Examination of Potential Applicability of Fourier Transform Infrared (FTIR) Spectroscopy for Routine Identification of Pathogenic Bacteria and Fungi and for Human Saliva-Based Detection of Viral Infection

Xin Di Zhu

Department of Food Science and Agricultural Chemistry

McGill University, Montreal

2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

SHORT TITLE

Implementation of FTIR spectroscopy for identification of microbial pathogens and for detection of viral infection

Abstract

Rapid identification of microbial and viral pathogens in clinical, veterinary and food laboratory settings remain challenging despite the availability of molecular typing methods, given that substantial resources are required to employ these methods on a routine basis. Over the past decade, the potential utility of Fourier transform infrared (FTIR) spectroscopy as a cost-effective technique for identification of microbial pathogens following routine culture has gained attention, although the feasibility of its routine implementation has yet to be fully addressed. While this culture-based approach cannot be extended to viral pathogens, a very limited amount of research has been reported on the detection of specific viral infections by FTIR analysis of various types of specimens. This thesis addresses the demonstration of FTIR spectroscopy for the identification of bacteria of veterinary or food safety importance and fungi, as well as its potential use as a screening method for SARS-CoV-2 infection employing saliva specimens. For the identification of cow mastitis-related bacterial pathogens, four spectral databases were developed by acquiring the FTIR spectra of isolates (*n*=582) grown on two growth media {tryptic soy agar (TSA) and Columbia blood agar (CBA)} using two modes of spectral acquisition {attenuated total reflectance (ATR) and transflectance (TR)}. Applying the prediction models developed for each of the four databases to a test set (*n*=98) yielded a rate of correct identification ranging from 96.91 to 93.81% at species level. To assess the capacity of FTIR spectroscopy for the identification of microbial pathogens, two databases were developed from different growth media and combined with FTIR sampling methods to predict the test sets. The latter results yielded high identification accuracy percentages from 98.98 to 92.78% for the four test sets at species level. Database interchangeability constructed using two commercially FTIR instruments from two different vendors was also assessed. Using the same database (n=361), the Summit Pro (ThermoFisher Nicolet, WI) FTIR (n=138) and the Cary 630 (Agilent Technologies, CA) FTIR spectrometer (n=305) achieved 99.5% and 96.2% at genus level, and 95.8% and 79.9% at species level, respectively. With regards to the change of growth culture medium, FTIR identification results were not affected at the species level identification. Also, the use of different FTIR instrument did not influence the results of identification at genus level. The identification of fungi has been challenging for decades even using molecular methods. Hence, the identification accuracy of *Aspergillus* spp. by quantitative real time polymerase chain reaction (RT-qPCR) and Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) were compared to FTIR spectroscopy-based approach using an in-house built FTIR database that was enlarged with additional fungal strains for further validation studies. Correct identification rates of 71.30%, 52%, and 92.31% were obtained for RT-qPCR, MALDI-TOF MS, and FTIR spectroscopy, respectively. RT-qPCR could be more suitable for the identification of *Aspergillus nigri* and *Aspergillus terrei*; and MALDI-TOF MS could be a good choice for identifying *Aspergillus fumigatus* and *Aspergillus flavus*; while FTIR spectroscopy provided correct identification rate of 96.6% using the enlarged database. Finally, due to the ongoing pandemic of coronavirus SARS-CoV-2 disease (Covid-19), the diagnostic efficacy of FTIR spectroscopy as a screening method was evaluated using 940 heat-

inactivated saliva samples (418 RT-qPCR positive for Covid-19 and 522 RT-qPCR negative). Multivariate analysis methods including random forest, k-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM) were used to develop algorithms for COVID-19 screening based solely on changes in the infrared spectral profiles of the saliva specimens. An independent test set yielded sensitivity rates from 82.9% to 85.4, specificity rates from 82.4% to 86.3%, accuracy rates from 84.8% to 85.9%, and precision rates from 79.6 to 83.4% from the KNN, ANN and SVM algorithms. This research study promoted FTIR spectroscopy as an alternative to molecular methods in the routine identification of microbial pathogens in the clinical setting.

Résumé

L'identification rapide des pathogènes microbiens et viraux dans les laboratoires cliniques, vétérinaires et alimentaires reste difficile malgré la disponibilité des méthodes de typage moléculaire, étant donné que des ressources importantes sont nécessaires pour utiliser ces méthodes sur une base régulière. Au cours de la dernière décennie, l'utilité potentielle de la spectroscopie infrarouge à transformée de Fourier (IRTF) en tant que technique rentable pour l'identification des pathogènes microbiens après une culture de routine a attiré l'attention. Bien que cette approche fondée sur la culture ne puisse pas être étendue aux agents pathogènes viraux, un nombre très limité de recherches ont été rapportées sur la détection d'infections virales spécifiques par l'analyse IRTF de divers types d'échantillons. Cette thèse porte sur la démonstration de la spectroscopie IRTF pour l'identification des bactéries d'importance vétérinaire ou de sécurité alimentaire et des champignons, ainsi que sur son utilisation potentielle comme méthode de dépistage de l'infection par le SARS-CoV-2 à l'aide d'échantillons de salive. Pour l'identification des pathogènes bactériens liés à la mammite des vaches, quatre bases de données spectrales ont été développées en acquérant les spectres IRTF d'isolats (n = 582) cultivés sur deux milieux de croissance {gélose tryptique de soja (TSA) et gélose au sang Columbia (CBA)} en utilisant deux modes d'acquisition spectrale {réflectance totale atténuée (ATR) et transflectance (TR)}. L'application des modèles de prédiction développés pour chacune des quatre bases de données à un ensemble de tests (n = 98) a donné un taux d'identification correcte allant de 96,9 à 93,8% au niveau de l'espèce. Pour évaluer la capacité de la spectroscopie IRTF pour l'identification des pathogènes microbiens, deux bases de données ont été développées à partir de milieux de croissance différents et combinées avec des méthodes d'échantillonnage IRTF pour prédire les ensembles d'essais. Ces derniers résultats ont donné des pourcentages élevés de précision d'identification de 99,0 à 92,8% pour les quatre ensembles d'essais au niveau de l'espèce. L'interchangeabilité des bases de données construites à l'aide de deux instruments IRTF commerciaux de deux fournisseurs différents a également été évaluée. En utilisant la même base de données (n = 361), le spectromètre IRTF Summit Pro (ThermoFisher Nicolet, WI) (n = 138) et le spectromètre IRTF Cary 630 (Agilent Technologies, CA) (n = 305) ont atteint 99,5% et 96,2% au niveau du genre, et 95,8% et 79,9% au niveau de l'espèce, respectivement. En ce qui concerne le changement de milieu de culture de croissance, les résultats de l'identification FTIR n'ont pas été affectés à l'identification au niveau de l'espèce. En outre, l'utilisation de différents instruments IRTF n'a pas influencé les résultats de l'identification au niveau du genre. L'identification des champignons a été difficile pendant des décennies, même en utilisant des méthodes moléculaires. Par conséquent, la précision d'identification d'*Aspergillus* spp. par réaction en chaîne de polymérase quantitative en temps réel (RT-qPCR) et la spectrométrie de masse à temps de vol par désorption/ionisation laser assistée par matrice (MALDI-TOF MS) ont été comparées à l'approche basée sur la spectroscopie IRTF à l'aide d'une base de données construite en interne qui a été élargie avec des souches fongiques supplémentaires pour d'autres études de validation. Des taux d'identification corrects de 71,3%, 52,0% et 92,3% ont été obtenus pour la spectroscopie RT-

qPCR, MALDI-TOF MS et IRTF, respectivement. La RT-qPCR pourrait être plus appropriée pour l'identification d'*Aspergillus nigri* et d'*Aspergillus terrei*; et MALDI-TOF MS pourrait être un bon choix pour identifier *Aspergillus fumigatus* et *Aspergillus flavus*; tandis que la spectroscopie IRTF a fourni un taux d'identification correct de 96,6% en utilisant la base de données élargie. Enfin, en raison de la pandémie actuelle de coronavirus SARS-CoV-2 (Covid-19), l'efficacité diagnostique de la spectroscopie IRTF en tant que méthode de dépistage a été évaluée à l'aide de 940 échantillons de salive inactivés par la chaleur (418 RT-qPCR positifs pour Covid-19 et 522 RT-qPCR négatifs). Des méthodes d'analyse multivariées, y compris la forêt aléatoire, le k-plus proche voisin (KNN), le réseau neuronal artificiel (ANN) et la machine à vecteurs de support (SVM) ont été utilisées pour développer des algorithmes de dépistage de la COVID-19 basés uniquement sur les changements dans les profils spectraux infrarouges des échantillons de salive. Un ensemble de tests indépendant a donné des taux de sensibilité de 82,9 à 85,4%, des taux de spécificité de 82,4 à 86,3%, des taux de précision de 84,8 à 85,9 % et des taux de précision de 79,6 à 83,4 % des algorithmes KNN, ANN et SVM. Cette étude de recherche a fait la promotion de la spectroscopie IRTF comme alternative aux méthodes moléculaires dans l'identification systématique des pathogènes microbiens en milieu clinique.

Acknowledgments

I wish to express my greatest gratitude and appreciation to Dr. Ashraf Ismail for his guidance, patience, and kindness. I would never be able to accomplish my research without him giving so many important advice throughout the development of the project. I appreciate for all the assistance and opportunities you offered me at every stage of the research. My deepest thanks to Dr. Mohamed Nassim Amiali and Dr. Jacqueline Sedman for her insightful comments and suggestions upon editing my work. Her immense knowledge and expertise encouraged me a lot for my research. I would also like to thank Dr. Jennifer Ronholm and her lab team for their generosity and patience during the bovine mastitis work.

I would like to offer my special thanks to Hayline Kim, Cailun Tanney and Mrs. Irene Iugovas for their treasured support at Health Canada which was influential in shaping my experiment methods and critiquing my results.

I am deeply grateful to my fellow colleagues of the McGill IR group (Lisa Lam, Tamao Tsutsumi, Ran Tao, Mazen Bahadi, and Ivan Tavcar). Their kind help and support have made my study and life at McGill a wonderful time.

Last but not least, I would like to extend my sincere thanks to my parents, friends and my boyfriend Guangxin Liu for their unwavering support and encouragement.

Contributions to knowledge

The general objective of the research presented in this thesis was to assess the feasibility of creating a robust FTIR spectroscopic methodology for routine identification of pathogenic microorganisms that would accommodate interlaboratory variations in culturing procedures and FTIR instrumentation, Identification of bacterial pathogens using FTIR spectroscopy has been widely reported in the literature since the 1990s. While various sample handling methods and modes of spectral acquisition have been employed therein, limited studies have been reported involving direct comparison of the results obtained with different sample handling methods. In addition, there is a knowledge gap regarding the interchangeability of infrared spectral databases developed for the same microbial pathogens but grown on different types of general-purpose growth media; filling this gap would provide an indication of the permissible latitude in the growth conditions employed in the methodology. In contrast to the numerous research papers related to bacterial identification, studies on fungal identification by FTIR spectroscopy are scarce, warranting a comprehensive investigation encompassing molds and yeasts from diverse sources. Viral pathogens are not amenable to FTIR spectroscopic study, except when a high-intensity source of infrared light can be employed, commonly necessitating access to an infrared beamline at a synchrotron facility. The possibility of detecting viral infection in a human host by FTIR spectroscopic analysis of appropriate biofluids had been investigated in a very limited number of studies prior to the Covid-19 pandemic. However, the pressing need for rapid whole-population screening methods to control the spread of Covid-19, the highly infectious disease caused by the novel coronavirus SARS-CoV-2, raised the question of whether viral infection could be detected by FTIR spectroscopic analysis of saliva specimens.

The challenges summarized above were addressed in the research presented in this thesis. The primary contributions to knowledge resulting from this research are listed below:

1. **Development and validation of an FTIR spectral database for the identification of common Gram-positive bovine mastitis pathogen.**

   Common bovine mastitis-related pathogens were collected, and their FTIR spectra recorded and placed in an infrared spectral database. The spectra were subdivided into a training and a test set. Discrimination between the different genera and species within each genus within the training set was undertaken by a multi-tier pairwise approach using a

PCA-LDA algorithm. The test set then was used to validate the discrimination model and assess its predictive performance.

2. **Comparison of ATR and TR spectral acquisition modes for the identification of microbial pathogens by FTIR spectroscopy.**

   Sample preparation steps required for spectral acquisition in the transmission mode represent a bottleneck in the conventional experimental procedure for microbial identification by FTIR spectroscopy. Transflection (TR) and attenuated total reflectance (ATR) are alternative modes of spectral acquisition that facilitate (TR) or eliminate (ATR) sample preparation, with each having additional advantages as well as inherent limitations. In the present work, a direct comparison of ATR-FTIR and TR-FTIR spectroscopy for the identification of *Staphylococcus* spp. and *Streptococcus* spp. at the species level was made by employing identical training and test sets.

3. **Comparison of different growth medium for the identification of microbial pathogens.**

   Many studies have reported that the identification accuracy of FTIR spectroscopy for microbial identification is highly dependent on the growth media. Bovine mastitis-associated pathogens were grown on two different agar culture media (CBA and TSA) prior to spectral acquisition, and the performance of the classification models developed with each set of cultures was compared. Database interchangeability was also examined by comparing the predictive accuracy attained when applying the resulting classification models to test sets of samples grown on each of the two media.

4. **Evaluation of ATR-FTIR spectroscopy for the discrimination of foodborne pathogens at the species level.**

   Infrared spectral databases comprising ATR-FTIR spectra of *Escherichia coli, Salmonella* spp., *Listeria* spp., and *Shigella* spp. isolated from food samples were created, and discrimination models were generated by HCA and PCA-LDA. Classification to the genus and species level was achieved for all genera, and spectral features serving as biomarkers were suggested for microbial differentiation.

5. **Investigation of interchangeability of spectral databases constructed using FTIR spectrometers from different manufacturer.**

   Two commercial models of FTIR spectrometers (Summit, Thermo Nicolet, WI and Cary 630, Agilent Technologies, CA) were employed to generate independent spectral databases.

The identification accuracy of microbial pathogens at the genus and species levels was established for each database. Database interchangeability was assessed by using the training sets from the two different spectrometer models for predicting both training sets. Genus-level classification could be achieved independent of the spectral database training set.

6. **Comparison of RT-qPCR, MALDI-TOF MS and FTIR spectroscopy for the accurate identification of fungi species.**

   Aspergillus spp. identification has been challenging for decades. RT-qPCR, MALDI-TOF MS and FTIR spectroscopy were evaluated and compared for the accurate identification of *Aspergillus* spp., with the latter method being the most performant.

7. **Development and validation of a fungal FTIR database with molds and yeasts from different source.**

   Mold and yeast strains isolated from clinical, food, and cannabis sources were collected to build a fungal FTIR database. The database was further validated, demonstrating accurate identification of strains regardless of the source of origin.

8. **Feasibility study of FTIR spectroscopy for Covid-19 diagnosis and development of a Covid-19 FTIR database.**

   FTIR spectroscopy was evaluated as a novel tool for Covid-19 diagnosis using heat-inactivated saliva specimens. Despite the complexity of the spectra of saliva, specific biomarkers were selected and used for the discrimination between Covid-19 negative and Covid-19 positive samples of saliva.

9. **Comparison of KNN, ANN and SVM algorithms for Covid-19 diagnosis using FTIR spectroscopy by heat-inactivated saliva.**

   Three machine learning algorithms (KNN, ANN and SVM) were applied and compared for the discrimination between Covid-19 negative and Covid-19 positive samples of saliva based on differences in their FTIR spectra. Although all three algorithms yielded comparable discrimination between Covid-19 negative and Covid-19 positive samples of saliva, KNN showed a minor advantage over ANN and SVM with greater identification accuracy and model performance.

Contribution of authors

With the exception of sample acquisition (microbial strains and saliva specimens), all work described in this thesis was carried out by the author.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACE2	Angiotensin converting enzyme 2
AFLP	Amplified fragment length polymorphism
ANG II	Angiotensin II
ANN	Artificial neural network
ANOVA	Analysis of variance
AOAC/IDF	Association of Analytical Chemists
API	Active pharmaceutical ingredient
ARIS	Automated Reading and Incubation System 2x System
ATR	Attenuated total reflectance
BHI	Brain heart infusion agar
BLAST	Basic Local Alignment Search Tool
BTS	Bruker bacterial test standard
Cary	Cary 630 FTIR spectrometer
CBA	Columbia blood agar
CFIA	Canadian Food Inspection Agency
CoNS	Coagulase-negative *Staphylococci*
Covid-19	Coronavirus disease 2019
DAOMC	Canadian Collection of Fungal Cultures
ELISA	Enzyme-linked immunosorbent assay
FCM	Fuzzy C-mean clustering
FIR	Far-infrared
FTIR	Fourier-transform infrared
FTIRM	Fourier-transform infrared microspectroscopy
GN	Gram-negative
GP	Gram-positive
HC	Health Canada

HCA	Hierarchical cluster analysis
HPFB	Health Products and Food Branch
HPLC	High-performance liquid chromatography
HUS	Hemolytic-uremic syndrome
ICA	Independent component analysis
IDFP	New Conidia ID-fungi plate
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IgM	Immunoglobulin M
IR	Infrared
IRE	Internal reflectance element
ISHAM	International Society of Human and Animal Mycology
ITS	Internal transcribed spacer
KMCA	k-Means cluster analysis
KNN	k-Nearest neighbor
LDA	Linear discriminant analysis
LR	Logistic regression
MALDI-TOF MS	Matrix-assisted laser desorption/ionization-time of flight mass spectrometry
MCC	Mathews' Correlation Coefficient
MIR	Mid-infrared
MLST	Multilocus sequence typing
MLVA	Multiple locus variable-number tandem repeat analysis
MRSA	Methicillin resistant *S. aureus*
MSC	Multiplicative scatter correction
NIR	Near-infrared
NMR	Nuclear magnetic resonance
PCA	Principal component analysis

PCR	Principal component regression
PFGE	Pulsed-field gel electrophoresis
PLSR	Partial least squares regression
PMF	Peptide mass fingerprint
QDA	Quadratic discriminant analysis
RFLP	Restriction fragment length polymorphism
ROC	Receiver operating characteristic
RT-qPCR	Multiplex quantitative real time polymerase chain reaction
SARS-CoV	Severe acute respiratory syndrome coronavirus
SDA	Sabouraud dextrose agar
Summit	Nicolet^{TM} Summit FTIR spectrometer
SVC	Support vector classifier
SVD	Singular value decomposition
SVM	Support vector machine
TR	Transflection
TSA	Tryptic soy agar
TSB	Tryptic soy broth
VTM	Viral transport medium
WGS	Whole genome sequencing
WHO	World Health Organization

Chapter 1. Introduction

1.1. General Introduction

Identification of microbial pathogens is of utmost concern for public health and infection prevention and control. They are inevitably employed in many research fields including biotechnology, medicine, genetic engineering, forensic, food science, disease diagnosis, and others. Current conventional and gold standard methods used in a routine microbial diagnostic laboratory setting for the identification of microbial pathogens are based on phenotypic and genotypic methods, where the latter especially dedicated for epidemiological and surveillance purposes due to their higher sensitivity and specificity compared to phenotypic methods. However, these methods have long turnaround time, and are generally labor intensive, costly, and requires specific reagents.

Presently, biophysical techniques such as Fourier-transform infrared (FTIR) spectroscopy and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) have gained more attention as they are low in cost, non-destructive, rapid, and easy to use for microbial identification. Nonetheless, MALDI-TOF MS may experience difficulties in sub-species level typing since it relies primarily on ribosomal protein sequences. FTIR spectroscopy, on the other hand, generates spectra based on the absorption of infrared light by the different chemical components (lipids, proteins, polysaccharides) of the whole microbial cell. As the entire spectral fingerprint is generated, closely related species could be differentiated using FTIR spectroscopy. Samples of different genera and species can be examined to find specific spectral biomarkers for rapid and accurate identification by FTIR spectroscopy. Although infrared spectroscopy is mainly used for analytical chemistry purposes, it has been evaluated for microbial pathogen identification for decades. Compared to conventional phenotypic and genotypic methods, FTIR spectroscopy is versatile, fast, non-invasive, and non-destructive, and it requires no reagents. Additionally, FTIR spectrometers are simple to operate and may have the potential to be automated (depending on the sample handling configuration) and implemented on-site. This latter advantage could facilitate routine inspection of pathogens in the food, animal feed, medicinal and bioengineering industries, and could also allow medical physicians to prescribe efficient treatment in a timely manner. Combined with appropriate multivariate statistical analysis tools, FTIR spectroscopy could be

comparable and even competitive to the gold standard methods for the identification of microorganisms. However, it is important to note that the use of different brands of FTIR instruments, modes of spectral acquisition, sample preparation and handling techniques, growth media compositions, and chemometric algorithms in research studies can affect the accuracy of microbial identification by FTIR spectroscopy.

In this doctoral research project study, we aimed to investigate the capabilities of FTIR spectroscopy for the identification of bovine mastitis-, food- and clinical-related microorganisms at species and strain level. In addition, following the sudden emergence of Covid-19, we were among several research groups worldwide to undertake an evaluation of the potential use of FTIR spectroscopy as a screening tool for Covid-19 while being the only group to base this work on the use of heat-inactivated saliva as biospecimen. We also attempted to standardize an FTIR spectroscopy-based methodology to expand the applicability of FTIR spectroscopy in a clinical microbiology laboratory setting.

1.2. Research Rationale and Objectives

The general objective of the research is to evaluate FTIR spectroscopy for routine identification of microbial pathogens (bacteria, molds, and yeasts) and for the detection of viral infection in biofluid specimens from human subjects, using the detection of SARS-CoV-2 infection in saliva specimens as a case study. The research aims to assess the capacity of FTIR spectroscopic methods to replace conventional methods of microbial identification or to be deployed in a pre-screening step to minimize the number of samples warranting costly genotypic analysis.

1.2.1. Specific Objectives

i. To create the first infrared spectral database dedicated to bovine mastitis-related pathogens and demonstrate its potential practical utility in diagnostic veterinary microbiology.

ii. To compare ATR-FTIR and TR-FTIR spectroscopy for the identification of bovine mastitis-related pathogens.

iii. To assess the interchangeability of spectral databases consisting of FTIR spectral profiles of the same microbial pathogens grown on two general-purpose agar media, namely, TSA and CBA, for the accurate identification of microbial pathogens by FTIR spectroscopy.

iv. To build a comprehensive spectral database comprising FTIR spectra of reference and wild-type strains of *E. coli*, *Salmonella* spp., *Listeria* spp., *Shigella* spp.

v. To evaluate the capabilities of FTIR spectroscopy for the identification of *E. coli*, *Salmonella* serogroups, *Listeria* spp., and *Shigella* spp. using an in-house developed database.

vi. To evaluate inter-instrument spectral database compatibility using two models of FTIR spectrometers from different manufacturers for the identification of microbial pathogens.

vii. To compare FTIR spectroscopy-based approach to RT-qPCR and MALD-TOF MS identification methods for the identification of *Aspergillus* spp.

viii. To create an infrared spectral dedicated to fungi by acquiring FTIR spectra of molds and yeasts from different sources for the accurate identification of fungal pathogens by FTIR spectroscopy

ix. To create an infrared spectral database consisting of FTIR spectra of heat-inactivated saliva specimens (found to be Covid-19 positive and Covid-19 negative by RT-qPCR) for the development of a discrimination between the two diagnostic states by FTIR spectroscopy.

x. To compare KNN, ANN and SVM algorithms for the effective discrimination between Covid-19 positive and Covid-19 negative saliva specimens based on FTIR spectra of the heat-inactivated specimens.

Chapter 2. Literature Review

2.1. Introduction

The identification of microorganisms is crucial for a magnitude of research purposes related to disease diagnosis, vaccine development, food safety, food processing, and pharmaceutical industry. Microorganisms such as bacteria, fungi viruses and phages could have beneficial or detrimental impact on health and safety. Pathogenic bacteria are etiological agents associated with 350 million cases of foodborne diseases. Among them, virus such as coronavirus (Covid-19) is currently the most discussed topic in many conversations throughout the world with this once-in-a-century pandemic [1]. The World Health Report 1996 - Fighting disease, fostering development, published by WHO, states that infectious diseases are the world's leading cause of premature death. Of about 52 million deaths from all causes in 1995, more than 17 million were due to infectious diseases, including about 9 million deaths in young children. Up to half the world's population of 5.72 billion are at risk of many endemic diseases [2]. In addition, millions of people are developing cancers as a direct result of preventable infections by bacteria and viruses; in Canada, infectious disease cause more than 200 thousand deaths annually [3]. The health effect and economic cost cannot be measured, but it is well acknowledged that the global impact with regards of health, trade, trust and development is enormous. There are about 1400 known species of microorganisms that are pathogenic to human, including bacteria, fungi, yeast and viruses, accounting for less than 1% of the total number of known existing microbial species on the planet [2]. Although many microbes cause public health concerns, some are very important and are essential to produce antibiotics, hormones, supplements, amino acids, therapeutics, and many food byproducts, such as yogurt, bread, and cheese. Moreover, microorganisms are also used for sewage treatment, waste composting and biodegradable plastic, which are essential for our ecological system. Additionally, the microbiome, which is essentially all colonizing microorganisms present in a human being play an important role in keeping us healthy by controlling digesting and boosting our immune system. Knowing all the pros and cons of microbes, identifying the microorganism is fundamental in order to prescribe corresponding treatment, further understand its physiology, and other scientific purposes.

The isolation, identification and classification of microorganisms are usually the first steps in microbiological studies. Hence, identification methods are inevitably employed in many research fields including biotechnology, medicine, genetic engineering, forensic, food science, disease diagnosis, and others. Current protocols used in conventional laboratories are generally based on diverse phenotypic and genotypic methods, where genotypic methods are especially used for epidemiological and surveillance purposes due to their higher accuracy compared to phenotypic methods. However, the time needed to identify the microorganisms employing genotypic methods could take days, which would potentially delay diagnosis or establish the origin of a foodborne outbreak. In addition of lacking speed, conventional genotypic methods are generally labor intensive, costly, and requires specific reagents [4].

Recent development in spectroscopic-based analytical methods has led to a new era in identification of microorganisms. Fourier transform infrared (FTIR) spectroscopy is an analytical tool that generates a unique fingerprint-like spectrum of each microbial species. This feature makes FTIR spectroscopy a promising tool for microbial identification and classification in microbiology diagnostic laboratories. FTIR spectrometer comprises an IR energy source emitting a broad band of distinct wavelengths that passes through an interferometer that is responsible for modulating the IR wavelengths. The modulated IR beam then passes through the sample where distinct wavelengths are partially absorbed, and the remaining intensity is measured by an infrared detector, which is subsequently converted into a transmittance IR spectrum. This IR spectrum is the ratio of the intensity recorded in the presence of the sample against the intensity reaching the detector in the absence of the sample. Compared to conventional phenotypic and genotypic methods, FTIR spectroscopy is versatile, fast, non-invasive, reagent-less, and non-destructive. This analytical technique is cost-effective and could generate detailed information on all biomolecules of microorganism, such as lipids, proteins, carbohydrates, and nucleic acids with a minimal amount of sample, allowing microbiologists to identify the microorganism in matter of minutes based on the spectral difference between different microbial isolates [5]. Combined with appropriate multivariate statistical algorithms, FTIR spectroscopy results could be comparable to standard phenotypic and genotypic methods. Additionally, this technique is overall easy to operate with the potential of being fully automated and implemented on-site. This would facilitate routine inspection of microbial pathogens in food, feed, medicinal and bioengineering industries, and would also

allow medical physicians to prescribe efficient treatment in a timely manner. The speed and easiness of microbial identification using FTIR spectroscopy may even allow testing for SARS-CoV-2 in order to prevent epidemiological outbreaks. This research aims to evaluate the ability of FTIR spectroscopy to identify and classify food-related microorganism at species and strain level, as well as its potential as diagnostic tool for SARS-CoV-2 virus using saliva as biospecimen. We attempted to promote FTIR spectroscopy as a potential cost-effective alternative diagnostic tool to genotypic method for the detection of Covid-19.

2.2. Common Foodborne Pathogens

According to WHO, almost 1 in 10 people (600 million of the world population) suffer from foodborne disease yearly, causing 420 000 deaths. Nearly half of the deceased are children under the age of 5 [6]. Food safety is generally recognized as a major health burden concern only in low- or middle-income countries; but the fact is, foodborne disease is an important issue worldwide, and it has been challenged due to globalization of the food supply. Canada, for example, is a well-developed country, and its food is considered amongst the safest in the world. Yet, according to the Public Health Agency of Canada, 1 in 8 people (4 million Canadians) get sick each year from contaminated food [7]. Causes of these foodborne illnesses can be bacteria, viruses, parasites, toxins, metals, or prions. While most people fully recover from mild symptoms or no symptom, some people may have more severe and possibly long-term or permanent side effects, especially for those having health issue history and elderly or immunocompromised individuals, as well as pregnant women and their unborn children. In extreme cases, foodborne illnesses can be fatal. Over 240 deaths and 11,500 hospitalizations occur each year in Canada, and keeping in mind that these numbers are an underestimation of people getting sick, as not everyone might go see a doctor and get tested when not feeling well [7]. The estimated economical cost to due to foodborne illnesses and related deaths is approximately $12-14 billion per year in Canada [8]. Estimates of the economic costs associated with foodborne disease are important to inform public health decision making. In 2008, 57 cases of listeriosis and 24 deaths in Canada were linked to contaminated delicatessen meat from one meat processing plant. Costs associated with the cases (including medical costs, nonmedical costs, and productivity losses) and those incurred by the implicated plant and federal agencies

responding to the outbreak were estimated to be nearly $242 million Canadian dollars, and these numbers are likely just a conservative estimate [9].

In Canada, the leading causes of foodborne illnesses are *Listeria monocytogenes* (*L. monocytogenes*), *Salmonella* spp., *Escherichia coli* (*E. coli*), *Campylobacter jejuni, Clostridium, Cyclospora, Hepatitis A, Shigella* and *Vibrio*, with *Staphylococcus* spp. emerging as a public health concern in recent years [10]. In general, bacteria can be classified as Gram-positive or Gram-negative. This classification is based on the reaction to the Gram stain test, where Gram-positive bacteria are violet, and Gram-negative bacteria pink due to their difference in cell wall composition. Despite some Gram-variable bacteria may be difficult to classy due to their variations in cell wall or other structural characteristics, Gram test is a useful tool and is still used in most laboratories as the first identification step. Gram-positives have cell walls composed mostly of peptidoglycan, which is responsible for the violet staining, whereas Gram-negatives have only a thin layer of peptidoglycan plus an outer lipopolysaccharide membrane. This outer membrane of Gram-negatives serves as a protective layer, allowing them to be more resistant to antibiotics than Gram-positives [11]. Though both groups of bacteria can cause disease, they require different treatments. While most foodborne pathogens are Gram-negatives, such as *E. coli* and *Salmonella* spp.; *Listeria* spp. and *Staphylococcus* spp. are Gram-positive food pathogens that cause severe damage to human body with higher mortality rate. Other than bacteria and viruses, the increasing number of fungal infections attracts consideration for an accurate diagnostic tool for routine identification of fungi during outbreaks. The research described in the present work will be focused on *E. coli, Salmonella* spp., *L. monocytogenes, Staphylococcus* spp., and several food-related fungi.

2.2.1. E. coli

E. coli is a Gram-negative, facultative anaerobic, rod-shaped bacterium of the genus *Escherichia* [12]. This is a large diverse group of bacteria that is commonly found in the gut of humans and warm-blooded animals. Enterohemorrhagic *E. coli* (EHEC), particularly serotype O157:H7, is a highly pathogenic subset of Shiga toxin-producing *E. coli* (STEC) that causes gastrointestinal illnesses, such as stomach cramps, vomiting, fever, aqueous or bloody diarrhea, kidney failure, and death [13]. Incubation period can range from 3 to 8 days, and most patients recover within 10 days. But in a small proportion of patients, particularly children and elderly,

the infection may lead to hemolytic-uremic syndrome (HUS), which is a potentially life-threatening complication with a case-fatality rate ranging from 3-5% [14]. Global morbidities and mortalities in *E. coli* foodborne illness are high. In 2010, 321,969,086 cases of *E. coli* foodborne illness contributing to 16.1% of global foodborne diseases was reported, along with 196,617 deaths due to *E. coli* which is 0.02% of global mortalities [14]. In Canada, *E. coli* is also one of the leading foodborne pathogens causing 88,000 illnesses, 925 hospitalizations, and 17 deaths per year [7].

Cattles are recognized as the main reservoir for *E. coli* O157:H7, although other mammals like sheep, goats, chickens, deer, pigs have also been known to carry it. Meat may be contaminated during slaughter and processing, when the infected animal's intestines or feces are in contact with the carcass. Ground meat is usually riskier because *E. coli* may be mixed within the meat in the grinding process. Other potential food source includes unpasteurized cheese and milk, fruits and vegetables, water, person-to-person contact. All these transmission routes mentioned above can be prevented by maintaining hygienic handling measures at all stages of the food chain. The strain *E. coli* O157:H7 differs from other *E. coli* since it is unable to ferment sorbitol, lack β-glucuronidase enzyme, and does not grow at temperature above 44℃ [15]. Hence, sorbitol MacConkey agar is very useful in the detection of *E. coli* O157:H7. Confirmation of *E. coli* isolates can be done by biochemical, enzymatic, or molecular method in hospitals.

Until recent years, serotype O157:H7 has been the top causative serotype of STEC related to foodborne illnesses. Nowadays, sporadic cases and outbreaks caused by non-O157 STEC strains have increased significantly, and these strains are now responsible for approximately 64% of STEC each year, namely O26, O45, O103, O111, O121 and O145 [16]. A characteristic that non-O157 STEC strains share with O157, is that they generally possess the adhesin intimin, *eae* gene, which is a pathogenic marker for all EHEC [17].

2.2.2. Salmonella

Salmonella are Gram-negative, non-sporulating, facultative anaerobic and predominantly motile. *Salmonella enterica* subsp. *enterica* is the only one to cause illness in humans. There are over 2000 serovars in this subspecies, while serovars *Typhimurium* and *Enteritidis* are the two most common strains associated with foodborne illness which are non-

typhoidal [18]. *Salmonella* is 1 of 4 key global causes of diarrheal diseases and contributes to 1 in 4 hospitalizations of all foodborne illnesses in Canada [7], and is the leading cause of food poisoning in the EU [19]. One recent massive Salmonellosis in Canada occurred in 2010 associated with headcheese [20]. Symptoms generally appear 5 to 72 hours after infection and last for 3 to 7 days without treatment, and include diarrhea, abdominal cramps, fever and vomiting. Reactive arthritis can also occur in 2-15% of *Salmonella* patients. Typhoid fever may be due to serovars *Typhimurium* and *Paratyphimurium*, which can, although not that common, be transmitted by food, and causes serious illness that may result in death [21].

Salmonella has been isolated from fruits and vegetables, eggs, chicken, pork, and even processed foods such as chicken nuggets, frozen pot pies, and nut butters. Pets, including cats, dogs, birds,and reptiles may also carry *Salmonella* [22]. Person-to-person transmission can also occur through fecal-oral route. Preventive measures for *Salmonella* are similar to other foodborne pathogens, which are to maintain proper hygienic habit and having control measures at all stages of the food chain. For most the part, treatment of salmonellosis includes rehydration and electrolyte replacement. Antimicrobial therapy is not recommended to avoid antimicrobial resistance.

In the last 20 years, there has been a worldwide emergence of multidrug-resistant phenotype among *Salmonella* serovars, especially for *Salmonella* serovar *Typhymurium* [22], alerting a need for a fast and specific diagnostic tool for identification at serovar level and for antimicrobial resistance susceptibility profile of *Salmonella*. For isolation, Rappaport-Vassiliadis broth followed by plating on xylose lysine desoxycholate agar can be used, and multiplex polymerase chain reaction (PCR) assay was also successfully employed in identifying serovars of *Salmonella* [23]. Although labor intensive and costly, whole genome sequencing has also been used and can accurately predict the serovar level [24].

2.2.3. L. monocytogenes

L. monocytogenes is a Gram-positive, facultative anaerobic, non-spore forming bacterium that is recognized as a human pathogen with 17% case-fatality rate [25]. It causes a disease known as listeriosis, which is uncommon but potentially fatal, usually caused by food contamination. *L. monocytogenes* has 13 serotypes, where 1/2a, 1/2b, and 4b are those associated with the majority of foodborne infections [26]. *L. monocytogenes* causes a serious public threat throughout the world.

In the United States, an estimated 1,600 people get listeriosis, and about 260 die each year [27]. In Canada, listeriosis is the leading cause of deaths related to foodborne illness, contributing 33 to 35% of known causes of foodborne deaths [7]. Symptoms may appear after 3 to 70 days after infection, and usually include fever and muscle aches. For people with weakened immune system, like elderly or immunocompromised individuals, *L. monocytogenes* can invade the central nervous system and cause meningitis or brain infection. Infection during pregnancy may lead to miscarriage, stillborn or infection of the newborn [25].

L. monocytogenes can be found in soil, water and animals, but most human infections are due to consumption of contaminated food. It can be found in a variety of raw foods, such as raw milk, meat and vegetable, as well as in foods that become contaminated after processing, such as soft cheese, hot dogs and deli meat, while rare but deadly cases include contaminated caramel apples and cantaloupe [28, 29]. Unlike many other bacteria, it can thrive and reproduce in contaminated food stored in the refrigerator. Furthermore, increasing evidence suggests that this bacterium is able to persist in food processing plants for years or even decades [30]. Therefore, the safest way to prevent listeriosis is to cook thoroughly raw food, avoid unpasteurized milk products and keep raw food separate from cooked ones.

There are several isolation methods that have been proven efficient for *Listeria*, such as FDA BAM and ISO 11290 methods, which involve enriched broth containing selective agents (acriflavin, naladixic acid and antifungal agent) and then plating onto selective agar (Oxford, PALCAM, MOX or LPM), and USDA and Association of Analytical Chemists (AOAC/IDF) method 993.12 with similar procedure with the former but have specific procedure for different type of food products [31]. For identification, the Christie, Atkins, Munch-Petersen (CAMP) test, antibody-based test such as ELISA and molecular typing techniques showed a high discrimination power for *Listeria* at species level [32].

2.2.4. Staphylococcus aureus (S. aureus)

S. aureus is Gram-positive, facultative anaerobic, non-spore-forming catalase-positive cocci. It was estimated that *S. aureus* causes 241 000 illnesses per year in the US, and costing $695 per case: this pathogen cost more than $160 million annually in the US. The ingestion of *S. aureus* enterotoxin causes symptoms including vomiting, nausea, cramps and diarrhea [33]. The illness can be self-limiting, recovering within one or two days, or life threatening for some population

groups, such as infants, elderly and immunocompromised people. Among the nine identified staphylococcal enterotoxins, type A and D are responsible for the majority of the outbreak associated illnesses [30].

Humans are considered the main reservoir of *S. aureus*. It can be found abundantly on the skin and nasal cavity of healthy individuals. Food can be contaminated during preparation if the food handler is infected with *S. aureus*. Furthermore, *S. aureus* has also been found in air, dust, sewage, and water, due to its ability to survive for long periods of time in dry state [26]. A variety of food has been associated with *S. aureus*, including ground beef, pork sausage, ground turkey, salmon steals, oysters, shrimp, cream pies, milk, and delicatessen salads. Since staphylococcal enterotoxins are highly heat stable, and generally requires heating above boiling point for at least one minute for them to lose serological activity [34].

The enterotoxin can be detected using bioassay methods, molecular biology, or immunological techniques, and several molecular typing methods can be used in combination. Pulsed-field gel electrophoresis (PFGE) and *spa* typing can be used to trace the origin of *S. aureus* contamination [35]. Multilocus sequence typing (MLST) is another molecular typing method that has helped in providing insights into the population structure of *S. aureus* [36].

S. aureus infection is also a major health concern in milking cows. *S. aureus* is responsible for around 5% to 70% of cow mastitis worldwide, causing 45% decrease in milk production per quarter, which highly impact the dairy sector and results in huge economical loss [37]. Not to mention the worldwide concern of methicillin resistant *S. aureus* (MRSA). Coagulase-negative *Staphylococci* (CoNS) are as pathogenic as *S. aureus* due to the upward prevalence of CoNS in bovine mastitis in the whole world [37]. *Staphylococcus* spp. cause many infections and health complications due to its combination of toxin-related virulence, invasiveness, targeting hosts, and antibiotic resistance. In addition to *Staphylococcus* spp., *Corynebacterium* spp., *Enterobacter* spp., *Klebsiella* spp., *Streptococcus* spp. and *Trueperalla pyogenes* are also found in intramammary infections of cows [38]. The Mastitis Pathogen Culture Collection in Canada contains more than 16,000 bacterial isolates including those mentioned above, and some of these bacteria genus will be investigated in this research.

2.2.5. Fungi

Fungi are eukaryotes, and as such, have a complex cellular organization. A typical fungal cell contains a true nucleus, mitochondria, and a complex system of internal membranes, including the endoplasmic reticulum and Golgi apparatus. Fungi differ from bacteria by the presence of chitin in their cell wall which could be identified in microscopic method throughout the use of lactophenol cotton blue stain. In brief, phenol acts as a disinfectant by killing the living fungi, lactic acid preserves the cell structure, and the cotton blue give color to the chitin present in the fungal cell wall making it visible and easier to interpret under microscope [39]. Gram stain is sometime used for fungi, especially yeasts, yet study has shown that intact yeast cells and broken ones may give completely different Gram stain result [40]. Fungi are broadly categorized into yeast, mold and mushroom, and part of this research will be focused on the accurate identification of several molds and yeasts.

Molds are multicellular fungi that reproduce by the formation of spores in either an asexual process or by sexual reproduction. Many of them can produce several types of spores, depending on the growth condition. Spores are formed in large numbers and are easily dispersed through the air. Food could be contaminated by spores, and these spores can grow and reproduce in adequate environmental conditions. Yeasts are unicellular fungi and oval or round shaped that are much larger than bacterial cells. They reproduce by an asexual process called binary fission or budding [41]. Molds and yeasts are both fungi that are useful in several industrial applications including food, medication, and beverage. *Saccharomyces cerevisiae*, for instance, also known as baker's yeast, is probably the most recognized for bread and wine production. In addition, some well-known species of fungi such as *Aspergillus nidulans*, *Aspergillus flavus*, *Aspergillus glaucus*, *Aspergillus oryzae*, *Aspergillus nomius*, *Penicillium griseofulvum*, *Bjerkandera adusta*, *Phanerochaete chrysosporium*, *Cladosporium cladosporioides*, and some other saprotrophic fungi, such as *Pleurotus abalones*, *Pleurotus ostreatus*, *Agaricus bisporus* and *Pleurotus eryngii* help in the bio-degradation of plastics, and therefore play an important role in our ecological system of helping degrading plants and animals. On the other hand, they are also responsible for food spoilage which is very commonly seen in households like rotten tomatoes and moldy bread. While the low pH of fruits is not optimal for bacterial growth, yeasts and molds live happily in it and cause spoilage. Although most of the fungi could be digested with no health complication, intoxication or adverse chronic effect could arise from the

presence of alfa-toxins produced by some fungal species. Among 1.5 million fungal species, approximately 300 are able to cause illnesses ranging from allergic reactions to life-threatening disease [42]. Mycotoxins, the secondary metabolites synthesized by fungi, are well known to be harmful to human and animals targeting kidney, liver, immune system and some are carcinogenic [43]. Although health consequences of mycotoxins have been relatively well described, the insignificant number of hospitalizations related to fungi contaminated food consumption may be due to underestimation of number of people being ill directly or indirectly due to ingesting molds and yeasts.

Many yeast species are naturally present in the environment and in human skin. *Candida* sp., for instance, are the most frequently isolated bloodstream pathogen in the United States, and the most common cause of fungal urinary tract infections [34]. However, even though yeast are opportunistic pathogens, some of them could be useful in food industry. Taking *Debaryomyces hansenii* as an example, this yeast is used for the fermentation of barrel-aged beers, but it is also known as *Candida famata,* which accounts to 2% of invasive candidiasis cases [44]. *Yarrowia lipolytica* or *Candida lipolytica* is capable of metabolize triglycerides and fatty acids as carbon source. This feature led it to be widely used in food industries for the production of citric acid and γ-decalactone [45]. Despite its expanding use in biotechnology, some yeast species have been recently found as an opportunistic pathogen that cause infections in premature newborns, immunocompromised and critically ill patients [46]. For example, among *Trichosporon* species, *Trichosporon asahii* has been recognized in causing candidiasis in clinical presentation and in histopathologic appearance [47]. Other fungal species such as *Mucor racemosus* havs also been recorded to cause disease in human. Despite causing opportunistic infection in diseased patients, *Mucor* species has been famous for its intense application in producing industrial enzymes. In the clinical field, yeast is also useful to study multidrug resistant and its anti-inflammatory activity of its secondary metabolites [48]. *Geotrichum* sp. consist over 100 species, in which most of them are desirable and are used for making cheese. Yet, *Saprochaeta capitata* and *Saprochaeta clavata* were all infection-causing species with high mortality previously classified under this genus [49]. Other environmental fungi such as *Cladosporium* species and *Penicillium commune* have not yet been reported to cause disease to human although they are often found as spoilage mold on cheese and vegetable [50]. One thing to note about *Penicillium commune*, is that this fungus produces both cyclopiazonic acid and regulovasine A and B as its main mycotoxins [51]. Even though no

case of infection has been reported to be caused by these fungi, long term exposure to large amount of any mold and mycotoxins may cause adverse health effects. Moreover, infections with microfungi resistant to antifungal drugs are an increasing concern. Accurate strain identification is therefore crucial to prescribe timely treatment and to perform further investigation.

Just like bacteria, fungi can be isolated and identified using different phenotypic and genotypic method. Conventional fungal identification is based on morphological and physiological test. Since these methods are time consuming, numerous DNA-based methods have been developed and used in conjunction with conventional methods in the clinical laboratory [52]. Despite being labor intensive and costly, PCR, 18S rRNA and 28S rRNA have shown to be promising genotypic methods for fungal identification. However, some of these commercially available methods may still provide limited result accuracy due to their limited database sequence [53].

2.3. Saliva as Diagnosis Biospecimen for SARS-CoV-2 (Covid-19)

Most laboratory diagnostic tests require collection of patients' specimens from the upper respiratory tract (e.g. nasopharyngeal and oropharyngeal swab), as well as lower respiratory specimens (eg. sputum or endotracheal aspirate or bronchoalveolar lavage), in addition to blood, feces, and urine [54]. Nevertheless, these collecting techniques may cause sneezing or coughing of the patient, increasing the exposure risk of healthcare staff to the disease, especially for highly infectious viruses like SARS-CoV-2. Moreover, these invasive colleting techniques are considered uncomfortable for patients as it may cause occasional bleedings that lead to further complications. In fact, a three-day-old baby and a one-and-a-half-year-old child died after inserting nasal swab to take samples for Covid-19 testing [55, 56]. On the other hand, saliva specimens have been reported with significant advantages. They are stable for diagnostic purpose for 24 hours in room temperature and for a week at 4℃ without coagulation [57]. Furthermore, saliva samples can be stored at -80℃ and still remain useful for scientific investigation [58]. Saliva can be easily self-collected by patients at home, without patient discomfort and minimizing the exposure of healthcare staff in the context of Covid-19. Saliva can be considered as the best specimen for diagnosis on humans from an ethical point of view. Since it is non-painful and non-stressful for patients, saliva collection can be used in large scale or epidemiological studies. Another thing worth mentioning is that saliva collection would waste less gloves and personal protection

equipment than collecting other body fluid specimens. In general, saliva collection technique is a stable and non-invasive method lower in cost, lower in risk of cross-contamination, possibility of being self-collected, easy to obtain, and needless of trained healthcare staff for collection.

Saliva has an important fraction made up of proteins, lipids, carbohydrates, salts, and non-protein nitrogen, although it is composed mainly of water. Saliva is excreted by major salivary glands, such as the parotid glands, submandibular glands, and sublingual glands. The salivary glands are surrounded by abundant capillaries, blood, and acini, and they have high permeability which allow them to exchange molecules. Therefore, some biomarkers in the blood can be ultimately secreted in saliva. It is well documented that saliva harbors a wide range of circulatory components, including minerals, electrolytes, buffers, enzymes and enzyme inhibitors, RNA and DNA, growth factors and cytokines, immunoglobulins, mucins and other glycoproteins [59]. The secretion and composition of saliva depend highly on the gland from which saliva is secreted, as well as the individual's age, gender, and type of stimulating factor [60, 61]. Body condition could also be reflected from the saliva. Emerging technologies have disclosed an increasing variety of diseases to be able to diagnose using saliva. Nowadays, successful diagnosis of bacterial, viral or fungal origin infections, as well as cardiovascular diseases, cancers such as breast, lung and pancreas cancer, gastrointestinal diseases, autoimmunological diseases, and developmental and genetic diseases using saliva specimens have been well documented [62-64].

During the (SARS-CoV) pandemic in 2003, saliva was already the focus of research for its use in diagnostic field in respiratory infections. Wang et al. examined samples of saliva and throat wash of 17 SARS patients and found that SARS-CoV RNA at highest amounts in saliva for all patients. Moreover, the authors were able to detect the presence of the virus in saliva samples at early stages even before the appearance of lung lesions, suggesting transmission by oral droplets by asymptomatic patients [65]. In another study involving rhesus macaques, after intranasal inoculations with SARS-CoV, the virus was detected in oral swabs of all animals, a few in lungs, but none in blood. This experiment suggests that coronavirus might appear in saliva before the infection reaches the lungs due to direct salivary gland infection [66]. Despite the little sample size, both studies indicated saliva collection as an accurate technique for the screening of SARS-CoV. In the identification of respiratory viruses, influenza, and respiratory syncytial virus, the diagnostic validity of saliva samples was reported to be comparable to nasopharyngeal swabs, demonstrating a 93% concordance to nasopharyngeal swabs with 90.8% sensitivity and 100%

specificity. In addition, coronavirus was detected only in saliva but not in nasopharyngeal aspirate for some patients. The authors of this study indicated that saliva can be used for the detection of respiratory viruses in sub-clinically infected patients. At the end, they assessed the average cost and time spent for the analysis of saliva samples and nasopharyngeal swabs and found that analyzing saliva samples was 2.26 times faster and 2.59 times cheaper, suggesting their use in bulk diagnosis and in research investigations [67].

The same scenario was observed with SARS-CoV-2. Several studies indicated that saliva samples have higher viral load or lower RT-PCR cycle threshold value compared to nasopharyngeal swab samples [68, 69]. In a study including 76 patients, Liu et al. showed that patients with severe symptoms tend to have higher viral load in saliva than patients with milder disease [70]. Asymptomatic patients also have detectable viral RNA in their oropharynx for at least five days, and more than 50% of asymptomatic patients had viral RNA detected in their saliva [71]. Key to the suitability of saliva as a specimen for Covid-19 screening is the presence of angiotensin-converting enzyme 2 (ACE2).

ACE2 is a vital protein that is critical to regulate normal body mechanism, such as blood pressure, wound healing and inflammation [72]. However, ACE2 also serves as the key cell-surface receptor in the human body for SARS-CoV-2. Similar to most other viruses, SARS-CoV-2 invades the host cells by attaching to the surface by recognizing the host cell surface receptor. The spike protein on the surface of SARS-CoV-2 binds easily to the host-cell receptor ACE2, providing the entry point for the virus to invade the host cells. ACE2 is present in many tissues including lungs, esophagus, colon, heart, blood vessels, kidney, liver, and bladder [73]. Several studies have shown that ACE2 is also present in salivary glands and the tongue. Expression of ACE2 in oral buccal and gingival tissue was found from paracarcinoma normal tissue. The same paper pointed out that ACE2 was highly enriched in epithelial cells of the tongue and also in epithelial cells, T cells, B cells, and fibroblasts of oral mucosa [74]. In another animal study involving rhesus macaques, ACE2 was expressed in epithelial cells lining minor salivary gland ducts, where the epithelial cells could also be detected in the sinonasal cavity, oral cavity, pharynx, larynx, trachea, and lungs [66]. Noteworthy, it was shown that the minor salivary glands express higher levels of ACE2 than that in the lungs. This latter statement suggests that the oral cavity is very susceptible to infection and that the salivary glands could be a major source of the virus in saliva. Furthermore, the same research group discovered that ACE2 epithelial cells of minor

salivary gland ducts were the first target cells for SARS-CoV-2 and that they attached to host cells as early as 48 hours after infection [66]. When the SARS-CoV-2 virus binds to ACE2, it prevents ACE2 from performing its normal function to regulate angiotensin II (ANG II) signaling. In the absence of ACE2 acting as brake for ANG II signaling, more ANG II are available to injure tissues [75]. Although viral load tends to be lower before symptom onset, SARS-CoV-2 RNA can be detected in the saliva before lung lesions emerge [76]. Table 2.1 summarized several studies investigating saliva and nasopharyngeal swab samples, as well as their comparison based on sensitivity and specificity, where it shows that the positive rate of SARS-CoV-2 detection in saliva can be up to 100%. Therefore, the confirmation of ACE2 expression in the epithelial cells of the salivary glands makes saliva a promising human specimen for SARS-CoV-2 detection investigation.

