Rapid Classification and Differentiation of Bacteria by Analytical Techniques

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2015

Nagla AlMasoud

School of Chemistry

Table of Contents

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Appendix

Word Count: 57,169 words

List of Figures

List of Tables

Abstract

The University of Manchester
Nagla AlMasoud
Doctor of Philosophy
Rapid Classification and Differentiation of Bacteria by Analytical Techniques
2015

Several traditional methods have been used to characterise bacteria, such as biochemical, morphological and molecular tests; however, these methods are time-consuming and not always reliable. Recently, modern analytical techniques have emerged as powerful tools offering high-throughput, reliable and rapid analysis in applications, such as clinical and microbiology studies. A variety of modern analytical techniques have been employed for bacterial characterisation, including matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS), liquid chromatography-mass spectrometry (LC-MS), Fourier transform infrared (FT-IR) spectroscopy and Raman spectroscopy. This thesis focused on developing a robust MALDI-TOF-MS methodology to generate mass spectra profiles for the discrimination of clinically-significant bacteria.

The data generated from MALDI-TOF-MS analysis are significantly influenced by a number of experimental factors, namely instrument settings, sample preparation, the choice of matrix, matrix additives and matrix preparation as well as sample-matrix deposition methods. The need to optimise experimental variables for bacterial analysis using MALDI-TOF-MS was evident despite the increased application of this analytical tool for clinical microbiology. Experimental optimisation revealed that the choice of matrix is the most important element in MALDI-TOF-MS analysis. Based on this study, a number of different matrices were used to obtain more reproducible mass spectra to classify bacterial samples using a rapid and effective approach. Studies in this thesis indicated that sinapinic acid (SA) is the best matrix for the analysis of proteins from intact bacteria, while 6-aza-2-thiothymine (ATT) and 2,5-dihydroxybenzoic acid (DHB) produced promising results for the analysis of lipid extracts from bacteria.

Analytical techniques in combination with multivariate analysis, such as principal components analysis (PCA) and principal component-discriminant function analysis (PC-DFA), were used for bacterial discrimination. Classification was initially undertaken using MALDI-TOF-MS analysis, and subsequently FT-IR spectroscopy, Raman spectroscopy and LC-MS were performed to confirm the classification results. Two main types of bacteria were used for this analysis: 34 strains from seven *Bacillus* and *Brevibacillus* species and 35 isolates from 12 *Enterococcus faecium* strains. The findings showed that the four analytical techniques provide clear discrimination between bacteria at these different levels. Classification of different *Bacillus* and *Brevibacillus* bacteria using MALDI-TOF-MS analysis of extracted lipids was confirmed by LC-MS data. In addition, MALDI-TOF-MS data based on extracted lipids and intact bacterial cell proteins were very similar. MALD-TOF-MS analysis of intact enterococci cells produced successful classification with 78% correct classification rate (CCR) at the strain level. FT-IR and Raman spectroscopic data produced very similar bacterial classification with CCR of 89% and 69% at the strain level, respectively. However, classification based on MALDI-TOF-MS data and that based on spectroscopic data were slightly different (Procrustes distance of 0.81, $p<0.001$, at the species level).

Overall, the findings in this thesis indicate the potential of MALDI-TOF-MS as a rapid, robust and reliable method for the classification of bacteria based on different bacterial preparations.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Dedication

To my beloved Husband, Mum, Dad, my Children, and my lovely Sisters and Brothers

Acknowledgments

First, I would like to thank Al-Mighty God for everything.

I would also like to express my gratitude to my supervisor Prof. Roy Goodacre for his supervision and for his support throughout my PhD study, providing me with ideas for experiments, valuable discussions and opportunities to further develop many of the skills that I have acquired during my academic career.

I extend my appreciation to Drs. Yun Xu and Elon Correa for their valuable help with the data analysis carried out in the work presented in this thesis.

I would also like to express my gratitude to Dr. Nicoletta Nicolaou-Markide for her help at the beginning of my PhD research. Special thanks also go to the members of my research group past and present; Dr. David Ellis, Dr. Nik Rattray, Dr. William Allwood, Dr. Catherine Winder, Dr. Lorna Ashton, Dr. Kat Hollywood, Dr. Drupad Trivedi, Dr. Haitham AlRabiah, Dr. Omar Alharbi, Dr. Howbeer Ali, Dr. Piotr Gromski, Ali Sayqal, Abdu Subaihi, Mekhala Spencer and Ewa Szula.

I would also like to express my appreciation, gratitude and thanks to my family; my mum, Fawzeya, who has given me happiness and has supported me with her prayers, my dad, Saad, who has always been my support and my role model, my mother in law, Monirah, for her warmth, kindness and continuous prayers despite her illness. My appreciation extends to my siblings, Norah, Mohammad, Ali, Hind and Fathmah for their continuous love, support and prayers. My gratitude also goes to my loving father in law, Abdulrahman, who sadly passed away last year after a battle with cancer.

I would also like to thank my close friend, Buthaina, who has been my support throughout my hard time adjusting to life away from home, and for her help with looking after my children when I was in the lab.

Most importantly, a huge thank you goes to my husband, Nasser, for his continuous support, patience, love, understanding and encouragement, not just in my academic career but also throughout our entire married life; without Nasser, I could not have achieved what I have today. Words cannot express my appreciation towards him. Finally, I would like to thank my little bundles of joy: Norah and Saad, for their endless love, entertainment, patience and joy that they bring our lives, making them complete.

Abbreviations and Acronyms

2D	Two dimensional
API	Atmospheric pressure ionisation
ACN	Acetonitrile
ANN	Artificial neural network
APCI	Atmospheric pressure chemical ionisation
ATT	6-aza-2-thiothymine
CA	Caffeic acid
CHCA	α-cyano-4-hydroxycinnamic acid
CID	Collision induced dissociation
CMPT	5-chloro-2-mercaptobenzothiazole
DFA	Discriminant function analysis
DGs	Diacylglycerides
DHAP	2,6-dihydroxyacetophenone
DHB	2,5-dihydroxybenzoic acid
DIMS	Direct infusion mass spectrometry
DNA	Deoxyribonucleic acid
EI	Electron ionisation
ELISA	Enzyme linked immunosorbent assay
EM	Electron multiplier
EMSC	Extended multiplicative signal correction
ESI	Electrospray ionisation
FA	Ferulic acid
FT-IR	Fourier transform infrared
FWHM	Full width at half maximum
GC	Gas chromatography
GRE	Glycopeptide resistant enterococci
HABA	2-(4-Hydroxyphenylazo) benzoic acid
HCA	Hierarchical cluster analysis
HILIC	Hydrophilic interaction liquid chromatography
HPLC	High performance liquid chromatography
HTP	High-throughput

INN	Dithranol
IR	Infrared
KE	Kinetic energy
L	Flight tube length
LB	Lysogeny broth
LC	Liquid chromatography
LDI	Laser desorption ionisation
m/z	Mass-to-charge ratio
MALDI	Matrix assisted laser desorption ionisation
MCP	Micro-channel plate
MVA	Multivariate analysis
MIR	Mid infrared
MS	Mass spectrometry
NA	Nutrient agar
NB	Nutrient broth
OD	Optical density
p	Probability
PA	Phosphatidic acid
PCs	Principal components
PC	Phosphatidylcholine
PCA	Principal component analysis
PC-DFA	Principal component-discriminant function analysis
PCoA	Principal coordinate analysis
PE	Phosphatidylethanolamine
PFGE	Pulsed field gel electrophoresis
PG	Phosphatidylglycerol
PLS	Partial least squares
PNA	*p*-nitroaniline
ppm	Parts per million
PS	Phosphatidylserine
QC	Quality control
RP	Reversed phase
S/N	Signal-to-noise ratio

SA	Sinapinic acid
SDS	Sodium dodecyl sulphate
T	Travel time
TEV	Total explained variance
TFA	Trifluoroacetic acid
THAP	2,4,6-trihydroxyacetophenone
TOF	Time of Flight
UHPLC	Ultra high performance liquid chromatography
UTI	Urinary tract infection
UV	Ultraviolet

Preface

A variety of analytical techniques, such as mass spectrometry and spectroscopy, have emerged as robust tools in many research areas, such as medical studies, drug development and discovery, environmental research and microorganism taxonomy. The studies carried out in this thesis aimed to develop a robust MALDI-TOF-MS data collection method for analysis of bacteria. The studies focused in particular on analysing different types of *Bacillus*, *Brevibacillus* and 12 *Enterococcus faecium* strains using different analytical techniques in combination with chemometrics. The outcomes of this investigation showed that these analytical techniques complemented each other for successful classification of bacteria. The analytical techniques used in these studies include: matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS), liquid chromatography in conjunction with mass spectrometry (LC-MS) in the positive ionisation mode, Fourier transform infrared (FT-IR) spectroscopy and Raman spectroscopy (RS). Data generated from these techniques were used to achieve the objectives of this thesis.

This thesis consists of six chapters. Chapter 1 is a general introduction to bacterial characterisation and analytical techniques. This chapter is followed by four results chapters and a final chapter on conclusions and future work. Chapter 2 is published in *Analytica Chimica Acta* and reports clear discrimination between 34 strains from the *Bacillus* genus encompassing seven different species based on analytical methods supported by multivariate analysis techniques. Chapter 3 is prepared as a manuscript to be submitted to *Analytical Chemistry*. The work therein focused on optimising the experimental conditions for detecting five lipids that were mixed together using MALDI-TOF-MS in combination with robust chemometrics to simplify and significantly reduce the huge number of potential experiments to be undertaken. Following on from this study, work in Chapter 4 was carried out using these optimum conditions to analyse lipids extracted from *Bacillus* bacteria using two different analytical techniques: MALDI-TOF-MS and LC-MS. Chapters 5 focused on classifying 35 isolates from *Enterococcus faecium* using a variety of different analytical techniques. Chapters 4 and 5 are prepared as manuscripts to be submitted to *Analytical and Bioanalytical Chemistry* and *Journal of Clinical Microbiology*, respectively.

Work carried out in the results chapters was in collaboration with colleagues and their work is acknowledged at the beginning of each chapter. These studies were challenging yet provided me with most enjoyable experiences. I gained valuable technical skills and expertise in scientific research, which will undoubtedly be useful in my future work. Last but not least, to my knowledge the findings and outcomes of the research carried out in this thesis have contributed useful methods to the classification of bacteria using modern analytical techniques.

Nagla AlMasoud

Chapter One

1. Introduction

1.1. Characterisation of bacteria

Successful bacterial characterisation requires the use of numerous techniques. The correct identification of bacteria is one of the fundamental foundations that link the fields of microbiology and infectious disease together (Janda and Abbott, 2002; Wilkins and Lay, 2005). Humans, animals and plants are in continuous contact with both saprophytic bacteria and harmful bacteria in their environmental surroundings, and hence the ability to identify and diagnose bacterial infections is absolutely vital (Peeling *et al.*, 2008; Sauer and Kliem, 2010). Diagnostics is required in many fields such as clinical microbiology, veterinary medicine and environmental studies (Coffey *et al.*, 1994; Sauer and Kliem, 2010). The importance of characterising and classifying bacteria lies in the ability to differentiate and group similar bacteria with each other making it easier for scientists to identify different bacterial types down to the species level, which has a particular significance in classification and characterisation studies (Yarza *et al.*, 2014; Nester, 2001; Parisi *et al.*, 2008).

Clinical, veterinary and environmental laboratories are forever striving to develop faster, low cost and reliable methods to characterise bacteria. Bacterial characterisation traditionally involves dividing and differentiating bacteria into groups depending on similarities in their cellular structures (Sintchenko *et al.*, 2007). In general, traditional methods, for example bacterial culture followed by staining and microscopic observations, are time-consuming and do not always provide conclusive findings (Luzzatto-Knaan *et al.*, 2015; Nomura, 2015; Wilkins and Lay, 2005). Therefore, newer technologies are being developed to reduce time and labour and improve classification results, and these methods include phenotypic and genotypic analytical techniques (Giebel *et al*., 2010; Wilkins and Lay, 2005), which rely on the analysis of phenotypic markers and genetic materials, respectively. Unlike genetic information, phenotypic markers used in bacterial characterisation may vary dramatically due to various environmental factors including conditions of cell culturing (growth times and medium), storage and sub-culturing. Bacteria can be characterised by means of simple methods, which facilitate the study of the different characteristics of cells such as their microscopic morphology and metabolic differences (Sauer and Kliem, 2010; Nester, 2001; Emerson *et al.*, 2008). An example of a method that relies on microscopic morphology is Gram staining a well-known simple staining method. The Danish physician, Hans Christian Gram, developed this staining method in the late 1800s and it can be used to classify bacteria based on their cell wall type into two main categories: Gram-positive and Gram-negative bacteria. These two main groups of bacteria are easily identified upon visual examination as Gram-positive bacteria stain purple whilst Gram-negative bacteria are counter-stained red (Beveridge, 2001).

Further technical developments to characterise bacteria were established due to the limitations of the traditional techniques, and these were classified into different broad areas, such as: deoxyribonucleic acid (DNA)-based methods (Olive and Bean, 1999; Olsen, 2000), immunoassay based methods (Emon and Lopez-Avila, 1992) and biochemical methods (Nester, 2001; Wilkins and Lay, 2005). DNA-based experimental methods usually require 24 hours to generate data since many steps are involved including: extraction and amplification of nucleic acid. Amplification is performed *in vitro* and is followed by gel electrophoresis and/or the use of hybridisation techniques for identification purposes. This technique is highly sensitive and accurate but relatively expensive in terms of equipment, reagents and training staff. Immunoassay and biochemical methods require more than 24 hours to obtain data and are less sensitive than DNA-based techniques (Nester, 2001; Davis and Mauer, 2010). Despite these disadvantages many researchers still find biochemical methods attractive due to their lower cost and ease of use.

In addition, other methods can be used to identify bacteria using nucleic acid based molecular techniques, such as polymerase chain reaction (PCR) (Holland *et al*., 2000; Emerson *et al.*, 2008; Das and Dash, 2014). This method has frequently been used to produce viable and robust data from bacterial samples due to its accuracy (Giebel *et al*., 2010; Wilkins and Lay, 2005).

Bacterial classification remains challenging due to the limitations of traditional techniques, such as staining and biochemical tests (Nomura, 2015; Sauer and Kliem, 2010), and therefore modern analytical techniques have the potential to provide more reliable classification. Further development and optimisation of these techniques for the purpose of bacterial classification constituted the motivation of this thesis. Modern analytical techniques include mass spectrometry (Gross, 2004;

Anhalt and Fenselau, 1975; Sauer and Kliem, 2010; Santos *et al.* 2105; Schumann *et al.* 2014; Karas an d Hillenkamp, 1988; Tanaka *et al.*, 1988; Claydon *et al.*, 1996) and spectroscopic techniques (Helm *et al.*, 1991; Kirschner *et al*., 2001; Kim *et al.*, 2005; Baker *et al.*, 2014; López-Díez and Goodacre, 2003; Naumann, 2001). These techniques provide powerful tools in the field of bacterial taxonomy due to their high specificity, analysis speed and cost effectiveness (Freiwald and Sauer, 2009; Sauer *et al.*, 2005; Aebersold and Mann, 2003). Modern analytical techniques such as MALDI-TOF-MS, LC-MS, Raman spectroscopy and FT-IR spectroscopy have undoubtedly gained importance in many microbiology laboratories including diagnostic laboratories due to the advantages that these techniques offer. The analytical techniques employed in this study are described in more detail in the sections below.

1.2. Analytical techniques

1.2.1. Mass spectrometry

Mass spectrometry (MS) is a powerful tool for analysing a variety of biological samples such as microbiological extracts, biofluids and tissue extracts, providing valuable information about the molecular mass of analytes and the contents of biological mixtures (Siuzdak, 1996; Sauer and Kliem, 2010; Watson and Sparkman, 2007). A mass spectrum is a two dimensional plot represented by the peak intensity of ions and their mass-to-charge ratio (*m/z*) (Williams and Fleming, 1995). The general concept behind mass spectrometry relies on two steps; analyte molecules are transformed into gas phase molecules with a single or multiple charges often at atmospheric pressure. Subsequently, the resultant ions travel through the mass analyser to the detector under vacuum. Mass spectrometers consist of three key parts: an ion source, a mass analyser and a detector (Figure 1.1) (Sauer and Kliem, 2010; Gross, 2004; de Hoffmann and Stroobant, 2007; Luzzatto-Knaan *et al*., 2015).

There are numerous mass spectrometry ionisation techniques with varying sensitivities. Examples of such techniques include matrix-assisted laser desorption/ionisation (MALDI) (Karas *et al.*, 1985; Krishnamurthy and Ross, 1996), electrospray ionisation (ESI) (Fenn *et al.*, 1989) and atmospheric pressure chemical ionisation (APCI). MALDI and ESI are the ionisation methods used in the present work.

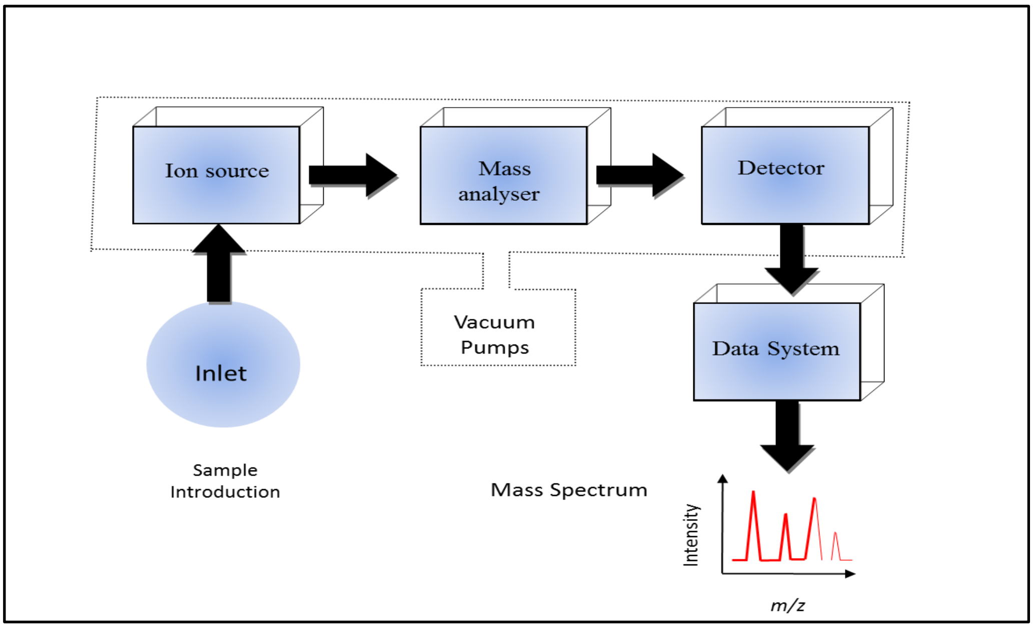


Figure 1.1: A flow diagram showing the principal components of a mass spectrometer.

A mass analyser is the heart of the mass spectrometer, which separates molecules according to their mass-to-charge (*m/z*) ratios. Mass analysers are known for their high sensitivity, accuracy, resolution and a broad mass range. There are many types of mass analysers and these include: time-of-flight (TOF) (Stephens, 1953) and Orbitrap mass analysers (Makarov, 2000). The TOF analyser has a broad *m/z* range capable of measuring the mass of low to high molecular mass compounds, such as peptides and proteins (Siuzdak, 1996; Kafka *et al.*, 2011; Watson and Sparkman, 2008), while the Orbitrap has a higher resolution (up to 120,000), high mass accuracy (less than 2 ppm) and has a wide dynamic range (Scigelova and Makarov, 2006).

The third component of a mass spectrometer is the detector, which is used to generate a signal from the passage of analyte ions with a specific *m/z* ratio. Microchannel plates (MCP) are the detectors usually used in mass spectrometers. Another infrequently used detector is the discrete dynode secondary electron multiplier (Siuzdak, 1996). Both detectors convert incoming ions into electrons, which are amplified by many orders of magnitude. MCPs are made of glass plates, which are densely aligned. Once an ion hits the front of the plate, electrons are released and cascade down the channels and exit out of the back. The amplified signals are then converted into a digital output using a computer. MALDI-TOF mass spectrometers use MCP detectors, while the Orbitrap detects the image currents that arise from the oscillation frequency of analyte ions as they circulate past the detecting electrodes (Scigelova and Makarov, 2006).

1.2.1.1.Matrix assisted laser desorption/ionisation mass spectrometry (MALDI-MS)

Laser desorption ionisation (LDI) is an analytical technique that relies on using a laser beam that targets the analytes and generates gas phase ions (de Hoffmann and Stroobant, 2007). LDI is a hard ionisation technique (Dreisewerd, 2003) hence it was not widely used as analytes degraded due to direct laser exposure. This prompted further development by introducing an energy-absorbing matrix, which led to a soft ionisation technique, matrix-assisted laser desorption/ionisation (MALDI) (de Hoffmann and Stroobant, 2007; Ellis *et al.*, 2007). MALDI-MS is a powerful, robust and a sensitive technique, which has routinely been used for the analysis of high molecular weight compounds such as proteins (Giebel *et al.*, 2010; Saenz *et al.*, 1999; Tanaka *et al.*, 1988; Karas and Hillenkamp, 1988; Burlingame *et al.*, 1996). This advanced analytical technique was first invented by Karas and co-workers in 1985 (Karas *et al.*, 1985). Their work was carried out on the analysis of a number of amino acids and dipeptides for the purpose of studying laser desorption. The developed technique allowed for singly protonated peaks to be observed with little or no fragmentation (Karas *et al.*, 1985).

MALDI-MS works by producing molecular ions when a laser beam is applied to analytes which are mixed prior to analysis with a matrix and air dried on a stainless steel MALDI plate. The matrix is a highly concentrated solution of low molecular weight organic molecules capable of absorbing laser energy. A nitrogen ultraviolet (UV) laser at 337 nm is typically used to excite the analyte/matrix mixture. Figure 1.2 below shows a schematic diagram of the MALDI-MS process (Giebel *et al.*, 2010; de Hoffmann and Stroobant, 2007; Sauer and Kliem, 2010).

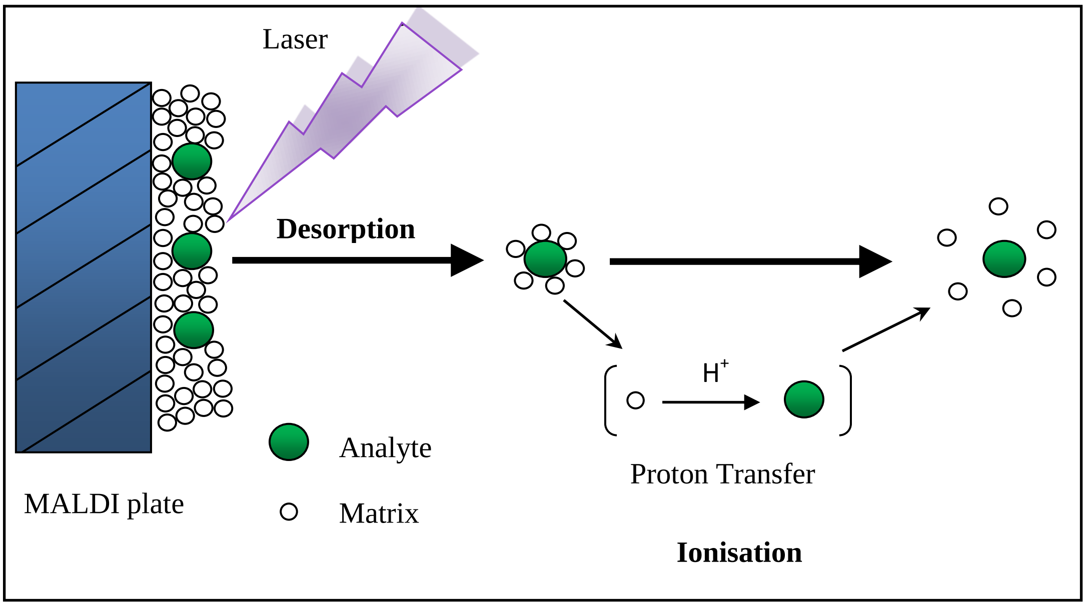


Figure 1.2: Schematic of the mechanism of matrix-assisted laser desorption/ionisation.

MALDI-MS is a soft ionisation method which can be used to analyse intact unicellular organisms and their components, such as proteins and lipids, a feature which is vital for many biological applications (Cramer *et al*., 2005; Saenz *et al*., 1999; van Baar, 2000; Lasch *et al*., 2009; Krishnamurthy and Ross, 1996; Sauer and Kliem, 2010; Claydon *et al*., 1996; Liu *et al*., 2007). MALDI-MS is capable of analysing heterogeneous samples without the need for laborious prior preparation and only small amounts of sample are required to perform routine analysis (de Hoffmann and Stroobant, 2007).

Matrices

In MALDI, sample and matrix molecules co-crystallise on a MALDI target plate, which is typically made of metal (Croxatto *et al*., 2012; Kafka *et al*., 2011; Nielen, 1999). The matrix is an essential component of this process as it absorbs UV laser energy directly (van Baar, 2000; Dreisewerd, 2003; Giebel *et al*., 2010) and then transfers it to analyte molecules, thereby protecting the analyte from laser-induced degradation (Dekker and Branda, 2011; Schumann *et al*., 2014).

Generally, matrices are acidic molecules (Lay Jr, 2000) that contain a conjugated double bond system capable of absorbing UV laser energy (Giebel *et al*., 2010; de Hoffmann and Stroobant, 2007). Choosing a suitable matrix for a specific application can be challenging (Marvin *et al.*, 2003). There is no single matrix, matrix deposition protocol or set of guidelines in the literature to assist in choosing the most compatible matrix and protocol for the analysis of microorganisms or other analytes. Hence, the matrix used in different applications is typically chosen on the basis of trial and error. The most suitable matrix used in any research study is highly dependent on certain factors, including the solubility of the matrix in different solvents and its ability to absorb laser energy at the wavelength used in the MALDI-MS device (Croxatto *et al.*, 2012; Kafka *et al.*, 2011; Ashcroft, 2003; Nielen, 1999).

Commonly employed matrices for the analysis of proteins and peptides are α-cyano-4-hydroxycinnamic acid (CHCA), sinapinic acid (SA) and ferulic acid (FA) (Giebel *et al*., 2010; Fenselau and Demirev, 2001). On the other hand, 2,5-dihydroxybenzoic acid (DHB), 2,4,6-thirhydroxacetophenone (THAP) and 6-aza-2-thiothymine (ATT) work well for the detection of low molecular weight compounds (Griffiths *et al*., 2012; Giebel *et al*., 2010; Stübiger *et al*., 2007; Shanta *et al.,* 2012). Figure 1.3 shows the distribution of two different matrices, SA and THAP with sample.

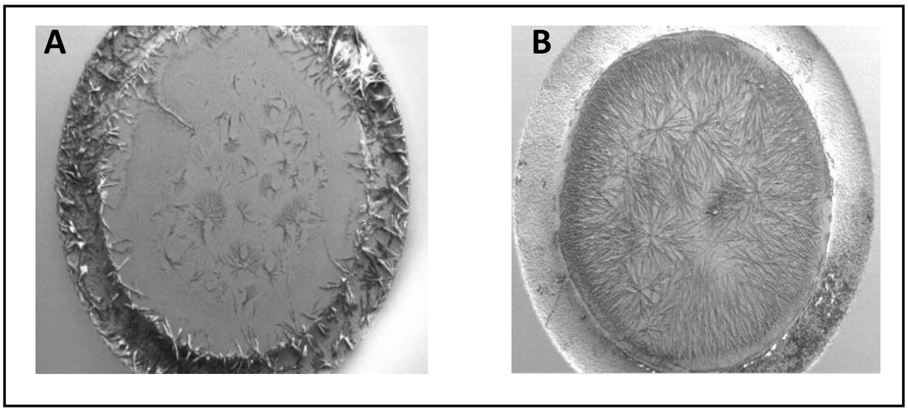
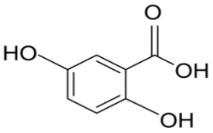
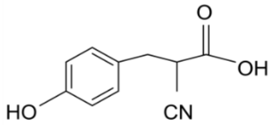
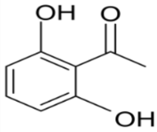
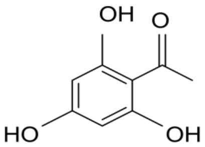
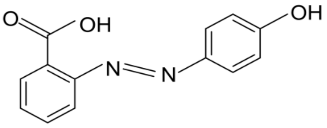
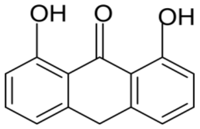
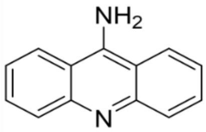
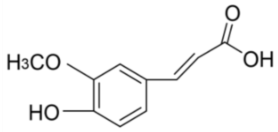
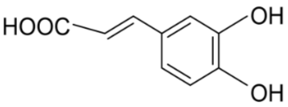
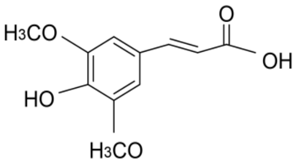


Figure 1.3: Distribution of two different matrices with sample: (A) THAP and (B) SA on MALDI stainless plate examined using scanning electron microscopy.

Table 1.1: Some of the commonly used matrices with their molecular formulae

Name of Matrix	Structure	Abbreviation
2,5 dihydroxybenzoic acid		DHB
α-cyano-4-hydroxycinnamic acid		CHCA
2,6 dihydroxyacatophenone		DHAP
2,4,6-thirhydroxacetophenone		THAP
2-4 (hydroxphenylaze) benzoic acid		HABA
1,8,9-trihydroxy-anthracene, dithranol		INN
9-aminoacridine		9-AA
ferulic acid		FA
caffeic acid		CA
sinapinic acid		SA

Sample deposition methods

There are several sample deposition methods that can be used for MALDI-MS analysis (Kussmann *et al.*, 1997). Co-crystallisation occurs when the matrix and analyte are homogenised causing the formation of what are known as sweet or hot spots (Zenobi and Knochenmuss, 1998). Sample deposition methods are usually easy to follow. The most common methods used for sample preparation include mix, overlay, underlay and sandwich methods. The mix method involves mixing the matrix and the analyte on the MALDI plate, the overlay method involves spotting the analyte first followed by the matrix, the underlay method is the opposite of the overlay method and the sandwich method involves sandwiching the sample between two layers of the matrix (Giebel *et al.*, 2010; Kafka *et al.*, 2011; Nielen, 1999; Liu *et al.*, 2007).

Time of flight (TOF) mass analyser

One of the most commonly used mass analysers with MALDI-MS is the time-of-flight (TOF) mass analyser (Ashcroft, 1997; El-Aneed *et al.*, 2009). Stephens was the first to publish the principles of TOF (Stephens, 1953). Briefly, the ions generated using the MALDI ion source are accelerated by means of high voltage and then travel along the flight tube to the detector (Ekman *et al.*, 2008; de Hoffmann and Stroobant, 2007). The ions are separated in the flight tube and reach the detector at different times; smaller ions are detected first followed by larger ones, which have lower velocity and therefore need more time to traverse the length of the flight tube. The time required for an ion to arrive at the detector is measured and its value is proportional to the *m/z* and kinetic energy (KE) of ions (Watson and Sparkman, 2008). A schematic diagram of the TOF method is shown in Figure 1.4 (Ekman et al., 2008; de Hoffmann and Stroobant, 2007). The time of flight can be calculated by using Equation 1.1:

$$\begin{aligned} \boldsymbol{t}_{\mathbf{TOF}} &= \frac{\boldsymbol{L}}{\boldsymbol{V}} \\ &= \mathbf{L}\sqrt{\frac{\boldsymbol{m}}{\boldsymbol{2qUa}}}\ \boldsymbol{\alpha}\sqrt{\boldsymbol{m/z}} \end{aligned} \qquad \mathbf{(1.1)}$$

where t_{TOF} represents the travel time (t) needed for an ion to fly from the ion source to the detector, L represents the flight tube length, v denotes the ion velocity after acceleration, m/z denotes the mass-to-charge ratio of the ion, U_a is the electric potential difference that causes the ion to accelerate after leaving the ion source, and q corresponds to the charge on the ion (de Hoffmann and Stroobant, 2007; Ekman *et a*l., 2008).

There are several advantages of using TOF as a mass analyser including having the highest mass range in comparison to other mass analysers (Cotter, 2013; El-Aneed *et al.*, 2009), having a very fast scan speed (Siuzdak, 1996), its simple design, being relatively inexpensive and being easy to use in conjunction with MALDI. However, despite the many advantages offered by TOF analysers, the main disadvantage of using this technique is its low mass resolution. The reflectron technique was developed to overcome this problem.

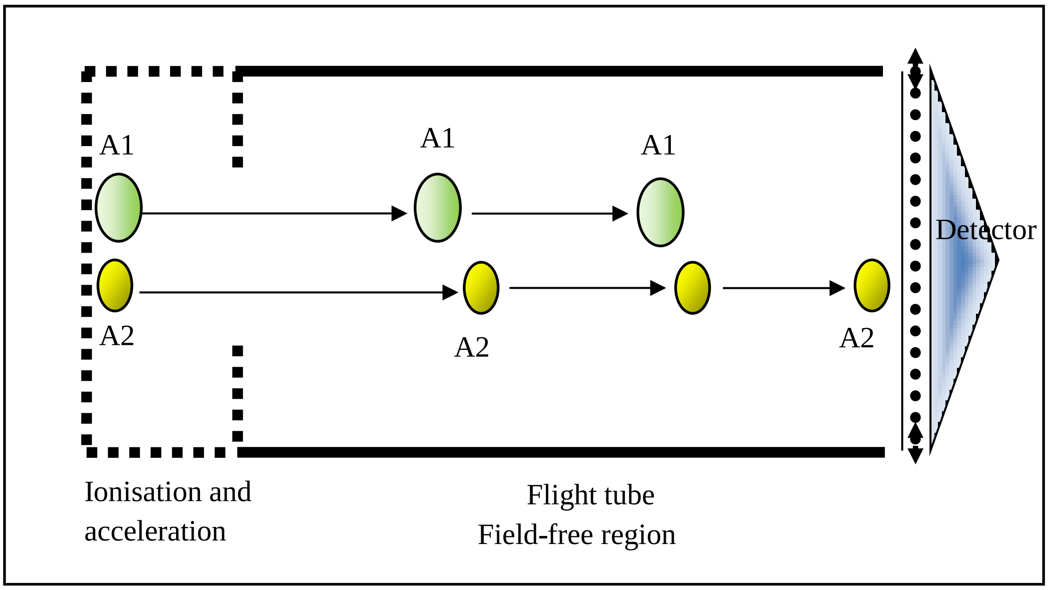


Figure 1.4: Schematic of a time-of-flight mass analyser, where A1 and A2 are ions of different *m/z* values, with A2 having the smaller *m/z*. The larger molecule (A1) requires longer time to reach the detector.

The reflectron

The reflectron and reflector refer to the same technology (Hillenkamp and Peter-Katalinic, 2013). Mamyrin and colleagues (1973) were the first to describe the use of this technique. Briefly, a series of grids and ring electrodes are placed after the field free region and act as an ion mirror, leading to increased length of the flight path and a decrease in variations in the kinetic energy of ions with the same *m/z* value (de Hoffmann and Stroobant, 2007; Siuzdak, 1996). Ions with higher KE travel more deeply into the opposing electrical field since they have higher velocities and they continue travelling until their KE reaches zero (Ekman *et al.*, 2008). Conversely, ions with lower KE are reflected more quickly towards the detector. Hence, ions with the same *m/z* but different initial KE reach the detector at the same time and are assigned the same *m/z* value. Using the reflectron results in increased resolution and prevents peak broadening due to the increased length of the flight path (Ekman *et al.*, 2008; de Hoffmann and Stroobant, 2007). The mechanism of this technique is summarised in Figure 1.5.

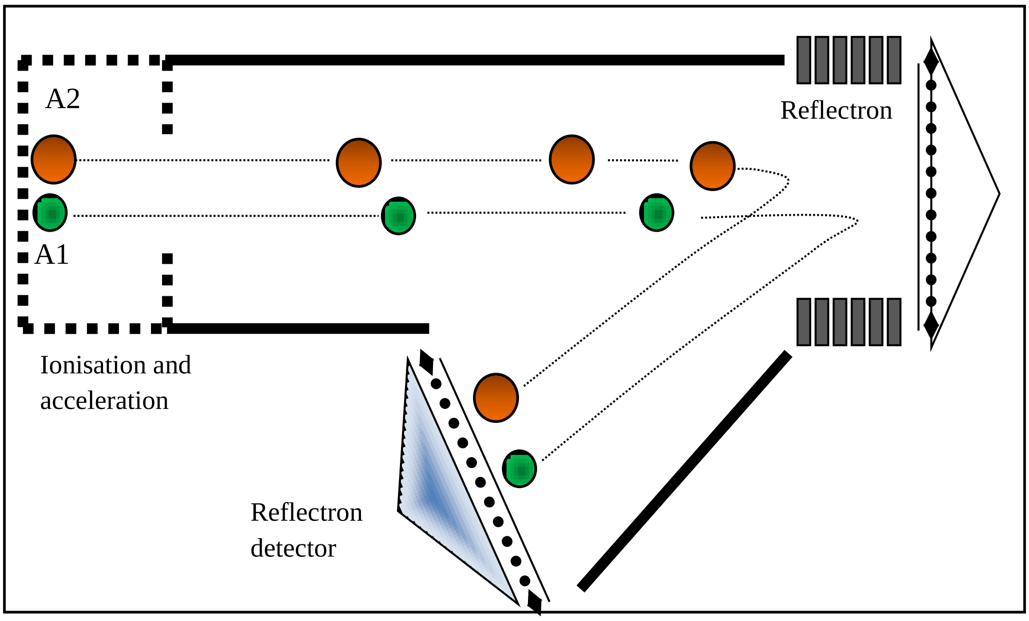


Figure 1.5: Schematic diagram of MALDI-TOF-MS reflectron mode: A1 and A2 correspond to ions with the same *m/z* but different kinetic energies (KE) making A1 travel faster than A2. The reflectron reduces variations in initial KE between ions of the same *m/z*. A1 and A2 arrive at the detector simultaneously.

MALDI-TOF-MS applications

Due to its many advantages, MALDI-TOF-MS has been applied in a number of research fields including cell biology, proteomics, lipidomics, medical research, health and safety and the food industry (Croxatto *et al.*, 2012; Cobo, 2013; Lay, 2001; Schiller *et al.*, 2004; Fuchs and Schiller, 2009; Stults, 1995; Santos *et al.* 2105). Despite the popularity of this technique, there are several factors that need to be taken into consideration for the successful application of MALDI-TOF-MS: these include sample concentration, the matrix, matrix solvent, deposition method and ionisation mode (Giebel *et al*., 2010; Kafka *et al.*, 2011; Šedo *et al.*, 2011).

One of the earliest studies that used MALDI-TOF-MS for the analysis of bacterial samples was carried out by Cain and colleague who reported that MALDI-TOF-MS can be used to distinguish between bacterial samples based on the proteins extracted from disrupted cells, leading to successful discrimination between Gram-positive and Gram-negative bacteria (Cain and Lubman, 1994). MALDI-TOF-MS was also used directly to analyse bacterial samples by Claydon and co-workers (1996) who examined 10 different bacterial species. Subsequently, other research groups used MALDI-TOF-MS for the analysis of different microbiological samples, such as whole cell analysis of different types of bacteria (Holland *et al*. 1996; Saenz *et al*., 1999; Lasch *et al.*, 2009; Krishnamurthy and Ross, 1996; Williams *et al.*, 2003) and analysis of bacterial lipid and protein extracts (Gidden *et al.*, 2009; Santos *et al.* 2105; Lasch *et al.*, 2014). One of the most recent studies that reported the application of this technique to microbiological research involved the use of MALDI-TOF-MS to optimise bacterial sample preparation and analysis protocols (Šedo *et al.*, 2011).

It is evident from the literature on MALDI-TOF-MS applications in different research fields, including microbiological research that this analytical tool can be used successfully to investigate a variety of biological processes at different cellular levels (van Baar, 2000; Santos *et al.* 2105; Gidden *et al.*, 2009; Kafka *et al.*, 2011).

1.2.1.2 Liquid chromatography-mass spectrometry (LC-MS)

Liquid chromatography (LC) can be used to separate compounds in complex mixtures. Separation is based on the affinity of molecules between two phases: a stationary phase and a mobile phase. Compounds that are attracted to the stationary phase will elute slowly, resulting in longer retention times, whereas compounds that are attracted to the mobile phase will elute quickly and hence have shorter retention times (Waston, 1999; Allwood and Goodacre, 2010).

LC has been widely used for analysing many low molecular mass compounds such as lipids, sugars and bile acids (Dunn, 2008). LC comes in different forms with the most common type used in lipidomics being reserved phase high performance liquid chromatography (RP-HPLC). Recent studies, however, have suggested the utility of hydrophilic interaction liquid chromatography (HILIC) (Buszewski and Noga, 2012), commonly used for the analysis of small polar molecules. The field of lipidomics takes advantage of the use of RP-HPLC in conjunction with MS (LC-MS) allowing lipids to be both separated then resolved based on their mass-to-charge ratio for better characterisation. In addition, advances in column chemistries and bonded phases make HPLC one of the best separation options for the analysis of various hydrophobic and hydrophilic compounds. The added bonus of using HPLC over gas chromatography (GC), for example, is that GC can only be used for volatile compounds whereas HPLC only requires the compounds to dissolve in a liquid medium (Iseman, 1993).

For many research fields such as lipidomics, there is a need to use a gradient of mobile phase for LC analysis. In RP-HPLC, for example, this phase starts with a high proportion of an aqueous phase and terminates with a high proportion of an organic solvent, such as methanol or acetonitrile. Polar compounds are eluted rapidly in this mobile phase followed by the elution of non-polar compounds (Allwood and Goodacre, 2010; Waston, 1999).

Finally, one of the most common soft ionisation techniques used in conjunction with LC is electrospray ionisation (ESI). This ion source is best used for analysed different types of metabolites such as amino acid (Tolstikov and Fiehn, 2002) and phospholipids (Allwood *et al.*, 2006). In general, there are different mass analysers that can be used with LC such as quadrupoles, TOF mass analysers and the Orbitrap (Allwood and Goodacre, 2010).

Electrospray ionisation (ESI)

Electrospray ionisation (ESI) was pioneered by Dole and co-workers in 1968 (Dole *et al.*, 1968), and was first coupled to a mass spectrometer in 1984 by Yamashita and Fenn. This technique is used as a soft ionisation method to generate ions from biological samples (de Hoffmann and Stroobant, 2007). The sample ions are transferred from a liquid solution to the gas phase by ion evaporation with little fragmentation (Ekman *et al.*, 2008). This is accomplished by generating a fine spray of highly charged droplets in an electric field, allowing analytes of high molecular mass to have the desired mass-to-charge ratio for the MS range of analysis (Gaskell, 1997). In this technique, a solution of the sample is allowed to pass through a stainless steel capillary (typical internal diameter of 75-100 μm) upon which a high voltage (3-6 kV) is applied to form a continuous spray of charged droplets in the form of a Taylor cone (Ashcroft, 1997). These charged droplets are subjected to either a drying inert gas (usually nitrogen) or heat to evaporate the solvent, resulting in the release of highly charged ions of the analyte (Watson and Sparkman, 2007; Siuzdak, 1996, Ekman *et al.*, 2008).

ESI has been used in the analysis of many biological samples including detection of proteins (Fenn *et al.*, 1989; Vaidyanathan *et al.*, 2004; Yates *et al.*, 2009), lipid characterisation (Brugger, 2014; Eberlin *et al.*, 2011) and analysis of intact microorganisms (Goodacre *et al.*, 1999). This is due to several advantages associated with this technique including its wide mass range (100 Da to 100 MDa) in both negative and positive ionisation modes, good sensitivity and the formation of multiple ions needed for the analysis of high molecular mass analytes such as proteins. Moreover, ESI is a soft ionisation technique with limited fragmentation of molecules under analysis (Sauer and Kliem, 2010; Ekman *et al.*, 2008). On the other hand, ESI also has a few disadvantages. For example, the signal of the analyte can be supressed because of competition between analytes for ionisation (ion suppression). Another complication associated with this technique is adduct formation, which renders ESI difficult to use with complex mixtures (Ekman *et al.*, 2008).

1.2.2. Vibrational spectroscopy

Vibrational spectroscopy involves the use of several methods to measure the vibrations and rotations of functional groups within a molecule. These vibrations are due to an exchange in energy as a result of the radiation interacting with the sample. As a result of this interaction, molecular energy is increased which can lead to the predication of three various transitions including: electronic excitation, vibrational change and rotational change (Dunn *et al.*, 2005). The incident radiation wavelength determines the type of event that will occur. Infrared (IR) spectroscopy and Raman spectroscopy are examples of vibrational spectroscopic techniques (Ellis *et al.*, 2007; Dunn *et al.*, 2005). These techniques can be used in many applications, such as the identification and characterisation of many biological samples including bacteria (Kirschner *et al.*, 2001; López-Díez and Goodacre, 2003; Guibet *et al.*, 2003). These two techniques are discussed in more detail below.

1.2.2.1. Fourier-transform infrared spectroscopy

Infrared (IR) spectroscopy is a very powerful tool that has been used for decades for the analysis of many biological molecules. This technique was first made available commercially in the 1940s, in which prisms were used to disperse infrared light. However, many medical and biological researchers avoided using this technology due to its low sensitivity, reproducibility and the prolonged periods of time required for the analysis of biological samples (Stuart, 1996; Dunn *et al.*, 2005). Developments in this field are largely due to the introduction of mathematical processes, which led to improvements in the quality of the IR spectra and significantly reduced the time of analysis. This improved type of IR spectroscopy is referred to Fourier transform infrared (FT-IR) spectroscopy (Stuart, 1996). FT-IR spectroscopy has become a popular method used for screening many biological and medical samples due to the many advantages that this technique offers, including the possibility of high throughput (fast analysis time) and most importantly its non-destructive effect on samples (Davis and Mauer, 2010).

In FT-IR spectroscopy, the samples of interest are subjected to an infrared beam, which allows the functional groups and polar bonds to absorb light in a specific region of the spectrum. A change in the dipole moment is observed as a result of this absorption by different bonds, for example C=O, O-H and N-H, leading to various vibrational characteristics such as bending, stretching and rotating (Stuart, 1996; Ellis *et al.*, 2007). On the other hand, some molecules such as O_2 and N_2 are not detected as there is no change in the dipole moment of their bonds (Stuart, 1996; Harrigan and Goodacre, 2003).

Infrared spectra are divided into three main regions; the far (less than 400 cm^{-1}), the mid (MIR) (4000-400 cm^{-1}), and the near-infrared (NIR) (14285-4000 cm^{-1}) regions.

Many applications employ the use of the mid-IR region area of the infrared spectrum since it provides rich information about the biochemical structure of compounds and the most fundamental vibrations are observed as sharp peaks in this region (Stuart 1996; Lu *et al*., 2011; Dunn *et al.*, 2005; Ellis *et al.*, 2007; Harrigan and Goodacre, 2003; Lasch *et al.*, 2002). Figure 1.6 shows a spectrum produced by FT-IR spectroscopy. FT-IR spectroscopy is known to be a relatively inexpensive, rapid, reproducible, and sensitive method that can be used for fingerprinting purposes (Stuart, 1996; Baker *et al.*, 2014). Moreover, FT-IR spectroscopy has the extremely useful advantage of high-throughput screening analysis of a large number of samples (hundreds/thousands per day). Due to these advantages, FT-IR spectroscopy has been used extensively in many research applications such as biological studies including identification of microorganisms (Naumann *et al.*, 1991; Helm *et al.*, 1991; Mariey *et al.*, 2001; Kirschner *et al.*, 2001). However, the two main disadvantages of using this spectroscopy technique are: (1) water is absorbed in the MIR however; this drawback can be avoided by drying the samples before analysis is carried out; (2) its limited specificity and sensitivity compared to other techniques available such as mass spectrometry coupled with chromatography or stand-alone high resolution mass spectrometry (Dunn *et al.*, 2005).

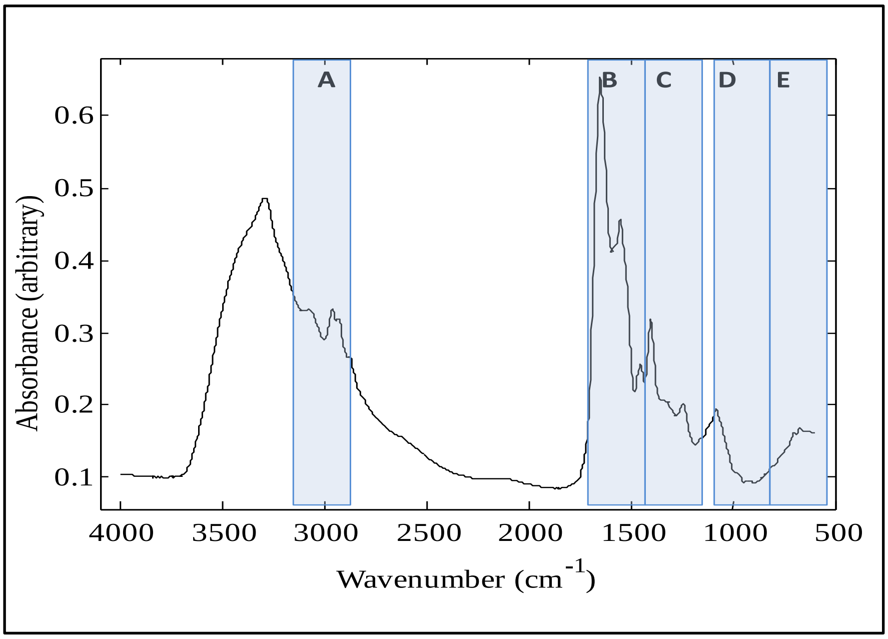


Figure 1.6: A typical Fourier transform infrared (FT-IR) spectrum obtained from a bacterial sample. Highlighted are the main characterisation regions, where A= fatty acids, B= amide region attributed to peptides and proteins, C= carboxylic group vibrations of polysaccharides, proteins and free amino acids, D= polysaccharides, and E= the fingerprint region.

1.2.2.2 . Raman spectroscopy

Raman spectroscopy is an important technique as it provides detailed information due to its sensitivity to chemical structure. Its use in biological studies has increased significantly due to the availability of modern lasers, allowing this spectroscopy technique to provide information on the structures of biological molecules including proteins, carbohydrates and lipids (Lu *et al.*, 2011; Huang *et al.*, 2010; Wen, 2007). The phenomenon of change in wavelength as a result of inelastic scattering of light by molecules was first observed by the Indian physicist and Noble Prize winner C.V. Raman in 1928 (Raman and Krishnan, 1928; Rajinder and Falk, 1998). Rayleigh scattering occurs when the energy remains constant on collision of a molecule or an atom with a photon. However, if nuclear motion is induced during the scattering process, a change in the wavelength of a scattered photon will occur resulting in inelastic scattering. In the process of inelastic scattering, the energy of the scattered photon is different from that of the incident photon, which results in the Raman effect (Raman, 1953; Tu and Chang, 2012). The term Raman scattering is used to refer to specific frequencies that are above or below the frequency that causes the incident beam to scatter. These features result in energy exchange due to photon-molecule collision, allowing molecules to either gain or lose the minimal amount of energy which is emitted from radiation at various frequencies. A lower scattering radiation frequency is linked to an increase in molecular energy, known as Stokes radiation. By contrast, anti-Stokes radiation results from the loss of molecular energy associated with the molecule being excited at a higher frequency. In general, Raman scattering has a weak effect such that typically only one in 10^6-10^8 photons scatter inelastically (Smith and Dent, 2013; Ellis *et al.*, 2013; Banwell, 1966).

Raman spectroscopy has been known and used in many different fields for the detection and characterisation of biological samples (Gaus *et al.*, 2006; Ashton *et al.*, 2011; López-Díez and Goodacre, 2003; Meisel *et al.*, 2014; Bocklitz *et al.*, 2009). In biological studies, the application of Raman spectroscopy has increased extensively from the late 1960s to the early 1970s for the determination of the structure of molecules. Raman spectroscopy has a significant advantage over FT-IR spectroscopy for studying biological analytes in aqueous solution since less interference occurs from water in this technique (Ellis *et al.*, 2007; Smith and Dent, 2013). Although Raman spectroscopy a great deal of useful information, however it has several limitations including its weak effect, highly focussed beams, which may damage sensitive samples, and interference with fluorescence in biological samples (Schie and Huser, 2013; Ashton and Goodacre, 2011).

1.2.3. Advantages and disadvantages of vibrational spectroscopy and mass spectrometry techniques

Vibrational spectroscopies, such as FT-IR and Raman spectroscopy, and mass spectrometry techniques, such as MALDI-TOF-MS and LC-MS, have a number of advantages and disadvantages in studying biological samples such as bacteria. These are summarised in Table 1.2 below.

Table1.2: Advantage and disadvantages of FT-IR spectroscopy, Raman spectroscopy, MALDI-TOF-MS and LC-MS techniques

Analytical technique	Advantages	Disadvantages
FT-IR spectroscopy	• Easy sample preparation • Simple to use • Sensitive technique • FT-IR spectra provide general information about bacteria • Inexpensive compared to several commonly used techniques • Rapid analysis	• May need expertise in chemometric analysis of data • Water band is very strong • Different conditions (e.g. growth time and culture medium) can cause variations in spectra
Raman spectroscopy	• Provides information on biological structures • Water band is negligible • Rapid • Reliable • Able to analysis small quantities of samples	• Raman effect is weak • Interference with fluorescence
MALDI-TOF-MS	• Rapid and specific detection of whole bacteria • Gentle ionisation technique • Ability to analyse high molecular weight compounds (e.g. protein) using a wide mass range • Sub-picomole sensitivity • Wide array of matrices	• MALDI matrix cluster ions obscure low *m/z* species (< 600) leading to matrix interference with small molecules • Homogeneity from spot to spot is variable • Adduct formation
LC-MS	• Separation and identification of any types of compounds present in bacteria • Gentle ionisation • Excellent detection • Reproducible • Quantitative	• Requirement of solvents for extraction • Time-consuming • Generation of complex data • Adduct formation • Ion suppression and competition for ionisation

Information coolected from (Sauer and Kliem, 2010; Siuzdak, 1996; Lartigue, 2013; Lay, 2001; Davis and Mauer, 2010; Naumann, 2001; Ferraro *et al.*, 2003; Huang *et al.*, 2010; Huang *et al.* 2004; van Baar, 2000).

1.3. Data analysis

Multivariate analysis (MVA) includes statistical methods which are used when visualising and analysing different variables in data obtained from more than one sample. In addition, relationships between various independent variables can be measured using these statistical techniques with initial assumption that all variables are equivalent in importance. Numerous MVA techniques have been used to analyse spectra generated from mass spectrometry (MS) and vibrational spectroscopy instruments. MVA is needed to simplify these data because MS and spectroscopy instruments generate complex data for analytes from various samples which are rich in information. For example, a spectrum that is generated from MALDI-TOF-MS consists of hundreds of *m/z* peaks which can be simplified using MVA (Manly, 1994; Goodacre, 2007). Therefore, MVA has been used for analysing data obtained from the analysis of bacterial samples using analytical techniques described in Section 1.2.

In general, MVA can be divided into two types of analysis: supervised and unsupervised analyses (Duda *et al*., 2012; Goodacre, 2007). The data generated from the analytical techniques used in the studies described in this thesis fall into a number of categories, an example of each category will be discussed in more detail below.

1.3.1 Unsupervised techniques

Unsupervised techniques are generally employed without the need for prior information regarding the classification and relationships between samples of interest. This technique describes the similarities and differences between the data generated using the analytical techniques. The data are subjected to a chemometric analysis algorithm used to analyse the data followed by visualisation of relationship patterns. Principal components analysis (PCA) is one of the most commonly used unsupervised techniques, which is discussed below (Goodacre *et al.*, 2007; Manly, 1994).

1.3.1.1 Principal Components Analysis (PCA)

In general, multivariate analysis starts with applying PCA to identify patterns in the data matrix obtained from MS and spectroscopy data (Goodacre, 2003; Boccard *et* *al.*, 2010). PCA can be used to simplify the collected spectral data (referred to as X data) by reducing the number of dimensions into smaller numbers known as principal components (PCs), (Jolliffe, 2014) while variability in the original data is represented by the PC indices. The first few PCs can often account for more than 90% of the total explained variance (TEV), where the first PC contains a large amount of data, which is subsequently reduced in successive PCs. The differences and similarities can be visualised and categorised rapidly by plotting the data in space either in 2D or 3D, which is defined by the PC scores (Manly, 1994; Goodacre, 200; Murtagh *et al.*, 2012). Figure 1.7 represents a schematic of a PCA plot.

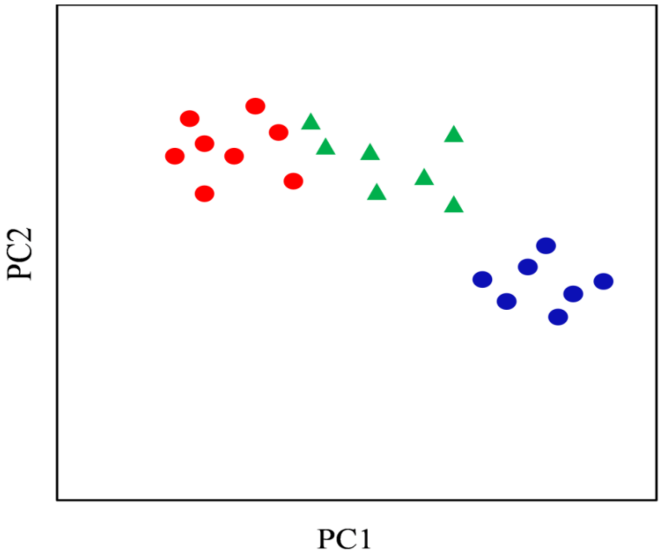


Figure 1.7: A schematic of a typical principal components analysis (PCA) plot. PCA reduces the number of variables by forming a new set of data called PCs. Three different groups are shown using different symbols for different classes.

1.3.2 Supervised techniques

Supervised techniques are also powerful tools for the analysis of complex data, but can only be used when the classes or values of responses are predictable (this information is also known as Y data), which are linked with sample input (also known as X data). The idea behind using supervised techniques is to construct a model that will show the association between the X and Y data. Discriminant function analysis (DFA) is one of many supervised methods that can be used to analyse data when prior biochemical information is known. Statistical algorithms obtained from supervised methods usually require *a priori* knowledge with regards to class structure. Hence, the statistical algorithms will be able to distinguish between samples based on these classes (Goodacre, 2007). However, validation of these supervised techniques is required in order to minimise any bias and over fitting of the data. The bootstrapping method is one of the tools that can be applied to validate the results that are generated from analytical methods that generate lots of information per sample. This method can be carried out by separating the samples into two different sets: a training set (used to generate the model) and a test set (used to validate the model). This process is repeated many times (hence the term bootstrap) so that the data are resample to allow statistical validation (Efron and Tibshirani, 1994; Mariey *et al.*, 2001).

1.3.2.1 Discriminant Function Analysis (DFA)

DFA is used to minimize within group variance from a group of samples from the same sample class and maximise between group variance of different classes. This chemometric technique is based on the true representation of the mean vectors of the PCs of the sample of interest, and it calculates the distance between the centres of each PC for grouping purposes (Johnson *et al.*, 2003; Kaderbhai *et al.*, 2003). Higher percentages of PCs that are correctly allocated to each group indicate better separation, which means that groups are different from each other (Manly, 1994). By contrast, closer distances between the centres of the vectors of each PC indicate more similarities identified between the groups. Figure 1.8 represents a schematic of a DFA plot.

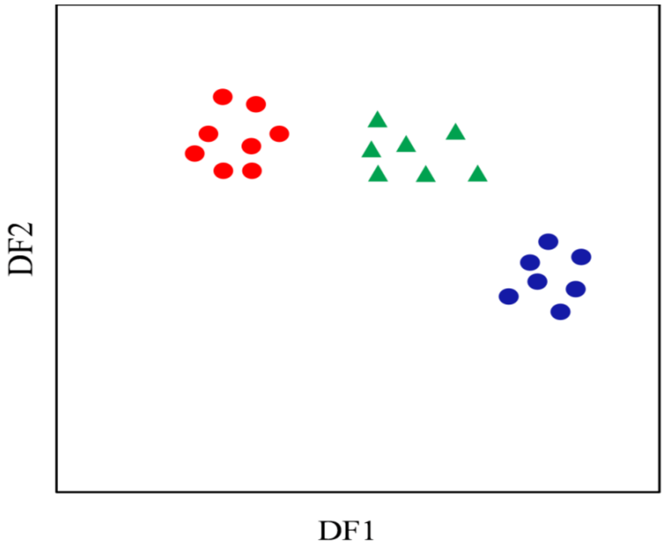


Figure 1.8: A schematic of a discriminant function analysis (DFA) plot; a supervised method that allows the reduction of intra- group variances and increases inter-group variances.

In addition, other chemometric analysis methods have been utilised in the present work including: partial least squares discriminant analysis (PLS-DA), hierarchical cluster analysis (HCA), Pareto optimality (PO), fraction factorial design (FFD) and Procrustes distance analysis as described in the relevant chapters to each of these methods.

1.4 Aims and Objectives

In response to the needs of clinicians and microbiologists, the main aim of this PhD research project was to develop research methods that can be used to classify and characterise a variety of bacteria including *Bacillus* spp. and *Enterococcus faecium*. To achieve this aim, a number of analytical techniques were used in combination with multivariate analysis (MVA). Specifically, the objectives of this research were:

(i) Assessment of various matrices and deposition methods to determine the optimal conditions for protein analysis using MALDI-TOF-MS. The optimised methods could then be applied to the analysis of proteins in intact bacteria from 34 *Bacillus* strains.

(ii) Assessment and exploratory analysis of experimental factors, including: choice of matrix, matrix additives, additive concentration, matrix preparation methods and matrix deposition methods, in order to find the optimal condition for the analysis of complex lipid mixture using MALDI-TOF-MS in combination with advanced chemometrics.

(iii) Analysis of lipids extracted from 33 strains of *Bacillus* bacteria to classify bacteria using MALDI-TO-MS and LC-MS in combination with advanced chemometrics.

(iv) Application of the optimised MALDI-TOF-MS protocols in (i) to analyse 35 isolates from *Enterococcus faecium*. FT-IR and Raman spectroscopies were also used as complementary analytical tools to confirm classification of enterococci.

1.5 References

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

Allwood, J. W. and Goodacre, R. 2010. An introduction to liquid chromatography mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis,* **21**, 33-47

Allwood W., J., Ellis, D. I., Heald, J. K., Goodacre, R. and Mur, L. A. J. 2006. Metabolomic approaches reveal that phosphatidic and phosphatidyl glycerol phospholipids are major discriminatory non-polar metabolites in responses by Brachypodium distachyon to challenge by Magnaporthe grisea. *The Plant Journal,* **46**, 351-368

Anhalt, J. P. and Fenselau, C. 1975. Identification of bacteria using mass spectrometry. *Analytical Chemistry,* **47**, 219-225

Ashcroft, A. E. 1997. *Ionisation methods in organic mass spectrometry,* Chapter 1- Cambridge, Royal Society of Chemistry, pp.15–16

Ashcroft, A. E. 2003. Protein and peptide identification: the role of mass spectrometry in proteomics. *Natural Product Reports,* **20**, 202-215

Ashton, L., Lau, K., Winder, C. L. and Goodacre, R. 2011. Raman spectroscopy: lighting up the future of microbial identification. *Future Microbiology,* **6**, 991-997

Ashton, L. and Goodacre, R., 2011. Application of deep UV resonanace Raman spectroscopy to bioprocessing, *Raman Spectroscopy-European Pharmaceutical Review*, **16**, 46-49

Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., Hughes, C., Lasch, P., Martin-Hirsch, P. L., Obinaju, B., Sockalingum, G. D., Sulé-Suso, J., Strong, R. J., Walsh, M. J., Wood, B. R., Gardner, P. and Martin, F. L. 2014. Using Fourier transform IR spectroscopy to analyse biological materials. *Nature Protocols,* **9**, 1771-1791

Banwell, C. N. 1966. *Fundamentals of molecular spectroscopy,* London, McGraw-Hill

Beveridge, T. J. 2001. Use of the Gram stain in microbiology. *Biotechnic and Histochemistry,* **76**, 111-118

Boccard, J., Veuthey, J.-L. and Rudaz, S. 2010. Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science,* **33**, 290-304

Bocklitz, T., Putsche, M., Stüber, C., Käs, J., Niendorf, A., Rösch, P. and Popp, J. 2009. A comprehensive study of classification methods for medical diagnosis. *Journal of Raman Spectroscopy,* **40**, 1759-1765

Brugger, B., 2014. Lipidomics: Analysis of the Lipid Composition of Cells and Subcellular Organelles by Electrospray Ionization Mass Spectrometry. *Annual Review of Biochemistry,* **83**, 79-98

Burlingame, A. L., Boyd, R. K. and Gaskell, S. J. 1996. Mass Spectrometry. *Analytical Chemistry,* **68**, 599-652

Buszewski, B. and Noga, S. 2012. Hydrophilic interaction liquid chromatography (HILIC) a powerful separation technique. *Analytical and Bioanalytical Chemistry,* **402**, 231-247

Cain, T. C. and Lubman, W. J., 1994, Differentiation of bacteria using protein profiles from matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, *Rapid Communication Mass Spectrometry,* **8**, 1026-1030

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. 1996. The rapid identification of intact microorganisms using mass spectrometry. *Nature Biotechnology,* **14**, 1584-1586

Cobo, F. 2013. Application of MALDI-TOF mass spectrometry in clinical virology: a review. *The Open Virology Journal,* **7**, 84

Coffey, A. G., Daly, C. and Fitzgerald, G. 1994. The impact of biotechnology on the dairy industry. *Biotechnology Advances,* **12,** 625-633

Cotter, R. 2013. High Energy Collisions on tandem time-of-flight mass spectrometers. *Journal of the American Society for Mass Spectrometry,* **24**, 657-674

Croxatto, A., Prod'hom, G. and Greub, G. 2012. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *Federation of European Microbiological Societies Microbiology Reviews,* **36**, 380-407

Cramer, R., Gobom, J. and Nordhoff, E. 2005. High-throughput proteomics using matrix-assisted laser desorption/ionization mass spectrometry. *Expert Review of Proteomics,* **2,** 407-420

Das, S. and Dash, H. R. 2014. *Microbial Biotechnology-A Laboratory Manual for Bacterial Systems*, India, Springer

de Hoffmann, E. and Stroobant, V. 2007. *Mass Spectrometry: Principles and Applications*, Chichester, John Wiley and Sons, pp.15-131

Dekker, J. P. and Branda, J. A. 2011. MALDI-TOF Mass Spectrometry in the Clinical Microbiology Laboratory. *Clinical Microbiology Newsletter,* **33**, 87-93

Dole, M., Mack, L., Hines, R., Mobley, R., Ferguson, L. and Alice, M. D. 1968. Molecular beams of macroions. *The Journal of Chemical Physics,* **49**, 2240-2249

Davis, R. and Mauer, L. 2010. Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic bacteria. In*:* Méndez-Vilas A. (Ed.), *Current research, technology and education topics in applied microbiology and microbial biotechnology, Volume II*. pp.1582-1594. Formatex Research Center: Badajoz, Spain.

Dunn, W. B. 2008. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology,* **5,** 011001

Dunn, W. B., Bailey, N. J. C. and Johnson, H. E. 2005. Measuring the metabolome: current analytical technologies. *Analyst,* **130**, 606-625

Duda, R. O., Hart, P. E. and Stork, D. G. 2012. Pattern classification, New York, John Wiley and Sons, pp.16-17

Dreisewerd, K. 2003. The desorption process in MALDI. *Chemical Reviews,* **103**, 395-426

Eberlin, L. S., Ferreira, C. R., Dill, A. L., Ifa, D. R. and Cooks, R. G. 2011. Desorption electrospray ionisation mass spectrometry for lipid characterisation and biological tissue imaging. *Biochimica et Biophysica Acta (BBA) Molecular and Cell Biology of Lipids,* **1811**, 946-960

Efron, B. and Tibshirani, R. J. 1994. *An introduction to the bootstrap*, Chapman and Hall/CRC press.

Ekman, R., Silberring, J., Westman-Brinkmalm, A. and Kraj, A. 2008. *Mass Spectrometry: Instrumentation, Interpretation, and Applications,* New Jersey, John Wiley and Sons, pp.26-69

El-Aneed, A., Cohen, A. and Banoub, J. 2009. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analysers. *Applied Spectroscopy Reviews,* **44**, 210-230

Ellis, D. I., Cowcher, D. P., Ashton, L., O'hagan, S. and Goodacre, R. 2013. Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool. *Analyst,* **138**, 3871-3884

Ellis, D. I., Dunn, W. B., Griffin, J. L., Allwood, J. W. and Goodacre, R. 2007. Metabolic fingerprinting as a diagnostic tool, *Pharmacogenomics,* **8**, 1243-66

Olsen, J. E. 2000. DNA-based methods for detection of food-borne bacterial pathogens. *Food Research International,* **33**, 257-266

Emerson, D., Agulto, L., Liu, H. and Liu, L. 2008. Identifying and characterising bacteria in an era of genomics and proteomics. *BioScience*, **58,** 925-936

Emon, J. M. V. and Lopez-Avila, V. 1992. Immunochemical methods for environmental analysis. *Analytical Chemistry*, **64**, 78A-88A

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

Ferraro, J. R., Nakamoto, K. and Brown, C. W. 2003. *Introductory raman spectroscopy, 2nd Ed.*, Academic Press: London, pp.1-27

Fenselau, C. and Demirev, P. A. 2001. Characterisation of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews,* **20**, 157-171

Freiwald, A. and Sauer, S. 2009. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols*, **4**, 732-742

Fuchs, B. and Schiller, J. 2009. Application of MALDI-TOF mass spectrometry in lipidomics. *European Journal of Lipid Science and Technology,* **111,** 83-98

Gaskell, S. J. 1997. Electrospray: principles and practice. *Journal of Mass Spectrometry,* **32**, 677-688

Gaus, K., Rösch, P., Petry, R., Peschke, K. D., Ronneberger, O., Burkhardt, H., Baumann, K. and Popp, J. 2006. Classification of lactic acid bacteria with UV-resonance Raman spectroscopy. *Biopolymers,* **82**, 286-290

Gidden, J., Denson, J., Liyanage, R., Ivey, D. M. and Lay Jr, J. O. 2009. Lipid compositions in *Escherichia coli* and *Bacillus subtilis* during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry. *International Journal of Mass Spectrometry,* **283**, 178-184

Giebel, R., Worden, C., Rust, S. M., Kleinheinz, G. T., Robbins, M. and Sandrin, T. R. 2010. Microbial fingerprinting using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS): applications and challenges. *Advances in Applied Microbiology,* **71**, 149-84

Goodacre, R. 2007. Metabolomics of a superorganism. *The Journal of Nutrition,* **137**, 259S-266S

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J. and Wulfert, F. 2007. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics,* **3**, 231-241

Goodacre, R., Heald, J. K. and Kell, D. B. 1999. Characterisation of intact microorganisms using electrospray ionisation mass spectrometry. *FEMS Microbiology Letters,* **176**, 17-24

Goodacre, R. 2003. Explanatory analysis of spectroscopic data using machine learning of simple interpretable rules. *Vibrational Spectroscopic*, **32**, 33-45

Gross, J. H. 2004. *Mass spectrometry: a textbook,* Germany, Springer, pp.2-5

Harrigan, G. G. and Goodacre, R. 2003. *Metabolic Profiling: its role in biomarker discovery and gene function analysis,* Norwell, Springer

Helm, D., Labischinski, H., Schallehn, G. and Naumann, D. 1991. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Journal of General Microbiology,* 1**37**, 69-79

Hillenkamp, F. and Peter-Katalinic, J. 2013. *MALDI MS: a practical guide to instrumentation, methods and applications*, Weinheim, John Wiley and Sons

Holland, R. D., Rafii, F., Heinze, T. M., Sutherland, J. B., Voorhees, K. J. And Lay, J. O. 2000. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometric detection of bacterial biomarker proteins isolated from contaminated water, lettuce and cotton cloth. *Rapid Communications in Mass Spectrometry,* **14**, 911-917

Holland, R. D., Wilkes, J. G., Rafii, F., Sutherland, J. B., Persons, C. C., Voorhees, K. J. and Lay, J. O. 1996. Rapid Identification of Intact Whole Bacteria Based on Spectral Patterns using Matrix-assisted Laser Desorption/Ionisation with Time-of-flight Mass Spectrometry. *Rapid Communications in Mass Spectrometry,* **10**, 1227-1232

Huang, W. E., Griffiths, R. I., Thompson, I. P., Bailey, M. J. and Whiteley, A. S. 2004. Raman Microscopic Analysis of Single Microbial Cells. *Analytical Chemistry,* **76**, 4452-4458

Huang, W. E., Li, M., Jarvis, R. M., Goodacre, R. and Banwart, S. A. 2010. Shining Light on the Microbial World: The Application of Raman Microspectroscopy, *Advances in Applied Microbiology.* **70**,153-186

Iseman, M. D. 1993. Treatment of Multidrug-Resistant Tuberculosis. *New England Journal of Medicine,* **329**, 784-791

Janda, J. M. and Abbott, S. L. 2002. Bacterial identification for publication: when is enough enough? *Journal of Clinical Microbiology,* **40**, 1887-1891

Johnson, H. E., Broadhurst, D., Goodacre, R. and Smith, A. R. 2003. Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry,* **62**, 919-928

Jolliffe, I. 2014. Principal Component Analysis. *Wiley StatsRef: Statistics Reference Online.* John Wiley and Sons, pp.1-2

Kaderbhai, N. N., Broadhurst, D. I., Ellis, D. I., Goodacre, R. and Kell, D. B. 2003. Functional genomics via metabolic footprinting: monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT–IR and direct injection electrospray mass spectrometry. *Comparative and Functional Genomics,* **4**, 376-391

Kafka, A. P., Kleffmann, T., Rades, T. and Mcdowell, A. 2011. The application of MALDI TOF MS in biopharmaceutical research. *International Journal of Pharmaceutics,* **417**, 70-82

Karas, M., Bachmann, D. and Hillenkamp, F. 1985. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry,* **57**, 2935-2939

Karas, M. and Hillenkamp, F. 1988. Laser desorption ionisation of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry,* **60**, 2299-2301

Kim, S., Reuhs, B. L. and Mauer, L. J. 2005. Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. *Journal of Applied Microbiology,* **99**, 411-417

Kirschner, C., Maquelin, K., Pina, P., Thi, N. N., Choo-Smith, L.-P., Sockalingum, G., Sandt, C., Ami, D., Orsini, F. and Doglia, S. 2001. Classification and identification of enterococci: a comparative phenotypic, genotypic, and vibrational spectroscopic study. *Journal of Clinical Microbiology,* **39**, 1763-1770

Krishnamurthy, T. and Ross, P. L. 1996. Rapid Identification of Bacteria by Direct Matrix-assisted Laser Desorption/Ionisation Mass Spectrometric Analysis of Whole Cells. *Rapid Communications in Mass Spectrometry,* **10**, 1992-1996

Kussmann, M., Nordhoff, E., Rahbek-Nielsen, H., Haebel, S., Rossel-Larsen, M., Jakobsen, L., Gobom, J., Mirgorodskaya, E., Kroll-Kristensen, A., Palm‖, L. and Roepstorff, P. 1997. Matrix-assisted laser desorption/ionisation mass spectrometry sample preparation techniques designed for various peptide and protein analytes. *Journal of Mass Spectrometry,* **32**, 593-601

Lasch, P., Beyer, W., Nattermann, H., Staemmler, M., Siegbrecht, E., Grunow, R. and Naumann, D. 2009. Identification of *Bacillus* anthracis by using matrix-assisted laser desorption ionisation-time of flight mass spectrometry and artificial neural networks. *Applied and Environmental Microbiology,* **75**, 7229-7242

Lasch, P., Haensch, W., Lewis, E. N., Kidder, L. H. and Naumann, D. 2002. Characterisation of colourectal adenocarcinoma sections by spatially resolved FT-IR microspectroscopy. *Applied Spectroscopy,* **56**, 1-9

Lasch, P., Fleige, C., Stammler, M., Layer, F., Nubel, U., Witte, W. and Werner, G. 2014. Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus francium* and *Staphylococcus aureus* strains. *Journal of Microbiology Methods*, **100**, 58-69

Lay Jr, J. O. 2001. MALDI-TOF mass spectrometry of bacteria. *Mass Spectrometry Reviews*, **20**, 172-194

Lay Jr, J. O. 2000. MALDI-TOF mass spectrometry and bacterial taxonomy. *Trac Trend Analytical Chemistry*, **19**, 507-516

Lartigue, M.-F. 2013. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterisation. *Infection, Genetics and Evolution***,** **13**, 230-235

Liu, H., Du, Z., Wang, J. and Yang, R. 2007. Universal sample preparation method for characterisation of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and environmental microbiology,* **73**, 1899-1907

López-Díez, E. C. and Goodacre, R. 2003. Characterisation of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry,* **76**, 585-591

Luzzatto-Knaan, T., Melnik, A. V. and Dorrestein, P. C. 2015. Mass spectrometry tools and workflows for revealing microbial chemistry. *Analyst*, **140**, 4949-4966

Makarov, A. 2000. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry,* **72**, 1156-1162

Manly, B. F. J. 1994. *Multivariate Statistical Methods: A Primer,* London, Chapman and Hall, pp. 12-17

Mariey, L., Signolle, J. P., Amiel, C. and Travert, J. 2001. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vibrational Spectroscopy,* **26,** 151-159

Marvin, L. F., Roberts, M. A. and Fay, L. B. 2003. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in clinical chemistry. *Clinica Chimica Acta,* **337**, 11-21

Meisel, S., Stöckel, S., Rösch, P. and Popp, J. 2014. Identification of meat-associated pathogens via Raman microspectroscopy. *Food Microbiology,* **38**, 36-43

Murtagh, F. and Heck, A. 2012. *Multivariate data analysis,* Springier and Business Media, PP.17-19

Naumann, D., Helm, D. and Labischinski, H. 1991. Microbiological characterisations by FT-IR Spectroscopy. *Nature,* **35**1, 81-82

Naumann, D. 2001. FT-infrared and FT-Raman spectroscopy in biomedical research. *Applied Spectroscopy Reviews*, **36**, 239-298

Nester, E. W. 2001. *Microbiology: a Human Perspective*, 3rd Edition, the University of Michigan, McGraw-Hill

Nicolaou, N., Xu, Y. and Goodacre, R. 2011. MALDI-MS and multivariate analysis for the detection and quantification of different milk species. *Analytical and Bioanalytical Chemistry,* **399**, 3491-3502

Nielen, M. W., 1999. MALDI time-of-flight mass spectrometry of synthetic polymer. *Mass Spectrometry Reviews*, **19**, 309-344

Nomura, F. 2015. Proteome-based bacterial identification using matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS): A revolutionary shift in clinical diagnostic microbiology. *Biochimica et Biophysica Acta (BBA),* **1854,** 528-537

Olive, D. M. and Bean, P. 1999. Principles and applications of methods for DNA-based typing of microbial organisms. *Journal of Clinical Microbiology,* **37**, 1661-1669

Parisi, D., Magliulo, M., Nanni, P., Casale, M., Forina, M. and Roda, A. 2008. Analysis and classification of bacteria by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry and a chemometric approach. *Analytical and Bioanalytical Chemistry*, **391**, 2127-2134

Peeling, R. W., Smith, P. G. and Bossuyt, P. M. 2008. A guide for diagnostic evaluations. *Nature Reviews Microbiology,* **8**, S2-S6

Rajinder, S. and Falk, R. 1998. Sir CV Raman and the Story of the Nobel Prize. *Current Science Bangalore,* **75**, 965-971

Raman, C. V. 1953. A new radiation. *Proceedings of the Indian Academy of Sciences Section A,* **37**, 333-341

Raman, C. V. and Krishnan, K. S. 1928. A new type of secondary radiation. *Nature,* **121**, 501-502

Lu, X., Al-Qadiri, H. M., Lin, M. and Rasco B. A., 2011. Application of mid-inferred and Raaman spectroscopy to the study of bacteria. *Food Bioprocess Technology*, **4**, 919-935

Saenz, A. J., Petersen, C. E., Valentine, N. B., Gantt, S. L., Jarman, K. H., Kingsley, M. T. and Wahl, K. L. 1999. Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis. *Rapid Communications in Mass Spectrometry,* **13**, 1580-1585

Sauer, S. and Kliem, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* **8**, 74-82

Sauer, S., Lange, B. M. H., Gobom, J., Nyarsik, L., Seitz, H. and Lehrach, H. 2005. Miniaturization in functional genomics and proteomics. *Nature Reviews Genetics,* **6**, 465-476

Santos, T., Capelo, J., Santos, H. M., Olverira, I., Mainho, C., Goncalves, A., Araujo, J. E. Poeta, P. and Igerjas, G. 2015. Use of MALDI-TOF mass spectrometry fingerprinting to characterise *Enterococcus* spp. and *Escherichia coli* strains. *Journal of Proteomics,* Epub ahead of print, **127**, 321-331

Schiller, J., Süß, R., Arnhold, J., Fuchs, B., Leßig, J., Müller, M., Petković, M., Spalteholz, H., Zschörnig, O. and Arnold, K. 2004. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research. *Progress in Lipid Research,* **43**, 449-488

Schumann, P. and Maier, T. 2014. Chapter 13 - MALDI-TOF mass spectrometry applied to classification and identification of bacteria. *Methods in Microbiology*. **41**, 275-306

Scigelova, M. and Makarov, A. 2006. Orbitrap mass analyser – overview and applications in proteomics. *Proteomics,* **6**, 16-21

Schie, I. W. and Huser, T., 2013, Methods and applications of Raman Microspectroscopy to single-cell analysis, *Applied Spectroscopy,* **67**, 813-829

Šedo, O., Sedláček, I. and Zdráhal, Z. 2011. Sample preparation methods for MALDI-MS profiling of bacteria. *Mass Spectrometry Reviews,* **30**, 417-434

Sintchenko, V., Iredell, J. R. and Gilbert, G. L. 2007. Pathogen profiling for disease management and surveillance. *Nature Reviews Microbiology,* **5**, 464-470

Siuzdak, G. 1996. *Mass spectrometry for biotechnology,* Elsevier Science, San Diego, CA, Academic Press, pp.4-54

Smith, E. and Dent, G. 2013. *Modern Raman spectroscopy: a practical approach,* John Wiley and Son, pp.1-80

Spengler, B. and Kaufmann, R. 1992. Gentle probe for tough molecules: matrix-assisted laser desorption mass spectrometry. *Analysis,* **20**, 91-101

Stübiger, G. and Belgacem, O. 2007. Analysis of lipids using 2, 4, 6-trihydroxyacetophenone as a matrix for MALDI mass spectrometry. *Analytical Chemistry,* **79**, 3206-3213

Shanta, S. R., Kim, T. Y., Hong, J. H., Lee, J. H., Shin, C. Y., Kim, K.-H., Kim, Y. H., Kim, S. K. and Kim, K. P. 2012. A new combination MALDI matrix for small molecule analysis: application to imaging mass spectrometry for drugs and metabolites. *Analyst,* **137**, 5757-5762

Stuart, B. 1996. *Modern infrared spectroscopy,* Chichester, John Wiley and Sons Ltd, pp.1-24

Stults, J. T. 1995. Matrix-assisted laser desorption/ionisation mass spectrometry (MALDI-MS). *Current Opinion in Structural Biology,* **5**, 691-698

Wolff, M. M. and Stephens, W. 1953. A pulsed mass spectrometer with time dispersion. *The Review of Scientific Instruments,* **24**, 616

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

Tolstikov, V. V. and Fiehn, O. 2002. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry,* **301**, 298-307

Tu, Q. and Chang, C. 2012. Diagnostic applications of Raman spectroscopy. *Nanomedicine: Nanotechnology, Biology and Medicine,* **8**, 545-558

Vaidyanathan, S., Kell, D. B. and Goodacre, R. 2004. Selective detection of proteins in mixtures using electrospray ionisation mass spectrometry: influence of instrumental settings and implications for proteomics. *Analytical Chemistry,* **76,** 5024-5032

Vaidyanathan, S., Rowland, J. J., Kell, D. B. and Goodacre, R. 2001. Discrimination of Aerobic Endospore-forming Bacteria via Electrospray-Ionisation Mass Spectrometry of Whole Cell Suspensions. *Analytical Chemistry,* **73**, 4134-4144

Van Baar, B. L. M. 2000. Characterisation of bacteria by matrix-assisted laser desorption/ionisation and electrospray mass spectrometry. *FEMS Microbiology Reviews,* **24**, 193-219

Watson, J. T. and Sparkman, O. D. 2008. Mass spectrometry/mass spectrometry. *Introduction to Mass Spectrometry.* John Wiley and Sons, pp.53-69

Wen, Z. Q. 2007. Raman spectroscopy of protein pharmaceuticals. *Journal of pharmaceutical sciences,* **96**, 2861-2878

Wilkins, C. L. and Lay, J. O. 2005. *Identification of microorganisms by mass spectrometry,* John Wiley and Sons, pp.303

Williams, D. H. and Fleming, I. 1995. *Spectroscopic methods in organic chemistry,* London, McGraw-Hill

Williams, T. L., Andrzejewski, D., Lay Jr, J. O. and Musser, S. M. 2003. Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. *Journal of the American Society for Mass Spectrometry,* **14**, 342-351

Yamashita, M. and Fenn, J. B. 1984. Electrospray ion-source - another variation on the free-jet theme. *Journal of Physical Chemistry,* **88**, 4451-4459

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzeby, J., Amann, R. and Rossello-Mora, R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology,* **12**, 635-645

Yates, J. R., Ruse, C. I. and Nakorchevsky, A. 2009. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual Review of Biomedical Engineering,* **11**, 49-79

Zenobi, R. and Knochenmuss, R. 1998. Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews,* **17**, 337-366

Chapter Two

Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the characterisation of *Bacillus* and *Brevibacillus* species

Najla AlMasoud,[a] Yun Xu,[a] Nicoletta Nicolaou[a] and Royston Goodacre [a]

[a]School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

Correspondence to Roy Goodacre: roy.goodacre@manchester.ac.uk

This work has been published as:

AlMasoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. 2014. Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the characterisation of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta*, **840**, 49-57

Yun Xu contributed to this work by creating the alingment and carrying out the associated data analysis. Nicoletta Nicolaou participated in optimising the operation of the MALDI-TOF-MS. Roy Goodacre contributed to this study with his supervision and guidance.

