Universidade de Aveiro Departamento de Química

2014

Bruno Manuel Santos Saraiva

Identificação de Bactérias por Espectrometria de Massa

Identification of Bacteria by Mass Spectrometry

Universidade de Aveiro Departamento de Química

2014

Bruno Manuel Santos Saraiva

Identificação de Bactérias por Espectrometria de Massa

Identification of Bacteria by Mass Spectrometry

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioquímica, mestrado em Bioquímica, ramo de Métodos Biomoleculares, realizada sob a orientação científica do Doutor Rui Vitorino, Investigador Auxiliar do Departamento de Química da Universidade de Aveiro e do Doutor António Barros, Investigador Auxiliar do Departamento de Química da Universidade de Aveiro

O Júri

presidente **Prof. Doutor Pedro Miguel Dimas Neves Domingues**

Professor Auxiliar com Agregação ao Departamento de Química da Universidade de Aveiro

Prof. Doutora Margarida Sâncio da Cruz Fardilha

Professora Auxiliar da Secção Autónoma das Ciências da Saúde

da Universidade de Aveiro

Doutor António de Sousa Barros

Investigador Auxiliar do Departamento de Química da Universidade de Aveiro

Agradecimentos

Gostaria de agradecer aos meus orientadores, Prof. Doutor Rui Vitorino e Doutor António Barros pela disponibilidade, orientação e conhecimentos cedidos, imprescindíveis à realização deste trabalho e pela sua contribuição para o meu desenvolvimento académico e pessoal.

Aos meus colegas de laboratório, pelo excelente ambiente de trabalho, por toda a disponibilidade para ajudar e pelos bons momentos passados.

À Tânia, por todo o apoio incondicional e por estar sempre presente.

Aos meus amigos e colegas de curso, em especial ao Bruno Valente pela boa disposição, companhia nos momentos de trabalho e por todo o apoio ao longo destes anos.

Por fim, um agradecimento muito especial aos meus pais, pois sem eles nada disto teria sido possível e à minha família pelo carinho e apoio sempre demonstrados.

Palavras-chave

Identificação de Bactérias; Espectrometria de Massa; MALDI-TOF; Algoritmos para identificação de Bactérias

Resumo

A identificação de bactérias é um dos principais componentes do diagnóstico de infeções patogénicas. De modo a se conseguir obter uma identificação podem ser aplicadas diferentes técnicas, tais como, diferenças de fenótipo, comparação de sequências de ADN e a comparação do conteúdo proteico das bactérias. Quando se compara a identificação bacteriana com recurso à espectrometria de massa MALDI-TOF com as metodologias alternativas, podemos destacar diversas vantagens: menor custo por análise, menos tempo para obtenção de resultados e um maior poder discriminatório.

Este trabalho tem como foco o desenvolvimento de uma nova aplicação capaz de identificar bactérias com recurso a espectrometria de massa. O trabalho foi iniciado com a extração proteica das amostras e a aquisição dos perfis de massa dessas bactérias. De seguida, prosseguiu-se com o desenvolvimento de uma aplicação para a identificação de bactérias com base na comparação dos perfis de massa da amostra e dos perfis contidos na base de dados.

Usando a aplicação desenvolvida conseguiu-se identificar corretamente tanto bactérias Gram-positivas como Gram-negativas. Quando uma identificação da estirpe não foi possível, a aplicação permitiu a identificação da espécie bacteriana e no caso de a base de dados não conter nenhuma entrada que correspondesse a estirpe ou espécie da amostra, os *scores* obtidos eram suficientemente baixos para uma eventual identificação ser descartada, resultando assim numa baixa taxa de falsos positivos.

Keywords Bacteria Identification; Mass Spectrometry; MALDI-TOF; Algorithms for Bacteria Identification;

Abstract

Identification of bacteria is a major part of the diagnosis of a pathogenic infection. In order to positively and confidently identify the bacteria, different techniques can be applied. These techniques are based on different principles, such as phenotypic differences, DNA sequences comparison, mass spectrometry (MALDI-TOF) and the protein content of the bacteria. From comparison of MALDI-TOF for bacteria identification with the other methodologies available, several advantages can be highlighted: lower cost *per* identification, faster results and higher discrimination power.

This work focus on the development of a new application capable of identifying bacteria using MALDI-TOF mass spectrometry. It started by the protein extraction of the samples and the acquisition of the mass profiles of those bacteria. This work proceeded with the development of an application to identify bacteria by comparing the mass profile of an unknown sample with the mass profiles of the bacteria in our database.

Using the developed application it was possible to identify both Gram-positive and Gram-negative bacteria to the strain-level. When an identification to strain-level was not possible, it was possible to identify the bacteria to the species-level and in the case the database did not contain an entry of the same species as the sample, the score values were low enough to disregard a possible identification, resulting in a low false-positive rate.

Index

Figures Index

Table Index

Abbreviations

α-Cyano – α-Cyano Hydroxicinnamic Acid

ACN – Acetonitrile

AOM – Acute Otitis Media

BLAST – Basic Local Alignment Tool

BSA – Bovine Serum Albumin

CFU – Colony Forming Unit

DHB – 2,5-Dihydroxibenzoic Acid

DNA – Deoxyribonucleic Acid

ESBL – Extended-spectrum Beta-lactamases

ESI-MS – Electrospray Ionization Mass Spectrometry

HPLC – High-pressure Liquid Chromatography

IR – Infra-Red

Lab. ID – Laboratory Identification

MALDI-TOF – Matrix-assisted Laser Desorption/Ionization – Time-of-flight

MDR – Multidrug Resistance

MS – Mass Spectrometry

PC – Principal Component

PCA – Principal Components Analysis

PCR – Polymerase Chain Reaction

RDP-II – Ribosomal Database Project

RNA – Ribonucleic Acid.

rRNA – ribosomal Ribonucleic Acid

TFA – Trifluoroacetic Acid

T-RFLP – Terminal Restriction Fragment Length Polymorphism

1. Introduction

Bacterial infections are a major concern in clinical environments. Diseases caused by bacterial infection require quick diagnosis for the appropriate management. For example, it is estimated that three quarters of the children will have an episode of Acute otitis media (AOM), an infection of the middle ear space which can be caused by different bacteria (1). Some of the bacteria associated to AOM are *Streptococcus pneumoniae, Haemophilus influenza* and *Moraxela Catarrhalis* (2) and depending on the type of bacteria, different treatments may be required. Moreover, other complications may emerge when it comes to bacterial infections: different strains of the bacteria do have different antibiotic resistance and hence the use of an incorrect treatment can stimulate the bacteria strains to develop even further resistance. Furthermore, the overall burden associated with bacterial infections, especially with antibiotic-resistant infections is very high. For instance, in USA the estimated annual cost of the treatment for these infections is around $16 billion dollars, based on the year 2000 reports, whereas in the European territory around 1,5 billion euros, based on 2007 reports (3).

In a clinical environment is sought for each single patient an accurate identification of bacteria causing the infection, namely, a pathogenic or simply a colonizer; to conceive the correct treatment, its extent and finally, the appropriate approach to seek, mitigate and/or eliminate the infection (4). The main requirements for a clinical bacteria identification system are reliability, the ability to differentiate between closely related species, the time needed to obtain a positive identification and the cost of the method (5). Those approaches are based in different traits such as phenotypic differences, genotyping and the content of the cell, among others (6). There is also a great interest in the development of new techniques that are more cost-effective, more reliable and less time-consuming. For that purpose and to ensure the correct diagnosis different techniques can be applied.

1.1 Methodologies for Bacteria Identification

In order to undertake a correct bacteria identification different methodologies have been used throughout the years, due to several factors some of them are more common on clinical environments than others. Some of the factors that are needed to consider when evaluating the usefulness of the different methods are, for example, the

accuracy, the false positive/negative rates and the cost *per* analysis. The time consuming required for each method is of concern, since the swiftness of the identification can have a major impact on clinical management (7). Different methods have been suggested throughout the years, with one of the early ones being the approach suggested by Abel *et al*. (8), which made use of gas chromatography (8). This study dates back to 1963 and focused on the lipid composition of the microorganisms. The results obtained showed that it was possible to identify the microorganisms using gas chromatography and they suggested that proteins or amino acids could also be used. The bacteria used for their study belongs to the *Schizomycetes* class, whereas they successfully obtained different lipid profiles for the different species, which suggests that this type of analysis could possible identify different bacteria (8). This method used a comparative algorithm to match the unknown organism methyl-ester profile with profiles on a database (7). However, this method was only accurate to the class level.

1.1.1 Phenotypic and Biochemical Methods

Traditionally, the most common methods used in clinical environments are based on the phenotype of the analysed bacteria and the monitoring of biochemical reactions (4). The classical phenotypic approach comprises data of morphological, physiological and biochemical features of different bacteria (9). The use of just one of these features had showed to be insufficient to produce a suitable identification. However the combination of all those features showed an increase in the identification reliability (9). The morphological features are based on the shape and size of the cell, Gram staining, among others, as well as colour, dimensions and form of the colony (9). Physiological and biochemical features often analysed are the growth of the culture on different temperatures, pH levels, salt concentrations or atmospheric conditions, growth in the presence of a set of substances (for example antimicrobial agents and metabolization of certain compounds) (9). One of the key problems using this approach is the reproducibility within and between different laboratories (9).

The analysis of biochemical processes is also commonly used and one of the suggested procedures available at Biolog, Inc. This system is based on the oxidation of 95 substrates on a 96-well plate, containing a redox dye, tetrazolium violet, that allows colorimetric determination of the increased respiratory processes when the bacteria are

consuming a carbon source (10). The bacterial samples are incubated in plates, being the results obtained after two different periods, 4h and 24h (10); an example of a plate after culture is seen on figure 1.

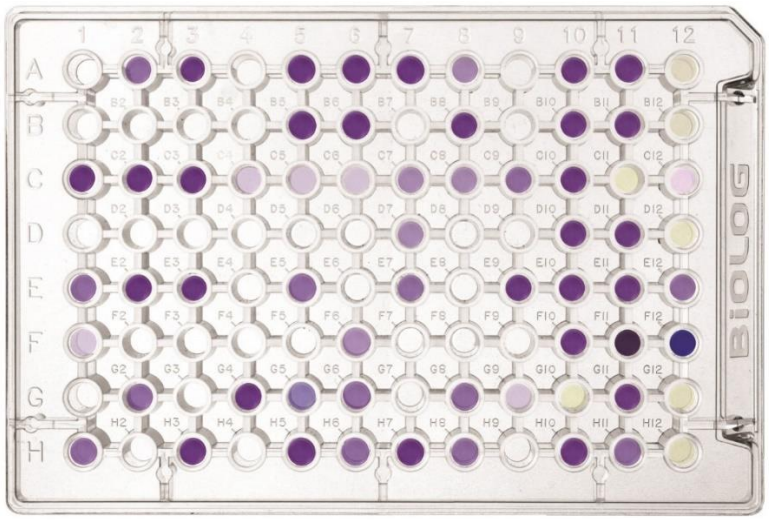


Figure 1 - A plate of the Biolog System after culture with a bacterial isolate. The wells contain bacteria cultures of the unknown isolate, each of them containing different substrates and a redox dye. The colour of each well correspond to the ability of the isolate to metabolize the substrate contained in that well. The identification is proposed by comparing the pattern of the plate with a reference database (11).

The Biolog system was evaluated by different study groups. Holmes *et al.* (10) tested the system in the identification of Gram-negative bacteria with clinical significance. Their results showed a correct identification to the genus level of 67%, with 93% of these organisms being also identified to the correct species level (10). Klinger *et al.* (12) also did a study to evaluate the performance of the Vitek and Biolog systems; the latter being another method based on biochemical tests. Both studies detected some errors which resulted in very low correct identification rates. An important aspect was the fact that reliable results were obtained with the Biolog system after 4h incubation when compared to 24h (10). Despite some of the positive aspects of using this type of techniques, they show major limitations such as: i) some organisms do not match to the pattern of biochemical reactions used in this type of analysis; ii) it cannot be used without

prior culture of the microorganisms, which impairs the identification of isolates that are hard to culture; iii) and the identification of some types of bacteria, such as anaerobes or mycobacteria, require additional equipment and expertise (4). To mitigate these problems different techniques have been used such as those based on the analysis of the bacterial genome. The identification of uncultivated bacteria usually relies on direct microscopy or immunologic assays (4). However, these two techniques also have serious limitations such as, direct microscopy needs a good number of observable cells; and immunologic assays may suffer from cross-reactivity and may be affected by the patient immunocompromised status (4).

1.1.2 Genomic Approaches

Currently the most common identification approaches relies on genome analysis, instead of phenotypic analysis. One of the mostly used targets in this genomic approach is the analysis of small subunit ribosomal RNA (16S rRNA) (13). The sequencing of the 16S subunit usually gives faster results when compared to the phenotypic methods, mainly when dealing with slow-growing bacteria (4). Furthermore, the sequencing of the 16S subunit is not affected by the presence or absence of housekeeping genes or by variability in the expression of some traits, which are some of the limitations of the phenotypic identification (14). The rRNA genes are transcribed from the ribosomal 30S operon and are later cleaved in 16S, 23S and 5S RNA molecules by RNase III (13). Since the 16S rRNA sequence is the most conserved sequence among the same species, it is a good target for genotyping (13). The usual workflow of this technique entails the extraction of the genetic material, amplification of the 16S rRNA gene, sequencing of the amplification product and comparison of the sequence to a reference database (figure 2). The 16S RNA sequences have been used for two different type of studies: identification and classification of isolated cultures of bacteria and the assessment of bacteria diversity in environmental samples (13). The results produced by these approaches are based on comparison of sample sequence and databases using comparative tools, such as BLAST (Basic Local Alignment Search Tool), an algorithm used to compare primary sequences, of amino-acids or nucleotides (13). However, the use of these databases is limited since there are no threshold values universally accepted from which one can obtain a correct identification. Moreover, the difference between the closest match-up and the next one can be of <0.5% score (13,14). Using the 16S rRNA sequencing for bacteria

identification, has shown over 90% correct identifications to the genus level and 65% to 83% correct identification to the species level (14). In all studies of bacteria identification using 16S rRNA sequencing none showed a match with a similarity over 99% (14); in fact even if one uses this threshold the identification may not be correct, due to the similarity of the 16S rRNA sequences within some strains of the same species (14). Additional problems arise from the use of this type of analysis. For instance, considering the *Aeromonas veronii* one can observe that it may contain up to 6 different 16S rRNA sequences which differ among themselves up to 1.5%, thus the intragenomic heterogeneity of the 16S rRNA gene between the aeromonads precludes the use of this technique as the single technique to achieve a correct identification (14). Despite some problems when using this approach in the identification of bacteria, the use of microarrays may provide a much more sensitive approach to the molecular species identification (14).

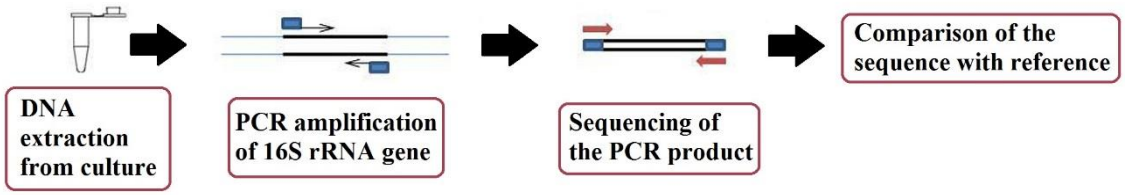


Figure 2 - Schematic representation of the workflow for bacteria identification using 16S rRNA sequencing.

The analysis of the 16S rRNA sequencing can be done using software packages, being MicroSeq and Ribosomal Database Project (RDP-II) the most widely used (4). Different studies evaluate the reliability of these software packages, as described in table 1. When analysing the results obtained by these studies it is not possible to compare the accuracy of both software packages as they were not done using the same conditions; nevertheless the results give an insight of the performance of these software packages.

Table 1 - Results of different studies regarding the accuracy of the MicroSeq and RDP-II software packages.

References	Bacteria isolate tested	Methods for determining of the isolates	Accuracy of the software
Tang *et al.*(7)	65 unusual aerobic Gram-negative bacilli from clinical specimens	Conventional phenotypic methods	MicroSeq: 89% to the species level and 100% to the genus level
Tang *et al.*(15)	52 coryneform Gram-positive bacilli	Conventional phenotypic methods	Microseq: 67% to the species level and 100% to the genus level
Patel *et al.*(16)	113 *Mycobacterium* isolates (18 different species)	Combination of phenotypical methods and biochemical assays	MicroSeq: 92.8%
Turenne *et al.*(17)	79 non-tuberculous mycobacterial isolates (ATCC type strains)	Not described	RDP-II: 41%

A different genomic approach to the bacteria identification is the analysis of the terminal restriction fragment length polymorphisms (T-RFLP). This technique uses one or both PCR primers, labelled with fluorescent dyes, to amplify the 16S rRNA sequence of the isolates. The resulting amplified sequences are then cleaved with restriction enzymes, resulting in fragments of the amplified sequence (18,19). After the cleavage, salts and primers are removed and a polyacrylamide gel electrophoresis is performed (18). The resulting gel will present a specific pattern based on the size of the fragments, which is caused by the differences in the 16S rRNA sequences of different species (13). The band pattern of the isolate is then compared to the reference band patterns of known bacteria and in the case of a positive match between the isolate pattern and references, one identification can be proposed. However, the mobility of the samples is affected by the used fluorescent dye and therefore comparisons should only be made between samples that used the same dye (20).

1.1.3 IR Spectroscopy

The infrared (IR) spectroscopy started to be used in the study of conformational structure of peptides in the 1950's and, since then, different applications on have been highlighted (5,21). Nevertheless, in the last two decades many studies were done using this approach for identification and characterization of bacteria (21). Basically, an IR spectrum is obtained by measuring the intensity of IR radiation before and after passing through a sample (21). Then, the identification of a bacterial isolate is accomplished by comparing the spectrum with spectra of a reference database (figure 3) (21).