Saliva also contains important biomarkers that could serve as a basis for detection of SARS-CoV-2. Antibodies including immunoglobulin A (IgA), IgG and IgM are released in saliva and serve as important biomarkers for the physiological changes in saliva. Although IgG and IgM are lesser in quantity than IgA, oral mucosal transudate obtained by swabbing buccal mucosa and tongue provides a richer source of antibodies IgG and IgM, including those against bacterial and viral pathogens [77]. Furin is an enzyme that is highly expressed in lung tissue and detected by immunostaining in human tongue epithelia, possibly providing a gain-of-function to infectivity of SARS-CoV-2. In the human body, furin activates many proprotein substrates including pathogenic agents, growth factors, receptors, and extracellular matrix proteins [78]. It has been previously identified that furin is implicated in viral infections by cleaving viral envelope glycoproteins, enabling the virus to further invade the host cells. A furin-like cleavage site in the spike protein of SARS-CoV-2 has been identified, and several studies demonstrated the critical role of the furin cleavage site insertion in SARS-CoV-2 replication and pathogenesis [79, 80]. The aforementioned biomarkers are those that have been well studied during the course of the Covid-19 pandemic, other salivary biomarkers having the potential for SARS-CoV-2 diagnosis include alanine aminotransferase, C-reactive protein, neutrophil, lactate dehydrogenase, and serum urea [81].

The presence of all these biomarkers in saliva infected by SARS-CoV-2 has led to a stronger focus on saliva sample collection for the diagnosis of COVID-19. Several research studies evaluated the diagnostic efficiency of saliva compared with nasopharyngeal swabs and their results are summarized in Table 2.1. Most of the studies reported high concordance with nasopharyngeal

swab diagnostic results, and no statistically significant difference was observed between nasopharyngeal or sputum specimens regarding viral load. The low concordance rate in sensitivity is probably due to differences in the clinical background, such as difference in sampling method, sampling tools, and the different sampling time in each study. In fact, higher viral loads were detected in the early morning versus bedtime, and several studies reported reducing viral load in saliva with time [69, 71, 82-85]. Cough-out saliva from throat is mainly sputum from the lower respiratory tract, and hence may provide a more accurate specimen for virus detection instead of saliva fluid secreted from the opening of salivary gland canals [86]. All these studies suggested saliva could be a reliable non-invasive specimen for the diagnosis and viral load monitoring of SARS- CoV-2. In brief, the advantages of using saliva as specimen for virus detection include avoiding the discomfort of patients during sample acquisition, the potential of self-collection of specimens outside of hospital, decrease of the infection risk of healthcare workers, and the cost-effectiveness compared to traditional nasopharyngeal swab. Furthermore, using saliva as diagnostic specimen enlarges the possibility of using strategies other than direct detection of the viral pathogen RNA. Alternatives would be the detection of important biomarkers such as antibodies, cytokines, chemokines, and other bio-analytes [87]. With the background of knowing the reformation of biomolecules in saliva that will develop once in contact with the SARS-CoV-2, FTIR spectroscopy may demonstrate high effectiveness as a rapid diagnostic method by identifying the presence of SARS-CoV-2 based on the molecular concentration and composition change in saliva.

In fact, several authors have demonstrated FTIR spectroscopy as an effective tool for the detection of SARS-CoV-2 and suggested its use as a pre-screening method prior to genotypic method. Biofluids including saliva, nasopharyngeal swab, and blood were employed as samples for Covid-19 detection, with saliva as the most used specimen, possibly due to ease of collection. With regards to FTIR sampling preparation technique used for the detection of Covid-19, most of studies applied ATR accessory except one study used transflectance mode. Furthermore, several different chemometrics algorithms were combined to FTIR spectroscopy and compared to each other. As the use of spectroscopic technique in the diagnosis of Covid-19 is recent, it is no wonder researchers were trying multiple approaches to reach the best conclusion. Most of these studies were able to achieve high rate of correct identification, despite using different analytical methods and relatively low number of samples ($n < 300$ for most), with the exception of one study using

over 1000 samples [88]. Nevertheless, the dataset of the latter study included a number of unbalanced healthy and Covid-19 positive population [88]. Another issue is the inconsistency of the collection method. While Barauna et al. used saliva cotton swab for spectral acquisition, Martinez et al. used self-collected saliva in sterile tube of fasting patients eight hours prior to saliva collection [88, 89]. On the other hand, without giving prior instructions, Nascimento et al. and Wood et al. asked for self-collected saliva, with the former in sterile tube, and the latter in viral transport medium [90, 91]. The non-uniformity saliva collection technique limits the parallel comparison of results among these studies. Yet the fact they all reached a high prediction rate implies the effectiveness of FTIR spectroscopy in the diagnosis of Covid-19. Recent studies of Covid-19 detection using FTIR spectroscopy are summarized in Table 2.2.

Table 2.1. Recent clinical findings using saliva as biospecimen for COVID-19 detection compared to nasopharyngeal swab.

Population		Method	Result reported	Ref.
Number (m/f)	Mean age (range)			
25 (17/8)	61.5 (39-85)	Drooling technique used for saliva collection	SARS-CoV-2 was detected in all 25 patients' salivary sample.	[92]
2 (2/0)	64 and 71	Pipetted saliva and drooling techniques, respectively	They both had negative respiratory swab test, but positive salivary samples at the same time.	[93]
119	53.5 (33.7-73.3)	Drooling technique used for saliva collection.	They recorded a high sensitivity (93%) and a specificity of 42% of the rapid test saliva. 57% of the false positive cases had their saliva positive also when analyzed with rRT-PCR, which means that the virus was actually present and that the nasopharyngeal swab was less sensitive in these cases.	[94]
15	N/A	Saliva	5/15 positive in saliva samples. Possibly a higher viral load in saliva.	[95]
70	N/A	Self-collected saliva	80.0% tested positive with swabs and 68.6% with saliva. Thirty-four participants (48.6%) tested positive on both swab and saliva samples	[96]
27	29 (16-60)	Saliva	SARS-CoV-2 RNA was detected in 20/27 (74%) available saliva; 7/11 (64%) in the asymptomatic and 13/16 (81%) in the symptomatic group (P=0.56).	[97]
31 (15/16)	60.6 (18-86)	Stimulated salivary collection by gentle massage of salivary gland.	13 cases tested positive for viral nucleic acid extraction. Out of the 13 cases, 4 tested positive for nucleic acid extraction of saliva For critically ill patients, saliva has a higher potential for detection of SARS-CoV-2 (75%, 3 out of 4).	[98]
58 (28/30)	38 (31-52)	Posterior oropharyngeal saliva	84.5% (49/58) tested positive in both nasopharyngeal swab and saliva, 10.3% (6/58) tested positive in nasopharyngeal swab only, and 5.2% (3/58) tested positive in saliva only.	[68]
1	N/A	Self-collected saliva	Viral load of 3.3×10^6 copies/mL (pooled nasopharyngeal and throat swabs) and 5.9×10^6 copies/mL (saliva).	[86]
32 (16/16)	41 (34-54)	Saliva	25/32 positive in saliva samples. Saliva in non-ICU and ICU patients took 13.33±5.27 and 16.50±6.19 days separately to converse to negative.	[99]
1 (0/1)	27-day-ol neonate	Saliva	The SARS-CoV-2 was detected in all of the neonate's clinical specimens, including blood, urine, stool, and saliva along with the upper respiratory tract specimens.	[82]

11	6.5 (27-day-old-16-years-ol)	Saliva	8 (73%) tested positive. Positivity in saliva samples was 80% in week 1 but dropped sharply to 33% in week 2 and 11% in week 3.	[100]
368 (195/173)	35 (18-75)	Self-collected specimen, spitting in tubes	Positive agreement between NPS and saliva was 93.8%. Negative agreement was 97.8% for NPS versus saliva	[101]
16 (6/10)	18-61 (interquartile range 22.75-53)	Posterior oropharyngeal saliva	Overall trend of lower Ct values from specimens collected in the early morning, with a gradual decrease of viral load towards nighttime, but reaching statistical significance only when compared with the specimens collected at bedtime. Eight out of 13 subjects had a higher viral load in the early morning than the rest.	[84]
10	69 (30-97)	Self-collected specimen, spitting in tubes	SARS-CoV-2 was detected in 8/10 patients in both nasopharyngeal and saliva samples, and in either sample only in 2/10 patients. The overall concordance rate of the virus detection was 97.4% (95%CI, 90.8-99.7).	[85]
53 (32/21)	63 (27-106)	Self-collected specimen, spitting in tubes	Sensitivity was 89% for nasopharyngeal swabs and 77% for saliva. Of 53 patients with paired specimens tested, 47 (89%) had at least one positive specimen. In 31 (66%) of these 47 patients, both nasopharyngeal swab and saliva were positive, in 11 (23%) only the nasopharyngeal swab was positive, and in 5 (11%) only saliva was positive.	[102]
35	N/A	Pure saliva	33/35 positive by NPS (sensitivity = 94.3 %) and 30/35 by pure saliva (sensitivity = 85.7 %), for an overall agreement of 117/124 (94.4 %).	[103]
95 (26/69)	42 (19-85)	Posterior oropharyngeal saliva	Overall agreement (95% CI) 78.9%.	[104]
45	N/A	Throat saliva	40% (18/45) for all sample sensitivity; 53.8% (14/26) for high viral load samples	[105]
156 (90/66)	47.8	Saliva, not sputum	47/49 samples were positive in saliva compared with the nasopharyngeal swab, resulting in a positive percent agreement of 96% (95% CI, 86.02% to 99.5%). A total of 105/106 samples had a negative saliva and NPS result, resulting in a negative percent agreement of 99% (95% CI, 94.86% to 99.98%).	[106]
44	N/A	Self-collected saliva	34 (77.3 %) had both samples positive, 3 (6.8 %) were only positive in saliva sample, and 7 (15.9 %) with only positive NP swabs. Saliva samples detected 37/44 (82.2 %) patients, while the NP swabs detected 41/44 (93.2 %) patients	[107]
34	N/A	Saliva	Positive and negative agreement with third-party laboratory results were reported as 97.1% and 96.5-98.2%, respectively. Limit of detection was established at 5 copies/µL. Stability through simulated shipping conditions found 100% concordance up to 56 hours after collection.	[108]
103 (66/37)	46 (18-87)	Self-collected specimen, spitting in tubes	All patients with severe disease (16/16, 100%) tested positive for viral RNA in their saliva, while 58 of 72 (78.4%) patients with mild disease tested positive (P 0.064).	[71]
149 (46/103)	40 (33-48.5)	Self-collected spitting, avoiding sputum	The sensitivity and specificity of RT-PCR using saliva samples were 94.4% (95% CI 86.4–97.8) and 97.62% (95% CI 91.7–99.3), respectively. There was an overall high agreement (96.1%) between the two tests.	[109]
200 (69/231)	36 (28-48)	Saliva sample voiding coughs	The sensitivity and specificity of the saliva sample RT-PCR were 84.2% (95% CI 60.4%–96.6%), and 98.9% (95% CI 96.1%–99.9%), respectively. An analysis of the agreement	[110]

			between the two specimens demonstrated 97.5% observed agreement (κ coefficient 0.851, 95% CI 0.723–0.979; $p < 0.001$)	
39 (14/25)	44 (18-82)	An enhanced saliva specimen (strong sniff, elicited cough, and collection of saliva/secretions)	Of the 216 patients, there was a 100% positive percent agreement (38/38 positive specimens) and 99.4% negative percent agreement (177/178 negative specimens).	[111]
160 (160/0)	27 (18-36)	Self-collected deep throat saliva sample	The detection rate for SARS-CoV-2 was higher in saliva compared to nasopharyngeal swab testing (93.1%, 149/160 vs 52.5%, 84/160, $p<0.001$). The concordance between the two tests was 45.6% (virus was detected in both saliva and NPS in 73/160), while 47.5% were discordant (87/160 tested positive for one while negative for the other).	[112]
18	N/A	Self-collected specimen, spitting in tubes	Saliva was positive for 15/18 patients, with a sensitivity and specificity of 83.3% and 99.1%	[113]
84 (45/39)	44 (20-79)	Rinse out saliva	Clinical sensitivity of nasopharyngeal specimens were 85%, throat 80%, midturbinate 62%, and saliva 38%-52%.	[114]
1 (1/0)	71	Sefl-collected specimen, spitting in tubes	Early morning saliva specimens were more likely to show positive results than those obtained later in the day.	[115]
12 (7/5)	62.5 (37-75)	Self-collection of coughed out saliva Specimens collected after median of 2 days of hospitalization.	SARS-CoV-2 (2019-nCoV) detected in the initial saliva samples of 11 out of the 12 patients (91.7%). Median viral load in first available specimens: 3.3×10^6copies/ mL. Viral load found to be highest in earliest available saliva specimens (83.3%).	[116]
23 (13/10)	62 (35-75)	Saliva sample by cough-ing and clearing the throat early morning before breakfast	Salivary viral load was highest during the first week after symptom onset and subsequently declined with time	[117]
32	N/A	Saliva	32 samples were found positive for both saliva and nasopharyngeal swab samples (N+S+), while 138 were negative for both (N−S−). 15 samples were positive for nasopharyngeal swab samples and negative for saliva samples (N+S−), and 11 samples were positive for saliva samples and negative for nasopharyngeal swab samples (N−S+). Overall, saliva and nasopharyngeal swab samples displayed 86.7% concordance with kappa coefficient as 0.625.	[118]
39	N/A	Self-collected saliva	The SARS-CoV-2 was detected in saliva specimens of 33/39 patients (84.6%; 95% CI: 70.0–93.1%). The SARS-CoV-2 was detected in 1 saliva specimen among 50 PCR negative nasopharyngeal swabs. The viral load of nasopharyngeal swabs is higher than that of saliva.	[69]

95 (57/38)	36 (4-92)	Posterior oropharyngeal saliva	The overall negative and positive percent agreement were 76.0% (95% CI 70.2–80.9%), 65.4% (95% CI 55.5–74.2%), 85.2% (95% CI 77.4–90.8%). Better positive percent agreement was observed in POPS-NP specimen obtained within 7 days (96.6%, 95% CI 87.3–99.4%) compared with after 7 days of symptom onset (75.0%, 95% CI 61.4–85.2%).	[119]
38 (21/17)	59 (23-91)	Saliva	SARS-CoV-2 was detected from the saliva but not in the nasopharyngeal swabs from eight matching samples (21%); while SARS-CoV-2 was detected only from nasopharyngeal swabs and not saliva from three matched samples.	[120]
1 (1/0)	44	Saliva	The viral RNA was detected in multiple types of specimens with extremely high titers in the saliva.	[121]
1924	N/A	Saliva	the sensitivity of RT-PCR using nasopharyngeal and saliva specimens were 86% and 92%, respectively, with specificities greater than 99.9%. The true concordance probability between them is 0.998	[122]
2 (0/2)	46 and 65	Self-collected saliva	The viral load was the highest in the nasopharynx (patient 1 = 8.41 $\log_{10}$ copies/mL; patient 2 = 7.49 $\log_{10}$ copies/mL), but it was also remarkably high in the saliva (patient 1 = 6.63 $\log_{10}$ copies/mL; patient 2 = 7.10 $\log_{10}$ copies/mL)	[123]
15	N/A	Oral swab	Out of 15 COVID-19 patients, 8 had positive oral swab (53.5%).	[83]
65 (40/25)	54 (39.5-62)	Patients produced saliva by coughing three to five times (wearing a mask) and spitting into a sterile container	37 out of 42 (88.09%) salivary samples detected SARS-CoV-2 as compared with the detection rates of throat swabs (45.24%, 19/24), and nasal swabs (76.19%, 32/42). Significantly higher viral loads were detected in saliva samples as compared with throat swabs.	[124]

Table 2.2. Recent studies using FTIR spectroscopy for the detection of Covid-19.

Sample	No. of samples	FTIR technique	Analysis method	Result reported	Ref.
Saliva	61 negatives, 20 positives	ATR-FTIR	GA-LDA	95% sensitivity and 89% specificity	[89]
Saliva	99 negatives, 138 positives	ATR-FTIR	URF, GA-LDA, SPA-LDA, PLS-DA, PSO-PLS-DA, consensus class	85% accuracy, 93% sensitivity, and 83% specificity	[90]
Saliva	1209 negatives, 255 positives	ATR-FTIR	Mann–Whitney test, Kruskal–Wallis test, MLRM	99.6% accuracy, 99.2% sensitivity, and 100% specificity	[88]
Saliva	28 negatives, 29 positives	Transflection-FTIR	Monte Carlo Double Cross Validation	93 % sensitivity and 82% specificity	[91]
Blood serum	11 negatives, 26 positives	ATR-FTIR	PLS, RF, SDT, DNN	96.30%–100% sensitivity and 91.67%–100% specificity	[125]
Blood plasma	160 positives (69 severe, 91 non-severe)	ATR-FTIR	PLS-DA	94.1% sensitivity and 69.2% specificity (classification of severeness)	[126]
Blood serum	20 negatives, 76 positives	ATR-FTIR	HCA, PCA, PLS-DA	87% sensitivity and 98% specificity	[127]
RNA extract of nasopharyngeal swab	180 negatives, 100 positives	ATR-FTIR	PCA, PLS, Logistic regression, SVM, Kernel SVM, DA	97.8% accuracy, 97% sensitivity and 98.3% specificity	[128]
Nasopharyngeal swab	92 negatives, 151 positives	ATR-FTIR	PLS, KNN	84% and 87% sensitivity, 66% and 64% specificity, and 76.9% and 78.4% accuracy	[129]

(ATR-FTIR: Attenuated Total Reflectance Fourier Transform Infrared; GA-LDA: Genetic Algorithm Linear Discriminant Analysis; URF: Unsupervised Random Forest; LDA: Linear Discriminant Analysis: SPA-LDA: Successive Projection Algorithm Linear Discriminant Analysis; PLS-DA: Partial Least Squares Discriminant Analysis; PSO-PLS-DA: Particle Swarm Optimization Partial Least Squares Discriminant Analysis; MLRM: Multiple Linear Regression Model; RF: Random Forest; SDT: Single Decision Tree; DNN: Deep Neural Networks; HCA: Hierarchical Cluster Analysis; PCA: Principal Component Analysis; SVM: Support Vector Machine; KNN: K-Nearest Neighbor

2.4. Identification Methods

In depth understanding of the major pathogens, their incidence and their major routes of infection, health officials will be able to implement preventive measure and provide thorough surveillance that could highly improve food safety. Identification of pathogens at the species level is essential to improve infection control during outbreaks. Therefore, a rapid, cost-effective, and accurate identification method for microbial pathogens in clinical settings and food industry is essential. In general, bacteria identification methods can be grouped into three categories: (1) phenotypic method, (2) genotypic method, and (3) spectroscopic method.

2.4.1. Phenotypic method

Phenotypic methods are the traditional method of microbial identification. Nowadays, although genotypic methods have been implemented in most laboratories, conventional phenotypic methods remain as the first identification method to proceed in diagnostic laboratory due to their low costs and little training of the personnel required. Phenotypic methods allow identification to the genus and species level, depending on the type of bacteria and the test used. In most cases, phenotypic identification uses a combination of more than one method. In broad, phenotypic methods can be classified into three categories, biotyping, serotyping, and phage typing.

In biotyping, bacterial cell's physiological aspects, such as colony and cell morphology, Gram's stain, cell wall and membrane composition, catalase test and a lot more are investigated. Morphological investigation can be performed by light and electron microscopy, providing information on membrane composition and flagella [130]. Environmental growth conditions can also be monitored, including different pHs, temperature, salt tolerance, oxygen requirement, antibiotic resistance, and bacteriocins susceptibility [131]. Among the tests mentioned above, fatty-acid composition of bacteria is stable and has been used to identify bacteria at the genus and species level. After bacteria growth, the fatty acids are extracted, and the methyl esters are determined by gas chromatography [132]. Another recent biochemical method, chromogenic substrates, utilizes specific enzymatic activities targeting different bacteria are increasingly used in clinical and food microbiology laboratories for the detection and identification of microorganisms. The incorporation of such substrates into a selective or non-selective growth medium does not need any further biochemical test, and can readily identify certain bacteria, such as *Bacillus* spp., *E. coli.*, *Staphylococcus* spp., *Streptococcus* spp., and yeast [133-137].

Serotyping is an immunoassay based on the agglutination of bacteria, namely antibody-antigen reactions, that can be employed to detect unique cellular determinants of specific microorganisms. One of the standard serotyping methods is enzyme-linked immunosorbent assay (ELISA). It is extensively used for *Salmonella* spp., *E. coli* and *Campylobacter* spp., and also several Gram-positive bacteria, such as *Listeria* spp. [138-140]. Serotyping can also target toxins, showing ability to detect botulinum neurotoxins [141].

Phage typing is based on the ability of a given bacterial phage to lyse bacteria cells. It is usually used to discriminate between *Salmonella* strains of the same serovar. This method can be used in a single or a mixed culture, as host specificity allows both detection and identification. However, ambiguous lysis reaction is a major drawback, and the assay requires careful coordination between reference laboratories to ensure reproducibility. In addition, this method is limited to the number of available phages [142].

Identification is done by the determination of the biochemical profile of a microorganism. Numerous multi-test system equipment is available in the market making phenotypic identification method fast and reliable, although these techniques require pure bacterial culture. The inoculation method, incubation time, and test reading techniques may lead to identification errors if specific procedures were not followed. The active pharmaceutical ingredient (API) strip is an example of a commercialized phenotypic identification system. The API system consists of strips of micro-tubes containing different kinds of dry substrates, and the reaction of the unknown with the substrates can be observed after incubation. This instrument was further refined with the use of Vitek automated system that miniaturized the process [143]. Biolog Inc. offers a second phenotypic identification system. The fundamental unit in this system is a 96-well plate that has different carbohydrate sources in each well, with a tetrazolium redox dye. If the microorganism is able to degrade the carbohydrate substrate, the well changes color [144]. Other commercially available kits include Crystal TM, and BD Phoenix TM. Most identification kits are simple to perform, but the interpretation of results could be subjective. Another limitation of phenotypic methods is that they are not always able to identify the bacteria down to the species and strain level. Mutating strains might show completely different physiological characteristics. In case of an outbreak, where the aim is to determine gene sequence, genotypic (molecular) methods could be

a better alternative to phenotypic methods in order to establish the identity of a microorganism and to perform epidemiological research.

Another phenotypic method that is worth mentioning is Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). Since 2010, the identification of bacterial pathogens has been revolutionized by its introduction, and this is now the method of choice for bacterial identification in most advanced clinical laboratories. MALDI-TOF MS uses the mass of proteins profile present in a microorganism as a signature for its identification, through the comparison of the mass spectrum from an unknown to a mass spectral reference database of well-defined microorganisms. MALDI-TOF-MS gained more attention as their ability to identify microorganisms at species level in a rapid and cost-effective manner. In MALDI-TOF MS, the sample is deposited within a matrix. After absorbing energy from ultraviolet light (nitrogen laser light, wavelength 337 nm), resulting a rapid heating, vaporization and ionization of a small part of the surface of matrix together with the sample. Since all ions are given the same kinetic energy, the time for ions to reach the detector defers based on their mass-to-charge ratio by the law of conservation of energy [145].

MALDI-TOF MS is based on ribosomal proteins detection, which generate a fingerprint spectrum, so called peptide mass fingerprint (PMF). For the identification purpose, PMF of the unknown is compared to a database, and assigned to the identity that the spectrum matches the most. In a comparison study, MALDI-TOF MS outperformed API and Automated Reading and Incubation System 2x System (ARIS) [146]. Another comparative study reported that MALDI-TOF MS showed significantly better performance (93.2%) over BD Phoenix (BD Diagnostic Systems, France) (75.6%) and Vitek-2 (bioMerieux, Marcy l'Etoile, France) (75.2%) for the identification of 234 CoNS belonging to 20 different species [147]. Despite showing 90% accuracy for clinical bacteria and fungal identification at species level, the robustness and discriminatory power highly depend on a unified protocol, such as sample preparation and data analysis. Yet, inherent difference in peak intensity or location tend to persist on the mass spectrum, due to acquisition time, environmental factors, and differences in devices or laboratories. Moreover, MALDI-TOF-MS may experience difficulties in identifying species with a low rate of differences in their ribosomal protein sequences. The most problematic identifications encountered include the viridans *Streptococci* and *Pneumococci*, the pathogenic *Shigella* spp., the commensal *E. coli*,

and anaerobic bacteria. Viridans *Streptococci* and *Pneumococci* were misidentified mainly due to an incomplete database reference library [148]. However, it remains challenging for MALDI-TOF to differentiate closely related species of *Streptococcus*, such as *S. pneumoniae*, *S. mitis*, and *S. parasanguinis*, and bacteria having almost identical mass spectrum, such as *E. coli* and *Shigella* [149]. Interestingly, FTIR spectroscopy has shown to correctly discriminate *Shigella sonnei* and *E. coli* O157:H7, due to the fact that FTIR acquire spectra over a broader range of biomarker features than MALDI-TOF MS [150]. FTIR spectroscopy is more known in its use for routine quantitative analysis such as adulteration and authentication of food [151]. However, FTIR spectroscopy has shown to have advantage over MALDI-TOF MS for bacteria identification in food sample, despite the fact MALDI-TOF MS is more widely recognized for identification. Schabauer et al. obtained 100% correct species identification rate for FTIR spectroscopy, and 90.5% for MALDI-TOF in the identification of *Streptococcus* spp. [152]. Currently, MALDI-TOF MS has shed light for other spectroscopic method, like FTIR spectroscopy and Raman spectroscopy, to be recognized as a diagnostic tool. Raman spectroscopy is also used to for identification of microorganisms. This instrument depends on a change in polarizability of a molecule and measures the relative frequencies in which a sample scatters radiation. However, the high potential of FTIR spectroscopy may outperform or provide similar accuracy to MALDI-TOF MS and Raman, with a significantly lower cost.

2.4.2. Genotypic method

Although phenotypic methods are preferred due to their low costs, the expression of microbial phenotype, including cell size and shape, cellular composition, antigen, biochemical activity and antimicrobial sensitivity is dependent on the media and growth medium used. In contrast, genotypic methods are used for the microbial genome study by exploring the genetic material in order to differentiate among closely related strains. Hence, the application of molecular biology techniques based on DNA or RNA analysis, are less subjective, less dependent on culture conditions, and more reliable [153]. Genotypic methods are considered nowadays as the ‘gold standard’ for identification due to their high level of accuracy.

Genotypic methods allow microbiologists to study gene expression of bacteria, typing species and subspecies that were previously thought to be indifferent from other species within the genus, for instance, the re-classification of *Enterobacter sakazakii* into a new genus, *Citrobacter*

sakazakii [154]. Overall, genotypic methods use either hybridization or sequencing techniques. In hybridization, microbiologists are able to explain how well two strands of DNA from different bacteria bind together, hence determining the relatedness of the two microorganisms. Sequencing of 16s rRNA is the target sequence of bacteria due to the high conservation by microbial species, and its vast amount of genetic information included [155]. The most common techniques are the DNA-DNA hybridization, PCR, PFGE, MLST, genetic fingerprinting (ribotyping), and 16s and 23s rRNA gene sequencing [4]. The application of nucleic acid amplification methods in routine detection of microorganisms has been limited due to the laborious standardization and validation procedures. Moreover, these methods are labor intensive and more technically challenging for technicians, and require more expensive equipment and supplies [153]. For instance, in PCR, the quality of the DNA template, the equipment, personal practice, the reaction conditions and the reaction materials, and the environment [156].

Whole genome sequencing (WGS) is the leading genotypic method by its ability to identify and characterize bacteria through very subtle differences between genome sequences. WGS allows for the identification of pathogens, the exact profiling of resistance genes, recognition of outbreak strains, non species-specific targeting without the requirement for continuous development of probes and primers, and the immediate design of PCR probes based on the generated genetic data in the event of an outbreak [148]. The entire genome sequence data offers a much higher resolution than other genotyping methods. Since the publication of the first-generation sequencing techniques in the late 1970s, a variety of rapid and cost-effective technologies have become available. Up to date, the first-generation shotgun sequencing, second generation massively parallel sequencing and third generation single-molecule sequencing are all in use by many laboratories. Due to its usefulness, WGS has been rapidly incorporated in PulseNet, a critical surveillance system targeting foodborne disease outbreaks, for foodborne pathogen subtyping in addition to PFGE, as many of the limitations encountered with PFGE could be easily overcome by WGS. In one circumstance, more than 50% of *Salmonella enteritidis* isolates show identical PFGE types, but WGS was able to discriminate these isolates and identified outbreaks that otherwise would not be detected by PFGE [157]. In another case, WGS successfully identified an *L. monocytogenes* isolate as part of a seasonal food product outbreak, whereas PFGE failed to associate it [158]. In general, WGS allows high discriminatory power and characterization of relatedness of isolates, which was not possible with PFGE. However, since WGS provides too detailed information on the entire

genome, abnormalities such as substitutions, insertions, deletions, duplications, and chromosome translocations are also included, which might lead to overfitting or data in excess for identification. Furthermore, although the time and costs of the method have significantly reduced over the last decade, it still may take up to 10 days to obtain a WGS report costing $1000 per sample [158]. Moreover, personnel training is also required to be able to interpret WGS data. Therefore, implementing WGS in routine surveillance require extremely high economic output and computational power to process and analyze the great amount of data.

Commercially available automated genotypic systems based on PCR include Riboprinter, which uses labeled ssDNA probe from the 16s rRNA codon, creating pattern for identification of the unknown [159]. MicroSeq 500 16S rDNA Bacterial Sequencing Kit offered by Applied Biosystems is another genotypic identification system. As its name implies, this latter provides the materials needed to sequence the first 500 base pair of the unknown microorganism's 16s rRNA codon [158]. Another genotypic method being marketed is the Bacterial Barcodes system. This system uses a sequence homologous to a repetitive sequence in the unknown bacterial genome. Then, the amplified sequence is separated by gel electrophoresis and visualized by giving a 'barcode' specific to that strain [160].

Genotyping method and phenotyping method have never been either or, but rather working in combination for a thorough understanding of a particular microorganism, especially during an outbreak or for epidemiological investigation, where cell morphology and biochemical reaction, as well as gene expression are all important factors to consider through. Even though genotypic methods are considered as the gold standard for microbial identification, especially due to their high sensitivity and specificity, spectroscopic methods are gaining more attention nowadays despite lacking standardized protocols.

2.4.3. Spectroscopic methods

Spectroscopic methods include a large number of techniques that use radiation to obtain information on the structure and properties of a sample. Basically, all spectroscopic identification methods share the same principle of emitting a beam of electromagnetic radiation onto a sample and observing how the sample would respond to such stimulus. Examples include UV-visible light, nuclear magnetic resonance (NMR), Raman and FTIR spectroscopy. UV-visible light spectroscopy is generally used for quantitative analysis or presence/absence analysis of

microorganisms [161]. NMR spectroscopy provide selectivity and specificity for identification as well as characterization. However, this method requires complex sample preparation and trained personnel to operate, it is also labor-intensive and costly. On the other hand, vibrational spectroscopic techniques are less expensive, and their fingerprinting capabilities provide accurate results along with several advantages including speed, non-destructive and easier to handle.

Vibrational spectroscopy including FTIR and Raman spectroscopy are based on the transition between quantized vibrational energy states of molecules due to the interaction between the material and the radiation from a light source. However, they differ in some key fundamental ways. Raman spectroscopy depends on a change in polarizability of a molecule and measures relative frequencies at which a sample scatters radiation, whereas FTIR spectroscopy depends on a change in the dipole moment and measures absolute frequencies at which a sample absorbs radiation. In other words, FTIR spectroscopy is sensitive to hetero-nuclear functional group vibrations but is unable to detect homo-nuclear diatomic molecules such as Cl_2, H_2, and O_2 [162]. Raman spectroscopy, on the other hand, is sensitive to homo-nuclear molecular bonds. Hence, FTIR and Raman spectroscopy are complementary to each other. Both requires little to no sample preparation, while Raman spectroscopy is relatively less affected by water content, which makes it advantageous more suitable for the analysis of complex heterogeneous materials. However, fluorescence may interfere with the ability of acquiring Raman spectra, which would not be an issue for FTIR spectroscopy. And since Raman technique requires highly stable laser sources and sensitive amplification equipment to detect weak signal, it is much more expensive compared with a FTIR spectrometer [163, 164]. Despite their differences, both techniques could be used for identification and characterization, and they are often used in a complementary way in research.

2.5. FTIR Spectroscopy

FTIR spectroscopy involves the interaction between electromagnetic radiation and matter. Covalent bonds absorb energy from electromagnetic radiation when the radiant energy matches the energy of that specific molecular vibration of the bond. The vibrational modes can be either stretching (change in bond length) or bending (change in bond angle); stretching can in turn be symmetrical (in-plane) or asymmetrical (out-of-plane), and the bending vibration is identified as rock (same direction) or deformation (opposite direction) [165]. As specific wavelengths are absorbed according to vibration by specific chemical bonds present in a sample, an infrared (IR)

spectrum is obtained by calculating the intensity of the IR radiation before and after it passes through a sample. The correlation between IR band positions and intensity vs. chemical structure of the sample provide qualitative information such as presenting functional groups, and quantitative information such as concentration of bacteria in a growth medium. For instance, when a bacteria sample is placed in the IR beam, selective wavelengths are absorbed yielding an IR spectrum that reflects the microorganism's chemical composition. This information is exploited to delineate taxonomic differences between the microorganisms, as well as to detect chemical changes within the microorganisms resulting from its exposure to stressful environments. This IR spectral information would be stored in a database as reference for further analysis. These features of FTIR spectroscopy allow differentiation of the microorganism at the genus, species, strains and serotypes levels [166]. Using an efficient database combined with appropriate analysis algorithms, FTIR spectroscopy could provide accurate results in a matter of minutes. The identification of subspecies and strains depends heavily on the availability of sufficient spectral database that encompasses microbial diversity.

The infrared portion of the electromagnetic spectrum can be broadly divided into three regions: (1) the near-infrared (NIR) (12820 – 4000 cm^{-1}, to which is poor in specific absorptions, can excite overtone or harmonic vibrations, and could be very useful for quantitative analysis; (2) the mid-infrared (MIR) (4000 – 400 cm^{-1}) provides structural information for most organic molecules, and may be used to study the fundamental vibrations and associated rotational-vibrational structure; (3) the far-infrared (FIR) (400 – 33 cm^{-1}) which is lying adjacent to the microwave region with lower energy may be used to study vibrations of molecules containing heavy atoms, molecular skeleton and crystal lattice [167]. With decreasing energy, the three infrared regions are useful for different applications. While NIR and MIR are often used separately or together to study microorganisms, FIR is more applied for medical use and rarely used for microorganism identification. This is because that in contrast to MIR region where most spectral features arise due to intramolecular vibrational modes, FIR spectral features can be due to a number of different types of transitions, such as torsional and ring-puckering modes, or intermolecular modes involving hydrogen bonds and charge-transfer species. Furthermore, water vapor shows strong FIR absorption, and the spectral features can be distorted by water vapor interference unless special precautions are taken during measurement [168]. Although FIR wavelength is too long to be perceived by the naked eyes, the body can experience its energy as a gentle radiant heat which

can penetrate up to 1.5 inches beneath the skin. The resulting epidermal temperature is higher when the skin is irradiated with FIR than thermal loads from shorter wavelengths like NIR and MIR. Even if the mechanism of the thermal effect and biological activities of FIR radiation are poorly understood, many instruments such as specialty lamps and saunas, which deliver pure FIR radiation, have become commercially available and widely used as a safe and effective treatment tool mainly for pain and stress relief [169]. FIR was also reported to be used for microbial decontamination in food products when combined with UV light [170].

The main difference between MIR and NIR is that absorption in MIR region corresponds to fundamental bands of molecular vibrations, whereas absorption in NIR, correspond to overtones and combinations of these fundamental bands. This characteristic makes NIR not as sensitive as MIR. NIR bands are approximately 10-100 times less intense than MIR bands. Furthermore, the broad overtone and combination bands makes it difficult to identify and associate them with specific chemical group. Therefore, efficient calibration techniques are often required for NIR analysis [171]. Another point is that diffusion of light is much greater in the NIR than in the MIR range. Hence, NIR spectra will be much more easily affected by factors which affect the diffusion of light such as the physical structure (size of aggregates, porosity), and the presence of water which changes the refractive index and therefore the diffusion of light [172]. However, NIR can be very useful in direct analysis of highly absorbing bulk and porous samples with less sample preparation required than MIR, and is best fitted for in-field analysis, with lesser specificity requirements. On the other hand, the MIR spectra (4000 – 400 cm^{-1}) is the most studied region for analysis for organic compounds as they possess characteristic absorbance frequencies, and primary molecular vibrations are all found in this range. Moreover, MIR generally shows a better specificity and reproducibility than FIR and NIR.

Typically, an infrared spectrum of a biological material presents characteristic bands due to lipids (3050–2800 cm^{-1}), amide region ascribed to proteins and peptides (1700–1500 cm^{-1}), the mixed region designated to carboxylic groups of proteins, free amino acids and polysaccharides (1500–1250 cm^{-1}), polysaccharides region (1200–900 cm^{-1}). Other spectra regions of interest include 1250-1200 cm^{-1} where phospholipids, DNA and RNA are found, and the fingerprint region (900–600 cm^{-1}) [173]. Notably, the fingerprint region represents bands composed of unique broad and complex contours which is specific to the molecular structure of the sample.

illustrates a transflectance spectra with several important regions highlighted, and Table 2.3 describes assignment of bands frequently found in microbial IR spectra.

Table 2.3. FTIR spectra characteristics band absorption.

Spectral Region	Wavenumber (cm^{-1})	Band Assignments
Fatty acids	2956	CH_3 asymmetric stretch
	2920	CH_2 asymmetric stretch
	2870	CH_3 asymmetric stretch
	2850	CH_2 asymmetric stretch
	1745-1735	C=O stretch (fatty acid esters)
Amide	1705	C=O stretch (esters, carboxylic groups)
	1652-1648	Amide I (C=O) different conformations
	1550-1548	Amide II (N-H, C-N)
	1460-1454	CH_2 bending
Mixed	1400-1398	C-O bending (carboxylate ions)
	1310-1240	Amide III (C-N)
	1240	P=O (phosphate)
	1222	P=O
	1114	C-O-P, P-O-P
Polysaccharide	1085	Sugar ring vibrations
	1052	C-O, C-O-C (polysaccharide)
Fingerprint	900-600	C-H bending

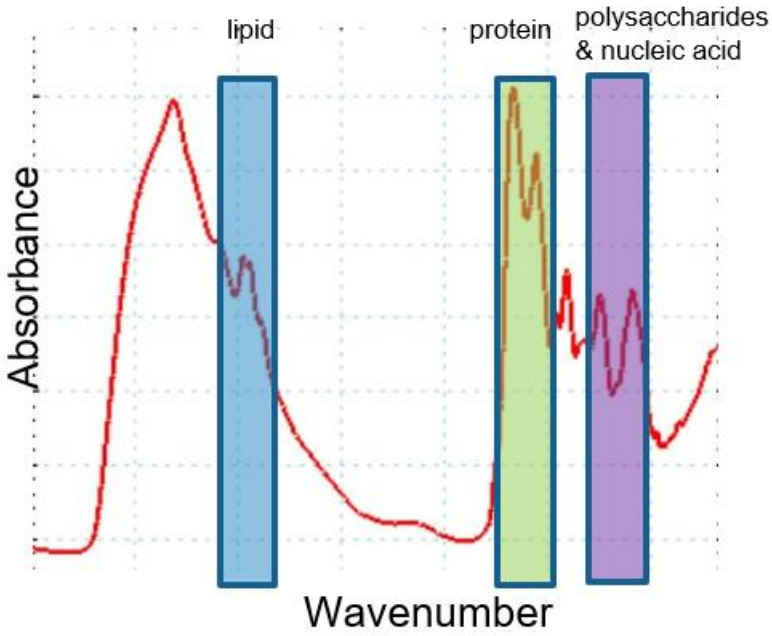


Figure 2.1. An example of a Transflectance spectra of a bacteria (*Staphylococcus aureus*) with lipid, protein, polysaccharides, and nucleic acid region highlighted.

The IR spectrum is generated based on the absorption proportion after the IR light passes through the sample. This extent of absorption is given by the Beer-Lambert Law, where T is the transmittance, I_S the intensity of the transmitted radiation after passing through the sample, I_R the intensity of incident radiation before reaching the sample.

$$T = \frac{I_S}{I_R}$$

A more general form of the Beer-Lambert law can be expressed by the absorbance (A) as following, where ε is the molar absorptivity (m^2mol^{-1} or $M^{-1}cm^{-1}$) or how strong a chemical species absorbs light at a given wavelength, L is the path length, which is the distance that the light travels through the chemical species, and c is molar concentration of chemical species.

$$A = \varepsilon Lc$$

In most cases, the IR spectrum is plotted by the intensity of absorbance (A) or transmittance (T) as a function of wavenumber, as their intensity at a given wavelength is directly proportional to the concentration of a sample [174]. A normal process when running an FTIR instrument include: (1) a source where the IR energy is emitted; (2) an interferometer so that the IR energy beam encodes into an interferogram; (3) a sample where the beam is either transmitted through or reflected off the surface of sample, depending on the type of analysis, and specific frequencies of energy are absorbed; (4) a detector that receives the beam for final measurement; (5) and a computer that digitized the signal for further analysis [175].

Figure 2.2 illustrates the principle of FTIR spectroscopy.

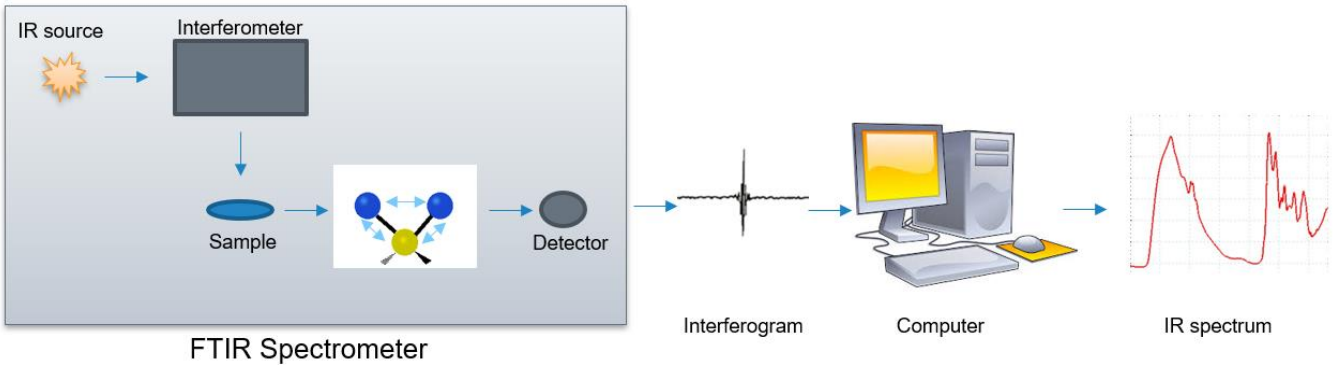


Figure 2.2. Working principle of an FTIR spectrometer.

Since the spectrum is obtained according to an absorbance percentage, a measurement with no sample as background spectrum is required as reference. Subtraction of the background spectra removes interferences of the instrument arising from noise and water vapor. Since each different material is a unique combination of atoms, no two compounds produce the exact same infrared spectrum. An infrared spectrum represents a fingerprint of a sample with absorption bands that correspond to the frequencies of vibrations between the bonds of the atoms that make up the material.

Due to its simplicity, FTIR spectroscopy is often used for characterization, quantification, identification, differentiation and classification of microorganisms, and it is utilized in an expansive range of application fields, including pharmaceuticals, clinical, food, environmental, and forensic industries [176]. Up to date, considerable studies have demonstrated the competitive performance of FTIR compared to conventional methods [177-179]. In one research done by our team, we have successfully achieved 100% and 99.7% of correct classification at genus and species level respectively, using FTIR spectroscopy for clinical yeast [180]. The advantages are particularly its simplicity to operate, no reagents required, non-destructive, non-invasive, rapid, automation potential, and most importantly, more cost-effective than MALDI-TOF MS and genotypic methods.

2.5.1. Spectral Acquisition Technique in FTIR

In general, there are two different sampling technique in FTIR spectroscopy, depending on the interaction of the IR beam with the sample, namely transmission and reflection. While transmission mode is common and is based on the measurement of the transmitted IR radiation, this method suffers from opacity problem, and it strictly requires sample to be 1 to 20 microns thick [181]. On the other hand, reflectance mode of FTIR have gained more attention in recent years. In brief, reflectance method relies on the reflection of the IR beam that is reflected after contacting the surface of the sample depending on the reflection process, and the two focused handling technique in reflectance are transflectance (TR) and attenuated total reflectance (ATR). A figure representation of the three different types of sampling techniques can be found in

Figure 2.3.

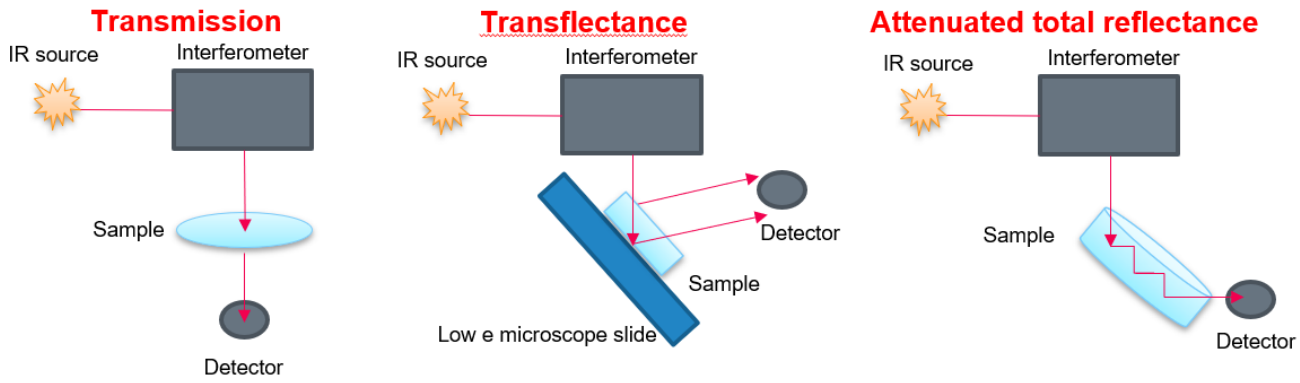


Figure 2.3. Common FTIR spectroscopy spectral acquisition modes, including Transmission, Transflectance, and Attenuated total reflectance.

2.5.1.1. Transflectance (reflection-absorption)

In transflectance (reflection-absorption), the sample is deposited on a highly reflective substrate and some of the IR beam passes through the surface layer, reflecting off the top layer of the substrate, and then passes through the sample a second time, doubling the pathlength and, hence increasing the sensitivity. Ag/SnO_2 coated glass slides are commonly used as transflectance substrates, providing the advantage of being cheap and robust [182].

2.5.1.2. Attenuated total reflectance (ATR)

In ATR, the IR beam enters the ATR crystal which is made of an optically dense material at a particular angle of incidence, and the light can be reflected internally. By doing so, the IR beam is internally reflected for several times, and this internal reflectance will result in the production of an evanescent wave which can extend into the sample shown in

Figure. When the sample absorbs this attenuated evanescent wave, a spectrum is obtained.

Note that the pathlength (L) is dependent on the physical characteristics of the internal reflection element material and the angle of incidence (θ). Therefore, fixed-pathlength (L) is not applicable in ATR. Instead, the effective pathlength is controlled by the depth of penetration (d_p) of the evanescent wave travel into the sample [183]. The depth of penetration (d_p) is calculated as follow:

$$d_p = \frac{\lambda}{2\pi n_{ATR}\{(sin^2\theta) - (\frac{n_{sample}}{n_{ATR}})^2\}^{\frac{1}{2}}}$$

Where λ is the wavelength of the IR beam, n_{sample} is the refraction index of the sample, and n_{ATR} is the refraction index of the ATR crystal. The wave penetrates into the sample with a depth of 0.5-2 µm [184]. The depth to which the wave penetrates depends on the angle of the incident IR beam and the refractive index of both the crystal material and the sample. Each time the sample absorbs the IR radiation, there is a change in the evanescent wave, and it is attenuated. The attenuated energy from each of the evanescent waves is then transferred back which exits the crystal and is measured by the detector to produce an IR spectrum. The result is an infrared spectrum that reflects the whole chemical composition of the sample. In order to attain a successful ATR spectrum, the sample must be in direct contact with the ATR crystal, which is also called an internal reflectance element (IRE), and the refractive index of the IRE must be significantly greater than that of the sample, otherwise internal reflectance will not occur [185]. Some common materials used for ATR crystals include diamond, ZnSe, and Ge. This spectral acquisition technique is widely used in the context of bacterial typing due to the associated versatility. Little or no sample preparation is required, thus bacterial cells can be placed directly on the ATR crystal surface. The main issue will be the lack of sensitivity due to the restricted depth of penetration. It is estimated that ATR can only detect molecules present in concentrations greater than 0.1% [186]. Furthermore, many ATR crystals absorb only in the MIR region, although most experiments only focus on the MIR region, it would be a problem if the research in question is based on a greater range of IR region. Despite its limitation, ATR is the premier type of sample preparation in use today for FTIR.

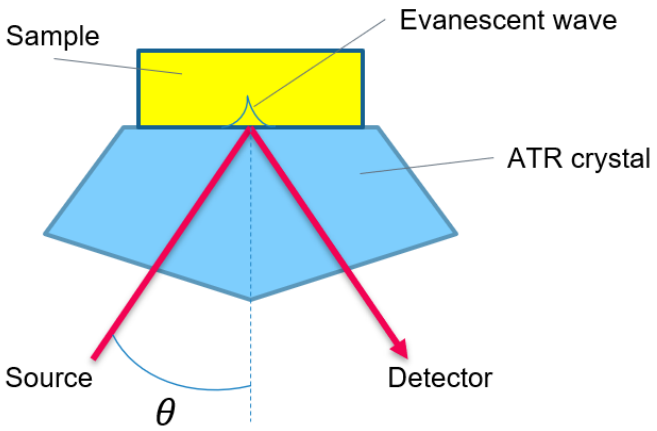


Figure 2.4. Illustration of ATR acquisition mode.

2.5.2. **Data Preprocessing Techniques**

The fact that functional groups representing different macromolecules could be quantified separately show that only few spectral signatures are sufficient for analysis and discrimination. However, background absorption and environmental factors such as water vapor humidity and the variability of sample composition may create outliers. The large amount of water within biological specimen also adds up unwanted interference to the sample spectra due to its strong O-H absorption caused by fundamental O-H stretching and H-O-H bending vibrations (Figure 2.5) [187].

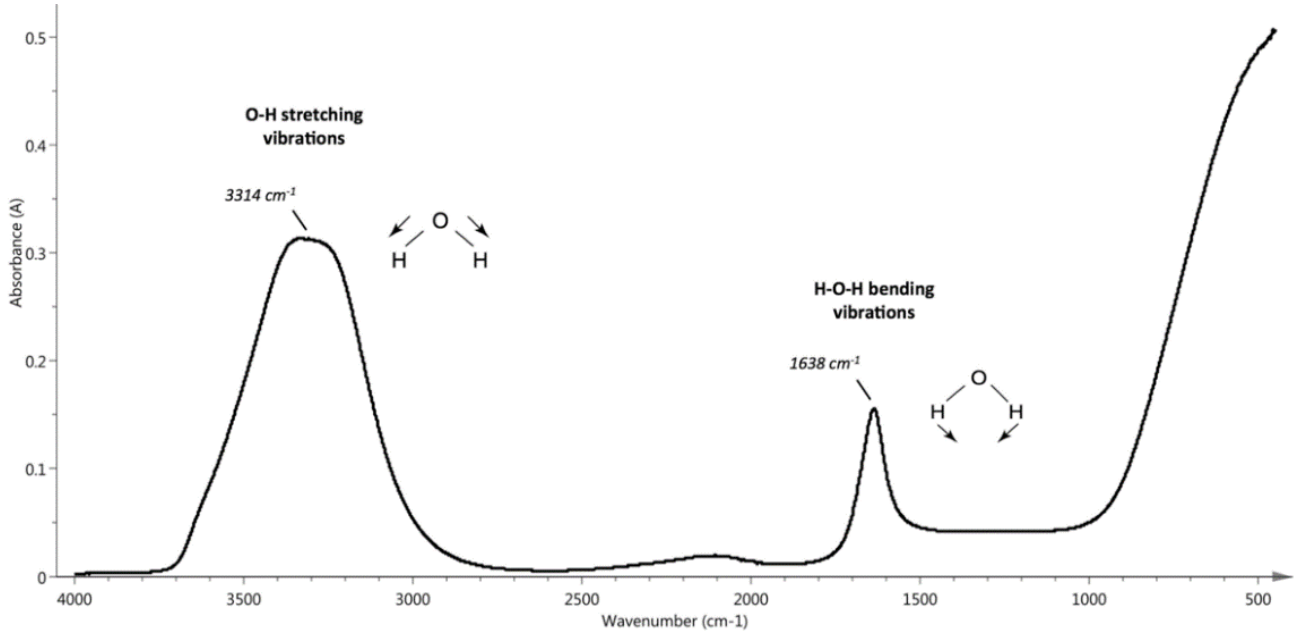


Figure 2.5. IR spectrum of liquid water on Attenuated total reflectance FTIR (ATR-FTIR).

