of clinical urine samples very probably explains the large shifts in both retention time and peak intensity that were evident in both positive and negative mode RP-LC-MS data (sections 4.3.1 and 4.3.2) from the clinical samples, as the MeCN in the mobile phase would have caused the precipitation of proteins onto the column, reducing the separation efficiency of the column. To avoid any subsequent precipitation of any protein onto the RP column, all clinical samples (and pooled samples) were diluted with an equal volume of MeCN to precipitate the protein present, before centrifugation and filtration prior to analysis.

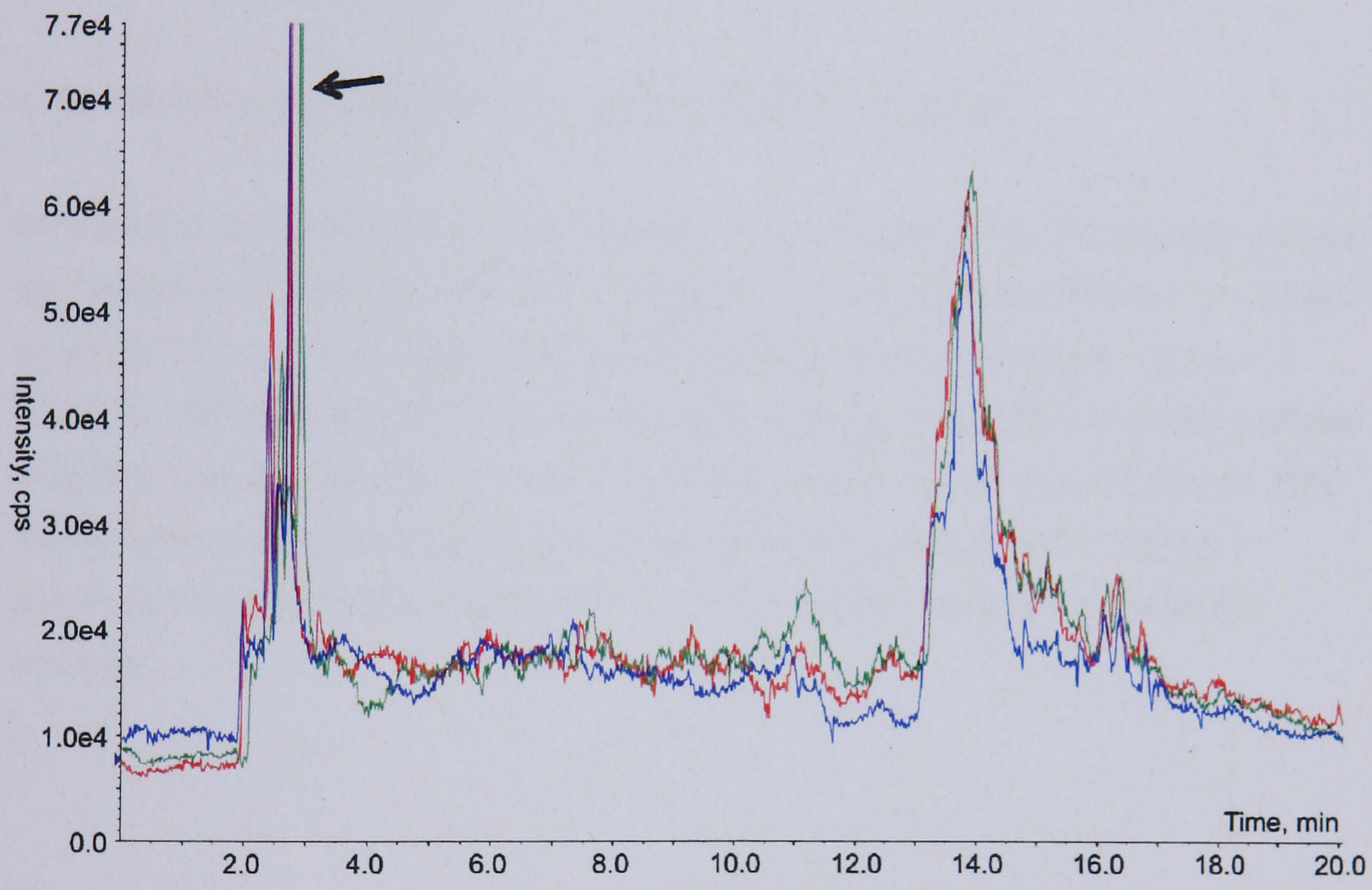


Figure 4.5.1. Three overlaid positive mode RP-LC-MS TICs of pooled urine aliquots from the beginning of data acquisition (blue TIC), mid-way through data acquisition (red TIC) and at the end of three days of data acquisition (green TIC).

Figure 4.5.1 shows three TICs from positive mode RP-LC-MS analysis of aliquots of pooled urine obtained following addition of MeCN, centrifugation and filtration. The blue TIC corresponds to a pooled urine aliquot that was analysed before any clinical samples, the red TIC corresponds to the analysis of a pooled urine aliquot carried out mid-way through data acquisition, and the green TIC corresponds to the analysis of an aliquot of pooled urine run at the end of data acquisition after three days of continual analysis. The three TICs all follow the same trend, although there are some

slight shifts in retention time and peak intensity evident. The most intense peaks (indicated with an arrow) show a retention time of ~2.75 min for the first two pooled samples analysed (blue and red TICs), and ~2.8 min for the final TIC. This gives a deviation of ~3 s, which is well below the retention time tolerance of ±0.5 min used for the metabolomics export script. Although the system appears to be much more stable after the precipitation of protein, the TICs shown in figure 4.5.1 exhibit fewer peaks than were originally observed for the pooled urine sample (without MeCN added) in the initial RP-LC-MS analyses (figure 4.3.1); this may be caused by using MeCN to precipitate protein, which could cause the co-precipitation of urinary metabolites.

4.5.1. Positive and negative mode RP-LC-ESI-MS analysis

All clinical urine samples (diluted with MeCN to precipitate protein) were re-analysed by positive and negative mode RP-LC-ESI-MS, with the random inclusion of aliquots of pooled urine and the analysis of multiple aliquots of some samples, chosen at random. The resulting raw data were extracted using the metabolomics export script (Applied Biosystems) to form a matrix for import into Excel (Microsoft Excel for Mac 2004), where information relating to each sample were added, before being subsequently imported into SIMCA P+ v11.5 (Umetrics, Sweden) for statistical analysis.

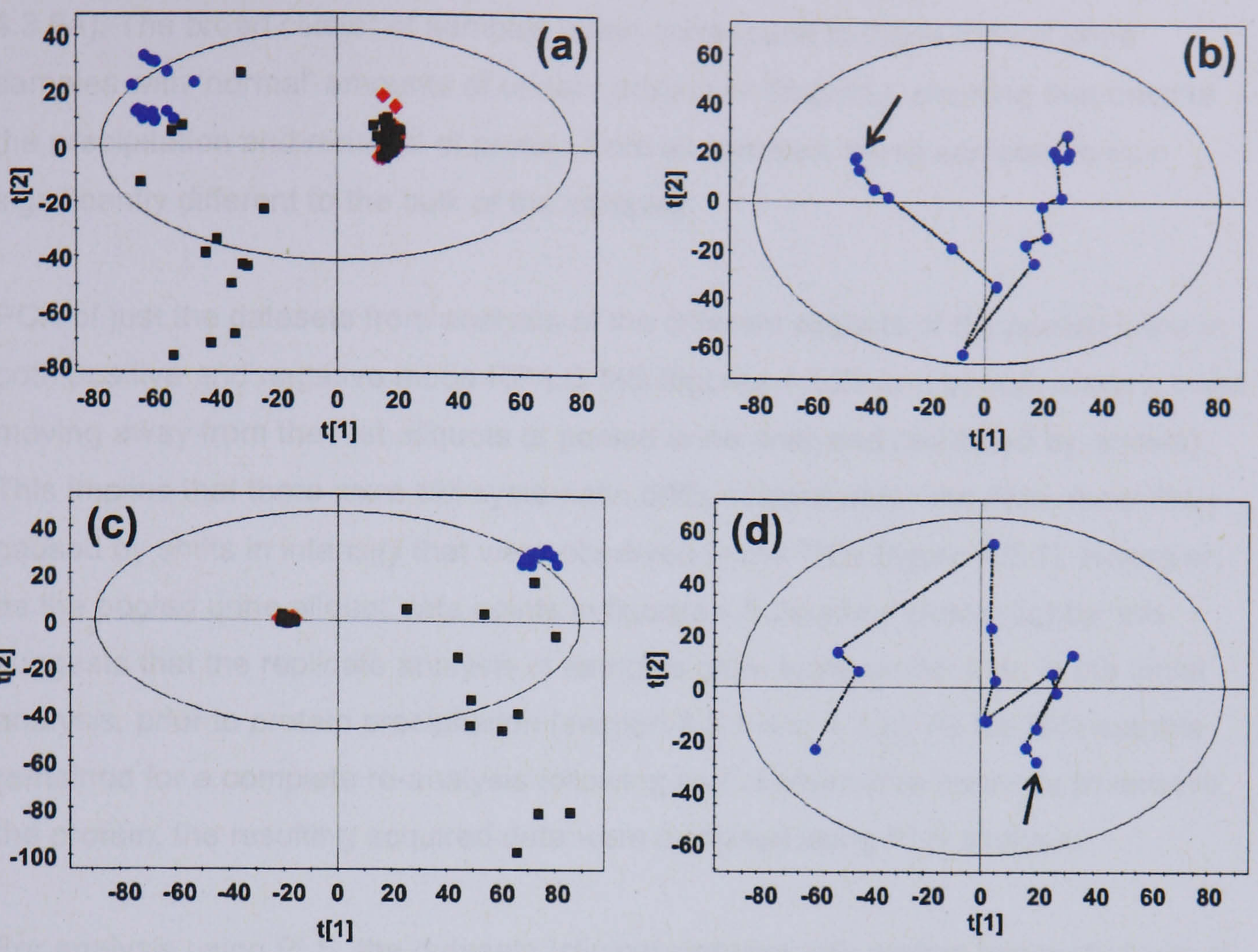


Figure 4.5.2. (a) PCA scores plot of positive mode RP-LC-MS data (■ = samples from males, ◆ = samples from females and ● = aliquots of pooled urine). (b) PCA scores plot of pooled samples only, the arrow indicates the first sample analysed, with the hashed line indicating subsequent sample analysis. (c) PCA scores plot of negative mode RP-LC-MS data (■ = samples from males, ◆ = samples from females and ● = aliquots of pooled urine). (d) PCA scores plot of aliquots of pooled urine only, the arrow indicates the first sample analysed, with the hashed line indicating subsequent sample analysis.

The PCA of both positive and negative mode RP-LC-MS data are shown in figure 4.5.2a and c respectively. Both PCA scores plots show a tight cluster near the centre of each plot that contains the bulk of the data, and another cluster of 12 samples (13 samples for figure 4.5.2a due to the analysis of two aliquots of one of the samples) spread out over a broad range. The pooled urine samples (blue dots) form clusters around the 95 % confidence limit for both positive and negative RP-LC-MS data PCA score plots. Compared to the initial PCA from positive mode RP-LC-MS data prior to protein precipitation (figure 4.3.2), the pooled urine aliquot data points exhibit a much tighter cluster, suggesting less variation between the samples; this further suggests that the precipitation of protein removed some of the original effects that were observed. The pooled urine samples analysed using negative mode RP-LC-MS show a tight cluster in the resulting scores plot (figure 4.5.2c), which, as for positive mode, was even tighter than the initial PCA of negative mode RP-LC-MS data (figure

4.3.6a). The broad cluster of samples again correspond to those clinical urine samples with 'normal' amounts of urinary protein (< 5mg/mL), showing that despite the precipitation and removal of protein from all samples, these samples remain significantly different to the bulk of the samples.

PCA of just the datasets from analysis of the different aliquots of the pooled urine in both positive and negative mode RP-LC-MS (figures 4.5.2b and d) both show a trend moving away from the first aliquots of pooled urine analysed (indicated by arrows). This implies that there were still systematic drifts evident within the data, most likely caused by shifts in intensity that were observed in the TICs (figure 4.5.1). However, as the pooled urine aliquot data points in figures 4.5.2a and c cluster tightly, this suggests that the replicate analysis of samples gave more similar than in the initial analysis, prior to protein precipitation (section 4.3.1 and 4.3.2). As too little sample remained for a complete re-analysis following use of alternative methods to remove the protein, the resulting acquired data were analysed using PLS analysis.

For analysis using PLS, the datasets (clinical samples with protein concentrations of less than 5 mg/mL were removed as these influenced the PLS models) were assigned two sets of different Y-variables. The first Y-variable split the dataset into two response variables: any samples from t=0 (admittance to A&E), and all other post-fracture samples (t=1 to 133 days), and was termed 'frac2'. The second Y-variable assigned three response variables to the dataset: those samples collected at t=0, any samples collected within the first three weeks post-fracture (t=1 to 21 days) and all remaining samples from 22 to 133 days post-fracture, with the Y-variable being termed 'frac3'. As ankle fractures formed the largest number of fractures (21 patients with a total of 51 samples), these were used to create a separate dataset for further analysis. The 'ankle' dataset were split into two response variables: samples from t=0 and all other samples obtained post-fracture (t=1 to 133 days). The resulting PLS analyses of positive and negative mode RP-LC-MS data with three different Y-variable-assigned datasets were optimised according to the scheme presented in chapter 3.5, and are presented in figure 4.5.3.

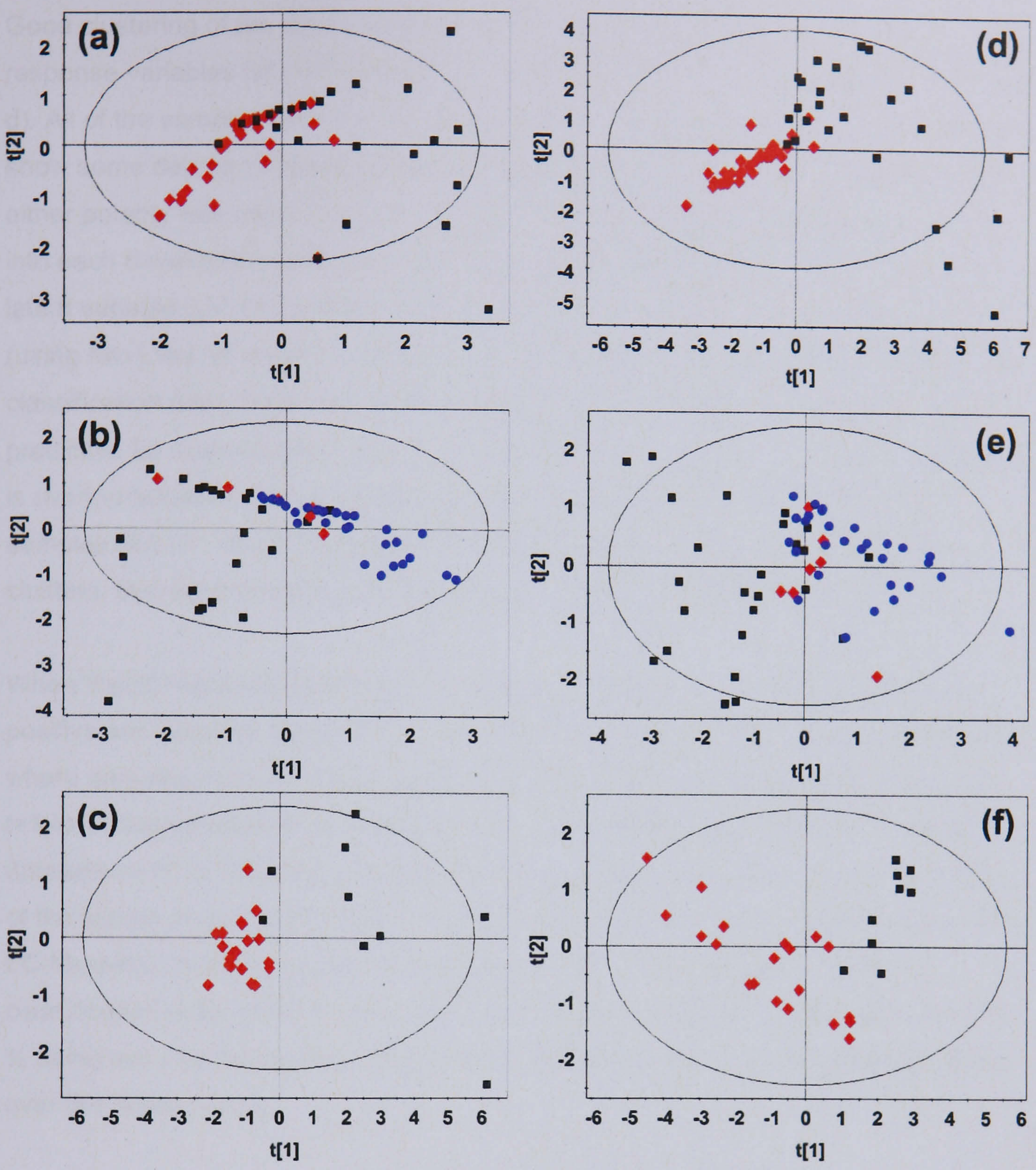


Figure 4.5.3. (a) PLS scores plot of positive mode RP-LC-MS data with the response variables 'frac2": ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (b) PLS scores plot of positive mode RP-LC-MS data with the response variables 'frac3': ■ = samples from time = 0, ◆ = samples from time = 1 to 21 days post-fracture, and ● = samples from time = 22 to 133 days post fracture. (c) PLS scores plot of positive mode RP-LC-MS data with the response variables 'ankle': ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (d) PLS scores plot of negative mode RP-LC-MS data with the response variables 'frac2": ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (e) PLS scores plot of negative mode RP-LC-MS data with the response variables 'frac3': ■ = samples from time = 0, ◆ = samples from time = 1 to 21 days post-fracture, and ● = samples from time = 22 to 133 days post fracture. (f) PLS scores plot of negative mode RP-LC-MS data with the response variables 'ankle': ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture.

Good clustering of the data was observed for separation according to 'frac2' response variables for both positive and negative RP-LC-MS data (figures 4.5.3a and d). All of the samples from t=0 (black squares), and the later samples (red diamonds) show some degree of separation between the two different response variables when either polarity was used to record the RP-LC-MS data. Importing external test sets into each developed model gave external classification results of 25 % (using one latent variable (LV)) for positive mode RP-LC-MS data (figure 4.5.3a), and 35 % (using two LVs) for negative mode RP-LC-MS data (figure 4.5.3d). The low external classification rates imply that the variables used to form each model are not overly predictive for discrimination between the two response variables. However, as there is overlap between the two clusters of data, and some of the external test set samples had LV values that placed them in the region of overlap between the two clusters, this would have somewhat reduced the external classification result.

When 'frac3' response variables were assigned, the resulting PLS scores plots for positive and negative mode RP-LC-MS data show a general trend across the first LV, where samples from t=0 (black squares) cluster on the left-hand side, samples from t=1 to 21 days post-fracture (red diamonds) cluster around '0', and the remaining datasets (t=22 to 133 days post-fracture, blue circles) cluster on the right-hand side of the scores plots (figure 4.5.3b for positive and figure 4.5.3e for negative mode RP-LC-Ms data). Importing external test sets for both models generated external classification rates of 45 % using one LV for positive mode RP-LC-MS data, and 75 % using two LVs for negative mode RP-LC-MS data, improving the predictive ability over the 'frac2' model.

The RP-LC-MS data for ankle fracture samples were assigned response variables for t=0 and t=1 to 133 days post fracture, due to the reduced number of samples available. The resulting PLS scores plots for the two ionisation mode RP-LC-MS datasets are shown in figures 4.5.3c (for positive mode) and 4.5.3f (for negative mode). Both models show clustering according to the assigned response variables, with samples from t=0 (black squares) clustering on the right-hand side, and samples from t=1 to 133 days post-fracture (red diamonds) clustering on the left-hand side. The external classification rates for the ankle fracture models were 90 % for positive mode RP-LC-MS data and 60 % for negative mode RP-LC-MS data, with both external classification rates being highest when one LV was utilised for prediction.

Table 4.5.1. Comparison of the top five variables for each developed model. Shaded cells show variables that are present for more than one developed model (coloured according to polarity).

Separation Method	Polarity	Statistical Model	Rank (VIP)	*m/z*	t_R	*Y Variable Discriminative for*
RP	Positive	frac2	1	**86.05**	12.89	⇑ for t = 0
			2	**107.95**	1.33	⇑ for t > 0
			3	**105.02**	14.32	⇑ for t = 0
			4	**86.05**	11.87	⇑ for t > 0
			5	**367.14**	2.95	⇑ for t > 0
		Ankle	1	**100.02**	1.98	⇑ for t = 0
			2	**90.52**	9.12	⇑ for t = 0
			3	**171.13**	17.75	⇑ for t = 0
			4	**205.12**	16.02	⇑ for t = 0
			5	**392.96**	13.15	⇑ for t = 0
		frac3	1	**198.05**	28.16	⇑ for t = 0
			2	**500.26**	21.54	⇑ for t > 21
			3	**105.02**	14.32	⇑ for t > 21
			4	**299.11**	2.90	⇑ for t > 21
			5	**120.07**	11.04	⇑ for t > 21
	Negative	frac2	1	**448.34**	20.17	⇑ for t = 0
			2	**473.02**	1.57	⇑ for t > 0
			3	**528.30**	21.69	⇑ for t = 0
			4	**326.11**	10.19	⇑ for t > 0
			5	**446.09**	10.17	⇑ for t > 0
		Ankle	1	**169.00**	28.87	⇑ for t > 0
			2	**448.34**	20.17	⇑ for t = 0
			3	**464.34**	18.64	⇑ for t = 0
			4	**465.29**	18.59	⇑ for t = 0
			5	**516.35**	20.17	⇑ for t = 0
		frac3	1	**448.34**	20.17	⇑ for t = 0
			2	**476.32**	21.46	⇑ for t >22
			3	**477.32**	21.47	⇑ for t > 22
			4	**326.11**	10.19	⇑ for t = 0 (>0 = 0)
			5	**276.02**	29.04	⇑ for t = 0

Table 4.5.1 presents the top five variables that gave the highest VIP values for each of the three developed PLS models using the two ionisation polarities, along with the Y variable they were discriminative for. For each of the statistical models, there are few variables present which are consistent across each model (shaded cells). The CID tandem MS analyses of the most important variables generated by each PLS model are presented in section 4.8.

4.6. Analysis of clinical samples by positive and negative mode HILIC-ESI-MS

All of the clinical samples were diluted by addition of an equal volume of MeCN, centrifuged and filtrated prior to analysis, then analysed by positive and negative mode HILIC-LC-ESI-MS, with the random injection of aliquots of a pooled urine sample and the random re-injection of aliquots of clinical urine samples. The resulting raw data were extracted using the metabolomics export script (Applied Biosystems) to form a matrix for import into Excel (Microsoft Excel for Mac 2004), where data relating to each sample were added, before being subsequently imported into SIMCA P+ v11.5 (Umetrics, Sweden) for statistical analysis.

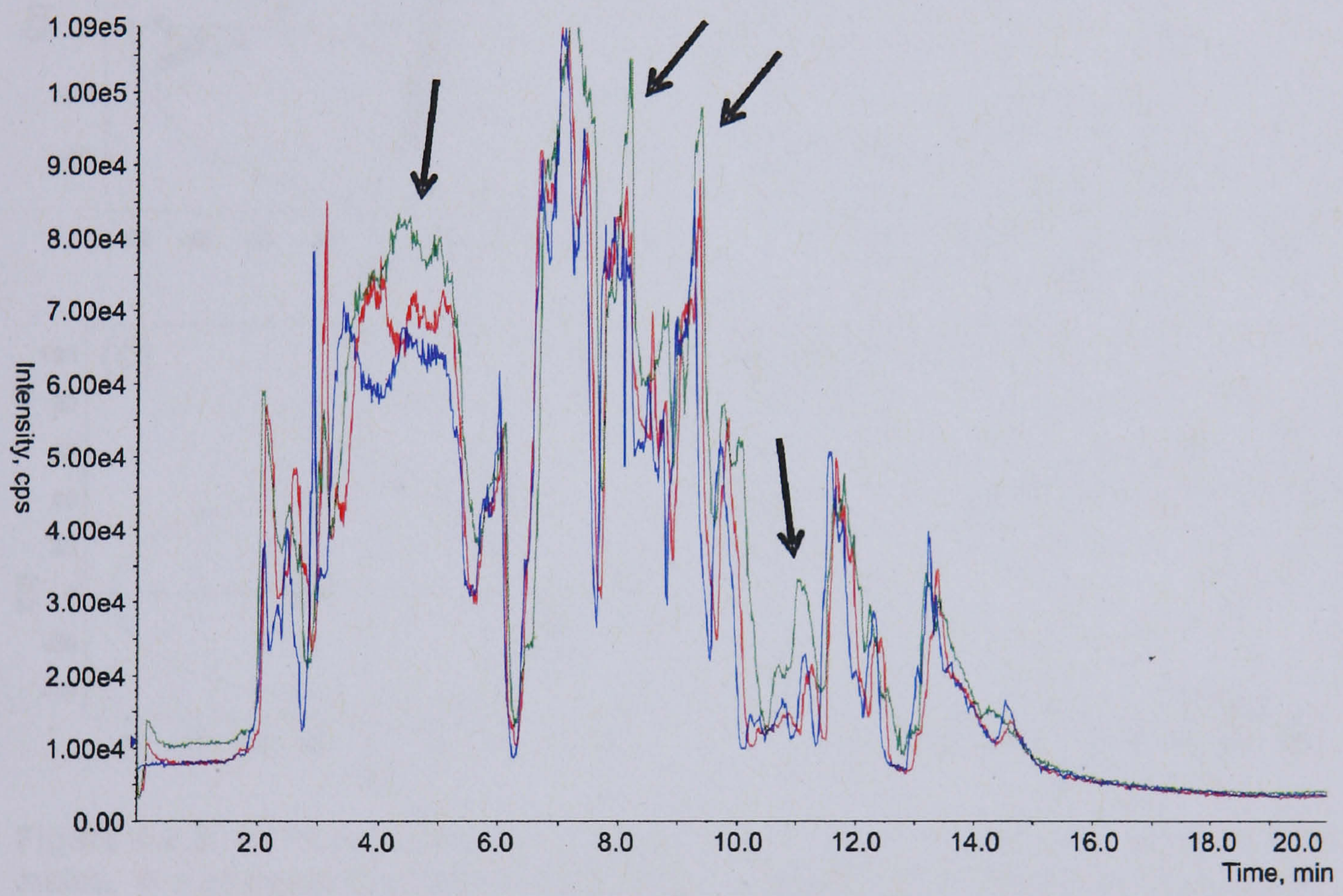


Figure 4.6.1. Three overlaid positive mode HILIC-MS TICs of aliquots of pooled urine samples from the beginning of data acquisition (green TIC), mid-way through data acquisition (red TIC) and at the end of three days of data acquisition (blue TIC). Arrows indicate areas where there are changes in intensity between the samples.

Figure 4.6.1 shows three superimposed TIC traces from positive mode HILIC-MS injections of aliquots of a pooled urine sample. The green TIC corresponds to an injection of an aliquot of pooled urine prior to the analysis of any clinical samples, the red TIC corresponds to the injection of an aliquot of pooled urine mid-way through data acquisition, and the blue TIC corresponds to the injection of an aliquot of pooled urine at the end of three days of data acquisition. The three TICs follow the same

trend, but there are some minor deviations in retention time, along with some shifts in the intensity of the peaks indicated by arrows. Any deviations in retention time are less than ±6 s, which is within the tolerance of ±0.5 min used in the metabolomics export script; normalising to total ion count should minimise the observed differences in intensity. Comparing the TICs of pooled samples from positive mode HILIC-MS data (figure 4.6.1) to that obtained using positive mode RP-LC-MS (figure 4.5.1), shows that HILIC separation provided a more 'information rich' chromatogram.

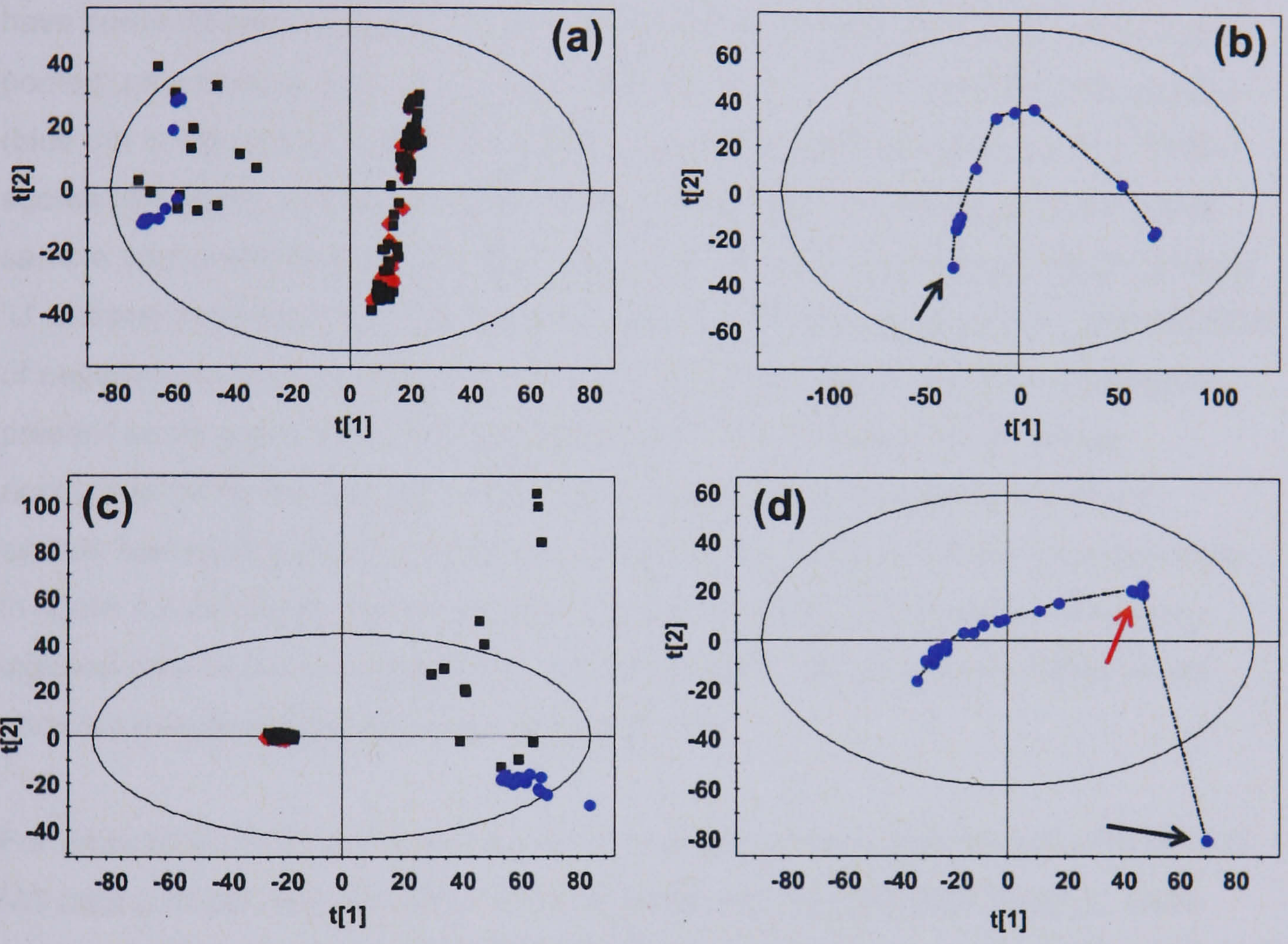


Figure 4.6.2. (a) PCA scores plot of positive mode HILIC-MS data (■ = samples from males, ◆ = samples from females and ● = aliquots of pooled urine). (b) PCA scores plot of aliquots of pooled urine only, the arrow indicates the first sample analysed, with the hashed line indicating subsequent sample analysis. (c) PCA scores plot of negative mode HILIC-MS data (■ = samples from males, ◆ = samples from females and ● = aliquots of pooled urine). (d) PCA scores plot of aliquots of pooled urine only, the black arrow indicates the first sample analysed, the red arrow indicates subsequent samples analysed prior to the analysis of clinical samples, with the hashed line indicating subsequent sample analysis.

The PCA of both positive and negative mode HILIC-MS data are presented in figure 4.6.2a and c respectively. Both PCA scores plots show a cluster containing the bulk of the data, with positive mode HILIC-MS data generating a larger 'linear' cluster (figure 4.6.2a) compared to a small tight cluster for negative mode HILIC-MS data

(figure 4.6.2c). The remaining clinical samples for both ionisation modes' HILIC-MS data form a loose cluster containing 12 data points (black squares) that, as for RP-LC-MS data, correspond to each of the clinical samples with low concentrations (less than 5 mg/mL) of urinary protein.

The datasets for clusters of pooled samples (blue dots) for each PCA scores plot show some differences. Positive HILIC-MS datasets for aliquots of pooled urine show two distinct clusters (figure 4.6.2a), meaning that there are two types of samples that have some different characteristics. The negative HILIC-MS datasets for aliquots of pooled urine show a tight cluster with only one data point deviating from the cluster (blue dot in the bottom right hand corner, outside the 95 % confidence limit, of the scores plot shown in figure 4.6.2c). When PCA of just the aliquots of pooled urine sample was undertaken, those from positive mode HILIC-MS (figure 4.6.2b) show a 'U' shaped trend from the first sample analysed (indicated by an arrow). For the PCA of negative mode HILIC-MS aliquots of pooled urine sample, the only observation present as an outlier from the tight cluster observed in figure 4.6.2c can be accounted for by the fact that it was the first aliquot injected before the LC-MS system had equilibrated (indicated by a black arrow in figure 4.6.2d). The red arrow in figure 4.6.2d shows the subsequent aliquots of pooled urine sample that were injected prior to the analysis of any clinical samples, with subsequent data points showing less deviation then originally observed.

For analysis by PLS, the same response variables were assigned as for the RP-LC-MS data (clinical samples with protein concentrations of less than 5mg/mL were removed as these influenced the PLS models). The first Y-variable split the dataset into two response variables: any samples from t=0 (admittance to A&E), and all other post-fracture samples (t=1 to 133 days), and was termed 'frac2'. The second Y-variable assigned three response variables to the dataset: those samples collected at t=0, any samples collected within the first three weeks post-fracture (t=1 to 21 days) and samples collected 22 to 133 days post-fracture, with the Y-variable being termed 'frac3'. As ankle fractures formed the largest number of fractures (21 patients with a total of 51 samples), these were used to create a separate dataset for further analysis. The 'ankle' dataset was split into two response variables: samples from t=0 and all other samples obtained post-fracture (t=1 to 133 days). The resulting PLS analyses of positive and negative mode HILIC-LC-MS data with three different Y-variable assigned datasets were optimised according to the scheme presented in chapter 3.5, and are presented in figure 4.6.3.

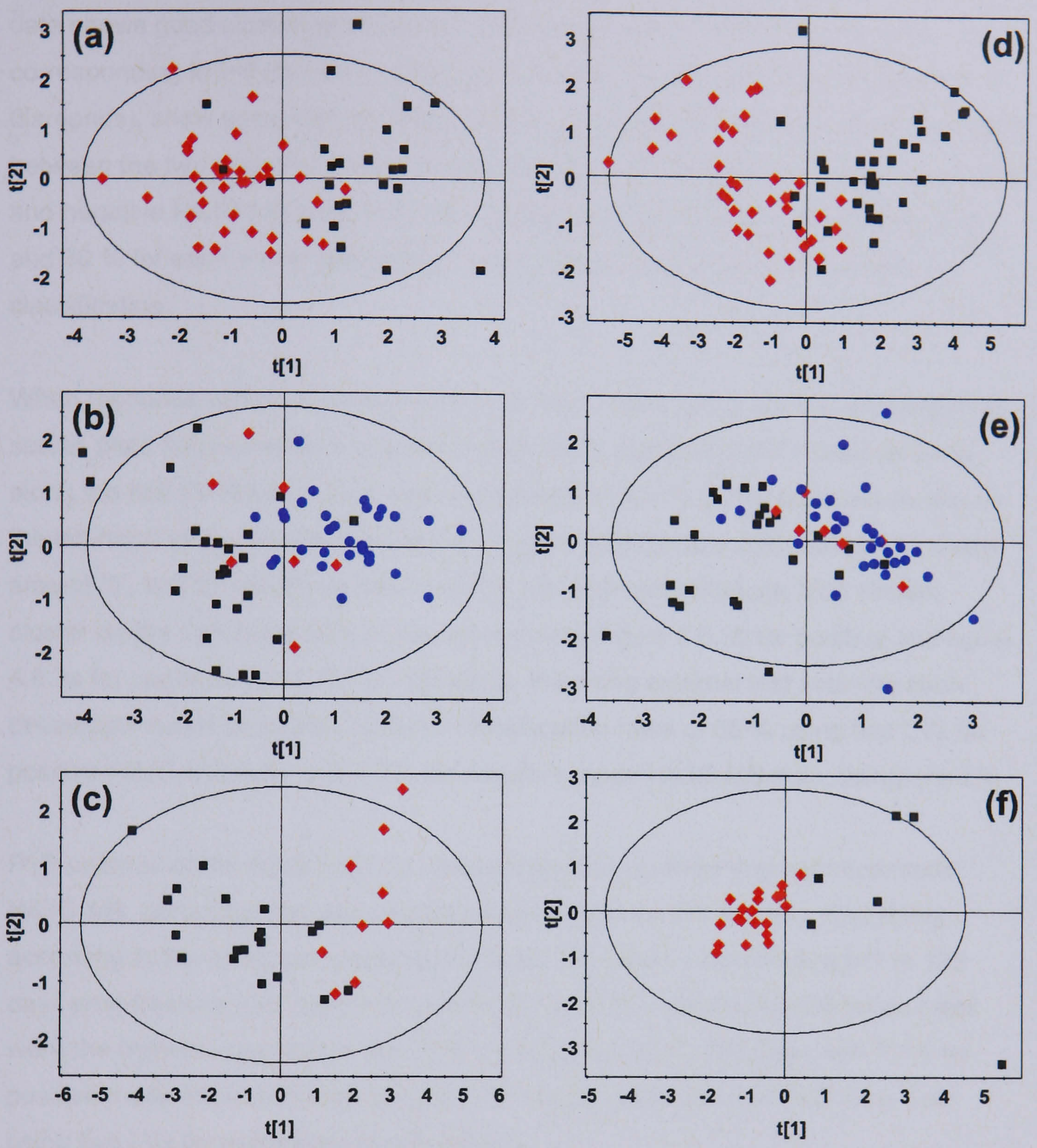


Figure 4.6.3. (a) PLS scores plot of positive mode HILIC-MS data with the response variables 'frac2": ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (b) PLS scores plot of positive mode HILIC-MS data with the response variables 'frac3': ■ = samples from time = 0, ◆ = samples from time = 1 to 21 days post-fracture, and ● = samples from time = 22 to 133 days post fracture. (c) PLS scores plot of positive mode HILIC-MS data with the response variables 'ankle': ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (d) PLS scores plot of negative mode HILIC-MS data with the response variables 'frac2": ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (e) PLS scores plot of negative mode HILIC-MS data with the response variables 'frac3': ■ = samples from time = 0, ◆ = samples from time = 1 to 21 days post-fracture, and ● = samples from time = 22 to 133 days post fracture. (f) PLS scores plot of negative mode HILIC-MS data with the response variables 'ankle': ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture.