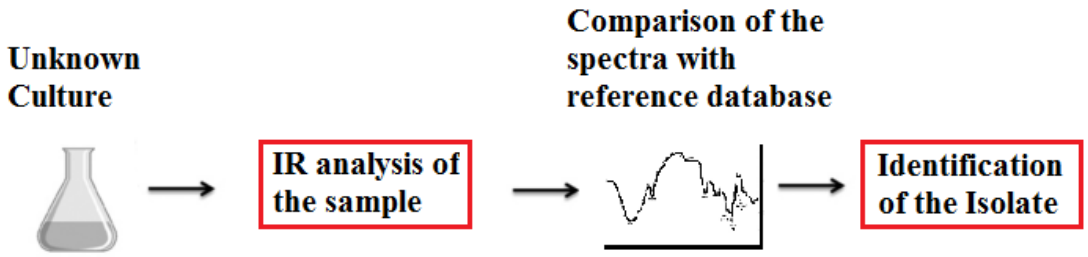


Figure 3 - Example workflow for identification of a bacterial isolate using IR spectroscopy. The spectra is obtained from the unknown sample and is then compared to a reference database. The most similar database entry should correspond to the correct identification.

Following this, Kirschner *et al.* (22) identified different strains of *Enterococci* isolated from urine and food. Using the IR they were able to obtain spectra for every sample, which was enough discriminatory to distinguish different strains and identify bacteria isolates (22). Nevertheless, identification of some strains obtained by the use of IR spectroscopy and phenotypic methods were not always in accordance, being in these cases the identification relied in the 16S rRNA sequencing (22). Comparison of results obtained with IR spectroscopy within those obtained with 16S rRNA, showed that the IR seems to be a very reliable technique for the identification of bacteria (22). Additionally, repetitive measures were performed over six months period in order to evaluate the reproducibility of the data obtained from IR spectroscopy, which highlight a consistency (22). A different approach is the use of IR microspectroscopy, which is the result of combining an IR spectrometer and a microscope, allowing the analysis of just a few

hundred cells (5). A study using this approach for the identification of bacteria collected more than 1570 spectra of bacterial colonies with sizes ranging from 30-150 μm, after 6h, 7h and 10h of culture (5). Their results showed that IR spectroscopy became more reliable with extended culture time due to the accumulation of products of biochemical reactions in the cells. However in the case of microspectroscopy resolution was lost with longer culture duration (5). It was obtained an accuracy of 100% for Gram-positive bacteria identification at the species level for *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Enterococcus faecalis*, and *Enterococcus faecium* (5). In the case of Gram-negative bacteria it was obtained an accuracy of 80% for bacteria from *Enterobacteriaceae* and *Pseudomonaceae* families (5). Their results were obtained comparing the analysis of IR spectroscopy spectra of bacteria colonies with 18h of culture (5). While these results show great promise, more tests with higher number of samples are required to obtain the proper conclusions. Still the rate of correct identifications obtained for Gram-negative bacteria were considerable lower when compared to what is sought for a clinical environment.

Aiming faster more reliable and cheaper analysis new techniques have been developed. Some of these new techniques focus on analysis of the protein content of cell to achieve a correct bacteria identification.

1.1.4 Analysis of the protein content of the cell

Proteomics have also been used in studies related to microbiology, either in the study of microbial pathogens, study of biomarkers or bacteria identification (23,24). Using different methodologies available for proteomic studies it is possible to identify and quantify the proteins in a sample. With the possibility of identification and quantification of protein content from bacteria, a deeper insight is envisioned about these organisms (25). While most of the protein analysis of bacteria are done with the objective of characterization, it is also possible to identify bacteria based on that type of analysis.

1.1.4.1 Protein Microarrays

Different methodologies can be applied to achieve a bacteria identification. One of the targets of these methodologies is the protein content of the cell, as used by Protein

Microarrays. This technique can be described as a miniaturization of thousands of assays on one small plate (figure 4) (26).

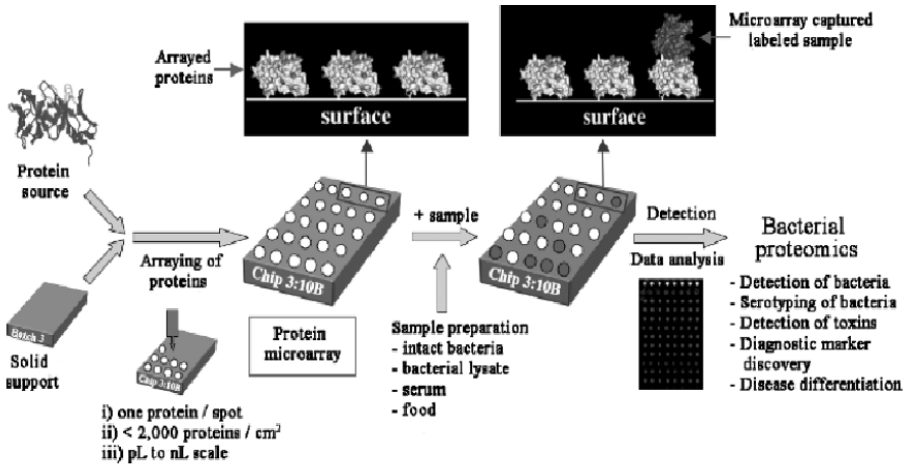


Figure 4 - Schematic representation of the workflow the protein microarray technology (27).

Protein microarrays can be divided in two major categories: analytical protein microarrays and functional protein microarrays (26). Analytical protein microarrays consist of antibodies that have high specificity to certain proteins, allowing the detection of their presence (26,28). Detection is either accomplished by direct labelling or using a reporter antibody in a sandwich like format, antibody-target-antibody(26). Functional protein microarrays consists on spotting a set of proteins in a given sample and testing their reactivity with the specific molecules. This allows the study of interactions between the sample proteins and, for example, other proteins or lipids (26).

This technology have been used in the detection of microbial pathogens, such as *Salmonella enterica* Serovar Typhimurium (28,29). Despite this technique not being the most popular tool for bacteria detection, a few studies were made using this method. An example is the study conducted by Howel *et al.* (30) were they propose a technique for the assembly of antibody microarrays in order to detect an *E. coli* strain and *Renibacterium salmoninarum* (30). Besides being able to successfully detect both bacteria, their results also show that the *E. coli* strain does not significantly bind to non-specific antibodies, making the method sensitive enough to produce accurate results (30).

Their work seems to indicate that protein microarrays can be a reliable tool to detect previously known bacteria, with the results being obtained in a short period of time. However, it requires previous knowledge of the bacteria in the analysis, which is not adequate for clinical purposes.

1.1.4.2 Mass Spectrometry

One of the most promising recent approaches for bacteria identification is based on mass spectrometry, through the analysis by MALDI-TOF (Matrix Laser Dissociation Ionization - Time of Flight). In fact this approach has the ability to provide results within a time frame of five minutes and with only a single colony which is in line with the requirements for clinical purposes: short time consuming; reproducibility; and reliability (26).

Comparing the use of mass spectrometry with the most common genomic approaches, for example 16S rRNA sequencing, several advantages can be highlighted such as: shorter analysis times after culture; lower cost *per* analysis; and higher discrimination power. Another advantage of using mass spectrometry to achieve a bacteria identification is that no amplification is needed after culture, when compared to genomic approaches. Nevertheless, the possibility of amplifying the genetic material is extremely advantageous when dealing with scarce amounts of sample (38). Another important aspect is that the most commonly used genes for bacteria identification are usually the most conserved along a bacteria strain or genus (38), which brings new difficulties when trying to identify bacteria isolates. For example, these genes can be so similar when compared among different strains of the same species which hinders their identification, resulting in false-positive hit (38). The use of mass spectrometry overcomes most of these problems, especially because all the proteins in the bacteria isolate are ready for analysis, conserved or not. For the cases in which analysis of just one protein does not provides enough information, a different protein or a set of proteins can be used (38). This is especially relevant when looking at the data provided by mass spectrometry of the protein content of a cell or even a cell fraction (38).

1.1.4.2.1 Techniques used for bacteria identification

Mass spectrometry can be roughly described as a process of detecting the different molecules on a sample, by analysing the mass of those molecules. The main components for a mass spectrometer are the ion source, analyser and detector. The ion source is responsible for the ionization of the molecules, being MALDI and Electrospray as predominant in nowadays. The analyser filters the ions produced by the ion source, according to their mass-to-charge rate, which reach to the detector, producing a mass spectrum. Some of the most commonly used analysers are the Orbitrap, Ion Trap, TOF (Time-of-flight) and the quadrupole.

For proteomics, in particular shotgun, the electrospray-ionization mass spectrometry, commonly referred as ESI-MS is the elected ion source. Electrospray mass spectrometry is based on a different ionization process. The sample, which must be in a solution, is passed through a capillary needle with an electric charge. The electrospray ionization is driven by the high voltage (2-6kV) applied (31). When the sample passes through the needle it becomes ionized and changes to a gas-state (32). This process can be described in three different steps: dispersion of a fine spray of charged droplets, evaporation of the solvent and ejection of the ion from the highly charged droplets (33). After the ions' ejection they go to the analyser which usually is a quadrupole analyser. The quadrupole consists of two pairs of electrified rods, with each pair having the same but opposite voltage (33). This creates an electric field inside the quadrupole and considering the movement of the ions when affected by an electric field is directly related to their mass/charge ratio, the ions will arrive to the detector at different times based on their *m/z* (33). The ions formed by the electrospray ionization can be the result of multiple-protonation, especially in the case of proteins or peptides, resulting in spectra that could be difficult to read (23,33,34). Another difficulty when using electrospray ionization with complex protein mixtures is the clogging of the needle of the electrospray (23). Considering that for electrospray analysis every sample must be in liquid state, make it ideal for the coupling to liquid chromatography. In addition, the use of liquid chromatography in the mass spectrometer improves the analysis by decreasing the sample complexity and minimizes some of the referred problems (23). A scheme of the usual workflow for this approach is seen on figure 5.

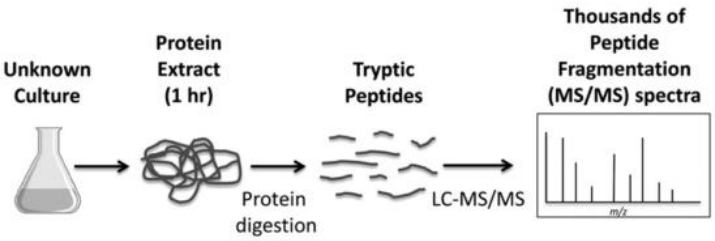


Figure 5 - Example workflow for LC-MS/MS analysis of an unknown culture isolate. Adapted from (35). Protein extraction is performed to the unknown culture, followed by the tryptic digestion. The resulting peptides are analyzed using LC-MS/MS and the proteins are identified. Finally the obtained proteome is compared with a reference database to achieve an identification (23).

For bacteria identification, the most common is the matrix-assisted laser desorption/ionization coupled with a time-of-flight analyser, usually referred as MALDI-TOF (36). In this technique the peptide mixture is mixed with a chemical reagent, the matrix, air dried and introduced in the mass spectrometer. The sample is then ionized with a laser leading to the formation of mainly monocharged peptide ions (36). This process occurs with the matrix absorbing the laser energy and transferring it to the acidified analyte, whereas the laser heating causes desorption of the matrix and $[M+H]^+$ ions of the analyte change to the gas phase (31). The forming ions are then radiated from the ionization chamber to the analyser and the time-of-flight is then measured. Since there is a direct relation between the time-of-flight and the mass-to-charge ratio, it is then possible to determine the mass-to-charge ratio of each ion (26). When dealing with microbial samples the matrix used can be α-cyano-hydroxycinnamic acid (α-cyano), sinapinic acid or 2,5-dihydroxicbenzoic acid (DHB) (37). The choice of the matrix depends on the focus of the analysis: α-cyano appears to be more suitable when focusing on proteins, whereas the sinapinic acid is more used for peptides and DHB as shown to be more efficient when focusing on oligosaccharides, glycopeptides and glycoproteins(38).

Using MALDI-TOF mass spectrometry it is possible to obtain a bacteria mass fingerprint which can be used to identify the isolate through comparison to a reference database (37). Different companies have developed many databases and software packages that are able to compare the sample mass fingerprint with the entries of the database (37). The first software developed for this purpose, which is no longer available,

was the MicrobeLynx. Several studies have been conducted using this technique and it was already shown that protein profiles can be obtained not only from crude lysates but even from whole cells or fractions (23). The major disadvantage of this technique is the strong dependence on sample preparation (31).

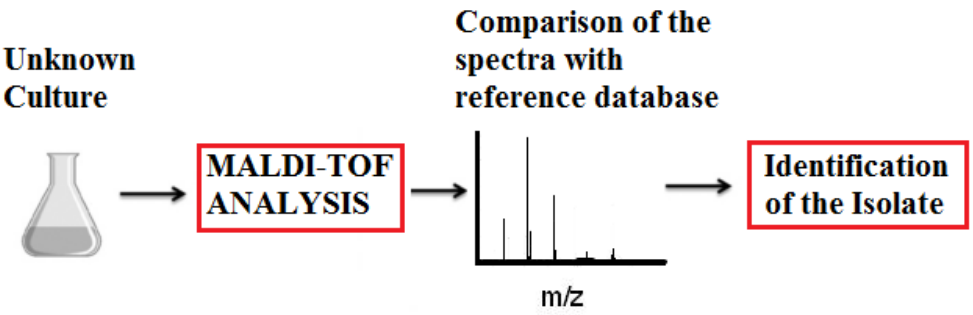


Figure 6 - Example workflow for identification of a bacterial isolate using MALDI-TOF mass spectrometry. Protein extraction is applied to the unknown sample, followed by the spectra acquisition using MALDI-TOF. The obtained spectra is then compared to a reference database and the most similar spectra should correspond to the correct identification.

In order to identify an unknown isolate, MALDI-TOF is used to obtain a bacteria mass fingerprint (34). Despite the fact that MALDI-TOF is the most common approach, MS/MS can also be used (34). Using this approach a peptide fragment fingerprint is acquired, for each select ion subjected to MS/MS, and by comparison of those peptide mass fingerprints with a reference database, the identification of the protein is obtained (34). In addition, using gel electrophoresis or liquid chromatography it is possible to overcome one of the biggest disadvantages of this technique: MALDI-TOF preferably ionizes small molecular weight molecules (<10Da) (34). Using a separation technique and selecting the proteins with high molecular weights it is possible to promote their digestion with an enzyme, for example trypsin, to obtain their peptides (34). Performing a MS/MS analysis it is possible to identify the sample proteins and with the identified proteins it should be possible to associate those proteins to specific strains of bacteria, hence achieving an identification of the bacterial sample.

1.1.5 Use of MALDI-TOF in Bacteria Identification

One of the first studies using MALDI-TOF to achieve a bacteria identification was conducted by Krishnamurthy *et al.* (39) where cell lysates of bacteria isolates were used and analysed by MALDI-TOF. The analysis was based on the spectra obtained and attempted in the annotation of characteristic peaks for each bacteria isolate (39). It was shown that identification of individual organisms down to the species level was possible and in some cases, even to the strain level (39). These results showed that it was possible to identify strain-specific traits for some bacteria such as *Bacillus anthracis, Brucella melitensis* and *Yersinia pestis* (39). In the case of *B. anthracis* it was also showed that differentiation of this strain and related *Bacilli* was possible (39). A study conducted by Edward-Jones *et al.* (40), in the year 2000, used MALDI-TOF to distinguish between methicillin-sensitive and methicillin-resistant *Staphylococcus aureus*. Instead of using protein extracts from the isolated bacteria they used the intact cells in mass spectrometry analysis. Most of the ions detected appeared in the range of 500-2000 *m/z* and allowed to clearly identify common peaks between methicillin-resistant and sensitive groups. In addition to the identification of characteristic peaks for each group, it was also possible to correctly categorize each bacteria has being methicillin-resistant or not (40).

Over the last decade more studies were made and the usefulness of this technique in bacteria identification was extended. A major breakthrough that enabled even better results was the development of bioinformatics software tools and reference databases. Indeed, this allowed a direct comparison of the unknown isolated bacteria spectra with reference spectra, without the prerequisite to obtain the reference spectra for every analysis or specific bacteria (23). Currently there are three identification databases available: Bruker Biotyper, SARAMIS and Andromas (23). These databases are now well established with the Bruker Biotyper containing more than 4500 unique species. The analysis is usually done using a Bruker MicroFlex (MALDI-TOF) and is based on species specific peaks present in standard reference mass spectra compared with the unknown sample. The SARAMIS database has two versions of the software with the first one being coupled with a Shimadzu Axima mass spectrometer as a single product. The second version is provided by bioMérieux as the VITEK MS system (23). The database of both versions contains about 3000 microorganisms (23). The Andromas software is distributed by Andromas SAS and contains around 700 bacterial strains (23). All of the systems have been compared with classic biochemical tests and 16S rRNA sequencing and all of them

were successfully shown to have enough precision to be considerable feasible (23). It should also be noted that the size and quality of the database plays an important role in the analysis and the reliability of the obtained results.

On the last few years this technique has been used with great success in bacteria identification. A study performed by van Veen *et al.* (41) tested the use of MALDI-TOF in the identification of several bacteria using the Bruker Biotyper software and included database (version 2.0). In this work they used 327 clinical isolates cultured from patient materials. The results obtained using the Bruker Biotyper were compared to those obtained using conventional methods and, in the case of discrepancies, the results were confirmed by 16S genes sequencing (41). Using MALDI-TOF they were able to identify 95.1% to the genus level and 85.6% were identified to the species level. After this initial procedure, they proceeded with a validation study where they analysed 980 clinical isolates. The results obtained showed that the overall performance of MALDI-TOF was significantly better that the conventional biochemical systems, having obtained 92.2% correct species identification with MALDI-TOF and only 83.1% with the conventional systems. In addition, MALDI-TOF gave lower genus identification errors (0.1) against the traditional biochemical tests (1.6%). Most of the misidentifications by MALDI-TOF spectrometry were associated with lack of reference spectra in the database, showing the need for building the most complete database possible (41).