Moreover, the low concentration levels of some analytes could hamper the determination of certain parameters in body fluids. Undesirable external factors such as particle size effects, scattering of light, morphological differences like surface roughness and detector artifacts can also result in poor quality spectra [188]. For this reason, IR spectroscopy for bacterial typing requires normally a high level of standardization, regarding growth and medium culture. Fortunately, modern FTIR software programs contain powerful algorithms for processing spectra and eliminating these effects. Spectroscopic data are composed of enormous amount of information, and multivariate analysis serves as a cut-off-pick-up tool, reducing high-dimensional data and picking up and retaining only essential information and comparing them against a database. This enables researchers to analyze very complicated and large datasets, and at the same time reducing

the dimensionality and complexity of the data allowing for the most meaningful information to be extracted. With automated analysis potential, multivariate methods allow analysis of multiple spectra simultaneously and interdependently, which facilitates comparison between spectra and to identify trends these may contain. Hence, pre-processing is an important first step in any workflow of spectral analysis and it is applied on spectral data to promote the linear relationship between spectral data and concentration of sample constituent. In short, pre-processing techniques are used for: (1) improvement of robustness and accuracy for quantitative and qualitative analysis, (2) improved interpretability, (3) detection and removal of outliers, noise and trends, (4) removal of inappropriate and unnecessary information from the data and (5) reduction of the scale of data mining step. To do so, many pre-processing techniques have been developed for easier interpretation of IR spectra, such as quality test, scatter correction, baseline correction and smoothing, subtraction, derivatives, and interpolation [189].

2.5.2.1. Quality Test

It is always recommended to closely examine the original spectra before deciding on what kind of process to be done, and it is always a good idea to retain the original data, so that if anything goes wrong, the original piece is still there to start over with. Quality test can be considered as outlier test, which involves defining a quality criterion, and based on this criterion, the program will automatically reject spectra that do not meet the set-up requirements. Some examples of criteria specification could be absorbance values in the amide I region, the signal-to-noise ratio, IR absorbance values of sharp water vapor features, and the presence or absence of optical fringes in the spectra of microorganisms [189].

2.5.2.2. Scatter Correction

Scatter correction is a statistical method that removes the scatter variation in the spectra that is caused by various particles in the sample. Multiplicative scatter correction (MSC), standard normal variate and de-trending are the most common scatter correction techniques nowadays, each having different roles including scatter correction, removal of multiplicative interferences, and removal of trend, respectively [188]. MSC is the most common pre-processing technique used for scatter correction. Briefly, MSC calculates the correction factor for the original spectra using reference spectra, which is usually the mean of spectrum acquired, and then corrects the original spectra using this correction factor by back transformation. It does not entirely eliminate the scatter

but decreases the inter-sample variance of the scatter by implementing an additive transformation of the individual spectrum into the mean spectrum.

Normalization is a scatter correction technique applied to scale up the spectra within its similar range. It aims to minimize the effects of varying optical path lengths on the spectra and the difference in sample quality, and scales and offset the spectrum at the same time of normalization. This technique increases the accuracy and efficiency of the spectral distance measurement modelling. Popular methods for this technique include Min-Max normalization, 1-norm, 2-norm, and standard normal variate [189]. Most of the time, the average of the standard deviation is used as the correction parameters for normalization. But in cases where the spectra are noisy, the median or the mean of the inner quartile range and the standard deviation of the inner quartile is suggested as correction parameters. In the case of bacterial IR spectra, the amide I band is often used as the internal standard for normalization and showed great efficiency.

2.5.2.3. Baseline Correction & Smoothing

Ideally, an IR spectrum should have a flat baseline that falls at zero absorbance or 100% transmittance. In reality, reflectance or transmittance IR spectra often contain unwanted background features or noise, which is caused due to scattering, external factors, such as illumination or temperature, causing variations in data acquisition. To acquire proper information from the spectral data, it is important to remove this noise from the signal before spectral comparison. Baseline correction is a pre-processing technique that eliminates the dissimilarities between spectra due to shifts in baseline, thus making an easier illustratable signal. It also generates more accurately predictable spectral parameters like band positions and intensity values [190]. One of the most common baseline correction techniques is offset correction, which is conducted by subtracting a linear horizontal baseline from the original spectrum and draws the baseline back to zero [162].

2.5.2.4. Derivatives and Deconvolution

IR spectra resolution can be well enhanced by spectral derivatives or deconvolution as it resolves and remove the overlapping bands. This pre-processing technique also reduce replicate variability, correct baseline shift, and amplify spectral variations [191]. Hence, it is very useful to obtain quantitative calibrations. The first derivatives are used on data to remove offset from standard spectra and to better resolve broad overlapping bands. The second derivatives also

remove offset, and additionally increase the spectra clarity by increasing the number of discriminative features and hence a net increase in spectral resolution. The most commonly used techniques for derivatives are Savitzky-Golay (SG) and Norris-Williams (NW) derivations. The SG technique provides an advantage over NW in that it can carry out smoothing, noise reduction and computing derivatives at the same time, which reduces certain noise level in the derivative, and thus, making spectra easier to interpret. The principle of NW technique, on the other hand, is to smoothen the spectral data based on a moving average over data points, and the gap between these data points is used to estimate the derivatives. Then, the finite difference is calculated based on this smoothing spectrum [188]. Yet, a problem with derivative spectra is that, as spectral variations are amplified, they contain more noise than the original spectra from which they are calculated. This suggests that a spectrum needs to have a good SNR to be able to apply derivatives [182]. The NW method is less prone to high-frequency noise compared to SG, as it uses both smoothing by moving average and gap size for derivative, and hence would be a better choice over the SG method. Whereas for deconvolution technique, the section having overlapping bands is Fourier transformed into a mathematical function called a spectrum. Then, the broad band spectrum is multiplied by an exponential function, depending on the optical retardation, to obtain a narrow band spectrum which resolves the overlapping bands and shows more peaks than in the original spectrum [182]. Derivation and deconvolution are very useful to analyze spectra of mixtures.

2.5.2.5. Other techniques

Other techniques include: spectral interference subtraction, which involves the removal or elimination of certain additive interferences from the input spectra; optimized scaling, which a theoretical-based method for the linear calibration of spectral data when it does not have a fixed intensity range; orthogonal signal correction, that is to remove the orthogonal variance to the component of interest from the dataset [188]. One can also select a specific spectral window subjectively according to the knowledge and experience of the investigator to reduce the data amount to be processed, which speed up the processing time. In general, this process involves mainly the focus of the fingerprint region, and the elimination of region between 1800 – 2750 cm^{-1} or between 3400 – 4000 cm^{-1} [189].

2.5.3. Statistical Analysis Techniques

After screening and filtrating out low-quality data, chemometric methods are used to reduce high-dimensional data and retain only essential spectral information. Spectroscopic data are composed of enormous amount of meaningless information, and those methods are able to analyze very complicated and large datasets, and at the same time reducing the dimensionality and complexity of the data allowing meaningful information to be extracted. Especially for vibrational spectroscopic datasets, multivariate methods allow analysis of multiple spectra simultaneously and interdependently, which facilitates comparison between spectra and to identify trends these may contain. Multivariate statistical analysis of FTIR spectra can be divided into two types: supervised methods and unsupervised methods [192]. In brief, unsupervised methods aim to extrapolate the spectral data without a prior knowledge, whereas supervised methods require prior knowledge of sample identity.

2.5.3.1. Unsupervised Methods

Unsupervised machine learning techniques are used to explore the hidden structures in a spectral dataset where an a priori class assignment is not available or not desired [193]. In here, there is no correct answers and there is no teacher, hence the term unsupervised. The result is not known, and the unsupervised techniques deal with the un-labelled data trying to find an underlying structure, pattern or trend of that data. Unsupervised learning will only have the original input data to work on. Some well-known unsupervised machine learning algorithms include singular value decomposition (SVD), principal component analysis (PCA), independent component analysis (ICA), distribution models, hierarchical clustering analysis (HCA), neural networks/deep learning, k-means cluster analysis (KMCA), and fuzzy C-mean clustering (FCM). In general, these techniques can be classified into dimensionality reduction, or classification, and clustering.

2.5.3.1.1. Dimensionality Reduction (Association) Technique

The most classical examples of dimensionality reduction techniques for spectra analysis are PCA and ICA. Both are simple, nonparametric methods for extracting relevant information from datasets, identifying patterns in data, and expressing the data in such a way to highlight their similarities and differences. In short, PCA reduce the dimensionality of the data to describe the variation present in a dataset, where the first principal component is a description of the maximum variance present in the dataset, the second describes the second most variance, and so on. It is one

of the most widely used multivariate methods because of its wide applicability in the multivariate problems. The main goal of PCA is to obtain the most important characteristics from data, whereas the goal of ICA is to find new components that are mutually independent in complete statistical sense [194].They both showed promising results, although some may argue one is better than the other, it all depends on the situation and the type of spectra in hand [195, 196]. Recently several variants of PCA are introduced, such as independent PCA (iPCA), which combines the advantages of both PCA and ICA [192]. This kind of tool can be useful for providing a method to separate spectra into groups, for instance, diseased and non-diseased, and it has also been used to reconstruct images.

2.5.3.1.2. Clustering Technique

Cluster analysis helps identify similarities between the spectra using the distances between spectra and aggregation algorithms. The most commonly used clustering techniques for IRS are KMCA, FCM and HCA. KMCA attempts to split data into “k” cluster groups of equal variances, where “k” represents the number of groups defined, and data points are clustered based on feature similarity. It uses the centroid of the cluster as the criterion to assign the cluster for each sample and is achieved by minimizing the sum of squared errors [188]. In IRS, KMCA has seen a number of uses to separate spectra into clusters, and as imaging tool, KMCA separate each spectrum acquired in the image and assign it to a cluster. Similar to KMCA, FCMA also assigns spectra to centroids in the datasets. However, unlike KMCA, the method is a soft clustering method, whereby assigning the samples to different clusters simultaneously with varying degrees of membership. Each point or spectrum in the dataset is assigned to a value from 0 to 1, the value closest to 1 being representative of the cluster center [197]. Therefore, by analyzing the centroid spectrum it is possible to extract chemical information which describes each reconstructed image. HCA clustering generally build a hierarchy of clusters that is normally presented in the form of a binary tree diagram, commonly known as dendrogram. There are two types of HCA, agglomerative and divisive as shown in Figure 2.6.

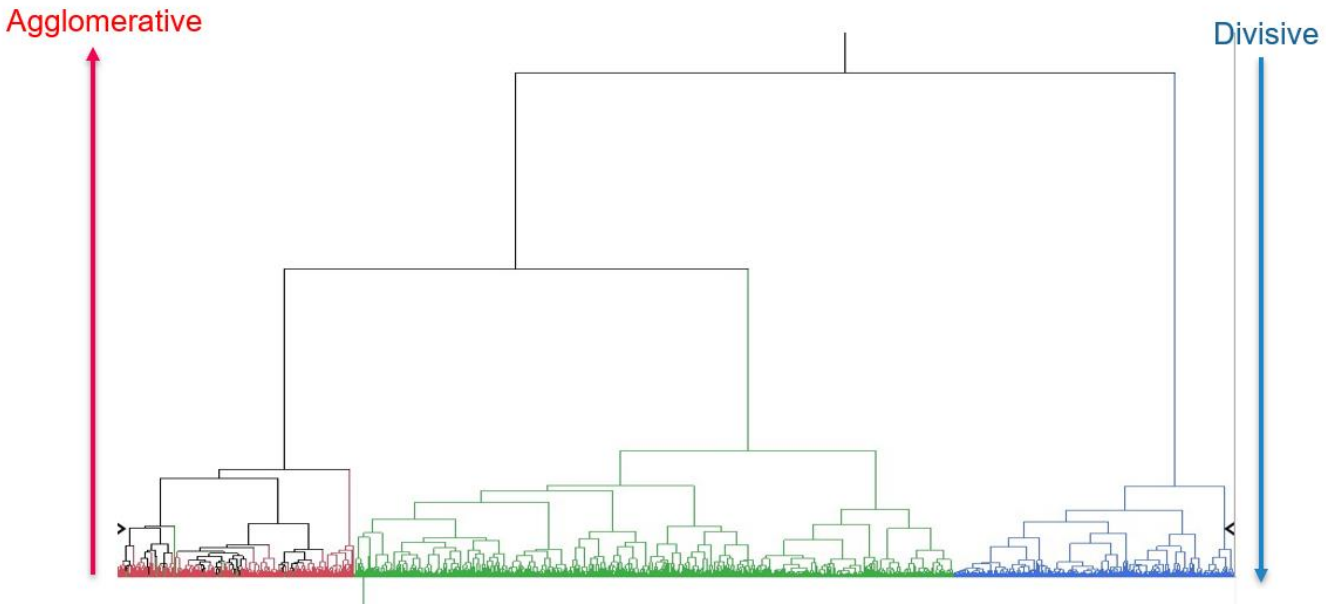


Figure 2.6. Conceptual dendrogram for agglomerative vs. divisive hierarchical clustering.

Briefly, this method starts out with each data point or spectrum in a separate group or cluster. The method then aims to group each data point together in an iterative process until there is only one cluster which contains all the data points, creating a dendrogram showing the linkage between each cluster [198]. There are two parameters to consider in hierarchical clustering: the linkage method and the distance metric. The linkage method determines the dissimilarity between two clusters of observations, and a number of different cluster linkage methods that have been developed. The most common types of methods are listed in Table 2.4 and Figure 2.7. Each linkage method uses different equation to calculate the inter-cluster distance. Several linkage methods could be used to compare the same dataset, and depending on the nature of the dataset, some methods may work better than the other.

Table 2.4. Linkage method description for agglomerative hierarchical clustering [199].

Linkage method	Description	Equation
Single linkage	The distance between two clusters is the minimum distance between an observation in one cluster, and an observation in the other cluster, i.e. the shortest distance between two points in each cluster. This method is especially useful when clusters are obviously separated.	$D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
Complete linkage	The distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster, i.e. the farthest distance between two points in each cluster. This method can be sensitive to outliers.	$D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
Average linkage	The distance between two clusters involves looking at the distances between all pairs and averages all of these distances, i.e. the average distance between points in each cluster	$D(c_1, c_2) = \frac{1}{\lvert c_1 \rvert \lvert c_2 \rvert} \sum_{x_1 \in c_1} \sum_{x_2 \in c_2} D(x_1, x_2)$
Centroid method	The distance between two clusters is the distance between the cluster centroids. This involves finding the mean vector location for each of the clusters and taking the distance between the two centroids.	$D(c_1, c_2) = D\left(\left(\frac{1}{\lvert c_1 \rvert} \sum_{x \in c_1} \vec{x}\right), \left(\frac{1}{\lvert c_2 \rvert} \sum_{x \in c_2} \vec{x}\right)\right)$
Ward method	The distance between two clusters is the sum of the squared deviations from points to centroids. This refers to the sum of the squared distance from each point to the mean of the merged clusters. It tries to find the distance that minimize the total within-cluster variance, and maximize the total between-cluster variance. The process resembles an ANOVA based approach It is an ANOVA based approach.	$TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D\left(x, \mu_{c_1 \cup c_2}\right)^2$

D = distance; c = cluster; x = observation, TD = total distance

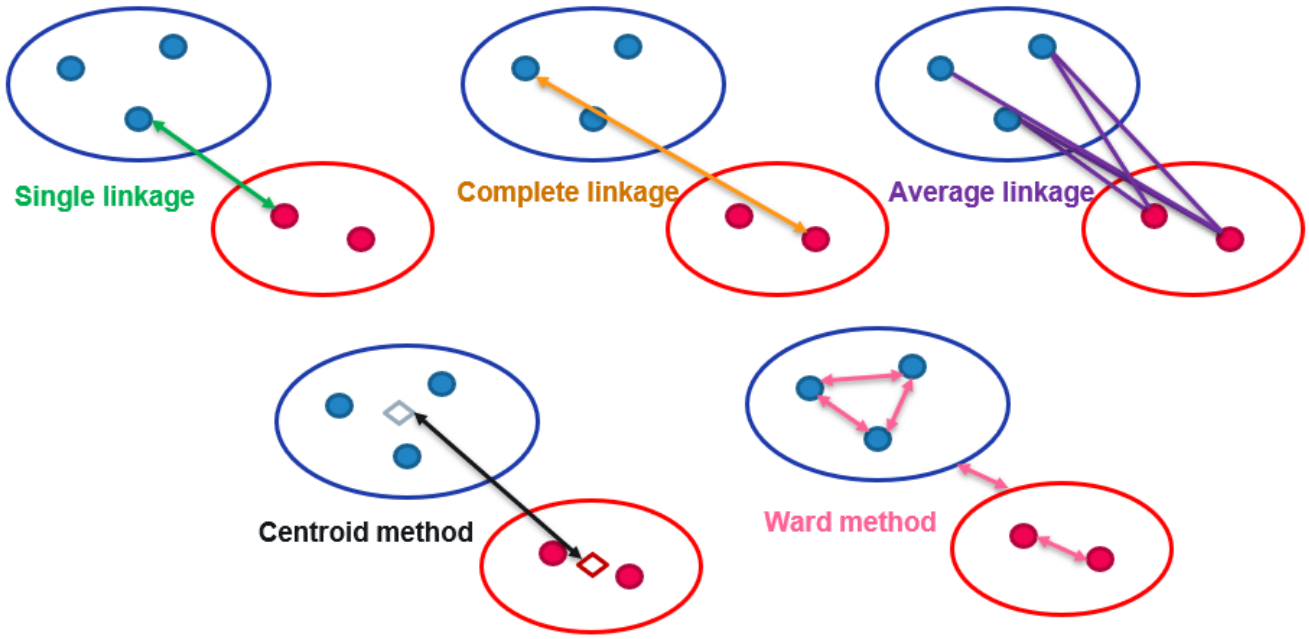


Figure 2.7. Representation of different cluster linkage methods in agglomerative hierarchical clustering.

The choice of clustering distance metric defines the closeness of the clusters. In other words, how similar two elements x and y are, and the different types of distance measures illustrated in Table 2.5 will influence the shape of the clusters.

Table 2.5. Definition of different distance measures for clustering analysis.

Clustering distance metric	Equation
Euclidean distance	$d_{euc}(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$
Manhattan distance	$d_{man}(A,B) = \sum_{i=1}^{n}\lvert(A_i - B_i)\rvert$
Pearson correlation distance	$d_{pear}(A,B) = 1 - \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2 \sum_{i=1}^{n}(B_i - B)^2}}$
Cosine correlation distance	$d_{cos}(A,B) = Cosine\ distance = 1 - cosine\ similarity$ $cosine\ similarity = \cos^{-1}\frac{\lvert\sum_{i=1}^{n} A_i B_i\rvert}{\sqrt{\sum_{i=1}^{n} {A_i}^2 \sum_{i=1}^{n} {B_i}^2}}$

Spearman correlation distance	$d_{spear}(A,B) = 1 - \frac{\sum_{i=1}^{n}(A'_i - \overline{A'})(B'_i - \overline{B'})}{\sqrt{\sum_{i=1}^{n}(A'_i - \overline{A'})^2 \sum_{i=1}^{n}(B'_i - \overline{B'})^2}}$

A and B = elements/vectors, n = length of vectors

Classical methods for distance measures are Euclidean and Manhattan distances. Other distance measures are classified as correlation-based distances, such as Pearson correlation, Spearman correlation, and cosine distance, in which cosine correlation is the one we will be using for our research purposes. This is because cosine distance measures the distance without accounting the magnitude of the vectors, which is very useful to analyze qualitative data. A visualized difference between the distance measured using Euclidean and cosine method is illustrated in Figure 2.8.

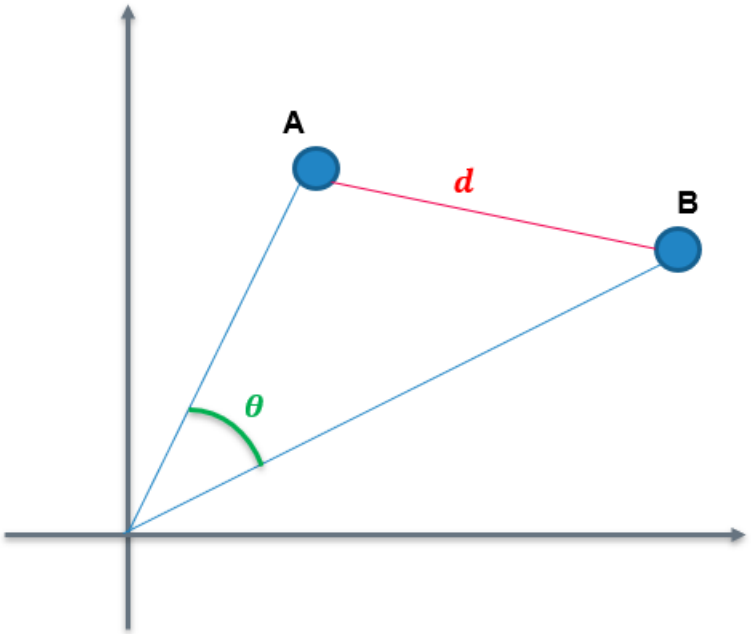


Figure 2.8. A figurative difference of Euclidean distance (d) and cosine similarity (θ) between two vectors A and B.

Among the clustering techniques, HCA is more prominently used for spectral data analysis, especially in microbiology, even though HCA is not well suited for large datasets.

2.5.3.2. Supervised Methods

Most of the practical machine learning uses supervised learning. Unlike unsupervised learning, supervised learning is guided by human intelligence, observation, and known outcomes. Hence, there is a teacher supervising the learning process, where the algorithm is being taught the

difference between right and wrong and is asked to mimic those results when new information is thrown. Supervised learning is typically done in the context of classification, when the goal is to map input to output labels, or regression, when the goal is to map input to a continuous output [200]. In other words, this technique groups the unknown samples into the known predefined groups according to their measured features. The main issue with supervised learning algorithms is that the term "correct" output is entirely determined from the training spectral data, therefore, noisy or incorrect data labels will clearly reduce the effectiveness of the "standard reference" spectra. Hence, the bigger the dataset, where all anomalies and edge cases are included, the more accurately the algorithm will work in each unique situation. Common algorithms in supervised learning include support vector classifier (SVC), quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), principal component regression (PCR), partial least squares regression (PLSR), logistic or linear regression, random forest, naïve Bayes, neural networks, regression trees, decision trees, and k-nearest neighbors.

2.5.3.2.1. Discriminant Analysis (Classification)

Discriminant analysis have two objectives: first, in a supervised way, it is used to describe and explain the differences among the groups, and second, similar to PCA, it is used to separate samples into different classes by maximizing the variance between classes and minimizing the variance within a class [30]. Two example techniques used for discriminant analysis are QDA and LDA. They are the most commonly used classification techniques of spectral imaging data from food and agricultural products [188]. QDA classifies the samples into the classes with quadratic-shaped boundaries and assuming that multivariate normal distribution is common in each class, whereas LDA on top of the assumptions by QDA, also assumes that covariance matrices of the classes are equal. The main disadvantage of LDA is that it does not hold well with the condition where the number of samples is less than that of number of variables, as in this condition, the inversion of covariance matrices becomes difficult. In some comparison literature, QDA-based models showed higher classification rates and quality performance than LDA [201]. However, a powerful analysis tool has been developed by combining PCA and LDA, and this technique was particularly helpful when the number of variables is large. Nonetheless, applying more than one technique on the same dataset is often recommended, as each technique's classification accuracy is mostly determined by the underlying structure of the data which can make one method more suitable than the other.

2.5.3.2.2. Regression Analysis

Regression is a statistical procedure which determines the relation between dependent variable and independent variable. PCR and PLSR are the best-known regression techniques used for spectral data analysis. PCR can analyze data with high multicollinearity between their variables. In regression, multicollinearity is a statistical procedure where several independent variables participating in multiple regression modelling are highly correlated to one another [188]. PCR reduces the standard error by adding bias to the regression estimates and is hoped that more reliable estimates will be achieved due to this overall effect. However, PCR model developed using the independent and dependent variables sometimes gives a random error or noise rather than giving the anticipated relationship. This kind of error can be avoided by choosing the optimal number of principal components. Another way to face the multicollinearity problem is to use PLSR. PLSR is a well-known chemometric tool that is used to estimate the biological and chemical properties of the sample from their spectral spectrum, especially when spectral data is massive. The core idea of using this method is to investigate the spectral variability as a function of a systematic conditional change. PLSR can be employed to construct predictive models for spectral response as a function of the target variable [202]. And this algorithm has become one of the dominant practices of multivariate calibration due to its high quality of the calibration model.

2.5.3.2.3. Artificial Neural Network (ANN)

ANN is the most popular deep learning technique in recent years and is primarily used for pattern recognition purposes. ANNs are robust and can handle unsupervised and supervised problems and can work with both qualitative and quantitative analysis. They are considered "nonparametric nonlinear regression estimators" because of their ability to determine relationships between one or more input, and one or more output, regardless of the form of the function defining the relationship between the two sets of variables [203]. It is a self-training system and intelligently constructed to optimize the processing power of its own network. ANN works like the human nervous system, where each neuron receives a signal from neighboring neurons, later executes them and finally gives out the output signal. The number of neurons used may vary from ten to several thousands and are based on the training set. As more data are fed in, the machine gets smarter and more efficient at interpreting future inputs. One key aspect of ANN is that each neuron can be formulated to utilize a single algorithm that could be useful for certain datasets but poor to others. And as weights are adjusted for each neuron, ANN learn by itself where to best analyze the

data for having the highest confidence output and continues to adjust neuron weights for more optimization of the network (Figure 2.9).

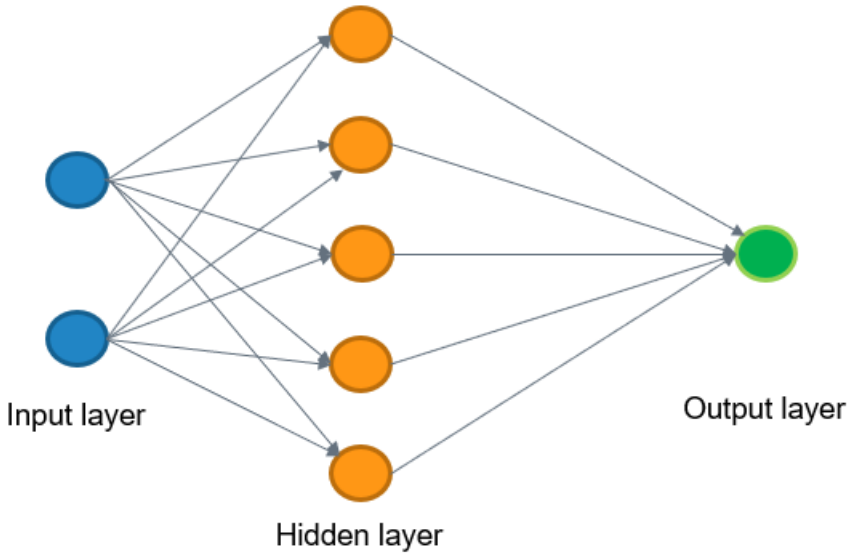


Figure 2.9 A general representation of an ANN analysis flowchart. The input layer represents the data that were fed into the algorithms; the hidden layer is where various types of mathematical computations are rigorously performed on the input data to recognize the pattern (there can be more than a single hidden layer); depending on the linkage and weigh of these linkages, the output layer tells the results of this analysis.

ANN-based analysis is often used in conjunction with HCA in IRS for microorganism identification [204]. Other than microbiology application, ANN has been applied in many functions that changes and ameliorates our everyday life, such as handwritten character recognition, facial recognition, speech recognition, and signature classification. ANNs have proven to be the most accurate of all systems for large deep learning problems with the only downside being the time it takes for training.

2.5.4. Application of FTIR in Microbiology

FTIR has emerged to become an essential analytical tool available to scientists to study various materials in various fields. To name a few, application of FTIR include food and environmental analysis, forensic science, semiconductor analysis, pharmaceutical, physiological and biological analysis, geological samples, and for multilayer compounds. It is widely used in industries as well as in research, due to its simplicity and reliability in terms of measurement, quality control and dynamic measurement. Especially in the field of microbiology, where FTIR is

useful for the identification of functional group and structure, identification of substances, studying reaction progress, detection of impurities, and quantitative analysis. In brief, application of FTIR in microbiology can be classified into three categories: (1) characterization, (2) quantification, (3) identification, differentiation, classification, and (4) the use of FTIR Microspectroscopy (FTIRM) for species identification from micro-colonies.

2.5.4.1. Microbial Characterization

Cells represent the fundamental biological unit from which the life of all living organisms depends. Hence, knowledge of their morphology and their biochemical processes is extremely important in order to counteract the onset of cell anomalies or pathological conditions. Biological samples contain macromolecules, such as nucleic acids, proteins, lipids and carbohydrates that have characteristic and well-defined IR vibrational modes. These bands can be used as markers for the biochemical response of cells and tissues to different treatments and pathologies. In FTIR spectra, each cellular component is at a peculiar position. The capability to extract specific information from each band of each spectrum is important for drawing useful conclusions on the process of interest and to advance knowledge. As

Table 2.3 has highlighted, most functional groups can be assigned according to their vibrational wavelength, and the macromolecule where the functional group is consisted of can be properly speculated. By observing changes in IR spectra, subtle changes caused by various biochemical processes, such as the occurrence of specific pathologies, benign and malignant ones, or by various cellular differentiation steps, can be detected. A more detailed analysis of these spectral features may reveal the presence of particular cell constituents. For instance, cell storage materials such as poly-β-hydroxy fatty acids can be determined on IR spectra in conditions of starvation. The release of CO_2 and the formation of endospore in bacterial cells can also be observed with IR spectra [193]. However, in many cases, FTIR spectroscopy alone is not sufficient for the characterization of microbiological cell due to overlapping absorbance bands, and it is often recommended to perform other techniques such as genotypic or other phenotypic method on top of FTIR spectroscopy [173].

2.5.4.2. Quantification Analysis

Since the intensities of the bands in MIR region are proportional to the concentration of their respective functional groups, quantitative analysis of FTIR is especially useful in the area of

food safety and quality. In terms of safety, pathogenic bacterial can be identified and quantified based on a spectral change or increase in certain peak intensity from a reference spectrum of an uncontaminated sample. The result can be verified by a thorough search in a spectral library for identification, and the concentration can be calculated based on Beer's law. For quality purposes, adulteration of food can be examined easily with FTIR [151]. An important application for quantitative analysis of FTIR is antibiotic susceptibility testing. Although many other promising physical techniques exist for such measurement, such as radiometry, microcalorimetry, bioluminescence and electrical impedance, FTIR spectroscopy is also useful in this field, since quantification of cell mass as a function of antibiotic treatment, as well as the detection of antibiotic-induced structural changes in microbial cells, is well within reach of its sensitivity and specificity [189]. For instance, for a protein synthesis inhibitor antibiotic, analyzing the various amide bands of the IR spectra would be enough to analyze the drug-induced changes in bacterial cells as shown in Figure 2.10 [205].

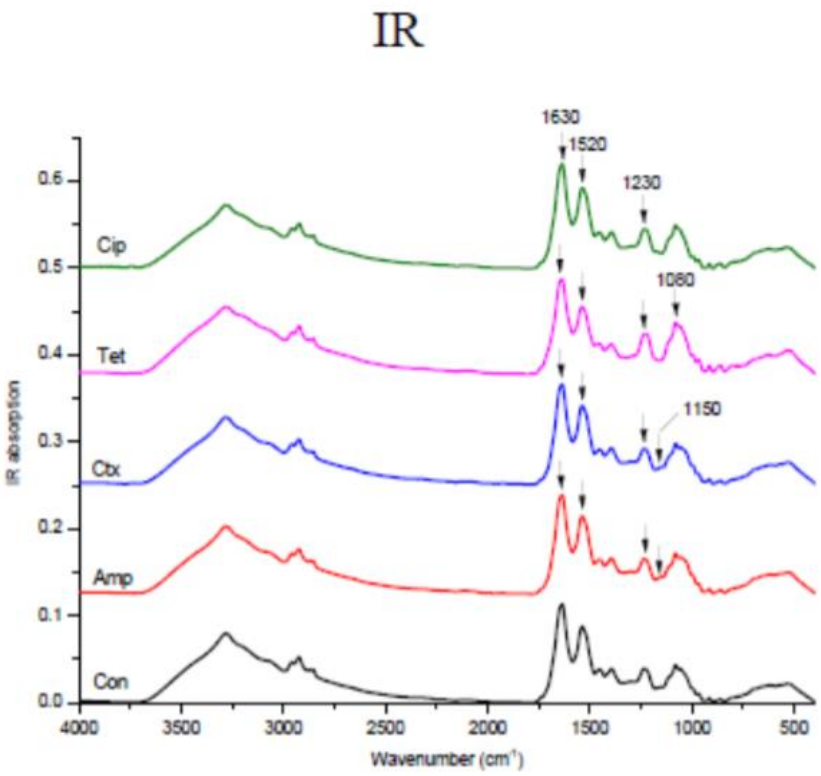


Figure 2.10. FTIR spectra of E. coli TOP10 without antibiotic (Con), and with different type of antibiotics added to the growing media (ampicillin, cefotaxin, tetracycline, ciprofloxacin).

Other fields that FTIR may be valuable could be assessing the mechanisms of microbial inactivation by food processing techniques, assessing the membrane properties in changing environments, and detecting and assessing stress-injured microorganisms [166].

2.5.4.3. Identification, Differentiation, Classification

Taxonomic level identification, differentiation and classification of microorganisms is important for epidemiological investigation, outbreak detection, source tracking and pathogen control. Identification involves describing an entity to a point where it uniquely stands out and could be picked out, whereas classification involves assigning the entity to a group according to a given criteria. Differentiation or discrimination is somewhat similar to classification; however, it focuses mainly on separating entities into groups depending on their properties that are focused on. These three terms often go hand-in-hand, as by identifying an entity, classification or differentiation conditions are often also met. FTIR spectroscopy had been successfully used for identification, differentiation and classification of a variety of microorganisms at genus, species, and sub-species levels since 1990s [206]. Differentiation by FTIR spectroscopy is based on the fact that the IR spectra are a reflection of the overall molecular composition. IR spectrum of intact microorganisms provides information on the structure and composition of the whole cell. Because microbial IR spectra are complex spectroscopic signals encoding the superposition of hundreds or even thousands of bands that cannot be resolved by any means, pattern recognition techniques have to be used which consider the spectra as fingerprints rather than a combination of discrete band intensities, frequencies and bandwidths [166]. Despite the fact, strains that differ in their molecular makeup will show relatively distinct IR band. The capacity of FTIR spectroscopy to identify and differentiate unknowns is strongly dependent on the quality and quantity of the reference library. After 20 years of research and as IR spectra have been continuously added to the spectrum library, the potential of FTIR spectroscopy for the identification, differentiation and classification of microorganisms is well documented, and the technique has been applied in many different fields like food microbiology, medical diagnostics and microbial ecology [207]. The spectrum library application ranges from the identification of clinical and food pathogens or food contaminants to starter and probiotic cultures in food as well as different kinds of environmental microorganisms. Different algorithms can be used to identify unknown microbial strains based on a reference database. It must be confirmed that these reference datasets contain representative numbers of spectra covering all relevant spectral types to be identified. Unknown microbial

samples will only be correctly identified with validated reference databases of microbial IR spectra. Identification is then achieved by comparing the IR spectrum of an unknown microorganism with all entries of the reference database. Some studies focused on the differentiation of only few species, others try to assemble large and comprehensive database for broad and general application in microbiology. In case of having small and specific databases, a differentiation rather than an identification of certain species is recommended, and additional methods can be applied for identification, such as matrix-assisted laser desorption/ionization time of flight mass spectrometry. And in these cases, HCA is used in conjunction for differentiation, which seems to work well for a limited number of strains and species. A thorough review of articles investigating in the identification, differentiation, and classification of pathogens by FTIR spectroscopy within the past decade can be found in Table 2.6.

Table 2.6. Articles (2014 – 2023) concerning the identification, differentiation, and classification of pathogen by FTIR spectroscopy.

Aim	Microorganism	No. of isolates	FTIR technique	Chemometrics	Results	Ref.
Develop and validate a model for typing *Acinetobacter baumannii* clinical isolates	*Acinetobacter baumannii*	77 in database; 148 in test set	ATR-FTIR	PLS-DA	100% at strain type level	[208]
Discriminate between *Bacillus* spp. and *Alicyclobacillus* spp. inoculated into apple juice, and the differentiation between their species	*Alicyclobacillus* and *Bacillus*	4 Alicyclobacillus ; 4 Bacillus	ATR-FTIR	PCA, SIMCA	Correct differentiation to genus and species level achieved of these 2 genera	[209]
Interpret, discriminate and classify *Aspergillus spp.* growing on peanut	*Aspergillus alliaceus, Aspergillus caelatus, Aspergillus flavus, Aspergillus parasiticus,* and *Aspergillus tamari*	N/A	ATR-FTIR	PLSR	98.5% correct, with only one misclassified sample	[210]
Assess FTIR spectroscopy and MALDI-TOF MS for routine microbial	*Campylobacter (C. jejuni, C. coli, C. lari, C. hyointestinalis* subsp.	174 in database; 40 in test set	Transmission -FTIR	HCA	88% (MALDI-TOF) and 75% (FTIR)	[211]

identification of food-related microorganisms	*hyointestinalis, C. fetus, C. concisus)*				correctness at species level	
Employ MALDI-TOF-MS, Raman, and FTIR spectroscopies combined with multivariate statistical analysis for differentiation of *Campylobacter* to subspecies level	*Escherichia coli*	11 isolates	Transmission -FTIR	PCA, PC-DFA, HCA	Correct classification achieved for all isolates	[205]
Demonstrate that FTIR spectroscopy can discriminate between uropathogenic *E. coli* (UPEC)	*Escherichia coli* and *Trueperella pyogene*SEP	95 isolates (63.2% training, 36.8% test)	Transmission -FTIR	PC-DFA, PLS-DA	Prediction accuracy: ST131 (91.19%), ST95 (86.58 %), ST127 (69.38 %), ST73 (39.15 %) and ST10 (30.15 %)	[212]
Assess if a database built from bacteria obtained from the uterus of Austrian and German dairy cows could be used to identify uterine bacteria from Argentinean dairy cows	*Escherichia coli Salmonella enteritidis*, *Pseudomonas ludensis*, and *Listeria monocytogenes*	55 in database; 25 in test set	Transmission -FTIR	LR	21/25 correct identification (84%)	[213]
Identify spectral windows to classify bacteria specific to poultry meat suspended individually in sterile water	*Escherichia coli, Salmonella sp., Enterobacter cloacae, Hafnia alvei, K. pneumoniae*, and *Proteus mirabilis*	4 isolates	ATR-FTIR	PCA, PLS-DA, SIMCA	100% correct classification of samples contaminated with *S. enteritidis* and *P. ludensis*	[214]
Assess the ability to distinguish between mixed genus bacteria	*Enterecoccus sp. Lactococcus sp.* and *Lactobacillus sp.*	4 isolates	Transmission -FTIR	PCA, LDA	When the mixing range was comparable (0.5:0.5 and 0.6:0.4 for Gram-positive and Gram-negative respectively),	[215]

					classification success rate was > 95%	
Compare biochemical identification methods with FTIR cluster method for identification of LAB isolated from Kaşar cheese	*Klebsiella (K. pneumoniae, K. variicola,* and *K. quasipneumoniae)*	157 in database; 83 in test set	Transmission -FTIR	HCA	>99% correlation to reference culture	[216]
Evaluate the discriminatory power of FTIR spectroscopy and MALDI-TOF MS for strain typing in comparison to WGS for potential integration into the routine diagnostic workflow	*Listeria monocytogenes* and *Salmonella* spp.	68 isolates	Transmission -FTIR	UPGMA	FTIR result was congruent with WGS 92.6% of isolates	[217]
Assess MIR spectroscopy as an alternative method for confirmation of foodborne pathogens	*Listeria innocua, Staphylococcus epidermidis, Salmonella spp., Shigella dysenteriae* and *Vibrio spp*	14 isolates	ATR-FTIR	PCA	Correct classification achieved for all isolates	[218]
Use SR-FTIR microspectroscopy for the identification and classification of foodborne pathogenic bacteria at the genus, species, and subspecies level	non-typhoid *Salmonella* serogroups and serotypes	10 isolates	SR-FTIR	PCA	Correct classification achieved for all isolates	[219]
Discriminate the most frequent and clinically relevant *Salmonella* serogroups and serotypes, correlating the discrimination obtained with the O-unit composition of somatic antigen	*Acinetobacter baumannii, Candida albicans, Enterobacter cloacae, Enterococcus faecalis, Enterococcus faecium, E. coli, Klebsiella pneumoniae,*	325 isolates	ATR-FTIR	PCA, PLSDA	99.6% at serogroup level	[220]

	Psuedomonas aeruginosa, Serratia marcescens, S. aureus, CoNS					
Identify *Staphylococcus aureus* independently of the culture growth stage	Methicillin-resistant *S. aureus*	141 isolates in database; 58 in test set	FTIR	PCA, LDA	100% sensitivity and 98% specificity	[221]
Develop a model for discrimination of heterogeneous vancomycin-intermediate *S. aureus* and vancomycin-intermediate *S. aureus*	*Candida krusei, Candida parapsilosis, Candida albicans, Candida glabrata*	59 isolates in database; 39 in test set	ATR-FTIR	PCA, PLS-DA	100% sensitivity and specificity	[222]
Ascertain the effect of sample preparation on the discriminatory capacity of ATR-FTIR spectroscopy of *Candida* species	*Penicillium implicatum, Penicillium aurantiogriseum, Penicillium notatum, Penicillium purpurogenum and Penicillium citrinum*	4 isolates	ATR-FTIR	PCA, KMC	Correct classification achieved for all isolates	[223]
Classify and discriminate five species from *Penicillium*, which were obtained from infected fruits	*Salmonella typhi*	5 isolates	FTIR-microscopy, ATR-FTIR	PCA	Correct classification achieved for all isolates	[224]
Evaluate ATR-FTIR spectroscopy to classify differentially expressed bio-molecular markers in *Salmonella typhi*-infected and healthy freeze-dried sera samples	Gram-positive and Gram-negative bacteria	25 isolates	ATR-FTIR	PCA, HCA	100% sensitivity and 96% specificity	[225]
Evaluate ATR-FTIR for routine yeast isolates for on-site	*Candida* spp., *Cryptococcus* spp., *Rhodotorula mucilaginosa,*	205 isolates in database; 573 in test set	ATR-FTIR	N/A	>98% correct identification	[226]

identification to the species level	*Saccharomyces cerevisiae*, *Geotrichum clavatum*, and *Trichosporon* spp.					
Differentiate four species of Shigella isolates from stool samples	*S. dysenteriae*, *S. flexneri*, *S. boydii*, and *S. sonnei*	91 isolates	Transmission -FTIR	PCA, HCA	100% correctness, sensitivity, and specificity	[227]
Rapid and reliable identification of biochemically confirmed typhoid and paratyphoid fever-associated *Salmonella* isolates.	*S.* Paratyphi A-C	359 isolates	Transmission -FTIR	PCA-LDA, ANN	Prediction accuracy: *Salmonella* Typhi (99.9%), *Salmonella* Paratyphi A (87.0%,), B (99.5%), and *Salmonella* Paratyphi C (99.0%)	[228]
Evaluate use of FTIR for the identification of *Legionella pneumophila* (Lp) at the serogroup level for diagnostic purposes and in outbreak events	*Legionella pneumophila* serogroups 1-15	133 isolates	Transmission -FTIR	PCA-LDA, HCA, ANN	95.49% correct identification	[229]
Evaluated FTIR for cluster analysis of *Burkholderia cenocepacia* epidemic strain ET12, isolated from adult cystic fibrosis patients.	*Burkholderia cenocepacia* epidemic strain ET12 and non ET12	12 isolates in database; 54 in test set	Transmission -FTIR	PCA, LDA, ANN	Up to 84.6% sensitivity, and up to 91.3% specificity	[230]
Evaluate the performance of the spectroscopic approach in identifying enterococci infections.	*Enterococcus faecium and Enterococcus faecalis*	60 isolates	ATR-FTIR	SIMCA, PLS-DA and SVM	Correct classification achieved for all isolates	[231]

(ATR-FTIR: Attenuated Total Reflectance Fourier Transform Infrared; PLS-DA: Partial Least Square-Discriminant Analysis; PCA: Principal Component Analysis; SIMCA: Soft Independent Modelling by Class Analogy; LR: LogiPLSR: Partial Least Square Regression; UV-Vis: Ultraviolet-Visible; NIR: Near Infrared; FPA-FTIR: Focal Plane Array Fourier Transform Infrared; ANN: Artificial Neural Network; HCA: Hierarchal Clustering Analysis; MALDI-TOF MS: Matrix Assisted Laser Desorption/Ionization Time Of Flight Mass Spectrometry; PC-DFA: Principal Component-Discriminant Function Analysis; SVM: Support Vector Machine; SFFS: Sequential Floating Forward Selection; MSA: Metabolomic Spectral Analysis; MLP: Multilayer Perceptron; LR: Logistic Regression; FO-FTIR: Fiber-optic Fourier Transform Infrared; CART: Classification And

Regression Trees; UPGMA: Unweighted Pair-Group Method with Arithmetic means; SR-FTIR: Synchroton Fourier Transform Infrared)

2.5.4.4. FTIR microscopy (FTIRM)

Since mid 1980s, the development of commercial visible IR microscopes let FTIRM become a valuable tool. The addition of a microscope as an accessory to conventional FTIR spectrometers has led to the possibility of analyzing intact tissue sections and even single cells at cellular resolution. While light microscopes provide information on shape, color, and contrast of a given sample, IRS may give information about structure and identity of complex samples at the molecular level. Thus, the combination of light microscopy with the sensitivity and specificity of IRS provides considerable additional information, in particular the possibility of visually obtaining structure information. This combination pushed the detection limit down to the sub-nano level and opened the field of spatial resolution to IRS. FTIRM is able to inspect limited areas on a surface such as an agar plate, and to obtain reflectance or transmittance spectra from samples constituted of few hundreds of cells, for example micro-colonies grown after only 6-10 h, which significantly reduces incubation time of cultures, thereby accelerating the identification process [232]. Although in general, FTIRM seems to be more difficult to apply under routine conditions as many researchers complained of have faced difficulties in standardizing the procedure, it has still gained increasing attention due to its ability to see both the spectral and the spatial information at the same time [233].

2.5.4.5. Disease Diagnosis

As many studies have studied, FTIR spectroscopy can diagnose infectious disease. Traditionally, FTIR spectroscopy is considered as the gold standard for kidney stone analysis [234]. Nowadays, its application in medical field extends far beyond that. Considerable amount of research papers has demonstrated superior performance of FTIR spectroscopy in viral disease diagnosis in terms of speed and cost [235-245]. It has been evaluated to detect biochemical components in saliva, serum, plasma, whole blood and urine, for the diagnosis of various type of cancer, periodontal disease, diabetes, chronic kidney disease, and burning moth syndrome [246]. Identification of biomarkers from the human conditions in saliva has been studied by various research groups. A general literature search was carried out to find relevant papers since 2010 to present using FTIR spectroscopy for disease diagnosis in human saliva specimen. Most of them were using ATR-FTIR spectroscopy as sampling method for spectral acquisition. Different type

of multivariate analysis techniques was employed, including PCA, PLS, LDA, QDA, support vector machine (SVM), ANN, logistic regression (LR), analysis of variance (ANOVA), receiver operating characteristic (ROC) curves and HCA, where PCA was the most used and often combined with other techniques. While sensitivity achieved in these studies were all above 90%, some of them obtained 100% accuracy in predicting diseased individual with cancer and diabetes [235, 236, 240]. Specificity was generally high with percentages all above 80. The difference in sensitivity and specificity among studies may be due to their use of divergent sampling method and chemometric techniques. Details can be found in Table 2.7..

Cancer has been investigated by several groups and has demonstrated that there are significant changes in secondary structure of proteins upon cancer development. Normal from cancer states were successfully distinguished from each other within the interval 900-1800 cm^{-1}. In esophageal cancer, notable differences between healthy and diseased patients were observed in the regions 1000-1150 cm^{-1}, 1350-1500 cm^{-1} and 1530-1600 cm^{-1}, each region corresponding to DNA/RNA, amide II and amide I, respectively [235]. Region 1000-1150 cm^{-1} has appeared consistently in other studies for the research in cancer, especially 1072 and 1074 cm^{-1} [236, 247]. These bands are related to the asymmetric and symmetric PO_2^- stretching from symmetric PO_2^- stretching from inorganic phosphates and phosphate group of phospholipids. This spectral feature is associated with the role of phosphates in DNA during diseases [247]. Other important band intensity changes observed include 2924 and 2854 cm^{-1}, corresponding to membranous lipid for oral cancer patients [236], 1460 and 1433-1302.9 cm^{-1}, corresponding to proteins and lipids in colon and breast cancer patients [238, 247]. Band shifts were observed at 1640-1655 cm^{-1}, 1547-1549 cm^{-1}, 1300-1310 cm^{-1} or 1300-1315 cm^{-1} for breast cancer [239]. These regions could all be assigned to amide. Lipid and fatty acid region (1735-1740 cm^{-1}, 1393-1406 cm^{-1}) band shifted were also observed in that study [239]. Prominent wavenumbers 964 cm^{-1}, 1024 cm^{-1}, 1411 cm^{-1}, 1577 cm^{-1} and 1656 cm^{-1} were remarked to separate lung cancer spectra from healthy spectra [248]. For head and neck cancer, shifts were observed at wavenumbers 1550 cm^{-1} and 1042 cm^{-1} [249]. Interestingly, FTIR spectroscopy have not only correctly diagnosed the diseased group out of saliva, but also identified the stage of cancer, and those who recovered from cancer in two different studies [235, 248].

In diabetic patients, 1640 cm^{-1} (amide I) and 1735 cm^{-1} (lipid ester) was more intense than normal group, tyrosine ring (1517 cm^{-1}) and proteins (1452 cm^{-1}) was altered, and amide II band

at 1550 cm^{-1} was less prominent in saliva of diabetic group [240]. Another research group noticed that 1545 cm^{-1} (amide II) and 1647 cm^{-1} (amide I) allowed to distinguish psoriatic and diabetic patients from the control group, and those bands were even useful to identify patients with different kinds of psoriasis [241]. Alteration and glycation of serum albumin and hemoglobin in patients with diabetes are well represented from the change of structure and activities in protein bands in their spectral signature [241]. Similar results were achieved by another group of researchers, concluding the diabetic characterizing spectra region were within wavenumbers 4000-2000 cm^{-1} [250]. However, one study using rats as experimental samples got 100% to 93.33% sensitivity of classifying non-diabetic, diabetic and insulin-treated diabetic rats with bands 1452 cm^{-1} and 836 cm^{-1} [251].

Periodontal disease diagnosis using FTIR spectroscopy in saliva was investigated by three research groups. Spectral range between 1230-1180 cm^{-1}, seems a promising tool for the diagnosis of periodontitis [243]. Aggressive and chronic periodontitis could be successfully differentiated from each other within 1800-950 cm^{-1} [244, 252]. The overall accuracy for the classification were 73.9% for distinguishing aggressive periodontitis from control, and 67.7% for chronic periodontitis and control [252]. Noteworthy, the smoking effect was also evaluated by the same research group, and they found that 3,700-1,850 cm^{-1} and 2,170-1,900 cm^{-1} (thiocyanate band) revealed better discrimination [244].

Thiocyanate (2052-2058 cm^{-1}) and nucleic acid (868-924 cm^{-1}) regions could be potentially used for the diagnosis of chronic kidney disease and burning mouth syndrome [245, 253]. The diagnosis accuracy of spectra could be hampered by excessive tobacco smoking, and thiocyanates might be important salivary maker in smokers that must be taken notes down if investigating other disease condition in the smoker patient [254].

Based on the research conducted, spectroscopy coupled with a multivariate analysis approach may represent a powerful tool for diagnosis by identifying salivary biomarkers through spectral bands. FTIR spectroscopy may provide novel insight to the current pandemic situation, with its advantages of being cost-effective, non-invasive, label-free and accurate in diagnosis, as demonstrated in Table 2.7.

Table 2.7.. Studies using FTIR spectroscopy for disease diagnosis using saliva (2010 – present).