Separation according to 'frac2' response variable for positive and negative HILIC-MS data shows good clustering of the two groups (figures 4.6.3a and d). Samples corresponding to t=0 (black squares), and samples from any time post-fracture (red diamonds), show some degree of separation, although there is some overlap present between the two clusters. External classification of the developed models for positive and negative HILIC-MS data using an independent test set generated results of 60 and 30 % for each model respectively, both using two LVs to gain maximum classification.

When response variables corresponding to 'frac3' were assigned, the resulting PLS scores plots for positive and negative mode HILIC-MS data show a general trend along the first LV (figure 4.6.3b and e). Samples from t=0 (black squares) cluster on the left-hand side, samples from t=1 to 21 days post-fracture (red diamonds) cluster around '0', and all remaining data (t=22 to 133 days post-fracture, blue circles) cluster on the right-hand side of the scores plots (figure 4.6.3b for positive and figure 4.6.3e for negative mode RP-LC-Ms data). Importing external test sets into each developed model generated external classification rates of 65 % using two LVs for positive HILIC-MS data, and 50 % for negative mode HILIC-MS data, using one LV.

PLS analysis of the ankle fracture sample cohort by positive and negative mode HILIC-MS, generated the scores plots shown in figures 4.6.3c and f. Clustering according to the assigned response variables t=0 (black squares) and t=1 to 133 days post-fracture (red diamonds) was observed. The external classification rates were the highest obtained for any PLS model using HILIC-MS data, with 70 % for positive mode HILIC-MS data and 80 % for negative mode HILIC-MS data, both using two LVs for maximum classification.

Table 4.6.1. Comparison of the top five variables for each developed model. Shaded cells show variables that are present for more than one developed model (coloured according to polarity).

Separation Method	Polarity	Statistical Model	Rank (VIP)	*m/z*	t_R	*Y Variable Discriminative for*
HILIC	Positive	frac2	1	**198.05**	3.45	⇑ for t > 0
			2	**254.14**	2.35	⇑ for t > 0
			3	**168.06**	2.95	⇑ for t > 0
			4	**392.03**	8.92	⇑ for t > 0
			5	**198.05**	5.33	⇑ for t > 0
		Ankle	1	**142.09**	9.38	⇑ for t > 0
			2	**350.18**	6.73	⇑ for t > 0
			3	**450.30**	2.90	⇑ for t = 0
			4	**451.31**	2.90	⇑ for t = 0
			5	**86.091**	19.55	⇑ for t > 0
		frac3	1	**198.05**	3.45	⇑ for t > 0
			2	**254.14**	2.35	⇑ for t > 0
			3	**152.06**	7.57	⇑ for t = 0
			4	**239.20**	7.30	⇑ for t < 21
			5	**796.55**	6.55	⇑ for t = 0
	Negative	frac2	1	**448.33**	2.90	⇑ for t > 0
			2	**562.33**	2.90	⇑ for t > 0
			3	**276.02**	5.53	⇑ for t = 0
			4	**498.32**	5.88	⇑ for t > 0
			5	**243.06**	7.38	⇑ for t = 0
		Ankle	1	**448.33**	2.90	⇑ for t > 0
			2	**464.31**	5.52	⇑ for t > 0
			3	**464.31**	3.00	⇑ for t > 0
			4	**186.12**	2.63	⇑ for t > 0
			5	**562.33**	2.90	⇑ for t > 0
		frac3	1	**151.06**	6.93	t = 0 > 1 > 2
			2	**448.33**	2.90	t = 0 > 1 > 2
			3	**562.33**	2.90	⇑ for t > 0
			4	**498.32**	5.88	⇑ for t > 0
			5	**243.06**	7.38	⇑ for t = 0

Table 4.6.1 presents the top five variables that gave the highest VIP values that were generated by each of the three PLS models for each ionisation polarity, and also each Y variable each variable is discriminatory for. For each of the three PLS models developed using data obtained from positive ionisation, there are two variables that were consistent between each of the 'frac2' and 'frac3; models (shaded cells). For the models developed using negative ionisation mode HILIC-MS data, there are more variables that are consistent throughout the models (blue shaded cells). One variable at *m/z* 448.33 with a retention time of 2.90 min was present in all three

models, and also had a mass that could correspond to a variable from the negative mode RP-LC-MS data at *m/z* 448.35 (44.6 ppm difference) with a retention time of 26.17 min. The shifts in retention time from 2.90 min (HILIC) to 26.17 min (RP) would appear to support the initial theory that this mass corresponds to the same compound, as a strongly retained compound on an RP column would be expected to elute very early from an HILIC column (2.90 min is close to the void). The variables presented in table 4.6.1 were analysed by CID tandem MS, and the results presented in section 4.8.

4.7. Analysis of ±RP and ±HILIC data by data fusion

Each of the four datasets (±HILIC- and ±RP-LC-MS) were imported into Excel (Microsoft Excel for Mac 2004) and concatenated one below the other, as outlined in chapter 3.5.3.

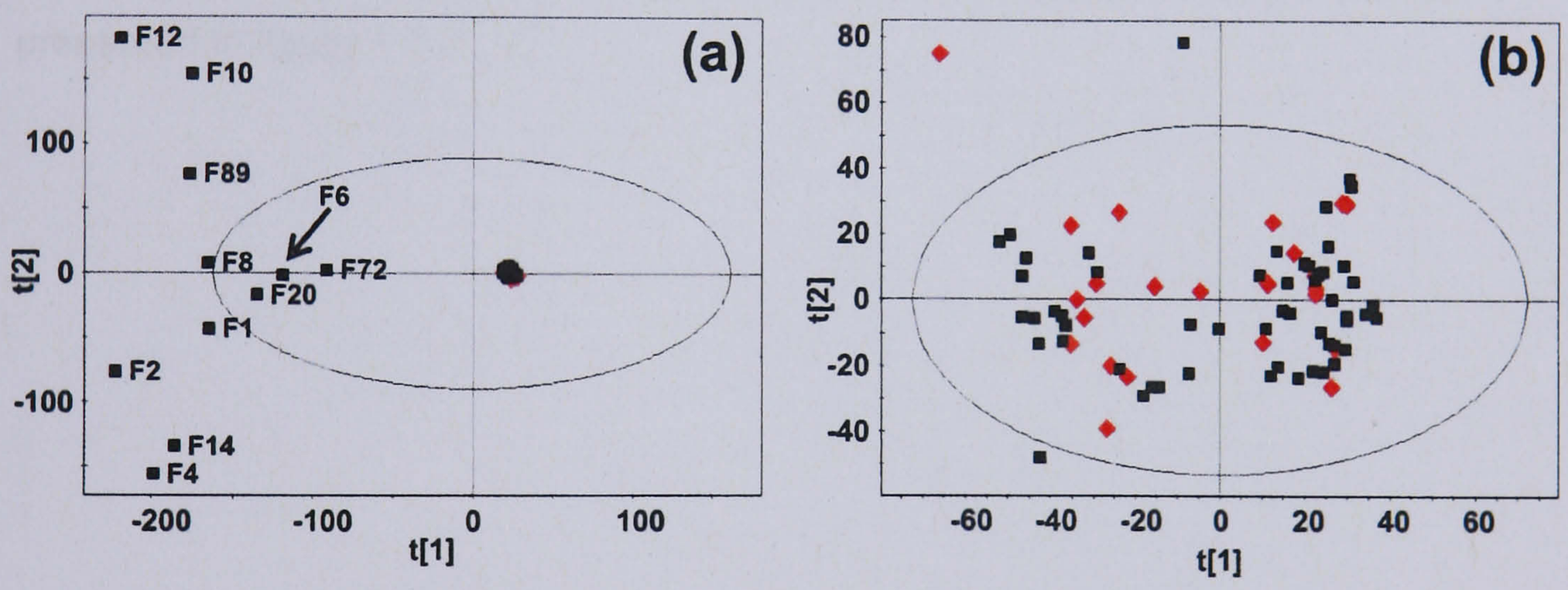


Figure 4.7.1. (a) PCA scores plot of concatenated data (■ = samples from males, ◆ = samples from females). (b) PCA scores plot concatenated data with all samples having normal levels of protein (less than 5 mg/mL, corresponding to the labelled samples in (a)) being removed.

Figure 4.7.1a shows the PCA of the concatenated dataset. The bulk of the data cluster in a tight group near the centre of the scores plot, whilst the remaining samples form a loose cluster in the left hand side of the scores plot. All of the labelled data points correspond to clinical samples with normal urinary protein concentrations (less than 5 mg/mL), as was observed for the PCA of each individual dataset (±RP, figures 4.5.2a and c, and ±HILIC, figures 4.6.2a and c). When the samples with normal protein concentrations were excluded from the dataset, the resulting PCA scores plot (figure 4.7.1b) shows no clear separation based upon gender (or any

other information available). There are only three data points outside the 95 % confidence limit, and none of these exhibited a large DModX value, and so remained to form the dataset for subsequent PLS analyses.

For analysis by PLS, the following response variables were assigned: The first Y-variable split the dataset into two response variables: samples from t=0 (admittance to A&E), and all post-fracture samples (t=1 to 133 days), and was termed 'frac2'. The second Y-variable assigned three response variables to the dataset: those samples collected at t=0, any samples collected within the first three weeks post-fracture (t=1 to 21 days) and all samples collected 22 to 133 days post-fracture, with the Y-variable being termed 'frac3'. As ankle fractures formed the largest number of fractures (21 patients with a total of 51 samples), these were used to create a separate dataset for further analysis. The 'ankle' dataset was split into two response variables: samples from t=0 and all post-fracture samples (t=1 to 133 days). The resulting PLS analyses of the concatenated dataset with three different Y-variables assigned were optimised according to the scheme presented in chapter 3.5, and are presented in figure 4.7.2:

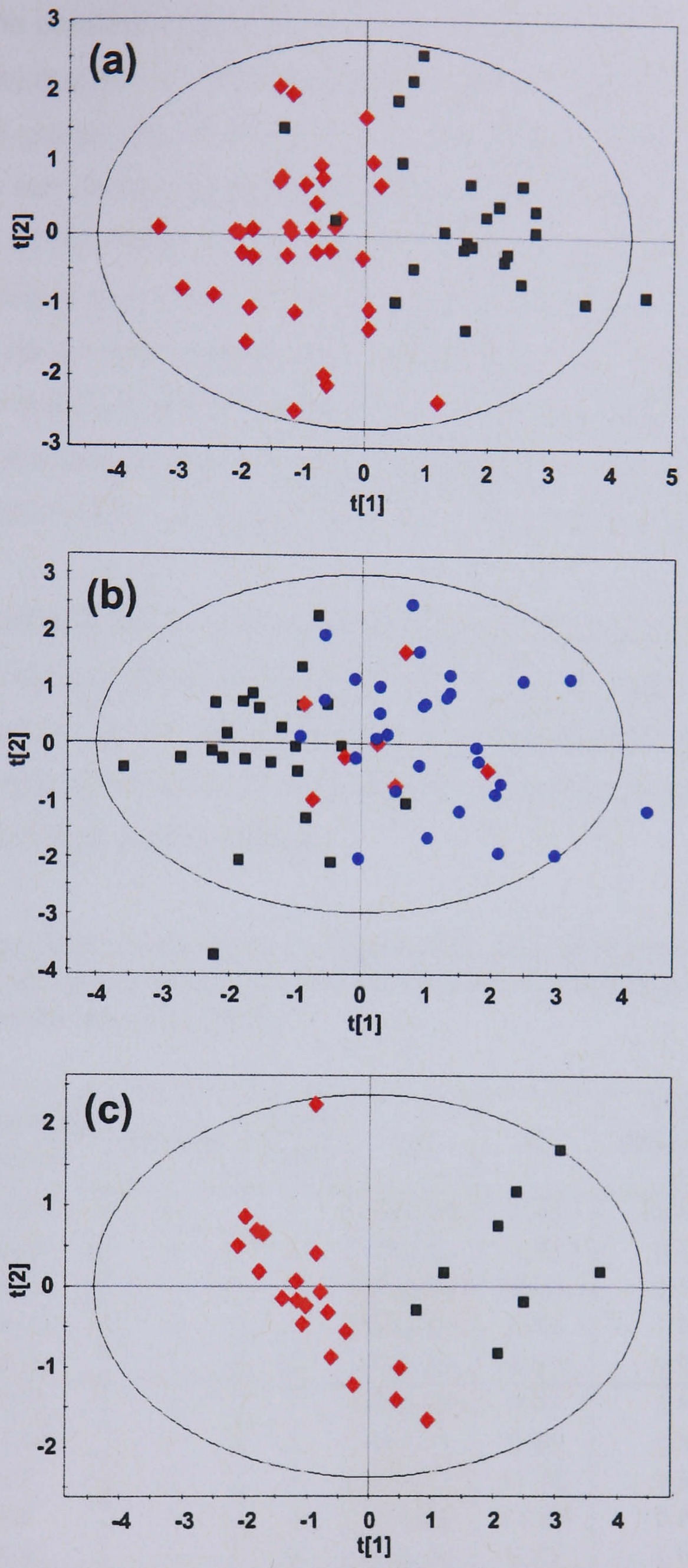


Figure 4.7.2. (a) PLS scores plot of concatenated data with the response variables 'frac2": ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture. (b) PLS scores plot of concatenated data with the response variables 'frac3': ■ = samples from time = 0, ◆ = samples from time = 1 to 21 days post-fracture, and ● = samples from time = 22 to 133 days post fracture. (c) PLS scores plot of concatenated data with the response variables 'ankle': ■ = samples from time = 0, ◆ = samples from time = 1 to 133 days post-fracture.

PLS analysis of the concatenated data according to 'frac2' response variables where t=0 (black squares) and t=1 to 133 days post-fracture (red diamonds) shows distinct clusters of the two groups (figure 4.7.2a). The external classification rate of 50 % using one LV was comparable to the best performing single dataset (positive mode HILIC-MS data at 60 %). When 'frac3' response variables were assigned, the resulting PLS scores plot (figure 4.7.2b) shows a trend across the first LV, where samples from t=0 (black squares) cluster on the left-hand side, and those from t=1 to 21 days (red diamonds) and 22-133 days post-fracture (blue dots) cluster on the right-hand side of the scores plot. The external classification rate of 50 % was equal to that for the 'frac2' model, but required two LVs for the highest classification rate.

The final PLS model was for the ankle dataset (figure 4.7.2c), where two clusters according to the assigned response variables were observed. Data points from t=0 (black squares) and t=1 to 133 days post-fracture (red diamonds) cluster apart with no overlap. The external classification rate of 80 % using one LV was the highest obtained using the concatenated dataset.

Table 4.7.1. Comparison of the top five variables for each of the three developed models. Shaded cells show variables that are present for more than one developed model (coloured according to polarity).

Statistical Model	Separation Method	Polarity	Rank (VIP)	*m/z*	t_R	*Y Variable Discriminative for*
frac2	HILIC	+	1	**198.05**	3.45	⇑ for t > 0
	HILIC	+	2	**392.03**	8.92	⇑ for t > 0
	RP	-	3	**448.34**	20.17	⇑ for t = 0
	HILIC	-	4	**448.33**	2.90	⇑ for t > 0
	HILIC	-	5	**230.02**	6.90	⇑ for t > 0
Ankle	HILIC	+	1	**450.30**	2.90	⇑ for t = 0
	HILIC	+	2	**451.31**	2.90	⇑ for t = 0
	RP	+	3	**171.13**	17.75	⇑ for t > 0
	RP	+	4	**205.11**	17.29	⇑ for t > 0
	HILIC	+	5	**350.18**	6.73	⇑ for t > 0
frac3	HILIC	+	1	**198.05**	3.45	⇑ for t > 0
	HILIC	+	2	**254.14**	2.35	⇑ for t > 0
	RP	-	3	**448.34**	20.17	⇑ for t = 0
	HILIC	-	4	**448.33**	2.90	t = 0 < 1 < 2
	HILIC	-	5	**243.06**	7.38	⇑ for t < 21

Table 4.7.1 presents the five most important variables from each of the three different statistical models. Each of the three models uses variables from each of the

separation methods and both ionisation polarities, along with the Y variable that each variable was discriminative for. The shaded cells represent variables that were found in all PLS models; each of the variables was observed in the individual PLS models for the separation method and polarity shown (tables 4.5.1 and 4.6.1). The CID tandem MS analysis of the variables presented in table 4.7.1 are discussed in section 4.8.

4.8. Variable analysis using CID tandem MS

Identifying the ions giving rise to the variables generated by each of the three developed PLS models (frac2, frac3 and ankle) for the two ionisation modes using both RP- and HILIC-MS in this chapter was carried out using CID tandem MS. A selection of four urine samples was chosen for analysis, based upon time of collection during fracture healing, and also type of fracture. As for the CID tandem MS analyses of data in chapter 3.6.4, the IDA setting was used in the Analyst QS software to analyse any variables added to the 'include' list.

When each of the variables detected was searched for in the HMDB[1] (Wishart *et al.*, 2007) and Metlin[2] databases, only one variable corresponded to a plausible metabolite; *m/z* 120 could correspond to the amino acid Thr. All other variables did not return any hits from either database search. The lack of more accurate mass measurements, along with only one potential metabolite structure from database searches, meant that both precursor and fragment ions shown in table 4.8.1 cannot be assigned atomic compositions. The only ion that was found in the literature was that at *m/z* 465, which in research by Lutz *et al.* (Lutz *et al.*, 2006), was postulated to correspond to a glucuronidated steroid.

---

[1] http://www.hmdb.ca (accessed November 2007)
[2] http://metlin.scripps.edu (accessed November 2007)

Table 4.8.1. A table showing precursor ions that were isolated and subjected to CID tandem MS, and *m/z* values of ions that were identified as ariables and corresponded to metabolites in either HMDB or Metlin. (*n/d* = not detected in CID analysis. *n/r* = not recorded, meaning no fragment ions were observed).

Separation Method	Statistical Model	Ionisation Polarity	tR	Presursor Ion Mass (*m/z*)	Product Ions (*m/z*)						Product Ion Spectrum in Appendix D
RP	frac2	Positive	*n/d*	*n/d*	*n/d*						-
		Negative	20.17	448.4	*n/r*						-
			21.69	528.3	478.4	155.0	113.0				1
	frac3	Positive	21.54	500.3	457.3	439.2	361.2	216.1	198.0	164.0	2
			2.90	299.1	253.1	164.9	90.1	72.1	45.0		3
			11.04	120.1	103.0	93.1	88.0				4
		Negative	21.46	476.3	279.3	253.3					5
	ankle	Positive	*n/d*	*n/d*	*n/d*						-
		Negative	20.17	516.4	448.4	427.0	385.0	249.0	205.0	155.0	6
			18.59	465.3	113.0						7
HILIC	frac2	Positive	*n/d*	*n/d*	*n/d*						-
		Negative	2.90	448.3	*n/r*						-
			2.90	562.3	448.4	113.0					8
	frac3	Positive	6.55	796.6	184.1						9
		Negative	2.90	448.3	*n/r*						-
			2.90	562.3	448.4	113.0					8
	ankle	Positive	2.90	450.3	433.2	415.3					10
		Negative	2.90	448.3	*n/r*						-
			2.90	562.3	448.4	113.0					8

Appendix D contains the CID tandem mass spectrums of precursor ions that produced fragment ions upon CID.

The precursor ion at *m/z* 528 shows a fragment ion at *m/z* 113 (appendix D1), suggesting that this compound is a glucuronide (Levsen *et al.*, 2005). Subtracting 176 Da (the mass increment corresponding to glucuronic acid) from the metabolites mass gives an RMM of 353 Da for this compound. Searching either database or the literature for this mass did not yield any results, making it impossible to propose an identity for this compound.

Appendix D2 shows the CID tandem MS of *m/z* 500; the precursor ion shows a peak at 2 Th higher, suggesting the presence of chlorine in the structure of this compound. There are two losses from fragment ions that could correspond to loss of water; the fragment ions at *m/z* 457 and 216 both have peaks 18 Th less. The fragment ion at *m/z* 90 from the precursor ion at *m/z* 299 also loses 18 Th, suggesting a loss of water from this ion too (appendix D3). However, no further information can be extracted from the data.

The fragment ion 17 Th lower than the precursor ion at *m/z* 120 could correspond to the loss of NH_3 from the compound (appendix D4). High energy CID tandem MS of Thr fragments to lose water from the precursor ion (giving *m/z* 102), then a subsequent loss of CO and CH_2O_2 from the fragment ion at *m/z* 102 to give *m/z* 74 and 56 respectively (Heerma and Kulik, 1988). However, at low energy CID tandem

MS, the loss of NH_3 from peptides/amino acids is more common. No plausible loss could be postulated for the fragment ion 27 Th lower than the precursor ion, although the loss of 32 Da (fragment ion at *m/z* 88) could correspond to the elements of MeOH.

The CID tandem MS of *m/z* 516 (appendix D6) shows six fragment ions. The fragment ion at *m/z* 448 could correspond to the precursor ion at *m/z* 448 that has the same retention time in negative mode RP-LC-MS data in the frac2 PLS model, suggesting that the ion at *m/z* 448 in the frac2 PLS model may be an in-source fragment of *m/z* 516. It is not possible to assign atomic compositions to the fragment ions, making it very difficult for a structure to be postulated.

The presence of *m/z* 113 in the CID tandem MS spectrum of the precursor ion at *m/z* 465 suggests that this compound may be a glucuronide (appendix D7). Lutz *et al.* used MS/MS transitions to the molecular anion of *m/z* 113, which is a characteristic fragment of glucuronic acid, to detect putative steroid glucuronides (Lutz *et al.*, 2006). A mass at *m/z* 465 was postulated to correspond to one of the following steroid glucuronides: androsterone, dihydrotestosterone, 3ß,17ß-dihydroxy-5-androsterone or epiandrosterone (Lutz *et al.*, 2006).

The negative ion mode CID tandem MS analysis of the precursor ion at *m/z* 562 (appendix D8) also produce an intense fragment ion at *m/z* 113, suggesting that this compound is also a glucuronide. The very low intensity fragment ion at *m/z* 448 corresponds to a loss of 114 Th, corresponding to the neutral loss of a glucuronide fragment. The precursor ion eluted at 2.9 min, which is the same retention time as the precursor ion at *m/z* 448, suggesting *m/z* 448 may have been an in-source fragment of a compound with an RMM of 563 Da. Interestingly, the positive mode CID tandem MS analysis of *m/z* 450 (appendix D10) eluted at 2.9 min, the same as *m/z* 448 (postulated in-source fragment) and 562; this could mean that *m/z* 450 is the positive ion in-source fragment (loss of 114 Th, a neutral glucuronide fragment) from the compound with an RMM of 563 Da. The fragment ions at *m/z* 433 and 415 could correspond to further losses of NH_3 and water from the positive mode in-source fragment.

The final product ion at *m/z* 796 (appendix D9) shows one fragment ion at *m/z* 184, due to the loss of 612 Th, meaning that the product ion spectrum was very uninformative for this compound.

4.8.1. Discussion

The CID tandem MS analysis of the most important variables from the developed PLS models produced some tandem MS spectra (of variable quality) containing fragment ions. However, the databases searched did not identify any possible compounds, and the lack of more accurate mass measurements means that assigning structures was not possible. The only tentative assignment is based upon work by Lutz *et al.* who postulated that the transition *m/z* 465 to 113 (as seen in the spectrum in appendix D7) corresponded to one of the following steroid glucuronides: androsterone, dihydrotestosterone, 3ß,17ß-dihydroxy-5-androsterone or epiandrosterone (Lutz *et al.*, 2006).

It is interesting that identical retention times for many of the precursor ions (most important variables) in both the RP- and HILIC-MS dataset PLS models were observed. The fragment ion mass of *m/z* 448 (from the precursor ion *m/z* 562), corresponding to the loss of 114 Th (neutral glucuronide fragment) suggests that *m/z* 448 (as a precursor ion) was formed as an in-source fragment of *m/z* 562 (also highlighted by the comparable importance of *m/z* 562 and 448 (in-source fragment) from the PLS models).

4.9. Discussion

The original aim of the work presented in this chapter was to profile the body's response to long-bone fractures, specifically to attempt to find biomarkers related to failed or delayed fracture healing. Unfortunately, at an early stage in sampling, a lack of long-bone fractures led to the largest sample cohort being from patients with ankle fractures.

The initial RP-LC-MS analysis of the clinical urine samples generated poor results, the reason for which only became apparent when the clinical samples were prepared for analysis by HILIC-MS. Upon addition of MeCN to the clinical urine samples, a large amount of precipitate formed, meaning that the RP column used would have been saturated with precipitate with the MeCN present in the mobile phase. The Bradford assay of all clinical samples confirmed that the precipitate was protein, with concentrations from 1-258 mg/mL (average = 166 mg/mL), compared to an average daily protein excretion of 150 mg (Pisitkun *et al.*, 2006; Gonzalez-Buitrago *et al.*, 2007). Proteomic analysis of the clinical urine samples identified 11 proteins, some of which were expected, but others were not, particularly those with high MWs related to an immunological response. Research by Gonzalez-Buitrago *et al.* showed that a "...*myriad of proteins*..." have been detected in urine, meaning that the presence of protein should not have been a surprise (Gonzalez-Buitrago *et al.*, 2007); what was a surprise was the very high levels of protein present in the clinical urine samples. After the precipitation of protein from the clinical urine samples, the RP- and HILIC-MS analyses generated much more satisfactory data for the metabonomic analysis.

Prior to the statistical analysis of the resulting data, it was revealed that none of the recruited patients' fractures had gone to non-union or had suffered delayed healing. Despite this, the data were still analysed to see if a metabonomic approach could profile the body's response to fracture healing. To add to the initial disappointment, the cohort of clinical samples obtained contained no clear groupings of time-setted samples. After the initial urine samples collected at t = 0, any subsequent urine samples were collected anything from 7 to 133 days post-fracture, with no separation into more discrete groups; the samples were spread more or less evenly from 7 to 133 days. Because of this, the whole dataset was split into three groups, those at t = 0, those under 21 days post-fracture, and all remaining samples post-fracture. For the ankle cohort, only two groups could be generated (those from t = 0, and all other samples post-fracture) due to the lack of samples.

Statistical analysis of each of the three models generated from the two ionisation modes and using both RP- and HILIC-MSs data (as well as the concatenated data) produced some variables that were consistent across the models.

Despite the many setbacks that limited the potential success of the work presented within this chapter, there were some results that show some promise for further analysis of clinical samples from fracture patients. Further to this, the proteomic work produced some very interesting results, which should be investigated further.

4.10. Conclusions

The work presented within this chapter used the 'metabonomics toolbox' generated in Chapter Three in an attempt to profile the body's response to a fracture. Despite none of the recruited patients' fractures going to non-union, the study was flawed from the offset by the lack of proper time-setted samples, something which was beyond my control. However, after the initial cause of poor RP analyses was found to be caused by large amounts of protein being present in the majority of the clinical urine samples, subsequent RP- and HILIC-MS analyses provided some PLS models with reasonable external classification results, as summarised in figure 4.10.1:

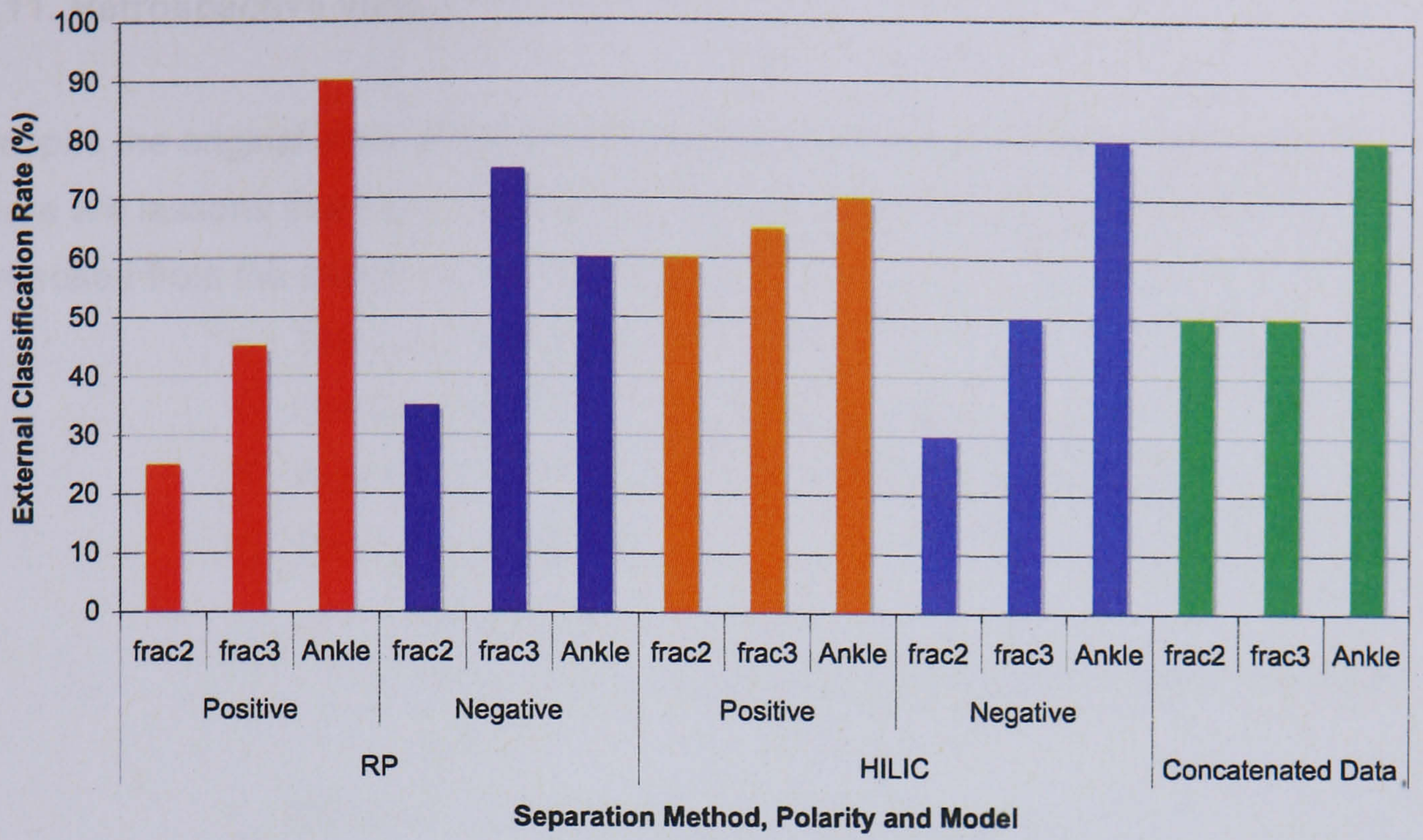


Figure 4.10.1. A graph comparing the external classification rates for each developed PLS model.

The external classification results are not as high as those presented in Chapter Three for the healthy volunteer samples, but the models relate to a more complex question than discrimination according to gender, time of collection and age, and would also be expected to be lower due to the lack of more discrete time-setted groups. Despite the high levels of protein, which was unexpected and has not been described in metabonomic literature before on the scale shown here, and the lack of many metabolite identifications (other than the possible detection of a steroid glucuronide), this work has produced some candidate biomarkers that could be related to the fracture healing process; obviously, the biomarkers are tenuous at best, due to the lack of clear, time-setted samples.

This work has also highlighted the conclusions in chapter 3.3, that proper design and implementation of a metabonomic study is vital if the end results are to stand any chance of being robust and of significance for hypothesis generation.

4.11. Retrospective view

Despite the original aims of the work undertaken within this chapter not being met, there are lessons that have been learnt. An 'ideal' experimental design can be proposed from the shortcomings highlighted:

- Thorough initial design of study
- Development of correct ethical guidelines
- Involvement of all parties before study commences
- Recruitment of patients
    - More patients required
    - More patient information obtained (as outlined in Chapter 3.8.1)
- Samples collected in the same manner
    - All samples treated in the same manner
    - Frozen within the same timescale
- Analysis of samples using a broad range of analytical methods
- Thorough statistical evaluation of resulting data
- Consultation with all parties of results obtained
- Re-analysis depending upon outcome
- Clinical testing based upon results obtained

Chapter Five

Structural determination of an extract from *Pseudomonas chlororaphis* PCL 1391

5.1. Introduction

5.1.1. Introduction to *Pseudomonas chlororaphis* PCL 1391 and tomato foot and root rot

Many plants have an intimate association with soil microorganisms in the rhizosphere[1]; some interactions can cause plant disease, whilst others can conversely protect against disease. Tomato foot and root rot is such a disease, which results in leaf yellowing, loss of turgidity and ultimately the death of the plant. Foot and root rot in tomato plants is caused by the fungal pathogen *Fusarium oxysporum f. sp. radicis-lycopersici*; the fungus colonises the xylem within plants and affects the transport of water throughout the plant, causing *Fusarium* wilt. The use of synthetic fungicides can (partially) suppress the effects of foot and root rot, however, there are both environmental and human/animal health questions about the desirability of long term usage of such synthetic fungicides. Beneficial plant-microbe interactions are therefore of interest, as these can provide *in situ* protection against pathogens, which could reduce man's reliance upon synthetic fungicides to aid the promotion of a healthy rhizosphere, increasing plant/crop productivity.

Various strains of *Pseudomonas* bacteria have been shown to possess biocontrol properties (Thomashow and Weller, 1995; Haas and Defago, 2005); pathogens are generally not completely removed from the rhizosphere by the actions of the biocontrol agents, but show reduced growth and lack/reduction of disease symptoms. A bacterial library isolated from the tomato rhizosphere from a commercial field in Andalucia, Spain has been analysed for the potential ability to cause disruption to the fungal pathogen that causes foot and root rot (Chin-A-Woeng *et al.*, 1998). The most active strain isolated was *Pseudomonas chlororaphis* PCL 1391, which was found to secrete many secondary antifungal metabolites (AFMs) such as hydrogen cyanide, chitinases, proteases and a hydrophobic compound, identified as phenazine-1-carboxamide (PCN) (Chin-A-Woeng *et al.*, 1998).

Phenazines are heterocyclic nitrogen-containing molecules that exhibit broad-spectrum antibiotic activity by inhibiting growth/metabolism (Turner and Messenger, 1986), and have been shown to be toxic to many organisms including bacteria, fungi and algae (Toohey *et al.*, 1965).

---

[1] The zone in soil immediately surrounding plant roots.

The effectivity of AFMs (such as PCN) for plant protection requires a delivery to the fungal pathogen for which bacteria need to be efficient colonisers of plant roots. More recently, *Pseudomonas* species have been shown to produce biosurfactants that are generally cyclic lipopeptides (CLPs) (Nielsen *et al.*, 2002). CLPs can be likened to the shape of a magnifying glass, where the handle corresponds to a fatty acid (FA) chain, and the glass to an amphiphilic ring structure consisting of amino acids (AAs). CLPs are thought to aid in the biocontrol of disease by causing disruption to lipid membranes/outer membrane structures (Coraiola *et al.*, 2006) by creating transmembrane pores, which when combined with AFMs, actively decrease fungal pathogen activity. CLPs have a broad range of structures (Desai and Banat, 1997), which explains the large and varied number of biological properties of *Pseudomonas* biosurfactants (Nielsen *et al.*, 2002).

5.1.2. CLP production

Whilst the ribosomal synthesis of proteins and peptides is template driven by the use of messenger and transfer RNA, the synthesis of CLPs by gram negative bacteria is non-ribosomal, and is catalysed by large peptide synthetases (Marahiel *et al.*, 1997; Nielsen *et al.*, 2002). Polypeptide chains are 'grown' using a thiotemplate mechanism, where 'domains' allow the sequential addition of specific AAs onto a growing product (lipopeptides or other antibiotics such as vancomycin). Many reactions, such as adenylation, thiolation, condensation and epimerisation (to name but a few) also occur, and allow modifications such as the cyclisation of polypeptide chains through the formation of an ester bond, or the addition of FA moieties (Marahiel *et al.*, 1997).

5.1.3. CLP analysis by CID tandem MS and amino acid analysis

Both CLPs and AFMs can be analysed using a variety of methods such as LC, MS(MS) (Yakimov *et al.*, 1999; Yang *et al.*, 2006), LC-MS(MS) (Chin-A-Woeng *et al.*, 1998), GC-MS (Yang *et al.*, 2007), NMR (Ptak *et al.*, 1980; Chin-A-Woeng *et al.*, 1998; Scott *et al.*, 2007), amino acid analysis (AAA) (Rodrigues, 2005) and by chemical tests (Wang *et al.*, 2003). As for metabonomics, no single technique can provide a complete picture of a CLP structure, rather a combination of approaches has to be utilised in determine to obtain the full structure.

The analysis of peptides (such as CLPs) using MS methods is well established[1], with peptide structural information being obtained upon CID, where bond cleavage occurs (principally the peptide bonds). The nomenclature for the fragmentation of peptides was originally proposed by Roepstoff and Fohlman (Roepstorff and Fohlman, 1984), with a subsequent modification to the nomenclature proposed by Johnson *et al.* (Johnson *et al.*, 1987).