MALDI-TOF had showed to be a very reliable technique for use in a clinical environment with many different studies being done in the last few years (23,34,42). Despite good results obtained when dealing with the clinical relevant bacteria, MALDI-TOF analysis have showed slightly worse accuracy when dealing with anaerobes and Gram-positive bacteria (23). The latter can be explained by the thick peptidoglycan in the cell wall of Gram-positive bacteria, which is assumed to hinder the ionization process. In order to overcome this limitation a different approach has been suggested by disruption of the cell wall and subsequent protein extraction (23).

An alternative approach of the use of mass spectrometry for bacteria identification is the electrospray ionization mass spectrometry. However this technique is not so successful as MALDI-TOF (23). A major difference between MALDI-TOF and ESI-MS is that MALDI can use intact cells and whereas, ESI cannot, since it leads to needle clogging and extremely complex spectra. In order to overcome this problem, instead of intact cells, cell lysates are usually used. The protein extracts are then exposed

to trypsin digestion and analysed on line with ESI-MS/MS (34). The obtained spectra are representative of the peptide composition of the bacteria isolate and is used to compare with a reference database (34). However since the spectra obtained from this approach tend to be extremely complex, the development of bioinformatics tools that are able to compare the unknown sample spectra with reference database, giving a correct identification as result, are much harder to develop (34). In both cases, the success of this approach is dependent of the development of enhanced algorithms and curated databases.

1.1.5.1 Role of the ribosomal proteins in the analysis

As already stated, to achieve a bacteria identification based on the spectra obtained by mass spectrometry, a comparison between the mass fingerprint of the samples and reference fingerprints in a database is performed. For the comparison of the spectra the approach usually consists on the detection of specific proteins that are typical of a certain strain or species of bacteria (34,43). This is achieved by searching for matches between peaks of the bacteria isolate spectra with the reference spectra (43). Ribosomal proteins are important for this approach since they can represent up to 20% of the weight of the cytosolic proteins (43). Besides the significant abundance of ribosomal proteins, another feature that make them suitable markers for use in this technique is the fact that rRNA is highly conservative along the same strain. Therefore ribosomal proteins of bacteria of the same strain should be similar enough to make them potential markers. However, for the ribosomal proteins to be considered potential markers it is also required that ribosomal proteins of bacteria of different strains or species have significant mass difference, making possible to distinguish between close species or strains (43).

In order to identify and assess the characteristics of the proteins detected in MALDI-TOF analysis when whole bacteria cells are analysed, Ruzhov and Fenselau (44) conducted a study to identify the proteins detected by MALDI of *Escherichia coli* K-12. Their results showed that the majority of the proteins were cytosolic proteins and mostly ribosomal proteins. In addition, a study of multivariate pattern recognition of markers of *Escherichia coli* in different growth phases, supported the major role of the ribosomal proteins in a mass spectrometry analysis (45). On this study it was used MALDI-TOF to analyse the protein content of 12 cell samples of *Escherichia coli* (45). When the cells were in their exponential phase it was possible to identify 27 proteins of which 15 were

ribosomal proteins. In the stationary phase it was possible to identify 18 proteins where 5 of them were ribosomal. Finally, in the decline phase 27 proteins were identified, with 6 being ribosomal proteins, as seen in table 2. These results showed that ribosomal proteins can be potential markers, especially when the cells are in their exponential phase (45). These results are in accordance with the previous studies whereas ribosomal proteins can represent 45% of the total mass of *Escherichia coli* and 21% of the proteins of the cell (45).

Table 2 – Ribosomal proteins identified during the stationary phase of a *Escherichia coli* culture on the study conducted by R. Momo *et all.*(45)

Molecular Weight (Da)	Protein Name	Protein Description	Accession Number
4309.30	RpmJ	50S Ribosomal Protein L36	Q9RSK0
5380.55	RpmH	50S Ribosomal Protein L34	P0A7P5
6240.06	RpmG	50S Ribosomal Protein L33	P0A7N9
6507.28	Rmf	Ribosomal modulation factor	P0AFW2
8854.33	RpsR	30S Ribosomal protein S18	P0A7T7
11579.52	RpsN	30S Ribosomal protein S14	P0AG59

1.1.5.2 Algorithms used in Bacteria Identification with Mass Spectrometry

When using mass spectrometry to identify bacteria a major aspect is the algorithm used for comparison of sample data with reference database. Throughout the years different algorithms were developed and different results were achieved, as seen on table 3. These algorithms can be classified as being Library-based or Bioinformatics-enabled (46). Library-based approaches can simply be described as using an algorithm

that compares spectra of unknown sample with spectra of known reference bacteria (46). While library-based approaches have been used in the majority of the studies with MALDI-TOF MS, Bioinformatics-enabled approaches have also been developed, especially due to the rapidly increasing number of bacteria with fully sequenced genomes (46). This second type of approach involves identification of the proteins in MALDI profiles and searching available genome databases of bacteria to find a match between those proteins and genome sequences. In order to achieve this identification MS/MS is usually applied (46). The advantage of this approach lays in the fact that it does not need to build libraries and the experimental conditions do not need to be standardized across laboratories, as in the case of library-based approaches (46). When comparing the performance of both approaches there is no clear consensus of which one performs better. However library-based approaches are more commonly used (46). An overview of both approaches is represented on figure 7.

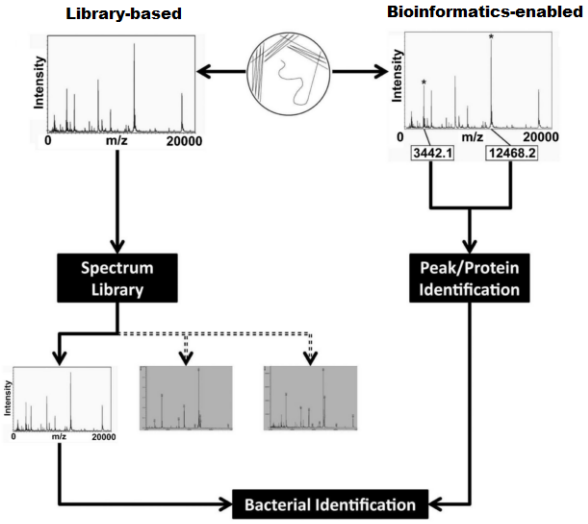


Figure 7 - Schematic representation of the different approaches in bacteria identification algorithms (46).

The algorithms are usually available in software packages similar to the ones already mentioned: Vitek MS system, SARAMIS and Bruker's MALDI Biotyper. The Vitek MS uses an identification matrix that has been computed by Advanced Spectrum Classifier, which is an algorithm based on supervised learning procedures and has a data set of more than 25000 binned reference spectra (the MS spectral accuracies are rounded to nearest integer atomic mass unit values; this promotes the reduction of false negatives)

(47). SARAMIS uses SuperSpectra to identify reference spectra, which were computed by weighting peaks in consensus spectra according to their specificity for the different taxonomic levels (47). Bruker's MALDI Biotyper uses Main Spectra as the reference spectra, which results from a computed consensus spectrum from multiple spectra. The consensus spectrum is compared to reference mass spectra and the identification result is linked to a scored computed by counting match peaks in the sample mass spectra (47). The results achieved with these software can also be found in table 3.

Table 3 - Results achieved using different algorithms for use with mass spectrometry in order to correctly identify bacteria

Algorithm / Software	**Method used by the algorithm**	**Sample**	**Sample Preparation Method**	**Results**	**Ref.**
Arnold's and Reilly's algorithm	Modified cross correlation of spectra	Five different strains of *Escherichia coli*	Whole cell mass spectrometry	Able to fully differentiate tested strains	(48)
Jarman's *et al*. algorithm	MALDI-MS fingerprint built using several spectra (the peaks that are considered part of the fingerprint are those that are present in more than 70% of the spectra). Identification is done by comparison of the sample fingerprint with a fingerprint database	*Bacillus atrophaeus* ATCC 49337, *Bacillus cereus* ATCC 14579T, *Escherichia coli* ATCC 33694, *Pantoea agglomerans* ATCC 33243, and *Pseudomonas putida* F1	Whole cell mass spectrometry	94% of the *B. atrophaeus* samples were correctly identified , 35% of the *B cereus* , 95% of the *E coli* and 100% of the *P. agglomerans* and *P. putida* samples	(49)
Manchester Metropolitan University Search Engine (MUSE)	Spectra correlation	212 of 35 strains from 20 species, three Gram-positive genera and nine Gram-negative genera	Whole cell mass spectrometry	79% correct identifications to the strain level, 84% to the species level and 89 to the genus level	(50)

Algorithm / Software	Method used by the algorithm	Sample	Sample Preparation Method	Results	Ref.
Tao's *et al.* algorithm	Statistical weight factor is assigned to each peak according to the frequency of appearance. To achieve an identification, the weight factors of each peak of the unknown isolate that matches a bacteria in the database are summed, yielding a weight factor score. The bacteria in the database that resulted in the higher score is proposed as the identification	10 strains of nine different bacteria species	Proteins extracted and separated by off-line HPLC	100% correct identification for unknown samples that corresponded to bacteria in the reference database	(51)
Hettick's *et al.* algorithm	Linear discrimination statistical analysis	Sixteen strains from 4 species of the genus *Mycobacterium*	Whole cell mass spectrometry	98% correct identification to the species level and 95% to the strain level	(52)
Genetic Algorithm embedded in ClinPro Tools software	Non-described	Six common human pathogenic bacterial species	Whole cell mass spectrometry	100% correct identification of unknown samples that corresponded to the reference database. Only 1 of 14 isolates of other bacteria, that were not in the reference database, was misidentified;	(53)
Bruker MALDI Biotyper	Non-described	>1600 isolates from over 100 different bacterial species	Whole cell mass spectrometry	84.1% correct identification at the species level and of the remaining 15.9%, 11.3% were identified at the genus level	(54)
SARAMIS software	MALDI-TOF fingerprint pattern with specific signals at strain, species, genera and family level. The fingerprint of the unknown isolate is compared with a reference database and an identification is proposed	317 isolates of 40 different species and 23 different genera	Whole cell mass spectrometry	97.2% correct identifications to the genus level and 93.4% to the species level	(55)

Algorithm / Software	Method used by the algorithm	Sample	Sample Preparation Method	Results	Ref.
Vitek MS System	Spectra correlation	1,129 isolates including 1003 routine isolates, 73 anaerobes and 53 bacterial enteropathogens	Whole cell mass spectrometry	93.2% of the routine isolates were correctly identified to the species level, 75.3% of the anaerobes and 15.1% of the enteropathogens were also correctly identified. The last result is explained by the lack of discrimination of *Shigella* strains, which were all identified as *Escherichia coli*	(56)

Despite the good results attained using the available software, they still have some drawbacks. For instance, dealing with an unknown isolate without one in the reference database, the identification will be as a close species, resulting in a false-positive result. Another major drawback is related to the format of file generated by the mass spectrometers and the requirements for the software, which may differ, making impossible to use some spectrometers with some software packages (57). With this in mind, there is a need for the development of new algorithmic approaches that could solve, or at least minimize, these weaknesses, allowing precise identification to the strain level, making the use of MALDI-TOF in bacteria identification a more reliable approach.

1.2 Comparison of the different approaches for Bacteria Identification

As described, there are many different approaches that can be used to positively identify bacterial samples. The most commonly used approaches mainly consist in phenotypical/biochemical differences, 16S rRNA sequencing, IR spectroscopy and (mass spectrometry) MALDI-TOF. Each one of these methodologies have different steps in their protocols with some differences. A schematic representation can be seen on figure 8.

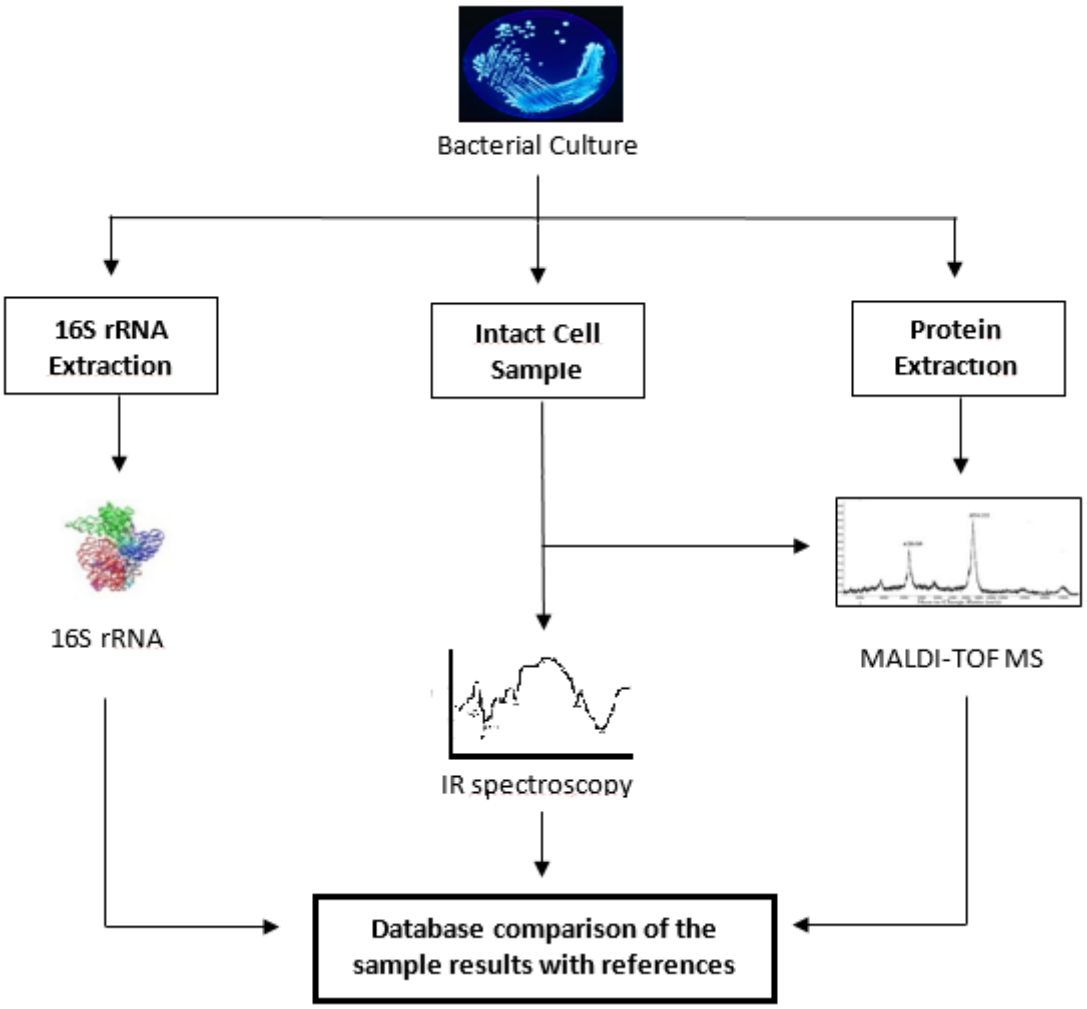


Figure 8 - Schematic representation of the workflow for the most common bacteria identification methodologies.

Regarding bacteria identification the most important features of the different methodologies are the accuracy, speed and cost of the analysis, as well as ease of use. It is important to compare these features of the different methods in order to understand the feasibility of their application in clinical environment (table 4). MALDI-TOF had showed a high accuracy when identifying bacteria based on the mass spectra of the isolates and comparing those spectra with reference databases (58). An average rate of overall correct identification when using MALDI-TOF is expected to be >90% (58). Another important aspect is that most of the misidentifications by mass spectrometry data comparison are due to problems with databases, either lack of reference data or misidentified references (58). When comparing this accuracy with the traditional methods it is possible to observe similar results. However those techniques are either highly dependent of the user skills or

require more time to provide results. Comparing the time consuming, after culture, required for bacteria identification, MALDI-TOF needs only six minutes to produce the result, whereas conventional techniques take five to forty-eight hours (54). Another important aspect of MALDI-TOF resides in the level of training, being high-technology is considered low-to-medium when compared with the alternative methods (54). In terms of the cost of the identification of bacteria isolates MALDI-TOF appears as the election method with low-cost involved (with an estimate cost *per* identification of <2€) (54). This cost is considerable low when comparing with commonly used methods for identification like Vitek system, which has an average cost of 6-8€, or the analytical profile index (API) system identification, with an average cost of around 6€ (54). The only technique that also presents a low cost besides mass spectrometry is the Gram staining. However, this technique solely distinguish between Gram-negative and Gram-positive bacteria, which in most cases is far from being enough to provide clinical valuable information and further analysis is often needed, rising the identification cost (54).

Table 4 - Advantages and disadvantages of the different methodologies of bacteria identification

Methodology	Cost	Average Accuracy	Time needed	Advantages / Disadvantages	References
Biochemical / Phenotypical methods	6-8€	~65%	24h-48h	Advantages: • Simple method • Accurate for some specific bacteria Disadvantages: • Time-consuming • Expensive • Lacks discrimination between different bacteria • Dependent of the analysis of the user	(10,54)
16S rRNA Sequencing	28-36€	65-83%	36h-72h	Advantages: • Good discrimination power at the genus level • Good discrimination power in some clinical relevant species • Possibility of amplification of the sample Disadvantages: • Time-consuming • Expensive • Some bacteria have more than one copy of 16S rRNA with different sequences • The similarity between sequences of different species may not allow the differentiation • 16S rRNA has to be extracted and separated fro bacteria isolate	(13,14, 59)

Methodology	Cost	Average Accuracy	Time needed	Advantages / Disadvantages	References
IR spectroscopy	Not available	~80%	Culture: 12h-24h After culture: ~20min	Advantages: • Fast methodology • Cost-effective • Good discrimination between Gram-positive species • Direct culture analysis Disadvantages: • Lacks discrimination between Gram-negative species • Optimization related to the culture time, which may differ in different bacteria species	(5,22)
MALDI-TOF MS	1.43€	>90%	Culture: 12-24h After culture: 4-8 min	Advantages: • Fast methodology • Inexpensive *per* analysis • High discrimination power • Intact cells can be used Disadvantages: • Worse performance when dealing with Gram-positive bacteria	(39,54, 60)

Overall, the MALDI-TOF approach is very suitable to use in clinical environment analysis since it has a low cost *per* analysis, a short time required to obtain results and an accuracy in the results obtained. In addition, this methodology allows the use of intact cells which makes the overall protocol simpler and less prone to contaminants and loss of sample. The major problem of this method relies in the poor discrimination of Gram-positive bacteria. Considering all features of bacteria identification using MALDI-TOF, one can assume that this approach can potentially produce very good results and further development should be done in order to optimize this technique.