Aim	Method	Result	Remarks	Ref
To identify early-stage oesophageal adenocarcinoma from healthy individuals	ATR-FTIR PCA-QDA	PCA-QDA model achived 100% accuracy for the inflammatory stage and high-quality metrics for other classes	1000 cm^{-1} to 1150 cm^{-1}, region associated with DNA/RNA seems effective for discrimination	[235]
To determine and differentiate the FTIR spectra of salivary exosomes from oral cancer patients and healthy individuals	ATR-FTIR PCA–LDA	Sensitivity of 100%, specificity of 89% and accuracy of 95%. The support vector machine (SVM) showed a training accuracy of 100% and a cross-validation accuracy of 89%	IR spectra of oral cancer patients were consistently different from healthy individuals at 1072 cm^{-1} (nucleic acids), 2924 cm^{-1} and 2854 cm^{-1} (membranous lipids), and 1543 cm^{-1} (transmembrane proteins).	[236]
To show how FTIR spectroscopy could be used to diagnose head and neck cancer at an earlier stage	FTIR PLS	Infrared wavenumbers 1650 cm^{-1}, 1550 cm^{-1}, and 1042 cm^{-1} were determined to discriminate between normal and cancer sputum	In cancer cases, the absorbance levels for 1550 cm^{-1} were increased relative to controls, whereas 1042 cm^{-1} absorbance was decreased suggesting changes to protein and glycoprotein structure within sputa cells	[249]
To determine the role of saliva in the early diagnosis of salivary gland tumor	ATR-FTIR	ATR-FTIR was able to track spectral variations between saliva samples from healthy volunteers and from salivary gland tumor patients	Most evident alterations occur in the region between ~900 and 1300 cm^{-1}	[255]
To apply ATR-FTIR onto saliva from patients with breast cancer, benign breast disease, and healthy matched controls to investigate its potential use in breast cancer diagnosis	ATR-FTIR	90% sensitivity and 80% specificity for discriminating breast cancer patients from controls. 80% sensitivity and 70% specificity to differentiate breast cancer patients from benign disease	The absorbance levels at wavenumber 1041 cm^{-1} were significantly higher in saliva of breast cancer patients compared with those of benign patients. 1433–1302.9 cm^{-1} band area was significantly higher in saliva of breast cancer patients than in control and benign patients	[247]
To identify and separate cancer from colitis in endoscopic colon biopsies	ATR-FTIR PCA	Sensitivity of FTIR detection for cancer achieved 97.6%	The relative intensity of amide II band to ~1643 cm^{-1} decreased in spectra of malignant colon tissues. The intensity of ~1460 cm^{-1} was weaker than that of ~1400 cm^{-1} peak in spectra of the cancerous samples. Peak at ~1460 cm^{-1} was stronger than or equal to that of ~1400 cm^{-1} in the spectra of colitis samples	[238]
To evaluate the performance of FTIR spectroscopy of the saliva for the diagnosis of cancer, namely, lung and breast cancer	FTIR	Statistically significant differences of lung cancer patients are observed at 1070–1240 cm^{-1}, while differences are observed for breast cancer patients in the	The amide I band in the normal group was found near 1655 cm^{-1} but was shifted in cancer patients to 1640 cm-1. The amide II band in the normal group had a maximum at 1547 cm^{-1}, while it was shifted to 1549 cm^{-1} for both major cancer patient groups. The	[239]

		entire spectral range studied	amide III band found in the normal patient group at 1300 cm^{-1} is shifted to 1310 cm^{-1} in lung cancer group, and 1315 cm^{-1} in breast cancer group. Shifts from 1740 to 1735 cm^{-1}, and 1393 to 1406 cm^{-1} for cancer patients were observed. The band 1240 cm^{-1} is shifted to 1242 cm^{-1} for breast cancer patients, and to 1244 cm^{-1} for lung cancer patients, while the band at 1075 cm^{-1} for the normal individuals is shifted to 1076–1078 cm^{-1} for both cancers.	
To evaluate FTIR spectroscopy as a method for identifying biochemical changes in sputum as biomarkers for detection of lung cancer	FTIR PCA HCA	Five prominent significant wavenumbers at 964 cm^{-1}, 1024 cm^{-1}, 1411 cm^{-1}, 1577 cm^{-1} and 1656 cm^{-1} separated cancer spectra from normal spectra into two distinct groups using multivariate analysis	PCA revealed that these wavenumbers were also able to distinguish lung cancer patients who had previously been diagnosed with breast cancer. No patterns of spectra groupings were associated with inflammation or other diseases of the airways	[248]
To predict diabetic status by analyzing the molecular and sub-molecular spectral signatures of saliva collected from subjects with diabetes and healthy controls	FTIR LDA	The overall accuracy based on infrared spectroscopy was 100% on the training set and 88.2% on the validation set.	The altered α-helix (1640 cm^{-1}) component is more obvious in the diabetic saliva spectra. The vibration of the tyrosine ring (1517 cm^{-1}) is altered in the diabetic group. The amide II band at 1550 cm^{-1} was less prominent in diabetic saliva than those from normal saliva, while the lipid ester band at 1735 cm^{-1} was more intense. The band at 1452 cm^{-1} also changed in diabetic group.	[240]
To analyze saliva proteomic components in psoriatic patients against diabetic patients and a control group using FTIR	ATR-FTIR PCA	Saliva spectra of the control group and palmoplantar psoriatic patients differ from plaque psoriasis and diabetic patient spectra due to the absence of the amide II band and the presence of different secondary protein-structure conformations	A prominent amide II band (1545 cm^{-1}) and amide I band (1647 cm^{-1}) allowed to distinguish the infrared salivary signature of psoriatic and diabetic patients from the control group and even from patients with different kinds of psoriasis	[241]
To characterize controlled and uncontrolled diabetic patients; clustering patients in groups low, medium, and high glucose levels; and finally performing the point estimation of a glucose value	ATR-FTIR SVM, ANN, LR	All the 540 spectra (100%) that make up the database were correctly characterized by studying the region 4000-2000 cm^{-1}	The region from 4000 to 2000 cm^{-1} lies mainly hydrogen bonding, which is a region ignored in most of the works since it is generally considered as a spectral silent region (2800-1800 cm^{-1})	[256]

To evaluate saliva of non-diabetic (ND), diabetic (D) and insulin-treated diabetic (D+I) rats to identify potential salivary biomarkers using ATR-FTIR	ATR-FTIR PCA-LDA, HCA	Classification of D rats was achieved with a sensitivity of 100%, and an average specificity of 93.33% and 100% using bands 1452 cm^{-1} and 836 cm^{-1}, respectively.	1452 cm^{-1} and 836 cm^{-1} spectral bands seems to be spectral biomarkers for diabetes, and highly correlated with glycemia. Both PCA-LDA and HCA classifications achieved an accuracy of 95.2% for the groups	[251]
To evaluate the diagnostic potential of periodontal disease by FTIR technique for saliva samples	FTIR	The leave-one-out cross-validation discrimination accuracy was 94.3%	Periodontal samples showed a larger raw IR spectrum than the control samples. Shape of the second derivative spectrum was clearly different between the periodontal and control samples.	[242]
To detect differences in composition of saliva supernatant in non-periodontitis individuals and patients with generalized aggressive periodontitis	ATR-FTIR PCA	Ten samples show in the analysis of variance of the two data sets a true difference (99.8%)	Spectral range between 1230 and 1180 cm^{-1}, or even of only two carefully selected wavelengths (1206 and 1196 cm^{-1}) is a promising tool for the analysis of saliva supernatant for the diagnosis of periodontitis	[243]
To determine the ability of FTIR spectroscopy to distinguish chronic periodontitis (CP) and aggressive periodontitis (AgP) patients by saliva samples and, to assess the potential confounding influence of smoking on discriminating disease-specific spectral signatures	FTIR HCA	Nonsmoker CP and AgP patients were discriminated from each other with high sensitivity and specificity. Successful differentiation was also obtained for the smoker CP and AgP groups. Thiocyanate levels successfully differentiated smokers from nonsmokers, irrespective of periodontal status, with 100% accuracy.	All smoker AgP samples were successfully discriminated from nonsmoker ones with 100% sensitivity and specificity for both spectral regions (thiocyanate band (2170–1900 cm^{-1}) and 3700–1850 cm^{-1} spectral region)	[244]
To characterize and determine specific spectral signatures in saliva from healthy, chronic periodontitis, and aggressive periodontitis patients using IR spectroscopy	FTIR LDA ANOVA	The overall accuracy for identifying the saliva samples as control or aggressive periodontitis was 73.9 % for the training set and 67.1 % for the validation set. The overall accuracy for classifying saliva samples as control or chronic periodontitis for the training set and the validation set was 67.7 % and 56.7 % respectively	Mean difference of IR spectra between control and periodontitis groups was significant for wavelengths 1087 cm^{-1}, 1240 cm^{-1} and 1652 cm^{-1} and 1740 cm^{-1}.	[252]
To compare salivary components between chronic kidney disease	ATR-FTIR ROC	92.8% sensitivity and 85.7% specificity for CKD detection	Thiocyanate (SCN-, 2052 cm^{-1}) and phospholipids/carbohydrates (924 cm^{-1}) could potentially be used as salivary	[201]

(CKD) patients and matched control subjects			biomarkers to differentiate CKD than control subjects	
To evaluate possible changes in saliva composition at the molecular level that can be associated with burning mouth syndrome (BMS)	FTIR PCA	Data obtained concludes the presence of alterations in saliva composition that may be directly related to BMS symptomatology	All bands showed the same or high intensity for the control group, except for the bands at 868 cm^{-1} and 2058 cm^{-1}, which corresponded respectively to nucleic acid and thiocyanate, and showed great intensity for patients with BMS	[245]

Limitations of FTIR spectroscopy exist. Although the method is fast, non-destructive and reagent-free, band overlapping restrictions causing inconclusive results of untreated samples cannot be ignored. Additionally, the sensitivity might not be as satisfactory as nucleic acid amplification-based diagnostics. However, it is easy to overcome these limitations by performing adequate spectral treatment and analysis, using effective multivariate analysis tool, and enlarging the spectral database.

In many FTIR spectroscopy studies, specific biomarkers were found in diseased patients according to their spectral bands and allowed successful discrimination between control and patients. As previously described, the binding of SARS-CoV-2 to ACE2 in salivary gland increase the amount of ANG II and may likely contributes to injury and inflammation in COVID-19 patients. This change in composition, as well as the presence of antibodies and other biomarkers triggered by the entry of the virus, could all be reflected on a spectral image. The proposed FTIR spectroscopic technique, combined with multivariate analytical tools, may not only allow the identification and classification of food-related microorganisms, but could also possibly monitor the chemical pathway to the progression of COVID-19 and identify any changes in the chemical structure of the virus that may occur.

2.6. Conclusion

Identification and classification of microorganisms is unquestionably important. The invention of more rapid and specific microbial detection and instrument automation allows to advance our knowledge of microorganism diversity. Traditional methods for microorganism detection and identification are still in use but are generally labor intensive and time consuming. Although MALDI-TOF MS, WGS and RT-PCR may lead to significant savings and in terms of specificity and accuracy compared to the conventional methods, the cost is still of concern, especially for developing countries. With implementation of FTIR spectroscopic-based methods, conclusive results with high confidence can be readily available, allowing for faster prescription of medication or triggering food recall action in a timely manner. For bacteria or fungi, FTIR spectroscopy is undoubtedly a promising analytical tool for discrimination even at the strain level. The utility of FTIR spectroscopy for species differentiation, especially between STEC from generic *E. coli*, and among *Salmonella* serogroups and *Staphylococcus* spp. will provide huge advancement in the microorganism identification field. For viruses, FTIR spectroscopy has the potential given its capacity of high discriminatory power. It is worth more attention in the investigation and validation of its potential as a screening tool to the current pandemic situation, in order to achieve governmental mass screening goals aimed at limiting the spread of the disease.

It is important to note that no single identification method will have 100% accuracy. Each method has its strengths and weaknesses, necessitating the use of multiple methods. Depending on the cost, available resources, the time that the microbiologist is prepared to wait, and the research question, choosing the appropriate techniques and understanding their limitations action is crucial to obtain the most precise result. FTIR spectroscopy requires in general lesser requirements in terms of finance and technician skills. As a result, an FTIR spectroscopic-based method, along with its advantages of rapid, non-destructive, label-free, and inexpensive, could be attractive for industries, hospitals, government surveillance microbiology laboratories for routine analysis. It can also be employed as a pre-screening step before undergoing tedious whole-genome sequencing methods, cutting down working load while attaining higher accuracy. The research into microbial quantification, identification and discrimination, as well as studies on microbial cellular modification in response to stress, by spectroscopic techniques and spectral imaging technologies will continuously be of interest in the future. The acceleration of developing portable spectroscopic

and imaging systems with simplicity and reliability, and the reduction of cost should facilitate adoption of these technologies in all research fields.

2.7. References

1. Manafi, M., *Fluorogenic and chromogenic enzyme substrates in culture media and identification tests.* Int J Food Microbiol, 1996. **31**(1-3): p. 45-58.
2. *Microbiology by numbers.* Nature Reviews Microbiology, 2011. **9**(9): p. 628-628.
3. Canada, S. *Table 13-10-0141-01 Deaths, by cause, Chapter I: Certain infectious and parasitic diseases (A00 to B99)*. 2020; Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310014101.
4. Aguilera, G., et al., *Identification and Typing Methods for the Study of Bacterial Infections: a Brief Review and Mycobacterial as Case of Study.* Archives of Clinical Microbiology, 2015. **7**: p. 3.
5. Kosa, G., et al., *FTIR spectroscopy as a unified method for simultaneous analysis of intra- and extracellular metabolites in high-throughput screening of microbial bioprocesses.* Microbial Cell Factories, 2017. **16**(1): p. 195.
6. organization, W.h. *Food Safety*. 2020 [cited 2020 March 20]; Available from: https://www.who.int/news-room/fact-sheets/detail/food-safety.
7. PHAC, I. *Food-related Illnesses, Hospitalizations and Deaths in Canada.* 2020 [cited 2020 October 20]; Available from: https://www.canada.ca/en/public-health/services/publications/food-nutrition/infographic-food-related-illnesses-hospitalizations-deaths-in-canada.html
8. Hoffmann, S., M.B. Batz, and J.G. Morris, Jr., *Annual cost of illness and quality-adjusted life year losses in the United States due to 14 foodborne pathogens.* J Food Prot, 2012. **75**(7): p. 1292-302.
9. !!! INVALID CITATION !!! [9].
10. Canada, G.o. *Canada's 10 least wanted foodborne pathogens*. 2011 [cited 2020 March 20]; Available from: http://www.wrha.mb.ca/community/seniors/files/CMP-18.pdf.
11. Exner, M., et al., *Antibiotic resistance: What is so special about multidrug-resistant Gram-negative bacteria?* GMS Hyg Infect Control, 2017. **12**: p. Doc05.
12. Athumani, L., *Isolation and Characterization of Escherichia coli from Animals, Humans, and Environment*. 2017.
13. Chekabab, S.M., et al., *The ecological habitat and transmission of Escherichia coli O157:H7.* FEMS Microbiology Letters, 2013. **341**(1): p. 1-12.
14. Kirk, M.D., et al., *World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis.* PLoS Med, 2015. **12**(12): p. e1001921.
15. Beneduce, L., G. Spano, and S. Massa, *Escherichia coli O157:H7 general characteristics, isolation and identification techniques.* Annals of Microbiology, 2003. **53**: p. 511-527.
16. Fan, R., et al., *High prevalence of non-O157 Shiga toxin-producing Escherichia coli in beef cattle detected by combining four selective agars.* BMC Microbiology, 2019. **19**(1): p. 213.
17. Kim, S.A., et al., *Rapid and simple method by combining FTA™ card DNA extraction with two set multiplex PCR for simultaneous detection of non-O157 Shiga toxin-producing Escherichia coli strains and virulence genes in food samples.* Lett Appl Microbiol, 2017. **65**(6): p. 482-488.
18. Gal-Mor, O., E.C. Boyle, and G.A. Grassl, *Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ.* Front Microbiol, 2014. **5**: p. 391.

19. EFSA. *Salmonella the most common cause of foodborne outbreaks in the European Union*. 2019; Available from: https://www.efsa.europa.eu/en/news/salmonella-most-common-cause-foodborne-outbreaks-european-union.
20. Taylor, J., et al., *An outbreak of salmonella chester infection in Canada: rare serotype, uncommon exposure, and unusual population demographic facilitate rapid identification of food vehicle.* J Food Prot, 2012. **75**(4): p. 738-42.
21. organization, W.h. *Salmonella (non-typhoidal).* . 2020 [cited 2020 March 20]; Available from: https://www.who.int/en/news-room/fact-sheets/detail/salmonella-(non-typhoidal).
22. Zhao, S., et al., *Characterization of Salmonella enterica serotype newport isolated from humans and food animals.* J Clin Microbiol, 2003. **41**(12): p. 5366-71.
23. Nair, A., et al., *Isolation and identification of Salmonella from diarrheagenic infants and young animals, sewage waste and fresh vegetables.* Vet World, 2015. **8**(5): p. 669-73.
24. Feng, X., et al., *Evaluation of real-time nanopore sequencing for Salmonella serotype prediction.* Food Microbiology, 2020. **89**: p. 103452.
25. Sauders, B.D., et al., *Diversity of Listeria species in urban and natural environments.* Appl Environ Microbiol, 2012. **78**(12): p. 4420-33.
26. FDA. *Bad Bug Book, Foodborne Pathogenic Microorganisms and Natural Toxins*. 2012 [cited 2020 March 20]; 2nd edition:[Available from: https://www.fda.gov/Food/FoodborneIllnessContaminants/CausesOfIllnessBadBugBook/.
27. Prevention, C.f.D.C.a. *Listeria – Listeriosis*. 2020 [cited 2020 March 20]; Available from: https://www.cdc.gov/listeria/.
28. Angelo, K., et al., *Multistate outbreak of Listeria monocytogenes infections linked to whole apples used in commercially produced, prepackaged caramel apples: United States, 2014–2015.* Epidemiology and Infection, 2017. **145**: p. 1-9.
29. FDA. *Environmental Assessment: Factors Potentially Contributing to the Contamination of Fresh Whole Cantaloupe Implicated in a Multi-State Outbreak of Listeriosis.* . 2011 [cited 2020 March 20]; Available from: https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm276247.htm.
30. Bintsis, T., *Foodborne pathogens.* AIMS Microbiol, 2017. **3**(3): p. 529-563.
31. Gasanov, U., D. Hughes, and P.M. Hansbro, *Methods for the isolation and identification of Listeria spp. and Listeria monocytogenes: a review.* FEMS Microbiology Reviews, 2005. **29**(5): p. 851-875.
32. Hitchins AD, J.K., Chen Y. *Detection of Listeria monocytogenes in Foods and Environmental Samples, and Enumeration of Listeria monocytogenes in Foods*. Bacteriological Analytical Manual 2017 [cited 2020 March 20]; Available from: https://www.fda.gov/food/laboratory-methods-food/bam-detection-and-enumeration-listeria-monocytogenes.
33. Byrd-Bredbenner, C., et al., *Food safety in home kitchens: a synthesis of the literature.* Int J Environ Res Public Health, 2013. **10**(9): p. 4060-85.
34. Bacon, R., et al., *Characteristics of Biological Hazards in Foods*. 2005. p. 157-195.
35. Hennekinne, J.A., M.L. De Buyser, and S. Dragacci, *Staphylococcus aureus and its food poisoning toxins: characterization and outbreak investigation.* FEMS Microbiol Rev, 2012. **36**(4): p. 815-36.
36. Fitzgerald, J.R., *Livestock-associated Staphylococcus aureus: origin, evolution and public health threat.* Trends Microbiol, 2012. **20**(4): p. 192-8.
37. Zecconi, A. and F. Scali, *Staphylococcus aureus virulence factors in evasion from innate*

immune defenses in human and animal diseases. Immunol Lett, 2013. **150**(1-2): p. 12-22.
38. Dufour, S., J. Labrie, and M. Jacques, *The Mastitis Pathogens Culture Collection.* Microbiol Resour Announc, 2019. **8**(15).
39. Leck, A., *Preparation of lactophenol cotton blue slide mounts.* Community Eye Health, 1999. **12**(30): p. 24.
40. Bianchi, D.E., *DIFFERENTIAL STAINING OF YEAST FOR PURIFIED CELL WALLS, BROKEN CELLS, AND WHOLE CELLS.* Stain Technol, 1965. **40**: p. 79-82.
41. Warnock, D.W., *61 - Fungi: Superficial, subcutaneous and systemic mycoses*, in *Medical Microbiology (Eighteenth Edition)*, D. Greenwood, et al., Editors. 2012, Churchill Livingstone: Edinburgh. p. 616-641.
42. Hawksworth, D.L., *The magnitude of fungal diversity: the 1.5 million species estimate revisited* *Paper presented at the Asian Mycological Congress 2000 (AMC 2000), incorporating the 2nd Asia-Pacific Mycological Congress on Biodiversity and Biotechnology, and held at the University of Hong Kong on 9-13 July 2000.* Mycological Research, 2001. **105**(12): p. 1422-1432.
43. Canada, G.o. *Natural toxins*. 2012 [cited 2020 December 21]; Available from: https://www.canada.ca/en/health-canada/services/food-nutrition/food-safety/chemical-contaminants/natural-toxins.html.
44. Garey, K.W., et al., *Treatment of Candida famata bloodstream infections: case series and review of the literature.* J Antimicrob Chemother, 2013. **68**(2): p. 438-43.
45. Gonçalves, F.A., G. Colen, and J.A. Takahashi, *Yarrowia lipolytica and its multiple applications in the biotechnological industry.* ScientificWorldJournal, 2014. **2014**: p. 476207.
46. Zieniuk, B. and A. Fabiszewska, *Yarrowia lipolytica: a beneficious yeast in biotechnology as a rare opportunistic fungal pathogen: a minireview.* World Journal of Microbiology and Biotechnology, 2018. **35**(1): p. 10.
47. Kremery, V., I. Krupova, and D.W. Denning, *Invasive yeast infections other than Candida spp. in acute leukaemia.* Journal of Hospital Infection, 1999. **41**(3): p. 181-194.
48. Meier, S.M., et al., *Proteomic and Metabolomic Analyses Reveal Contrasting Anti-Inflammatory Effects of an Extract of Mucor Racemosus Secondary Metabolites Compared to Dexamethasone.* PLOS ONE, 2015. **10**(10): p. e0140367.
49. García-Ruiz, J.C., et al., *Invasive infections caused by Saprochaete capitata in patients with haematological malignancies: report of five cases and review of the antifungal therapy.* Rev Iberoam Micol, 2013. **30**(4): p. 248-55.
50. Bensch, K., et al., *The genus Cladosporium.* Stud Mycol, 2012. **72**(1): p. 1-401.
51. Frisvad, J. and R. Samson, *Polyphasic taxonomy of Penicillium subgenus Penicillium. A guide to identif cation of food and airborne terverticillate Penicillia and their mycotoxins.* Studies in Mycology, 2004. **2004**: p. 1-173.
52. Banerjee, S., et al., *Identification of fungal pathogens in a patient with acute myelogenic leukemia using a pathogen detection array technology.* Cancer Biol Ther, 2016. **17**(4): p. 339-45.
53. Theel, E.S., et al., *Dermatophyte identification using matrix-assisted laser desorption ionization-time of flight mass spectrometry.* J Clin Microbiol, 2011. **49**(12): p. 4067-71.
54. organization, W.h. *Laboratory testing for coronavirus disease (COVID-19) in suspected human cases: interim guidance.* . 2020; Available from: WHO/COVID-19/laboratory/2020.5.

55. Ratnadip, C. *3-Day-Old Baby Dies In Tripura Covid Ward, Family Blames Swab For Test*. NDTV 2020 [cited 2020 October 6]; Available from: https://www.ndtv.com/india-news/tripura-agartala-3-day-old-baby-dies-in-tripura-covid-ward-family-blames-swab-for-test-2279083.
56. Tuqa, K. *Coronavirus: Saudi Arabian child dies after COVID-19 test swab breaks in his nose*. Al Arabiya 2020 [cited 2020 October 6]; Available from: https://english.alarabiya.net/en/coronavirus/2020/07/15/Coronavirus-Saudi-child-dies-due-to-a-COVID-19-test-swab-breaking-in-his-nose.
57. Kaufman, E. and I.B. Lamster, *The Diagnostic Applications of Saliva— A Review.* Critical Reviews in Oral Biology & Medicine, 2002. **13**(2): p. 197-212.
58. Chiappin, S., et al., *Saliva specimen: a new laboratory tool for diagnostic and basic investigation.* Clin Chim Acta, 2007. **383**(1-2): p. 30-40.
59. Zhang, C.Z., et al., *Saliva in the diagnosis of diseases.* Int J Oral Sci, 2016. **8**(3): p. 133-7.
60. Chojnowska, S., et al., *Human saliva as a diagnostic material.* Advances in Medical Sciences, 2018. **63**(1): p. 185-191.
61. Fatima, S., et al., *COMPOSITION AND FUNCTION OF SALIVA: A REVIEW.* WORLD JOURNAL OF PHARMACY AND PHARMACEUTICAL SCIENCES, 2020. **9**: p. 1552-1567.
62. Hamid, H., et al., *COVID-19 Pandemic and Role of Human Saliva as a Testing Biofluid in Point-of-Care Technology.* Eur J Dent, 2020. **14**(S 01): p. S123-s129.
63. Khurshid, Z., et al., *Role of Salivary Biomarkers in Oral Cancer Detection.* Adv Clin Chem, 2018. **86**: p. 23-70.
64. Abdul Rehman, S., et al., *Role of Salivary Biomarkers in Detection of Cardiovascular Diseases (CVD).* Proteomes, 2017. **5**(3).
65. Wang, W.K., et al., *Detection of SARS-associated coronavirus in throat wash and saliva in early diagnosis.* Emerg Infect Dis, 2004. **10**(7): p. 1213-9.
66. Liu, L., et al., *Epithelial cells lining salivary gland ducts are early target cells of severe acute respiratory syndrome coronavirus infection in the upper respiratory tracts of rhesus macaques.* J Virol, 2011. **85**(8): p. 4025-30.
67. To, K.K.W., et al., *Saliva as a diagnostic specimen for testing respiratory virus by a point-of-care molecular assay: a diagnostic validity study.* Clin Microbiol Infect, 2019. **25**(3): p. 372-378.
68. Chen, J.H., et al., *Evaluating the use of posterior oropharyngeal saliva in a point-of-care assay for the detection of SARS-CoV-2.* Emerg Microbes Infect, 2020. **9**(1): p. 1356-1359.
69. Williams, E., et al., *Saliva as a Noninvasive Specimen for Detection of SARS-CoV-2.* J Clin Microbiol, 2020. **58**(8).
70. Liu, Y., et al., *Viral dynamics in mild and severe cases of COVID-19.* Lancet Infect Dis, 2020. **20**(6): p. 656-657.
71. Nagura-Ikeda, M., et al., *Clinical Evaluation of Self-Collected Saliva by Quantitative Reverse Transcription-PCR (RT-qPCR), Direct RT-qPCR, Reverse Transcription-Loop-Mediated Isothermal Amplification, and a Rapid Antigen Test To Diagnose COVID-19.* J Clin Microbiol, 2020. **58**(9).
72. Chris Tikellis and M.C. Thomas, *Angiotensin-Converting Enzyme 2 (ACE2) Is a Key Modulator of the Renin Angiotensin System in Health and Disease.* Int J Pept, 2012. **2012**: p. 256294.
73. Pedrosa MS, S.C., Nogueira FN. , *Salivary glands, saliva and oral findings in COVID-19*

infection. . Pesqui Bras Odontopediatria Clín Integr, 2020. **2020; 20(supp1):e0104**.
74. Xu, H., et al., *High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa.* International Journal of Oral Science, 2020. **12**(1): p. 8.
75. Sriram K., I.P., Loomba R. . *What is the ACE2 receptor, how is it connected to coronavirus and why might it be key to treating COVID-19? The experts explain.* . 2020 [cited 2020 October 8]; Available from: https://theconversation.com/what-is-the-ace2-receptor-how-is-it-connected-to-coronavirus-and-why-might-it-be-key-to-treating-covid-19-the-experts-explain-136928
76. Baghizadeh Fini, M., *Oral saliva and COVID-19.* Oral Oncol, 2020. **108**: p. 104821.
77. Corstjens, P.L., W.R. Abrams, and D. Malamud, *Detecting viruses by using salivary diagnostics.* J Am Dent Assoc, 2012. **143**(10 Suppl): p. 12s-8s.
78. Thomas, G., *Furin at the cutting edge: from protein traffic to embryogenesis and disease.* Nat Rev Mol Cell Biol, 2002. **3**(10): p. 753-66.
79. Coutard, B., et al., *The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade.* Antiviral Res, 2020. **176**: p. 104742.
80. Johnson, B.A., et al., *Furin Cleavage Site Is Key to SARS-CoV-2 Pathogenesis.* bioRxiv, 2020.
81. Mardani, R., et al., *Laboratory Parameters in Detection of COVID-19 Patients with Positive RT-PCR; a Diagnostic Accuracy Study.* Arch Acad Emerg Med, 2020. **8**(1): p. e43.
82. Han, M.S., et al., *Sequential Analysis of Viral Load in a Neonate and Her Mother Infected With Severe Acute Respiratory Syndrome Coronavirus 2.* Clin Infect Dis, 2020. **71**(16): p. 2236-2239.
83. Zhang, W., et al., *Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes.* Emerg Microbes Infect, 2020. **9**(1): p. 386-389.
84. Hung, D.L., et al., *Early-Morning vs Spot Posterior Oropharyngeal Saliva for Diagnosis of SARS-CoV-2 Infection: Implication of Timing of Specimen Collection for Community-Wide Screening.* Open Forum Infect Dis, 2020. **7**(6): p. ofaa210.
85. Iwasaki, S., et al., *Comparison of SARS-CoV-2 detection in nasopharyngeal swab and saliva.* J Infect, 2020. **81**(2): p. e145-e147.
86. Cheng, V.C.C., et al., *Escalating infection control response to the rapidly evolving epidemiology of the coronavirus disease 2019 (COVID-19) due to SARS-CoV-2 in Hong Kong.* Infect Control Hosp Epidemiol, 2020. **41**(5): p. 493-498.
87. Helmerhorst, E.J., C. Dawes, and F.G. Oppenheim, *The complexity of oral physiology and its impact on salivary diagnostics.* Oral Dis, 2018. **24**(3): p. 363-371.
88. Martinez-Cuazitl, A., et al., *ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients.* Scientific Reports, 2021. **11**(1): p. 19980.
89. Barauna, V.G., et al., *Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity.* Analytical Chemistry, 2021. **93**(5): p. 2950-2958.
90. Nascimento, M.H.C., et al., *Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis.* Anal Chem, 2022. **94**(5): p. 2425-2433.
91. Wood, B.R., et al., *Infrared Based Saliva Screening Test for COVID-19.* Angew Chem Int Ed Engl, 2021. **60**(31): p. 17102-17107.
92. Azzi, L., et al., *Saliva is a reliable tool to detect SARS-CoV-2.* J Infect, 2020. **81**(1): p. e45-e50.

93. Azzi, L., et al., *Two cases of COVID-19 with positive salivary and negative pharyngeal or respiratory swabs at hospital discharge: A rising concern.* Oral Dis, 2021. **27 Suppl 3**(Suppl 3): p. 707-709.
94. Azzi, L., et al., *Rapid Salivary Test suitable for a mass screening program to detect SARS-CoV-2: A diagnostic accuracy study.* Journal of Infection, 2020. **81**.
95. Bosworth, A., et al., *Rapid implementation and validation of a cold-chain free SARS-CoV-2 diagnostic testing workflow to support surge capacity.* J Clin Virol, 2020. **128**: p. 104469.
96. Caulley, L., et al., *Salivary Detection of COVID-19.* Ann Intern Med, 2021. **174**(1): p. 131-133.
97. Chau Van Vinh, N., et al., *The Natural History and Transmission Potential of Asymptomatic Severe Acute Respiratory Syndrome Coronavirus 2 Infection.* Clinical Infectious Diseases, 2020. **71**(10): p. 2679-2687.
98. Chen, L., et al., *Detection of SARS-CoV-2 in saliva and characterization of oral symptoms in COVID-19 patients.* Cell Prolif, 2020. **53**(12): p. e12923.
99. Fang, Z., et al., *Comparisons of viral shedding time of SARS-CoV-2 of different samples in ICU and non-ICU patients.* J Infect, 2020. **81**(1): p. 147-178.
100. Han, M.S., et al., *Viral RNA Load in Mildly Symptomatic and Asymptomatic Children with COVID-19, Seoul, South Korea.* Emerg Infect Dis, 2020. **26**(10): p. 2497-2499.
101. Hanson, K.E., et al., *Self-Collected Anterior Nasal and Saliva Specimens versus Health Care Worker-Collected Nasopharyngeal Swabs for the Molecular Detection of SARS-CoV-2.* J Clin Microbiol, 2020. **58**(11).
102. Jamal, A.J., et al., *Sensitivity of Nasopharyngeal Swabs and Saliva for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2.* Clinical Infectious Diseases, 2020. **72**(6): p. 1064-1066.
103. Landry, M.L., J. Criscuolo, and D.R. Peaper, *Challenges in use of saliva for detection of SARS CoV-2 RNA in symptomatic outpatients.* J Clin Virol, 2020. **130**: p. 104567.
104. Leung, E.C., et al., *Deep throat saliva as an alternative diagnostic specimen type for the detection of SARS-CoV-2.* J Med Virol, 2021. **93**(1): p. 533-536.
105. Mak, G.C., et al., *Evaluation of rapid antigen test for detection of SARS-CoV-2 virus.* J Clin Virol, 2020. **129**: p. 104500.
106. McCormick-Baw, C., et al., *Saliva as an Alternate Specimen Source for Detection of SARS-CoV-2 in Symptomatic Patients Using Cepheid Xpert Xpress SARS-CoV-2.* J Clin Microbiol, 2020. **58**(8).
107. Migueres, M., et al., *Saliva sampling for diagnosing SARS-CoV-2 infections in symptomatic patients and asymptomatic carriers.* J Clin Virol, 2020. **130**: p. 104580.
108. Miller, M., et al., *Validation of a Self-administrable, Saliva-based RT-qPCR Test Detecting SARS-CoV-2*. 2020.
109. S, N.V., et al., *Saliva is a reliable, non-invasive specimen for SARS-CoV-2 detection.* Braz J Infect Dis, 2020. **24**(5): p. 422-427.
110. Pasomsub, E., et al., *Saliva sample as a non-invasive specimen for the diagnosis of coronavirus disease 2019: a cross-sectional study.* Clin Microbiol Infect, 2021. **27**(2): p. 285.e1-285.e4.
111. Gary W. Procop, et al., *A Direct Comparison of Enhanced Saliva to Nasopharyngeal Swab for the Detection of SARS-CoV-2 in Symptomatic Patients.* J Clin Microbiol, 2020. **58**(11).
112. Rao, M., et al., *Comparing Nasopharyngeal Swab and Early Morning Saliva for the Identification of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).* Clin

Infect Dis, 2021. **72**(9): p. e352-e356.
113. Skolimowska, K., et al., *Non-invasive saliva specimens for the diagnosis of COVID-19: caution in mild outpatient cohorts with low prevalence.* Clin Microbiol Infect, 2020. **26**(12): p. 1711-1713.
114. Sutjipto, S., et al., *The Effect of Sample Site, Illness Duration, and the Presence of Pneumonia on the Detection of SARS-CoV-2 by Real-time Reverse Transcription PCR.* Open Forum Infect Dis, 2020. **7**(9): p. ofaa335.
115. Tajima, Y., Y. Suda, and K. Yano, *A case report of SARS-CoV-2 confirmed in saliva specimens up to 37 days after onset: Proposal of saliva specimens for COVID-19 diagnosis and virus monitoring.* J Infect Chemother, 2020. **26**(10): p. 1086-1089.
116. To, K.K., et al., *Consistent Detection of 2019 Novel Coronavirus in Saliva.* Clin Infect Dis, 2020. **71**(15): p. 841-843.
117. To, K.K., et al., *Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study.* Lancet Infect Dis, 2020. **20**(5): p. 565-574.
118. Uwamino, Y., et al., *Accuracy and stability of saliva as a sample for reverse transcription PCR detection of SARS-CoV-2.* J Clin Pathol, 2021. **74**(1): p. 67-68.
119. Wong, S.C.Y., et al., *Posterior Oropharyngeal Saliva for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).* Clin Infect Dis, 2020. **71**(11): p. 2939-2946.
120. Wyllie, A.L., et al., *Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs.* medRxiv, 2020: p. 2020.04.16.20067835.
121. Yang, J.R., et al., *Persistent viral RNA positivity during the recovery period of a patient with SARS-CoV-2 infection.* J Med Virol, 2020. **92**(9): p. 1681-1683.
122. Yokota, I., et al., *Mass Screening of Asymptomatic Persons for Severe Acute Respiratory Syndrome Coronavirus 2 Using Saliva.* Clin Infect Dis, 2021. **73**(3): p. e559-e565.
123. Yoon, J.G., et al., *Clinical Significance of a High SARS-CoV-2 Viral Load in the Saliva.* J Korean Med Sci, 2020. **35**(20): p. e195.
124. Zheng S, et al., *Hock-a-loogie saliva as a diagnostic specimen for SARS-CoV-2 by a PCR-based assay: A diagnostic validity study.* Clin Chim Acta, 2020. **511**: p. 177-180.
125. Guleken, Z., et al., *Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy, and machine learning classifications.* Talanta, 2022. **237**: p. 122916.
126. Banerjee, A., et al., *Rapid Classification of COVID-19 Severity by ATR-FTIR Spectroscopy of Plasma Samples.* Analytical Chemistry, 2021. **93**(30): p. 10391-10396.
127. Zhang, L., et al., *Fast Screening and Primary Diagnosis of COVID-19 by ATR–FT-IR.* Analytical Chemistry, 2021. **93**(4): p. 2191-2199.
128. Kitane, D.L., et al., *A simple and fast spectroscopy-based technique for Covid-19 diagnosis.* Scientific Reports, 2021. **11**(1): p. 16740.
129. Nogueira, M.S., et al., *Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning.* Scientific Reports, 2021. **11**(1): p. 15409.
130. Donelli, G., C. Vuotto, and P. Mastromarino, *Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism.* Microb Ecol Health Dis, 2013. **24**.
131. Pradhan, P. and J.P. Tamang, *Phenotypic and Genotypic Identification of Bacteria Isolated From Traditionally Prepared Dry Starters of the Eastern Himalayas.* Front Microbiol,

2019. **10**: p. 2526.
132. Günther, S.K., et al., *Microbiological Control of Cellular Products: The Relevance of the Cellular Matrix, Incubation Temperature, and Atmosphere for the Detection Performance of Automated Culture Systems.* Transfus Med Hemother, 2020. **47**(3): p. 254-263.
133. Benjamin Caballero, P.F., Fidel Toldra, *Encyclopedia of Food Sciences and Nutrition*. 2003: Academic Press; 2 edition (May 28 2003). 6000.
134. Wu, J., et al., *Rapid Detection of Escherichia coli in Water Using Sample Concentration and Optimized Enzymatic Hydrolysis of Chromogenic Substrates.* Curr Microbiol, 2018. **75**(7): p. 827-834.
135. Engels, W., M.A. Kamps, and C.P. van Boven, *Rapid and direct staphylocoagulase assay that uses a chromogenic substrate for identification of Staphylococcus aureus.* J Clin Microbiol, 1981. **14**(5): p. 496-500.
136. Bascomb, S. and M. Manafi, *Use of enzyme tests in characterization and identification of aerobic and facultatively anaerobic gram-positive cocci.* Clin Microbiol Rev, 1998. **11**(2): p. 318-40.
137. D. H, P., S. Orenga, and S. Chatellier, *Yeast identification--past, present, and future methods.* Med Mycol, 2007. **45**(2): p. 97-121.
138. Fratamico, P.M., et al., *Advances in Molecular Serotyping and Subtyping of Escherichia coli.* Front Microbiol, 2016. **7**: p. 644.
139. Strid, M.A., et al., *Antibody responses to Campylobacter infections determined by an enzyme-linked immunosorbent assay: 2-year follow-up study of 210 patients.* Clin Diagn Lab Immunol, 2001. **8**(2): p. 314-9.
140. Palumbo, J.D., et al., *Serotyping of Listeria monocytogenes by enzyme-linked immunosorbent assay and identification of mixed-serotype cultures by colony immunoblotting.* J Clin Microbiol, 2003. **41**(2): p. 564-71.
141. Cheng, L.W. and L.H. Stanker, *Detection of botulinum neurotoxin serotypes A and B using a chemiluminescent versus electrochemiluminescent immunoassay in food and serum.* J Agric Food Chem, 2013. **61**(3): p. 755-60.
142. Baggesen, D.L., et al., *Phage typing of Salmonella Typhimurium - is it still a useful tool for surveillance and outbreak investigation?* Euro Surveill, 2010. **15**(4): p. 19471.
143. Sutton, S., *Microbial identification in a GXP environment--which system is best?* Journal of GXP Compliance, 2012. **vol. 16, no. 2, Spring 2012, p. 55**.
144. Klingler, J.M., et al., *Evaluation of the Biolog automated microbial identification system.* Appl Environ Microbiol, 1992. **58**(6): p. 2089-92.
145. Singhal, N., et al., *MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis.* Front Microbiol, 2015. **6**: p. 791.
146. Savage, E., et al., *Evaluation of Three Bacterial Identification Systems for Species Identification of Bacteria Isolated from Bovine Mastitis and Bulk Tank Milk Samples.* Foodborne Pathog Dis, 2017. **14**(3): p. 177-187.
147. Dupont, C., et al., *Identification of clinical coagulase-negative staphylococci, isolated in microbiology laboratories, by matrix-assisted laser desorption/ionization-time of flight mass spectrometry and two automated systems.* Clin Microbiol Infect, 2010. **16**(7): p. 998-1004.
148. Váradi, L., et al., *Methods for the detection and identification of pathogenic bacteria: past, present, and future.* Chem Soc Rev, 2017. **46**(16): p. 4818-4832.
149. Salman, A., et al., *Detection of antibiotic resistant Escherichia Coli bacteria using infrared*

microscopy and advanced multivariate analysis. The Analyst, 2017. **142**: p. 2136-2144.
150. Yang, L., et al., *Rapid Differentiation and Identification of Shigella sonnei and Escherichia coli O157: H7 by Fourier Transform Infrared Spectroscopy and Multivariate Statistical Analysis.* Advanced Materials Research, 2014. **926-930**: p. 1116-1119.
151. Rodriguez-Saona, L. and M. Allendorf, *Use of FTIR for Rapid Authentication and Detection of Adulteration of Food.* Annual review of food science and technology, 2011. **2**: p. 467-83.
152. Schabauer, L., et al., *Novel physico-chemical diagnostic tools for high throughput identification of bovine mastitis associated gram-positive, catalase-negative cocci.* BMC Vet Res, 2014. **10**: p. 156.
153. Tim, S. *Automated Microbial Identification: A Comparison of USP and EP Approaches*. 2013; Available from: https://www.americanpharmaceuticalreview.com/Featured-Articles/140520-Automated-Microbial-Identification-A-Comparison-of-USP-and-EP-Approaches/.
154. Iversen, C., et al., *The taxonomy of Enterobacter sakazakii: proposal of a new genus Cronobacter gen. nov. and descriptions of Cronobacter sakazakii comb. nov. Cronobacter sakazakii subsp. sakazakii, comb. nov., Cronobacter sakazakii subsp. malonaticus subsp. nov., Cronobacter turicensis sp. nov., Cronobacter muytjensii sp. nov., Cronobacter dublinensis sp. nov. and Cronobacter genomospecies 1.* BMC Evol Biol, 2007. **7**: p. 64.
155. Farrance, C.E., *Identification of Microorganisms*, in *Pharmaceutical Microbiological Quality Assurance and Control*. 2019. p. 265-328.
156. Malorny, B., et al., *Standardization of diagnostic PCR for the detection of foodborne pathogens.* Int J Food Microbiol, 2003. **83**(1): p. 39-48.
157. den Bakker, H.C., et al., *Rapid whole-genome sequencing for surveillance of Salmonella enterica serovar enteritidis.* Emerg Infect Dis, 2014. **20**(8): p. 1306-14.
158. Fontana, C., et al., *Use of the MicroSeq 500 16S rRNA gene-based sequencing for identification of bacterial isolates that commercial automated systems failed to identify correctly.* J Clin Microbiol, 2005. **43**(2): p. 615-9.
159. Jeffrey, M., et al., *Validation of an Enhanced Method of Bacterial Ribotyping: For Improved Efficiency and Identification of Stressed Isolates.* Pharmaceutical Technology, 2004. **28**: p. 156-166.
160. Healy, M., et al., *Microbial DNA typing by automated repetitive-sequence-based PCR.* J Clin Microbiol, 2005. **43**(1): p. 199-207.
161. Özdemir, T., *FT-IR, Laser-Raman, UV-Vis, and NMR Spectroscopic Studies of Antidiabetic Molecule Nateglinide.* Journal of Spectroscopy, 2018. **2018**: p. 1-12.
162. Kandpal L.M., C.B.K., *A Review of the Applications of Spectroscopy for the Detection of Microbial Contaminations and Defects in Agro Foods.* Journal of Biosystems Engineering 2014. **2014; 39(3): 215-226**.
163. Exline, D. *Comparison of Raman and FTIR Spectroscopy: Advantages and Limitations.* . Industry Resources 2013 [cited 2020 December 28]; Available from: https://gatewayanalytical.com/resources/publications/comparison-raman-and-ftir-spectroscopy-advantages-and-limitations/.
164. Wartewig, S., *IR and Raman Spectroscopy: Fundamental Processing*. 2006: Wiley.
165. Rees, O.J., *Fourier Transform Infrared Spectroscopy: Developments, Techniques and Applications*. 2010: Nova Science Pub Inc; UK ed. edition (Sept. 30 2010). 215.
166. Alvarez-Ordóñez Avelino, M.P., *Fourier Transform Infrared Spectroscopy in Food*

Microbiology. 1 ed. SpringerBriefs in Food, Health, and Nutrition. 2012: Springer New York, NY. VI, 55.
167. Thompson, J.M., *Infrared Spectroscopy*. 2018, New York: Pan Stanford Publishing Pte. Ltd.
168. PerkinElmer. *A Combined Mid-IR/Far-IR Spectrometer for Analytical and Research Applications.* . Technical note Infrared Spectroscopy 2008 [cited 2020 December 28]; Available from: https://www.perkinelmer.com/lab-solutions/resources/docs/TCH_MidFarIR.pdf.
169. Vatansever, F. and M.R. Hamblin, *Far infrared radiation (FIR): its biological effects and medical applications.* Photonics Lasers Med, 2012. **4**: p. 255-266.
170. Erdoğdu, S.B. and H. Ekiz, *Effect of ultraviolet and far infrared radiation on microbial decontamination and quality of cumin seeds.* J Food Sci, 2011. **76**(5): p. M284-92.
171. Sun, D.-W., *Infrared Spectroscopy for Food Quality Analysis and Control*. 1st Edition ed. 2009: Academic Press.
172. bellon maurel, V. and A. McBratney, *Near-Infrared (NIR) and Mid-Infrared (MIR) Spectroscopic Techniques for Assessing the Amount of Carbon Stock in Soils—Critical Review and Research Perspectives.* Soil Biology & Biochemistry - SOIL BIOL BIOCHEM, 2011. **43**: p. 1398-1410.
173. Ellis, D.I., G.G. Harrigan, and R. Goodacre, *Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy*, in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, G.G. Harrigan and R. Goodacre, Editors. 2003, Springer US: Boston, MA. p. 111-124.
174. IUPAC. *Compendium of Chemical Terminology*. 1997; 2nd edition:[Available from: https://www.ufjf.br/baccan/files/2011/05/goldbook-IUPAC1.pdf.
175. Corporation, T.N. *Introduction to Fourier Transform Infrared Spectrometry. ISO 9001*. A Thermo Electron business 2001; Available from: http://assets.thermofisher.com/TFS-Assets/MSD/brochures/introduction-fourier-transform-infrared-spectroscopy-br50555.pdf.
176. Wang, Y., et al., *Differentiation in MALDI-TOF MS and FTIR spectra between two closely related species Acidovorax oryzae and Acidovorax citrulli.* BMC Microbiology, 2012. **12**(1): p. 182.
177. Suntsova, A., et al., *Identification of microorganisms by Fourier-transform infrared spectroscopy.* Bulletin of Russian State Medical University, 2018. **7**: p. 50-57.
178. Quintelas, C., et al., *An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing.* Biotechnol J, 2018. **13**(1).
179. Erukhimovitch, V., et al., *FTIR microscopy as a method for identification of bacterial and fungal infections.* Journal of Pharmaceutical and Biomedical Analysis, 2005. **37**(5): p. 1105-1108.
180. Lam, L.M.T., et al., *Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy.* Journal of clinical microbiology, 2019. **57**(5): p. e01739-18.
181. M.Beasley, M., et al., *Comparison of transmission FTIR, ATR, and DRIFT spectra: implications for assessment of bone bioapatite diagenesis.* Journal of Archaeological Science, 2014. **46**: p. 16-22.
182. Pilling, M. and P. Bassan, *Comparison of Transmission and Transflectance Mode FTIR Imaging of Biological Tissue.* The Analyst, 2015. **140**.
183. Perro, A., et al., *Combining Microfluidics and FT-IR Spectroscopy: Towards spatially*

resolved information on chemical processes. Reaction Chemistry & Engineering, 2016. **1**.
184. Khoshhesab, Z.M., *Reflectance IR Spectroscopy, Infrared Spectroscopy – Materials Science, Engineering and Technology*, ed. T. Theophile. 2012.
185. PerkinElmer. *FT-IR Spectroscopy: Attenuated Total Reflectance (ATR).* . Technical note FT-IR Spectroscopy 2005 [cited 2020 April 3]; Available from: https://www.utsc.utoronto.ca/~traceslab/ATR_FTIR.pdf.
186. Smith, B.C., *Fundamentals of Fourier transform infrared spectroscopy*. 2nd ed. 2011, Boca Raton, FL: CRC Press. xiii, 193 p.
187. BH, S., *Infrared Spectroscopy: Fundamentals and Applications*. 2004: Wiley. 248.
188. Ravikanth, L., et al., *Extraction of Spectral Information from Hyperspectral Data and Application of Hyperspectral Imaging for Food and Agricultural Products.* Food and Bioprocess Technology, 2017. **10**(1): p. 1-33.
189. Lasch, P. and D. Naumann, *Infrared Spectroscopy in Microbiology*, in *Encyclopedia of Analytical Chemistry*. p. 1-32.
190. Smith, B.C., *Fundamentals of Fourier Transform Infrared Spectroscopy*. 2nd Edition ed. 2011. 207 Pages.
191. Davis, R., G. Paoli, and L.J. Mauer, *Evaluation of Fourier transform infrared (FT-IR) spectroscopy and chemometrics as a rapid approach for sub-typing Escherichia coli O157:H7 isolates.* Food Microbiol, 2012. **31**(2): p. 181-90.
192. Lee, L.C., C.-Y. Liong, and A.A. Jemain, *A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum.* Chemometrics and Intelligent Laboratory Systems, 2017. **163**: p. 64-75.
193. Naumann, D., *FT-INFRARED AND FT-RAMAN SPECTROSCOPY IN BIOMEDICAL RESEARCH.* Applied Spectroscopy Reviews, 2001. **36**(2-3): p. 239-298.
194. Tibaduiza, D.A., L.E. Mujica, and J. Rodellar, *Damage classification in structural health monitoring using principal component analysis and self-organizing maps.* Structural Control and Health Monitoring, 2013. **20**(10): p. 1303-1316.
195. Meksiarun, P., et al., *Comparison of multivariate analysis methods for extracting the paraffin component from the paraffin-embedded cancer tissue spectra for Raman imaging.* Scientific Reports, 2017. **7**(1): p. 44890.
196. Virtanen, K.A., et al., *Functional brown adipose tissue in healthy adults.* N Engl J Med, 2009. **360**(15): p. 1518-25.
197. Wu, D. and D.-W. Sun, *Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part II: Applications.* Innovative Food Science & Emerging Technologies, 2013. **19**: p. 15-28.
198. Byrne, L., et al., *A multi-country outbreak of Salmonella Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012.* Euro Surveill, 2014. **19**(31): p. 6-13.
199. University, T.P.S. *14.4 Agglomerative Hierarchical Clustering.* STAT505 Applied Multivariate Statistical Analysis. 2021 [cited 2020 December 28]; Available from: https://online.stat.psu.edu/stat505/lesson/14/14.4.
200. Brownlee, J., *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-end*. 2016: Jason Brownlee.
201. Rodrigues, R.P., et al., *Differential Molecular Signature of Human Saliva Using ATR-FTIR Spectroscopy for Chronic Kidney Disease Diagnosis.* Braz Dent J, 2019. **30**(5): p. 437-445.
202. Gomes, A., et al., *The successive projections algorithm for interval selection in PLS.*