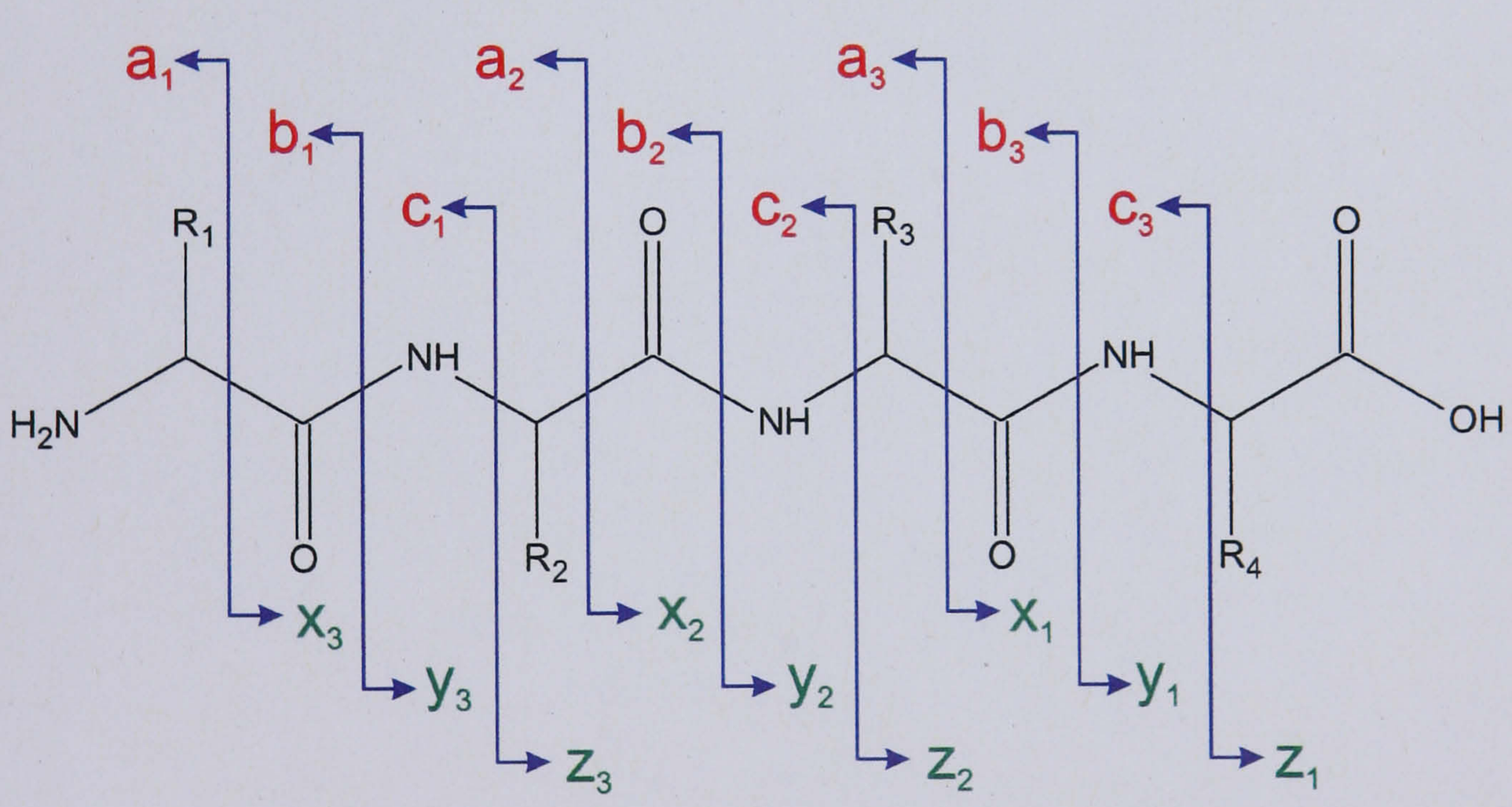


Figure 5.1.1. Peptide fragmentation nomenclature. (Johnson *et al.*, 1987)

Upon fragmentation of a peptide, if the charge remains on the N-terminus, the resulting fragments are described as either a_n, b_n or c_n ions; if the charge remains on the C-terminus after fragmentation, then the ions are described as either x_n, y_n or z_n ions. The subscript 'n' refers to the number of AAs in the fragment. For spectra of

[1] 22,062 search results for the phrase 'peptide mass spectrometry' using http://www.pubmed.com (accessed November 2007)

peptides obtained by low energy CID tandem MS of protonated peptides, the ions that are most commonly observed, correspond to b and y fragmentation pathways due to the peptide bond being the weakest, and therefore the easiest to break.

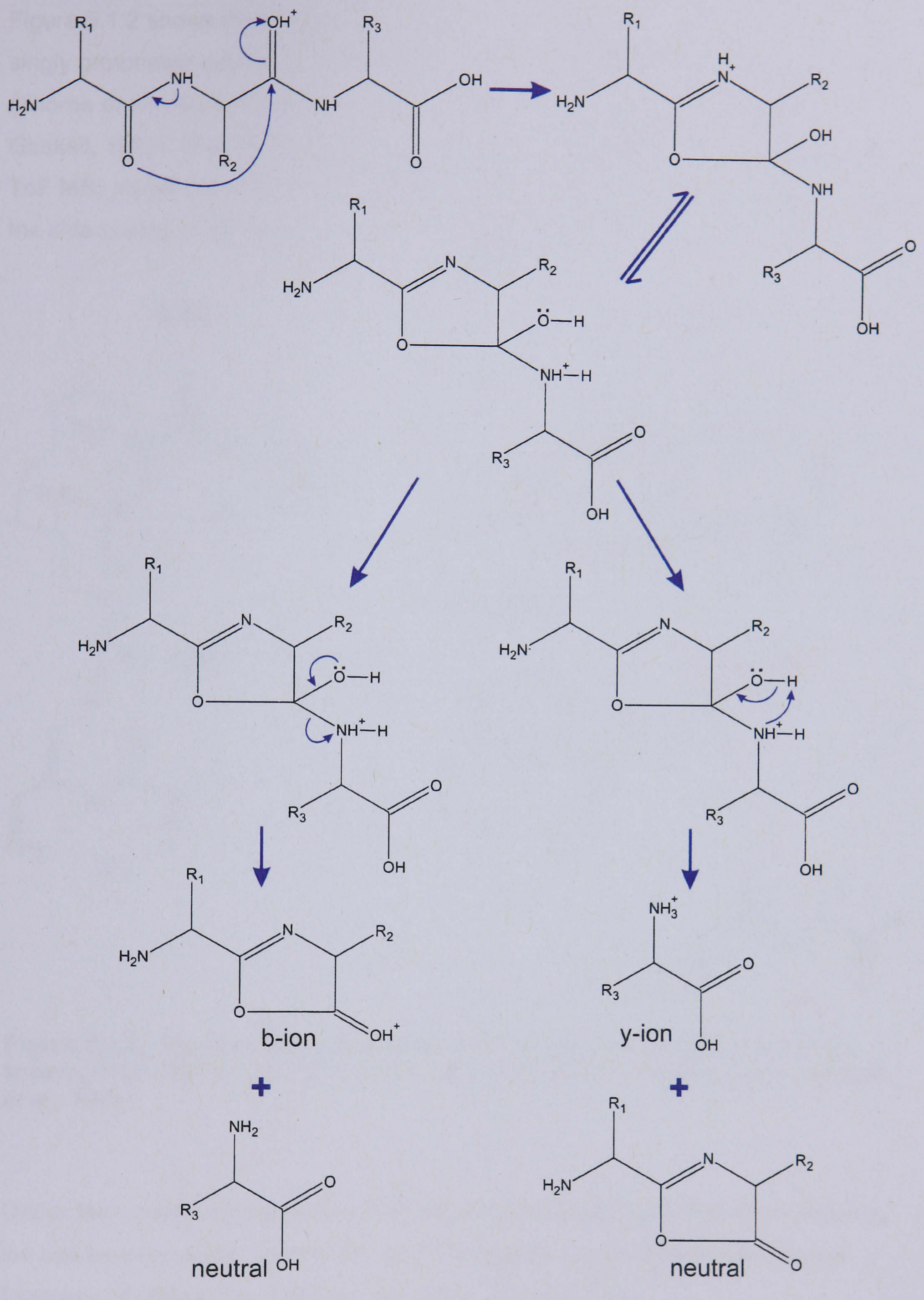


Figure 5.1.2. Proposed formation of b and y ions by CID tandem MS from many papers, principally (Thorne *et al.*, 1990; Kenny *et al.*, 1992; Yalcin *et al.*, 1995; Summerfield and Gaskell, 1997).

Figure 5.1.2 shows the proposed mechanism for the formation of b and y ions from singly protonated peptides, as developed over a range of literature, but principally (Thorne *et al.*, 1990; Kenny *et al.*, 1992; Yalcin *et al.*, 1995; Summerfield and Gaskell, 1997). The use of high energy CID (≥ 1 keV, such as that afforded by ToF-ToF MS) allows the formation of other fragment ions, which arise from cleavages of the side chains of AAs to form d_n, v_n or w_n fragments.

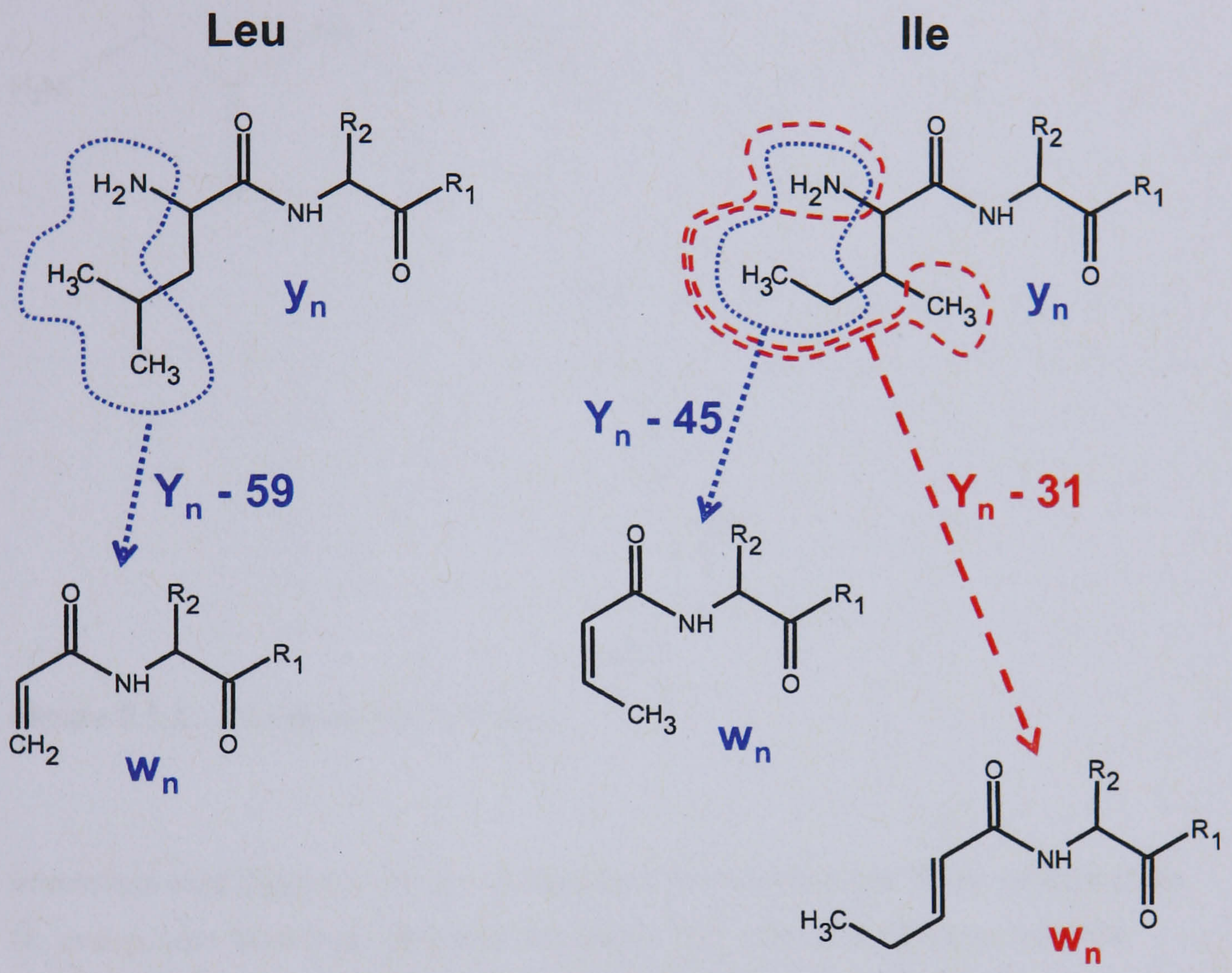


Figure 5.1.3. Proposed fragmentation scheme for the isomeric AAs Leu and Ile, showing their differentiation by the formation of different w_n fragment ions (Johnson *et al.*, 1988).

Under the correct conditions, the formation of w_n fragment ions from the relevant y_n ion can be very useful, as isomeric AAs Leu and Ile can be differentiated by the formation of different w_n fragment *m/z* values (figure 5.1.3).

There are other fragment ions that can typically be found at low *m/z* values in tandem mass spectra (below *m/z* 150), namely immonium ions.

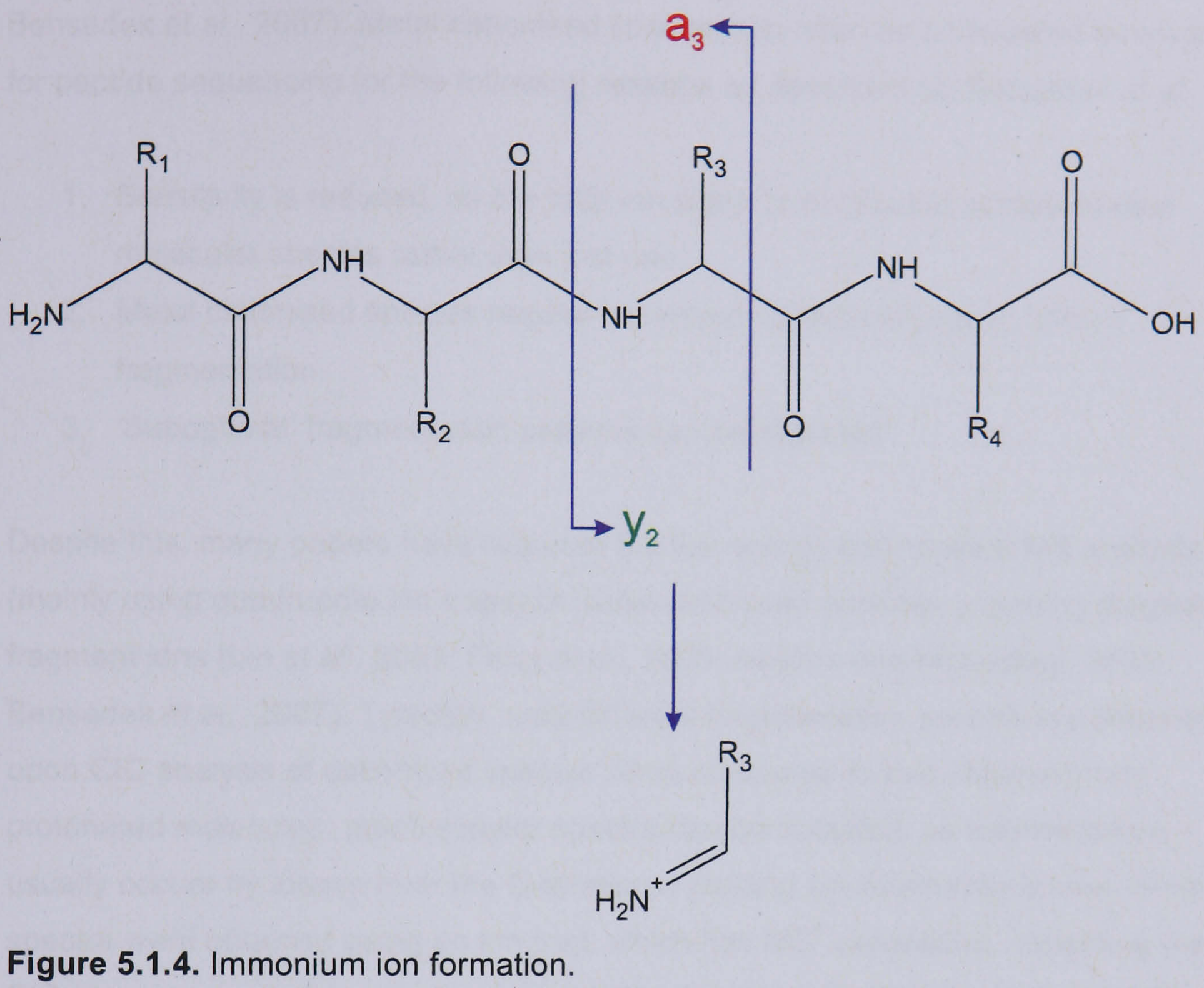


Figure 5.1.4. Immonium ion formation.

Immonium ions (figure 5.1.4) are formed by a double cleavage of the peptide chain (a_n and y_n type cleavage), and are indicative of the individual AAs that form the peptide chain. Some immonium ions are much less stable and thus less often observed than others.

The most commonly-utilised precursor ions are protonated, however, the use of sodiated molecules as precursors has been reported (Grese *et al.*, 1989; Teesch and Adams, 1990; Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004; Bensadek *et al.*, 2007). Metal cationised species can often be considered a nuisance for peptide sequencing for the following reasons as described by Bensadek *et al.*:

1. Sensitivity is reduced, as the total ion signal is distributed across several molecular species rather than just one.
2. Metal cationised species require increased collision energy to obtain fragmentation.
3. 'Suboptimal' fragmentation patterns can be obtained[1].

Despite this, many papers have reported the low energy CID tandem MS analysis (mainly using quadrupole ion traps) of metal cationised peptides providing diagnostic fragment ions (Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004; Bensadek *et al.*, 2007). Typically, very different fragmentation spectra are obtained upon CID analysis of cationised species when compared to that obtained from protonated molecules; much simpler spectra can be obtained, as fragmentation usually occurs by losses from the C-terminus, yielding predominantly b ions. Simple spectra were obtained using an ion trap, which has MS^n capabilities, repeating the CID process, producing structural information by sequentially generating b ions (Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004; Bensadek *et al.*, 2007). Analysis using a Q-o-ToF provides a much different fragmentation spectrum when compared to that obtained using an ion trap; increased fragmentation is usually obtained as the MS/MS step allows the generation of many more fragments, however, this can lead to harder spectra to interpret.

For C-terminus sequencing using cationised species, peptides of less than ten residues are generally used as it has been shown to be harder to obtain cationised species above 1 kDa using ESI. The use of MALDI has been reported for the ionisation of peptides up to 16 residues in length, but again, difficulty in forming cationised molecules generally results in protonated species fragmented instead (Newton and McLuckey, 2004).

[1] Meaning that fragment ions cannot be interpreted or are structurally uninformative.

5.1.4. Amino acid analysis

Amino acid analysis (AAA) is a process that can determine the quantities of individual AAs in a sample. The sample is first hydrolysed using a strong acid for a period of 24 h at a temperature of 110 °C. This releases free AAs into solution by hydrolysing the peptide bonds. After hydrolysis, the sample is dried and reconstituted in a rehydration fluid that contains L-homo-Arg (0.01 M) as internal standard.

Any free AAs within the reconstituted solution cannot be detected using fluorescence without first being derivatised.

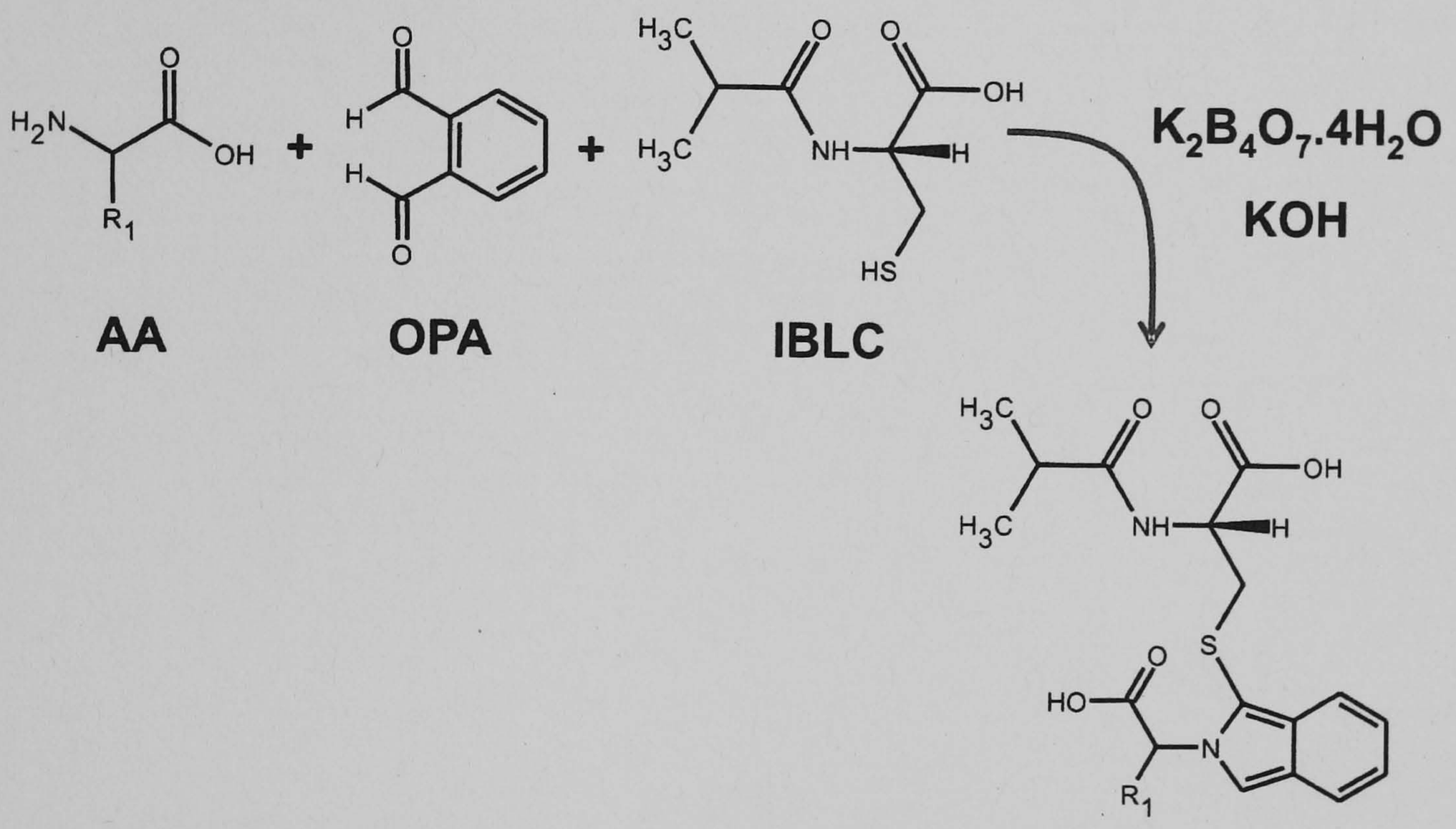


Figure 5.1.5. Reaction scheme for the derivatisation of free AAs (Bruckner *et al.*, 1994).

The derivatisation is carried out prior to reversed phase analysis of the resulting derivatised AAs using a C_{18} column. Free AAs in solution are reacted with *N*-iso-L-butyryl-L-cycteine (IBLC) and *o*-phthaldialdehyde (OPA) (figure 5.1.5); this is carried out on-line within the HPLC system. This method of AAA allows the routine detection of the L and D isomers of the following 12 AAs: Asp, Glu, Ser, Thr, Arg, Ala, Tyr, Val, Met, Phe, Leu and Ile.

5.1.5. Aims

The aims of this work were to identify the components in a fraction isolated using HPLC of an ethyl acetate extract of *P. chlororaphis* PCL 1391 spent growth medium, provided by the Department of Biology, University of Leiden, the Netherlands. This fraction was selected for the presence of surface tension reducing ability. Here, the same methods as used in a metabonomic study were exploited, but in a very different manner. Both ESI- and MALDI-MS and tandem MS (along with chemical methods) were used for structural identification. In addition, racemic amino acid analysis was used to provide extra information on amino acid composition of the compounds in the fraction.

5.2. Results

5.2.1. Sample information

A sample collected following HPLC fractionation of the supernatant obtained from an extract of *Pseudomonas chlororaphis* PCL 1391 was provided by the Institute of Biology, Leiden University, the Netherlands. Researchers at Leiden are studying PCL 1391 in an attempt to unravel the mode of action of the antifungal metabolite phenazine-1-carboxamide, which is produced by this strain of bacteria; it was discovered that PCL 1391 also produced an unknown biosurfactant.

Analysis of the HPLC fraction was undertaken as part of a long-standing collaboration between the JTO group at the University of York and Leiden University in the area of plant-microbe interactions. Initial findings by Leiden suggested that the unknown biosurfactant may be similar to the cyclic lipopeptide massetolide C (figure 5.2.1), as a BLAST[1] search in Leiden had suggested that the PCL 1391 synthase sequence was similar to that of the synthase that produces massetolide C.

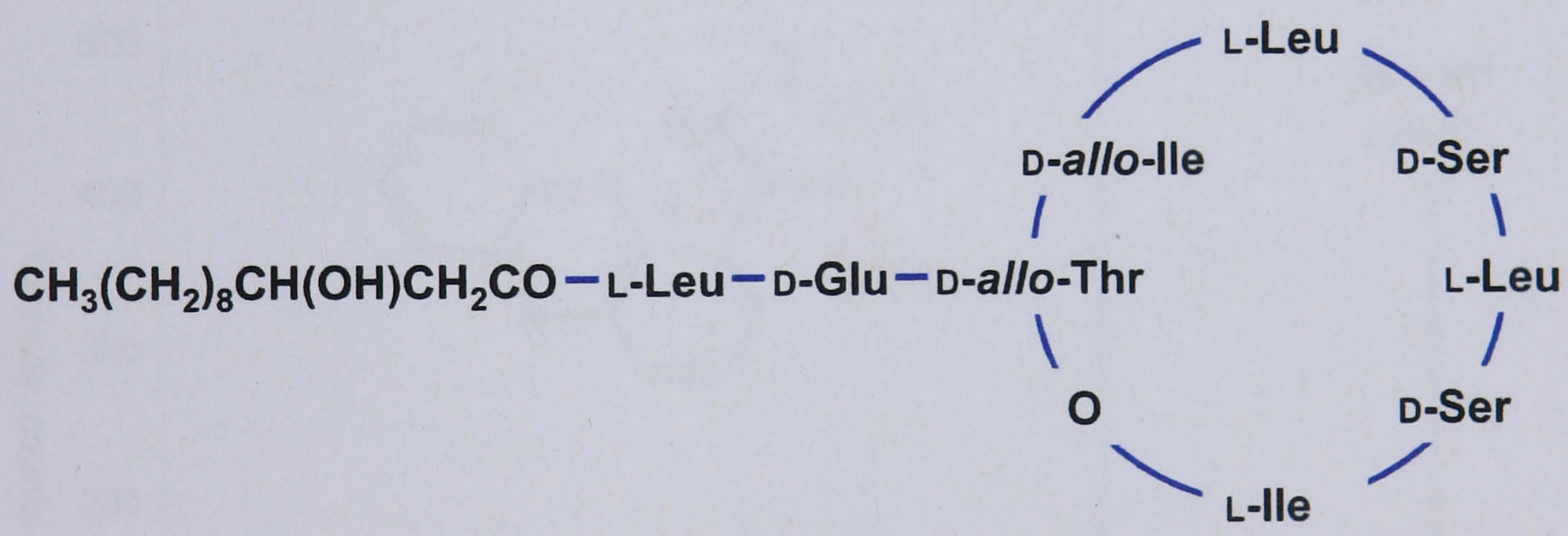


Figure 5.2.1. Structure of the CLP massetolide C.

[1] Basic Local Alignment Search tool that finds regions of similarity between nucleotide or amino acid sequences.

5.2.2. ESI-Q-o-ToF MS analysis of PCL 1391 extract

The HPLC fraction was received in a glass vial and had been dried under vacuum in a centrifugal evaporator. The sample was reconstituted in 300 µL of MeOH before a dilution series was created. Initial positive ion mode ESI-Q-o-ToF MS studies showed that a dilution to 25 % of the original stock solution gave a satisfactory signal.

PCL 1391 had already been shown to produce the secondary antifungal metabolite phenazine-1-carboxamide (PCN) (Chin-A-Woeng *et al.*, 1998), this was expected to be present in the reconstituted solution as it co-eluted with the unknown biosurfactant (shown by HPLC UV chromatograms, in Leiden). Chin-A-Woeng *et al.* identified PCN by MS and NMR; their MS/MS analysis of a protonated molecule at *m/z* 224 produced two fragment ions at *m/z* 207 and 179, which were identified as losses of NH_3 and the carboxamide group respectively. A peak at *m/z* 224 was identified in the ESI-MS of the reconstituted solution and was postulated to correspond to PCN, which has a nominal mass of 223.

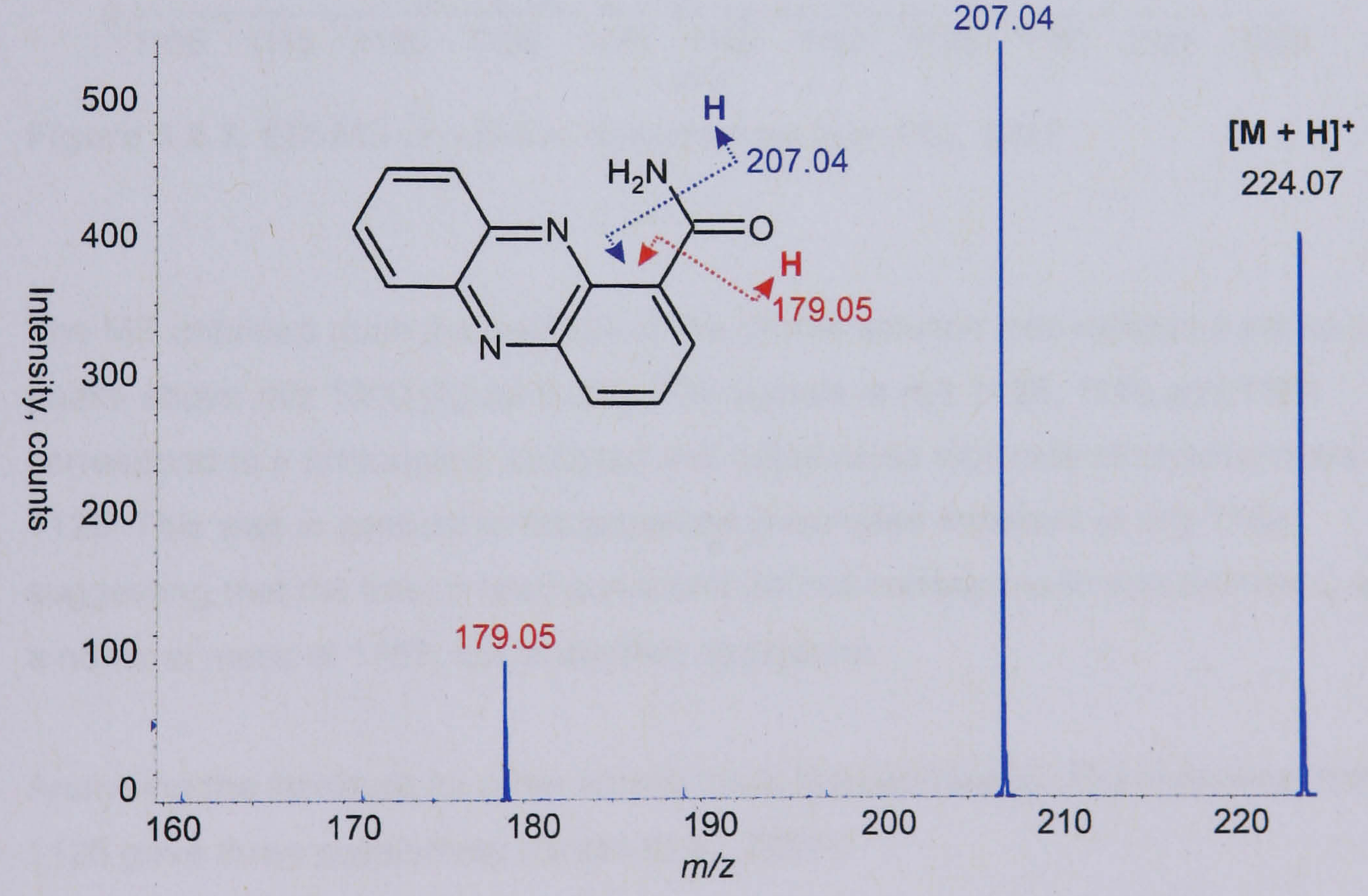


Figure 5.2.2. Product ion spectrum of proposed PCN peak at *m/z* 224.

The product ion spectrum of *m/z* 224 contained only two fragment ion peaks (figure 5.2.2). The peak at *m/z* 207 corresponds to a loss of 17 Th, which is equivalent to the loss of NH_3; the less abundant fragment ion at *m/z* 179 corresponds to the loss of the

carboxamide group, leaving the very stable phenazine ring. These MS/MS results are identical to those of Chin-A-Woeng *et al.*, suggesting that the antifungal metabolite PCN was present in the HPLC fraction.

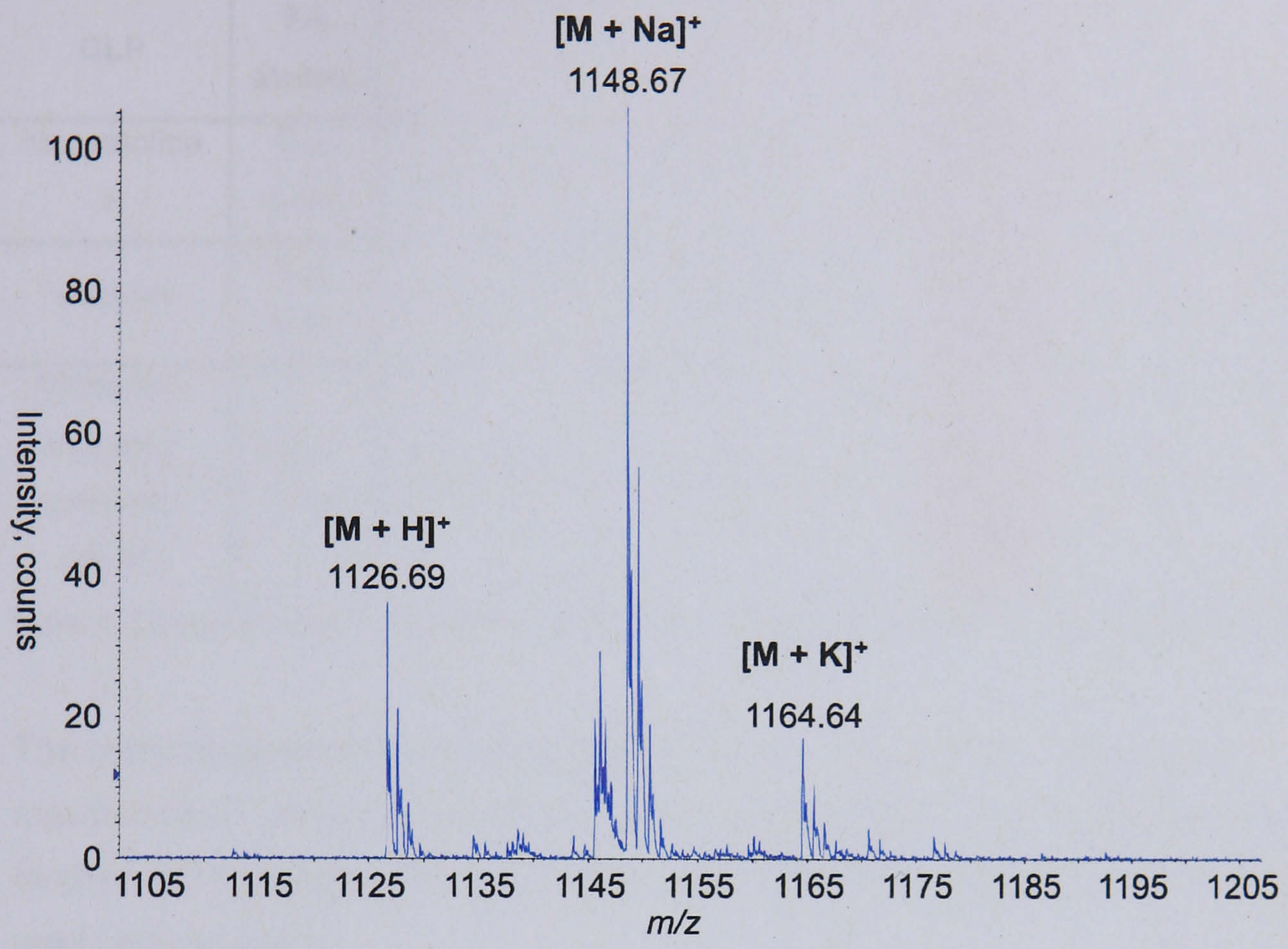


Figure 5.2.3. ESI-MS of putative biosurfactant from PCL 1391.

The MS obtained upon the analysis of the diluted solution also exhibits a series of peaks above *m/z* 1000 (figure 5.2.3). The signals at *m/z* 1126, 1148 and 1164 correspond to a protonated, sodiated and potassiated molecule of nominal mass 1125. This was in contrast to the expected protonated molecule at *m/z* 1168, suggesting that the fractionated surfactant did not correspond to massetolide C with a nominal mass of 1167, but to another compound.

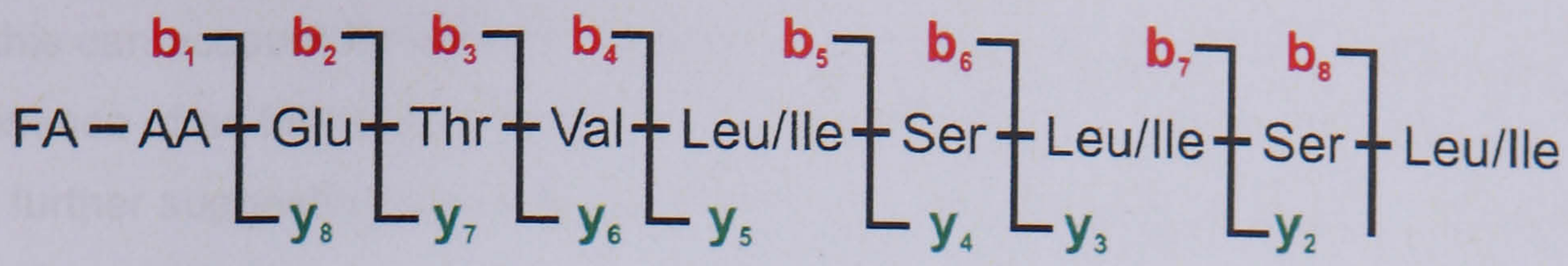
Analysing the literature for other known cyclic lipopeptides (CLPs) of nominal mass 1125 gave three possibilities (Gross *et al.*, 2007):

Table 5.2.1. CLPs reported within the literature with nominal mass 1125 (Gross *et al.*, 2007). Each of the three CLPs cyclise through an ester bond between the AAs in positions 3 and 9. The pink shaded cells highlight the only differences between the three CLPs.