Abstract

Over the past few decades there has been an increased interest in using various analytical techniques for detecting and identifying microorganisms. More recently there has been an explosion in the application of matrix assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) for bacterial characterisation, and here we optimise this approach in order to generate reproducible MS data from bacteria belonging to the genera *Bacillus* and *Brevibacillus*. Unfortunately MALDI-TOF-MS generates large amounts of data and is prone to instrumental drift. To overcome these challenges we have developed a pre-processing pipeline that includes baseline correction, peak alignment followed by peak picking that in combination significantly reduces the dimensionality of the MS spectra and corrects for instrument drift. Following this two different prediction models were used which are based on support vector machines and these generated satisfactory prediction accuracies of approximately 90%.

2.1 Introduction

Bacillus are rod-shaped aerobic Gram-positive bacteria that are able to sporulate. These bacteria are normally found in the soil, plants, and can be transferred to meat and dairy products where they can spoil food making them unfit for human consumption (Granum, 1997). Even though most of these bacteria are harmless saphrophytes there still remains a few toxic members of this genus, such as *B. subtilis* and *B. cereus*, which are often associated with food-borne infections, (Drobniewski, 1993) along with the more notorious *B. anthracis* the casual agent of anthrax. Whilst *B. sphaericus* is toxic to insects and is used for biocontrol of mosquitoes (Singer, 1991). *B. subtilis* is the most scientifically defined member of the *Bacillus* genus and has thus been used as a model organism in many genetic research studies. Other members of this *B. subtilis* group are less defined and are harder to identify such as *B. licheniformis* and *B. amyloliquefaciens*, because they are very similar microorganisms (Fritze, 2004; Granum, 1997). The *B. cereus* group contains a number of different bacteria, with some leading to negative health implications in humans, and as discussed above have sometimes been linked to food poisoning (Granum and Lund, 1997; Priest *et al*., 2004; Ghelardi *et al*., 2002).

The unequivocal identification is vital steps in medical therapy and the food industry and this is usually performed at the genotypic or phenotypic level. A number of traditional methods have so far been used to identify microorganisms, such as cell culturing with differential staining (Wilkins and Lay, 2005), polymerase chain reaction (PCR) (Hill and Wachsmuth, 1996; Gulledge *et al*., 2010; Zara *et al*., 2006, Vidal-Quist *et al*., 2009) and enzyme linked immunosorbent assays (ELISA) (Engvall, 1977). Whilst these approaches formed the foundations of knowledge and understanding in microorganism research, these methods are very time consuming, costly and labour intensive, hence more rapid detection methods are continually needed (Sauer and Kliem, 2010). In addition to rapid testing, methods that provide molecular-specific information are also preferred as these may allow one to relate any markers to specific microbiological function.

Modern methods for the identification of microorganisms have recently focussed on mass spectrometry as these are rapid and provide molecular information on the bacteria under investigation. Whilst pyrolysis mass spectrometry was used for bacterial analysis in the past (Goodacre and Kell, 1996), current methods are based on electrospray-ionisation (ESI-MS) (Goodacre *et al.*, 1999; Vaidyanathan *et al.*, 2001) and the more popular method of matrix-assisted laser desorption ionisation (MALDI-MS) (Sauer and Kliem, 2010; Lay Jr, 2000; Claydon *et al.*, 1996; Krishnamurthy *et al.*, 2000). MALDI-TOF-MS is easy to use, provides rapid results, and has been used for identification and taxonomy of microorganisms (Welham *et al*., 1998; Fenselau and Demirev, 2001; Lay Jr, 2000). The maturity of this analytical technique has benefitted its application to a wide range of areas such as proteomics (Demirev *et al*., 1999; Ryzhov and Fenselau; 2001), intact-cell mass spectrometry (ICMS) (Holland *et al*., 1996; Claydon *et al*., 1996; Krishnamurthy and Ross, 1996; Lasch *et al*., 2008) and in the area of lipidomics (Schiller *et al*., 2004; Gidden *et al*., 2009; Fuchs and Schiller, 2009).

MALDI-TOF-MS on bacteria (and indeed other complex samples) results in a multivariate spectral pattern, which usually provides information on the protein content of the bacterium under analysis. This protein profile or barcode can be matched against MALDI-TOF-MS profiles/barcodes that have been previously collected under identical conditions and stored within (usually) organism specific databases (Lasch *et al*., 2009; Freiwald and Sauer, 2009; Demirev *et al*., 1999; Fenselau and Demirev, 2001). This matching may involve the generation of dendrograms from hierarchical cluster analyses (HCA) (Lasch *et al*., 2009; Vargha *et al*., 2006) or ordination plots from principal component analysis (PCA) (Toh-Boyo *et al*., 2012; Goodacre, 2003) or discriminant analysis (DA) (Nicolaou *et al*., 2012; Lopez-Diez and Goodacre, 2004).

The aim of this study was to generate a reproducible MALDI-TOF-MS protocol for measuring the protein spectra from bacteria. In order to establish this we used a set of 34 well-characterised bacteria belonging to the genus *Bacillus*. In a series of experiments we optimised the matrix and the sample preparation method used using first a mixture of pure proteins followed by the analysis of a subset of these bacilli, before the optimised method was used on the full set of 34 bacteria.

2.2 Materials and Methods

2.2.1 Compounds

Trifluoroacetic acid (TFA), acetonitrile (ACN), sinapinic acid (SA), caffeic acid (CA), 2,5 dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHCA), ferulic acid (FA), 2,4,6-trihydroxyacetophenone monohydrate (THAP), 2-4 hydroxphenylaze benzoic acid (HABA), 2,6 dihydroxyacatophenone (DHAP), 9-aminoacridine (9-AA) and dithranol (INN) from Sigma Aldrich (Dorset, UK) were used.

14 g of nutrient agar (Fisher Scientific Ltd., Loughborough, UK) was dissolved and mixed thoroughly in a bottle containing 500 mL of water. This bottle was then autoclaved at 121^{o}C for 15 min and subsequently used for the bacterial cultures.

2.2.2 Standard protein samples for MALDI-TOF-MS

Five different proteins were mixed together at the same concentration (20 µM) to find the optimum matrix and deposition method for pure protein analysis. These proteins (molecular weight provided in parentheses) included: insulin (5,735)**,** cytochrome *c* (12,362), apomyoglobin (16,952), aldolase (39,212) and Albumin (66,437) were acquired from Sigma Aldrich.

2.2.3 Bacterial culturing

General information of the 34 strains of *Bacillus* is provided in Table 2.1 and these belonged to two genera (*Bacillus* and *Brevibacillus*) and seven different species. The cells were cultured on nutrient agar and were incubated at 37^{o}C for 24 h. Bacterial strains were cultivated aerobically three times under these conditions to makes sure that the cultures were axenic, and to maintain a stable phenotype. After this was established single bacterial colonies were then cultured on nutrient agar and also incubated at 37^{o}C for 24 h. Five biological replicates were prepared for each strains. After growth the biomass of each sample was carefully collected using two full sterilised plastic loops (equivalent to about 20 µL). This biomass was then centrifuged for 3 min at 13000 × *g*. The pellets containing the bacteria were then washed twice with 1 mL of sterile distilled water to remove residual culture media, centrifuged again to remove the supernatant, and the pellet was then stored at -80^{o}C until further analysis.

Table 2.1: The 34 *Bacillus* species and strains used in this work

Sample no.	Species	Strain no.	Key colour used in figures
1	*B. sphaericus*	7134^{T}	Yellow
2		B0408*	
3		B0219	
4		B0769	
5		B1147	
6	*Br. laterosporus*	B0043	Blue
7		B0262	
8	*B. subtilis*	B0014^{T*}	Black
9		B0044	
10		B0098	
11		B0099	
12		B0410	
13		B0501	
14		B1382	
15	*B. cereus*	B0002^{T*}	Green
16		B0550	
17		B0702	
18		B0712	
19		B0851	
20	*B. amyloliquefaciens*	B0177^{T}	Red
21		B0168*	
22		B0175	
23		B0251	
24		B0620	
25	*B. megaterium*	B0010^{T*}	Pink
26		B0056	
27		B0057	
28		B0076	
29		B0621	
30	*B. licheniformis*	B0252^{T*}	Cyan
31		B0242	
32		B0755	
33		B1081	
34		B1379	

^{T} indicates the type strain; *indicates strains used for preliminary optimisation experiments.

2.2.4 Optimisation of MALDI-TOF-MS

Optimisation of sample preparation was carried out in order to identify the most appropriate matrix preparation and deposition method for the analysis of bacteria. Initial experiments optimised the matrix and deposition method on mixtures of pure proteins (Supplementary Information Table S2.1 illustrates the four different sample preparation methods for MALDI-TOF-MS). Briefly, 10 different matrices were used to find the most compatible matrix for MALDI-TOF-MS analysis and these included DHB, CHCA, SA, FA, THAP, CA, HABA, DHAP, 9-AA and INN. At the same time four different depositions methods (mix, overlay, underlay and sandwich) were investigated for protein sample preparation. The optimised conditions involved using SA as the matrix and the mix method for sample deposition and this was subsequently used for bacterial analysis. We note of course that the five proteins chosen are a substitute for bacterial analysis and we did not assume that the best protein preparation method would be the optimal method for bacteria so we tested the top three matrices and preparation methods on a small subset of bacteria (the strains used for preliminary optimisation experiments were marked with * in Table 2.1); SA with the mix method was indeed the best method (data not shown for this optimisation).

2.2.5 Bacterial sample preparation

Preliminary experiments also suggested that it was important to optimise the appropriate amount of biomass for MALDI-TOF-MS. The defrosted pellet from above was diluted at various levels in water containing 0.1% TFA (250, 500, 1000, 1500, 4000 µL; data not shown except for 1000 µL water containing 0.1% TFA). The optimum pellet dilution was established at 1000 µL and this was subsequently used.

For MALDI-TOF-MS analysis of the bacteria 10 mg SA was dissolved in 500 µL of ACN and 500 µL of water containing 2 % TFA. 10 µL from the above bacterial sample and 10 µL of matrix were mixed together (Table S2.1) and vortexed for 10 s before. 2 µL from the resultant mixture was spotted on a MALDI-TOF-MS stainless steel target plate. This was allowed to dry at room temperature (*ca*. 22 °C) for 1 h.

2.2.6 MALDI-TOF-MS

Samples were analysed in batches using an AXIMA-Confidence (Shimadzu Biotech, Manchester, U.K) mass spectrometer. This MALDI-TOF-MS device contained a nitrogen pulsed UV laser with a wavelength of 337 nm as described previously (Nicolaou *et al.*, 2011). The power of the laser at the laser head used was set to 140 mV. Each profile contained 20 shots, and 100 profiles were collected using a circular raster pattern. The MS was operated in positive ion source and linear TOF was used over the range from 1000-80,000 *m*/*z*. The collection time for each sample was ~2 min and each biological sample was analysed four times (technical replicates). A single biological replicate for each of the 34 bacteria was analysed each day, and the analysis time took 5 days of machine time during a 2 week period. The result of this analysis generated 680 MALDI-TOF-MS spectra: 34 bacteria × 5 biological replicates × 4 technical replicates. The MALDI device was calibrated using the protein mixture mentioned above.

2.2.7 Data analysis

2.2.7.1 Pre-processing

MATLAB 2010a (The Math Works, Natick, MA, USA) was used for pre-processing and data analysis. Baseline corrections were first performed on the spectra by using asymmetric least squares (AsLS) (Eilers, 2004). In addition, the interpolation and alignment of MALDI-TOF-MS spectra in the *m*/*z* axis were required in order to integrate all the spectra in a unified coordinate system and also reduce the amount of ambiguities of assigning peaks from different samples collected over the 2 week period (see below). This was achieved by firstly interpolating all the spectra into a common *m*/*z* domain which is from 1,000 to 13,000 *m*/*z* with an interval of 0.1078 *m*/*z* and then an algorithm named interval correlation optimised shifting (icoshift) (Tomasi *et al*., 2011) was used to correct *m/z* drifting across different samples. Peak picking was then performed on the aligned spectra to detect mass peaks in each spectrum and this was performed using intensity weighted variance (IWV) algorithm as described by Jarman (Jarman *et al*., 2003). The detected peaks of all the samples were then aligned together with a drift tolerance threshold of ±1 *m*/*z*. After this peak picking and alignment process, a total number of 243 unique mass peaks were detected and resulted in a peak table matrix of dimensions 680 × 243 which was used for further data analysis. The peak intensities were firstly $\log_{10}$-scaled and then normalised so that the sum of squares of each row (i.e. a sample) equals 1.

2.2.7.2 Multivariate analysis

Two different types of analysis were performed on the data: one was a semi-quantitative analysis and the other a qualitative analysis.

The semi-quantitative analysis was performed on the $\log_{10}$-scaled and normalised peak intensity table matrix. Principal component analysis (PCA) was performed first to reveal the “natural” pattern of the data and then support vector machines (SVM), with a linear kernel, was used for supervised classification. The SVM models were validated by using a bootstrap replacement procedure coupled with cross-validation for the model parameter selection (see below). In this process the data were first split into a training set and a test set via a bootstrapping resampling based on the biological replicates; i.e., all the samples from the same biological replicates were considered as one during the resampling. Considering the random nature of this bootstrapping process, the number of samples selected in the training and test sets varied between the different 1000 iterations, on average 63.3% of the samples would be in the training set and 36.7% in the test. Next a *k*-fold cross-validation was performed on the training set where *k* is the number of unique biological replicates in the training set, the error penalty parameter *C* within the SVM varied from 1 to 10^6 and the one which yielded the lowest cross-validation error was chosen to build the SVM model. The model was then applied to the test set generated via the bootstrapping selection in order to calculate the predictive accuracy of the test set. This bootstrap procedure was repeated 1,000 times and the collected predictive accuracies for the *test set only* were then averaged. This can be considered as an unbiased estimation of the generalisation performance of the SVM model. Two types of classification were carried out: one was to classify the samples on species level (7 classes); and the other was to classify the samples on strain level (34 classes). Both types of classification followed the same validation procedure as described above.

The qualitative analysis on the data focused on the presence/absence of certain feature (i.e. mass peaks) while ignoring the intensities of the peaks. The peak table matrix was converted into a binary format: if a peak had been detected in one particular sample the corresponding element in the matrix was set to 1 and 0 if otherwise; the threshold for presence/absence was set to be 3× standard deviation of baseline signals. Principal coordinate analysis (PCoA) was used as a counterpart of PCA in the qualitative analysis and the Jaccard distance was used to measure the dissimilarity between the samples. A distance matrix ***D*** was calculated which contains the Jaccard distance between every pair of samples. PCoA was then applied to ***D*** to obtain a scores matrix and this scores matrix can be interpreted in the same way as the scores matrix obtained from PCA. For supervised classification, a naïve Bayesian classifier and SVM with a Jaccard kernel (Nemmour and Chibani, 2008) were applied to the data. Both classifiers were validated using exactly the same bootstrapping procedure as described above and the classifications were again performed on both species and strain level.

2.3 Results and discussion

2.3.1 MALDI-TOF-MS optimisation

Initially a mixture contain five different proteins was used to obtain the optimum conditions for protein analysis using MALDI-TOF-MS. At this stage 10 matrices were used to determine the most suitable matrix and four sample preparations procedure when performed. Good protein detection was seen for SA, CA and FA, whilst others such as DHAP and 9-AA were not suitable matrices for protein analysis. Results obtained from this study showed that SA was the most suitable matrix for protein analysis (Tables S2.2-S2.5). This finding was supported by other workers analysis (Beavis *et al*., 1989; Giebel *et al*., 2010; Gantt *et al*., 1999; Toh-Boyo *et al*., 2012; Pineda *et al*., 2003; Smole *et al*., 2002), and this may be due its classification as a hot matrix, (Zenobi and Knochenmuss, 1998). In addition, as discussed by Vaidyanathan (Vaidyanathan *et al*., 2002), the reason behind SA's compatibility coud be its high level of homogeneity and crystallisation with the solvent when SA is mixed with bacteria.

During the matrix optimisation the most appropriate sample deposition method for protein analysis was also assessed. Four methods were used (see Table S2.1 for details) and it was found that the 'mix method' where sample and matrix are pre-mixed prior to spotting on the MALDI target plate was best. This deposition method was very reproducible and caused improved desorption and ionisation in comparison with other deposition methods. Tables S2.2-S2.5 (see SI) summarises the data obtained from analysing the 5-way protein mixture using the 10 different matrices and the 4 different deposition methods.

After this the top 3 matrices (SA, CA and FA) were assessed on a subset of 6 bacteria comprising the type strain from each species. SA with the mix method was also the best method in terms of the number of protein peaks routinely detected in replicate analyses and in terms of the reproducibility of signal (as judged by PCA; data not shown). Thus SA with the mix method was used for all bacterial analyses.

2.3.2 Bacillus MALDI-TOF-MS spectra

Typical MALDI-TOF-MS spectra of *B. cereus* B0712 obtained SA with the mix method for both the raw MS data and after baseline correction and alignment are shown in Figure 2.1. It is clear from the raw data from this bacterium (and indeed all the bacteria analysed; data not shown) that significant baseline artefacts are observed which were unavoidable. Spectra were therefore pre-processed using the following routine: (i) baseline correction was performed using AsLS on the raw MS profiles; (ii) this was followed by spectral alignment using icoshift; (iii) finally, following this step these spectra were scaled so that the sum of square of each spectrum equals to 1. Typical normalised and scaled spectra of all 7 type species from these bacilli are shown in Figure 2.2A-G.

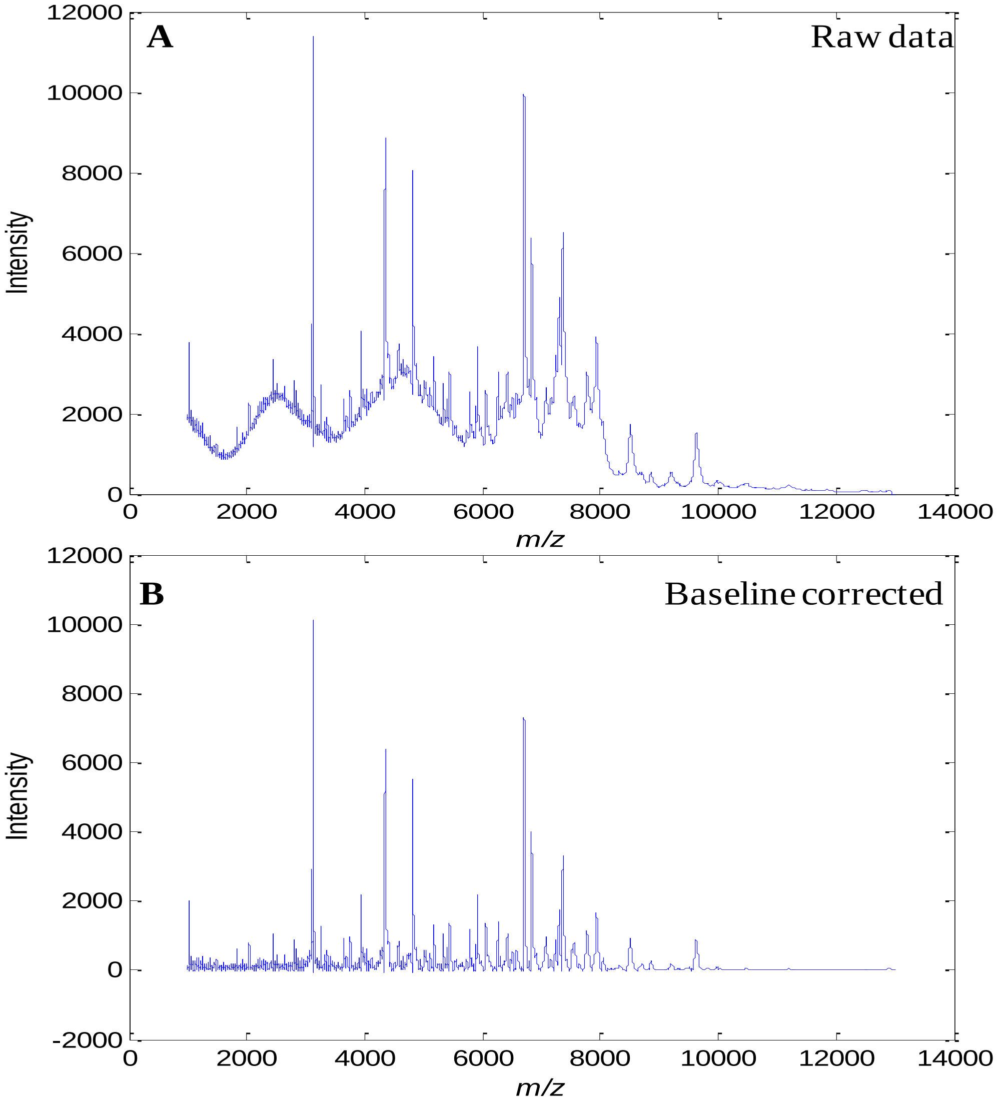


Figure 2.1: Differences between MALDI mass spectra obtained from the analysis of *B. cereus* B0712 (A) before and (B) after baseline correction.

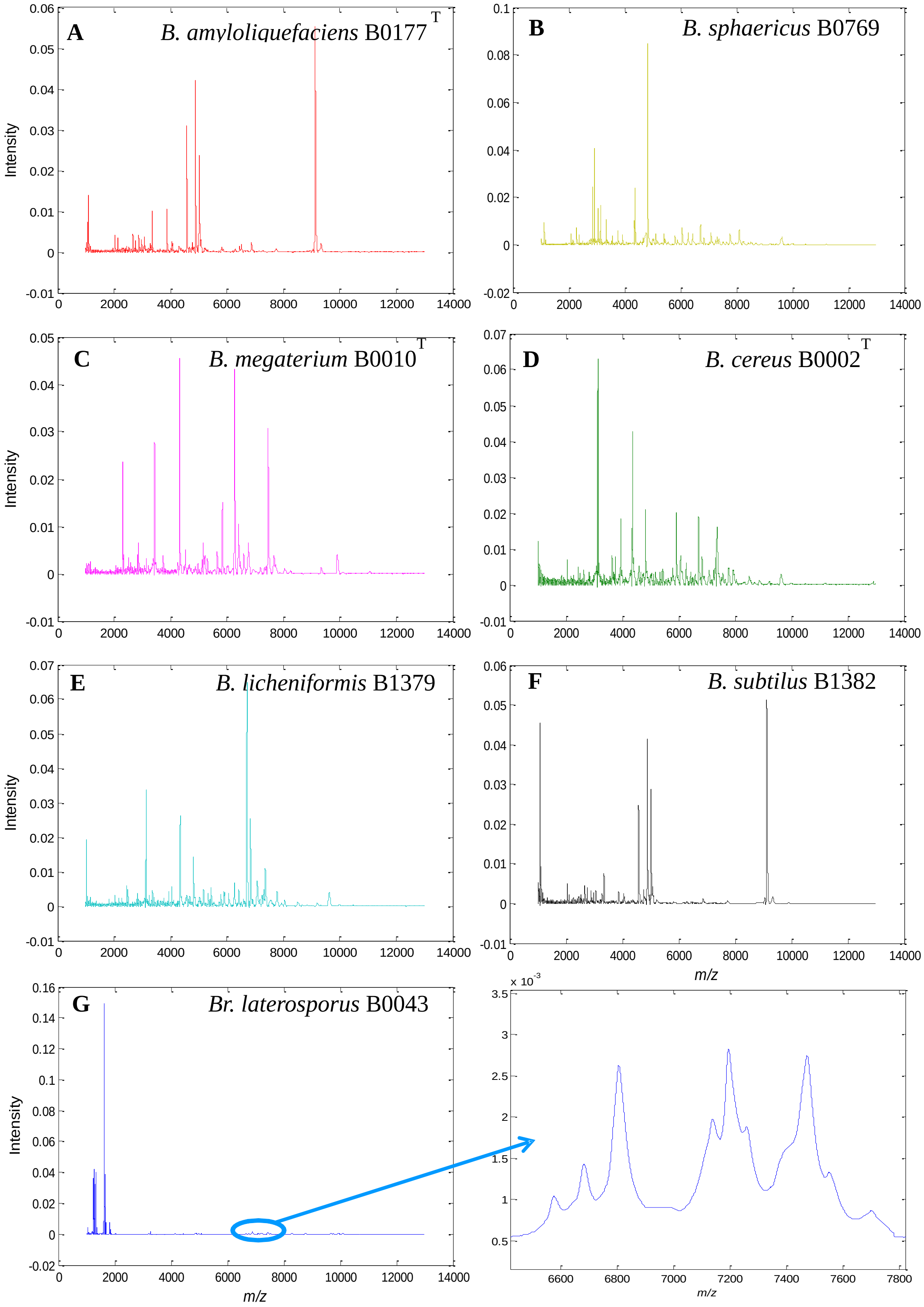


Figure 2.2: Typical MALDI-TOF-MS spectra of (A) *B. amyloliquefaciens* B0177, (B) *B. sphaericus* B0769, (C) *B. megaterium* B0010^T, (D) *B. cereus* B0002, (E) *B. licheniformis* B1379, (F) *B. subtilus* B1382 and (G) *Br. laterosporus* B0034. The panel to the right of (G) is a zoomed in region (highlighted with an ellipse) of the MALDI-TOF-MS spectrum from *Br. laterosporus* B0034. These spectra have been baseline corrected and normalised.

It is known that sample preparation for bacterial analysis is important and this has been discussed before for the analysis of *Bacillus* species (Lasch *et al*., 2009; Lasch *et al*., 2008). It can be seen that these MALDI-TOF-MS spectra are generally distinct from one another and possess good signal-to-noise in the *m*/*z* 1000-13,000 range used. Whilst some spectra are clearly very different, *Br. laterosporus* (which belongs to a different genera) compared with the other *Bacillus* species, it is very difficult to use only visual inspection to identify these different bacteria. Therefore chemometric methods are needed for spectral analysis.

The spectra that were generated from MALDI-TOF-MS are very high dimensional nature and each spectrum contains 0.1078 *m*/*z* intervals after interpolation with ion counts at each value. It is clear from the spectra (Figure 2.2) that much of this information is redundant (i.e. noise), such that direct computation using PCA would be both puerile, as many spurious correlations may be found, as well as being computational intense.

Therefore we used peak picking to select only those *m*/*z* which had arisen from real signals. In this process the intensity weighted variance (IWV) algorithm was used and resulted in a peak table comprising 243 features from the bacteria analysis of 680 samples. This matrix was of dimensions 680 × 243 and significantly reduced from the full spectra (680 × 111,339) and was used for further data analysis.

The scores plots of the first 3 PCs from PCA performed on the peak table matrix are provided in Figure 2.3 and the loadings plot of the first 2 PCs are provided in Figure 2.4. The variables with their absolute loadings (either PC 1 or PC 2) greater than 0.1 are labelled in Figure 2.4 along with their corresponding *m*/*z*. Four main clusters (Figure 2.3) can be observed: (1) the first contained *B.megaterium* and *B. cereus*; (2) comprised *B. subtilus*, *B. amyloliquefaciens* and *B. licheniformis*; (3) contained only *B. sphaericus*; and (4) was also a single member cluster of *Br. laterosporus* (see Figure 2.3A for an annotated 3-D representation). The MALDI-TOF-MS spectra obtained from the analysis of *Br. laterosporus* (Figure 2.2G) were very different to the spectra from the other *Bacillus* species and this was reflected in PCA clusters (Figure 2.3). As can be seen *Br. laterosporus* strains were significantly different in PC2 (Figure 2.3B and 2.3D) which is why when PC2 *versus* PC3 were plotted the groupings of the other 3 clusters were revealed. This was perhaps not surprising as this species belonged to a different bacilli genus, namely *Brevibacillus*.

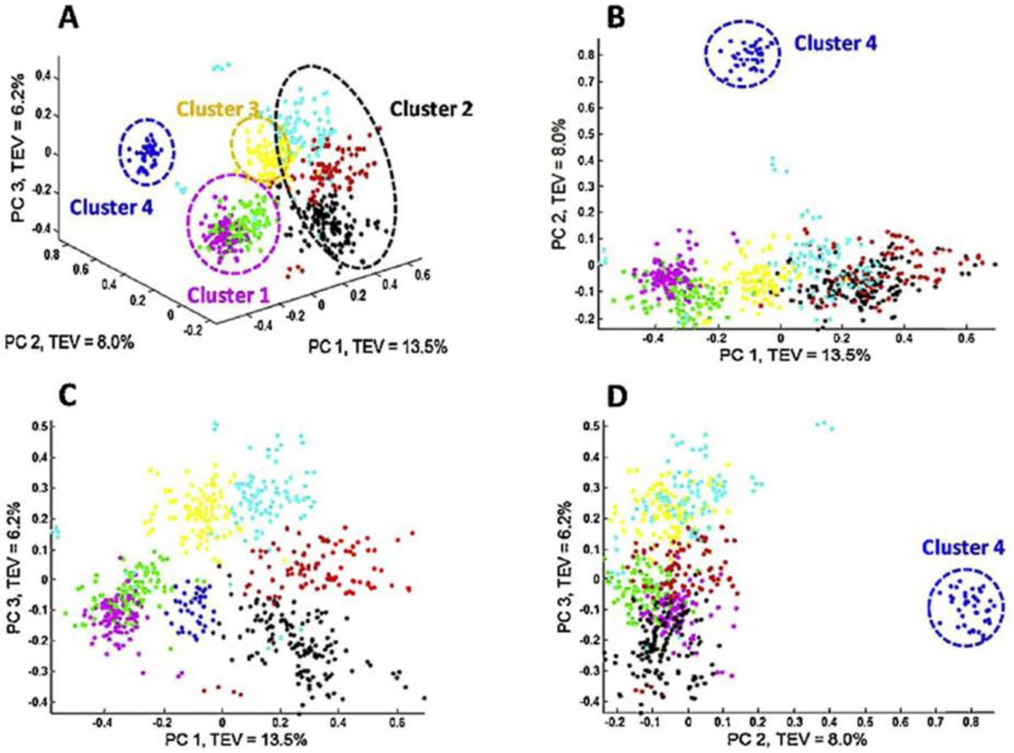


Figure 2.3: PCA scores plots from the peak table matrix after pre-processing the MS data. Multiple principal components are plotted: (A) PC1 *vs*. PC2 *vs*. PC3, (B) PC1 *vs*. PC2, (C) PC1 *vs*. PC3, and (D) PC2 vs. PC3. The colours represent the different species see Table 2.1 for annotations. TEV = total explained variance for the PC score plotted.

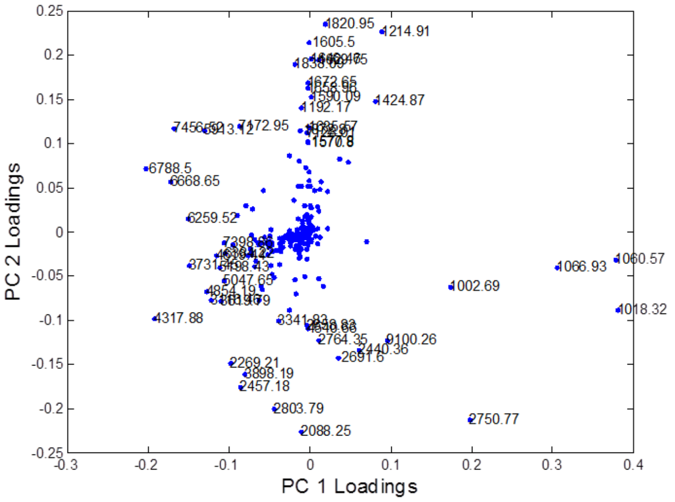


Figure 2.4: PCA loadings plots from the peak table matrix after pre-processing the MS data.

The reason for choosing this set of bacilli is that these species have previously been analysed using a range of classification approaches including miniaturised biochemical test Analytical Profile Index (API), genotyping using 16S rDNA sequencing and an alternative physciochemical methods to MALDI-MS called Raman spectroscopy that measures molecular vibrations of functional groups. Based on the API tests these bacteria have been placed into four different groups (Logan and Berkeley, 1984) consisting of: (I) *B. cereus*, (II) *Br. laterosporus*, (III) *B. sphaericus*, (IV) *B. megaterium*, *B. subtilis*, *B. licheniformis*, and *B. amyloliquefaciens*. Slightly different clusters were also previously found from 16S rDNA analysis: clusters (I), (II) and (III) from the API were also seen, but the *B. subtilis* group (comprising *B. subtilis*, *B. licheniformis*, and *B. amyloliquefaciens*) was split from *B. megaterium*; in addition, although clustered separated *B. cereus* and *B. megaterium* were relatively close relatives at the genetic level (Goodacre *et al.*, 2000; Lopez-Diez and Goodacre, 2004). The clusters generated from our MALDI-TOF-MS analysis is therefore highly congruent with both phenotypic (API) and phylogenetic markers (16S rDNA), as well as other biophysical characterisation methods based on UV resonance Raman spectroscopy (Lopez-Diez and Goodacre, 2004).

The results above used the quantitative data from the peak intensities, or at least the $\log_{10}$ of the signal to try and make the data appear normally distributed. In preliminary analyses we also attempted square root scaling and this produced similar results; for brevity we report only $\log_{10}$ here. As detailed in the materials and methods we also processed the data so that they were considered qualitative in nature; that is to say, we encoded the mass ions as being present (1) or absent (0). The purpose of employing such a strategy is to test whether such greatly simplified information is still sufficient to discriminate different types of bacteria, either on species level or strain level. Moreover, this would compensate for the fact that MALDI-TOF-MS is not considered truly quantitative. We, and others, have observed differences in the ion intensities of proteins from intact bacteria (Holland *et al.*, 1996) and this significant variation in the peak intensities can be due to various analytical reasons. These are most likely due to small changes in bacteria growth, sample handling and the formation of different co-crystals with the matrix ‘spot’ (Ellis *et al.*, 2007; Cohen and Gusev, 2002). If this qualitative approach were successful, it would suggest that the characterisation of the bacteria based on the MALDI-TOF-MS spectra is in fact not sensitive to such variations and would suggest that MALDI-TOF-MS, as an analytical platform, is robust for bacterial analyses. Moreover, those features which had high probabilities of occurrence in some types of bacteria while absent or much rarer in other types could have significant biological implications and perhaps worth further investigation. Therefore PCoA was performed on the binary peak table matrix and resulted in a highly similar pattern (Figure 2.5) to the one showed in the PCA scores plot (Figure 2.3). This had suggested that based on the information of presence/absence of the features, it was indeed possible to discriminate bacteria on species level.

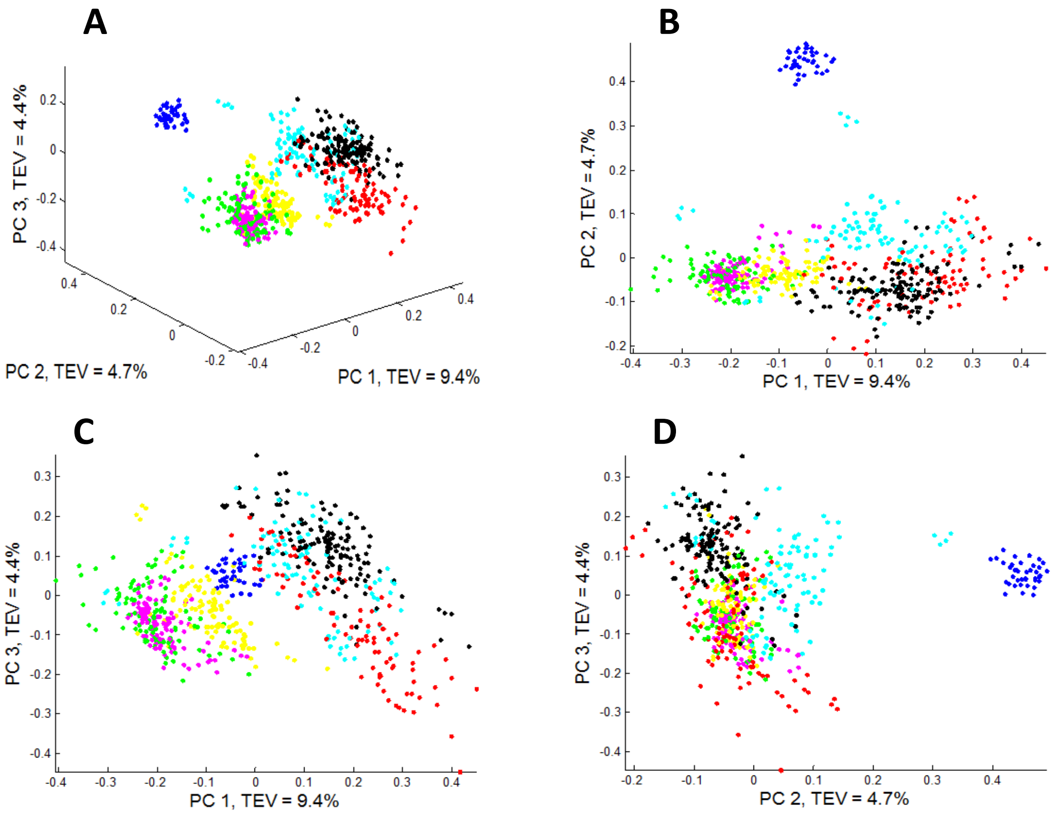


Figure 2.5: PCoA scores plots of the data obtained to show clusters of present and absent peaks using the Jaccard distance model. Multiple principal components are plotted: (A) PC1 vs. PC2 vs. PC3, (B) PC1 vs. PC2, (C) PC1 vs. PC3, and (D) PC2 vs. PC3. The colours represent the different species see Table 2.1 for annotations. TEV = total explained variance for the PC score plotted.

2.3.3 Automated identification of *Bacillus* from their MALDI-TOF-MS spectra

The next stage was to assess whether the information from the MALDI-TOF-MS data were discriminative enough to allow identification using supervised learning methods. The results of these classifications performed at the species level (i.e., 7 classes to be predicted) are given in Tables 2.2 and 2.3 using support vector machines (SVM) for the semi-quantitative and qualitative data, respectively. While prediction accuracies at the strain level (i.e., 34 classes prediction) are provided in SI Tables S2.6 and S2.7. It is very interesting to see that the SVM with Jaccard kernel (i.e., the SVM model based on the presence/absence information) and the SVM with linear kernel gave almost identical prediction accuracies. This suggests that the qualitative information on protein content is sufficient to effect accurate classification, rather than the level of the proteins in the bacterial cells.

Table 2.2: Prediction accuracies of the seven species from *Bacillus* using DAG-SVM with the linear kernel model

	B. am	***B. ce***	***Br. la***	***B. li***	***B. me***	***B. sp***	***B. su***
B. am	92.56%	0.13%	0.00%	0.11%	0.58%	0.95%	5.68%
B. ce	3.37%	83.37%	0.00%	0.12%	11.27%	1.82%	0.05%
Br. la	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
B. li	5.28%	1.41%	0.00%	80.26%	2.65%	3.93%	6.47%
B. me	0.10%	9.22%	0.00%	0.00%	90.67%	0.01%	0.00%
B. sp	1.37%	1.91%	0.00%	2.41%	0.09%	94.23%	0.01%
B. su	6.46%	0.00%	0.00%	2.13%	0.00%	0.00%	91.42%

B. am: B. amyloliquefaciens, B. ce: B. cereus, Br. la: Br. laterosporus, B. li: B. licheniformis, B. me: B. megaterium, B. sp: B. sphaericus and B. su: B. subtilis.

Table 2.3: Prediction accuracies of the seven species from *Bacillus* using DAG-SVM with the Jaccard kernel model

	B. am	***B. ce***	***Br. la***	***B. li***	***B. me***	***B. sp***	***B. su***
B. am	91.29%	0.23%	0.00%	0.14%	0.85%	1.36%	6.14%
B. ce	3.09%	81.75%	0.00%	0.02%	12.26%	2.67%	0.22%
Br. la	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
B. li	5.67%	2.57%	0.00%	78.64%	1.62%	3.71%	7.79%
B. me	0.04%	8.79%	0.00%	0.01%	91.12%	0.04%	0.00%
B. sp	1.06%	4.17%	0.00%	1.98%	0.30%	92.45%	0.05%
B. su	4.23%	0.00%	0.00%	1.46%	0.00%	0.00%	94.31%

B. am: B. amyloliquefaciens, B. ce: B. cereus, Br. la: Br. laterosporus, B. li: B. licheniformis, B. me: B. megaterium, B. sp: B. sphaericus and B. su: B. subtilis.

For the species classification models, the SVM with a linear kernel had an average correct classification rate (CCR) of 89.27% and the SVM with the Jaccard kernel providing 88.92% average CCR. The naïve Bayesian classifier accuracy was slightly worse (77.69% average CCR). For all classification models *Br. laterosporus* was never mis-classified which is perhaps unsurprising as it is a difference genus. *B. cereus* and *B. megaterium* were sometimes misclassified as each other, which was also to be expected as these are phylogenetically similar (Logan and Berkeley, 1984). Finally, the *B. subtilis* group comprising *B. amyloliquefaciens, B. licheniformis* and *B. subtilis* which are similar at the biochemical and genetic level (Wang *et al*., 2007) were also occasionally misclassified as each other. If these were taken as a single group the classification for these three species (e.g. in Table 2.3) would increase from 91.29%, 78.64%, 94.31% to 97.57%, 92.10% and 100% for *B. amyloliquefaciens*, *B. licheniformis* and *B. subtilis*, respectively. The fact that such observations were consistent across all the classification models indicates this is a model independent general trend and a reflection of the phenotypic characteristics being measured using MALDI-TOF-MS.

The CCRs of the classification models for strain (*n*=34) classification is as expected much worse than those at the species level. The average CCR for these models ranged from 45.88% to 54.04% (SI Tables S2.7 and S2.6) for the qualitative and semi-quantitative models. As expected the misclassification of these bacterial strains usually occurred within the same species but to different strains. These may seem poor but considering the fact that there were 34 strains analysed this is a large number of classes and the expected CCR from a random classification model would be only 2.9%. Therefore the prediction accuracies of these models were still very impressive. It was also notable that the semi-quantitative classifier was ~9% better than the qualitative model which suggests that unlike the species classification the information on the peak intensities might also be required to achieve better discrimination between the strains.

2.4 Concluding remarks

MALDI-TOF-MS is gaining popularity for microbial classification and identification (Patel, 2013; Croxatto *et al.*, 2012; Marvin *et al.*, 2003; Wieser *et al.*, 2012). These results in information on the protein content of the organism under study and this proteomic barcode can be used to characterise the bacteria under investigation. However, in order to generate a consistent barcode, the analytical approach must be optimised and tested. In this study we assessed 10 different matrices with 4 different sample preparation approaches. These 40 conditions were first applied to protein mixtures and the top 3 matrices-preparation methods were then assessed for reproducibility and for the generation of information rich protein profiles on 6 bacteria. This established that sinapinic acid with the mixed sample preparation approach was the preferred method, which is in agreement with other studies (Ryzhov *et al*., 2000; Gantt *et al*., 1999).

This matrix was then used on all 34 bacilli and each bacteria was grown 5 times and each of these biological replicates were analysed 4 times (technical replicates). These 680 MALDI-TOF-MS spectra were collected over a period of 2 weeks. Due to the extended mass range over which the spectra were collected (1000-13,000 *m*/*z*) a drift in the *m*/*z* X-axis was observed which if not corrected would adversely affect bacterial characterisation. This was successfully overcome by aligning the peaks using interval correlation optimised shifting. Preprocessing also involved using asymmetric least squares for baseline removal. Chemometric classifiers were then used on these data and the same data after peak picking using intensity weighted variance. This peak picking reduced the dimensionality of the MS data from a massive 680 samples × 111,339 *m*/*z* channels (75,710,520 data points) to a mere 680 × 243 (165240 data points) and this process did not negatively affect classification.

Classification accuracies at *Bacillus* species level were ~90% for the 7 species under analysis and this was robustly tested using bootstrap analysis. The few misclassifications that were made could be readily explained by very close species similarity of the *B. subtilis* group (*viz. B. amyloliquefaciens, B. licheniformis* and *B. subtilis*). In conclusion we have developed a robust MALDI-TOF-MS data collection and data analysis pipeline that we shall now expand to the analysis of other bacterial groups.

2.5 References

Beavis, R. C., Chait, B. T. and Fales, H. M. 1989. Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. *Rapid Communications in Mass Spectrometry*, **3**, 432-435

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. 1996. The rapid identification of intact microorganisms using mass spectrometry. *Nature Biotechnology*, **14**, 1584-1586

Cohen, L. and Gusev, A. 2002. Small molecule analysis by MALDI mass spectrometry. *Analytical and Bioanalytical Chemistry,* **373**, 571-586

Croxatto, A., Prod'hom, G. and Greub, G. 2012. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *Federation of European Microbiological Societies Microbiology Reviews*, **36**, 380-407

Demirev, P. A., Ho, Y.-P., Ryzhov, V. and Fenselau, C. 1999. Microorganism Identification by Mass Spectrometry and Protein Database Searches. *Analytical Chemistry*, **71**, 2732-2738

Drobniewski, F. A. 1993. *Bacillus cereus* and related species. *Clinical Microbiology Reviews*, **6**, 324-338

Eilers, P. H. 2004. Parametric time warping. *Analytical Chemistry*, **76**, 404-411

Ellis, D. I., Dunn, W. B., Griffin, J. L., Allwood, J. W. and Goodacre, R. 2007. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, **8**, 1243-1266

Engvall, E. 1977. Quantitative Enzyme Immunoassay (ELISA) In Microbiology. *Medical Biology*, **55**, 193-200

Fenselau, C. and Demirev, P. A. 2001. Characterisation of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews*, **20**, 157-171

Freiwald, A. and Sauer, S. 2009. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols*, **4**, 732-742

Fritze, D. 2004. Taxonomy of the genus *Bacillus* and related genera: The aerobic endospore-forming bacteria. *Phytopathology*, **94**, 1245-1248

Fuchs, B. and Schiller, J. 2009. Application of MALDI-TOF mass spectrometry in lipidomics. *European Journal of Lipid Science and Technology*, **111**, 83-98

Gantt, S. L., Valentine, N. B., Saenz, A. J., Kingsley, M. T. and Wahl, K. L. 1999. Use of an internal control for matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry analysis of bacteria. *Journal of the American Society for Mass Spectrometry*, **10**, 1131-1137

Ghelardi, E., Celandroni, F., Salvetti, S., Barsotti, C., Baggiani, A. and Senesi, S. 2002. Identification and Characterisation of toxigenic *Bacillus cereus* strains responsible for two food-poisoning outbreaks. *FEMS Microbiology Letters,* **208**, 129-134

Gidden, J., Denson, J., Liyanage, R., Ivey, D. M. and Lay, J. O. 2009. Lipid compositions in *Escherichia coli* and *Bacillus subtilis* during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry. *International Journal of Mass Spectrometry,* **283**, 178-184

Giebel, R., Worden, C., Rust, S. M., Kleinheinz, G. T., Robbins, M. and Sandrin, T. R. 2010. Microbial fingerprinting using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS): applications and challenges. *Advances in Applied Microbiology,* **71**, 149-84

Griffiths, R. L. and Bunch, J. 2012. A survey of useful salt additives in matrix-assisted laser desorption/ionisation mass spectrometry and tandem mass spectrometry of lipids: introducing nitrates for improved analysis. *Rapid Communications in Mass Spectrometry,* **26**, 1557-1566

Goodacre, R. 2003. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrational Spectroscopy,* **32**, 33-45

Goodacre, R., Heald, J. K. and Kell, D. B. 1999. Characterisation of intact microorganisms using electrospray ionisation mass spectrometry. *Federation of European Microbiological Societies Microbiology Letters,* **176**, 17-24

Goodacre, R. and Kell, D. B. 1996. Pyrolysis mass spectrometry and its applications in biotechnology. *Current Opinion in Biotechnology,* **7**, 20-28

Goodacre, R., Shann, B., Gilbert, R. J., Timmins, M., Mcgovern, A. C., Alsberg, B. K., Logan, N. A. and Kell, D. B. 2000, Detection of the dipicolinic acid biomarker in *Bacillus* spores using Curie point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy, *Analytical Chemistry*, **72**, 119-127

Granum, P. E. 1997. *In Food Microbiology: Fundamentals and Frontiers,* Washington DC, ASM Press, pp. 327-336

Granum, P. E. and Lund, T. 1997. *Bacillus cereus* and its food poisoning toxins. *FEMS Microbiology Letters,* **157**, 223-228

Gulledge, J. S., Luna, V. A., Luna, A. J., Zartman, R. and Cannons, A. C. 2010. Detection of low numbers of *Bacillus anthracis* spores in three soils using five commercial DNA extraction methods with and without an enrichment step. *Journal of Applied Microbiology,* **109**, 1509-1520

Hill, W. E. and Wachsmuth, K. 1996. The polymerase chain reaction: Applications for the detection of foodborne pathogens. *Critical Reviews in Food Science and Nutrition*, **36**, 123-173

Holland, R. D., Wilkes, J. G., Rafii, F., Sutherland, J. B., Persons, C. C., Voorhees, K. J. and Lay, J. O. 1996. Rapid Identification of Intact Whole Bacteria Based on Spectral Patterns using Matrix-assisted Laser Desorption/Ionisation with Time-of-flight Mass Spectrometry. *Rapid Communications in Mass Spectrometry,* **10**, 1227-1232

Jarman, K. H., Daly, D. S., Anderson, K. K. and Wahl, K. L. 2003. A new approach to automated peak detection. *Chemometrics and Intelligent Laboratory,* **69**, 61-76

Krishnamurthy, T., Rajamani, U., Ross, P. L., Jabhour, R., Nair, H., Eng, J., Yates, J., Davis, M. T., Stahl, D. C. and Lee, T. D. 2000. Mass spectral investigations on microorganisms. *Journal of Toxicology Toxin Reviews,* **19**, 95-117

Krishnamurthy, T. and Ross, P. L. 1996. Rapid Identification of Bacteria by Direct Matrix-assisted Laser Desorption/Ionisation Mass Spectrometric Analysis of Whole Cells. *Rapid Communications in Mass Spectrometry,* **10**, 1992-1996

Lasch, P., Beyer, W., Nattermann, H., Stammler, M., Siegbrecht, E., Grunow, R. and Naumann, D. 2009. Identification of *Bacillus anthracis* by Using Matrix-Assisted Laser Desorption Ionisation-Time of Flight Mass Spectrometry and Artificial Neural Networks. *Applied Environmental Microbiology journal,* **75**, 7229-7242

Lasch, P., Nattermann, H., Erhard, M., Staemmler, M., Grunow, R., Bannert, N., Appel, B. and Naumann, D. 2008. MALDI-TOF mass spectrometry compatible inactivation method for highly pathogenic microbial cells and spores. *Analytical Chemistry,* **80**, 2026-2034

Lay Jr, J. O. 2000. MALDI-TOF mass spectrometry and bacterial taxonomy. *Trac Trend Analytical Chemistry,* **19**, 507-516

Logan, N. and Berkeley, R. 1984. Identification of *Bacillus* strains using the API system. *Journal of General Microbiology,* **130**, 1871-1882

Lopez-Diez, E. C. And Goodacre, R. 2004. Characterisation of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry,* **76**, 585-591

Marvin, L. F., Roberts, M. A. and Fay, L. B. 2003. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in clinical chemistry. *Clinica Chimica Acta,* **337**, 11-21

Nemmour, H. and Chibani, Y. New jaccard-distance based support vector machine kernel for handwritten digit recognition. *ICTTA*, 2008. *3rd International Conference.* IEEE Computer Society, pp.1-4

Nicolaou, N., Xu, Y. and Goodacre, R. 2011. MALDI-MS and multivariate analysis for the detection and quantification of different milk species. *Analytical and Bioanalytical Chemistry,* **399**, 3491-3502

Nicolaou, N., Xu, Y. and Goodacre, R. 2012. Detection and quantification of bacterial spoilage in milk and pork meat using MALDI-TOF-MS and multivariate analysis. *Analytical Chemistry,* **84**, 5951-5958

Patel, R. 2013. Matrix-assisted laser desorption ionisation-time of flight mass spectrometry in clinical microbiology. *Clinical Infectious Diseases,* **57**, 564-572

Pineda, F. J., Antoine, M. D., Demirev, P. A., Feldman, A. B., Jackman, J., Longenecker, M. and Lin, J. S. 2003. Microorganism identification by matrix-assisted laser/desorption ionisation mass spectrometry and model-derived ribosomal protein biomarkers. *Analytical Chemistry,* **75**, 3817-3822

Priest, F. G., Barker, M., Baillie, L. W., Holmes, E. C. and Maiden, M. C. 2004. Population structure and evolution of the *Bacillus cereus* group. *Journal of Bacteriology.,* **186,** 7959-7970

Ryzhov, V. And Fenselau, C. 2001. Characterisation of the protein subset desorbed by MALDI from whole bacterial cells. *Analytical Chemistry,* **73,** 746-750

Ryzhov, V., Hathout, Y. and Fenselau, C. 2000. Rapid Characterisation of spores of *Bacillus* cereus group bacteria by matrix-assisted laser desorption-ionisation time-of-flight mass spectrometry. *Applied Environmental Microbiology,* **66**, 3828-3834

Sauer, S. and Kliem, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* **8**, 74-82

Schiller, J., Süß, R., Arnhold, J., Fuchs, B., Leßig, J., Müller, M., Petković, M., Spalteholz, H., Zschörnig, O. and Arnold, K. 2004. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research. *Progress in Lipid Research,* **43**, 449-488

Singer, S. 1991. Introduction to the study of *Bacillus* sphaericus as a mosquito control agent. *Bacterial Control of Mosquitoes and Black Flies.* Netherlands, Springer, pp. 221-227

Smole, S. C., King, L. A., Leopold, P. E. and Arbeit, R. D. 2002. Sample preparation of Gram-positive bacteria for identification by matrix assisted laser desorption/ionisation time-of-flight. *The Journal of Microbiological Methods,* **48**, 107-115

Toh-Boyo, G. M., Wulff, S. S. and Basile, F. 2012. Comparison of sample preparation methods and evaluation of intra-and intersample reproducibility in bacteria MALDI-MS profiling. *Analytical Chemistry,* **84**, 9971-9980

Tomasi, G., Savorani, F. and Engelsen, S. B. 2011. icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A,* **1218**, 7832-7840

Vaidyanathan, S., Rowland, J. J., Kell, D. B. and Goodacre, R. 2001. Discrimination of aerobic endospore-forming bacteria via electrospray-ionisation mass spectrometry of whole cell suspensions. *Analytical Chemistry,* **73,** 4134-4144

Vaidyanathan, S., Winder, C. L., Wade, S. C., Kell, D. B. and Goodacre, R. 2002. Sample preparation in matrix-assisted laser desorption/ionisation mass spectrometry of whole bacterial cells and the detection of high mass (>20 kDa) proteins. *Rapid Communications in Mass Spectrometry,* **16,** 1276-1286

Vargha, M., Takáts, Z., Konopka, A. and Nakatsu, C. H. 2006. Optimisation of MALDI-TOF MS for strain level differentiation of Arthrobacter strains. *Journal of Microbiological Methods,* **66**, 399-409

Vidal-Quist, J. C., Castañera, P. and González-Cabrera, J. 2009. Simple and rapid Method for PCR Characterisation of large *Bacillus thuringiensis* strain collections. *Current Microbiology,* **58**, 421-425

Wang, L.-T., Lee, F.-L., Tai, C.-J. and Kasai, H. 2007. Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA–DNA hybridization in the *Bacillus subtilis* group. *International Journal of Systematic and Evolutionary Microbiology,* **57**, 1846-1850

Welham, K. J., Domin, M. A., Scannell, D. E., Cohen, E. and Ashton, D. S. 1998. The Characterisation of micro-organisms by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry,* **12**, 176-180

Wieser, A., Schneider, L., Jung, J. and Schubert, S. 2012. MALDI-TOF MS in microbiological diagnostics/identification of microorganisms and beyond (mini review). *Applied Microbiology Biotechnology,* **93**, 965-974

Wilkins, C. L. and Lay, J. O. 2005. *Identification of microorganisms by mass spectrometry,* New Jersey, John Wiley and Sons, pp.303

Zara, G., Zara, S., Mangia, N., Garau, G., Pinna, C., Ladu, G. and Budroni, M. 2006. PCR-based methods to discriminate *Bacillus thuringiensis* strains. *Annals of Microbiology,* **56**, 71-76

Zenobi, R. and Knochenmuss, R. 1998. Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews,* **17**, 337-366

2.6 Supplementary Information

2.6.1 Data Processing and Peak Picking

As these data were collected over a relatively long period we also needed to develop spectra pre-processing and peak picking algorithms that allowed robust and reproducible profiles to be generated. This process is detailed below:

In liquid chromatography mass spectrometry (LC-MS) or gas chromatography mass spectrometry (GC-MS) any chromatographic shifts (due to unavoidable changes in retention times of analytes that are being separated) can be aligned based on the mass spectrum and the results of such alignment checked (indeed guided) using the unique fragmentation of the analytes within the aligned spectra. However such orthogonal data do not exist in MALDI-TOF-MS as generally no fragmentation is used; indeed even with TOF-TOF configurations this would not be possible due to the large *m*/*z* used. Thus it is not possible analytically to establish if our alignment and peak picking process was successful, and excessive misalignment of the peaks could have undesired effect on the ability to effect accurate bacterial identification from the MALDI-TOF-MS spectra. We therefore performed PCA on both the $\log_{10}$-scaled peak table matrix and the $\log_{10}$-scaled raw spectra. The results showed that the scores plot obtained from the PCA performed on the peak table matrix were highly similar to that obtained from the raw data (data not shown). To quantify the level of similarity Procrustes analysis (Gower and Dijksterhuis 2004) was performed on the two sets of PC scores using the first 3 PCs and a Procrustes error of 0.2474 was obtained. Given there were 680 samples and merely 3 variables (PC scores), such Procrustes error is considered very low. For comparison, if the order of the samples was randomly permuted the Procrustes error was always greater than 0.99. Thus we can conclude that the patterns represented by these two types of data are highly comparable and this suggests that the information in the raw data had indeed been faithfully translated to the very much smaller peak table matrix.

Table S 2.1: Details of the four different sample preparation methods for MALDI-TOF-MS (A) Mix, (B) overlay, (C) underlay and (D) sandwich

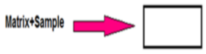
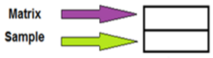
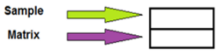
Deposition method	Sample preparation
(A) Mix 	10 µL of the prepared protein mixture were added to 10 µL of each matrix in an Eppendorf tube. The sample was then mixed by vortexing to ensure thorough mixing. 2 µL of the resultant matrix/protein mixture was applied to the MALDI plate and allowed to dry. Once the liquid had evaporated the plate was then ready for analysis.
(B) Overlay 	1 µL of the protein mix sample was applied to the MALDI plate and was allowed to dry. Following evaporation, 1 µL of matrix was added to the protein sample.
(C) Underlay 	1 µL of matrix was applied to the MALDI plate and was allowed to dry. Following evaporation, 1 µL of the protein mix sample was added to the matrix.
(D) Sandwich Matrix Sample Matrix	0.5 µL of matrix was applied to the MALDI plate and was then removed. 1 µL of the protein sample was subsequently added to the plate which was allowed to dry. 1 µL of matrix was finally added after evaporation of the protein sample.

Table S 2.2: MALDI-TOF-MS sample preparation optimisation results from 10 matrices combined with 5 different proteins, using the mix method for preparation

Matrix\Types of protein	Insulin	Cytochrome *c*	Apomyoglobin	Aldolase	Albumin
SA	√	√	√	√	√
CA	√	√	√	√	
DHB	√	√	√		
FA	√	√	√		
HABA	√		√		
CHCA	√	√			
9-AA	√				
THAP	√				
DHAP	√	√	√		
1,8,9Anthractral	√				

The “tick” sign indicates the detection of a particular protein

Table S 2.3: MALDI-TOF-MS sample preparation optimisation results from 10 matrices combined with 5 different proteins, using the overlay method for preparation

Matrix\Types of protein	Insulin	Cytochrome *c*	Apomyoglobin	Aldolase	Albumin
SA	√	√	√	√	
CA	√	√	√		
DHB	√	√	√		
FA	√	√	√		
HABA	√				
CHCA	√				
9-AA	√				
THAP	√	√	√		
DHAP	√				
1,8,9Anthractral					

The “tick” sign indicates the detection of a particular protein

Table S 2.4: MALDI-TOF-MS sample preparation optimisation results from 10 matrices combined with 5 different proteins, using the underlay method for preparation

Matrix\Types of protein	Insulin	Cytochrome *c*	Apomyoglobin	Aldolase	Albumin
SA	√				
CA	√				
DHB					
FA	√	√			
HABA	√				
CHCA	√				
9-AA	√				
THAP	√				
DHAP	√				
1,8,9Anthractral					

The “tick” sign indicates the detection of a particular protein

Table S 2.5: MALDI-TOF-MS sample preparation optimisation results from 10 matrices combined with 5 different proteins, using the sandwich method for preparation

Matrix\Types of protein	Insulin	Cytochrome *c*	Apomyoglobin	Aldolase	Albumin
SA	√	√	√	√	
CA	√	√	√		
DHB	√	√	√		
FA	√	√	√		
HABA					
CHCA	√				
9-AA	√				
THAP	√	√	√		√
DHAP					
1,8,9Anthractral					

The “tick” sign indicates the detection of a particular protein

Table S 2.6: Prediction accuracies of the 34 *Bacillus* strains using DAG-SVM with Linear kernel models

	am1	am2	am3	am4	am5	ce1	ce2	ce3	ce4	ce5	la1	la2	li1	li2	li3	li4	li5	me1	me2	me3	me4	me5	sp1	sp2	sp3	sp4	sp5	su1	su2	su3	su4	su5	su6	su7
am1	15.76%	11.60%	0.24%	49.11%	18.31%	0.00%	1.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.44%	0.19%	0.00%	0.00%	0.00%	0.00%	0.00%	0.51%	0.22%	0.00%	0.00%	0.00%	0.00%	1.86%	0.00%	0.01%	0.00%	0.50%	0.03%
am2	16.67%	45.30%	14.40%	9.96%	11.76%	0.01%	0.16%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.12%	0.39%	0.00%	0.00%	0.00%	0.00%	0.30%	0.00%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	0.00%	0.00%	0.60%	0.00%
am3	28.37%	15.31%	17.84%	12.93%	9.07%	0.03%	0.23%	0.00%	0.00%	0.39%	0.00%	0.15%	0.87%	0.00%	0.03%	0.43%	4.31%	0.00%	0.00%	0.22%	0.00%	0.00%	0.97%	0.00%	0.00%	0.00%	0.08%	3.13%	0.00%	1.34%	2.30%	0.03%	0.89%	1.10%
am4	29.09%	3.64%	1.09%	53.85%	0.00%	0.20%	0.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.40%	0.61%	0.00%	0.00%	0.00%	0.00%	0.00%	2.48%	0.14%	0.00%	0.05%	0.24%	0.25%	1.70%	0.00%	1.30%	2.26%	2.28%	0.00%
am5	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ce1	0.00%	0.00%	0.28%	0.00%	0.00%	54.15%	0.02%	22.30%	10.22%	1.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.05%	0.00%	0.06%	2.38%	5.29%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%	0.07%	0.00%	0.00%
ce2	0.32%	1.19%	4.90%	0.03%	0.11%	5.35%	71.31%	0.02%	7.31%	0.00%	0.17%	0.50%	0.00%	0.00%	0.00%	0.10%	0.06%	2.77%	1.45%	0.01%	2.38%	0.13%	0.15%	0.06%	0.25%	0.26%	0.00%	0.85%	0.09%	0.15%	0.00%	0.07%	0.00%	0.01%
ce3	0.03%	0.00%	0.01%	0.00%	0.02%	28.17%	0.06%	31.37%	34.78%	0.03%	0.00%	0.08%	0.00%	0.00%	0.00%	0.03%	0.97%	0.01%	0.01%	0.85%	2.12%	0.00%	0.00%	0.00%	0.24%	0.01%	0.00%	0.01%	0.00%	0.00%	0.00%	1.17%	0.01%	0.01%
ce4	0.00%	0.00%	0.00%	0.00%	0.00%	6.71%	0.25%	44.58%	42.27%	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.52%	0.00%	0.00%	0.02%	2.13%	0.00%	0.00%	0.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ce5	0.00%	0.00%	0.13%	0.07%	0.00%	5.13%	5.36%	0.33%	21.81%	54.71%	0.00%	0.00%	0.01%	0.05%	0.00%	0.07%	4.99%	0.01%	2.01%	0.04%	1.18%	0.00%	1.88%	0.03%	0.03%	0.32%	0.80%	0.00%	0.00%	0.00%	0.20%	0.56%	0.30%	0.00%
la1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	65.43%	34.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
la2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	55.06%	44.94%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
li1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	89.73%	8.81%	1.06%	0.40%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
li2	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	19.07%	36.97%	23.56%	0.00%	12.98%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.41%
li3	0.10%	0.00%	0.00%	0.12%	0.00%	0.03%	0.00%	0.05%	0.01%	0.00%	0.20%	0.00%	7.43%	20.19%	68.73%	2.49%	0.49%	0.00%	0.03%	0.02%	0.00%	0.00%	0.00%	0.01%	0.00%	0.08%	0.00%	0.00%	0.01%	0.00%	0.00%	0.03%	0.00%	0.00%
li4	3.98%	0.06%	0.23%	6.04%	0.00%	0.00%	0.10%	0.00%	0.00%	0.48%	0.00%	0.00%	9.99%	12.25%	14.97%	25.56%	15.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.04%	0.03%	0.00%	0.00%	0.18%	1.89%	0.00%	0.04%	2.83%	6.19%	0.02%
li5	2.36%	7.35%	2.47%	4.72%	0.00%	1.33%	0.00%	1.55%	0.15%	2.56%	0.00%	0.00%	13.64%	11.31%	2.74%	22.69%	23.54%	0.00%	0.00%	0.51%	0.00%	0.00%	0.15%	0.01%	0.00%	0.00%	0.10%	2.03%	0.00%	0.00%	0.09%	0.69%	0.00%	0.00%
me1	0.00%	0.35%	0.00%	0.00%	0.00%	0.00%	0.43%	0.43%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	94.86%	0.00%	0.00%	0.18%	3.53%	0.00%	0.00%	0.00%	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
me2	0.00%	0.00%	0.00%	0.00%	14.48%	4.36%	0.00%	0.14%	0.00%	0.25%	0.00%	0.00%	0.00%	0.01%	0.00%	0.01%	0.00%	0.05%	60.67%	0.78%	18.69%	0.00%	0.00%	0.00%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%
me3	0.00%	0.00%	0.00%	0.00%	0.03%	0.11%	0.00%	1.06%	0.38%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.23%	0.00%	0.05%	88.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
me4	0.00%	0.00%	0.00%	0.00%	0.00%	2.09%	0.22%	3.42%	3.43%	0.02%	0.00%	0.00%	0.00%	0.45%	0.00%	0.00%	0.00%	25.75%	15.43%	0.00%	49.11%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.05%
me5	0.03%	5.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	0.00%	93.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
sp1	0.00%	0.00%	0.00%	0.31%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	99.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
sp2	0.63%	0.00%	0.03%	3.62%	0.00%	0.00%	0.06%	1.12%	0.63%	1.42%	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	6.50%	0.00%	0.00%	0.00%	0.00%	0.00%	1.00%	74.11%	0.00%	0.30%	9.53%	0.92%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%
sp3	0.00%	0.00%	0.27%	0.00%	0.00%	0.11%	0.91%	1.62%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.33%	0.00%	0.00%	0.00%	0.04%	0.08%	0.00%	8.14%	5.73%	73.94%	0.07%	5.06%	0.08%	0.51%	1.96%	0.72%	0.00%	0.41%	0.00%
sp4	0.08%	0.00%	0.00%	0.00%	0.00%	3.28%	0.13%	0.51%	0.23%	0.33%	0.13%	0.03%	0.00%	0.00%	0.11%	0.01%	0.14%	2.91%	0.02%	0.01%	0.81%	0.00%	8.34%	0.08%	1.96%	74.77%	3.92%	0.00%	0.00%	0.16%	1.93%	0.00%	0.00%	0.10%
sp5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.91%	0.00%	0.00%	97.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
su1	0.04%	0.86%	0.39%	0.06%	0.00%	0.45%	0.01%	0.16%	0.00%	0.00%	0.00%	0.01%	5.62%	0.00%	0.00%	0.46%	0.01%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.20%	70.04%	0.54%	0.00%	0.38%	2.10%	15.97%	2.58%
su2	1.16%	0.00%	0.00%	1.81%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.03%	0.00%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.50%	0.26%	0.00%	0.00%	0.00%	10.19%	34.36%	50.98%	0.00%	0.00%	0.53%	0.00%
su3	0.00%	0.00%	1.39%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.28%	39.47%	57.22%	0.32%	0.04%	0.84%	0.37%
su4	0.00%	0.00%	0.58%	0.78%	0.00%	0.16%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%	0.00%	0.00%	0.00%	0.00%	0.05%	0.24%	0.37%	17.54%	0.17%	18.98%	1.20%	0.66%	59.19%
su5	0.17%	8.85%	**0.00%**	0.23%	0.00%	0.12%	0.00%	0.00%	0.00%	0.27%	0.00%	0.00%	0.56%	0.00%	0.00%	0.48%	2.93%	0.00%	0.00%	0.09%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	1.56%	4.80%	7.84%	0.00%	1.17%	65.33%	5.21%	0.39%
su6	0.78%	0.20%	0.20%	0.33%	0.03%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.12%	0.28%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.82%	0.00%	0.11%	0.78%	2.39%	80.76%	1.15%
su7	0.00%	0.00%	2.45%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.18%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.99%	0.00%	0.26%	52.81%	0.03%	1.24%	39.04%

The different colours represent the species level identifications

Table S 2.7: Prediction accuracies of the 34 *Bacillus* strains using DAG-SVM with Jaccard kernel model

	am1	am2	am3	am4	am5	ce1	ce2	ce3	ce4	ce5	la1	la2	li1	li2	li3	li4	li5	me1	me2	me3	me4	me5	sp1	sp2	sp3	sp4	sp5	su1	su2	su3	su4	su5	su6	su7
am1	15.76%	11.60%	0.24%	49.11%	0.00%	0.00%	1.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.44%	0.19%	0.00%	0.00%	0.00%	0.00%	0.00%	0.51%	0.22%	0.00%	0.00%	0.00%	0.00%	1.86%	0.00%	0.01%	18.31%	0.50%	0.03%
am2	16.67%	45.30%	0.32%	9.96%	0.00%	0.01%	0.16%	0.00%	0.00%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.00%	0.12%	0.39%	0.00%	0.00%	0.00%	0.00%	0.30%	0.00%	0.00%	0.00%	0.00%	0.00%	14.40%	0.00%	0.00%	0.00%	11.76%	0.60%	0.00%
am3	9.07%	17.84%	1.10%	12.93%	0.03%	0.03%	0.23%	0.00%	0.00%	0.39%	0.00%	0.15%	0.87%	0.00%	0.03%	0.43%	4.31%	0.00%	0.00%	0.22%	0.00%	0.00%	0.97%	0.00%	0.00%	0.00%	0.08%	3.13%	0.00%	1.34%	2.30%	28.37%	0.89%	15.31%
am4	29.09%	3.64%	1.09%	53.85%	0.00%	0.20%	0.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.40%	0.61%	0.00%	0.00%	0.00%	0.00%	0.00%	2.48%	0.14%	0.00%	0.05%	0.24%	0.25%	1.70%	0.00%	1.30%	2.26%	2.28%	0.00%
am5	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ce1	0.00%	0.00%	0.28%	0.00%	0.00%	54.15%	0.02%	22.30%	10.22%	1.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.05%	0.00%	0.06%	2.38%	5.29%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%	0.07%	0.00%	0.00%
ce2	0.32%	1.19%	4.90%	0.03%	0.11%	5.35%	71.31%	0.02%	7.31%	0.00%	0.17%	0.50%	0.00%	0.00%	0.00%	0.10%	0.06%	2.77%	1.45%	0.01%	2.38%	0.13%	0.15%	0.06%	0.25%	0.26%	0.00%	0.85%	0.09%	0.15%	0.00%	0.07%	0.00%	0.01%
ce3	0.03%	0.00%	0.01%	0.00%	0.02%	28.17%	0.06%	31.37%	34.78%	0.03%	0.00%	0.08%	0.00%	0.00%	0.00%	0.03%	0.97%	0.01%	0.01%	0.85%	2.12%	0.00%	0.00%	0.00%	0.24%	0.01%	0.00%	0.01%	0.00%	0.00%	0.00%	1.17%	0.01%	0.01%
ce4	0.00%	0.00%	0.00%	0.00%	0.00%	6.71%	0.25%	44.58%	42.27%	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.52%	0.00%	0.00%	0.02%	2.13%	0.00%	0.00%	0.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ce5	0.00%	0.00%	0.13%	0.07%	0.00%	0.07%	1.18%	0.33%	5.36%	54.71%	0.00%	0.00%	8.33E-05	0.05%	0.00%	5.13%	4.99%	0.01%	2.01%	0.04%	21.81%	0.00%	1.88%	0.03%	0.03%	0.32%	0.80%	0.00%	0.00%	0.00%	0.20%	0.56%	0.30%	0.00%
la1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	65.43%	34.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
la2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	55.06%	44.94%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
li1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	89.73%	8.81%	1.06%	0.40%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
li2	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	19.07%	36.97%	23.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.98%	0.00%	0.00%	7.41%
li3	0.10%	0.00%	0.00%	0.12%	0.00%	0.03%	0.00%	0.05%	8.33E-05	0.00%	0.20%	0.00%	7.43%	20.19%	68.73%	2.49%	0.49%	0.00%	0.03%	0.02%	0.00%	0.00%	0.00%	8.33E-05	0.00%	0.08%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.03%	0.00%	0.00%
li4	3.98%	0.06%	0.23%	6.04%	0.00%	0.00%	0.10%	0.00%	0.00%	0.48%	0.00%	0.00%	6.19%	0.03%	14.97%	25.56%	15.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.04%	12.25%	0.00%	0.00%	0.18%	1.89%	0.00%	0.04%	2.83%	9.99%	0.02%
li5	2.36%	11.31%	2.47%	4.72%	0.00%	1.33%	0.00%	1.55%	13.64%	2.56%	0.00%	0.00%	0.15%	0.00%	2.74%	23.54%	0.69%	0.00%	0.00%	0.51%	0.00%	0.00%	0.15%	0.01%	0.00%	0.00%	0.10%	2.03%	0.00%	0.00%	0.09%	22.69%	7.35%	0.00%
me1	0.00%	0.35%	0.00%	0.00%	0.00%	0.00%	0.43%	0.43%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	94.86%	0.00%	0.00%	0.18%	3.53%	0.00%	0.00%	0.00%	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
me2	0.00%	0.00%	0.00%	0.00%	14.48%	4.36%	0.00%	0.14%	0.00%	0.25%	0.00%	0.00%	0.00%	0.01%	0.00%	8.33E-05	0.00%	0.05%	60.67%	0.78%	18.69%	0.00%	0.00%	0.00%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%
me3	0.00%	0.00%	0.00%	0.00%	0.03%	0.11%	0.00%	1.06%	0.38%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.23%	0.00%	0.05%	88.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
me4	0.00%	0.00%	0.00%	0.00%	0.00%	2.09%	0.22%	3.42%	3.43%	25.75%	0.00%	0.00%	0.00%	0.45%	0.00%	0.00%	0.00%	0.02%	15.43%	0.00%	49.11%	0.00%	0.00%	0.00%	6.25E-05	0.01%	8.33E-05	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.05%
me5	0.03%	5.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.32%	0.00%	0.00%	0.00%	93.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
sp1	0.00%	0.00%	0.00%	0.31%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	99.69%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
sp2	0.63%	0.00%	0.03%	3.62%	0.00%	0.00%	0.06%	1.12%	0.63%	1.42%	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	6.50%	0.00%	0.00%	0.00%	0.00%	0.00%	1.00%	74.11%	0.00%	0.30%	9.53%	0.92%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%
sp3	0.00%	0.00%	0.27%	0.00%	0.00%	0.11%	0.91%	1.62%	0.00%	0.00%	0.00%	0.00%	0.00%	8.14%	0.00%	0.33%	0.00%	0.00%	5.73%	0.04%	0.08%	0.00%	0.00%	0.00%	73.94%	0.07%	0.00%	0.08%	0.51%	1.96%	0.72%	0.00%	0.41%	5.06%
sp4	0.08%	0.00%	0.00%	3.92%	0.00%	3.28%	0.13%	0.51%	0.23%	0.33%	0.13%	0.03%	0.00%	8.34%	0.11%	0.01%	0.14%	2.91%	0.02%	0.01%	0.81%	0.00%	0.00%	0.08%	1.96%	74.77%	0.00%	0.00%	0.00%	0.16%	1.93%	0.00%	0.00%	0.10%
sp5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.91%	0.00%	0.00%	97.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
su1	0.04%	0.86%	0.39%	0.06%	0.00%	0.45%	6.25E-05	0.16%	0.00%	0.00%	0.00%	8.33E-05	5.62%	0.00%	0.00%	0.46%	0.01%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.20%	70.04%	0.54%	0.00%	0.38%	2.10%	15.97%	2.58%
su2	1.16%	0.00%	0.00%	1.81%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.03%	0.00%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.50%	0.26%	0.00%	0.00%	0.00%	10.19%	34.36%	50.98%	0.00%	0.00%	0.53%	0.00%
su3	0.00%	0.00%	1.39%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.28%	39.47%	57.22%	0.32%	0.04%	0.84%	0.37%
su4	0.00%	0.00%	0.58%	0.78%	0.00%	0.16%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	17.54%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%	0.00%	0.00%	0.00%	0.00%	0.05%	0.24%	0.37%	0.01%	0.17%	18.98%	1.20%	0.66%	59.19%
su5	0.17%	8.85%	7.84%	0.23%	0.00%	0.12%	0.00%	0.00%	0.00%	0.27%	0.00%	0.00%	0.56%	0.00%	0.00%	0.48%	2.93%	0.00%	0.00%	0.09%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.00%	1.56%	4.80%	0.00%	0.00%	1.17%	65.33%	5.21%	0.39%
su6	0.78%	0.20%	0.20%	0.33%	0.03%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.12%	0.28%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.82%	0.00%	0.11%	0.78%	2.39%	80.76%	1.15%
su7	0.00%	0.00%	2.45%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3.18%	0.00%	0.00%	0.00%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.00%	0.00%	8.33E-05	0.00%	0.00%	0.99%	0.00%	0.26%	52.81%	0.03%	1.24%	39.04%

The different colours represent the species level identifications

2.6.2 Supplementary References

Gower, J. C. and Dijksterhuis, G. B. 2004. *Procrustes problems*, *Volume* 3, Oxford, Oxford University Press

Chapter Three

Fractional Factorial Design of MALDI-TOF-MS sample preparations for the optimized detection of phospholipids and acylglycerols

Najla AlMasoud, [a] *Elon Correa,* [a] *Drupad K Trivedi,* [a] *and Royston Goodacre* [a]*

[a] *School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.*

* *Correspondence to Roy Goodacre: roy.goodacre@manchester.ac.uk*

This work has been published as:

AlMasoud, N., Correa, E., Trivedi, D.K. and Goodacre, R. 2016. Fractional Factorial Design of MALDI-TOF-MS sample preparations for the optimized detection of phospholipids and acylglycerols. *Analytical Chemistry,* **88**, 6301-6308.

Leon Correa contributed to this work by creating the fractional factorial design (FFD) and carrying out associated data analysis. Drupad Trivedi provided assistance in reporting the finding of this study. Royston Goodacre contributed to this study via supervision and guidance.

Abstract

Matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) has successfully been used for the analysis of high molecular weight compounds, such as proteins. By contrast, analysis of low molecular weight compounds with this technique has been less successful due to interference from matrix peaks which have a similar mass to the target analyte(s). Recently, a variety of modified matrices and matrix additives have been used to overcome these limitations. An increased interest in lipid analysis arose from the feasibility of correlating these components with many diseases, e.g. atherosclerosis and metabolic dysfunctions. Lipids have a wide range of chemical properties making their analysis difficult with traditional methods. MALDI-TOF-MS shows excellent potential for sensitive and rapid analysis of lipids, and therefore this study focuses on computational-analytical optimisation of the analysis of five lipids (4 phospholipids and 1 acylglycerol) in complex mixtures using MALDI-TOF-MS with fractional factorial design (FFD) and Pareto optimality (PO). Five different experimental factors were investigated using FFD which reduced the number of experiments performed by identifying 720 key experiments from a total of 8064 possible analyses. Factors investigated included: matrices, matrix preparations, matrix additives, additive concentrations and deposition methods. This led to a significant reduction in time and cost of sample analysis with near optimal conditions. We discovered that the key factors to produce high quality spectra were the matrix and use of appropriate matrix additives.