1.3 Objectives of this work

MALDI-TOF has shown considerable promise as a reliable method for bacteria identification. However it still has some drawbacks that hinder its establishment as the new standard for bacteria identification. For instance, the available software packages for the identification of bacteria using MALDI-TOF spectra still have a low amount of entries in database, which often results in misidentifications. Besides, the identification of Gram-

positive bacteria using the already available software packages has a considerable lower success identification rate, which does not make this approach reliable in a clinical environment.

With this in mind, this work focused on the development of a library-based application capable of identifying unknown bacteria samples. The application should be able to identify bacteria by comparing the unknown sample mass profile with a database containing mass profiles of known bacteria. It should identify both Gram-negative and Gram-positive bacteria with the correct identification rate desired for a clinical environment. Also, the application should have a low false-positive rate, providing an accurate alternative to the already available methods and pushing this approach to become the standard of bacteria identification in a clinical environment.

2. Experimental Procedures

2.1 Protein Extraction

The samples used on this work were provided by Hospital Infante D. Pedro, EPE - Aveiro. The bacteria were collected from infected patients and were cultivated in a plaque with the proper medium. After the growth of the bacteria, a sample of the culture was collected and suspended on a saline solution.

With the sample on a saline solution the protein extraction protocol was started. The first step was to subject the sample to a light vortex, followed by 2 minutes of centrifugation at 15000g. After the centrifugation the supernatant was removed and the pellet was resuspended on 300µL of milli-Q H_2O. This step was followed by the usage of the vortex until the pellet was completely mixed with the solvent. Next, 700µL of EtOH (ethanol) were added, followed by 30 seconds on the vortex. With the EtOH added, a new 2 minutes of centrifugation at 15000g was performed. When the centrifugation was finished the supernatant was removed and 120µL of a 1:1 mixture of 70% formic acid and acetonitrile were added, followed by vortex until the pellet was resuspended. After that, the samples were centrifuged for 2 minutes at 15000g and the supernatant was saved, which corresponds to the protein extract.

2.2 Protein Quantification

The protein quantification was achieved using two different methods, the DC Protein Assay of Biorad and the measure of the absorption of the samples at 280nm wavelength. The protein quantification using the DC Protein Assay of Biorad was done using the standard protocol for microplates. The assay was started by preparing solutions of BSA (Bovine Serum Albumin) protein at different concentrations, in order to be used as the standards of the protocol. The chosen concentrations were 0.2, 0.4, 0.6, 0.8, 1.0 and 1.4 mg/mL. The assay proceed by adding 5µL of each standard and sample to different spots of the microplate. Then to each spot 20µL of reagent of A were added, followed by the addition of 200µL of reagent B. The microplate was then lightly agitated in order to mix the reagents and the samples/standards. The plate was then left to incubate for 15 minutes at room temperature in the dark. After the 15 minutes, the absorption at 750nm

wavelength was read. The concentration values of the samples were estimated by using the standards values to build the calibration line, from which the concentration of the samples was computed.

The second method used to assess the sample protein concentration was the measurement of the absorption at 280nm wavelength. The first step was the preparation of the standards. The chosen standards were solutions of 0.2, 0.4, 0.6, 0.8 and 1.0 mg/mL of myoglobin. 40µL of each sample and standard were added to a quartz optical cuvette, followed by the addition of 760µL of the same solution used to prepare the samples (1:1 solution of 70% formic acid and acetonitrile). After this step, the absorption values at 280nm wavelength were read. To estimate the concentrations of the samples, the absorption values of the standards were used to build the calibration line, from which the concentration of the samples was computed.

2.3 Spectra acquisition

The spectra of the samples were obtained using a 4800 MALDI TOF/TOF AB Sciex mass spectrometer operating at middle mass setting in positive mode. The selected mass range was 3000-20000 *m/z* with a focus mass of 10000 Da. On average, the spectra were acquired with 1050 shots *per* spectra. The matrix used in the MALDI-TOF analysis was the α-cyano. The matrix was prepared by dissolving 5mg of α-cyano *per* mL of a 1:1 solution of milli-Q H_2O and Acetonitrile, followed by the addition of trifluoroacetic acid (TFA) in order to achieve a final concentration of 0.1% TFA.

The MALDI-TOF plate was prepared by mixing each sample with the matrix (1:1 proportion) and adding 0.5µL of each sample + matrix to four spots of the plate. The spectra were acquired using the MALDI-TOF spectrometer in the positive mode and the selected mass range was 3000 *m/z* to 20000 *m/z*.

3. Results and Discussion

This work started with the extraction of the protein fraction of the bacterial samples. This was followed by the quantification of the total protein of the extract to ensure that the extraction was provide enough protein content to allow the acquisition of high-quality mass spectra of the samples. The next step was the analysis of the bacteria mass profiles to assess the distinction power between the different strains. For this part of the work, two strains of *Escherichia coli* and one strain of *Klebsiella pneumoniae* were used.

The final part of this work was the development of the library-based application and the subsequent assessment. The application was tested by analysing the output results in three different situations: i) the sample loaded was present in the database, ii) the sample loaded had a similar entry in the database and iii) the database did not contain any entry similar to the loaded sample. Moreover, the application was also tested using some of its features.

3.1 Sample Preparation

The first step of this work was the extraction of the protein content of some bacteria samples. Following the extraction of two samples (two different strains of *Klebsiella pneumoniae*), the quantification of the protein content of the extracts was performed in order to assess if there was enough protein content for the spectra acquisition. The first method used was based on the DC Biorad protein assay for protein quantification. Considering the protein extracts are in an acidic solution (70% formic acid in a 1:1 solution with acetonitrile) it was necessary to neutralize the sample to make it compatible with the DC Biorad protein assay. To achieve this, a solution of ammonia was added in a 3:1 proportion (3 mL of ammonia for each 1 mL of protein extract). However, the absorption values of the samples at 750nm with the reagents provided were too low, which gave negative values thus indicating that the protein concentration after adding the reagents was too low. In order to estimate the concentration, two times the initial volume of the sample was used, maintaining the same ammonia proportion; however the obtained absorption values were still too low. These results could be explained by the protein content of the sample which might not be compatible with the used method.

In order to estimate the concentration of proteins of the sample, the absorption of the sample at 280nm was used comparing the absorption of solutions of known concentration of myoglobin. The concentration of the standards were 0.05, 0.10, 0.20, 0.40, 0.80 and 1.00 mg/mL of myoglobin in the same buffer as protein extracts. The samples used were three new samples provided by the local hospital and consisted of two different strains of *Escherichia coli* and one strain of *Klebsiella pneumoniae*. The results of this test showed a concentration of total protein content of 0.414 mg/mL for the *K. pneumoniae* - 164092 sample and 0.805 mg/mL and 1.299 mg/mL, for the *E. coli* – 163906 and for the *E. coli* – 163962 respectively. These results showed that the extraction method was performing accordingly and the protein extract should have enough protein content to produce the high quality spectra one would need.

After the protein quantification of protein extracts the acquisition of the mass spectra was started, using a MALDI-TOF spectrometer at middle-mass settings. While acquiring the spectra, some problems about the quality of obtained spectra were detected. It was concluded that this was caused by freezing of the samples prior to the protein extraction, resulting in spectra with close to none information, as one can see in figure 9.

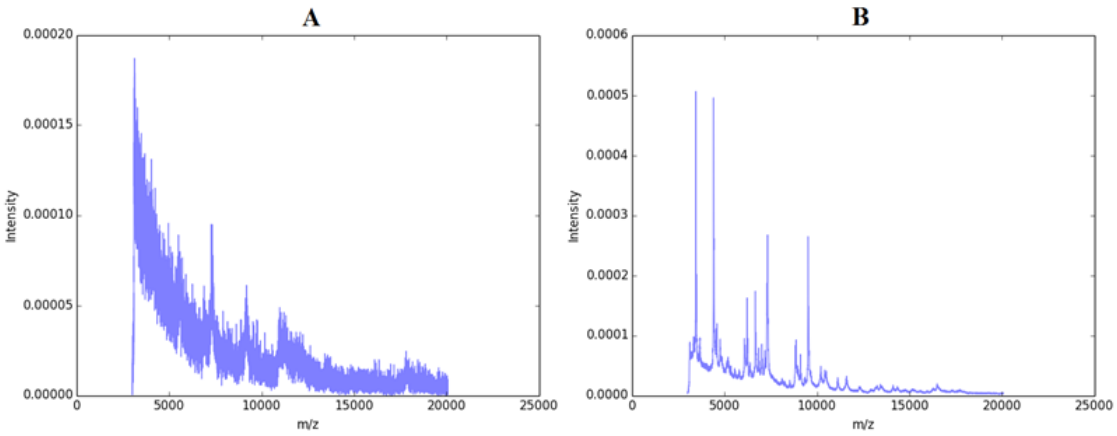


Figure 9 - Comparison of the quality of the spectra of a frozen sample (A) versus the quality of a fresh sample (B). The values of intensity are normalized to the sum of all intensity values.

As can be observed at figure 9, the first spectrum doesn't provide the information necessary to be considered the bacteria mass profile. So, considering that, the spectra that

were similar to the spectrum A weren't included in the bacteria mass profile database. As soon as this was realised the high quality of the acquired spectra was guaranteed by the usage of fresh samples without any freezing.

3.2 MALDI data pre-treatment

Prior to the development of the application there was a need to evaluate if MALDI-TOF spectra had the potential to differentiate bacteria based on their mass profiles. For this, two strains of *Escherichia coli* and one of *Klebsiella pneumoniae* were chosen (table 5). All of them show Multi-Drug Resistance (MDR) and are producers of Extended-spectrum Beta-Lactamases (ESBL). These strains were selected by considering two major factors: pathological infections caused by *E. coli* are among the most common and *K. pneumoniae* commonly shows resistance to the common antibiotics. These two characteristics make these samples interesting subjects for the preliminary study of the identification based on their mass profiles. With the samples selected, the extraction of their protein content was performed using the formic acid/acetonitrile method and the mass profiles were acquired using MALDI-TOF with a matrix of α-Cyano Hydroxicinnamic Acid. After the acquisition of the mass profiles, it was used several data pre-treatment/processing to highlight the differences between the profiles and potentially enhance the differentiation of the profiles.

Table 5 - Characteristics of the bacteria samples and sample preparation

Lab. ID	Bacteria Species	ESBL producer (Y/N)	MDR (Y/N)	CFU's/mL	Neutralized (Y/N)	Digested (Y/N)
163906	*Escherichia coli*	Y	Y	> 2,3 X 10^9	N	N
163962	*Escherichia coli*	Y	Y	> 2,3 X 10^9	N	N
164092	*Klebsiella pneumoniae*	Y	Y	> 2,3 X 10^9	N	N

The first step of the data treatment was the normalization of the intensity values. The normalization allows the comparison of different spectra independently of the intensity values obtained in the spectra acquisition. In order to do this, every value of intensity corresponding to each *m/z* values was divided by the sum of all the intensities of that sample. After the normalization, the intensity values were aligned to the same *m/z* values and only the intensities corresponding to the interval of 3000-18000 *m/z* were selected and used in further analysis. Using the normalized intensities the first plots were obtained (figure 10).

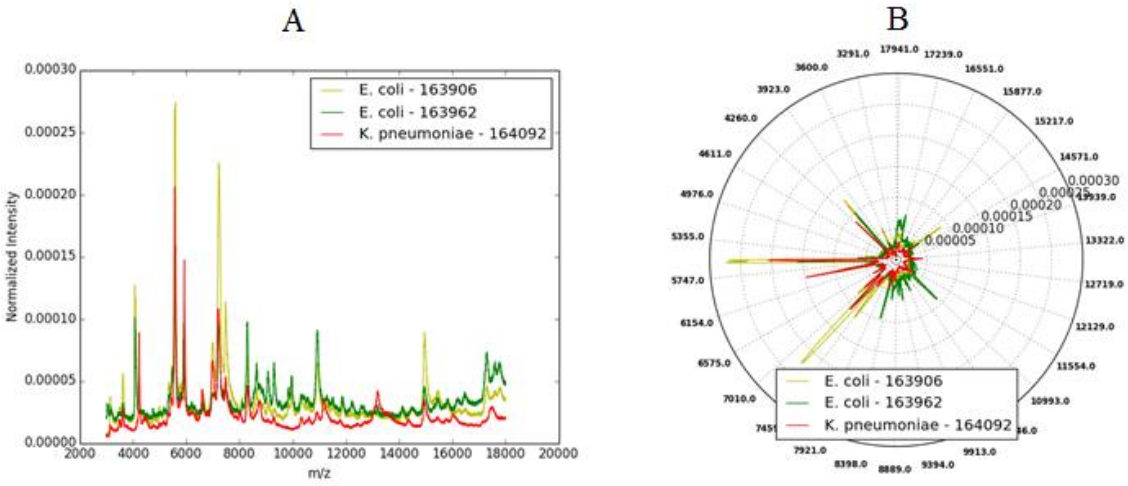


Figure 10 - Line plot (A) and radial plot (B) of the normalized intensities for each *m/z.*

In the first plot (A), a simple line plot for the three bacterial samples, already shows differences between the different strains. In figure 10-B, a radial plot was made, which enhances the visualization of the differences between the three species.

Looking at the plots in figure 10 it is possible to observe that the samples spectra have high intensity peaks at different *m/z* values. In order to further distinguish those peaks, the lower intensity values of each sample were temporarily removed. To perform this, different percentile values were computed: 0.10, 0.50 and 0.90. A percentile is the value of intensity below which a certain percentage of the data can be found. For example, a 0.10 percentile should be the intensity value at which every intensity lower than that corresponds to the first 10% of the data. This concept is usefull in the treatment of MALDI-TOF data, by calculating a certain percentile and removing all the values below

that. Since the lower intensity values are removed plotting the resulting data highlights the higher intensity peaks for the three samples. The results of these plots can be seen on figures 11 (for 0.10 percentile), 12 (for 0.50 percentile) and 13 (for 0.90 percentile).

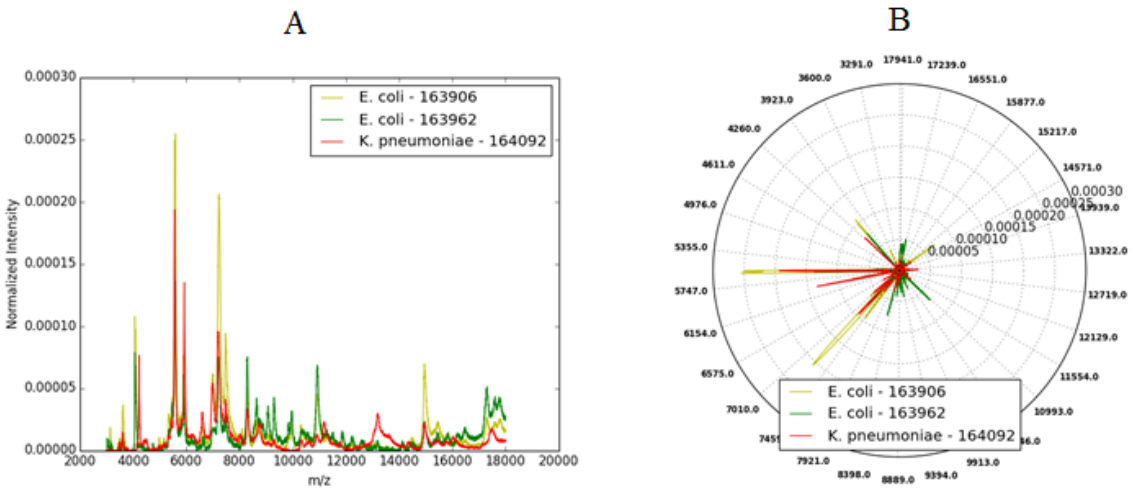


Figure 11 –Line plot (A) and radial plot (B) of the normalized intensities for each *m/z* removing the values below the 0.10 percentile.

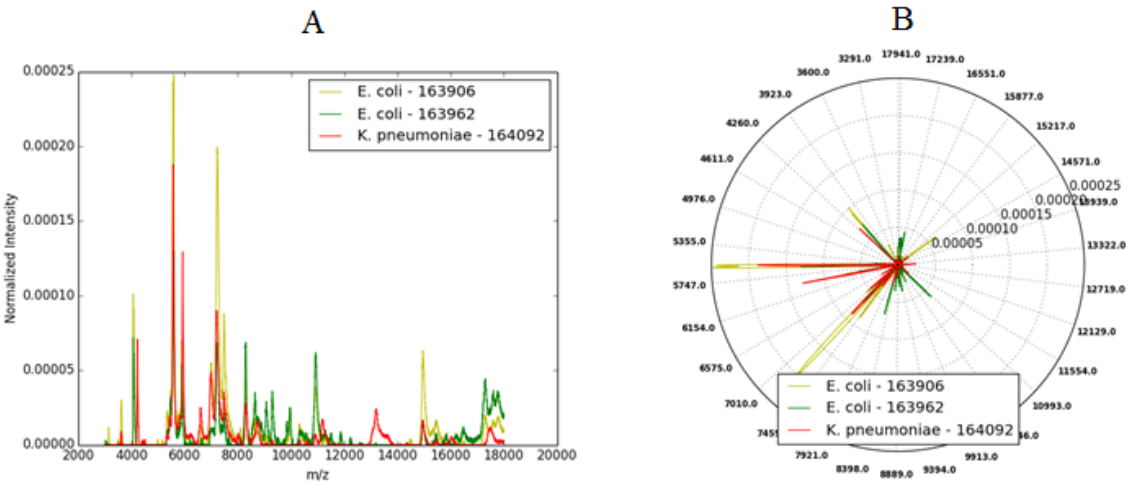


Figure 12 - Line plot (A) and radial plot (B) of the normalized intensities for each *m/z* removing the values below the 0.50 percentile.