CLP	FA Moiety	AA Sequence Number 1	2	3	4	5	6	7	8	9
Massetolide F	C_{10}, 3-OH	L-Leu	D-Glu	D-*allo*-Thr	D-Val	L-Leu	D-Ser	L-Leu	D-Ser	L-Leu
Viscosin	C_{10}, 3-OH	L-Leu	D-Glu	D-*allo*-Thr	D-Val	L-Leu	D-Ser	L-Leu	D-Ser	L-Ile
White line inducing principle (WLIP)	C_{10}, 3-OH	L-Leu	D-Glu	D-*allo*-Thr	D-Val	D-Leu	D-Ser	L-Leu	D-Ser	L-Ile

(Leu = Leucine, Glu = Glutamic acid, Thr = Threonine, Ser = Serine, Ile = Isoleucine.)

The protonated molecule observed at *m/z* 1126, which could correspond to massetolide F, viscosin, WLIP or to an as yet unreported CLP, was analysed by CID tandem MS to produce fragment ions to aid in elucidating the structure of the unknown molecule.



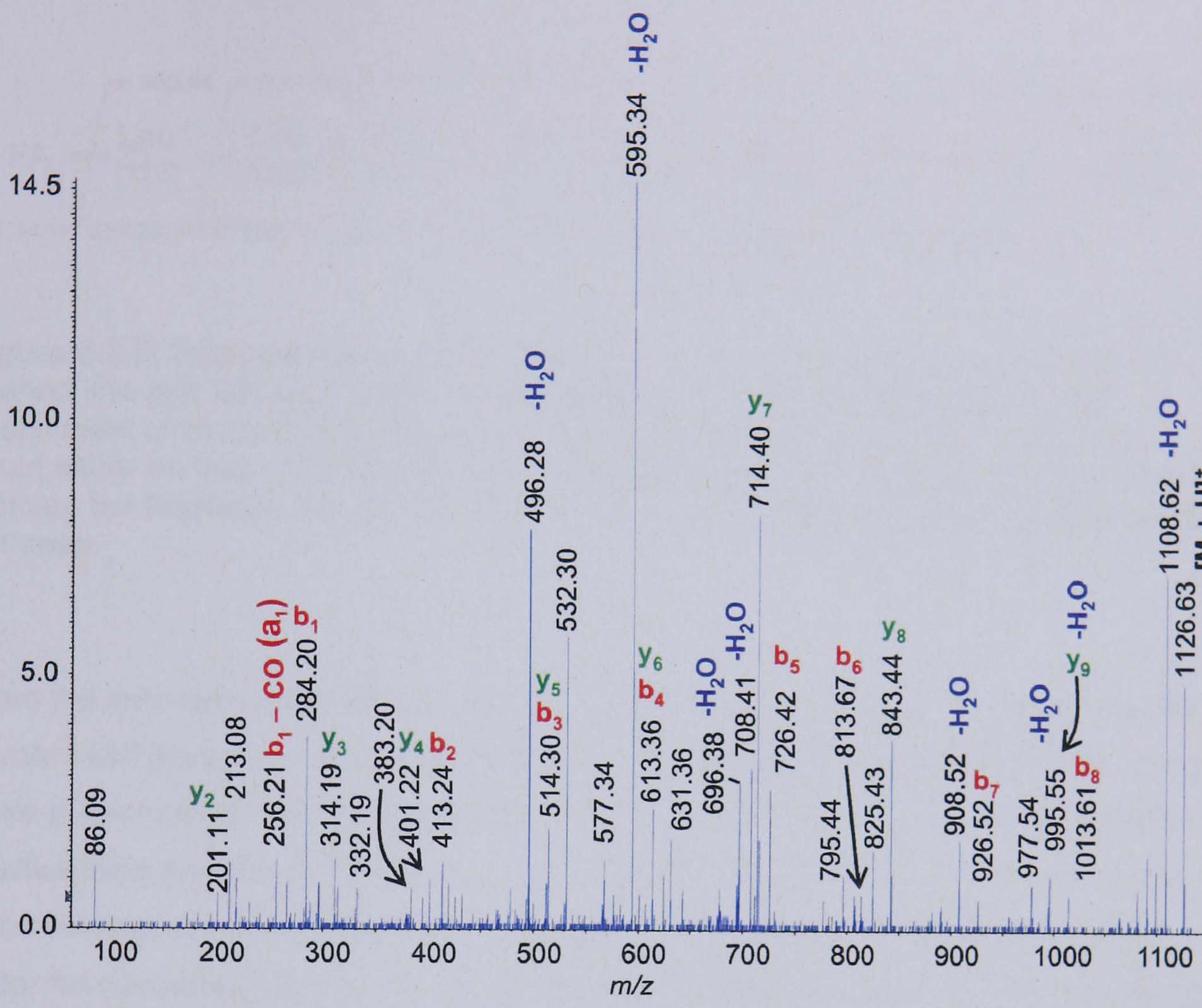


Figure 5.2.4. Product ion spectrum of the protonated molecule at *m/z* 1126, with both y and b ion AA losses identified ('AA' in the sequence indicates unknown amino acid).

On CID product ion tandem MS, many fragment ions were generated across a broad range of *m/z* values, with the structural interpretation of the fragment ions not being straightforward due to the cyclic nature of the putative CLP at *m/z* 1126 (figure 5.2.4). Two bond cleavages must occur to generate the y and b ions, with ring opening occurring principally at the ester bond. Losses corresponding to the AAs Leu and/or Ile, Glu, Thr, Val and Ser were found within the spectrum. Starting with the protonated precursor ion at *m/z* 1126, losses corresponding to Leu/Ile, Ser, Leu/Ile, Ser, Leu/Ile, Val, Thr and Glu are observed, which correspond to the b ions shown from b_8 to b_1 and a_1. A complementary partial y ion series from y_8 to y_2 was also observed. A common feature of positive ionisation, low energy CID, is the facile loss of water during fragmentation across a peptide backbone (Ballard and Gaskell,

1993); this can account for many of the peaks 18 Th less than the b or y series ions. The presence of an immonium ion at *m/z* 86 corresponds to the isomeric AAs Leu and Ile, further suggesting the presence of either or both of these AAs within the CLP structure.

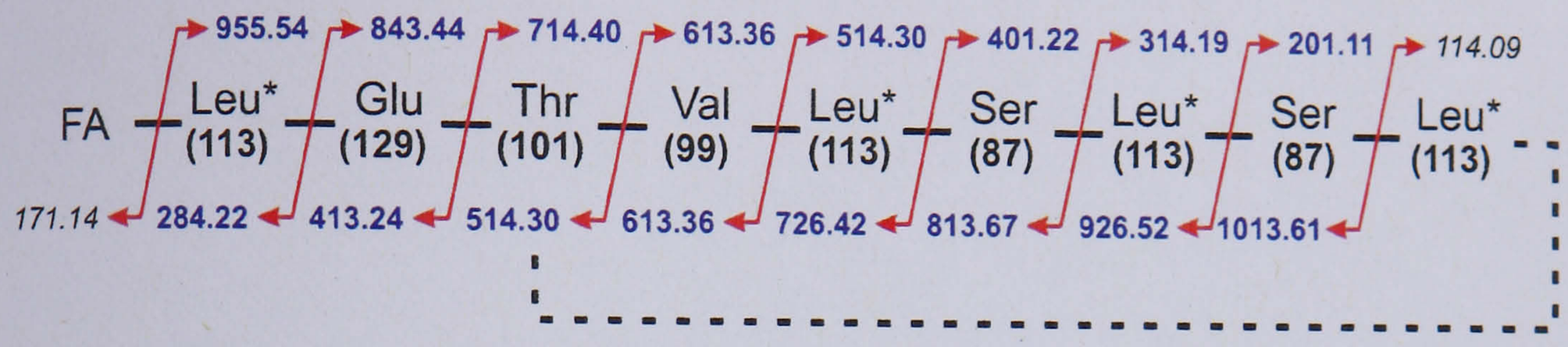


Figure 5.2.5. Proposed structure of fragment ions (proposed cyclisation shown by hashed line and values in parenthesis correspond to AA residual masses) and assignment of biosurfactant from the CID tandem MS analysis of *m/z* 1126. Leu* could either be Leu or Ile and FA corresponds to fatty acid. Values in bold and blue indicate the fragment ions present in the product ion spectrum; values in black were not seen.

From the structural information obtained upon the fragmentation of the precursor ion at *m/z* 1126 (figure 5.2.4), an AA sequence was proposed (figure 5.2.5). Each of the three published CLP structures (table 5.2.1) are known to cyclise between the AAs in position nine and three (corresponding to Leu/Ile and D-*allo*-Thr). For an ester bond to be formed, an –OH group needs to be present on one of the AA side chains in order for condensation to occur. This means that given the proposed structure (figure 5.2.5), cyclisation could occur through the AAs Ser or Thr in positions three, six and eight. Condensation through Ser in positions six and eight is unlikely due to steric hindrance, leaving only three as the most likely site for cyclisation. From the product ion spectrum (figure 5.2.4), the C-terminal fragments y_8 and y_7 are more intense than lower *m/z* y ions; this suggests that the cleavage between Leu-Glu (positions one and two) and Glu-Thr (positions two and three) was easier than that of Thr-Val (positions three and four). This would require cleavage of two bonds, rather than one, consistent with cyclisation occurring between the C-terminal carboxylic acid and the OH in the side chain of Thr (as in the three reported CLP structures).

As the literature relating to sodiated peptide fragmentation suggests that the fragmentation obtained is simpler to interpret than protonated peptide fragmentation (Grese *et al.*, 1989; Teesch and Adams, 1990; Ballard and Gaskell, 1993; Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004), the sodiated molecule at *m/z*

1148 was analysed by CID tandem MS to try and simplify the resulting product ion spectrum of the cyclic CLP.

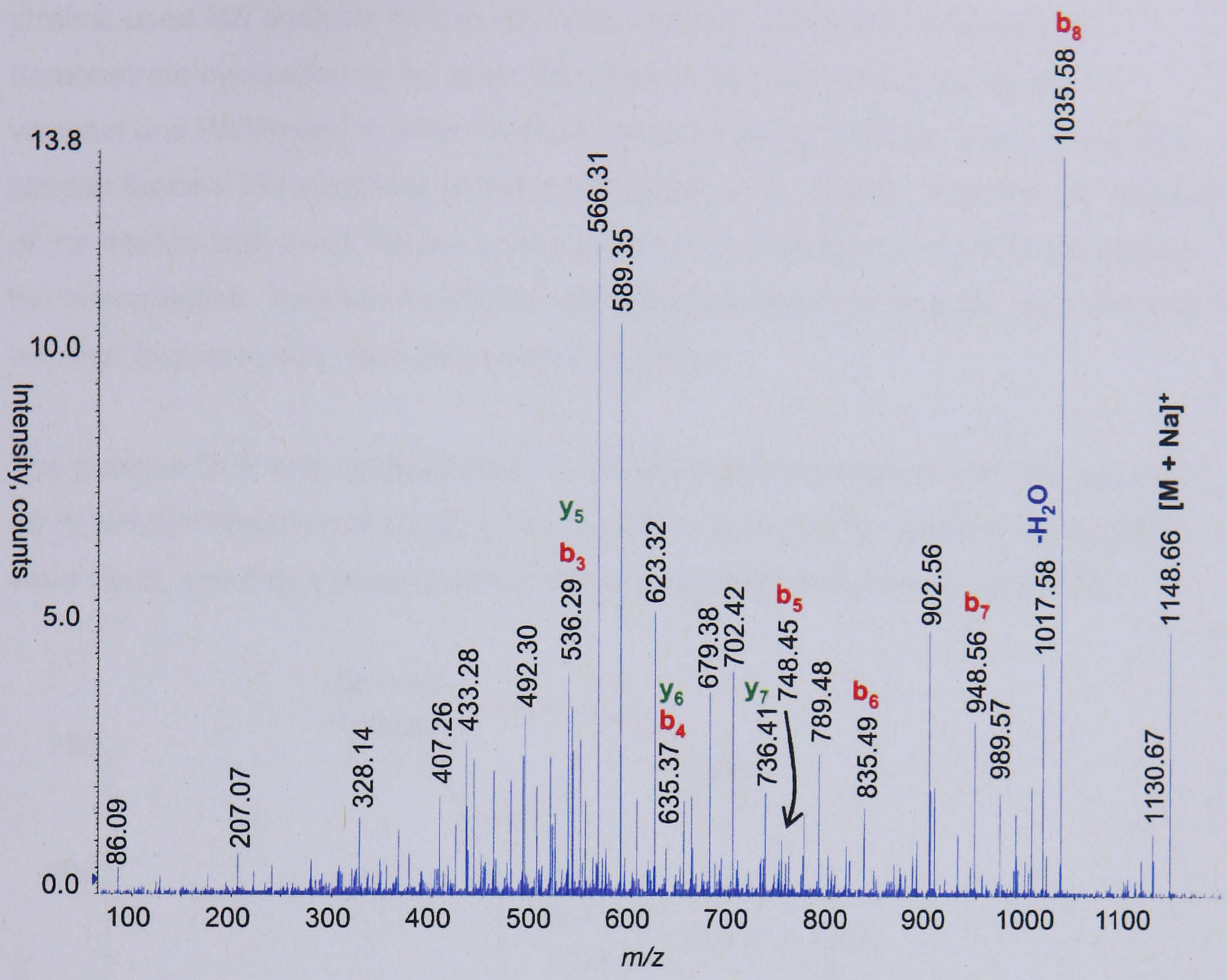


Figure 5.2.6. Product ion spectrum of the sodiated molecule at *m/z* 1148.

The product ion spectrum of the more intense sodiated molecule at *m/z* 1148 contains many fragments across a broad range of *m/z* values, which at first glance yields an even more complex spectrum than that of the protonated molecule (figure 5.2.4). In contrast to the literature, where sodiated peptides produce fragment ions almost exclusively from the C-terminus, there are many peaks present that cannot be readily identified, with the exception of a few b and y ions that correspond to the sodiated mass of the fragment ions already identified in figure 5.2.4.

As more evidence was required to assign the structure of the CLP, and given the complexity of the resulting product ion spectrum from subjecting the CLP to CID tandem MS in its cyclic form, it was decided to open the ring to form a linear peptide, hopefully simplifying the resulting product ion spectrum.

5.2.3. ESI-Q-o-ToF MS analysis of CLP after treatment with base

Research by Kuiper *et al.* into biosurfactants produced by other *Pseudomonas* strains, used MS analysis before and after mild base treatment followed, to demonstrate cyclisation by an ester bond (as in the structures massetolide F, viscosin and WLIP) and to subsequently provide a linear peptide, which provided a simpler tandem MS spectrum to interpret (Kuiper *et al.*, 2004). Upon the MS analysis of the treated surfactant, Kuiper *et al.* observed an increase of 17 Da to the mass of the biosurfactant, consistent with the addition of ammonia across the ester bond; C-terminal fragment ions were also shifted by 16 Th.

The putative CLP with nominal mass 1125 was therefore treated with mild base (1:1 35 % NH_4OH:MeOH) overnight, which would be expected to cleave the proposed ester bond, creating a linear peptide with a protonated molecule at *m/z* 1143.

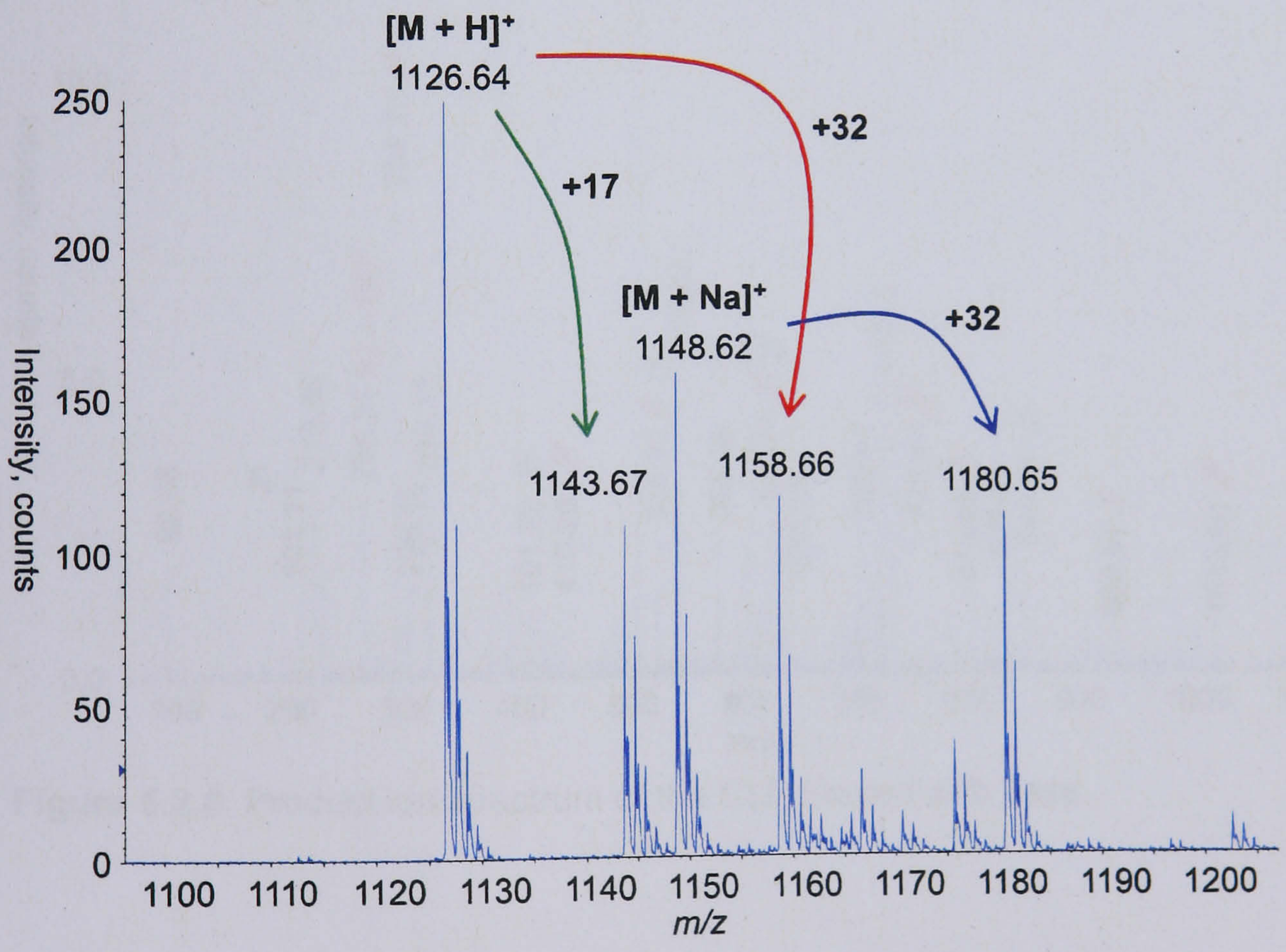


Figure 5.2.7. ESI-Q-o-ToF MS of CLP after treatment with ammonium hydroxide.

The resulting MS analysis of the base treated CLP shows five intense peaks at *m/z* 1126, 1143, 1148, 1158 and 1180 (figure 5.2.7). From the previous MS analysis of untreated CLP (figure 5.2.3), the peaks at *m/z* 1126 and 1148 correspond to the protonated and sodiated molecules of cyclic CLP with nominal mass 1125. The peak

at *m/z* 1143 is 17 Th higher that the protonated untreated CLP peak, meaning that despite the presence of unreacted CLP, the addition of ammonia across the ester bond appeared to have been partially successful. The two remaining peaks at *m/z* 1158 and 1180 have an increase of 32 Th from the protonated and sodiated untreated CLP peaks; this has been described before by Joao Rodrigues from the JTO group, where an increase of 32 Th corresponded to the addition of MeOH across the ester bond of a related CLP, rather than ammonia as expected (Rodrigues, 2005).

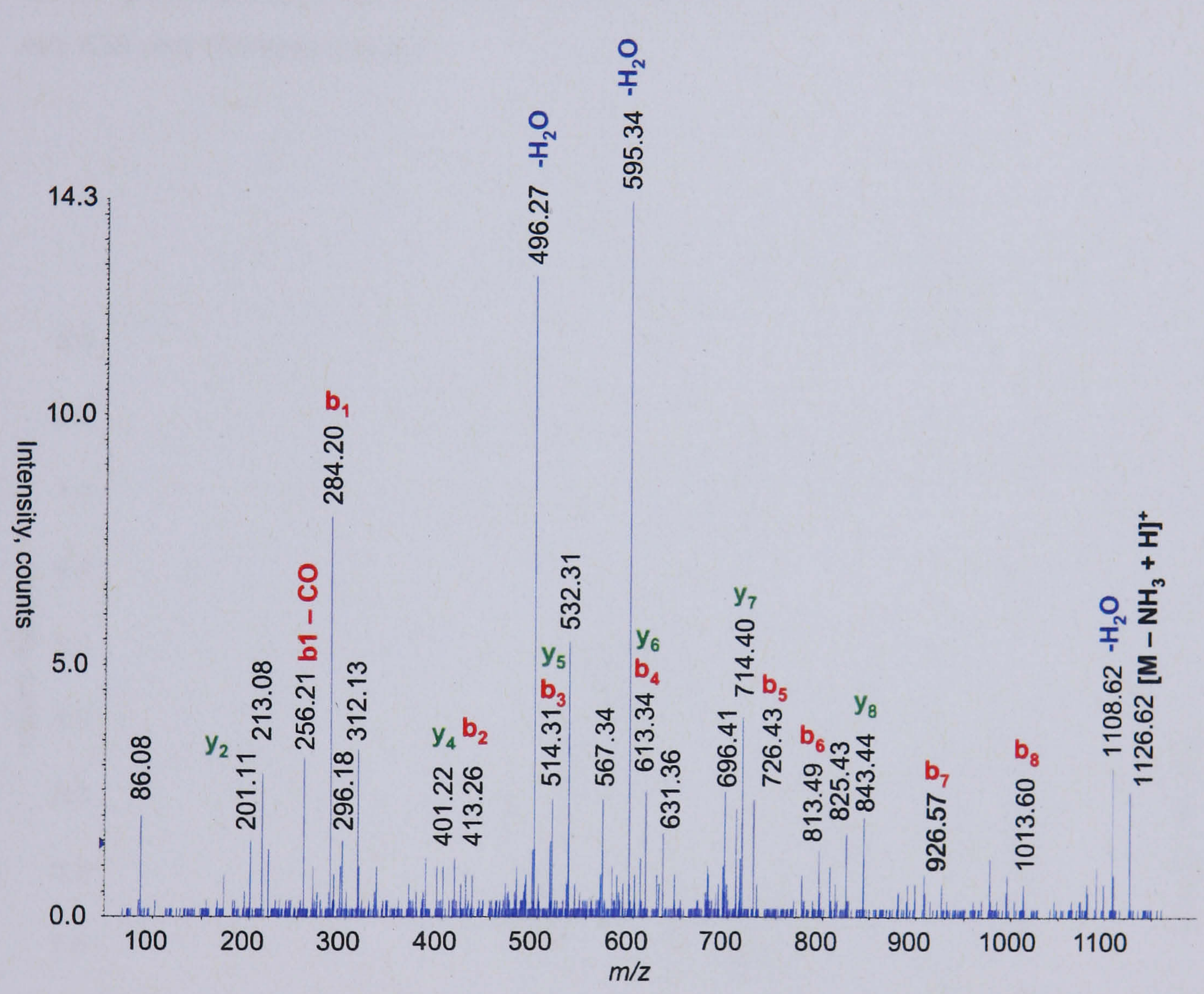


Figure 5.2.8. Product ion spectrum of the CLP treated with base.

The product ion spectrum resulting from CID tandem MS analysis of *m/z* 1143 (figure 5.2.8) showed a more complex spectrum than was expected for a ring-opened CLP. Upon isolation of the precursor ion at *m/z* 1143 with no collision energy (CE) applied, the peak immediately disappeared to yield the peak at *m/z* 1126 (figure 5.2.8). When the CE was increased, the fragmentation was extensive, generating a spectrum almost identical to that of the untreated CLP at *m/z* 1126 (figure 5.2.4).

As the peak at *m/z* 1126 was immediately observed upon isolation of the supposed ring opened CLP at *m/z* 1143, it suggests that the increase of 17 Th was caused by an ammonium adduct, rather than a ring-opened compound in which ammonia is added across the ester bond. Kuiper *et al.* did not isolate an ammonia adduct, as subsequent CID tandem MS analysis of their treated biosurfactant showed an increase of 16 Th for any observed C-terminus fragment ions (Kuiper *et al.*, 2004). Conversely, no change in any of the observed b and y fragment ions was observed here; the isobaric fragments b_3 & y_5, and b_4 & y_6 should have not been present in the resulting spectrum, as the y_5 and y_6 fragment ions increased from *m/z* 514 and 613 to *m/z* 530 and 629 respectively.

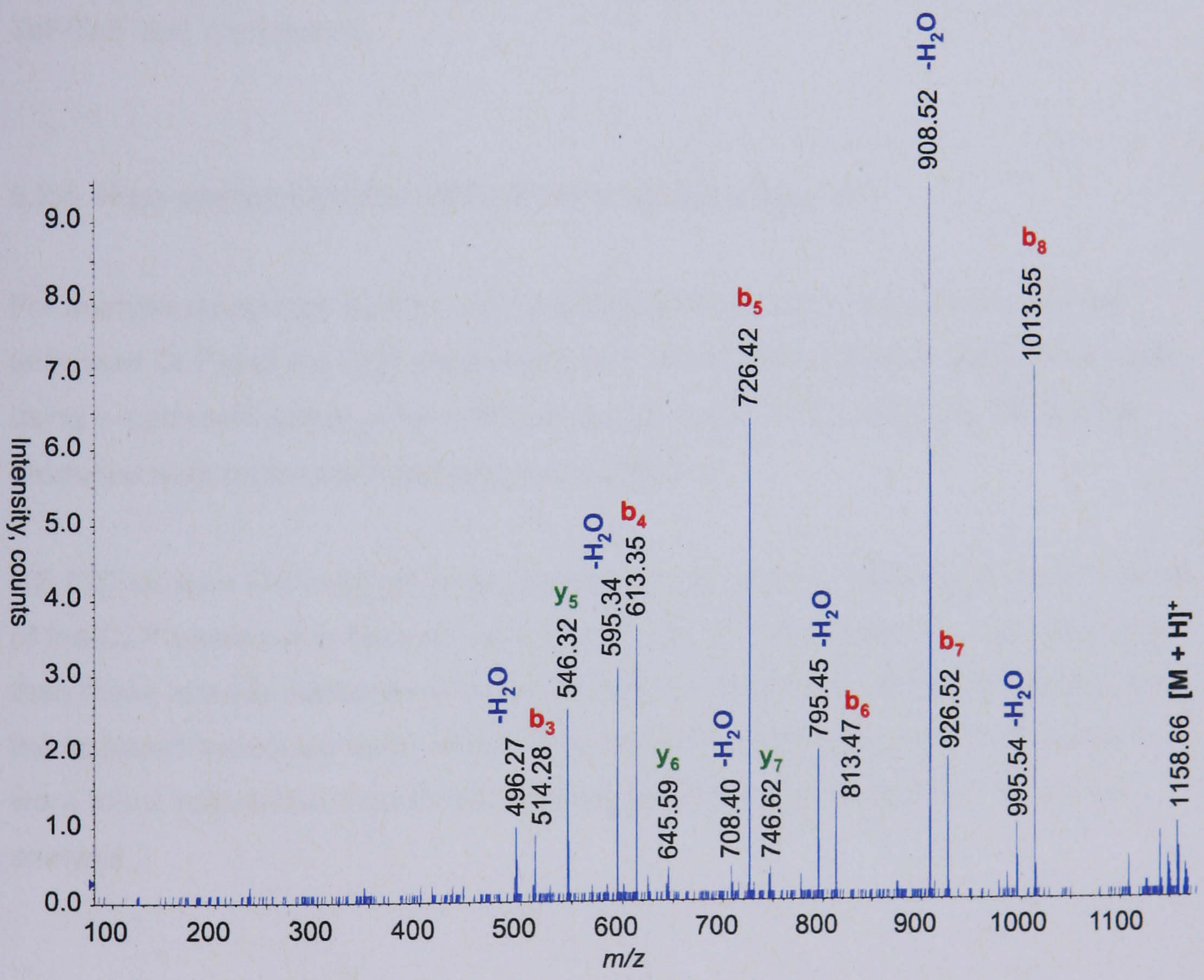


Figure 5.2.9. Product ion spectrum of protonated lipopeptide observed at 32 Th higher than the cyclic lipopeptide.

CID product ion analysis of the protonated molecule at *m/z* 1158 (figure 5.2.9) yielded a much simpler spectrum, with easily assignable b and y ion series (as expected for a linear peptide). The increase of 32 Th arising from transesterification, which results in the addition of elements of methanol across the ester bond, can be

confirmed by a shift of 32 Th for each of the observed y ions. Despite the addition of methanol being an unintended side reaction, it was of use, providing a ring opened CLP. Even though incomplete b and y ion series were observed, the presence of b_{3-8} and y_{5-7} ions aids in strengthening the proposed AA sequence assignment (figure 5.2.5).

CID tandem MS analysis of the sodiated ring opened CLP at *m/z* 1180 did not yield any additional information than the protonated precursor. To confirm the proposed AA sequence (and to attempt to elucidate the Leu/Ile uncertainty), a full range of fragment ions (including w ions) should be obtained to assign all of the AAs present, along with the FA moiety mass. For this reason, high energy CID using a MALDI-ToF/ToF was undertaken.

5.2.4. High energy MALDI-ToF/ToF MS analysis of the CLP

For analysis using high energy (HE) CID MALDI-ToF/ToF, an aliquot of both the untreated CLP and the CLP treated with mild base were analysed using MALDI-MS, using ∝-cyano-4-hydroxycinammic acid as the matrix. As for ESI-MS, MALDI-MS produced both protonated and sodiated molecules.

HE-CID tandem MS analysis of the protonated molecule of CLP at *m/z* 1126 and that of the CLP treated with base at *m/z* 1158 did not yield any additional fragment ions than those already observed using LE-CID tandem MS. Because of this, peaks for the sodiated molecules were analysed to see if their resulting product ion spectra were more informative than those obtained at low energy using ESI-Q-o-ToF MS analysis.

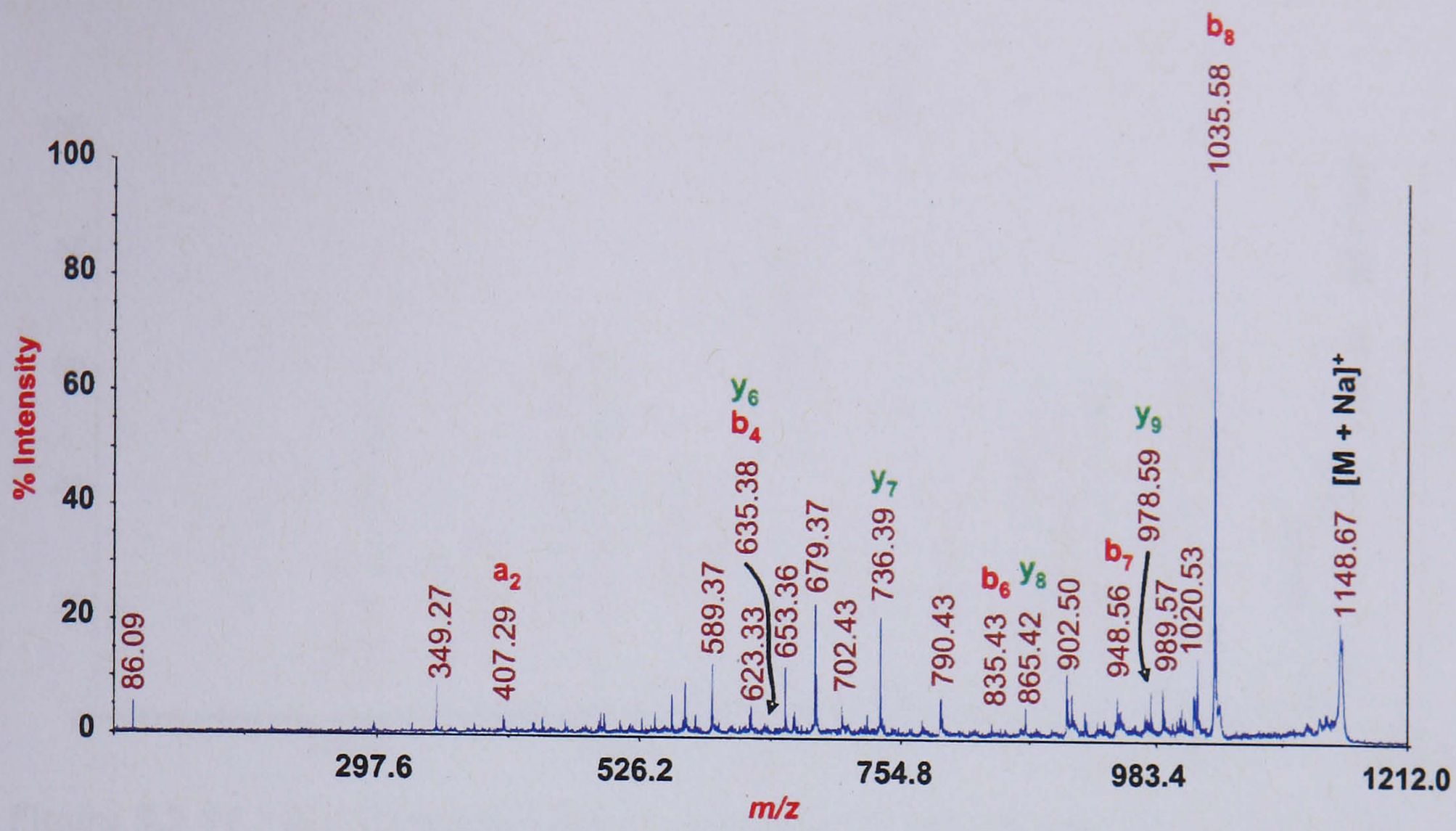


Figure 5.2.10. HE-CID tandem MS analysis of sodiated CLP at *m/z* 1148.

The HE product ion spectrum from the sodiated molecule at *m/z* 1148 yielded a simpler spectrum than the LE-CID analysis, with fewer fragment ion peaks being observed (figure 5.1.10). Whilst this should have made interpretation easier, there were some fragment ions present for which no plausible structure could be proposed. An extensive search of the literature failed to highlight any previous studies where MALDI-ToF/ToF had been used for the analysis of sodiated peptides by HE-CID tandem MS. Following the literature on LE-CID analysis of sodiated peptides, it was expected that much structural information would gbe obtained, possibly from the C-terminus (Grese *et al.*, 1989; Teesch and Adams, 1990; Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004); however, the lack of any particularly structurally useful ions observed here may perhaps be accounted for by the closed ring structure.

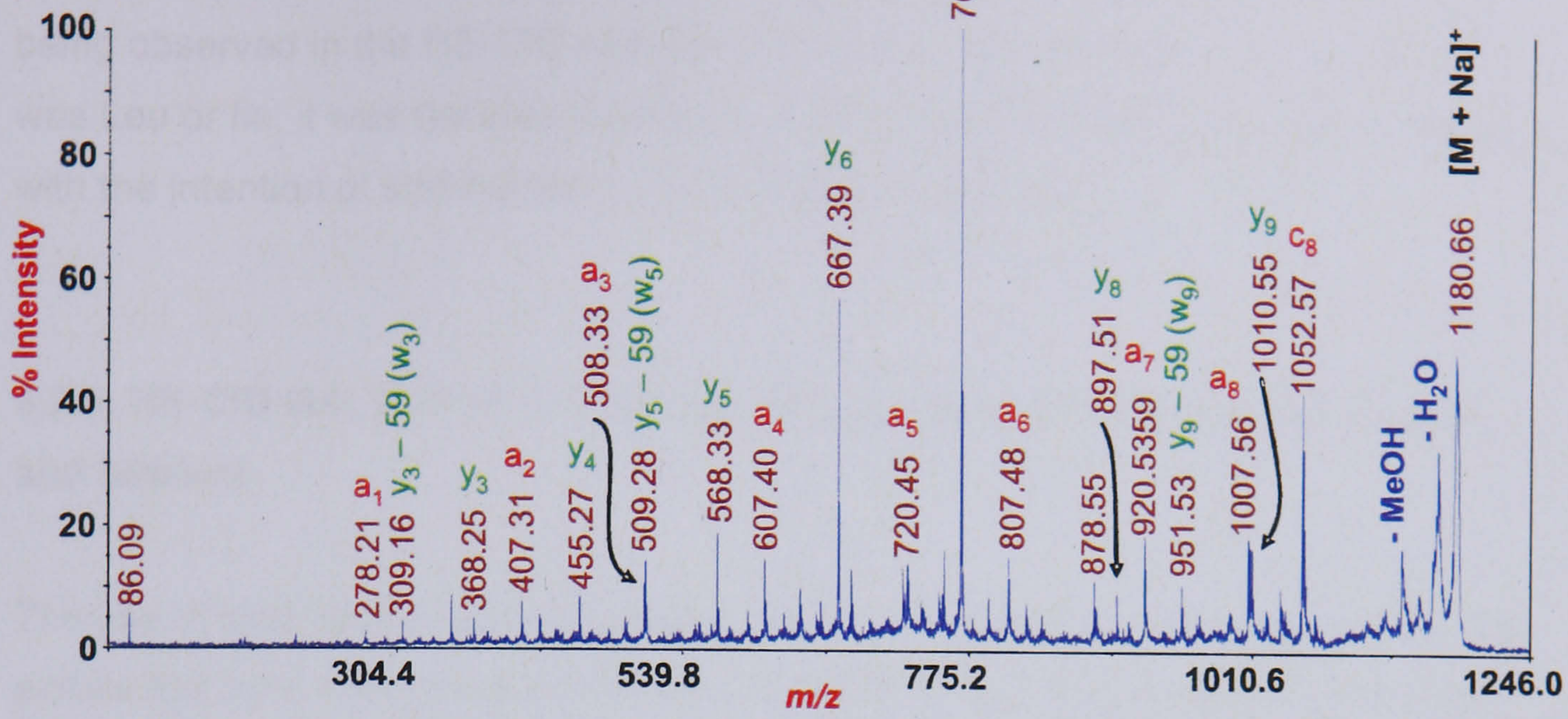


Figure 5.2.11. HE-CID tandem MS of sodiated CLP treated with base at *m/z* 1180.