3.1 Introduction

Lipids, among other cellular components such as proteins, carbohydrates and nucleic acids are the most fundamental components found in bacterial cells (Prescher and Bertozzi, 2005). These cellular components have many important functions such as storing energy, cell signalling, as well as comprising the lipid bilayer needed to protect the organism from its environment (Vance and Vance, 2008). The structures of these cellular components are varied due to different combinations of building blocks that they are composed of, and these different polar head groups and acyl chains allow differentiation between bacterial species (Shu *et al*., 2012). Moreover, one important property of lipids is that they are hydrophobic; hence, they are usually dissolved in organic solvents such as chloroform, dichloromethane and hexane rather than aqueous solutions (Cliff *et al*., 2012).

Development in lipid research has accelerated due to the availability of modern analytical technologies such as electrospray ionisation (ESI) coupled with mass spectrometry (MS), often with prior lengthy separation using liquid chromatography (Goodacre *et al*., 2004). Lipidomics involves the analysis of lipids and aims to explore their roles in health and disease. This field has gained an increased interest over the last decade by academics and clinical researchers in different fields as a vital means for studying many medical conditions (Mattila *et al*., 2008; Kenny *et al*., 2010) including biomarkers for cancer (Lee *et al*., 2012; Zemski Berry *et al*., 2011) and microbiological diseases such as anthrax (Li *et al*., 2013).

Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS), (Lay *et al*., 2012; Schiller *et al*., 1999) has also been used by researchers in the field of lipidomics due to its high sensitivity, ease of automation and rapidity. This technique is a powerful tool for analysing microorganisms and biomolecules such as lipids, (Batoy *et al*., 2009; Zemski Berry *et al*., 2011; Stübiger and Belgacem, 2007; Jackson *et al*., 2005; Ryzhov *et al*., 2000; Stübiger *et al*., 2010) carbohydrates (Choi *et al*., 2009) and proteins (Lasch *et al*., 2009; Nicolaou *et al*., 2012) since it provides useful information about molecular masses (*m/z*). The importance of understanding the usual distribution of lipids in samples of interest is fundamental for the development of treatments and understanding of the disease of interest. MALDI-TOF-MS offers the opportunity to analyse lipid mixtures making this technology attractive to researchers interested in understanding underlying healthy and disease states using their chosen biological systems. Moreover, one of the most fundamental advantages of using this technique is that it is a soft ionisation method leading to the production of little or no ion fragmentation. Many lipidomic studies have utilised additives which are mixed to the matrix solution when analysing lipids using MALDI-TOF-MS. Examples of these additives include (amongst others): lithium chloride, (Jackson *et al.*, 2005; Griffiths *et al.*, 2013) potassium chloride (Griffiths and Bunch, 2012) sodium acetate (Stübiger and Belgacem, 2007) and calcium chloride (Müller *et al.*, 2001). The purpose of using some of these salts is sometimes to simplify the spectra and reduce the background noise. Furthermore, additives allow adducts of interest to be favoured and the concentrations of other adducts to be reduced.

The aim of this study was to optimise experimental conditions for the detection of a mixture of five different lipids via MALDI-TOF-MS using combinations of five different experimental conditions: matrices, matrix additives, additive concentrations, deposition methods as well as matrix preparation methods. This was followed by the use of robust chemometrics to simplify the huge number of possible experiments with a view to using these optimum conditions to analyse lipids extracted from bacteria.

3.2 Materials and Methods

3.2.1 Chemical solvents and acids

Acetonitrile (ACN), methanol, trifluoroacetic acid (TFA) and isopropanol (C_3H_7OH), chloroform ($CHCl_3$), ethanol (EtOH) and HPLC grade water (H_2O) were purchased from Sigma-Aldrich (Dorset, UK).

3.2.2 Matrices

Eight different matrices were used in this study: 2,4,6-trihydroxyacetophenone (THAP), 2,5-dihydroxybenzoic acid (2,5 DHB), 2,6-dihydroxyacetophenone (DHAP), 5-chloro-2-mercaptobenzothiazole (CMBT), 6-aza-2-thiothymine (ATT), 2-4-hydroxybenzeneazo benzoic acid (HABA), dithranol (INN) and *p*-nitroaniline (PNA). All matrices purchased from Sigma-Aldrich (Gillingham, UK).

3.2.3 Matrices additives

All matrix additives were also purchased from Sigma-Aldrich (Gillingham, UK) including: diammonium citrate, ammonium chloride (NH_4Cl), potassium acetate (CH_3CO_2K), potassium chloride (KCl), potassium nitrate (KNO_3), sodium acetate ($C_2H_3NaO_2$), sodium nitrate ($NaNO_3$), sodium dodecyl sulfate ($NaC_{12}H_{25}SO_4$), lithium nitrate ($LiNO_3$), calcium chloride ($CaCl_2$) and EDTA ammonium.

3.2.4 Lipid standards

Lipids are typically found in bacterial membranes and the choice of the lipids analysed in this study reflects this. Five different lipids were used in this study: 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1'-rac-glycerol) (PG), L-α-hosphatidyl-ethanolamine dioleoyl (PE), 1,2-distearoyl-*sn*-glycero-3-phospho-(1'-*rac*-glycerol) (PG), 1,2-di-(13Z-docosenoyl)-*sn*-glycero-3-phosphocholine (PC) and 1,2 diacyl-3-O-(*α*-D-galactosyl1-6)-*β*-D-galactosyl-sn-glycerol (DGDG), purchased from Avanti Polar Lipids Inc. (Delfzyl, The Netherlands). These were named as follows: lipid B, lipid C, lipid E, lipid F and lipid G, respectively. Each lipid was dissolved in (1:6) MeOH:$CHCl_3$ (*v/v*). This was followed by mixing the lipid together to form a lipid mixture. Each lipid within the mixture was at an equal concentration (1.08 mM). The reason behind the lipids nomenclature was that initially seven lipids were used, however; lipids A [L-α-phosphatidylglycerol (*E. coli*) and D (L-α-phosphatidylethanolamine (*E. coli*)] mass spectra did not have definitive and reproducible unique peaks that could be assigned to these lipids. Therefore, these two lipids were excluded from any further experiments (data not shown).

3.2.5 Preparation of matrices and matrix additives

Each matrix was prepared at 10 mg/mL with six different matrix preparations including: MeOH:0.1% TFA (80:20, *v/v*), MeOH:H_2O (70:30, *v/v*), $CHCl_3$:MeOH:H_2O (40:40:20 ,*v/v*), H_2O:ACN, (50:50 *v/v*), ACN:H_2O:C_3H_7OH (20:20:50, *v/v*) and EtOH:H_2O, (90:10 *v/v*). All the matrix additives were diluted in MeOH:H_2O (80:20, *v/v*). In addition, six different concentrations were used for matrix additives *viz*. 0, 5, 10, 20, 40, and 80 mM.

3.2.6 Spotting of lipid mixture for MALDI-TOF-MS

Details of the matrices and matrix additives along with the preparation of matrices and matrix additives are provided in SI (see Tables S3.1 and S3.2). Three analytical replicates for each experiment (matrix, matrix additive, additive concentration, matrix preparation method and sample deposition method) were prepared for MALDI-TOF-MS analysis.

Three different deposition methods (AlMasoud *et al.*, 2014 (Chapter 2)) were used while ensuring the same amount of analyte is used for MALDI-TOF-MS analysis: (i) dried droplet method (mix method), where the analyte and the matrix are first mixed at equal volumes (1 µL each) followed by spotting 2 µL of the resultant mixture onto a MALDI plate and allowing the mixture to dry; (ii) thin layer method (underlay method), where 1µL of the matrix was applied onto a MALDI plate and was allowed to dry, then 1 µL of the analyte was added to the matrix and allowed to dry; and (iii) overlay method, in which 1 µL of the analyte was applied onto a MALDI plate and allowed to dry, then 1 µL of matrix was spotted and the mixture allowed to dry.

3.2.7 Preparation of lipid extracts from bacterial samples and human serum

Gram-positive (*Bacillus. cereus*, *Bacillus. subtilis*) and Gram-negative (*Escherichia coli*, and *Pseudomonas aeruginosa*) bacteria were grown in LB media for 10 h at 37°C. Lipids were extracted from quenched bacterial samples as described in SI, and the procedure used for lipid extraction from artificial human serum (Sigma-Aldrich, Dorset, United Kingdom) is also described in SI. For MALDI-TOF-MS analysis of the bacterial lipid extracts, the samples were reconstituted in 100 µL of 80:20 (*v/v*) methanol:HPLC water. The extracted lipid pellet from artificial human serum was reconstituted in 100µL of HPLC grade water on the day of analysis where MALDI protocol was followed for sample preparation.

3.2.8 Operating the MALDI-TOF-MS

720 experiments were carried out in 18 batches using an AXIMA-Confidence (Shimadzu Biotech, Manchester, UK) mass spectrometer. This MALDI-TOF-MS device contained a nitrogen pulsed UV laser with a wavelength of 337 nm as described previously (AlMasoud *et al*., 2014 (Chapter 2)). The power of the laser was set to 100 mV. Each profile contained 20 shots, and 78 profiles were collected using a circular raster pattern. The MS was operated in positive ion mode and reflectron TOF was used over the range 100-1000 *m/z.* Approximately 40 experiments were carried out each day and the analysis time took ~4 months. The results of this analysis generated 2160 MALDI-TOF-MS spectra: 720 experiments × 3 technical replicates.

3.2.9 Data analysis

3.2.9.1 Fractional Factorial Design

Five factors were tested in this study (matrix type, matrix preparation, type of matrix additive, additive concentration and sample deposition method) to detect lipids using MALDI-TOF-MS. Considering all the parameters under study, the full factorial design or the total number of unique experiments that could be generated is:

[(8 matrices) × (11 matrix additives) × (6 matrix preparation methods) × (5 additive concentrations) × (3 deposition methods)**] + [**(8 matrices) × (6 matrix preparation methods) × (3 deposition methods)**] = 8064 experiments**

> (The product on the left hand side of "+" corresponds to samples where a matrix additive was used (i.e., 5, 10, 20, 40, 80 mM), whereas the product on

the right hand side of "+" corresponds to samples where no matrix additive was used (i.e., 0 mM)).

This is a large number of experiments to perform in the laboratory, and an exhaustive analysis of the search space would be unnecessarily laborious, time-consuming and expensive. Moreover, many experiments are probably redundant in terms of indicating which combinations of mixtures enhance the MALDI-TOF-MS signal, as these conditions will have multiple interacting factors. Therefore, in order to determine which experiments were to be carried out, a fractional factorial design (FFD) was used to filter the search space. FFD is based on a principle known as sparsity-of-effects (SOE) (Mukerjee and Wu, 2006). This principle assumes that the main effects with low-order interactions dominate a system. FFD selects only a subset (fraction) from a full experimental run. This significantly low fraction of experiments is expected to be sufficient to understand the underlying problem (Quinn and Keough, 2002). FFD was computed using MATLAB (The MathWorks, Inc., Natick, Massachusetts, USA) version 2012b. The FFD algorithm returned 720 (Table S3.3) (see list in the attached material: Chapter 3_SI)) suggested experiments (less than 10% of the full design) to be performed. Each of these 720 MALDI experiments was then assessed in the laboratory.

3.2.9.2 Data pre-processing

MATLAB was used for pre-processing and data analysis. Initially, the spectra were baseline corrected using statistics-sensitive non-linear iterative peak-clipping algorithm (SNIP) (Ryan *et al.*, 1988), then the peaks were aligned using the "msalign" function from MATLAB and Bioinformatics Toolbox Release 2012b (The MathWorks, Inc., Natick, Massachusetts, United States) (Monchamp *et al.*, 2007). This was followed by: (1) peak detection and all known expected peaks for each lipid (Table S3.4) and (2) for each individual lipid the intensities of its expected peaks are then summed up to represent the evidence of detection of that respective lipid on the spectrum analysed (i.e., the higher the sum of the expected peaks, the higher the evidence that the lipid has been detected).

3.2.9.3 Multiple objectives measured

Despite the complexity of MALDI data, even more so in lipidomics, various aspects need to be considered. Therefore, *four* different multiple objectives were measured to evaluate the quality of each proposed experimental solution (combination of factors under study). These measurements synthesise highly desirable objectives that are in general difficult to achieve in MALDI experiments, namely: (i) high reproducibility, (ii) high signal to noise ratio, (iii) high peak intensity of each lipid under study and (iv) detection of all lipids included in the mixture, since that the correct identification of all lipids present in the sample is a fundamental requirement for systematically searching for the simplest possible global optimal solution.

(Objective function 1) High reproducibility of the spectra: spectra reproducibility was evaluated based on the correlation coefficient between the three generated analytical replicates using the following formula:

$$\mathrm{Rep}(e_k) = \frac{1}{3}\left(\frac{\mathrm{cov}(e_{k_1}, e_{k_2})}{\sqrt{\mathrm{var}(e_{k_1})\mathrm{var}\,(e_{k_2})}} + \frac{\mathrm{cov}(e_{k_1}, e_{k_3})}{\sqrt{\mathrm{var}(e_{k_1})\mathrm{var}\,(e_{k_3})}} + \frac{\mathrm{cov}(e_{k_2}, e_{k_3})}{\sqrt{\mathrm{var}(e_{k_2})\mathrm{var}\,(e_{k_3})}}\right) \qquad (\textbf{Equation 3.1})$$

where, e_k represents the k^{th} replicates of experiment e (in this work k = 1, 2, 3). The spectra generated from MALDI-TOF-MS suggested that they were highly reproducible (Figure S3.5).

(Objective function 2) High signal-to-noise ratio: the signal to noise ratio (Snr) was based on extracting all identified and valid peaks from the original spectrum with the undesired noise remaining and was computed based on the following formula:

$$\mathrm{Snr} = \frac{\sum \text{extracted signal}}{\sum(\text{ spectrum} - \text{extracted signal})} \qquad (\textbf{Equation 3.2})$$

$$\mathrm{Snr} = \frac{\sum \text{ signal}}{\sum \text{noise}}$$

(Objective function 3) Number of lipids detected: a lipid is considered to have been detected in the mixture when at least one of its unique characteristic peaks are present in the spectrum. This objective function then counts how many different lipids (out of the 5 composing the mixture) have been positively detected within the MALDI-TOF-MS spectrum of lipid mixture.

(Objective function 4) High peak intensity of each lipid: for each lipid, the sum of all of its known expected peaks (Table S3.4) was considered as a single value or evidence score indicating the strength of the detection of that lipid. For instance, taking lipid B (see Table S3.4) as an example, the sum of its known expected peaks would be:

$$\text{Lipid B (intensity)} = \underbrace{\textbf{peak@688} + \textbf{peak@710} + \textbf{peak@726}}_{\text{Evidence for lipid B detection}} \qquad (\textbf{Equation 3.3})$$

As this study analysed 5 different lipids, this objective function (#4) is actually subdivided into the sum of 5 different lipid intensities as shown in Equation 3.4.

3.2.9.4 Pareto optimality

The Pareto optimality (PO) principle was first introduced by Smilde (Smilde *et al*., 1986). PO is defined by experiments having better results for some objectives in conjunction with possibly not as good results for other objectives. The aim of PO was to identify the samples (spectra) or points for which no other sample is better than them for at least one of the objectives measured.

As there are *four* objectives to be satisfied in this research (reproducibility, high signal to noise ratio, high peak intensity and the number of lipids detected), we used the PO approach to identify relevant solutions. In PO, a solution is achieved when each objective, in our case *four* main objectives (note that peak intensity is subdivided into five objective functions one of each lipid under study); is optimised to the extent that it is acceptable to the decision maker and without other objectives suffering as a result of this process if further optimisation were to take place (Ramı *et al*., 2002). In PO, a solution is considered valid and said to be “in the Pareto front” if no other solution dominates that solution in all objectives being measured. Otherwise, the solution is said to be dominated, is not in the Pareto front and is rejected. As at the end of the process there may be many solutions in the Pareto front, the decision of which solution is the best depends on the user expectations and requirements for each objective measured (Figure S3.2 illustrates PO). The objective function used for computation of PO is then given by the following multi-objective function:

$$= \frac{\sum \text{signal}}{\sum \text{noise}}$$

$$\mathrm{Rep}(e_k) = \frac{1}{3}\left(\frac{\mathrm{cov}(e_{k1}, e_{k2})}{\sqrt{\mathrm{var}(e_{k1})\mathrm{var}(e_{k2})}} + \frac{\mathrm{cov}(e_{k1}, e_{k3})}{\sqrt{\mathrm{var}(e_{k1})\mathrm{var}(e_{k3})}} + \frac{\mathrm{cov}(e_{k2}, e_{k3})}{\sqrt{\mathrm{var}(e_{k2})\mathrm{var}(e_{k3})}}\right)$$

count of different lipids positively detected

$$f_{multi-objective} = snr + rep + \#_{lipids\ detected} + int_{lipidB} + int_{lipidC} + int_{lipidE} + int_{lipidF} + int_{lipidG}$$

Lipid B(intensity) = peak@688 + peak@710 + peak@726

(**Equation 3.4**)

3.3 Results and Discussion

3.3.1 Systematic matrix optimisation

There has been wide interest in addressing questions related to the role of lipids and their biological function due to their importance in cells (Gidden *et al*., 2009; Dong *et al*., 2013), using modern analytical techniques such as MALDI-TOF-MS, (Schiller *et al*., 2004; Cornett *et al*., 2007; Fuchs and Schiller, 2009) since MALDI-TOF-MS is a very powerful technique for lipid analysis. For this purpose a mixture of 5 lipids were selected due to their importance in bacterial cell membranes and these included: 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1'-rac-glycerol) PG (lipid B), L-α-phosphatidylethanolamine dioleoyl PE (lipid C), 1,2-distearoyl-*sn*-glycero-3-phospho-(1'-*rac*-glycerol) (lipid E), 1,2-di-(13Z-docosenoyl)-*sn*-glycero-3-phosphocholine PC (lipid F) and 1,2 diacyl-3-O-(*α*-D-galactosyl1-6)-*β*-D-galactosyl-*sn*-glycerol (lipid G). Our central focus was to optimise MALDI-TOF-MS for lipids using different factors: matrix, matrix preparation, matrix additive, additive concentration and deposition method. Therefore different experimental combinations resulting in different lipid preparations were spotted directly onto the wells of the MALDI plates and MALDI-TOF-MS measurements taken producing corresponding mass spectra.

MALDI-TOF-MS data are quite often challenging to interpret due to the complexity of the spectra acquired. The complexity of the data is due to the presence of different ion species formed from the lipid molecule including $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$ and $[M-H]^-$ (Zemski Berry *et al.*, 2011). Hence, a selection of eight different commonly used matrices, including: THAP (Stübiger and Belgacem, 2007; Lee *et al.*, 2013), DHB (Griffiths and Bunch, 2012; Lee *et al.*, 2013; Petkovic *et al.*, 2009; Marto *et al.*, 1995), DHAP (Gorman *et al.*, 1996; Stübiger and Belgacem, 2007), CMBT (Zhou *et al.*, 2010), ATT (Shanta *et al.*, 2012; Stübiger *et al.*, 2010), HABA (Przybylski *et al.*, 2010), INN (Le *et al.*, 2012) and PNA (Ham *et al.*, 2005) were used to facilitate the analysis of the lipid mixture. This selection of matrices was based on a literature survey conducted to establish which matrices have been reported to work well for lipid analysis using MALDI-TOF-MS. Some of these matrices were shown to increase background noise and others to decrease it. These finding were also reported previously in the literature (Calvano *et al.*, 2011; Schiller *et al.*, 1999).

3.3.2 Fractional Factorial Design used to identify optimum conditions for MALDI-TOF-MS

The significance of different factors (matrix, matrix preparation, matrix additive, additive concentration and deposition method) was assessed using 18 peaks, which were directly assigned to the corresponding lipids in the mixture. Table S3.4 shows the assignment of these peaks. These 18 peaks were extracted from the results of the 720 experimental protocols selected by FFD. Figure 3.1 illustrates the 3D scores plot of MALDI-TOF-MS data using multiple objectives measured for 720 × 3 = 2160 spectra generated in this study. The different characteristics (shapes and colours) shown in this 3D plot represent different matrices. For instance, the red and blue squares represent 6-aza-2-thiothymine (ATT) and 2,6-dihydroxyacetophenone (DHAP), respectively. Moreover, the size of the shapes is proportional to the quality of the spectra according to Equation 3.4 (the bigger, the better).

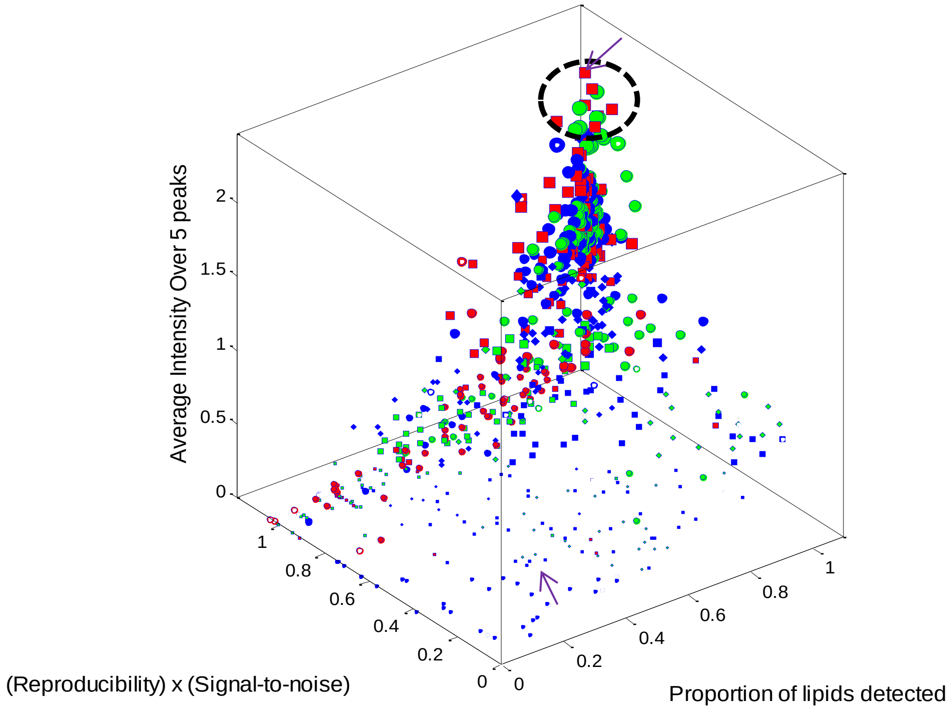


Figure 3.1: 3D scores plot of MALDI-TOF-MS data using multiple objectives measured for all 2160 spectra. The different characteristics (shapes and colours) represent different experimental conditions. The size of the shapes is proportional to the quality of the spectra; i.e. the bigger, the better. A key was not used to represent the different characteristics as 720 experiments conditions making this impossible. The purple arrow at the top of the 3D score plot indicates the top solution and the bottom arrow indicates one of the worse solutions.

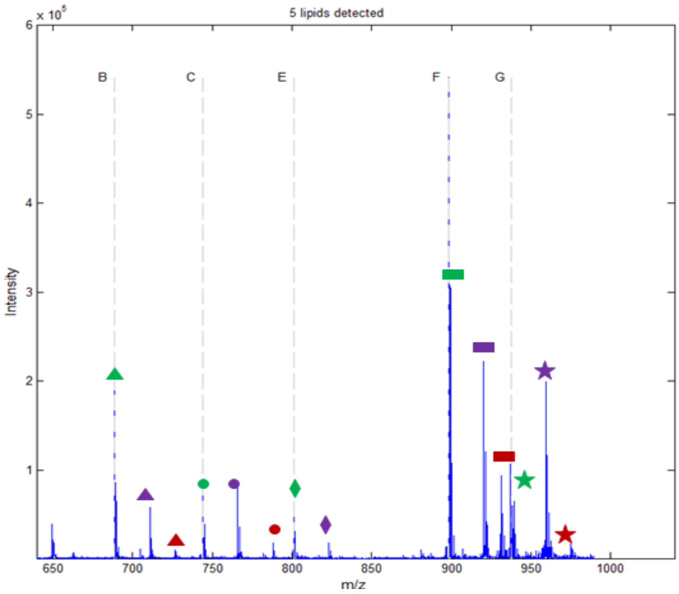
Referring back to Figure 3.1, a red square at the top of the 3-D scores plot clearly seen. This square represents the overall best experiment from the 720 experiments carried out and its corresponding spectrum is shown in Figure 3.2. This optimised experimental setup was further used to analyze lipid extracts from four different bacterial biological samples encompassing both Gram-positive and Gram-negative bacteria (*B. cereus*, *B. subtilis*, *E. coli*, *P. aeruginosa*) as well as human serum. The spectra for these five real-world samples are shown in Figure S3.6 (A-E) and it is clear that this solution yields high quality spectra.

By contrast, the experiments that failed to detect the five lipids are shown towards the bottom of the plot in Figure 3.1. In this Figure, a significant amount of experiments concentrated at the top of the 3D plot. The highlight of these experiments (circled in Figure 3.1) showed that ATT, DHB and THAP matrices were found in this region. Hence, these matrices are the most compatible with the lipids analysed. Experiments shown to be less suitable for the analysis of the lipid mixture also yielded very poor quality spectra when used to analyze lipid extracts from bacterial and serum as shown in Figure S3.6 (F-J).

3.3.3 MALDI-TOF-MS spectra of a lipid mixture

In this study, it was shown that using MALDI-TOF-MS, it is possible to analyse lipids in a mixture, but their detectability changed significantly (Müller *et al*., 2001) when changing the type of matrix, matrix additives and the concentration of some matrix additives. Comparing the matrices that were mentioned in the previous section, it can be noted that ATT was the most compatible for the analysis of the five lipids in a mixture since it was the softest, (Harvey, 2014) produced good shot-to-shot and sample-to-sample reproducibility (Lecchi *et al*., 1995), in turn causing a substantially lower amount of fragmentation, reducing the background noise and increasing both the signal-to-noise ratio and reproducibility, which produced an overall optimum experiment. These observations were also reported by Stubiger *et al*. (Stübiger *et al*., 2010). Moreover, it can be argued that other matrices, such as PNA and CMBT, could be used as alternatives in some cases as some of the lipids were detectable, and matrices such as DHAP were not used for further lipid analysis as poor spectra were produced regardless of whether matrix additives were used or not (Stübiger and Belgacem, 2007). However, reproducibility with PNA was low and hence it was decided that ATT was to be used in future lipid studies.

Figure 3.2 shows the positive ion MALDI-TOF-MS spectrum of the overall optimum experiment when using the combination: mix deposition method, ATT as a matrix, H_2O/ACN (50:50) matrix preparation vehicle. This combination was selected in the current experimental design guided by Pareto optimality (described in Section 3.3.5). Explaining such selection would require extensive theoretical speculation, which is beyond the scope of this work. The *m/z* range was from 650 to 1000, and the peaks that corresponded to each lipid in the mixture were assigned and summarised in the table shown within Figure 3.2. The green, purple and red symbols in the spectrum and table corresponded to H^+, Na^+ adduct and K^+ adduct of each lipid, respectively. These adducts are usually detected when conducting such experiments. The spectrum corresponds to the precursor ions $[M + H]^+$ (*m/z* 688), $[M + Na]^+$ (*m/z* 710) and $[M + K]^+$ (*m/z* 726) for lipid B, and the protonated peak represents the most intense peak for lipid B. Moreover, the peaks at *m/z* 744, 766 and 782 correspond to lipid C and also represent $[M + H]^+$, $[M + Na]^+$ and $[M + K]^+$, respectively, and this time the sodium adduct peak dominated. These observations were also noticed with lipids E, F and G, with the exception of $[M + K]^+$ peak not being detected for lipid E. The ability to identify protonated molecules simplifies the interpretation of spectra (Ham *et al*., 2005).



Lipid	$[M+ H]^+$	$[M+ Na]^+$	$[M+ K]^+$
B	688 ▲	710 ▲	726 ▲
C	744 ⬮	766 ⬮	782 ⬮
E	801 ◆	823 ◆	- ◆
F	898 ■	920 ■	936 ■
G	937 ★	959 ★	975 ★

Figure 3.2: Typical and best MALDI-TOF-MS spectrum of a lipid mixture, detected *m/z* from 650 to 1000. 14 lipid peaks are highlighted in the spectrum showing the main lipids and lipid-adducts that were detected using the top conditions, where lipid (B) is PG, (C) is PE, (E) is PG, (F) is PC and (G) is DGDG.

On the other hand, some poor complex MALDI-TOF-MS spectra were produced, which may be due to the immiscibility of the matrix solution and the lipid mixture, making crystallisation inhomogeneous, (Stübiger and Belgacem, 2007) or due to failure of the one or more the four different multiple objectives measured. Figure 3.3B shows an example of a poor spectrum produced for the lipid mixture when using the combination: underlay deposition method, PNA dissolved in chloroform, MeOH and H_2O with 80 mM lithium nitrate as a matrix additive (Figure 3.3B). Furthermore, the spectrum slightly improved and was less complex when the concentration of the additive was reduced to 40 Mm (Figure 3.3C). From this observation, it can be seen in Figure 3.3C that some of the lipids were detected using MALDI-TOF-MS, such as peaks at *m/z* 750, 904 and 943 which correspond to $[M + Li]^+$ for lipid C, F and G, respectively, with the exception of lipid B and E, which were not detected. As discussed above the positions of experiments within the 3-D scores plot are vital hence, the position of the experimental condition in Figure 3.3A was interesting. For example, when 40 mM lithium nitrate was used (Figure 3.3C) compared with the position of the same experiment carried out using 80 mM lithium nitrate (Figure 3.3B) instead, with 40 mM lithium nitrate being higher position than that 80 mM lithium nitrate. Surprisingly, when the matrix PNA was added to the lipid mixture alone without the matrix additive (Figure 3.3D), the spectrum was less complex as the protonated peak was easily detected.

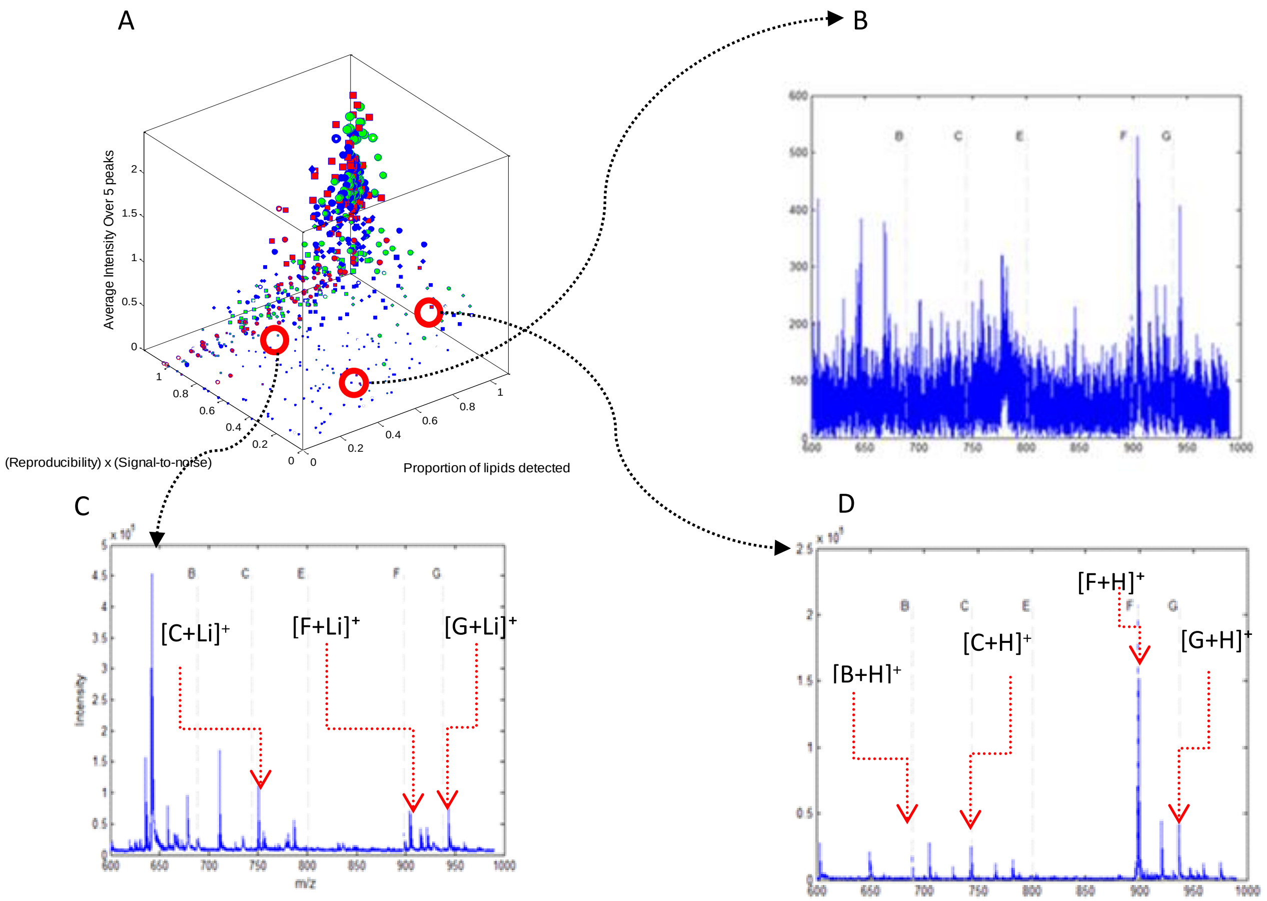


Figure 3.3: (A) 3D scores plot of MALDI-TOF-MS, (B) MALDI-TOF-MS spectrum for an experiment that failed to detect lipid peaks when using PNA as the matrix and 80 mM lithium nitrate as the matrix additive, (C) the experiment slightly improved when the concentration of the matrix additive was reduced to 40 mM and (D) PNA was used without matrix additive.

3.3.4 Additives to reduce the complexity of data

A number of research groups have used matrix additives for the analysis a of variety biological compounds. These additives include: ammonium acetate (Griffiths and Bunch, 2012; Stübiger *et al.*, 2010) and citrate (Zhu and Papayannopoulos, 2003) for analysing phosphopeptides and proteins; lithium and caesium chlorides (Wang *et al.*, 2000) for analysing polymers; and tetraamine spermine (Asara and Allison, 1999) and polyamine (Vandell and Limbach, 1999) for the analysis of oligonucleotides. Interest in using additives has increased due to the quality of MALDI-TOF-MS spectra produced upon their addition (Griffiths and Bunch, 2012; Zhou *et al.*, 2010). Hence, this study included the addition of matrix additives to reduce the complexity of MALDI-TOF-MS spectra generated with some of the matrices used in lipid analysis.

The matrix additives used in this study contained different cations including: Na^+, Li^+, K^+ and Ca^{+2}, which participated in the formation of adducts with the lipids. Lipid detection was affected significantly by the addition of some additives including: sodium nitrate, sodium acetate and diammonium citrate, more so than others such as EDTA ammonium (Müller *et al*., 2001). The best conditions for the experiments showed that the use of matrix additives is not always necessary as all the five lipid peaks were detected with the only difference in intensity of the adducts. Moreover, there appeared to be no obvious effect on the spectra when using different concentrations of some of the matrix additives especially when using concentrations between 10 to 40 mM (Figures S3.3).

On the other hand, these matrix additives were found to be useful in reducing the complexity of the data with some matrices, such as dithranol. Figure 3.4A and B show the spectra of the lipid mixture before and after the addition of sodium nitrite (10 mM), respectively. These spectra reiterated that the use of additives can indeed generate spectra with more useful information. Figure 3.4A shows that the spectrum generated from the analysis of the lipid mixture without the addition of an additive was extremely complex and the peaks were not detectable. However, Figure 3.4B shows that some of the lipid peaks were detectable such as sodiated peaks at 710, 766, 823, 920 and 959 *m/z* corresponding to lipids B, C, E, F and G, respectively. By contrast, these peaks were undetectable in Figure 3.4A when the additive was not added.

Three different anions were the centre points for lipid analysis when choosing the additives, these anions included: nitrates, acetates and chlorides. We have observed that the addition of sodium nitrate/acetate led to an increase in the abundance of Na^+ adducts in comparison to K^+ adducts and the protonated peaks were also detected (data not shown). Moreover, the addition of potassium nitrate led to an increase in the intensity of K^+ adducts relative to Na^+ adducts. In addition, when lithium nitrate was added to some of the matrix solutions, such as PNA, this led to reduction in spectral complexity and ease of the identification of some peaks. However, high concentrations of matrix additives resulted in poor spectra as no or few lipids were detected. Moreover, the addition of ammonium chloride to some of matrix solutions resulted in a decrease in abundance of Na$^+$ and K$^+$ adducts, in line with a previous study carried out by Griffiths *et al*. (Griffiths and Bunch, 2012). The use of calcium chloride resulted in poor spectral quality compared to other additives (Müller *et al*., 2001) as Ca^{+2} adducts cannot be detected.

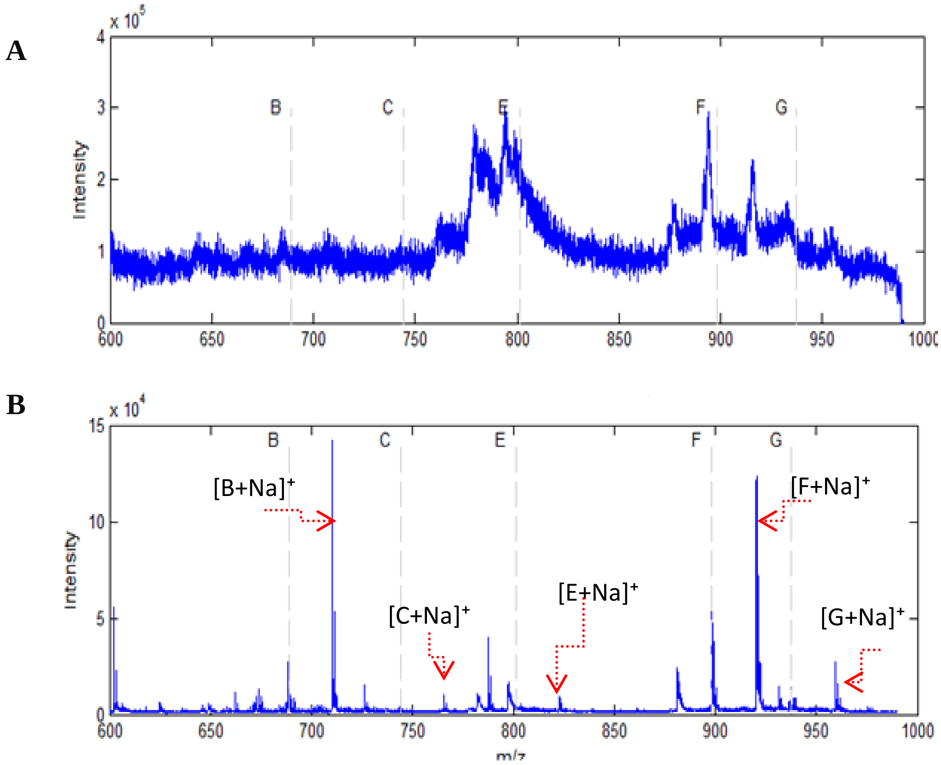


Figure 3.4: MALDI-TOF-MS spectra for the lipid mixture using dithranol as a matrix, (A) without a matrix additive, and (B) with 10 mM of sodium nitrate. The experiment resulted in an improved spectrum when sodium nitrite was added to the matrix solution.

Table 3.1A shows different experiments, which consisted of a combination of different factors. The green boxes represent the combination of factors that allow the detection of lipid peaks. In contrast, the red boxes represent the experiments that combined factors, which failed to detect lipid peaks. For example ATT, DHB and THAP are useful to analyse lipid samples with or without matrix additives, whereas, as some of the matrices such as CMPT and dithranol are not able to detect some of the lipid peaks without matrix additives. CMPT can detect some of the lipids in the presence of an additive such as 40 mM of sodium acetate, however; the spectra generated remain poor (Figure S3.4). Referring back to Table 3.1A it can be seen that some of the combinations enable the detection of lipids. However, in-depth analysis reveals that the intensities of each lipid vary between one experiment to another (Table S3.4) (see list in the attached material: Chapter 3_SI). This table shows the individual and combined FX ($\boldsymbol{f}_{\boldsymbol{multi-objective}}$) values for each objective for each experiment. Although, most of the experiments generated a signal, it can be noted that some of the objectives have higher scores than others as expected.

Our observations in Figure 3.5 below indicate that the choice of matrix and matrix additive are the best predictors of peak intensities from the five experiments parameters (factors) that were investigated in this study, whereas other factors, such as the concentration of the additive, the deposition method and occasionally the matrix preparation method, are less important in the detection the lipid peaks using MALDI-TOF-MS analysis.

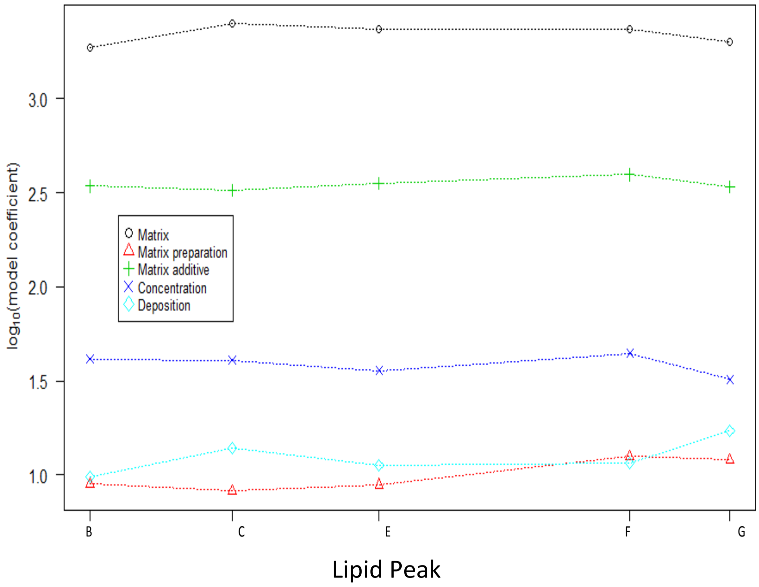


Figure 3.5: A factor impact model showing the impact of each factor considered in the experimental design. It indicates that the choice of matrix and matrix additive are the best predictors of peak intensity, whereas other factors are less important to optimise for the detection of these lipid species using MALDI-TOF-MS. Key: black line is the matrix, red line is the matrix preparation, green line is the matrix additive, blue line is the concentration of matrix additive, and cyan line is the deposition method.

3.3.5 Pareto Optimality

The aim of this method was to identify the optimal experimental settings based on at least one of the objectives which were used for the optimisation process. Table 3.1B shows the first overall optimum experiment that was identified using PO. In this table, the column (1) which is represented by a blue colour shows the FX score value which is the computed score for the multiple objectives. FX score value was generated by measuring the overall objective contributions. As with other methods, this method also needed to be validated.

The validation was carried out in different ways:

(i) Using the top experiment seen in the 3-D plot (Figure 3.1), which is represented by a red square; this was shown to be reproducible as this square represents an average of three experiments with the same overall optimum conditions which have shown good reproducibility and high signal to noise ratio (see Figure S3.5).

(ii) The lipid mixture without the additive produced the spectrum shown in (Figure 3.4A) which was in position 257 of 260 samples based on the Pareto optimality; on the other hand, this spectrum was improved upon the addition of the additive to the matrix and is shown in (Figure 3.4B), leading to a change in the position of the sample to position 177.

(iii) The first overall optimum experiments were repeated again (Table 3.1B) and column (2) which is represented by a yellow colour shows the FX score values for the repeated experiments carried out for validation. The results achieved had a better score than the maximum value, which indicates that these newly tested conditions can be used for subsequent experiments.

Table 3.1: (A) Examples of experiments carried out using different factors. The green boxes show the best combination of factors for lipid peak detections, whereas the red boxes shows the combination of factors that failed to detect lipid peaks. (B) The first overall optimum experiments that was selected using Pareto Optimality. (1) Represents by the blue column shows the FX score value which is computed score for the multiple objectives. (2) Represents by the yellow column shows the FX score value for the repeated experiment carried out for validation and is also computed score for the multiple objectives.

A Matrix	Matrix Prep.	Matrix Additive	Conc.	Matrix Dep.	B	C	E	F	G	FX value
ATT	H_2O/ACN	None	None	Mix	🟩	🟩	🟩	🟩	🟩	4.277
ATT	Chloroform/MeOH/H_2O	Sodium nitrate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.109
DHB	ACN/H_2O/isopropanol	Sodium nitrate	40 mM	Underlay	🟩	🟩	🟩	🟩	🟩	3.972
DHB	MeOH/H_2O	Sodium acetate	5 Mm	Underlay	🟩	🟩	🟩	🟩	🟩	3.852
DHB	EtOH/H_2O	Sodium nitrate	5 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.829
DHB	EtOH/H_2O	Sodium dodecyl sulphate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.787
DHB	ACN/H_2O/isopropanol	Sodium dodecyl sulphate	5 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.782
THAP	MeOH/H_2O	Sodium nitrate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.831
THAP	Chloroform/MeOH/H_2O	None	None	Mix	🟩	🟩	🟩	🟩	🟩	3.602
Dithranol	MeOH/H_2O	None	None	Overlay	🟥	🟥	🟥	🟥	🟥	1.029
Dithranol	MeOH/H_2O	Sodium nitrate	10 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.775
CMPT	H_2O/ACN	Sodium dodecyl sulphate	40 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.567
HABA	EtOH/H_2O	None	None	Mix	🟥	🟥	🟥	🟩	🟥	1.267
DHAP	H_2O/ACN	Potassium acetate	10 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.430

B Matrix	Matrix pre.	Matrix additive	Matrix additive conc.	Matrix Dep.	(1) FX value	(2) FX value
ATT	(H_2O+ACN)	None	None	Mix	4.27	4.95

3.4 Concluding remarks

In this study, we presented evidence for the feasibility of translating complex data generated from lipid analysis using MALDI-TOF-MS to more simplified spectra which yields useful information about the lipids being analysed. Reproducibility and robustness were achieved when using fractional factorial design and Pareto optimality combined with MALDI-TOF-MS analysis, which had the desired effect of significantly reducing the experimental search space. Indeed, the use of FFD showed that the choice of matrix, matrix preparation, choice of matrix additive, additive concentration and deposition method for MALDI-TOF-MS analysis could be optimised for lipid detection in a mixture. This resulted in the number of possible experimental conditions being reduced from 8064 to 720.

This study showed that for lipid analysis using MALDI-TOF-MS, the key factors to obtain quality spectra are the choice of matrix and matrix additives. For the analysis of the five target lipid species analysed the overall optimum conditions were achieved when using: mix deposition method, ATT as a matrix, H_2O/ACN (50:50, *v/v*) matrix preparation without the addition of a matrix additive. Hence, this would suggest that if the correct matrix is used for MALDI-TOF-MS analysis of lipids, a matrix additive is often not required. However, this should not be generalised as the matrix dithranol required the addition of an additive and gave acceptable results for lipid detection.

Although this study showed the utility of MALDI-TOF-MS in the analysis of lipid mixtures, applying this technique for the analysis of low molecular weight compounds suffers from several limitations including the observed interference of matrix peaks with low molecular weight analyte peaks, the presence of analyte isobaric peaks, as well as the complexity of spectra arising from unfractionated biological samples. These limitations can still be overcome when suitable technologies used to resolve analytes are applied in conjunction with mass spectrometry. These technologies include liquid chromatography (Pitt, 2009) and ion mobility (Lanucare *et.al*, 2014), with each having its own specific application.

In conclusion, we have shown that using FFD and PO it is possible to optimise the detection of lipids in an artificial mixture containing 5 lipid species and future analyses will concentrate on applying these conditions to real biological systems.

3.5 References

AlMasoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. 2014. Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the Characterisation of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta*, **840**, 49-57

Asara, J. and Allison, J. 1999. Enhanced detection of phosphopeptides in matrix-assisted laser desorption/ionisation mass spectrometry using ammonium salts. *Journal of the American Society for Mass Spectrometry,* **10**, 35-44

Batoy, S. M. A., Borgmann, S., Flick, K., Griffith, J., Jones, J. J., Saraswathi, V., Hasty, A. H., Kaiser, P. and Wilkins, C. L. 2009. Lipid and phospholipid profiling of biological samples using MALDI Fourier transform mass spectrometry. *Lipids,* **44**, 367-371

Calvano, C. D., Zambonin, C. G. and Palmisano, F. 2011. Lipid fingerprinting of Gram-positive lactobacilli by intact cells-matrix-assisted laser desorption/ionisation mass spectrometry using a proton sponge based matrix. *Rapid Communications in Mass Spectrometry,* **25,** 1757-1764

Choi, S.-S., Lee, H. M., Jang, S. and Shin, J. 2009. Comparison of ionisation behaviors of ring and linear carbohydrates in MALDI-TOFMS. *International Journal of Mass Spectrometry,* **279**, 53-58

Cliff, J. B., Kreuzer, H. W., Ehrhardt, C. J. and Wunschel, D. S. 2012. *Chemical and physical signatures for microbial forensics*, New York, Springer, pp.35-36

Cornett, D. S., Reyzer, M. L., Chaurand, P. and Caprioli, R. M. 2007. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods,* **4**, 828-833

Dong, W., Shen, Q., Baibado, J. T., Liang, Y., Wang, P., Huang, Y., Zhang, Z., Wang, Y. and Cheung, H.-Y. 2013. Phospholipid analyses by MALDI-TOF/TOF mass spectrometry using 1,5-diaminonaphthalene as matrix. *International Journal of Mass Spectrometry,* **343–344**, 15-22

Fuchs, B. and Schiller, J. 2009. Application of MALDI-TOF mass spectrometry in lipidomics. *European Journal of Lipid Science and Technology,* **111**, 83-98

Gidden, J., Denson, J., Liyanage, R., Ivey, D. M. and Lay, J. O., Jr. 2009. Lipid compositions in Escherichia coli and *Bacillus subtilis* during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry. *International Journal of Mass Spectrometry,* **283**, 178-184

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G. and Kell, D. B. 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology,* **22**, 245-252

Gorman, J. J., Ferguson, B. L. and Nguyen, T. B. 1996. Use of 2, 6-Dihydroxyacetophenone for analysis of fragile peptides, disulphide bonding and small proteins by matrix-assisted laser desorption/ionisation. *Rapid Communications in Mass Spectrometry,* **10**, 529-536

Griffiths, R. L. and Bunch, J. 2012. A survey of useful salt additives in matrix-assisted laser desorption/ionisation mass spectrometry and tandem mass spectrometry of lipids: introducing nitrates for improved analysis. *Rapid Communications in Mass Spectrometry,* **26**, 1557-1566

Griffiths, R. L., Sarsby, J., Guggenheim, E. J., Race, A. M., Steven, R. T., Fear, J., Lalor, P. F. and Bunch, J. 2013. Formal lithium fixation improves direct analysis of lipids in tissue by mass spectrometry. *Analytical Chemistry,* **85,** 7146-7153

Ham, B. M., Jacob, J. T. and Cole, R. B. 2005. MALDI-TOF MS of phosphorylated lipids in biological fluids using immobilized metal affinity chromatography and a solid ionic crystal matrix. *Analytical Chemistry,* **77**, 4439-4447

Harvey, D. J. 2014. Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionisation mass spectrometry: An update for 2009–2010. *Mass Spectrometry Reviews***, 34**, 268-422

Jackson, S. N., Wang, H.-Y. J. and Woods, A. S. 2005. In situ structural Characterisation of phosphatidylcholines in brain tissue using MALDI-MS/MS. *Journal of the American Society for Mass Spectrometry,* **16,** 2052-2056

Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., Mccowan, L., Roberts, C., Cooper, G. J., Kell, D. B. and Baker, P. N. 2010. Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. *Hypertension,* **56**, 741-749

Lay, J. O., Gidden, J., Liyanage, R., Emerson, B. and Durham, B. 2012. Rapid Characterisation of lipids by MALDI MS. Part 1: Bacterial taxonomy and analysis of food oils. *Lipid Technology,* **24**, 11-14

Lasch, P., Beyer, W., Nattermann, H., Staemmler, M., Siegbrecht, E., Grunow, R. and Naumann, D. 2009. Identification of *Bacillus* anthracis by using matrix-assisted laser desorption ionisation-time of flight mass spectrometry and artificial neural networks. *Applied and Environmental Microbiology,* **75**, 7229-7242

Le, C. H., Han, J. and Borchers, C. H. 2012. Dithranol as a MALDI matrix for tissue imaging of lipids by Fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry,* **84**, 8391-8398

Lecchi, P., Le, H. and Pannell, L. K. 1995. 6-Aza-2-thiothymine: a matrix for MALDI spectra of oligonucleotides. *Nucleic Acids Research,* **23**, 1276

Lee, G., Son, J. and Cha, S. 2013. Selective or class-wide mass fingerprinting of phosphatidylcholines and cerebrosides from lipid mixtures by MALDI mass spectrometry. *Bulletin of the Korean Chemical Society,* **34**, 2143-2147

Lee, G. K., Lee, H. S., Park, Y. S., Lee, J. H., Lee, S. C., Lee, J. H., Lee, S. J., Shanta, S. R., Park, H. M., Kim, H. R., Kim, I. H., Kim, Y. H., Zo, J. I., Kim, K. P. and Kim, H. K. 2012. Lipid MALDI profile classifies non-small cell lung cancers according to the histologic type. *Lung Cancer,* **76**, 197-203

Li, M., Yang, L., Bai, Y. and Liu, H. 2013. Analytical methods in lipidomics and their applications. *Analytical Chemistry,* **86**, 161-175

Marto, J. A., White, F. M., Seldomridge, S. and Marshall, A. G. 1995. Structural Characterisation of phospholipids by matrix-assisted laser desorption/ionisation Fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry,* **67**, 3979-3984

Mattila, I., Seppänen-Laakso, T., Suortti, T. and Orešič, M. 2008. Application of lipidomics and metabolomics to the study of adipose tissue, Humana Press

Monchamp, P., Andrade-Cetto, L. and Zhang, J. Y. 2007. *Systems Bioinformatics: An Engineering Case-Based Approach (chapter 4),* Artech House, INC.

Mukerjee, R. and Wu, C. F. J. 2006. *A Modern Theory of Factorial Design*, Springer, pp.9

Müller, M., Schiller, J., Petković, M., Oehrl, W., Heinze, R., Wetzker, R., Arnold, K. and Arnhold, J. 2001. Limits for the detection of (poly-)phosphoinositides by matrix-assisted laser desorption and ionisation time-of-flight mass spectrometry (MALDI-TOF MS). *Chemistry and Physics of Lipids,* **110**, 151-164

Nicolaou, N., Xu, Y. and Goodacre, R. 2012. Detection and quantification of bacterial spoilage in milk and pork meat using MALDI-TOF-MS and multivariate analysis. *Analytical Chemistry,* **84**, 5951-5958

Petkovic, M., Schiller, J., Muller, M., Suss, R., Arnold, K. and Arnhold, J. 2009. Detection of adducts with matrix clusters in the positive and negative ion mode MALDI-TOF mass spectra of phospholipids. *Zeitschrift für Naturforschung, B, A Journal of Chemical Sciences,* **64**, 331

Prescher, J. A. and Bertozzi, C. R. 2005. Chemistry in living systems. *Nature Chemical Biology,* **1**, 13-21

Przybylski, C., Gonnet, F., Bonnaffé, D., Hersant, Y., Lortat-Jacob, H. and Daniel, R. 2010. HABA-based ionic liquid matrices for UV-MALDI-MS analysis of heparin and heparan sulphate oligosaccharides. *Glycobiology,* **20**, 224-234

Quinn, G. P. and Keough, M. J. 2002. *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, pp.257

Ramı, AMP, X, K, J. and VLACH, M. 2002. Pareto-optimality of compromise decisions. *Fuzzy Sets and Systems,* **129**, 119-127

Ryan, C., Clayton, E., Griffin, W., Sie, S. and Cousens, D. 1988. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms,* **34**, 396-402

Ryzhov, V., Hathout, Y. and Fenselau, C. 2000. Rapid characterisation of spores of *bacillus cereus* group bacteria by matrix-assisted laser desorption-ionisation time-of-flight mass spectrometry. *Applied and Environmental Microbiology,* **66**, 3828-3834

Schiller, J., Arnhold, J., Benard, S., Müller, M., Reichl, S. and Arnold, K. 1999. Lipid analysis by matrix-assisted laser desorption and ionisation mass spectrometry: a methodological approach. *Analytical Biochemistry,* **267**, 46-56

Schiller, J., Süß, R., Arnhold, J., Fuchs, B., Leßig, J., Müller, M., Petković, M., Spalteholz, H., Zschörnig, O. and Arnold, K. 2004. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research. *Progress in Lipid Research,* **43**, 449-488

Shanta, S. R., Kim, T. Y., Hong, J. H., Lee, J. H., Shin, C. Y., Kim, K.-H., Kim, Y. H., Kim, S. K. and Kim, K. P. 2012. A new combination MALDI matrix for small molecule analysis: application to imaging mass spectrometry for drugs and metabolites. *Analyst,* **137**, 5757-5762

Shu, X., Liang, M., Yang, B., Li, Y., Liu, C., Wang, Y. and Shu, J. 2012. Lipid fingerprinting of *Bacillus* spp. using online MALDI-TOF mass spectrometry, *Analytical Methods,* **4**, 3111-3117

Smilde, A. K., Knevelman, A. and Coenegracht, P. M. J. 1986. Introduction of multi-criteria decision making in optimisation procedures for high-performance liquid chromatographic separations. *Journal of Chromatography A,* **369**, 1-10

Stübiger, G. and Belgacem, O. 2007. Analysis of lipids using 2, 4, 6-trihydroxyacetophenone as a matrix for MALDI mass spectrometry. *Analytical Chemistry,* **79**, 3206-3213

StüBiger, G., Belgacem, O., Rehulka, P., Bicker, W., Binder, B. R. and Bochkov, V. 2010. Analysis of oxidized phospholipids by MALDI mass spectrometry using 6-aza-2-thiothymine together with matrix additives and disposable target surfaces. *Analytical Chemistry,* **82**, 5502-5510

Vance, J. E. and Vance, D. E. 2008. *Biochemistry of lipids, lipoproteins and membranes,* Hungary, Elsevier, pp. 1-39

Vandell, V. E. and Limbach, P. A. 1999. Polyamine co-matrices for matrix-assisted laser desorption/ionisation mass spectrometry of oligonucleotides. *Rapid Communications in Mass Spectrometry,* **13**, 2014-2021

Wang, Y., Rashidzadeh, H. and Guo, B. 2000. Structural effects on polyether cationisation by alkali metal ions in matrix-assisted laser desorption/ionisation. *Journal of the American Society for Mass Spectrometry,* **11**, 639-643

Zemski Berry, K. A., Hankin, J. A., Barkley, R. M., Spraggins, J. M., Caprioli, R. M. and Murphy, R. C. 2011. MALDI imaging of lipid biochemistry in tissues by mass spectrometry, *Chemical Reviews,* **111**, 6491-6512

Zhou, P., Altman, E., Perry, M. B. and Li, J. 2010. Study of matrix additives for sensitive analysis of lipid A by matrix-assisted laser desorption ionisation mass spectrometry. *Applied and Environmental Microbiology,* **76**, 3437-3443

Zhu, X. and Papayannopoulos, I. A. 2003. Improvement in the detection of low concentration protein digests on a MALDI TOF/TOF workstation by reducing α-cyano-4-hydroxycinnamic acid adduct ions. *Journal of Biomolecular techniques: Journal of Biomolecular Techniques,* **14**, 298

Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. 2009, *Clin Biochem Rev*. **30**, 19–34

Lanucara, F., Holman, S. W., Gray, C. J. and Eyers, C. E. 2014. The power of ion mobility-mass spectrometry for structural characterisation and the study of conformational dynamics. *Nat. Chem,* **6**, 281-294

3.6 Supplementary information

3.6.1 Preparation of lipid extracts from bacterial samples

Nutrient agar (NA) plate cultures were prepared for four bacterial species (*Bacillus. cereus, Bacillus. subtilis, Escherichia coli* MG1655*, and Pseudomonas aeruginosa*). NA was prepared from a preparatory mixture (beef extract 3 g/L, peptone 5 g/L, NaCl 8 g/L and agar 2 at 12 g/L) (Lab-M, Bury, UK) following the manufacturer's instructions (28 g in 1 L of deionised water) and the agar was autoclaved (at 121ºC and 15 psi for 15 min) before use. A single colony from the agar culture was transferred to the LB media (50 mL) in a 250 mL flask which was incubated overnight at 37ºC with shaking at 200 rpm. LB media was prepared by mixing 10 g of NaCl, 5 g of yeast extract (Amersham Life Sciences, Cleveland, USA) and 10 g of tryptone (Formedia, Hunstanton, UK) dissolved in 1 L of distilled water. The broth was then autoclaved (at 121ºC and 15 psi for 45 min) before use.

15 mL from each LB media culture was quenched using 30 mL 60% methanol (-48ºC) and mixed quickly. The mixture was then centrifuge for 10 min at 4800 ×*g* at -8ºC [1]. This was followed by removing the supernatant rapidly and leaving the pellet in the centrifuge tube.

Biomass of the bacteria was mixed with 2 mL HPLC grade chloroform:methanol (2:1) pre-chilled at -20ºC and 1 mL of HPLC water. The mixture was then centrifuge at 4800 ×*g* for 3 min at -8ºC (Winder *et.al*. 2008). This was followed by tranfering the bottom chloroform-based layer which contains most of the lipids into fresh 2 mL micro-centrifuge tubes, then the samples were left to evaporate on a hot plate overnight at 40ºC before storage at -80ºC. For MALDI-TOF-MS analysis of the bacterial lipid extracts, the samples were reconstituted in 100µL of 80:20 (*v/v*) methanol:HPLC water.

3.6.2 Lipid extraction from human serum

The lipids from the artificial human serum were extracted by adding 400 µL of cold MeOH to 100 µL of serum. The mixture was homogenised using a vortex for 10 seconds. This followed by centrifugation for 15 min at 13500 *g*. The supernatant was then transferred to an Eppendorf tube and dried overnight in vacuum concentrator.

The extracted lipid pellet from artificial human serum was reconstituted in 100µL of HPLC grade water on the day of analysis where MALDI protocol was followed for sample preparation.

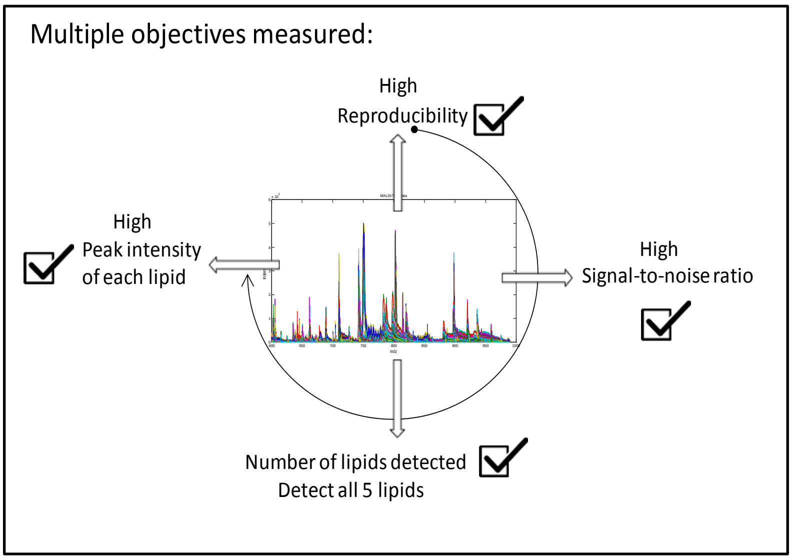


Figure S 3.1: Maximising peak intensity is not the only objective. This figure depicts the four objectives that were simultaneously measured and optimised in this study.

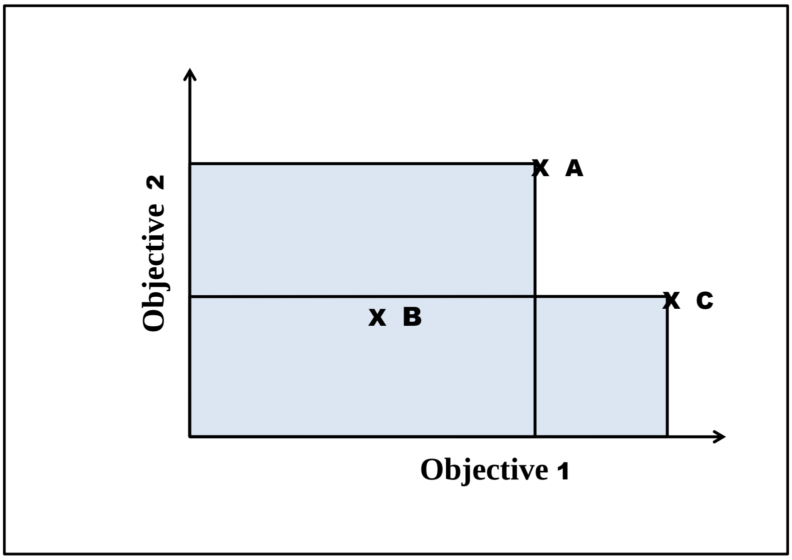
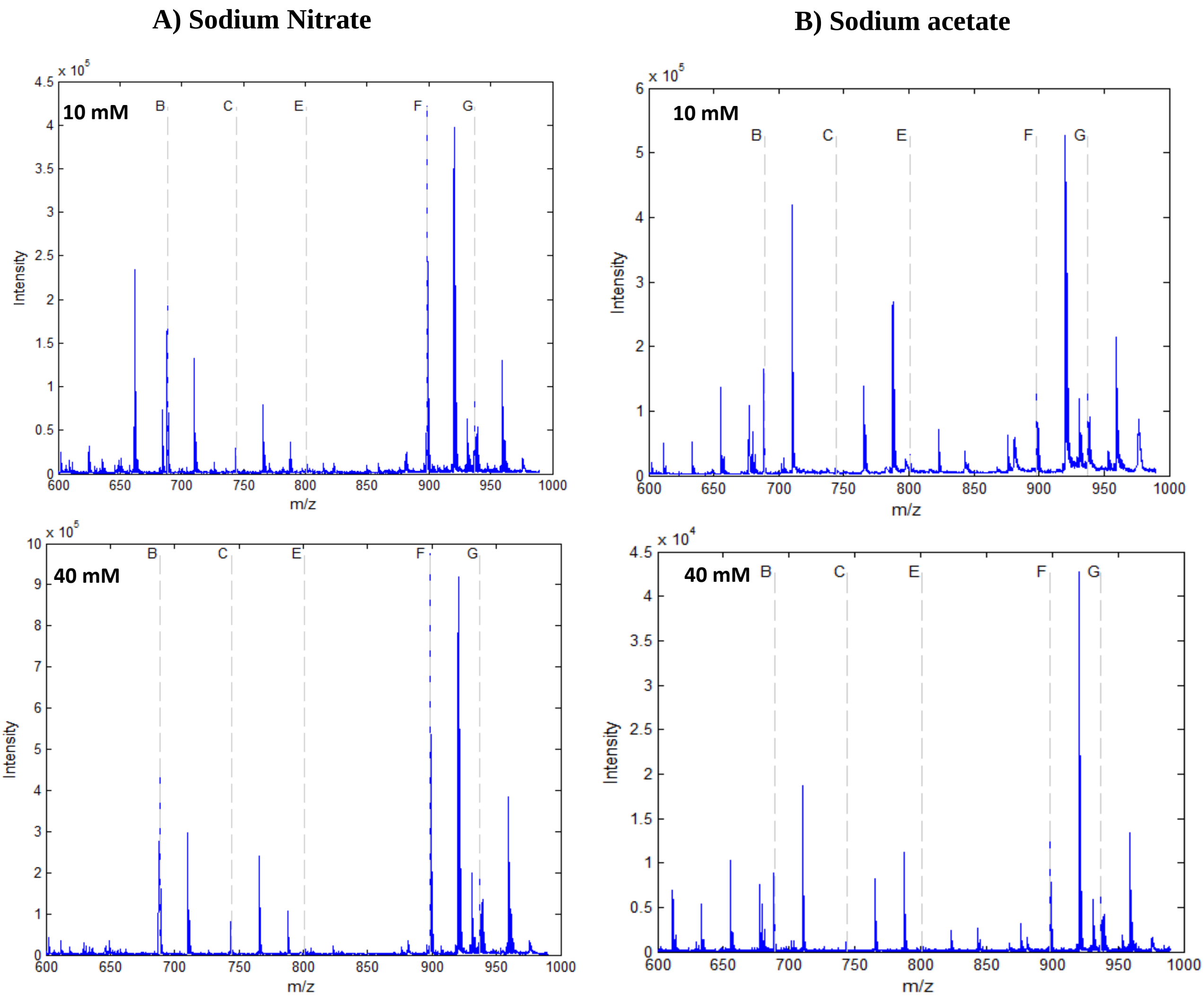


Figure S 3.2: Pareto optimality example for two objective functions. Solution A dominates solution B in all objectives. Solution C also dominates solution B in all objectives. Therefore, solution B is not in the Pareto front and is ignored. However, none of the solutions A or C dominates the others in all objectives and both are considered to be in the Pareto front, the choice of which solution to choose depends on the user requirements for each objective measured.



A) Sodium Nitrate

B) Sodium acetate

Figure S 3.3: Positive ion mode spectra for lipids in mixture detected using MALDI-TOF-MS in the presence of matrix additives: (A) sodium nitrate and (B) sodium acetate at two different concentrations. ATT was used as a matrix.

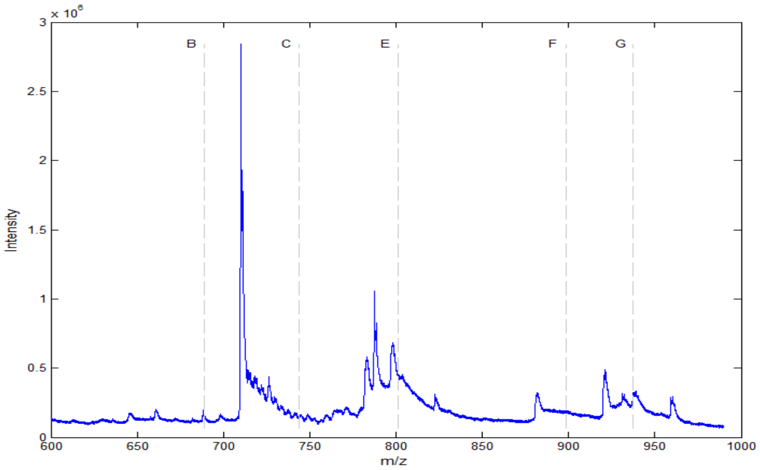


Figure S 3.4: Positive ion mode spectra for lipids in mixture detected via MALDI-TOF-MS using CMPT matrix and 40 mM of sodium acetate.

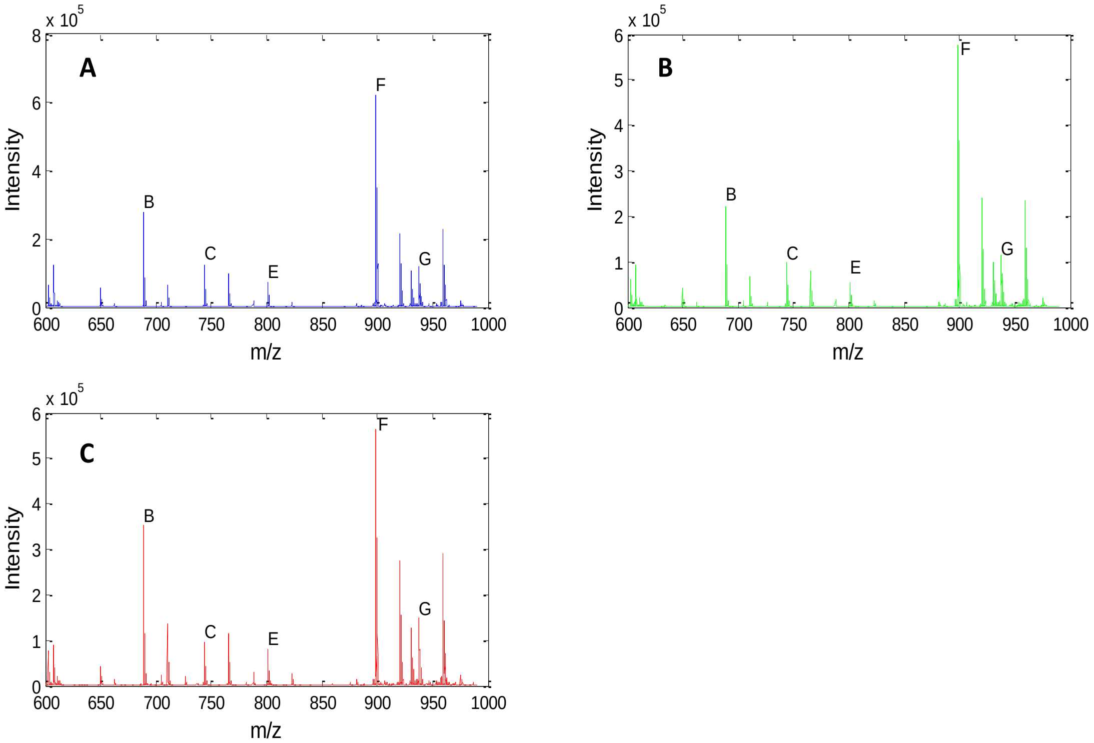


Figure S 3.5: Typical MALDI-TOF-MS spectra for a lipid mixture using the overall optimum conditions. Reproducibility checks were carried out to prove that the experiments identified using Pareto optimality is repeatable for 3 analytical replicates (A, B and C).

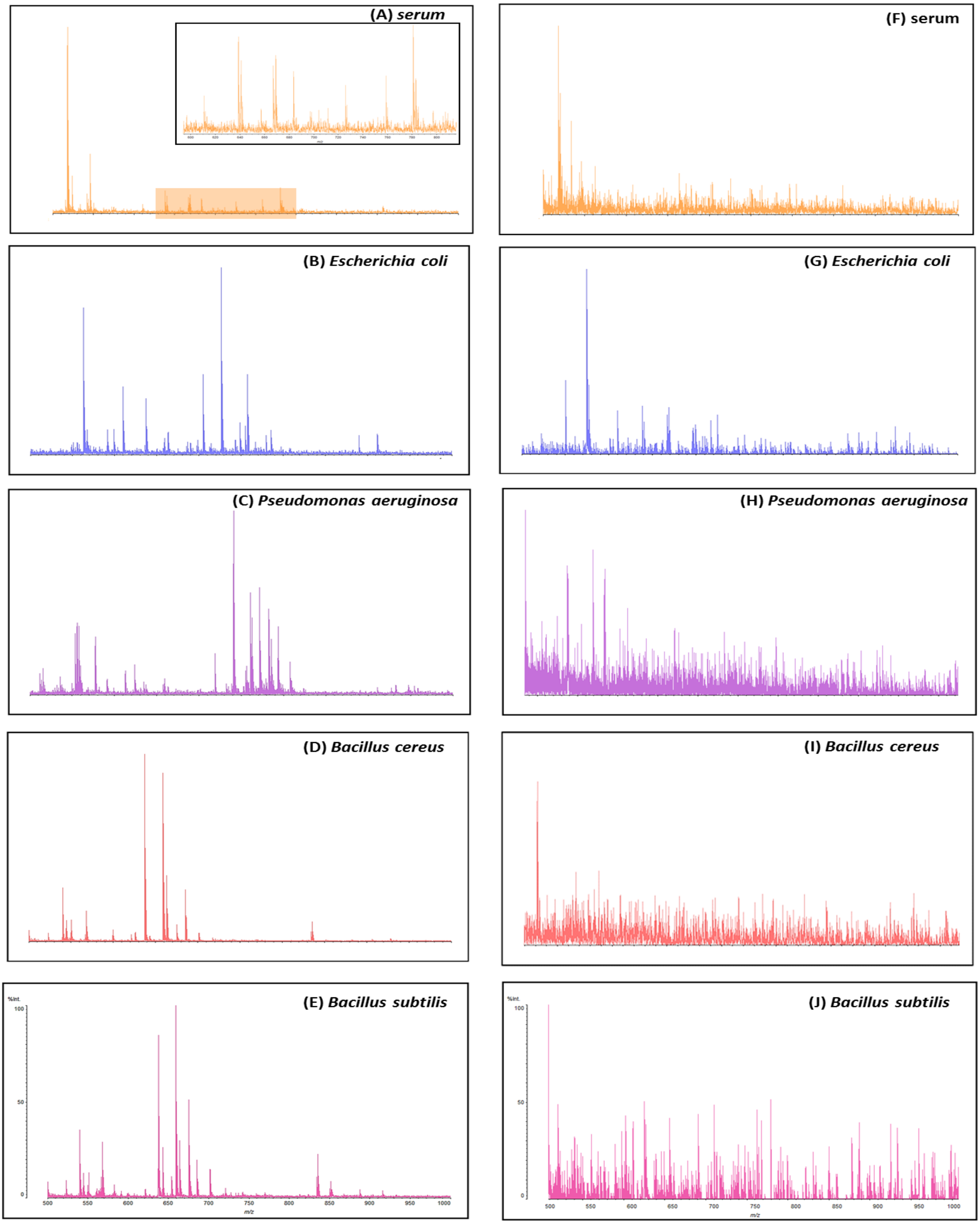


Figure S 3.6: Typical MALDI-TOF-MS spectra for lipid extracts from examples for bacterial samples using the most optimal set of conditions (A-E) and one of the least optimal set of conditions (F-J). These two conditions are in Figure 3.1 are indicated by purple arrow.

Table S3.1. The concentrations of matrices in the different matrix/matrix additive mixtures used for MALDI-TOF-MS analysis.

Matrix	MW	Initial concentration of the matrix in mM (10 mg/mL)	Concentration of matrix with*				
			5 mM additive	10 mM additive	20 mM additive	40 mM additive	80 mM additive
2',4',6'-Trihydroxyacetophenone (THAP)	168.15	59.50	58.01	56.53	53.55	47.60	35.70
2,5-Dihydroxybenzoic acid (2,5 DHB)	154.12	64.90	63.28	61.66	58.41	51.92	38.94
2′,6′-Dihydroxyacetophenone	152.15	65.70	64.06	62.42	59.13	52.56	39.42
5-chloro-2-mercaptobenzothiazole (CMBT)	201.7	49.60	48.36	47.12	44.64	39.68	29.76
6-aza-2-thiothymine (ATT)	143.17	69.80	68.06	66.31	62.82	55.84	41.88
2-(4'-Hydroxybenzeneazo)benzoic acid (HABA)	242.23	41.30	40.27	39.24	37.17	33.04	24.78
Dithranol	226.22	44.20	43.10	41.99	39.78	35.36	26.52
p-nitroaniline (PNA)	138.12	72.40	70.59	68.78	65.16	57.92	43.44

***** The concentrations of matrix are calculated using (*mv* × *mc* / 1000);

where: *mv* is the volume of the matrix added (see column 3 of Table S-2) and
mc is the initial concentration of the matrix (see column 3 of Table S-1)

Table S3.2. Volumes used to prepare the matrix/matrix additive mixtures.

Concentration of matrix additives (mM)	Additive (µL)	Matrix (µL)
5	25	975
10	50	950
20	100	900
40	200	800
80	400	600

Table S 3.3: The 720 performed experiments that were identified using fractional factorial design (enclosed sheet: Chapter 3_SI).

Table S 3.4: Mass peak assignments for the lipid mixture detected using MALDI-TOF-MS

No.	Lipid	Measured *m/z*	Assignment
1	Lipid B	688.98	[lipid B +H]$^+$
		710.93	[lipid B +Na]$^+$
		726.90	[lipid B +K]$^+$
		*	Li adduct
		*	Ca adduct
2	Lipid C	744.59	[lipid C +H]$^+$
		766.11	[lipid C +Na]$^+$
		782.09	[lipid C +K]$^+$
		750.69	[lipid C +Li]$^+$
		*	Ca adduct
3	Lipid E	801.17	[lipid E +H]$^+$
		823.12	[lipid E +Na]$^+$
		839.19	[lipid E +K]$^+$
		*	Li adduct
		*	Ca adduct
4	Lipid F	898.41	[lipid F +H]$^+$
		920.37	[lipid F +Na]$^+$
		936.39	[lipid F +K]$^+$
		904.38	[lipid F +Li]$^+$
		*	Ca adduct
5	Lipid G	937.96	[lipid G +H]$^+$
		959.27	[lipid G +Na]$^+$
		975.37	[lipid G +K]$^+$
		943.50	[lipid G +Li]$^+$
		*	Ca adduct

(*) Indicates peaks that were not detect in these conditions. These peaks were not included in the data analysis

Table S 3.4: Illustrations of the values for each individual objective that was computed for the Pareto optimisation of experiments carried out using different factors (first 5 columns). Each objective is normalised to be a number between 0 and 1. Therefore, the number of lipids identified is also normalised to be between 0 and 1 (*e.g.*, for the objective "number of lipids identified" if 5 lipids have been identified the score is 5/5 = 1) (enclosed sheet: Chapter 3_SI).

Chapter Four

Classification of *Bacillus* and *Brevibacillus* species using rapid analysis of lipids by mass spectrometry

Najla AlMasoud[a], Yun Xu[a], Drupad K Trivedi[a], Simona Salivo[b], Tom Abban[b], Nicholas J W Rattray[a], Ewa Szula[a], Haitham AlRabiah[a,c], Ali Sayqal[a] and Royston Goodacre[a]*

[a] School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

[b] Shimadzu, Kratos Analytical Ltd. Wharfside, Trafford Wharf Road, Manchester, M17 1GP, UK.

[c] Present address: Department of Pharmaceutical Chemistry, College of Pharmacy, King Saud University, P.O. Box 2457, Riyadh 11451, Saudi Arabia.

[*] Correspondence to Royston Goodacre: roy.goodacre@manchester.ac.uk

This chapter is a manuscript of an article submitted to *Analytical and Bioanalytical Chemistry*.

Drupad Trivedi and Nicholas Rattray participated in operating the HPLC-MS and deconvoluting the data. Yun Xu contributed to data analysis. Simona Salivo and Tom Abban participated in operating MALDI-TOF-TOF-MS 7090 mass spectrometer (Shimadzu Kratos, Manchester, UK). Haitham Alrabiah, Ewa Szula and Ali Sayqal provided technical assistance. Royston Goodacre contributed to this work through supervision and guidance of the study.

Abstract

Bacillus are aerobic spore-forming bacteria that are known to lead to specific diseases, such as anthrax and food poisoning. This study focuses on the characterisation of these bacteria by the detection of lipids extracted from 33 well-characterised strains from the *Bacillus* and *Brevibacillus* genera, with the aim to discriminate between the different species. For the purpose of analysing the lipids extracted from these bacterial samples, two rapid physicochemical techniques were used: matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) and liquid chromatography in conjunction with mass spectrometry (LC-MS). The average correct classification rate (CCR) for the 7 species of bacteria was 62.23% and 77.03% based on MALDI-TOF-MS and LC-MS data, respectively. The Procrustes distance for the two datasets was 0.0699 indicating that the results from the two techniques were very similar. In addition, we also compared these bacterial lipid MALDI-TOF-MS profiles to protein profiles also collected by MALDI-TOF-MS on the same bacteria (Procrustes distance 0.1006). The level of discrimination between lipids and proteins was equivalent and this further indicates the potential of MALDI-TOF-MS analysis as a rapid, robust and reliable method for the classification of bacteria based on different bacterial chemical components.