Microchemical Journal, 2013. **110**: p. 202–208.
203. Zhou, Y., S. Zheng, and G. Zhang, *Artificial neural network based multivariable optimization of a hybrid system integrated with phase change materials, active cooling and hybrid ventilations.* Energy Conversion and Management, 2019. **197**: p. 111859.
204. Sudha, P., R. Duraisamy, and P. Manikandan, *Enhanced Artificial Neural Network for Protein Fold Recognition and Structural Class Prediction.* Gene Reports, 2018. **12**.
205. Xuan Nguyen, N.T., et al., *Detection of molecular changes induced by antibiotics in Escherichia coli using vibrational spectroscopy.* Spectrochim Acta A Mol Biomol Spectrosc, 2017. **183**: p. 395-401.
206. Helm, D., et al., *Classification and identification of bacteria by Fourier-transform infrared spectroscopy.* J Gen Microbiol, 1991. **137**(1): p. 69-79.
207. Sabbatini, S., et al., *Infrared spectroscopy as a new tool for studying single living cells: Is there a niche?* Biomedical Spectroscopy and Imaging, 2017. **6**: p. 85-99.
208. Sousa, C., et al., *Development of a FTIR-ATR based model for typing clinically relevant Acinetobacter baumannii clones belonging to ST98, ST103, ST208 and ST218.* Journal of Photochemistry and Photobiology B: Biology, 2014. **133**: p. 108-114.
209. Al-Holy, M., et al., *Discrimination between Bacillus and Alicyclobacillus Isolates in Apple Juice by Fourier Transform Infrared Spectroscopy and Multivariate Analysis.* Journal of Food Science, 2015. **80**.
210. Kaya-Celiker, H., et al., *Discrimination of moldy peanuts with reference to aflatoxin using FTIR-ATR system.* Food Control, 2014. **44**: p. 64–71.
211. Muhamadali, H., et al., *Chicken, beams, and Campylobacter: Rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry.* The Analyst, 2015. **141**.
212. Jaureguiberry, M., et al., *Identification of Escherichia coli and Trueperella pyogenes isolated from the uterus of dairy cows using routine bacteriological testing and Fourier transform infrared spectroscopy.* Acta Veterinaria Scandinavica, 2016. **58**(1): p. 81.
213. Grewal, M.K., P. Jaiswal, and S.N. Jha, *Detection of poultry meat specific bacteria using FTIR spectroscopy and chemometrics.* Journal of food science and technology, 2015. **52**(6): p. 3859-3869.
214. Al-Deen, R., A. Azizieh, and L. Al-Ameer, *Identification of Enterobacteriaceae foodborne bacteria in Syrian foods by PCR and FTIR-ATR techniques.* Advances in Environmental Biology, 2014. **8**: p. 1233-1237.
215. Turhan Kara, İ. and O. beyde, *Determination of lactic acid bacteria in Kaşar cheese and identification by Fourier transform infrared (FTIR) spectroscopy.* African Journal of Biotechnology, 2015. **14**: p. 2891-2902.
216. Dinkelacker, A.G., et al., *Typing and Species Identification of Clinical Klebsiella Isolates by Fourier Transform Infrared Spectroscopy and Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry.* J Clin Microbiol, 2018. **56**(11).
217. Moreirinha, C., et al., *MIR spectroscopy as alternative method for further confirmation of foodborne pathogens Salmonella spp. and Listeria monocytogenes.* J Food Sci Technol, 2018. **55**(10): p. 3971-3978.
218. Wang, Y.-D., et al., *Discrimination of foodborne pathogenic bacteria using synchrotron FTIR microspectroscopy.* Nuclear Science and Techniques, 2017. **28**(4): p. 49.
219. Campos, J., et al., *Discrimination of non-typhoid Salmonella serogroups and serotypes by Fourier Transform Infrared Spectroscopy: A comprehensive analysis.* Int J Food Microbiol,

2018. **285**: p. 34-41.
220. RG, R., et al., *Discrimination of Staphylococcus aureus Strains from Coagulase-Negative Staphylococci and Other Pathogens by Fourier Transform Infrared Spectroscopy.* Analytical Chemistry, 2020. **92**(7): p. 4943-4948.
221. Wongthong, S., et al., *Attenuated total reflection: Fourier transform infrared spectroscopy for detection of heterogeneous vancomycin-intermediate Staphylococcus aureus.* World J Microbiol Biotechnol, 2020. **36**(2): p. 22.
222. Pebotuwa, S., et al., *Influence of the Sample Preparation Method in Discriminating Candida spp. Using ATR-FTIR Spectroscopy.* Molecules, 2020. **25**(7).
223. Saif, F., et al., *FTIR spectroscopy and microscopy as methods for identification and discrimination of penicillium species of fruit*. Vol. 2244. 2020. 020002.
224. Naseer, K., et al., *Use of ATR-FTIR for detection ofSalmonella typhiinfection in human blood sera.* Infrared Physics & Technology, 2020. **110**: p. 103473.
225. Steenbeke, M., et al., *Exploring the possibilities of infrared spectroscopy for urine sediment examination and detection of pathogenic bacteria in urinary tract infections.* Clin Chem Lab Med, 2020. **58**(10): p. 1759-1767.
226. Lam, L.M.T., et al., *Multicenter Evaluation of Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) Spectroscopy-Based Method for Rapid Identification of Clinically Relevant Yeasts.* Journal of Clinical Microbiology, 2022. **60**(1): p. e01398-21.
227. Pakbin, B., et al., *FTIR differentiation based on genomic DNA for species identification of Shigella isolates from stool samples.* Sci Rep, 2022. **12**(1): p. 2780.
228. Cordovana, M., et al., *Classification of Salmonella enterica of the (Para-)Typhoid Fever Group by Fourier-Transform Infrared (FTIR) Spectroscopy.* Microorganisms, 2021. **9**(4).
229. Pascale, M.R., et al., *Use of Fourier-Transform Infrared Spectroscopy With IR Biotyper® System for Legionella pneumophila Serogroups Identification.* Front Microbiol, 2022. **13**: p. 866426.
230. Barker, K.R., et al., *Fourier Transform Infrared Spectroscopy for Typing Burkholderia cenocepacia ET12 Isolates.* Microbiol Spectr, 2021. **9**(3): p. e0183121.
231. Nitrosetein, T., et al., *Attenuated Total Reflection Fourier Transform Infrared Spectroscopy combined with chemometric modelling for the classification of clinically relevant Enterococci.* J Appl Microbiol, 2021. **130**(3): p. 982-993.
232. Alvarez-Ordóñez, A., et al., *Fourier transform infrared spectroscopy as a tool to characterize molecular composition and stress response in foodborne pathogenic bacteria.* J Microbiol Methods, 2011. **84**(3): p. 369-78.
233. Wenning, M., et al., *Identification and differentiation of food-related bacteria: A comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry.* J Microbiol Methods, 2014. **103**: p. 44-52.
234. Kravdal, G., D. Helgø, and M.K. Moe, *Infrared spectroscopy is the gold standard for kidney stone analysis.* Tidsskr Nor Laegeforen, 2015. **135**(4): p. 313-4.
235. Maitra, I., et al., *Attenuated total reflection Fourier-transform infrared spectral discrimination in human bodily fluids of oesophageal transformation to adenocarcinoma.* Analyst, 2019. **144**(24): p. 7447-7456.
236. Zlotogorski-Hurvitz, A., et al., *FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer.* J Cancer Res Clin Oncol, 2019. **145**(3): p. 685-694.
237. Ferreira, I.C.C., et al., *Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR)*

Spectroscopy Analysis of Saliva for Breast Cancer Diagnosis. J Oncol, 2020. **2020**: p. 4343590.
238. Xiang, L., et al., *Identification of colitis and cancer in colon biopsies by Fourier Transform Infrared spectroscopy and chemometrics.* ScientificWorldJournal, 2012. **2012**: p. 936149.
239. Bel'skaya, L.V., E.A. Sarf, and I.A. Gundyrev, *Study of the IR Spectra of the Saliva of Cancer Patients.* Journal of Applied Spectroscopy, 2019. **85**(6): p. 1076-1084.
240. Scott, D.A., et al., *Diabetes-related molecular signatures in infrared spectra of human saliva.* Diabetol Metab Syndr, 2010. **2**: p. 48.
241. Bottoni, U., et al., *Infrared Saliva Analysis of Psoriatic and Diabetic Patients: Similarities in Protein Components.* IEEE Transactions on Biomedical Engineering, 2016. **63**(2): p. 379-384.
242. Fujii, S., et al., *Diagnosis of Periodontal Disease from Saliva Samples Using Fourier Transform Infrared Microscopy Coupled with Partial Least Squares Discriminant Analysis.* Anal Sci, 2016. **32**(2): p. 225-31.
243. Beyer-Hans, K.M., et al., *Salivary Fingerprinting of Periodontal Disease by Infrared-ATR Spectroscopy.* Proteomics Clin Appl, 2020. **14**(3): p. e1900092.
244. Simsek Ozek, N., et al., *Differentiation of Chronic and Aggressive Periodontitis by FTIR Spectroscopy.* J Dent Res, 2016. **95**(13): p. 1472-1478.
245. Rodrigues, L.M., et al., *Analysis of saliva composition in patients with burning mouth syndrome (BMS) by FTIR spectroscopy.* Vibrational Spectroscopy, 2019. **100**: p. 195-201.
246. N., K., S. Ali, and J. Qazi, *ATR-FTIR spectroscopy as the future of diagnostics: a systematic review of the approach using bio-fluids.* Applied Spectroscopy Reviews, 2021. **56**(2): p. 85-97.
247. Ferreira, I.C.C., et al., *Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR) Spectroscopy Analysis of Saliva for Breast Cancer Diagnosis.* Journal of Oncology, 2020. **2020**: p. 4343590.
248. Lewis, P.D., et al., *Evaluation of FTIR spectroscopy as a diagnostic tool for lung cancer using sputum.* BMC Cancer, 2010. **10**: p. 640.
249. Menzies, G.E., et al., *Fourier transform infrared for noninvasive optical diagnosis of oral, oropharyngeal, and laryngeal cancer.* Transl Res, 2014. **163**(1): p. 19-26.
250. Sánchez-Brito, M., et al., *A machine-learning strategy to evaluate the use of FTIR spectra of saliva for a good control of type 2 diabetes.* Talanta, 2021. **221**: p. 121650.
251. Caixeta, D.C., et al., *Salivary molecular spectroscopy: A sustainable, rapid and non-invasive monitoring tool for diabetes mellitus during insulin treatment.* PLOS ONE, 2020. **15**(3): p. e0223461.
252. Saranya, K.K.N., et al., *Molecular signatures in infrared spectra of saliva in healthy, chronic and aggressive periodontitis.* Vibrational Spectroscopy, 2020. **111**: p. 103179.
253. Novais, A., et al., *Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing.* Eur J Clin Microbiol Infect Dis, 2019. **38**(3): p. 427-448.
254. Derruau, S., et al., *Shedding light on confounding factors likely to affect salivary infrared biosignatures.* Anal Bioanal Chem, 2019. **411**(11): p. 2283-2290.
255. Czeslawa, P., et al., *Saliva as a first-line diagnostic tool: A spectral challenge for identification of cancer biomarkers.* Journal of Molecular Liquids, 2020. **307**: p. 112961.
256. Sanchez-Brito, M., et al., *A machine-learning strategy to evaluate the use of FTIR spectra of saliva for a good control of type 2 diabetes.* Talanta, 2020. **221**: p. 121650.

Chapter 3. Evaluation of ATR-FTIR spectroscopy and Transflection-FTIR spectroscopy as a tool for rapid identification of bovine mastitis related Gram-positive cocci in different growing medium

3.1. Abstract

Bovine mastitis is the inflammation of the mammary-gland caused by pathogenic infection, inducing abnormal and decreased milk production. It causes huge economic loss that can be detrimental to farmers worldwide. *Staphylococcus* spp. and *Streptococcus* spp. are Gram-positive cocci that are among the most prevalent pathogens causing mastitis. Traditional methods for identification of bovine mastitis pathogens provide ambiguous results. In this study, we aim to develop and evaluate the use of Fourier-transform infrared (FTIR) spectroscopy for bovine mastitis pathogens identification. Four databases of bovine mastitis-related pathogens including 440 strains as training, 142 strains as validation, and 98 strains as test set have been developed using different FTIR sampling method (attenuated total reflectance (ATR) and transflectance (TR)) and growth media namely tryptic soy agar (TSA) and Columbia blood agar (CBA) by principal component analysis and linear discriminant analysis (PCA-LDA). Outlier samples were investigated using hierarchical cluster analysis (HCA) and subjected to other identification methods for identity confirmation. A correct classification rate range of 94-97% was achieved for all databases at the species level. FTIR database compatibility was assessed using combined databases grown under different growth media resulting in a high identification rate of ~93% at species level. Therefore, FTIR spectroscopy has the potential applicability for routine identification of *Staphylococcus* spp. and *Streptococcus* spp. as causative agents in bovine mastitis.

3.2. Introduction

Bovine mastitis is a major endemic disease in dairy farms. It causes a reduction in milk yield and may alter the milk quality, making it watery in appearance, showing flakes, clots, or pus. Also, bovine mastitis can lead to shedding bacteria and toxins in the milk [1]. Furthermore, it leads to a decline in nutrient content, such as potassium, lactoferrin and casein [2]. Bovine mastitis is a significant disease in dairy cattle that causes substantial economic loss that can be detrimental to farmers worldwide. Dairy farm associated costs include waste of milk due to antibiotic residues contamination, reduction in yield, veterinary costs, labor or personnel costs, and reduced longevity of the infected cows and occasional deaths [3]. In Canada, the total cost of bovine mastitis in an average herd is approximately CAD$ 662 per cow per year mostly associated with a low milk yield, followed by excessive expenses in control and medication with a total annual cost of CAD$ 400 million for Canadian dairy farmers [4].

Several bacteria can cause bovine mastitis [5]. Gram-positive cocci, including contagious pathogens such as *Staphylococcus aureus* (*S. aureus*) and *Streptococcus dysgalactiae* (*S. dysgalactiae,* and environmental pathogens like some coagulase-negative *Staphylococci* (CoNS) and *Streptococcus uberis* (*S. uberis*) are the major cause of bovine mastitis. *S. aureus* is responsible for around 5% to 70% of cow mastitis worldwide, and causes chronic mastitis infection, resulting in a 45% decrease in milk production per quarter [6]. Prevalence of CoNS is increasing in many countries. Some CoNS most found in bovine mastitis are *S. chromogenes, S. simulans, S. haemolyticus, S. xylosus, S. hyicus* and *S. epidermidis* [5, 7]. *S. uberis* is also gaining attention throughout the world as it causes both clinical and subclinical mastitis. Furthermore, major concerns have been raised on the overuse of antimicrobials and the development of antimicrobial resistance, reducing therefore the effectiveness of existing treatments of bovine mastitis. Rapid and cost-effective analysis methods are needed for pathogen identification at the species level in order to prescribe targeted treatment, optimize use of antibiotics, and develop effective control strategies. However, differentiation between bovine mastitis pathogen species is sometimes challenging [5].

In routine diagnostic laboratories, the identification of gram-positive, catalase-negative cocci is still mainly based on biochemical tests and serological grouping. However, these methods are rather time-consuming, labor-intensive and may give uncertain results due to a lack of mastitis-associated species in the database or misinterpretation resulting in ambiguous identification results.

Furthermore, Lancefield-group antisera do not react with every streptococcal species [8]. Traditional phenotypic plates used for bovine mastitis such as Accumast, Minnesota Easy System, SSGN, and SSGNC Quad plates yield low rates of identification, with Accumast performing accuracy between 73.46% and 89.57% [9]. Moreover, isolates belonging to the same species may vary in expression on culture plate, and their interpretation is mainly subjective. All that said, methods currently used in routine diagnosis for bovine mastitis are prone to error and may give uncertain result,

Identification of microorganisms at the species level can be achieved by genotypic-based methods, but the detection and simultaneous identification on-site is limited [9]. In many studies of bovine mastitis, there is no species-level identification investigated for *Staphylococcus* species (other than *S. aureus*), *Streptococcus* spp. and other Gram-positive cocci. Accurate identification at the species level is essential for epidemiological studies. The cost and the lack of accuracy of traditional methods are among the probable causes for the lack of speciation of these groups of microorganisms. In the light of the growing threat of antibiotic resistant bacteria, fast and proper identification systems are not only crucial for determination of the role of certain bacterial species in bovine mastitis, but also for choosing the right therapeutic treatment [9]. For targeted therapies, fast, easy-handling, and accurate identification methods, allowing the discrimination of bacteria at least at species level, are urgently needed.

Presently, biophysical techniques such as Fourier-Transform Infrared (FTIR) spectroscopy and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) have gained more attention due to its low cost, non-destructive and rapid and easy identification. MALDI-TOF MS is based on the molecular mass of small peptides and ribosomal proteins, whereas FTIR spectroscopy is based on the specific fingerprint like patterns of bacterial infrared of the chemical functional group signals from carbohydrates (Polysaccharides), proteins and peptides and lipids. The low cost per sample, simplicity, and time-efficiency of time of analysis of the experimental procedure make MALDI-TOF MS and FTIR spectroscopy attractive options for bacterial identification and typing [9]. Nonetheless, MALDI-TOF MS lacks robustness and the absence of a user-friendly software, which makes it difficult for routine use [10]. Furthermore, it can experience difficulties in identifying species having minor differences in their ribosomal protein sequences [11]. FTIR spectroscopy has been shown to have advantage over MALDI-TOF

MS for bacteria identification of bovine mastitis sample, in which 100% correct species identification was attained for FTIR spectroscopy, and 90.5% for MALDI-TOF MS [12]. However, MALDI-TOF MS only started being implemented in microbiology laboratories recently [13]. Due to its high throughput capacity and discriminatory power, FTIR spectroscopy might represent an interesting method for the identification of bovine mastitis pathogens. FTIR is a metabolome-based method, which is able to distinguish the broad diversity of microbiota down to the species and subspecies level [4]. In contrast to conventional methods, FTIR spectroscopy enables a high sample throughput, delivers high spectral quality data in short time and represents a rapid, inexpensive, and reliable identification method for microorganisms. It has been successfully employed for studying complex microbial communities in raw milk and dairy products [9]. Because of its high discriminatory capacity, FTIR spectroscopy can be employed for strain typing purposes [9].

We therefore aimed to establish an accurate species-specific identification system for differential diagnosis of Gram-positive cocci (*Streptococcus spp.*, *S. aureus* and common CoNS species) and other related species associated bovine mastitis using FTIR spectroscopy using two different spectral acquisition techniques; attenuated total reflectance (ATR) and transflection (TR), and evaluate the effect of different culture medium composition (Tryptic soy agar and Columbia blood agar) to on the accuracy of microbial identification by FTIR spectroscopy.

3.3. Materials and Methods

3.3.1. Bacterial isolates

A total of 680 bacterial strains representing 7 genera and 30 species were included in this study. The strains were isolated from bovine clinical mastitis, and kindly provided by 'Oplait Regroupement pour un lair de qualité optimale' (Oplait Saint-Hyacinthe, Canada). The identification of all strains was confirmed at Oplait by phenotypical and genotypical tests as previously described [14, 15] and stored in 10% glycerol at -80℃. For FTIR analysis, strains were subculured in tryptic soy broth (TSB) and cultured on tryptic soy agar (TSA; BD Difco, Le Pont de Claix, France) at 35℃ for 24 h. 582 isolates were used to develop the FTIR spectral database. The isolates were randomly assigned to a training (n = 440) or validation (n = 142) set. A total of 98 isolates were used for prediction and are summarized in Table 3.1..

Table 3.1.. Species and number of strains used for the development of the FTIR spectral database and validation set.

Species	Training	Validation	Test	Total strains
Enterobacter spp.	10	3		13
Escherichia coli	173	55		228
Klebsiella oxytoca	4	2		6
Klebsiella pneumoniae	3	2		5
Klebsiella spp.	21	10		31
Corynebacterium spp.	24	8		32
Staphylococcus aureus	44	10	22	76
CoNS				
S. arlettae	3	1		4
S. capitis	8	1	1	10
S. caprae	2	0		2
S. chromogenes	8	5	4	17
S. cohnii	9	4	2	15
S. devriesei	6	3	1	10
S. epidermidis	2	2	6	10
S. equorum	6	3	1	10
S. gallinarum	8	1	1	10
S. haemolyticus	4	1	4	9
S. hominis	5	2	1	8
S. hyicus	3	1	5	9
S. pasteuri	4	1		5
S. aprophyticus	8	4	2	15

S. sciuri	12	2	1	14
S. simulans	10	2	5	17
S succinus	8	2		10
S vitulinus	5	1		6
S. warneri	5	4	1	10
S. xylosus	3	1	6	10
Streptococcus dysgalactiae	17	4	18	39
Streptococcus uberis	15	4	17	36
Trueperella pyogenes	10	3		13
Total	**440**	**142**	**98**	**680**

3.3.2. FTIR spectroscopy spectral acquisition methods

FTIR spectroscopy generates spectra based on the absorption of the infrared light by the different chemical composition (lipids, proteins, polysaccharides) of the whole bacterial cell. As the entire spectral fingerprint is generated by FTIR spectroscopy, many details could be revealed, and even closely related species could be differentiated through proper analysis techniques. Microbial samples with different Gram-level, genera and species can be examined to find specific biomarkers associated with specific spectral markers for building the identification/discrimination models. With a large database and adequate analysis methods, FTIR spectroscopy could identify a microorganism in a matter of minutes (Figure 3.1. Schematic representation of (A) ATR-FTIR sampling technique and (B) Trans-FTIR sampling technique for the procedure of identification using FTIR spectroscopy). Due to its discriminatory power, FTIR spectroscopy is suitable to identify and discriminate closely related bacterial species from different genera [16]. The advantages are particularly its simplicity to operate, no reagents required, non-destructive, noon-invasive, rapid, and most importantly cost-effective [17]. In general, transmission FTIR and ATR-FTIR are the two sampling techniques in FTIR spectroscopy. While transmission FTIR is more common, it suffers from opacity problem and strictly requires sample to be 1 to 20 microns thick [18]. On the other hand, transflection reflectance technique has gained more attention in recent years in our laboratory. The latter spectral acquisition technique includes attenuated total reflectance (ATR) and transflection (TR) and are employed this work. In brief, ATR spectra of a sample are recorded by the attenuation of an evanescent waves that is produced when light is internally reflected with a crystal). In transflectance (transflection) mode, also called reflection-absorption, the sample is deposited onto an infrared reflective substrate (e.g. polished metal or low-E glass), IR beam passes through the sample and reflects off the reflective surface layer, and

then it passes through the sample a second time, doubling the pathlength and hence increasing the sensitivity [18]. The ATR-FTIR spectra differ significantly from those acquired in TR-FTIR mode. A general schematic representation of the two sampling techniques can be found in Figure 3.1. A main issue of ATR-FTIR will be the lack of sensitivity due to the restricted depth of penetration. It is estimated that ATR-FTIR can only detect molecules present in concentrations greater than 0.1% [19]. As for TR, the double absorption of incident light through the sample that takes place increases the intensity of the IR signal, hence resulting in higher absorbance compared to ATR. Nevertheless, while ATR-FTIR simply needs the bacteria in direct contact with the IRE surface followed by spectral measurement, TR-FTIR requires an additional step to deposit and dry the bacteria on low emissivity glass substrates initially followed by spectra acquisition. Furthermore, some microorganisms, such as yeasts, are not suitable for TR-FTIR as they tend to crack and fall-off the reflective substrate when dried.

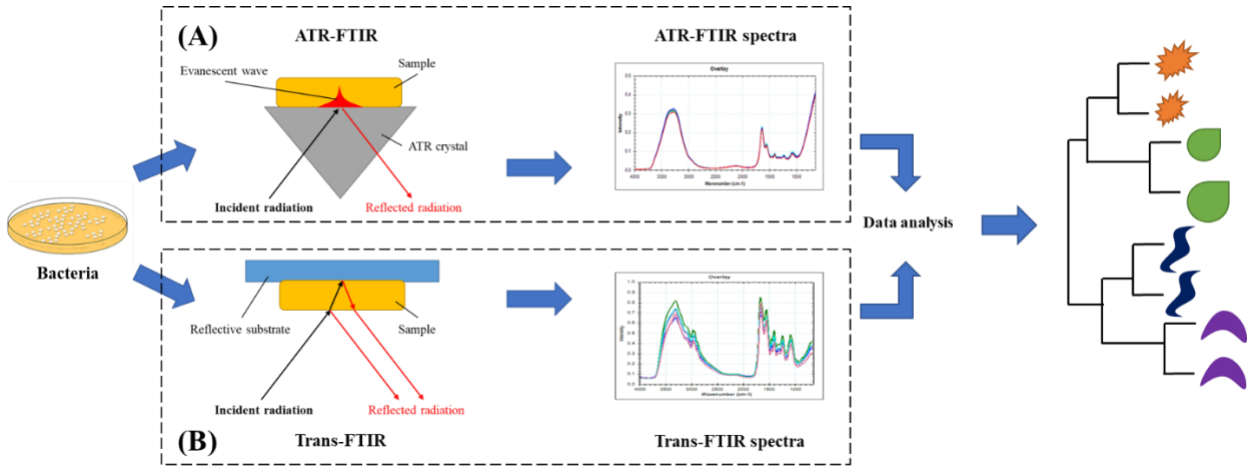


Figure 3.1. Schematic representation of (A) ATR-FTIR sampling technique and (B) Trans-FTIR sampling technique for the procedure of identification using FTIR spectroscopy.

3.3.3. Microbial growth

All isolates were subcultured onto TSA and Columbia agar with 5% sheep blood (CBA; Oxoid Australia Pty. Ltd) for 24 to 48 h at 35℃ prior to FTIR measurements.

All FTIR spectra were recorded in the region between 700 cm^{-1} and 4000 cm^{-1} in absorbance mode with Happ-Genzel apodization using a Nicolet^{TM} Summit (Thermo Fisher Scientific Waltham, US) FTIR spectrometer. For each FTIR spectrum, 32 scans were co-added and averaged at a spectra resolution of 8 cm^{-1} and ratioed against a background spectrum collected

from a clean sampling surface. To ensure sample reproducibility, triplicate spectra was acquired from individual bacteria colonies on the same agar plate and averaged.

For ATR-FTIR spectra acquisition, a loopful of bacteria was harvested with a 1-mm diameter sterile inoculating loop and deposited directly onto the ATR crystal of the Summit FTIR spectrometer to acquire an ATR-FTIR spectrum. ATR crystal was cleaned thoroughly with 70% ethanol and wiped dry. For TR-FTIR spectral acquisition, the ATR crystal was removed from the FTIR spectrometer and replaced by a holder with a window in the middle for the IR beam to pass. The bacterial samples were smeared as a thin uniform layer on a reflective glass slide (Low-e Microscope Slides, Kevley Technologies) and allowed to dry prior to TR-FTIR spectral acquisition.

Four spectral databases were obtained, namely ATR-FTIR database from TSA (ATR-TSA-FTIR), ATR-FTIR database from CBA (ATR-TSA-FTIR), TR-FTIR database from TSA (TR-TSA-FTIR), and TR-FTIR database from CBA (TR-CBA-FTIR).

For database compatibility, ATR-TSA-FTIR and ATR-CBA-FTIR databases were combined together forming a large ATR-FTIR database used to evaluate the prediction efficiency of all bacterial spectra acquired in ATR mode. The same protocol was followed for the TR-FTIR database as well as for the prediction. A total of 6 databases were obtained (ATR-TSA-FTIR, ATR-TSA-FTIR, Trans-TSA-FTIR, Trans-CBA-FTIR, ATR-TSA/CBA-FTIR, and Trans-TSA/CBA-FTIR).

3.3.4. Spectral data analysis

Prior to statistical analysis, preprocessing was applied to all spectra using commercially available spectral analysis software OMNIC (Thermo Scientific™, USA). Only high-quality spectra were considered for the analysis. Low-quality spectra are those with a weak water absorbance band (broad region of 3600-3200 cm^{-1}) of <0.15 for ATR-FTIR spectra, and absorbances of >1.2 in the spectral region between 1700 and 1600 cm^{-1} assigned to the amide I band of proteins for TR-FTIR spectra. Spectra with significant absorption from atmospheric water vapor were also excluded.

The mean of the spectra was calculated from the remaining spectra based from the replicate measurements for each growth medium and spectral acquisition mode. Spectra were subjected to vector normalization and their first derivative calculated by an in-house-built software. All spectra

were also made compatible by interpolation so that they consist of the same number of wavenumbers datapoints per spectral file. The relevant wavenumber ranges were narrowed down to 700-1800 cm^{-1} and 2800-3000 cm^{-1} in order to perform a forward region selection algorithm to discriminate among the species based on specific spectral region [20]. The combination of selected wavenumber regions producing the best separation of classes was subsequently chosen for validation purposes.

Spectra were also exported as an x-y CSV matrix using an in-house written software in MATLAB (The MathWorks Inc., Natick, MA, US) The CSV matrix was imported directly into JMP Statistical Discovery software, version 16 to perform data analysis. Unsupervised hierarchical cluster analysis (HCA) using the Ward's algorithm with Euclidean distances, principal component analysis (PCA) and linear discriminant analysis (LDA) were selected for spectral analysis. PCA-LDA is a supervised statistical analysis method that assigns each unknown spectrum to a predefined class (database IR spectra) that was created previously. The algorithm creates a class prediction to which the unknown spectrum can be assigned.

3.3.5. IR spectra libraries

A multi-level approach was used to establish the four FTIR identification database by dividing the identification process into several consecutive steps allows optimal classification at each taxonomic level, hence improving the identification accuracy. For this purpose, HCA was employed to all database strains to determine the grouping based on their similarity. Depending on the clusters, all strains were divided into groups, and they were further separated into smaller groups in the following levels: (i) level-1, strains were grouped into 2 main classes, separating Gram-positive and Gram-negative strains; (ii) level-2, *Staphylococcus* spp. was clustered out from the other Gram-positive bacteria; (iii) level-3, further differentiation of coagulase positive *Staphylococcus* (*S. aureus*) and CoNS, as well as *Corynebacterium* spp., *Streptococcus* spp. and *Trueperella pyogenes* (*T. pyogenes)*. CoNS species and *Streptococcus* spp. required an additional level to be split into single species. The structure of the FTIR spectral databases is shown in Figure 3.2. Specific steps for CoNS species identification are shown in A.1 (Appendix 1).

The misclassified isolates were re-grown and subjected to MALDI-TOF MS identification for confirmation at Health Canada laboratory. PCA was then used to derive 30 principal components ensuring 99% of the variability is considered by the analysis. Lower number of PC

scores were also used to study the impact each score on the classification accuracy of each model. Subsequently, linear discriminant analysis (LDA) was performed for the training and validation of the database.

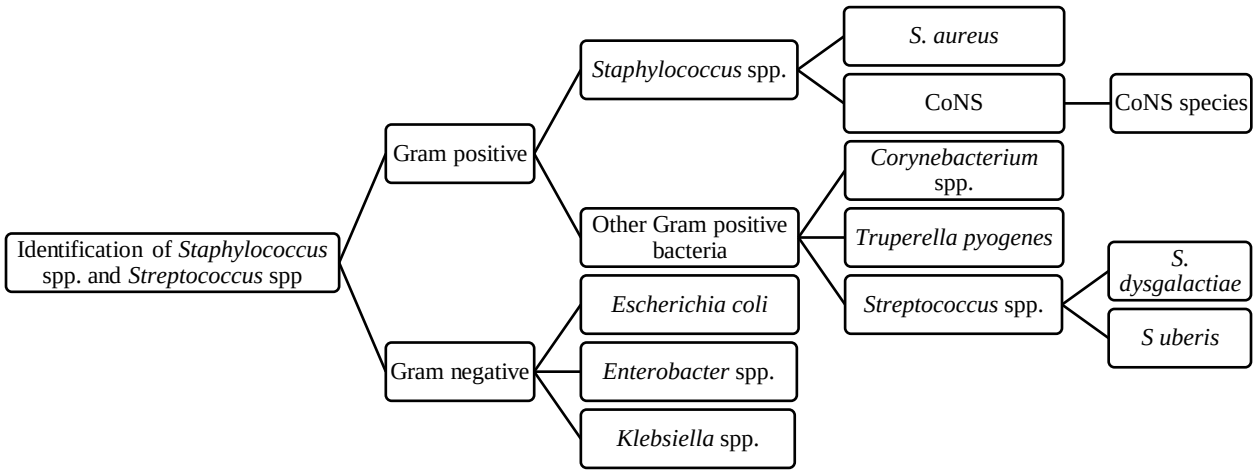


Figure 3.2. Structure of FTIR spectral databases for the identification of cow mastitis related Gram-positive cocci.

3.3.6. Identification of cow mastitis related Gram-positive cocci bacteria

Ninety-eight cow mastitis bacterial strains (not included in the spectra databases) were used as an external validation set to evaluate the identification accuracy of all 6 FTIR databases. Database compatibility and spectra interchangeability were also evaluated by using the combined ATR-TSA/CBA-FTIR database to identify the validation strains from their perspective ATR-FTIR spectra. TR-TSA/CBA-FTIR database were similarly employed to identify the external validation sets from their perspective TR-FTIR spectra. At the end, two groups of ATR-acquired prediction spectra were attained, one group from bacteria grown of CBA (ATR-CBA-FTIR), and one group from bacteria cultured on TSA agar plates (ATR-TSA-FTIR); two groups of TR-acquired prediction spectral sets were also similarly obtained, one group from bacteria cultured on CBA agar plates (TR-CBA-FTIR), and one from bacteria cultured on TSA agar plates (TR-TSA-FTIR). The 98 isolates used were 22 *S. aureus* (10104354; 10113752; 10200018; 10200025; 10200070; 10200094; 10200117; 10200186; 10200223; 10200230; 10200339; 10200346; 10200667; 10200742; 10200797; 10201145; 10201152; 10201817; 10201824; 10201855; 10201947; 10202456), 1 *S. capitis* (11304197), 4 *S. chromogenes* (10100134; 10100158; 10101834;

10103661), 2 *S. cohnii* (10605431; 11214199), 1 *S. devriesei* (11009283), 6 *S. epidermidis* (10203255; 10203613; 10205761; 10206539; 10214404; 10405260), 1 *S. equorum* (10101247), 1 *S. gallinarum* (10501412), 4 *S. haemolyticus* (10102930; 10115831; 10116807; 10116982), 1 *S. hominis* (20213459), 5 *S. hyicus* (11004585; 11005971; 11510345; 11511038; 11513759), 2 *S. saprophyticus* (20608750; 20707316), 1 *S. sciuri* (10203361), 5 *S. simulans* (10300404; 10312964; 10501252; 10509425; 10709382), 1 *S. warneri* (10204979), 6 S. *xylosus* (10104334; 10209974; 10606438; 10607084; 10608371; 10608678), 18 S. *dysgalactiae* (10201107; 11301769; 11301776; 11301783; 11303831; 11306610; 20103071; 20103637; 20104757; 20106140; 21307447;21307997; 21310041; 21313783; 21902970; 22201898; 22204820; 22205476) and 17 *S. uberis* (10112106; 10112434; 10115510; 10115534; 10202227; 10303474; 10311202; 10311219; 10311233; 10311868; 11610038; 11800255; 11800590; 11806684; 11900474; 11901327; 11910190).

Identification of each strain is achieved by pairwise identification as shown in Figure 3.2, by comparing test set spectra against each group within Gram, genus, and species level. The correct identification of *S. aureus* against other CoNS was also evaluated. Specific identification of commonly isolated CoNS species from bovine mastitis was also investigated, including *S. chromogenes, S. epidermidis, S. haemolyticus, S. hyicus, S. simulans,* and *S. xylosus*. All other CoNS species were grouped and considered as 'Other CoNS'. The prediction results along with prediction probability was generated. Prediction probability >0.8 indicates a reliable identification (high confidence), values between 0.7999 and 0.6001 represent probable correct identification (medium confidence), whereas <0.6 is regarded as non-identifiable.

3.4. Results and Discussion

3.4.1. FTIR spectra analysis

ATR-FTIR and TR-FTIR spectra were acquired in triplicate from 680 bacterial isolates (30 species from 7 genera) grown on two different culture media plates (TSA and CBA). First derivative and vector normalization was carried out on the average of the triplicate spectra from the same strain. A total of 582 strains were used for database construction for each of the 4 spectral databases. Ninety-eight isolates were used as unknowns (not included in the databases).

3.4.2. Development of the IR spectral database

The different functional groups present in the biochemical structure of bacteria strains contribute to distinct absorbance patterns in the IR spectrum. For bacteria, different content of peptidoglycan layer, lipoproteins, phospholipids, proteins, and lipopolysaccharides could be responsible for the significant differences in the IR spectral patterns. These macromolecules generally have strong absorption in the infrared spectral region between 900 and 1800 cm^{-1}, commonly assumed to be dominated by chemical groups related to lipids, protein, carboxylic side chains of proteins, free amino acids, polysaccharides, RNA/DNA and phospholipid constituents [21, 22]. The 900–700 cm^{-1} region is referred to as the 'fingerprint region' which contains weak but specific absorbance characteristic of bacteria. The spectral region between 3000–2800 cm^{-1} is associated with C-H absorption from lipids, carbohydrates, and proteins. Therefore, the spectral regions 700-1800 cm^{-1} and 2800-3000 cm^{-1} were selected for interrogation by an in-house built software to identify narrow wavenumber ranges with the selected broad regions. The narrow spectral regions can enhance the performance of the classification models by eliminating spectral regions associated with noise or other extraneous spectral interference.

Although there is a general similarity between spectra of different genera and species within, a unique pattern for each genus with different peaks can be observed, suggesting possible separation of genera at the level-2 (Figure 3.3**).** To ensure optimal identification accuracy of the spectral database and to prevent overfitting, a multitier process is employed for genus and species differentiation. At tier 1 or level-1, two main cluster groups were observed, separating Gram-positive and Gram-negative strains. This grouping was observed in all four spectral databases over a wide wavenumber region of 1200-1500 cm^{-1}, as illustrated in Figure 3.4 for ATR-TSA-FTIR strains.

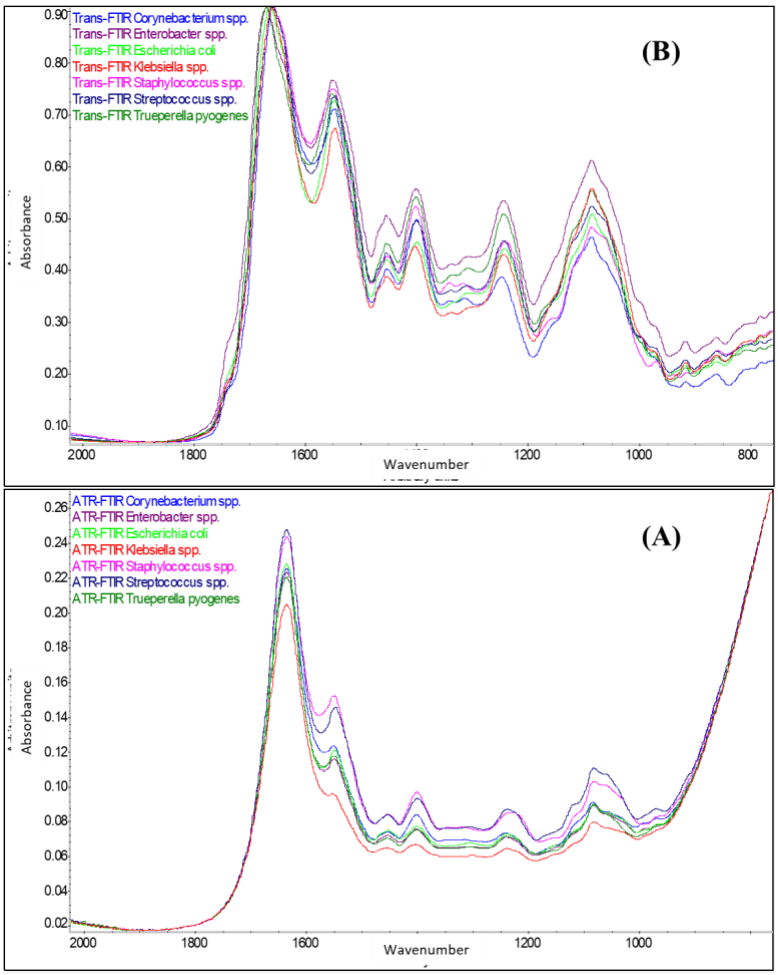


Figure 3.3. (A) ATR-TSA-FTIR spectra and (B) TR-TSA- FTIR spectra in the region of 700-2000 cm^{-1} of the averaged seven bacteria genera.

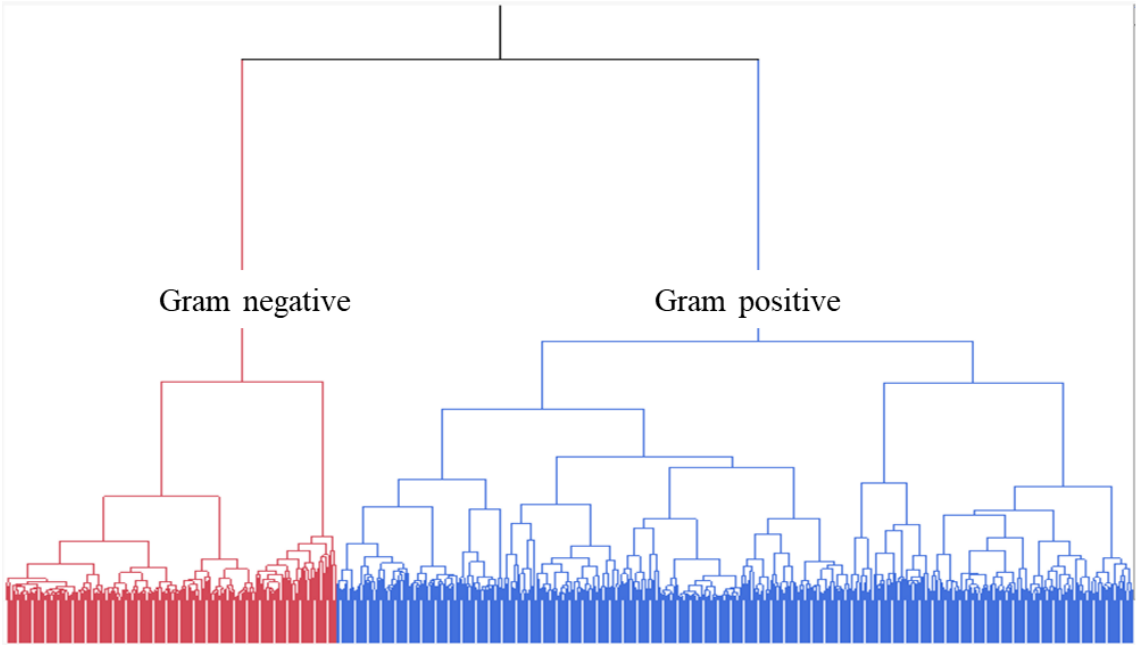


Figure 3.4. HCA of ATR-TSA-FTIR database training strains over a broad wavenumber region 1266-1518 cm^{-1} for the differentiation of Gram-positive and Gram-negative bacteria.

3.4.2.1. Genus Level Differentiation

Reducing spectra data to a group of PC scores is considered a more effective approach to class discrimination than empirical region selection. The PC scores are responsible for the variability among genera and species groups. They can be easily derived so that the analysis time can be dramatically decreased relative to the use of spectral wavenumber ranges. Figure 3.5 shows the use of the first 3 PC scores in the differentiation of multiple genera. The 3D score plot shows *Escherichia coli* (*E. coli*), *Enterobacter* spp. and *Klebsiella* spp. from Gram-negative cluster further separates from each other at the second level (Figure 3.5A). Gram-positive genera differentiation required additional steps. Staphylococcus spp. aggregated together and discriminated from the other Gram-positive genera in all four databases (Figure 3.5B), requiring two supplementary levels to be split into species. The other three Gram-positive genera differentiated at the third level (Figure 3.5C). Figure 3.6 illustrates the spectral analysis for database construction at the genus level for ATR-TSA-FTIR strains by HCA. Similar discrimination efficacy was also observed for TR-TSA-FTIR, ATR-CBA-FTIR and TR-CBA-FTIR strains.

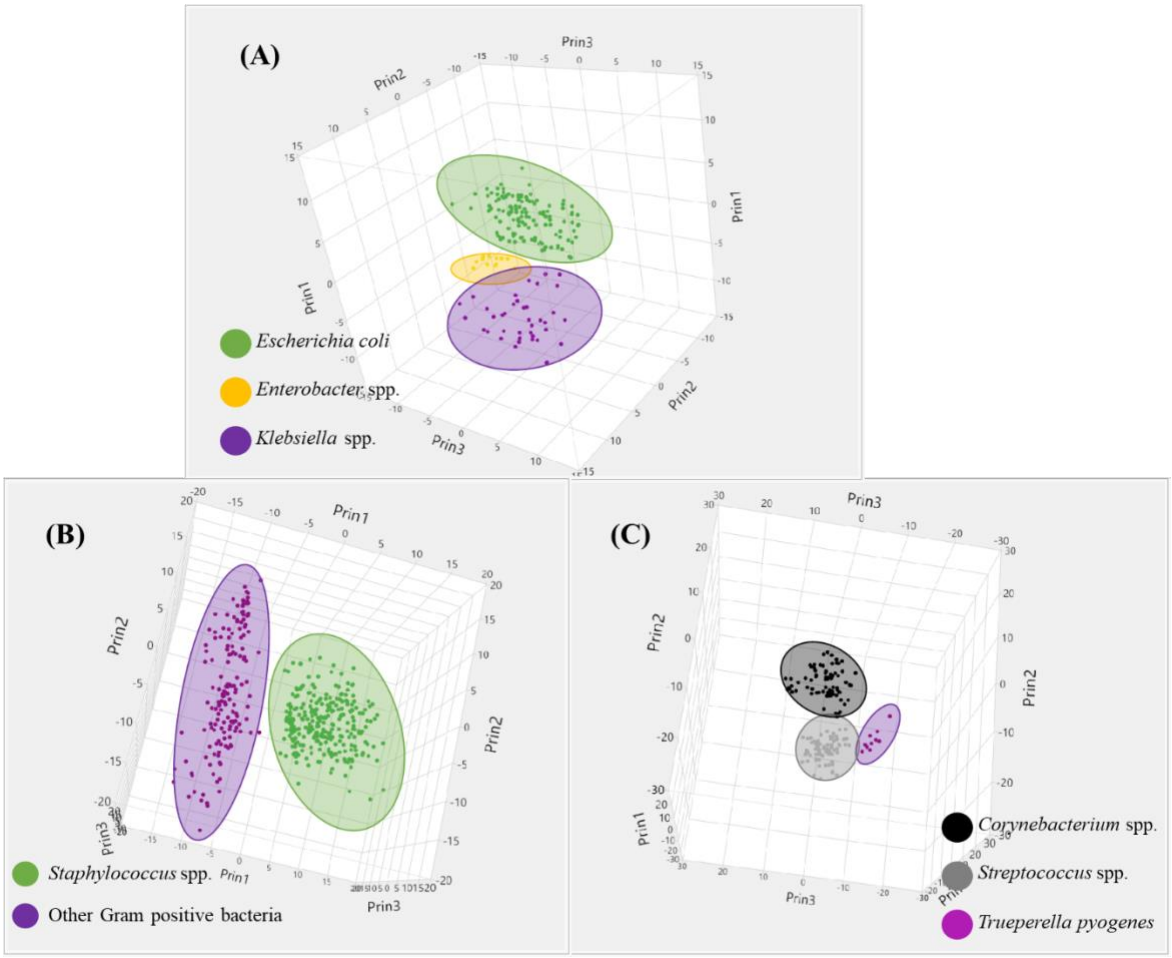


Figure 3.5. 3D score plot of PCA of (A) genera differentiation of between *Escherichia coli* (*E. coli*), *Enterobacter* spp. and *Klebsiella* spp. genera (B) *Staphylococcus* spp. against other Gram-positive bacteria and (C) *Corynebacterium* spp., *Streptococcus* spp. and *Trueperella pyogenes* genus differentiation by ATR-TSA-FTIR database strains. PC1, PC2 and PC3 totally expressed of 81.6% (PC1 47.9% PC2 27.7%, PC3 6%), 60% (PC1 26.7%, PC2 22.8%, PC3 10.5%), and 67.8% (PC1 33.8%, PC2 20.2%, PC3 13.8%), the variation for (A) Gram-negative bacteria strains, (B) *Staphylococcus* spp, against other Gram-positive bacteria, and (C) differentiation of the resting three Gram-positive genera, respectively.

3.4.2.2. *Staphylococcus* Outlier Detection

Differentiation of *S. aureus* and CoNS spectra were successfully achieved by employing the spectra data derived from using TSA and CBA growth media and using both spectral acquisition techniques (ATR-FTIR and TR-FTIR). The results obtained were due to the selection of specific regions within the broad wavenumber region 900-1500 cm^{-1}. Only 3 *S. aureus* strains (11007852, 11511212, 10602379) clustered within CoNS in HCA dendrogram of spectral database sets, similar pattern was observed in the constellation plot as shown in Figure 3.6. The three *S.*

aureus misclassified as CoNS from the external validation set were re-grown and identified by MALDI-TOF-MS as a confirmation of the results. The identification results of both methods and MALDI-TOF MS are summarized in Table 3.2. The three *S. aureus* 11007852, 11511212, 10602379 were predicted as *S. hyicus*, *S. hominis*, and *S. aureus* by the ATR-TSA-FTIR database, respectively. From the ATR-CBA-FTIR database, they were identified as *S. hyicus*, *S. chromogenes*, and *S. hyicus*, respectively. MALDI-TOF MS results were in compliance with the latter. Accordingly, the FTIR-based predictions were not a misclassification error, apart from one sample (*S. aureus* 10602379) misidentified using the ATR-TSA-FTIR database. However, it should be noted that the misidentified *S. aureus* 10602379, was identified by MALDI-TOF-MS as *S. hyicus* was found to be coagulase positive. The three non-*S. aureus* outliers (11007852, 11511212, 10602379) were excluded from the four spectral databases.

One remark is that *Staphylococcus* sp. 11511212 was predicted as *S. hominis* with 0.7309 probability in ATR-TSA-FTIR database, whereas its second hit, which was *S. chromogenes*, consistent with both the ATR-CBA-FTIR results and MALDI-TOF MS results, though with 0.21 probability. Although coagulase positive staphylococci generally entail *S. aureus* only, *Staphylococcus intermedius*, *Staphylococcus pseudintermedius*, *Staphylococcus delphini*, *Staphylococcus schleiferi* subsp. *coagulans*, and *S. hyicus* have also been described to be coagulase positive [23]. *S. hyicus* coagulase expression is strain dependent and could be coagulase positive. From the biochemical assay, the strain *Staphylococcus* sp. 10602379 showed coagulase positive activity, which lead to its prediction as *S. aureus* from the ATR-TSA-FTIR database. In another study, coagulase positive *S. hyicus* was also misidentified as *S. aureus* from a septic patient [24]. Nonetheless, the agreement of ATR-CBA-FTIR database with MALDI-TOF MS suggest the use of CBA as a growth medium may yield spectra with higher discriminatory power over TSA-grown spectra at species level. Identification accuracy deviated due to differences in growth medium was also observed by Wenning et al. for food pathogens [25] and Oust et al. for *Lactobacillus* spp [26].

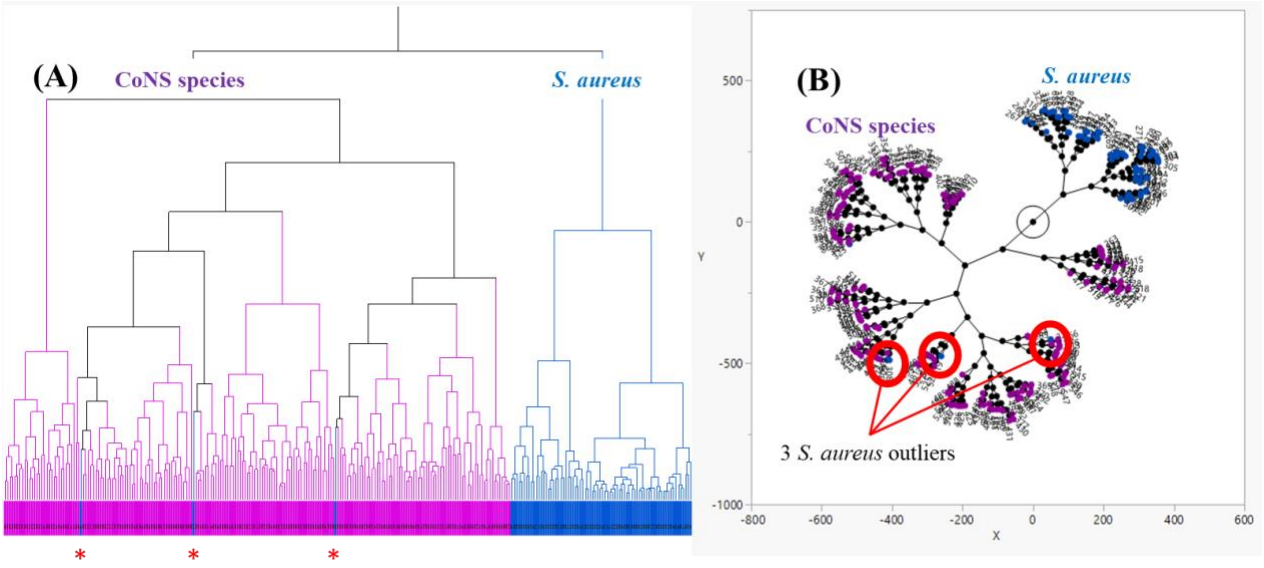


Figure 3.6. HCA of ATR-TSA-FTIR Staphylococcal spectra; (B) Constellation diagram of ATR-TSA-FTIR Staphylococcal spectra. Complete separation between *S. aureus* and CoNS species can be visualized in the HCA diagram (A) except for 3 *S. aureus* isolates (pointed out in red stars). A similar pattern is observed in the constellation diagram (B), where the 3 *S. aureus* isolates is clearly grouped with CoNS species.