Interestingly, the HE-CID tandem MS analysis of the sodiated CLP treated with base at *m/z* 1180 produced a very informative spectrum, with many C and N-terminal fragment ions being present across a broad range of *m/z* values (figure 5.2.11). HE-CID causes less intense ions arising by loss of water from AAs, therefore reducing the complexity of resulting product ion spectra. However, other so called 'satellite ions' formed from side chain cleavages can be observed; this can be of great benefit when differentiation between isobaric AAs Leu and Ile is sought.

A full series of fragment ions from the N-terminus (a_n series) were obtained, which when combined with the near complete y_n series (from the C-terminus), shows that the AA sequence corresponds to that proposed earlier (figure 5.2.5). The FA moiety has a mass of 171 Da, which is consistent with a saturated hydroxylated C_{10} FA chain (Gross *et al.*, 2007).

Perhaps even more surprising from a sodiated spectrum (given the literature), was the presence of three fragment ions that correspond to a loss of 59 Th from y_9, y_5 and y_3, which is consistent with the positioning of three Leu/Ile AAs in the proposed sequence (figure 5.2.5). A loss of 59 Th giving w_9, w_5 and w_3 fragment ions, is consistent with side chain losses from Leu. If the AAs present were Ile, then the losses from the y ions y_9, y_5 and y_3 would have been 15 and 41 Th (Johnson *et al.*, 1988).

Despite the demonstration that three out of the four possible Leu/Ile combinations were Leu, the remaining terminal AA was not identified due to no y_1 or w_1 peaks being observed in the HE-CID spectrum. In order to ascertain if the remaining AA was Leu or Ile, it was decided to attempt to add a larger group during ester cleavage, with the intention of shifting the y_1 ion to slightly higher *m/z*.

5.2.5. HE-CID MALDI-ToF/ToF tandem MS analysis of CLP treated with base and butanol

The use of mild base would be expected to cleave an ester bond, adding ammonia across the bond and giving an increase of 17 Th in the resulting MS (Kuiper *et al.*, 2004); in this case, however, the addition of MeOH was observed with an increase of 32 Th being evident. In order to attempt to obtain a y_1 ion upon HE-CID tandem MS analysis of a ring opened CLP, it was decided to add a larger group that would hopefully allow the generation a more stable y_1 fragment ion.

The CLP was treated overnight with mild base (35 % ammonium hydroxide), this time mixed with butanol (BuOH) instead of MeOH. An increase of 74 Th would be expected in the resulting MS, consistent with ring opening and the addition of BuOH across the ester bond.

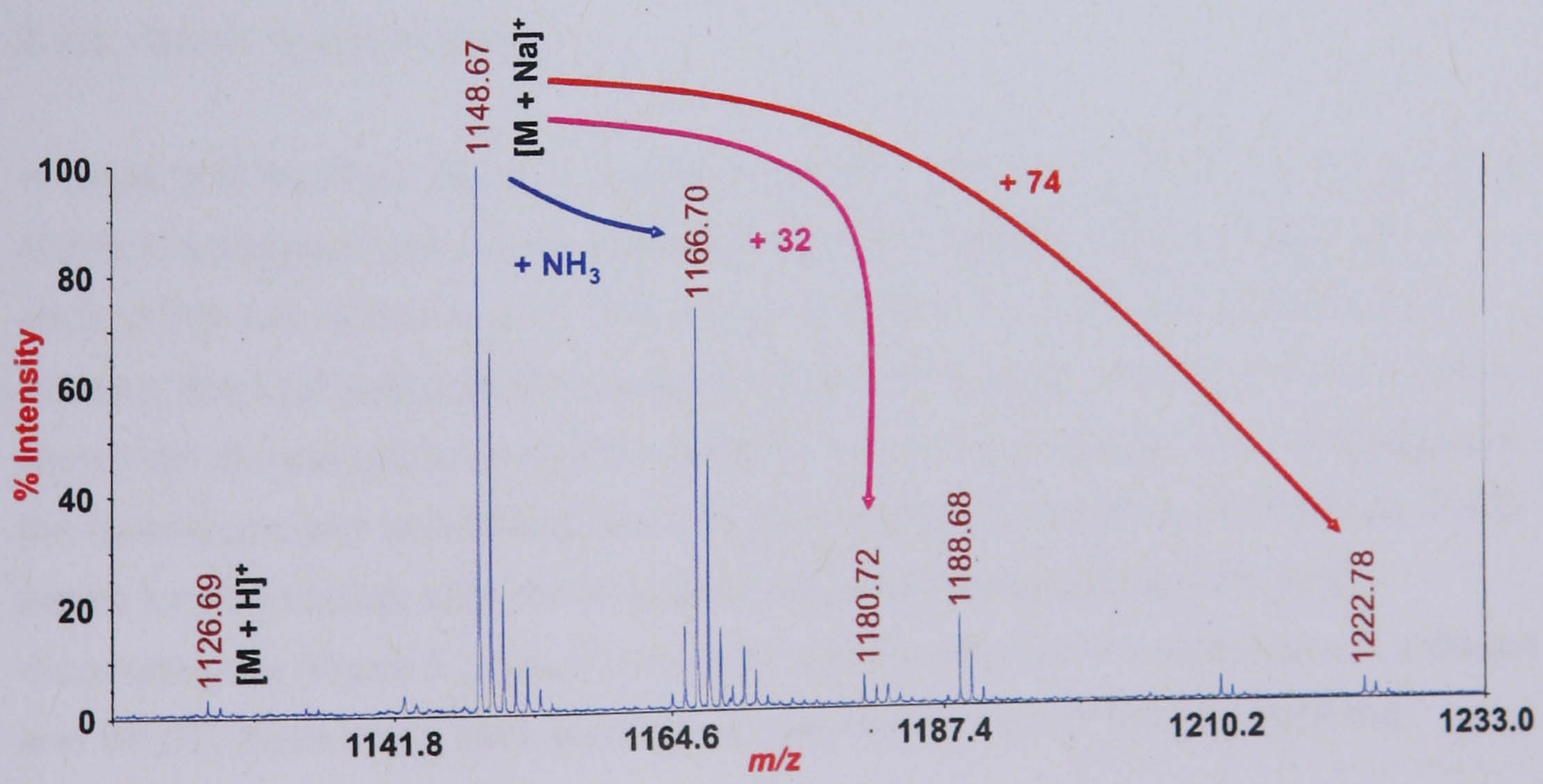


Figure 5.2.12. MALDI-ToF/ToF MS of CLP treated with mild base and butanol.

The resulting MALDI-MS of the CLP treated with mild base and BuOH exhibits more peaks than were expected. The most intense peak at *m/z* 1148 is consistent with sodiated CLP (the corresponding protonated peak at *m/z* 1126 has a very low intensity), with other sodiated molecules at 17, 32 and 74 Th higher, corresponding to the addition of ammonia, MeOH and BuOH respectively. The addition of BuOH gave rise to a low intensity peak at *m/z* 1222, suggesting that perhaps the bulky nature of the alkyl chain prevented the efficient addition of BuOH across the ester bond due to too much steric hindrance. The peak at *m/z* 1180 corresponds to the addition of MeOH across the ester bond; as only 5 µL of a methanolic solution of CLP had been added to 1 mL of mild base and BuOH mixture, it was thought that BuOH would react with much greater efficiency. The sodiated molecule at *m/z* 1166 again corresponds to the adduction of ammonia, which has a much higher intensity than for the previous ring cleavage using methanol (figure 5.2.7).

HE-CID tandem MS analysis of the three peaks at *m/z* 1166, 1180 and 1222 failed to produce any spectra with increased levels of fragmentation (not shown). As the species deriving from the addition of BuOH was least intense, the ion was too weak to allow a reasonable signal to noise ratio in the product ion spectrum, meaning that adding a larger group did not aid in obtaining the necessary y_1 ion needed for differentiation between Leu and Ile in position nine.

5.2.6. Amino acid analysis

In order to fit the final pieces of the jigsaw, it was decided to use amino acid analysis (AAA) to elucidate both the AA composition, and to determine the configuration of each of the AAs (either D or L). This requires that each of the AAs present be free in solution; the CLP was therefore hydrolysed to free the AAs. The free AAs in solution were then derivatised to allow their detection using fluorescence. The RP analysis of the hydrolysed and derivatised AAs using the method by Penkman (Penkman, 2005) lasted for 95 minutes, with seven substantial peaks present in the resulting chromatogram (figure 5.2.13). From the three reported CLPs (massetolide F, viscosin and WLIP), there could have been up to nine peaks present from the different possible AA combinations.

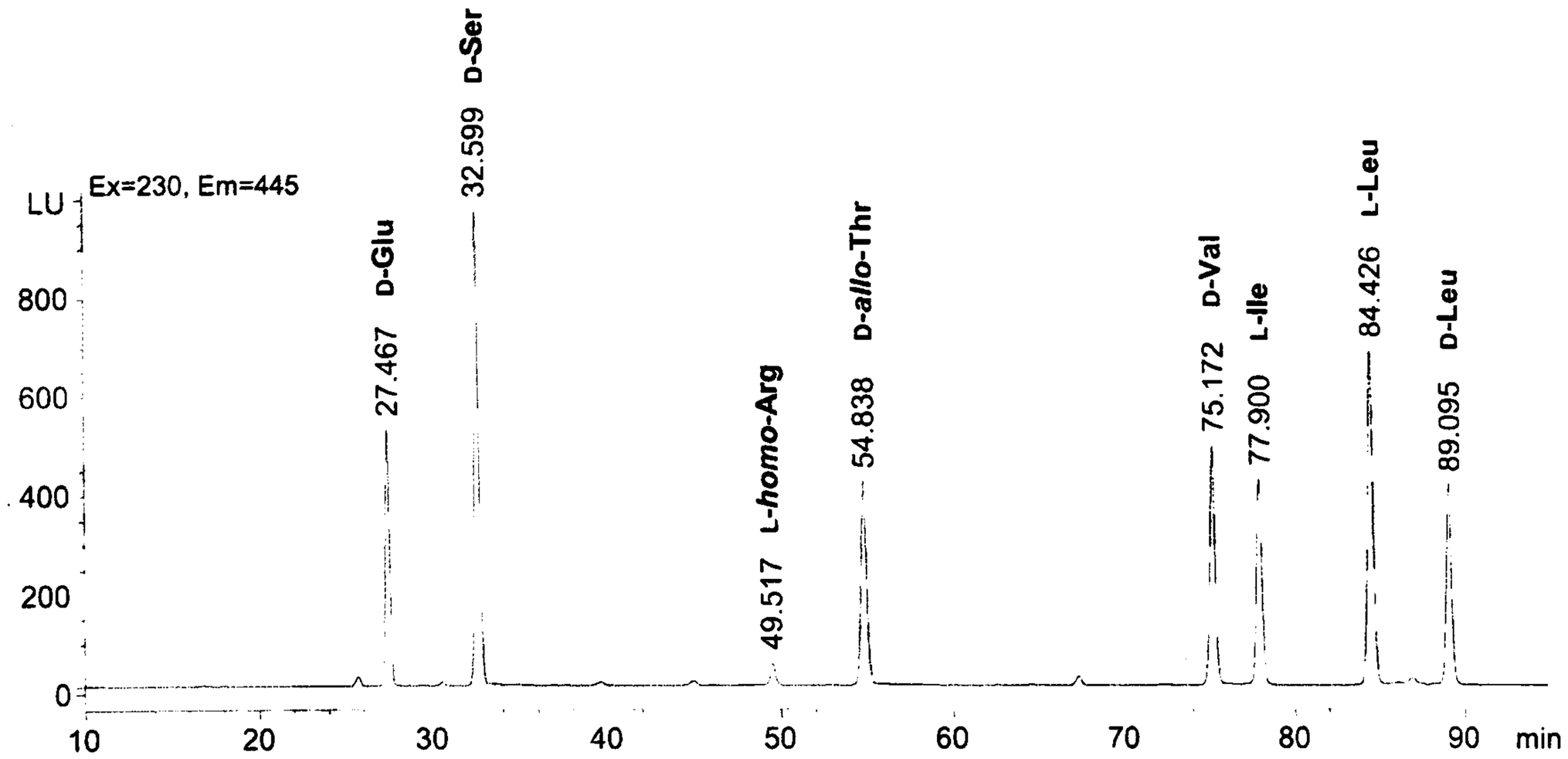


Figure 5.2.13. Resulting UV fluorescence chromatogram from the AAA of the CLP.

From the retention times of each of the seven substantial peaks, the following AAs are present within the CLP structure: D-Glu, D-Ser, D-*allo*-Thr, D-Val, L-Ile, L-Leu and D-Leu. The internal standard at 49.5 min was used to calculate the concentrations of each of the AAs present in order to evaluate the stoichiometry (figure 5.2.14).

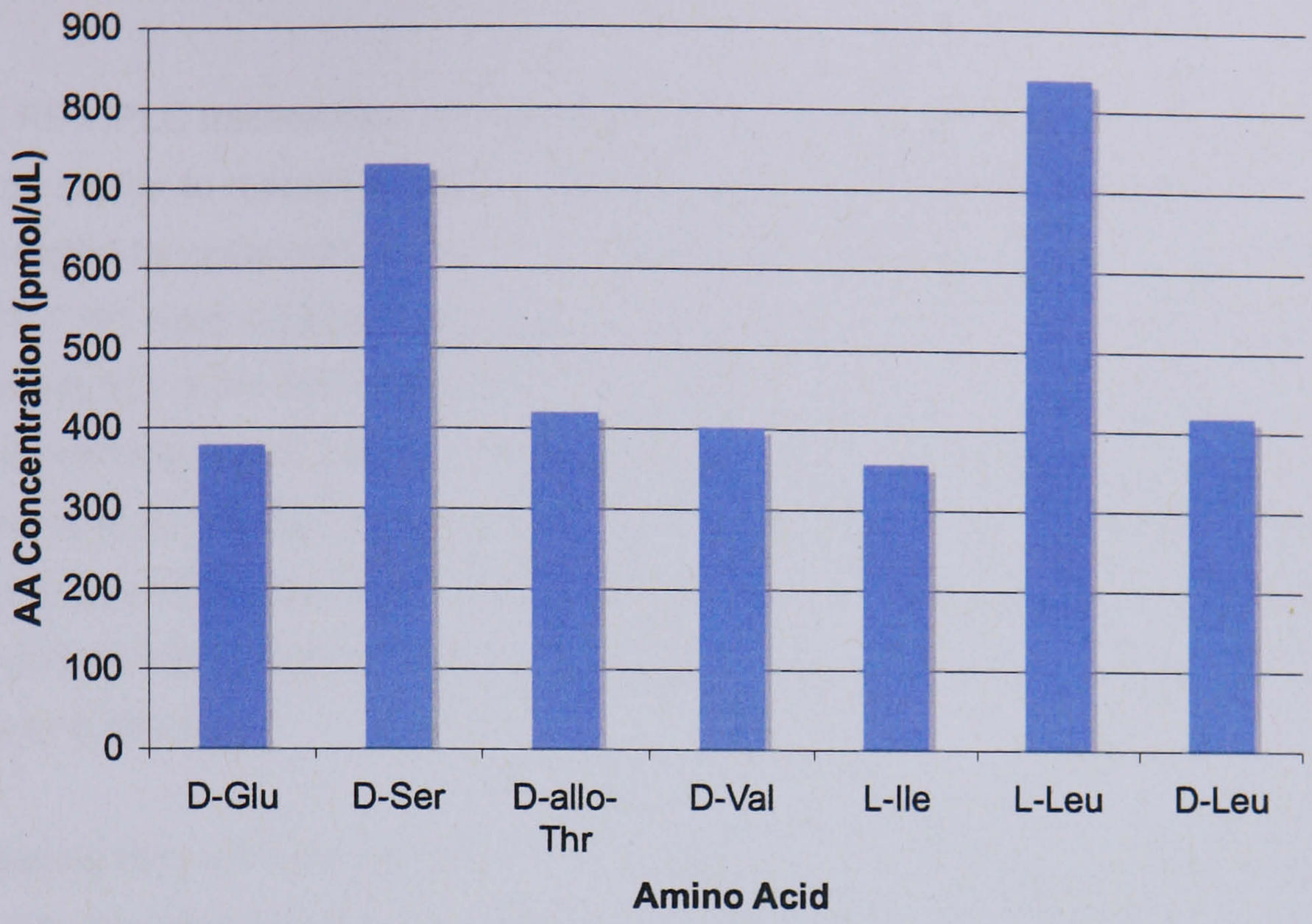


Figure 5.2.14. Graphical representation of the amount of each AA detected.

Despite there being some fluctuations between the absolute amounts of the AAs shown in figure 5.2.14, it is clear to see that two of the AAs (D-Ser and L-Leu) are present at higher concentrations than the other AAs. Dividing all of the amounts by the smallest, gives the ratio of AAs to be 1.1 : 2.0 : 1.2 : 1.1 : 1.0 : 2.3 : 1.2, which when rounded to the nearest whole number is 1 : 2 : 1 : 1 : 1 : 2 : 1 (totalling nine AAs), giving the stoichiometry to be one of each of the AAs D-Glu, D-*allo*-Thr, D-Val, L-Ile, and D-Leu with two of each of the AAs D-Ser and L-Leu; WLIP has the same AA composition as that determined here.

5.3. Conclusions

An RP-HPLC fraction containing a surfactant (initially proposed by our collaborators to be similar to massetolide C on basis of the preliminary mass) and an AFM extracted by collaborators from the culture supernatant of *Pseudomonas chlororaphis* PCL 1391 were supplied for analysis and subsequent structural identification. CID tandem MS of the AFM using ESI-Q-o-ToF was consistent with phenazine-1-carboxamide, which has previously been reported as produced by this strain (Chin-A-Woeng *et al.*, 1998). The identification of the surfactant was less straightforward than that of the AFM. The initial LE-CID tandem MS of the surfactant yielded substantial structural information consistent with a CLP, but was difficult to fully interpret due to the ring structure.

Treating the CLP with mild base to cleave the ester bond gave an addition of 32 Th across the ester bond, which was due to the addition of methanol (Rodrigues, 2005) rather than ammonia as expected (Kuiper *et al.*, 2004). LE-CID tandem MS analysis of the ring-opened lipopeptide gave a simpler spectrum with fewer fragment ions than the cyclic structure, so that structural information was incomplete.

HE-CID-MALDI-ToF/ToF allowed differentiation between Leu and Ile for three out of the four possible Leu or Ile assignments from the loss of 59 Th (w_n fragment ions), as well as providing a complete series of a_n ions and near-complete series of y_n ions. These results were in contrast to the literature on sodiated peptides (Grese *et al.*, 1989; Teesch and Adams, 1990; Lin *et al.*, 2001; Feng *et al.*, 2003; Newton and McLuckey, 2004; Bensadek *et al.*, 2007), where C-terminal fragment ions were reported to be the predominant species. To the author's knowledge, the observation of w_n fragment ions from sodiated precursor ions has not previously been described in the literature, perhaps due to the difficulty in obtaining sodiated peptides with a free amino terminus, which would preferentially protonate.

As the y_1 ion, which would have allowed the differentiation of the final Leu/Ile AA, was not observed (even after the addition of BuOH to increase the mass of the y_1 fragment ion), it was decided that AAA be used to attempt to complete the identification of the CLP produced by PCL 1391.

Table 5.3.1. Comparison of MS and AAA results.

	Detection Method		
Amino Acid	**AAA**	**MS**	
Glutamic acid	1	1	
Serine	2	2	
Threonine	1	1	
Valine	1	1	
Isoleucine	1	0	1*
Leucine	3	3	

** Differentiation between one AA (either Leu/Ile) could not be obtained.*

Table 5.3.1 shows the complementary results obtained by the MS and AAA methods. The use of AAA allowed the Leu/Ile in position nine to be identified as Ile, as well as the D and L configurations of the AAs in the CLP. Combining all of the information obtained from both analysis methods allows the structure of the CLP to be proposed (figure 5.3.1).

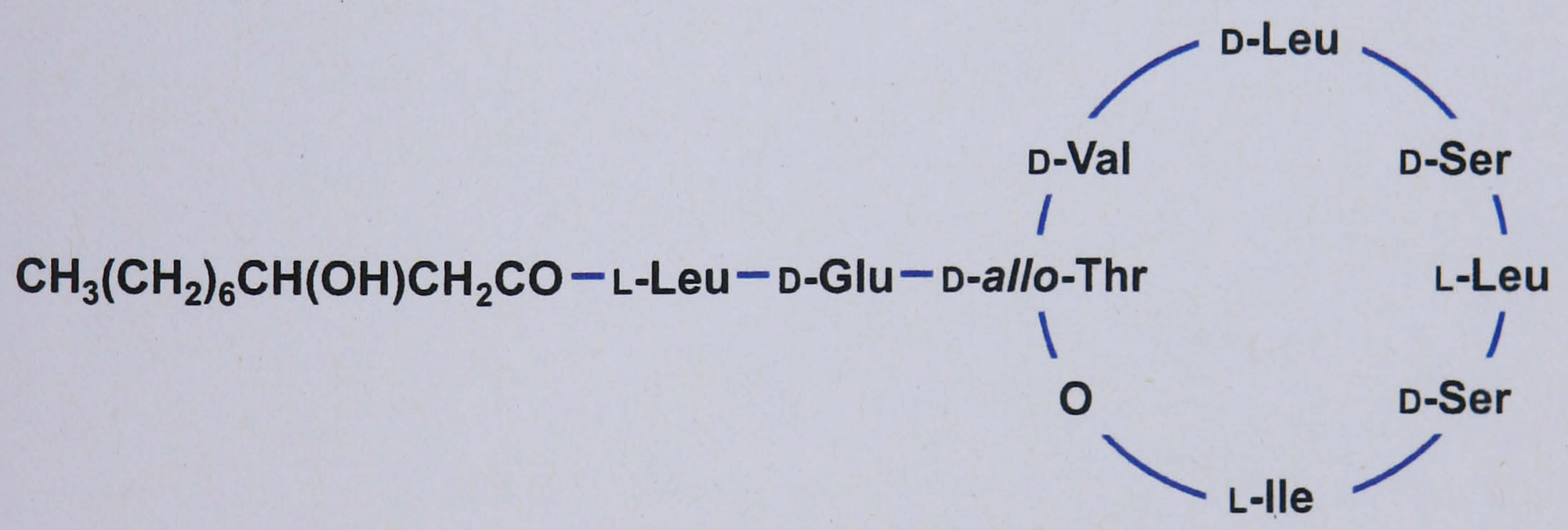


Figure 5.3.1. Structure of the CLP as determined by MS and AAA. The position of the hydroxy group was not determined, but is proposed to be a 3-hydroxylated FA by homology. (Coraiola *et al.*, 2006).

The structure as proposed from MS and AAA, corresponds to the structure of white line inducing principle (WLIP). WLIP has previously been reported by Coraiola *et al.* (Coraiola *et al.*, 2006) as being produced by *Pseudomonas chlororaphis*, having the ability to permeabilise membranes; something that is likely to be required if PCL 1391 is to counteract the effects of tomato foot and root rot caused by the fungal pathogen *Fusarium oxysporum f. sp. radicis-lycopersici*.

Chapter Six

Conclusions and proposals for future work

6.1. Conclusions

The primary focus of the work presented within this thesis was the development of a robust LC-MS 'metabonomic toolbox', where all current thoughts for metabonomic studies were brought together and critically compared and evaluated. A to date rarely-used HILIC separation technique was successfully applied for the metabonomic analysis of both cohorts of human urine samples (those obtained from volunteers within the Department of Chemistry, and those from patients who had suffered a bone fracture), showing that different compounds were detected when using this method compared to the commonly-employed RP separation approach. This was an expected result, and confirmed the fears that the sole use of RP separation for metabonomic studies of urine is likely to mean that a potentially large wealth of information has been missed and not detected. As more metabonomic studies utilise a HILIC separation method, it can only be hoped that the wider scientific community (not restricted to metabonomics) adopts this technology to increase the coverage obtained for the analysis of biofluids when using LC-MS analytical methodologies.

It was shown that concatenating LC-MS datasets (±RP-LC-MS and ±HILIC-MS data) together generated robust statistical models according to the external CV values, and by also comparing the individual PLS models to the concatenated PLS models. This is a method that shows great promise for future metabonomic studies, where LC-MS and NMR platforms are used (or even combined) to obtain a more representative picture of the compounds contained within a biofluid. The use of such a data analysis method may highlight potential metabolites, or groups of metabolites, that may fail to be found when only statistically analysing data from a single dataset (e.g. +RP-LC-MS); further to this, data concatenation also provides an idea of which variables are the most influential across all models, rather than just one. One problem that data fusion could encounter is the use of variables from different datasets to form a model when, in fact, they correspond to the same compound; the resulting variables should therefore be checked for any similarities in retention times and *m/z* values to identify any such variables.

Other conclusions from the work presented in Chapter Three are that the popular use of normalising to creatinine could potentially skew results, as the levels of excreted creatinine have been shown to be easily perturbed by many factors, such as illness. Comparing normalisation to creatinine (from healthy volunteers) to normalisation

using total ion count, showed that the two methods were comparable in as much as the most important variables for each discriminatory model developed were consistent. Overall, the work presented in Chapter Three has shown that the field of metabonomics is still very much in its infancy and lacks a common theme whereby samples are carefully obtained, analysed and reported; this is something which *will* hopefully come with maturity.

The use of the developed metabonomic toolbox for the analysis of clinical urine samples to profile the body's response to bone fractures (with the goal of identifying potential biomarkers of delayed or failed fracture healing) did develop discriminatory models, which highlighted potential biomarkers that could be related to the fracture healing process. However, the study was flawed from the offset by the lack of time-setted samples, and the fact that none of the patients suffered failed fracture healing. Drawing together the conclusions from Chapters Three and Four, as well as the retrospective views, highlights that careful planning is required before samples are collected to ensure that any subsequent analysis stands the greatest possible chance of being a success.

Perhaps the most interesting piece of work from Chapter Four stemmed from the sample issues, which was found to correspond to very high levels of protein within the clinical urine samples. The protein corresponded to a broad range of molecular weight and function. The larger proteins corresponded were related to the body's immune system, which suggests that the presence of the large amounts of protein found in the clinical urine samples may have been caused by some kind of immunological response to either stress or resulting treatment. This is certainly an area where more research needs to be undertaken in an attempt to understand why so much protein was detected.

Both Chapters Three and Four showed that the most common problem of metabonomic analyses is the subsequent identification of potential biomarkers highlighted from statistical studies. This 'poor' biomarker identification could have been resolved by the combination of higher mass accuracy instrumentation and NMR after the fractionation and purification of potential biomarkers. For biomarkers that were proposed, the analysis of standards using LC-MS/MS would have confirmed if the proposed compounds were correct, allowing further justification to be sought by relating the identified biomarker to the biological process in question.

An additional, piece of work is also presented in this thesis. An ethyl acetate extract (provided by the Department of Biology, University of Leiden, the Netherlands) of *Pseudomonas chlororaphis* PCL 1391 spent growth medium, containing an unknown biosurfactant, was successfully analysed using ESI-Q-o-ToF MS(MS), MALDI-ToF/ToF MS(MS) and racemic amino acid analysis. The in-depth study elucidated both the amino acid sequence and the stereochemistry of the amino acids present within the cyclic lipopeptide. The high energy MALDI-ToF/ToF CID tandem MS analysis of the sodiated cyclic lipopeptide molecule generated characteristic w_n fragments, which allowed the differentiation between the isomeric amino acids Leu and Ile; the use of sodium cationised non-tryptic peptides has seldom been reported in the literature, as sodiated tryptic peptides are commonly considered as a nuisance rather than a benefit. All of the analytical methods together allowed the structure of the cyclic lipopeptide to be postulated as corresponding to white line inducing principle, WLIP.

6.2. Future work

Future work that could stem from the work presented in this thesis is as follows:

- An in-depth study into normalisation methods: is the use of normalising to total ion count acceptable? Are there any other endogenous metabolites that could be used which are not perturbed by illness or other factors?

- The statistical models presented within Chapters Three and Four should be studied in more depth. The sole use of external classification results does not give an in-depth view of the data; the number of variables that were used to develop each model should be assessed, along with their relative weightings. A further way of probing the variables detected is to quantify which variables are most persistent, not only for each model, but across all models.

- The refinement and development of data fusion techniques for the analysis of LC-MS data, and also for the inclusion of complementary NMR data is needed. However, LC-MS datasets cannot just be concatenated to NMR data as the data is different. Hierarchical analysis works by first generating PLS models for data which is the same, so there would be two different models for LC-MS and NMR data for example, and then compares the scores (linear combinations of the original data) from each developed model by PLS. This allows the comparison of two different types of data and should allow more in-depth metabonomic studies to be achieved where all of the data obtained from multiple platforms can be analysed both separately and together, hopefully giving a global metabolite profile.

- The use of higher mass accuracy MS/MS and also NMR to elucidate the most important variables as produced by each of the developed models from both Chapters Three and Four.

- The unexpectedly high levels of protein detected in the clinical urine samples should be investigated further. The data presented in Chapter Four should first be linked back to the original patient files to see if there is anything that can be correlated to an increase in the excretion of protein. Further simple tests could be carried out: for fracture patients admitted to accident and emergency, their levels of protein could be detected using a simple dip-stick

method. This would allow a comparison of the levels found within this study to other samples; it would be very interesting to see if there is a correlation, or if the clinical samples were unique in their protein concentrations.

- The study of urine samples from patients suffering a bone fracture should be re-designed as outlined in the conclusions/retrospective work from Chapter Four to obtain biofluids from patients at multiple, regularly spaced time points with the inclusion of more patients who suffered delayed fracture healing. It would also be interesting to analyse serum samples using metabonomics to see if the same conclusions can be drawn.

- The work presented in this thesis has prompted a joint study with Hull University into the metabonomic study of properly time-setted urine, serum, plasma, saliva and cancer cells from patients at various stages of colonic cancer.

References

Al-Dehaimi, A. W., A. Blumsohn and R. Eastell (1999). *Clinical Chemistry* **45**(5): 676-681.

Albert Koulman. (2007). *Rapid Communications in Mass Spectrometry* **21**(3): 421-428.

Alpert, A. J. (1990). *Journal of Chromatography A* **499**: 177.

Anderson, H., O. N. Jensen, E. P. Moiseeva and E. F. Eriksen (2003). *Journal of Bone Mineral Research* **18**(2): 195-203.

Antti, H., T. M. D. Ebbels, H. C. Keun, M. E. Bollard, O. Beckonert, J. C. Lindon, J. K. Nicholson and E. Holmes (2004). *Chemometrics Intelligent Laboratory Systems* (73): 139-149.

Asaba, Y., K. Hiramatsu, Y. Matsui, A. Harada, Y. Nimura, N. Katagiri, T. Kobayashi, T. Takewaka, M. Ito, S. Niida and K. Ikeda (2006). *Bone* **39**(6): 1276-1282.

Ault, J. (2007). *PhD. Thesis (University of York).*

Bajad, S., M. Coumar, R. Khajuria, O. P. Suri and K. L. Bedi (2003). *European Journal of Pharmaceutical Sciences* **19**(5): 413-421.

Bajad, S. and V. Shulaev (2007). *Trends in Analytical Chemistry* **26**(6): 625-636.

Bakhtiar, R., L. Ramos and L. Tse (2002). *Journal of Liquid Chromatography and Related Technololgies* **25**(4): 507-540.

Bales, J. R., D. P. Higham, I. Howe, J. K. Nicholson and P. J. Sadler (1984). *Clinical Chemistry* **30**(3): 426-432.

Ballard, K. D. and S. J. Gaskell (1993). *Journal of the American Society for Mass Spectrometry* **4**(6): 477-481.

Balogh, M. P. (2004). *Spectroscopy* **19**(10): 44-52.

Bandu, M. L., T. Grubbs, M. Kater and H. Desaire (2006). *International Journal of Mass Spectrometry* **251**(1): 40-46.

Barratt, J. and P. Topham (2007). *Canadian Medical Association Journal* **177**(4): 361-368.

Beattie, I., K. Joncour and K. Lawson (2005). *Separation Science Redefined*: 22-30.

Becker, J. M., G. A. Caldwell and E. A. Zachgo (1996). *Biotechnology (Second Edition).* San Diego, Academic Press: 119-124.

Behnam, K., S. S. Murray, J. P. Whitelegge and E. J. Brochmann (2002). *Journal of Orthopaedic Research* (20): 1190-1196.

Bensadek, D., F. Monigatti, J. A. J. Steen and H. Steen (2007). *International Journal of Mass Spectrometry* **268**(2-3): 181-189.

Berkel, G. J. v., K. G. Asano and P. D. Schnier (2001). *Journal of the American Society for Mass Spectrometry*: 853-862.

Bertram, H. C., C. Hoppe, B. O. Petersen, J. Duus, C. M. Igaard and K. F. Michaelsen (2007). *British Journal of Nutrition* **97**(04): 758-767.

Bischoff, R. and B. Barroso (2002). *Recent Applications in LC-MS*: 2-8.

Bischoff-Ferrari, H. A. and B. Dawson-Hughes (2007). *Bone* **41**(1, Supplement 1): S13-19.

Bleakney, W. (1929). *Physical Review* **34**: 157-160.

Bock, K., C. Kohle, S. Helmut and L. Packer (2005). *Methods in Enzymology*, Academic Press. **Volume 400**: 57-75.

Boeniger, M. F., L. K. Lowry and J. Rosenberg (1993). *American Industrial Hygiene Association Journal* **54**(10): 615-627.

Bollard, M. E., E. Holmes, J. C. Lindon, S. C. Mitchell, D. Branstetter, W. Zhang and J. K. Nicholson (2001). *Analytical Biochemistry* **295**(2): 194-202.

Bollard, M. E., H. C. Keun, O. Beckonert, T. M. D. Ebbels, H. Antti, A. W. Nicholls, J. P. Shockcor, G. H. Cantor, G. Stevens and J. C. Lindon (2005). *Toxicology and Applied Pharmacology* **204**(2): 135-151.

Bonfiglio, R., R. C. King, T. V. Olah and K. Merkle (1999). *Rapid Communications in Mass Spectrometry* **13**: 1175-1185.

Borth, W. (1992). *Federation of American Societies for Experimental Biology Journal.* **6**(15): 3345-3353.

Bradford, M. M. (1976). *Analytical Biochemistry* **72**(1-2): 248-254.

Brinkman, J. W., D. d. Zeeuw, J. J. Duker, R. T. Gansevoort, I. P. Kema, H. L. Hillege, P. E. d. Jong and S. J. L. Bakker (2005). *Clinical Chemistry* **51**(11): 2181-2182.

Broadhurst, D. and D. Kell (2006). *Metabolomics* **2**(4): 171-196.

Brown, M., W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasic and D. B. Kell (2005). *Metabolomics* **1**(1): 39-51.

Bruce, S. J., P. Jonsson, H. Antti, O. Cloarec, J. Trygg, S. L. Marklund and T. Moritz (2007). *Analytical Biochemistry* **372**(2): 237-249.

Bruckner, H., S. Haasmann, M. Langer, T. Westhauser, R. Wittner and H. Godel (1994). *Journal of Chromatography A* **666**(1-2): 259-273.

Burton, K. I., J. R. Everett, M. J. Newman, F. S. Pullen, D. S. Richards and A. G. Swanson (1997). *Journal of Pharmaceutical and Biomedical Analysis* **15**(12): 1903-1912.

Byles, M., M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes and J. Trygg (2006). *Journal of Chemometrics* **20**(8-10): 341-351.

Byrd, G. D. and M. W. Ogden (2003). *Journal of Mass Spectrometry*(38): 98-107.

Calvo, M. S., D. R. Eyre and C. M. Gundberg (1996). *Endocrine Reviews*(17): 333-368.

Castro-Perez, J. M (2007). *Drug Discovery Today* **12**(5-6): 249-256.

Chambers, E., D. M. Wagrowski-Diehl, Z. Lu and J. R. Mazzeo (2007). *Journal of Chromatography B* **852**(1-2): 22-34.

Chang, D., C. D. Banack and S. L. Shah (2007). *Journal of Magnetic Resonance* **187**(2): 288-292.

Chapurlat, R. D., P. Garnero, G. Brart, P. J. Meunier and P. D. Delmas (2000). *Bone* **27**(2): 283-286.

Chen, C., F. J. Gonzalez and J. R. Idle (2007). *Drug Metabolism Reviews* **39**(2): 581 - 597.

Chen, H. s., T. Rejtar, V. Andreev, E. Moskovets and B. L. Karger (2005). *Analytical Chemistry* **77**(8): 2323-2331.

Chen, M. and R. Hofestadt (2005). *J Biomedical Informatics* **38**(3): 173-175.

Chin-A-Woeng, T. F. C., G. V. Bloemberg, I. H. M. Mulders, L. C. Dekkers and B. J. J. Lugtenberg (2000). *Molecular Plant-Microbe Interactions* **13**(12): 1340-1345.

Chin-A-Woeng, T. F. C., G. V. Bloemberg, A. J. van der Bij, K. M. G. M. van der Drift, J. Schripsema, B. Kroon, R. J. Scheffer, C. Keel, P. A. H. M. Bakker, H.-V. Tichy, F. J. de Bruijn, J. E. Thomas-Oates and B. J. J. Lugtenberg (1998). *Molecular Plant-Microbe Interactions* **11**(11): 1069-1077.

Chrambach, A., R. A. Reisfeld, M. Wyckoff and J. Zaccari (1967). *Analytical Biochemistry* **20**(1): 150-154.

Christian, M. T. and A. R. Watson (2004). *Current Paediatrics* **14**(7): 547-555.

Churchwell, M. I., N. C. Twaddle, L. R. Meeker and D. R. Doerge (2005). *Journal of Chromatography B* **825**(2): 134-143.

Churms, S. C. (1996). *Journal of Chromatography A*(720): 75-91.

Compton, S. J. and C. G. Jones (1985). *Analytical Biochemistry* **151**(2): 369-374.

Constantinou, M. A., E. Papakonstantinou, M. Spraul, S. Sevastiadou, C. Costalos, M. A. Koupparis, K. Shulpis, A. Tsantili-Kakoulidou and E. Mikros (2005). *Analytica Chimica Acta* **542**(2): 169-177.

Constanzer, M. L., C. M. Chavez-Eng, I. Fu, E. J. Woolf and B. K. Matuszewski (2004). *Journal of Chromatography B.* **807**(2): 243-250.