4.1 Introduction

Classification of bacteria has recently received increasing attention, most likely arising from public health concerns, environmental monitoring, food safety monitoring, taxonomic identification and differentiation of pathogenic species from non-pathogenic species, as well for the identification of biological threat agents (Priest and Austin, 1993; Irudayaraj *et al*., 2002; Sauer *et al*., 2008; Lopez-Diez and Goodacre, 2004). Bacteria can be classified using various physicochemical approaches based on different methods that rely either on analysis of (i) whole bacterial cells (Lasch *et al*., 2009; Claydon *et al*., 1996, Wilkins and Lay, 2005; Gaia *et al*., 2011; AlMasoud *et al*., 2014) or (ii) extracts of different compounds including (as in the current study) lipids (Allwood *et al*., 2014; Gidden *et al*., 2009; Shu *et al*., 2012b); each of these methods has its advantages and disadvantages.

Lipids are important components in bacterial cell membranes as they form lipid bilayers responsible for cell integrity (Vance and Vance, 2008; Zhang *et al*., 2011). These cell components have various structures and several factors can affect lipid synthesis such as culture media, temperature and physical dynamics during cell growth (Cliff *et al*., 2012). Complex lipids, just like proteins, can be used to identify and characterise bacteria (Calvano *et al*., 2011; Fahy *et al*., 2011). Interest in the analysis of lipid profiles from bacterial cells for taxonomic identification is increasing (Gidden *et al*., 2009). Not only do lipids play a structural role in the integrity of cell membranes but they also contribute to other cellular processes such as metabolic and signalling pathways (Wymann and Schneiter, 2008; Van Meer *et al*., 2008).

Early studies that aimed to resolve lipid species traditionally used different chromatographic techniques such as thin layer chromatography (TLC) (Wenk, 2005). This approach has disadvantages such as limited resolution and sensitivity which negatively affect many lipidomic applications (Wenk, 2005). Therefore, an armoury of techniques has been used to address many of these issues, which has led to the use of mass spectrometry technology, including direct infusion mass spectrometry (DIMS) (Goodacre *et al*., 2002) and liquid chromatography-mass spectrometry (LC-MS) (Wedge *et al*., 2011), which have been extensively used to analyse lipid samples enabling the detection of different types of lipids. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) has also been used for lipidomic analysis to overcome the limitations seen with other traditional methods and to analyse samples containing complex mixtures of lipids enabling classification and identification of bacteria (Lasch *et al.*, 2009; Gidden *et al.*, 2009; Bernardo *et al.*, 2002). The main advantages of MALDI-TOF-MS include: (i) it uses soft ionisation causing minimal analyte degradation, (ii) offers the possibility to analyse a range of complex molecules in complex mixtures such as bacterial samples, (iii) requires minimal sample preparation, and (iv) yields mass spectra that contain specific chemical features and fingerprints that can be used to identify and characterise bacterial species (Fenselau and Demirev, 2001; Lay, 2001).

The aim of this study was to classify 33 strains of bacteria belonging to 7 species – namely *B. amyloliquefaciens, B. cereus, Br. laterosporus, B. licheniformis, B. megaterium, B. sphaericus* and *B. subtilis* – based on MALDI-TOF-MS of extracted bacterial lipids. The results of which were evaluated and validated using LC-MS to confirm the bacterial classification based on MALDI-TOF-MS analysis.

4.2 Materials and Methods

4.2.1 Chemicals and solvents

Chemicals used were of a high purity grade and included the following: HPLC grade chloroform (Sigma-Aldrich, Dorset, UK), ethanol (Sigma-Aldrich), HPLC water (Sigma-Aldrich), HPLC grade methanol (Fisher Scientific Ltd., Loughborough, UK) and 99.99% pure formic acid (VWR International, East Grinstead, UK). Two different matrices were used in this study: 6-Aza-2-thiothymine (ATT) and 2,5-dihydroxybenzoic acid (DHB) (both from Sigma-Aldrich).

4.2.2 Microorganisms

Seven bacterial species (i.e., *B. amyloliquefaciens, B. cereus, Br. laterosporus, B. licheniformis, B. megaterium, B. sphaericus* and *B. subtilis*) were used in this study. Table 4.1 gives details of these 33 strains from these *Bacillus* and *Brevibacillus* genera and these were used previous from MALDI-TOF-MS on bacterial proteins (AlMasoud *et al*., 2014 (Chapter 2)).

4.2.3 Bacterial cultivation

Using sterile plastic loops bacterial strains were cultivated three times for 24 h at 37ºC on nutrient agar (NA) to generate axenic colonies and to maintain a stable phenotype. NA contained beef extract 3 g/L, peptone 5 g/L, NaCl 8 g/L and agar no. 2 at 12 g/L from Lab-M (Bury, UK) and was prepared following the manufacturer's instructions (28 g in 1 L of deionised water) and subsequently autoclaved (121ºC and 15 psi for 15 min) before Petri dishes were prepared.

Table 4.1: The 33 *Bacillus* and *Brevibacillus* species and strains were used in this study

Sample no.	**Species**	**Strain no.**	**Key colour in Figures**
1	*B. amyloliquefaciens*	B0177^T	Red
2		B0168	
3		B0175	
4		B0251	
5		B0620	
6	*B. cereus*	B0002^T	Green
7		B0550	
8		B0702*	
9		B0712	
10		B0851	
11	*B. licheniformis*	B0252^T	Blue
12		B0242	
13		B0755	
14		B1081	
15		B1379	
16	*B. megaterium*	B0056	Cyan
17		B0057	
18		B0076	
19		B0621	
20	*B. sphaericus*	7134^T	Pink
21		B0408	
22		B0219	
23		B0769	
24		B1147	
25	*B. subtilis*	B0014^T	Yellow
26		B0044	
27		B0098	
28		B0099*	
29		B0410	
30		B0501	
31		B1382	
32	*Br. laterosporus*	B0043*	Black
33		B0262	

^{T} indicates the type strain; * indicates strains used for preliminary optimisation experiments for time points

4.2.4 Optimisation of collection time points of collection for three different species of *Bacillus* and *Brevibacillus* for LC-MS

Three different species were used at the beginning of this work to choose the optimal collection time points; these species were *B. cereus* B0702, *B. subtilis* B0099 and *Br. laterosporus* B0043. An axenic colony was collected from each culture and inoculated in 600 mL of nutrient broth, (prepared according to the manufacturer's instructions (Oxoid Ltd., Basingstoke, UK) in 2 L flasks and then incubated for 24 h at 37ºC at 200 rpm. Optical density measurements (OD) at 600 nm were collected at 6 different time points (4, 6, 8, 10, 14 and 18 h) using a Biomate 5 spectrophotometer (Thermo, Hemel Hempstead, UK). For each species, three biological replicates were prepared in the same way.

4.2.4.1 Quenching

Samples were collected at the six different time points (4, 6, 8, 10, 14 and 18 h). From each culture, 15 mL was quenched using 30 mL of 60% cold methanol (-48ºC, chilled on dry ice) and mixed rapidly. This was followed by centrifugation of the quenched culture for 10 min at 4800 ×*g* at -8ºC (Winder *et al*., 2008). The supernatant was removed quickly then the rest was centrifuged again for 2 min and the remaining supernatant removed, leaving the pellet containing the bacterial cells in the centrifuge tube. The pellets were stored at -80ºC until lipid extraction was performed (AlRabiah *et al*. 2014). Figure S4.1A illustrates this process.

4.2.4.2 Lipid extraction

Bacterial pellets were mixed with 2 mL HPLC grade chloroform:methanol (2:1) pre-chilled at -20ºC (Winder *et al*., 2008). The samples were mixed using a laboratory shaker for 15 min, and 1 mL of cold HPLC water was then added to the mixtures. This was followed by centrifugation at 4800 ×*g* for 3 min at -8ºC. A biphasic system was generated, with the bottom chloroform based layer containing most of the lipids. The lipid layers were transferred to fresh 2 mL micro-centrifuge tubes (Allwood *et al*., 2014). The samples were left to evaporate on a hot plate at 40ºC to complete dryness prior to storage at -80ºC (Figure S4.1B). These samples were reconstituted in 80:20 methanol:water (*v/v*) at 100 µL per 0.1 OD_{600} and then analysed using LC-MS.

4.2.5 Collections of *Bacillus* and *Brevibacillus* strains for LC-MS and MALDI-TOF-MS analysis

In total, 33 strains were collected for LC-MS and MALDI-TOF-MS after 10 h of culturing at 37ºC and 200 rpm. Five biological replicates were collected for each strain.

A- Preparing extracted samples for MALDI-TOF-MS

For MALDI analysis of the extracted lipids, the samples were reconstituted in 80:20 methanol:HPLC water (*v/v*) (Tables S4.1-S4.5). 10 mg of DHB was dissolved in 900 µL ethanol and 100 µL sterile deionised water, and 10 mg of ATT was dissolved in 500 µL acetonitrile and 500 µL of sterile deionised water. 10 µL of the extracted lipid samples was mixed with 10 µL of either matrix and then 2 µL of the matrix/samples mixture was applied to a MALDI stainless steel plate and allowed to dry at room temperature (*ca*. 22 °C).

Samples were analysed in batches using an AXIMA-Confidence mass spectrometer (Shimadzu Biotech, Manchester, UK) equipped with a nitrogen pulsed UV laser (wavelength 337 nm) (AlMasoud *et al*., 2014 (Chapter 2)) set at 100 mV; each profile was produced using 20 laser shots, and 78 profiles were collected using a circular raster pattern. The instrument was operated in positive ionisation mode using the reflectron TOF over the mass-to-charge ratio (*m*/*z*) range 100-1600. Each biological sample was analysed in 4 technical replicates. A single biological replicate of each of the 33 bacterial strains was analysed each day. Before sample analysis, the MALDI instrument was calibrated using polyethylene glycol using the following *m/z* values: 613.7, 657.75, 710.80, 746.86, 789.91, 833.96, 878.02, 922.07, 966.12, 1010.18, 1054.23, 1098.28, 1142.34, 1186.39, 1230.44, 1274.50, 1318.55 and 1362.60.

Sample preparation of MALDI TOF-TOF

Sample preparation was carried out as follows. Samples were reconstituted in 1:1 chloroform/methanol (*v/v*). DHB was used as matrix, and was prepared in methanol (10 mg/mL) containing 10 mM NaCl. A sample droplet (0.35 µL) was placed onto a MALDI target spot, followed by an equal amount of matrix solution.

MALDI TOF-TOF analysis

The samples were analysed on a MALDI 7090 mass spectrometer (Shimadzu Kratos, Manchester, UK) with a solid state UV-laser (355 nm) operating at a 2 kHz acquisition repetition rate. The instrument was operated at an acceleration voltage of 20 keV, and a pulsed extraction function to improve mass resolution was carefully applied. The low mass rejection and the focus mass were set to 300 and 800 Da, respectively. The instrument was operated in the reflectron mode. To enhance the signal-to-noise ratio, 100 single shots were averaged for each mass spectrum. The laser intensity was adjusted for each experiment to obtain the best signal-to-noise ratio and to maximize the number and intensity of structural fragments. Positive mode spectra of all analytes were recorded. Helium gas was used for high-energy CID (20 keV) MS/MS experiments. All mass spectrometric data were acquired and analyzed by using the MALDI Solution software (Shimadzu Kratos, Manchester, UK).

B- LC-MS analysis

An Accela UHPLC system (Thermo-Fisher Ltd., Hemel Hempsted, UK) coupled to an electrospray LTQ-Orbitrap XL hybrid mass spectrometry system (Thermo-Fisher, Bremen, Germany) was used to analyse the samples. Samples were reconstituted in 80:20 methanol:HPLC water based on 100 µL per OD_{600} of 0.1 (Tables S4.1-S4.5). The mixture was vortexed and centrifuged at 11500 *g* for 30 s. Quality control (QC) samples were prepared by mixing an equal volume of each extracted sample and vortexing the mixture thoroughly. The mixtures were then transferred to 100 µL analytical vials (Dunn *et al*., 2011). All samples were run in positive ESI mode since LC-MS was used to confirm the results obtained from MALDI-TOF-MS which was also operated in the positive ionisation mode.

First, three biological replicates were analysed over 5 days and the remaining two biological replicates were analysed over a further 3 day period to account for the large number of samples. Briefly, 10 µL of extracted sample was injected onto a Hypersil GOLD UHPLC C_{18} analytical column (length 100 mm, diameter 2.1 mm, particle size 1.9 µm, Thermo-Fisher Ltd.). The flow rate used for UHPLC was

400 µL/min. The two solvents used for LC were water with 0.1% formic acid (solvent A) and methanol with 0.1% formic acid (solvent B). The following settings were used for chromatographic separation in positive ionisation mode: 100% A held for 1 min, 0-100% B over 11 min, 100% B held for 8 min, returning to 100% A over 2 min (total run 22 min). The column was conditioned prior to analysis by running 50:50 water:methanol gradient in isocratic conditions for 3 h at 50ºC followed by 30 min of initial gradient conditions. Xcalibur software (Thermo-Fisher Ltd.) was used to operate the Thermo LTQ-Orbitrap XL MS system using the same method described by Wedge *et al.* (Wedge *et al.*, 2011). The LTQ-Orbitrap MS was calibrated according to the manufacturer's instructions. Orbitrap data was obtained at a resolution of 30,000 (Full width at half maximum (FWHM) defined at *m/z* 400).

The batch programme involved the use 20 injections of QC samples for each individual analytical block. These were used for column conditioning. The analysis batch then followed where five injections of extracted samples were followed by a QC injection. These steps were repeated until all the samples were analysed and the run was concluded by performing three QC injections.

Orbitrap MS*n* analysis parameters

Direct infusion of samples was carried out onto a LTQ-Orbitrap XL hybrid mass spectrometry system (Thermo-Fisher, Bremen, Germany) in order to conduct MSn experiments. Samples were injected at a constant flow of 10 µL/min into ESI probe. A full scan of sample was followed by trapping the ion of interest in ion trap for 30 msec and collision induced fragmentation was carried out with varied CID levels (between 35 to 200 arbitrary units). This was repeated until no more fragmentation could be carried out for the pre-cursor ion in each cycle.

4.2.6 Processing raw data and using UHPLC-MS profiles

Xcalibur software's file conversion option was used to convert the raw data profiles obtained using UHPLC-MS into a NetCDF format (Dunn *et al.*, 2008). XCMS, a free package for R available from (http://masspec.scripps.edu/xcms/xcms.php), was used to deconvolve the peaks using in-house deconvolution parameters fit for high resolution mass spectrometric data collected. Once the peaks were deconvoluted, a Microsoft Excel sheet (XY) matrix was produced containing spectral features including: the retention time and *m/z* ratios. The total numbers of mass spectral features from the LC-MS data was 2618. After deconvolution, lipid identification was carried out using Taverna Workbench version 2.4 (Wedge *et al*., 2011).

4.2.7 Statistical analysis of data

4.2.7.1 Analysis of MALDI-TOF-MS data

All data pre-processing and data analysis were carried out using MATLAB 2012a (The MathWorks, Natick, MA, USA). MALDI-TOF-MS spectra were subjected to the following pre-processing steps: (i) baseline correction using asymmetric least squares (AsLS) (Eilers, 2004) of the raw MS data and (ii) normalisation carried out by dividing the baseline corrected spectrum with the square root of the sum of squares of the spectrum (Brereton, 2003). Multivariate analysis included principal components discriminant function analysis (PC-DFA) and partial least squares for discriminant analysis (PLS-DA). PLS-DA is a linear model representing a supervised method used for classification. For PLS-DA with 1,000 bootstraps was performed. In this process the data were split into two different sets: a training set and a test set using bootstrap re-sampling based on biological replicates as described before (AlMasoud *et al*., 2014 (Chapter 2)).

In order to identify the most significant lipids features, PLS-DA loadings plot was used. The lipid maps online database was used to identify the lipid peaks based on accurate mass information from MALDI-TOF-MS analysis (http://www.lipidmaps.org/).

4.2.7.2 Analysis of LC-MS data

PC-DFA and PLS-DA were also performed on LC-MS data and PLS-DA modelling was also validated using bootstrap resampling. As also described above loadings plots were generated to identify the most significant lipid features at both species and strain levels.

4.2.7.3 Comparison of two analytical techniques

MALDI-TOF-MS and LC-MS results were then compared using the Procrustean test (Peres-Neto and Jackson, 2001). The test was based on Procrustes analysis which is an effective approach for assessing the similarities and differences between different ordination spaces from cluster analyses and has been used previously for the assessment of different analytical techniques (AlRabiah *et al*. 2014). In Procrustes analysis, the similarity between two sets of multivariate data sets, i.e. two matrices with same number of rows, was measured in terms of the Procrustes distance, which ranges between 0 and 1, where 0 indicates a perfect match and 1 indicates nothing in common. The Procrustes test on the two datasets was based such Procrustes distance. Give two data matrices, a Procrustes distance was calculated (named observed Procrustes distance) and this distance was then compared against a null distribution generated by n permutations. In each permutation, the order of the rows in one matrix (e.g. MALDI-TOF-MS lipid) was randomly permuted while that of the other (e.g. LC-MS lipid) remained the same; a Procrustes distance was then calculated. A total number of n Procrustes distances were calculated from n different random permutations and formed the null distribution. An empirical *p*-value was then calculated by counting the cases where the Procrustes distance from the null distribution was lower than the observed Procrustes distance. In this study, we compared three data sets, i.e. MALDI- TOF-MS lipids, MALDI- TOF-MS protein and LC-MS lipids using Procrustes test. For each test, 1,000 permutations were performed and the observed Procrustes distance and the associated p-values were reported.

4.3 Results and Discussion

Traditional phenotypic methods such as biochemical tests (Wilkins and Lay, 2005) are used routinely to discriminate between different microorganisms. These methods, however, are not always reliable, are generally laborious, time-consuming and provide limited information compared to modern analytical techniques (Wenk, 2005; Allwood and Goodacre, 2010; AlRabiah *et al*. 2014). For the purpose of this lipidomics study, two complementary analytical techniques were used to analyse lipids extracted from 33 *Bacillus* and *Brevibacillus* strains – MALDI-TOF-MS and LC-MS. The findings of this work show that the use of MALDI-TOF-MS to classify bacteria based on lipid extracts is promising and can be a useful analytical tool for research carried out in the lipidomics field.

At the beginning of this work, three different species (*B. cereus* B0702, *B. subtilis* B0099 and *Br. laterosporus* B0043) were analysed using LC-MS to determine the optimal time point for collecting bacterial samples based on the quality of separation determined using LC-MS data. Our observations show that samples collected after 10 h of cell culture (Figure S4.2), generated better separation for the three species due to there being a sufficient amount of biomass that is needed for lipid extraction, which was evident from the optical density (OD) (data not shown).

4.3.1 MALDI-TOF-MS lipid profiles

Recently, we optimised the experimental conditions for the detection of lipid mixtures using MALDI-TOF-MS analysis and fractional factorial design. Our observations suggested that ATT and DHB were the most compatible matrices with lipid mixture. Initially, as routine practice in our laboratory when conducting MALDI-TOF-MS experiments, pilot tests are performed before analysing samples. In this case, two different species, *B. cereus* and *B. subtilis*, were used and analysed with MALDI-TOF-MS using two different matrices ATT (Shanta *et al*., 2012; Stübiger *et al*., 2010) and DHB (Griffiths and Bunch, 2012; Schiller *et al*., 2004) as these were found to be the most compatible matrices with the lipid mixture. Figure S4.3 shows the principal components analysis (PCA) scores plot of *B. cereus* and *B. subtilis* using these matrices and the results suggested that DHB provided better separation between the bacteria based on the total expline variance (TEV) values generated in PC1 dimension in the PCA plots, which were higher at around 84% compared to 54% achieved with ATT. Previous studies showed that DHB is more compatible with lipids than other matrices as DHB matrix peaks do not cause complications when interpreting data (Schiller *et al*., 2001; Zhou *et al*., 2010). A relatively good separation between bacterial samples was still generated using ATT; however, due to the huge number of samples, only the better performing matrix (DHB) was used in order to generate more reliable data for all of the 33 *Bacillus* and *Brevibacillus* strains.

Lipids were extracted from *Bacillus* and *Brevibacillus* using chloroform:methanol (2:1) since this method was used previously with successful outcomes (Allwood *et al*., 2014, Shu *et al*., 2012b, Shu *et al*., 2012a). MALDI-TOF-MS spectra of lipids extracted from all the seven species, *B. amyloliquefaciens* B0177*, B. cereus* B0002, *Br. laterosporus* B0034*, B. licheniformis* B1379*, B. megaterium* B0010^T, *B. sphaericus* B0769 and *B. subtilis* B1382, are shown in Figure 4.1. In general, MALDI-TOF-MS generated high quality data due to high signal-to-noise ratios over the *m/z* range of acquisition.

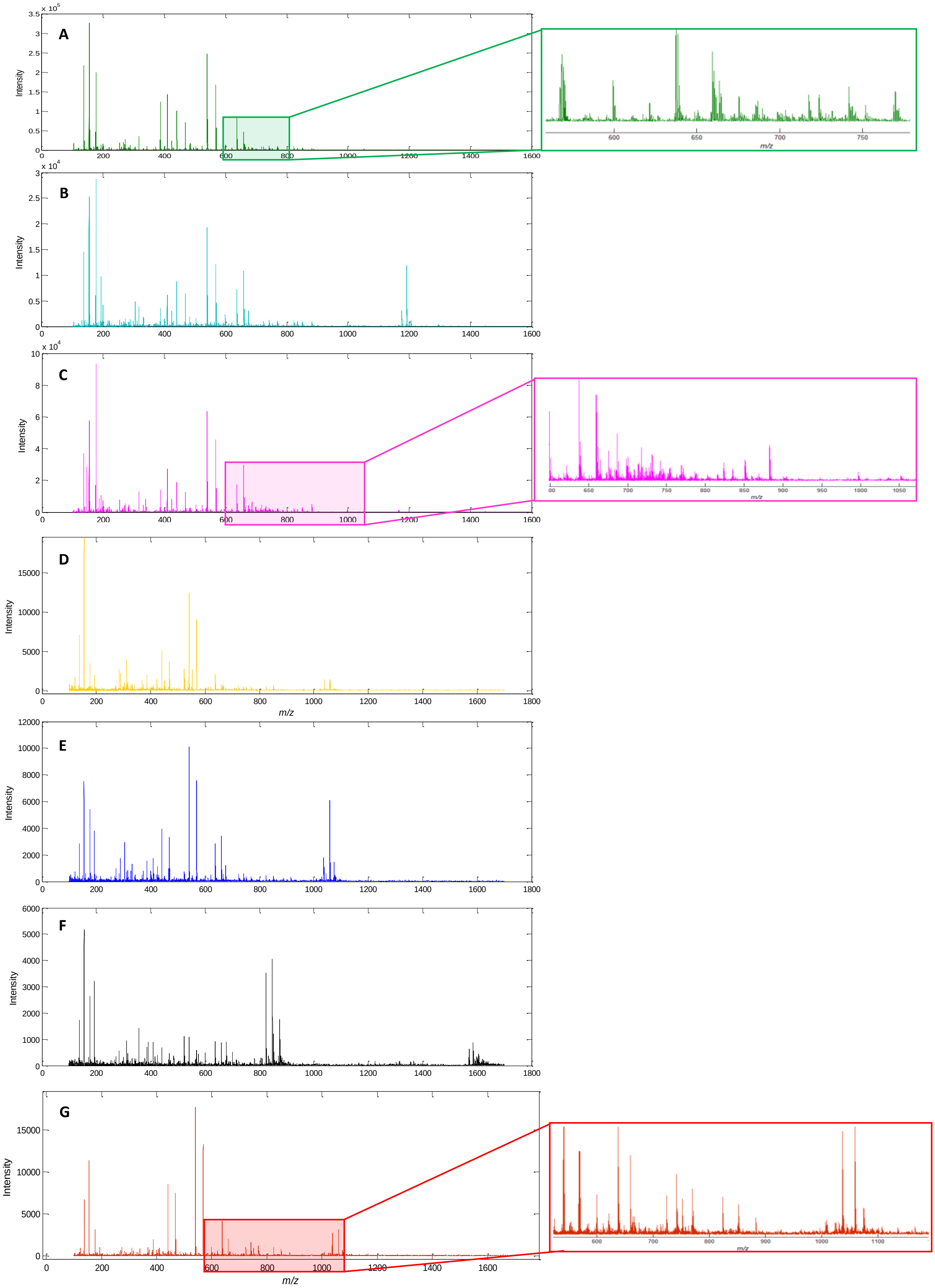


Figure 4.1: Typical MALDI-TOF-MS after pre-processing of lipids extracted from seven species: (A) *B. cereus* B0002, (B) *B. megaterium* B0056, (C) *B. sphaericus* B0769, (D) *B. subtilus* B1382, (E) *B. licheniformis* B1379, (F) *Br. laterosporus* B0034 and (G) *B. amyloliquefaciens* B0177.

At first glance, MALDI-TOF-MS spectra for the 7 species from *Bacillus* and *Brevibacillus* appeared to have different patterns in *m/z* range 200-1600 (Figure 4.1). Some parts of the spectra were amplified to show peaks that cannot be visualised due to low intensities in comparison to the more dominant peaks. These spectra are rich in information and lipids were detected across a broad range mainly below *m/z* 1600. Some of the peaks remained the same for the 7 species; for example, lipids at *m/z* values of 568, 637 and 851. On the other hand, other parts of the spectra are unique to each species, such as the region between *m/z* 1500 and 1600 in *Br. laterosporus*. Visual inspection of the MALDI-TOF-MS spectra revealed features that can be used to discriminate between some of the species. *Br. laterosporus* was characterised by significantly different spectra compared to the other species, most likely due to the expected differences between bacterial genera (Logan and Berkeley, 1984). It is important to note that the biomass concentration was the same for the 7 species analysed in this study. However, the signal-to-noise ratios seemed to be different from one spectrum to another; this is possibly due to the ionisation efficiency of analytes under MALDI-TOF-MS analysis and can possibly be assessed using different matrices.

Figure 4.1(A-G) shows that during the growth of bacterial strains in nutrient broth; they produced lipids represented by the detection of various peaks on different spectra. These peaks, which were readily detectable by a simple MALDI-TOF-MS analysis, may represent significant lipids that can be used as a fingerprint for each type of bacteria. The Lipid Maps database (http://www.lipidmaps.org/) was used to assign the most abundant lipid peaks and the probable assignments for the seven species are listed in Table 4.2. Table 4.2 also shows that sodium and potassium adducts can be seen in the MALDI mass spectra owing to the nature of the biological samples, which are rich in these cations. Most of the significant lipids detected using MALDI-TOF-MS were identified using advanced chemometrics such as PLS-DA modelling; these lipids are highlighted in bold in Table 4.2. Lipids detected in these species are a broad set of naturally occurring molecules. Several studies have confirmed that phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) are the most abundant phospholipids in bacteria such as *Bacillus* spp. (Shu *et al*., 2012a; Epand and Epand, 2011; Dowhan, 1997) and *Escherichia coli* (Shu *et al*., 2012a). *Bacillus* has also been reported to produce other categories of lipids such as digalactosyldiacylglycerol (DGDG) (Gidden *et al.*, 2009), phosphatidylcholine (PC) (Pomerantsev *et al.*, 2003) and fatty acids (FA) (Kaneda, 1977). These significant lipid features were subjected to MS/MS analysis as well as MS*n* analysis on MALDI-TOF-TOF as well as Orbitrap MS respectively in order to obtain structural information to validate putative assignments. It was noted that not all lipid features that were in significant abundance required for MS*n* analysis. Table 2 includes lipid features presents in seven species classified in this study. Identification of lipids was based on accurate mass match on LipidMaps, followed by verification of their presence reported in literature and further confirmation by MS*n* analysis.

With regards to the structural identification carried out by means of tandem MS, the high energy-CID MS/MS (MALDI-TOF-MS) and MS*n* (Orbitrap) spectra exhibited the characteristic fragmentation of the polar head group of the phospholipids. Specifically, ions equivalent to $[M - 43]^+$, $[M - 141]^+$ and $[M - 163]^+$, corresponding to the loss of ethanolamine, ethanolamine phosphate and sodiated ethanolamine phosphate, respectively, were consistently observed in the tandem MS spectra of PE-lipids. In the MS/MS and MS*n* spectra of PA-lipids (one single species has been found), the loss of phosphate ($[M - 98]^+$) and potassium phosphate group ($[M - 136]^+$) have been observed accordingly.

Out of 17 lipids, six lipids were assigned definite identification based on their fragmentation pattern whereas five lipids were observed in insufficient quantities to be able to perform fragmentation. There were six lipids that were only identified based on their accurate mass as their fragmentation pattern did not follow a lipid-like fragmentation. Putatively identified lipids were assigned identification based on either their match on LipidMaps or previous reports of successful fragmentation by other authors using various fragmentation techniques.

Figure 4.1A shows a zoomed in area that contains mass peaks between around *m/z* 600 and 800 in the *B. cereus* spectrum, representing lipids consisting of different numbers of carbon, from different categories such as PE, PG and PC. The spectrum generated from *B. megaterium* (Figure 4.1B) seemed to be similar to *B. cereus* based on the existence of PE, PG and PC, while *B. megaterium* produced a visibly unique peak at around 1206 *m/z*. Figure 4.1D shows the mass spectrum of *B. subtilis*, where fewer peaks were detected compared to *B. cereus* and *B. megaterium*. Notably, the spectrum in Figure 4.1G, which represents *Br. laterosporus*, is largely dominated by peaks at *m/z* 1224, 1315, 1335, 1367, 1570 and 1584, a series of peaks that can be used to identify this species; the fact that this species is different is perhaps not surprising as these bacteria are from a different genera.

Table 4.2: List of probable and definite identification of the seven Bacillus species using MSn fragmentation results. If a peak was detected for a particular lipid, this is illustrated with a colour matching the different species.

m/z	**Matched *m/z***	**Delta**	**Probable assignment**	**Definite identification**	***Br.la***	***B .ce***	***B .me***	***B .su***	***B .am***	***B.li***	***B .sp***
659.5	659.2	0.2406	[Lyso-PI(20:4) + K]$^+$	[Lyso-PI(20:4) + K]$^+$	■	■	■	■	■	■	■
675.5~	675.4	0.0793	[1][PG(27:0) +Na]$^+$	-	■	■	■			■	
678.5	678.5	0.0069	[PC(28:0) +Na]$^+$	-	■						
637.5	637.4	0.0925	[PG(26:1)+H]$^+$	-	■	■	■	■	■	■	■
685.5	685.4	0.0794	[PA(32:1) +K]$^+$	[PA(32:1) +K]$^+$			■	■			■
686.5~	686.4	0.0269	[1][PE (30:0) +Na]$^+$	[PE (30:0) +Na]$^+$	■						■
700.5~	700.4	0.0112	[1][PE(31:0) + Na]$^+$	[PE(31:0) + Na]$^+$	■						■
710.5	710.5	0.0269	[PE(32:2) +Na]$^+$	[PE(32:2)+Na]$^+$		■	■				
714.5~	714.5	0.0044	[1][PE(32:0) +Na]$^+$	[PE(32:0) +Na]$^+$	■						
741.5^	741.4	0.0323	[PG(32:2) + Na]$^+$	-		■	■	■	■	■	■
766.6~	766.5	0.5357	[2][PE (36:2) +Na]$^+$	-	■						
768.5^~	768.5	0.0514	[1][PE(36:1) + Na]$^+$	-		■					
823.5	823.5	0.0459	[PG(38:3) + Na]$^+$	-		■	■		■	■	■
846.6	846.6	0.0008	[PC(41:7) + H]$^+$	-	■						
851.6^~	-	-	[1][LPG(32:0) +H]$^+$	-	■	■	■	■	■	■	■
882.6^	882.6	0.0017	[PC(42:7) +Na]$^+$	-		■	■	■	■	■	■
915.7^~	915.6	0.0985	[1][DGDG(32:0) +Na]$^+$	-						■	

Bacteria: *B. amy, B. amyloliquefaciens; B. cer, B. cereus; Br. lat, Br. laterosporus; B. lic, B. licheniformis; B. meg, B. megaterium; B. sph, B. sphaericus; B. sub, B. subtilis.* Lipids: phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidic acid (PA), phosphatidylglycerol (PG), L-alpha-lysophosphatidylinositol (LPI), lyso- phosphatidylglycerol (LPG) and digalactosyldiacylglycerol (DGDG).

^indicates *m/z* of lipid-like features that were not in high enough quantities to be able to successfully fragment into product ions.
~indicates *m/z* of features that have been putatively assigned an identification based on previously reports in literature.
[1] (Gidden, *et. al* , 2009), [2] (Shu, *et al*. 2012)

However, visual inspection is laborious and unreliable; consequently, advanced chemometric methods were required to extract more information from the MS data in a reproducible, objective and automated manner. We have previously shown that after the optimisation of MALDI-TOF-MS in combination with advanced chemometrics, this analytical technique can become a robust and rapid tool that enables the classification of a large number of *Bacillus* and *Brevibacillus* bacterial strains based on their proteins (AlMasoud *et al*., 2014 (Chapter 2)). Multivariate analysis has proven vital for extracting information when analysing samples using different analytical techniques such as pyrolysis mass spectrometry (PyMS), Fourier transform infrared (FT-IR) spectroscopy and Raman spectroscopy to discriminate between bacterial samples (Goodacre *et al*., 2000; Timmins *et al*., 1998; Goodacre *et al*., 1998). There are different statistical methods that can be used to assess the information generated from the MALDI-TOF-MS spectra, enabling discrimination between the 7 species. One such method is PC-DFA and in Figure 4.2A a three dimensional DFA scores plot shows four major clusters detected based on the data: (1) *B. megaterium* and *B. cereus*; (2) *B. subtilus*, *B. amyloliquefaciens* and *B. licheniformis*; (3) *B. sphaericus*; and (4) *Br. laterosporus*. Figure 4.2B shows that *Br. laterosporus* is well separated in the first DF and is therefore completely chemically different from the other six *Bacillus* species, which confirms differences seen in the MALDI-TOF-MS spectra. These large lipid differences in all the *Br. laterosporus* species dominated both plots, and therefore another PC-DFA plot was generated for *Bacillus* species only. This resulted in more separation between the 6 *Bacillus* species (Figure 4.2 C). Most notably, *B. licheniformis* could be separated from *B. subtilus* and *B. amyloliquefaciens*, since *B. amyloliquefaciens* was shown to similar to *B. subtilis*, which is expected as these two species are phylogenetically very closely related (Priest *et al*., 1987).

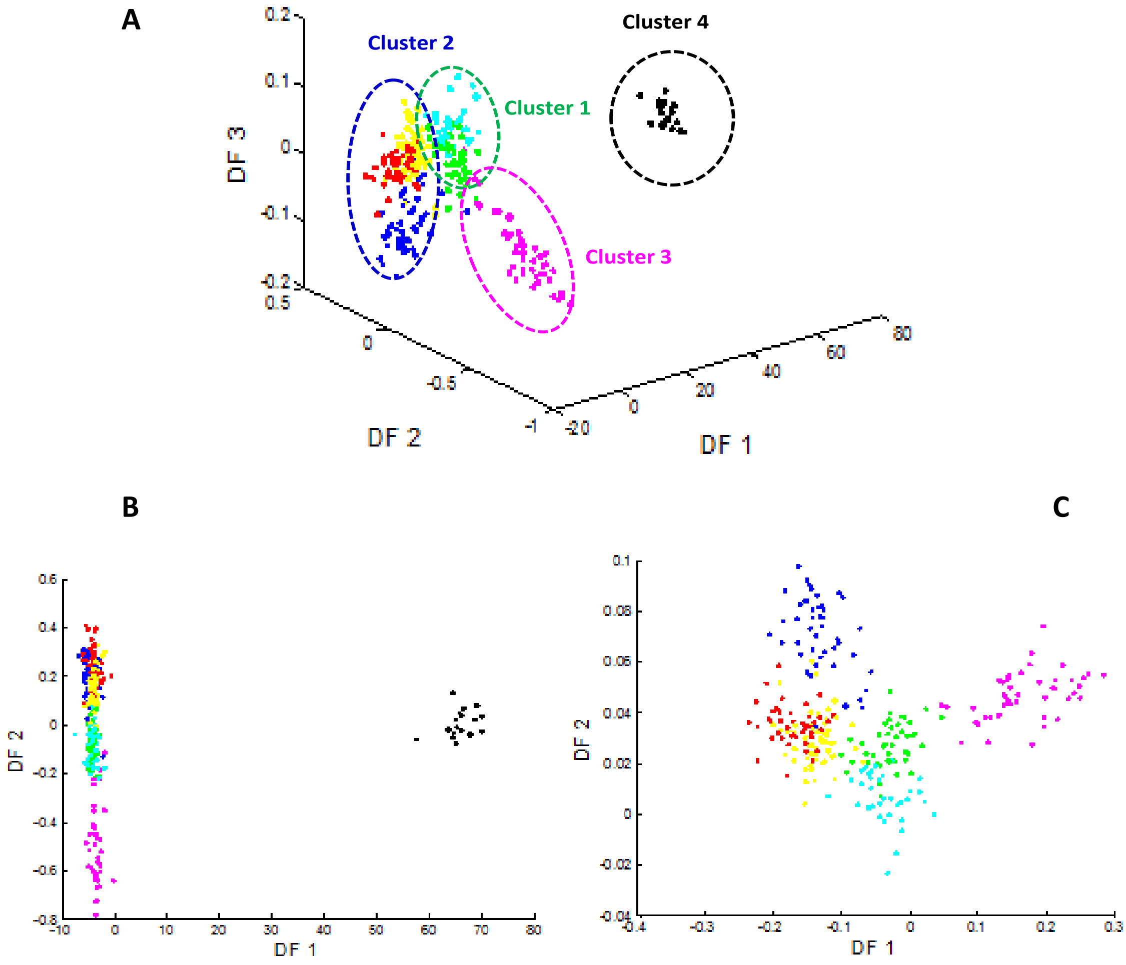


Figure 4.2: DFA scores plots after pre-processing MALDI-TOF-MS data. Different DFA plots were generated for seven species including: (A) DF1 *vs*. DF2 *vs*. DF3; (B) DF1 *vs*. DF2; (C) DF1 *vs*. DF2 of six species with (*Br. laterosporus* removed) DF1 *vs*. DF2. Different colours represent different species; Table 4.1 shows the annotations.

These *Bacillus* species were examined previously using different types of techniques, such as the analytical profile index (API) (Logan and Berkeley, 1984) and genotyping using 16S rDNA sequencing (Goodacre *et al*., 2000). API test used for bacterial classification based on miniaturized biochemical test, using API tests, four main groups were observed including: group I with only *B. cereus*, group II containing only *Br. laterosporus*, group III containing only *B. sphaericus* and a large group IV consisting of *B. subtilis*, *B. licheniformis*, *B. amyloliquefaciens*, and *B. megaterium*. By contrast, phylogenetic analysis using 16S rDNA sequencing detected five different clusters (1) *B. sphaericus*, (2) *Br. laterosporus*, (3) *B. subtilis*, *B. licheniformis*, and *B. amyloliquefaciens*, (4) *B. megaterium*, and (5) *B. cereus*.

The next stage in the present study was to assess whether the MALDI lipid profiles contained enough information to all identification of the different bacteria analyses. Therefore, automated classification prediction accuracies for the 7 species and 33 strains were calculated based on the MALDI-TOF-MS data using multiple PLS-DA models. Tables 4.3 and S4.6 summarise the classification of *Bacillus* and *Brevibacillus* bacteria at the species level (i.e., 7 classes) and at the strain level (i.e., 33 classes), respectively. The average correct classification rate (CCR) for the 7 species was 62.23% whereas the CCR for the 33 strains was 15.67%. Interestingly, prediction of *Br. laterosporus* based on MALDI-TOF-MS data was more accurate compared to the other species. Moreover, if *B. amyloliquefaciens, B. subtilis* and *B. licheniformis* are considered as one class, as these species have the same phylogenetic origin (Logan and Berkeley, 1984), the prediction accuracies for the three species increases from 60.26%, 67.76% and 59.60% to 92.28%, 91.61% and 87.45%, respectively.

Table 4.3: Prediction accuracy of seven species from *Bacillus* using PLS-DA based on MALDI-TOF-MS data

	B.am	***B.ce***	***B.li***	***B.me***	***B.sp***	***B.su***	***Br.la***
B.am	60.26	5.85	5.80	1.25	0.50	26.22	0.01
B.ce	0.23	63.36	2.45	18.02	6.30	9.55	0.00
B.li	8.52	10.51	59.60	2.36	1.58	17.34	0.00
B.me	2.84	36.54	0.87	43.40	0.95	15.31	0.00
B.sp	0.02	19.17	0.16	0.23	80.11	0.21	0.00
B.su	12.50	9.63	3.56	6.35	0.10	67.76	0.00
Br.la	2.26	3.15	1.33	2.60	1.60	5.30	83.66

B. am: B. amyloliquefaciens, B. ce: B. cereus, Br. la: Br. laterosporus, B. li: B. licheniformis, B. me: B. megaterium, B. sp: B. sphaericus and *B. su: B. subtilis.*

Heat map plots from confusion matrices were generated using the PLS models from the 7 species and 33 strains (Figure S4.4A and B respectively). In these figures, warm colours (e.g. red) are indicative of species or strains of high percentage class membership assignments using MALDI-TOF-MS data, while cold colours (e.g., blue) represent low percentage class membership assignment. It can be seen that the colours on diagonal “tiles” were generally much warmer than off-diagonal “tiles”, indicating high agreement between predicted and known classes.

The same bacterial species were previously classified based on MALDI-TOF-MS analysis of proteins from intact bacterial cells [4]. The overall classification based on protein analysis was highly similar to that based on lipid analysis. However, the quality of classification carried out based on protein analysis from intact cells was superior with CCR values of over 80% at the species level (average CCR of 89%). This may be explained by the better quality of spectra obtained for proteins using MALDI-TOF-MS or the inherent differences in gene products between bacteria compared to those of metabolites, such as lipids. The case of misclassification of *Bacillus megaterium* with *Bacillus cereus* based on lipid profiles is interesting as these two species were very distinctly classified using protein profiles (CCR of 91% and 83%, respectively), indicating that the protein profiles were different whereas the lipid profiles were similar.

4.3.2 Interpretation of LC-MS lipid profiles

Although MALDI-TOF-MS is a robust and rapid analytical technique, interference of matrix peaks with low molecular weight analyte peaks, especially those of lipids below 300 *m/z*, and inability of MALDI-TOF-MS to discriminate between isobaric peaks (which have the same *m/z*) present a potential limitation to this chemotaxonomic technique. Therefore, LC-MS analysis was carried out on the same samples to complement and confirm the classification of bacteria based on MALDI-TOF-MS analysis. Although the mass accuracy (< 10 ppm) of TOF analysers is high (~15000 FWHM in reflectron mode), it is recognised that Orbitrap mass analysers have higher mass accuracy (sub-ppm) and resolution (>100,000 FWHM), allowing the identification of lipids to be more accurate and robust. The high mass accuracy and resolution of the Orbitrap combined with resolution of analytes by HPLC can reduce the observed interference between the different lipid species and other components of the samples. These factors considered together are expected to lead to better classification and identification by LC-MS.

The LC-MS findings suggest that *Bacillus* species produced many different lipid categories such as: phosphatidylcholine (PC), phosphatidylethanolamine (PE), diradylglycerolipid (DG), glycerophosphoglycerol (PG), phosphatidic acids (PA), glycerophosphoinositol (PI), ceramide (Cer) as well as free fatty acids (FFA).

Methyl-branched fatty acids were observed in the lipid profiles of *Bacillus* species; these include dimethyl tetradecanoic acid (C15), methyl hexadecanoic acid (C17), 13-methyl pentadecanoic acid (C16) and menaquinones in line with previous reports (Kaneda, 1977; Kaneda, 1972). A summary of these putative lipid categories is shown in Table S4.8 (in the enclosed material: Chapter 4_SI). Table S4.8 shows that the main lipids detected in LC-MS were most likely PE, PC and free fatty acid (FFA), in addition to a small number of PA.

In order to compare classifications based on LC-MS data with those generated from MALDI-TOF-MS, PC-DFA was also applied to LC-MS data. Figure 4.3A shows a PC-DFA scores plot in three dimensions. It can be noted that four main clusters were detected: (1) *B. megaterium* and *B. cereus,* (2) *B. subtilus*, *B. amyloliquefaciens* and *B. licheniformis,* (3) *B. sphaericus* and (4) *Br. laterosporus*. These observations were in agreement with MALDI-TOF-MS analysis based on these bacterial lipids (Figure 4.2A). Moreover, this observation also is similar to the previous work that we carried out based on whole cell analysis of proteins using MALDI-TOF-MS (AlMasoud *et al*., 2014 (Chapter 2)), Raman spectroscopy (Lopez-Diez and Goodacre, 2004) and direct infusion ESI-MS (Vaidyanathan *et al*., 2001). Figure 4.3B shows that *Br. laterosporus* again is significantly different from the other strains when DF1 *vs*. DF3 is plotted. Therefore, *Br. laterosporus* was again excluded from data analysis and this resulted in the separation of *B. licheniformis* from *B. subtilus* and *B. amyloliquefaciens* (Figure 4. 3C).

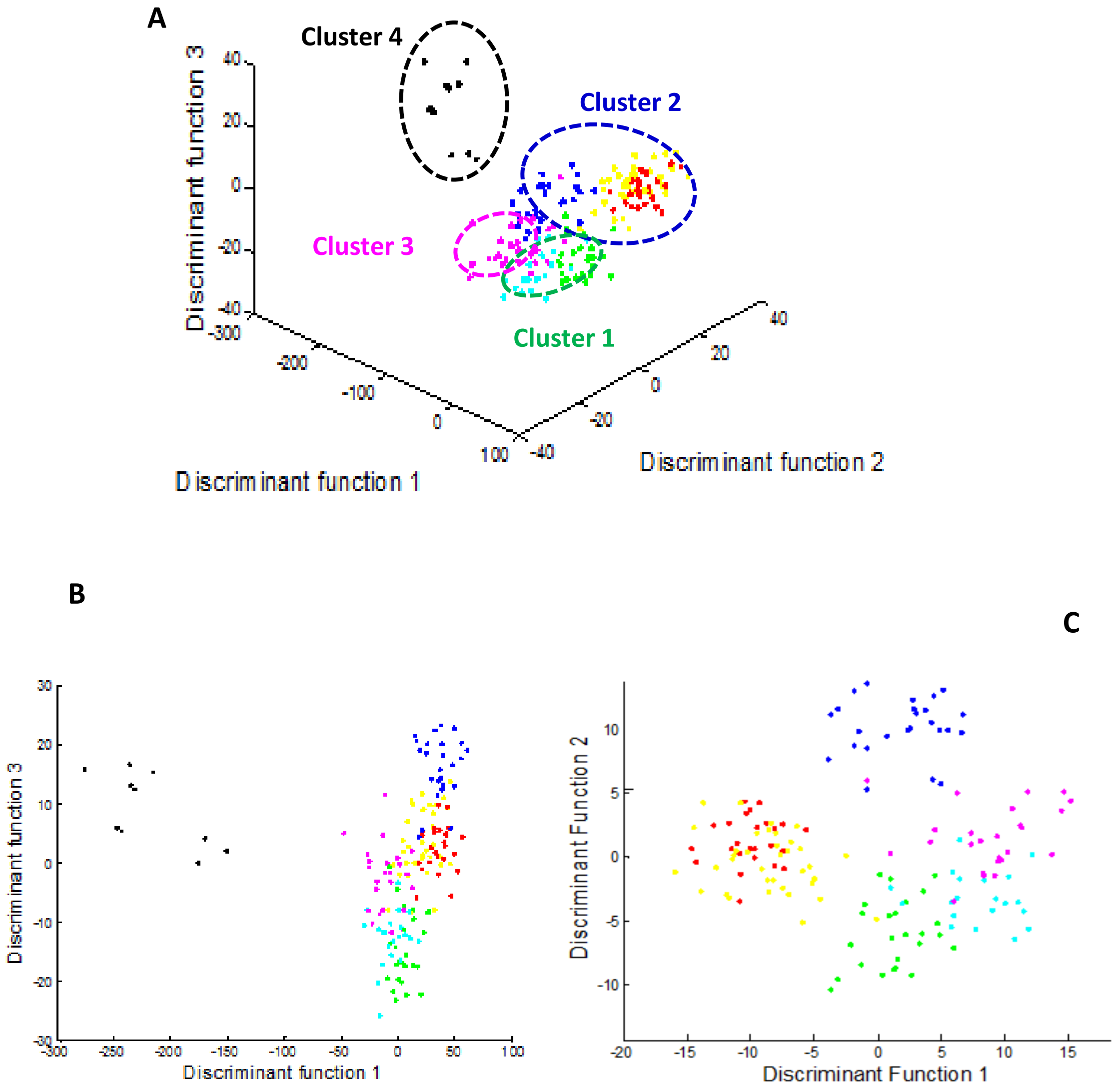


Figure 4.3: DFA scores plots after pre-processing the LC-MS data. Different DFA plots were generated for seven species including: (A) DF1 *vs*. DF2 *vs*. DF3; (B) DF1 *vs*. DF3 and (C) DF1 *vs*. DF2 for the six species (again *Br. laterosporus* was not included). Different colours represent different species; Table 4.1 shows the annotations.

In order to effect bacterial classification from these LC-MS lipid profiles data analysis was carried out used using a PLS-DA model for the 7 species (i.e. 7 classes) and 33 strains (i.e. 33 classes). Tables 4.4 and S4.7 show the prediction accuracies for the 7 species and 33 strains, respectively. Table 4.4 shows that qualitative information based on lipids is appropriate for accurate classification of bacteria. This model provided an average correct classification rate (CCR) of 77.03% and 15.2 % for the 7 species and 33 strains, respectively. Looking back at Table 4.3 which was generated from MALDI-TOF-MS data using the PLS-DA model, it can be observed that the results from these two analytical techniques overlapped and most of the species reflected higher predication accuracies based on LC-MS data due to the high sensitivity of LC-MS compared to MALDI-TOF-MS.

Table 4.4: Prediction accuracy of seven species from *Bacillus* using PLS-DA based on the LC-MS data

	B.am	***B.ce***	***B.li***	***B.me***	***B.sph***	***B.su***	***Br.la***
B.am	93.85	1.09	0.18	0.16	0.10	4.62	0.00
B.ce	5.41	71.93	1.63	17.75	0.59	2.70	0.00
B.li	1.50	0.40	84.03	0.95	4.39	8.73	0.00
B.me	3.04	35.64	3.22	38.41	8.86	10.82	0.01
B.sp	2.93	9.95	1.92	6.44	77.13	1.62	0.01
B.su	4.20	0.61	4.64	1.69	1.05	87.81	0.01
Br.la	0.01	0.03	0.00	1.13	7.28	0.11	91.43

B. am: B. amyloliquefaciens, B. ce: B. cereus, Br. la: Br. laterosporus, B. li: B. licheniformis, B. me: B. megaterium, B. sp: B. sphaericus and *B. su: B. subtilis.*

The findings in Table 4.4 can be summarised in three points:

(i) *Br. laterosporus* did not match other species, which is not surprising because these bacteria are from a different genus.

(ii) Some species, including *B. cereus* and *B. megaterium*, are sometimes misclassified since they are phylogenetically related (Lopez-Diez and Goodacre, 2004).

(iii) *B. subtilis* is sometimes misclassified with *B. licheniformis* and *B. amyloliquefaciens*.

Furthermore, heat maps of the confusion matrices were generated in order to visualise the classification of *Bacillus* strains. Figure S4.5 A and B show the heat maps generated for the 7 species and 33 strains, respectively. Comparing the two heat maps that were generated from MALDI-TOF-MS (Figure S4.4A) and LC-MS (Figure S4.5A) when 7 classes (species) are used, it can be seen that both techniques were robust at the species level and both techniques show that *B. megaterium* can be misclassified with *B. cereus*. Moreover, when 33 strains were compared, it can be seen that all the strains from *Br. laterosporus* showed the highest prediction accuracies. In addition, *B. subtilis* B0044 and *B. subtilis* B0098 overlapped and gave the mixed classification results in both heat maps. These observations from LC-MS confirm that MALDI-TOF-MS is indeed a very useful and robust analytical technique which generates classifications similar to LC-MS.

LC-MS has relatively high resolution and sensitivity, and it allows quantitative analysis to be performed. Figure S4.6 shows the relative levels of examples of the most significant lipids (based on the PCA loadings plot) in the seven species classified in this study. Table S4.9 (see attached material: Chapter 4_SI) shows a list of the putative assignment of the significant lipids. Again, based on the levels of these lipids, *Br. laterosporus* was observed to be significantly different in comparison with the other species, particularly based on fatty acid content (Figure S4.6 A-D). Different lipids can be used to distinguish between species; for example, PE (14:1(9Z)/15:0) in Figure S4.6G could be used to distinguish *B. subtilis* from *B. amyloliquefaciens* and *licheniformis*. Moreover, the level of 7-hydroxy-10E, 16-heptadecadien-8-ynoic acid from the FA category was relatively high in *B. licheniformis* compared to other species (Figure S4.6H). Significant lipids were also identified in the remaining 33 strains and are shown in Figure S4.7 A-D. Table S4.10 (see enclosed material: Chapter 4_SI) lists the putatively assignment of examples of significant lipids in the 33 strains. Looking back at Figure S4.7A, it can be noted that the existence of an unknown lipid is significantly higher in all the strains from *Br. laterosporus* compared to the remaining strains from *Bacillus*. Moreover, Figure S4.7 B-D confirms that *B. subtilis* B0044 and *B. subtilis* B0098 are highly similar and this is most likely due to producing similar amounts of lipids.

4.3 Comparison of two analytical techniques

The objective of this step was to compare the patterns of *Bacillus* and *Brevibacillus* bacteria based on lipid extracts to those based on protein analysis which has already been carried out previously using MALDI-TOF-MS (AlMasoud *et al*., 2014 (Chapter 2)). In order to assess the similarities in the patterns that were generated from the two analytical techniques used for analysing lipids and proteins from *Bacillus* and *Brevibacillus* samples, three datasets were compared: MALDI-TOF-MS and LC-MS were used for the analysis of lipids and MALDI-TOF-MS for protein analysis. This led to the use of Procrustean test. Table 4.5 shows the similarity between data obtained from DFA plots for the 7 species (highlighted in bold) and the 33 strains (in normal font). Table 4.5 highlights the following observations:

(i) MALDI-TOF-MS lipid profiles and LC-MS lipid profiles had the highest similarity level with a Procrustes distance of 0.0699 and a *p*-value of <0.001 (i.e. not a single case where the permuted data obtained a lower Procrustes distance than that of the data without permutation). These findings were encouraging because this indicated bacteria were successfully classified using MALDI-TOF-MS analysis of lipids.

(ii) MALDI-TOF-MS protein profiles and both lipid based experiments (MALDI-TOF-MS and LC-MS) were significantly similar with Procrustes errors of 0.1006 and 0.1081 (*p*<0.001), respectively. However, the errors are higher compared to that highlighted in point (i), which was expected as different compounds were compared (i.e. lipids and proteins), and as a result this observation supports the validity of our work.

(iii) Data based on the 33 strains generated higher Procrustes errors comparing to data on the 7 species and this is as expected because of the larger number of strains compared to the number of species, hence the more complex data and high similarity within a bacterial species. Nevertheless, the *p*-values were still very significant (*p*<0.001).

Table 4.5: Similarity between three different data sets for species and strain levels using Procrustes distance

	MALDI-MS (lipid)	MALDI-MS (protein)	LC-MS (lipid)
MALDI-MS (lipid)	-	-	-
MALDI-MS (protein)	**0.1006 (*p*<0.001)** 0.3443 (*p*<0.001)	-	-
LC-MS (lipid)	**0.0699 (*p*<0.001)** 0.3262 (*p*<0.001)	**0.1081 (*p*<0.001)** 0.4717 (*p*<0.001)	-

The values that are highlighted in bold correspond to 7 classes (7 species) and those in normal font correspond to 33 classes (33 strains)

4.4 Conclsion and remarks

MALDI-TOF-MS in an analysing bimolecular compounds, has proven to be useful for discriminating between different microorganisms, and its use in bacterial profiling is common in clinical microbiology laboratories (Sauer and Kliem, 2010; Carbonnelle *et al*., 2011). Our study involved the use of two analytical techniques, MALDI-TOF-MS and LC-MS, to analyse 33 strains from 7 bacterial species belonging to the *Bacillus* (*n*=6 species) and *Brevibacillus* (*n*=1) genera. The spectral information generated using MALDI-TOF-MS on lipids extracted from the 33 strains and 7 species was highly informative and was useful in discriminating between the bacteria at the sub-species level. In order to validate these findings LC-MS data were used to evaluate and confirm results obtained from the simple and rapid MALDI-TOF-MS analysis for bacterial classification. The results obtained from the two analytical techniques based on the 7 bacterial species showed that these data were highly similar, which was supported by the use of Procrustes distance analysis. The calculated Procrustes distance was 0.0699 for the two datasets indicating very high similarity between MALDI-TOF-MS and LC-MS data. Finally, MALDI-TOF-MS data based on analysis of extracted lipids and previous analysis of proteins, from intact bacteria analysis of the same species, were also very similar (Procrustes distance was 0.1006). These findings suggest that MALDI-TOF-MS can be used reliably as a powerful routine clinical tool for the robust classification and reliable identification of bacteria based on lipids or proteins. However; direct MALDI-TOF-MS analysis for lipids from intact bacterial cells is expected to reduce the scope of analysis and the quality of data would be compromised due to the interference from different cell components such as proteins, which can lead to ion suppression and spectra dominated by highly abundant proteins considering the analysis would be carried out using MALDI-TOF-MS. The extra steps required for lipid extraction and sample preparation are therefore a necessary inconvenience to acquire better quality data. For future work, it would be interesting to compare lipid analysis on intact cells with lipid analysis after extraction on the same bacteria.

4.5 References

Allwood, J. W., Alrabiah, H., Correa, E., Vaughan, A., Xu, Y., Upton, M. and Goodacre, R. 2014. A workflow for bacterial metabolic fingerprinting and lipid profiling: application to Ciprofloxacin challenged *Escherichia coli*. *Metabolomics***,** **11**, 1-16

Allwood, J. W. and Goodacre, R. 2010. An introduction to liquid chromatography–mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis,* **21**, 33-47

AlRabiah, H., Xu, Y., Rattray, N.J., Vaughan, A.A., Gibreel, T., Sayqal, A., Upton, M., Allwood, J.W. and Goodacre, R. 2014. Multiple metabolomics of uropathogenic E. coli reveal different information content in terms of metabolic potential compared to virulence factors. Analyst, **139**, 4193-4199

AlMasoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. 2014. Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the characterisation of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta,* **840**, 49-57

Bernardo, K., Pakulat, N., Macht, M., Krut, O., Seifert, H., Fleer, S., Hünger, F. and Krönke, M. 2002. Identification and discrimination of *Staphylococcus aureus* strains using matrix-assisted laser desorption/ionisation-time of flight mass spectrometry. *Proteomics*, **2**, 747-753

Brereton, R. G. 2003. *Chemometrics: data analysis for the laboratory and chemical plant*, Chichester, John Wiley and Sons, pp. 489

Calvano, C. D., Zambonin, C. G. and Palmisano, F. 2011. Lipid fingerprinting of Gram-positive *lactobacilli* by intact cells matrix-assisted laser desorption/ionisation mass spectrometry using a proton sponge based matrix. *Rapid Communications in Mass Spectrometry,* **25**, 1757-1764

Carbonnelle, E., Mesquita, C., Bille, E., Day, N., Dauphin, B., Beretti, J.-L., Ferroni, A., Gutmann, L. and Nassif, X. 2011. MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory. *Clinical Biochemistry,* **44**, 104-109

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. 1996. The rapid identification of intact microorganisms using mass spectrometry. *Nature Biotechnology,* **14**, 1584-1586

Cliff, J. B., Kreuzer, H. W., Ehrhardt, C. J. and Wunschel, D. S. 2012. *Chemical and physical signatures for microbial forensics,* Alexandria, VA, USA Springer, pp.35-36

Dowhan, W. 1997. Molecular basis for membrane phospholipid diversity: why are there so many lipids? *Annual Review of Biochemistry,* **66**, 199-232

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-Mcintyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B. and Goodacre, R. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols,* **6**, 1060-1083

Dunn, W. B., Broadhurst, D., Brown, M., Baker, P. N., Redman, C. W. G., Kenny, L. C. and Kell, D. B. 2008. Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *Journal of Chromatography B,* **871**, 288-298

Eilers, P. H. C. 2004. Parametric Time Warping. *Analytical Chemistry,* **76**, 404-411

Epand, R. M. and Epand, R. F. 2011. Bacterial membrane lipids in the action of antimicrobial agents. *Journal of Peptide Science,* **17**, 298-305

Fahy, E., Cotter, D., Sud, M. and Subramaniam, S. 2011. Lipid classification, structures and tools. *Biochimica et Biophysica Acta,* **1811**, 637-647

Fenselau, C. and Demirev, P. A. 2001. Characterisation of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews,* **20**, 157-171

Gaia, V., Casati, S. and Tonolla, M. 2011. Rapid identification of Legionella spp. by MALDI-TOF MS based protein mass fingerprinting. *Systematic and Applied Microbiology,* **34**, 40-44

Gidden, J., Denson, J., Liyanage, R., Ivey, D. M. and Lay Jr, J. O. 2009. Lipid compositions in *Escherichia coli* and *Bacillus subtilis* during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry. *International Journal of Mass Spectrometry,* **283**, 178-184

Goodacre, R., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B. and Rooney, P. J. 1998. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology,* **144**, 1157-1170

Goodacre, R., Shann, B., Gilbert, R. J., Timmins, E. M., Mcgovern, A. C., Alsberg, B. K., Kell, D. B. and Logan, N. A. 2000. Detection of the dipicolinic acid biomarker in *Bacillus* spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Analytical Chemistry,* **72**, 119-127

Goodacre, R., Vaidyanathan, S., Bianchi, G. and Kell, D. B. 2002. Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst,* **127**, 1457-1462.

Griffiths, R. L. and Bunch, J. 2012. A survey of useful salt additives in matrix-assisted laser desorption/ionisation mass spectrometry and tandem mass spectrometry of lipids: introducing nitrates for improved analysis. *Rapid Communications in Mass Spectrometry,* **26**, 1557-1566.

Hashimoto, M., Seki, T., Matsuoka, S., Hara, H., Asai, K., Sadaie, Y. and Matsumoto, K. 2012. Induction of extracytoplasmic function sigma factors in *Bacillus subtilis* cells with defects in lipoteichoic acid synthesis. *Microbiology***,** **159**, 23-35

Irudayaraj, J., Yang, H. and Sakhamuri, S. 2002. Differentiation and detection of microorganisms using Fourier transform infrared photoacoustic spectroscopy. *Journal of Molecular Structure,* **606**, 181-188

Kaneda, T. 1972. Positional distribution of fatty acids in phospholipids from *Bacillus subtilis*. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism,* **270**, 32-39

Kaneda, T. 1977. Fatty acids of the genus *Bacillus*: an example of branched-chain preference. *Bacteriological Reviews,* **41**, 391

Lasch, P., Beyer, W., Nattermann, H., Stämmler, M., Siegbrecht, E., Grunow, R. and Naumann, D. 2009. Identification of *Bacillus anthracis* by using matrix-assisted laser desorption ionisation-time of flight mass spectrometry and artificial neural networks. *Applied and Environmental Microbiology,* **75**, 7229-7242

Lay, J. O. 2001. MALDI-TOF mass spectrometry of bacteria. *Mass Spectrometry Reviews,* **20**, 172-194

Logan, N. and Berkeley, R. 1984. Identification of *Bacillus* strains using the API system. *Journal of General Microbiology,* **130**, 1871-1882

Lopez-Diez, E. C. and Goodacre, R. 2004. Characterisation of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry,* **76**, 585-591

Pomerantsev, A., Kalnin, K., Osorio, M. and Leppla, S. 2003. Phosphatidylcholine-specific phospholipase C and sphingomyelinase activities in bacteria of the *Bacillus cereus* group. *Infection and Immunity,* **71**, 6591-6606

Priest, F. G. and Austin, B. 1993. *Modern bacterial taxonomy,* Springer Science and Business Media

Sauer, S. and Kliem, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* **8**, 74-82

Sauer, S., Freiwald, A., Maier, T., Kube, M., Reinhardt, R., Kostrzewa, M. and Geider, K. 2008. Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS ONE,* **3**, e2843

Schiller, J., Süß, R., Arnhold, J., Fuchs, B., Leßig, J., Müller, M., Petković, M., Spalteholz, H., Zschörnig, O. and Arnold, K. 2004. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research. *Progress in Lipid Research,* **43**, 449-488

Schiller, J., Zschörnig, O., Petkovic, M., Müller, M., Arnhold, J. and Arnold, K. 2001. Lipid analysis of human HDL and LDL by MALDI-TOF mass spectrometry and 31P-NMR. *Journal of Lipid Research,* **42**, 1501-1508

Shanta, S. R., Kim, T. Y., Hong, J. H., Lee, J. H., Shin, C. Y., Kim, K.-H., Kim, Y. H., Kim, S. K. and Kim, K. P. 2012. A new combination MALDI matrix for small molecule analysis: application to imaging mass spectrometry for drugs and metabolites. *Analyst,* **137**, 5757-5762

Shu, X., Li, Y., Liang, M., Yang, B., Liu, C., Wang, Y. and Shu, J. 2012a. Rapid lipid profiling of bacteria by online MALDI-TOF mass spectrometry. *International Journal of Mass Spectrometry,* **321–322**, 71-76

Shu, X., Liang, M., Yang, B., Li, Y., Liu, C., Wang, Y. and Shu, J. 2012b. Lipid fingerprinting of *Bacillus* spp. using online MALDI-TOF mass spectrometry. *Analytical Methods,* **4**, 3111-3117

Stübiger, G., Belgacem, O., Rehulka, P., Bicker, W., Binder, B. R. and Bochkov, V. 2010. Analysis of Oxidized Phospholipids by MALDI Mass Spectrometry Using 6-Aza-2-thiothymine Together with Matrix Additives and Disposable Target Surfaces. *Analytical Chemistry,* **82**, 5502-5510

Timmins, É. M., Howell, S. A., Alsberg, B. K., Noble, W. C. and Goodacre, R. 1998. Rapid differentiation of closely relatedCandida species and strains by pyrolysis-mass spectrometry and fourier transform-infrared spectroscopy. *Journal of Clinical Microbiology,* **36**, 367-374

Vaidyanathan, S., Rowland, J. J., Kell, D. B. and Goodacre, R. 2001. Discrimination of aerobic endospore-forming bacteria via electrospray-ionisation mass spectrometry of whole cell suspensions. *Analytical Chemistry,* **73**, 4134-4144

Van Meer, G., Voelker, D. R. and Feigenson, G. W. 2008. Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology,* **9**, 112-124

Vance, J. E. and Vance, D. E. 2008. *Biochemistry of lipids, lipoproteins and membranes,* Amsterdam, Elsevier, pp.1-39

Wedge, D. C., Allwood, J. W., Dunn, W., Vaughan, A. A., Simpson, K., Brown, M., Priest, L., Blackhall, F. H., Whetton, A. D., Dive, C. and Goodacre, R. 2011. Is Serum or Plasma More Appropriate for Intersubject Comparisons in Metabolomic Studies? An Assessment in Patients with Small-Cell Lung Cancer. *Analytical Chemistry,* **83**, 6689-6697

Wenk, M. R. 2005. The emerging field of lipidomics. *Nature Reviews Drug Discovery,* **4**, 594-610

Wilkins, C. L. and Lay, J. O. 2005. *Identification of microorganisms by mass spectrometry,* New Jersey, John Wiley and Sons, pp.303

Winder, C. L., Dunn, W. B., Schuler, S., Broadhurst, D., Jarvis, R., Stephens, G. M. and Goodacre, R. 2008. Global Metabolic Profiling of *Escherichia coli* Cultures: an Evaluation of Methods for Quenching and Extraction of Intracellular Metabolites. *Analytical Chemistry,* **80**, 2939-2948

Wymann, M. P. and Schneiter, R. 2008. Lipid signalling in disease. *Nature Reviews Molecular Cell Biology,* **9**, 162-176

Zhang, J. I., Talaty, N., Costa, A. B., Xia, Y., Tao, W. A., Bell, R., Callahan, J. H. and Cooks, R. G. 2011. Rapid direct lipid profiling of bacteria using desorption electrospray ionisation mass spectrometry. *International Journal of Mass Spectrometry,* **301**, 37-44

Zhou, P., Altman, E., Perry, M. B. and Li, J. 2010. Study of matrix additives for sensitive analysis of lipid A by matrix-assisted laser desorption ionisation mass spectrometry. *Applied and Environmental Microbiology,* **76**, 3437-3443

4.6 Supplementary information

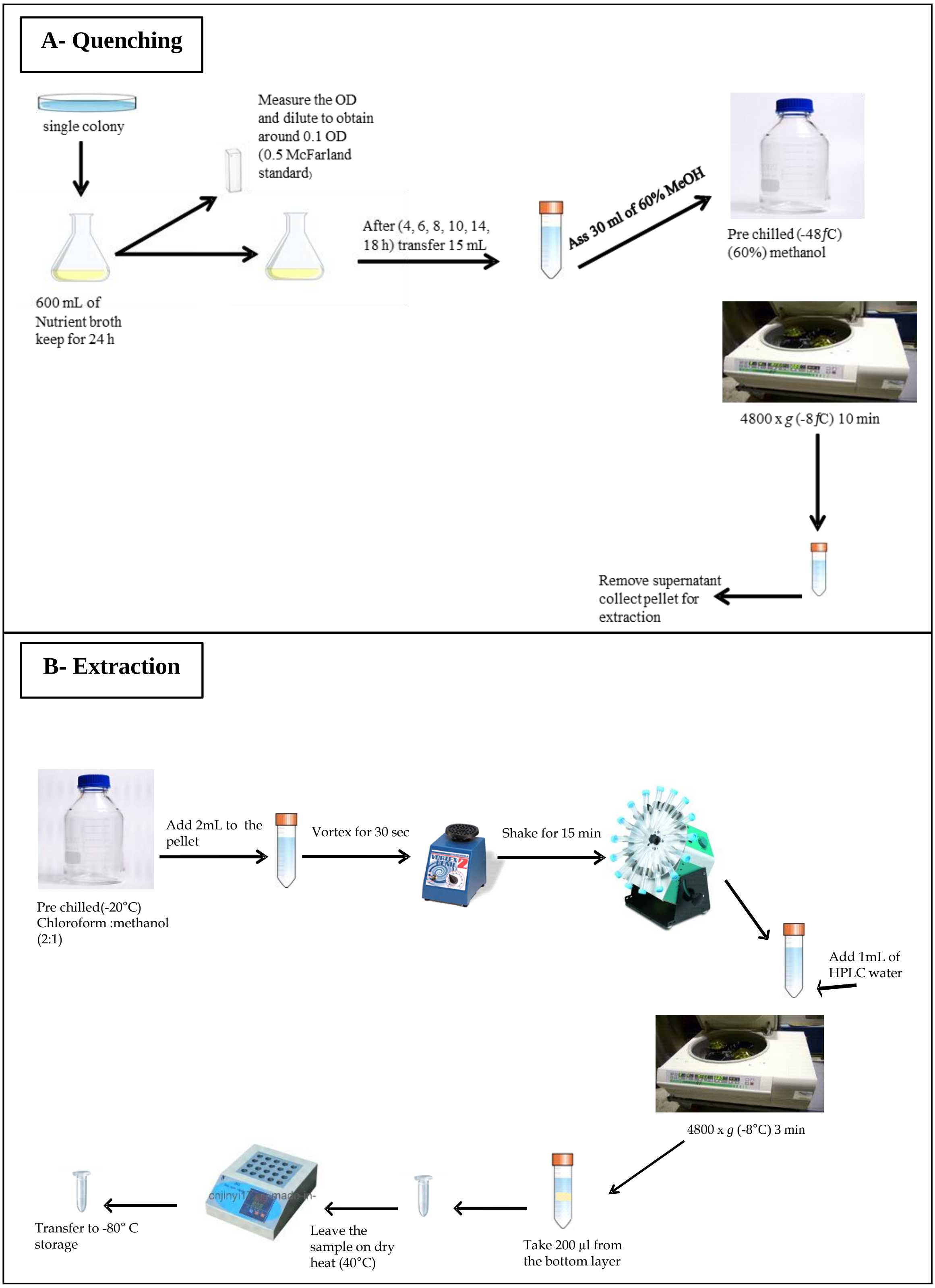


Figure S 4.1: (A) Schematic representing sample quenching using methanol (-48°C). (B) Schematic representing extraction of each sample for UHPLC-MS and MALDI-MS analysis using (2:1) chloroform:methanol.

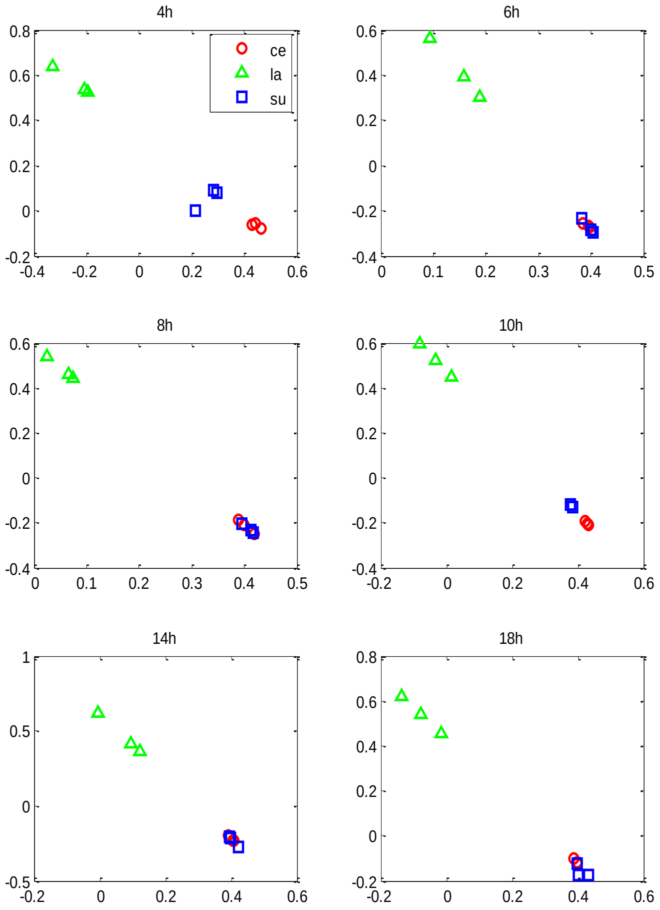


Figure S 4.2: UHPLC-MS Parallel factor analysis (PARAFAC2) at 4, 6, 8, 10, 14 and 18 h for three different species; where *B. cereus*= ce, *B. subtilis*= su and *Br. laterosporus*= la.

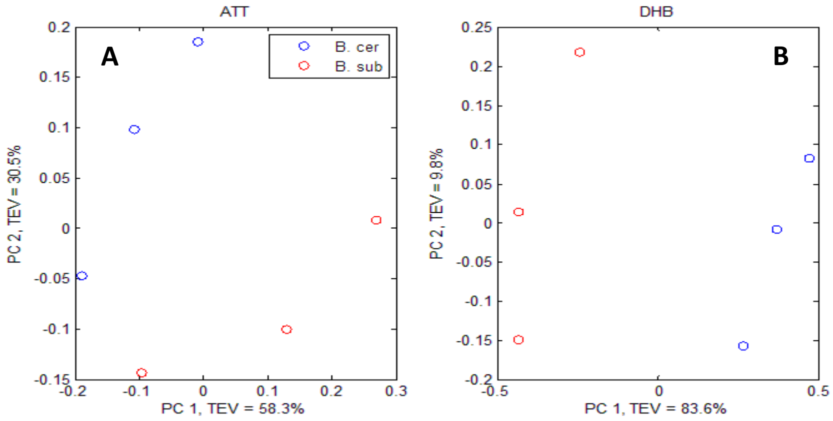
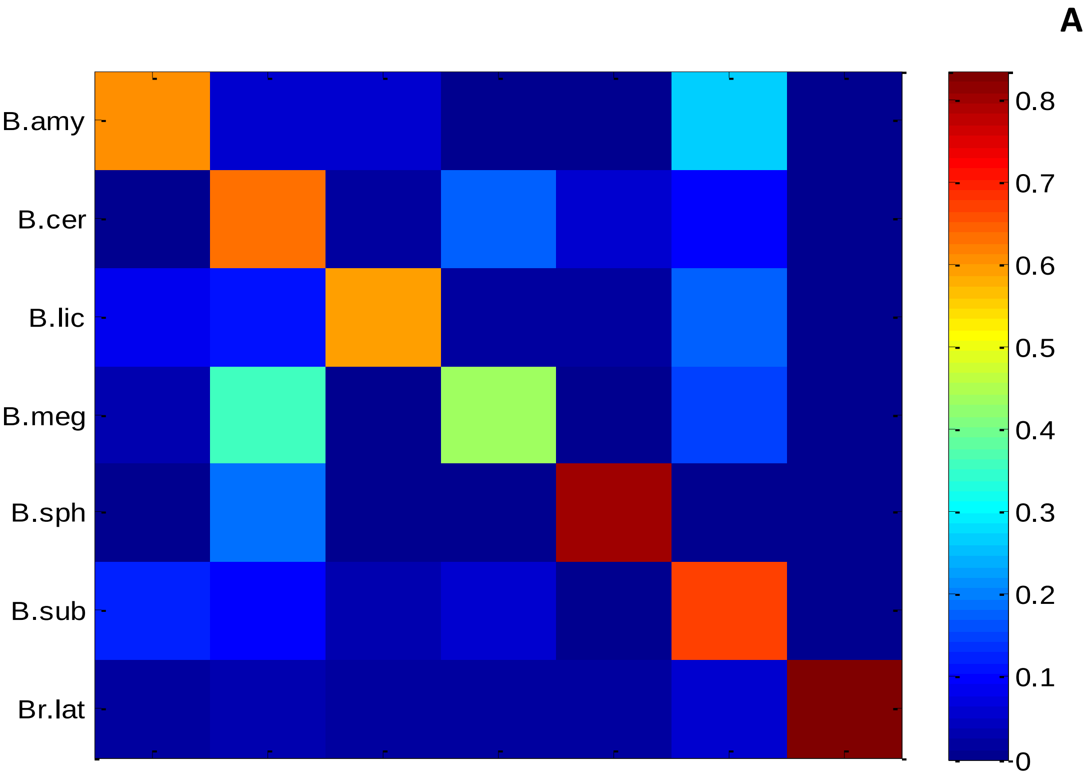


Figure S 4.3: PCA scores plots for two different *Bacillus* species: *B. cereus* and *B. subtilis* using two different matrices: (A) ATT and (B) DHB.



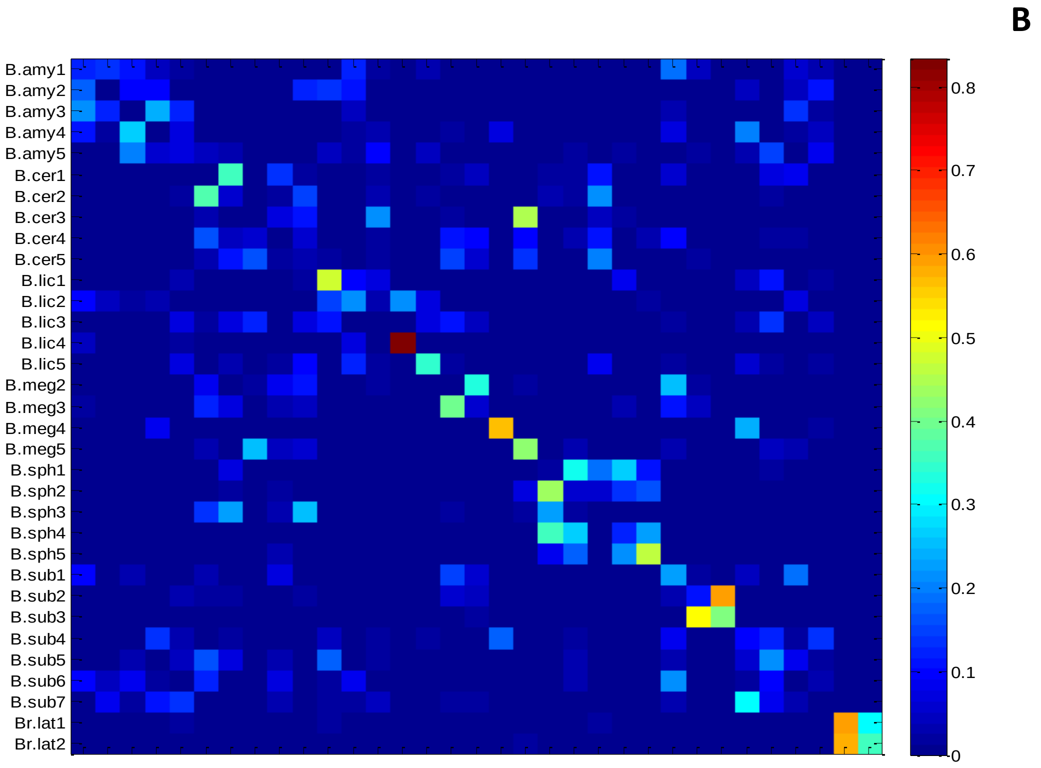
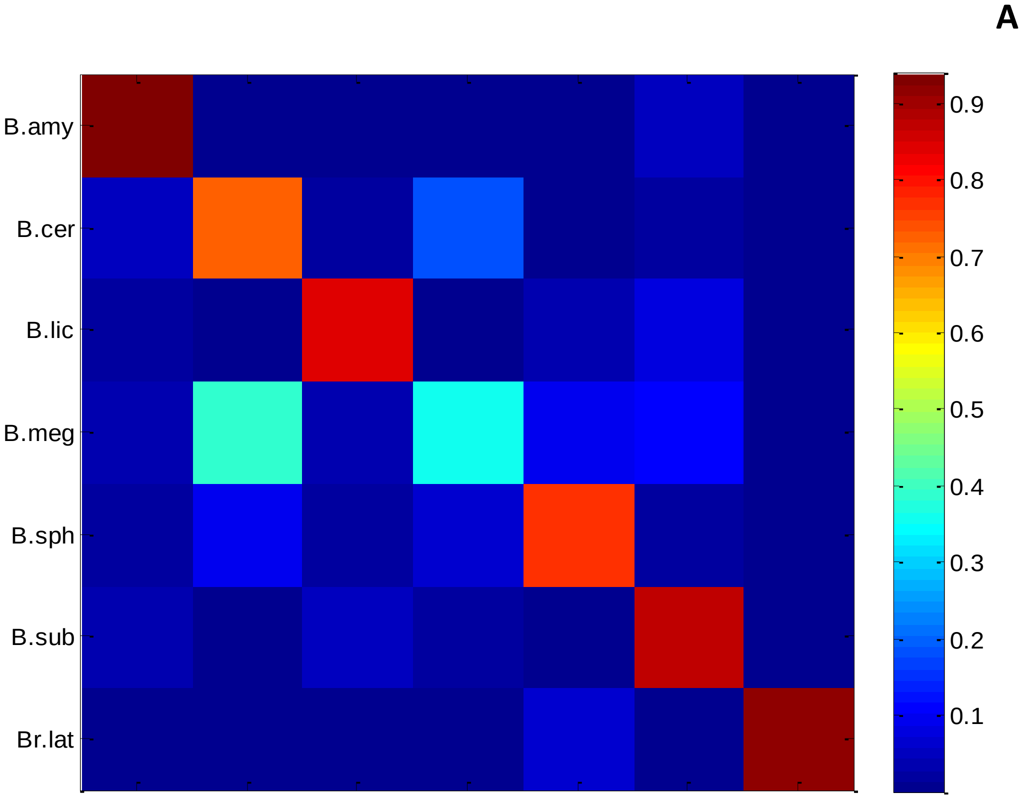


Figure S 4.4: Heat maps of the confusion matrices from: (A) 7 species and (B) 33 strains from *Bacillus* generated from PLS-DA on the MALDI-TOF-MS data.