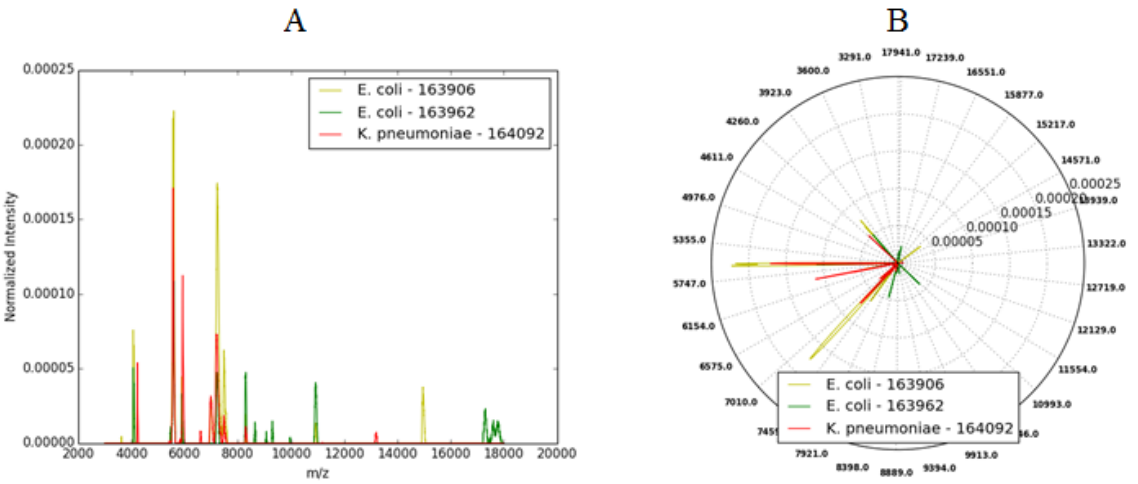


Figure 13 - Line plot (A) and radial plot (B) of the normalized intensities for each *m/z* removing the values below the 0.90 percentile.

The histogram analysis (distribution of intensity values) of each sample, leads to a better understanding if: i) the sample have a good amount of significant peaks; ii) the data is more comprised in the lower intensity values (which are harder to distinguish between the samples); iii) it has the possibility of being a discrimination factor. Besides the simple plot of the histograms (figure 14-A), it is possible to plot the cumulative histogram (figure 14-B) which provides us an easier way to analyse the distribution of the intensity values.

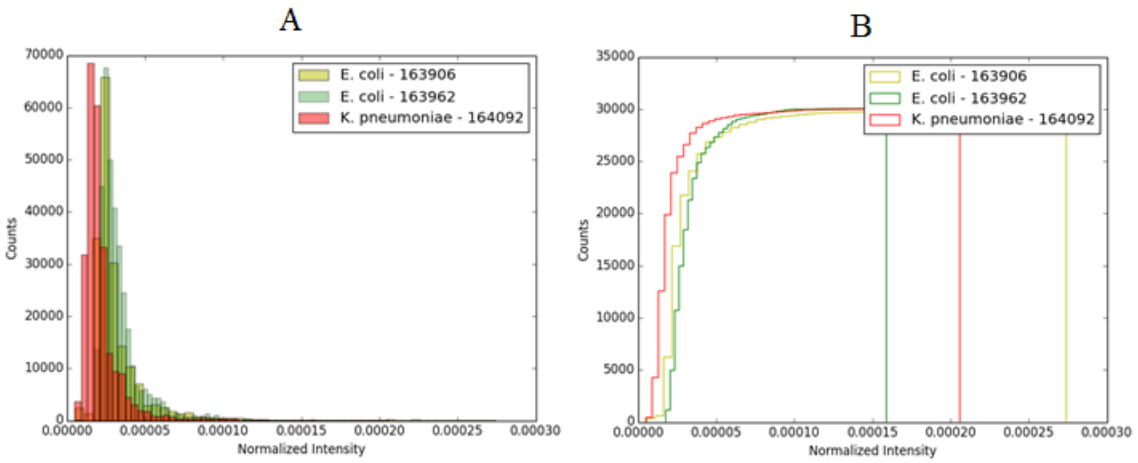


Figure 14 – Histograms (A) and Cumulative density distribuction function (B) of the intensity values for the three samples.

As observed in the line and radial plots, removing part of the data highlights the differences between the three different samples. Thus, percentile values were used: 0.10, 0.50 and 0.90 to build the plots depicted in figure 15.

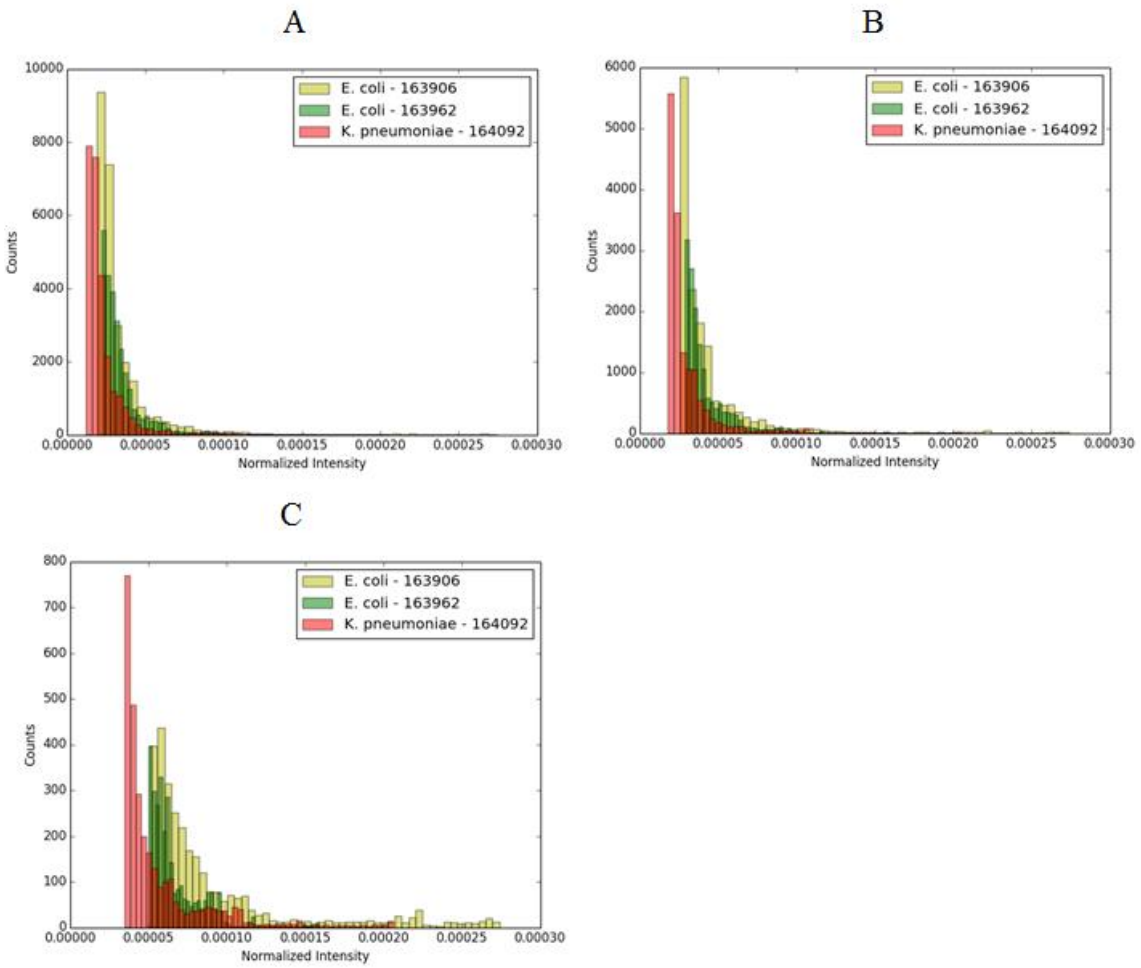


Figure 15 - Histograms of the intensity values of the three samples removing the values below 0.10 (A), 0.50 (B) and 0.90 (C) percentile.

Looking at the plots, it is possible to observe that the removal of an higher amount of the data enables the distinction between the different strains with lower overlaping, especially with the removal of the values that are below the 0.90 percentile. However, the frequency of the higher intensity values are still relatevely lower and harder to distinct between the different strains.

One of the problems that may arise when comparing the mass profiles of bacteria is caused by the alignment of the peaks made by the mass spectrometer, which often

causes the intensity peaks to appear deviated by a few *m/z* units. To avoid this problem a different approach can be performed using buckets of data, instead of using every intensity value. This approach consists of defining the range of buckets, which in this case a range of 750 values was selected, and assigning the intensity value for each bucket, which corresponds to the sum of all the individual intensities in that bucket. With this approach, the possibility of errors generated by the alignment of the intensities to the same *m/z* is reduced. The exact value of the *m/z* for each intensity is no longer relevant, since it is only required to know the sum value of the intensities for that range. In figure 16-A, the plot of this approach can be seen. While some buckets have a significantly higher value for a specific strain, it is still not easy to distinguish between the different strains. In order to enhance the differences between the different strains it is possible to apply the same treatment as before and remove intensity values below a certain percentile. The result of this can be seen on figure 16.

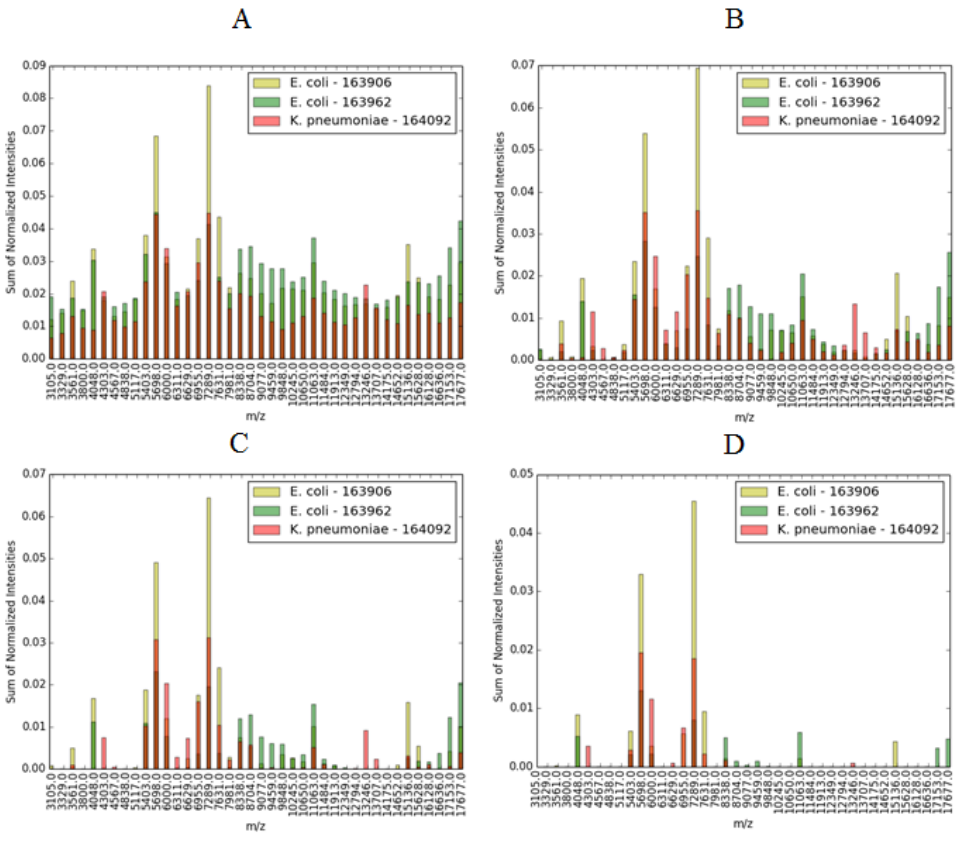


Figure 16 - Plot of the bucket intensity values, for a range of 750 individual intensity values in each data bucket (A). The *m/z* value corresponds range centroid. On B, C and D intensity values below the 0.10, 0.50 and 0.90 percentile were removed from the spectra.

Looking at these plots it is observable that removing data below 0.10 percentile highlights the differences between the different species. Moreover, removing the data below 0.50 percentile seems to help in the species distinction. However, when the data below 0.90 percentile was removed, a considerable amount of information is lost. Taking this into account, it seems that for this bucket size the optimum percentile to use is this time, 0.50 percentile, which is different than the previous plots, where better results were obtained using 0.90 percentile. This could be explained by the amount of information lost when combining the bucket approach and the removal of data below 0.90 percentile.

The last approach consisted of plotting the intensity values for two different strains against each other. If intensities were the same for two strains for the same *m/z*, the points would always appear along the diagonal line. This can be seen when a strain is plotted against itself, as shown in figure 17-A. To facilitate the analysis, the points were colored as a function of *m/z* ratio. The result of these plots can be seen in figure 17.

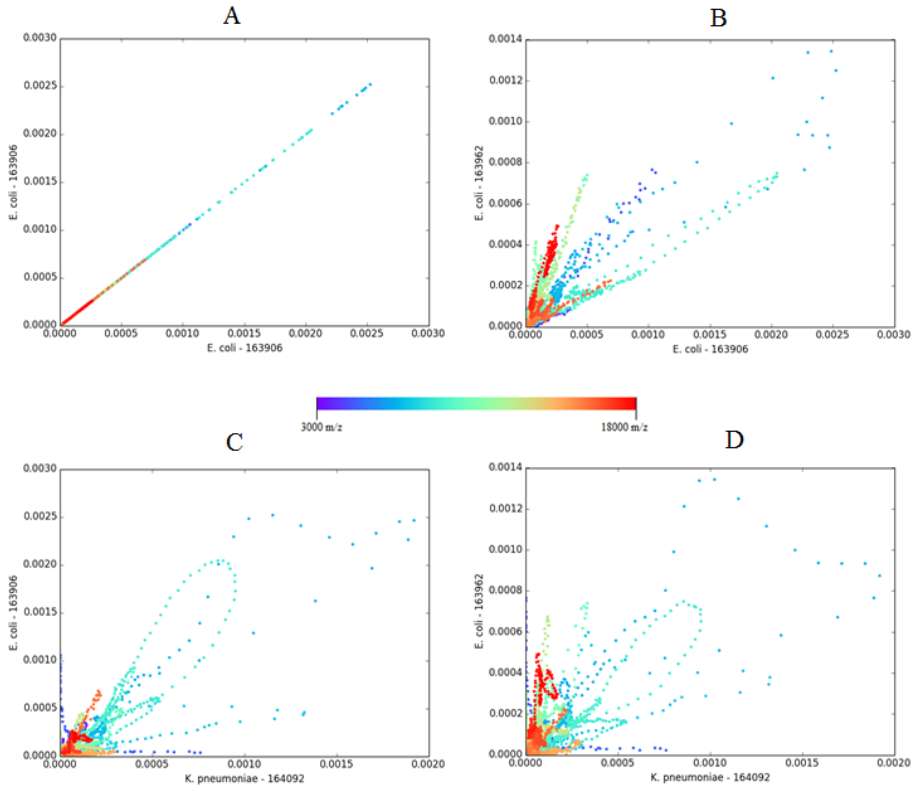


Figure 17 – Scatter plots of the: *E. coli* – 163906 intensities vs. itself (A); *E. coli* – 163906 intensities vs *E. coli* – 163962 (B); *K. pneumoniae* – 164092 intensities vs. *E. coli* – 163906 (C); *K. pneumoniae* – 164092 intensities vs. *E. coli* – 163962 (D).

Taking into account the results from this data pre-processing it is possible to conclude that the distinction of the bacteria based on their mass profiles has a high potential for the detection of markers that could help in the identification of unknown samples. Considering the results obtained, it was decided that the developed application would use a library-based approach (comparison of the unknown sample profile with the profiles of the database). In addition, the analysis of the different tests that were made showed that the comparison of full spectrum would provide the desired results for bacteria identification and that the removal of data below a certain threshold is a reliable tool for further distinguish the bacteria strains. With this in mind, an application was developed to accomplish the comparison of the mass profiles, offering the possibility to select a percentile value (threshold) from which it removes all the intensity values below that percentile. The metric selected for the comparison of spectra was the Pearson's correlation coefficient which returns a score based on similarity of signals (vectors).

3.3 Analysis of the bacteria samples

Considering that the developed application will use a library-based approach, it was required the creation of a database containing the mass profiles of known bacteria. The samples used to obtain the mass profiles were provided by the local hospital. Using these samples 353 high quality spectra were collected. These spectra correspond to twenty-three different species of bacteria and fifty-six different strains. Besides the information of the species the hospital provided the information of whether these bacteria were producers of ESBL and if they presented MDR. These information for each bacteria, along with the number of each individual strain spectra is provided in table 6.