Coraiola, M., P. Lo Cantore, S. Lazzaroni, A. Evidente, N. S. Iacobellis and M. Dalla Serra (2006). *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1758**(11): 1713-1722.

Cordero, M. M., J. J. Houser and C. Wesdemiotis (1993). *Analytical Chemistry* **65**(11): 1594-1601.

Creaser, C. S. and J. W. Stygall (1998). *Trends Analytical Chemistry* **17**(10): 583-592.

Crockford, D. J., J. C. Lindon, O. Cloarec, R. S. Plumb, S. J. Bruce, S. Zirah, P. Rainville, C. L. Stumpf, K. Johnson, E. Holmes and J. K. Nicholson (2006). *Analytical Chemistry* **78**: 4398-4408.

Davis, R. A., A. J. Charlton, J. Godward, S. A. Jones, M. Harrison and J. C. Wilson (2007). *Chemometrics and Intelligent Laboratory Systems* **85**(1): 144-154.

Dawson, P. H. and C. Lambert (1975). *International Journal of Mass Spectrometry and Ion Physics* (16): 269-280.

de Bruijn, I., M. J. D. de Kock, M. Yang, P. de Waard, T. A. van Beek and J. M. Raaijmakers (2007). *Molecular Microbiology* **63**(2): 417-428.

DeGaudio, A., R. Spina, A. DiFilippo and M. Feri (1999). *Critical Care in Medicine* **27**(10): 2105-2108.

DeLaurier, A., B. Jackson, D. Pfeiffer, K. Ingham, M. A. Horton and J. S. Price (2004). *Research in Vetinary Science*(77): 29-39.

Dempster, A. J. (1921). *Physical Review* **18**(6): 415-422.

Desai, J. D. and I. M. Banat (1997). *Microbiology and Molecular Biology Reviews* **61**(1): 47-64.

Dettmer, K., P. A. Aronov and B. D. Hammock (2007). *Mass Spectrometry Reviews* **26**(1): 51-78.

Dihazi, H. and G. A. Muller (2007). *Expert Review of Proteomics* **4**(1): 39-50.

Dimitriou, R., E. Tsiridis and P. V. Giannoudis (2005). *Injury* **36**(12): 1392-1404.

Doblare, M., J. M. Garcia and M. J. Gomez (2004). *Engineering and Fracture Mechanics*(71): 1809-1840.

Dole, M., L. L. Mach, R. L. Hines, R. C. Mobley, L. D. Ferguson and M. B. Alice (1968). *Journal of Chemical Physics*(49): 2240-2247.

Drexler, D. M., J. H. M. Feyen and M. Sanders (2004). *Drug Discov Today* **1**(1): 17-23.

Dunckley, T., K. D. Coon and D. A. Stephan (2005). *Drug Discovery Today*. **10:** 326-334.

Dunn, W. B. and D. I. Ellis (2005). *Trends Analytical Chemistry*(24): 285-294.

Ebeling, P. R. and K. Åkesson (2001). *Best Practices in Research in Clinical Rheumatology* **15**(3): 385-400.

Edwards, E. (2007). *PhD. Thesis (University of York).*

Einhorn, T. A. (2005). *Journal of Orthopaedic Trauma* **19**(10): 4-6.

El-Faramawy, A., K. W. M. Siu and B. A. Thomson (2005). *Journal of the American Society for Mass Spectrometry*(16): 1702-1707.

Eyre, D. R. (1996). Biochemistryical Basis of Collagen Metabolites and Bone Turnover Markers. *Principles of Bone Biology*, Academic Press.

Felitsyn, N. M., G. N. Henderson, M. O. James and P. W. Stacpoole (2004). *Clinica Chimica Acta*(350): 219-230.

Feng, W. Y., S. Gronert, K. A. Fletcher, A. Warres and C. B. Lebrilla (2003). *International Journal of Mass Spectrometry* **222**(1-3): 117-134.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). *Science*(246): 64-71.

Fiehn, O., J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey and L. Willmitzer (2000). *Nature - Biotechnology*(18): 1157-1161.

Fiehn, O., D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L. Sumner, R. Goodacre, N. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J. Lindon and S.-A. Sansone (2007). *Metabolomics* **3**(3): 175-178.

Finch, J. L., A. J. Brown and E. Slatopolsky (1999). *Journal of the American Society for Nephrology*(10): 980-985.
Finch, J. L., A. S. Dusso, T. Pavlopoulos and E. A. Slatopolsky (2001). *Journal of the American Society for Nephrology*(12): 1468-1474.
Fisher, M. C., C. Meyer, G. Garber and C. N. Dealy (2005). *Bone* **37**: 741-750.
Fledelius, C., A. H. Johnsen, P. A. C. Cloos, M. Bonde and P. Qvist (1997). *Journal of Biological Chemistry* **272**(15): 9755-9763.
Fligge, T. A., K. Bruns and M. Przybylski (1998). *Journal of Chromatography B*(706): 91-100.
Fong, K. W. Y. and T. W. D. Chan (1999). *Journal of the American Society for Mass Spectrometry*(10): 72-75.
Forbes, M. W., M. Sharifi, T. Croley, Z. Lausevic and R. E. March (1999). *Journal of Mass Spectrometry*(34): 1219-1239.
Forina, M., S. Lanteri and M. Casale (2007). *Journal of Chromatography A* **1158**(1-2): 61-93.
Forshed, J., H. Idborg and S. P. Jacobsson (2007a). *Chemometrics and Intelligent Laboratory Systems* **85**(1): 102-109.
Forshed, J., R. Stolt, H. Idborg and S. P. Jacobsson (2007b). *Chemometrics and Intelligent Laboratory Systems* **85**(2): 179-185.
Fridman, E. and E. Pichersky (2005). *Current Opinions in Chemical Biology*(8): 242-248.
Fura, A., T. W. Harper, H. Zhang, L. Fung and W. C. Shy (2003). *Journal of Pharmacological and Biomedical Analysis*(32): 513-522.
Garnero, P. and P. D. Delmas (2003). *Bone* **32**(1): 20-26.
Gavaghan, C. L., J. K. Nicholson, S. C. Connor, I. D. Wilson, B. Wright and E. Holmes (2001). *Analytical Chemistry* **291**(2): 245-252.
Geeraerts, F., L. Schimpfessel and R. Crokaert (1978). *Journal of Chromatography A*(145): 63-71.
George, S. K., M. T. Dipu, U. R. Mehra, P. Singh, A. K. Verma and J. S. Ramgaokar (2006). *Journal of Chromatography B* **832**: 134-137.
Gika, H. G., G. A. Theodoridis, J. E. Wingate and I. D. Wilson (2007). *Journal of Proteome Research***6**(8): 3291-3303.
Gill, S., M. Pop, R. DeBoy, P. Eckburg, PTurnbaugh, B. Samuel, J. Gordon, D. Relman, C. Fraser-Liggett and K. Nelson (2006). *Science* **312**(5778): 1355-1359.
Gonzalez-Buitrago, J. M., L. Ferreira and I. Lorenzo (2007). *Clinica Chimica Acta* **375**(1-2): 49-56.
Goodacre, R. (2007). *Journal of Nutrition* **137**(1): 259S-266.
Gottfries, J., M. Sjogren, B. Holmberg, L. Rosengren, P. Davidsson and K. Blennow (2004). *Chemometrics and Intelligent Lab Systems*(73): 47-53.
Grazioli, V., E. Casari, M. Murone and P. A. Bonini (1993). *Journal of Chromatography A*(615): 59-66.
Grese, R. P., R. L. Cerny and M. L. Gross (1989). *Journal of the American Chemical Society***111**(8): 2835-2842.
Griffin, J. L. (2006). *Philisophical Transactions B***361**: 147-161.
Griffin, J. L. and M. E. Bollard (2004). *Current Drug Metabolism* **5**(5): 389-398.
Gritti, F. and G. Guiochon (2006). *Journal of Chromatography A* **1136**(2): 192-201.
Gross, H., V. O. Stockwell, M. D. Henkels, B. Nowak-Thompson, J. E. Loper and W. H. Gerwick (2007). *Chemistry & Biology* **14**(1): 53-63.
Gu, H., H. Chen, Z. Pan, A. U. Jackson, N. Talaty, B. Xi, C. Kissinger, C. Duda, D. Mann, D. Raftery and R. G. Cooks (2007). *Analytical Chemistry* **79**(1): 89-97.
Gullberg, J., P. Jonsson, A. Nordstrom, M. Sjostrom and T. Moritz (2004). *Analytical Biochemistry* **331**(2): 283-295.
Guneral, F. and C. Bachmann (1994). *Clinical Chemistry* **40**(6): 862-866.

Haas, D. and G. Defago (2005). *Nature - Review of Microbiology***3**: 307-319.
Hamer, I., J. P. Paccaud, D. Belin, C. Maeder and J. L. Carpentier (1998). *Biochemistry. J.* **329**(1): 183-190.
Handl, J., J. Knowles and D. B. Kell (2005). *Bioinformatics* **21**(15): 3201-3212.
Heavner, D. L., W. T. Morgan, S. B. Sears, J. D. Richardson, G. D. Byrd and M. W. Ogden (2006). *Journal of Pharmaceutical and Biomedical Analysis* **40**(4): 928-942.
Heer, M., N. Baecker, C. Mika, A. Boese and R. Gerzer (2005). *Acta Astronautica* ***56****(9-12): 801-808.*
Heerma, W., J. Boon, C. Versluis, J. Kruijtzer, L. Hofmeyer and R. Liskamp (1997). *Journal of Mass Spectrometry* **32**(7): 697-704.
Heerma, W. and W. Kulik (1988). *Biological Mass Spectrometry* **16**(1-12): 155-159.
Heldon, S., J. Cals, F. Kessels, P. Brink, G. J. Dinant and P. Geusens (2006). *Osteoporosis International***17**: 348-354.
Hemström, P. and K. Irgum (2006). *Journal of Separation Science* **29**(12): 1784-1821.
Henle, P., G. Zimmermann and S. Weiss (2005). *Bone* **37**(6): 791-798.
Herrmann, M., D. Klitscher, T. Georg, J. Frank, I. Marzi and W. Herrmann (2002). *Clinical Chemistry* **48**(12): 2263-2266.
Hilliard, L. M., T. M. Osicka, S. P. Clavant, P. J. Robinson, D. J. Nikolic-Paterson and W. D. Comper (2006). *Journal of Laboratory and Clinical Medicine* **147**(1): 36-44.
Hodson, M. P., G. J. Dear, A. D. Roberts, C. L. Haylock, R. J. Ball, R. S. Plumb, C. L. Stumpf, J. L. Griffin and J. N. Haselden (2007). *Analytical Biochemistry* **362**(2): 182-192.
Hofmann, U., M. Schwab, S. Seefried, C. Marx, U. M. Zanger, M. Eichelbaum and T. E. Murdter (2003). *Journal of Chromatography B* **791**(1-2): 371-380.
Hogendoorn, E., P. v. Zoonen and F. Hernandez (2003). *Recent Applications in Multidimensional Chromatography*: 2-9.
Hopfgartner, G. and E. Varesio (2005). *Trends Analytical Chemistry* **24**(7): 583-589.
Hotelling, H. (1933). *Journal of Educational Psychology* **24**: 417-441.
Hu, J., K. R. Coombes, J. S. Morris and K. A. Baggerly (2005). *Brief in Functional Genomics and Proteomics* **3**(4): 322-331.
Huebner, J. L. and V. B. Kraus (2006). *Osteoarthritis and Cartilage* **14**: 923-930.
Hulme, A. N., H. McNab, D. A. Peggie and A. Quye (2005). *Phytochemistry* **66**(23): 2766.
Husková, R., P. Chrastina, T. Adam and P. Schneiderka (2004). *Clinica Chimica Acta*(350): 99-106.
Iadarola, P., G. Cetta, M. Luisetti, L. Annovazzi, B. Casado, J. Baraniuk, C. Zanone and S. Vigilo (2005). *Electrophoresis*(26): 1-15.
Idborg, H., P. Edlund and S. P. Jacobsson (2004). *Rapid Communications in Mass Spectrometry* **18**: 944-954.
Idborg, H., L. Zamani, P. Edlund, I. Schuppe-Koistinen and S. Jacobsson (2005a). *Journal of Chromatography B* **828**(1-2): 9-13.
Idborg, H., L. Zamani, P. Edlund, I. Schuppe-Koistinen and S. P. Jacobsson (2005b). *Journal of Chromatography B*(828): 14-20.
Igarashi, A. and M. Yamaguchi (2001). *International Journal of Molecular Medicine*(8): 433-438.
Igarashi, A. and M. Yamaguchi (2002). *International Journal of Molecular Medicine*(9): 503-508.
Ikegami, T. and N. Tanaka (2004). *Current Opinions in Chemical Biology*(8): 527-533.
Iribarne, J. V. and B. A. Thomson (1976). *Journal of Chemical Physics*(64): 2287-2294.

Ishizuka, N., H. Kobayashi, H. Minakuchi, K. Nakanishi, K. Hirao, K. Hosoya, T. Ikegami and K. Tanaka (2002). *Journal of Chromatography A* **960**: 85-96.
Izquierdo, P., M. Roses and E. Bosch (2006). *Journal of Chromatography A* **1107**: 96-103.
Jennings, K. R. (1968). *International Journal or Mass Spectrometry and Ion Physics* **1**(3): 227-235.
Jiang, C. and L. Luo (2004). *Analytica Chimica Acta* **506**(2): 171-175.
Jiang, H., J. Jiang, P. Hu and Y. Hu (2002). *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **769**(1): 169-176.
Johnson, R. S., S. A. Martin and K. Biemann (1988). *International Journal of Mass Spectrometry and Ion Processes* **86**: 137-154.
Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson (1987). *Analytical Chemistry* **59**(21): 2621-2625.
Jokiranta, T. S., A. Solomon, M. K. Pangburn, P. F. Zipfel and S. Meri (1999). *Journal of Immunology* **163**(8): 4590-4596.
Jones, A. W. and L. Karlsson (2005). *Human & Experimental Toxicology* **24**: 615-622.
Jones, C. B. (2005). *Journal of Orthopaedic Trauma* **19**(10): 1-3.
Jonscher, K. R. and J. R. Yates (1997). *Analytical Biochemistry*(244): 1-15.
Jordan, A. (1994). *Contraception* **49**(3): 189-201.
Juraschek, R., T. Dulcks and M. Karas (1999). *Journal of the American Society for Mass Spectrometry*(10): 300-308.
Kakonen, S., J. Hellman, M. Karp, P. Laaksonen, K. J. Obrant, H. K. Vaananen, T. Lovgren and K. Pettersson (2000). *Clinical Chemistry* **46**(3): 332-337.
Karas, M., D. Bachmann, U. Bahr and F. Hillenkamp (1987). *International Journal of Mass Spectrometry and Ion Processes*(78): 53-68.
Katajamaa, M. and M. Oresic (2005). *BMC Bioinformatics*(6): 179-186.
Katajamaa, M. and M. Oresic (2007). *Journal of Chromatography A* **1158**(1-2): 318-325.
Katja Dettmer, P. A. A. B. D. H. (2007). *Mass Spectrometry Reviews* **26**(1): 51-78.
Kato, A., E. G. Seo, T. A. Einhorn, J. E. Bishop and A. W. Norman (1998). *Bone* **23**(2): 141-146.
Kazmi, S. A., S. Ghosh, D. Shin, D. W. Hill and D. F. Grant (2006). *Metabolomics* **2**(2): 75-83.
Kell, D. B. (2004). *Current Opinions in Microbiology*(7): 296-307.
Kell, D. B. (2007). *Expert Review of Molecular Diagnostics* **7**(4): 329-333.
Kenney, B. and J. P. Shockcor (2003). *PharmaGenomics*. **November/December:** 56-63.
Kenny, P., K. Nomoto and R. Orlando (1992). *Rapid Communications in Mass Spectrometry* **6**(2): 95-97.
Kettaneh, N., A. Berglund and S. Wold (2005). *Computational Statistics & Data Analysis* **48**(1): 69-85.
Keun, H. C. (2005). *Pharmacology and Therapeutics* **109**(1-2): 92-106.
Kind, T., V. Tolstikov, O. Fiehn and R. H. Weiss (2007). *Analytical Biochemistry* **363**(2): 185-195.
King, R., R. Bonfiglio, C. Fernandez-Metzler, C. Miller-Stein and T. Olah (2000). *Journal of the American Society for Mass Spectrometry*(11): 942-950.
Kluge, H. J. and G. Bollen (1992). *Nuclear Instruments and Methods*(70): 473-481.
Knott, L. and A. J. Bailey (1998). *Bone* **22**(3): 181-187.
Kochhar, S., D. M. Jacobs, Z. Ramadan, F. Berruex, A. Feurholz and L. B. Fay (2006). *Analytical Biochemistry* **352**(2): 274-281.
Kopka, J. (2006). *Journal of Biotechnology* **124**: 312-322.
Koppenaal, D. W., C. J. Barinaga, M. B. Denton, R. P. Sperline, G. M. Heiftje, G. D. Schilling, F. J. Andrade and J. H. Barnes (2005). *Analytical Chemistry*(1): 419-427.

Korner, R., M. Wilm, K. Morand, M. Schubert and M. Mann (1996). *Journal of the American Society for Mass Spectrometry*(7): 150-156.
Kuiper, I., E. L. Lagendijk, G. V. Bloemberg and B. J. J. Lugtenberg (2004a). *Molecular Plant-Microbe Interactions* **17**(1): 6-15.
Kuiper, I., E. L. Lagendijk, R. Pickford, J. P. Derrick, G. E. M. Lamers, J. E. Thomas-Oates, B. J. J. Lugtenberg and G. V. Bloemberg (2004b). *Molecular Microbiology* **51**(1): 97-113.
Kulik, W. and W. Heerma (1991). *Biological Mass Spectrometry* **20**: 553-559.
Lafontaine, M., C. Champmartin, P. Simon, P. Delsaut and C. Funck-Brentano (2006). *Toxiology Letters* **162**(2): 181-185.
Lahoz, C., R. Peña, J. Mostaza, J. Jiménez, E. Subirats, X. Pintó, M. Taboada and A. López-Pastor (2003). *Atherosclerosis* **168**(2): 289-295.
Lamers, R., J. v. Nesselrooij, V. B. Kraus, J. M. Jordan, J. B. Renner, A. D. Dragomir, G. Luta, J. v. d. Greef and J. DeGroot (2005). *Osteoarthritis and Cartilage* **7**: 56-62.
Lauridsen, M., S. H. Hansen, J. W. Jaroszewski and C. Cornett (2007). *Analytical Chemistry* **79**(3): 1181-1186.
LeBeau, M. A., M. L. Miller and B. Levine (2001). *Forensic Science International*(119): 161-167.
Lee, S. H., S. O. Kim and B. C. Chung (1998). *Journal of Chromatography B: Biomedical Sciences and Applications* **719**(1-2): 1-7.
Lenz, E., J. Bright, R. Knight, I. Wilson and H. Major (2004). *Analyst* **129**: 535-541.
Lenz, E. M., R. E. Williams, J. Sidaway, B. W. Smith, R. S. Plumb, K. A. Johnson, P. Rainville, J. Shockcor, C. L. Stumpf, J. H. Granger and I. D. Wilson (2007). *Journal of Pharmaceutical and Biomedical Analysis* **44**(4): 845-852.
Lenz, E. M. and I. D. Wilson (2007). *Journal of Proteome Research***6**(2): 443-458.
Lenz, E. M., I. D. Wilson, J. A. Timbrell and J. K. Nicholson (2000). *Biomarkers* **5**(6): 424-435.
Leu, C., E. Luegmayr, L. P. Freedman, G. A. Rodan and A. A. Reszka (2006). *Bone* **38**: 628-636.
Levsen, K., H.-M. Schiebel, B. Behnke, R. Dotzer, W. Dreher, M. Elend and H. Thiele (2005). *Journal of Chromatography A* **1067**(1-2): 55-72.
Li, X., R. J. Quigg, J. Zhou, J. T. Ryaby and H. Wang (2005). *Journal of Cell Biochemistry* (95): 189-205.
Liebich, H. M., A. Pickert and J. Woll (1981). *Journal of Chromatography* **217**: 255-262.
Lin, T., A. H. Payne and G. L. Glish (2001). *Journal of the American Society for Mass Spectrometry* **12**(5): 497-504.
Lindon, J. C., E. Holmes, M. E. Bollard, E. G. Stanley and J. K. Nicholson (2004). *Biomarkers* **9**(1): 1-31.
Lindon, J. C., J. K. Nicholson, J. C. Lindon and J. K. Nicholson *TrAC Trends in Analytical Chemistry* **In Press, Accepted Manuscript**: 246.
Lofman, O., P. Magnusson, G. Toss and L. Larsson (2005). *Clinica Chimica Acta*(356): 67-75.
Lopes-Virella, M., G. Virella, M. Debeukelaer, C. J. Owens and J. A. Colwell (1979). *Clinica Chimica Acta* **94**(1): 73-81.
Lu, G., J. Wang, X. Zhao, H. Kong and G. Xu (2006). *Chinese J Chromatgraphy* **24**(2): 109-113.
Lugtenberg, B., A. deWeger and B. Schippers (1994). *BCPC Monograph* **57**: 293-302.
Lutz, U., R. W. Lutz and W. K. Lutz (2006). *Analytical Chemistry* **78**: 4564-4571.
Mamyrin, B. A. (2001). *International Journal of Mass Spectrometry*(206): 251-266.

Mandelin, J., M. Hukkanen, T. Li, M. Korhonen, M. Liljestrom, T. Sillat, R. Hanemaaijer, J. Salo, S. Santavirta and Y. T. Konttinen (2005). *Bone* **38**(6): 769-777.

Mann, M. and M. Wilm (1995). *Trends in Biochemical Sciences* (20): 219-224.

Marahiel, M. A., T. Stachelhaus and H. D. Mootz (1997). *Chemical Reviews* **97**(7): 2651-2674.

Matsumoto, I. and T. Kuhara (1996). *Mass Spectrometry Reviews* (15): 43-57.

Mawhinney, D. B., E. I. Hamelin, R. Fraser, S. S. Silva, A. J. Pavlopoulos and R. J. Kobelski (2007). *Journal of Chromatography B* **852**(1-2): 235.

McLafferty, F. (1968). *Journal of the American Chemical Society***90**(17): 4745-4746.

Metabolomics volume 3 (2007). *Metabolomics* **3**(3): 175-256.

Miller, M. G. (2007). *Journal of Proteome Research***6**(2): 540-545.

Miller, P. D. (2005). *Current Osteoporosis Reports* (3): 103-110.

Minakuchi, H., K. Nakanishi, N. Soga, N. Ishizuka and K. Tanaka (1997). *Journal of Chromatography A* **762**: 135-146.

Minakuchi, H., K. Nakanishi, N. Soga, N. Ishizuka and K. Tanaka (1998). *Journal of Chromatography A* **797**: 121-131.

Minakuchi, H., K. Nakanishi, N. Soga, N. Ishizuka and N. Tanaka (1996). *Analytical Chemistry* **68**: 3498-3501.

Minshull, J., J. E. Ness, C. Gustafsson and S. Govindarajan (2005). *Current Opinions in Chemical Biology*(9): 202-209.

Monton, M. R. N. and T. Soga (2007). *Journal of Chromatography A* **1168**(1-2): 237-246.

Moravcova, D., P. Jandera, J. Urban and J. Planeta (2004). *Journal of Separation Science* **27**: 789-800.

Moro, L., C. Modricky, N. Stagni, F. Vittur and B. d. Bernard (1984). *Analyst* **109**: 1621-1622.

Mukherjee, A. K. *Letters in Applied Microbiology* **45**(3): 330-335.

Mukhopadhyay, R. (2006). *Analytical Chemistry*: 4255-4259.

Munson, M. and F. Field (1966). *Journal of the American Chemical Society***88**: 2621-2629.

Nancollas, G. H., R. Tang, R. J. Phipps, Z. Henneman, S. Gulde, W. Wu, A. Mangood, R. G. G. Russell and F. H. Ebetino (2006). *Bone* **38**: 617-627.

Need, A. G. (2006). *Clinica Chimica Acta* **368**: 48-52.

Newman, D. J., M. J. Pugia, J. A. Lott, J. F. Wallace and A. M. Hiar (2000). *Clinica Chimica Acta* **294**(1-2): 139-155.

Newton, K. A. and S. A. McLuckey (2004). *Journal of the American Society for Mass Spectrometry* **15**(4): 607-615.

Ngai, H. H. Y., W. H. Sit, P. P. Jiang, V. Thongboonkerd and J. M. F. Wan (2007). *Journal of Proteome Research***6**(8): 3313-3320.

Nicholson, J., J. Conelly and E. Holmes (2002). *Nature Reviews - Drug Discovery* **1**(2): 153-161.

Nicholson, J., E. Holmes, J. Lindon and I. Wilson (2004). *Nature - Biotechnology* **22**(10): 1268-1274.

Nicholson, J., E. Holmes and I. Wilson (2005). *Nature Reviews - Microbiology* **3**(5): 431-438.

Nicholson, J. K., J. C. Lindon and E. Holmes (1999). *Xenobiotica* **29**(11): 1181-1189.

Nicholson, J. K. and I. D. Wilson (2003). *Nature Reviews - Drug Discovery* **2**(8): 668-676.

Nielsen, J. and S. Oliver (2005). *Trends in Biotechnology* **23**(11): 544-546.

Nielsen, T. H., D. Sorensen, C. Tobiasen, J. B. Andersen, C. Christophersen, M. Givskov and J. Sorensen (2002). *Applied Environmental Microbiology* **68**(7): 3416-3423.

Nielsen, T. H., C. Thrane, C. Christophersen, U. Anthoni and J. Sorensen (2000). *Journal of Applied Microbiology* **89**(6): 992-1001.
Niemela, O. (2007). *Clinica Chimica Acta* **377**(1-2): 39.
Nier, A. (1947). *Reviews of Scientific Instrumentation* **18**: 398-411.
Niessen, W. M. A. (2003). *Journal of Chromatography A* **1000**: 413-436.
Nordstrom, A., G. O'Maille, C. Qin and G. Siuzdak (2006). *Analytical Chemistry* **78**: 3289-3295.
Nork, S. E. (2005). *Journal of Orthopaedic Trauma* **19**(10): 7-9.
Nourse, B. D. and R. G. Cooks (1990). *Analytica Chimica Acta* (228): 1-21.
Obrant, K. J., K. K. Ivaska, P. Gerdhem, S. L. Alatalo, K. Pettersson and H. K. Vaananen (2005). *Bone*(36): 786-792.
Oldiges, M., S. Lütz, S. Pflug, K. Schroer, N. Stein and C. Wiendahl (2007). *Applied Microbiology and Biotechnology* **76**(3): 495-511.
Oliver, S. G., M. K. Winson, D. B. Kell and D. B. Baganz (1998). *Trends in Biotechnology* (16): 373-378.
Pan, Z., H. Gu, N. Talaty, H. Chen, N. Shanaiah, B. Hainline, R. Cooks and D. Raftery (2007). *Analytical and Bioanalytical Chemistry* **387**(2): 539-549.
Papale, M., M. C. Pedicillo, B. J. Thatcher, S. Di Paolo, L. L. Muzio, P. Bufo, M. T. Rocchetti, M. Centra, E. Ranieri and L. Gesualdo *Journal of Chromatography B* **856**(1-2): 205-213.
Pappin, D., P. Hojrup and A. Bleasby (1993). *Current Biology* **3**(6): 327-332.
Pascual, M., G. Steiger, S. Sadallah, J. P. Paccaud, J. L. Carpentier, R. James and J. A. Schifferli (1994). *Journal of Experimental Medicine* **179**(3): 889-899.
Paul, W. and H. Steinwedel (1953). *Naturforsch*(8a): 448-451.
Pauling, L., A. B. Robinson, R. Teranishi and P. Cary (1971). *Proceedings of the National Academy of Sciences* **68**(10): 2374-2376.
Pearson, K. (1901). *Philosophy Magazine* **2**(11): 559-572.
Penkman, K. (2005). *Ph.D. Thesis (University of Newcastle)*.
Perkins, D., D. Pappin, D. Creasy and J. Cottrell (1999). *Electrophoresis* **20**(18): 3551-3567.
Pham-Tuan, H., L. Kaskavelis, C. A. Daykin and H.-G. Janssen (2003). *Journal of Chromatography B* **789**(2): 283-301.
Pisitkun, T., R. Johnstone and M. A. Knepper (2006). *Molecular Cell Proteomics* **5**(10): 1760-1771.
Pizzolato, T. M., M. J. L. de Alda and D. Barcelo (2007). *TrAC Trends in Analytical Chemistry* **26**(6): 609-624.
Plumb, R. S., J. H. Granger, C. L. Stumpf, K. A. Johnson, B. W. Smith, S. Gaulitz, I. D. Wilson and J. Castro-Perez (2005). *Analyst*(130): 844-849.
Plumb, R. S., C. L. Stumpf, J. H. Granger, J. Castro-Perez, J. H. Haselden and G. J. Dear (2003). *Rapid Communications in Mass Spectrometry* **17**: 2632-2638.
Politi, L., L. Morini, A. Groppi, V. Poloni, F. Pozzi and A. Polettini (2005). *Rapid Communications in Mass Spectrometry*(19): 1321-1331.
Premstaller, A., H. Oberacher, W. Walcher, A. M. Timperio, L. Zolla, J. P. Chervet, N. Cavusoglu, A. van Dorsselaer and C. G. Huber (2001). *Analytical Chemistry* **73**(11): 2390-2396.
Price, K. E., S. S. Vandaveer, C. E. Lunte and C. K. Larive (2005). *Journal of Pharmacological and Biomedical Analysis* **38**(5): 904-909.
Ptak, M., A. Heitz, M. Guinand and G. Michel (1980). *Biochemistryical and Biophysical Research Communications* **94**(4): 1311-1318.
Qvist, P., S. Christgau, B. J. Pedersen, A. Schlemmer and C. Christiansen (2002). *Bone* **31**(1): 57-61.
Rainville, P. D., C. L. Stumpf, J. P. Shockcor, R. S. Plumb and J. K. Nicholson (2007). *Journal of Proteome Research***6**(2): 552-558.

Ramadan, Z., D. Jacobs, M. Grigorov and S. Kochhar (2006). *Talanta* **68**(5): 1683-1691.
Ramsay, S. L., P. J. Meikle, J. J. Hopwood and P. R. Clements (2005). *Analytical Biochemistry*(345): 30-46.
Rassi, Z. E. (1996). *Journal of Chromatography A*(720): 93-118.
Rezzi, S., Z. Ramadan, L. B. Fay and S. Kochhar (2007a). *Journal of Proteome Research***6**(2): 513-525.
Rezzi, S., Z. Ramadan, F. P. J. Martin, L. B. Fay, P. vanBladeren, J. C. Lindon, J. K. Nicholson and S. Kochhar (2007b). *Journal of Proteome Research***6**(11): 4469-4477.
Robertson, D. G., M. D. Reily and J. D. Baker (2007). *Journal of Proteome Research***6**(2): 526-539.
Robertson, D. G., M. D. Reily, R. E. Sigler, D. F. Wells, D. A. Paterson and T. K. Braden (2000). *Toxicological Sciences* **57**(2): 326-337.
Robins, S. P. (1995). *Acta Orthopaedica Scandanavica* (66): 171-175.
Robinson, P. and M. MacDonell (2004). *Environmental Toxicology Pharmacology*(18): 201-213.
Rodrigues, J. (2005). *PhD. Thesis (University of York).*
Roepstorff, P. and J. Fohlman (1984). *Biological Mass Spectrometry* **11**(11): 601-601.
Roy, S. M., M. Anderle, H. Lin and C. H. Becker (2004). *International Journal of Mass Spectrometry*(238): 163-171.
Rubinacci, A., R. Melzi, M. Zampino, A. Soldarini and I. Villa (1999). *Clinical Chemistry* **45**(9): 1510-1516.
Bonfiglio, R., T. Olah, R. King and K Merkle (1999). *Rapid Communications in Mass Spectrometry* **13**(12): 1175-1185.
Ryan, D. and K. Robards (2006). *Analytical Chemistry* **78**(23): 7954-7958.
Ryu, H., H. D. Rosas, S. M. Hersch and R. J. Ferrante (2005). *Pharmacology and Therapeutics*(108): 193-207.
Samuelson, O. and E. Sjöström (1952). *Sven. Kem. Tidskr.*
Sanchez-Ponce, R. and F. P. Guengerich (2007). *Analytical Chemistry* **79**(9): 3355-3362.
Sanders, B. D., R. L. Slotcavage, D. L. Scheerbaum, C. J. Kochansky and T. G. Strein (2005). *Analytical Chemistry* **77**(8): 2332-2337.
Sangster, T., H. Major, R. Plumb, A. Wilson and I. Wilson (2006). *The Analyst*(131): 1075-1078.
Sangster, T. P., J. E. Wingate, L. Burton, F. Teichert and I. D. Wilson (2007). *Rapid Communications in Mass Spectrometry* **21**(18): 2965-2970.
Sansone, S., T. Fan, R. Goodacre, J. L. Griffin, N. W. Hard, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. v. d. Werf, B. v. Ommen and O. Fiehn (2007). *Nature - Biotechnology* **25**(8): 846-848.
Sarti, A., A. DeGaudio, A. Messineo, M. Cuttini and A. Ventura (2001). *Critical Care in Medicine* **29**(8): 1626-1629.
Satoh, T., H. Tsuno, M Ianaga and Y. Kammei (2005). *Journal of the American Society for Mass Spectrometry*(16): 1969-1975.
Saude, E. J. and B. D. Sykes (2007). *Metabolomics* **3**(1): 19-27.
Schmidt, C. W. (2004). *Environmental Health Perspectives* **112**(7): 411-415.
Schneider, U., E. A. Schober, N. A. Streich and S. J. Breusch (2002). *Clinica Chimica Acta*(324): 81-88.
Schnell, N., K.-D. Entian, U. Schneider, F. Gotz, H. Zahner, R. Kellner and G. Jung (1988). *Nature* **333**(6170): 276-278.
Schoenau, E. and F. Rauch (2003). *Journal of Laboratory Medicine*(27): 32-42.
Schram, K. H. (1998). *Mass Spectrometry Reviews* (17): 131-251.
Scott, C. D. (1974). *Science*(186): 226-233.

Scott, W. R. P., S.-B. Baek, D. Jung, R. E. W. Hancock and S. K. Straus (2007). *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1768**(12): 3116-3126.

Seebeck, P., H. J. Bail, C. Exner, H. Schell, R. Michel, H. Amthauer, H. Bragulla and G. N. Duda (2005). *Bone*(37): 669-677.

Severns, A. E., Y. Lee, S. D. Nelson, E. E. Johnson and J. M. Kabo (2003). *Clinical Orthopaedics and Related Research* (424): 231-238.

Shirley, D., D. Marsh, G. Jordan, S. McQuaid and G. Li (2005). *Journal of Orthopaedic Research* (23): 1013-1021.

Simpson, R. C., W. B. Emary, I. Lys, R. J. Cotter and C. C. Fenselau (1991). *Journal of Chromatography A*(536): 143-153.

Siuzdak, G. (1994). *Procedures of the National Academy of Science* **91**: 11290-11297.

Smith, C. A., E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak (2006). *Analytical Chemistry* **78**: 779-778.

Smith, M. T., R. Vermeulen, G. Li, L. Zhang, Q. Lan, A. E. Hubbard, M. S. Forrest, C. McHale, X. Zhao, L. Gunn, M. Shen, S. M. Rappaport, S. Yin, S. Chanock and N. Rothman (2005). *Chemical-Biological Interactions* (153-154): 123-127.

Soria, A., B. Wright, D. Goodall and J. Wilson (2007). *Electrophoresis* **28**(6): 950-964.

Srivastava, A. K., S. Mohan, F. R. Singer and D. J. Baylink (2002). *Bone* **31**(1): 62-69.

Stafford, G. (2002). *Journal of the American Society for Mass Spectrometry*(13): 589-596.

Stokvis, E., H. Rosing and J. H. Beijnen (2005). *Rapid Commun. Mass Spectrom.*(19): 401-407.

Sturm, G., H. Haberlein, T. Bauer, T. Plaum and D. J. Stalker (1991). *Journal of Chromatography: Biomedical Applications* **562**(1-2): 351-362.

Summerfield, S. and S. Gaskell (1997). *International Journal of Mass Spectrometry and Ion Processes* **165-166**: 509-521.

Sumner, L. W. (2006). *Biotechnology in Agricultural Forestry* **57**: 21-32.

Sumner, L. W., P. Mendes and R. A. Dixon (2003). *Phytochemistry*(62): 817-836.

Sumpton, D. (2007). *PhD. Thesis (University of York).*

Svante Wold, N. K. K. T. (1996). *Journal of Chemometrics* **10**(5-6): 463-482.

Svec, F. (2004). *Journal of Separation Science*(27): 747-766.

Svec, F. and J. M. J. Frechet (1992). *Analytical Chemistry* **64**(7): 820-822.

Svoboda, P. and H. Kasai (2004). *Analytical Biochemistry*(334): 239-250.

Tambong, J. T. and M. Höfte (2001). *European Journal of Plant Pathology* **107**(5): 511-521.

Tanaka, K., H. Waki, Y. Ido, S. Akita, Y. Yoshida and T. Yoshida (1988). *Rapid Communications in Mass Spectrometry*(2): 151-153.

Tang, H. and Y. Wang (2006). *Progression in Biochemistry & Biophysics* **33**(5): 401-417.

Tang, L. and P. Kebarle (1993). *Analytical Chemistry* **65**(24): 3654-3668.

Tang, S., W. Zhou, N. S. Sheerin, R. W. Vaughan and S. H. Sacks (1999). *Journal of Immunology* **162**(7): 4336-4341.

Taylor, P. J. (2005). *Clinical Biochemistryistry* **38**(4): 328.

Teahan, O., S. Gamble, E. Holmes, J. Waxman, J. K. Nicholson, C. Bevan and H. C. Keun (2006). *Analytical Chemistry* **78**: 4307-4318.

Teas, J., J. E. Cunningham, J. H. Fowke, D. Nitcheva, C. P. Kanwat, R. J. Boulware, D. W. Sepkovic, T. G. Hurley and J. R. Hebert (2005). *Cancer Detection and Prevention Journal* **29**(6): 494-500.

Teesch, L. M. and J. Adams (1990). *Journal of the American Chemical Society***112**(11): 4110-4120.

Terpos, E., M. Politou and A. Rahemtulla (2005). *Blood Reviews* (19): 125-142.

Thomashow, L. and D. Weller (1995). Current concepts in the use of introduced bacteria for biological disease control: mechanisms and antifungal metabolites., *Chapman & Hall.*

Thompson, J. (1899). *Philosophical Magazine* **48**(5): 123-126.