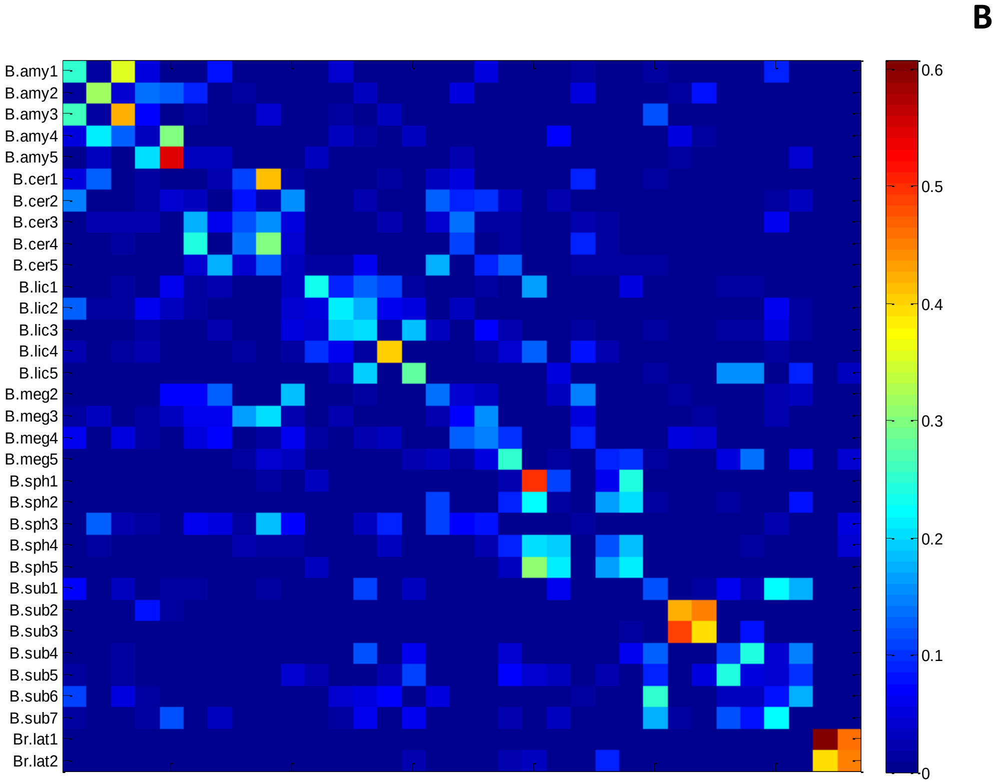


Figure S 4.5: Heat maps of the confusion matrices from: (A) 7 species and (B) 33 strains from *Bacillus* were generated from PLS-DA on the LC-MS data.

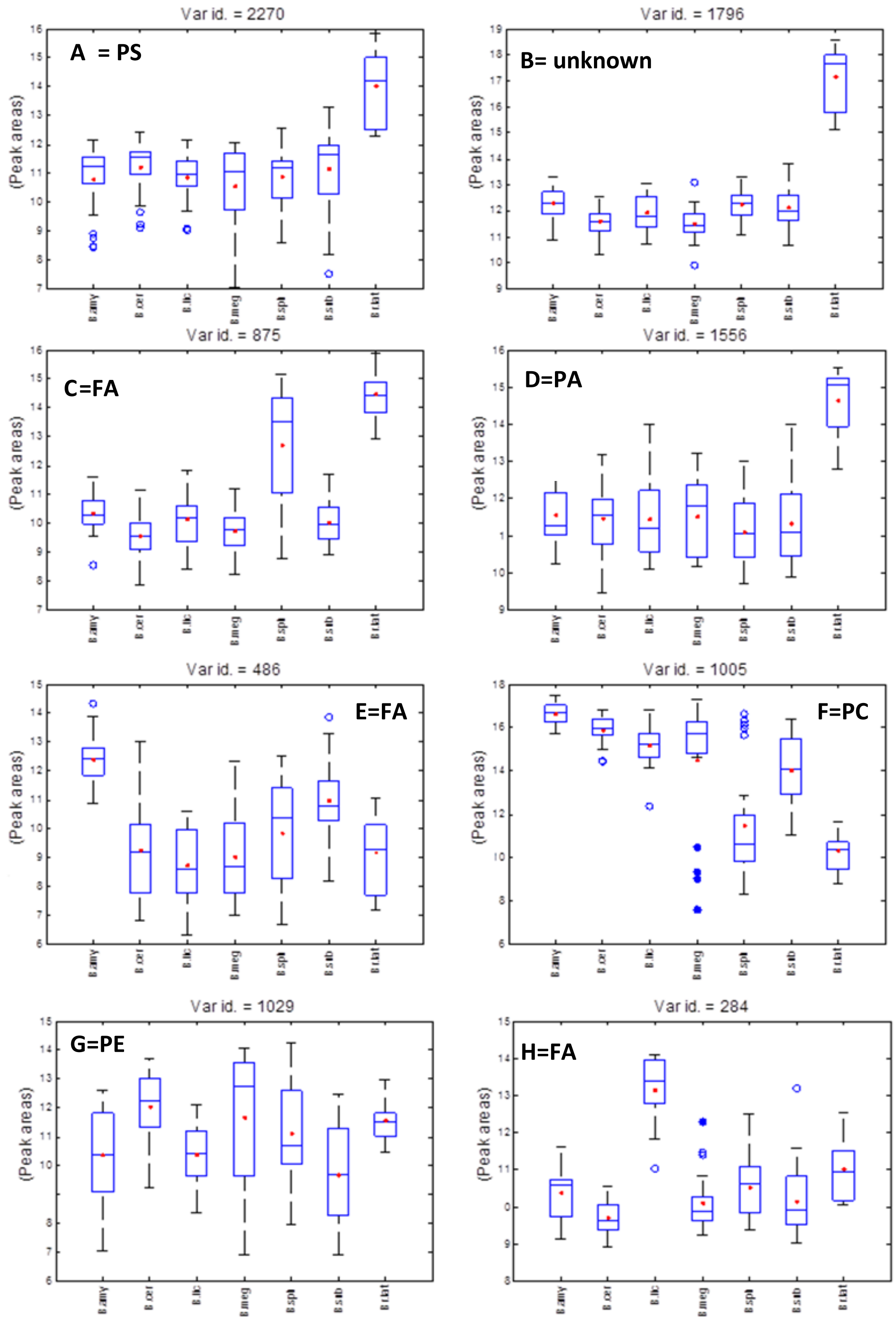


Figure S 4.6: Box-whisker plots for the seven species from *Bacillus* representing the relative concentration levels of different lipids (A-H). Each box plot indicates a different type of lipid; for more details see attached Excel sheet (Lipids-SI-TabsS3-S5.xls). X-axis coding: *B. amy: B. amyloliquefaciens*, *B. cer: B. cereus*, *Br. lat: Br. laterosporus*, *B. lic: B. licheniformis*, *B. meg: B. megaterium*, *B. sph: B. sphaericus* and *B. sub: B. subtilis*. For more information see Table S4.9 (enclosed material: Chapter 4_SI). Phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidic acid (PA) and free fatty acid (FFA).

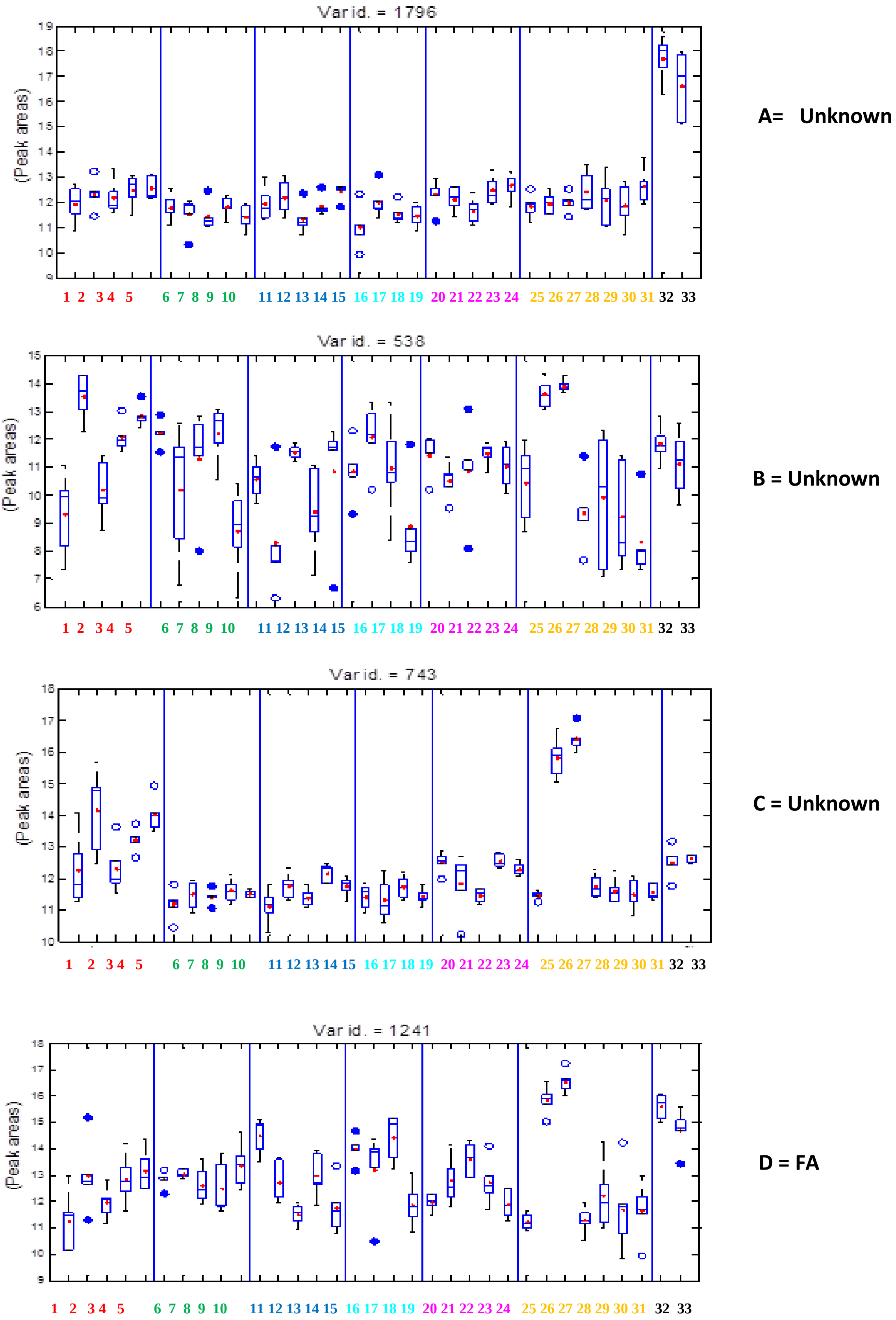


Figure S 4.7 A-D: Box-whisker plots (A-D) for the 33 strains representing the concentration levels of different lipids. Each box plot indicates a different type of lipids; for more details see Table S4.10 (see enclosed material: Chapter 4_SI). Each colour represents different strains as indicated in Table 4.1 on the manuscript. FA, fatty acids.

Table S 4.1: Normalised sample for the first biological replicates reconstitution volumes used prior to LC-MS analysis

Sample number	Strain no.	Absorbance at OD 600 nm (average of 3 technical replicates)	Class	Normalised reconstitution volume (µL)*
1	B0056	2.53	1	361.43
2	B0057	1.57	1	224.29
3	B0076	1.74	1	248.57
4	B0621	1.75	1	250.00
5	B0002^T	3.57	2	510.71
6	B0550	2.56	2	366.43
7	B0702	3.32	2	475.00
8	B0712	3.30	2	472.14
9	B0851	2.76	2	394.29
10	B0014^T	1.66	3	237.86
11	B0044	1.84	3	263.57
12	B0098	2.11	3	301.43
13	B0099	1.42	3	202.86
14	B0410	1.94	3	277.14
15	B0501	2.07	3	295.71
16	B1382	1.52	3	217.86
17	B0177^T	1.39	3	198.57
18	B0168	1.76	4	252.14
19	B0175	1.36	4	194.29
20	B0251	1.53	4	219.29
21	B0620	1.23	4	176.43
22	B0252^T	2.5	5	357.14
23	B0242	2.16	5	309.29

Table S 4.1: Continued

24	B0755	1.99	5	285.00
25	B1081	1.84	5	262.86
26	B1379	1.64	5	235.00
27	7134^{T}	1.00	6	143.57
28	B0408	1.38	6	197.14
29	B0219	3.1	6	442.86
30	B0769	0.95	6	136.43
31	B1147	0.86	6	123.57
32	B0043	0.88	7	126.43
33	B0262	1.38	7	197.86

*(Minimum volume (100 µL)/Minimum OD (0.70)) × Sample OD

Table S 4.2: Normalised sample for the second biological replicates reconstitution volumes used prior to LC-MS analysis

Sample number	Strain no.	Absorbance at OD 600 nm (average of 3 technical replicates)	Class	Normalised reconstitution volume (µL)*
1	B0056	2.48	1	354.29
2	B0057	1.83	1	261.43
3	B0076	1.54	1	220.71
4	B0621	1.83	1	261.43
5	B0002^T	3.23	2	461.43
6	B0550	2.47	2	353.57
7	B0702	3.24	2	462.86
8	B0712	3.16	2	451.43
9	B0851	2.54	2	362.86
10	B0014^T	1.47	3	210.00
11	B0044	2.02	3	289.29
12	B0098	1.8	3	257.14
13	B0099	1.21	3	173.57
14	B0410	1.47	3	210.71
15	B0501	1.87	3	267.14
16	B1382	1.30	3	186.43
17	B0177^T	1.235	3	176.43
18	B0168	1.43	4	204.29
19	B0175	1.28	4	182.86
20	B0251	1.27	4	182.14
21	B0620	0.98	4	140.71
22	B0252^T	2.27	5	324.29
23	B0242	2.05	5	292.86

Table S 4.2: Continued

24	B0755	1.97	5	282.14
25	B1081	1.44	5	205.71
26	B1379	1.65	5	235.71
27	7134^{T}	1	6	142.86
28	B0408	0.76	6	108.57
29	B0219	2.71	6	387.14
30	B0769	0.81	6	116.43
31	B1147	0.91	6	130.00
32	B0043*	1.28	7	183.57
33	B0262	0.85	7	122.14

*(Minimum volume (100 µL)/Minimum OD (0.70)) × Sample OD

Table S 4.3: Normalised sample for the third biological replicates reconstitution volumes used prior to LC-MS analysis

Sample number	Strain no.	Absorbance at OD 600 nm (average of 3 technical replicates)	Class	Normalised reconstitution volume (µL)*
1	B0056	2.36	1	337.14
2	B0057	1.34	1	192.14
3	B0076	1.34	1	192.14
4	B0621	1.84	1	263.57
5	B0002^T	3.26	2	466.43
6	B0550	2.39	2	342.14
7	B0702	3.025	2	432.14
8	B0712	2.98	2	425.71
9	B0851	2.45	2	350.71
10	B0014^T	1.76	3	252.14
11	B0044	1.77	3	253.57
12	B0098	1.59	3	227.86
13	B0099	1.24	3	177.14
14	B0410	0.70	3	100.00
15	B0501	1.82	3	260.00
16	B1382	1.35	3	193.57
17	B0177^T	1.17	3	167.86
18	B0168	1.43	4	204.29
19	B0175	1.17	4	167.14
20	B0251	1.19	4	170.00
21	B0620	0.99	4	141.43
22	B0252^T	2.13	5	304.29
23	B0242	1.95	5	278.57

Table S 4.3: Continued

24	B0755	2.0	5	287.86
25	B1081	1.46	5	208.57
26	B1379	1.25	5	178.57
27	7134^{T}	1.01	6	144.29
28	B0408	1.10	6	157.86
29	B0219	2.47	6	352.86
30	B0769	0.84	6	120.00
31	B1147	0.84	6	120.00
32	B0043	1.23	7	176.43
33	B0262	1.26	7	180.71

*(Minimum volume (100 µL)/Minimum OD (0.70)) × Sample OD

Table S 4.4: Normalised sample for the fourth biological replicates reconstitution volumes used prior to LC-MS analysis

Sample number	**Strain no.**	**Absorbance at OD 600 nm (average of 3 technical replicates)**	**Class**	**Normalised reconstitution volume (µL)***
1	B0056	3.30	1	471.43
2	B0057	3.35	1	478.57
3	B0076	2.81	1	401.43
4	B0621	1.80	1	257.14
5	B0002^T	3.41	2	487.14
6	B0550*	3.21	2	458.57
7	B0702	3.56	2	508.57
8	B0712	3.46	2	494.29
9	B0851	3.29	2	470.00
10	B0014^T	1.98	3	282.86
11	B0044	2.12	3	302.86
12	B0098*	2.16	3	308.57
13	B0099	1.38	3	197.14
14	B0410	1.36	3	194.29
15	B0501	2.20	3	314.29
16	B1382	1.345	3	192.14
17	B0177^T	1.66	3	237.14
18	B0168	1.86	4	265.71
19	B0175	1.60	4	228.57
20	B0251	1.54	4	220.00
21	B0620	0.96	4	137.14
22	B0252^T	2.42	5	345.71
23	B0242	2.19	5	312.86

Table S 4.4: Continued

24	B0755	2.24	5	320.00
25	B1081	2.11	5	301.43
26	B1379	1.61	5	230.00
27	7134^{T}	0.95	6	135.71
28	B0408	0.83	6	118.57
29	B0219	3.69	6	527.14
30	B0769	1.28	6	182.86
31	B1147	1.12	6	160.00
32	B0043*	1.67	7	238.57
33	B0262	1.98	7	282.86

*(Minimum volume (100 µL)/Minimum OD (0.70)) × Sample OD

Table S 4.5: Normalised sample for the fifth biological replicates reconstitution volumes used prior to LC-MS analysis

Sample number	**Strain no.**	**Absorbance at OD 600 nm (average of 3 technical replicates)**	**Class**	**Normalised reconstitution volume (µL)***
1	B0056	4.37	1	624.29
2	B0057	3.54	1	505.71
3	B0076	3.64	1	520.00
4	B0621	2.05	1	292.86
5	B0002^T	3.52	2	502.86
6	B0550*	3.99	2	570.00
7	B0702	3.86	2	551.43
8	B0712	3.66	2	522.86
9	B0851	3.69	2	527.14
10	B0014^T	2.05	3	292.86
11	B0044	2.16	3	308.57
12	B0098*	2.41	3	344.29
13	B0099	1.37	3	195.71
14	B0410	1.32	3	188.57
15	B0501	2.58	3	368.57
16	B1382	1.57	3	224.29
17	B0177^T	1.83	3	261.43
18	B0168	1.68	4	240.00
19	B0175	1.88	4	268.57
20	B0251	1.54	4	220.00
21	B0620	0.82	4	117.14
22	B0252^T	2.55	5	364.29
23	B0242	2.15	5	307.14

Table S 4.5: Continued

24	B0755	2.20	5	314.29
25	B1081	2.45	5	350.00
26	B1379	1.66	5	237.14
27	7134^{T}	1.58	6	225.71
28	B0408	1.75	6	250.00
29	B0219	4.04	6	577.14
30	B0769	1.79	6	255.71
31	B1147	1.07	6	152.86
32	B0043*	1.78	7	254.29
33	B0262	1.24	7	177.14

*(Minimum volume (100 µL)/Minimum OD (0.70)) × Sample OD

Table S 4.6: Prediction accuracies of the 33 *Bacillus* strains from MALDI-TOF-MS data using PLS-DA

	B.amy1	B.amy2	B.amy3	B.amy4	B.amy5	B.cer1	B.cer2	B.cer3	B.cer4	B.cer5	B.lic1	B.lic2	B.lic3	B.lic4	B.lic5	B.meg2	B.meg3	B.meg4	B.meg5	B.sph1	B.sph2	B.sph3	B.sph4	B.sph5	B.sub1	B.sub2	B.sub3	B.sub4	B.sub5	B.sub6	B.sub7	Br.lat1	Br.lat2
B.amy1	12.59%	14.04%	11.68%	4.92%	1.70%	0.02%	0.00%	0.13%	0.20%	0.37%	0.00%	12.21%	1.74%	0.11%	3.44%	0.20%	1.30%	0.00%	0.00%	0.00%	0.02%	0.22%	1.01%	0.00%	18.96%	4.39%	0.00%	0.75%	0.22%	6.22%	3.42%	0.00%	0.00%
B.amy2	17.43%	0.00%	9.30%	9.14%	1.30%	0.04%	0.65%	0.07%	0.00%	11.74%	14.09%	10.64%	0.00%	0.63%	0.00%	0.25%	0.09%	0.02%	0.90%	0.00%	0.29%	0.00%	0.20%	0.00%	0.04%	0.72%	1.28%	5.20%	0.00%	4.68%	11.18%	0.00%	0.00%
B.amy3	21.59%	12.47%	0.04%	24.63%	12.32%	1.02%	0.02%	0.00%	0.29%	0.00%	0.09%	5.13%	0.18%	0.04%	0.82%	0.00%	0.02%	0.00%	0.11%	0.00%	0.51%	0.60%	0.16%	0.04%	3.33%	0.00%	0.20%	0.00%	1.22%	13.14%	1.89%	0.00%	0.00%
B.amy4	11.40%	2.35%	26.13%	0.00%	6.66%	0.00%	0.70%	0.07%	0.07%	0.46%	0.15%	2.39%	2.79%	0.09%	0.28%	1.31%	0.04%	7.68%	0.24%	0.02%	0.02%	1.17%	0.04%	0.04%	7.77%	0.02%	0.20%	20.76%	0.00%	1.96%	5.00%	0.02%	0.04%
B.amy5	0.27%	0.65%	19.96%	5.26%	7.74%	4.68%	3.33%	0.00%	0.76%	0.43%	5.15%	2.01%	9.62%	0.00%	3.96%	0.38%	0.54%	0.00%	0.60%	0.00%	1.79%	0.60%	1.41%	0.22%	0.47%	2.44%	0.00%	3.80%	15.19%	0.07%	8.55%	0.00%	0.00%
B.cer1	0.00%	0.00%	0.27%	0.00%	0.02%	0.09%	35.88%	0.79%	13.67%	2.57%	0.00%	0.00%	1.96%	0.00%	0.32%	1.78%	5.07%	0.00%	0.83%	1.85%	2.01%	11.05%	0.18%	0.00%	5.68%	0.14%	0.00%	0.05%	7.31%	8.30%	0.05%	0.00%	0.00%
B.cer2	0.02%	0.00%	0.00%	0.00%	1.55%	36.76%	5.64%	0.00%	2.52%	14.48%	0.00%	0.00%	3.52%	0.00%	1.57%	0.20%	0.97%	0.00%	0.77%	3.50%	2.01%	21.05%	0.20%	0.31%	0.66%	0.58%	0.00%	0.49%	2.55%	0.51%	0.00%	0.00%	0.00%
B.cer3	0.00%	0.00%	0.00%	0.00%	0.00%	2.73%	0.89%	1.10%	6.99%	10.80%	0.00%	0.00%	21.96%	0.00%	0.00%	1.38%	0.11%	0.34%	45.55%	0.00%	0.91%	4.51%	1.31%	0.00%	0.00%	0.00%	0.00%	0.53%	0.32%	0.08%	0.19%	0.00%	0.17%
B.cer4	0.00%	0.00%	0.28%	0.00%	1.20%	16.56%	3.98%	6.20%	0.00%	6.14%	0.04%	0.00%	1.62%	0.00%	0.20%	10.83%	10.22%	0.02%	10.00%	0.07%	3.71%	10.62%	0.87%	3.23%	9.17%	0.48%	0.00%	0.13%	1.92%	1.97%	0.39%	0.00%	0.00%
B.cer5	0.11%	0.05%	0.00%	0.16%	0.02%	2.62%	10.89%	16.37%	2.03%	3.86%	2.41%	0.05%	1.56%	0.00%	0.79%	15.36%	5.82%	0.52%	13.37%	0.18%	0.05%	20.28%	0.38%	0.02%	0.09%	1.35%	0.02%	0.18%	0.00%	0.77%	0.56%	0.00%	0.00%
B.lic1	0.00%	0.16%	0.83%	0.20%	2.82%	0.07%	0.13%	0.00%	0.07%	1.45%	47.61%	10.20%	7.32%	0.00%	0.00%	0.40%	0.31%	0.09%	0.09%	0.11%	0.20%	0.65%	8.01%	1.07%	0.00%	0.02%	0.04%	4.34%	11.43%	0.74%	1.48%	0.02%	0.00%
B.lic2	9.42%	4.90%	2.00%	2.68%	1.26%	0.00%	0.00%	0.00%	0.09%	0.00%	15.10%	21.45%	3.09%	21.52%	7.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.26%	0.00%	1.31%	0.04%	0.00%	0.00%	0.39%	0.07%	7.51%	1.28%	0.00%	0.00%
B.lic3	0.79%	0.24%	0.07%	0.77%	7.02%	1.34%	6.93%	12.28%	0.31%	6.84%	11.45%	1.29%	1.10%	0.00%	6.80%	10.57%	4.78%	0.02%	0.55%	0.07%	0.24%	0.46%	0.90%	0.83%	1.43%	1.12%	0.07%	3.60%	13.53%	0.18%	4.28%	0.04%	0.00%
B.lic4	4.97%	0.57%	0.07%	0.02%	1.86%	0.00%	0.00%	0.00%	0.09%	0.00%	1.24%	7.21%	0.00%	83.40%	0.00%	0.02%	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.35%	0.00%	0.00%	0.00%
B.lic5	1.17%	0.00%	0.98%	0.04%	6.96%	1.13%	2.92%	0.17%	1.70%	9.20%	0.43%	13.03%	1.96%	0.02%	34.08%	2.38%	0.36%	0.09%	0.51%	0.09%	0.13%	9.11%	0.38%	0.02%	1.40%	0.30%	0.38%	6.51%	2.06%	0.32%	2.00%	0.04%	0.00%
B.meg2	0.55%	0.13%	0.20%	0.04%	0.18%	7.96%	0.51%	1.73%	8.05%	11.64%	0.00%	0.04%	1.55%	0.00%	0.38%	0.29%	33.13%	0.04%	2.02%	0.00%	0.00%	1.09%	1.13%	0.11%	25.66%	2.59%	0.16%	0.16%	0.00%	0.47%	0.02%	0.00%	0.04%
B.meg3	1.91%	0.46%	0.09%	0.00%	0.37%	11.93%	6.74%	0.00%	3.30%	4.87%	0.65%	0.00%	1.09%	0.00%	0.39%	39.81%	6.37%	0.00%	0.09%	0.39%	0.00%	0.17%	3.87%	0.89%	11.17%	3.98%	0.04%	0.15%	0.67%	0.24%	0.00%	0.15%	0.09%
B.meg4	0.48%	0.22%	0.33%	8.94%	0.18%	0.31%	0.24%	0.13%	0.18%	0.51%	0.09%	0.62%	0.33%	0.00%	0.07%	0.11%	0.84%	57.20%	0.22%	0.18%	0.33%	0.22%	0.18%	0.07%	0.11%	0.15%	0.00%	23.98%	1.21%	0.00%	2.47%	0.00%	0.00%
B.meg5	0.04%	0.00%	0.11%	0.00%	0.20%	2.95%	0.42%	25.21%	4.66%	5.51%	0.00%	0.00%	0.09%	0.00%	0.11%	1.18%	0.80%	0.60%	42.39%	0.00%	2.88%	0.29%	0.09%	0.22%	2.86%	0.07%	0.00%	0.54%	4.98%	3.48%	0.18%	0.00%	0.00%
B.sph1	0.00%	0.00%	0.00%	0.00%	0.00%	0.86%	7.42%	0.00%	0.00%	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	1.65%	32.06%	18.74%	26.20%	11.18%	0.00%	0.02%	0.00%	0.02%	1.56%	0.00%	0.00%	0.00%	0.04%
B.sph2	0.00%	0.00%	0.00%	0.16%	0.00%	0.78%	2.56%	0.62%	1.43%	0.09%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.40%	44.07%	5.64%	5.26%	13.97%	16.24%	0.00%	0.00%	0.00%	0.02%	0.76%	0.80%	0.02%	0.00%	0.00%
B.sph3	0.00%	0.00%	0.11%	0.00%	0.20%	13.38%	23.09%	0.78%	3.68%	25.62%	0.63%	0.04%	0.00%	0.00%	0.26%	1.70%	0.04%	0.15%	2.00%	22.51%	1.61%	1.13%	0.22%	1.13%	0.00%	0.00%	0.00%	0.00%	0.52%	0.87%	0.13%	0.00%	0.04%
B.sph4	0.09%	0.00%	0.04%	0.00%	0.02%	0.00%	0.00%	0.00%	0.56%	1.14%	0.04%	0.06%	0.00%	0.00%	0.00%	0.67%	0.00%	0.00%	0.09%	35.25%	26.22%	0.00%	12.41%	23.10%	0.00%	0.00%	0.00%	0.00%	0.17%	0.02%	0.00%	0.00%	0.00%
B.sph5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	3.50%	0.40%	0.54%	0.04%	0.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.40%	8.66%	17.83%	0.25%	21.95%	46.10%	0.00%	0.02%	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%
B.sub1	9.26%	0.00%	3.66%	0.34%	0.06%	3.75%	0.82%	0.00%	6.57%	0.17%	0.00%	0.19%	1.03%	0.00%	0.93%	14.92%	5.99%	0.00%	1.12%	0.00%	0.06%	0.06%	0.04%	0.00%	23.21%	2.15%	0.15%	4.63%	0.84%	19.32%	0.58%	0.00%	0.00%
B.sub2	0.80%	0.00%	0.00%	0.82%	2.91%	2.55%	2.40%	0.00%	0.28%	1.84%	0.00%	0.00%	1.25%	0.00%	0.17%	6.15%	5.14%	0.00%	0.00%	0.00%	0.00%	0.45%	0.00%	0.00%	3.52%	10.43%	59.89%	0.84%	0.04%	0.30%	0.09%	0.00%	0.00%
B.sub3	0.13%	0.02%	0.00%	0.00%	0.07%	0.00%	0.13%	0.00%	0.51%	0.70%	0.00%	0.00%	0.53%	0.00%	0.04%	1.23%	2.40%	0.00%	0.00%	0.00%	0.00%	0.04%	0.04%	0.00%	0.84%	51.50%	41.28%	0.33%	0.00%	0.00%	0.07%	0.00%	0.00%
B.sub4	1.28%	1.09%	0.22%	14.28%	3.71%	0.02%	1.84%	0.11%	0.15%	0.59%	4.88%	0.48%	2.13%	0.00%	1.80%	0.00%	0.00%	17.19%	0.82%	0.02%	2.04%	0.13%	0.00%	0.04%	8.03%	0.76%	0.00%	9.55%	12.78%	1.82%	14.11%	0.00%	0.00%
B.sub5	0.13%	0.00%	2.67%	0.04%	4.72%	16.31%	6.53%	0.00%	3.18%	0.02%	18.17%	0.20%	1.37%	0.00%	0.46%	0.00%	0.02%	0.00%	0.26%	0.79%	3.25%	0.02%	0.00%	0.15%	3.00%	0.11%	0.55%	6.36%	21.24%	8.59%	1.70%	0.00%	0.00%
B.sub6	9.84%	4.96%	9.09%	1.55%	0.62%	12.15%	0.29%	0.02%	7.71%	0.64%	2.55%	8.31%	0.00%	0.16%	0.18%	0.04%	0.00%	0.00%	0.62%	0.04%	2.88%	0.31%	0.00%	0.00%	21.21%	0.27%	0.04%	2.59%	9.22%	0.84%	3.72%	0.00%	0.00%
B.sub7	0.71%	7.86%	2.35%	11.46%	13.26%	0.00%	0.02%	0.69%	3.40%	0.47%	1.49%	1.68%	4.63%	0.00%	0.30%	1.36%	1.87%	1.03%	1.01%	0.00%	0.26%	0.15%	0.00%	0.09%	2.76%	0.11%	0.00%	30.90%	8.10%	3.25%	0.65%	0.00%	0.02%
Br.lat1	0.11%	0.00%	0.46%	0.00%	1.48%	0.04%	0.02%	0.00%	0.04%	0.17%	1.35%	1.11%	0.02%	0.00%	0.00%	0.00%	1.20%	0.00%	0.00%	0.00%	0.33%	2.17%	0.04%	0.00%	0.00%	0.67%	0.00%	0.02%	0.26%	0.61%	0.00%	59.10%	30.66%
Br.lat2	0.00%	0.00%	0.09%	0.00%	0.73%	0.09%	0.00%	0.00%	0.11%	0.02%	0.33%	0.07%	0.00%	0.00%	0.44%	0.18%	0.11%	0.00%	1.51%	0.02%	0.00%	0.86%	0.00%	0.33%	0.00%	0.16%	0.00%	0.09%	0.95%	0.24%	0.00%	57.40%	36.13%

The different colours represent the species level identification

Table S 4.7: Prediction accuracies of the 33 *Bacillus* strains from LC-MS data using PLS-DA

	B.amy1	B.amy2	B.amy3	B.amy4	B.amy5	B.cer1	B.cer2	B.cer3	B.cer4	B.cer5	B.lic1	B.lic2	B.lic3	B.lic4	B.lic5	B.meg2	B.meg3	B.meg4	B.meg5	B.sph1	B.sph2	B.sph3	B.sph4	B.sph5	B.sub1	B.sub2	B.sub3	B.sub4	B.sub5	B.sub6	B.sub7	Br.lat1	Br.lat2
B.amy1	24.86%	1.30%	36.05%	5.64%	0.13%	0.42%	8.08%	0.04%	0.00%	0.00%	0.00%	3.97%	0.06%	0.77%	0.00%	0.00%	0.24%	5.07%	0.00%	0.00%	0.00%	1.02%	0.00%	0.00%	1.72%	0.00%	0.00%	0.00%	0.16%	9.06%	0.23%	0.00%	0.00%
B.amy2	1.40%	31.88%	3.91%	14.09%	12.60%	8.61%	0.00%	1.67%	0.15%	0.00%	0.00%	0.81%	3.63%	0.09%	0.00%	0.53%	4.96%	0.00%	0.24%	0.00%	0.00%	5.18%	0.33%	0.33%	0.00%	1.33%	8.01%	0.00%	0.00%	0.09%	0.00%	0.00%	0.00%
B.amy3	25.97%	1.77%	41.97%	7.06%	0.07%	1.04%	0.00%	0.03%	4.01%	0.08%	0.04%	1.12%	0.04%	2.89%	0.00%	0.00%	0.09%	0.68%	0.00%	0.00%	0.00%	0.46%	0.00%	0.00%	11.71%	0.00%	0.06%	0.04%	0.00%	0.44%	0.00%	0.00%	0.00%
B.amy4	5.32%	21.16%	12.60%	3.26%	29.96%	0.44%	0.58%	0.11%	0.00%	0.93%	0.19%	3.45%	1.67%	0.65%	2.93%	0.00%	0.00%	0.11%	0.00%	0.06%	6.79%	0.31%	0.31%	0.11%	0.06%	5.63%	1.74%	0.49%	0.00%	0.25%	0.45%	0.00%	0.00%
B.amy5	0.55%	3.62%	0.18%	20.04%	54.57%	3.19%	3.07%	0.00%	0.07%	0.00%	2.88%	0.19%	0.00%	0.00%	0.00%	0.07%	2.68%	0.29%	0.28%	0.00%	0.03%	0.00%	0.00%	0.14%	0.09%	1.40%	0.22%	0.92%	0.07%	0.00%	4.59%	0.00%	0.00%
B.cer1	5.06%	12.92%	0.90%	1.12%	0.84%	0.22%	2.33%	11.25%	41.13%	1.67%	0.86%	0.08%	0.04%	1.45%	0.06%	3.66%	5.16%	0.18%	0.11%	0.00%	0.23%	8.69%	0.06%	0.09%	1.34%	0.44%	0.00%	0.12%	0.00%	0.00%	0.00%	0.00%	0.00%
B.cer2	14.74%	0.07%	0.00%	1.64%	3.91%	2.96%	0.50%	8.30%	2.50%	16.00%	0.39%	0.24%	2.30%	0.56%	0.74%	13.18%	8.83%	10.04%	3.08%	0.00%	2.09%	0.85%	0.69%	0.00%	0.39%	0.47%	0.06%	0.00%	0.00%	1.07%	3.75%	0.00%	0.00%
B.cer3	0.24%	2.50%	2.72%	2.07%	0.17%	17.49%	6.08%	11.73%	15.80%	4.76%	0.37%	0.47%	0.20%	2.25%	0.00%	3.99%	13.52%	1.80%	1.54%	0.00%	0.06%	2.70%	0.98%	0.23%	0.04%	0.16%	0.15%	0.00%	0.09%	6.56%	0.00%	0.00%	0.00%
B.cer4	0.06%	0.35%	1.12%	0.00%	0.06%	24.22%	0.20%	14.06%	30.02%	4.57%	0.06%	0.00%	0.03%	0.00%	0.00%	0.04%	10.92%	0.48%	1.70%	0.00%	0.00%	8.55%	1.33%	0.71%	0.86%	0.44%	0.00%	0.00%	0.07%	0.15%	0.00%	0.00%	0.00%
B.cer5	0.05%	0.00%	0.63%	0.15%	0.00%	3.86%	17.76%	4.23%	13.02%	3.25%	1.22%	1.81%	5.70%	0.17%	0.00%	17.83%	0.79%	9.50%	12.43%	0.11%	0.22%	1.79%	1.34%	1.12%	1.34%	0.80%	0.00%	0.13%	0.25%	0.56%	0.00%	0.00%	0.00%
B.lic1	0.38%	0.03%	0.96%	0.09%	6.40%	1.72%	2.16%	0.09%	0.00%	3.51%	23.06%	8.75%	12.73%	11.36%	1.16%	0.03%	0.00%	1.03%	0.00%	17.04%	0.15%	0.00%	0.09%	5.64%	0.00%	0.00%	0.74%	1.43%	1.41%	0.00%	0.00%	0.00%	0.00%
B.lic2	12.99%	1.70%	1.27%	5.90%	3.33%	1.71%	0.14%	0.04%	0.00%	4.02%	5.66%	21.39%	17.49%	6.03%	5.26%	0.15%	3.41%	0.06%	0.04%	0.00%	0.04%	0.61%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.58%	5.92%	1.74%	0.00%	0.00%
B.lic3	0.68%	0.13%	0.90%	0.99%	0.00%	0.00%	2.82%	0.10%	0.00%	5.02%	4.36%	19.56%	20.45%	1.36%	18.86%	3.05%	0.00%	7.58%	2.32%	0.00%	0.00%	1.22%	0.00%	0.00%	1.69%	0.00%	0.00%	1.85%	1.03%	5.02%	1.37%	0.00%	0.00%
B.lic4	2.52%	0.20%	1.88%	2.31%	0.00%	0.27%	0.09%	1.11%	0.00%	0.96%	10.35%	6.25%	1.54%	40.48%	0.06%	0.00%	0.26%	1.10%	4.49%	13.17%	0.41%	8.50%	1.99%	0.39%	0.00%	0.00%	0.00%	0.00%	0.00%	1.24%	0.00%	0.00%	0.00%
B.lic5	0.66%	0.00%	0.00%	0.14%	0.11%	0.00%	0.00%	0.00%	0.04%	0.00%	0.24%	2.07%	19.71%	0.06%	27.75%	0.58%	0.24%	0.07%	0.11%	0.10%	5.44%	0.21%	0.06%	0.00%	1.32%	0.03%	0.00%	16.00%	15.87%	0.06%	8.64%	0.00%	2.94%
B.meg2	0.00%	0.53%	0.00%	0.25%	7.51%	6.67%	13.26%	0.93%	0.55%	18.96%	0.04%	0.13%	1.76%	0.22%	0.46%	13.69%	4.62%	3.25%	0.07%	0.00%	3.49%	15.18%	0.00%	0.00%	0.00%	1.10%	0.40%	0.00%	0.32%	2.55%	3.19%	0.00%	0.00%
B.meg3	1.28%	3.64%	0.53%	1.55%	3.10%	6.10%	6.61%	16.50%	20.46%	2.48%	0.34%	2.61%	0.03%	0.30%	0.11%	2.04%	7.59%	15.23%	0.06%	0.19%	0.06%	5.11%	0.15%	0.00%	0.00%	0.06%	1.05%	0.00%	0.00%	2.63%	0.51%	0.00%	0.00%
B.meg4	6.45%	0.00%	4.76%	1.41%	0.06%	4.77%	7.31%	0.22%	1.31%	6.08%	1.85%	0.70%	2.77%	3.64%	0.16%	0.67%	13.19%	15.17%	9.83%	0.00%	0.00%	9.17%	0.60%	0.00%	0.00%	5.39%	4.42%	0.00%	0.05%	0.00%	0.00%	0.00%	0.00%
B.meg5	0.00%	0.00%	0.00%	0.00%	0.00%	0.48%	0.31%	1.86%	4.69%	3.31%	0.00%	0.06%	0.32%	0.65%	2.76%	2.98%	1.38%	5.27%	25.24%	0.26%	1.74%	0.09%	8.60%	10.10%	1.20%	0.00%	0.44%	5.57%	13.55%	0.60%	6.01%	0.00%	3.92%
B.sph1	0.00%	0.04%	0.82%	0.20%	0.00%	0.00%	0.14%	0.00%	1.41%	0.00%	3.02%	0.00%	0.00%	0.23%	0.00%	0.00%	0.00%	0.00%	2.30%	49.86%	10.99%	0.09%	6.47%	23.80%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
B.sph2	0.04%	0.00%	0.67%	0.94%	0.74%	0.00%	0.06%	0.06%	0.00%	0.26%	0.20%	0.00%	0.00%	0.39%	0.00%	11.05%	0.11%	0.15%	9.03%	22.28%	1.30%	0.33%	16.68%	20.85%	1.23%	0.29%	0.00%	1.29%	0.07%	0.09%	8.03%	0.00%	0.00%
B.sph3	0.11%	12.77%	2.22%	1.00%	0.00%	6.00%	4.89%	1.18%	18.12%	7.59%	0.00%	0.09%	2.91%	9.25%	0.00%	10.81%	7.01%	8.54%	0.09%	0.04%	0.37%	0.95%	0.15%	0.28%	0.07%	0.00%	0.50%	0.00%	0.00%	2.03%	0.00%	0.00%	4.90%
B.sph4	0.11%	0.99%	0.07%	0.86%	0.00%	0.06%	0.00%	2.02%	1.44%	1.22%	0.00%	0.10%	0.06%	2.90%	0.00%	0.06%	0.04%	2.83%	9.49%	20.64%	19.54%	0.04%	11.78%	18.28%	0.06%	0.00%	0.04%	0.00%	1.77%	0.25%	0.00%	0.49%	3.92%
B.sph5	0.00%	0.00%	0.94%	0.00%	0.00%	0.00%	0.07%	0.00%	0.24%	0.00%	3.09%	0.00%	0.06%	0.31%	0.00%	0.00%	0.04%	0.08%	3.34%	30.65%	21.14%	0.30%	16.30%	20.99%	0.13%	0.00%	0.39%	0.72%	0.00%	0.22%	0.80%	0.00%	0.00%
B.sub1	7.51%	0.11%	3.10%	0.08%	1.09%	1.23%	0.69%	0.03%	1.69%	0.50%	0.00%	0.47%	10.72%	0.11%	2.95%	0.00%	0.00%	0.12%	0.87%	0.11%	5.71%	0.27%	0.17%	0.43%	11.94%	0.04%	0.99%	6.00%	2.46%	22.24%	17.32%	0.00%	0.00%
B.sub2	0.03%	0.42%	0.00%	8.46%	1.81%	0.00%	0.59%	0.00%	0.09%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%	0.05%	0.80%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	42.60%	45.11%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
B.sub3	0.00%	0.50%	0.00%	0.59%	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	0.00%	0.00%	0.40%	0.03%	0.46%	0.00%	0.00%	0.08%	0.00%	0.00%	1.07%	0.04%	49.06%	39.47%	0.04%	7.94%	0.00%	0.00%	0.00%	0.00%
B.sub4	0.00%	0.00%	1.60%	0.82%	0.28%	0.00%	0.00%	0.84%	0.00%	0.00%	0.15%	0.12%	12.04%	0.00%	5.95%	0.00%	0.00%	0.00%	4.38%	0.00%	0.10%	0.00%	0.00%	6.28%	12.61%	0.04%	0.00%	10.77%	24.38%	3.87%	14.77%	0.00%	0.00%
B.sub5	1.78%	0.00%	1.07%	0.13%	0.24%	0.04%	0.00%	0.08%	0.25%	3.89%	2.08%	0.44%	0.04%	1.98%	11.05%	0.15%	0.57%	0.10%	7.47%	4.20%	3.49%	0.00%	2.69%	0.24%	9.22%	0.07%	4.88%	24.18%	4.84%	3.90%	10.02%	0.00%	0.00%
B.sub6	11.17%	0.00%	5.36%	1.39%	0.00%	0.06%	0.19%	0.39%	0.00%	0.31%	0.00%	3.93%	5.24%	6.79%	0.05%	4.82%	0.71%	0.09%	0.00%	0.00%	0.00%	1.10%	0.00%	0.09%	25.12%	0.46%	0.00%	3.11%	3.31%	8.18%	17.80%	0.00%	0.00%
B.sub7	1.19%	0.00%	0.11%	1.37%	11.58%	0.00%	2.99%	0.11%	0.00%	0.46%	0.00%	1.37%	5.75%	0.00%	6.03%	0.00%	0.11%	0.17%	2.28%	0.00%	3.13%	0.34%	0.00%	0.63%	17.41%	1.73%	0.00%	11.89%	8.28%	22.18%	0.91%	0.00%	0.00%
Br.lat1	0.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	60.78%	46.08%
Br.lat2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.96%	0.00%	0.00%	0.00%	1.96%	2.94%	0.00%	0.00%	8.82%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	39.22%	45.10%

The different colours represent the species level identification

Table S 4.8: shows the putative lipid categories of *Bacillus* bacteria detected using LC-MS (enclosed sheet: Chapter 4_SI).

Table S 4.9: Excel sheet shows the putative significant lipid categories of 7 species detected using LC-MS (enclosed sheet: Chapter 4_SI).

Table S 4.10: Excel sheet shows the putative significant lipid categories of 33 strains detected using LC-MS (enclosed sheet: Chapter 4_SI).

Chapter Five

Rapid classification of *Enterococcus faecium* strains using phenotypic analytical techniques

Najla AlMasoud,[a] *Yun Xu,*[a] *David Ellis,*[a] *Paul Rooney,*[b] *Jane F. Turton,*[c] *Royston Goodacre* [a #]

[a] *School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.*

[b] *Belfast City Hospital, Belfast, 51 Lisburn Rd, BT9 7AB, UK.*

[c] *National Infection Service, Public Health England, London, NW9 5EQ, UK.*

[#] *Correspondence to Roy Goodacre:* roy.goodacre@manchester.ac.uk

This chapter is a manuscript of an article submitted to *Analyste*.

Yun Xu contributed to data analysis. Paul Rooney provided the bacterial samples studied in this work and confirmed the identity of the bacteria to the species level. Jane Turton provided the pulsed field gel electrophoresis data. David Ellis provided assistance in reporting the findings of this study. Royston Goodacre contributed to this work through supervision and guidance of the study.

Abstract

Recent clinical isolates of glycopeptide resistant enterococci (GRE) were used to compare three rapid phenotyping and analytical techniques. Fourier transform infrared (FT-IR) spectroscopy, Raman spectroscopy and matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) –were used to classify 35 isolates from 12 *Enterococcus faecium* strains, which had been previously analysed by pulsed-field gel electrophoresis (PFGE).The results show that the three analytical techniques provide clear discrimination among enterococci at both the strain and isolate levels. FT-IR and Raman spectroscopic data produced very similar bacterial classification, also reflected in the Procrustes distance between the datasets (0.2125-0.2411, p<0.001); however, FT-IR data provided superior prediction accuracy to Raman data, with correct classification rates (CCR) of 89% and 69% at the strain level, respectively. MALDI-TOF-MS produced slightly different classification of these enterococci, also with high classification accuracy (78%). Classification data from the three analytical techniques were consistent with PFGE data especially in the case of strains identified as unique by PFGE. This study presents phenotypic techniques as a complementary approach to current methods with a potential for high-throughput point-of-care screening enabling rapid and reproducible classification of clinically relevant enterococci.

5.1 Introduction

Enterococcus are a highly significant genus of bacteria, which cause important clinical infections including urinary tract infections (UTIs), endocarditis, meningitis, bacteremia, wound infections, pelvic and intra-abdominal infections amongst others. Some of these Gram-positive cocci were originally classified as *Streptococcus* spp. until genomic analysis by Schleifer and Kilpper-Balz in 1984 demonstrated the requirement for a separate genus classification (Schleifer and Kilpper-Bälz, 1984). This well-known genus is part of the normal intestinal microflora of humans and other animals (Kayser *et al*., 2011). *Enterococcus* are also part of the lactic acid bacteria (LAB) group present in foods, and whilst they are able to spoil fresh meats (Hayes *et al*., 2003), they are important in ripening and development of certain foods (i.e. dairy products), as well as being used as probiotics in humans (Franz *et al*., 2003).

The majority of human clinical isolates of enterococci belong to tow species, *Enterococcus faecalis* and *Enterococcus faecium* (McCracken *et al*., 2013). In addition to their prevalence and pathogenicity, another very important factor associated with enterococci is the high level of antimicrobial resistance, particularly resistance to glycopeptide antibiotics (such as vancomycin, teicoplanin and telavancin); resistant strains are referred to as GRE (glycopeptide-resistant enterococci) (Woodford, 1998; Arias and Murray, 2012).

There is a constant requirement to develop analytical methods for the classification of bacteria, which can be used in clinical diagnostics and food quality control. These methods should ideally be rapid, reproducible, and easy to use and automate, in addition to having high resolution and sensitivity (Altekruse *et al*., 1997). Over a decade ago, it was common to use chemical methods, such as polymerase chain reaction (PCR) for identification of specific DNA sequences and recognition by antibodies via enzyme-linked immunosorbent assay (ELISA), to characterise whole bacteria. Although these techniques are sensitive and specific, they are time-consuming and their use is limited by the complexity of preparation procedures and the requirement for specific primers and antibodies (Engvall, 1977; Yolken, 1980; Ke *et al*., 1999; Reen 1994). Nowadays, modern analytical techniques, such as matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) (Claydon *et al.*, 1996; Quintela-Baluja *et al.*, 2013), Fourier transform infrared (FT-IR) spectroscopy (Goodacre *et al.*, 1998, Helm *et al.*, 1991; Naumann *et al.*, 1991; Burgula *et al.*, 2007) and Raman spectroscopy (Huang *et al.*, 2010; Beekes *et al.*, 2007) are also used for the characterisation of bacteria. High dimensional and information rich datasets are produced from these techniques, which have also directly led to the requirement of robust and reliable chemometrics methods to assist with data deconvolution and in-depth analysis (Ellis *et al.*, 2013). This saw the introduction, acceptance and use of chemometrics, such as discriminant function (DFA) and hierarchical cluster analyses (HCA) (Gutteridge *et al.*, 1985; Davis and Mauer, 2010; López-Díez and Goodacre, 2004).

Previously, MALDI-TOF-MS has shown promising results for bacterial classification and characterisation (AlMasoud *et al.*, 2014 (Chapter 2); Claydon *et al.*, 1996; Lasch *et al.*, 2014). FT-IR and Raman spectroscopy complement each other for bacterial classification; both are robust metabolic fingerprinting techniques and need little sample preparation (Sauer and Kliem, 2010; Ellis and Goodacre, 2006; Marvin *et al.*, 2003; Lay, 2001). FT-IR is used by many researchers since it is not only rapid but also represents a high-throughput and non-destructive method, allowing the analysis of intact bacteria and producing unique, reproducible and distinct biochemical fingerprints (Argyri *et al.*, 2013). Moreover, Raman spectroscopy shares similar advantages to FT-IR and also has the additional advantage of water being a very weak Raman scatter (Smith and Dent, 2013; Ferraro *et al.*, 2003) and producing complementary information to its related vibrational spectroscopic technique, FT-IR spectroscopy.

Here, the aim was to use these three distinct phenotypic approaches (namely MALDI-TOF-MS, FT-IR spectroscopy and Raman spectroscopy) in combination with rigorous chemometric analysis of the resultant datasets to classify 35 clinically relevant isolates of *Enterococcus faecium*, which had been previously analysed by pulsed-field gel electrophoresis (PFGE). This was carried out in order to compare the results from, and determine the efficiency of, these analytical techniques for the rapid classification of enterococci. In future, this may allow clinical diagnostic laboratories to analyse multiple bacteria samples rapidly for infection control purposes in point-of-care setting within hospitals, clinics, or GP surgeries which would significantly accelerate diagnosis, ensure the correct antimicrobial therapies were used if required, and eliminate the delay associated with sending strains to reference laboratories when analysing patient samples.

5.2 Materials and Methods

5.2.1 General Chemicals

Trifluoroacetic acid (TFA), acetonitrile (ACN), sinapinic acid (SA), α-cyano-4-hydroxycinnamic acid (CHCA), and ferulic acid (FA) were purchased from Sigma-Aldrich (Dorset, UK).

5.2.2 Enterococci isolates

35 isolates from enterococci were previously analysed using pulsed-field gel electrophoresis (PFGE) in Public Health England’s National Reference Laboratory. Table 5.1 summarises information on the 35 clinical isolates within 12 groups (12 strains (12 PFGE-defined types)) including: EC04, EC09, EC10, EC13, EC14, EC15, EC19, EC20, UNI 156, UNI 178, UNI 191 and UNI 214. These bacterial samples were collected and sourced from clinical isolates from a hospital in Belfast UK. Following an increased number of enterococcal infections on a surgical ward in this hospital, appropriate infection control arrangements required identification of all patients carrying enterococci. Patients submitted fecal samples to the lab in the hospital, from which bacteria were cultured. The isolates were then sent to the reference laboratory where PFGE typing was performed and strains were allocated to a recognised EC*n* group or described as “unique” (UNI*n*).

5.2.3 Media

Two different types of media were used to culture the enterococci: Lysogeny Broth (LB) and Nutrient Agar (NA). LB was prepared by mixing 5 g of yeast extract (Amersham Life Sciences, Cleveland, USA), 10 g of tryptone (Formedia, Hunstanton, UK) and 10 g of NaCl dissolved in 1 L of distilled water and the broth was then autoclaved (at 121ºC and 15 psi for 45 min). NA was prepared from a preparatory mixture (beef extract 3 g/L, peptone 5 g/L, NaCl 8 g/L and agar 2 at 12 g/L) (Lab-M, Bury, UK) following the manufacturer’s instructions (28 g in 1 L of deionised water) and the broth was autoclaved (at 121ºC and 15 psi for 15 min).

Table 5.1: The 35 *Enterococcus faecium* isolates used in this study

No.	Isolates	Strains
1	139	EC10
2	151	EC10
3	144	EC13
4	149	EC13
5	152	EC13
6	154	EC13
7	155	EC13
8	167	EC13
9	177	EC13
10	185	EC13
11	194	EC14
12	203	EC14
13	190	EC15
14	223	EC15
15	224	EC15
16	173	EC19
17	174	EC19
18	175	EC19
19	192	EC20
20	198	EC20
21	204	EC20
22	109	EC04
23	170	EC04
24	179	EC04
25	193	EC04
26	133	EC09
27	160	EC09
28	211	EC09
29	205	EC09
30	219	EC09
31	233	EC09
32	156	UNI
33	178	UNI
34	191	UNI
35	214	UNI

UNI= UNIQUE (named in Public Health England’s National Reference Laboratory)

5.2.4 Bacterial isolates

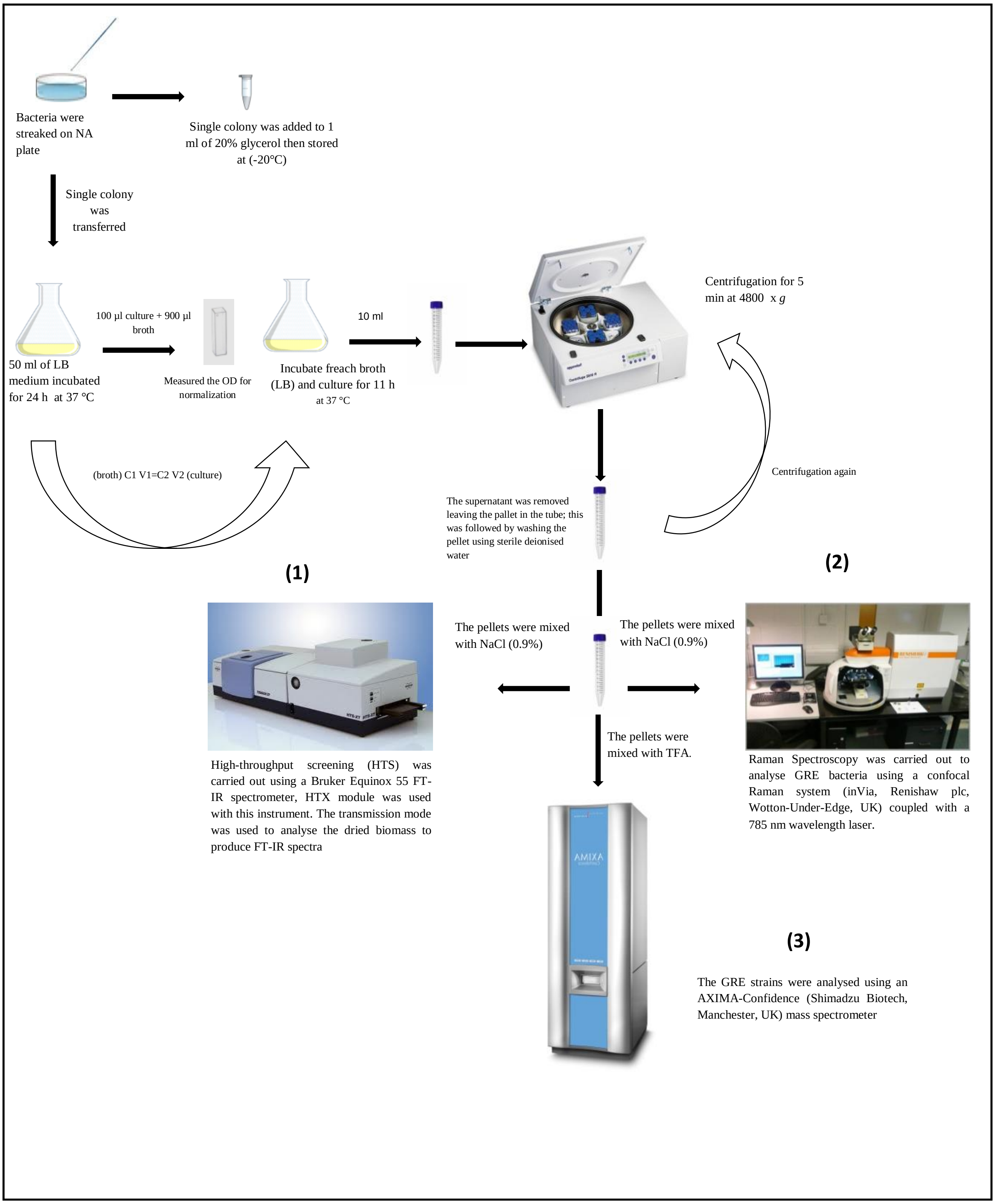
35 isolates of enterococci were shipped as slopes from Microbiology Laboratory, Royal Victoria Hospital, Belfast UK. Enterococci were streaked on NA plates to obtain single colonies. This was followed by collecting and transferring the bacteria to 1 mL of 20% [v/v] glycerol working inoculum stocks and these were then stored at -80ºC.

The samples analysed by the three techniques (*viz.* MALDI-TOF-MS, FT-IR and Raman) were collected from the same flask to avoid any variations between different preparations that may affect results obtined using the different anlaytical platforms. First, enterococci were cultured on NA plates for 24 h at 37ºC. A single colony from the agar culture was used to inoculate with 50 mL of LB in a 250 mL flask which was incubated overnight at 37ºC with shaking at 200 rpm. This was followed by measuring the optical density (OD) at 600 nm using a Biomate 5 spectrophotometer (Thermo, Hemel Hempstead, UK) for each of the isolates. The volume of analysed bacteria was then normalised to account for variance in cell biomass in the different replicate cultures (4 biological replicates were prepared for each isolate). Second, the new cultures were incubated at 37ºC for 11 h. Then, 10 mL from each flask was collected and centrifuged at 4800 *g* for 5 min and the pellet washed three times with sterile deionised water. Figure 1 illustrates the preparation process.

For vibrational spectroscopic analysis, the collected pellets were suspended in suitable volumes of NaCl (0.9% (w/v)) (depending on the OD). Then, 15 µL was spotted onto a silicon plate (Bruker Ltd., Coventry, UK) and was allowed to dry at 40ºC for 45 min before analysis with FT-IR spectroscopy. For Raman spectroscopy, 4 µL of each sample was spotted onto a stainless steel plate and then allowed to dry at 40ºC for 45 min.

For MALDI-TOF-MS, three different matrices were tested to find the most compatible matrix with enterococci; these matrices were: SA, FA and CHCA. In addition, 3 different deposition methods as described previously (AlMasoud *et al.*, 2014 (Chapter 2)) were tested to find the best method for depositing the samples: mix, overlay and underlay (data not shown). SA matrix and the mix deposition method were found to be the optimal combination for MALDI-TOF-MS analysis. On the day of analysis of the samples, the biomass was suspended in 1000 µL of 2%

TFA then vortexed for 3 min. An equal volume of 10 µL of bacterial suspended and matrix were vortexed for 2 s and 2 µL of this mixture spotted onto a MALDI stainless steel plate and allowed to dry at ambient temperature.


Bacteria were streaked on NA plate
Single colony was added to 1 ml of 20% glycerol then stored at (-20°C)
Single colony was transferred
50 ml of LB medium incubated for 24 h at 37 °C
100 µl culture + 900 µl broth
Measured the OD for normalization
Incubate freach broth (LB) and culture for 11 h at 37 °C
10 ml
Centrifugation for 5 min at 4800 x *g*
(broth) C1 V1=C2 V2 (culture)
Centrifugation again
(1)
(2)
(3)
The pellets were mixed with NaCl (0.9%)
The pellets were mixed with NaCl (0.9%)
The pellets were mixed with TFA.

The supernatant was removed leaving the pallet in the tube; this was followed by washing the pellet using sterile deionised water

High-throughput screening (HTS) was carried out using a Bruker Equinox 55 FT-IR spectrometer, HTX module was used with this instrument. The transmission mode was used to analyse the dried biomass to produce FT-IR spectra

Raman Spectroscopy was carried out to analyse GRE bacteria using a confocal Raman system (inVia, Renishaw plc, Wotton-Under-Edge, UK) coupled with a 785 nm wavelength laser.

The GRE strains were analysed using an AXIMA-Confidence (Shimadzu Biotech, Manchester, UK) mass spectrometer

Figure 5.1: Schematic of sample preparation for: (1) FT-IR spectroscopy, (2) Raman spectroscopy and (3) MALDI-TOF-MS analysis of bacterial isolates

5.2.5 Fourier transform infrared (FT-IR) spectroscopy

Sodium dodecyl sulfate (SDS) was used to wash a silicon FT-IR spectroscopy plate (Bruker Ltd., Coventry, UK) which contained 96 locations/spots. This was followed by washing the plate using deionised water and allowing it to dry at room temperature (Patel *et al*., 2008). High-throughput screening (HTS) was carried out using a Bruker Equinox 55 FT-IR spectrometer. The HTX^{TM} module described by Winder *et al.* (Winder *et al*., 2006) was used with this instrument. Transmission mode was used to analyse the dried biomass to produce FT-IR spectra. The parameters used for FT-IR analysis included the following: spectra were collected from the wavenumber range between 4000 and 600 cm^{-1}, resolution was 4 cm^{-1} and each spectrum was the average of 64 co-adds. Spectral acquisition and subtracting the background were achieved using Opus software (Bruker Ltd.). Four biological replicates, each in four analytical replicates were analysed and analysis was performed in three machine runs, resulting in 1680 FT-IR spectra.

5.2.6 Raman Spectroscopy

This was carried out using a confocal Raman system (inVia, Renishaw plc., Wotton-Under-Edge, UK) coupled with a 785 nm wavelength laser. A power intensity of ~30 mW was applied on the samples at an exposure time of 20 s. Four biological replicates and seven different locations within each sample spot were analysed, resulting in a total of 980 Raman spectra.

5.2.7 MALDI spectrometry

The enterococci isolates were analysed using an AXIMA-Confidence MALDI-TOF-MS (Shimadzu Biotech, Manchester, UK), equipped with a nitrogen pulsed UV laser with a wavelength of 337 nm. The parameters of this device were set as follows: 140 mV laser power, 91 acquired profiles with each profile containing 20 shots, linear TOF, positive ionisation mode, and mass-to-charge (*m/z*) range of 1000-18000. The spectra were collected using a circular raster pattern. The MALDI-TOF-MS device was calibrated using a protein mixture: insulin (5,735 Da), cytochrome *c* (12,362 Da), and apomyoglobin (16,952 Da) (Sigma-Aldrich). Each of 4 biological replicates from the 35 isoaltes was analysed in four technical replicates on four different days; this led to the generation of a total of 560 MALDI-TOF-MS spectra (35 isolates × 4 biological replicates × 4 analytical replicates).

5.3 Data analysis

5.3.1 Data pre-processing

Opus software was used to export FT-IR data into ASCII format; the data were then transferred into MATLAB 2012a (The Mathworks Inc., MA, US). All FT-IR spectra were corrected using standard normal variate (SNV) to remove any light scattering effect. The analytical replicates were averaged to reduce the number of redundant samples. Due to the large number of samples, 8 separate (96 spot silicon) sampling plates were used; therefore, it was necessary to correct for the subtle differences in signals from different silicon plates. This was achieved by using a piece-wise direct standardisation (PDS) model (Wang *et al.*, 1991). The PDS model was built on two different 'refreance' isolates which were spotted on every plate. The pre-processed FT-IR spectra were then subjected to multivariate analysis (MVA, see below). Raman spectra were also normalised using standard normal variate (SNV) and then subjected to MVA.

MALDI-TOF-MS data were pre-processed as follows: (i) the baseline was corrected using asymmetric least squares (AsLS) (Eilers, 2004), and (ii) spectra were normalised by dividing each individual baseline corrected spectrum with the square root of the sum of squares of the spectrum (Brereton, 2003). The pre-processed MALDI-TOF-MS data were subjected to the same data analysis flow as Raman and FT-IR spectral data.

5.3.2 Multivariate data analysis

A flowchart of multivariate data analysis is provided in Figure 5.2. For all three datasets, two types of classification were performed: one at the strain level (i.e. 12 classes), and the other at the isolate level (i.e. 35 classes).

For cluster analyses principal components-discriminant function analysis (PC-DFA) (Manly, 2004; Harrigan *et al.*, 2004; Gromski *et al.*, 2015) was applied to reduce the dimensionality of the data and discriminate samples from the designated classes. The PC-DFA scores of each class were then averaged and subjected to hierarchical cluster analysis (HCA) (Hastie *et al.*, 2009). Dendrograms from each analysis were generated to illustrate the relative relatedness of these bacteria.

Partial least squares-discriminant analysis (PLS-DA) (Barker and Rayens, 2003), with 1,000 bootstrapping validations (Efron and Tibshirani, 1994), was also applied to obtain a validated supervised classification model for discriminating different strains or isolates. In each bootstrapping process, the data were randomly split into two different sets: a training set and a test set. A PLS-DA model was trained on the training set and then applied to the test set to predict the class membership of the samples in the test set. This process was repeated 1,000 times and the results were recorded and averaged to produce a $c \times c$ confusion matrix (c is the number of designated classes, either 12 (strains) or 35 (isolates)), in which the element at the i^{th} row, j^{th} column is the percentage of samples in class i being predicted as class j on average. In order to assess the statistical significance of the predictive performance of the PLS-DA models, a corresponding permutation test within each bootstrapping resampling was also performed. This means that in addition to building the PLS-DA model using the known class membership, another model (called the 'null' model) was also built using a randomly permuted class membership. The results of the null models were also recorded and from this the null distribution was obtained. An empirical *p*-value was calculated by counting the number of cases where the null model had obtained better predictive accuracy than the real model and dividing the obtained number by the total number of bootstrapping resampling (i.e. 1,000 in this study).

Finally, similarities between the three different datasets (FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS data) were measured using Procrustes analysis (Gower and Dijksterhuis, 2005). Procrustes analysis is an excellent approach for assessing the differences and similarities between different ordination space from cluster analyses and has been used previously for the assessment of different analytical techniques (AlRabiah *et al*. 2014). The distances were calculated based on the averaged PC-DFA scores for the biological replicates.

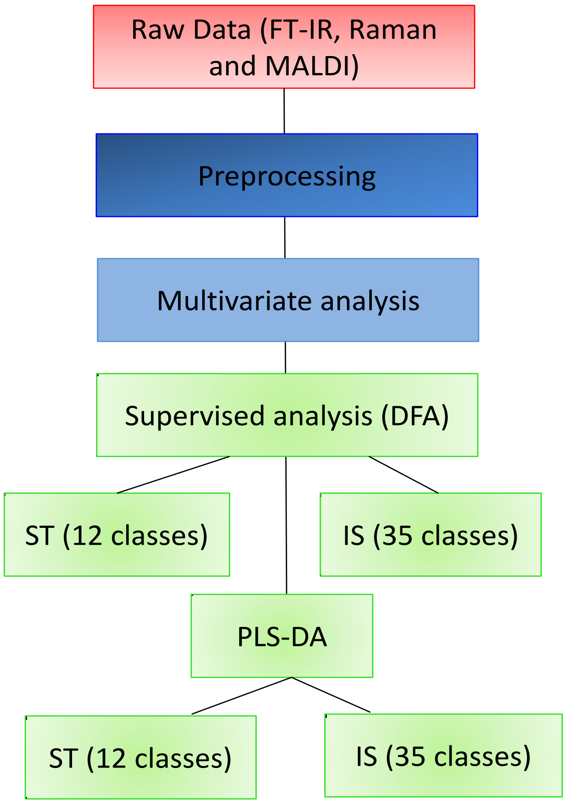


Figure 5.2: Workflow of data analysis undertaken for FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS. The data were first pre-processed then MVA was applied using PC-DFA at both the (ST) strain (12 classes) and (IS) isolate (35 classes) levels. This was followed by PLS-DA.

5.4 Results and Discussions

The number of *Enterococcus* infections is continually on the rise, leading to a requirement for improvement and development of sensitive, reliable and rapid methods to analyse these important and clinically relevant microorganisms (Quintela-Baluja *et al*., 2013). Accurate and rapid identification and classification of these pathogenic bacteria can assist in successful antibiotic treatment of infections. In this study, three analytical techniques were applied for the discrimination and classification of 35 isolates from 12 *Enterococcus faecium* strains. These clinical strains acquired from the surgical ward of the hospital were previously analysed using PFGE. Table 5.1 shows all 35 isolates belonging to 12 strains including: EC04, EC09, EC10, EC13, EC14, EC15, EC19, EC20 UNI 156, UNI 178, UNI 191 and UNI 214. The PFGE results (Figure S5.1) were compared to results obtained in this study using FT-IR spectroscopy (Naumann, 1984, Goodacre *et al*., 1998), Raman spectroscopy (Ferraro *et al*., 2003; Ashton *et al*., 2011; Muhamadali *et al*., 2015) and MALDI-TOF-MS (Cramer *et al*., 2005; Quintela-Baluja *et al*., 2013; Carbonnelle *et al*., 2011). We believe that these analytical techniques offered an improvement in the classification of bacteria due to their higher chemical resolution.

5.4.1 Classification using FT-IR spectroscopy

FT-IR spectroscopy is a rapid, robust and highly reproducible analytical technique with considerable potential for routine use in high-throughput clinical screening (Mariey *et al*., 2001; Ellis and Goodacre, 2006). This technique has been successfully used to discriminate bacteria to species and strain levels since pioneering work at the Robert Koch Institute in Berlin was published by Dieter Naumann and co-workers from the mid-1980s onwards (Naumann *et al*., 1988; Naumann *et al*., 1991; Helm *et al*., 1991). In this study, four biological replicates of bacterial isolates were analysed in four analytical replicates and analysis was performed in three machine runs, resulting in a total of 1680 FT-IR spectra. The three machine replicates measurements were performed in order to evaluate the reproducibility of the FT-IR technique. Typical spectra based on four biological replicates of 12 enterococcus strains (EC04, EC09, EC10, EC13, EC14, EC15, EC19, EC20, UNI 156, UNI 178, UNI 191 and UNI 214) are provided in Figure 5.3A. The infrared spectra contain different distinct regions that can be used to characterise bacterial samples. These have been well documented previously and include: wavenumbers around 3000-2850 cm^{-1} corresponding to fatty acids, at 1705-1454 cm^{-1} related to amide I and II regions attributed to peptides and proteins, and around 1085-1052 cm^{-1} corresponding to polysaccharides (Winder and Goodacre 2004; Kim *et al*., 2005; Naumann *et al*., 1991; Ellis *et al*., 2003).

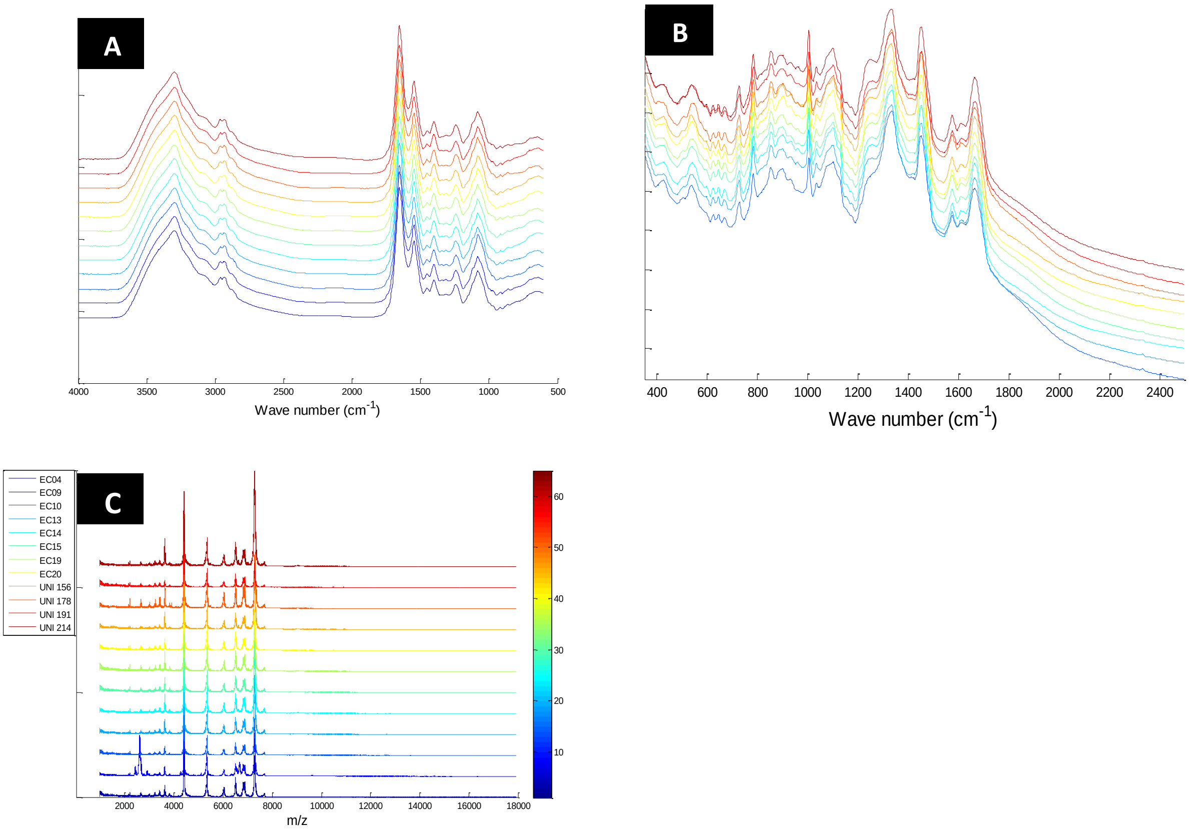


Figure 5.3: Typical spectra from (A) FT-IR spectroscopy, (B) Raman spectroscopy and (C) MALDI-TOF-MS for the 12 *Enterococcus faecium* strains (EC04, EC09, EC10, EC13, EC14, EC15, EC19, EC20, UNI 156, UNI 178, UNI 191 and UNI 214). The spectra from each analytical technique were plotted after pre-processing.

Discrimination between the bacterial strains based on visual inspection of the spectra was difficult because these strains are very similar phenotypically. Therefore, in order to develop a classification model to distinguish between bacterial samples based on similarities in the spectral data, multivariate analysis was used to reduce the high dimensionality of the data (Goodacre *et al*., 1998). First, PC-DFA was applied using 40 principal components (PCs) to the 12 strains (i.e. 12 classes) and 35 isolates (i.e. 35 classes) using the pre-processed FT-IR spectra (Figure. 5.4A and 5.5A, respectively). Figure 5.4A shows a clear separation between the 12 strains, displaying 4 main clusters; Cluster 1 containing only (EC10), Cluster 2 includes (EC20 and UNI 156), Cluster 3 (UNI 191, EC04 and EC15) and Cluster 4 formed a large group and is a combination of (EC13, EC19, EC14, EC09, UNI 214 and UNI 178). Each cluster is represented by a different colour in the figure. As described above, HCA was undertaken using spectral data in order to simplify the DFA plot and to illustrate the related strains. Cluster analysis was based on averaged DFA scores (12 classes/strains), using Ward's linkage as shown in Figure 5.4B. Clusters seen in Figure 5.4A are reflected in the HCA dendrogram plot (Figure 5.4B).

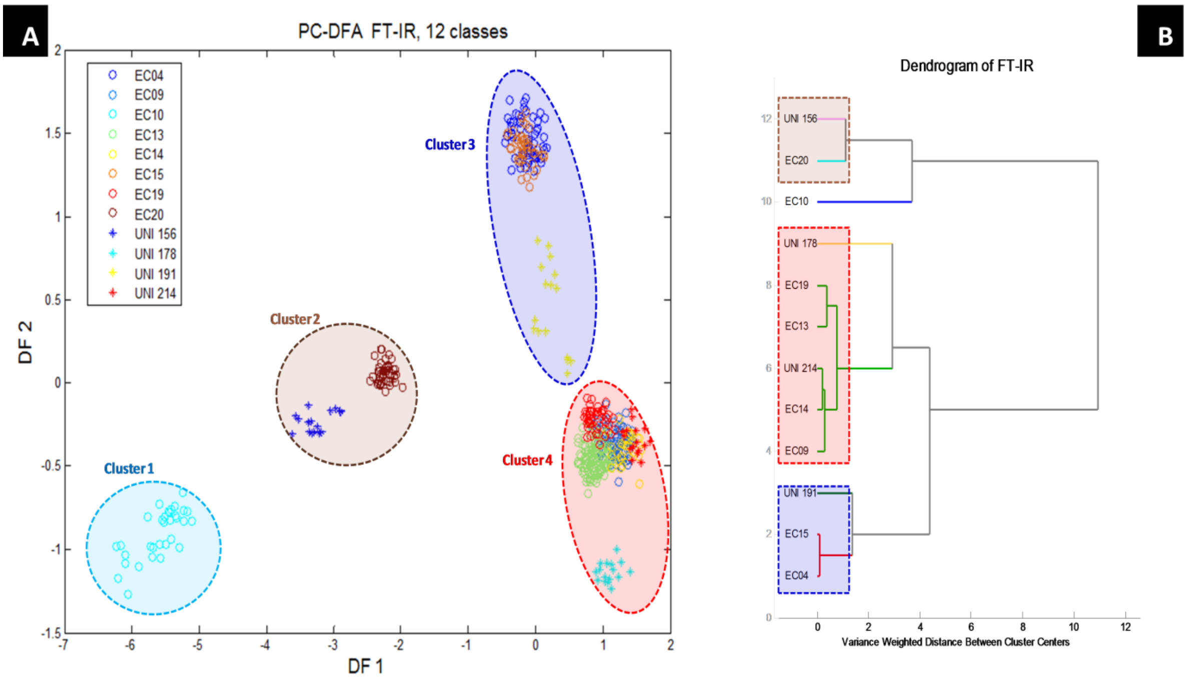


Figure 5.4: (A) Discriminant function analysis (DFA) scores plot from FT-IR data after pre-processing, illustrating the relationship between the 12 enterococci. (B) Cluster analysis on averaged PC-DFA scores (12 classes/strains) using Ward's linkage.

PC-DFA was subsequently performed for all the 35 isolates and the results are provided in Figure 5.5. Clear separation between all 35 isolates was observed despite the fact that there were a much higher number of classes to be separated than the number of strains. For example, clear separation was observed between the two representatives of EC10 (139 and 151). Furthermore, results generated using PFGE correlated well with FT-IR spectroscopic data. For example, the isolates UNI 156 and UNI 178 were seen as unique by both techniques. In addition, the three representatives of EC20 (192, 198 and 204) and EC19 (173, 174 and 175) clustered together and were not differentiated using FT-IR spectroscopy, which was also observed in the PFGE results, (Figure 5.5B). This implies that these isolates with each of these groups are highly similar to each other phenotypically and genetically. Finally, two more clusters were observed, with one cluster containing all the EC04, EC15 and UNI 191 and the remainder of the isolates forming another cluster.

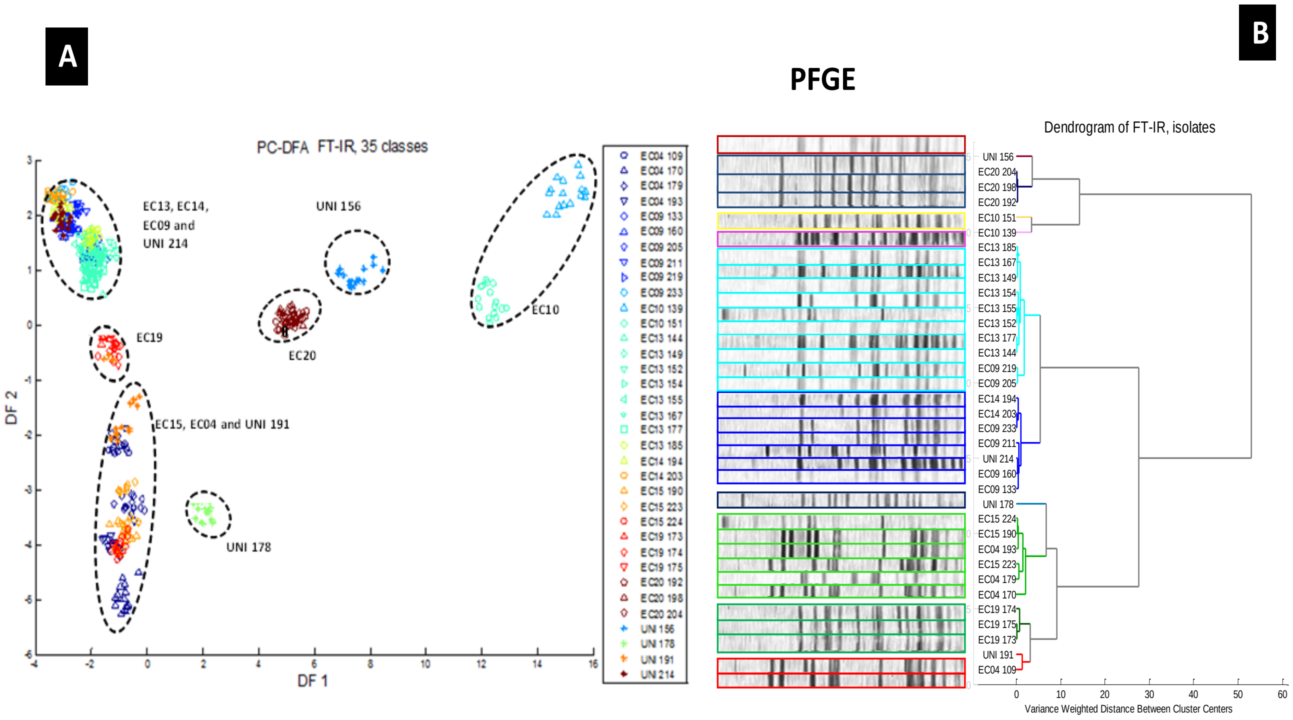


Figure 5.5: (A) PC-DFA plot from FT-IR data after pre-processing which illustrates the relationship between the 35 enterococcus isolates. (B) Hierarchical cluster analysis on averaged PC-DFA scores (35 classes/isolates) using Ward's linkage (right) and PFGE results (left). Each strain is represented by the same colour in both PFGE and dendrogram of FT-IR data.