Table 3.2. Identity confirmation of the 3 *S. aureus* outliers by MALDI-TOF MS and ATR-FTIR database.

Bacteria ID			**11007852**	**11511212**	**10602379**
Health Canada MALDI-TOF MS[1]	1st score	Result	*S. hyicus*	*S. chromogenes*	*S. hyicus*
		Score	1.763	1.996	2.359
	2nd score	Result	*S. chromogenes*	NA	*S. chromogenes*
		Score	1.684	NA	1.759
Health Canada biochemical assay	Oxidase		Negative	Negative	Negative
	Catalase		Positive	Positive	Positive
	Coagulase		Negative	Negative	Positive[2]
ATR-TSA-FTIR database			*S. hyicus* (0.9996)[3]	*S. hominis* (0.7309)[3]/*S. chromogenes* (0.21)[3]	*S. aureus* (1.0000)[3]
ATR-CBA-FTIR database			*S. hyicus* (1.0000)[3]	*S. chromogenes* (0.9997)[3]	*S. hyicus* (1.0000)[3]

[1]For MALDI-TOF MS, a score value ≥2.000 indicates a reliable species identification, values between 1.999 and 1.700 represent probable correct identification, whereas <1.699 is regarded as non-reliable
[2]*S. hyicus* coagulase expression is strain-dependent
[3]Prediction probability ≥0.8 indicates a reliable identification (high confidence), values between 0.7999 and 0.6000 represent probable correct identification (medium confidence), whereas <0.6 is regarded as non-identifiable

3.4.2.3. *Staphylococcus aureus* differentiation compared to Multilocus sequence typing (MLST)

It was reported that sequence types (ST) 352 and ST151 of *S. aureus* were the two main clonal populations in bovine globally, and that they are associated with the majority of cases of bovine mastitis [27]. The differentiation of ST151 and ST352 were completed for 21 *S. aureus*

strains by ATR-FTIR spectroscopy grown one TSA and CBA separately, showing promising discriminatory power of FTIR spectroscopy at strain-specific level (Figure 3.7).

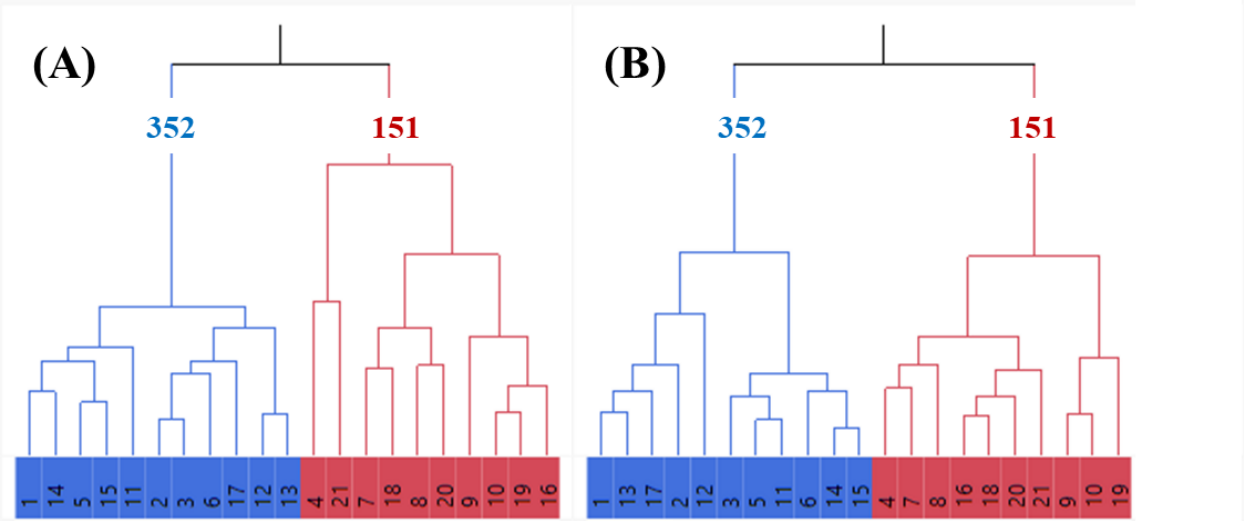


Figure 3.7. *S. aureus* sequence type differentiation (ST352 and ST151) based on spectral differences among the strains acquired by (A) ATR-TSA-FTIR spectra and (B) ATR-CBA-FTIR spectra.

3.4.2.4. *Streptococcus* species differentiation

Distinct clusters are observed for the two species of *Streptococcus*, *S. dysgalactiae* and *S. uberis* grown on either TSA or CBA media, using either spectral acquisition method (Figure 3.8). However, the total variation expressed by PC1, PC2 and PC3 is 54.8% and 70.9% for TSA spectra and CBA spectra, respectively. This may again suggest that the use CBA as the growth medium can provide more spectral information for the discrimination among different taxonomical groups.

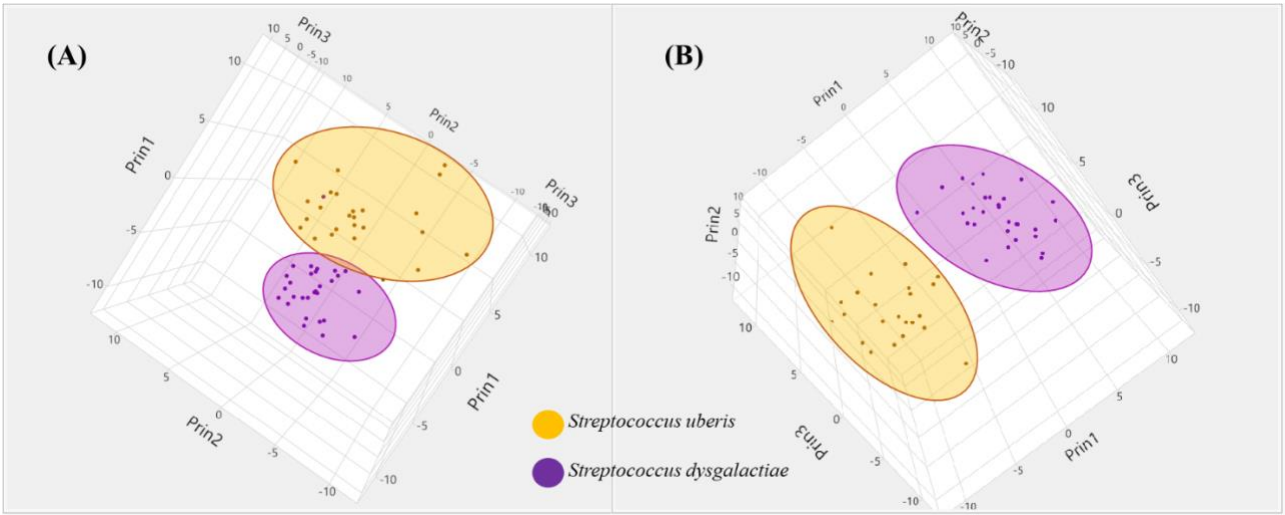


Figure 3.8. 3D score plot of PCA of differentiation between *Streptococcus dysgalactiae* and *Streptococcus uberis* in (A) ATR-TSA-FTIR database and (B) ATR-CBA-FTIR database. PC1, PC2 and PC3 totally expressed of 54.8% (PC1 24.9%, PC2 19.6%, PC3 10.3%) and 70.9% (PC1 39.2%, PC2 21.8%, PC3 9.9%) the variation for (A) ATR-TSA-FTIR database and (B) ATR-CBA-FTIR database, respectively.

3.4.3. Database accuracy dependence on variations in cultivation medium and sampling method.

The final reference spectral databases for the identification comprised 582 spectra for each of the four databases (ATR-TSA-FTIR, ATR-CBA-FTIR, TR-TSA-FTIR, and TR-CBA-FTIR) belonging to 30 species from 7 genera. Using the collected data set of FTIR spectra, a PCA-LDA database model was employed for the identification of *Staphylococcus* spp. and *Streptococcus* spp. Ninety-eight bovine mastitis strains were used as unknown to evaluate the influence of cultivation medium and sampling technique of FTIR spectroscopy on the identification accuracies.

Prediction of bovine mastitis pathogens resulted 100-99.5% accuracy at genus level, 100-98.4% for identifying *S. aureus* against CoNS (*S. aureus* vs. CoNS), and 96.4-94.8% at the species level (Table 3.3). Summarizing FTIR spectroscopy data for the 98 strains, species-level prediction using the two TSA databases obtained a an overall accuracy of 95.9% and 93.8% for TR-FTIR and ATR-FTIR, respectively. For database and test set grown on CBA, the species level identification was 96.9% and 95.9% for TR-FTIR and ATR-FTIR, respectively. Overall, TR-FTIR prediction for species all yielded better results compared ATR-FTIR, regardless of the growth medium. However, ATR-FTIR performed better at both predicting genus (100%) and *S. aureus* vs. CoNS (100%) than TR-FTIR (100-99% and 98.4%, respectively), with one *S. uberis* (10112106)

strain that was misidentified as *S. dysgalactiae* with high confidence (0.9998) at the genus level for the TR-CBA-FTIR database. For *S. aureus* vs. CoNS, ATR-FTIR correctly identified all isolates for both media, except for one isolate from CBA, which was non-identifiable (0.5278). On the other hand, each of the two TR-FTIR (TR-TSA-FTIR and TR-CBA-FTIR) databases had one misidentified isolate for *S. aureus* vs. CoNS: one *S. simulans* (10312964) with low confidence level (0.7233), and one *S. aureus* (10200117), respectively. In terms of FTIR-based species identification, a total of 30 strains from 6 common CoNS (*S. chromogenes, S. epidermidis, S. haemolyticus, S. hyicus, S. simulans, S. xylosus*) were investigated. For TSA databases ATR-TSA-FTIR and TR-TSA-FTIR, the identification accuracy was 89.6% (26/29) and 86.7% (26/30), respectively. Using the ATR-CBA-FTIR and TR-CBA-FTIR databases, correct identification was 27/30 (90%) and 27/29 (93.1%), respectively. For all species-level misidentification of those 6 common CoNS, they were misidentified. All other CoNS in the external test set were correctly identified by the 4 databases with high confidence level. For *Streptococcus* species identification, TR-FTIR databases performed well with 100% correctness, whereas ATR-TSA-FTIR database was useful for the prediction of 91.4% (32/35) prediction for all strains, and ATR-CBA-FTIR database correctly identified 94.1% (32/34).

In general, the identification rate from the 4 databases were comparable with correct prediction over 94% at the species level. TR-FTIR database had slightly higher identification accuracy with a single non-identifiable strain and 3 lower confidence identification. While with ATR-FTIR database resulted 3 non-identifiable and 2 low confidence identification. The slight discrepancy is due to the use of two different sampling methods of FTIR spectroscopy. In ATR-FTIR, the spectra are obtained from the attenuated evanescent wave due to the absorbed energy by the sample. The penetration depth of the evanescent wave beyond the crystal surface and into the sample is typically of the order of a few microns (0.5-5 μm), and thus, the thickness of the sample does not affect the spectra [28]. Contrarily, in TR-FTIR, the incident light is transmitted through the sample and reflected back by the reflective substrate, consequently increasing the intensity of the reflection signal [29]. But at the same time, due to the double transmission feature of TR, the thickness of the sample will highly influence the spectra. As a result, TR spectra have generally higher absorbance and contain more information than ATR spectra. In our study, TR had more better identification results than ATR-FTIR. As TR-FTIR spectra encompass more spectral information, its wider spectral range has proven useful for bacterial strain typing [30] and the use

of spectral library searches using the whole spectral region [31]. Furthermore, the threshold established enables confidence level comparable to MALDI-TOF MS. Transmission-FTIR and ATR-FTIR techniques are also widely employed for microorganism identification, with reported accuracies ranging from 84.4% to 100%, 69.5% to 75.3% and 98.4% to 100% for bacteria [32-34], molds [35], and yeasts [36] respectively.

However, most of the reported studies focus on strain typing and identification of *S. aureus* [33, 37, 38] and *Streptococcus pneumoniae* [39]. The identification efficiency of CoNS species, *S. dysgalactiae* and *S. uberis* is rarely evaluated. In our study, species level identification was successfully achieved based on PCA-LDA calculation using either ATR-FTIR or TR-FTIR spectroscopy.

As described above, spectral acquisition techniques and growth media can influence the infrared spectral profiles and spectral intensities and therefore can impact the prediction results of test sets. For strains grown on TSA and CBA, TR-FTIR database yielded higher correct classification results at the species level than ATR-FTIR but performed with lower rates compared to ATR at genus level and for the discrimination of *S. aureus* vs. CoNS. The fact that spectra contain more information in TR mode may make it more robust for species or subspecies level identification. Nevertheless, ATR mode is also very effective providing near comparable results for species level identification as the TR mode. Despite the automation potential and the possibility of keeping the used e-glass slide of TR-FTIR, ATR-FTIR is slightly faster as it does not require the smearing and drying step. The influence of the growth medium chosen (TSA and CBA) were evaluated, and both were robust enough by achieving over 94% accuracy, with CBA over 96%. Overall, strain identification was slightly higher using the CBA as a growth medium than TSA for both ATR-FTIR (95.9% vs. 93.8%) or TR-FTIR (96.9% vs. 95.9%). Furthermore, most of the misidentification from ATR-CBA-FTIR and TR-CBA-FTIR were either non-identifiable or had low confidence level. Wenning et al. [25] found that bacteria grown on TSA had considerably higher identification accuracy than CBA. Other studies have suggested that the variation in growth medium has minimal influence on the identification and discrimination of both genus, species, and strains for FTIR spectroscopy [40-43]. Larger variation in cultivation conditions may only impact the separation of strains, and quantitative data, while the qualitative properties remain unchanged [41, 43]. Unlike other comparable identification techniques, such as MALDI-TOF MS which

produce spectra focusing mainly on ribosomal proteins, FTIR spectroscopy records the whole-organism spectra. In theory, FTIR spectra of an organism include signatures of all main building blocks of cells (e.g., proteins, lipids, polysaccharides). Even though qualitative information is identical for the same bacterial strain, depending on the incubation time, growth temperature and nutrient supplied by the cultivation medium, biochemical composition could vary depending on gene expression in closely related isolates. This IR fingerprinting feature may make FTIR spectroscopy less specific due to intraspecies biodiversity. Hence, the variability in identification accuracy is produced when the same strains are grown on different medium. Despite of that, identification can be improved using appropriate standardized sampling method, standardized cultivation medium as well as enlarging the database thereby increasing the coverage of each species biodiversity and filling the gaps in the database. In general, *S. aureus*, CoNS species and *Streptococcus* spp. isolates achieved high identification accuracy in the 4 databases, with TR-CBA-FTIR being the most outstanding. Both spectral acquisition techniques have shown their identification efficiency for bovine mastitis pathogen identification using either TSA or CBA as growth media. This result illustrates that FTIR spectroscopy has a strong potential for identifying Gram-positive cocci, which is demonstrated by the results shown in Table 3.3 for all four databases, despite they were collected using two different sampling methods and two different cultivation media.

Table 3.3. Identification accuracy of 98 isolates from four different FTIR spectral databases, constructed from two different growing medium (TSA and CBA) and two different spectral acquisition method (ATR-FTIR and TR-FTIR).

Species	No of strains	Identification accuracy (%)			
		ATR-TSA-FTIR Database	ATR-CBA-FTIR Database	TR-TSA-FTIR Database	TR-CBA-FTIR Database
		TSA test set	CBA test set	TSA test set	CBA test set
Genus					
Staphylococcus spp.	63	63/63	63/63	63/63	63/63
Streptococcus spp.	35	35/35	35/35	35/35	34/35
Total	98	98/98	98/98	98/98	97/98
Identification accuracy		100%	100%	100%	98.98%
Overall identification accuracy		100%		99.49%	
Coagulase *Staphylococcus*					
Staphylococcus aureus	22	22/22	22/22	22/22	21/22
CoNS	41	41/41	40/40[3]	40/41[6]	41/41
Total	63	63/63	62/62	62/63	62/63
Identification accuracy		100%	100%	98.41%	98.41%
Overall identification accuracy		100%		98.41%	
Species					
S. aureus	22	22/22	22/22	22/22	21/22
S. chromogenes	4	4/4	4/4	2/4	4/4
S. epidermidis	6	5/5[1]	6/6	6/6	5/6[7]
S. haemolyticus	4	4/4	2/4[4]	4/4	4/4
S. hyicus	5	4/5[2]	4/5	4/5	4/4[8]
S. simulans	5	5/5	5/5	5/5	4/5[9]
S. xylosus	6	4/6	6/6	5/6	6/6
Other CoNS	11	11/11	11/11	11/11	11/11
Streptococcus dysgalactiae	18	16/18	16/17[5]	18/18	18/18
Streptococcus uberis	17	16/17	17/17	17/17	17/17
Total	98	91/97	93/97	94/98	94/97
Identification accuracy		93.81%	95.88%	95.92%	96.91%
Overall identification accuracy		94.85%		96.41%	

[1]One *S. epidermidis* (10203613) was non-identifiable (0.5885).
[2]One *S. hyicus* (11513759) was misidentified with low confidence (0.7079).
[3]One *S. simulans* (10300404) was non-identifiable (0.5278).
[4]One *S. haemolyticus* (10116807) was misidentified with low confidence (0.6849).
[5]One *Streptococcus dysgalactiae* (21307997) was non-identifiable (0.5042), and another (11301783) was misidentified with low confidence (0.6991).
[6]One *S. simulans* (10312964) was misidentified with low confidence (0.7233).
[7]One *S. epidermidis* (10203255) was misidentified with low confidence (0.6744).
[8]One *S. hyicus* (11004585) was non-identifiable (0.5820).
[9]One *S. simulans* (10509425) was misidentified with low confidence (0.7183).

3.4.4. Evaluation of database compatibility

Following successful identification of the 98 isolates by all four spectral databases, the database interchangeability was evaluated. To achieve this, the 582 strains from training and validation grown from both TSA and CBA were combined together according to their sampling method (ATR or TR), forming 2 large databases each with 1164 strains (ATR-TSA/CBA-FTIR and TR-TSA/CBA-FTIR). The four test sets were then used for prediction according to their spectral acquisition method. Overall, correct identification was achieved at 100% at genus level, 100-95.2% for *S. aureus* vs. CoNS, and 99-92.8% at species level for all four test sets (Table 3.4). ATR-TSA/CBA-FTIR database was in general slightly less performant when it was used separately according to the growth medium for species (93.9% vs. 94.8%, respectively) and for *S. aureus* vs. CoNS identification (97.2% vs. 100%, respectively). On the other hand, compared to TR-TSA-FTIR and TR-CBA-FTIR, the TR-TSA/CBA-FTIR database had significantly improved performance for genus (100% vs. 99.5%), *S. aureus* vs. CoNS (100% vs. 98.1%), and at the species level (96.9% vs. 97.4%). All strains were correctly identified at genus level by both databases, same was for *S. aureus* vs. CoNS, except for 3 *S. simulans* (10312964, 10501252, 10300404) from the ATR-CBA test set that were misclassified as *S. aureus*. For the identification of the 6 common CoNS, the TR-CBA and TR-TSA test set achieved 100% (30/30) and 93.3% (28/30) of correct classification, whereas the ATR-CBA and ATR-TSA achieved 83.3% (25/30) and 86.2% (25/29) of correct classification using the two combine databases, respectively. Misidentified *S. chromogenes*, *S. haemolyticus*, *S. hyicus, S. simulans*, and *S. xylosus* strains were predicted as 'Other CoNS' instead of their correct species, except for one *S. chromogenes* (10103661) from ATR-CBA test set that was predicted as *S. hyicus*. All other CoNS were correctly identified with high confidence level. For *Streptococcus* species identification, ATR-CBA test set achieved complete identification (35/35), followed by TR-CBA test set with 97.1% (34/35) correct identification, TR-TSA test set with 94.3% (33/35) correct identification, and ATR-TSA test set with 91.3% (32/35) correct identification. Nonetheless, the identification results for *Streptococcus* spp. were comparable by both FTIR spectral acquisition methods. Lower rate of non-identifiable and low confidence results was observed when predicting an isolate from the test set using the combined infrared spectral databases.

Although the effect of growth media on the identification of strains by FTIR spectroscopy was reported, evaluation of the compatibility of ATR-FTIR and TR-FTIR database between

different growth media for bacteria is scarce. Based on our results, FTIR prediction results from the two different spectral acquisition methods showed to be robust enough for identification by creating a database built by two different cultivation media at the genus and species level for Gram-positive cocci. Using a growth medium-constructed database different than the one that unknowns was grown may affected the accuracy of the identification. Impact of growth medium on strain discrimination was evaluated previously by many researchers [40-43]. In a nutshell, bacterial IR spectra are indeed influenced by the cultivation conditions. The spectral difference between CBA and TSA spectra of the same strains is illustrated in A.2. It has been confirmed previously by many researchers that different metabolites are produced when the same bacteria isolate is grown on different medium [44-46]. Ingredients in TSA and CBA media are very different (A.3), and since FTIR spectroscopy records all information present in the whole organism, variation in the composition of the microbial cells are detected. Apparent positional shifts of peaks are observed between the spectra, which may be due to fluctuation in relative contribution of biomolecules. The variation in peaks in turn affects the discriminating peaks that were selected during spectral analysis for the database construction. For an improvement of identification, including more species and strains from bovine sources would be optimal to contribute to the biodiversity and associated spectra variability in the database. Nevertheless, identification of 4 test sets using the two combined databases yielded a high rate of identification. This result indicates that CBA and TSA spectra, combined together, had sufficient spectral information, and that the feature selection wavenumber was broad enough to enable species level discrimination for bacteria spectra grown either of the medium used for building the database.

It is very important to evaluate the identification accuracy of an unknown microorganism from a database with different cultivation conditions for FTIR spectroscopy. By showing the compatibility of FTIR database created using different media, one large database comprising microbial pathogens grown on popular media could be employed. This study demonstrated that with a large-enough database comprising bacterial strains grown on common cultivation media, identification of *Staphylococcus* spp. and *Streptococcus* spp. will be achieved with either choice of agar medium. The bacteria spectra were indeed affected for ATR-FTIR, but at a very minor level that would not impact the identification accuracy. In fact, identification by TR-FTIR had even improved by when combining the databases built using spectra of bacteria grown on the two media. This study demonstrates that both ATR-FTIR and TR-FTIR spectroscopy coupled with

PCA-LDA were able to successful identify the external set of microorganisms by extracting species specific spectral signatures out of the total amount of information and processing them effectively. Identifying the identity of an unknown microorganism against a combined database built based on bacteria grown on different medium did not affect the identification results. FTIR spectroscopy of for both spectral acquisition techniques demonstrated good applicability for routine microbial diagnostics for bovine mastitis.

Table 3.4. Evaluation of combined TSA- and CBA-FTIR database for the identification accuracy of prediction sets.

Species	No of strains	Identification accuracy (%)			
		ATR-TSA/CBA-FTIR Database		TR-TSA/CBA-FTIR Database	
		TSA test set	CBA test set	TSA test set	CBA test set
Genus					
Staphylococcus spp.	63	63/63	63/63	63/63	63/63
Streptococcus spp.	35	35/35	35/35	35/35	35/35
Total	98	98/98	98/98	98/98	98/98
Identification accuracy		100%	100%	100%	100%
Overall identification accuracy		100%		100%	
Coagulase *Staphylococcus*					
Staphylococcus aureus	22	22/22	22/22	22/22	22/22
CoNS	41	41/41	38/41[3]	41/41	41/41
Total	63	63/63	60/63	63/63	63/63
Identification accuracy		100%	95.24%	100%	100%
Overall identification accuracy		97.62%		100%	
Species					
S. aureus	22	22/22	22/22	22/22	22/22
S. chromogenes	4	2/3[1]	3/4	3/4	4/4
S. epidermidis	6	6/6	6/6	6/6	6/6
S. haemolyticus	4	4/4	2/4	4/4	4/4
S. hyicus	5	5/5	4/5	5/5	5/5
S. simulans	5	4/5	5/5	5/5	5/5
S. xylosus	6	4/6[2]	5/6[4]	5/6	6/6
Other CoNS	11	11/11	11/11	11/11	11/11
Streptococcus dysgalactiae	18	15/18	18/18	17/18	17/18
Streptococcus uberis	17	17/17	17/17	16/17	17/17
Total	98	90/97	93/98	94/98	97/98
Identification accuracy		92.78%	94.90%	95.92%	98.98%
Overall identification accuracy		93.85%		97.45%	

[1]One *S. chromogenes* (10100158) was non-identifiable (0.5298).
[2]One *S. xylosus* (10607084) was misidentified with low confidence (0.7941).
[3]One *S. simulans* (10300404) was misidentified with low confidence (0.6984).
[4]One *S. xylosus* (10607084) was misidentified with low confidence (0.6837).

3.5. Conclusion

Current methods for mastitis-associated microbial pathogen identification is tedious, time consuming, expensive and reported to have low accuracy compared to genotypic-based methods. FTIR spectrum obtained from a single microbial colony corresponds to all the biomolecules present in the biomass at the time of spectral acquisition. The differences in the biochemical composition of each isolate at different taxonomic level yields a unique spectral profile for each isolate. This in principle makes it possible to differentiate among genera, species, and strain types due to the differences in composition and quantity of each biomolecule in the biomass. Although extensive studies have showed FTIR spectroscopy's potential for the bacteria identification, few studies were aimed at the identification of *Streptococcus* spp. and *Staphylococcus* spp. from bovine mastitis, and especially the identification of CoNS species. In this study, the spectral database were constructed by growing 680 isolates of bacteria associated with cow mastitis. The bacteria were grown on two different growth media and the spectra acquired using two different spectral acquisition methods. Discrimination and identification of *Streptococcus* spp. and *Staphylococcus* spp. to the species level was evaluated as a function of each variable. At genus level and coagulase type of *Staphylococcus*, ATR-FTIR outperformed TR-FTIR in general (100% vs. 98.41%). However, TR-FTIR spectra of bacteria grown on CBA provided a higher identification accuracy compared to ATR-FTIR of the same bacteria grown on TSA at the species level. Overall, either TSA and CBA using either ATR or TR mode for spectral acquisition yielded high identification accuracy (≥93.9%). For database compatibility, prediction of either of the 4 test sets by the combined databases yielded high rate of classification (≥92.8%). Cultivation conditions should be standardized when using FTIR spectroscopy for bacterial identification. Yet, flexible protocol is also possible by using a growing medium other than the database built if other growth parameters are standardized (e.g., growth time and incubation temperature). Combining all findings together, FTIR spectroscopy has shown its potential for routine identification of bovine mastitis Gram-positive cocci identification.

3.6. References

1. Dworecka-Kaszak, B., et al., *High prevalence of Candida yeast in milk samples from cows suffering from mastitis in poland.* ScientificWorldJournal, 2012. **2012**: p. 196347.
2. GM, J. *Understanding the Basics of Mastitis*. Virginia Cooperative Extension 2009 [cited 2022 March 10]; Available from: http://pubs.ext.vt.edu/404/404-233/404-233.html.
3. Abebe, R., et al., *Bovine mastitis: prevalence, risk factors and isolation of Staphylococcus aureus in dairy herds at Hawassa milk shed, South Ethiopia.* BMC Vet Res, 2016. **12**(1): p. 270.
4. Aghamohammadi, M., et al., *Herd-Level Mastitis-Associated Costs on Canadian Dairy Farms.* Front Vet Sci, 2018. **5**: p. 100.
5. Wilson, D.J., R.N. Gonzalez, and H.H. Das, *Bovine mastitis pathogens in New York and Pennsylvania: prevalence and effects on somatic cell count and milk production.* J Dairy Sci, 1997. **80**(10): p. 2592-8.
6. Zecconi, A. and F. Scali, *Staphylococcus aureus virulence factors in evasion from innate immune defenses in human and animal diseases.* Immunol Lett, 2013. **150**(1-2): p. 12-22.
7. Condas, L.A.Z., et al., *Prevalence of non-aureus staphylococci species causing intramammary infections in Canadian dairy herds.* J Dairy Sci, 2017. **100**(7): p. 5592-5612.
8. Tamime, A.Y. and R.K. Robinson, *6 - Microbiology of yoghurt and related starter cultures*, in *Tamime and Robinson's Yoghurt (Third Edition)*, A.Y. Tamime and R.K. Robinson, Editors. 2007, Woodhead Publishing. p. 468-534.
9. *<19-1570.pdf>*.
10. Sauget, M., et al., *Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria?* Trends Microbiol, 2017. **25**(6): p. 447-455.
11. Salman, A., et al., *Detection of antibiotic resistant Escherichia Coli bacteria using infrared microscopy and advanced multivariate analysis.* The Analyst, 2017. **142**: p. 2136-2144.
12. Schabauer, L., et al., *Novel physico-chemical diagnostic tools for high throughput identification of bovine mastitis associated gram-positive, catalase-negative cocci.* BMC Vet Res, 2014. **10**: p. 156.
13. Novais, A., et al., *Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing.* Eur J Clin Microbiol Infect Dis, 2019. **38**(3): p. 427-448.
14. Cheng, H.R. and N. Jiang, *Extremely rapid extraction of DNA from bacteria and yeasts.* Biotechnol Lett, 2006. **28**(1): p. 55-9.
15. Pradhan, P. and J.P. Tamang, *Phenotypic and Genotypic Identification of Bacteria Isolated From Traditionally Prepared Dry Starters of the Eastern Himalayas.* Front Microbiol, 2019. **10**: p. 2526.
16. Quintelas, C., et al., *An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing.* Biotechnol J, 2018. **13**(1).
17. Wang, Y., et al., *Differentiation in MALDI-TOF MS and FTIR spectra between two closely related species Acidovorax oryzae and Acidovorax citrulli.* BMC Microbiol, 2012. **12**: p. 182.
18. Beasley, M.M., et al., *Comparison of transmission FTIR, ATR, and DRIFT spectra: implications for assessment of bone bioapatite diagenesis.* Journal of Archaeological Science, 2014. **46**: p. 16-22.
19. Smith, B.C., *Fundamentals of Fourier Transform Infrared Spectroscopy*. 2nd Edition ed. 2011. 207 Pages.

20. *Microbiology by numbers.* Nature Reviews Microbiology, 2011. **9**(9): p. 628-628.
21. Naumann, D., D. Helm, and H. Labischinski, *Microbiological characterizations by FT-IR spectroscopy.* Nature, 1991. **351**(6321): p. 81-82.
22. Naumann, D., *FT-INFRARED AND FT-RAMAN SPECTROSCOPY IN BIOMEDICAL RESEARCH.* Applied Spectroscopy Reviews, 2001. **36**(2-3): p. 239-298.
23. Kosecka-Strojek, M., A. Buda, and J. Międzobrodzki, *Chapter 2 - Staphylococcal Ecology and Epidemiology*, in *Pet-To-Man Travelling Staphylococci*, V. Savini, Editor. 2018, Academic Press. p. 11-24.
24. Casanova, C., et al., *Staphylococcus hyicus bacteremia in a farmer.* J Clin Microbiol, 2011. **49**(12): p. 4377-8.
25. Wenning, M., et al., *Identification and differentiation of food-related bacteria: A comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry.* J Microbiol Methods, 2014. **103**: p. 44-52.
26. Oust, A., et al., *Evaluation of the robustness of FT-IR spectra of lactobacilli towards changes in the bacterial growth conditions.* FEMS Microbiol Lett, 2004. **239**(1): p. 111-6.
27. Thomas, A., et al., *Prevalence and distribution of multilocus sequence types of Staphylococcus aureus isolated from bulk tank milk and cows with mastitis in Pennsylvania.* PLoS One, 2021. **16**(3): p. e0248528.
28. Smith, B.C., *Fundamentals of Fourier transform infrared spectroscopy*. 2nd ed. 2011, Boca Raton, FL: CRC Press. xiii, 193 p.
29. Korte, E.H., et al., *Infrared ellipsometric view on monolayers: towards resolving structural details.* Anal Bioanal Chem, 2002. **374**(4): p. 665-71.
30. Tsutsumi, T., et al., *243. Transflection Fourier Transform Infrared Spectroscopy as a Real-Time Strain Typing Technique: A Vancomycin-Resistant Enterococcus faecium (VRE) Typing Prospective Study.* Open Forum Infectious Diseases, 2019. **6**(Suppl 2): p. S138-S138.
31. Estupinan Mendez, D. and T. Allscher, *Advantages of External Reflection and Transflection over ATR in the Rapid Material Characterization of Negatives and Films via FTIR Spectroscopy.* Polymers (Basel), 2022. **14**(4).
32. Costa, F.S.L., et al., *Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii.* Analytical Methods, 2016. **8**(39): p. 7107-7115.
33. XIE, Y., et al., *RAPID IDENTIFICATION AND CLASSIFICATION OF STAPHYLOCOCCUS AUREUS BY ATTENUATED TOTAL REFLECTANCE FOURIER TRANSFORM INFRARED SPECTROSCOPY.* Journal of Food Safety, 2012. **32**(2): p. 176-183.
34. Grewal, M.K., P. Jaiswal, and S.N. Jha, *Detection of poultry meat specific bacteria using FTIR spectroscopy and chemometrics.* Journal of food science and technology, 2015. **52**(6): p. 3859-3869.
35. Salman, A., et al., *Detection of Fusarium oxysporum Fungal Isolates Using ATR Spectroscopy.* Spectroscopy: An International Journal, 2012. **27**: p. 109708.
36. Lam, L.M.T., et al., *Multicenter Evaluation of Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) Spectroscopy-Based Method for Rapid Identification of Clinically Relevant Yeasts.* J Clin Microbiol, 2022. **60**(1): p. e0139821.
37. Johler, S., et al., *High-resolution subtyping of Staphylococcus aureus strains by means of Fourier-transform infrared spectroscopy.* Syst Appl Microbiol, 2016. **39**(3): p. 189-194.

38. Lamprell, H., et al., *Discrimination of Staphylococcus aureus strains from different species of Staphylococcus using Fourier transform infrared (FTIR) spectroscopy.* Int J Food Microbiol, 2006. **108**(1): p. 125-9.
39. Vaz, M., et al., *Serotype discrimination of encapsulated Streptococcus pneumoniae strains by Fourier-transform infrared spectroscopy and chemometrics.* J Microbiol Methods, 2013. **93**(2): p. 102-7.
40. Lefier, D., et al., *Effect of sampling procedure and strain variation in Listeria monocytogenes on the discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variates analysis.* FEMS Microbiol Lett, 1997. **147**(1): p. 45-50.
41. Samuels, A., et al., *Infrared Spectra of Bacillus subtilis Spores: The Effect of Growth Media.* 2003: p. 10.
42. Baldauf, N.A., et al., *Effect of selective growth media on the differentiation of Salmonella enterica serovars by Fourier-Transform Mid-Infrared Spectroscopy.* J Microbiol Methods, 2007. **68**(1): p. 106-14.
43. Oust, A., et al., *Evaluation of the robustness of FT-IR spectra of lactobacilli towards changes in the bacterial growth conditions.* FEMS Microbiology Letters, 2004. **239**(1): p. 111-116.
44. Ratiu, I.A., et al., *The effect of growth medium on an Escherichia coli pathway mirrored into GC/MS profiles.* J Breath Res, 2017. **11**(3): p. 036012.
45. Fouchard, S., et al., *Influence of growth conditions on Pseudomonas fluorescens strains: A link between metabolite production and the PLFA profile.* FEMS Microbiology Letters, 2005. **251**(2): p. 211-218.
46. Šajbidor, J., *Effect of Some Environmental Factors on the Content and Composition of Microbial Membrane Lipids.* Critical Reviews in Biotechnology, 1997. **17**(2): p. 87-103.

Connecting Statement

In the previous chapter, the influence of the growth medium and the mode of spectral acquisition on the capability of FTIR spectroscopy to discriminate among Gram-positive cocci associated with bovine mastitis was found to be minor. In addition, interchanging database for prediction of the identity of an unknown microorganism by FTIR spectroscopy showed promising results. In the next chapter, two commercially available FTIR instruments manufactured by different companies were evaluated for the identification of common food pathogens.

Chapter 4. Assessing the compatibility of two portable FTIR instruments for FTIR-based spectroscopic identification of foodborne pathogens

4.1. Abstract

Foodborne illnesses are caused by microbial hazards, posing major threat to public health and the economy. Rapid pathogen identification is important for food safety reasons by avoiding spread of foodborne illness and outbreak proliferation. Routine food microbiology laboratories rely on conventional methods that are costly and lengthy. Fourier-transform infrared (FTIR) spectroscopy is a biophysical method that is gaining attention due to its speed, accuracy, and cost for microbial identification. The current study analyzed the FTIR spectra of common food pathogen and suggested potential biomarkers for genus and species identification. Furthermore, 2 commercially available FTIR instruments (from different instrument manufacturers) have been evaluated for spectral database interchangeability through a cross validation study. Spectral databases developed using the Nicolet^{TM} Summit FTIR spectrometer (Summit), Summit-FTIR (n=138) and Cary 630 (n=305) FTIR spectrometer (Cary-630). For each spectral database an external test set provided 99.5% and 96.2% at genus level, and 95.8% and 79.9% at species level, correct identification respectively. The use of different FTIR instrumentation did not appreciably influence the identification accuracy at the genus level. However, loss of sensitivity was significant at the species discrimination level when spectral database from one of the instruments models was employed to identify the test microorganisms from spectra acquired from the other spectrometer. This study shows that infrared databases of foodborne microbial pathogens must be constructed using a single instrument model if correct identification to species-level identification is required.

4.2. Introduction

Over the past decades, there has been an increase in incidence of foodborne illnesses caused by bacterial pathogens. An estimated of 600 million people fall ill every year, resulting in 420 000 deaths and US $110 billion loss [1]. Among all foodborne disease-causing bacteria, *Listeria monocytogenes* (*L. monocytogenes*), *Escherichia coli* (*E. coli*) O157, *Salmonella enterica* serogroups and *Shigella* spp. are among the most common and important pathogens causing public health problems annually worldwide. Outbreaks of sliced cold beef ham with *L. monocytogenes* [2], venison products with *E. coli* O157 [3], watermelon contaminated with *Salmonella* that caused a multi-country outbreak in Europe [4], and the shigellosis outbreak in a Mariscos San Juan restaurant [5] were reported in literature. Even though most foodborne illnesses are self-limiting, there is possibility of causing life threatening complications if left untreated for certain population groups (e.g., immunocompromised groups). It is important to know that not all bacteria belonging to the same genus would be pathogenic to human. For instance, in the genus *Listeria*, only *L. monocytogenes* is an opportunistic pathogen to human, with *L. ivanovii* and *L. seeligeri* occasionally reported to cause human infections [6]. Although all four species of *Shigella* cause illnesses, most *E. coli* and *Salmonella* species are not harmful for humans. Therefore, accurate and fast identification of foodborne pathogen at genus as well as species level is crucial for detection and early prevention of outbreaks.

Conventional methods for food pathogen identification include phenotypic methods (biotyping, serotyping and phage typing) and genotypic method (pulsed-field gel electrophoresis (PFGE) and polymerase chain reaction (PCR)) [7-10]. Despite being very effective, these existing classification methods are labor-intensive, time consuming, expensive, and often require highly trained personnel and specialized laboratory to carry out the experiment. Moreover, many of the conventional detection methods require extensive sample preparation and long incubation time, which may take in total up to weeks to differentiate and identify microorganisms [11]. As a result, there is an urgent need to develop a rapid, reliable and cost-effective technique, for the detection of foodborne pathogenic microorganisms.

Infrared spectroscopy has been used for the purpose of microorganism identification since 1950s [12]. More recently numerous reports of the FTIR spectroscopy were cited, demonstrating the capability to identify variations in the biochemical composition of microorganisms through changes in spectral absorption bands associated with functional groups, such as lipids, proteins,

nucleic acids and polysaccharides [13]. With a comprehensive spectral database and adequate spectral analysis methods, FTIR spectroscopy can yield microbial identification results in a matter of minutes. This technique has been shown to possess high discriminatory power for some pathogens even at the subspecies levels [6, 14, 15]. Due to its simplicity, FTIR spectroscopy is often used for characterization, quantification, identification, differentiation and classification of microorganisms, and it is utilized in an expansive range of applications, including pharmaceuticals, clinical, food, environmental, and forensic industries [16]. Other advantages are its simplicity to operate, no reagents requirements, is non-destructive, non-invasive, rapid, and most importantly cost-effective [17]. Current technology advancements allowed commercially available portable FTIR instruments to be lighter and available for field-based usage, without compromising their functionality compared to their benchtop counterparts. Thermo Scientific™ Nicolet™ Summit FTIR spectrometer (Summit) and Agilent Cary 630 FTIR spectrometer (Cary) are two differently manufacturers of FTIR spectrometers. These instruments all operate in the mid-infrared spectral region and are capable for both attenuated total reflectance (ATR) FTIR and transflectance (TR) FTIR spectra acquisition.

The objective of this study is to compare the intra-instrument (Summit prediction set vs. Summit database) prediction correctness to that of inter-instrument (Cary prediction set vs Summit database) prediction, and to evaluate the compatibility of these two FTIR instruments in terms of general use for bacteria identification. In our previous work, we have demonstrated that TSA) and CBA are reliable cultivation media for bacterial strains for use in infrared spectral database construction, as change in agar medium did not significantly influence the identification accuracy. Therefore, we did not limit the use of a single growth medium for the prediction set. The aim of this study is to investigate the microbial identification compatibility of intra- and inter-FTIR instruments.

4.3. Materials and Methods

4.3.1. Bacterial strains

The training and validation spectral data sets comprises ATR-FTIR spectra (in triplicate) of 361 isolates from 41 microbial species of *Enterobacter* spp. (n = 27), *Enterococcus* spp. (n = 7), *E. coli* (n = 159), *Klebsiella pneumoniae* (n = 5), *Listeria* spp. (n = 77), *Pseudomonas* spp. (n = 11), *Salmonella* spp. (n = 42), *Shigella* spp. (n = 25), and *Staphylococcus* spp. (n = 8). An external spectral data set (test set) was created independently of the training and validation spectral sets and was used to test the predictive accuracy of the classification models. The prediction set includes a total of 443 isolates belonging to four bacteria genera represented in the spectral database: including *E. coli* (n = 51), *Listeria* spp. (n = 207), *Salmonella* spp. (n = 132), and *Shigella* spp. (n = 53). All isolates were obtained from the strain collection of Health Canada (HC) microbiology lab (Longueuil, Quebec, Canada) and Canadian Food Inspection Agency (CFIA) (Ottawa, Ontario, Canada) (Table 4.1). Strains were stored in 25% glycerol broth vials and frozen at -80℃ using the microbank system (Microbanks, Pro-Lab Diagnostics, Richmond Hill, Canada).

All isolates were previously identified and confirmed by MALDI-TOF Biotyper mass spectrometer (Bruker Daltonics, Germany) and reference biochemical methods at the Health Canada and CFIA microbiology laboratories. Serotyping of the *Salmonella* serogroups and *Shigella* serotypes were carried out at the HC laboratory.

4.3.2. Sample preparation for FTIR analysis

The prediction strains were divided into two sets: training and prediction sets. Bacterial isolates from the training set and one prediction set were incubated directly from the microbank cryotube on Tryptic soy agar (TSA; BD Difco; Le Pont de Claix, France) for 18-24 h at 35-37°C. The second prediction set were incubated on Brain heart infusion agar (BHI; BD Difco; Detroit, USA) for 18-24 h at 35-37°C. For purity purpose, all isolates were subcultured twice, and samples suspected being contaminated were rejected.

Table 4.1. List of bacterial strains employed in the construction of the ATR-FTIR spectral databases. Spectral were acquired on two separate ATR-FTIR spectrometers.

Strains		Training set[1]	Validation set[2]	
			Nicolet™ Summit FTIR[3]	Cary 630 FTIR[4]
Enterobacter				
	sakazakii	5		
	cloacae	10		
	kobei	2		
	spp.	11		
Enterococcus spp.		7		
Escherichia coli			36	15
	O157:H7	30		
	Other	129		
Klebsiella pneumoniae		5		
Listeria				
	grayi	7	2	4
	innocua	8	1	13
	ivanovii	9	1	5
	monocytogenes	36	8	150
	murrayi	1		1
	seeligeri	8		5
	welshimeri	8		7
	spp.			10
Pseudomonas spp.		10		
Salmonella				
	serogroup B	10	11	32
	serogroup C1	2	33	3
	serogroup C2C3	6	4	20
	serogroup D	4	3	6
	serogroup E	6	2	3
	Other serogroups	14	2	13
Shigella				
	boydii	5	8	3
	dysenteriae	5	7	4
	flexneri	7	10	7
	sonnei	8	10	4
Staphylococcus				
	aureus	3		
	capitis	1		
	epidermidis	1		
	saprophyticus	1		
	xylosus	2		
Total (spectra)		**361 (1068)**	**138 (407)**	**305 (940)**

[1]Isolates from training set came from HC, and they were all grown on TSA and acquired by Summit.

[2]Isolates from validation set came from CFIA.

[3]Although the database was built based on bacterial spectra acquired on TSA, the validation set acquired by Summit were grown on BHI.

[4]The Cary validation set were grown on the same cultivation medium as the database, that is, TSA.

4.3.3. Spectroscopic measurements

FTIR spectrometers were acquired from two different manufacturers and used for spectral acquisition. Specification of each of the two instruments can be found in Table 4.2. A loopful of bacterial cells is transferred using a sterile disposable 1-μL loop to the surface of the ATR-FTIR accessory. Spectra were acquired using OMNIC™ Paradigm Software for the Summit, and the Agilent Microlab software for Cary-630 spectrometer. Prior to the collection of each spectrum of a microorganism, a background spectrum of the clean surface of the ATR was obtained. The spectrum of the sample is ratioed against the spectrum of the clean ATR surface to obtain a transmittance spectrum which is then transformed to an absorbance spectrum prior to addition to the ATR-FTIR spectral database. For each sample of the Summit, three replicates spectra were collected from three separate colonies acquired from the same plate. For samples of the Cary, three to four replicates spectra were collected from separate colonies acquired from the same plate. A total of 32 scans were coadded at 8 cm^{-1} resolution, with zero order filling factor and Happ-Genzel apodization of the interferogram. The spectra were acquired over wavenumber range between 4000 and 650 cm^{-1}. The final spectral database comprised a total of 1068 spectra and was used for the defilement of classification models.

Table 4.2. ATR-FTIR spectrometer specifications employed in this study[1].

Manufacturer	**Agilent Technologies**	**Thermo Fisher Scientific**
Model name	Cary 630 FTIR	Nicolet™ Summit FTIR
Dimensions	22.9cm x 15.2cm	32cm x 53cm
Beam diameter	7mm	6mm
Weight	5kg	12.9kg
Spectral collection software	MicroLab	Omnic Paradigm

[1]ATR accessory (Everest, SPECAC, UK) was used with the Nicolet™ Summit FTIR spectrometer. Both ATR accessories used a diamond ATR crystal**.**

4.3.4. Statistical analysis

Prior to statistical analysis, all spectra were transferred to first derivative, and vector normalized (between 1800 and 900 cm^{-1}) to minimize baseline drift and variation in sample thickness respectively. Outlier detection stemming from spectral artifacts were identified by hierarchical cluster analysis (HCA) using Ward's algorithm as the linkage type. The remaining spectra were subdivided randomly into training and test sets with equal fractioning at species level,

and the replicate of each strain were assigned to the same set. The training set were employed in the construction of the databases. Multitier classification models were built in a pairwise fashion. Briefly, spectra were classified into separate genera or species groups based on the spectral distances derived from the hierarchical cluster analysis (Figure 4.1 and Figure 4.2). Forward region selection algorithm was used in conjunction with principal component analysis (PCA) by employing the in-house written software along with JMP Pro 16 (Statistics Discovery Software; SAS Institute, Cary, NC, USA) for the selection of spectral regions contributing to the discrimination among genus and species groups. The regions between 3100-2800 cm^{-1} and 1800-900 cm^{-1} were employed as the starting point for the identification of narrower spectral regions that can be employed for effective discrimination between the different genera and species. The combination of narrower spectral regions producing the most efficient separation of classes was subsequently chosen to assess the performance of the identification using the independent spectral test set. Principal components (PCs) derived from PCA were used together with linear discriminant analysis (LDA) for development of additional classification models.

The performance of the classification model was assessed using the independent test set. It is important to note that the spectra in the test set were not employed in the spectral region selection nor building of the PCA or HCA-models. Pairwise identification at genus and species level was achieved by a multitier approach (Figure 4.1), where each spectrum from the test set is assigned to one of the pairs at each tier each step identifying the group in which the “unknown” spectra belong, until identification at species level is attained.

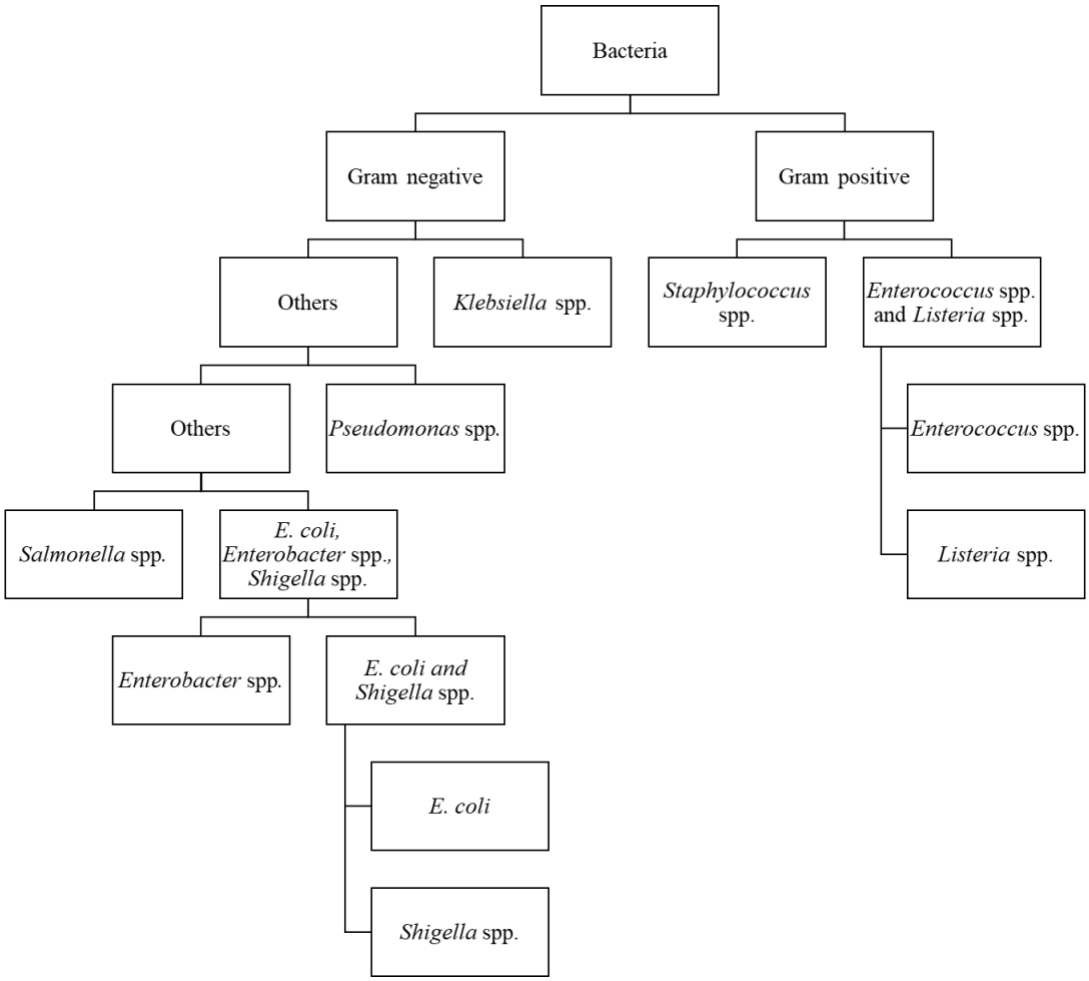


Figure 4.1.. Identification of an unknown based on a multitier pair-wise approach at the Gram and genus level.