Thorne, G. C., K. D. Ballard and S. J. Gaskell (1990). *Journal of the American Society for Mass Spectrometry* **1**(3): 249-257.

Tolstikov, V. V., A. Lommen, K. Nakanishi, N. Tanaka and O. Fiehn (2003). *Analytical Chemistry* **75**(23): 6737-6740.

Toohey, J. I., C. D. Nelson and G. Krotkov (1965). *Canadian Journal of Botany* **43**: 1151-1155.

Tran, H., A. Ficke, T. Asiimwe, M. Hofte and J. M. Raaijmakers *New Phytologist* **175**(4): 731-742.

Trygg, J., E. Holmes and T. Lundstedt (2007). *Journal of Proteome Research* **6**(2): 469-479.

Tswett, M. (1906). *Berichte der Deutschen botanischen Gesellschaft*: 316.323.

Turner, J. M. and A. J. Messenger (1986). *Advanced Microbiology and Physiology* **27**: 211-275.

Tyan, Y.-C., H.-R. Guo, C.-Y. Liu and P.-C. Liao (2006). *Analytica Chimica Acta* **579**(2): 158-176.

Ullsten, S., R. Danielsson, D. Backstrom, P. Sjoberg and J. Bergquist (2006). *Journal of Chromatography A* **1117**: 87-93.

van den Berg, R., H. Hoefsloot, J. Westerhuis, A. Smilde and M. van der Werf (2006). *BMC Genomics* **7**(1): 142-151.

van Ravenzwaay, B., G. C.-P. Cunha, E. Leibold, R. Looser, W. Mellert, A. Prokoudine, T. Walk and J. Wiemer (2007). *Toxicology Letters* **172**(1-2): 21-28.

Various (1999). Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism - American Society for Bone and Mineral Research, *Lippincott Williams and Wilkins.*

Verhoeven, N. M., G. S. Salomons and C. Jakobs (2005). *Clinica Chimica Acta*(361): 1-9.

Vidotto, C., D. Fousert, M. Akkermann, A. Griesmacher and M. M. Muller (2003). *Clinica Chimica Acta*(335): 27-32.

Viguet-Carrin, S., P. Garnero and P. D. Delmas (2006). *Osteoporosis International* **17**: 319-336.

Villas-Boas, S. G., S. Mas, M. Akesson, J. Smedsgaard and j. Nielson (2004). *Mass Spectrometry Reviews* (24): 613-646.

Vortkamp, A., S. Pathi, G. M. Peretti, E. M. Caruso, D. J. Zaleske and C. J. Tabin (1998). *Mechanical Developments* (71): 65-76.

Wagner, S., K. Scholz, M. Sieber, M. Kellert and W. Voelkel (2007). *Analytical Chemistry* **79**(7): 2918-2926.

Walgren, J. L. and D. C. Thompson (2004). *Toxiology Letters*(149): 377-385.

Wang, S. and C. Liao (2004). *Journal of Chromatography A*(1051): 213-219.

Wang, W., Q. Li, L. Hasvold, B. Steiner, D. A. Dickman, H. Ding, A. Clairborne, H.-J. Chen, D. Frost, R. C. Goldman, K. Marsh, Y.-H. Hui, B. Cox, A. Nilius, D. Balli, P. Lartey, J. J. Plattner and Y. L. Bennani (2003). *Bioorganic & Medicinal Chemistry Letters* **13**(3): 489-493.

Wang, X., W. Li and H. T. Rasmussen (2005a). *Journal of Chromatography A* **1083**(1-2): 58-62.

Wang, X., X. Zhang, Z. Li and X. Yu (2005b). *Journal of Zhejiang University Science* **9**: 926-930.

Want, E. J., B. F. Cravatt and G. Siuzdak (2005). *ChemBiochemistry* **6**(11): 1941-1951.

Want, E. J., A. Nordstrom, H. Morita and G. Siuzdak (2007). *Journal of Proteome Research* **6**(2): 459-468.

Weckwerth, W. and K. Morgenthal (2005). *Drug Discovery Today Targets*(22): 1551-1558.

Weisman, S. M. and V. Matkovic (2005). *Clinical Therapy* (27): 299-308.

Wienkoop, S., M. Glinski, N. Tanaka, V. Tolstikov, O. Fiehn and W. Weckwerth (2004). *Rapid Communications in Mass Spectrometry* **18**(6): 643-650.

Williams, R. E., E. M. Lenz, J. S. Lowden, M. Rantalainen and I. D. Wilson (2005). *Molecular Biosystems* **1**(2): 166-175.

Willis, C. M., S. M. Church, C. M. Guest, W. A. Cook, N. McCarthy, A. J. Bransbury, M. R. T. Church and J. C. T. Church (2004). *British Medical Journal* **329**(7468): 712-720.

Wilm, M. and M. Mann (1994). *International Journal of Mass Spectrometry and Ion Processes* (136): 167-180.

Wilm, M. and M. Mann (1996). *Analytical Chemistry* (68): 1-8.

Wilson, I. D., R. Plumb, J. Granger, H. Major, R. Williams and E. M. Lenz (2005). *Journal of Chromatography B* (817): 67-76.

Windig, W. and W. F. Smith (2007). *Journal of Chromatography A* **1158**(1-2): 251.

Wishart, D. S., D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel and L. Querengesser (2007). *Nucleic Acids Research.* **35**(Supplement 1): D521-526.

Woitge, H. W., M. Pecherstorfer, Y. Li, A. Keck, E. Horn, R. Ziegler and M. J. Seibel (1999). *Journal of Bone Mineral Research* (14): 792-801.

Woitge, H. W., C. Scheidt-Nave, C. Kissling, G. Leidig-Bruckner, K. Meyer, A. Grauer, S. H. Scharla, R. Zeigler and M. J. Seibel (1998). *Journal of Clinical Endocrinology Metabolism* **83**: 68-75.

Wold, S., M. Josefson, J. Gottfries and A. Linusson (2004). *Journal of Chemometrics* **18**(3-4): 156-165.

Wold, S., M. Sjostrom and L. Eriksson (2001). *Chemometrics and Intelligent Laboratory Systems* **58**(2): 109-130.

Wraighte, P. J. and B. E. Scammell (2006). *Surgery* **24**(6): 198-206.

Wuhrer, M., C. A. M. Koeleman, A. M. Deelder and C. H. Hokke (2004). *Analytical Chemistry*(76): 833-838.

Yakimov, M. M., W.-R. Abraham, H. Meyer, G. Laura and P. N. Golyshin (1999). *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1438**(2): 273-280.

Yalcin, T., C. Khouw, I. G. Csizmadia, M. R. Peterson and A. G. Harrison (1995). *Journal of the American Society for Mass Spectrometry* **6**(12): 1165-1174.

Yamashita, M. and J. B. Fenn (1984). *Journal of Physical Chemistry* (88): 4451-4459.

Yang, J., G. Xu, Y. Zheng, H. Kong, C. Wang, X. Zhao and T. Pang (2005). *Journal of Chromatography A* (1084): 214-221.

Yang, S.-Z., D.-Z. Wei and B.-Z. Mu (2006). *Journal of Biochemistry and Biophysical Methods* **68**(1): 69.

Yang, S.-Z., D.-Z. Wei and B.-Z. Mu (2007). *Journal of Biochemistry and Biophysical Methods* **70**(3): 519.

Yokoyama, Y., K. Yamasaki and H. Sato (2005). *Journal of Chromatography B* (816): 333-338.

Yu, H., E. H. Cooper, J. A. Settle and T. Meadows (1983). *Burns* **9** (5): 339-349.

Yu-Yahiro, J. A., R. H. Michael, N. H. Dubin, K. M. Fox, M. Sachs, W. G. Hawkes, J. R. Hebel, S. I. Zimmerman, J. Shapiro and J. Magaziner (2001). *Journal of the American Geriatrics Society* **49**(7): 877-883.

Zerefos, P., J. Prados, S. Kossida, A. Kalousis and A. Vlahou (2007). *Journal of Chromatography B* **853**(1-2): 20-30.

Zhengzheng, P. and R. Daniel (2007). *Analytical and Bioanalytical Chemistry* **V387**(2): 525-527.

Zhengzheng, P., G. Haiwei, T. Nari, C. Huanwen, S. Narasimhamurthy, E. H. Bryan, R. G. Cooks and R. Daniel (2007). *Analytical and Bioanalytical Chemistry* **V387**(2): 539-549.

Zhou, H., P. S. T. Yuen, T. Pisitkun, P. A. Gonzales, H. Yasuda, J. W. Dear, P. Gross, M. A. Knepper and R. A. Star (2006). *Kidney International* **69**(8): 1471-1476.

Zimmermann, G., P. Henle, M. Kusswetter, A. Moghaddam, A. Wentzensen, W. Richter and S. Weiss (2005). *Bone*(36): 779-785.

Zor, T. and Z. Selinger (1996). *Analytical Biochemistry* **236**(2): 302-308.

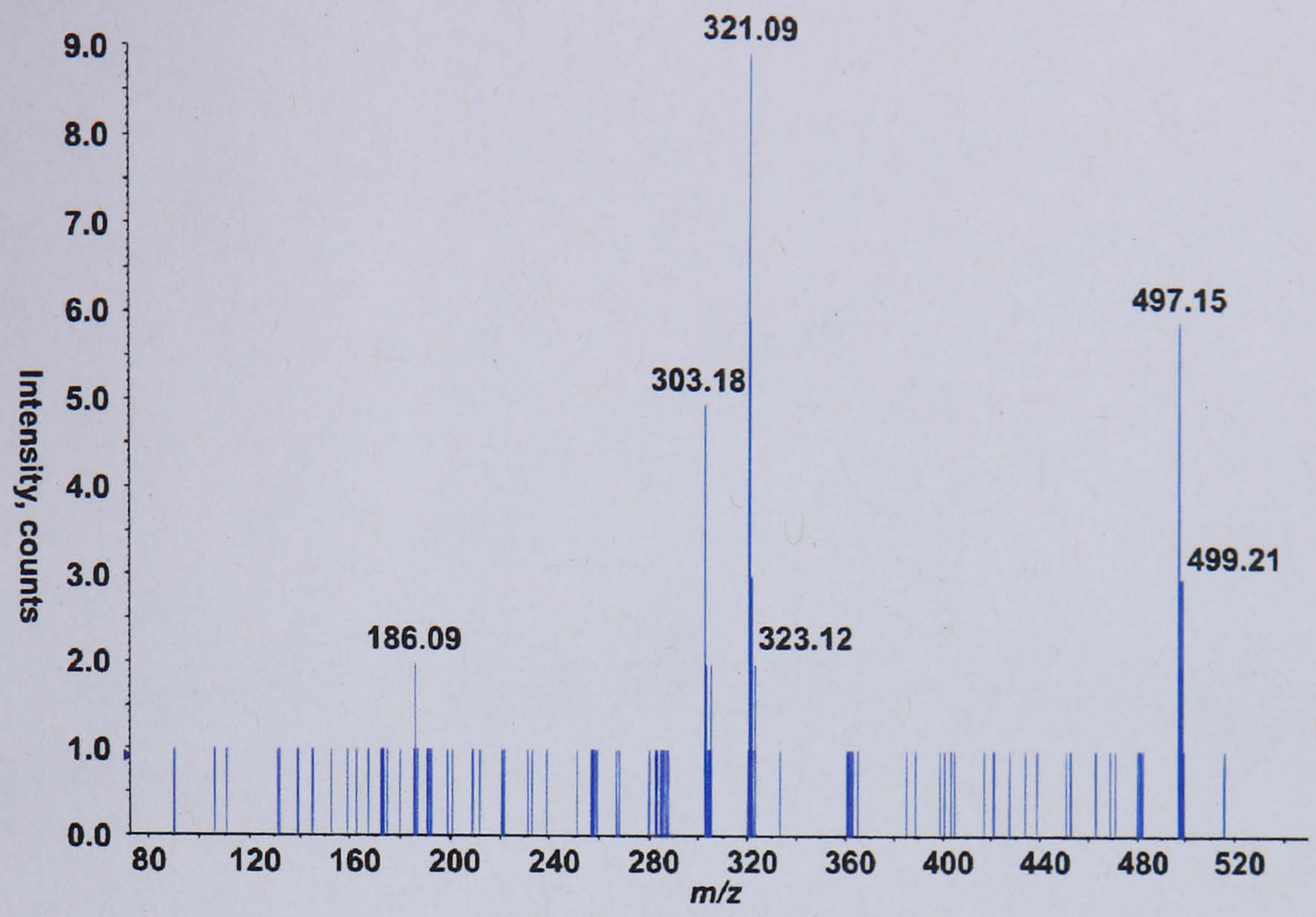
Appendices

Appendix A

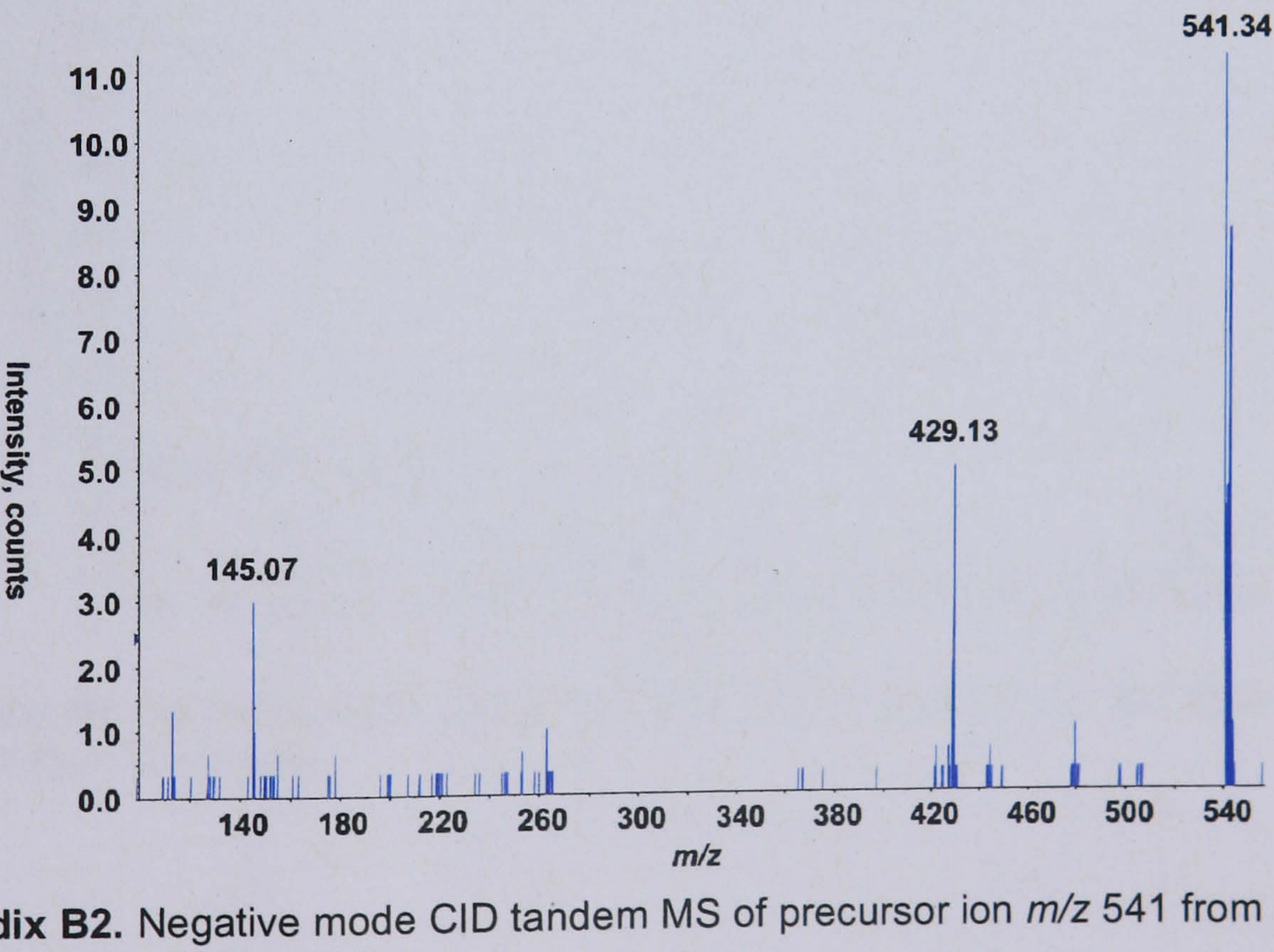
Randomly Assigned Code	Initial Code	Sex	Age	Smoker	Age Group	Time	Number of Aliquots
004	F01A	FEMALE	38	NO	36-40	AM	12
005	F01B	FEMALE	38	NO	36-40	PM	12
019	F02A	FEMALE	22	NO	21-25	AM	16
023	F02B	FEMALE	22	NO	21-25	PM	15
018	F04A	FEMALE	24	NO	21-25	AM	16
024	F04B	FEMALE	24	NO	21-25	PM	10
007	F05A	FEMALE	51-55	NO	51-55	AM	10
015	F05B	FEMALE	51-55	NO	51-55	PM	10
051	F09A	FEMALE	27	NO	26-30	AM	10
038	F10A	FEMALE	26-30	NO	26-30	AM	7
039	F10B	FEMALE	26-30	NO	26-30	PM	10
012	F13A	FEMALE	40	NO	36-40	AM	16
013	F13B	FEMALE	40	NO	36-40	PM	16
037	F15A	FEMALE	47	YES	46-50	AM	16
047	F15B	FEMALE	47	YES	46-50	PM	10
049	F20A	FEMALE	24	NO	21-25	AM	10
050	F20B	FEMALE	24	NO	21-25	PM	16
009	M01A	MALE	23	NO	21-25	AM	14
010	M01B	MALE	23	NO	21-25	PM	10
034	M02A	MALE	39	NO	36-40	AM	14
043	M02B	MALE	39	NO	36-40	PM	15
030	M03A	MALE	34	NO	31-35	AM	12
041	M03B	MALE	35	NO	31-35	PM	11
035	M04A	MALE	42	NO	41-45	AM	10
042	M04B	MALE	42	NO	41-45	PM	10
008	M05A	MALE	41-45	NO	41-45	AM	9
011	M05B	MALE	41-45	NO	41-45	PM	10
031	M07A	MALE	35	NO	31-35	AM	10
036	M07B	MALE	35	NO	31-35	PM	14
057	M09A	MALE	22	NO	21-25	AM	10
058	M09B	MALE	22	NO	21-25	PM	12
055	M11A	MALE	30	NO	26-30	AM	10
056	M11B	MALE	30	NO	26-30	PM	10
022	M12A	MALE	28	NO	26-30	AM	10
028	M12B	MALE	28	NO	26-30	PM	14
002	M13A	MALE	30	NO	26-30	AM	16
053	M13B	MALE	30	NO	26-30	PM	10
016	M14A	MALE	23	NO	21-25	AM	10
027	M14B	MALE	23	NO	21-25	PM	14
017	M15A	MALE	26-30	NO	26-30	AM	14
026	M15B	MALE	26-30	NO	26-30	PM	16
060	M20A	MALE	22	NO	21-25	AM	16
061	M20B	MALE	22	NO	21-25	PM	10
045	M21A	MALE	26	YES	26-30	AM	10
046	M21B	MALE	26	YES	26-30	PM	8
048	M22A	MALE	36	NO	36-40	AM	10
054	M22B	MALE	36	NO	36-40	PM	10
021	M23A	MALE	52	NO	51-55	AM	10
029	M23B	MALE	52	NO	51-55	PM	14

Randomly Assigned Code	Initial Code	Sex	Age	Smoker	Age Group	Time	Number of Aliquots
032	M24A	MALE	23	NO	21-25	AM	10
044	M24B	MALE	23	NO	21-25	PM	15
062	RAND1400	MALE	25	NO	21-25	PM	14
033	UNK2A	FEMALE	UNK	NO	UNK	AM	14
040	UNK2B	FEMALE	UNK	NO	UNK	PM	14
003	UNK4A	MALE	61	NO	61-65	PM	15
001	UNK4B	MALE	61	NO	61-65	AM	9
006	UNK5A	FEMALE	45	NO	41-45	AM	16
014	UNK5B	FEMALE	45	NO	41-45	PM	10
020	X03A	MALE	25	NO	21-25	AM	10
025	X03B	MALE	25	NO	21-25	PM	14
052	X06A	FEMALE	50	YES	46-50	AM	10
059	X06B	FEMALE	50	YES	46-50	PM	10

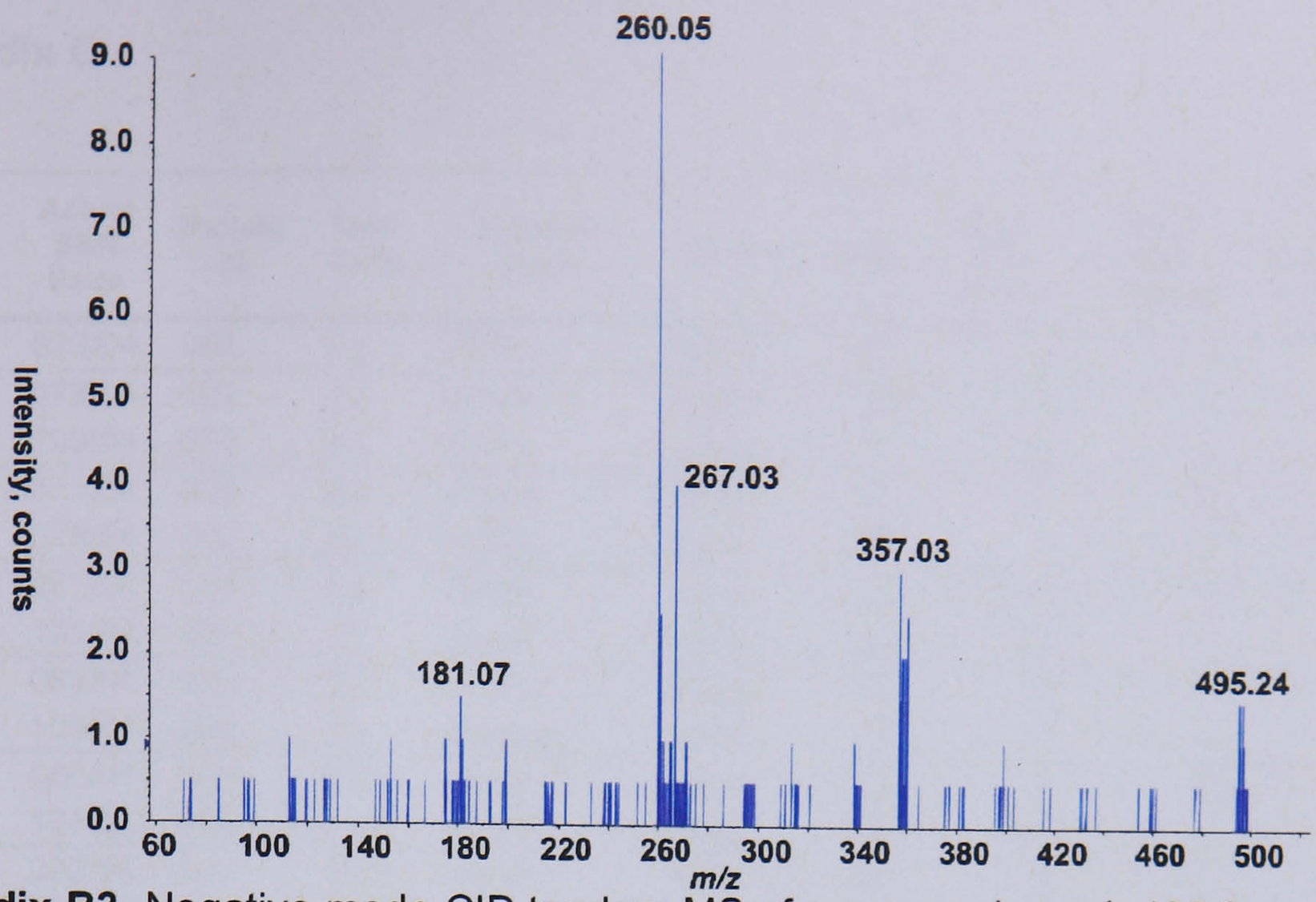
Appendix B



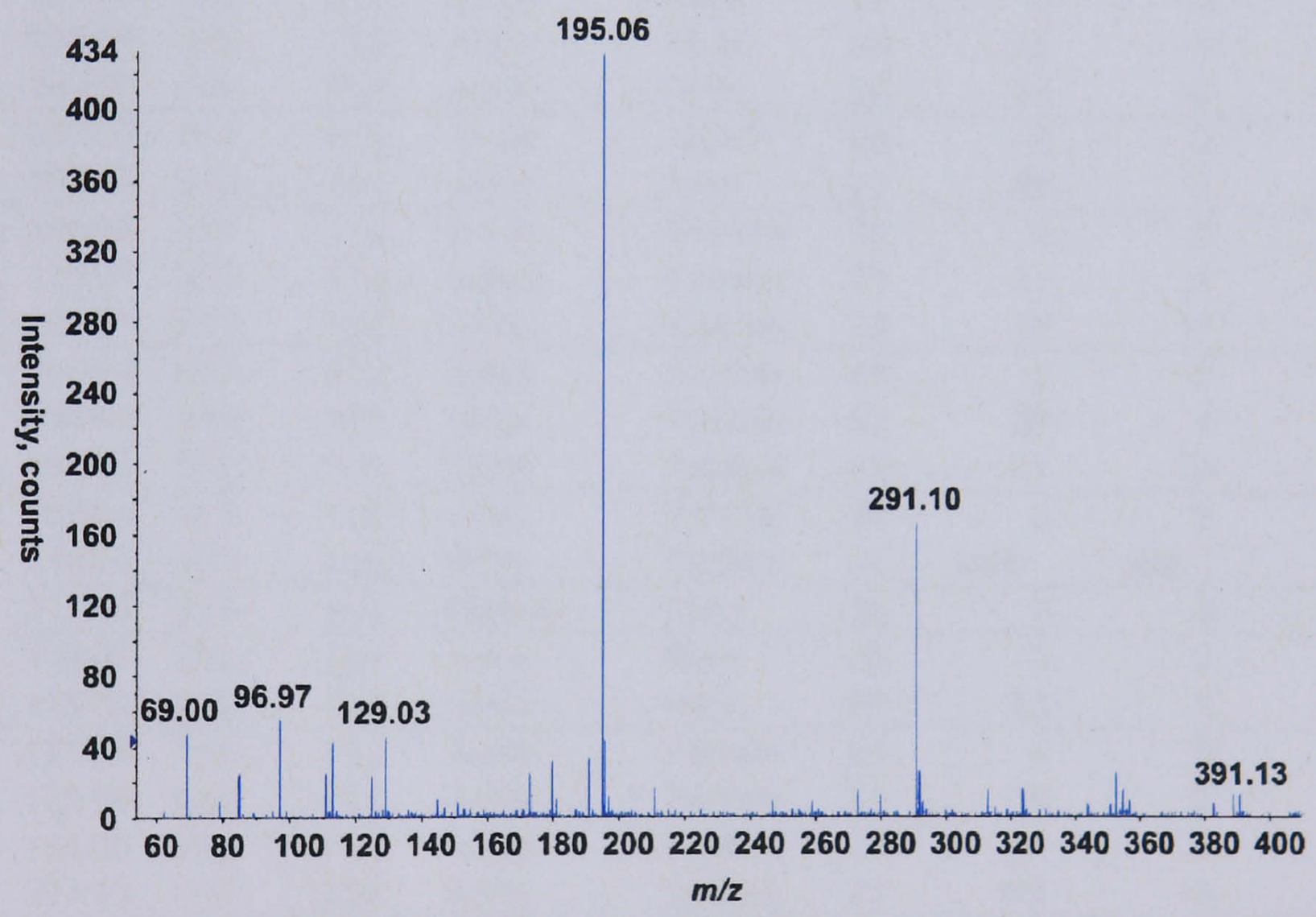
Appendix B1. Positive mode CID tandem MS of precursor ion *m/z* 497 from RP-LC-MS data gender PLS model. The presence of peaks at 2 Da higher, suggests that this metabolite contains chlorine.



Appendix B2. Negative mode CID tandem MS of precursor ion *m/z* 541 from RP-LC-MS data gender PLS model.



Appendix B3. Negative mode CID tandem MS of precursor ion *m/z* 495 from RP-LC-MS data gender PLS model.



Appendix B4. Negative mode CID tandem MS of precursor ion *m/z* 391 from RP-LC-MS data age PLS model.

Appendix C

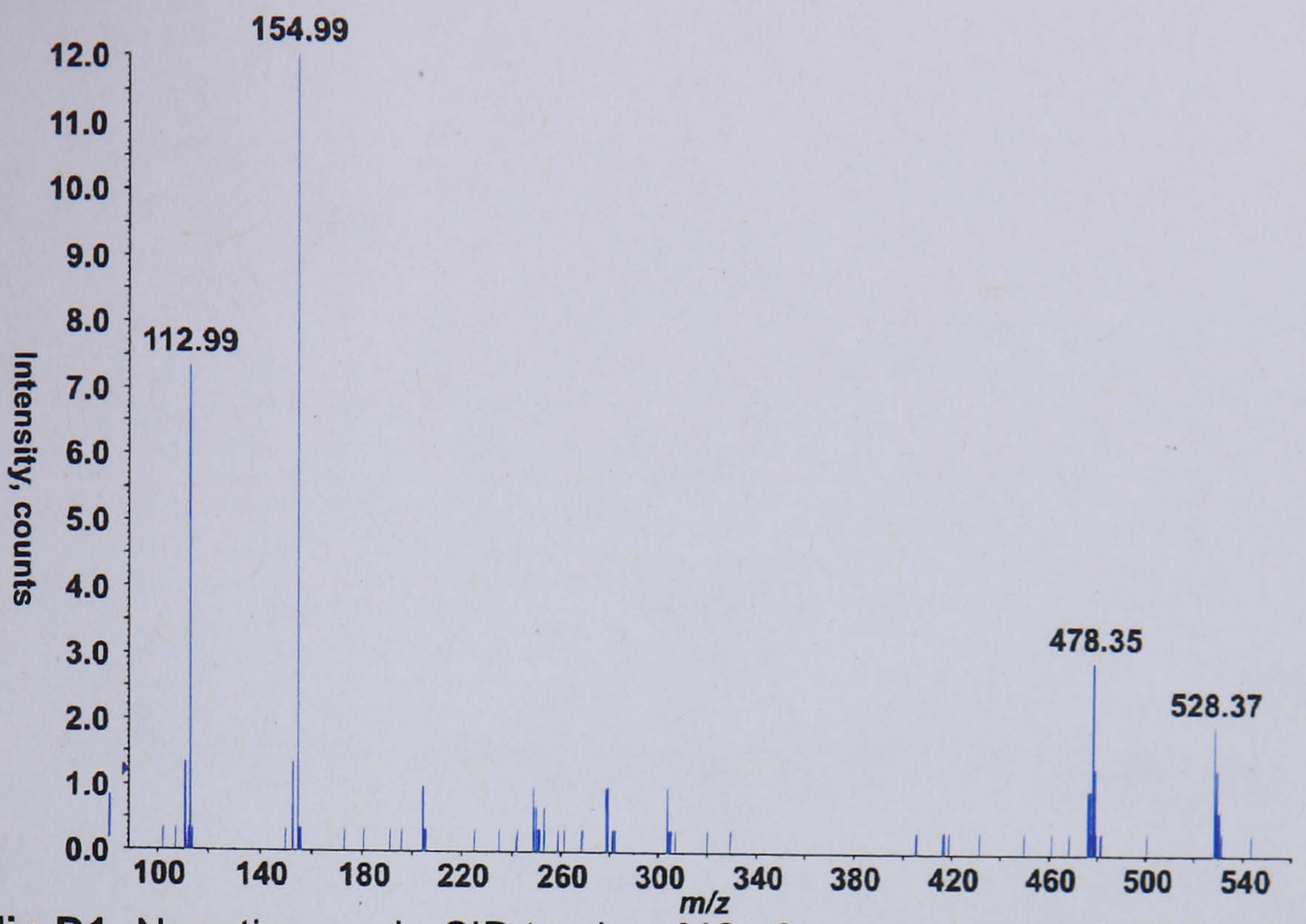
Internal S&N Code	Actual S&N Code	Patient ID	New Code	Fracture Type	Gender	Age	Days after Fracture	Weeks after Fracture	Protein Concentration (mg/mL)
72/04	073/04	001	F1	Pilon	Male	26	0	0	4
74/04	075/04	002	F2	Ankle	Male	27	0	0	5
98/04	099/04	002	F3	Ankle	Male	27	39	6	166
76/04	077/04	003	F4	Ankle	Male	47	0	0	3
78/04	079/04	003	F5	Ankle	Male	47	7	1	79
80/04	081/04	004	F6	Ankle	Male	33	0	0	*nr*
100/04	101/04	004	F7	Ankle	Male	33	34	5	165
82/04	083/04	005	F8	Radius	Male	35	0	0	1
102/04	103/04	005	F9	Radius	Male	35	28	4	236
84/04	085/04	006	F10	Ankle	Male	33	0	0	2
130/04	131/04	006	F11	Ankle	Male	33	40	6	156
86/04	087/04	007	F12	Ankle	Male	22	0	0	1
132/04	133/04	007	F13	Ankle	Male	22	35	5	*nr*
88/04	089/04	008	F14	Ankle	Male	28	0	0	1
205/05	206/05	008	F15	Ankle	Male	28	61	9	192
193/05	194/05	008	F16	Ankle	Male	28	79	11	221
90/04	091/04	009	F17	Ankle	Male	24	0	0	4
96/04	097/04	009	F18	Ankle	Male	24	22	3	160
248/05	249/05	009	F19	Ankle	Male	24	83	12	178
92/04	093/04	010	F20	Ankle	Male	20	0	0	1
181/05	180/05	010	F21	Ankle	Male	20	48	7	197
104/04	105/04	011	F22	Ankle	Female	23	0	0	179
173/05	174/05	011	F23	Ankle	Female	23	31	4	192
195/05	196/05	011	F24	Ankle	Female	23	74	11	131
106/04	107/04	012	F25	Ankle	Female	43	0	0	179
134/04	134/04	012	F26	Ankle	Female	43	29	4	198
187/05	188/05	012	F27	Ankle	Female	43	70	10	181
108/04	109/04	013	F28	Wrist	Female	24	0	0	186
136/04	136/04	013	F29	Wrist	Female	24	*unk*	*unk*	183
110/04	111/04	014	F30	Clavicle	Male	35	0	0	223
118/04	119/04	015	F31	Ankle	Male	26	0	0	148
179/05	179/05	015	F32	Ankle	Male	26	33	5	186
120/04	121/04	016	F33	Ankle	Female	23	0	0	196
135/04	135/04	016	F34	Ankle	Female	23	16	2	224
151/05	155/05	016	F35	Ankle	Female	23	44	6	*nr*
323/05	324/05	016	F36	Ankle	Female	23	123	18	169
122/04	123/04	017	F37	Ankle	Male	35	0	0	161
136/04	330/05	017	F38	Ankle	Male	35	133	19	159
124/04	125/04	018	F39	Radial Head	Male	20	0	0	182
126/04	127/04	019	F40	Ankle	Male	23	0	0	167
177/05	178/05	019	F41	Ankle	Male	23	28	4	170
189/05	190/05	019	F42	Ankle	Male	23	56	8	173
345/05	346/05	019	F43	Ankle	Male	23	126	18	173
128/04	129/04	020	F44	Ankle	Female	20	0	0	202
152/05	153/05	020	F45	Ankle	Female	20	35	5	175
256/05	257/05	020	F46	Ankle	Female	20	77	11	151
209/05	210/05	021	F47	Ankle	Female	23	0	0	148
215/05	215/05	021	F48	Ankle	Female	23	39	6	191
343/05	344/05	021	F49	Ankle	Female	23	95	14	*nr*

Internal S&N Code	Actual S&N Code	Patient ID	New Code	Fracture Type	Gender	Age	Days after Fracture	Weeks after Fracture	Protein Concentration (mg/mL)
207/05	208/05	022	F50	Wrist	Male	23	0	0	211
201/05	202/05	023	F51	Ankle	Male	19	0	0	169
254/05	253/05	023	F52	Ankle	Male	19	35	5	172
337/05	338/05	023	F53	Ankle	Male	19	91	13	165
166/05	167/05	024	F54	Wrist	Male	20	0	0	***nr***
246/05	247/05	024	F55	Wrist	Male	20	29	4	217
156/05	169/05	025	F56	Wrist	Male	22	0	0	160
197/05	198/05	025	F57	Wrist	Male	22	7	1	203
262/05	263/05	025	F58	Wrist	Male	22	35	5	205
156/05	157/05	026	F59	Radial Head	Male	25	0	0	204
249/05	241/05	026	F60	Radial Head	Male	25	14	2	242
160/05	161/05	027	F61	Ankle	Male	24	0	0	158
162/05	163/05	028	F62	Tib/Fib	Male	26	0	0	157
164/05	165/05	029	F63	Ankle	Male	44	0	0	2
315/05	316/05	029	F64	Ankle	Male	44	50	7	185
170/05	216/05	030	F65	Tib/Fib	Male	29	0	0	164
273/05	273/05	030	F66	Tib/Fib	Male	29	21	3	238
199/05	200/05	031	F67	Wrist	Male	36	0	0	248
303/05	304/05	031	F68	Wrist	Male	36	***unk***	***unk***	206
331/05	332/05	031	F69	Wrist	Male	36	***unk***	***unk***	217
158/05	159/05	032	F70	Clavicle	Male	***unk***	0	0	205
191/05	192/05	033	F71	Tib/Fib	Female	33	0	0	124
211/05	212/05	034	F72	Ankle	Male	30	0	0	3
213/05	214/05	035	F73	Ankle	Female	36	0	0	177
301/05	302/05	035	F74	Ankle	Female	36	41	6	135
202/05	204/05	036	F75	Wrist	Female	44	0	0	224
242/05	243/05	036	F76	Wrist	Female	44	***unk***	***unk***	215
266/05	267/05	036	F77	Wrist	Female	44	***unk***	***unk***	214
311/05	312/05	036	F78	Wrist	Female	44	45	6	253
327/05	328/05	036	F79	Wrist	Female	44	73	10	200
185/05	184/05	037	F80	Wrist	Male	42	0	0	191
244/05	245/05	037	F81	Wrist	Male	42	10	1	213
254/05	255/05	037	F82	Wrist	Male	42	24	3	218
183/05	183/05	038	F83	Fibula	Male	32	0	0	185
301/05	301/05	038	F84	Fibula	Male	32	28	4	***nr***
175/05	176/05	039	F85	Ulna	Male	24	0	0	183
299/05	300/05	039	F86	Ulna	Male	24	28	4	161
335/05	336/05	039	F87	Ulna	Male	24	42	6	180
270/05	271/05	040	F88	Radius/Ulna	Male	20	0	0	217
264/05	265/05	040	F89	Radius/Ulna	Male	20	18	3	2
272/05	272/05	041	F90	Wrist	Male	19	0	0	197
321/05	322/05	041	F91	Wrist	Male	19	27	4	188
260/05	261/05	042	F92	Wrist	Female	41	0	0	175
250/05	251/05	043	F93	Tibial Plateau	Male	38	0	0	174
317/05	318/05	043	F94	Tibial Plateau	Male	38	8	1	258
347/05	348/05	043	F95	Tibial Plateau	Male	38	15	2	180
333/05	334/05	043	F96	Tibial Plateau	Male	38	43	6	235