The PLS-DA classification achieved an average correct classification rate (CCR) of 89.4% at the strain level and 54.3% at the isolate level, both with an empirical *p*-value of <0.001, i.e. not a single case where the null model obtained better results, indicating that the predictive accuracies were highly significant. The null distributions are provided in Figure S5.2A and B at the two levels.

The confusion matrices of strain and isolate classification are presented in Table 5.2 and Table S5.1, respectively. Most of the 12 strains showed high prediction accuracies, for example EC04, EC10, EC13 and EC20 had accuracies of around 89.9%, 99.7%, 99.8% and 99.2%, respectively. However, EC14 and UNI 214 had lower prediction accuracies of 47.3% and 58.9%, respectively. The confusion matrix showed that there was a certain level of overlap between (EC14 and EC09) and (UNI 214 and EC19).

Table 5.2: The prediction accuracies of the 12 enterococci strains using FT-IR spectroscopy data

Class Known / Predicted	EC04	EC09	EC10	EC13	EC14	EC15	EC19	EC20	UNI 156	UNI 178	UNI 191	UNI 214
EC04	89.9%	0.5%	0.0%	0.0%	0.4%	8.3%	0.1%	0.0%	0.0%	0.0%	0.7%	0.1%
EC09	0.1%	90.3%	0.0%	1.3%	4.8%	0.0%	3.5%	0.0%	0.0%	0.0%	0.0%	0.0%
EC10	0.0%	0.1%	99.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%
EC13	0.0%	0.0%	0.0%	99.8%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
EC14	0.1%	48.9%	0.0%	1.1%	47.3%	1.0%	1.4%	0.1%	0.0%	0.0%	0.1%	0.0%
EC15	6.8%	1.4%	0.0%	0.0%	0.5%	91.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%
EC19	1.6%	9.3%	0.0%	0.2%	3.6%	0.0%	83.5%	0.0%	0.0%	0.0%	0.0%	1.8%
EC20	0.0%	0.1%	0.0%	0.0%	0.0%	0.7%	0.0%	99.2%	0.0%	0.0%	0.0%	0.0%
UNI 156	0.4%	0.0%	0.0%	0.5%	0.0%	0.0%	0.1%	0.9%	98.1%	0.0%	0.0%	0.0%
UNI 178	0.0%	5.3%	0.0%	0.1%	0.0%	0.0%	0.4%	0.0%	0.0%	93.9%	0.2%	0.0%
UNI 191	6.5%	0.9%	0.0%	25.2%	0.0%	1.3%	0.0%	0.0%	0.0%	0.0%	66.1%	0.1%
UNI 214	1.9%	13.4%	0.0%	1.0%	0.1%	0.0%	20.4%	0.0%	0.0%	0.0%	4.2%	58.9%

Furthermore, in-depth analysis of the confusion matrix for the classification of isolates (Figure 5.6) showed that classification of unique strains was generally in line with PFGE results. In Figure 5.6, high percentage class membership assignments are represented by warm colours (e.g. red), indicating agreement between predicted classes and known classes. It is also interesting to see that representatives from strains EC19 and EC20 formed two “squares” of “tiles” on the diagonal line, in which the colours were similar to each other. Results from Figure 5.6 suggest that the PLS-DA model was not able to differentiate the isolates within EC19 and EC20, yet another observation that is consistent with PFGE results. On the other hand, all representatives of EC04 and EC09 (160 and 133) were unique in the FT-IR spectroscopy profile using the PLS-DA model but not in the PFGE profiles. This is most likely due to PFGE providing genetic information (Turabelidze *et al*., 2000; Bannerman *et al*., 1995) while FT-IR spectroscopy describes phenotypes (Davis and Mauer, 2010; Alvarez-Ordóñez *et al*., 2011). This implies that isolates from EC19 and EC20 may be highly conserved phenotypically, whereas those from EC04 and EC09 are not, and such subtle differences in phenotypes were detected by FT-IR spectroscopy. Our observations showed that FT-IR spectroscopy appears to be a very promising analytical approach for discrimination of enterococci bacteria at the strain and isolate levels. In line with the results presented in this study, work carried out by Guibet *et al*. showed that clear discrimination and classification of enterococci strains can be achieved using FT-IR spectroscopy (Guibet *et al*., 2003).

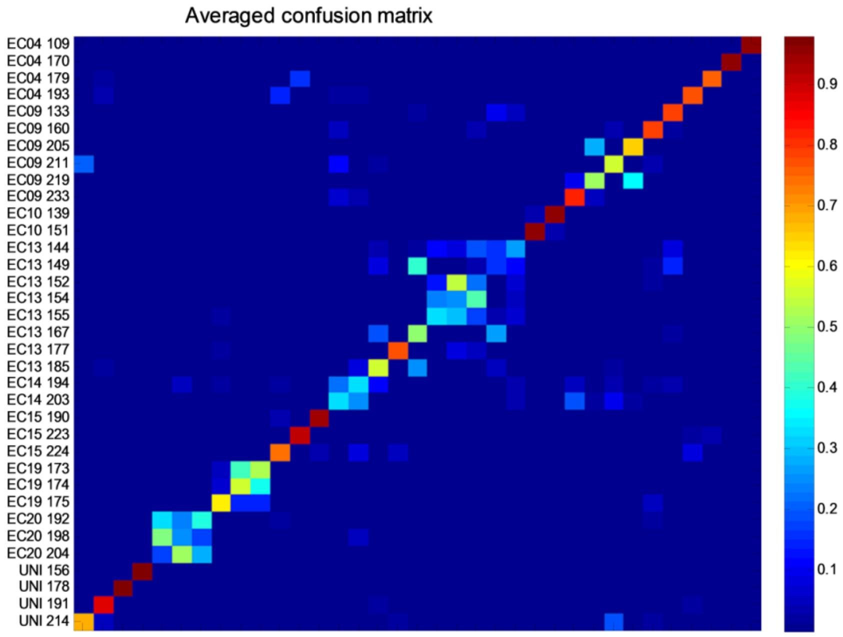


Figure 5.6: PLS-DA trained on 35 classes (i.e. 35 isolates) from FT-IR spectral data. High percentage class membership assignments are represented by warm colours (e.g. red) whilst the cold colours (e.g. blue) represent low percentage class membership assignments. The diagonal “tiles” are much warmer than off-diagonal “tiles”, which indicates agreement between predicted classes and known classes.

5.4.2 Classification using Raman spectroscopy

In addition to the FT-IR spectroscopy technique used in this study, Raman spectroscopy was used as a complementary technique (Kirschner *et al*., 2001; van de Vossenberg *et al*., 2013; Ch. Schroder *et al*., 2015). As expected, the two techniques generated different spectra. These two approaches are complementary due to the selection rules, whereby infrared causes a change in the net dipole moment in a particular functional group, induced by molecular vibrations, whereas Raman causes a change in the polarisation of bond within a molecule. Therefore, bonds within a molecule are infrared or Raman active with the result being that the two techniques can provide complementary (bio) chemical information (Ferraro *et al*., 2003; Goodacre *et al*., 2002).

Raman spectra of the 12 *Enterococcus faecium* strains are shown in Figure 5.3B. Raman spectra for these strains appeared almost indistinguishable and no differences waere detected on visual inspection. Moreover, some specific peaks which were identified in these spectra included: peaks at around 720 cm^{-1}, 854 cm^{-1}, 1004 cm^{-1}, 1336 cm^{-1}, 1451 cm^{-1} and 1663 cm^{-1}, which correspond to adenine, tyrosine, phenylalanine, guanine, protein and amide I, respectively (Uzunbajakava *et al*., 2003; Huang *et al*., 2010).

PC-DFA scores plot of pre-processed Raman spectra for the 12 enterococci at the strain level is shown in Figure 5.7A. The figure shows classification results similar to those seen with FT-IR spectroscopy data. There was an obvious overlap between the two spectroscopic techniques, especially with representatives of EC10. However, EC20 overlapped with UNI 156 in FT-IR spectroscopy data, whereas EC20 was closer to UNI 178 based on Raman spectroscopy data. These observations can be seen in the HCA dendrogram based on Raman data (Figure 5.7B), which was quite similar to the HCA results generated from FT-IR data. Looking back at the dendrogram in Figure S5.1 based on PFGE data, visual inspection showed that there were some similarities between results generated via spectroscopic techniques and those based on PFGE; for example, EC04 and EC15 were shown to overlap in both sets of results (Figure S5.1).

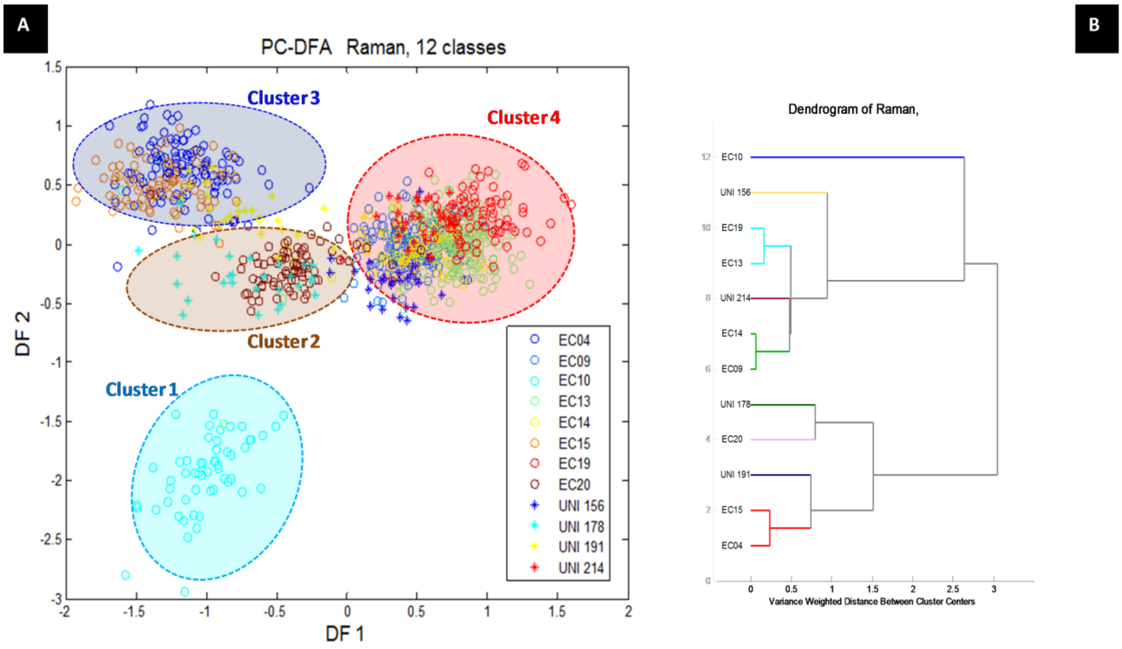


Figure 5.7: PC-DFA plot from Raman data after pre-processing, illustrating the relationship between the 12 enterococci. (B) Cluster analysis on averaged DFA scores (12 classes/strains) using Ward's linkage.

As with FT-IR data, Raman spectroscopy data on the 35 isolates were also submitted to PC-DFA and HCA (Figure S5.3A and B, respectively). The results suggested that Raman spectroscopy was also successful in discriminating the two representatives of EC10 (139 and 151), which was also the case using FT-IR analysis (Figure 5.5). Furthermore, in order to ensure the classification is robust, the data were analysed using a heat map based on PLS-DA (Figure S5.3C). The results suggested that all the isolates indicated as unique (UNI) by PFGE were also unique in the PLS-DA model generated using Raman spectroscopy data.

In addition, chemometric-based identification was carried out using PLS-DA at both the strain and isolate levels and the predictive accuracies were calculated based on 1,000 bootstrapping resampling using Raman spectral data. The null distribution was obtained (Figure S5.2C and D) at both the strain (12 classes) and isolate levels (35 classes) resulting in an average correct classification rates (CCR) of 69.3% (*p*<0.001) and 21.1% (*p*<0.001), respectively. The CCR from FT-IR data was higher at both levels compered to Raman data possibly due to the higher reproducibility of FT-IR data. The prediction accuracies were also generated at both the strain level (Table 5.3) and the isolate level (Table S5.2); these results suggested that Raman spectroscopy can also be used as a robust technique for bacterial discrimination. In-depth analysis showed that Raman spectroscopy generated around 70% prediction accuracy at the strain level which is lower than that of FT-IR spectroscopy (nearly 90%). This is most likely due to the low concentration of cells used for analysis: the infrared interrogation beam used was *ca*. 1 mm and passes completely through the dried bacterial film; while the Raman microscope delivers a highly focussed laser beam with an interrogation volume of ~1 pL and therefore measures very few bacteria. To overcome this limitation with Raman, bacteria can be analysed directly from the agar plates or surface-enhanced Raman spectroscopy (SERS) can be used as an alternative technique (Cotton *et al*., 1991; Nabiev *et al*., 1994; Jarvis and Goodacre, 2008), but this is an area for future study.

Table 5.3: The prediction accuracies of the 12 enterococci strains using Raman spectroscopy data

Class Known / Predicted	EC04	EC09	EC10	EC13	EC14	EC15	EC19	EC20	UNI 156	UNI 178	UNI 191	UNI 214
EC04	71.5%	2.3%	0.6%	2.1%	0.3%	21.5%	1.4%	0.1%	0.0%	0.0%	0.1%	0.0%
EC09	0.6%	69.3%	0.0%	17.4%	7.4%	0.9%	2.8%	0.9%	0.0%	0.0%	0.1%	0.5%
EC10	0.1%	7.0%	88.8%	3.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%
EC13	1.0%	6.9%	0.4%	82.3%	0.5%	0.7%	4.4%	1.2%	2.1%	0.4%	0.0%	0.1%
EC14	0.2%	77.6%	0.0%	10.0%	7.3%	0.5%	0.6%	3.7%	0.0%	0.0%	0.0%	0.1%
EC15	33.3%	2.3%	0.0%	4.4%	0.2%	58.5%	0.8%	0.3%	0.0%	0.1%	0.1%	0.0%
EC19	0.4%	4.9%	0.0%	25.5%	0.0%	0.0%	68.9%	0.0%	0.0%	0.0%	0.2%	0.0%
EC20	0.5%	3.3%	0.2%	2.4%	0.7%	2.2%	0.2%	90.2%	0.1%	0.1%	0.0%	0.0%
UNI 156	1.2%	4.1%	0.0%	42.6%	0.1%	0.1%	1.5%	16.1%	34.1%	0.0%	0.0%	0.0%
UNI 178	4.0%	14.4%	1.3%	14.5%	0.2%	4.7%	10.4%	4.9%	0.0%	45.4%	0.2%	0.0%
UNI 191	39.1%	11.9%	0.0%	9.8%	4.7%	19.3%	4.1%	5.6%	0.0%	0.1%	5.4%	0.0%
UNI 214	6.0%	41.7%	0.0%	31.5%	2.1%	0.3%	12.4%	0.0%	0.0%	0.0%	0.1%	5.8%

5.4.3 Classification using MALDI-TOF mass spectrometry

As described in the Materials and Methods section, four biological replicates were analysed in four analytical replicates for each bacterial strain, resulting in 560 MALDI-TOF-MS spectra. The spectra for all 35 enterococci isolates were pre-processed before data analysis. The typical pre-processed positive ion mode MALDI-TOF-MS spectra for all 12 *E. faecium* strains (EC04, EC09, EC10, EC13, EC14, EC15, EC19, EC20, UNI 156, UNI 178, UNI 191 and UNI 214) are provided in Figure 5.3C. The generated MALDI-TOF-MS spectra were of high quality with high signal-to-noise ratios in the acquisition *m/z* range 1000-18,000 and a high number of peaks for each studied strains. There are many factors that can affect MALDI-TOF-MS results and some of these can differ from lab to another, such as the type of medium used (Lay Jr, 2000; Shu *et al*., 2012), sample handling, type of matrix (Giebel *et al*., 2010), sample deposition method (Dreisewerd, 2003), solvents (Williams *et al*., 2003), instrument settings (Freiwald and Sauer, 2009; Williams *et al*., 2003) and the type of data analysis chosen (Gromski *et al*., 2015; Gromski *et al*., 2014). These can inadvertently affect MALDI-TOF-MS results and subsequent PC-DFA and HCA.

MALDI-TOF-MS spectra are not readily interpretable from enterococci at strain and isolate levels, as they are similar phenotypically and MALDI-TOF-MS spectra show only two dimensions (*m*/*z* × intensity). Therefore, as is the case for the vibrational spectroscopy techniques, robust multivariate analysis methods were employed for this purpose. The results of PC-DFA using 12 classes (12 strains) in a three-dimensional plot of DF1 *vs* DF2 *vs* DF3 and a two-dimensional plot of DF2 *vs* DF3 are shown in Figure 5.8A and B, respectively. Four main clusters were observed in the PC-DFA plots; Cluster 1 contains only UNI 178; Cluster 2 contains EC20; Cluster 3 consists of EC04, EC10, EC15 and UNI 191 and Cluster 4 formed a large group of (EC13, EC19, EC14, EC09, UNI 214 and UNI 156). Results from the HCA dendrogram (Figure 5.8C) confirmed the separation between the 12 classes (i.e. 12 strains). This indicated that UNI 178 is phenotypically very different from the other strains based on MALDI-TOF-MS data.

PC-DFA was also applied to data from the 35 isolates; the results showed that isolates 160 and 219 (both from EC09) were very different from the other isolates and dominated the plot (data not shown). The HCA dendrogram plot of data on the 35 isolates showed that these two strains also dominated the corresponding dendrogram (Figure S5.4B). Therefore, another PC-DFA was carried out with the two dominating strains removed and the HCA results are shown in Figure S5.4D. It appears that all representatives of EC20 (204, 198 and 192) overlap with each other, which was also observed in FT-IR and Raman spectroscopy data, with the exception that isolates 192 slightly differed from the other two representatives (204 and 198) in the HCA dendrogram when using Raman data (Figure S5.3B). However, analysis using PFGE typing showed that isolates 192 and 198 clustered more closely with each other than with 204.

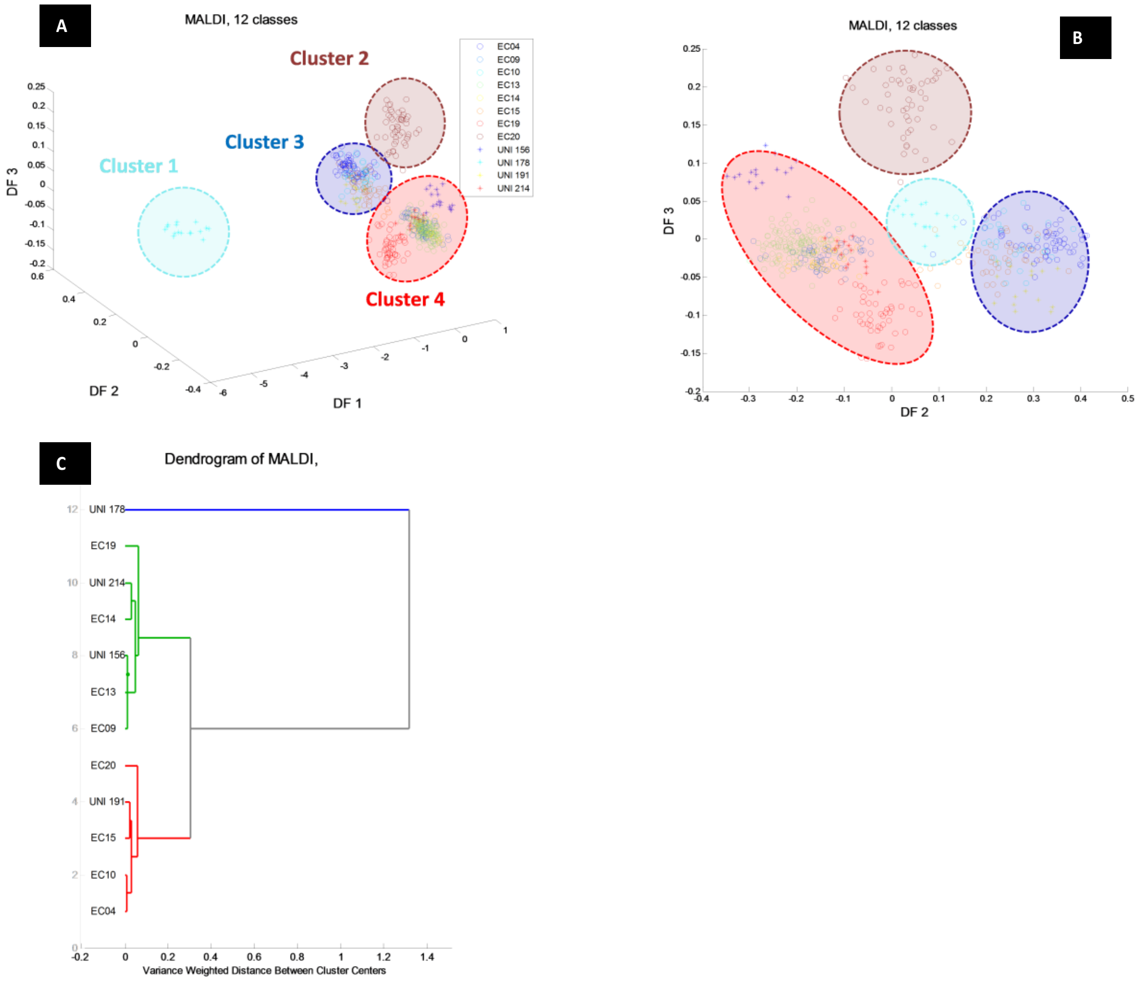


Figure 5.8: (A) 3-D PC-DFA plot from MALDI-TOF-MS data after pre-processing, illustrating the relationship between the 12 enterococci strains. (B) DFA plot for DF2 *vs* DF3. (C) Hierarchical cluster analysis on averaged DFA scores from MALDI-TOF-MS data of the 12 strains from enterococci bacteria using Ward's linkage.

Furthermore, PLS-DA model applied to MALDI-TOF-MS data achieved an averaged CCR of 78.2% (p<0.001) and 35.7% (p<0.001) for 12 (strains) and 35 (isolates) classes, respectively. When PLS-DA was undertaken with 33 isolates (with isolates 160 and 219 removed), the average CCR increased to 53.95% (p<0.001). The prediction accuracies for the 12 classes (strains) are shown in Table 5.4 and those for the 35 classes (isolates) are shown in Table S5.3. Table 5.4 shows that discrimination between most of the strains (12 calsses) using MALDI-TOF-MS data achieved high correct classification rates, except for EC14 and UNI 191, which had rather low classification rates.

Table 5.4: The prediction accuracies of the 12 enterococci strains using MALDI-TOF-MS data

Class Known / Predicted	EC04	EC09	EC10	EC13	EC14	EC15	EC19	EC20	UNI 156	UNI 178	UNI 191	UNI 214
EC04	93.8%	1.0%	0.1%	0.4%	0.3%	3.8%	0.1%	0.1%	0.0%	0.0%	0.3%	0.0%
EC09	1.0%	71.5%	0.1%	11.0%	13.7%	0.3%	0.5%	0.2%	0.2%	0.0%	0.1%	1.5%
EC10	0.4%	3.7%	83.1%	4.4%	0.1%	2.6%	1.6%	2.8%	0.0%	0.0%	1.0%	0.3%
EC13	0.3%	2.5%	0.0%	95.8%	0.7%	0.6%	0.0%	0.0%	0.1%	0.0%	0.1%	0.0%
EC14	0.0%	58.7%	0.0%	15.0%	25.8%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
EC15	7.8%	2.7%	0.9%	0.8%	1.3%	77.0%	0.8%	0.4%	0.2%	0.0%	8.0%	0.1%
EC19	0.0%	0.4%	0.0%	1.8%	0.0%	0.0%	97.7%	0.0%	0.0%	0.0%	0.0%	0.0%
EC20	0.1%	0.2%	0.1%	1.1%	0.1%	0.4%	0.0%	97.6%	0.5%	0.0%	0.0%	0.0%
UNI 156	0.0%	10.8%	0.0%	18.7%	0.5%	0.1%	0.0%	12.1%	56.1%	0.0%	0.0%	1.7%
UNI 178	0.6%	4.4%	0.0%	8.2%	0.0%	3.1%	0.4%	0.0%	0.0%	83.2%	0.0%	0.0%
UNI 191	51.6%	0.4%	0.7%	0.0%	0.2%	29.1%	4.7%	0.0%	0.0%	0.0%	13.2%	0.0%
UNI 214	0.6%	20.6%	0.8%	13.9%	1.4%	0.9%	1.1%	0.0%	2.8%	0.0%	0.0%	57.9%

Confusion matrices for the 35 classes and the 33 classes (when 160 and 219 isolates were removed) are shown in Figure S5.4A and C, respectively. From these matrices, it can be seen that all the isolates identified by the reference laboratory as unique (UNI), which included isolates 156, 178, 191 and 214, were also classified as unique based on MALDI-TOF-MS data using PLS-DA modelling. Moreover, EC20 and EC19 were assigned the same classification in PFGE typing, and this was in agreement with MALDI-TOF-MS, FT-IR spectroscopy and Raman spectroscopy data. In addition, based on MALDI-TOF-MS data (Figure S5.4A and C), representatives of EC13 (152, 154 and 155) belonged to the same cluster, and isolates 177 from EC13 was significantly different from the remaining EC13 isolates; this was also observed in FT-IR and PFGE data. Looking back at Figure S5.4C, it can be seen that all the strains from EC04 were unique in MALDI-TOF-MS and FT-IR profiles when using PLS-DA modelling.

5.4.4 Procrustes distance of three analytical techniques

Analytical techniques such as FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS are currently used in clinical research studies worldwide and many reports have been published showing advantages of using such techniques (Beekes *et al*., 2007; De Carolis *et al*., 2012; Risch *et al*., 2010; Carbonnelle *et al*., 2011). Kirschner *et al*. (2001) demonstrated accurate identification and classification of 18 strains from 6 different species belonging to enterococci using vibrational spectroscopic techniques in combination with chemometrics. This study suggested that FT-IR and Raman spectroscopy can offer potential alternatives to the conventional typing tests due to their speed and ease of use. In another previous study it was also shown that 59 clinical bacterial strains associated with urinary tract infections (UTIs) could be identified using FT-IR and Raman spectroscopy (Goodacre *et al*., 1998). As an alternative to vibrational spectroscopic techniques, MALDI-TOF-MS is a relatively new technique which has shown very promising results in agreement with methodologies carried out in microbiological laboratories, and therefore has been used for the identification and classification of bacterial species (Benagli *et al*., 2011; Sauer and Kliem, 2010; Bizzini and Greub, 2010).

Previous studies have generally focused on the application of just one or two analytical techniques for the classification of *Enterococcus* spp. bacteria. However, to generate complementary data and more comprehensive analysis, this study combines three different analytical techniques – FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS – to analyse whole bacterial cells. Successful classification was demonstrated at the strain (i.e. 12 classes) and isolate (i.e. 35 classes) level based on data generated by the three analytical platforms. Using Procrustes analysis, similarity between the patterns detected by these three platforms. In order to assess the overall information content in the spectra that has been revealed by the cluster analysis from the scores plots, Procrustes analysis was employed to assess the overall similarity between the patterns detected by these three platforms. The results are presented in terms of Procrustes distance (Table 5.5A and B), where the Procrustes distance varies from 0 to 1; the lower the distance, the higher the similarity between the results. The comparisons were made using averaged PC-DFA scores. For each dataset, there were two sets of PC-DFA scores, one for the strain level (12 classes) and the other for isolate classification (35 classes). For each set of PC-DFA scores, the scores were then averaged according to their strain label and isolate label to give two sets of *averaged* PC-DFA scores.

Table 5.5: Shows the similarity between three different data sets using Procrustes distance

(A) PC-DFA on strain level

Averaging on ST level	FT-IR (IS)	FT-IR (ST)	Raman (IS)	Raman (ST)	MALDI (IS)	MALDI (ST)
FT-IR (IS)	-					
FT-IR (ST)	0.0858	-				
Raman (IS)	0.2125	0.2933	-			
Raman (ST)	0.2314	0.3187	0.1502	-		
MALDI (IS)	0.8602	0.889	0.899	0.8202	-	
MALDI (ST)	0.9125	0.8846	0.9149	0.8988	0.1812	-

(B) PC-DFA on isolate level

Averaging on IS level	FT-IR (IS)	FT-IR (ST)	Raman (IS)	Raman (ST)	MALDI (IS)	MALDI (ST)
FT-IR (IS)	-					
FT-IR (ST)	0.1085	-				
Raman (IS)	0.2112	0.2446	-			
Raman (ST)	0.2411	0.3168	0.1132	-		
MALDI (IS)	0.8593	0.8719	0.8196	0.8001	-	
MALDI (ST)	0.8975	0.8608	0.8841	0.8703	0.0681	-

(ST) and (IS) indicate the PC-DFA calculated at the strain (12 classes) and isolate (35 classes) levels, respectively.

The findings in Table 5.5 can be summarised as follows:

(i) The patterns in the PC-DFA scores at the isolate level and strain level were highly similar to each other for all the three analytical platforms. The Procrustes distances varied from 0.0681 to 0.1812. This suggested that the variation originated from different strain is the main factor in PC-DFA, i.e. the differences between different strains were significantly higher than those between different isolates.

(ii) The two vibrational spectroscopic techniques (FT-IR and Raman) generated highly similar results both at the strain and isolate classification levels, with the corresponding Procrustes distances varying from 0.2112 to 0.3187.

(iii) However, the results generated by MALDI-TOF-MS were significantly different from those generated by the two spectroscopic techniques, and the corresponding Procrustes distances varied were all above 0.8. Such differences can be mainly attributed to data on isolate UNI 178, which appeared to be very different to other isolates in the MALDI-TOF-MS dataset.

Table 5.6 shows a summative comparison of the 4 main clusters identified based on the three analytical techniques using PC-DFA plots of the 12 strains of *E. faecium* (12 classes). It can be seen from this table that despite the large Procrustes distances between data generated by MALDI-TOF-MS and those generated by the other two techniques, the main identified clusters patterns observed in all three datasets were still largely consistent.

Table 5.6: Shows the 4 main clusters that were observed from the three different analytical techniques based on the PC-DFA plots of 12 classes (12 strains)

	Cluster 1	**Cluster 2**		**Cluster 3**		**Cluster 4**	
FT-IR	EC10	EC20	UNI 156	EC04/EC15/UNI 191		EC13/EC19/EC14/ EC09/UNI 214	UNI 178
Raman	EC10	EC20	UNI 178	EC04/EC15/UNI 191		EC13/EC19/EC14/ EC09/UNI 214/ UNI 156	
MALDI	UNI 178	EC20		EC04/EC15/UNI 191	EC10	EC13/EC19/EC14/ EC09/UNI 214/ UNI 156	

5.5 Concluding and remarks

The results obtained from the two vibrational spectroscopic techniques demonstrated that good discrimination can be achieved at both the strain and isolate levels and the detected patterns from the two techniques were highly similar. In addition, bacterial classification results from MALDI-TOF-MS were generally consistent with these vibrational spectroscopic techniques. However, UNI 178 was detected to be very different in MALDI-TOF-MS data, which differed from the other two analytical techniques employed in this study.

The results obtained using these spectroscopic phenotyping approaches were mostly consistent with previous results obtained from experiments carried out using the genotypic classification method of PFGE. Some of the results differed when directly comparing our analytical approach with results from the molecular approach and these differences may be due to comparing phenotypic differences from whole-organism fingerprinting with genotypic differences using PFGE.

In conclusion, in this work, we have presented an assessment of several analytical phenotyping methods as a complementary approach to currently used molecular methods. All the described methods provided excellent identification, which is in general agreement with results from genotypic baseline methods, and therefore, allowed high level of discrimination down to the strain level with sufficient resolution at the sub-strain level. We believe that these physicochemical techniques have excellent potential to become high-throughput point-of-care screening tools for rapid and reproducible classification and identification.

5.6 References

AlMasoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. 2014. Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the Characterisation of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta,* **840**, 49-57

AlRabiah, H., Xu, Y., Rattray, N.J., Vaughan, A.A., Gibreel, T., Sayqal, A., Upton, M., Allwood, J.W. and Goodacre, R. 2014. Multiple metabolomics of uropathogenic *E. coli* reveal different information content in terms of metabolic potential compared to virulence factors. *Analyst*, **139**, 4193-4199

Altekruse, S., Cohen, M. and Swerdlow, D. 1997. Emerging foodborne diseases. *Emerging Infectious Diseases,* **3**, 285-294

Alvarez-Ordóñez, A., Mouwen, D. J. M., López, M. and Prieto, M. 2011. Fourier transform infrared spectroscopy as a tool to characterise molecular composition and stress response in foodborne pathogenic bacteria. *Journal of Microbiological Methods,* **84**, 369-378

Argyri, A. A., Jarvis, R. M., Wedge, D., Xu, Y., Panagou, E. Z., Goodacre, R. and Nychas, G.-J. E. 2013. A comparison of Raman and FT-IR spectroscopy for the prediction of meat spoilage. *Food Control,* **29**, 461-470

Arias, C. A. and Murray, B. E. 2012. The rise of the Enterococcus: beyond vancomycin resistance. *Nature Reviews Microbiology,* **10**, 266-278

Ashton, L., Lau, K., Winder, C. L. and Goodacre, R. 2011. Raman spectroscopy: lighting up the future of microbial identification. *Future Microbiology,* **6**, 991-997

Bannerman, T. L., Hancock, G. A., Tenover, F. C. and Miller, J. M. 1995. Pulsed-field gel electrophoresis as a replacement for bacteriophage typing of Staphylococcus aureus. *Journal of Clinical Microbiology,* **33**, 551-555

Barker, M. and Rayens, W. 2003. Partial least squares for discrimination. *Journal of Chemometrics,* **17,** 166-173

Beekes, M., Lasch, P. and Naumann, D. 2007. Analytical applications of Fourier transform-infrared (FT-IR) spectroscopy in microbiology and prion research. *Veterinary Microbiology,* **123**, 305-319

Benagli, C., Rossi, V., Dolina, M., Tonolla, M. and Petrini, O. 2011. Matrix-assisted laser desorption ionisation-time of flight mass spectrometry for the identification of clinically relevant bacteria. *PLoS One,* **6**, e16424

Bizzini, A. and Greub, G. 2010. Matrix-assisted laser desorption ionisation time-of-flight mass spectrometry, a revolution in clinical microbial identification. *Clinical Microbiology and Infection,* **16**, 1614-1619

Brereton, R. G. 2003. *Chemometrics: data analysis for the laboratory and chemical plant*, Chichester, John Wiley and Sons, pp. 489

Burgula, Y., Khali, D., Kim, S., Krishnan, S. S., Cousin, M. A., Gore, J. P., Reuhs, B. L. and Mauer, L. J. 2007. Review of mid-infrared fourier transform-infrared spectroscopy applications for bacterial detection. *Journal of Rapid Methods and Automation in Microbiology,* **15**, 146-175

Carbonnelle, E., Mesquita, C., Bille, E., Day, N., Dauphin, B., Beretti, J.-L., Ferroni, A., Gutmann, L. and Nassif, X. 2011. MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory. *Clinical Biochemistry,* **44**, 104-109

Ch. Schroder, U., Beleites, C., Assmann, C., Glaser, U., Hubner, U., Pfister, W., Fritzsche, W., Popp, J. and Neugebauer, U. 2015. Detection of vancomycin resistances in enterococci within 3 1/2 hours. *Scientific Reports,* **5**, 8217

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. 1996. The rapid identification of intact microorganisms using mass spectrometry. *Nature Biotechnology,* **14**, 1584-1586

Cotton, T. M., Kim, J. H. and Chumanov, G. D. 1991. Application of surface-enhanced Raman spectroscopy to biological systems. *Journal of Raman Spectroscopy,* **22,** 729-742

Cramer, R., Gobom, J. and Nordhoff, E. 2005. High-throughput proteomics using matrix-assisted laser desorption/ionisation mass spectrometry, *Expert Review Proteomics,* **2**, 407-20

Davis, R. and Mauer, L. 2010. Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic bacteria. In*:* Méndez-Vilas A. (Ed.)*, Current research, technology and education topics in applied microbiology and microbial biotechnology, Volume II*. pp.1582-1594. Formatex Research Center: Badajoz, Spain.

De Carolis, E., Posteraro, B., Lass-Flörl, C., Vella, A., Florio, A. R., Torelli, R., Girmenia, C., Colozza, C., Tortorano, A. M., Sanguinetti, M. and Fadda, G. 2012. Species identification of Aspergillus, Fusarium and Mucorales with direct surface analysis by matrix-assisted laser desorption ionisation time-of-flight mass spectrometry. *Clinical Microbiology and Infection,* **18**, 475-484

Dreisewerd, K. 2003. The desorption process in MALDI. *Chemical Reviews,* **103,** 395-426

Efron, B. and Tibshirani, R. J. 1994. *An introduction to the bootstrap*, Chapman and Hall/CRC press.

Eilers, P. H. C. 2004. Parametric Time Warping. *Analytical Chemistry,* **76**, 404-411

Ellis, D. I., Cowcher, D. P., Ashton, L., O'hagan, S. and Goodacre, R. 2013. Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool. *Analyst,* **138**, 3871-3884

Ellis, D. I. and Goodacre, R. 2006. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst,* **131**, 875-885

Ellis, D. I., Harrigan, G. G. and Goodacre, R. 2003. Metabolic fingerprinting with Fourier transform infrared spectroscopy. *Metabolic Profiling: its Role in Biomarker Discovery and Gene Function Analysis.* Verlag US, Springer

Engvall, E. 1977. Quantitative enzyme immunoassay (ELISA) in microbiology. *Medical Biology,* **55**, 193-200

Ferraro, J. R., Nakamoto, K. and Brown, C. W. 2003. *Introductory raman spectroscopy, 2nd Ed.*, Academic Press: London, pp.1-27

Franz, C. M., Stiles, M. E., Schleifer, K. H. and Holzapfel, W. H. 2003. Enterococci in foods—a conundrum for food safety. *International Journal of Food Microbiology,* **88**, 105-122

Freiwald, A. and Sauer, S. 2009. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols,* **4**, 732-742

Giebel, R., Worden, C., Rust, S. M., Kleinheinz, G. T., Robbins, M. and Sandrin, T. R. 2010. Microbial fingerprinting using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS): applications and challenges. *Advances in Applied Microbiology,* **71**, 149-84

Goodacre, R., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B. and Rooney, P. J. 1998. Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology,* **144**, 1157-1170

Goodacre, R., Radovic, B. S. and Anklam, E. 2002. Progress toward the rapid nondestructive assessment of the floral origin of European honey using dispersive Raman spectroscopy. *Applied Spectroscopy,* **56**, 521-527

Goodacre, R., Timmins, E. M., Rooney, P. J., Rowland, J. J. and Kell, D. B. 1996. Rapid identification of streptococcus and *enterococcus* species using diffuse reflectance-absorbance fourier transform infrared spectroscopy and artificial neural networks. *Federation of European Microbiological Societies Microbiology Letters,* **140**, 233-239

Gower, J. C. and Dijksterhuis, G. B. 2005. *Procrustes problems*, Oxford, Oxford University Press, Psychometrika, Vol. 70, NO. 4, 799–801

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L. and Goodacre, R. 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta,* **879**, 10-23

Gromski, P. S., Xu, Y., Correa, E., Ellis, D. I., Turner, M. L. and Goodacre, R. 2014. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Analytica Chimica Acta,* **829**, 1-8

Guibet, F., Amiel, C., Cadot, P., Cordevant, C., Desmonts, M. H., Lange, M., Marecat, A., Travert, J., Denis, C. and Mariey, L. 2003. Discrimination and classification of Enterococci by Fourier transform infrared (FT-IR) spectroscopy. *Vibrational Spectroscopy,* **33**, 133-142

Gutteridge, C. S., Valus, L. and Macfie, H. J. H. 1985. 14 - Numerical methods in the classification of micro-organisms by pyrolysis mass spectrometry. *In:* Priest, M. G. J. G. (ed.) *Computer-Assisted Bacterial Systematics.* London: Academic Press

Harrigan, G. G., Laplante, R. H., Cosma, G. N., Cockerell, G., Goodacre, R., Maddox, J. F., Luyendyk, J. P., Ganey, P. E. and Roth, R. A. 2004. Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicology Letters,* **146**, 197-205

Hastie, T., Tibshirani, R. and Friedman, J. 2009. The Elements of Statistical Learnin. New York, Springer, pp. 520

Hayes, J. R., English, L. L., Carter, P. J., Proescholdt, T., Lee, K. Y., Wagner, D. D. and White, D. G. 2003. Prevalence and antimicrobial resistance of enterococcus species strainsd from retail meats. *Applied and Environmental Microbiology,* **69**, 7153-7160

Helm, D., Labischinski, H., Schallehn, G. and Naumann, D. 1991. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Journal of General Microbiology,* **137**, 69-79

Huang, W. E., Li, M., Jarvis, R. M., Goodacre, R. and Banwart, S. A. 2010. Shining Light on the Microbial World: The application of Raman microspectroscopy. *In: Advances in Applied Microbiology,* **70**, 153-186

Jarvis, R. M. and Goodacre, R. 2008. Characterisation and identification of bacteria using SERS. *Chemical Society Reviews,* **37**, 931-936

Kayser, F. H., Bienz, K. A. and Eckert, J. 2011. *Medical microbiology*, New York, Thieme Medical Publishers, pp.234-235

Kim, S., Reuhs, B. L. and Mauer, L. J. 2005. Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. *Journal of Applied Microbiology,* **99**, 411-417

Kirschner, C., Maquelin, K., Pina, P., Thi, N. N., Choo-Smith, L.-P., Sockalingum, G., Sandt, C., Ami, D., Orsini, F. and Doglia, S. 2001. Classification and identification of enterococci: a comparative phenotypic, genotypic, and vibrational spectroscopic study. *Journal of Clinical Microbiology,* **39**, 1763-1770

Ke, D., Picard, F. J., Martineau, F., Ménard, C., Roy, P. H., Ouellette, M. and Bergeron, M. G. 1999. Development of a PCR assay for rapid detection of enterococci. *Journal of Clinical Microbiology,* **37**, 3497-3503

Lasch, P., Fleige, C., Stämmler, M., Layer, F., Nübel, U., Witte, W. and Werner, G. 2014. Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus faecium* and *Staphylococcus aureus* strains. *Journal of Microbiological Methods,* **100**, 58-69

Lay, J. O. 2001. MALDI-TOF mass spectrometry of bacteria. *Mass Spectrometry Reviews,* **20**, 172-194

Lay, J. O. 2000. MALDI-TOF mass spectrometry and bacterial taxonomy. *Trac Trend Analytical Chemistry,* **19**, 507-516

López-Díez, E. C. and Goodacre, R. 2004. Characterisation of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry,* **76**, 585-591

Manly, B. F. 2004. *Multivariate statistical methods: a primer,* CRC Press, pp. 12-17

Mariey, L., Signolle, J. P., Amiel, C. and Travert, J. 2001. Discrimination, classification, identification of microorganisms using FT-IR spectroscopy and chemometrics. *Vibrational Spectroscopy,* **26**, 151-159

Marvin, L. F., Roberts, M. A. and Fay, L. B. 2003. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in clinical chemistry. *Clinica Chimica Acta,* **337**, 11-21

Mccracken, M., Wong, A., Mitchell, R., Gravel, D., Conly, J., Embil, J., Johnston, L., Matlow, A., Ormiston, D. and Simor, A. 2013. Molecular epidemiology of vancomycin-resistant enterococcal bacteraemia: results from the Canadian Nosocomial Infection Surveillance Program, 1999–2009. *Journal of Antimicrobial Chemotherapy,* **68**, 1505–9

Muhamadali, H., Chisanga, M., Subaihi, A. and Goodacre, R. 2015. Combining Raman and FT-IR Spectroscopy with Quantitative Isotopic Labeling for Differentiation of E. coli Cells at Community and Single Cell Levels. *Analytical Chemistry,* **87**, 4578-4586

Nabiev, I., Chourpa, I. and Manfait, M. 1994. Applications of Raman and surface-enhanced Raman scattering spectroscopy in medicine. *Journal of Raman Spectroscopy,* **25**, 13-23

Naumann, D. 1984. Some ultrastructural information on intact, living bacterial cells and related cell-wall fragments as given by FT-IR. *Infrared Physics,* **24**, 233-238

Naumann, D., Fijala, V., Labischinski, H. and Giesbrecht, P. 1988. The rapid differentiation and identification of pathogenic bacteria using Fourier transform infrared spectroscopic and multivariate statistical analysis. *Journal of Molecular Structure,* **174**, 165-170

Naumann, D., Helm, D. and Labischinski, H. 1991. Microbiological characterisations by FT-IR spectroscopy. *Nature,* **351**, 81-82

Patel, S. A., Currie, F., Thakker, N. and Goodacre, R. 2008. Spatial metabolic fingerprinting using FT-IR spectroscopy: investigating abiotic stresses on Micrasterias hardyi. *Analyst,* **133**, 1707-1713

Quintela-Baluja, M., Böhme, K., Fernández-No, I. C., Morandi, S., Alnakip, M. E., Caamaño-Antelo, S., Barros-Velázquez, J. and Calo-Mata, P. 2013. Characterisation of different food-strainsd *Enterococcus* strains by MALDI-TOF mass fingerprinting. *Electrophoresis,* **34**, 2240-2250

Risch, M., Radjenovic, D., Han, J. N., Wydler, M., Nydegger, U. and Risch, L. 2010. Comparison of MALDI TOF with conventional identification of clinically relevant bacteria. *Swiss Medical Weekly,* **140**, w13095

Reen, D. 1994. Enzyme-Linked Immunosorbent Assay (ELISA). *In:* WALKER, J. (ed.) *Basic Protein and Peptide Protocols.* Pp. 461-466, Humana Press

Sauer, S. and Kliem, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* **8**, 74-82

Schleifer, K. H. and Kilpper-Bälz, R. 1984. Transfer of *Streptococcus faecalis* and *Streptococcus faecium* to the Genus *Enterococcu*s nom. rev. as *Enterococcus faecalis* comb. nov. and *Enterococcus faecium* comb. nov. *International Journal of Systematic Bacteriology,* **34**, 31-34

Smith, E. and Dent, G. 2013. *Modern Raman spectroscopy: a practical approach*, John Wiley and Son, pp.1-80

Shu, X., Li, Y., Liang, M., Yang, B., Liu, C., Wang, Y. and Shu, J. 2012. Rapid lipid profiling of bacteria by online MALDI-TOF mass spectrometry. *International Journal of Mass Spectrometry,* **321–322**, 71-76

Turabelidze, D., Kotetishvili, M., Kreger, A., Morris, J. G. and Sulakvelidze, A. 2000. Improved pulsed-field gel electrophoresis for typing vancomycin-resistant enterococci. *Journal of Clinical Microbiology,* **38**, 4242-4245

Uzunbajakava, N., Lenferink, A., Kraan, Y., Willekens, B., Vrensen, G., Greve, J. and Otto, C. 2003. Nonresonant Raman imaging of protein distribution in single human cells. *Biopolymers,* **72**, 1-9

Van De Vossenberg, J., Tervahauta, H., Maquelin, K., Blokker-Koopmans, C. H., Uytewaal-Aarts, M., Van Der Kooij, D., Van Wezel, A. P. and Van Der Gaag, B. 2013. Identification of bacteria in drinking water with Raman spectroscopy. *Analytical Methods,* **5**, 2679-2687

Wang, Y., Veltkamp, D. J. and Kowalski, B. R. 1991. Multivariate instrument standardization. *Analytical Chemistry,* **63**, 2750-2756

Winder, C. L., Gordon, S. V., Dale, J., Hewinson, R. G. and Goodacre, R. 2006. Metabolic fingerprints of Mycobacterium bovis cluster with molecular type: implications for genotype–phenotype links. *Microbiology,* **152**, 2757-2765

Winder, C. L. and Goodacre, R., 2004. Comparison of diffuse-reflectance absorbance and attenuated total reflectance FT-IR for the discrimination of bacteria. *Aanlyst*, **129**, 1118–1122

Williams, T. L., Andrzejewski, D., Lay Jr, J. O. and Musser, S. M. 2003. Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. *Journal of the American Society for Mass Spectrometry,* **14,** 342-351

Woodford, N. 1998. Glycopeptide-resistant enterococci: a decade of experience. *Journal of Medical Microbiology,* **47**, 849-862

Yolken, R. H. 1980. Enzyme-linked immunosorbent assay (ELISA): a practical tool for rapid diagnosis of viruses and other infectious agents. *The Yale Journal of Biology and Medicine,* **53**, 85-92

5.7 Supplementary information

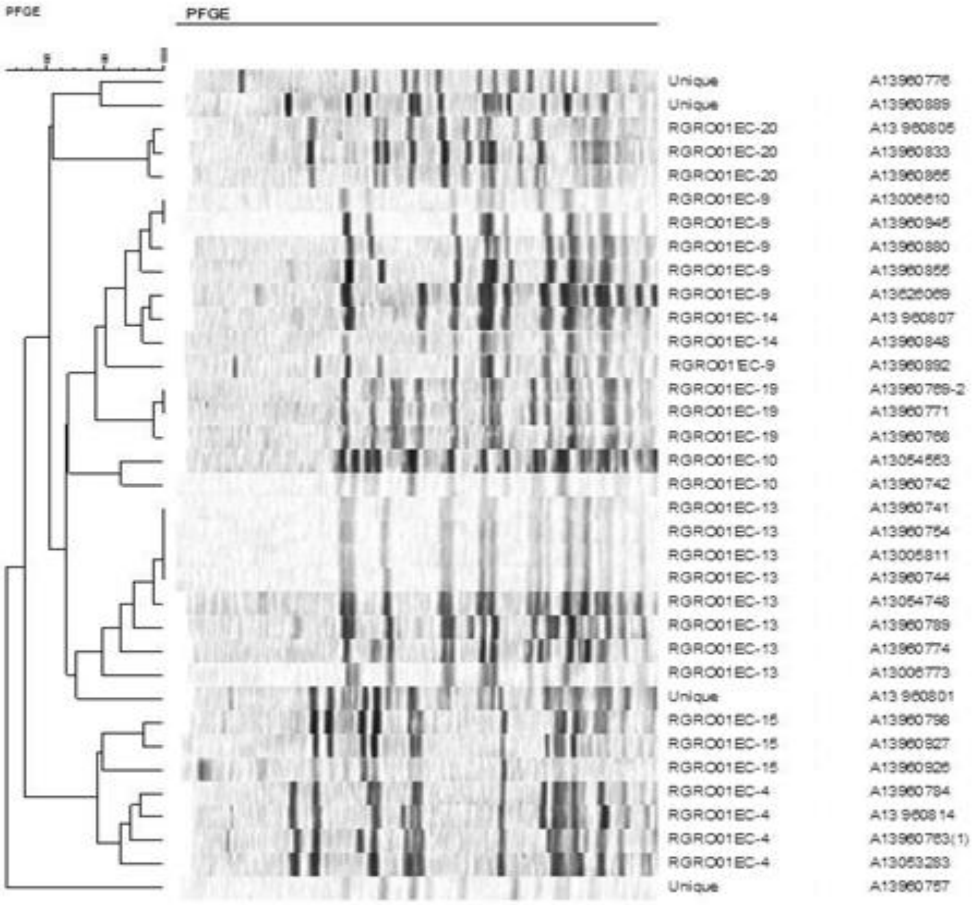


Figure S 5.1: Dendrogram generated from pulsed field gel electrophoresis (PFGE) of the 35 enterococci isolates. The top strain A13960776 is strain 178; the others follow in the sequence: 214, 192, 198, 204, 160, 233, 211, 205, 133, 194, 203, 219, 174, 175, 173, 139, 151, 154, 155, 149, 152, 144, 185, 177, 167, 191, 190, 223, 224, 179, 193, 170, 109, and 156.

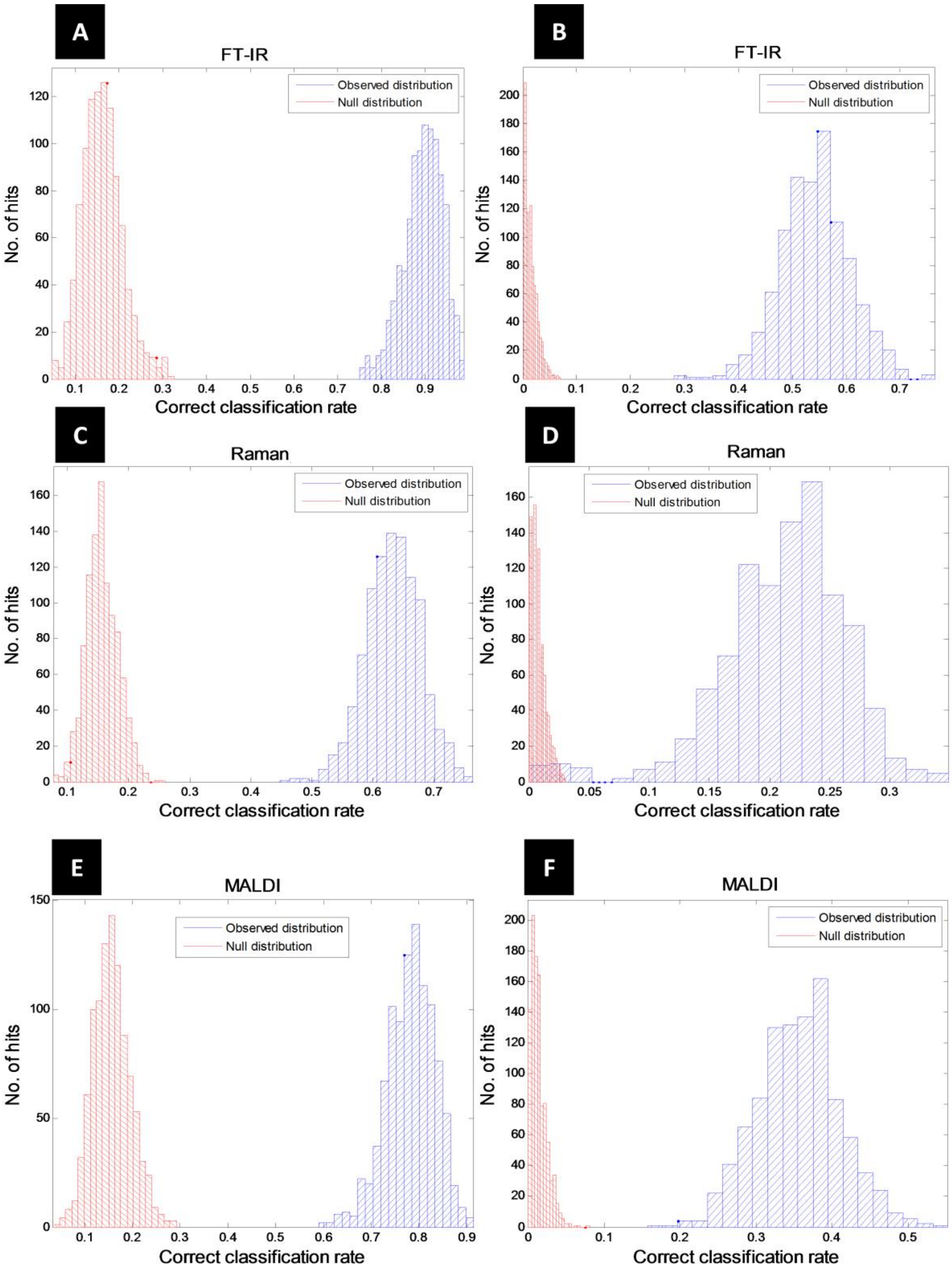


Figure S 5.2: The predictive accuracies expressed as correct classification rates (CCRs) generated from FT-IR spectroscopy data (A-B), Raman spectroscopy data (C-D) and MALDI-TOF-MS data (E-F) based on 1,000 bootstrapping re-sampling (blue bars). The null distribution (red bars) was obtained by permuting the order of the labels and conducting the same PLS-DA classification procedure. Not a single case out of 1,000 bootstrap cases had a model using permuted labels that obtained a higher CCR than the one using the known labels (A, C and E) at the strain level (12 classes) based on FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS data; the mean CCRs = 89.4%, 69.3% and 78.2%, respectively. At the isolate level (35 classes), based on FT-IR spectroscopy, Raman spectroscopy and MALDI-TOF-MS data, the mean CCRs were 54.3%, 21.1%, and 35.7%, respectively.

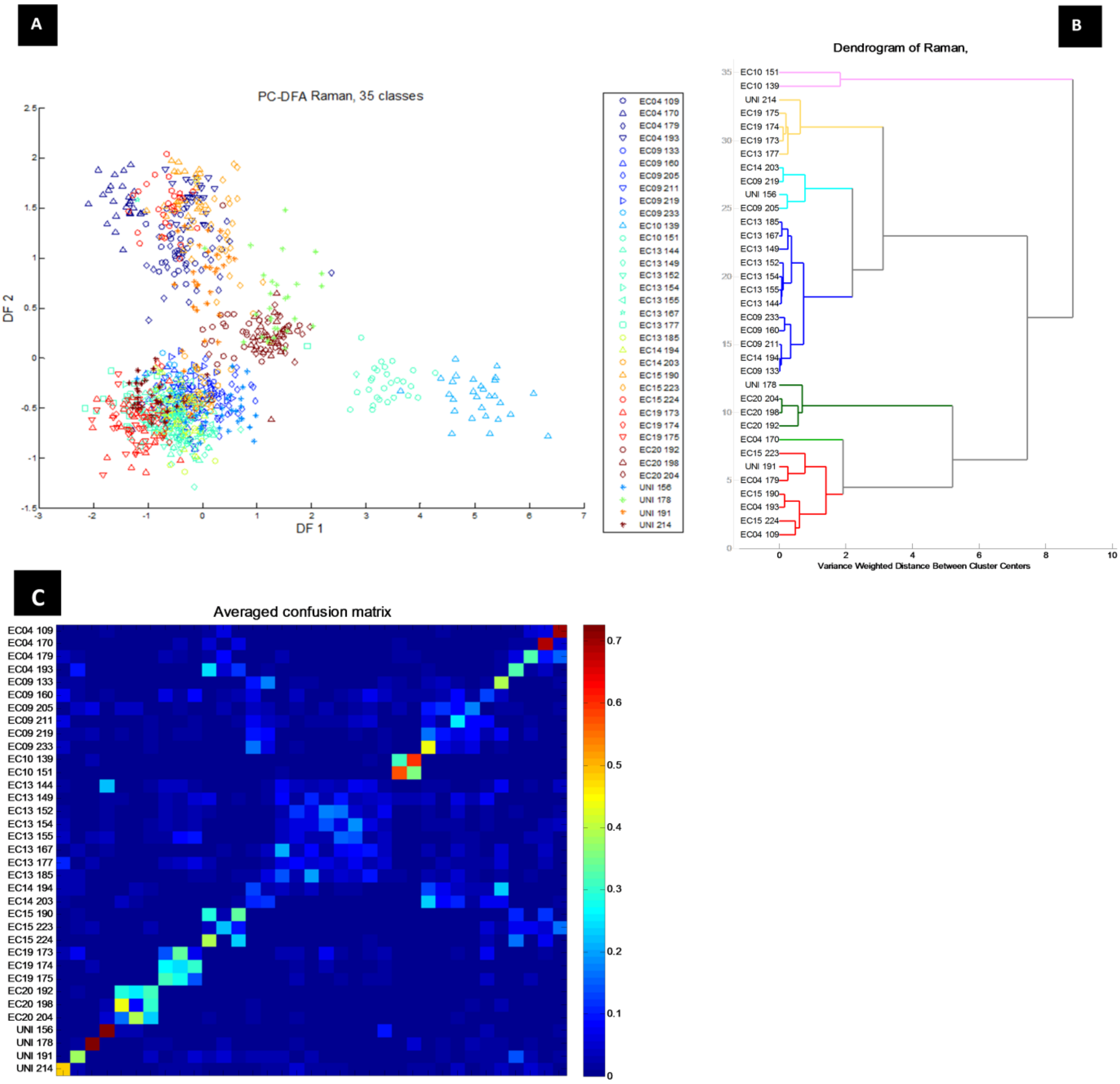


Figure S 5.3: (A) PC-DFA plot of Raman spectroscopy data after pre-processing illusting the relationship between the 35 isolates. (B) Hierarchical cluster analysis on averaged PC-DFA scores (35 classes), using the Ward’s linkage algorithm. (C) PLS-DA trained on 35 classes (i.e. 35 isolates) generated from Raman spectroscopy data.

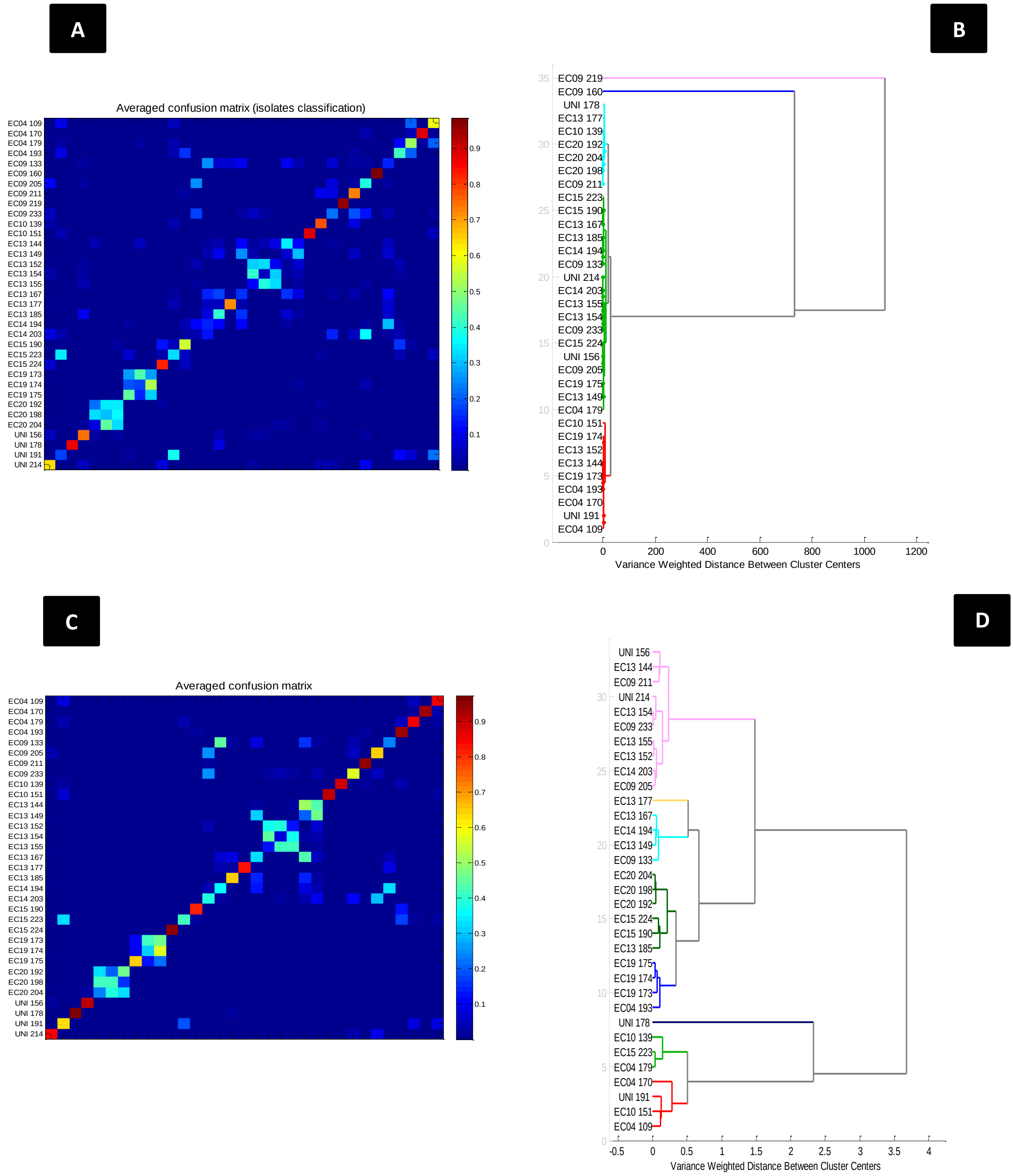


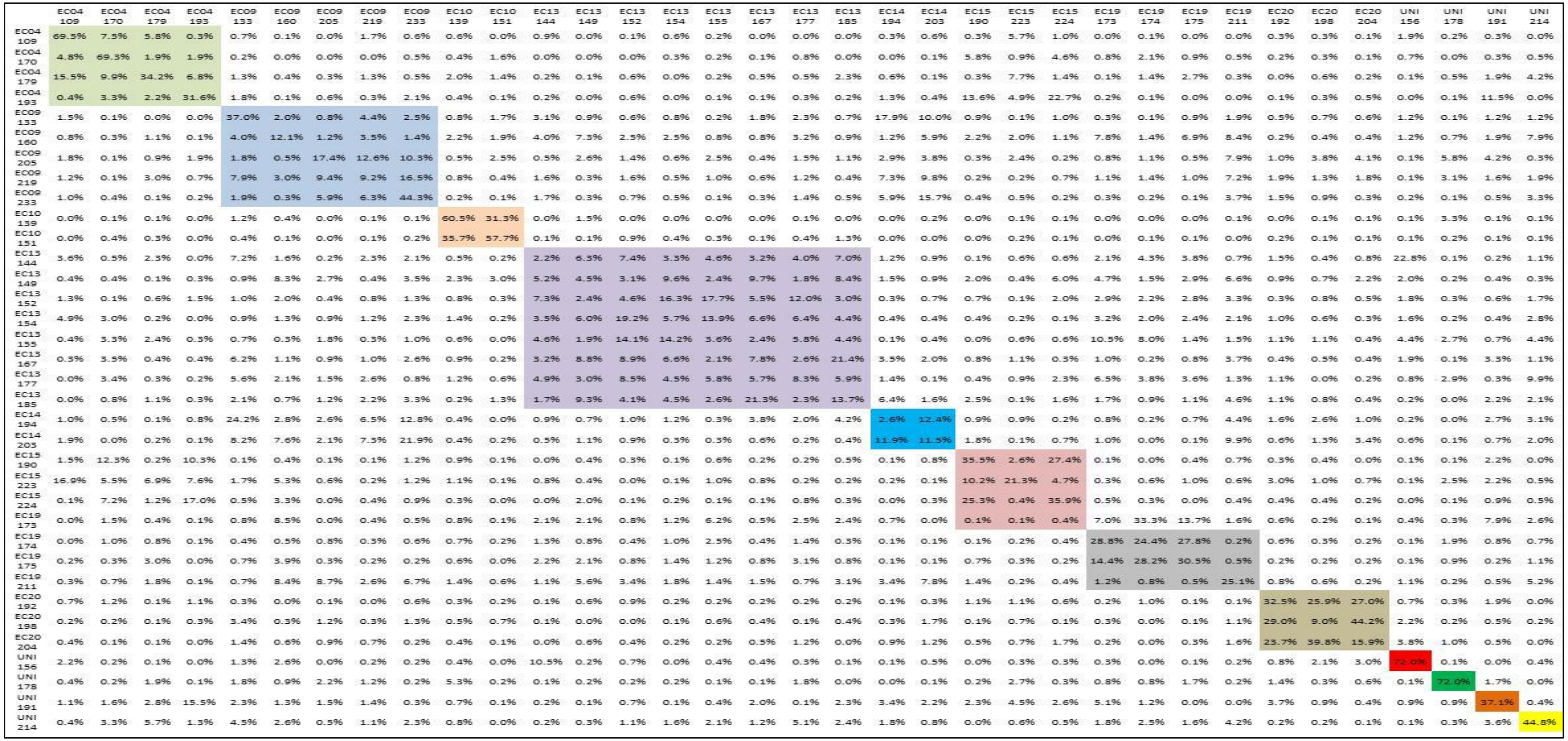
Figure S 5.4: (A) PLS-DA trained based on MALDI-TOF-MS data for the 35 strains (i.e. 35 classes). (B) Hierarchical cluster analysis based on averaged DFA scores of 35 isolates (i.e. 35 classes) using Ward's linkage. (C) PLS-DA results trained based on MALDI-TOF-MS data for 33 isolates (i.e. 33 classes) where species 160 and 219 were removed. (D) Hierarchical cluster analysis based on averaged mean DFA scores of the 33 isolates (i.e. 33 classes) using Ward's linkage.

Table S 5.1: The prediction accuracies of the 35 enterococci isolates using a PLS-DA model generated from FT-IR spectroscopy data

	EC04 109	EC04 170	EC04 179	EC04 193	EC09 133	EC09 160	EC09 205	EC09 211	EC09 219	EC09 233	EC10 139	EC10 151	EC13 144	EC13 149	EC13 152	EC13 154	EC13 155	EC13 167	EC13 177	EC13 185	EC14 194	EC14 203	EC15 190	EC15 223	EC15 224	EC19 173	EC19 174	EC19 175	EC20 192	EC20 198	EC20 204	UNI 156	UNI 178	UNI 191	UNI 214
EC04 109	95.0%	0.8%	0.9%	1.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.3%	0.0%	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
EC04 170	0.2%	96.4%	0.1%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.1%	0.0%	0.6%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%
EC04 179	1.5%	0.2%	76.3%	0.4%	0.0%	0.1%	0.0%	0.8%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.1%	0.3%	0.1%	0.2%	0.0%	0.0%	0.0%	15.6%	0.3%	0.0%	0.1%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	2.9%	0.1%
EC04 193	0.0%	0.3%	0.0%	76.8%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%	1.8%	0.6%	0.1%	15.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%	0.0%
EC09 133	0.1%	0.0%	0.0%	0.0%	78.3%	1.2%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	6.0%	9.4%	0.0%	0.0%	0.0%	3.0%	0.0%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC09 160	0.0%	0.0%	0.4%	0.0%	2.0%	79.3%	0.5%	3.2%	0.0%	0.8%	0.0%	0.0%	0.0%	0.3%	3.9%	1.1%	0.4%	0.2%	0.1%	0.1%	0.9%	5.4%	0.0%	0.4%	0.0%	0.1%	0.0%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%
EC09 205	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	64.9%	0.9%	28.3%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.8%	0.0%	0.4%	0.9%	0.1%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.5%	0.6%	0.0%	0.0%	0.2%	0.0%
EC09 211	0.0%	0.0%	0.8%	0.4%	0.0%	3.7%	1.3%	56.5%	1.3%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.2%	2.1%	0.5%	12.0%	0.0%	0.0%	0.2%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	19.9%
EC09 219	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	35.5%	0.7%	50.9%	10.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.9%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	0.1%	0.1%
EC09 233	0.0%	0.0%	0.0%	0.0%	0.3%	0.1%	0.1%	0.3%	5.4%	82.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	3.1%	7.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.5%
EC10 139	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	95.6%	3.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%	0.1%
EC10 151	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.2%	96.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC13 144	0.0%	0.0%	0.2%	0.0%	7.8%	0.0%	0.0%	0.0%	0.1%	0.9%	0.0%	0.0%	27.1%	16.2%	18.8%	7.7%	12.2%	2.2%	1.2%	3.4%	0.2%	0.0%	0.0%	0.0%	0.0%	1.1%	0.8%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC13 149	0.0%	0.0%	0.0%	0.0%	14.8%	2.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	11.9%	15.4%	4.0%	1.0%	0.6%	40.8%	0.6%	8.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC13 152	0.0%	0.0%	0.1%	0.0%	0.0%	2.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.5%	0.4%	21.9%	54.3%	13.3%	0.2%	0.8%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC13 154	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%	0.0%	0.1%	0.2%	0.0%	0.0%	0.0%	5.8%	0.4%	43.1%	25.3%	23.0%	0.5%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
EC13 155	0.0%	0.0%	0.1%	0.0%	0.0%	1.2%	0.0%	0.4%	0.3%	0.0%	0.0%	0.0%	7.2%	3.3%	18.2%	30.6%	32.9%	0.9%	0.4%	0.5%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.4%	1.6%	0.9%	0.0%	0.0%	0.0%	0.0%	0.1%	0.9%
EC13 167	0.0%	0.0%	0.0%	0.0%	2.1%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.3%	26.7%	0.1%	0.5%	0.0%	49.7%	0.0%	19.7%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC13 177	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.8%	0.2%	0.0%	1.3%	0.0%	0.1%	0.9%	0.6%	5.4%	7.7%	0.2%	0.4%	77.5%	1.3%	0.0%	0.0%	0.1%	0.5%	0.3%	0.1%	0.0%	1.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
EC13 185	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%	0.0%	0.0%	0.0%	0.0%	0.3%	5.4%	0.0%	0.0%	0.1%	24.6%	0.6%	55.8%	8.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	2.1%	0.0%
EC14 194	0.0%	0.1%	0.0%	1.2%	3.2%	1.9%	0.4%	4.4%	0.3%	5.7%	0.0%	0.0%	3.9%	0.3%	0.0%	0.0%	0.0%	0.6%	0.0%	11.5%	32.3%	21.7%	0.3%	1.4%	1.8%	0.0%	0.1%	3.0%	0.0%	5.1%	0.1%	0.0%	0.0%	0.4%	0.0%
EC14 203	0.0%	0.0%	0.0%	0.2%	0.1%	1.2%	1.6%	9.5%	1.7%	18.7%	0.0%	0.0%	4.1%	0.0%	0.1%	0.1%	0.2%	1.1%	0.0%	0.7%	25.7%	33.1%	0.4%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.6%	0.3%	0.0%	0.0%	0.0%	0.0%
EC15 190	0.0%	0.2%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	0.2%	94.4%	0.1%	3.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC15 223	0.5%	0.0%	3.6%	2.1%	0.0%	0.4%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.6%	0.0%	0.0%	0.0%	0.3%	91.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%
EC15 224	0.0%	0.4%	0.0%	8.3%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.6%	0.0%	8.7%	0.3%	3.6%	0.1%	73.7%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC19 173	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	53.4%	41.9%	4.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC19 174	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	36.8%	55.8%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
EC19 175	0.0%	0.0%	0.4%	0.0%	0.0%	4.7%	0.2%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.5%	0.3%	0.1%	0.1%	0.1%	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	15.3%	14.6%	62.2%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
EC20 192	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.8%	0.1%	1.6%	0.0%	0.0%	0.5%	38.8%	23.4%	32.3%	0.0%	0.0%	0.0%	0.1%
EC20 198	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	0.2%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	5.7%	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	18.2%	25.8%	48.1%	0.0%	0.0%	0.0%	0.0%
EC20 204	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%	0.0%	0.1%	0.2%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.4%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	28.0%	50.8%	18.1%	0.0%	0.0%	0.0%	0.0%
UNI 156	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.3%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.3%	0.0%	0.1%	97.9%	0.0%	0.0%	0.0%
UNI 178	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	98.0%	0.2%	0.0%
UNI 191	0.0%	0.0%	0.6%	3.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	1.5%	0.0%	3.1%	0.1%	0.0%	0.0%	1.1%	0.2%	0.1%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	88.5%	0.2%
UNI 214	0.0%	0.1%	0.4%	0.0%	0.0%	2.5%	0.0%	19.0%	1.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%	0.3%	0.2%	1.7%	0.0%	0.0%	0.1%	0.0%	0.0%	0.1%	0.4%	0.0%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	5.2%	67.6%

The different colours represent the strain level identification

Table S 5.2: The prediction accuracies of the 35 enterococci isolates using a PLS-DA model generated from Raman spectroscopy data