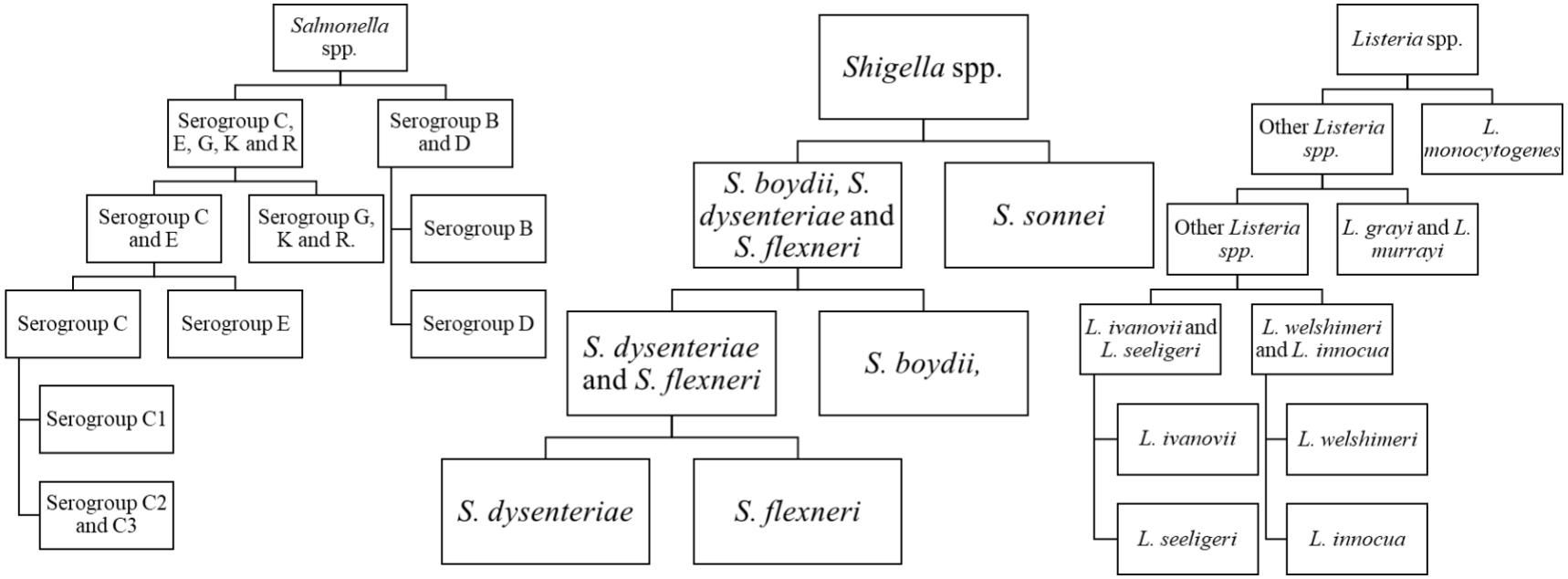


Figure 4.2. Identification of an unknown based on a multitier pair-wise approach at the species or serogroup level.

4.4. Results and Discussion

The aim of this study was to evaluate the identification accuracy of foodborne pathogens at genus and species level by ATR-FTIR spectroscopy. ATR-FTIR spectroscopy was assessed as a technique for the identification of an unknown microorganism based on spectra acquired using different ATR-FTIR spectrometers (from different manufacturers). A total of 1068 spectra of 361 isolates and 9 genera, grown on TSA were acquired by Summit ATR-FTIR spectrometer to generate a spectral database in the development of multivariate-based classification models. A total of 407 spectra of 138 strains were grown on BHI and spectra recorded by Summit ATR-FTIR spectrometer to evaluate impact of changing growth media on the predicative accuracy of the classification models. The second validation set included 940 spectra from 305 strains (used in the development of the classification models) and grown on TSA. The spectra were acquired using a different ATR-FTIR spectrometer (Cary-630) to test the impact of a change in spectrometer on the predicative accuracy.

Spectral markers for discrimination between foodborne pathogens were identified using the region selection algorithms and exploited for the bacteria at the genus and species level. An infrared spectrum of a microorganism is comprised of infrared absorption bands of lipids (3050–2800 cm^{-1}), amide band absorption regions ascribed to proteins and peptides (1700–1500 cm^{-1}), and region that reflects overlapping absorptions from side chain amino acid moieties, amide III and nucleic acids (1500–1250 cm^{-1}). Other spectral regions of interest include 1250-1200 cm^{-1}, where phospholipids, DNA and RNA absorption take place. Polysaccharides absorption region is between 1200 and 900 cm^{-1}). The fingerprint region (900–600 cm^{-1}) is rarely used for microorganism analysis due to the absorption cut-off of the infrared substrate [18, 19].

4.4.1. Spectral analysis and Development of the IR spectral database specific to Foodborne pathogens

4.4.1.1. Classification model development at the Gram and genus level

Discrimination between Gram-positive and Gram-negative bacteria was achieved through principal component analysis of the whole training set using the combined wavenumber range between 900-1800 cm^{-1} and 2800-3000 cm^{-1} (Figure 4.3). Two well-defined clusters were formed, and there were no sample outliers nor overlap detected. The major differences between Gram-positive and Gram-negative spectra were the absorbance level around 1050, 1200, 1400, 2850 and

2930 cm^{-1}, as shown in the grey-shaded area in Figure 4.4. The distinct spectral bands of Gram-positive bacteria at 1050, 1200, and 1400 cm^{-1} (Figure 4.4A) can be attributed to carbohydrates, phosphate and carboxylates, respectively [18]. These absorption bands stem from the thicker peptidoglycan and the presence of teichuronic acid in Gram-positive bacteria cell walls. Similarly, Gram-discrimination was also possible from the changes in the relative intensities within the same regions [20]. In the region between 2850 and 3000 cm^{-1} assigned to C-H stretching vibrations, a significant increase in band absorptions at 2850 and 2930 cm^{-1} (relative to the amide I intensity) is seen in Gram-negative (Figure 4.4B), reflecting the higher lipopolysaccharide content in the cell wall of Gram-negative bacteria. Rapid discrimination between Gram types using FTIR has been consistently investigated since the 1980s [20].

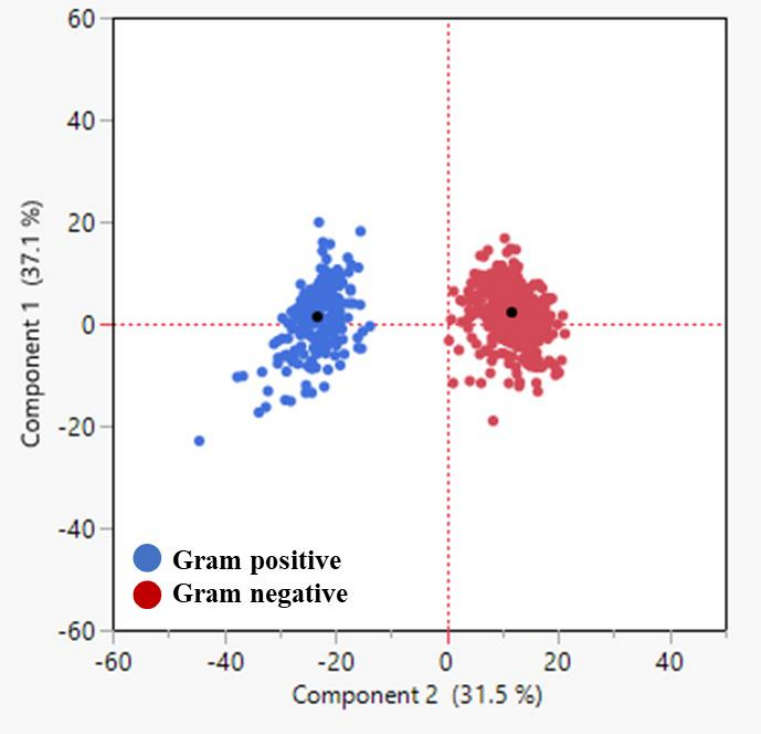


Figure 4.3. Plot of PC1 vs PC2 generated by PCA of first-derivative/vector normalized spectral data of Gram positive and Gram negative bacteria using a broad spectral range (900-1800 cm^{-1}).

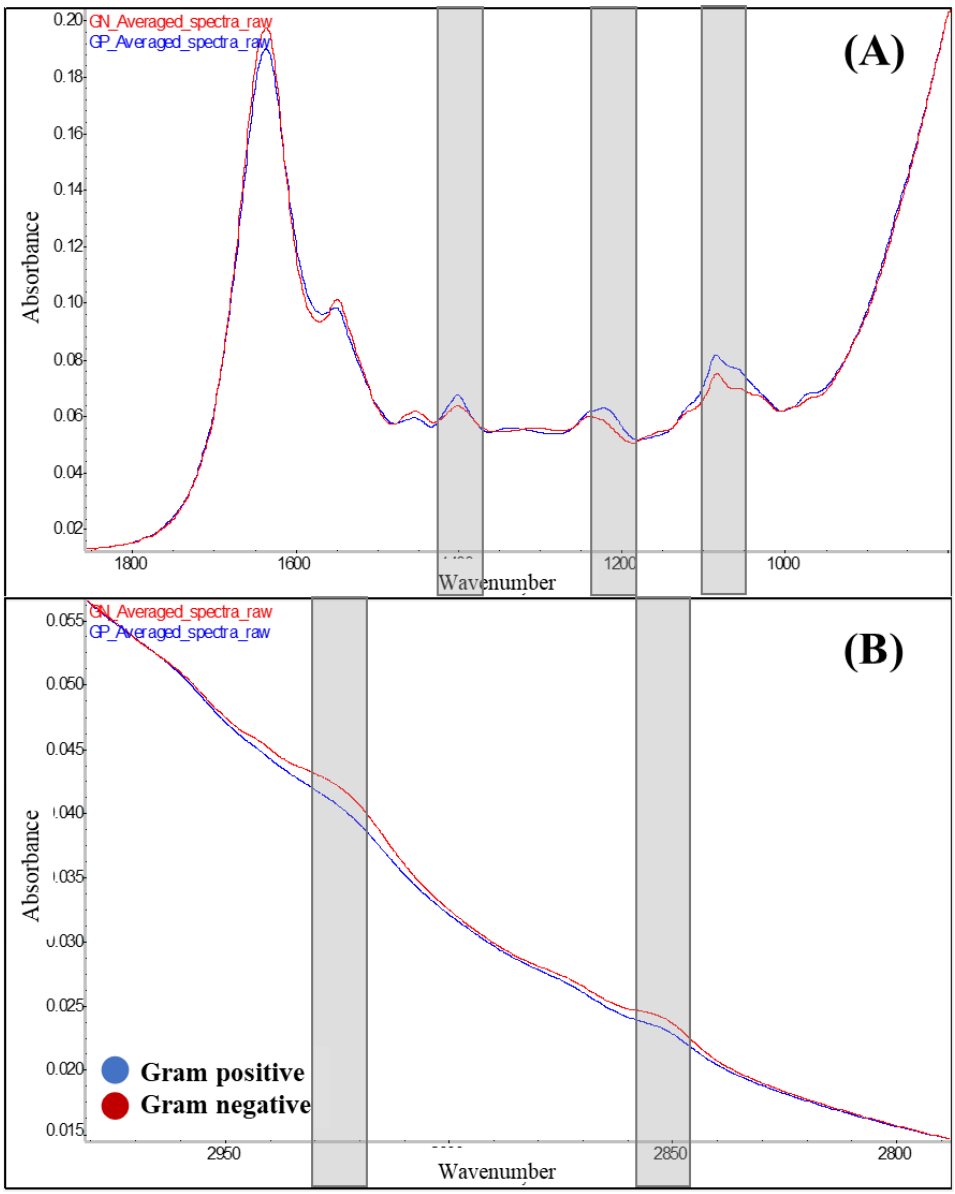


Figure 4.4. Superposition of averaged raw Gram-positive and Gram-negative spectra, in the region (A) 900-1800 cm^{-1}, and (B) 2800-3000 cm^{-1}.

Representative ATR-FTIR spectra between 1800 and 900 cm^{-1} of *Listeria* spp., *E. coli*, *Salmonella* spp. and *Shigella* spp. are shown in Figure 4.5. The infrared spectrum of *Listeria* spp. had a band at ~1440 cm^{-1} that is higher than those of the three other bacteria belonging to the family Enterobacteriaceae. As a matter of fact, *Listeria* spp. could be readily distinguished from the other three microbial genera from the colony morphology on agar plates. *Listeria* spp. colonies are significantly smaller and dryer, whereas the other three genera tend to have larger and moistier colonies. Although *E. coli*, *Salmonella* spp., and *Shigella* spp. are all from the family

Enterobacteriaceae, *E. coli* and *Shigella* spp. are more genetically similar than *Salmonella* spp. There are some gas-producing *Shigella* that resembles *E. coli*, and there are lactose-negative, non gas-producing and non-motile *E. coli* strains that are similar to *Shigella* [21]. The argument has been made that almost all *Shigella* strains could be regarded as metabolically inactive biogroups of *E. coli* [22]. On the other hand, although previous study revealed that *E. coli* has >800 genes absent from the *Salmonella* spp., and that >1100 *Salmonella* genes lack counterparts in *E. coli*, 23 of the 46 *Salmonella* O antigens are identical or very similar to an *E. coli* O antigen [23]. Nonetheless, FTIR spectroscopy was able to correctly differentiate all genera studied. The level of similarity among genera could be observed in the polysaccharide absorption region between 900 and 1200 cm^{-1} (Figure 4.5).

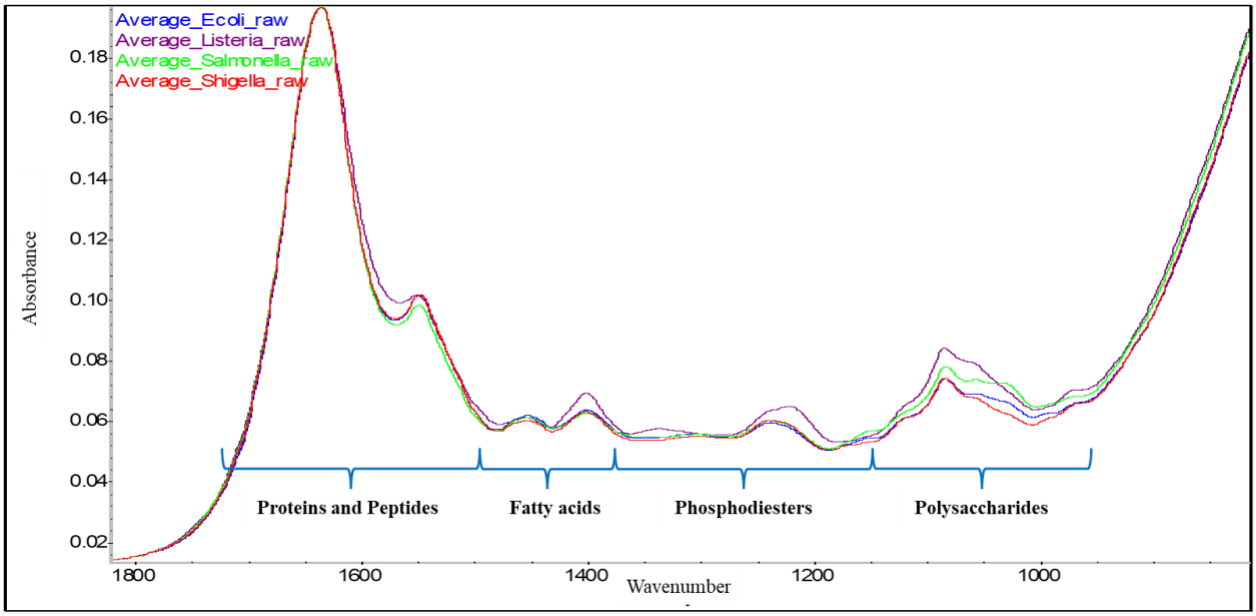


Figure 4.5. Representative absorbance ATR-FTIR spectra of *Listeria* spp., *E. coli, Salmonella* spp. and *Shigella* spp.

Figure 4.6 is an overview of species differentiation of the four genera. Each genus had identified to have different wavenumber range for the optimal separation of species, which is shaded in grey. The polysaccharide regions 900-1200 cm^{-1} (grey-shaded area in Figure 4.6A, C, D) contributed to most of species differentiation among Gram negative genera (*E. coli, Salmonella* spp., and *Shigella* spp.), as well as between *L. monocytogenes* and non-monocytogenes *Listeria* spp. (darker grey-shaded area in Figure 4.6B). The DNA and RNA regions (1200-1250 cm^{-1}), and around absorbance peaks at wavenumber 1400 cm^{-1} (lighter grey-shaded area in Figure 4.6), which

could be attributed to -CH_3 in proteins, lipids, polyesters, or COO^- in amino acid side chains and carboxylated polysaccharides, further contributed to the separation of other *Listeria* spp. [24].

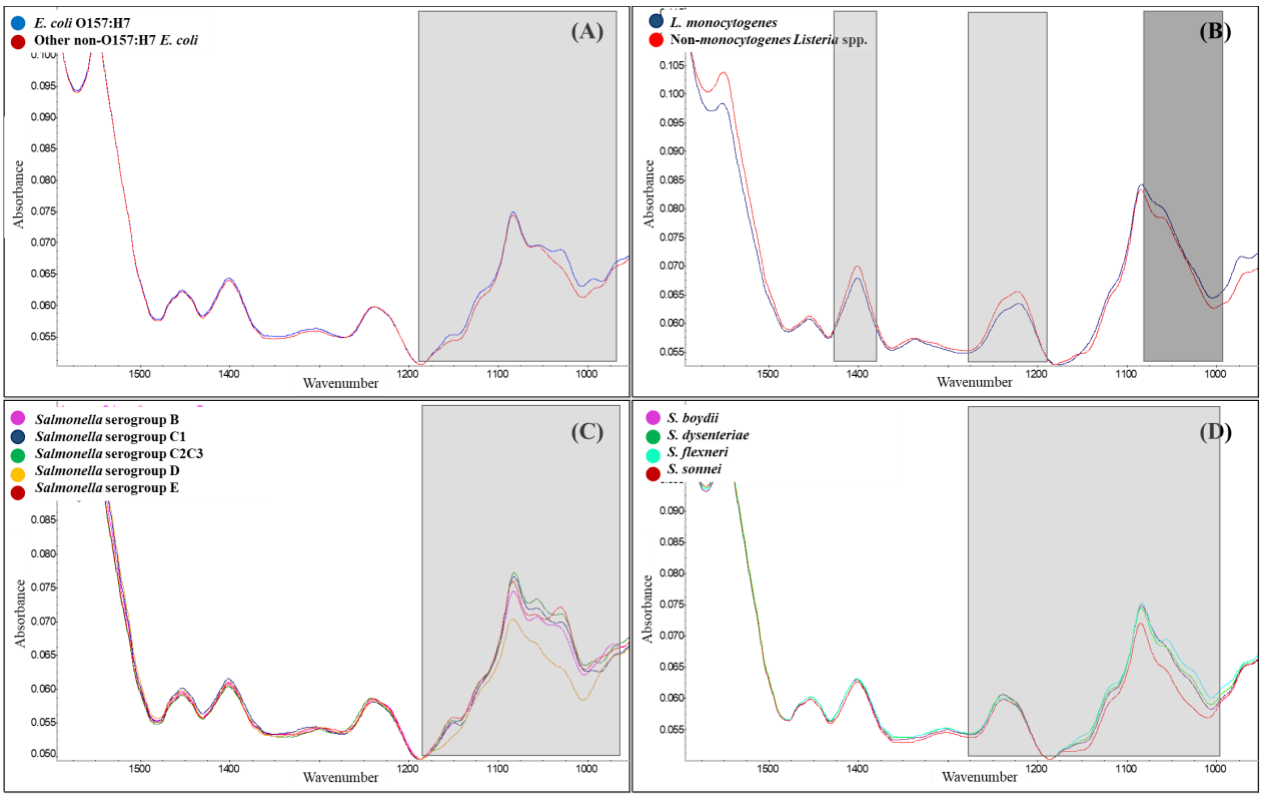


Figure 4.6. Absorbance ATR-FTIR spectra of (A) *E. coli*, (B) *Listeria* spp., (C) *Salmonella* serogroups, and (D) *Shigella* spp., and their corresponding discrimination area shaded in grey.

4.4.1.2. Classification of *Listeria* spp.

A total of 231 ATR-FTIR spectra of *Listeria* spp. were used for the development of the database, with 108 isolates from *L. monocytogene*s and 123 *Listeria* strains of *L. grayi, L. innocua, L. ivanovii, L.murrayi, L. seeligeri, L. welshimeri*. The spectral window between 1800 and 900 cm^{-1} was selected for classifier development. Region selection of the preprocessed spectra allowed for identification of specific wavenumber regions that contributes the most for differentiation of *Listeria* species (Figure 4.6B). There are two groups within the *Listeria* genus, namely *Listeria* sensu stricto, which include *L. monocytogene*s, *L. innocua, L. ivanovii, L. seeligeri, L. welshimeri*, and *Listeria* sensu lato, including *L. grayi, L.murrayi* and 10 other species [25]. Due to the pathogenicity and health interest, separation of *L. monocytogenes* from the rest of the *Listeria* species was investigated first. Interestingly, although 1200-1500 cm^{-1} seems to play crucial role

through examining raw spectra as shown in Figure 6B, the region most effective for the discrimination of *L. monocytogenes* from the other species were found between 950-1000 cm^{-1}. This region could be assigned to the carbohydrate absorption regions [26]. This latter results are in concordance with other studies reported in the literature [6, 27]. Hierarchical cluster analysis of the 231 *Listeria* spp. spectra using wavenumber regions 950-1000 cm^{-1} is shown in Figure 4.7.

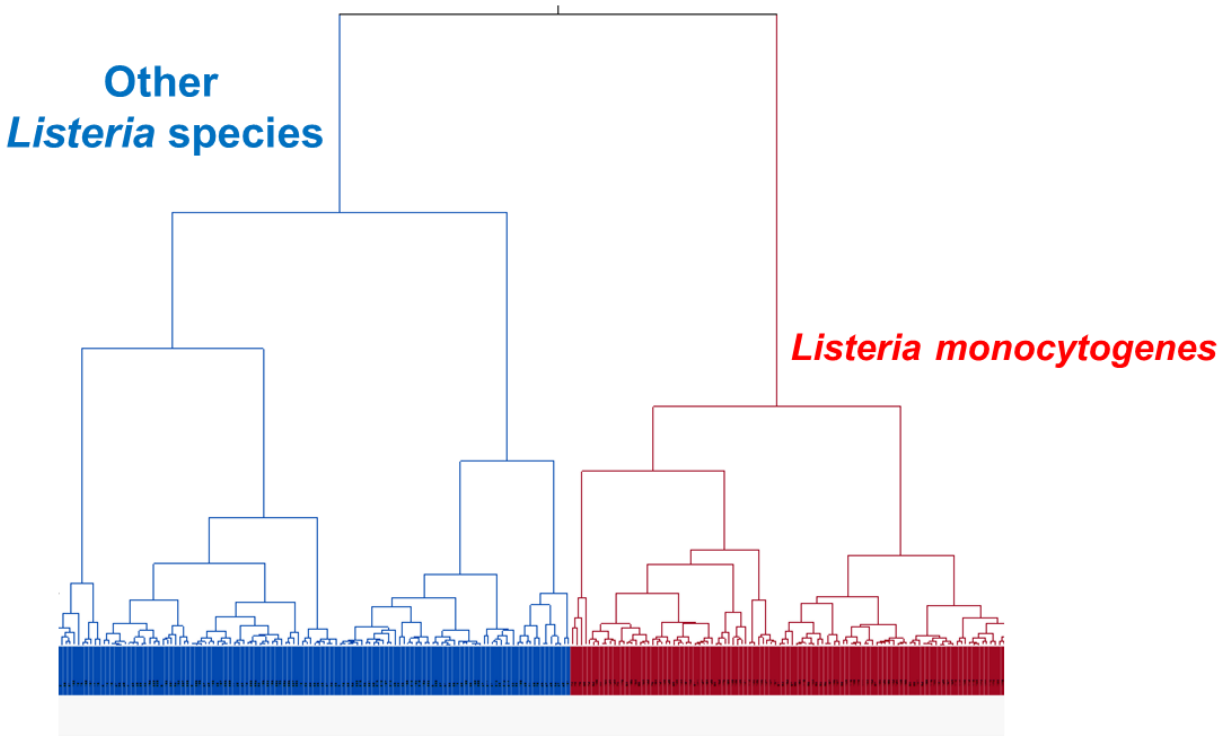


Figure 4.7. Hierarchical cluster analysis of the 231 *Listeria* spp. spectra using the wavenumber region between 950 and 1000 cm^{-1}.

After separating out *L. monocytogenes*, the rest of the species discrimination process was based on a decision tree. At each step, one species or a group of species presenting similar spectral patterns was separated using a set of wavenumber ranges as shown in Figure 4.8 *L. grayi* and *L. murrayi* both belong to *Listeria* sensu lato groups, and hence they were differentiated from other species easily by FTIR spectroscopy using the 1180-1230 cm^{-1} region which could be attributed to phospholipids and nucleic acid varibaility. Note that the two species have high level of gene similarities, their separation was not investigated in this study. While *L. grayi* and *L. murrayi* belong to the same group, they can be distinguished from other sensu lato species as a result of positive Voges-Proskauer test and motility [28]. However, *L. grayi* and *L. murrayi* can also be differentiated from *Listeria* sensu strictu species by its ability to ferment D-mannitol [25]. These

properties have led many researchers to propose *L. grayi* and *L. murrayi* could belong to the genus, *Murraya* [29]. Of the remaining four species belonging of the sensu stricto group, *L. ivanovii* is more closely related to *L. seeligeri*, as they both cause hemolysis and are unable to produce α-mannosidase, the opposite of what *L. innocua* and *L. welshimeri* can do [29]. Therefore, group *L. ivanovii-L. seeligeri* spectra were differentiated from group *L. innocua-L. welshimeri* (Figure 4.8B) using the combined region 1060-1070 cm^{-1} and 1180-1230 cm^{-1}. Finally, two clear clusters of *L. ivanovii* and *L. seeligeri* in the 3D PCA plot were obtained using the region 1060-1070 cm^{-1}. The same region was also used for the discimination between *L. innocua* and *L. welshimeri* in Figure 4.8D. The discrimination by gene similarities among species has been successfully revealed by FTIR spectroscopy employing a multitier pairwise discrimination strategy. The wavenumber regions were employed to build the classification models and used for validation purposes.

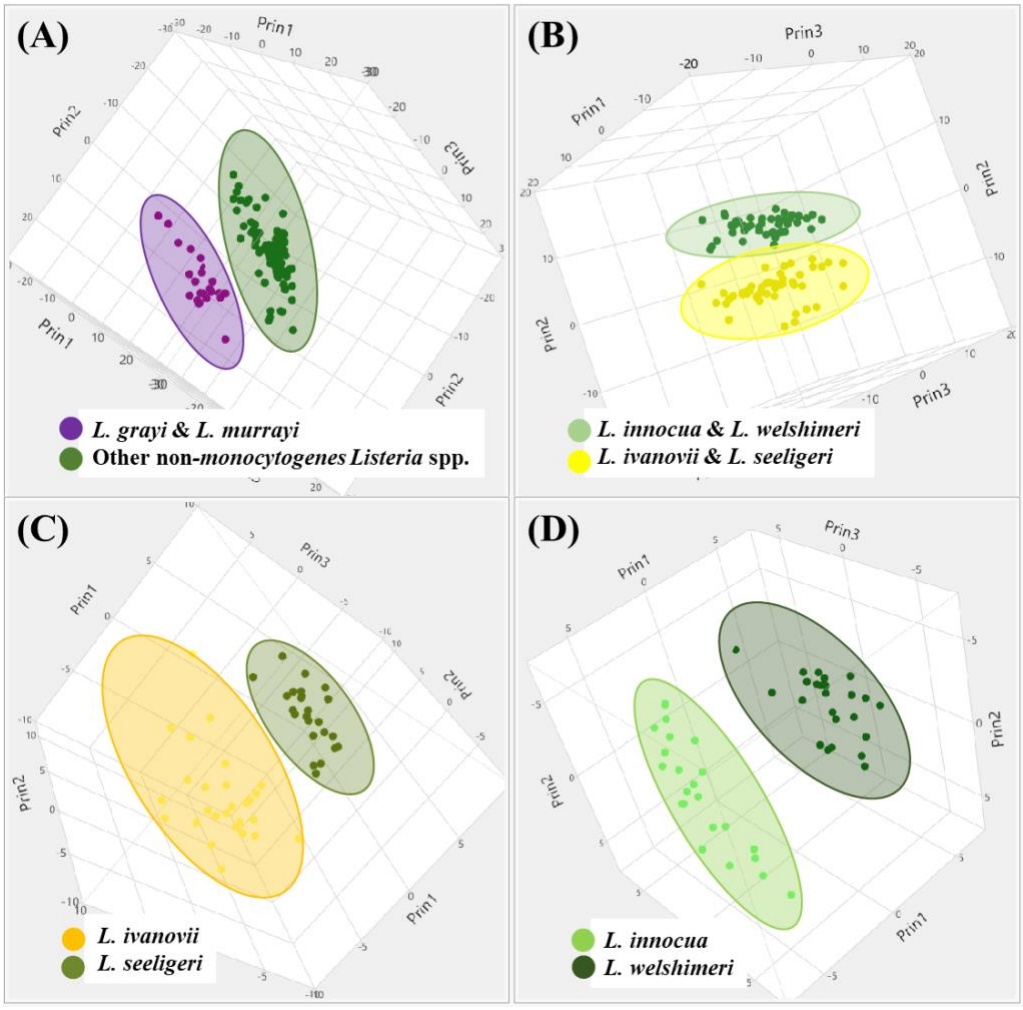


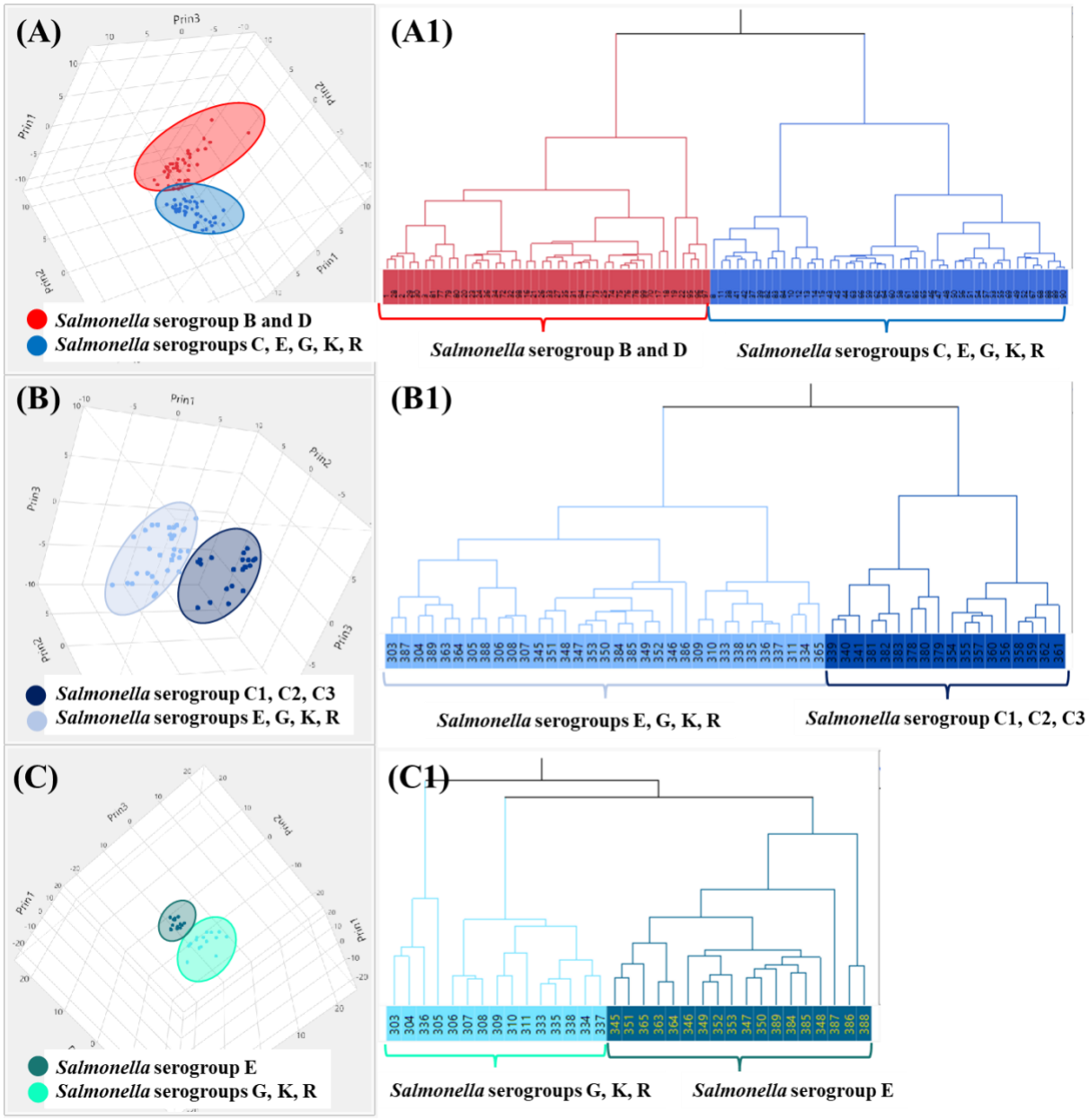
Figure 4.8. 3D score plot of PCA of non-monocytogenes *Listeria* spp. PC1, PC2 and PC3 totally expressed of 63.7% (PC1 29%, PC2 21.2%, PC3 13.5%), 82.1% (PC1 42.3%, PC2

29.3%, PC3 10.5%), 78.6% (PC1 42%, PC2 23.2%, PC3 13.4%), and 63.8% (PC1 35.6%, PC2 16%, PC3 12.2%), the variation for (A) *L. grayi* and *L. murrayi* versus other non-monocytogenes *Listeria* spp., (B) *L. innocua and L. welshimeri*, against *L. ivanovii* and *L. seeligeri*, (C) *L. ivanovii* and *L. seeligeri*, and (D) *L. innocua and L. welshimeri*.

4.4.1.3. Classification of Salmonella enterica serogroups

A total of 126 *Salmonella* spectra were included in the construction of the database, including *Salmonella enterica* serogroups B, C, D, E, G, K, and R, in which serogroups B, C1, C2, D and E cause approximately 99% of *Salmonella* infections in humans and warm-blooded animals [30]. Additionally, four out of five most common serotypes with antibiotic resistance, Enteritidis, Typhimurium, Newport and Heidelberg belonging to serogroups D, B, C2 and B, respectively. Thus, serogroups B, C, D and E are of main interest for classification. Serotype classification was not considered in this study due to the low spectral representation of each serovar. The ability of transmission-based FTIR spectroscopy for differentiation of *Salmonella* serovars has been reported in the literature [31, 32].

The differentiation of *Salmonella* serogroups was achieved between 1800 and 900 cm^{-1}. Classification of *Salmonella* serogroups was also performed in a pairwise manner resulting in multiple PCA and HCA separation plots shown in Figure 4.9. Serogroup B and D were first separated from the rest through the use of 1120-1130 cm^{-1} and 1185-1195 cm^{-1} regions (Figure 4.9A); serogroup C was then differentiated from the other groups through using 1070-1080 cm^{-1}, 1225-1275 cm^{-1}, 1350-1360 cm^{-1} regions (Figure 4.9B); after which, serogroup E was segregated from the rest by using combined spectral regions 1100-1130 cm^{-1}, 1190-1220 cm^{-1}, 1265-1280 cm^{-1}, 1320-1380 cm^{-1}, 1420-1450 cm^{-1} (Figure 4.9C). Finally, serogroup B and D were differentiated through using the wavenumber regions 1130-1140 cm^{-1} and 1305-1315 cm^{-1} (Figure 4.9D) and serogroup C1 was further separated from C2 and C3 by using the spectral regions 940-950 cm^{-1}, 1180-1225 cm^{-1}, 1260-1360 cm^{-1}, 1400-1450 cm^{-1} (Figure 4.9E). It is worth mentioning that comparing to *Listeria* spp., *Salmonella* serogroups required more spectral windows over a wider wavenumber region for discrimination, suggesting the need of higher discriminatory power for *Salmonella enterica* serogroups. From the wavenumber regions used for producing the 3D PCA plot, most of the discriminating spectral regions are found within the region 900-1200 cm^{-1}, assigned to the polysaccharide region, and within 1200-1500 cm^{-1}, assigned to phospholipids, DNA and RNA, and carbohydrates [26, 33].


(A)
Prin3
Prin1
Prin2
Salmonella serogroup B and D
Salmonella serogroups C, E, G, K, R
(A1)
Salmonella serogroup B and D
Salmonella serogroups C, E, G, K, R
(B)
Prin1
Prin2
Prin3
Salmonella serogroup C1, C2, C3
Salmonella serogroups E, G, K, R
(B1)
303 387 304 389 363 364 305 388 306 308 307 345 351 348 347 353 350 384 385 349 352 346 386 309 310 333 338 335 336 337 311 334 365 339 340 341 381 382 383 378 380 379 354 355 357 360 356 358 359 362 361
Salmonella serogroups E, G, K, R
Salmonella serogroup C1, C2, C3
(C)
Prin3
Prin2
Prin1
Prin1
Salmonella serogroup E
Salmonella serogroups G, K, R
(C1)
303 304 336 305 306 307 308 309 310 311 333 335 338 334 337 345 351 365 363 364 346 349 352 353 347 350 369 384 385 348 387 386 388
Salmonella serogroups G, K, R
Salmonella serogroup E

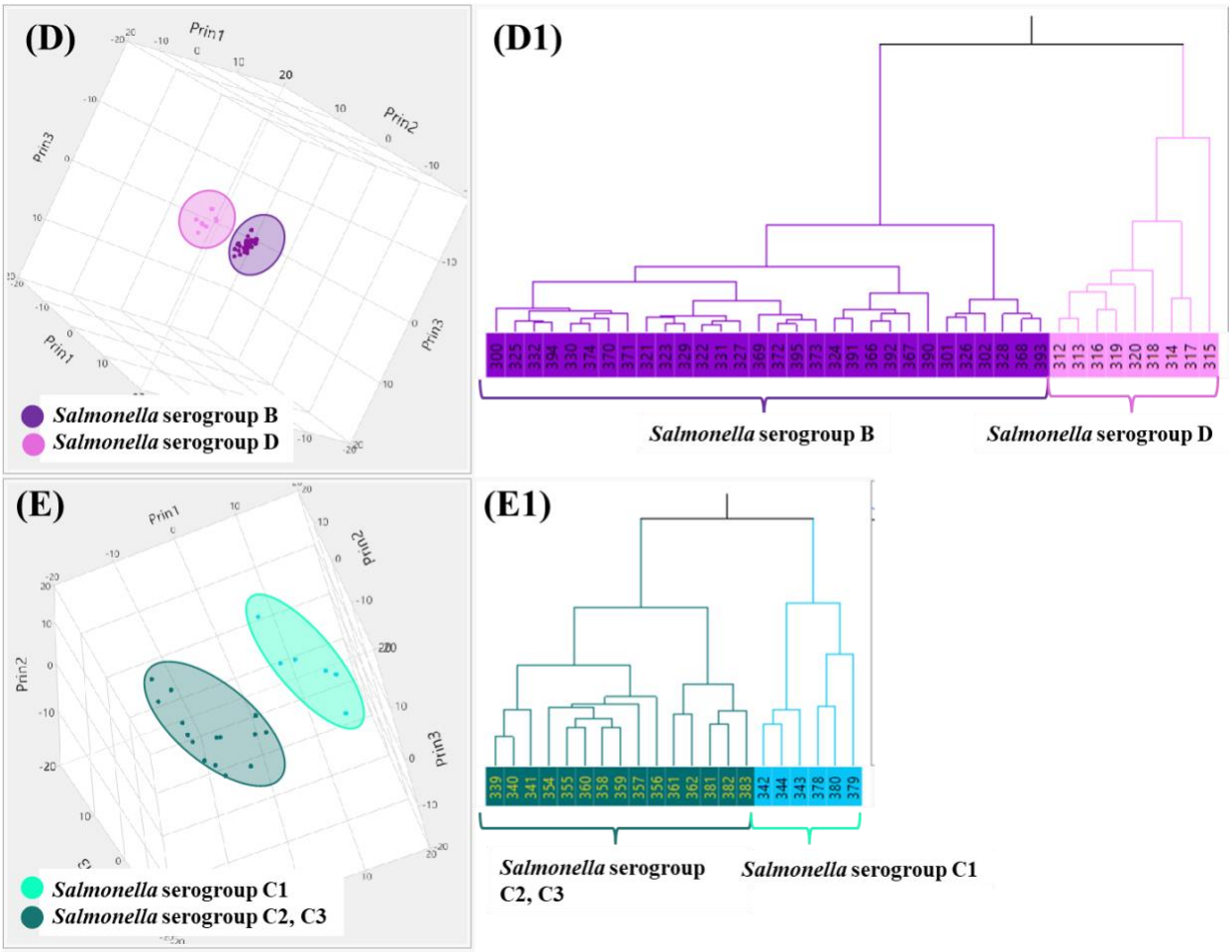


Figure 4.9. 3D score plot of PCA and HCA of Salmonella serogroups. PC1, PC2 and PC3 totally expressed of 95.11% (PC1 46.7% PC2 42.7%, PC3 5.71%), 84.2% (PC1 51.5%, PC2 22.6%, PC3 10.1%), 79.5% (PC1 32.9%, PC2 25.8%, PC3 20.8%), 97.3% (PC1 68%, PC2 24.1%, PC3 5.2%), and 76.4% (PC1 41.3%, PC2 22.3%, PC3 12.8%), the variation for (A) *Salmonella* serogroups B and D versus serogroups C, E, G, K, R, (B) *Salmonella* serogroups C1, C2, C3 versus serogroups E, G, K, R, (C) *Salmonella* serogroups E against serogroups G, K, R, (D) *Salmonella* serogroups B versus serogroups D, and (E) *Salmonella* serogroups C1 versus serogroups C2, C3.

As mentioned previously, the 1500-900 cm^{-1} spectral region comprises overlapping bands attributed to carbohydrates, phospholipids, polysaccharides and nucleic acids, contributing the most to the separation of the *Salmonella* serogroups. These regions were also reported to be useful by others in differentiating among serovars for *Salmonella* [34, 35]. In fact, due to the diverse carbohydrate composition of O antigens among *Salmonella enterica* serogroups, and differences in surface cell composition a significant impact can be observed in the IR spectra resulting in sucessful differentiation between the serogroups by ATR-FTIR spectroscopy (Figure 4.10).

Serogroup B has *abe* gene as its dideoxyhexose, whereas serogroup D has a *tyv* side-branch sugar instead, accounting for the biochemcial difference. Moreover, serogroup B could be distinguished from other serogroups by having an abequose side branch on the galactose residue of the 3-sugar main chain [33]. Despite the difference between serogroups B and D, their similarity as shown in grey blocks in Figure 4.10has made them group together from the rest of the *Salmonella* serogroups. While serogroup C was originally divided into C1, C2 and C3 on serological grounds, they are now treated as C1 and C2-C3, as C1 gene cluster (presence of O:6,7 epitopes) is quite different from C2 and C3 (presence of O:6,8 epitopes), which in turn, are identical [33, 36]. Compared to other serogroups, C2C3 has a completely different order of glycoxyltransferase genes, with a acetyltransferase gene in between. Almost all genes order in the central region of the O-antigen gene clusters are unique to serogroup C2-C3. For serogroup E, it lacks the entire CDP-sugar pathway genes that could be found in B, C and D [37]. The one thing in common of serogroups studied in this research is that they all have galactose as their first sugar in their O antigen. These differences in *Salmonella* O antigen gene clusters were successfully been employed for serogroup discrimination by ATR-FTIR spectroscopy.

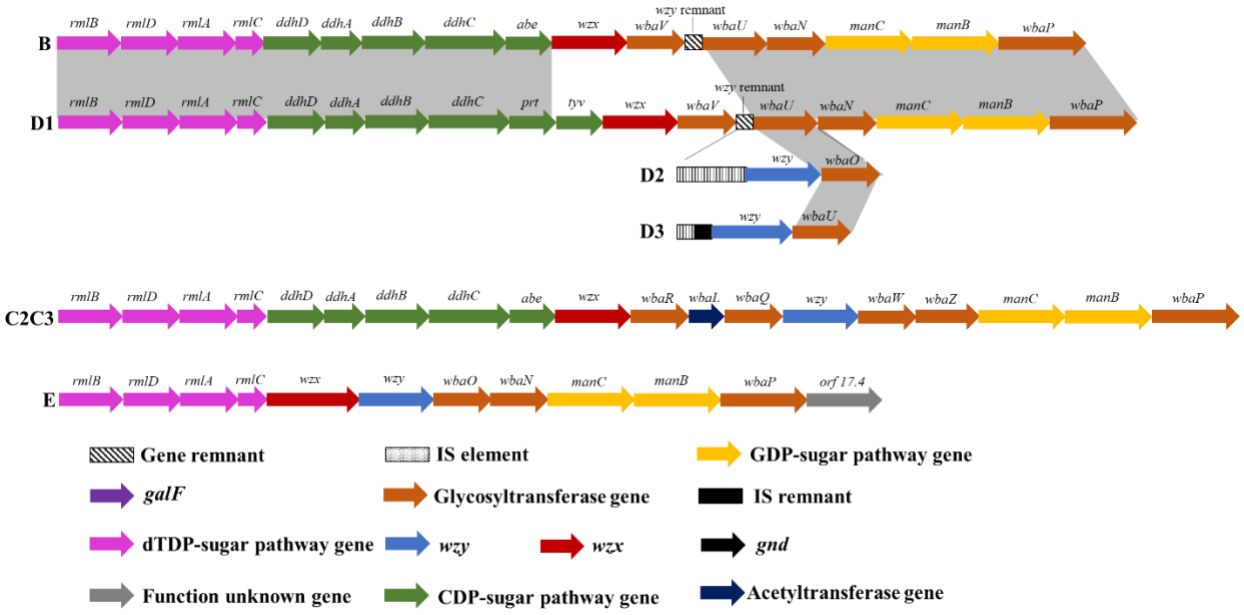


Figure 4.10. O antigen gene clusters of *Salmonella* serogroups (represented in the infrared spectral database) with regions of sequence similarity between gene clusters shaded in grey blocks.

Overall, serogroups differentiation was difficult if only the 1200-900 cm^{-1} region was used. Wider regions search for spectral windows improved the discriminatory power for the classification of *Salmonella enterica* serogroup spectra. This finding is in agreement with many studies where wider spectral windows search proved more effective than using narrower bands for discrimination among the serogroups [31, 38, 39].

4.4.1.4. Classification of E coli O157:H7

Enteroinvasive *E. coli* (EIEC), enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC), and enterohemorrhagic *E. coli* (EHEC) are common pathotypes of *E. coli*, in which *E. coli* O157:H7 belongs to EHEC. Even though pathogenicity of EPEC resembles salmonellosis and those of EIEC resembles *Shigella*, *E. coli* was correctly separately from both *Salmonella* and *Shigella* [22]. Furthermore, within the 477 spectra of *E. coli* included in the spectral database, all 90 *E.coli* O157:H7 were correctly separated using unsupervised HCA over the wavenumber region of 1030-1040 cm^{-1}, 1270-1280 cm^{-1}, 1345-1360 cm^{-1} as shown in Figure 4.11A. The differentiation region lies again within the mixed region of 900-1200 cm^{-1}, assigned to C-O-C stretch and deformation or C-O ring vibrations in polysaccharides. The region between 1270 and 1360 cm^{-1} is assigned to the amide III component of proteins [40].

Currently, molecular methods are the gold standard for serotyping of *E. coli*. These methods are based on their differences in their O, H, and K surface antigens, and requires tedious and expensive equipment to perform due to the limited sensitivity and specificity. For *E. coli*, more than 180 somatic (O), flagellar (H), and capsular (K) antigens have been proposed up to now [41]. It was reported that a rare kind of hexose sugar 4-acetamido-4,6-dideoxy-D-mannose is unique to *E. coli* O157, and may explain the difference of their spectral profile from other *E. coli* serogroups [42]. Furthermore, only a few polyprenol phosphate glycosyltransferases in *E. coli* have been identified until now, which includes *wbdN* in *E. coli* O157 [43]. It has also been observed that *E. coli* O157:H7 was missing 0.53 Mb of DNA compared to non-pathogenic *E. coli* [44]. O-antigen provides important pathotype information which is crucial in the detection and serotyping of *E. coli* by conventional methods. The same is observed for FTIR spectroscopy, where the carbohydrates region 900-1200 cm^{-1} discriminates serogroup O157 from other polysaccharide- O-antigens as shown in the PCA plot of Figure 4.11B. Interestingly, spectral differences can also be

observed in region between 1250 and 1360 cm^{-1}, which can be attributed to protein. This possibly reflects the difference of H-antigen among *E. coli* serotypes, thereby demonstrating the high discriminatory power of FTIR spectroscopy.

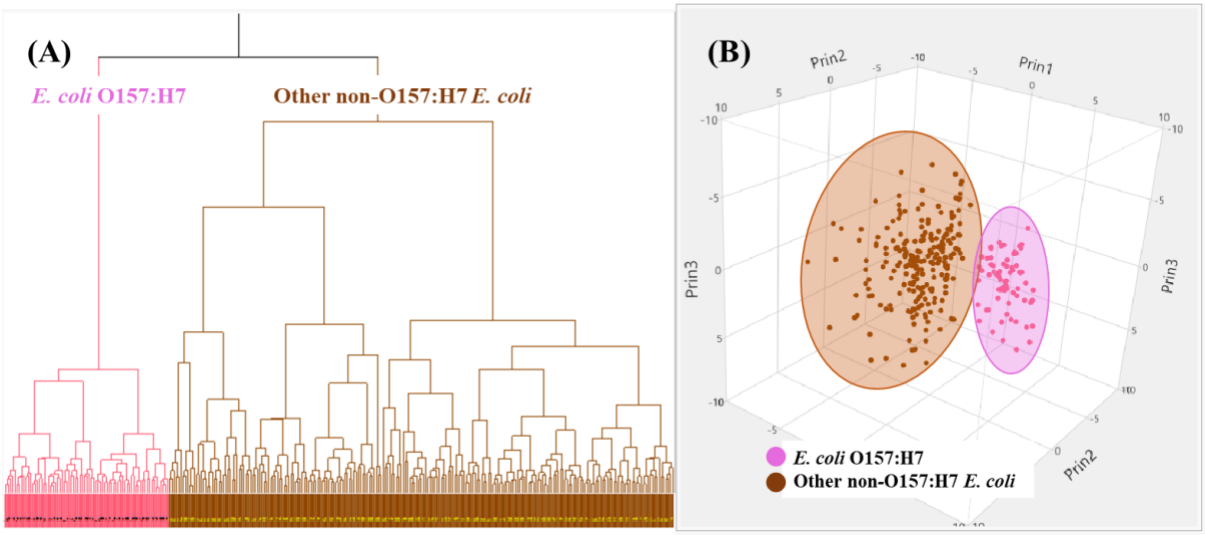


Figure 4.11. (A) HCA and (B) PCA showing the discrimination between *E. coli* O157-H7 and other *E. coli*. PC1, PC2 and PC3 account for 77.5%variablity (PC1 32.1% PC2 28.4%, PC3 17%).

4.4.1.5. Classification of *Shigella* spp.

Shigellosis is a common cause of diarrhea. In recent years, approximately 880 cases of shigellosis have been reported annually in Canada [45]. All four species of *Shigella* are included in our study. A total of 75 *Shigella* spectra has been collected for inclusion in the IR spectral database. *Shigella* spp. and *E. coli* share many common characteristics, such as similar aggregate biochemical reactions, and identical lipopolysaccharide O antigens of *Shigella* (except *S. sonnei*) to one or more of *E. coli* serotypes [46]. Many molecular methods such as 16S rRNA gene sequencing and MALDI-TOF MS are unable to differentiate *Shigella* spp. from *E. coli* [47] Despite being challenging to differentiate between these two genera and accurately classifying the four species of *Shigella*, FTIR spectroscopy accurately differentiated *Shigella* spp. from *E. coli*, and also attained species level discrimination using the similar stepwise classification method with corresponding wavenumber regions as used previously.

Again, the classification of *Shigella* spp. was also done by a stepwise method, as shown in the 3D PCA plot of Figure 4.12. A canonical variate analysis has been processed and a canonical plot was generated (Figure 4.13) to visualize the differences in the spectral distances among

Shigella spp. The different structure of the lipopolysaccharide O antigen repeats, of *Shigella* spp. has been reflected by the spectral regions employed for species discrimination. *S. sonnei* was first to separate as it differs from the other three *Shigella* spp. in having a major deletion of O antigen gene cluster between *galF* and *gnd* [48]. As expected, *S. sonnei* was correctly separated using regions 1005-1015 cm^{-1},1150-1155 cm^{-1}, and 1350-1360 cm^{-1} (Figure 4.12A), in which can be attributed to sugar–phosphate vibrations, C-O of polysaccharides, and COO- group in amino acid side chains and carboxylated polysaccharides, respectively [49]. *S. boydii* was later differentiated from *S. dysenteriae* and *S. flexneri* using spectral differences between 1060-1070 cm^{-1}, 1140-1145 cm^{-1}, 1180-1190 cm^{-1}, 1280-1305 cm^{-1} (Figure 4.12B). The 1060-1070 cm^{-1} region can be assigned to the P=O symmetric stretching in DNA, RNA and phospholipids, whereas 1140-1190 cm^{-1} belongs to the broad polysaccharide region dominated by ring vibrations of C-O-C and C-O, the 1280-1305 cm^{-1} falls in the amide III band absorption of proteins [24]. Lastly, *S. dysenteriae* and *S. flexneri* were separated using the 1065-1070 cm^{-1}, 1135-1155 cm^{-1}, 1185-1195 cm^{-1}, and 1300-1360 cm^{-1} spectral regions (Figure 4.12C). The first three regions again falls within the various polysaccharides absorption region [49]. In general, discrimination among the *Shigella* species are mainly associated with spectral changes between 900-1250 cm^{-1}, which could be due to vibrations along the sugar-phosphate chain and sensitivity to the conformation of the nucleic acid backbone, and 1250-1500 cm^{-1}, where glycosidic bond rotation and sugar puckering modes are observed [50].