Table 6 - Characteristics of the bacteria contained in the library based spectra database

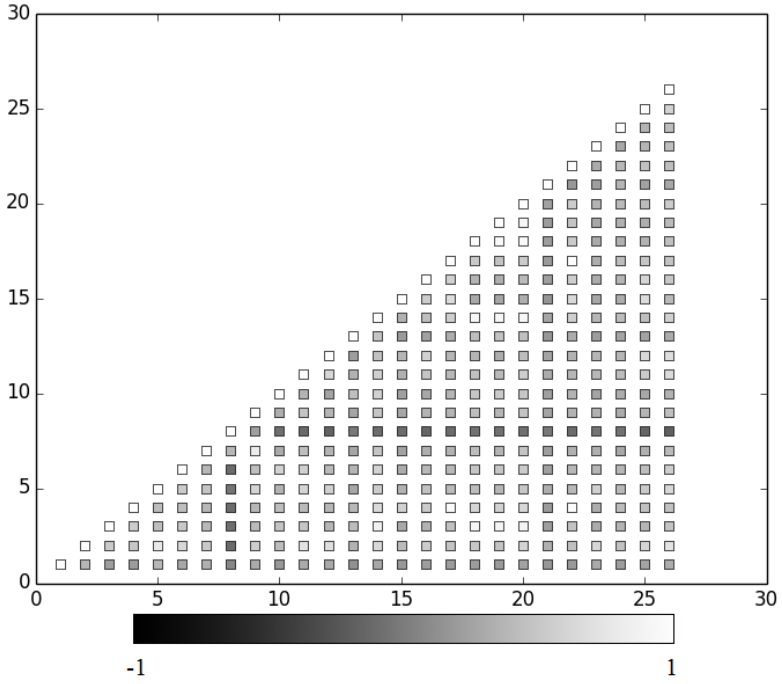
Lab. ID	Species	Gram	ESBL	MDR	Nº Spectra
519231	*Acinetobacter baumannii*	Negative	N	Y	2
533734	*Acinetobacter baumannii*	Negative	N	N	4
529317	*Citrobacter feundii*	Negative	N	N	4
550235-2	*Citrobacter morganii*	Negative	N	Y	4
519091	*Enterobacter aerogenes*	Negative	N	N	8
541243-1	*Enterobacter clocae*	Negative	N	N	6
153192	*Enterococcus faecalis*	Positive	N	N	14
182061	*Enterococcus faecalis*	Positive	N	N	4
312494	*Enterococcus faecalis*	Positive	N	Y	13

Lab. ID	Species	Gram	ESBL	MDR	Nº Spectra
518903	*Enterococcus faecalis*	Positive	N	N	10
519025	*Enterococcus faecalis*	Positive	N	N	13
519214	*Enterococcus faecalis*	Positive	N	N	3
154216	*Enterococcus faecium*	Positive	N	Y	13
518480	*Enterococcus faecium*	Positive	N	Y	12
520117	*Enterococcus faecium*	Positive	N	Y	11
163906	*Escherichia coli*	Negative	Y	Y	3
163962	*Escherichia coli*	Negative	Y	Y	1
198249	*Escherichia coli*	Negative	N	N	12
552338-1	*Escherichia coli*	Negative	N	N	6
518971	*Haemophilus influenza*	Negative	N	N	3
343349	*Haemophilus parainfluenzae*	Negative	N	N	4
520726	*Haemophilus parainfluenzae*	Negative	N	N	4
150499	*Klebsiella oxytoca*	Negative	N	N	1
164092	*Klebsiella pneumoniae*	Negative	Y	Y	3
552007	*Morganella morganii*	Negative	N	Y	6
152505	*Proteus mirabillis*	Negative	N	Y	3
523996	*Proteus mirabillis*	Negative	N	Y	2
505814	*Proteus vulgaris*	Negative	N	N	4
190894	*Providencia stuartii*	Negative	N	N	4
543246	*Providencia stuartii*	Negative	N	N	6
197648	*Pseudomonas aeruginosa*	Negative	N	N	12
154700	*Serratia liquefaciens*	Negative	N	N	1
364827	*Serratia marcescens*	Negative	N	N	4
542200	*Serratia marcescens*	Negative	N	N	8
542731	*Serratia marcescens*	Negative	N	N	4
188547	*Staphylococcus aureus*	Positive	N	Y	4
189021	*Staphylococcus aureus*	Positive	N	N	4
517900	*Staphylococcus aureus*	Positive	N	Y	14
518247	*Staphylococcus aureus*	Positive	N	Y	14
532357	*Staphylococcus aureus*	Positive	N	Y	5
534055	*Staphylococcus aureus*	Positive	N	Y	4
534216	*Staphylococcus aureus*	Positive	N	Y	4
534422	*Staphylococcus aureus*	Positive	N	Y	4
534448	*Staphylococcus aureus*	Positive	N	Y	4
535259	*Staphylococcus aureus*	Positive	N	Y	6
537177	*Staphylococcus aureus*	Positive	N	Y	4
538157	*Staphylococcus aureus*	Positive	N	Y	4
538716	*Staphylococcus aureus*	Positive	N	Y	4
539446-1	*Staphylococcus aureus*	Positive	N	Y	7
539521	*Staphylococcus aureus*	Positive	N	Y	4
540057	*Staphylococcus aureus*	Positive	N	Y	4
540838	*Staphylococcus aureus*	Positive	N	Y	4
541860	*Staphylococcus aureus*	Positive	N	N	12
198003	*Staphylococcus epidermidis*	Positive	N	N	12
541938	*Staphylococcus haemolyticus*	Positive	N	N	12
157098	*Streptococcus agalactiae*	Positive	N	N	8

While the hospital did not provided an identification to the strain level, each strain was identified by a different Lab. ID. Looking at the table it is possible to see some disparity related to the amount of collected spectra. This can be explained by the sample preparation on the hospital. Some of the samples were frozen prior to the protein extraction which resulted in spectra without the proper quality to be considered the mass profile of the corresponding bacteria; for those cases, the spectra were discarded.

3.3.1 Pearson's correlation coefficient of the entries of the built database

Before comparing samples with the built library-based database, it is required a complete differentiation between the entries in the database. Thus, the Pearson's correlation coefficient of the spectra for every possible pairs of strains was computed. In order to ensure the readability of the map, only the spectra of 26 samples were used. The results of this test are provided on figure 18.



Number	Lab. ID	Species
1	150499	Klebsiella oxytoca
2	152505	Proteus mirabilis
3	153192	Enterococcus faecalis
4	154216	Enterococcus faecium
5	154700	Serratia liquefaciens
6	157098	Streptococcus agalactiae
7	163906	Escherichia coli
8	163962	Escherichia coli
9	164092	Klebsiella pneumoniae
10	182061	Enterococcus faecalis
11	197648	Pseudomonas aeruginosa
12	198003	Staphylococcus epidermidis
13	198249	Escherichia coli
14	312494	Enterococcus faecalis
15	517900	Staphylococcus aureus
16	518247	Staphylococcus aureus
17	518480	Enterococcus faecium
18	518903	Enterococcus faecalis
19	519025	Enterococcus faecalis
20	519214	Enterococcus faecalis
21	519231	Acinetobacter baumannii
22	520117	Enterococcus faecium
23	523996	Proteus mirabilis
24	533734	Acinetobacter baumannii
25	541860	Staphylococcus aureus
26	541938	Staphylococcus haemolyticus

Figure 18 – Correlation coefficients map of the entries in the bacteria database. The correlations values are represented in a grayscale with a value of correlation equal to -1 (the minimun value possible, where there is an inverse correlation) corresponding to black and the value 1 (maximum correlation, totally overlap) corresponding to white. The table gives the corresponding bacteria to the x and y coordinates of the plot.

Looking at figure 18, it is possible to observe a considerable difference among all the entries in the database. A higher correlation is observed when the entries of database are closely related. For example the correlation of the third entry with entries eighteen, nineteen and twenty, which are all strains of *Enterococcus faecalis*. However, the correlation of the third entry against the eight entry, which corresponds to the correlation of a strain of *Enterococcus faecalis* and a strain of *Escherichia coli*, the color of the plot is close to black, which indicates an extremely low value of correlation. When the remaining data is compared, the same behavior is observed, closer entries have high correlation score, whereas lower correlation scores are obtained for non-related species. The results of this test show that the Pearson's correlation provides a good measurement of similarities in mass profiles. Thus, it is expected that the comparison of a sample against the database should provide a correct identification of the bacteria when there is an entry on the database of the corresponding bacteria.

3.3.2 Principal Components Analysis of the database

A different analysis that can be done to the database to assess the variability of the spectra is the Principal Components Analysis (PCA). PCA is one of the most used tools in exploratory data analysis and can be described as a way of identifying/detect patterns in data and expressing that data to enhance the visualization of similarities and differences in a dataset. This is achieved by computing a number of principle components (PC) from the data and expressing the data as a function of those principal components (61). The plot of that data should show which samples are close to each other and which are different.

A PCA was performed using the spectra obtained from the twenty-six of the collected samples. For this PCA only the first three principal components were used and to ensure the readability of the plots only twenty-six strains were used (the same samples used for the Pearson's correlation coefficients map). The results of this PCA can be seen on figure 19.

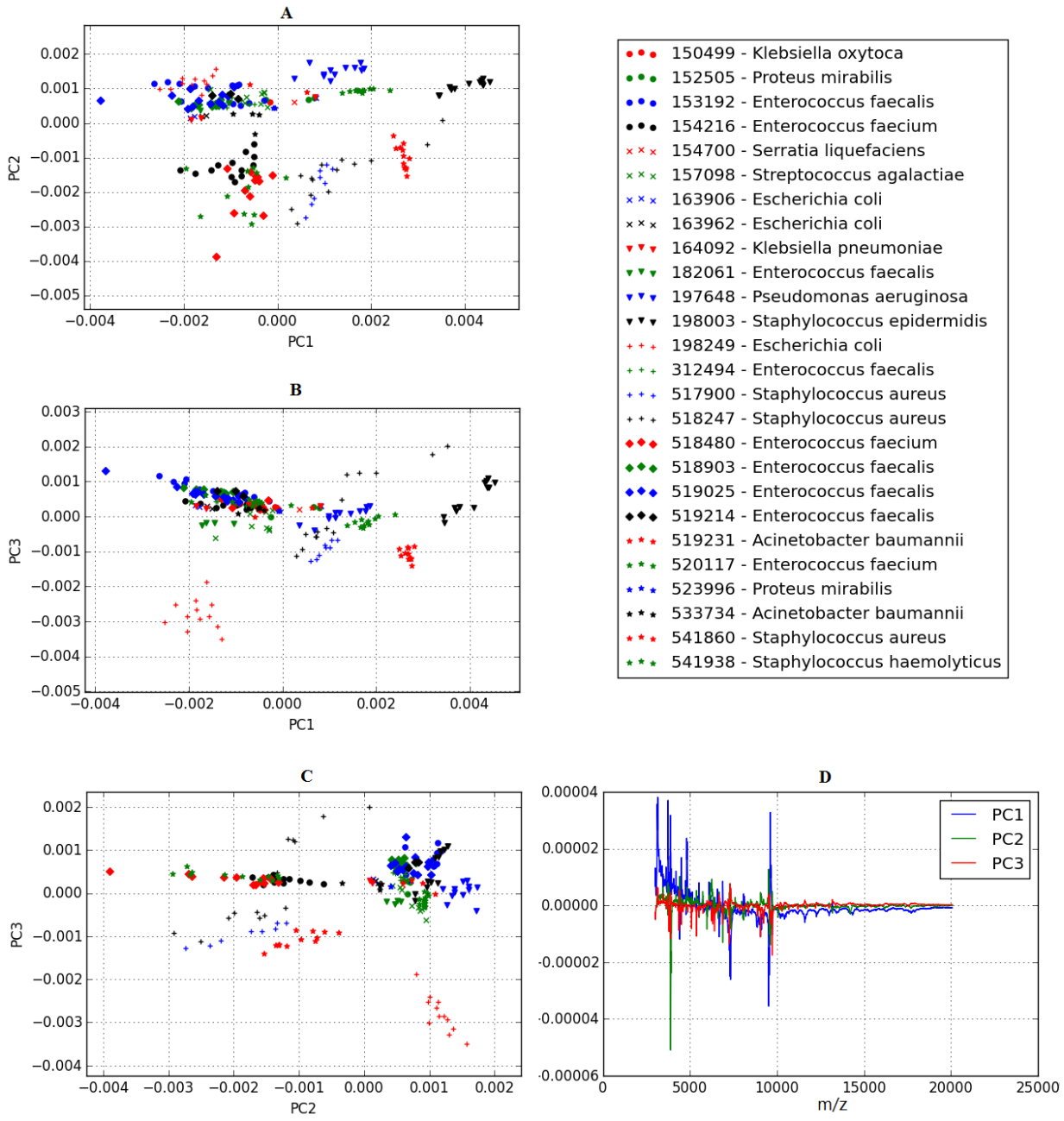


Figure 19 – Scores scatter plots: PC1 vs. PC2 (A); PC1 vs. PC3 (B); PC2 vs. PC3 (C); Loadings profile plot of the first three principal components (D).

Looking at the results given by the PCA it is possible to observe a differentiation for the different strains. While in some of the plots some strains overlap (showing no differences), looking at the other plots of the projected into other principal components it is possible to observe that they no longer overlap, which shows that the strains have characteristic mass profiles. The results also shows that the most overlapping strains correspond to strains of the same species, for example, the strains of *Enterococcus*

faecalis, which is expected considering how closely related they are. However, it should be possible to enhance the differentiation by using a higher number of principal components.

PCA can also be used to confirm identifications given by the library-based approaches. By analyzing the PCA loadings it is possible to observe which *m/z* ratios have a higher contribution on the samples differentiation. Considering the spectra were acquired from protein extracts of bacterial samples, those *m/z* ratios should correspond to specific proteins which could be characteristic of certain strains and thus could be used to confirm one identification. For example, looking at figure 19-B it is possible to see that the strain *Escherichia coli* – 198249 is clearly separated in the negative quadrant of PC1 and PC3. This should mean that the proteins causing the distinction of this strain should correspond to overlapping negative peaks of the loadings of PC1 and PC2. A closer look to the loadings (figure 20) reveal a negative peak in the range of 7268 to 7273 *m/z*.

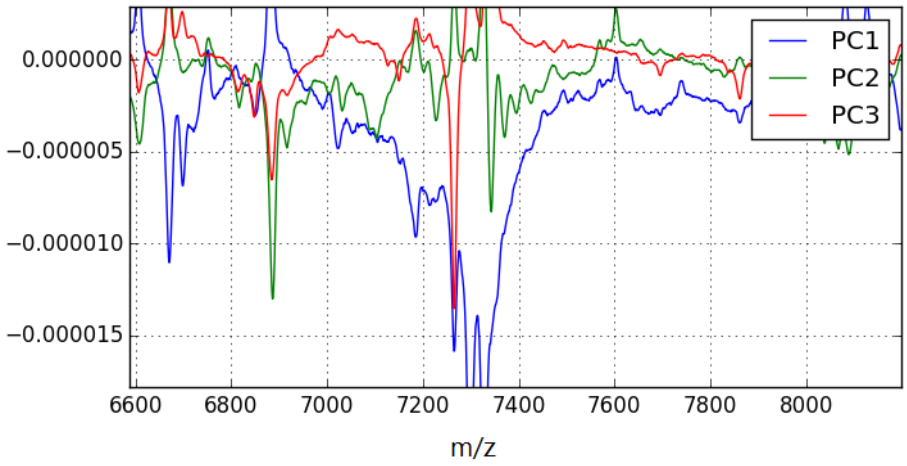


Figure 20 – Loadings profile of three PCs, in the *m/z* range of 6600 to 8200 *m/z*.

A search on UniprotKB database, a protein database, using the species *Escherichia coli* and the term ribosomal as the query of the search, reveals that there is a candidate protein with a mass of 7273 Da which should correspond to the negative peak of the loadings of PC1 and PC3. This protein is the ribosomal protein rpmC, 50S ribosomal protein L29 and is identified as being a protein that several strains of *Escherichia coli* produce. This type of analysis could be used to confirm possible

identifications by searching if the peaks responsible for the differentiation correspond to proteins produced by the strain proposed as the correct identification.

3.4 Bacteria Identification using the developed application

After the analysis of the collected bacterial samples and the analysis of the possible spectra treatment to enhance the differentiation of different bacteria profiles, development of the biotyping application started. The programming language selected was Python 2.7 and it was decided that the application would use a library-based approach. For the graphic user interface the framework selected was the Kivy framework. Using Python along with Kivy allows the application to run on every operative system, including mobile operative systems, such as Android and iOS. The method used for the identification of unknown samples is based on the Pearson's correlation coefficient, which indicates the similarity between two datasets.

The profile of an unknown sample is compared with every entry in the database and the hit with the higher correlation value should correspond to the correct species, in the case it is contained in the database. If the database does not contain an entry of the corresponding bacteria the correlation score should be low enough to be discarded without any doubt. With the application finished, the analysis of the results given by our application was performed.

3.4.1 General presentation of the developed application

A major component of this work was the development of the application and algorithms for the identification of the bacteria species. The initial screen of the developed application can be seen on figure 21.

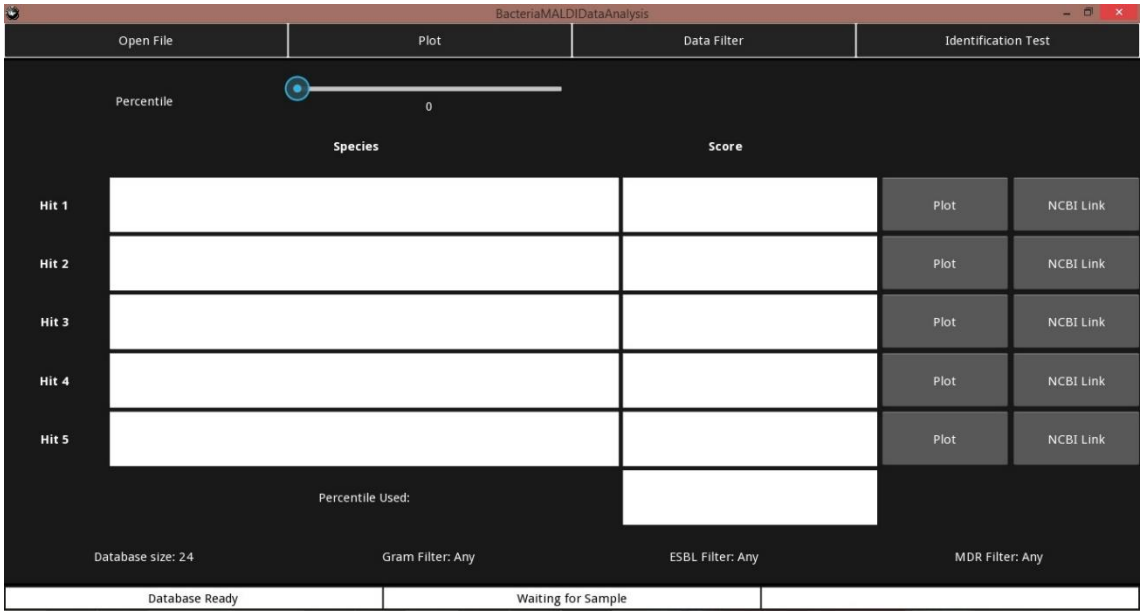


Figure 21 - Initial screen of the application, showing the table results and analysis options.