	EC04 109	EC04 170	EC04 179	EC04 193	EC09 133	EC09 160	EC09 205	EC09 219	EC09 233	EC10 139	EC10 151	EC13 144	EC13 149	EC13 152	EC13 154	EC13 155	EC13 167	EC13 177	EC13 185	EC14 194	EC14 203	EC15 190	EC15 223	EC15 224	EC19 173	EC19 174	EC19 175	EC19 211	EC20 192	EC20 198	EC20 204	UNI 156	UNI 178	UNI 191	UNI 214
EC04 109	69.5%	7.5%	5.8%	0.3%	0.7%	0.1%	0.0%	1.7%	0.6%	0.6%	0.0%	0.9%	0.0%	0.1%	0.6%	0.2%	0.0%	0.0%	0.0%	0.3%	0.6%	0.3%	5.7%	1.0%	0.0%	0.1%	0.0%	0.0%	0.3%	0.3%	0.1%	1.9%	0.2%	0.3%	0.0%
EC04 170	4.8%	69.3%	1.9%	1.9%	0.2%	0.0%	0.0%	0.0%	0.5%	0.4%	1.6%	0.0%	0.0%	0.0%	0.3%	0.2%	0.1%	0.8%	0.0%	0.0%	0.1%	5.8%	0.9%	4.6%	0.8%	2.1%	0.9%	0.5%	0.2%	0.3%	0.1%	0.7%	0.0%	0.3%	0.5%
EC04 179	15.5%	9.9%	34.2%	6.8%	1.3%	0.4%	0.3%	1.3%	0.5%	2.0%	1.4%	0.2%	0.1%	0.6%	0.0%	0.2%	0.5%	0.5%	2.3%	0.6%	0.1%	0.3%	7.7%	1.4%	0.1%	1.4%	2.7%	0.3%	0.0%	0.6%	0.2%	0.1%	0.5%	1.9%	4.2%
EC04 193	0.4%	3.3%	2.2%	31.6%	1.8%	0.1%	0.6%	0.3%	2.1%	0.4%	0.1%	0.2%	0.0%	0.6%	0.0%	0.1%	0.1%	0.3%	0.2%	1.3%	0.4%	13.6%	4.9%	22.7%	0.2%	0.1%	0.0%	0.0%	0.1%	0.3%	0.5%	0.0%	0.1%	11.5%	0.0%
EC09 133	1.5%	0.1%	0.0%	0.0%	37.0%	2.0%	0.8%	4.4%	2.5%	0.8%	1.7%	3.1%	0.9%	0.6%	0.8%	0.2%	1.8%	2.3%	0.7%	17.9%	10.0%	0.9%	0.1%	1.0%	0.3%	0.1%	0.9%	1.9%	0.5%	0.7%	0.6%	1.2%	0.1%	1.2%	1.2%
EC09 160	0.8%	0.3%	1.1%	0.1%	4.0%	12.1%	1.2%	3.5%	1.4%	2.2%	1.9%	4.0%	7.3%	2.5%	2.5%	0.8%	0.8%	3.2%	0.9%	1.2%	5.9%	2.2%	2.0%	1.1%	7.8%	1.4%	6.9%	8.4%	0.2%	0.4%	0.4%	1.2%	0.7%	1.9%	7.9%
EC09 205	1.8%	0.1%	0.9%	1.9%	1.8%	0.5%	17.4%	12.6%	10.3%	0.5%	2.5%	0.5%	2.6%	1.4%	0.6%	2.5%	0.4%	1.5%	1.1%	2.9%	3.8%	0.3%	2.4%	0.2%	0.8%	1.1%	0.5%	7.9%	1.0%	3.8%	4.1%	0.1%	5.8%	4.2%	0.3%
EC09 219	1.2%	0.1%	3.0%	0.7%	7.9%	3.0%	9.4%	9.2%	16.5%	0.8%	0.4%	1.6%	0.3%	1.6%	0.5%	1.0%	0.6%	1.2%	0.4%	7.3%	9.8%	0.2%	0.2%	0.7%	1.1%	1.4%	1.0%	7.2%	1.9%	1.3%	1.8%	0.1%	3.1%	1.6%	1.9%
EC09 233	1.0%	0.4%	0.1%	0.2%	1.9%	0.3%	5.9%	6.3%	44.3%	0.2%	0.1%	1.7%	0.3%	0.7%	0.5%	0.1%	0.3%	1.4%	0.5%	5.9%	15.7%	0.4%	0.5%	0.2%	0.3%	0.2%	0.1%	3.7%	1.5%	0.9%	0.3%	0.2%	0.1%	0.5%	3.3%
EC10 139	0.0%	0.1%	0.1%	0.0%	1.2%	0.4%	0.0%	0.1%	0.1%	60.5%	31.3%	0.0%	1.5%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.2%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	0.1%	0.1%	3.3%	0.1%	0.1%
EC10 151	0.0%	0.4%	0.3%	0.0%	0.4%	0.1%	0.0%	0.1%	0.2%	35.7%	57.7%	0.1%	0.1%	0.9%	0.4%	0.3%	0.1%	0.4%	1.3%	0.0%	0.0%	0.0%	0.2%	0.1%	0.0%	0.1%	0.1%	0.0%	0.2%	0.1%	0.1%	0.1%	0.2%	0.1%	0.1%
EC13 144	3.6%	0.5%	2.3%	0.0%	7.2%	1.6%	0.2%	2.3%	2.1%	0.5%	0.2%	2.2%	6.3%	7.4%	3.3%	4.6%	3.2%	4.0%	7.0%	1.2%	0.9%	0.1%	0.6%	0.6%	2.1%	4.3%	3.8%	0.7%	1.5%	0.4%	0.8%	22.8%	0.1%	0.2%	1.1%
EC13 149	0.4%	0.4%	0.1%	0.3%	0.9%	8.3%	2.7%	0.4%	3.5%	2.3%	3.0%	5.2%	4.5%	3.1%	9.6%	2.4%	9.7%	1.8%	8.4%	1.5%	0.9%	2.0%	0.4%	6.0%	4.7%	1.5%	2.9%	6.6%	0.9%	0.7%	2.2%	2.0%	0.2%	0.4%	0.3%
EC13 152	1.3%	0.1%	0.6%	1.5%	1.0%	2.0%	0.4%	0.8%	1.3%	0.8%	0.3%	7.3%	2.4%	4.6%	16.3%	17.7%	5.5%	12.0%	3.0%	0.3%	0.7%	0.7%	0.1%	2.0%	2.9%	2.2%	2.8%	3.3%	0.3%	0.8%	0.5%	1.8%	0.3%	0.6%	1.7%
EC13 154	4.9%	3.0%	0.2%	0.0%	0.9%	1.3%	0.9%	1.2%	2.3%	1.4%	0.2%	3.5%	6.0%	19.2%	5.7%	13.9%	6.6%	6.4%	4.4%	0.4%	0.4%	0.4%	0.2%	0.1%	3.2%	2.0%	2.4%	2.1%	1.0%	0.6%	0.3%	1.6%	0.2%	0.4%	2.8%
EC13 155	0.4%	3.3%	2.4%	0.3%	0.7%	0.3%	1.8%	0.3%	1.0%	0.6%	0.0%	4.6%	1.9%	14.1%	14.2%	3.6%	2.4%	5.8%	4.4%	0.1%	0.4%	0.0%	0.6%	0.6%	10.5%	8.0%	1.4%	1.5%	1.1%	1.1%	0.4%	4.4%	2.7%	0.7%	4.4%
EC13 167	0.3%	3.5%	0.4%	0.4%	6.2%	1.1%	0.9%	1.0%	2.6%	0.9%	0.2%	3.2%	8.8%	8.9%	6.6%	2.1%	7.8%	2.6%	21.4%	3.5%	2.0%	0.8%	1.1%	0.3%	1.0%	0.2%	0.8%	3.7%	0.4%	0.5%	0.4%	1.9%	0.1%	3.3%	1.1%
EC13 177	0.0%	3.4%	0.3%	0.2%	5.6%	2.1%	1.5%	2.6%	0.8%	1.2%	0.6%	4.9%	3.0%	8.5%	4.5%	5.8%	5.7%	8.3%	5.9%	1.4%	0.1%	0.4%	0.9%	2.3%	6.5%	3.8%	3.6%	1.3%	1.1%	0.0%	0.2%	0.8%	2.9%	0.3%	9.9%
EC13 185	0.0%	0.8%	1.1%	0.3%	2.1%	0.7%	1.2%	2.2%	3.3%	0.2%	1.3%	1.7%	9.3%	4.1%	4.5%	2.6%	21.3%	2.3%	13.7%	6.4%	1.6%	2.5%	0.1%	1.6%	1.7%	0.9%	1.1%	4.6%	1.1%	0.8%	0.4%	0.2%	0.0%	2.2%	2.1%
EC14 194	1.0%	0.5%	0.1%	0.8%	24.2%	2.8%	2.6%	6.5%	12.8%	0.4%	0.0%	0.9%	0.7%	1.0%	1.2%	0.3%	3.8%	2.0%	4.2%	2.6%	12.4%	0.9%	0.9%	0.2%	0.8%	0.2%	0.7%	4.4%	1.6%	2.6%	1.0%	0.2%	0.0%	2.7%	3.1%
EC14 203	1.9%	0.0%	0.2%	0.1%	8.2%	7.6%	2.1%	7.3%	21.9%	0.4%	0.2%	0.5%	1.1%	0.9%	0.3%	0.3%	0.6%	0.2%	0.4%	11.9%	11.5%	1.8%	0.1%	0.7%	1.0%	0.0%	0.1%	9.9%	0.6%	1.3%	3.4%	0.6%	0.1%	0.7%	2.0%
EC15 190	1.5%	12.3%	0.2%	10.3%	0.1%	0.4%	0.1%	0.1%	1.2%	0.9%	0.1%	0.0%	0.4%	0.3%	0.1%	0.6%	0.2%	0.2%	0.5%	0.1%	0.8%	35.5%	2.6%	27.4%	0.1%	0.0%	0.4%	0.7%	0.3%	0.4%	0.0%	0.1%	0.1%	2.2%	0.0%
EC15 223	16.9%	5.5%	6.9%	7.6%	1.7%	5.3%	0.6%	0.2%	1.2%	1.1%	0.1%	0.8%	0.4%	0.0%	0.1%	1.0%	0.8%	0.2%	0.2%	0.2%	0.1%	10.2%	21.3%	4.7%	0.3%	0.6%	1.0%	0.6%	3.0%	1.0%	0.7%	0.1%	2.5%	2.2%	0.5%
EC15 224	0.1%	7.2%	1.2%	17.0%	0.5%	3.3%	0.0%	0.4%	0.9%	0.3%	0.0%	0.0%	2.0%	0.1%	0.2%	0.1%	0.1%	0.8%	0.3%	0.0%	0.3%	25.3%	0.4%	35.9%	0.5%	0.3%	0.0%	0.4%	0.4%	0.4%	0.2%	0.0%	0.1%	0.9%	0.5%
EC19 173	0.0%	1.5%	0.4%	0.1%	0.8%	8.5%	0.0%	0.4%	0.5%	0.8%	0.1%	2.1%	2.1%	0.8%	1.2%	6.2%	0.5%	2.5%	2.4%	0.7%	0.0%	0.1%	0.1%	0.4%	7.0%	33.3%	13.7%	1.6%	0.6%	0.2%	0.1%	0.4%	0.3%	7.9%	2.6%
EC19 174	0.0%	1.0%	0.8%	0.1%	0.4%	0.5%	0.8%	0.3%	0.6%	0.7%	0.2%	1.3%	0.8%	0.4%	1.0%	2.5%	0.4%	1.4%	0.3%	0.1%	0.1%	0.1%	0.2%	0.4%	28.8%	24.4%	27.8%	0.2%	0.6%	0.3%	0.2%	0.1%	1.9%	0.8%	0.7%
EC19 175	0.2%	0.3%	3.0%	0.0%	0.7%	3.9%	0.3%	0.2%	0.2%	0.6%	0.0%	2.2%	2.1%	0.8%	1.4%	1.2%	0.8%	3.1%	0.8%	0.1%	0.1%	0.7%	0.3%	0.2%	14.4%	28.2%	30.5%	0.5%	0.2%	0.2%	0.2%	0.1%	0.9%	0.2%	1.1%
EC19 211	0.3%	0.7%	1.8%	0.1%	0.7%	8.4%	8.7%	2.6%	6.7%	1.4%	0.6%	1.1%	5.6%	3.4%	1.8%	1.4%	1.5%	0.7%	3.1%	3.4%	7.8%	1.4%	0.2%	0.4%	1.2%	0.8%	0.5%	25.1%	0.8%	0.6%	0.2%	1.1%	0.2%	0.5%	5.2%
EC20 192	0.7%	1.2%	0.1%	1.1%	0.3%	0.0%	0.1%	0.0%	0.6%	0.3%	0.2%	0.1%	0.6%	0.9%	0.2%	0.2%	0.2%	0.2%	0.2%	0.1%	0.3%	1.1%	1.1%	0.6%	0.2%	1.0%	0.1%	0.1%	32.5%	25.9%	27.0%	0.7%	0.3%	1.9%	0.0%
EC20 198	0.2%	0.2%	0.1%	0.3%	3.4%	0.3%	1.2%	0.3%	1.3%	0.5%	0.7%	0.1%	0.0%	0.0%	0.1%	0.6%	0.4%	0.1%	0.4%	0.3%	1.7%	0.1%	0.7%	0.1%	0.3%	0.0%	0.1%	1.1%	29.0%	9.0%	44.2%	2.2%	0.2%	0.5%	0.2%
EC20 204	0.4%	0.1%	0.1%	0.0%	1.4%	0.6%	0.9%	0.7%	0.2%	0.4%	0.1%	0.0%	0.6%	0.4%	0.2%	0.2%	0.5%	1.2%	0.0%	0.9%	1.2%	0.5%	0.7%	1.7%	0.2%	0.0%	0.3%	1.6%	23.7%	39.8%	15.9%	3.8%	1.0%	0.5%	0.0%
UNI 156	2.2%	0.2%	0.1%	0.0%	1.3%	2.6%	0.0%	0.2%	0.2%	0.4%	0.0%	10.5%	0.2%	0.7%	0.0%	0.4%	0.4%	0.3%	0.1%	0.1%	0.5%	0.0%	0.3%	0.3%	0.3%	0.0%	0.1%	0.2%	0.8%	2.1%	3.0%	72.0%	0.1%	0.0%	0.4%
UNI 178	0.4%	0.2%	1.9%	0.1%	1.8%	0.9%	2.2%	1.2%	0.2%	5.3%	0.2%	0.1%	0.2%	0.2%	0.2%	0.1%	0.1%	1.8%	0.0%	0.0%	0.1%	0.2%	2.7%	0.3%	0.8%	0.8%	1.7%	0.2%	1.4%	0.3%	0.6%	0.1%	72.0%	1.7%	0.0%
UNI 191	1.1%	1.6%	2.8%	15.5%	2.3%	1.3%	1.5%	1.4%	0.3%	0.7%	0.1%	0.2%	0.1%	0.7%	0.1%	0.4%	2.0%	0.1%	2.3%	3.4%	2.2%	2.3%	4.5%	2.6%	5.1%	1.2%	0.0%	0.0%	3.7%	0.9%	0.4%	0.9%	0.9%	37.1%	0.4%
UNI 214	0.4%	3.3%	5.7%	1.3%	4.5%	2.6%	0.5%	1.1%	2.3%	0.8%	0.0%	0.2%	0.3%	1.1%	1.6%	2.1%	1.2%	5.1%	2.4%	1.8%	0.8%	0.0%	0.6%	0.5%	1.8%	2.5%	1.6%	4.2%	0.2%	0.2%	0.1%	0.1%	0.3%	3.6%	44.8%

The different colours represent the strain level identification

Table S 5.3: The prediction accuracies of the 35 enterococci isolates using a PLS-DA model generated from MALDI-TOF-MS data

	EC04 109	EC04 170	EC04 179	EC04 193	EC09 133	EC09 160	EC09 205	EC09 211	EC09 219	EC09 233	EC10 139	EC10 151	EC13 144	EC13 149	EC13 152	EC13 154	EC13 155	EC13 167	EC13 177	EC13 185	EC14 194	EC14 203	EC15 190	EC15 223	EC15 224	EC19 173	EC19 174	EC19 175	EC20 192	EC20 198	EC20 204	UNI 156	UNI 178	UNI 191	UNI 214
EC04 109	59.6%	1.1%	20.9%	0.8%	0.3%	0.0%	0.1%	0.1%	0.0%	0.5%	0.3%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	5.5%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	9.5%	0.0%
EC04 170	1.9%	87.0%	4.0%	0.0%	0.0%	0.0%	3.7%	0.1%	0.0%	0.0%	0.0%	1.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.2%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%
EC04 179	23.1%	1.5%	49.9%	11.5%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	0.7%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.6%	0.0%	1.0%	0.0%	0.0%	0.5%	3.2%	0.2%	0.0%	2.0%	1.2%	0.0%	0.2%	0.1%	0.4%	0.0%	2.9%	0.1%
EC04 193	1.5%	0.0%	19.0%	42.0%	0.0%	0.0%	0.0%	2.7%	0.0%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.3%	0.8%	16.5%	4.2%	1.2%	0.0%	0.0%	0.0%	0.3%	1.0%	0.4%	0.0%	0.0%	9.5%	0.2%
EC09 133	0.1%	0.0%	0.3%	0.0%	15.5%	0.0%	2.2%	2.0%	0.0%	5.7%	0.4%	0.0%	2.9%	11.1%	0.3%	0.0%	0.0%	8.0%	8.0%	7.7%	27.5%	1.3%	1.2%	0.1%	0.7%	0.5%	0.0%	0.9%	0.0%	0.0%	0.1%	2.3%	0.0%	0.6%	0.5%
EC09 160	0.1%	0.3%	0.0%	0.1%	0.0%	97.2%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.5%	0.0%	0.2%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.2%	0.0%	0.0%	0.0%
EC09 205	0.0%	0.3%	0.1%	0.0%	2.0%	0.0%	36.4%	2.3%	0.0%	9.6%	0.0%	0.0%	0.1%	0.0%	0.2%	1.0%	1.4%	1.0%	0.0%	1.8%	1.1%	26.6%	1.4%	0.0%	0.1%	0.5%	0.1%	0.4%	0.0%	0.0%	0.0%	2.1%	0.0%	0.0%	11.4%
EC09 211	0.0%	0.0%	0.1%	0.4%	0.3%	0.0%	2.3%	69.8%	0.0%	10.1%	11.3%	0.0%	0.0%	0.0%	0.0%	0.4%	0.1%	0.2%	0.0%	0.1%	0.0%	1.9%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.3%
EC09 219	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.3%	0.0%	96.2%	0.0%	0.0%	0.1%	0.1%	0.0%	0.3%	0.0%	0.0%	0.1%	0.0%	0.9%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.3%	0.6%
EC09 233	2.1%	0.0%	0.0%	0.1%	4.9%	0.0%	14.6%	21.0%	0.0%	18.9%	0.3%	0.1%	1.0%	0.0%	1.2%	2.9%	5.4%	2.0%	0.0%	0.2%	0.7%	17.7%	0.2%	0.0%	1.3%	0.0%	0.1%	0.1%	0.2%	0.0%	0.0%	0.5%	0.0%	0.2%	4.4%
EC10 139	0.6%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	7.5%	0.0%	0.0%	80.0%	0.3%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	2.9%	0.0%	0.9%	0.4%	0.9%	0.2%	0.2%	0.9%	0.0%	0.0%	1.2%	2.9%
EC10 151	4.8%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.6%	89.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.3%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%	0.0%
EC13 144	0.0%	0.0%	0.0%	0.7%	4.8%	0.0%	0.3%	0.2%	0.0%	2.0%	0.1%	0.2%	9.8%	38.7%	8.2%	2.8%	2.1%	11.1%	0.0%	4.2%	1.7%	1.6%	0.2%	0.0%	0.7%	0.9%	5.4%	0.4%	0.1%	0.1%	3.2%	0.5%	0.0%	0.0%	0.2%
EC13 149	0.0%	0.0%	0.8%	0.0%	8.9%	0.0%	0.0%	4.8%	0.0%	0.7%	0.1%	0.0%	28.4%	6.3%	1.0%	2.1%	3.4%	24.6%	0.1%	13.0%	2.8%	0.0%	0.2%	0.1%	0.0%	0.0%	0.5%	0.3%	0.0%	0.0%	1.1%	0.1%	0.0%	0.1%	0.5%
EC13 152	0.1%	0.0%	0.0%	2.8%	0.1%	0.0%	0.6%	0.5%	0.0%	1.3%	0.5%	0.0%	5.7%	3.1%	10.7%	34.1%	29.7%	0.0%	0.0%	0.3%	0.0%	4.9%	0.3%	0.4%	0.0%	0.4%	1.3%	0.4%	0.0%	0.1%	0.1%	1.7%	0.0%	0.7%	0.3%
EC13 154	0.1%	0.0%	0.0%	3.4%	0.0%	0.0%	0.3%	1.1%	0.0%	1.0%	0.0%	0.0%	3.5%	2.1%	29.1%	5.9%	40.7%	0.8%	0.1%	0.6%	0.0%	1.0%	1.2%	0.0%	0.1%	1.0%	1.2%	0.5%	0.0%	0.2%	0.1%	2.7%	0.0%	0.0%	3.3%
EC13 155	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	2.1%	1.2%	0.0%	1.2%	0.0%	0.0%	1.2%	0.4%	32.6%	41.6%	9.7%	0.2%	0.0%	0.2%	0.0%	2.7%	0.0%	0.7%	0.0%	0.1%	0.2%	0.0%	0.0%	0.0%	0.0%	3.8%	0.0%	0.0%	0.9%
EC13 167	0.0%	0.0%	0.3%	0.0%	10.6%	0.0%	0.9%	3.7%	0.0%	2.4%	0.0%	0.0%	9.0%	19.1%	0.3%	0.1%	0.2%	13.0%	2.5%	21.3%	11.9%	0.3%	0.0%	1.6%	0.1%	0.1%	0.0%	0.8%	0.6%	0.0%	0.2%	0.9%	0.0%	0.0%	0.1%
EC13 177	0.1%	0.5%	1.2%	0.0%	6.1%	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	0.2%	0.1%	0.3%	0.0%	0.0%	0.0%	3.7%	70.4%	7.2%	6.1%	0.0%	0.1%	2.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.1%
EC13 185	0.1%	0.0%	0.9%	0.0%	5.5%	0.0%	0.6%	0.5%	0.0%	1.3%	0.0%	0.5%	1.7%	8.8%	0.3%	0.1%	0.1%	13.8%	0.2%	42.3%	9.3%	0.5%	1.3%	0.0%	0.0%	0.0%	0.0%	0.2%	0.3%	0.6%	0.2%	10.5%	0.0%	0.3%	0.0%
EC14 194	0.1%	0.0%	0.2%	1.8%	27.0%	0.0%	3.7%	0.9%	0.0%	0.4%	0.0%	0.0%	0.7%	3.7%	0.0%	0.3%	0.0%	10.2%	1.3%	13.4%	15.6%	12.5%	3.1%	0.1%	0.0%	1.0%	1.4%	1.6%	0.0%	0.3%	0.0%	0.2%	0.0%	0.4%	0.0%
EC14 203	0.0%	0.0%	0.0%	2.9%	2.1%	0.0%	36.5%	6.5%	0.0%	13.8%	0.0%	0.0%	0.7%	0.3%	0.5%	1.1%	0.3%	0.0%	0.0%	0.2%	12.9%	4.3%	2.8%	0.3%	0.0%	0.2%	0.2%	0.0%	0.0%	0.0%	0.5%	0.3%	0.0%	3.3%	10.2%
EC15 190	0.0%	0.0%	2.1%	16.3%	0.4%	0.0%	0.2%	1.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.3%	0.4%	0.1%	0.0%	0.0%	0.1%	1.4%	0.2%	57.0%	5.1%	12.9%	0.2%	0.0%	0.0%	1.4%	0.0%	0.0%	0.1%	0.0%	0.3%	0.0%
EC15 223	4.9%	0.4%	0.6%	3.6%	0.9%	0.0%	0.3%	0.1%	0.0%	0.7%	3.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	0.0%	0.0%	0.0%	4.9%	32.4%	3.3%	0.3%	0.0%	7.2%	0.5%	0.9%	0.2%	0.2%	0.0%	34.7%	0.0%
EC15 224	0.0%	0.2%	0.0%	0.3%	1.4%	0.0%	0.4%	0.0%	0.0%	0.5%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	5.7%	0.3%	81.9%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	1.1%	0.0%	0.0%	7.3%
EC19 173	0.0%	0.2%	0.0%	0.1%	0.2%	0.0%	0.4%	0.1%	0.0%	0.0%	0.6%	0.0%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%	0.1%	0.0%	24.6%	47.2%	24.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%	0.7%
EC19 174	0.4%	0.2%	0.1%	0.4%	0.0%	0.0%	3.3%	0.0%	0.0%	0.3%	0.2%	0.2%	1.5%	0.0%	0.3%	1.0%	0.1%	0.0%	0.0%	0.1%	0.0%	0.1%	0.1%	0.1%	0.0%	53.2%	20.1%	17.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%
EC19 175	0.3%	0.2%	0.0%	0.0%	0.8%	0.0%	0.1%	0.0%	0.0%	0.4%	0.5%	0.0%	1.2%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.5%	0.0%	0.0%	0.3%	0.0%	30.6%	15.0%	47.3%	0.6%	0.5%	0.0%	0.5%	0.0%	0.9%	0.0%
EC20 192	0.0%	0.0%	0.2%	0.2%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.9%	0.1%	0.0%	1.4%	0.1%	2.0%	0.0%	0.0%	0.7%	32.3%	34.9%	24.2%	0.2%	0.0%	0.1%	0.0%
EC20 198	0.0%	0.1%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	32.4%	30.8%	34.5%	0.1%	0.0%	0.1%	0.0%
EC20 204	0.0%	0.0%	0.2%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	2.8%	0.0%	0.4%	2.1%	1.3%	0.0%	0.0%	0.4%	0.0%	0.4%	0.0%	0.3%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	31.5%	43.6%	12.0%	3.0%	0.0%	0.4%	0.2%
UNI 156	0.3%	0.0%	0.0%	0.0%	0.9%	0.0%	2.6%	0.4%	0.0%	0.6%	0.0%	0.0%	0.0%	0.0%	1.7%	1.8%	1.6%	0.9%	0.0%	4.7%	0.3%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	1.1%	1.6%	0.5%	1.7%	73.6%	0.0%	0.0%	5.4%
UNI 178	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	8.3%	0.0%	0.0%	1.0%	0.0%	0.1%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	89.2%	0.0%	0.0%
UNI 191	19.7%	0.1%	6.4%	11.9%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.1%	0.3%	36.4%	0.0%	0.2%	0.1%	2.0%	0.0%	0.1%	0.0%	0.0%	0.0%	21.2%	0.0%
UNI 214	0.0%	0.0%	0.1%	0.2%	0.5%	0.0%	11.4%	0.1%	0.0%	5.0%	4.5%	0.0%	0.1%	0.0%	0.1%	0.2%	1.5%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	8.9%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	4.9%	0.0%	0.0%	61.9%

The different colours represent the strain level identification

Chapter Six

6. Conclusion and Future work

6.1 General discussion

Rapid discrimination and classification of bacteria is vital for the diagnosis and timely treatment of bacterial infections in human and veterinary medicine (Sauer and Kliem, 2010; Peeling *et al.*, 2006). However, bacterial discrimination has presented an analytical challenge, which has motivated scientists to investigate different approaches continually to identify and discriminate microorganisms rapidly and reliably (Wilkins and Lay, 2005). Traditional methods ranging from serological and physiological assessment to modern genomic methods (Freiwald and Sauer, 2009; Nester, 2001) have gained importance in a wide range of fields in addition to clinical applications. These methods are based on the analysis of morphological characteristics and molecular features; however, these tend to be laborious and time-consuming (Luzzatto-Knaan *et al.*, 2015; Wilkins and Lay, 2005; Nomura, 2015). Moreover, the applications of such methods are usually limited; for example, methods such as enzyme-linked immunosorbent assay (ELISA) are normally used for specific types of bacteria whereas the Analytical Profile Index (API) test is usually applied for a small range of bacteria (Freiwald and Sauer, 2009).

Relatively modern technological approaches, such as mass spectrometry and vibrational spectroscopy, have emerged as essential tools for the analysis of bacterial samples, providing rapid and significantly reliable results based on the measurement of a range of chemical compounds within a given bacterial sample. Mass spectrometry (MS) is a popular example of modern analytical techniques, which has been used routinely to characterise bacteria, generating rich information at both the species and strain levels (Siuzdak, 1996; Luzzatto-Knaan *et al.*, 2015; de Hoffmann and Stroobant 2007; Claydon *et al.*, 1996). Table 6.1 lists the advantages and disadvantages of different methods used to characterise bacteria while Figure 6.1 provides a schematic illustration of the processes of such methods.

Table 6.1: Some of the most commonly used methods for the characterisation of different types of bacteria

Methods	Advantages	Disadvantages
PCR	• Very sensitive • Precise • Accurate	• DNA sequence to be analysed must have been previously identified • Time-consuming • Expensive equipment • Small amount of contamination within samples interferes with experiments
ELISA	• Specific • Sensitive • Reliable	• Need specific antibody • Laborious
PFGE	• Generates stable and reproducible bands • Detection of large DNA molecules	• Time-consuming • High cost per sample
FT-IR spectroscopy	• Easy sample preparation • Simple to use • Sensitive technique • FT-IR spectra provide general information about bacteria • Rapid analysis • High-throughput screening of multiple samples	• May need expertise in chemometric analysis of data • Water band is very strong. • Different conditions (e.g. growth time and culture medium) can cause variations in spectra
Raman spectroscopy	• Provides information on biological structures • Water band is very weak • Rapid • Able to analyse small quantities of samples	• Raman effect is weak • Interference with fluorescence.
MALDI-TOF-MS	• Rapid and specific detection of whole bacteria • Ability to analyse high molecular mass compounds (e.g. proteins) using a wide mass range • Gentle ionisation technique • Sub-picomole sensitivity • Wide array of matrices	• MALDI matrix cluster ions obscure low *m/z* species (<600) leading matrix interference with small molecules • Homogeneity from spot to spot is variable
LC-MS	• Separates and identifies any type of compounds present in bacteria • Reproducible • Quantitative	• Need solvents for extraction. • Time-consuming • Generates complex data • Adduct formation

Information combined from (Sauer *et al*., 2008; Lequin, 2005; Herschleb *et al*., 2007; Durmaz *et al*., 2009; Gan and Patel, 2013; Garibyan and Avashia, 2013; Sauer and Kliem, 2010).

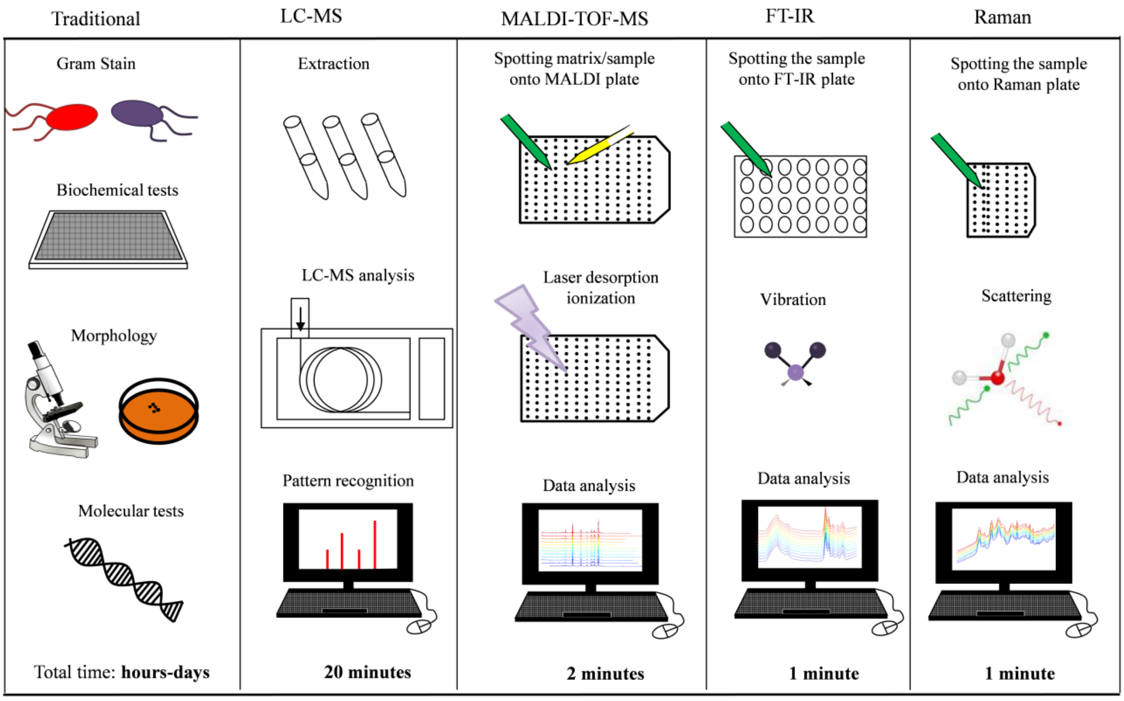


Figure 6.1: Illustration of various methods used for bacterial characterisation including: traditional methods, mass spectrometry and vibrational spectroscopy.

The main motivation in this research work was to identify useful modern analytical techniques that can be applied to discriminate bacteria at both species and strain levels. In general, a comprehensive analytical approach to characterise bacteria should ideally fulfil the following criteria: provide a standard protocol for bacterial characterisation, detect uncharacterised bacteria and match them to closely related species, detect bacterial samples at high speed and sensitivity with high throughput analysis, rely on cost-effective methodology and have the ability to process the data to a high standard using chemometrics (Sauer and Kliem, 2010).

The introduction of modern biomolecular analytical tools that were initially based on mass spectrometry and spectroscopic technologies provided an excellent balance for the characterisation of bacteria samples (Krásný *et al*., 2013; Helm *et al*., 1991; Mariey *et al*., 2001; Naumann *et al*., 1991). Interestingly, some of these analytical techniques, such as MALDI-TOF-MS and LC-MS, generate both genotypic and phenotypic information that can provide a complete set of valuable data (Sauer, 2007; Wilkins and Lay, 2005; Claydon *et al*., 1996; Cobo, 2013; Goodacre *et al*., 1999; Fenn *et al*., 1989; El-Aneed *et al*., 2009). At present, it is evident from many research studies, including this research, that MALDI-TOF-MS has gained greater popularity for the analysis of bacterial samples. A few years after MALDI-TOF-MS was introduced in the field of bacterial classification, it has become a standard analytical tool in most clinical microbiology laboratories as it provides a rapid, accurate and cost-effective method for successful analysis and characterisation of microorganisms (Welker and Moore, 2011; Claydon *et al.*, 1996; de Hoffmann and Stroobant, 2007).

One of the greatest advantages of using MALDI-TOF-MS is its ability to analyse proteins and lipids in complex mixtures, in the presence of salts and buffers, facilitating the analysis of non-purified extracts of biomolecules (i.e. proteins and lipids) and whole (intact) cell samples (Saenz *et al.*, 1999; Schumann and Maier, 2014). Different factors can affect the quality of MALDI-TOF-MS experiments including: sample preparation, sample handling, the choice of matrix, matrix deposition methods and cell lysis methods. Additionally, instrument performance settings, such as the mass range and choice of mode (linear or reflector), can also have a direct effect on the MALDI spectra generated. Different culture conditions and solvent extraction processes can affect MALDI-TOF mass spectra (Liu *et al.*, 2007; Giebel *et al.*, 2010; Kafka *et al.*, 2011; Saenz *et al.*, 1999; Valentine *et al.*, 2005).

Choosing the matrix to match the analyte is a key element in MALDI-TOF-MS analysis. The most commonly used matrices are sinapinic acid (SA), ferulic acid (FA), alpha-cyano-4-hydroxycinnamic acid (CHCA), and 2,5-dihydroxybenzoic acid (DHB) (Marvin *et al.*, 2003; Giebel *et al.*, 2010; Saenz *et al.*, 1999; Schumann and Maier, 2014; Šedo *et al.*, 2011; Lartigue, 2013). Optimising the protocol for MALDI-TOF-MS sample preparation for bacterial analysis is vital to generate a sufficient number of clear and identifiable peaks with a high signal-to-noise (S/N) ratio that enable reliable discrimination of bacteria.

The main challenge faced in this thesis was to develop a robust MALDI-TOF-MS data collection system to discriminate between different types of bacteria based on analysis of different biomolecular compounds (such as proteins and lipids mentioned above), in combination with multivariate analysis. This system resulted in the rapid analysis of bacterial samples, with analysis time of approximately 2 min per sample, with accurate bacterial discrimination compared to other commonly used techniques, such as PCR. This research was carried out on a number of bacterial samples including *Bacillus* spp. and enterococci.

In Chapter 2 of this thesis, a number of MALDI-TOF-MS experimental conditions were optimised including various matrices and matrix deposition methods. This was carried out to develop a standard experimental procedure for the analysis of a protein mixture, followed by applying these optimum conditions to the analysis of proteins from intact bacterial samples from 34 strains and 7 different species belonging to *Bacillus* and *Brevibacillus* genera. The results obtained indicated that the strains were successfully classified using MALDI-TOF-MS with correct classification rate of 90% for the 7 species. This indicates that MALDI TOF-MS can be a powerful, rapid and accurate analytical tool for analysing and classifying bacteria, with minimum sample preparation. Therefore, the remainder of the studies carried out in this thesis relied on the use of MALDI-TOF-MS in addition to other analytical tools.

In Chapter 3, lipids were the main focus of analysis by MALDI-TOF-MS. This powerful analytical tool provided promising and reliable results for analysing a lipid mixture containing 5 different lipids. This work involved the optimisation of experimental conditions for detecting lipids within mixtures using MALDI-TOF-MS. A fractional factorial design (FFD) was applied to the data generated in this study in order to significantly reduce the number of potential experiments from 8064 to just 720, which in turn reduced the workload significantly. In addition, using FFD, it was possible to explore multiple MALDI-TOF-MS parameters and optimise the system for the detection of specific analytes of interest. A number of different conditions were investigated in relation to analysis of the lipid mixture via MALDI-TOF-MS including: matrices, matrix preparations, matrix additives, additive concentrations and matrix deposition methods. The significance of each variable was investigated using the FFD exercise, which aided in the discovery of exactly which combinations enabled the detection of the five lipid peaks successfully. The work in this chapter showed that the choice of matrix and the presence of matrix additives were the key factors in producing high quality spectra. Moreover, ATT and DHB were shown to be the best matrices that can be used to analyse lipid samples using MALDI-TOF-MS.

In Chapter 4, the results above were used to direct MALDI-TOF-MS for the analysis of 7 of different species featuring 33 strains from *Bacillus* and *Brevibacillus* genera based on lipid extracts, which confirmed that this technique can be used to identify extracted lipids (putatively) and classify bacterial samples based on lipid analysis. These findings contradicted some review articles where MALDI-TOF-MS was reported to have less successful outcomes with low molecular mass biomolecules, due to the interference between matrix peaks and lipid peaks. In this chapter, LC-MS (as a gold standard tool) was used to evaluate and confirm the results obtained using MALDI-TOF-MS. Confirmatory results from these two powerful analytical techniques indicated that the classification of *Bacillus* and *Brevibacillus* species based on extracted lipids was possible. Furthermore, Procrustes distance analysis was employed and suggested that classification of *Bacillus* and *Brevibacillus* bacteria based on these two analytical techniques was highly similar with a Procrustes distance of 0.0699 (p<0.001). Moreover, the results obtained based on protein and lipid analysis (Chapters 2 and 4, respectively) were also very similar (Procrustes distance of 0.1006, p<0.001). This strongly suggests that MALDI-TOF-MS could be used reliably as a routine clinical tool to classify and identify bacteria based on the analysis of lipids or proteins and that both biomolecular species yield a similar level of differentiation.

Following the successful development of MALDI-TOF-MS for bacterial classification, clinical samples from Belfast Hospital (35 isolates from 12 *Enterococcus faecium* strains) were analysed using three different modern analytical techniques (Chapter 5 of this thesis). *Enterococcus faecium* was chosen as a clinical sample as it frequently causes infections in babies in the Neonatal Unit of this hospital. In addition to MALDI-TOF-MS, FT-IR and Raman spectroscopic techniques were used to analyse these clinical relevant samples. Again, MALDI-TOF-MS provided promising discrimination results at the strain level providing a clear distinction between the 12 classes of bacteria. In addition, FT-IR spectroscopy generated high quality and promising results in the discrimination of *E. faecium* strains and isolates. Raman spectroscopy was also able to discriminate the 35 isolates and provided relatively similar results to FT-IR spectroscopy. However, the correct classification rates (CCRs) using Raman spectroscopy was low compared to FT-IR spectroscopy.

The characteristics of the various analytical techniques that were used to analyse bacteria in this research are summarised in Table 6.2.

Table 6.2: Characteristics of MALDI-TOF-MS, LC-MS, FT-IR spectroscopy and Raman spectroscopy relevant for analysing bacterial samples*

Analytical Techniques / Characteristics	MALDI-MS	LC-MS	FT-IR	Raman
Cost running per-sample	Medium	High	Low	Low
Automation	Yes	Yes	Yes	No#
Sample preparation	Minimum/Moderate (depending on biomolecule)	Moderate	Minimum	Minimum
Amount of the samples	2 µL	10 µL	20 µL	3 µL
Analysis time (spectral acquisition)	2 min	20 min	1 min	1 min
Reproducibility	Medium	Good	Good	Poor
Sensitivity	High	High	High	Medium
Destructive of sample	Yes	Yes	No	No
Size of generated dataset	Average	Large	Small	Small
Complexity of data analysis	Average	Complex	Simple	Average

**This table was generated from work carried out on each individual analytical technique in this thesis.*

# Ideally, Raman spectroscopy can be used in automated mode for running samples; however, in this study, non--automated mode was used as the stage tended to move and come out of focus during analysis.

Modern analytical techniques enable accurate and rapid analysis of a wide range of biological and clinical samples such as bacteria. However, it is vital that analytical techniques are continuously optimised and developed to meet the increasing demand to analyse bacterial samples, in particular those related to disease and infection.

Many analytical techniques have been used individually for the analysis of bacterial samples. However, from the work carried out in this thesis and as seen in Table 6.2, the use of two or more analytical techniques can provide complementary information, thus offering more in-depth insights into bacterial characterisation and classification. The main characteristics of each analytical technique described in this thesis are highlighted in Table 6.2. Despite the few limitations MALDI-TOF-MS has, this analytical technique offers excellent and promising results for analysing bacterial samples. Though it is recommended that the findings from MALDI-TOF-MS analysis are confirmed using other useful analytical techniques such as FT-IR spectroscopy and LC-MS. Our research complemented previous studies as it showed that in contrast to other analytical techniques, Raman spectroscopy provides relatively lower quality data.

The main hindrance to the application of MALDI-TOF-MS as a technique for the analysis of different types of bacteria at both species and strain levels is reproducibility. This is mainly due to the large number of factors on which the quality and the reproducibility of MALDI-TOF-MS spectra depend; namely, the matrix, matrix additives, matrix solvents, deposition methods, concentration of cells and spectral variation in sample handling. Furthermore, the absence of standard protocols for the analysis of different types of bacteria is another contributing factor. In this thesis, the optimisation of some of these parameters was pursued with the aim of contributing to building up standard protocol.

From the findings of this thesis, it can be speculated that MALD-MS analysis can be used for bacterial typing and classification exercises with relevance to applications in the clinic. This may be of significance to diagnostic and therapeutic applications in addition to other uses in taxonomical studies. However, using FT-IR and Raman spectroscopic methods may be hindered by the relatively lower accuracy of identification (especially for Raman spectroscopy) and more importantly the lack of available reference databases of spectroscopic data. Further, lipid analyses using MALDI-MS is of particular interest as this type of application can be of clinical relevance with such advantages as higher resolution of analysis than proteins. Understanding the existence and distribution of lipid species in normal biological samples compared to trends associated with disease states are key to further characterisation of such diseases as cancer and cardiovascular health problems and development of effective therapies.

Figure 6.2 summarises the results of this thesis (Chapters 2, 3, 4 and 5) on bacterial discrimination using various analytical techniques: MALDI-TOF-MS, LC-MS, FT-IR and Raman spectroscopies.

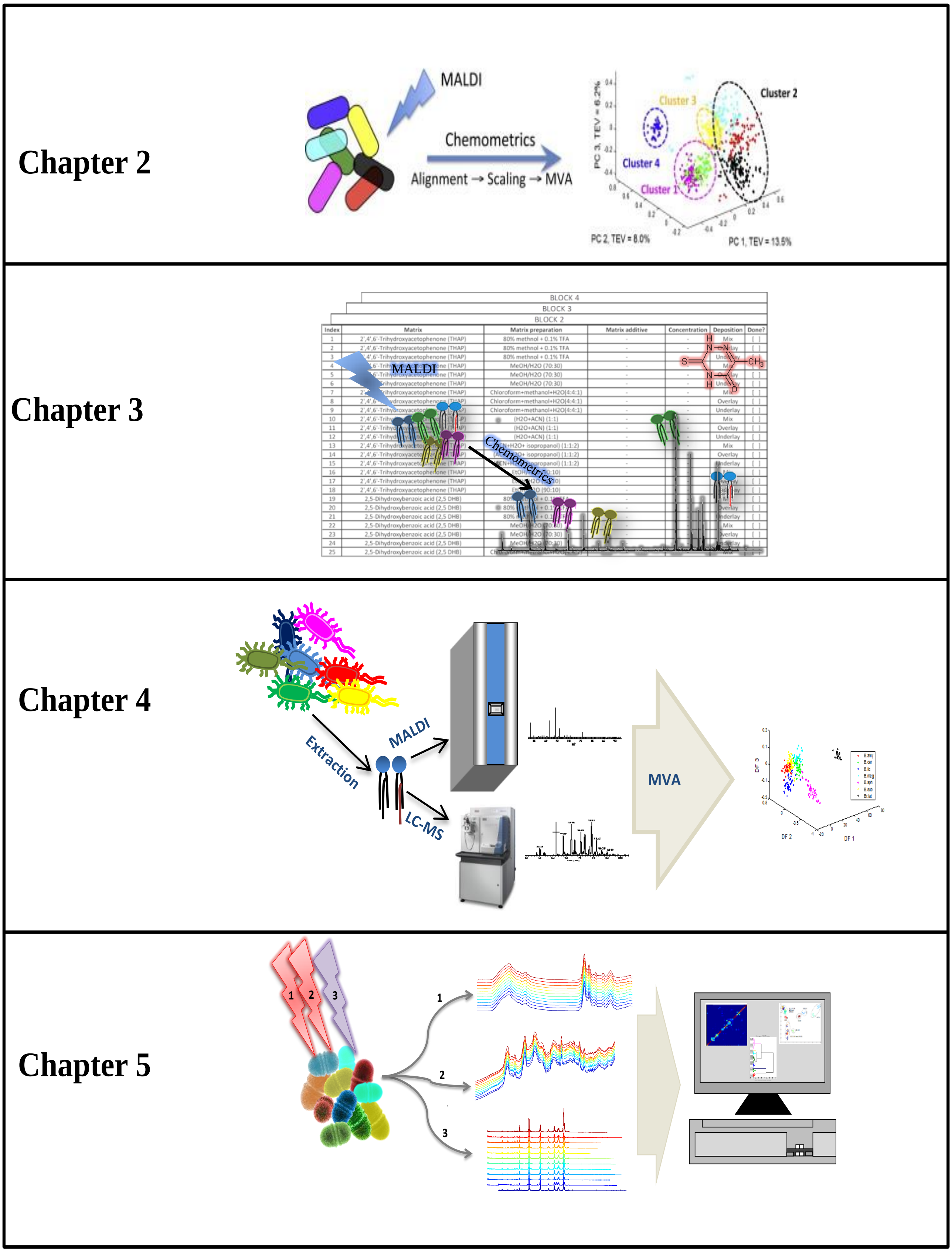


Figure 6.2: A summary of the results of this thesis (Chapters 2, 3, 4 and 5) on bacterial discrimination using various analytical techniques including; MALDI-TOF-MS, LC-MS, FT-IR and Raman spectroscopies.

6.2 Future work

Microbiology has developed over the past decades from traditionally classifying microorganisms by morphology to more comprehensive genotypic, phenotypic and biochemical characterisations. Arguably, the findings described in this study have the potential to contribute to the field of microbiological classification and clinical diagnosis of bacteria. As an outlook for the future, the work described in this thesis can be extended to analysing different types of bacteria of clinical relevance and can be useful as a high throughput approach for the analysis of a large number of bacterial samples.

Moreover, the *Enterococcus faecium* bacteria examined in this thesis can also be analysed based on lipid extracts using MALDI-TOF-MS to investigate the usefulness of lipid extracts *versus* whole-cell based approaches.

Tandem mass spectrometry, such as MALDI-TOF-MS/MS and LC-MS/MS, can also be used for bacterial analysis as these can provide more detailed structural information and more features on which to base bacterial classification and identification using database searching.

In addition, as the proteomic and lipidomic fields gain more importance in the analysis of different types of bacteria, a ‘universal library’ of bacterial classification features can be built based on MALDI-TOF-MS data for the two genera investigated in this thesis, and could also be extended to other clinically relevant bacteria.

6.3 Outlook

The results generated from the work carried out in this thesis indicate that modern analytical techniques are extremely useful tools for providing in-depth information to classify and discriminate different types of bacteria successfully down to the isolate, species and strain levels. This work focused on analysing bacterial samples using MALDI-TOF-MS, which was found to be a powerful tool, and contributed to the optimisation of the experimental and analytical procedures relevant to this technique which can be of benefit to further studies. The work undertaken in this study strongly suggests that this analytical technique has huge potential in many biological and clinical applications.

6.4 References

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. 1996. The rapid identification of intact microorganisms using mass spectrometry. *Nature Biotechnology,* **14**, 1584-1586

Cobo, F. 2013. Application of MALDI-TOF mass spectrometry in clinical virology: a review. *The Open Virology Journal,* **7**, 84

de Hoffmann, E. and Stroobant, V. 2007. *Mass Spectrometry: Principles and Applications,* Chichester, Wiley, pp.15-131,

Durmaz, R., Otlu, B., Koksal, F., Hosoglu, S., Ozturk, R., Ersoy, Y., Aktas, E., Gursoy, N. C. and Caliskan, A. 2009. The optimisation of a rapid pulsed-field gel electrophoresis protocol for the typing of *Acinetobacter baumannii, Escherichia coli* and *Klebsiella* spp. *Japanese Journal of Infectious Diseases,* **62**, 372-7

El-Aneed, A., Cohen, A. and Banoub, J. 2009. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analysers. *Applied Spectroscopy Reviews,* **44**, 210-230

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

Freiwald, A. and Sauer, S. 2009. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nature Protocols,* **4**, 732-742

Goodacre, R., Heald, J. K. and Kell, D. B. 1999. Characterisation of intact microorganisms using electrospray ionisation mass spectrometry. *FEMS Microbiology Letters,* **176**, 17-24

Garibyan, L. and Avashia, N. 2013. Polymerase Chain Reaction. *Journal of investigative Dermatology,* **133**, e6

Gan, S. D. and Patel, K. R. 2013. Enzyme Immunoassay and Enzyme-Linked Immunosorbent Assay. *Journal of Investigative Dermatology,* **133**, e12

Giebel, R., Worden, C., Rust, S. M., Kleinheinz, G. T., Robbins, M. and Sandrin, T. R. 2010. Microbial fingerprinting using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS): applications and challenges. *Advances in Applied Microbiology,* **71**, 149-84

Helm, D., Labischinski, H., Schallehn, G. and Naumann, D. 1991. Classification and identification of bacteria by Fourier-transform infrared spectroscopy. *Journal of General Microbiology,* **137,** 69-79

Herschleb, J., Ananiev, G. and Schwartz, D. C. 2007. Pulsed-field gel electrophoresis. *Nature Protocols,* **2**, 677-684

Kafka, A. P., Kleffmann, T., Rades, T. and Mcdowell, A. 2011. The application of MALDI TOF MS in biopharmaceutical research. *International Journal of Pharmaceutics,* **417**, 70-82

Krásný, L., Hynek, R. and Hochel, I. 2013. Identification of bacteria using mass spectrometry techniques. *International Journal of Mass Spectrometry,* **353**, 67-79

Lartigue, M.-F. 2013. Matrix-assisted laser desorption ionisation time-of-flight mass spectrometry for bacterial strain characterisation. *Infection, Genetics and Evolution***,** **13**, 230-235

Liu, H., Du, Z., Wang, J. and Yang, R. 2007. Universal sample preparation method for characterisation of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology,* **73**, 1899-1907

Lequin, R. M. 2005. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clinical Chemistry,* **51**, 2415-2418

Luzzatto-Knaan, T., Melnik, A. V. and Dorrestein, P. C. 2015. Mass spectrometry tools and workflows for revealing microbial chemistry. *Analyst,* **140**, 4949-4966

Mariey, L., Signolle, J. P., Amiel, C. and Travert, J. 2001. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vibrational Spectroscopy,* **26**, 151-159

Marvin, L. F., Roberts, M. A. and Fay, L. B. 2003. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in clinical chemistry. *Clinica Chimica Acta,* **337**, 11-21

Naumann, D., Helm, D. and Labischinski, H. 1991. Microbiological characterisations by FT-IR spectroscopy. *Nature,* **351**, 81-82

Nomura, F. 2015. Proteome-based bacterial identification using matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS): A revolutionary shift in clinical diagnostic microbiology. *Biochimica et Biophysica Acta (BBA),* **1854,** 528-537

Nester, E. W. 2001. *Microbiology: a Human Perspective*, 3rd Edition, the University of Michigan, McGraw-Hill

Peeling, R. W., Smith, P. G. and Bossuyt, P. M. M. 2006. A guide for diagnostic evaluations. *Nature Reviews Microbiology,* **4**, S2-S6

Sauer, S. 2007. The essence of DNA sample preparation for MALDI mass spectrometry. *Journal of Bioc hemical and Biophysical Methods,* **70**, 311-318

Sauer, S., Freiwald, A., Maier, T., Kube, M., Reinhardt, R., Kostrzewa, M. and Geider, K. 2008. Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS ONE,* **3**, e2843

Sauer, S. and Kliem, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nature Reviews Microbiology,* **8**, 74-82

Saenz, A. J., Petersen, C. E., Valentine, N. B., Gantt, S. L., Jarman, K. H., Kingsley, M. T. and Wahl, K. L. 1999. Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis. *Rapid Communications in Mass Spectrometry,* **13**, 1580-1585

Schumann, P. and Maier, T. 2014. MALDI-TOF Mass Spectrometry Applied to Classification and Identification of Bacteria. *In:* Michael Goodfellow, I. S. and Jongsik, C. *Methods in Microbiology,* San Diego, CA: Elsevier Academic Press, pp. 275–306

Šedo, O., Sedláček, I. and Zdráhal, Z. 2011. Sample preparation methods for MALDI-MS profiling of bacteria. *Mass Spectrometry Reviews,* **30**, 417-434

Valentine, N., Wunschel, S., Wunschel, D., Petersen, C. and Wahl, K. 2005. Effect of culture conditions on microorganism identification by matrix-assisted laser desorption ionisation mass spectrometry. *Applied and Environmental Microbiology,* **71**, 58-64

Welker, M. and Moore, E. R. B. 2011. Applications of whole-cell matrix-assisted laser-desorption/ionisation time-of-flight mass spectrometry in systematic microbiology. *Systematic and Applied Microbiology*, **34**, 2-11

Wilkins, C. L. and Lay, J. O. 2005. *Identification of microorganisms by mass spectrometry,* Hoboken New Jersey, John Wiley and Sons, pp.303

Appendix: Published work in the original format

Publication 1

AlMasoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. 2014. Optimisation of matrix assisted desorption/ionisation time of flight mass spectrometry (MALDI-TOF-MS) for the characterisation of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta,* **840**, 49-57

Publication 2

AlMasoud, N., Correa, E., Trivedi, D.K. and Goodacre, R. 2016. Fractional Factorial Design of MALDI-TOF-MS sample preparations for the optimized detection of phospholipids and acylglycerols. *Analytical Chemistry,* **88**, 6301-6308

Analytica Chimica Acta 840 (2014) 49–57

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS) for the characterization of *Bacillus* and *Brevibacillus* species

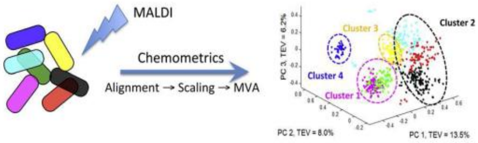
Najla AlMasoud, Yun Xu, Nicoletta Nicolaou, Royston Goodacre [*]

School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

HIGHLIGHTS

- Optimization of MALDI-TOF-MS for characterizing *Bacillus* and *Brevibacillus* species.
- Development of a suitable chemometric workflow for processing raw MALDI-TOF-MS data.
- Classification of 7 species from bacteria achieved high accuracy (~90%).
- Allowed to dry at room temperature (*ca.* 22 °C) for 1 h.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 12 April 2014
Received in revised form 16 June 2014
Accepted 19 June 2014
Available online 23 June 2014

Keywords:
Optimization
Bacteria
Mass spectrometry
Alignment
Classification

ABSTRACT

Over the past few decades there has been an increased interest in using various analytical techniques for detecting and identifying microorganisms. More recently there has been an explosion in the application of matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF-MS) for bacterial characterization, and here we optimize this approach in order to generate reproducible MS data from bacteria belonging to the genera *Bacillus* and *Brevibacillus*. Unfortunately MALDI-TOF-MS generates large amounts of data and is prone to instrumental drift. To overcome these challenges we have developed a preprocessing pipeline that includes baseline correction, peak alignment followed by peak picking that in combination significantly reduces the dimensionality of the MS spectra and corrects for instrument drift. Following this two different prediction models were used which are based on support vector machines and these generated satisfactory prediction accuracies of approximately 90%.

1. Introduction

Bacillus are rod-shaped aerobic Gram-positive bacteria that are able to sporulate. These bacteria are normally found in the soil, plants, and can be transferred to meat and dairy products where they can spoil food making them unfit for human consumption [1]. Even though most of these bacteria are harmless saprophytes there still remains a few toxic members of this genus, such as *Bacillus subtilis* and *Bacillus cereus*, which are often associated with food-borne infections [2], along with the more notorious *Bacillus anthracis* the casual agent of anthrax. Whilst *Bacillus sphaericus* is toxic to insects and is used for biocontrol of mosquitoes [3]. *B. subtilis* is the most scientifically defined member of the *Bacillus* genus and has thus been used as a model organism in many genetic research studies. Other members of this *B. subtilis* group are less defined and are harder to identify such as *Bacillus licheniformis* and

[*] Corresponding author. Tel.: +44 161 3064480.
E-mail address: roy.goodacre@manchester.ac.uk (R. Goodacre).

http://dx.doi.org/10.1016/j.aca.2014.06.032
0003-2670/© 2014 Elsevier B.V. All rights reserved.

Bacillus amyloliquefaciens because they are very similar microorganisms [1,4]. The *B. cereus* group contains a number of different bacteria, with some leading to negative health implications in humans, and as discussed above have sometimes been linked to food poisoning [5–7].

The unequivocal identification of bacterial is a vital step in medical therapy and the food industry and this is usually performed at the genotypic or phenotypic level. A number of traditional methods have so far been used to identify microorganisms, such as cell culturing with differential staining [8], polymerase chain reaction (PCR) [9–12] and enzyme linked immunosorbent assays (ELISA) [13]. Whilst these approaches formed the foundations of knowledge and understanding in microorganism research, these methods are very time consuming, costly and labour intensive, hence more rapid detection methods are continually needed [14]. In addition to rapid testing, methods that provide molecular-specific information are also preferred as these may allow one to relate any markers to specific microbiological function.

Modern methods for the identification of microorganisms have recently focussed on mass spectrometry as these are rapid and provide molecular information on the bacteria under investigation. Whilst pyrolysis mass spectrometry was used for bacterial analysis in the past [15], current methods are based on electrospray-ionization (ESI-MS) [16,17] and the more popular method of matrix-assisted laser desorption ionization (MALDI-MS) [14,18–20]. MALDI-TOF-MS is easy to use, provides rapid results, and has been used for identification and taxonomy of microorganisms [18,21,22]. The maturity of this analytical technique has benefitted its application to a wide range of areas such as proteomics [23–25], intact-cell mass spectrometry (ICMS) [19,26–29] and in the area of lipidomics [30–32].

MALDI-MS on bacteria (and indeed other complex samples) results in a multivariate spectral pattern, which usually provides information on the protein content of the bacterium under analysis. This protein profile or barcode can be matched against MALDI-MS profiles/barcodes that have been previously collected under identical conditions and stored within (usually) organism specific databases [22,23,33,34]. This matching may involve the generation of dendrograms from hierarchical cluster analyses (HCA) [33,35] or ordination plots from principal component analysis (PCA) [36,37] or discriminant analysis (DA) [38,39].

The aim of this study was to generate a reproducible MALDI-TOF-MS protocol for measuring the protein spectra from bacteria. In order to establish this we used a set of 34 well-characterised bacteria belonging to the genus *Bacillus*. In a series of experiments we optimised the matrix and the sample preparation method used using first a mixture of pure proteins followed by the analysis of a subset of these bacilli, before the optimised method was used on the full set of 34 bacteria.

2. Materials and methods

2.1. Compounds

Trifluoroacetic acid (TFA), acetonitrile (ACN), sinapinic acid (SA), caffeic acid (CA), 2,5-dihydroxybenzoic acid (DHB), α-cyano-4-hydroxycinnamic acid (CHAH), ferulic acid (FA), 2,4,6-trihydroxyacetophenone monohydrate (THAP), 2-(4-hydroxyphenylazo)benzoic acid (HABA), 2,6-dihydroxyacatophenone (DHAP), 9-aminoacridine (9-AA) and dithranol (INN) from Sigma–Aldrich (Dorset, UK) were used.

14 g of nutrient agar (Fisher Scientific Ltd. Loughborough, UK) was dissolved and mixed thoroughly in a bottle containing 500 mL of water. This bottle was then autoclaved at 121 °C for 15 min and subsequently used for the bacterial cultures.

2.2. Standard protein samples for MALDI-TOF-MS

Five different proteins were mixed together at the same concentration (20 μM) to find the optimum matrix and deposition method for pure protein analysis. These proteins (molecular weight provided in parentheses) included: insulin (5735), cytochrome *c* (12,362), apomyoglobin (16,952), aldolase (39,212) and albumin (66,430) and were acquired from Sigma–Aldrich.

2.3. Bacterial culturing

General information of the 34 strains of *Bacillus* is provided in Table 1 and these belonged to two genera (*Bacillus* and *Brevibacillus*) and seven different species. The cells were cultured on nutrient agar and were incubated at 37 °C for 24 h. Bacterial strains were cultivated aerobically three times under these conditions to make sure that the cultures were axenic, and to maintain a stable phenotype. After this was established single bacterial colonies were then cultured on nutrient agar and also incubated at 37 °C for 24 h. Five biological replicates were prepared for each isolate. After growth the biomass of each sample was carefully collected using two full sterilised plastic loops (equivalent to about 20 μL). This biomass was then centrifuged for 3 min at 13,000 × *g*. The pellets containing the bacteria were then washed twice with 1 mL of

Table 1
The 34 *Bacillus* species and strains used in this work.

Sample no.	Species	Strain no.	Key colour used in figures
1	*B. sphaericus*	7134[T]	Yellow
2		B0408[*]	
3		B0219	
4		B0769	
5		B1147	
6	*Br. laterosporus*	B0043	Blue
7		B0262	
8	*B. subtilis*	B0014[T,*]	Black
9		B0044	
10		B0098	
11		B0099	
12		B0410	
13		B0501	
14		B1382	
15	*B. cereus*	B0002[T,*]	Green
16		B0550	
17		B0702	
18		B0712	
19		B0851	
20	*B. amyloliquefaciens*	B0177[T]	Red
21		B0168[*]	
22		B0175	
23		B0251	
24		B0620	
25	*B. megaterium*	B0010[T,*]	Pink
26		B0056	
27		B0057	
28		B0076	
29		B0621	
30	*B. licheniformis*	B0252[T,*]	Cyan
31		B0242	
32		B0755	
33		B1081	
34		B1379	

[T] Indicates the type strain.
[*] Indicates strains used for preliminary optimization experiments.

distilled water to remove residual culture media, centrifuged again to remove the supernatant, and the pellet was then stored at −80 °C until further analysis.

2.4. Optimization of MALDI-TOF-MS

Optimization of sample preparation was carried out in order to identify the most appropriate matrix preparation and deposition method for the analysis of bacteria. Initial experiments optimised the matrix and deposition method on mixtures of pure proteins (Supplementary Information Table S1 illustrates the four different sample preparation methods for MALDI-TOF-MS). Briefly, 10 different matrices were used to find the most compatible matrix for MALDI-TOF-MS analysis and these included DHB, CHAH, SA, FA, THAP, CA, HABA, DHAP, 9-AA and INN. At the same time four different depositions methods (mix, overlay, underlay and sandwich) were investigated for protein sample preparation. The optimised conditions involved using SA as the matrix and the mix method for sample deposition and this was subsequently used for bacterial analysis. We note of course that the five proteins chosen are a substitute for bacterial analysis and we did not assume that the best protein preparation method would be the optimal method for bacteria so we tested the top three matrices and preparation methods on a small subset of bacteria (the type of strain from each species is marked with 'T' and the strains used for preliminary optimization experiments were marked with '*' in Table 1); SA with the mix method was indeed the best method (data not shown for this optimization).

2.5. Bacterial sample preparation

Preliminary experiments also suggested that it was important to optimise the appropriate amount of biomass for MALDI-MS; which one can think of as the amount of matrix:analyte ratio. The defrosted pellet from above (which contained ~10^{10} CFU (colony forming units)) was diluted at various levels in water containing 0.1% TFA (250, 500, 1000, 1500 and 4000 μL; data not shown except for 1000 μL water containing 0.1% TFA). The optimum pellet dilution was established at 1000 μL and this was subsequently used.

For MALDI-TOF-MS analysis of the bacteria 10 mg SA was dissolved in 500 μL of ACN and 500 μL of water containing 2% TFA. 10 μL from the above bacterial sample and 10 μL of matrix were mixed together (Table S1) and vortexed for 10 s before. 2 μL from the resultant mixture was spotted on a MALDI-TOF-MS stainless steel target plate. This was allowed to dry at room temperature (*ca.* 22 °C) for 1 h.

2.6. MALDI-TOF-MS

Samples were analysed in batches using an AXIMA-Confidence (Shimadzu Biotech, Manchester, UK) mass spectrometer. This MALDI-TOF-MS device contained a nitrogen pulsed UV laser with a wavelength of 337 nm as described previously [40]. The power of the laser at the laser head used was set to 140 mV. Each profile contained 20 shots, and 100 profiles were collected using a circular raster pattern. The MS was operated in positive ion source and linear TOF was used over the range from 1000 to 80,000 m/z. The collection time for each sample was ~3 min and each biological sample was analysed four times (technical replicates). A single biological replicate for each of the 34 bacteria was analysed each day, and the analysis time took 5 days of machine time during a 2 week period. The result of this analysis generated 680 MALDI-TOF-MS spectra: 34 bacteria × 5 biological replicates × 4 technical replicates. The MALDI device was calibrated using the protein mixture mentioned above.

3. Data analysis

3.1. Pre-processing

MATLAB 2010a (The Math Works, Natick, MA, USA) was used for pre-processing and data analysis. Baseline corrections were first performed on the spectra by using asymmetric least squares (AsLS) [41]. In addition, the interpolation and alignment of MALDI-TOF-MS spectra in the m/z axis were required in order to integrate all the spectra in a unified coordinate system and also reduce the amount of ambiguities of assigning peaks from different samples collected over the 2 week period (see below). This was achieved by firstly interpolating all the spectra into a common m/z domain which is from 1000 to 13,000 m/z with an interval of 0.1078 m/z and then an algorithm named interval correlation optimized shifting (icoshift) [42] was used to correct m/z drifting across different samples. Peak picking was then performed on the aligned spectra to detect mass peaks in each spectrum and this was performed using intensity weighted variance (IWV) algorithm as described by Jarman et al. [43]. The detected peaks of all the samples were then aligned together with a drift tolerance threshold of ±1 m/z. After this peak picking and alignment process, a total number of 243 unique mass peaks were detected and resulted in a peak table matrix of dimensions 680 × 243 which was used for further data analysis. The peak intensities were firstly $\log_{10}$-scaled and then normalised so that the sum of squares of each row (i.e. a sample) equals 1.

3.2. Multivariate analysis

Two different types of analysis were performed on the data: one was a semi-quantitative analysis and the other a qualitative analysis.

The semi-quantitative analysis was performed on the $\log_{10}$-scaled and normalised peak intensity table matrix. Principal components analysis (PCA) was performed first to reveal the 'natural' pattern of the data and then support vector machines (SVM), with a linear kernel, was used for supervised classification. The SVM models were validated by using a bootstrap replacement procedure coupled with cross-validation for the model parameter selection (see below). In this process the data were first split into a training set and a test set via a bootstrapping resampling based on the biological replicates; i.e. all the samples from the same biological replicates were considered as one during the resampling. Considering the random nature of this bootstrapping process, the number of samples selected in the training and test sets varied between the different 1000 iterations, on average 63.3% of the samples would be in the training set and 36.7% in the test. Next a k-fold cross-validation was performed on the training set where k is the number of unique biological replicates in the training set, the error penalty parameter C within the SVM varied from 1 to 10^6 and the one which yielded the lowest cross-validation error was chosen to build the SVM model. The model was then applied to the test set generated via the bootstrapping selection in order to calculate the predictive accuracy of the test set. This bootstrap procedure was repeated 1000 times and the collected predictive accuracies for the test set only were then averaged. This can be considered as an unbiased estimation of the generalisation performance of the SVM model. Two types of classification were carried out: one was to classify the samples on species level (7 classes); and the other was to classify the samples on strain level (34 classes). Both types of classification followed the same validation procedure as described above.

The qualitative analysis on the data focused on the presence/absence of certain feature (i.e. mass peaks) while ignoring the intensities of the peaks. The peak table matrix was converted into a binary format: if a peak had been detected in one particular sample the corresponding element in the matrix was set to 1 and 0 if otherwise; the threshold for presence/absence was set to be 3× standard deviation of baseline signals. Principal coordinate analysis (PCoA) was used as a counterpart of PCA in the qualitative analysis and the Jaccard distance was used to measure the dissimilarity between the samples. A distance matrix ***D*** was calculated which contains the Jaccard distance between every pair of samples. PCoA was then applied to ***D*** to obtain a scores matrix and this scores matrix can be interpreted in the same way as the scores matrix obtained from PCA. For supervised classification, a naïve Bayesian classifier and SVM with a Jaccard kernel [44] were applied to the data. Both classifiers were validated using exactly the same bootstrapping procedure as described above and the classifications were again performed on both species and strain level.

4. Results and discussion

4.1. MALDI-TOF-MS optimization

Initially a mixture contain five different proteins was used to obtain the optimum conditions for protein analysis using MALDI-TOF-MS. At this stage 10 matrices were used to determine the most suitable matrix and four sample preparation procedure when performed. Good protein detection was seen for SA, CA and FA, whilst others such as DHAP and 9-AA were not suitable matrices for protein analysis. Results obtained from this study showed that SA was the most suitable matrix for protein analysis (Tables S2–S5). This finding was supported by other workers analysis [36,45–48], and this may be due its classification as a hot matrix, which causes less protein fragmentation [49]. In addition, as discussed by Vaidyanathan [24], the reason behind SA's compatibility lies in its high level of homogeneity and crystallisation with the solvent when SA is mixed with bacteria.

During the matrix optimization the most appropriate sample deposition method for protein analysis was also assessed. Four methods were used (see Table S1 for details) and it was found that the 'mix method' where sample and matrix are pre-mixed prior to spotting on the MALDI target plate was best. This deposition method was very reproducible and caused improved desorption and ionization in comparison with other deposition methods. Tables S2–S5 (see SI) summarises the data obtained from analysing the 5-way protein mixture using the 10 different matrices and the 4 different deposition methods.

After this the top 3 matrices (SA, CA and FA) were assessed on a subset of 6 bacteria comprising the type strain from each species. SA with the mix method was also the best method in terms of the number of protein peaks routinely detected in replicate analyses and in terms of the reproducibility of signal (as judged by PCA; data not shown). Thus SA with the mix method was used for all bacterial analyses.

4.2. Bacillus MALDI-TOF-MS spectra

Typical MALDI-TOF-MS spectra of *B. cereus* B0712 obtained SA with the mix method for both the raw MS data and after baseline correction and alignment are shown in Fig. 1. It is clear from the raw data from this bacterium (and indeed all the bacteria analysed; data not shown) that significant baseline artefacts are observed which were unavoidable. Spectra were therefore pre-processed using the following routine: (i) baseline correction was performed using AsLS on the raw MS profiles; (ii) this was followed by spectral alignment using icoshift (Fig. 1) and (iii) finally, following this step these spectra were scaled so that the sum of square of each spectrum equals to 1. Typical normalised and scaled spectra of all 7 type strains from these bacilli are shown in Fig. 2A–G.

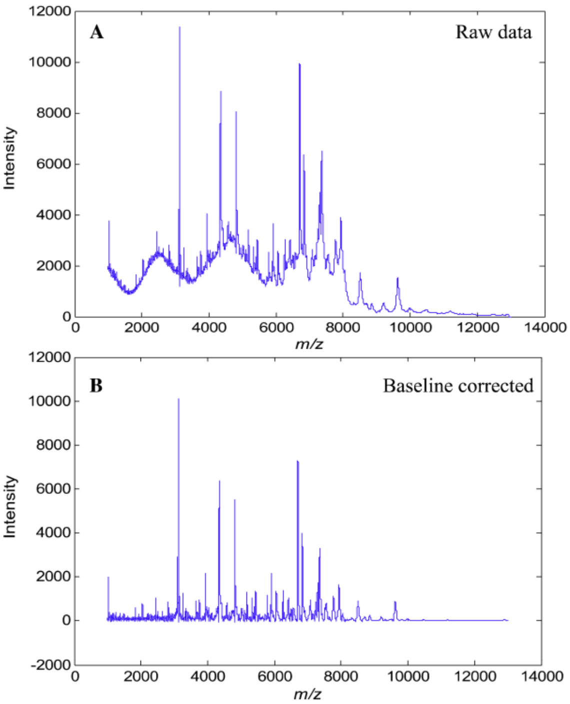


Fig. 1. Differences between MALDI mass spectra obtained from the analysis of *B. cereus* B0712 (A) before and (B) after baseline correction.

It is known that sample preparation for bacterial analysis is important and this has been discussed before for the analysis of *Bacillus* species [17,29,33]. It can be seen that these MALDI-TOF-MS spectra are generally distinct from one another and possess good signal-to-noise in the *m*/*z* 1000–13,000 range used. Whilst some spectra are clearly very different, *Brevibacillus laterosporus* (which belongs to a different genera) compared with the other *Bacillus* species, it is very difficult to use only visual inspection to identify these different bacteria. Therefore chemometric methods are needed for spectral analysis.

The spectra that were generated from MALDI-TOF-MS are very high dimensional nature and each spectrum contains 0.1078 *m*/*z* intervals after interpolation with ion counts at each value. It is clear from the spectra (Fig. 2) that much of this information is redundant (i.e. noise), such that direct computation using PCA would be both puerile, as many spurious correlations may be found, as well as being computational intense.

Therefore we used peak picking to select only those *m*/*z* which had arisen from real signals. In this process the intensity weighted variance (IWV) algorithm was used and resulted in a peak table comprising 243 features from the bacteria analysis of 680 samples. This matrix was of dimensions 680 × 243 and significantly reduced from the full spectra (680 × 111,339) and was used for further data analysis.

The scores plots of the first 3 PCs from PCA performed on the peak table matrix are provided in Fig. 3 and the loadings plot of the first 2 PCs are provided in Fig. 4. The variables with their absolute loadings (either PC1 or PC2) greater than 0.1 are labelled in Fig. 4 along with their corresponding *m*/*z*. Four main clusters can be

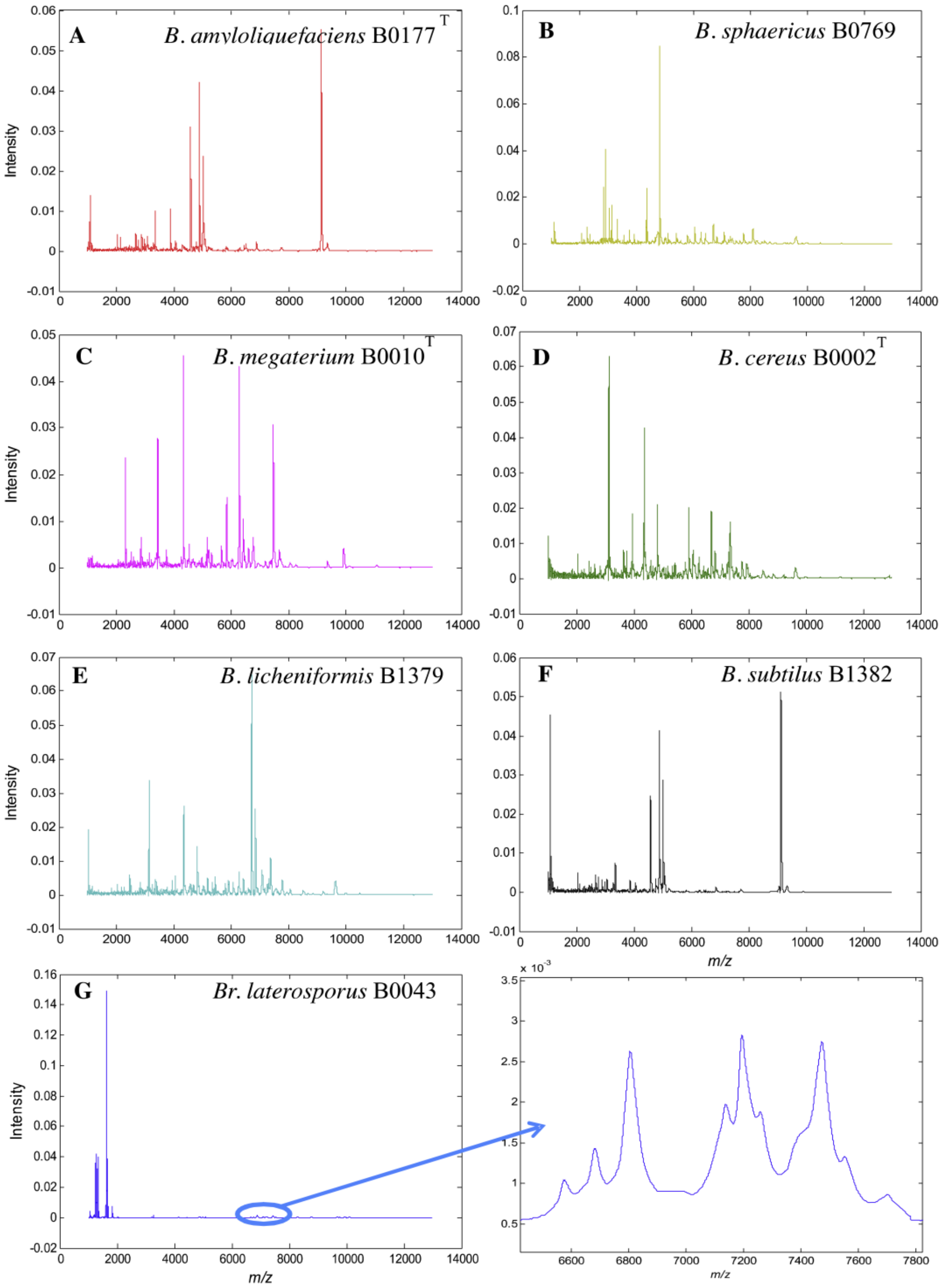


Fig. 2. Typical MALDI-TOF-MS spectra of (A) *B. amyloliquefaciens* B0177, (B) *B. sphaericus* B0769, (C) *B. megaterium* B0010^T, (D) *B. cereus* B0002, (E) *B. licheniformis* B1379, (F) *B. subtilus* B1382 and (G) *Br. laterosporus* B0034. The panel to the right of (G) is a zoomed in region (highlighted with an ellipse) of the MALDI-TOF-MS spectrum from *Br. laterosporus* B0034. These spectra have been baseline corrected and normalized, so that the sum of each squared spectrum equals to 1.

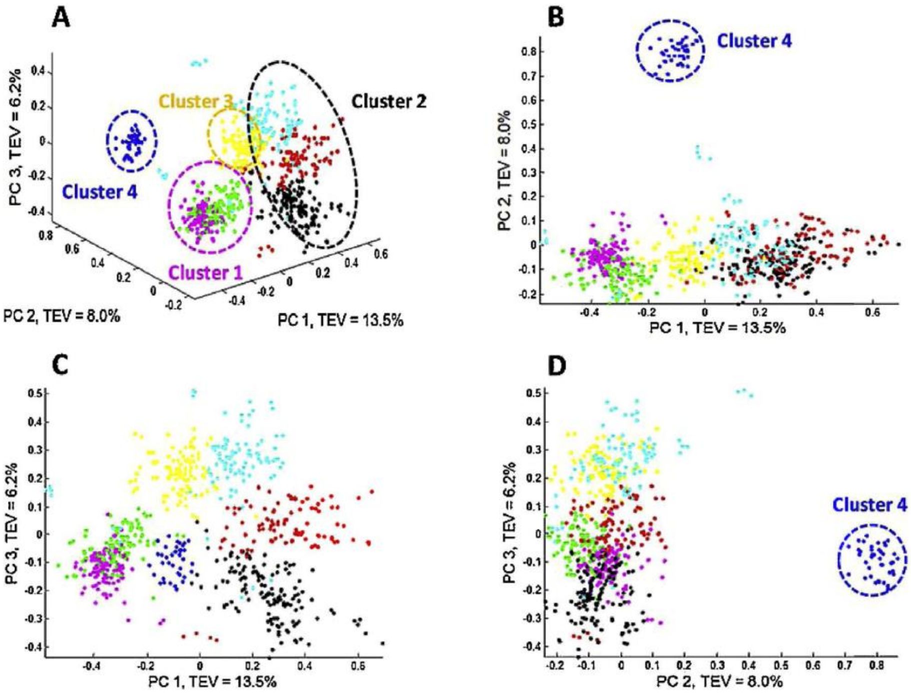


Fig. 3. PCA scores plots from the peak table matrix after pre-processing the MS data. Multiple principal components are plotted: (A) PC1 *vs.* PC2 *vs.* PC3; (B) PC1 *vs.* PC2; (C) PC1 *vs.* PC3 and (D) PC2 *vs.* PC3. The colours represent the different species see Table 1 for annotations. TEV: total explained variance for the PC score plotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observed: (1) the first contained *Bacillus megaterium* and *B. cereus*; (2) comprised *B. subtilus*, *B. amyloliquefaciens* and *B. licheniformis*; (3) contained only *B. sphaericus*; and (4) was also a single member cluster of *Br. laterosporus* (see Fig. 3A for an annotated 3-D representation). The MALDI-TOF-MS spectra obtained from the analysis of *Br. laterosporus* (Fig. 2G) were very different to the spectra from the other *Bacillus* species and this was reflected in PCA clusters (Fig. 3). As can be seen *Br. laterosporus* strains were significantly different in PC2 (Fig. 3B and D) which is why when PC2 *vs.* PC3 were plotted the groupings of the other 3 clusters were revealed. This was perhaps not surprising as this species belonged to a different bacilli genus, namely *Brevibacillus*.

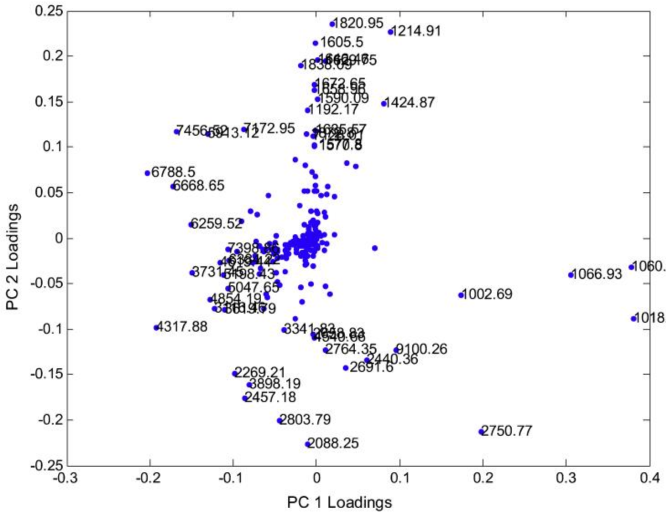


Fig. 4. PCA Loadings plots from the peak table matrix after pre-processing the MS data.

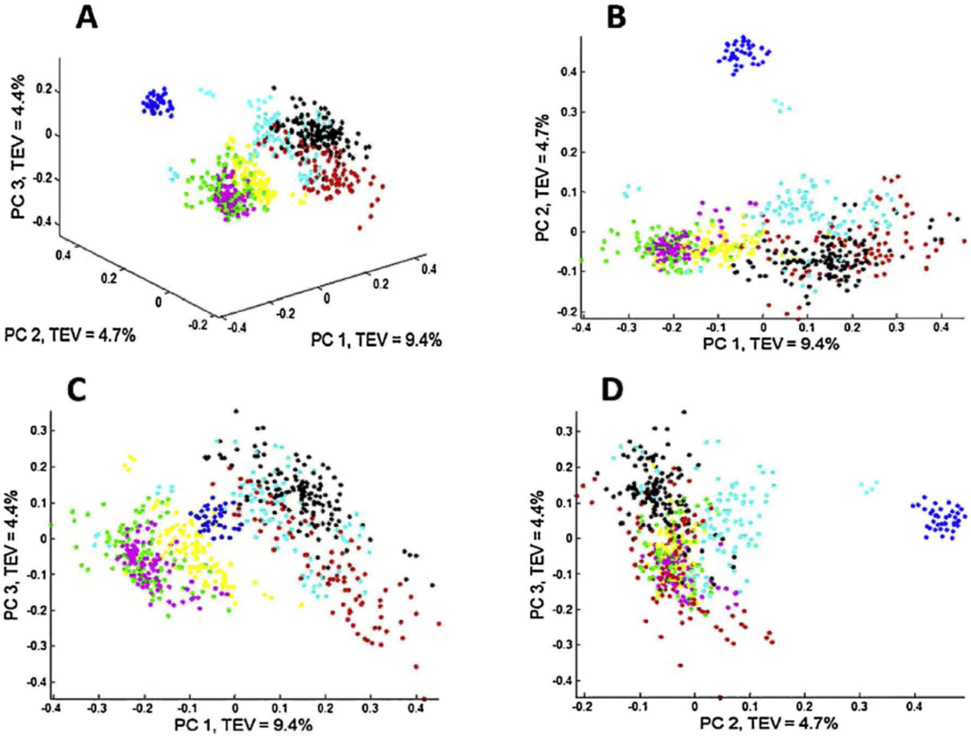


Fig. 5. PCoA scores plots of the data obtained to show clusters of present and absent peaks using the Jaccard distance model. Multiple principal components are plotted: (A) PC1 *vs.* PC2 *vs.* PC3; (B) PC1 *vs.* PC2; (C) PC1 *vs.* PC3 and (D) PC2 *vs.* PC3. The colours represent the different species see Table 1 for annotations. TEV: total explained variance for the PC score plotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The reason for choosing this set of bacilli is that these species have previously been analysed using a range of classification approaches including miniaturised biochemical test Analytical Profile Index (API), genotyping using 16S rDNA sequencing and an alternative physciochemical methods to MALDI-MS called Raman spectroscopy that measures molecular vibrations of functional groups. Based on the API tests these bacteria have been placed into four different groups [50] consisting of: (I) *B. cereus*; (II) *Br. laterosporus*; (III) *B. sphaericus*; (IV) *B. megaterium*, *B. subtilis*, *B. licheniformis* and *B. amyloliquefaciens*. Slightly different clusters were also previously found from 16S rDNA analysis: clusters (I); (II) and (III) from the API were also seen, but the *B. subtilis* group (comprising *B. subtilis*, *B. licheniformis* and *B. amyloliquefaciens*) was split from *B. megaterium*; in addition, although clustered separated *B. cereus* and *B. megaterium* were relatively close relatives at the genetic level [28,39]. The grouping generated from our MALDI-TOF-MS analysis is therefore highly congruent with both phenotypic (API) and phylogenetic markers (16S rDNA), as well as other biophysical characterization methods based on UV resonance Raman spectroscopy [39].

The results above used the quantitative data from the peak intensities, or at least the $\log_{10}$ of the signal to try and make the data appear normally distributed. In preliminary analyses we also attempted square root scaling and this produced similar results; for brevity we report only $\log_{10}$ here. As detailed in the materials and methods we also processed the data so that they were considered qualitative in nature; that is to say, we encoded the mass ions as being present (1) or absent (0). The purpose of employing such a strategy is to test whether such greatly simplified information is still sufficient to discriminate different types of bacteria, either on species level or strain level. Moreover, this would compensate for the fact that MALDI-TOF-MS is not considered truly quantitative. We, and others, have observed differences in the ion intensities of proteins from intact bacteria [19,26] and this significant variation in the peak intensities can be due to various analytical reasons. These are most likely due to small changes in bacteria growth, sample handling and the formation of different co-crystals with the matrix 'spot' [51,52]. If this qualitative approach were successful, it would suggest that the characterization of the bacteria based on the MALDI-TOF-MS spectra is in fact not sensitive to such variations and would suggest that MALDI-TOF-MS, as an analytical platform, is robust for bacterial analyses. Moreover, those features which had high probabilities of occurrence in some types of bacteria while absent or much rarer in other types could have significant biological implications and perhaps worth further investigation. Therefore PCoA was performed on the binary peak table matrix and resulted in a highly similar pattern (Fig. 5) to the one showed in the PCA scores plot (Fig. 3). This had suggested that based on the information of presence/absence of the features, it was indeed possible to discriminate bacteria on species level.

4.3. Automated identification of Bacillus from their MALDI-TOF-MS spectra

The next stage was to assess whether the information from the MALDI-TOF-MS data were discriminative enough to allow identification using supervised learning methods. The results of

Table 2
Prediction accuracies of the seven species from *Bacillus* using DAG-SVM with the linear kernel model.

	B. am (%)	*B. ce* (%)	*Br. la* (%)	*B. li* (%)	*B. me* (%)	*B. sp* (%)	*B. su* (%)
B. am	92.56	0.13	0.00	0.11	0.58	0.95	5.68
B. ce	3.37	83.37	0.00	0.12	11.27	1.82	0.05
Br. la	0.00	0.00	100.00	0.00	0.00	0.00	0.00
B. li	5.28	1.41	0.00	80.26	2.65	3.93	6.47
B. me	0.10	9.22	0.00	0.00	90.67	0.01	0.00
B. sp	1.37	1.91	0.00	2.41	0.09	94.23	0.01
B. su	6.46	0.00	0.00	2.13	0.00	0.00	91.42

B. am: *B. amyloliquefaciens*, *B. ce*: *B. cereus*, *Br. la*: *Br. laterosporus*, *B. li*: *B. licheniformis*, *B. me*: *B. megaterium*, *B. sp*: *B. sphaericus* and *B. su*: *B. subtilis*.

these classifications performed at the species level (i.e. 7 classes to be predicted) are given in Tables 2 and 3 using support vector machines (SVM) for the semi-quantitative and qualitative data, respectively. While prediction accuracies at the strain level (i.e. 34 class prediction) are provided in SI Tables S6 and S7.