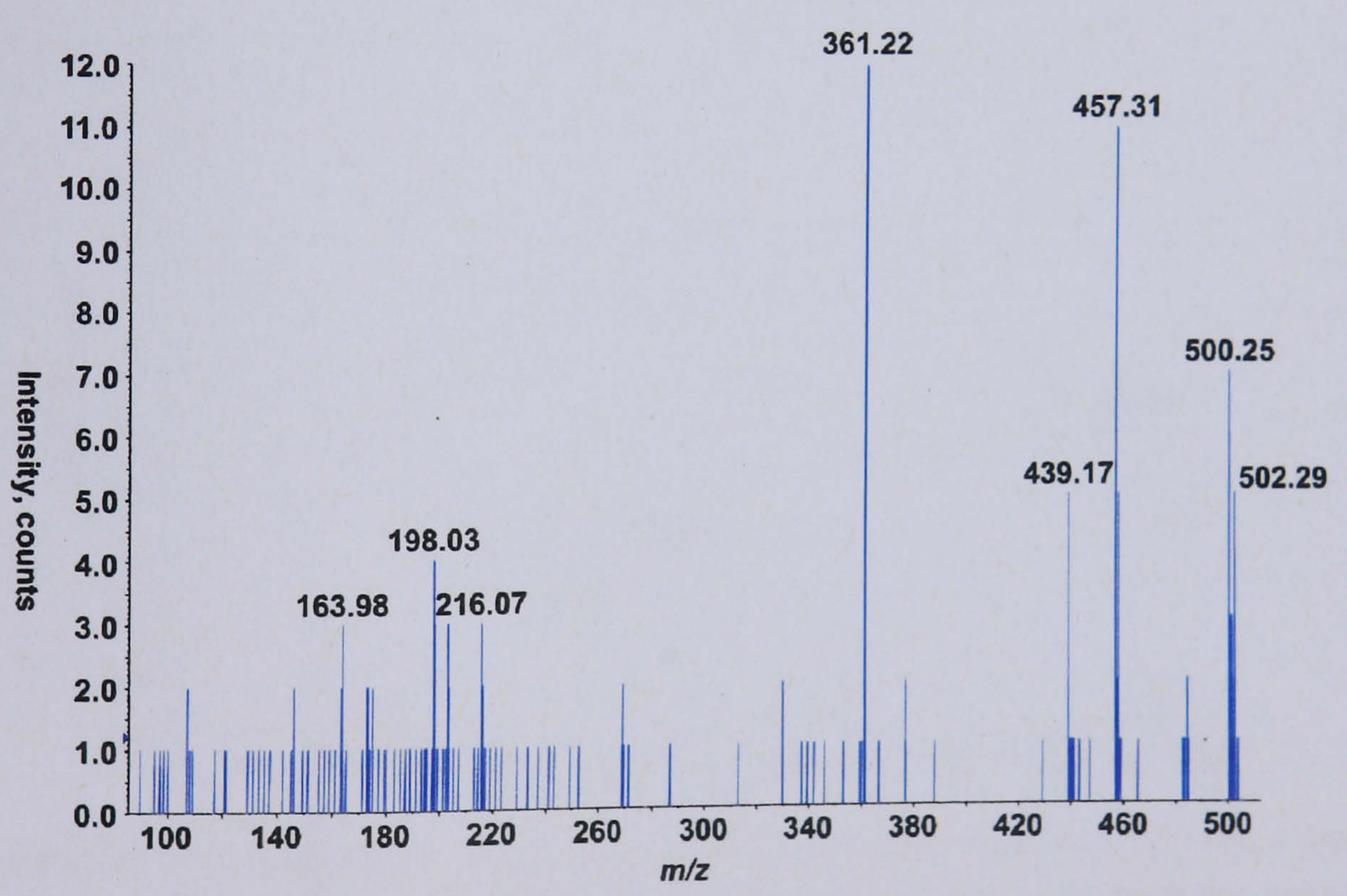
Internal S&N Code	Actual S&N Code	Patient ID	New Code	Fracture Type	Gender	Age	Days after Fracture	Weeks after Fracture	Protein Concentration (mg/mL)
268/05	269/05	044	F97	Radius	Male	30	0	0	177
325/05	326/05	044	F98	Radius	Male	30	13	2	233
339/05	340/05	044	F99	Radius	Male	30	41	6	192
258/05	259/05	045	F100	Fibula	Male	25	0	0	195
305/05	306/05	045	F101	Fibula	Male	25	7	1	201
341/05	342/05	045	F102	Fibula	Male	25	35	5	180
274/05	274/05	046	F103	Wrist	Female	32	0	0	206
307/05	308/05	046	F104	Wrist	Female	32	***unk***	***unk***	***nr***
274/05	314/05	047	F105	Gr. Tuberosity of Humerus	Female	40	0	0	217
317/05	320/05	047	F106	Gr. Tuberosity of Humerus	Female	40	48	7	200
309/05	310/05	048	F107	Clavicle	Male	34	0	0	193

Where ***unk*** = no data available, and ***nr*** = not recorded due to lack of sample.

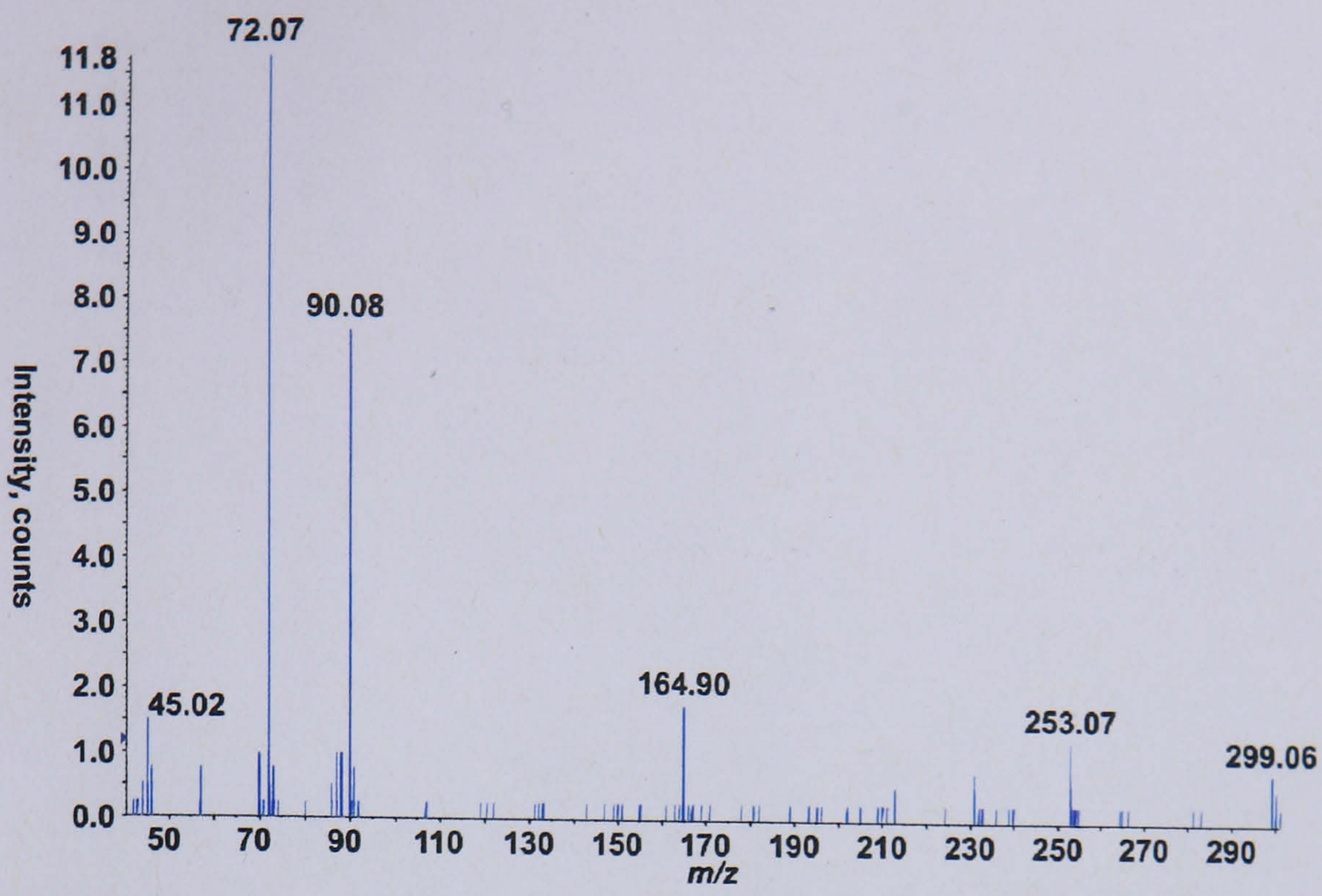
Appendix D



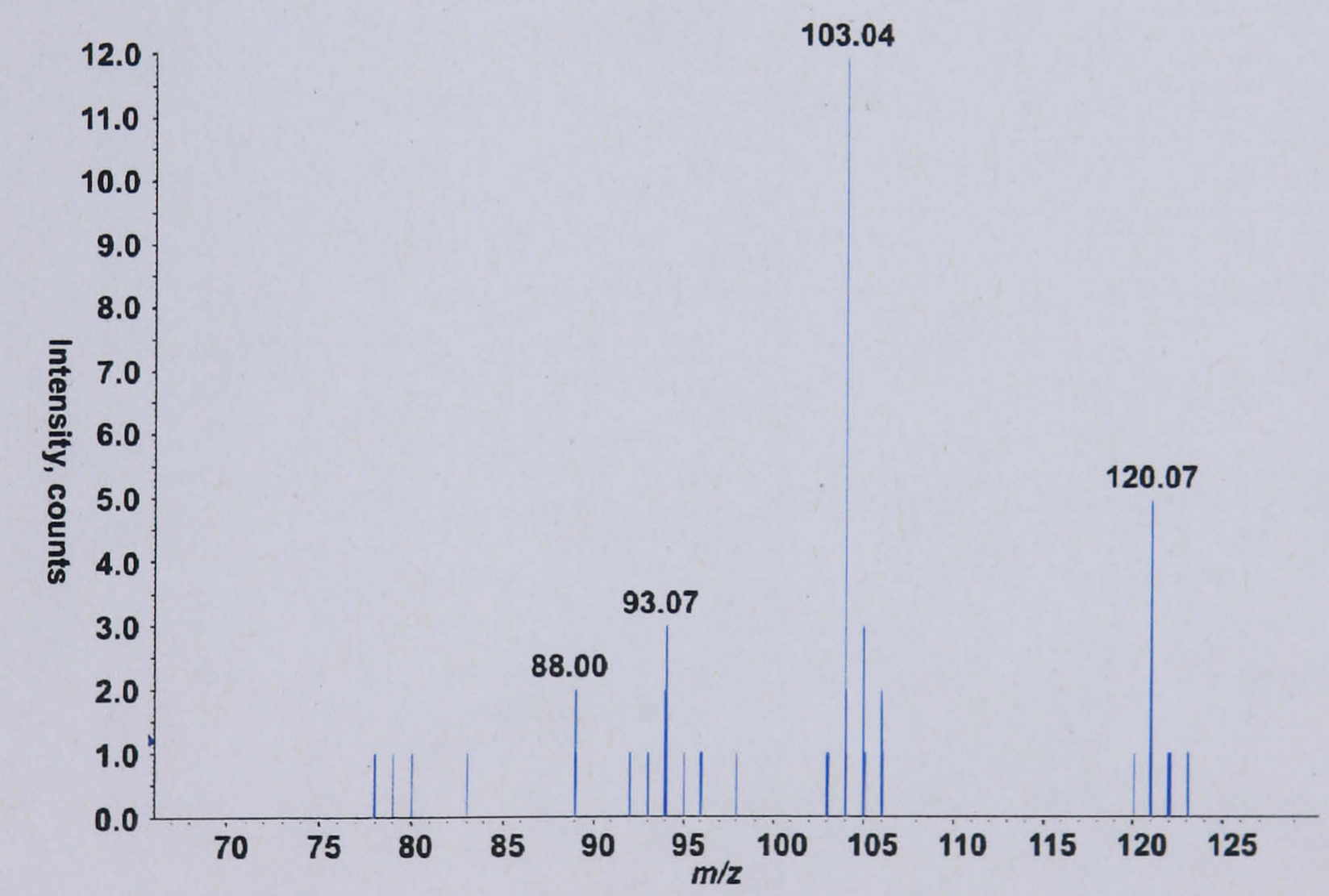
Appendix D1. Negative mode CID tandem MS of precursor ion *m/z* 528 from RP-LC-MS data frac2 PLS model. The peak at *m/z* 113 suggests the presence of a glucuronide, which would give the RMM of this metabolite to be 353.



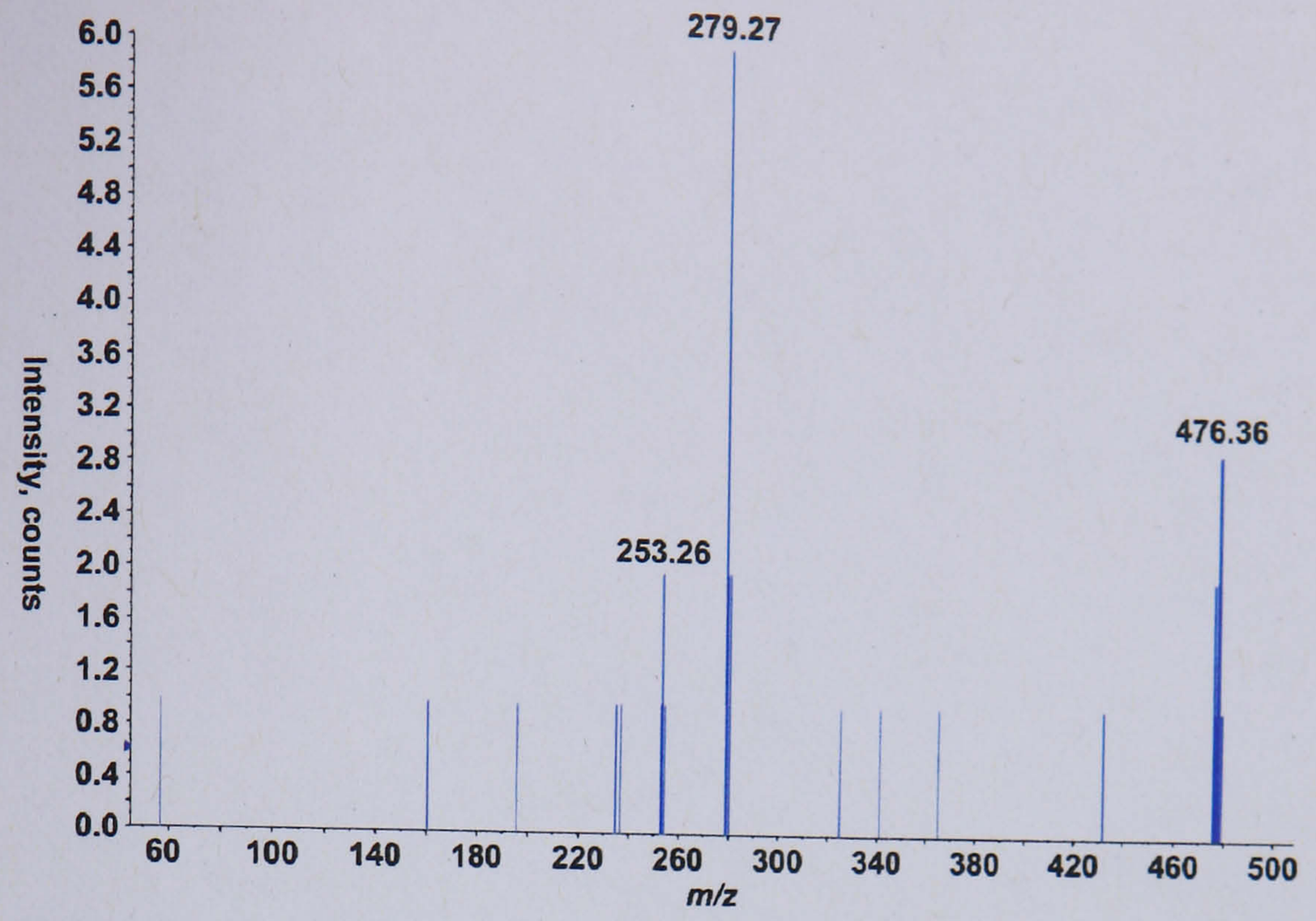
Appendix D2. Positive mode CID tandem MS of precursor ion *m/z* 500 from RP-LC-MS data frac3 PLS model. The presence of peaks at 2 Da higher suggests that chlorine is present within this metabolite.



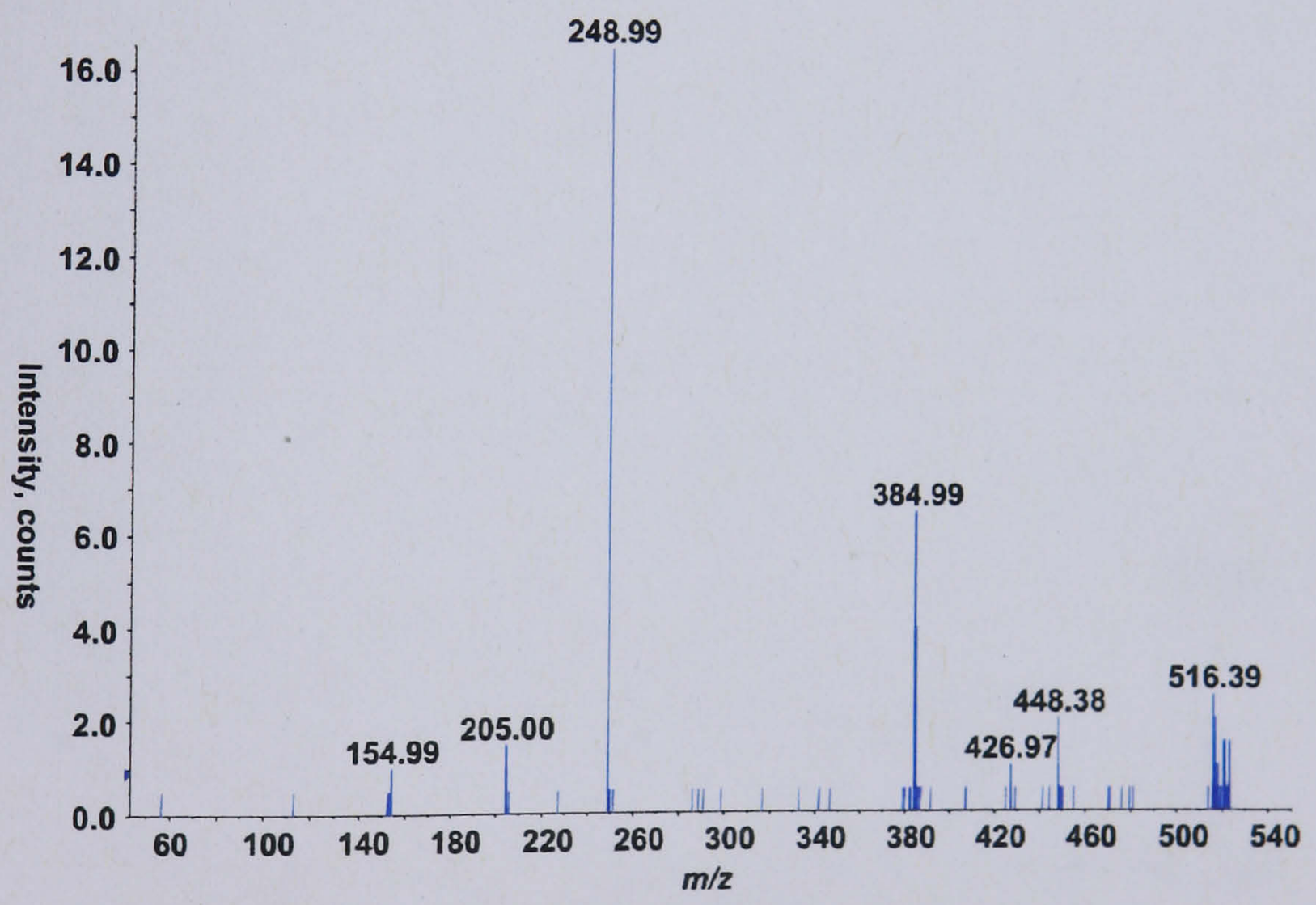
Appendix D3. Positive mode CID tandem MS of precursor ion *m/z* 299 from RP-LC-MS data frac3 PLS model. The peak at 18 Da less than the fragment ion at *m/z* 90 corresponds to the loss of water from this ion.



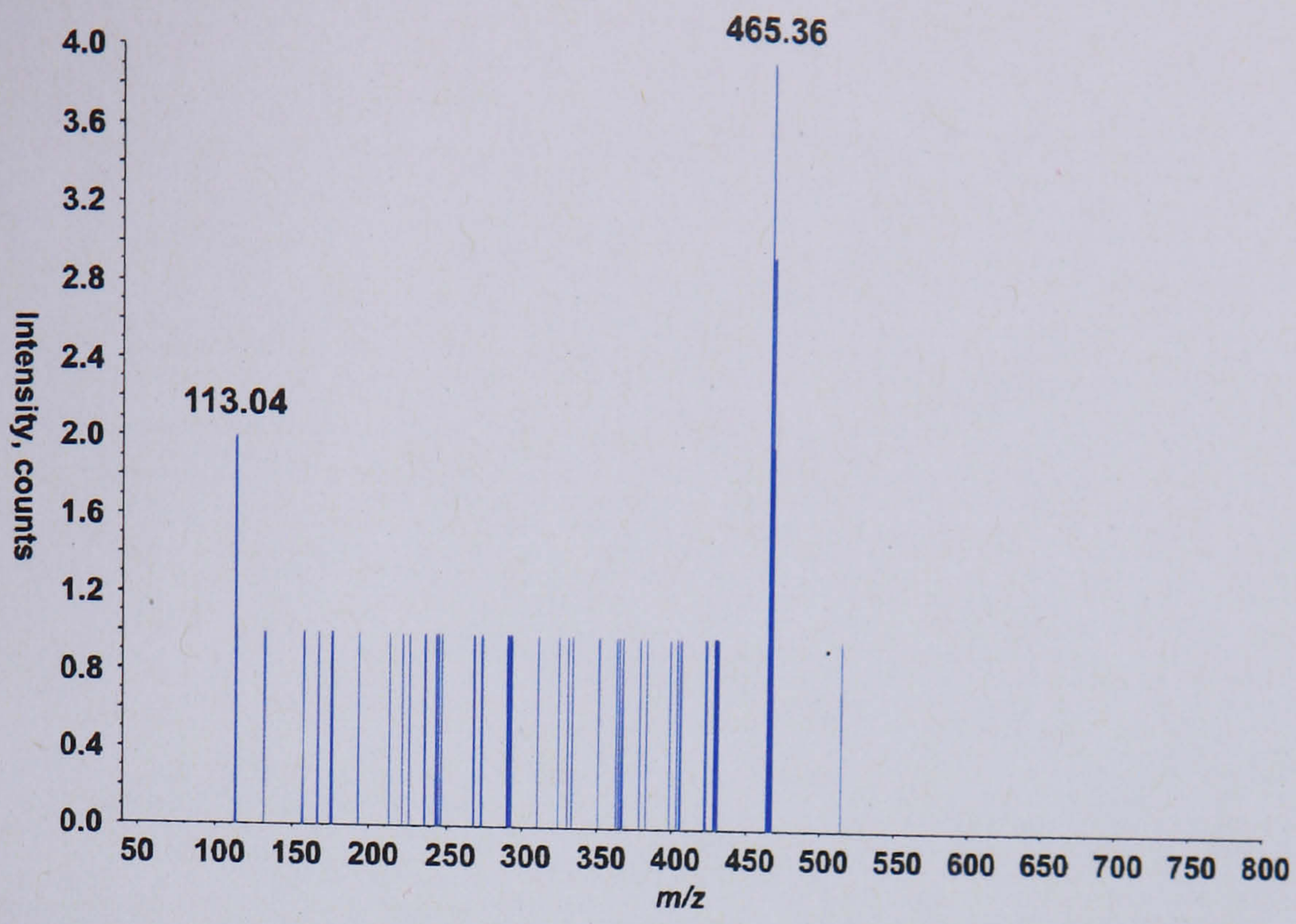
Appendix D4. Positive mode CID tandem MS of precursor ion *m/z* 120 from RP-LC-MS data frac3 PLS model. The peak at 17 Da lower than the precursor ion could correspond to the loss of NH_3.



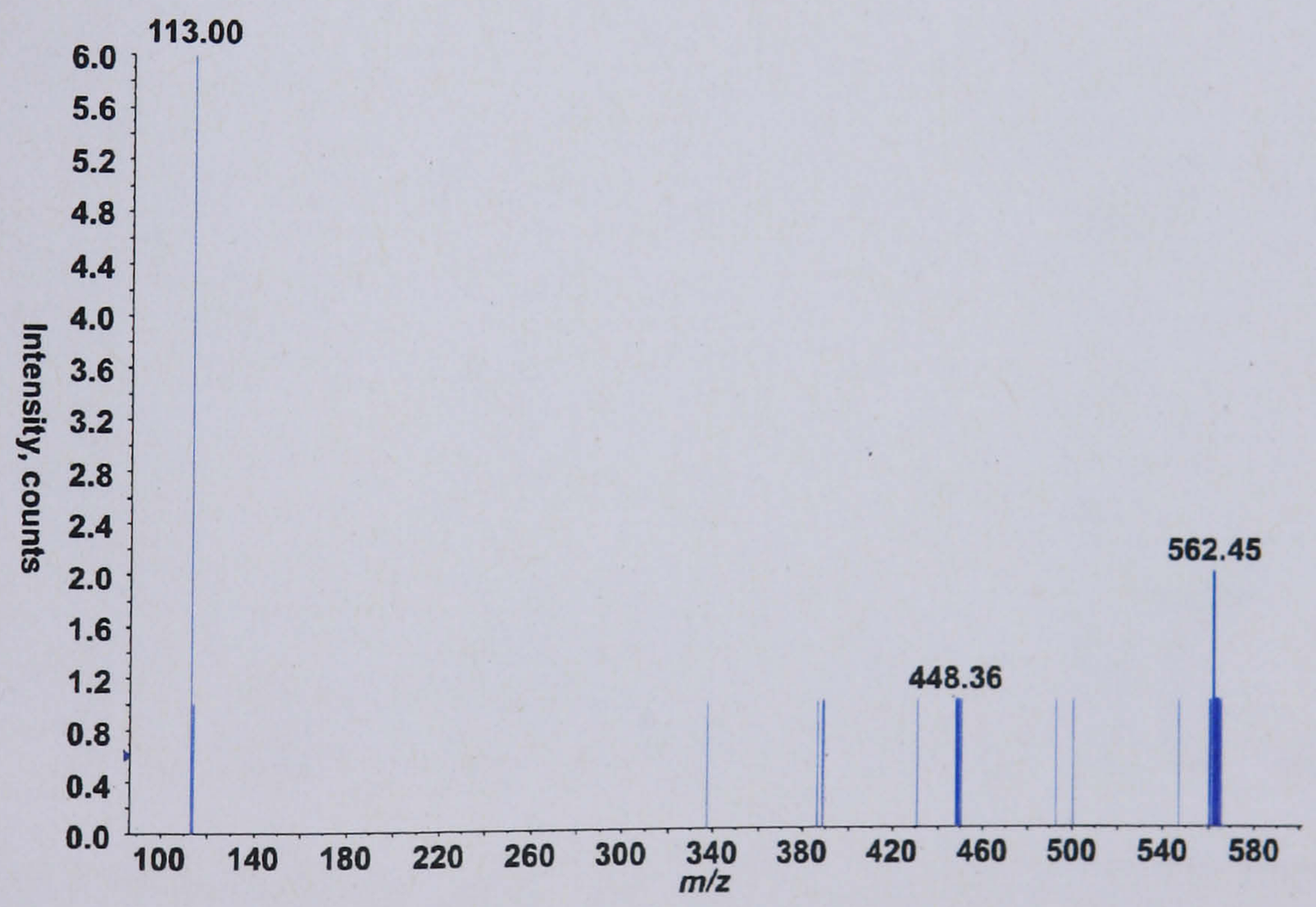
Appendix D5. Negative mode CID tandem MS of precursor ion *m/z* 476 from RP-LC-MS data frac3 PLS model.



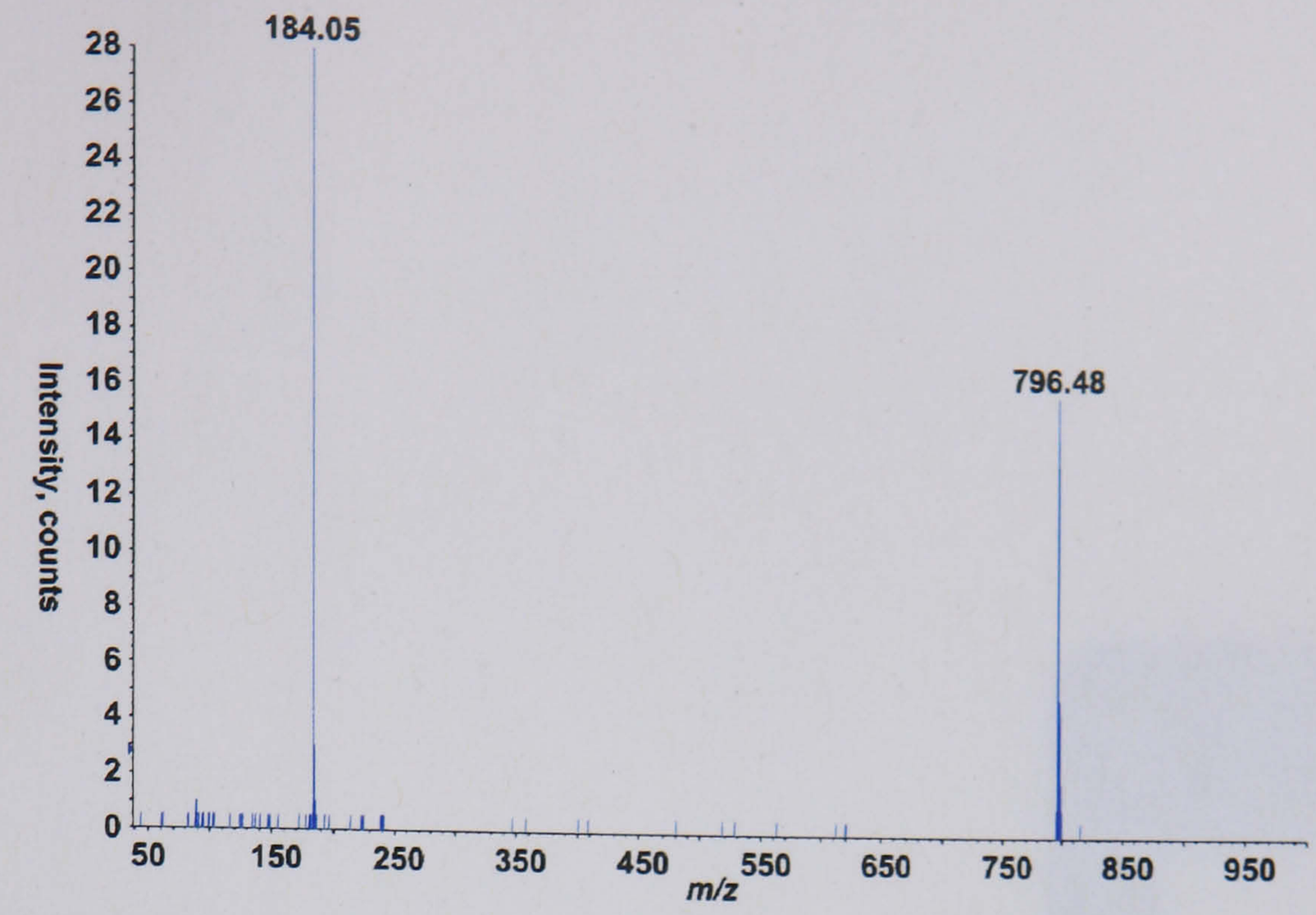
Appendix D6. Negative mode CID tandem MS of precursor ion *m/z* 516 from RP-LC-MS data ankle PLS model. The fragment ion at *m/z* 448 could correspond to the precursor ion of the same mass and retention time seen in the negative mode RP-LC-MS data for the frac2 PLS model, suggesting that the ion is in fact an in-source fragment.



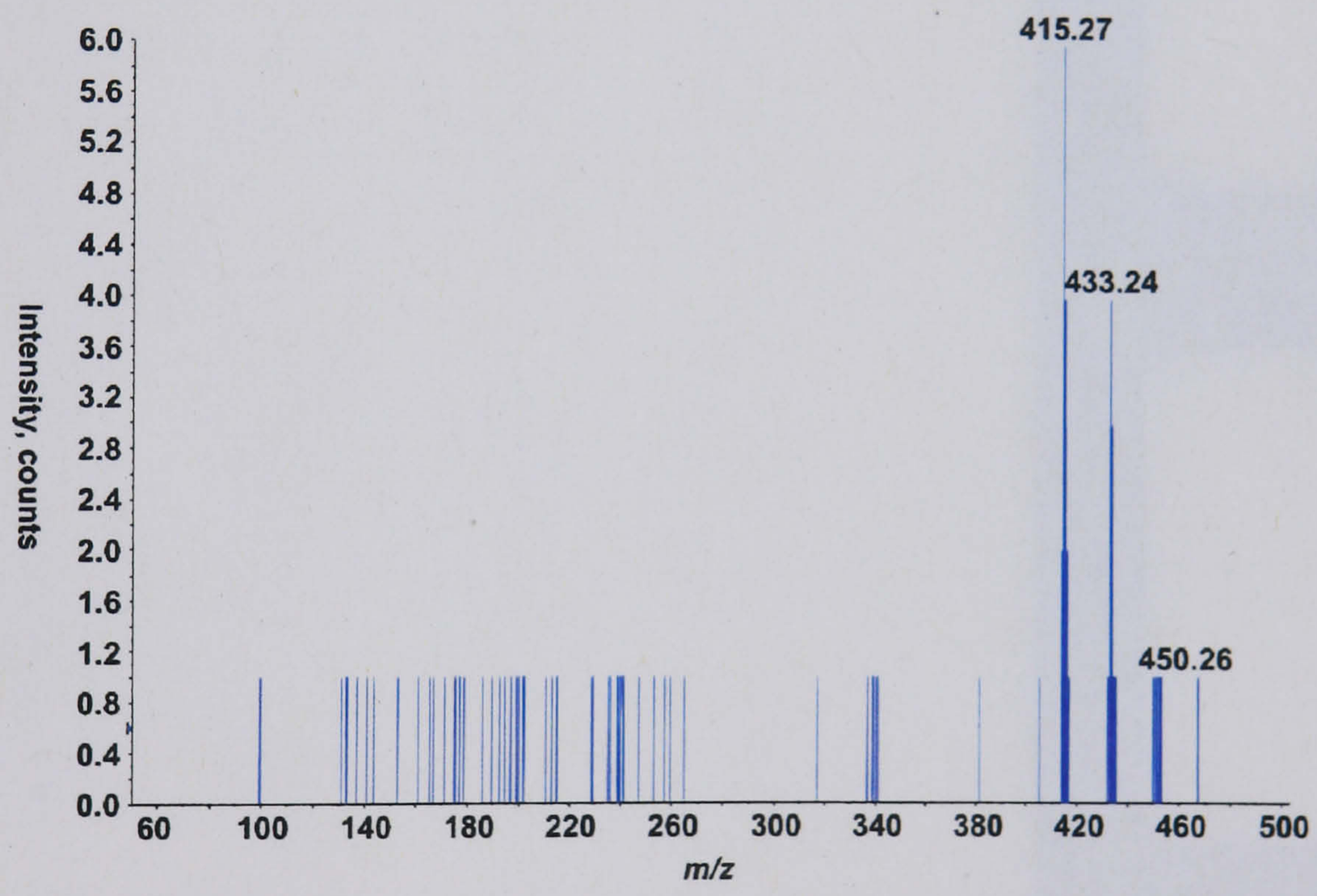
Appendix D7. Negative mode CID tandem MS of precursor ion *m/z* 465 from RP-LC-MS data ankle PLS model. The fragment ion at *m/z* 113 could correspond to a fragment from a glucuronide moiety.



Appendix D8. Negative mode CID tandem MS of precursor ion *m/z* 562 from HILIC-LC-MS data frac2, frac3 and ankle PLS models. The fragment ion at *m/z* 113 could correspond to a fragment from a glucuronide moiety, and the very low intensity fragment ion at *m/z* 448 is 114 Da less than the precursor ion, corresponding to a neutral loss of a glucuronide moiety. The fragment ion at *m/z* 448 could also correspond to an in-source fragment, as other precursor ions of the same mass and retention time were detected in negative ionisation mode data from both RP- and HILIC-MS data.



Appendix D9. Positive mode CID tandem MS of precursor ion *m/z* 796 from HILIC-LC-MS data frac3 PLS model.



Appendix D10. Positive mode CID tandem MS of precursor ion *m/z* 450 from HILIC-LC-MS data ankle PLS model. The mass of the precursor ion is 2 Da higher than *m/z* 448 fragment ion (and possible in-source fragment) detected from negative mode HILIC-MS data (see appendix D8); this could correspond to a protonated in-source fragment from the anionic glucuronide compound detected at *m/z* 562 (adding a loss of 114 Da to *m/z* 450, would correspond to an RMM of 563). The fragment ions at *m/z* 433 and 415 could correspond to the loss of NH_3 and water respectively.