The application has a file loading option to load the sample mass spectrometry data, with the type of file used corresponding to the text file given by MALDI-TOF. The file should contain a column containing the *m/z* values and another column corresponding to the intensity values of those *m/z* values. An example can be seen on figure 22. The application is also prepared to load more than one file in case of replicas.

Figure 22 - Example of the file used as input in the application.

The application also contains a function to filter the database based on meta-data provided: Gram of bacteria, Multi-Drug Resistance and if they are producters of Extended-Spectrum Beta-Lactamases. The identification test button compares the sample data with the data contained in the database, giving a score (corresponding to the Pearson's correlation coefficient) for each entry of the database. The top five hits (entries of the database with top 5 scores) of this test are shown in a table on the screen. With the test finished, the application provides options to show the plot of both sample and corresponding hit and also has a button to open an internet browser page of the corresponding bacteria at NCBI page. An example of the screen with a finished test and the plot of the sample with one of the hits can be seen on figure 23.

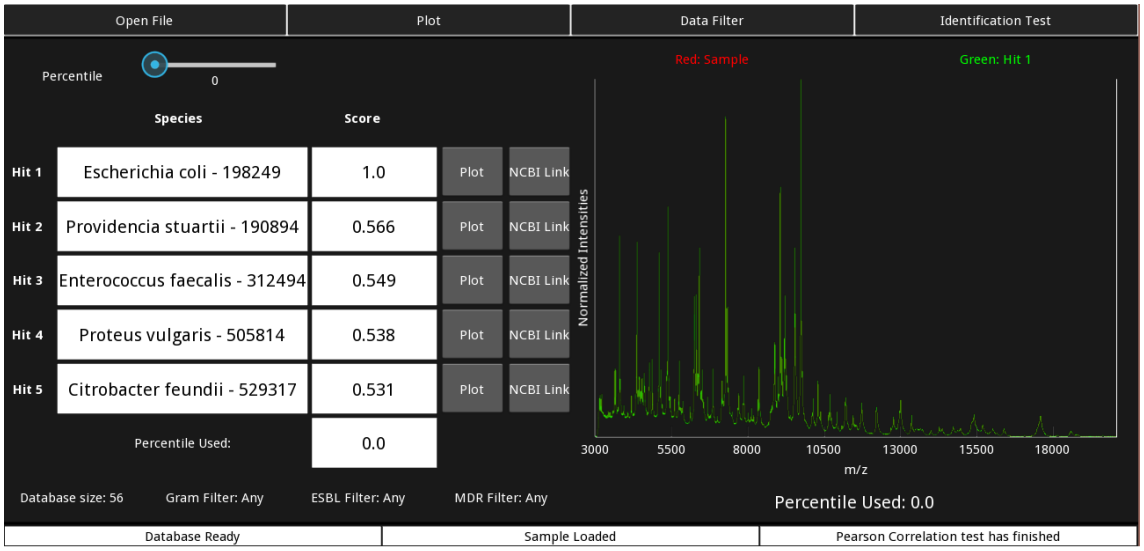


Figure 23 - Example of the screen after a finished test and the plot of the mass spectrum of the sample and the second hit.

Another option in application is data filter based on a percentile value. This option serves the purpose of removing all values below a given percentile value. Thus, the noise values of both samples and database entries are removed and only the higher intensity peaks are compared. While for the most part comparison of whole spectra is the ideal procedure, the isolation of higher intensity peaks can be useful to enhance the differentiation between close hits and provide a more reliable identification.

3.4.2 Test using a sample that is already in the database

The first test using the developed application consisted of testing the results using entries of the database as the "unknown" samples. Considering the previous results this should return a correlation score of 1 to the corresponding species. To perform this test three different species were selected, one *Escherichia coli*, one *Klebsiella pneumoniae* and one *Enterococcus faecium*. The results of these tests can be seen on figure 24.

A

Hit 1	Escherichia coli - 163962	1.0
Hit 2	Escherichia coli - 163906	0.430
Hit 3	Citrobacter freundii - 550235-2	0.258
Hit 4	Citrobacter feundii - 529317	0.254
Hit 5	Klebsiella pneumoniae - 164092	0.250
	Percentile Used:	0.0

B

Hit 1	Klebsiella pneumoniae - 164092	1.0
Hit 2	Escherichia coli - 163906	0.848
Hit 3	Citrobacter feundii - 529317	0.708
Hit 4	Enterobacter cloacae - 541243-1	0.697
Hit 5	Serratia liquefaciens - 154700	0.660
	Percentile Used:	0.0

C

Hit 1	Enterococcus faecium - 154216	1.0
Hit 2	Enterococcus faecium - 520117	0.965
Hit 3	Enterococcus faecium - 518480	0.962
Hit 4	Enterococcus faecalis - 312494	0.712
Hit 5	Enterococcus faecalis - 519214	0.683
	Percentile Used:	0.0

Figure 24 - Results given by the application when loading the entry of the database *Escherichia coli* – 163962 (A), *Klebsiella pneumoniae* – 164092 (B) and *Enterococcus faecium* – 154216 (C) as the sample.

The results seen in the figure 24-A shows that the application is working as expected. The entry of the database loaded as the sample returns a score of 1.0, which corresponds to a perfect overlap of the values. However, looking at the rest of hits it is possible to see that other two strains of *Escherichia coli* contained in the database are also given as positive identifications. However, the scores for other hits are considerable lower than what would be expected for a positive identification, considering how closely related they are in terms of taxonomy.

Analyzing results of the figure24-B it shows that the application returned a score of 1.0 to the entry of database loaded as the sample. The second hit corresponds to a different species. However the score of this hit could be considered high enough to rise some doubts about the proper the identification. This could possibly be resolved by the utilization of percentile filter as seen later on this work.

The figure 24-C shows results when it was loaded one of the strains of *Enterococcus faecium* contained in database. Once again the corresponding entry has a score of 1.0. Looking at the other hits it is observable that hit #2 and hit #3 have values close to 1.0 and since they correspond to different strains of same species as the loaded entry these values are acceptable. However, if the loaded data did not correspond to any entry of database it might be needed higher differentiation. This could possibly be achieved by using the percentile filter option as shown later on this work. Looking at the other hits, #4 and #5, it is possible to see high correlation values (0.712 and 0.683) which corresponds to bacteria of the same genus. Considering these bacteria are closely related to the loaded entry, the scores obtained are in accordance to what was expected.

The results obtained in this test show that the developed application is capable of identifying matches between sample and entries of database. Considering that entries of database were used as testing samples, the first hit always corresponded to the correct identification with a score of 1.0 (maximum score possible). However looking at the other hits it shows that application is also capable of identifying other strains of the same species. So when database does not contain any entry of same strain as the sample, it should be able to achieve a reliable identification to species-level.

3.4.3 Blind test using a sample with an entry of the database corresponding to the same strain

The next test performed consisted of using samples that had entries on database of the same strain. In order to do this test, few biological replicas of some entries on database were removed and used as sample to test. Then, it was evaluated if results given were the correct bacteria identification. This procedure is important to ensure that the application gives a confident identification. The samples consisted of three different strains of bacteria: one of *Enterococcus faecalis* - 153192, one of *Pseudomonas*

aeruginosa - 197648 and one of *Staphylococcus epidermis* - 198003. The results of this test can be seen in the figure 25.

A

Hit 1	Enterococcus faecalis - 153192	0.972
Hit 2	Enterococcus faecalis - 519025	0.888
Hit 3	Enterococcus faecalis - 518903	0.863
Hit 4	Enterococcus faecalis - 519214	0.854
Hit 5	Enterococcus faecalis - 312494	0.814
	Percentile Used:	0.0

B

Hit 1	Pseudomonas aeruginosa - 197648	0.987
Hit 2	Proteus mirabilis - 152505	0.769
Hit 3	Haemophilus parainfluenzae - 343349	0.766
Hit 4	Proteus vulgaris - 505814	0.756
Hit 5	Haemophilus influenzae - 518971	0.739
	Percentile Used:	0.0

C

Hit 1	Staphylococcus epidermidis - 198003	0.976
Hit 2	Staphylococcus aureus - 538716	0.751
Hit 3	Staphylococcus haemolyticus - 541938	0.730
Hit 4	Proteus mirabilis - 152505	0.699
Hit 5	Haemophilus parainfluenzae - 343349	0.691
	Percentile Used:	0.0

Figure 25 - Results given by the application when removing a biological replica of the database of *Enterococcus faecalis* – 153192 (A), *Pseudomonas aeruginosa* - 197648 (B) and *Staphylococcus epidermis* – 198003 (C) and loading it as the sample.

The figure 25-A corresponds to the test when the loaded sample was *Enterococcus faecalis* – 153192. Looking at the given results it is possible to observe that hit #1 corresponds to the loaded sample with a score of 0.972. The rest of hits correspond to four other strains of *Enterococcus faecalis* contained in database. The scores for these hits are comprised between 0.888 and 0.814 which are lower than the score given to the correct strain. This indicates that even with closely related samples, as in this case, the application should be able to differentiate strains of same species. However, in case of the corresponding strain was not comprised in database it should be able to identify the bacteria to the species level if there is an entry of the same species in the database.

The figure 25-B corresponds to the results for the test using *Pseudomonas aeruginosa* - 197648 as the loaded sample. The hit #1 corresponded to the correct identification with a score of 0.987. Looking at the other hits it is possible to see that the

score for the first hit is considerable higher than the others, with hit #2 having a value of 0.769. Analyzing the entries in database, it is possible to see that there are no other entries of the same strain, species or even genus. This should explain the lower values of hits #2 to #5. These results indicates that the algorithm used to compare the sample with the database entries provide accurate results with a low rate of false positives.

The test shown on figure 25-C was made using one of the biological replicas of *Staphylococcus epidermis* – 198003. Once again, the score obtained for hit #1 is close to 1, which corresponds to the correct identification. The scores for the rest of the hits were all below 0.751 which usually is not considered high enough to provide a reliable identification. Looking at the identification proposed on hit #3 and #5 it is observable that they correspond to bacteria of the same genus: *Staphylococcus haemolyticus* and *Staphylococcus aureus*, respectively. However, scores for these hits are low enough to be discarded.

The results of this test indicate that when there is an entry of the database of the same strain of the sample, the application should be able to achieve a correct identification. In the three examples shown in the figure 25 the correct identification had scores of at least 0.972, while the rest of the hits had considerable lower values with the exception being made to the case of *Enterococcus faecalis*. However these scores can be explained by the existence of entries in the database of different strains of the same species as our testing sample.

3.4.4 Blind test using a sample with no entry on the database corresponding to the same strain

The last test performed consisted of testing the application using a sample with no entry on the database corresponding to the sample strain. The goal of this test is to verify if the application is capable of providing an identification to the species level, even in the case the database does not contain an entry of the same strain. Also, this test should provide some insight in case of absence of one entry in database of the same species, the scores are low enough to the hits be discarded as possible identifications. For this test the removed samples from the database were a strain of *Klebsiella oxytoca*, a strain of *Serratia liquefaciens* and a strain of *Enterococcus faecalis*. The results of this test can be seen on figure 26.

A

Hit 1	Enterobacter cloacae - 541243-1	0.459
Hit 2	Staphylococcus aureus - 539446-1	0.444
Hit 3	Proteus mirabilis - 152505	0.432
Hit 4	Serratia liquefaciens - 154700	0.427
Hit 5	Haemophilus parainfluenzae - 343349	0.417
	Percentile Used:	0.0

B

Hit 1	Proteus mirabilis - 152505	0.821
Hit 2	Haemophilus parainfluenzae - 343349	0.766
Hit 3	Haemophilus influenzae - 518971	0.751
Hit 4	Staphylococcus aureus - 539446-1	0.727
Hit 5	Serratia marcescens - 364827	0.703
	Percentile Used:	0.0

C

Hit 1	Enterococcus faecalis - 518903	0.924
Hit 2	Enterococcus faecalis - 519025	0.917
Hit 3	Enterococcus faecalis - 519214	0.910
Hit 4	Enterococcus faecalis - 312494	0.892
Hit 5	Proteus vulgaris - 505814	0.658
	Percentile Used:	0.0

Figure 26 - Results given by the application when using *Klebsiella oxytoca* – 150499 (A), *Serratia liquefaciens* - 154700 (B) and *Enterococcus faecalis* - 153192 (C) as our sample without having an entry on the database of the same strain.

Looking at figure 26-A it is possible to see the results given by the developed application when *Klebsiella oxytoca* – 150499 is the loaded sample without having an entry of the database corresponding to this strain. All of the five hits given by the application have considerable low scores which are not enough to be considered as possible identifications. Considering that there is no entry of the database for the same strain or even for the same species, these results correspond to what was expected.

When *Serratia liquefaciens* – 154700 was used as the sample (figure 26-B) the scores obtained were higher than when used *Klebsiella oxytoca* – 150499. However, with the exception of the first hit, they are low enough to be discarded as possible identifications. The hit #1 has a score value of 0.821 and while it's considerable lower to the scores obtained for correct identifications it could raise some doubts to the user. In this case the usage of the percentile filter could probably help to enhance the results and show more clearly that *Proteus mirabillis* was not the correct identification.

The last sample used for this test was *Enterococcus faecalis* – 153192 (Figure 26-C). In this case, the hits #1 to #4 corresponded to bacteria of the same species of the sample, with score ranging from 0.924 (hit #1) to 0.892 (hit #4). Hit #5 was a bacteria of

the same genus as the sample, however the score value (0.604) was lower than the scores obtained for the correct species.

The results obtained in this test seem to indicate that when the database does not contain an entry of the same species, the score values are low enough to be discarded which should lower the rate of providing false positive results. However, when the database contains entries of the same species, the top hits correspond to strains of the same species of the sample. With this, it should be possible to achieve an identification to the species level even if the database does not contain an entry of the same strain as the sample.

3.4.5 Usage of a percentile value to enhance the results

The developed application works by comparing the full spectra of the sample against a database containing spectra of different bacteria. As seen in prior tests, this approach is usually able to successfully identify the bacterial sample to the species level and even to the strain level. However, sometimes the obtained scores of different hits are close enough to raise doubts on what would be the correct identification, so an option to enhance the results could be necessary. To fulfill this need the application contains an option to filter the mass spectrometry data based on a percentile value selected by the user. This works by calculating the intensity value corresponding to the selected percentile and removing the data below that percentile. Doing this removes the lower values of intensities, which results in an isolation of the higher intensity peaks. This is often useful to obtain a more reliable identification. An example of the usage of the percentile filter can be seen on figure 27.

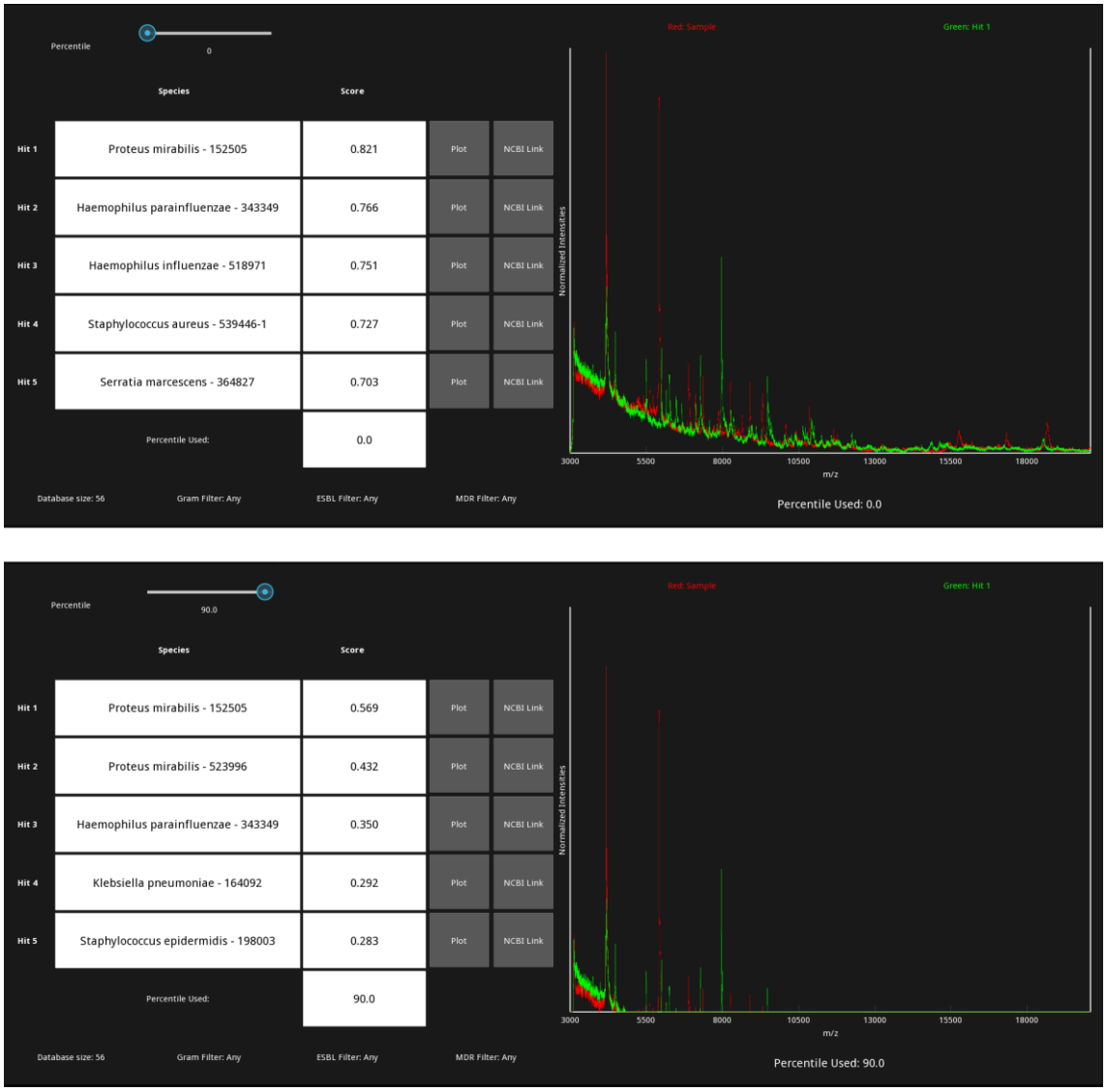


Figure 27 - Example of the usage of the percentile filter option with 0.90 percentile, when using *Serratia liquefaciens* - 154700 as our sample without having an entry on the database of the same strain and the comparison with no filter.