It is very interesting to see that the SVM with Jaccard kernel (i.e. the SVM model based on the presence/absence information) and the SVM with Linear kernel gave almost identical prediction accuracies. This suggests that the qualitative information on protein content is sufficient to effect accurate classification, rather than the level of the proteins in the bacterial cells.

For the species classification models, the SVM with a linear kernel had an average correct classification rate (CCR) of 89.27% and the SVM with the Jaccard kernel providing 88.92% average CCR. The naive Bayesian classifier accuracy was slightly worse (77.69% average CCR). For all classification models *Br. laterosporus* was never mis-classified which is perhaps unsurprising as it is a difference genus. *B. cereus* and *B. megaterium* were sometimes misclassified as each other, which was also to be expected as these are phylogenetically similar [50]. Finally, the *B. subtilis* group comprising *B. amyloliquefaciens*, *B. licheniformis* and *B. subtilis* which are similar at the biochemical and genetic level [53] were also occasionally misclassified as each other. If these were taken as a single group the classification for these three species (e.g. in Table 2) would increase from 91.29%, 78.64%, 94.31% to 97.57%, 92.10%, 100% for *B. amyloliquefaciens*, *B. licheniformis* and *B. subtilis*, respectively. The fact that such observations were consistent across all the classification models indicates this is a model independent general trend and a reflection of the phenotypic characteristics being measured using MALDI-TOF-MS.

The CCRs of the classification models for strain ($n = 34$) classification is as expected much worse than those at the species level. The average CCR for these models ranged from 45.88% to 54.04% (SI Tables S7 and S6) for the qualitative and semi-quantitative models. As expected the misclassification of these bacterial strains usually occurred within the same species but to different strains. These may seem poor but considering the fact that there were 34 strains analysed this is a large number of classes and the expected CCR from a random classification model would be only 2.9%. Therefore the prediction accuracies of these models were still very impressive. It was also notable that the semi-quantitative classifier was ~9% better than the qualitative model which suggests that unlike the species classification the information on the peak intensities might also be required to achieve better discrimination between the strains.

Table 3
Prediction accuracies of the seven species from *Bacillus* using DAG-SVM with the Jaccard kernel model.

	B. am (%)	*B. ce* (%)	*Br. la* (%)	*B. li* (%)	*B. me* (%)	*B. sp* (%)	*B. su* (%)
B. am	91.29	0.23	0.00	0.14	0.85	1.36	6.14
B. ce	3.09	81.75	0.00	0.02	12.26	2.67	0.22
Br. la	0.00	0.00	100.00	0.00	0.00	0.00	0.00
B. li	5.67	2.57	0.00	78.64	1.62	3.71	7.79
B. me	0.04	8.79	0.00	0.01	91.12	0.04	0.00
B. sp	1.06	4.17	0.00	1.98	0.30	92.45	0.05
B. su	4.23	0.00	0.00	1.46	0.00	0.00	94.31

B. am: *B. amyloliquefaciens*, *B. ce*: *B. cereus*, *Br. la*: *Br. laterosporus*, *B. li*: *B. licheniformis*, *B. me*: *B. megaterium*, *B. sp*: *B. sphaericus* and *B. su*: *B. subtilis*.

5. Concluding remarks

MALDI-TOF-MS is gaining popularity for microbial classification and identification [54–57]. This results in information on the protein content of the organism under study and this proteomic barcode can be used to characterise the bacteria under investigation. However, in order to generate a consistent barcode the analytical approach must be optimised and tested. In this study we assessed 10 different matrices with 4 different sample preparation approaches. These 40 conditions were first applied to protein mixtures and the top 3 matrices-preparation methods were then assessed for reproducibility and for the generation of information rich protein profiles on 6 bacteria. This established that sinapinic acid with the mixed sample preparation approach was the preferred method, which is in agreement with other studies [46,58].

This matrix was then used on all 34 bacilli and each bacteria was grown 5 times and each of these biological replicates were analysed 4 times (technical replicates). These 680 MALDI-TOF-MS spectra were collected over a period of 2 weeks. Due to the extended mass range over which the spectra were collected (1000–13,000 m/z) significant drift in the m/z X-axis was observed which if not corrected would adversely affect bacterial characterization. This was successfully overcome by aligning the peaks using interval correlation optimized shifting. Preprocessing also involved using asymmetric least squares for baseline removal. Chemometric classifiers were then used on these data and the same data after peak picking using intensity weighted variance. This peak picking reduced the dimensionality of the MS data from a massive 680 samples × 111,339 m/z channels (75,710,520 data points) to a mere 680 × 243 (165,240 data points) and this process did not negatively affect classification.

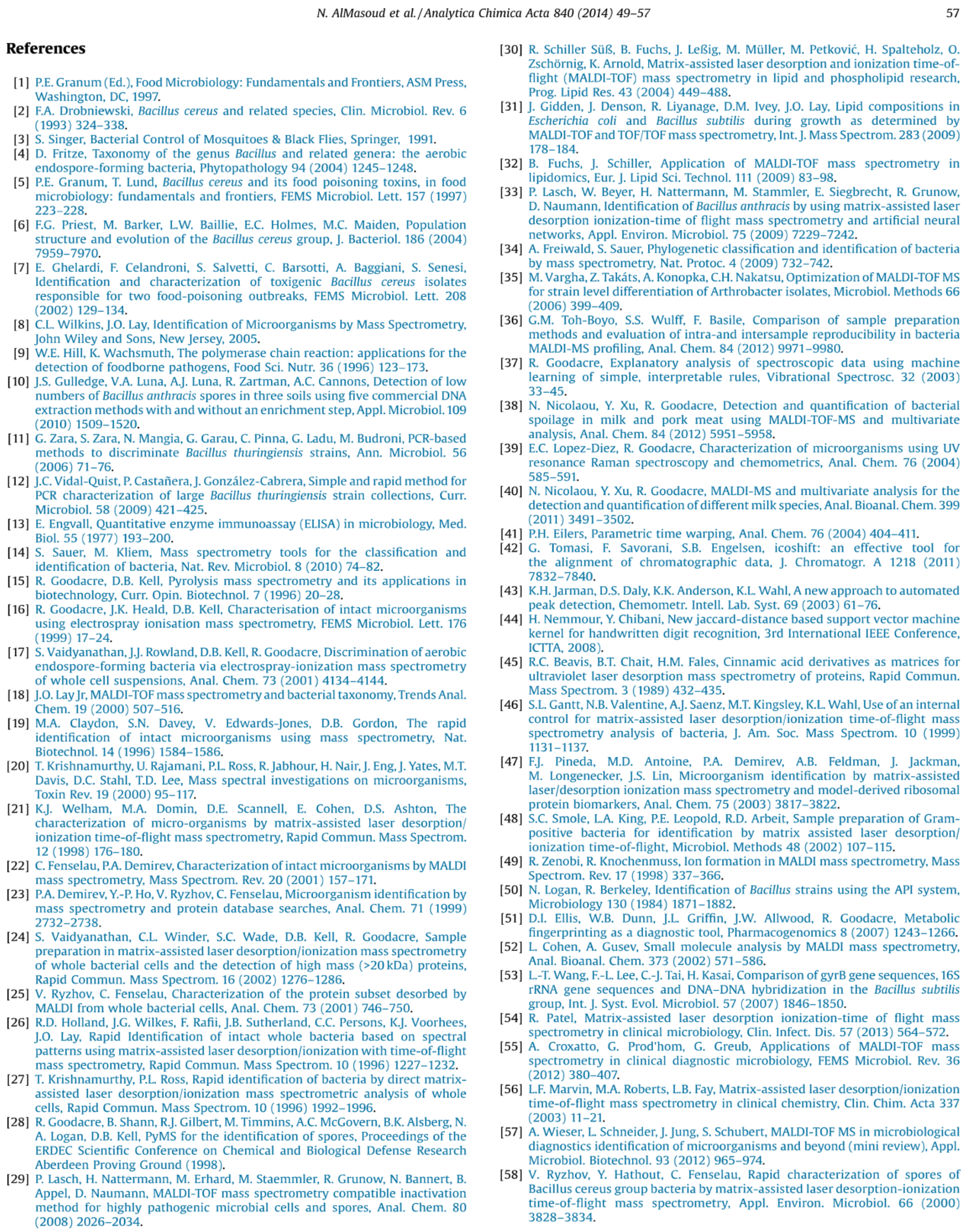
Classification accuracies at *Bacillus* species level were ~90% for the 7 species under analysis and this was robustly tested using bootstrap analysis. The few misclassifications that were made could be readily explained by very close species similarity of the *B. subtilis* group (*viz. B. amyloliquefaciens*, *B. licheniformis* and *B. subtilis*). In conclusion we have developed a robust MALDI-TOF-MS data collection and data analysis pipeline that we shall now expand to the analysis of other bacterial groups.

Acknowledgments

NM thanks The Saudi Ministry of higher education and Princess Nora bint Abdul Rahman University for funding. YX thanks the Cancer Research UK (including Experimental Cancer Medicine Centre award) and the Wolfson Foundation for funding, and NN is indebted to UMIP for financial support. RG thanks BBSRC for support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aca.2014.06.032.



References

[1] P.E. Granum (Ed.), Food Microbiology: Fundamentals and Frontiers, ASM Press, Washington, DC, 1997.

[2] F.A. Drobniewski, *Bacillus cereus* and related species, Clin. Microbiol. Rev. 6 (1993) 324–338.

[3] S. Singer, Bacterial Control of Mosquitoes & Black Flies, Springer, 1991.

[4] D. Fritze, Taxonomy of the genus *Bacillus* and related genera: the aerobic endospore-forming bacteria, Phytopathology 94 (2004) 1245–1248.

[5] P.E. Granum, T. Lund, *Bacillus cereus* and its food poisoning toxins, in food microbiology: fundamentals and frontiers, FEMS Microbiol. Lett. 157 (1997) 223–228.

[6] F.G. Priest, M. Barker, L.W. Baillie, E.C. Holmes, M.C. Maiden, Population structure and evolution of the *Bacillus cereus* group, J. Bacteriol. 186 (2004) 7959–7970.

[7] E. Ghelardi, F. Celandroni, S. Salvetti, C. Barsotti, A. Baggiani, S. Senesi, Identification and characterization of toxigenic *Bacillus cereus* isolates responsible for two food-poisoning outbreaks, FEMS Microbiol. Lett. 208 (2002) 129–134.

[8] C.L. Wilkins, J.O. Lay, Identification of Microorganisms by Mass Spectrometry, John Wiley and Sons, New Jersey, 2005.

[9] W.E. Hill, K. Wachsmuth, The polymerase chain reaction: applications for the detection of foodborne pathogens, Food Sci. Nutr. 36 (1996) 123–173.

[10] J.S. Gulledge, V.A. Luna, A.J. Luna, R. Zartman, A.C. Cannons, Detection of low numbers of *Bacillus anthracis* spores in three soils using five commercial DNA extraction methods with and without an enrichment step, Appl. Microbiol. 109 (2010) 1509–1520.

[11] G. Zara, S. Zara, N. Mangia, G. Garau, C. Pinna, G. Ladu, M. Budroni, PCR-based methods to discriminate *Bacillus thuringiensis* strains, Ann. Microbiol. 56 (2006) 71–76.

[12] J.C. Vidal-Quist, P. Castañera, J. González-Cabrera, Simple and rapid method for PCR characterization of large *Bacillus thuringiensis* strain collections, Curr. Microbiol. 58 (2009) 421–425.

[13] E. Engvall, Quantitative enzyme immunoassay (ELISA) in microbiology, Med. Biol. 55 (1977) 193–200.

[14] S. Sauer, M. Kliem, Mass spectrometry tools for the classification and identification of bacteria, Nat. Rev. Microbiol. 8 (2010) 74–82.

[15] R. Goodacre, D.B. Kell, Pyrolysis mass spectrometry and its applications in biotechnology, Curr. Opin. Biotechnol. 7 (1996) 20–28.

[16] R. Goodacre, J.K. Heald, D.B. Kell, Characterisation of intact microorganisms using electrospray ionisation mass spectrometry, FEMS Microbiol. Lett. 176 (1999) 17–24.

[17] S. Vaidyanathan, J.J. Rowland, D.B. Kell, R. Goodacre, Discrimination of aerobic endospore-forming bacteria via electrospray-ionization mass spectrometry of whole cell suspensions, Anal. Chem. 73 (2001) 4134–4144.

[18] J.O. Lay Jr, MALDI-TOF mass spectrometry and bacterial taxonomy, Trends Anal. Chem. 19 (2000) 507–516.

[19] M.A. Claydon, S.N. Davey, V. Edwards-Jones, D.B. Gordon, The rapid identification of intact microorganisms using mass spectrometry, Nat. Biotechnol. 14 (1996) 1584–1586.

[20] T. Krishnamurthy, U. Rajamani, P.L. Ross, R. Jabhour, H. Nair, J. Eng, J. Yates, M.T. Davis, D.C. Stahl, T.D. Lee, Mass spectral investigations on microorganisms, Toxin Rev. 19 (2000) 95–117.

[21] K.J. Welham, M.A. Domin, D.E. Scannell, E. Cohen, D.S. Ashton, The characterization of micro-organisms by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 12 (1998) 176–180.

[22] C. Fenselau, P.A. Demirev, Characterization of intact microorganisms by MALDI mass spectrometry, Mass Spectrom. Rev. 20 (2001) 157–171.

[23] P.A. Demirev, Y.-P. Ho, V. Ryzhov, C. Fenselau, Microorganism identification by mass spectrometry and protein database searches, Anal. Chem. 71 (1999) 2732–2738.

[24] S. Vaidyanathan, C.L. Winder, S.C. Wade, D.B. Kell, R. Goodacre, Sample preparation in matrix-assisted laser desorption/ionization mass spectrometry of whole bacterial cells and the detection of high mass (>20 kDa) proteins, Rapid Commun. Mass Spectrom. 16 (2002) 1276–1286.

[25] V. Ryzhov, C. Fenselau, Characterization of the protein subset desorbed by MALDI from whole bacterial cells, Anal. Chem. 73 (2001) 746–750.

[26] R.D. Holland, J.G. Wilkes, F. Rafii, J.B. Sutherland, C.C. Persons, K.J. Voorhees, J.O. Lay, Rapid Identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 10 (1996) 1227–1232.

[27] T. Krishnamurthy, P.L. Ross, Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells, Rapid Commun. Mass Spectrom. 10 (1996) 1992–1996.

[28] R. Goodacre, B. Shann, R.J. Gilbert, M. Timmins, A.C. McGovern, B.K. Alsberg, N. A. Logan, D.B. Kell, PyMS for the identification of spores, Proceedings of the ERDEC Scientific Conference on Chemical and Biological Defense Research Aberdeen Proving Ground (1998).

[29] P. Lasch, H. Nattermann, M. Erhard, M. Staemmler, R. Grunow, N. Bannert, B. Appel, D. Naumann, MALDI-TOF mass spectrometry compatible inactivation method for highly pathogenic microbial cells and spores, Anal. Chem. 80 (2008) 2026–2034.

[30] R. Schiller Süß, B. Fuchs, J. Leßig, M. Müller, M. Petković, H. Spalteholz, O. Zschörnig, K. Arnold, Matrix-assisted laser desorption and ionization time-of-flight (MALDI-TOF) mass spectrometry in lipid and phospholipid research, Prog. Lipid Res. 43 (2004) 449–488.

[31] J. Gidden, J. Denson, R. Liyanage, D.M. Ivey, J.O. Lay, Lipid compositions in *Escherichia coli* and *Bacillus subtilis* during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry, Int. J. Mass Spectrom. 283 (2009) 178–184.

[32] B. Fuchs, J. Schiller, Application of MALDI-TOF mass spectrometry in lipidomics, Eur. J. Lipid Sci. Technol. 111 (2009) 83–98.

[33] P. Lasch, W. Beyer, H. Nattermann, M. Stammler, E. Siegbrecht, R. Grunow, D. Naumann, Identification of *Bacillus anthracis* by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks, Appl. Environ. Microbiol. 75 (2009) 7229–7242.

[34] A. Freiwald, S. Sauer, Phylogenetic classification and identification of bacteria by mass spectrometry, Nat. Protoc. 4 (2009) 732–742.

[35] M. Vargha, Z. Takáts, A. Konopka, C.H. Nakatsu, Optimization of MALDI-TOF MS for strain level differentiation of Arthrobacter isolates, Microbiol. Methods 66 (2006) 399–409.

[36] G.M. Toh-Boyo, S.S. Wulff, F. Basile, Comparison of sample preparation methods and evaluation of intra-and intersample reproducibility in bacteria MALDI-MS profiling, Anal. Chem. 84 (2012) 9971–9980.

[37] R. Goodacre, Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules, Vibrational Spectrosc. 32 (2003) 33–45.

[38] N. Nicolaou, Y. Xu, R. Goodacre, Detection and quantification of bacterial spoilage in milk and pork meat using MALDI-TOF-MS and multivariate analysis, Anal. Chem. 84 (2012) 5951–5958.

[39] E.C. Lopez-Diez, R. Goodacre, Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics, Anal. Chem. 76 (2004) 585–591.

[40] N. Nicolaou, Y. Xu, R. Goodacre, MALDI-MS and multivariate analysis for the detection and quantification of different milk species, Anal. Bioanal. Chem. 399 (2011) 3491–3502.

[41] P.H. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411.

[42] G. Tomasi, F. Savorani, S.B. Engelsen, icoshift: an effective tool for the alignment of chromatographic data, J. Chromatogr. A 1218 (2011) 7832–7840.

[43] K.H. Jarman, D.S. Daly, K.K. Anderson, K.L. Wahl, A new approach to automated peak detection, Chemometr. Intell. Lab. Syst. 69 (2003) 61–76.

[44] H. Nemmour, Y. Chibani, New jaccard-distance based support vector machine kernel for handwritten digit recognition, 3rd International IEEE Conference, ICTTA, 2008).

[45] R.C. Beavis, B.T. Chait, H.M. Fales, Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins, Rapid Commun. Mass Spectrom. 3 (1989) 432–435.

[46] S.L. Gantt, N.B. Valentine, A.J. Saenz, M.T. Kingsley, K.L. Wahl, Use of an internal control for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis of bacteria, J. Am. Soc. Mass Spectrom. 10 (1999) 1131–1137.

[47] F.J. Pineda, M.D. Antoine, P.A. Demirev, A.B. Feldman, J. Jackman, M. Longenecker, J.S. Lin, Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers, Anal. Chem. 75 (2003) 3817–3822.

[48] S.C. Smole, L.A. King, P.E. Leopold, R.D. Arbeit, Sample preparation of Gram-positive bacteria for identification by matrix assisted laser desorption/ionization time-of-flight, Microbiol. Methods 48 (2002) 107–115.

[49] R. Zenobi, R. Knochenmuss, Ion formation in MALDI mass spectrometry, Mass Spectrom. Rev. 17 (1998) 337–366.

[50] N. Logan, R. Berkeley, Identification of *Bacillus* strains using the API system, Microbiology 130 (1984) 1871–1882.

[51] D.I. Ellis, W.B. Dunn, J.L. Griffin, J.W. Allwood, R. Goodacre, Metabolic fingerprinting as a diagnostic tool, Pharmacogenomics 8 (2007) 1243–1266.

[52] L. Cohen, A. Gusev, Small molecule analysis by MALDI mass spectrometry, Anal. Bioanal. Chem. 373 (2002) 571–586.

[53] L.-T. Wang, F.-L. Lee, C.-J. Tai, H. Kasai, Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA–DNA hybridization in the *Bacillus subtilis* group, Int. J. Syst. Evol. Microbiol. 57 (2007) 1846–1850.

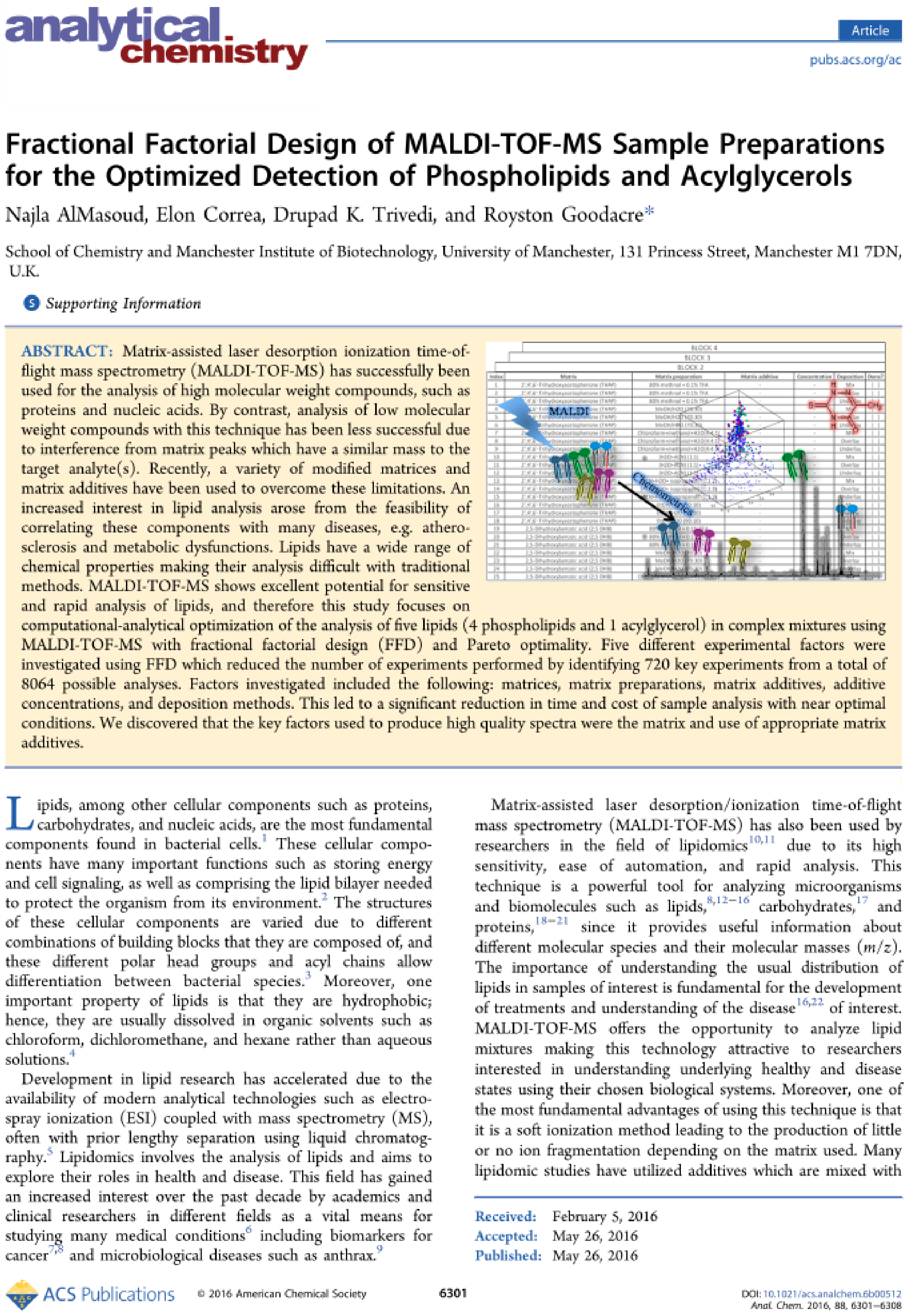
[54] R. Patel, Matrix-assisted laser desorption ionization-time of flight mass spectrometry in clinical microbiology, Clin. Infect. Dis. 57 (2013) 564–572.

[55] A. Croxatto, G. Prod'hom, G. Greub, Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology, FEMS Microbiol. Rev. 36 (2012) 380–407.

[56] L.F. Marvin, M.A. Roberts, L.B. Fay, Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry, Clin. Chim. Acta 337 (2003) 11–21.

[57] A. Wieser, L. Schneider, J. Jung, S. Schubert, MALDI-TOF MS in microbiological diagnostics identification of microorganisms and beyond (mini review), Appl. Microbiol. Biotechnol. 93 (2012) 965–974.

[58] V. Ryzhov, Y. Hathout, C. Fenselau, Rapid characterization of spores of Bacillus cereus group bacteria by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry, Appl. Environ. Microbiol. 66 (2000) 3828–3834.



analytical chemistry

Article

pubs.acs.org/ac

Fractional Factorial Design of MALDI-TOF-MS Sample Preparations for the Optimized Detection of Phospholipids and Acylglycerols

Najla AlMasoud, Elon Correa, Drupad K. Trivedi, and Royston Goodacre*

School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester M1 7DN, U.K.

Supporting Information

ABSTRACT: Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS) has successfully been used for the analysis of high molecular weight compounds, such as proteins and nucleic acids. By contrast, analysis of low molecular weight compounds with this technique has been less successful due to interference from matrix peaks which have a similar mass to the target analyte(s). Recently, a variety of modified matrices and matrix additives have been used to overcome these limitations. An increased interest in lipid analysis arose from the feasibility of correlating these components with many diseases, e.g. atherosclerosis and metabolic dysfunctions. Lipids have a wide range of chemical properties making their analysis difficult with traditional methods. MALDI-TOF-MS shows excellent potential for sensitive and rapid analysis of lipids, and therefore this study focuses on computational-analytical optimization of the analysis of five lipids (4 phospholipids and 1 acylglycerol) in complex mixtures using MALDI-TOF-MS with fractional factorial design (FFD) and Pareto optimality. Five different experimental factors were investigated using FFD which reduced the number of experiments performed by identifying 720 key experiments from a total of 8064 possible analyses. Factors investigated included the following: matrices, matrix preparations, matrix additives, additive concentrations, and deposition methods. This led to a significant reduction in time and cost of sample analysis with near optimal conditions. We discovered that the key factors used to produce high quality spectra were the matrix and use of appropriate matrix additives.

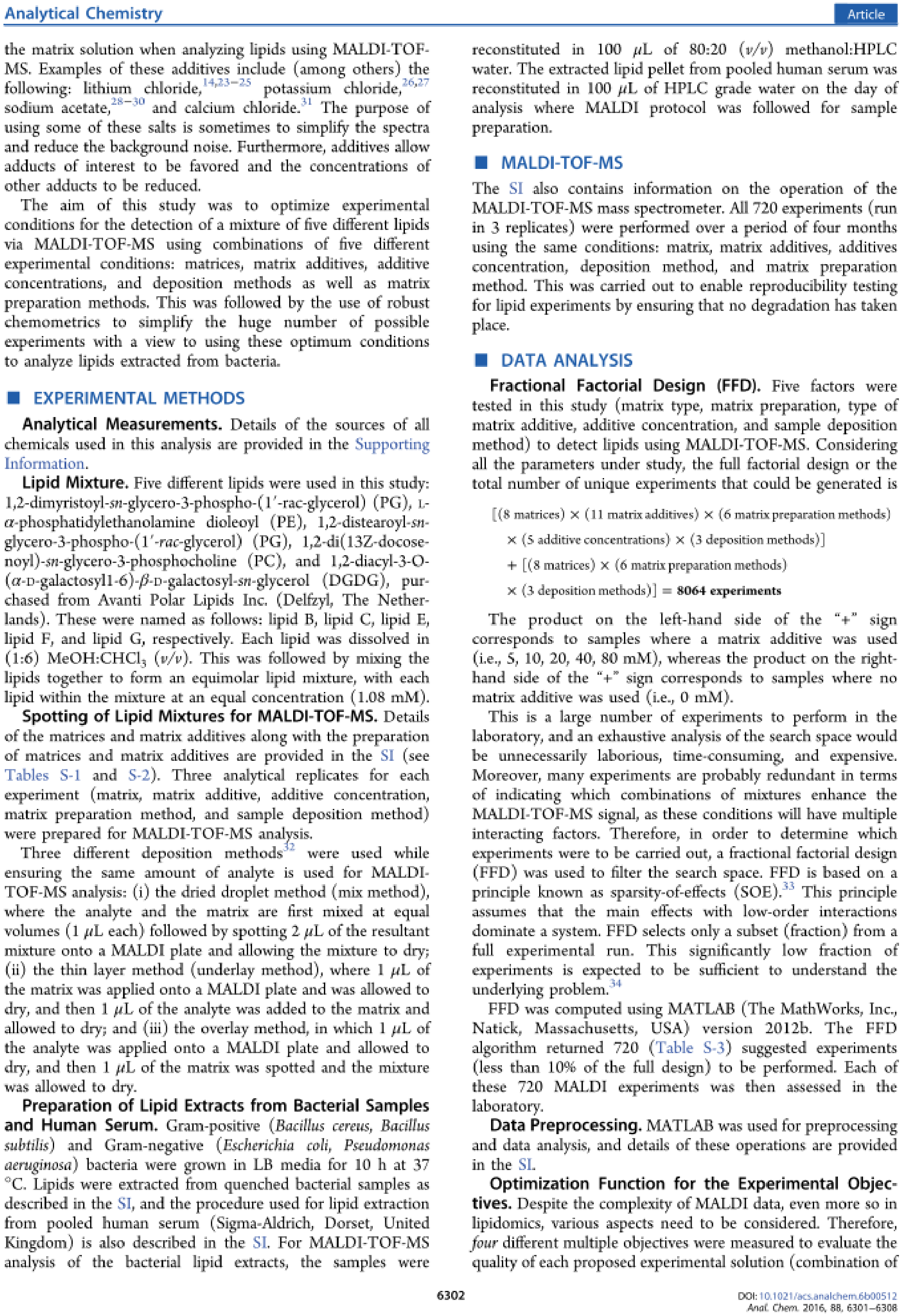
Lipids, among other cellular components such as proteins, carbohydrates, and nucleic acids, are the most fundamental components found in bacterial cells.[1] These cellular components have many important functions such as storing energy and cell signaling, as well as comprising the lipid bilayer needed to protect the organism from its environment.[2] The structures of these cellular components are varied due to different combinations of building blocks that they are composed of, and these different polar head groups and acyl chains allow differentiation between bacterial species.[3] Moreover, one important property of lipids is that they are hydrophobic; hence, they are usually dissolved in organic solvents such as chloroform, dichloromethane, and hexane rather than aqueous solutions.[4]

Development in lipid research has accelerated due to the availability of modern analytical technologies such as electrospray ionization (ESI) coupled with mass spectrometry (MS), often with prior lengthy separation using liquid chromatography.[5] Lipidomics involves the analysis of lipids and aims to explore their roles in health and disease. This field has gained an increased interest over the past decade by academics and clinical researchers in different fields as a vital means for studying many medical conditions[6] including biomarkers for cancer[7,8] and microbiological diseases such as anthrax.[9]

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) has also been used by researchers in the field of lipidomics[10,11] due to its high sensitivity, ease of automation, and rapid analysis. This technique is a powerful tool for analyzing microorganisms and biomolecules such as lipids,[8,12−16] carbohydrates,[17] and proteins,[18−21] since it provides useful information about different molecular species and their molecular masses (m/z). The importance of understanding the usual distribution of lipids in samples of interest is fundamental for the development of treatments and understanding of the disease[16,22] of interest. MALDI-TOF-MS offers the opportunity to analyze lipid mixtures making this technology attractive to researchers interested in understanding underlying healthy and disease states using their chosen biological systems. Moreover, one of the most fundamental advantages of using this technique is that it is a soft ionization method leading to the production of little or no ion fragmentation depending on the matrix used. Many lipidomic studies have utilized additives which are mixed with

Received: February 5, 2016
Accepted: May 26, 2016
Published: May 26, 2016

DOI: 10.1021/acs.analchem.6b00512



the matrix solution when analyzing lipids using MALDI-TOF-MS. Examples of these additives include (among others) the following: lithium chloride,[14,23−25] potassium chloride,[26,27] sodium acetate,[28−30] and calcium chloride.[31] The purpose of using some of these salts is sometimes to simplify the spectra and reduce the background noise. Furthermore, additives allow adducts of interest to be favored and the concentrations of other adducts to be reduced.

The aim of this study was to optimize experimental conditions for the detection of a mixture of five different lipids via MALDI-TOF-MS using combinations of five different experimental conditions: matrices, matrix additives, additive concentrations, and deposition methods as well as matrix preparation methods. This was followed by the use of robust chemometrics to simplify the huge number of possible experiments with a view to using these optimum conditions to analyze lipids extracted from bacteria.

■ EXPERIMENTAL METHODS

Analytical Measurements. Details of the sources of all chemicals used in this analysis are provided in the Supporting Information.

Lipid Mixture. Five different lipids were used in this study: 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) (PG), L-*α*-phosphatidylethanolamine dioleoyl (PE), 1,2-distearoyl-*sn*-glycero-3-phospho-(1′-*rac*-glycerol) (PG), 1,2-di(13Z-docosenoyl)-*sn*-glycero-3-phosphocholine (PC), and 1,2-diacyl-3-O-(*α*-D-galactosyl1-6)-*β*-D-galactosyl-*sn*-glycerol (DGDG), purchased from Avanti Polar Lipids Inc. (Delfzyl, The Netherlands). These were named as follows: lipid B, lipid C, lipid E, lipid F, and lipid G, respectively. Each lipid was dissolved in (1:6) MeOH:$CHCl_3$ (*v*/*v*). This was followed by mixing the lipids together to form an equimolar lipid mixture, with each lipid within the mixture at an equal concentration (1.08 mM).

Spotting of Lipid Mixtures for MALDI-TOF-MS. Details of the matrices and matrix additives along with the preparation of matrices and matrix additives are provided in the SI (see Tables S-1 and S-2). Three analytical replicates for each experiment (matrix, matrix additive, additive concentration, matrix preparation method, and sample deposition method) were prepared for MALDI-TOF-MS analysis.

Three different deposition methods[32] were used while ensuring the same amount of analyte is used for MALDI-TOF-MS analysis: (i) the dried droplet method (mix method), where the analyte and the matrix are first mixed at equal volumes (1 *μ*L each) followed by spotting 2 *μ*L of the resultant mixture onto a MALDI plate and allowing the mixture to dry; (ii) the thin layer method (underlay method), where 1 *μ*L of the matrix was applied onto a MALDI plate and was allowed to dry, and then 1 *μ*L of the analyte was added to the matrix and allowed to dry; and (iii) the overlay method, in which 1 *μ*L of the analyte was applied onto a MALDI plate and allowed to dry, and then 1 *μ*L of the matrix was spotted and the mixture was allowed to dry.

Preparation of Lipid Extracts from Bacterial Samples and Human Serum. Gram-positive (*Bacillus cereus, Bacillus subtilis*) and Gram-negative (*Escherichia coli, Pseudomonas aeruginosa*) bacteria were grown in LB media for 10 h at 37 °C. Lipids were extracted from quenched bacterial samples as described in the SI, and the procedure used for lipid extraction from pooled human serum (Sigma-Aldrich, Dorset, United Kingdom) is also described in the SI. For MALDI-TOF-MS analysis of the bacterial lipid extracts, the samples were reconstituted in 100 *μ*L of 80:20 (*v*/*v*) methanol:HPLC water. The extracted lipid pellet from pooled human serum was reconstituted in 100 *μ*L of HPLC grade water on the day of analysis where MALDI protocol was followed for sample preparation.

■ MALDI-TOF-MS

The SI also contains information on the operation of the MALDI-TOF-MS mass spectrometer. All 720 experiments (run in 3 replicates) were performed over a period of four months using the same conditions: matrix, matrix additives, additives concentration, deposition method, and matrix preparation method. This was carried out to enable reproducibility testing for lipid experiments by ensuring that no degradation has taken place.

■ DATA ANALYSIS

Fractional Factorial Design (FFD). Five factors were tested in this study (matrix type, matrix preparation, type of matrix additive, additive concentration, and sample deposition method) to detect lipids using MALDI-TOF-MS. Considering all the parameters under study, the full factorial design or the total number of unique experiments that could be generated is

$$\begin{aligned}&[(8\text{ matrices})\times(11\text{ matrix additives})\times(6\text{ matrix preparation methods})\\&\quad\times(5\text{ additive concentrations})\times(3\text{ deposition methods})]\\&\quad+[(8\text{ matrices})\times(6\text{ matrix preparation methods})\\&\quad\times(3\text{ deposition methods})]=\textbf{8064 experiments}\end{aligned}$$

The product on the left-hand side of the "+" sign corresponds to samples where a matrix additive was used (i.e., 5, 10, 20, 40, 80 mM), whereas the product on the right-hand side of the "+" sign corresponds to samples where no matrix additive was used (i.e., 0 mM).

This is a large number of experiments to perform in the laboratory, and an exhaustive analysis of the search space would be unnecessarily laborious, time-consuming, and expensive. Moreover, many experiments are probably redundant in terms of indicating which combinations of mixtures enhance the MALDI-TOF-MS signal, as these conditions will have multiple interacting factors. Therefore, in order to determine which experiments were to be carried out, a fractional factorial design (FFD) was used to filter the search space. FFD is based on a principle known as sparsity-of-effects (SOE).[33] This principle assumes that the main effects with low-order interactions dominate a system. FFD selects only a subset (fraction) from a full experimental run. This significantly low fraction of experiments is expected to be sufficient to understand the underlying problem.[34]

FFD was computed using MATLAB (The MathWorks, Inc., Natick, Massachusetts, USA) version 2012b. The FFD algorithm returned 720 (Table S-3) suggested experiments (less than 10% of the full design) to be performed. Each of these 720 MALDI experiments was then assessed in the laboratory.

Data Preprocessing. MATLAB was used for preprocessing and data analysis, and details of these operations are provided in the SI.

Optimization Function for the Experimental Objectives. Despite the complexity of MALDI data, even more so in lipidomics, various aspects need to be considered. Therefore, *four* different multiple objectives were measured to evaluate the quality of each proposed experimental solution (combination of

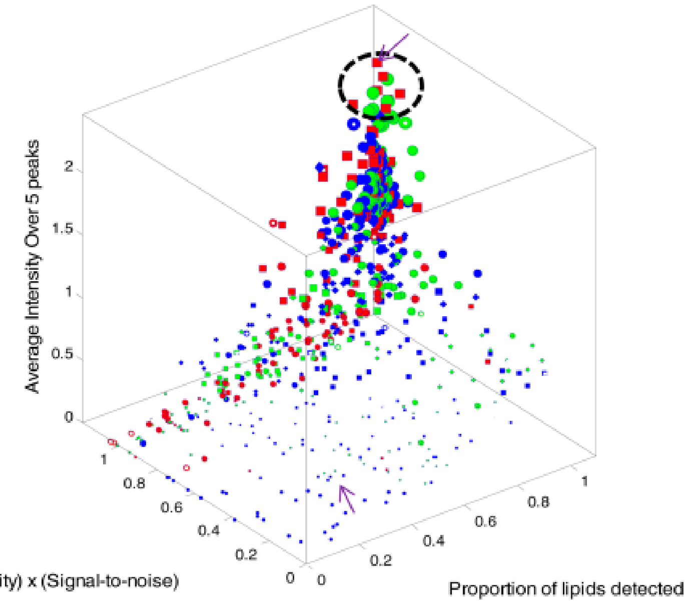


Figure 1. 3D scores plot of MALDI-TOF-MS data using multiple objectives measured for all 2160 spectra. The different characteristics (shapes and colors) represent different experimental conditions. The size of the shapes is proportional to the quality of the spectra; i.e. the bigger, the better. A key was not used to represent the different characteristics as 720 experiments conditions making this impossible. The purple arrow at the top of the 3D score plot indicates the top solution, and the bottom arrow indicates one of the worse solutions. These two solutions (indicated with arrows) were used on real biological samples (Figure S-8).

factors under study). These measurements synthesize highly desirable objectives that are in general difficult to achieve in MALDI experiments, namely: (i) high reproducibility, (ii) high signal-to-noise ratio, (iii) high peak intensity of each lipid under study, and (iv) detection of all lipids included in the mixture, since the correct identification of all lipids present in the sample is a fundamental requirement for systematically searching for the simplest possible global optimal solution. Figure S-1 depicts the four objectives simultaneously measured and optimized in this study.

Also provided in the SI are the following computations for the four objective functions:

- Objective function 1 – used to estimate highly reproducible spectra
- Objective function 2 – used to assess signal-to-noise ratio
- Objective function 3 – number of lipids detected
- Objective function 4 – estimation of the high peak intensity of each lipid

Pareto Optimality. The Pareto optimality (PO) principle was first introduced by Smilde et al.[35] Pareto optimality is defined by experiments having better results for some objectives in conjunction with possibly not as good results for other objectives. The aim of PO was to identify the samples (spectra) or points for which no other sample is better than them for at least one of the objectives measured.

As there are four objectives to be satisfied in this research (reproducibility, high signal-to-noise ratio, high peak intensity, and the number of lipids detected), we used the PO approach to identify relevant solutions. In PO, a solution is achieved when each objective, in our case four main objectives (note that peak intensity is subdivided into five objective functions one of each lipid under study), is optimized to the extent that it is acceptable to the decision maker and without other objectives suffering as a result of this process if further optimization were to take place.[36] In PO, a solution is considered valid and said to be "in the Pareto front" if no other solution dominates that solution in all objectives being measured. Otherwise, the solution is said to be dominated, is not in the Pareto front, and is rejected. As at the end of the process there may be many solutions in the Pareto front, the decision of which solution is the best depends on the user's expectations and requirements for each objective measured (Figure S-2 illustrates Pareto optimality). The objective function used for computation of PO is then given by the following multiobjective function.

$$f_{multi-objective} = \underbrace{snr}_{\boxed{= \frac{\sum \text{signal}}{\sum \text{noise}}}} + \underbrace{rep}_{\boxed{\mathrm{Rep}(x_k) = \frac{1}{3}\left(\frac{\mathrm{cov}(x_{k1}, x_{k2})}{\sqrt{\mathrm{var}(x_{k1})\mathrm{var}(x_{k2})}} + \frac{\mathrm{cov}(x_{k1}, x_{k3})}{\sqrt{\mathrm{var}(x_{k1})\mathrm{var}(x_{k3})}} + \frac{\mathrm{cov}(x_{k2}, x_{k3})}{\sqrt{\mathrm{var}(x_{k2})\mathrm{var}(x_{k3})}}\right)}} + \underbrace{\#_{lipids\ detected}}_{\boxed{\text{count of different lipids positively detected}}} + \underbrace{int_{lipidB}}_{\boxed{\text{Lipid B(intensity) = peak@688 + peak@710 + peak@726}}} + int_{lipidC} + int_{lipidE} + int_{lipidF} + int_{lipidG} \qquad (1)$$

thiothymine (ATT) and 2,6-dihydroxyacetophenone (DHAP), respectively. Moreover, the size of the shapes is proportional to the quality of the spectra according to eq 1 (the bigger, the better).

Referring to Figure 1, a red square at the top of the 3D scores plot is clearly seen. This square represents the overall best experiment from the 720 experiments carried out, and its corresponding spectrum is shown in Figure 2. This optimized experimental setup was further used to analyze lipid extracts from four different bacterial biological samples encompassing both Gram-positive and Gram-negative bacteria (*B. cereus, B. subtilis, E. coli, P. aeruginosa*) as well as human serum. The spectra for these five real-world samples are shown in Figure S-8A-E, and it is clear that this solution yields high quality spectra.

By contrast, the experiments that failed to detect the five lipids are shown toward the bottom of the plot in Figure 1. In this figure, a significant number of experiments are concentrated at the top of the 3D plot. The highlight of these experiments (circled in Figure 1) showed that ATT, DHB, and THAP matrices were found in this region. Hence, these matrices are the most compatible with the lipids analyzed. Experiments shown to be less suitable for the analysis of the lipid mixture also yielded very poor quality spectra when used to analyze lipid extracts from bacteria and serum as shown in Figure S-8F-J.

MALDI-TOF-MS Spectra of a Lipid Mixture. In this study, it was shown that with MALDI-TOF-MS, it is possible to analyze lipids in a mixture, but their detectability changed significantly[31] when changing the type of matrix, matrix additives, and the concentration of some matrix additives. Comparing the matrices that were mentioned in the previous section, it can be noted that ATT was the most compatible for the analysis of the five lipids in a mixture since it was the softest[43] and produced good shot-to-shot and sample-to-sample reproducibility,[49] in turn causing a substantially lower amount of fragmentation, reducing the background noise, and increasing both the signal-to-noise ratio and reproducibility, which produced an overall optimum experiment. These observations were also reported by Stübiger et al.[43] Moreover, it can be argued that other matrices, such as PNA and CMBT, could be used as alternatives in some cases as some of the lipids were detectable, and matrices such as DHAP were not used for further lipid analysis as poor spectra were produced regardless of whether matrix additives were used or not.[30] However, reproducibility with PNA was low, and hence it was decided that ATT was to be used in future lipid studies.

Figure 2 shows the positive ion MALDI-TOF-MS spectrum of the overall optimum experiment when using the following combination: mix deposition method, ATT as a matrix, H_2O:ACN (50:50) matrix preparation vehicle. The m/z range was from 600 to 1000, and the peaks that corresponded to each lipid in the mixture were assigned and summarized in the table shown within Figure 2. The green, purple, and red symbols in the spectrum and table correspond to H^+, Na^+ adduct, and K^+ adduct of each lipid, respectively. These adducts are usually detected when conducting such experiments. The spectrum corresponds to the precursor ions $[M + H]^+$ (m/z 688), $[M + Na]^+$ (m/z 710), and $[M + K]^+$ (m/z 726) for lipid B, and the protonated peak represents the most intense peak for lipid B. Moreover, the peaks at m/z 744, 766, and 782 correspond to lipid C and also represent $[M + H]^+$, $[M + Na]^+$, and $[M + K]^+$, respectively, and this time the sodium adduct peak dominates. These observations were also noticed with lipids E, F, and G, with the exception of the $[M + K]^+$ peak not being detected for lipid E. The ability to identify protonated molecules simplifies the interpretation of spectra.[47]

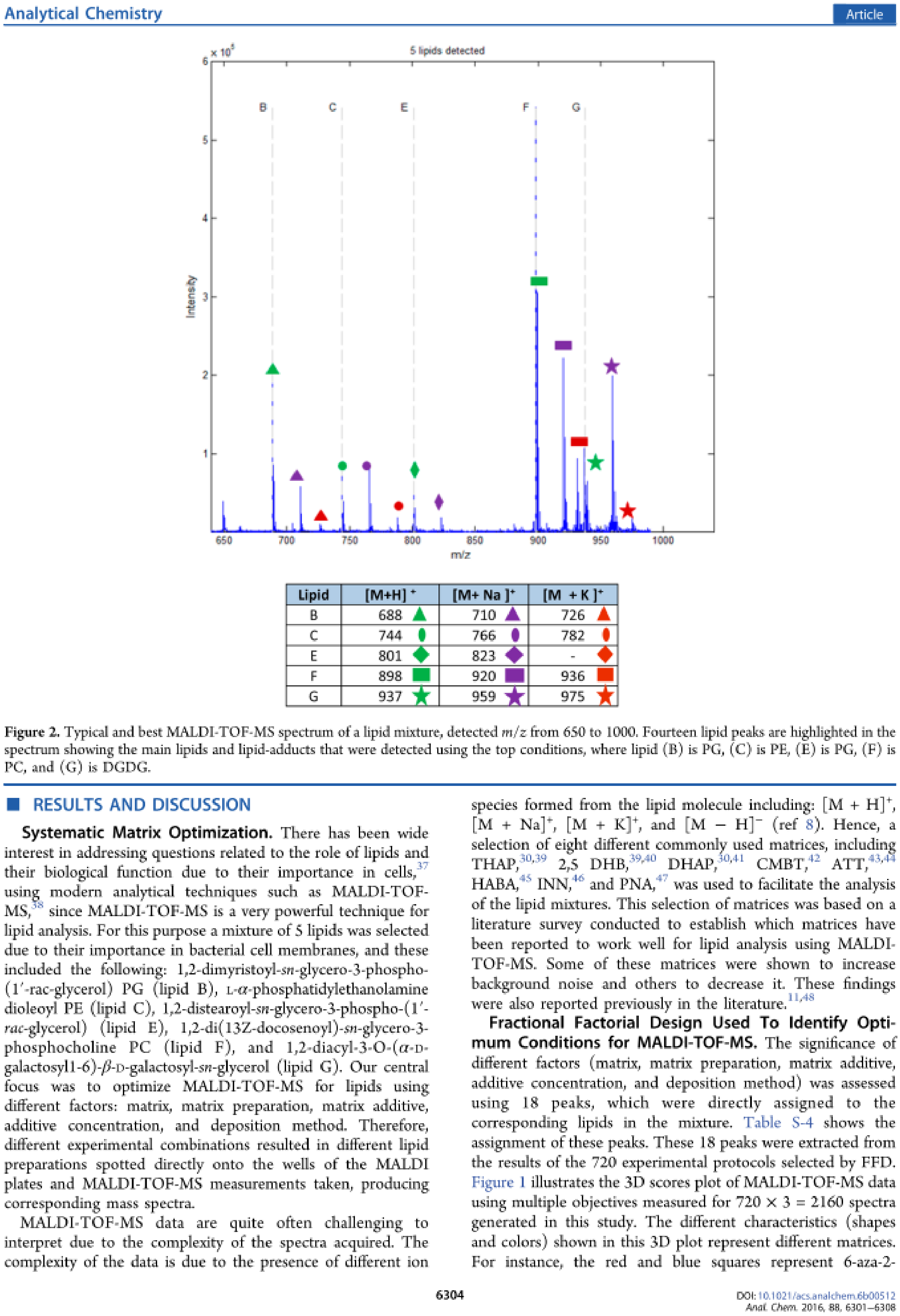
On the other hand, some poor complex MALDI-TOF-MS spectra were produced, which may be due to the immiscibility of the matrix solution and the lipid mixture, making crystallization inhomogeneous,[30] or due to failure of one or more of the four different multiple objectives measured. Figure S-3B shows an example of a poor spectrum produced for the lipid mixture when using the following combination: underlay deposition method, PNA dissolved in chloroform, MeOH, and H_2O with 80 mM lithium nitrate as a matrix additive. Furthermore, the spectrum slightly improved and was less complex when the concentration of the additive was reduced to 40 mM. From this observation, it can be seen in Figure S-3C that some of the lipids were detected using MALDI-TOF-MS, such as peaks at m/z 750, 904, and 943 which correspond to $[M + Li]^+$ for lipids C, F, and G, respectively, with the exception of lipids B and E, which were not detected. As discussed above, the positions of experiments within the 3D scores plot (Figure 1) are vital; hence, the position of the experimental condition in Figure S-3A was interesting. For example, when 40 mM lithium nitrate was used (Figure S-3C) compared with the position of the same experiment carried out using 80 mM lithium nitrate (Figure S-3B) instead, 40 mM lithium nitrate had a higher position than that of 80 mM lithium nitrate. Surprisingly, when the matrix PNA was added to the lipid mixture alone without the matrix additive (Figure S-3D), the spectrum was less complex as the protonated peak was easily detected.

Additives To Reduce the Complexity of Data. A number of research groups have used matrix additives for the analysis of a variety biological compounds. These additives include the following: ammonium acetate[40,43] and citrate[50] for analyzing phosphopeptides and proteins; lithium and cesium chlorides[51] for analyzing polymers; and tetraamine spermine[52] and polyamine[53] for the analysis of oligonucleotides. Interest in using additives has increased due to the quality of MALDI-TOF-MS spectra produced upon their addition.[40,42] Hence, this study included the addition of matrix additives to reduce the complexity of MALDI-TOF-MS spectra generated with some of the matrices used in lipid analysis.

The matrix additives used in this study contained different cations including Na^+, Li^+, K^+, and Ca^{2+}, which participated in the formation of adducts with the lipids. Lipid detection was affected significantly by the addition of some additives including sodium nitrate, sodium acetate, and diammonium citrate, more so than others such as EDTA ammonium.[29] The best conditions for the experiments showed that the use of matrix additives is not always necessary as all five lipid peaks were detected with the only difference being the intensity of the adducts. Moreover, there appeared to be no obvious effect on the spectra when using different concentrations of some of the matrix additives especially when using concentrations between 10 and 40 mM (Figure S-4).

On the other hand, these matrix additives were found to be useful in reducing the complexity of the data with some matrices, such as dithranol. Figure 3A and B show the spectra of the lipid mixture before and after the addition of sodium nitrate (10 mM), respectively. These spectra reiterated that the use of additives can indeed generate spectra with more useful information. Figure 3A shows that the spectrum generated from the analysis of the lipid mixture without the addition of an

DOI: 10.1021/acs.analchem.6b00512
Anal. Chem. 2016, 88, 6301−6308



× 10^5
5 lipids detected
B
C
E
F
G
Intensity
6
5
4
3
2
1
650
700
750
800
850
900
950
1000
m/z

Lipid	[M+H] ⁺	[M+ Na]⁺	[M + K]⁺
B	688 ▲	710 ▲	726 ▲
C	744 ⬮	766 ⬮	782 ⬮
E	801 ◆	823 ◆	- ◆
F	898 ■	920 ■	936 ■
G	937 ★	959 ★	975 ★

Figure 2. Typical and best MALDI-TOF-MS spectrum of a lipid mixture, detected *m*/*z* from 650 to 1000. Fourteen lipid peaks are highlighted in the spectrum showing the main lipids and lipid-adducts that were detected using the top conditions, where lipid (B) is PG, (C) is PE, (E) is PG, (F) is PC, and (G) is DGDG.

■ RESULTS AND DISCUSSION

Systematic Matrix Optimization. There has been wide interest in addressing questions related to the role of lipids and their biological function due to their importance in cells,[37] using modern analytical techniques such as MALDI-TOF-MS,[38] since MALDI-TOF-MS is a very powerful technique for lipid analysis. For this purpose a mixture of 5 lipids was selected due to their importance in bacterial cell membranes, and these included the following: 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) PG (lipid B), L-*α*-phosphatidylethanolamine dioleoyl PE (lipid C), 1,2-distearoyl-*sn*-glycero-3-phospho-(1′-*rac*-glycerol) (lipid E), 1,2-di(13Z-docosenoyl)-*sn*-glycero-3-phosphocholine PC (lipid F), and 1,2-diacyl-3-O-(*α*-D-galactosyl1-6)-*β*-D-galactosyl-*sn*-glycerol (lipid G). Our central focus was to optimize MALDI-TOF-MS for lipids using different factors: matrix, matrix preparation, matrix additive, additive concentration, and deposition method. Therefore, different experimental combinations resulted in different lipid preparations spotted directly onto the wells of the MALDI plates and MALDI-TOF-MS measurements taken, producing corresponding mass spectra.

MALDI-TOF-MS data are quite often challenging to interpret due to the complexity of the spectra acquired. The complexity of the data is due to the presence of different ion species formed from the lipid molecule including: $[M + H]^+$, $[M + Na]^+$, $[M + K]^+$, and $[M - H]^-$ (ref 8). Hence, a selection of eight different commonly used matrices, including THAP,[30,39] 2,5 DHB,[39,40] DHAP,[30,41] CMBT,[42] ATT,[43,44] HABA,[45] INN,[46] and PNA,[47] was used to facilitate the analysis of the lipid mixtures. This selection of matrices was based on a literature survey conducted to establish which matrices have been reported to work well for lipid analysis using MALDI-TOF-MS. Some of these matrices were shown to increase background noise and others to decrease it. These findings were also reported previously in the literature.[11,48]

Fractional Factorial Design Used To Identify Optimum Conditions for MALDI-TOF-MS. The significance of different factors (matrix, matrix preparation, matrix additive, additive concentration, and deposition method) was assessed using 18 peaks, which were directly assigned to the corresponding lipids in the mixture. Table S-4 shows the assignment of these peaks. These 18 peaks were extracted from the results of the 720 experimental protocols selected by FFD. Figure 1 illustrates the 3D scores plot of MALDI-TOF-MS data using multiple objectives measured for 720 × 3 = 2160 spectra generated in this study. The different characteristics (shapes and colors) shown in this 3D plot represent different matrices. For instance, the red and blue squares represent 6-aza-2-

DOI: 10.1021/acs.analchem.6b00512
Anal. Chem. 2016, 88, 6301−6308

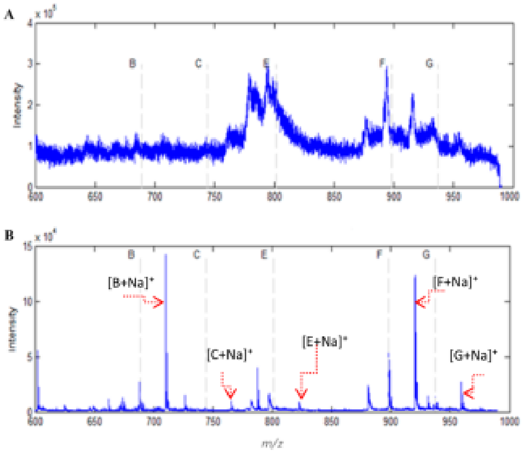


Figure 3. MALDI-TOF-MS spectra for the lipid mixture using dithranol as a matrix (A) without a matrix additive and (B) with 10 mM of sodium nitrate. The experiment resulted in an improved spectrum when sodium nitrate was added to the matrix solution.

additive was extremely complex, and the peaks were not detectable. However, Figure 3B shows that some of the lipid peaks were detectable such as sodiated peaks at 710, 766, 823, 920, and 959 *m/z* corresponding to lipids B, C, E, F, and G, respectively. By contrast, these peaks were undetectable in Figure 3A when the additive was not added.

Three different anions were the center points for lipid analysis when choosing the additives; these anions included the following: nitrates, acetates, and chlorides. We have observed that the addition of sodium nitrate/acetate led to an increase in the abundance of Na^+ adducts in comparison to K^+ adducts, and the protonated peaks were also detected (data not shown). Moreover, the addition of potassium nitrate led to an increase in the intensity of K^+ adducts relative to Na^+ adducts. In addition, when lithium nitrate was added to some of the matrix solutions, such as PNA, this led to reduction in spectral complexity and ease of the identification of some peaks. However, high concentrations of matrix additives resulted in poor spectra as no or few lipids were detected. Moreover, the addition of ammonium chloride to matrix solutions resulted sometimes in a decrease in abundance of Na^+ and K^+ adducts, in line with a previous study carried out by Griffiths et al.[39] In addition, ammonium adducts were not observed in this set of experiments. The use of calcium chloride resulted in poor spectral quality compared to other additives as Ca^{2+} adducts cannot be detected either in a singly and doubly charged state.

Table 1A shows different experiments, which consisted of a combination of different factors. The green boxes represent the combination of factors that allow the detection of lipid peaks. In contrast, the red boxes represent the experiments that combined factors, which failed to detect lipid peaks. For example ATT, DHB, and THAP are useful to analyze lipid samples with or without matrix additives, whereas some of the matrices such as CMPT and dithranol are not able to detect some of the lipid peaks without matrix additives. CMPT can detect some of the lipids in the presence of an additive such as 40 mM of sodium acetate; however, the spectra generated remain poor (Figure S-5). Referring to Table 1A it can be seen that some of the combinations enable the detection of lipids. However, in-depth analysis reveals that the intensities of each lipid vary from one experiment to another and these are shown in Table S-5. This table shows the individual and average combined FX ($f_{\text{multiobjective}}$) values for each objective for each

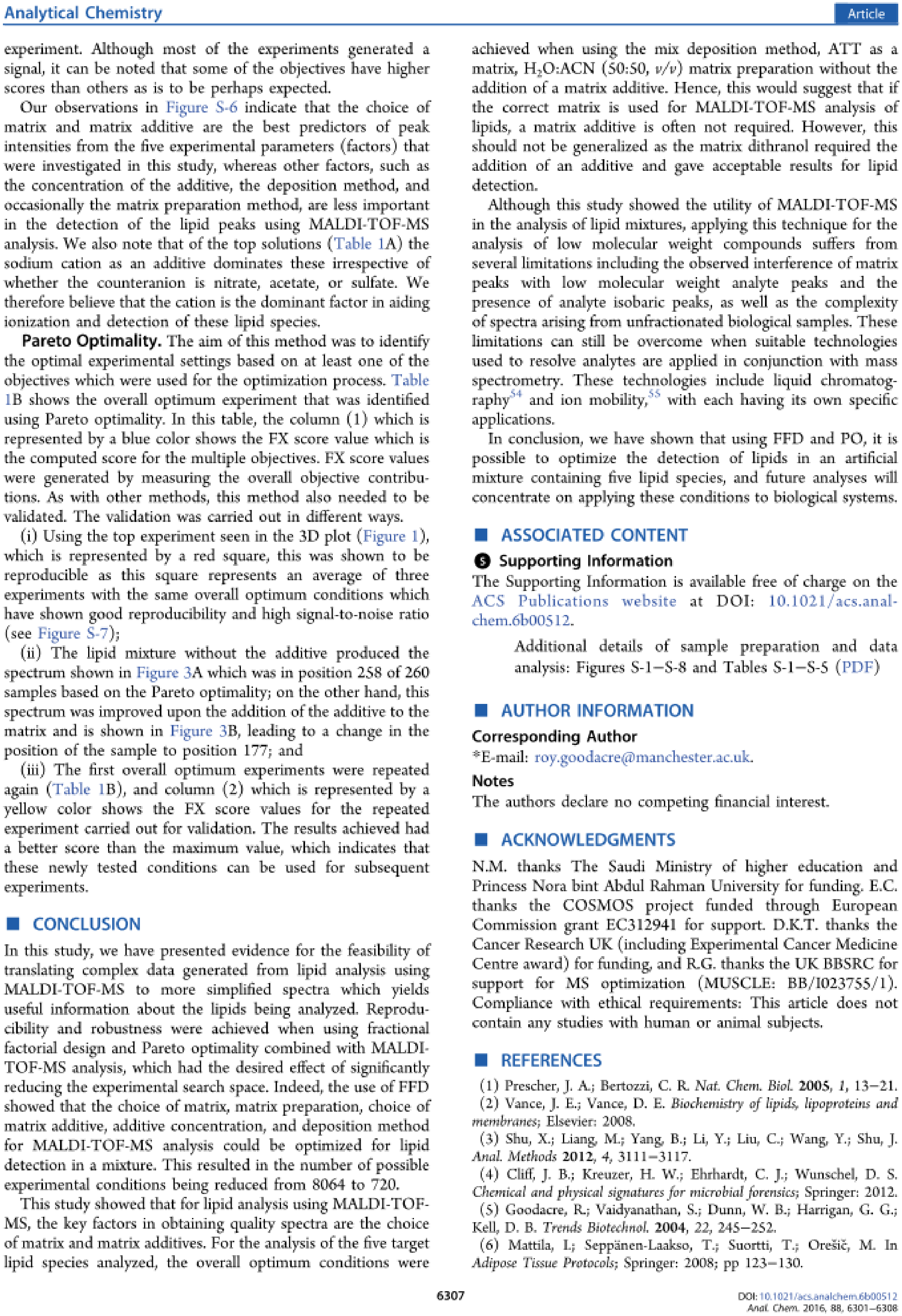
Table 1. (A) Examples of Experiments Carried out Using Different Factors[a] and (B) the First Overall Optimum Experiment That Was Selected Using Pareto Optimality[b]

A / Matrix	Matrix Prep.	Matrix Additive	Conc.	Matrix Dep.	B	C	E	F	G	FX value
ATT	H_2O/ACN	None	None	Mix	🟩	🟩	🟩	🟩	🟩	4.277
ATT	Chloroform/MeOH/H_2O	Sodium nitrate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.109
DHB	ACN/H_2O/isopropanol	Sodium nitrate	40 mM	Underlay	🟩	🟩	🟩	🟩	🟩	3.972
DHB	MeOH/H_2O	Sodium acetate	5 Mm	Underlay	🟩	🟩	🟩	🟩	🟩	3.852
DHB	EtOH/H_2O	Sodium nitrate	5 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.829
DHB	EtOH/H_2O	Sodium dodecyl sulfate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.787
DHB	ACN/H_2O/isopropanol	Sodium dodecyl sulfate	5 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.782
THAP	MeOH/H_2O	Sodium nitrate	10 mM	Mix	🟩	🟩	🟩	🟩	🟩	3.831
THAP	Chloroform/MeOH/H_2O	None	None	Mix	🟩	🟩	🟩	🟩	🟩	3.602
Dithranol	MeOH/H_2O	None	None	Overlay	🟥	🟥	🟥	🟥	🟥	1.029
Dithranol	MeOH/H_2O	Sodium nitrate	10 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.775
CMPT	H_2O/ACN	Sodium dodecyl sulfate	40 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.567
HABA	EtOH/H_2O	None	None	Mix	🟥	🟥	🟥	🟩	🟥	1.267
DHAP	H_2O/ACN	Potassium acetate	10 mM	Overlay	🟩	🟩	🟩	🟩	🟩	2.430

B / Matrix	Matrix pre.	Matrix additive	Matrix additive conc.	Matrix Dep.	(1) FX value	(2) FX value
ATT	(H_2O+ACN)	None	None	Mix	4.27	4.95

[a]The green boxes show the best combination of factors for lipid peak detections, whereas the red boxes show the combination of factors that failed to detect lipid peaks. [b]The (1) represented by the blue column shows the FX score value computed for the multiple objectives. The (2) represented by the yellow column shows the FX score value for the repeated experiment carried out for validation and is also the computed score for the multiple objectives.

DOI: 10.1021/acs.analchem.6b00512
Anal. Chem. 2016, 88, 6301−6308



experiment. Although most of the experiments generated a signal, it can be noted that some of the objectives have higher scores than others as is to be perhaps expected.

Our observations in Figure S-6 indicate that the choice of matrix and matrix additive are the best predictors of peak intensities from the five experimental parameters (factors) that were investigated in this study, whereas other factors, such as the concentration of the additive, the deposition method, and occasionally the matrix preparation method, are less important in the detection of the lipid peaks using MALDI-TOF-MS analysis. We also note that of the top solutions (Table 1A) the sodium cation as an additive dominates these irrespective of whether the counteranion is nitrate, acetate, or sulfate. We therefore believe that the cation is the dominant factor in aiding ionization and detection of these lipid species.

Pareto Optimality. The aim of this method was to identify the optimal experimental settings based on at least one of the objectives which were used for the optimization process. Table 1B shows the overall optimum experiment that was identified using Pareto optimality. In this table, the column (1) which is represented by a blue color shows the FX score value which is the computed score for the multiple objectives. FX score values were generated by measuring the overall objective contributions. As with other methods, this method also needed to be validated. The validation was carried out in different ways.

(i) Using the top experiment seen in the 3D plot (Figure 1), which is represented by a red square, this was shown to be reproducible as this square represents an average of three experiments with the same overall optimum conditions which have shown good reproducibility and high signal-to-noise ratio (see Figure S-7);

(ii) The lipid mixture without the additive produced the spectrum shown in Figure 3A which was in position 258 of 260 samples based on the Pareto optimality; on the other hand, this spectrum was improved upon the addition of the additive to the matrix and is shown in Figure 3B, leading to a change in the position of the sample to position 177; and

(iii) The first overall optimum experiments were repeated again (Table 1B), and column (2) which is represented by a yellow color shows the FX score values for the repeated experiment carried out for validation. The results achieved had a better score than the maximum value, which indicates that these newly tested conditions can be used for subsequent experiments.

■ CONCLUSION

In this study, we have presented evidence for the feasibility of translating complex data generated from lipid analysis using MALDI-TOF-MS to more simplified spectra which yields useful information about the lipids being analyzed. Reproducibility and robustness were achieved when using fractional factorial design and Pareto optimality combined with MALDI-TOF-MS analysis, which had the desired effect of significantly reducing the experimental search space. Indeed, the use of FFD showed that the choice of matrix, matrix preparation, choice of matrix additive, additive concentration, and deposition method for MALDI-TOF-MS analysis could be optimized for lipid detection in a mixture. This resulted in the number of possible experimental conditions being reduced from 8064 to 720.

This study showed that for lipid analysis using MALDI-TOF-MS, the key factors in obtaining quality spectra are the choice of matrix and matrix additives. For the analysis of the five target lipid species analyzed, the overall optimum conditions were achieved when using the mix deposition method, ATT as a matrix, H_2O:ACN (50:50, *v*/*v*) matrix preparation without the addition of a matrix additive. Hence, this would suggest that if the correct matrix is used for MALDI-TOF-MS analysis of lipids, a matrix additive is often not required. However, this should not be generalized as the matrix dithranol required the addition of an additive and gave acceptable results for lipid detection.

Although this study showed the utility of MALDI-TOF-MS in the analysis of lipid mixtures, applying this technique for the analysis of low molecular weight compounds suffers from several limitations including the observed interference of matrix peaks with low molecular weight analyte peaks and the presence of analyte isobaric peaks, as well as the complexity of spectra arising from unfractionated biological samples. These limitations can still be overcome when suitable technologies used to resolve analytes are applied in conjunction with mass spectrometry. These technologies include liquid chromatography[54] and ion mobility,[55] with each having its own specific applications.

In conclusion, we have shown that using FFD and PO, it is possible to optimize the detection of lipids in an artificial mixture containing five lipid species, and future analyses will concentrate on applying these conditions to biological systems.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b00512.

Additional details of sample preparation and data analysis: Figures S-1−S-8 and Tables S-1−S-5 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: roy.goodacre@manchester.ac.uk.

Notes

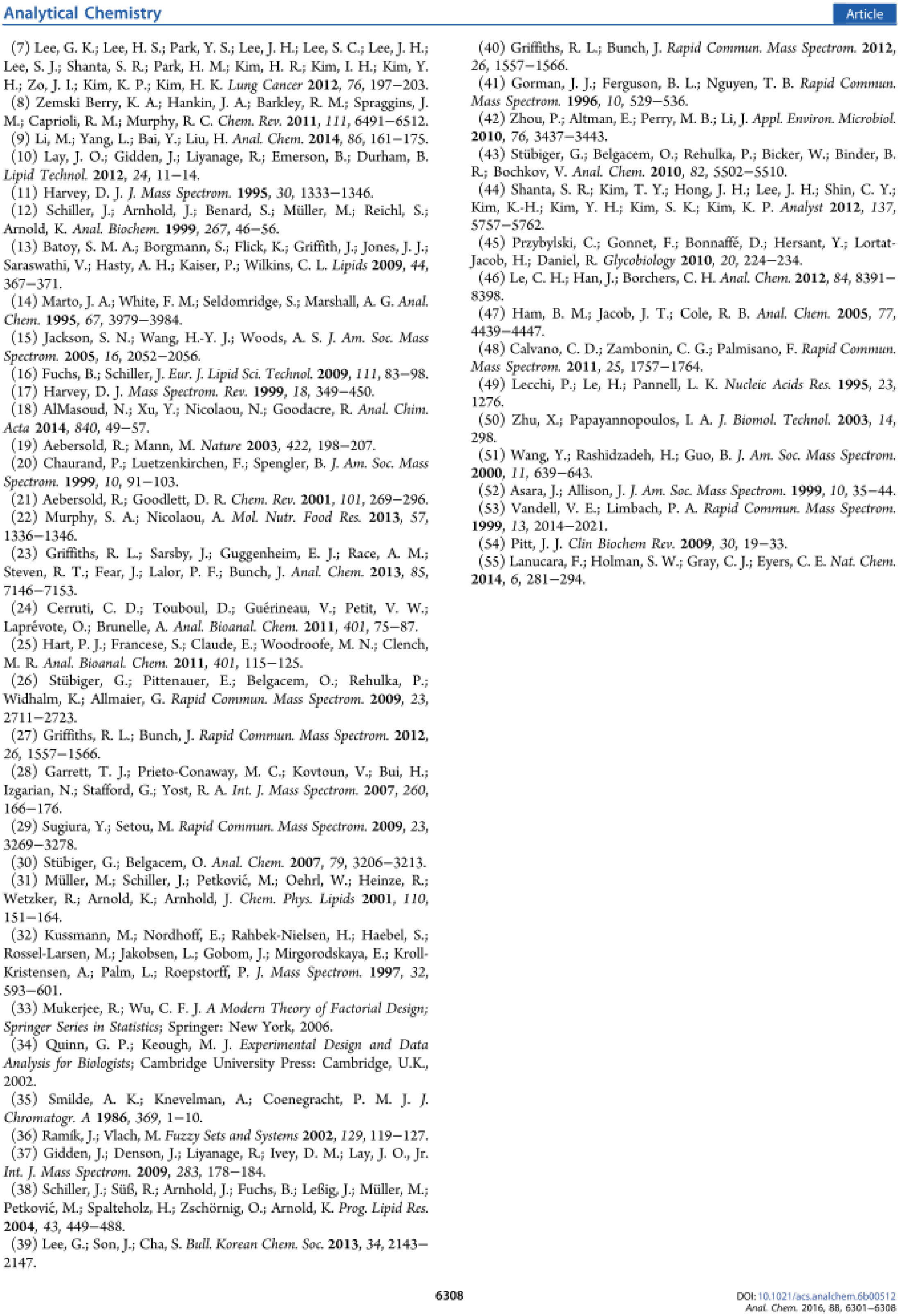
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

N.M. thanks The Saudi Ministry of higher education and Princess Nora bint Abdul Rahman University for funding. E.C. thanks the COSMOS project funded through European Commission grant EC312941 for support. D.K.T. thanks the Cancer Research UK (including Experimental Cancer Medicine Centre award) for funding, and R.G. thanks the UK BBSRC for support for MS optimization (MUSCLE: BB/I023755/1). Compliance with ethical requirements: This article does not contain any studies with human or animal subjects.

■ REFERENCES

(1) Prescher, J. A.; Bertozzi, C. R. *Nat. Chem. Biol.* **2005**, *1*, 13−21.
(2) Vance, J. E.; Vance, D. E. *Biochemistry of lipids, lipoproteins and membranes*; Elsevier: 2008.
(3) Shu, X.; Liang, M.; Yang, B.; Li, Y.; Liu, C.; Wang, Y.; Shu, J. *Anal. Methods* **2012**, *4*, 3111−3117.
(4) Cliff, J. B.; Kreuzer, H. W.; Ehrhardt, C. J.; Wunschel, D. S. *Chemical and physical signatures for microbial forensics*; Springer: 2012.
(5) Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. *Trends Biotechnol.* **2004**, *22*, 245−252.
(6) Mattila, I.; Seppänen-Laakso, T.; Suortti, T.; Orešič, M. In *Adipose Tissue Protocols*; Springer: 2008; pp 123−130.



(7) Lee, G. K.; Lee, H. S.; Park, Y. S.; Lee, J. H.; Lee, S. C.; Lee, J. H.; Lee, S. J.; Shanta, S. R.; Park, H. M.; Kim, H. R.; Kim, I. H.; Kim, Y. H.; Zo, J. I.; Kim, K. P.; Kim, H. K. *Lung Cancer* **2012**, *76*, 197−203.

(8) Zemski Berry, K. A.; Hankin, J. A.; Barkley, R. M.; Spraggins, J. M.; Caprioli, R. M.; Murphy, R. C. *Chem. Rev.* **2011**, *111*, 6491−6512.

(9) Li, M.; Yang, L.; Bai, Y.; Liu, H. *Anal. Chem.* **2014**, *86*, 161−175.

(10) Lay, J. O.; Gidden, J.; Liyanage, R.; Emerson, B.; Durham, B. *Lipid Technol.* **2012**, *24*, 11−14.

(11) Harvey, D. J. *J. Mass Spectrom.* **1995**, *30*, 1333−1346.

(12) Schiller, J.; Arnhold, J.; Benard, S.; Müller, M.; Reichl, S.; Arnold, K. *Anal. Biochem.* **1999**, *267*, 46−56.

(13) Batoy, S. M. A.; Borgmann, S.; Flick, K.; Griffith, J.; Jones, J. J.; Saraswathi, V.; Hasty, A. H.; Kaiser, P.; Wilkins, C. L. *Lipids* **2009**, *44*, 367−371.

(14) Marto, J. A.; White, F. M.; Seldomridge, S.; Marshall, A. G. *Anal. Chem.* **1995**, *67*, 3979−3984.

(15) Jackson, S. N.; Wang, H.-Y. J.; Woods, A. S. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2052−2056.

(16) Fuchs, B.; Schiller, J. *Eur. J. Lipid Sci. Technol.* **2009**, *111*, 83−98.

(17) Harvey, D. J. *Mass Spectrom. Rev.* **1999**, *18*, 349−450.

(18) AlMasoud, N.; Xu, Y.; Nicolaou, N.; Goodacre, R. *Anal. Chim. Acta* **2014**, *840*, 49−57.

(19) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.

(20) Chaurand, P.; Luetzenkirchen, F.; Spengler, B. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 91−103.

(21) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−296.

(22) Murphy, S. A.; Nicolaou, A. *Mol. Nutr. Food Res.* **2013**, *57*, 1336−1346.

(23) Griffiths, R. L.; Sarsby, J.; Guggenheim, E. J.; Race, A. M.; Steven, R. T.; Fear, J.; Lalor, P. F.; Bunch, J. *Anal. Chem.* **2013**, *85*, 7146−7153.

(24) Cerruti, C. D.; Touboul, D.; Guérineau, V.; Petit, V. W.; Laprévote, O.; Brunelle, A. *Anal. Bioanal. Chem.* **2011**, *401*, 75−87.

(25) Hart, P. J.; Francese, S.; Claude, E.; Woodroofe, M. N.; Clench, M. R. *Anal. Bioanal. Chem.* **2011**, *401*, 115−125.

(26) Stübiger, G.; Pittenauer, E.; Belgacem, O.; Rehulka, P.; Widhalm, K.; Allmaier, G. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 2711−2723.

(27) Griffiths, R. L.; Bunch, J. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1557−1566.

(28) Garrett, T. J.; Prieto-Conaway, M. C.; Kovtoun, V.; Bui, H.; Izgarian, N.; Stafford, G.; Yost, R. A. *Int. J. Mass Spectrom.* **2007**, *260*, 166−176.

(29) Sugiura, Y.; Setou, M. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 3269−3278.

(30) Stübiger, G.; Belgacem, O. *Anal. Chem.* **2007**, *79*, 3206−3213.

(31) Müller, M.; Schiller, J.; Petković, M.; Oehrl, W.; Heinze, R.; Wetzker, R.; Arnold, K.; Arnhold, J. *Chem. Phys. Lipids* **2001**, *110*, 151−164.

(32) Kussmann, M.; Nordhoff, E.; Rahbek-Nielsen, H.; Haebel, S.; Rossel-Larsen, M.; Jakobsen, L.; Gobom, J.; Mirgorodskaya, E.; Kroll-Kristensen, A.; Palm, L.; Roepstorff, P. *J. Mass Spectrom.* **1997**, *32*, 593−601.

(33) Mukerjee, R.; Wu, C. F. J. *A Modern Theory of Factorial Design; Springer Series in Statistics*; Springer: New York, 2006.

(34) Quinn, G. P.; Keough, M. J. *Experimental Design and Data Analysis for Biologists*; Cambridge University Press: Cambridge, U.K., 2002.

(35) Smilde, A. K.; Knevelman, A.; Coenegracht, P. M. J. *J. Chromatogr. A* **1986**, *369*, 1−10.

(36) Ramík, J.; Vlach, M. *Fuzzy Sets and Systems* **2002**, *129*, 119−127.

(37) Gidden, J.; Denson, J.; Liyanage, R.; Ivey, D. M.; Lay, J. O., Jr. *Int. J. Mass Spectrom.* **2009**, *283*, 178−184.

(38) Schiller, J.; Süß, R.; Arnhold, J.; Fuchs, B.; Leßig, J.; Müller, M.; Petković, M.; Spalteholz, H.; Zschörnig, O.; Arnold, K. *Prog. Lipid Res.* **2004**, *43*, 449−488.

(39) Lee, G.; Son, J.; Cha, S. *Bull. Korean Chem. Soc.* **2013**, *34*, 2143−2147.

(40) Griffiths, R. L.; Bunch, J. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1557−1566.

(41) Gorman, J. J.; Ferguson, B. L.; Nguyen, T. B. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 529−536.

(42) Zhou, P.; Altman, E.; Perry, M. B.; Li, J. *Appl. Environ. Microbiol.* **2010**, *76*, 3437−3443.

(43) Stübiger, G.; Belgacem, O.; Rehulka, P.; Bicker, W.; Binder, B. R.; Bochkov, V. *Anal. Chem.* **2010**, *82*, 5502−5510.

(44) Shanta, S. R.; Kim, T. Y.; Hong, J. H.; Lee, J. H.; Shin, C. Y.; Kim, K.-H.; Kim, Y. H.; Kim, S. K.; Kim, K. P. *Analyst* **2012**, *137*, 5757−5762.

(45) Przybylski, C.; Gonnet, F.; Bonnaffé, D.; Hersant, Y.; Lortat-Jacob, H.; Daniel, R. *Glycobiology* **2010**, *20*, 224−234.

(46) Le, C. H.; Han, J.; Borchers, C. H. *Anal. Chem.* **2012**, *84*, 8391−8398.

(47) Ham, B. M.; Jacob, J. T.; Cole, R. B. *Anal. Chem.* **2005**, *77*, 4439−4447.

(48) Calvano, C. D.; Zambonin, C. G.; Palmisano, F. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1757−1764.

(49) Lecchi, P.; Le, H.; Pannell, L. K. *Nucleic Acids Res.* **1995**, *23*, 1276.

(50) Zhu, X.; Papayannopoulos, I. A. *J. Biomol. Technol.* **2003**, *14*, 298.

(51) Wang, Y.; Rashidzadeh, H.; Guo, B. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 639−643.

(52) Asara, J.; Allison, J. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 35−44.

(53) Vandell, V. E.; Limbach, P. A. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 2014−2021.

(54) Pitt, J. J. *Clin Biochem Rev.* **2009**, *30*, 19−33.

(55) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. *Nat. Chem.* **2014**, *6*, 281−294.

DOI: 10.1021/acs.analchem.6b00512
Anal. Chem. 2016, 88, 6301−6308

Other publications by the author

In addition to the work undertaken in this thesis, the author also contributed to the following publications:

- Brown, A.R., Correa, E., Xu, Y., AlMasoud, N., Pimblott, S.M., Goodacre, R. and Lloyd, J.R. 2015. Phenotypic characterisation of *Shewanella oneidensis* MR-1 exposed to X-radiation. *PLoS ONE,* **10**, e0131249.

- Muhamadali, H., Weaver, D., Subaihi, A., AlMasoud, N., Trivedi, D.K., Ellis, D.I., Linton, D. and Goodacre, R. 2016. Chicken, beams, and *Campylobacter*: rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry. *Analyst.* **141**, 111-122.

- Sayqal, A., Xu, Y., Trivedi, D., AlMasoud, N., Ellis, D.I. and Goodacre, R. 2016. Metabolic fingerprinting of *Pseudomonas putida* DOT-T1E strains: understanding the influence of divalent cations in adaptation mechanisms following exposure to toluene. *Metabolites* **6**, 14

- Sayqal, A., Xu, Y., Trivedi, D.K., AlMasoud, N., Ellis, D.I., Rattray, N.J.W. and Goodacre, R. 2016. Metabolomics analysis reveals the participation of efflux pumps and ornithine in the response of *Pseudomonas putida* DOT-T1E cells to challenge with propranolol. *PLoS ONE* **11**: e0156509.

- Sayqal, A., Xu, Y., Trivedi, D.K., AlMasoud, N., Ellis, D.I., Muhamadali, H., Rattray, N.J.W., Webb, C. & Goodacre, R. (2016) Metabolic analysis of the response of*Pseudomonas putida* DOT-T1E strains to toluene using Fourier transform infrared spectroscopy and gas chromatography mass spectrometry. *Metabolomics* **12**: 112.

Awards and recognitions

In addition to the work undertaken in this thesis, the author also participated in different conferences and received the following recognitions:

1- Award for one of the best PhD final year talks at the school of chemistry, University of Manchester, UK.

2- Poster prize in the Metabomeeting 2015, Cambridge, UK.