Looking at figure 27 it is possible to see that without any percentile filter, the application returns a score that could raise doubts about it being the correct identification. While the sample used corresponded to *Serratia liquefaciens* – 154700, the application gave a score of 0.821 to *Proteus mirabillis* – 152505. However with the use of the percentile filter option, the score value for that hit gets lower (0.569) which is lower enough to be discarded as a possible identification. Also, looking at the plots it is possible to observe that after the percentile filter, the sample and hit #1 have many different high intensity peaks, which confirms that *Proteus mirabillis* does not correspond to the correct

identification. This option could also help when the application returns two scores with close values. In this case it is expected that the usage of this option will keep the correct hit score close to the original score and lower the score of the incorrect hit. An example of this can be seen on figure 28.

Figure 28 - Example of the usage of the percentile filter option with 0.90 percentile, when using *Enterococcus faecium* - 154216 as our sample and the comparison with no filter. The plots correspond to the sample and hit #2 to show how percentile filter enhances the results.

In this case, using one of the biological replicas of *Enterococcus faecium* – 154216 with no percentile filter the application gives three strains of *Enterococcus faecium* as the first three hits. While the score for the correct hit is higher than the scores of hit #2 and

#3, the identification of the correct strain could raise some questions. However, using a 0.90 percentile the score for the correct strain maintains the score value, while the hit #2 and #3 scores are lower, which indicates that the hit #1 correspond to correct strain. This can be explained by looking at the plots, which correspond to the sample versus hit #2. When applying the percentile filter the higher intensity peaks are isolated, resulting in a lower correlation between the sample and the incorrect hits. These results show that this option can be used to avoid possible false possible and to provide a more accurate result.

3.4.6 Usage of collected meta-data to enhance the results

Another option the developed application provides is the filter of the database using the meta-data provided by the hospital. By doing this the chances of getting incorrect identifications should be lower. The meta-data provided by the hospital consisted on the Gram of the bacteria, if the bacteria showed MDR and if the bacteria was producer of ESBL. An example of the usage of the data-filter option can be seen on figure 29.

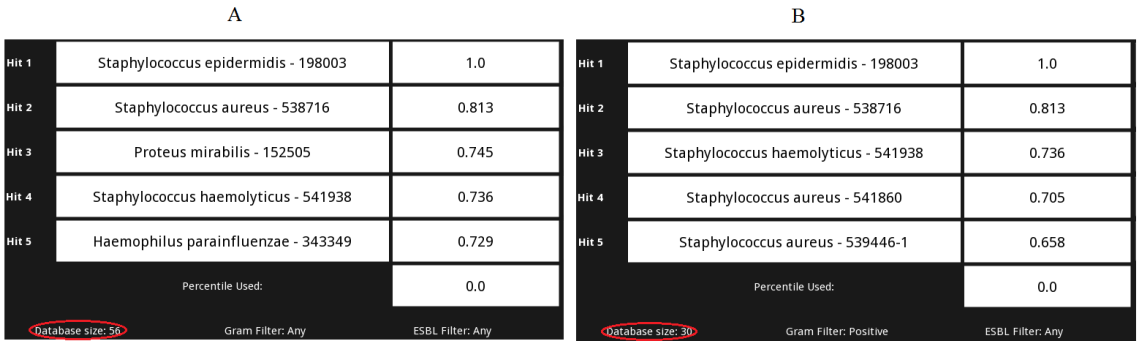


Figure 29 - Example of the usage of the database filter option, showing the results of an identification test without the database filter (A) and after filtering the database to contain only Gram-positive bacteria (B).

Looking at the figure 29 it is possible to see that on the first test (A), without the database filter, the hit #3 corresponds to a strain of *Proteus mirabillis*, a Gram-negative bacteria, which is not possible since the sample used was a strain of *Staphylococcus epidermis*, a Gram-positive bacteria. Applying the database filter, all the possible identifications correspond to Gram-positive bacteria. In this specific case, the application

returned a score of 1.0 so it is certain that the hit #1 was the correct identification. However, there might be cases of having closer score values for hit #1 and #2, which could lead to doubts on what was the correct identification. If those hits had different meta-data filtering the database with the meta-data corresponding to the sample would eliminate the incorrect identification and allowing a reliable identification of our sample. This function of the application, along with the percentile filter, should provide the user with the needed tools to enhance the results given by the application and achieve a reliable identification.

3.4.7 Evaluation of the obtained results

Using the developed application it was possible to properly identify bacterial samples based on their mass spectra. The results obtained so far showed that the algorithm used seems to properly identify bacteria at the strain-level if the database contains an entry of the same strains as the sample. In this case the score should be a value in the interval of 0.95 to 1.0. If database does not contain any entry for the same strain but contains one entry of the same species, it should be able to achieve an identification to the species level, with the scores value ranging from 0.90 to 0.95. However a few considerations should be taken. In order to evaluate the effectiveness of the application a larger database is required along with a larger pool of samples. While the results obtained until now are promising, further tests should be performed.

When comparing results obtained with this application with those already available alternatives some points can be highlighted: during these tests, the developed application seem to return scores low enough to be discarded as possible identifications when there are no entries on database corresponding to the same species. This should result in a minimal rate of false-positives. Another drawback is related to the identification of Gram-positive bacteria, showing a lower rate of correct identification when compared to the rate of correct identifications of Gram-negative bacteria. However, the current application was able to correctly identify both Gram-positive and Gram-negative bacteria without encountering any incorrect identification.

The current application also offers the user some options to enhance the results and achieve a reliable identification. Both the use of percentile filter and meta-data filter seems to enhance the differentiation of correct identification and the other given hits.

These options should allow the user to be able to identify a bacterial sample even when the initial scores raise some doubts. A possible addition to the application could be the matching of the higher intensity peaks to known proteins of the proposed bacteria identification to confirm the results given by the current analysis.

4. Conclusions

During this work, the extraction of protein fraction of bacterial samples was successful along with the acquisition of mass profiles for those bacteria using MALDI-TOF mass spectrometry. With those profiles a database was created with the purpose of comparing it with unknown samples to provide a correct bacteria identification. For this process a proposed application was developed with an algorithm to compare the mass spectrometry data of the unknown samples with the data contained in our database.

Using developed application, the identification of bacterial samples based on their mass profiles has clearly show its high potential. The identification can be made to the strain level if the database contains one entry for the same strain as sample. In the case the database does not contain an entry of the same species of the sample the scores should be low enough to consider that it was not possible to achieve a reliable identification. This should be enough to avoid false-positive results, which may provide incorrect identifications. Another problem with the current alternatives is the identification of Gram-positive bacteria, having a lower rate of success when compared to the identification of Gram-negative bacteria. The developed application showed the same results for both Gram-positive and Gram-negative bacteria.

In the future, more samples should be added to database to allow identification of more different strains of bacteria and an exact evaluation of the false-positive rate should be made.

5. References

1. Alter SJ, Vidwan NK, Sobande PO, Omoloja A, Bennett JS. Common Childhood Bacterial Infections. Curr Probl Pediatr Adolesc Health Care. 2011;41(10):256–83.

2. Chen Y-J, Hsieh Y-C, Huang Y-C, Chiu C-H. Clinical manifestations and microbiology of acute otitis media with spontaneous otorrhea in children. J Microbiol Immunol Infect. 2013;46(5):382–8.

3. Spicknall IH, Foxman B, Marrs CF, Eisenberg JNS. A modeling framework for the evolution and spread of antibiotic resistance: literature review and model categorization. Am J Epidemiol. 2013 Aug 15;178(4):508–20.

4. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen K-Y. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect. Blackwell Publishing Ltd; 2008;14(10):908–34.

5. Sandt C, Madoulet C, Kohler a, Allouch P, De Champs C, Manfait M, et al. FT-IR microspectroscopy for early identification of some clinically relevant pathogens. J Appl Microbiol. 2006 Oct;101(4):785–97.

6. Van Belkum A, Durand G, Peyret M, Chatellier S, Zambardi G, Schrenzel J, et al. Rapid clinical bacteriology and its future impact. Ann Lab Med. 2013 Jan;33(1):14–27.

7. Tang Y, Ellis NM, Hopkins MK, Smith H, Dodge DE, Persing DH, et al. Comparison of Phenotypic and Genotypic Techniques for Identification of Unusual Aerobic Pathogenic Gram-Negative Bacilli Comparison of Phenotypic and Genotypic Techniques for Identification of Unusual Aerobic Pathogenic Gram-Negative Bacilli. J Clin Microbiol. 1998;36(12):3674–9.

8. Abel K, Deschmertzing H, Peterson JI. Classification of Microorganisms By Analysis of Chemical Composition. I. Feasibility of Utilizing Gas Chromatography. J Bacteriol. 1963 May;85:1039–44.

9. Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001 Jan;25(1):39–67.

10. Holmes B, Costas M, Ganner M. Evaluation of Biolog system for identification of some gram-negative bacteria of clinical importance. J Clin Microbiol. 1994 Aug;32(8):1970–5.

11. http://www.techno-path.co.uk/_EN_/biotechnologies/pharmaceutical/rapid-microbial-identification.aspx#.UqIbPPRdU08.

12. Klingler JM, Stowe RP, Obenhuber DC, Groves TO, Mishra SK, Pierson DL. Evaluation of the Biolog automated microbial identification system. Appl Environ Microbiol. 1992 Jun;58(6):2089–92.

13. Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: Small subunit ribosomal {RNA} sequence analysis and beyond. Microbiol Res. 2011;166(2):99–110.

14. Janda J, Abbott S. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007;45(9):2761–4.

15. Tang YW, Von Graevenitz a, Waddington MG, Hopkins MK, Smith DH, Li H, et al. Identification of coryneform bacterial isolates by ribosomal DNA sequence analysis. J Clin Microbiol. 2000 Apr;38(4):1676–8.

16. Patel JB, Leonard DG, Pan X, Musser JM, Berman RE, Nachamkin I. Sequence-based identification of Mycobacterium species using the MicroSeq 500 16S rDNA bacterial identification system. J Clin Microbiol. 2000 Jan;38(1):246–51.

17. Turenne C, Tschetter L. Necessity of Quality-Controlled 16S rRNA Gene Sequence Databases: Identifying Nontuberculous Mycobacterium Species. J Clin Microbiol. 2002;39(10):3637–48.

18. Sibley CD, Peirano G, Church DL. Molecular methods for pathogen and microbial community detection and characterization: current and potential application in diagnostic microbiology. Infect Genet Evol. Elsevier B.V.; 2012 Apr;12(3):505–21.

19. Bottero MT, Dalmasso A. Animal species identification in food products: evolution of biomolecular methods. Vet J. Elsevier Ltd; 2011 Oct;190(1):34–8.

20. Tu O, Knott T, Marsh M, Bechtol K, Harris D, Barker D, et al. The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA. Nucleic acids research. 1998. p. 2797–802.

21. Beekes M, Lasch P, Naumann D. Analytical applications of Fourier transform-infrared (FT-IR) spectroscopy in microbiology and prion research. Vet Microbiol. 2007;123(4):305–19.

22. Kirschner C, Maquelin K, Pina P. Classification and identification of enterococci: a comparative phenotypic, genotypic, and vibrational spectroscopic study. J Clin Microbiol. 2001;39(5):1763–70.

23. Krásný L, Hynek R, Hochel I. Identification of bacteria using mass spectrometry techniques. Int J Mass Spectrom. 2013;(0):-.

24. Cash P. Proteomics in medical microbiology. Electrophoresis. Wiley Subscription Services, Inc., A Wiley Company; 2000;21(6):1187–201.

25. Chaerkady R, Pandey A. Applications of Proteomics to Lab Diagnosis. Annu Rev Pathol Mech Dis. 2008;3(1):485–98.

26. Sutandy FXR, Qian J, Chen C-S, Zhu H. Overview of protein microarrays. Curr Protoc Protein Sci. 2013 Apr;Chapter 27(April):Unit 27.1.

27. Wingren C, Borrebaeck C. Protein microarray technologies for detection and identification of bacterial and protein analytes. Princ Bact Detect Biosens …. 2008;715–29.

28. Dufva M, Christensen CB V. Diagnostic and analytical applications of protein microarrays. Expert Rev Proteomics. 2005 Jan;2(1):41–8.

29. Taitt C, Shubin Y. Detection of Salmonella enterica serovar typhimurium by using a rapid, array-based immunosensor. Appl …. 2004;70(1):152–8.

30. Howell SW, Inerowicz HD, Regnier FE, Reifenberger R. Patterned Protein Microarrays for Bacterial Detection. 2003;(21):436–9.

31. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009 Jan;11:49–79.

32. Ho CS, Lam CWK, Chan MHM, Cheung RCK, Law LK, Lit LCW, et al. Electrospray ionisation mass spectrometry: principles and clinical applications. Clin Biochem Rev. 2003 Jan;24(1):3–12.

33. Ho C, Lam C. Electrospray ionisation mass spectrometry: principles and clinical applications. Clin …. 2003 Jan;24(1):3–12.

34. Intelicato-Young J, Fox A. Mass spectrometry and tandem mass spectrometry characterization of protein patterns, protein markers and whole proteomes for pathogenic bacteria. J Microbiol Methods. 2013 Mar;92(3):381–6.

35. Tracz D, McCorrister S, Chong P. A simple shotgun proteomics method for rapid bacterial identification. J Microbiol Methods. 2013;

36. Armengaud J. Microbiology and proteomics, getting the best of both worlds! Environ Microbiol. 2013 Jan;15(1):12–23.

37. Lartigue M-F. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterization. Infect Genet Evol. 2013 Jan;13(null):230–5.

38. Fenselau C, Demirev P a. Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrom Rev. 2002;20(4):157–71.

39. Krishnamurthy T, Ross PL, Rajamani U. Detection of pathogenic and non-pathogenic bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom. 1996 Jan;10(8):883–8.

40. Edwards-Jones V, Claydon M a, Evason DJ, Walker J, Fox a J, Gordon DB. Rapid discrimination between methicillin-sensitive and methicillin-resistant Staphylococcus aureus by intact cell mass spectrometry. J Med Microbiol. 2000 Mar;49(3):295–300.

41. Van Veen SQ, Claas ECJ, Kuijper EJ. High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. J Clin Microbiol. 2010 Mar;48(3):900–7.

42. Biswas S, Rolain J-M. Use of MALDI-TOF mass spectrometry for identification of bacteria that are difficult to culture. J Microbiol Methods. 2013 Jan;92(1):14–24.

43. Pineda FJ, Antoine MD, Demirev P a, Feldman AB, Jackman J, Longenecker M, et al. Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. Anal Chem. 2003 Aug 1;75(15):3817–22.

44. Ryzhov V, Fenselau C. Characterization of the protein subset desorbed by MALDI from whole bacterial cells. Anal Chem. 2001 Feb 15;73(4):746–50.

45. Momo R, Povey J, Smales C. MALDI-ToF mass spectrometry coupled with multivariate pattern recognition analysis for the rapid biomarker profiling of Escherichia coli in different growth phases. Anal …. 2013 Oct;405(25):8251–65.

46. Sandrin T. MALDI TOF MS profiling of bacteria at the strain level: a review. Mass Spectrom Rev. 2012;32(3):188–217.

47. Welker M. Proteomics for routine identification of microorganisms. Proteomics. 2011 Aug;11(15):3143–53.

48. Arnold RJ, Reilly JP. Fingerprint matching of E. coli strains with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of whole cells using a modified correlation approach. Rapid Commun Mass Spectrom. 1998 Jan;12(10):630–6.

49. Jarman KH, Cebula ST, Saenz a J, Petersen CE, Valentine NB, Kingsley MT, et al. An algorithm for automated bacterial identification using matrix-assisted laser desorption/ionization mass spectrometry. Anal Chem. 2000 Mar 15;72(6):1217–23.

50. Bright J, Claydon M. Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. J Microbiol Methods. 2002;48:127–38.

51. Tao L, Yu X, Snyder a P, Li L. Bacterial identification by protein mass mapping combined with an experimentally derived protein mass database. Anal Chem. 2004 Nov 15;76(22):6609–17.

52. Hettick JM, Kashon ML, Slaven JE, Ma Y, Simpson JP, Siegel PD, et al. Discrimination of intact mycobacteria at the strain level: a combined MALDI-TOF MS and biostatistical analysis. Proteomics. 2006 Dec;6(24):6416–25.

53. Hsieh S-Y, Tseng C-L, Lee Y-S, Kuo A-J, Sun C-F, Lin Y-H, et al. Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. Mol Cell Proteomics. 2008 Feb;7(2):448–56.

54. Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain JM, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis. 2009 Aug 15;49(4):543–51.

55. Carbonnelle E, Grohs P, Jacquier H, Day N, Tenza S, Dewailly A, et al. Robustness of two MALDI-TOF mass spectrometry systems for bacterial identification. J Microbiol Methods. Elsevier B.V.; 2012 May;89(2):133–6.

56. Martiny D, Busson L, Wybo I, El Haj RA, Dediste A, Vandenberg O. Comparison of the Microflex LT and Vitek MS systems for routine identification of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol. 2012 Apr;50(4):1313–25.

57. Sauer S, Kliem M. Mass spectrometry tools for the classification and identification of bacteria. Nat Rev Microbiol. Nature Publishing Group; 2010 Jan;8(1):74–82.

58. Bizzini a, Greub G. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. Clin Microbiol Infect. 2010 Nov;16(11):1614–9.

59. Justesen US, Skov MN, Knudsen E, Holt HM, Søgaard P, Justesen T. 16S rRNA gene sequencing in routine identification of anaerobic bacteria isolated from blood cultures. J Clin Microbiol. 2010 Mar;48(3):946–8.

60. Holland RD, Duffy CR, Rafii F, Sutherland JB, Heinze TM, Holder CL, et al. Identification of bacterial proteins observed in MALDI TOF mass spectra from whole cells. Anal Chem. 1999 Aug 1;71(15):3226–30.

61. Smith LI. A tutorial on Principal Components Analysis. Cornell Univ USA. 2002;