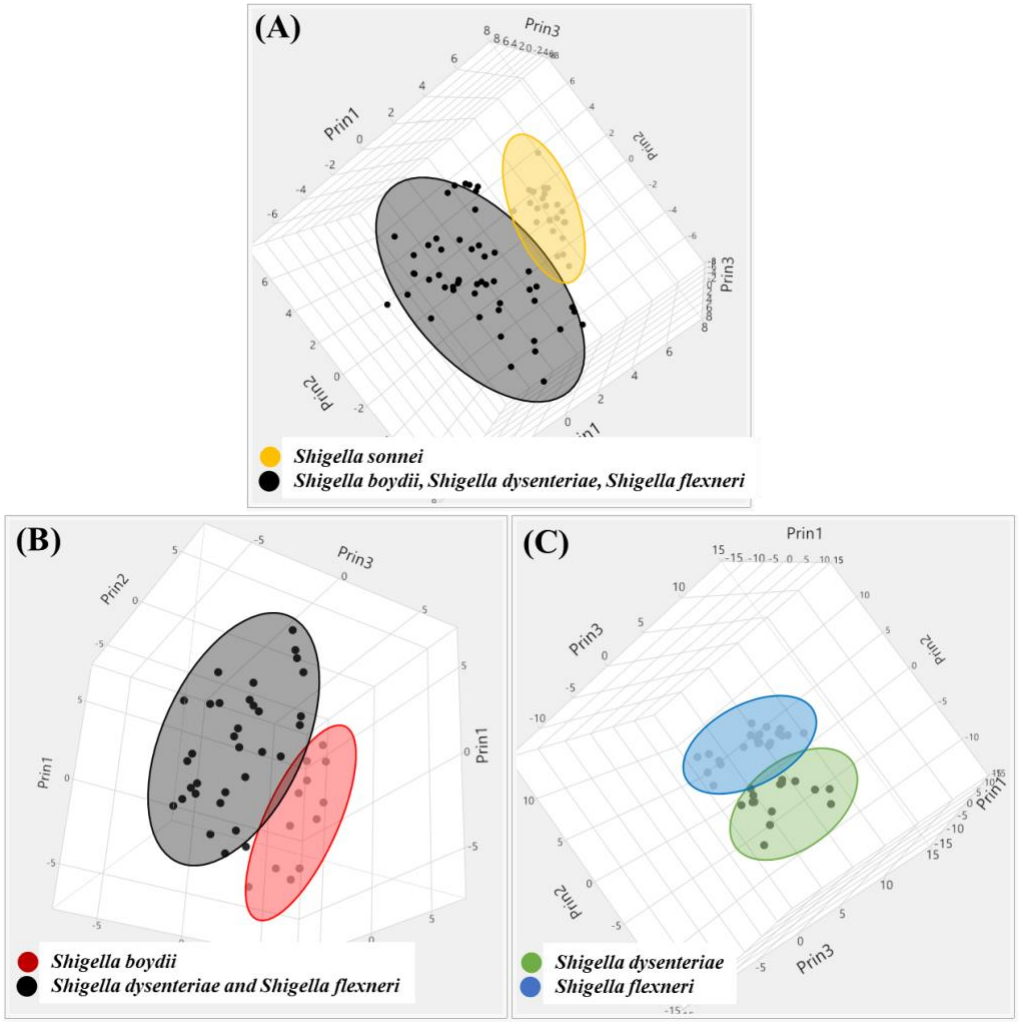


Figure 4.12. 3D score plot of PCA of *Shigella* spp. PC1, PC2 and PC3 totally expressed of 88.4% (PC1 44.1% PC2 33.4%, PC3 10.9%), 71.2% (PC1 32.1%, PC2 22.6%, PC3 16.5%), and 70.3% (PC1 39.2%, PC2 19.5%, PC3 11.6%), the variation for (A) *S. sonnei* versus other *Shigella* spp., (B) *S. boydii* against *S. dysesnteriae* and *S. flexneri,* and (C) *S. dysesnteriae* versus *S. flexneri.*

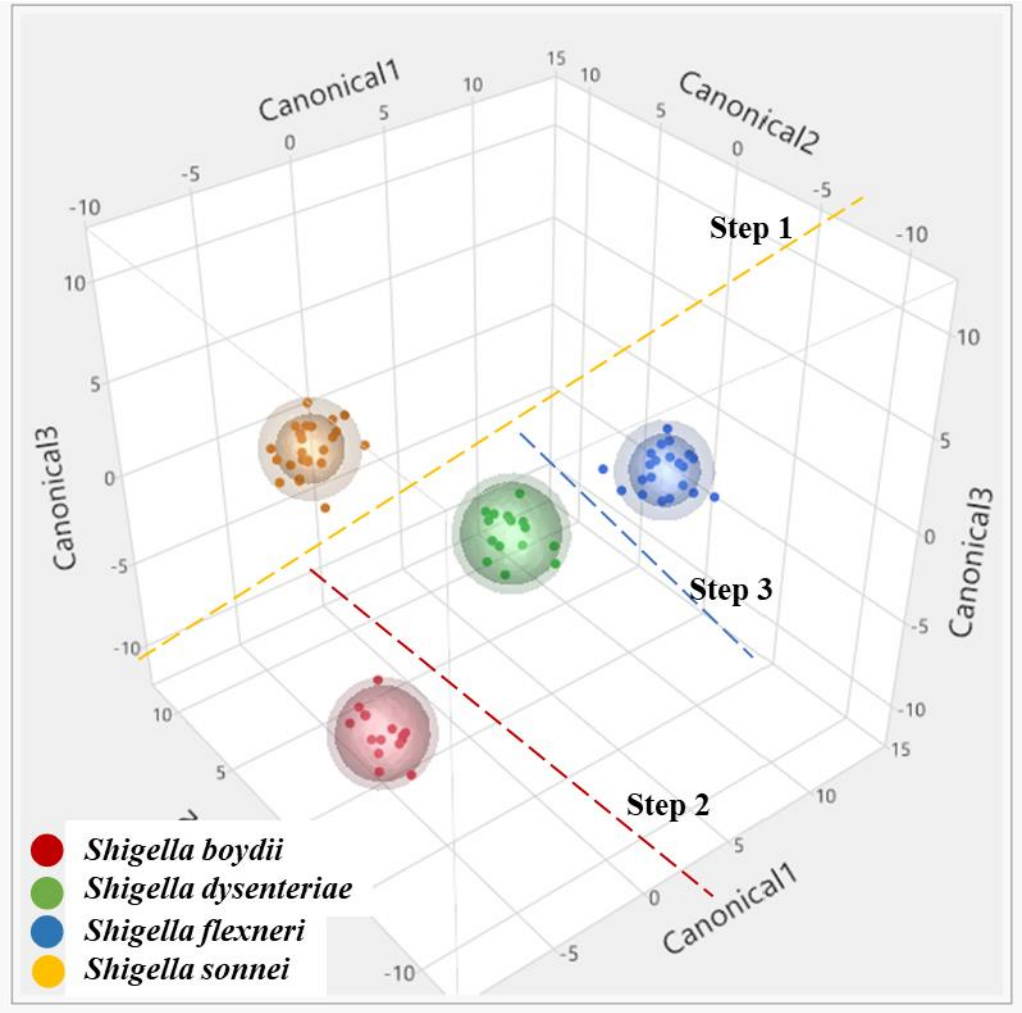


Figure 4.13. 3D Canonical plot of *Shigella* spp.

Using the region of 980-1500 cm^{-1}, the five serotypes of *S. boydii* were clearly separated from each other. As shown in the HCA dendrogram in Figure 4.14. *S. boydii* serotype 6 was first differentiated out, signifying a higher level of heterogeneity from the other 4 serotypes. Then, following the degree of similarity, serotype 14 branched out, lastly serotypes 5, 4 and 9. The increasing level of difference among *S. boydii* serotypes can be noticed within the O antigen gene clusters in Figure 4.15, with gene cluster similarity among serotypes shaded in grey blocks. ATR-FTIR spectroscopy appears to have high discriminatory power and possibility in providing subtyping capacity for *Shigella* spp.

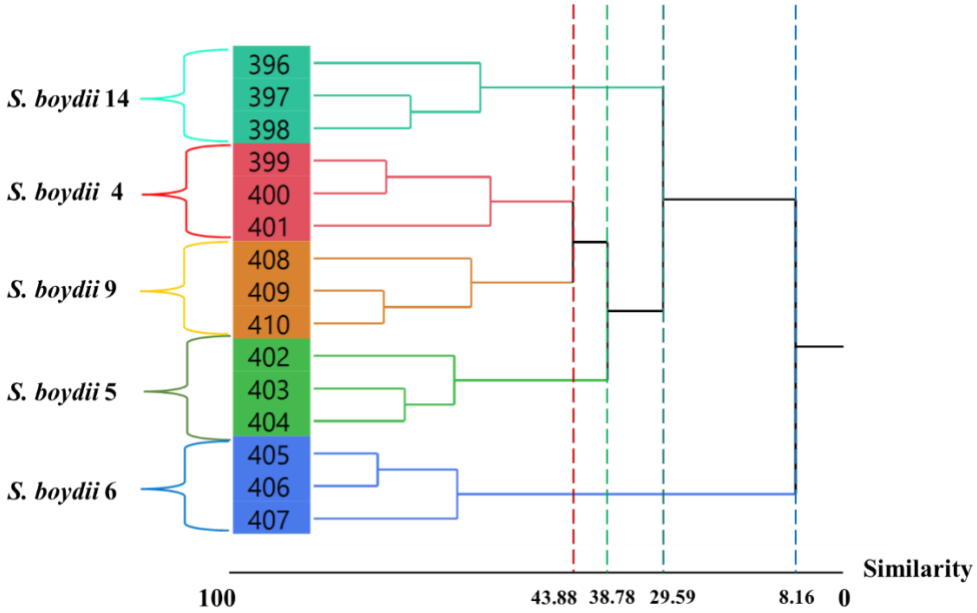


Figure 4.14.. HCA differentiation of *Shigella boydii* serotypes over broad wavenumber region 980-1500 cm^{-1}.

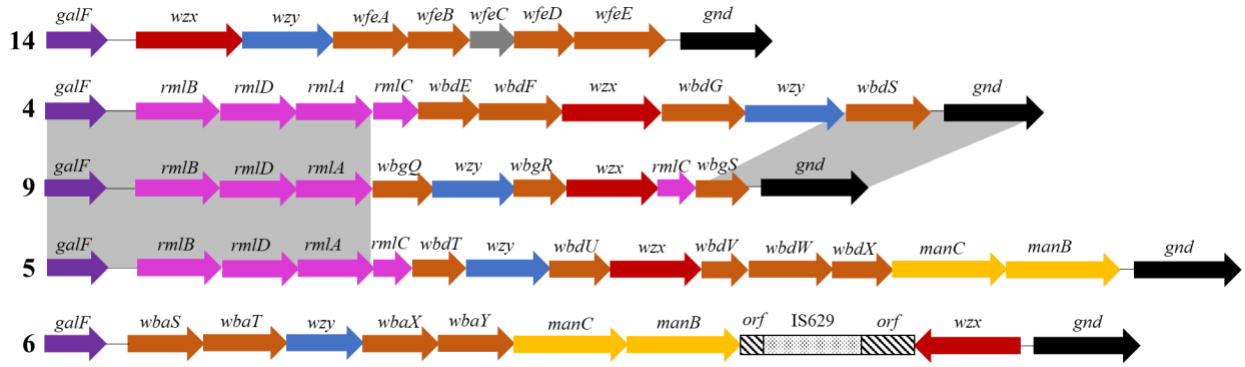


Figure 4.15. O antigen gene clusters of *Shigella boydii* serotypes included in the database with regions of sequence similarity between gene clusters shaded in grey blocks. Gene key is listed in Figure .

4.4.2. Validation of the spectral reference database

As shown in our previous study, bacteria identification by ATR-FTIR spectroscopy is robust enough despite using different culture media when the database was built using tryptic soy agar (TSA) and spectra acquired on the Summit ATR-FTIR. As such, we used BHI, which is the general agar CFIA uses the most, as cultivation medium for the validation set. This also provided the means of examining spectral compatibility of another FTIR spectrometer instrument from another manufacturer. For optimum inter-instrument identification results, the bacteria strains

were also grown on the same agar (TSA) as the strains used to build the database when they were scanned by Cary-630.

A total of 1068 spectra, comprising 361 bacterial strains were obtained by using Summit from HC for the spectral database construction. Two cross-validation sets were obtained from CFIA: (1) 407 spectra belonging to 138 strains were acquired using Summit, and (2) 940 spectra belonging to 305 strains, acquired on the Cary 630. All the spectra were first evaluated visually to remove significant outliers, and then preprocessed for data analysis. A pairwise comparison model based on PCA-LDA was applied for the identification of the two validation sets. Detailed results of genus and species level identification are listed in Table 4.3 and Table 4.4, respectively. At the genus level, despite using a different growth medium (CBA) than the database (from TSA), the Summit had a correct identification rate 99.5% compared to 96.2% for Cary 630. The results of misidentification and no identification were 0.25% and 0.25% for Summit, respectively, with only 1 spectrum of *E. coli* misidentified as *Shigella*. The Cary had relatively higher misidentification (2.8%) and no identification rate (1.1%), in which the *E. coli* was the most misidentified as *Shigella* spp.

At the species level, the Summit validation set was able to achieve a high identification accuracy of 95.8%, although the validation set were grown on a different medium than the reference strains used to build the database. On the other hand, Cary achieved only 79.9% correct identification despite using the same medium as the database. The error rate was 2.5%, and 1.7% no identification for the Summit, with most of the error were associated with *Salmonella* serogroup B and C2C3. The Cary had 15.4% misidentification rate and 4.7% no identification, with incorrectness found in almost all species evaluated in this study, except for *L. grayi* and *L. murrayi, Salmonella* serogroup D and *S. sonnei*.

Isolates from the validation sets provided by CFIA and cultured on BHI were used to evaluate the identification and robustness of a TSA-built database. Result down to species level were promising. Only 1 *E. coli* isolate was misidentified as *Shigella*; 2 *Salmonella* isolates were predicted as C1, and the other 3 as 'other serogroups', where in fact, they belonged to serogroup C2C3; all the 3 misidentified *Salmonella* serogroup B were predicted as serogroup D; and finally, a single isolate of *S. flexneri* was misidentified as *S. dysenteriae*. Despite the validation and

reference strains were grown on a different medium, these results demonstrate again the robustness of FTIR spectroscopy.

Table 4.3. Identification results of validation set on genus level by FTIR instruments.

Species	Nicolet™ Summit[1]				Cary 630[2]			
	No. of spectra (No. of strains)	Correct identification (%)[3]	Misidentification (%)[4]	No identification (%)[5]	No. of spectra (No. of strains)	Correct identification (%)	Misidentification (%)	No identification (%)
Escherichia coli	102 (36)	100 (98.04)	1 (0.98)	1 (0.98)	50 (15)	37 (74)	7 (14)	6 (12)
Listeria spp.	36 (12)	36 (100)	0 (0)	0 (0)	585 (195)	572 (97.78)	11 (1.88)	2 (0.34)
Salmonella spp.	164 (55)	164 (100)	0 (0)	0 (0)	252 (77)	245 (97.22)	5 (1.98)	2 (0.79)
Shigella spp.	105 (35)	105 (100)	0 (0)	0 (0)	53 (18)	50 (94.34)	3 (5.66)	0 (0)
Total	407 (138)	405	1	1	940 (305)	904	26	10
Identification accuracy		**99.50%**	**0.25%**	**0.25%**		**96.17%**	**2.77%**	**1.06%**

[1]Although the database was built based on bacterial spectra acquired on TSA, the validation set acquired by Summit were grown on BHI.
[2]The Cary validation set were grown on the same cultivation medium as the database, that is, TSA.
[3]Correct identification signify that the percentage of correct prediction must be $\geq$0.8000.
[4]Misidentification signify that the percentage of incorrect prediction must be $\geq$0.8000.
[5]No identification signify that the percentage of prediction result is $<$0.8000.

Table 4.4. Species level identification of the validation set on two FTIR instruments.

Species	Nicolet™ Summit				Cary 630			
	No. of spectra (No. of strains)	**Correct identification (%)**	**Misidentification (%)**	**No identification (%)**	**No. of spectra (No. of strains)**	**Correct identification (%)**	**Misidentification (%)**	**No identification (%)**
Escherichia coli	102 (36)	100 (98.04)	1 (0.98)	1 (0.98)	50 (15)	37 (74)	7 (14)	6 (12)
Listeria								
grayi	6 (2)	6 (100)	0 (0)	0 (0)	12 (4)	11 (91.67)	0 (0)	1 (8.33)
innocua	3 (1)	3 (100)	0 (0)	0 (0)	39 (13)	26 (66.67)	10 (25.64)	3 (7.69)
ivanovii	3 (1)	3 (100)	0 (0)	0 (0)	14 (5)	5 (35.71)	9 (64.29)	0 (0)
monocytogenes	24 (8)	24 (100)	0 (0)	0 (0)	451 (150)	391 (86.70)	57 (12.64)	3 (0.67)
murrayi					3 (1)	3 (100)	0 (0)	0 (0)
seeligeri					15 (5)	5 (33.33)	10 (66.67)	0 (0)
welshimeri					21 (7)	13 (61.91)	5 (23.81)	3 (14.28)
spp.					30 (10)	30 (100)	0 (0)	0 (0)
Salmonella								
serogroup B	36 (11)	30 (83.33)	3 (8.33)	3 (8.33)	101 (32)	81 (80.2)	14 (13.86)	6 (5.94)
serogroup C1	92 (33)	92 (100)	0 (0)	0 (0)	9 (3)	3 (33.33)	6 (66.67)	0 (0)
serogroup C2C3	14 (4)	9 (64.29)	5 (35.71)	0 (0)	68 (20)	51 (75)	8 (11.76)	9 (13.24)
serogroup D	9 (3)	9 (100)	0 (0)	0 (0)	20 (6)	19 (95)	0 (0)	1 (5)
serogroup E	6 (2)	6 (100)	0 (0)	0 (0)	11 (3)	7 (63.64)	3 (27.27)	1 (9.09)
other serogroups	7 (2)	7 (100)	0 (0)	0 (0)	43 (13)	33 (76.74)	5 (11.63)	5 (11.63)
Shigella								
boydii	24 (8)	24 (100)	0 (0)	0 (0)	9 (3)	1 (11.11)	6 (66.67)	2 (22.22)
dysenteriae	21 (7)	21 (100)	0 (0)	0 (0)	11 (4)	6 (54.55)	4 (36.36)	1 (9.09)
flexneri	30 (10)	27 (90)	1 (3.33)	2 (6.67)	21 (7)	17 (80.95)	1 (4.76)	3 (14.29)
sonnei	30 (10)	30 (100)	0 (0)	0 (0)	12 (4)	12 (100)	0 (0)	0 (0)
Total	407 (138)	390	10	7	940 (305)	751	145	44
Identification accuracy		**95.82%**	**2.46%**	**1.72%**		**79.89%**	**15.43%**	**4.68%**

To the best of our knowledge, no study has ever evaluated the database compatibility for the identification of foodborne pathogen between two FTIR instruments from different manufacturers. In this study, Summit FTIR spectrometer was used as the standard instrument for bacteria identification, and hence, it was used for the acquisition of reference strains from HC for the database construction. For the validation sets, one was acquired using our standard Summit spectrometer, the other set was acquired using a Cary-630. Overall, genus level reached an acceptable identification correctness of 96.2% in a inter-instrument evaluation study. *E. coli* had the lowest identification accuracy at only 74%. The 7% misidentified isolates were all predicted as *Shigella*. Nonetheless, only 3 *Shigella* spectra were mistaken as *E. coli*. It is worth mentioning that misidentifying *E. coli* as *Shigella*, is far less dangerous than the other way around, as *Shigella* may produce more serious complications and cause a lot more deaths than *E. coli* throughout the world, especially in underdeveloped countries [51]. At the species level, *S. sonnei* had outstandingly 100% correctness. This is probably because that *S. sonnei* has an atypical O antigen compared to the other three species as discussed previously, lacking the entire dTDP-sugar pathway and GDP-sugar pathway gene cluster [48]. Identification of the *Salmonella* serogroups was not easy inter-instruments. For *Salmonella* serovar isolates belonging to serogroup B, C2C3 and D, identification result was 80.2%, 75%, and 95%, respectively. Interestingly, identification of *Salmonella* serogroup C2C3 yielded higher correct classification rate for inter-instrument (75%) than intra-instruments (64.3%). Serovar level identification was also promising using FTIR spectroscopy in previous studies [34, 52, 53].

Identification of *L. monocytogenes* is crucial compared to other *Listeria* spp. due to its clinical complications and mortality. Intra-instrument identification yielded 100% correctness, whereas inter-instrument was only 86.7%. No misidentification for *L. grayi* and *L. murrayi* in both validation sets was observed and may be attributed to the significant differences from the other *Listeria* spp. both phenotypically and genotypically. The intra-instrument high identification correctness and lower accuracy inter-instrument prediction signify that although using the same parameters for spectral acquisition, the optical design used by different companies may influence spectral output. As shown in Figure 4.16, there are visual peak differences between two spectra acquired by the two FTIR spectrometers, despite being acquired from the exact same *L. monocytogenes* isolate at the same time.

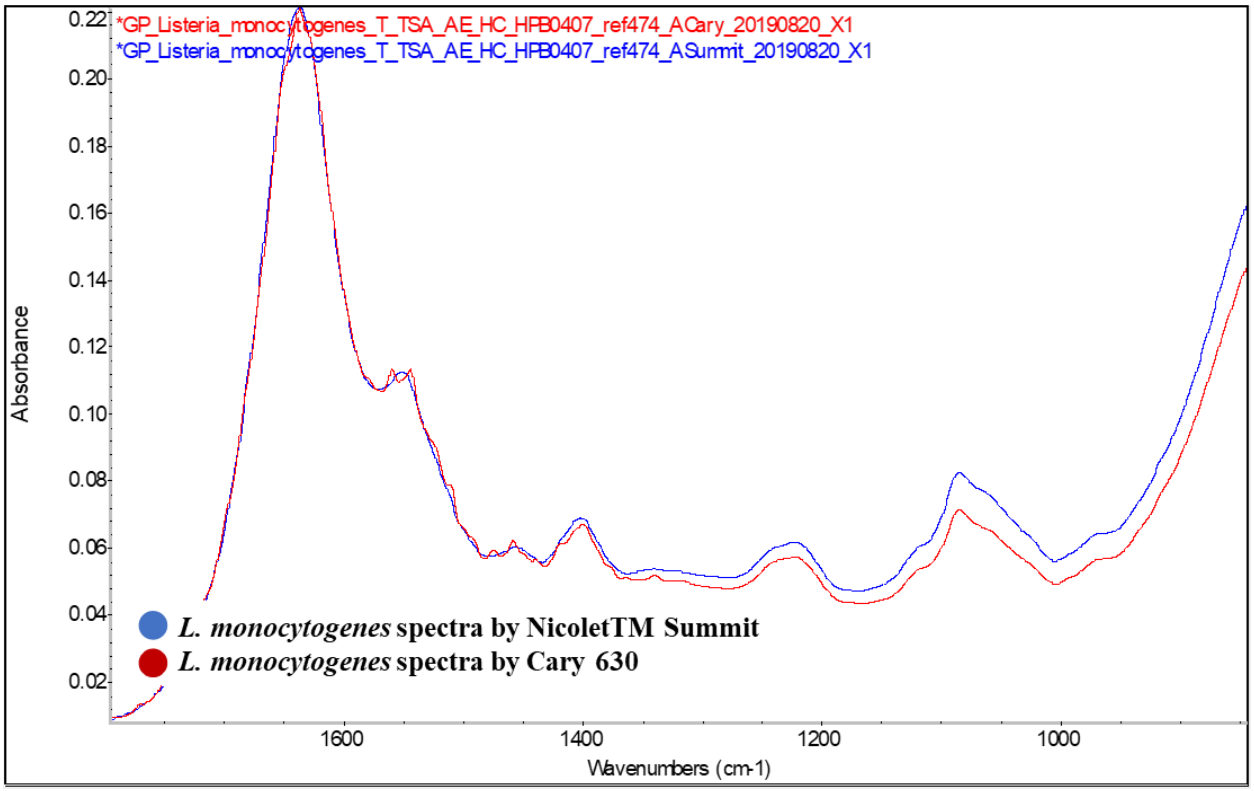


Figure 4.16. Raw FTIR spectra of the same *L. monocytogenes* isolate grown on TSA but acquired on different FTIR instruments from different manufacturers.

It is important to investigate the effect of growth media interchangeability in terms of correct identification rate for bacteria, with the objective of having a media independent spectral database for identification of isolates grown on different media. In this and previous studies done by our group and others, significant differences in the FTIR spectra of the same isolate grown on different media are shown. Nonetheless, TSA would be a good choice as a general growth medium for the database construction, and that it is robust enough to be used for the prediction of CBA- and BHI-grown validation set. Additionally, evaluation of identification compatibility between different manufactured FTIR instrument is also valuable, as it will reduce the financial struggle of purchasing an additional instrument when the FTIR database is built based on another instrument. It is noteworthy that both Summit and Cary-630 have achieved excellent performance on bacteria identification when they were used individually [54, 55]. In here, we achieved high accuracy (96.2%) at genus level identification regardless of the instrument manufacturer. Lower accuracy (11.1% to 100%) was achieved at species level. On the other hand, intra-instrument identification achieved 99.5% at genus level and 64.3% to 100% at species level. The spectral regions found to

be the best for discrimination among genus and species for the ATR-FTIR Summit may not work efficiently for Cary spectra, and vice versa.

Overall, FTIR spectroscopy is shown to be a useful technique for the identification of foodborne microbial pathogenic regardless of media and instruments used for genus-level discrimination and identification. At the species level, the cultivation medium seems not to have large influence on the identification correctness, but the change in instruments may make the species prediction doubtful.

4.5. Conclusion

In the present study, we analyzed selected foodborne pathogens and created a spectral database, followed by comparing the identification accuracy when using two manufactured FTIR instrument from different manufacturers. In general, FTIR spectroscopy has shown to be useful tor the rapid identification in routine. FTIR spectroscopy is rapid, low cost, reagent-free, and requires no sample preparation after incubation. The effect of alteration in growth medium on the identification accuracy is minimal at species level, and the use of different FTIR instrument does not pose notable misidentification errors at the genus level. To our knowledge, no prior study has compared two commercial FTIR instrument from different manufacturers for database compatibility in terms of pathogen identification. The database created in this study can be used as a first-line tool and readily applied for routine identification of food pathogens along with traditional genotypic methods after receiving regulatory approval.

4.6. Reference

1. WHO, *Food Safety.* 2023. Available from: https://www.who.int/health-topics/food-safety#tab=tab_1
2. Maurella, C., et al., *Outbreak of febrile gastroenteritis caused by Listeria monocytogenes 1/2a in sliced cold beef ham, Italy, May 2016.* Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 2018. **23**(10).
3. Smith-Palmer, A., et al., *Outbreak of Escherichia coli O157 Phage Type 32 linked to the consumption of venison products.* Epidemiol Infect, 2018. **146**(15): p. 1922-1927.
4. Byrne, L., et al., *A multi-country outbreak of Salmonella Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012.* Euro Surveill, 2014. **19**(31): p. 6-13.
5. Marler, B. *Mariscos San Juan Employee a Shigella Victim Too*. 2015 June 23, 2022]; Available from: https://www.foodpoisonjournal.com/foodborne-illness-outbreaks/mariscos-san-juan-employee-a-shigella-victim-too/.
6. Rebuffo, C.A., et al., *Reliable and rapid identification of Listeria monocytogenes and Listeria species by artificial neural network-based Fourier transform infrared spectroscopy.* Appl Environ Microbiol, 2006. **72**(2): p. 994-1000.
7. Ferrari, R.G., P.H.N. Panzenhagen, and C.A. Conte-Junior, *Phenotypic and Genotypic Eligible Methods for Salmonella Typhimurium Source Tracking.* Frontiers in Microbiology, 2017. **8**.
8. Hilliard, A., et al., *Genomic Characterization of Listeria monocytogenes Isolates Associated with Clinical Listeriosis and the Food Production Environment in Ireland.* Genes (Basel), 2018. **9**(3).
9. Li, B., H. Liu, and W. Wang, *Multiplex real-time PCR assay for detection of Escherichia coli O157:H7 and screening for non-O157 Shiga toxin-producing E. coli.* BMC Microbiol, 2017. **17**(1): p. 215.
10. Joensen, K.G., et al., *Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data.* J Clin Microbiol, 2015. **53**(8): p. 2410-26.
11. Cho, I.H. and S. Ku, *Current Technical Approaches for the Early Detection of Foodborne Pathogens: Challenges and Opportunities.* Int J Mol Sci, 2017. **18**(10).
12. Lam, L.M.T., et al., *Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy.* Journal of clinical microbiology, 2019. **57**(5): p. e01739-18.
13. Naumann, D., D. Helm, and H. Labischinski, *Microbiological characterizations by FT-IR spectroscopy.* Nature, 1991. **351**(6321): p. 81-82.
14. Rodrigues, C., et al., *A Front Line on Klebsiella pneumoniae Capsular Polysaccharide Knowledge: Fourier Transform Infrared Spectroscopy as an Accurate and Fast Typing Tool.* mSystems, 2020. **5**(2): p. e00386-19.
15. Kümmel, J., et al., *Staphylococcus aureus Entrance into the Dairy Chain: Tracking S. aureus from Dairy Cow to Cheese.* Frontiers in Microbiology, 2016. **7**.
16. Alvarez-Ordóñez, A., et al., *Fourier transform infrared spectroscopy as a tool to characterize molecular composition and stress response in foodborne pathogenic bacteria.* J Microbiol Methods, 2011. **84**(3): p. 369-78.
17. Wang, Y., et al., *Differentiation in MALDI-TOF MS and FTIR spectra between two closely related species Acidovorax oryzae and Acidovorax citrulli.* BMC Microbiology, 2012. **12**(1): p. 182.

18. Ellis, D.I., G.G. Harrigan, and R. Goodacre, *Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy*, in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, G.G. Harrigan and R. Goodacre, Editors. 2003, Springer US: Boston, MA. p. 111-124.
19. Maquelin, K., et al., *Identification of medically relevant microorganisms by vibrational spectroscopy.* J Microbiol Methods, 2002. **51**(3): p. 255-71.
20. Naumann, D., et al., *The rapid differentiation and identification of pathogenic bacteria using Fourier transform infrared spectroscopic and multivariate statistical analysis.* Journal of Molecular Structure, 1988. **174**: p. 165-170.
21. Nataro, J., et al., *Escherichia, Shigella and Salmonella. Manual of Clinical Microbiology*. 2007, SM Press, Washington DC.
22. Fukushima, M., K. Kakinuma, and R. Kawaguchi, *Phylogenetic analysis of Salmonella, Shigella, and Escherichia coli strains on the basis of the gyrB gene sequence.* Journal of clinical microbiology, 2002. **40**(8): p. 2779-2785.
23. McCleland, M., et al., *Latreille.* P., Courtney, L., Prowollik, S., Ali, J., Dante, M., Du, F. et al, 2001: p. 852-856.
24. Kamnev, A.A., et al., *Fourier Transform Infrared (FTIR) Spectroscopic Analyses of Microbiological Samples and Biogenic Selenium Nanoparticles of Microbial Origin: Sample Preparation Effects.* Molecules, 2021. **26**(4).
25. Orsi, R.H. and M. Wiedmann, *Characteristics and distribution of Listeria spp., including Listeria species newly described since 2009.* Applied microbiology and biotechnology, 2016. **100**(12): p. 5273-5287.
26. Lasch, P. and D. Naumann, *Infrared Spectroscopy in Microbiology*, in *Encyclopedia of Analytical Chemistry*. p. 1-32.
27. Romanolo, K.F., et al., *Rapid Identification and Classification of Listeria spp. and Serotype Assignment of Listeria monocytogenes Using Fourier Transform-Infrared Spectroscopy and Artificial Neural Network Analysis.* PLOS ONE, 2015. **10**(11): p. e0143425.
28. Weller, D., et al., *Listeria booriae sp. nov. and Listeria newyorkensis sp. nov., from food processing environments in the USA.* Int J Syst Evol Microbiol, 2015. **65**(Pt 1): p. 286-292.
29. den Bakker, H.C., et al., *Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss.* BMC Genomics, 2010. **11**(1): p. 688.
30. Brenner, F.W., et al., *Salmonella nomenclature.* Journal of clinical microbiology, 2000. **38**(7): p. 2465-2467.
31. Preisner, O.E., et al., *Discrimination of Salmonella enterica serotypes by Fourier transform infrared spectroscopy.* Food Research International, 2012. **45**(2): p. 1058-1064.
32. Cordovana, M., et al., *Classification of Salmonella enterica of the (Para-)Typhoid Fever Group by Fourier-Transform Infrared (FTIR) Spectroscopy.* Microorganisms, 2021. **9**(4): p. 853.
33. Reeves, P.R., et al., *Genetics and Evolution of the Salmonella Galactose-Initiated Set of O Antigens.* PLOS ONE, 2013. **8**(7): p. e69306.
34. Baldauf, N.A., et al., *Effect of selective growth media on the differentiation of Salmonella enterica serovars by Fourier-Transform Mid-Infrared Spectroscopy.* J Microbiol Methods, 2007. **68**(1): p. 106-14.
35. Baldauf, N.A., et al., *Differentiation of selected Salmonella enterica serovars by Fourier transform mid-infrared spectroscopy.* Appl Spectrosc, 2006. **60**(6): p. 592-8.

36. Fuche, F.J., et al., *Salmonella Serogroup C: Current Status of Vaccines and Why They Are Needed.* Clinical and vaccine immunology : CVI, 2016. **23**(9): p. 737-745.
37. Liu, B., et al., *Structural diversity in SalmonellaO antigens and its genetic basis.* FEMS Microbiology Reviews, 2014. **38**(1): p. 56-89.
38. Männig, A., et al., *Differentiation of Salmonella enterica serovars and strains in cultures and food using infrared spectroscopic and microspectroscopic techniques combined with soft independent modeling of class analogy pattern recognition analysis.* J Food Prot, 2008. **71**(11): p. 2249-56.
39. Campos, J., et al., *Discrimination of non-typhoid Salmonella serogroups and serotypes by Fourier Transform Infrared Spectroscopy: A comprehensive analysis.* Int J Food Microbiol, 2018. **285**: p. 34-41.
40. Davis, R., G. Paoli, and L.J. Mauer, *Evaluation of Fourier transform infrared (FT-IR) spectroscopy and chemometrics as a rapid approach for sub-typing Escherichia coli O157:H7 isolates.* Food Microbiol, 2012. **31**(2): p. 181-90.
41. Robins-Browne, R.M. and E.L. Hartland, *Escherichia coli as a cause of diarrhea.* J Gastroenterol Hepatol, 2002. **17**(4): p. 467-75.
42. Perry, M.B., L. MacLean, and D.W. Griffith, *Structure of the O-chain polysaccharide of the phenol-phase soluble lipopolysaccharide of Escherichia coli 0:157:H7.* Biochem Cell Biol, 1986. **64**(1): p. 21-8.
43. Liu, B., et al., *Structure and genetics of Escherichia coli O antigens.* FEMS microbiology reviews, 2020. **44**(6): p. 655-683.
44. Lim, J.Y., J. Yoon, and C.J. Hovde, *A brief overview of Escherichia coli O157:H7 and its plasmid O157.* Journal of microbiology and biotechnology, 2010. **20**(1): p. 5-14.
45. Canada, G.o. *Surveillance of shigellosis (Shigella)*. 2020 June 10, 2022]; Available from: https://www.canada.ca/en/public-health/services/diseases/shigella/surveillance.html.
46. Chattaway, M.A., et al., *Identification of Escherichia coli and Shigella Species from Whole-Genome Sequences.* Journal of clinical microbiology, 2017. **55**(2): p. 616-623.
47. Devanga Ragupathi, N.K., et al., *Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies.* New microbes and new infections, 2017. **21**: p. 58-62.
48. Liu, B., et al., *Structure and genetics of Shigella O antigens.* FEMS Microbiol Rev, 2008. **32**(4): p. 627-53.
49. Davis, R. and L.J. Mauer, *Fourier Transform Infrared (FT-IR) Spectroscopy: A Rapid Tool for Detection and Analysis of Foodborne Pathogenic Bacteria*. 2010. p. 1582-1594.
50. Banyay, M., M. Sarkar, and A. Gräslund, *A library of IR bands of nucleic acids in solution.* Biophys Chem, 2003. **104**(2): p. 477-88.
51. Khalil, I.A., et al., *Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990-2016.* Lancet Infect Dis, 2018. **18**(11): p. 1229-1240.
52. De Lamo-Castellví, S., A. Männing, and L.E. Rodríguez-Saona, *Fourier-transform infrared spectroscopy combined with immunomagnetic separation as a tool to discriminate Salmonella serovars.* Analyst, 2010. **135**(11): p. 2987-92.
53. Muntean, C.M., et al., *Identification of Salmonella Serovars before and after Ultraviolet Light Irradiation by Fourier Transform Infrared (FT-IR) Spectroscopy and Chemometrics.* Analytical Letters, 2021. **54**(1-2): p. 150-172.
54. Lam, L.M.T., et al., *Reagent-Free Identification of Clinical Yeasts by Use of Attenuated*

Total Reflectance Fourier Transform Infrared Spectroscopy. J Clin Microbiol, 2019. **57**(5).
55. Xu, J.L., et al., *Characterisation and Classification of Foodborne Bacteria Using Reflectance FTIR Microscopic Imaging.* Molecules, 2021. **26**(20).

Connecting Statement

FTIR spectroscopy-based method for the identification of pathogenic bacteria was developed and shown to be effective at genus-level identification using a single spectral database acquired on one ATR-FTIR spectrometer and used in the identification of isolates from spectra acquired on a second ATR-FTIR spectrometer from a different manufacturer. The effect of changes in media on the predictive accuracy of a spectral database generated from isolates grown on a single growth medium type was also assessed. Good predictive accuracy was observed at the genus-level, but limited predictive performance was observed at the species-level identification. In the next chapter, the focus shifts from these aspects of bacterial identification to fungal identification by FTIR spectroscopy, which has been much less studied in the literature. Examination of the efficacy of ATR-FTIR spectroscopy in the identification of fungal strains, including molds and yeasts, was undertaken in the following study and compared to both multiplex RT-qPCR and MALDI-TOF MS.

Chapter 5. Comparison of PCR, MALDI-TOF MS and FTIR spectroscopy for the identification of *Aspergillus* spp., and evaluation of an in-house built FTIR database for mold and yeast prediction

5.1. Abstract

Fungi are complicate in structure. Current methods for fungal identification rely heavily on mycologist using subjective morphologic methods to examine fungal colonies and fungal structure characteristics, and sometimes phenotypic methods when necessary. This poses risk to erroneous identification. Nowadays, new methods are being employed in fungi identification and have shown high prediction accuracy. Here we compared the identification accuracy of Multiplex quantitative real time polymerase chain reaction (RT-qPCR), Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS), and Fourier-transform infrared (FTIR) spectroscopy for *Aspergillus* spp. Identification. Furthermore, two different growth media and commercial libraries will be compared for MALDI-TOF MS-based identification. Finally, an in-house built FTIR spectral database of fungal strains from multiple genera was constructed and further validated. The three methods investigated yielded 71.3%, 52% and 92.3% correct identification for Multiplex RT-qPCR, MALDI-TOF MS and ATR-FTIR spectroscopy, respectively. PCR may be more suitable for identification of *Aspergillus* strains of *A. nigri* and *A. terrei*; MALDI-TOF MS was effective for identifying *A. fumigatus* and *A. flavus*; while FTIR spectroscopy correctly identified all the *Aspergillus* species investigated, and this method attained 96.6% correctness after enlarging the database with additional fungal strains. Therefore, FTIR spectroscopy can serve as a valuable technique for the identification of different fungal species.

5.2. Introduction

Due to the wide distribution in the environment, the number of yeast and mold infection in humans and contamination in food products is increasing. *Aspergillus* is well recognized as one of the most economically important genera in fungi [1]. Some *Aspergillus* species have the ability to produce mycotoxins as second metabolites that may induce carcinogenic effects in animals and humans [2]. Furthermore, *Aspergillus* can also cause a wide range of infections including cutaneous manifestations, otomycosis, and invasive infections such as pulmonary aspergillosis [3]. In addition to their pathogenic importance in medical fields, *Aspergillus* also pose serious economical issues in food, pharmaceutical, and cosmetic industries. In agriculture and the food industry, they are responsible for the spoilage of raw materials and processed foods and may cause serious health issues from the mycotoxins. Fast and accurate identification of fungi isolates is therefore important to initiate appropriate antifungal regiment.

Currently, the identification of fungi is based mainly on their macroscopic and microscopic features, in addition to phenotypic or biochemical tests when appropriate. However, these methods are often complicated by morphological divergence even among isolates of the same species. Additionally, they are time-consuming, laborious, and sometimes not accurate, and require a thorough knowledge and expertise in the morphological analysis of fungi. These drawbacks led to the exploration of new methods to obtain better and more reliable results. Polymerase chain reaction (PCR)-based techniques, Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS), and Fourier-transform infrared (FTIR) spectroscopic assays are currently being utilized or under active investigation for fungal discrimination and identification. PCR has been used as an aid in the diagnosis of invasive aspergillosis and is currently the recommended method by WHO [4]. However, a lack of standardization has limited both its acceptance as a diagnostic tool and multicenter clinical evaluations, preventing its inclusion in disease-defining criteria. Furthermore, molecular identification of filamentous fungi at the species level can be difficult and misclassification may occur. In fact, it is possible that a strain may be misclassified at the species level, especially regarding closely-related species [5]. For MALDI-TOF MS, although it has gained popularity in the clinical microbiology laboratory for the identification of bacteria and yeast species, its use for mold identification is limited to date. This is due to several factors including limited fungal entries in commercially available databases, the difficulty to obtain good quality mass spectra, and to the use of a non standardized pre-treatment

of the samples to break the fungal cell wall, which is thicker and more robust than that of bacteria [6]. In fact, the first commercial mold database consist only 89 entries corresponding to 18 different species of the genus *Aspergillus*, considering that more than 300 species have already been described.

FTIR spectroscopy could be a valuable alternative for characterization and identification of fungi. This technique is based on the measurement of fundamental molecular vibrational modes and can be used to determine the chemical composition of organic compounds. When an infrared radiation is absorbed by molecular bonds the energy absorbed by the sample results in bending, stretching, and twisting of the bonds leading to characteristic transmittance and reflectance patterns [7]. The result is presented in form of an IR spectrum, where each spectral band can be studied depending on its frequency and intensity. The overall spectral comparison showed the main functional groups from lipids, carbohydrates, nucleic acids, polysaccharides, proteins, simple sugars, phospholipids generate in part a "molecular fingerprint" of the microorganism [7]. Up to date, only few studies on the application of FTIR spectroscopy for identifying fungi are available even though promising results was reported [8]. Same as MALDI-TOF MS, the sample preparation step for FTIR spectroscopy is crucial for acquiring a reproducible infrared spectrum with acceptable quality. In fact, since fungi are not unicellular like bacteria and their spores can be easily spread in air, they need longer cultivation time and multistage sample preparation procedure to ensure safe handling, making the preparation protocol for fungi much more complex and sensitive than for bacteria and yeasts. The sample preparation used in this study for the three identification methods is after through literature search and optimization of existing protocols.

RT-qPCR, MALDI-TOF MS and FTIR-based methods aforementioned could be reliable alternatives to morphological fungi characterization. Although these techniques have been previously investigated, no comparative study has been reported yet. In this present work, our aim is to compare Multiplex RT-qPCR, MALDI-TOF MS, and FTIR spectroscopy for the identification of fungi. Furthermore, two different growth media and commercial libraries will be compared for MALDI-TOF MS. An in-house built FTIR spectral database will be enlarged to include additional fungal strains and further validated.

5.3. Material Methods

5.3.1. Fungi Preparation

Aspergillus isolates used in this study are listed in Table 5.1. All 93 *Aspergillus* isolates corresponding to 10 species (*Aspergillus aculeatus* (n = 5), *Aspergillus flavus* (n = 31), *Aspergillus fumigatus* (n = 5), *Aspergillus japonicus* (n = 3), *Aspergillus niger* (n = 20), *Aspergillus niveus* (n = 1), *Aspergillus oryzae* (n = 2), *Aspergillus parasiticus* (n = 14), *Aspergillus terreus* (n = 6), *Aspergillus uvarum* (n = 6)) came from the Canadian Collection of Fungal Cultures (DAOMC) directed by Agriculture and Agri-food Canada (AAFC) in lyophilized form. These strains were initially identified by morphology with conventional methods, including macro- and microscopic means. All fungal isolates were cultured on Sabouraud dextrose agar (SDA; BD Difco; Detroit, USA) in an incubator at 25°C for 5 days.

Table 5.1.. Fungal isolates used in this study.

Fungal species	No. of isolates	Used for ATR-FTIR	Used for RT-qPCR	Used for MALDI-TOF MS
Aspergillus aculeatus	5	5	5	2
Aspergillus flavus	31	31	31	7
Aspergillus fumigatus	5	5	5	3
Aspergillus japonicus	3	3	3	0
Aspergillus niger	20	20	19	6
Aspergillus niveus	1	1	1	0
Aspergillus oryzae	2	2	2	1
Aspergillus parasiticus	14	14	14	5
Aspergillus terreus	6	6	6	0
Aspergillus uvarum	6	6	6	1
Total	**93**	**93**	**92**	**25**

5.3.2. Multiplex real-time qPCR

Fungal DNA extraction was done by methods previously described [9-11]. Mycelia were cut from the fungal colonies with a sterile pipette tip. Three mycelia plugs were put in a 1.5 mL Eppendorf tube and were prepared in triplicate for each sample. Four hundred microliter of YPG culture media (1.5% glucose, 1% peptone, 0.5% yeast extract; w/v) supplemented with 0.5% (w/v) pre-treated sand (Cat S5631; Sigma-Aldrich Canada, Oakville, ON) were added to each sample to maximize and facilitate the grinding and collection of the mycelia. The sand was pre-treated by soaking in 50× volumes of 100 mM Tris buffer (pH 7.5) for 4 h and rinsed with water. The reaction mixture tubes were then shaken at 150 rpm at 25°C for 48 h. Finally, mycelia were pelleted after by centrifugation at 14,000 rpm for 2 min. The supernatant was removed, and the collected mycelia were kept in a -20°C freezer for less than one week.

Four hundred microliter of lysis buffer (100 mM Tris base, 50 mM EDTA, 1% SDS (w/v), 1% N-lauroyl sarcosine sodium salt (w/v) and 10 µg mL−1 RNase A; pH 8.2) were added to each Eppendorf tube containing the mycelia collected previously. The mycelium pellet is then ground with a pellet pestle driven by a cordless motor at 1000 rpm for 5 s. The Eppendorf tubes were inverted several times, and then 100 µL of potassium acetate solution (3.0 M, pH 6.5) was added. The tubes are inverted several times before centrifuging for 2 min at 14,000 rpm. Four hundred microliter aliquot of supernatant was transferred into a new 1.5 mL Eppendorf tube containing 500 µL of isopropanol, which was then inverted several times again, and centrifuged for another 2 minutes at 14,000 rpm to precipitate DNA. The supernatant was removed, and the DNA pellet was washed with 750 µL of 70% (v/v) ethanol. After centrifugation at 14,000 rpm for 1 min, the ethanol was removed, and the DNA pellet was air dried for 20 min. The DNA pellet was dissolved in 50 µL Tris-EDTA buffer (pH 8.0). A QIAamp DNA Mini spin column kit (Qiagen, Hildren, Germany) was used to purify fungal DNA. In the case that further DNA purification was required, the DNeasy PowerClean Pro CleanUp Kit (Qiagen) was used according to the manufacturer's protocol. The extracted DNA concentration was measured using Quant-iT™ PicoGreen™ dsDNA Reagent (Invitrogen) according to the manufacturer's protocol to reach a DNA concentration of 10 ng/ µL using serial dilution.

The set of primers for DNA template and hybridisation probes for *Aspergillus* spp. to run the PCR were as described previously [12] (Table 5.2), and were synthesised by Integrated DNA Technologies (Coralville, USA). Each PCR run included a negative control consisting of water without DNA template to monitor any contamination. The PCR master mix (Brilliant III Ultra-Fast qPCR Master Mix, Agilent: 600880) contains the mutant Taq DNA polymerase, dNTPs, Mg^{2+} and a buffer specially formulated for fast cycling. Each PCR assay consisted of 10 µL of master mix, 15 µM of each primer, 15 µM probe, 0.3 µL Rox reference dye (Brilliant III Ultra-Fast qPCR Master Mix, Agilent: 600880), and 1 µL of sample DNA. Ultrapure sterile water was added to a final volume of 20 µL. PCR amplification and detection of amplification was performed on a QuantStudio 5 RT-qPCR instrument (ThermoFisher Scientific). The thermal cycling conditions were conducted as follows: 10 min of pre-denaturation at 95°C followed by 40 cycles of denaturation at 95°C for 25 s, annealing at 58°C for 30 s, and extension at 72°C for 35 s for fluorescence reading. Results were analyzed using the QuantStudio 5 Design & Analysis Software (Thermo Fisher Scientific), and considered positive if they had a Ct value <40 [13].

Table 5.2. Nucleic acid sequences of primers and probes set used for multiplex real-time PCR identification in this study [12].

PCR	Target species	Primers / probes	Sequence (5′ → 3′)	Target loci	Product name	Product size	Purification	Yield guarantee
Multiplex real-time PCR primer	*Aspergillus spp*.	benAF3	TCG GTG TAG TGA CCC TTG G	*β-tubulin (benA)*	100 nmole DNA Oligo	254 ~ 272 bp	Standard desalting	35 nmoles
		benAR2	GCT GGA GCG YAT GAA CGT CT					
Hydrolysis probe	Ascomycetes	Asco1F9	/5TET/AV ACG AAG T/ZEN/T GTC GGG RC/3IABkFQ/	*β-tubulin (benA)*	100 nm PrimeTime® 5' TET™ /ZEN™/ 3' IB® FQ	18 bp	High-performance liquid chromatography (HPLC)	10 nmoles
	Section *Fumigati*	Fumi1R2	/56-FAM/CG GCA ACA T/ZEN/C TCA CGA TCT GAC TCG C/3IABkFQ/		100 nm PrimeTime® 5' 6-FAM™ /ZEN™/ 3' IB® FQ	26 bp		15 nmoles
	Section *Nigri*	Nig1R26	/56-FAM/AC TTC AGC A/ZEN/G GCT AGC GGT AAC AAG T/3IABkFQ/			26 bp		15 nmoles
	Section *Flavi*	Flavi1F18	/56-FAM/CG GTC AGG A/ZEN/G TTG CAA AGC GTT TTC A/3IABkFQ/			26 bp		15 nmoles
	Section *Terrei*	Terrei1R29	/56-FAM/AC CAT CCT G/ZEN/G GAC AGA TTC TYC ACG C/3IABkFQ/			26 bp		15 nmoles

5.3.3. MALDI-TOF MS

Identification performance of a new Conidia ID-fungi plate (IDFP) for MALDI-TOF MS was first compared with the traditional SDA growing medium. To do so, 5 *Aspergillus* spp. were randomly selected and subcultured on SDA and IDFP in parallel at 25°C for 5 days until sufficient

mycelial growth is observed. Samples were prepared and analyzed following the method described [14]. In brief, the fungal material was resuspended in a 1.5 mL Eppendorf tube with 300 µl of pure water by scraping off approximately 1 cm in diameter of the fungi from the medium plate with a sterile disposable loop and homogenized to create a turbid suspension. Samples were then mixed with 900 µl absolute ethanol (Sigma-Aldrich; Merck KGaA) and centrifuged at 14,000 rpm for 2 min. The supernatant was discarded without disturbing the pellet, and the Eppendorf tube was air dried for 5 min. Then, 50 µl of 70% formic acid was added to the pellet and vortexed. Fifty microliter of 100% acetonitrile was added after that. The suspension was subsequently centrifuged for 2 mins at 14,000 rpm. One microliter of the supernatant was spotted onto a 96-spot polished steel target plate (Bruker Daltonics) in triplicates and allowed to dry. Thereafter, 1 µl HCCA matrix (α-cyano-4-hydroxy-cinnamic acid solution in 50% acetonitrile and 2.5% trifluoroacetic acid) was pipetted onto the sample spot and dried at room temperature.

The acquisition and analysis of mass spectra was performed by a Microflex LT mass spectrometer (Bruker Daltonik GmbH) using the MALDI Biotyper software package (version 3.0). The spectra were recorded with default parameter settings, that is a positive linear mode at a laser frequency of 60 Hz within a mass range of 2000 to 20,000 Da. The acceleration voltage, extraction voltage, lens voltage and delayed extraction time were set as 20 kV, 18.5 kV, 6.0 kV and 150 ns, respectively. For each spectrum, 240 laser shots in 40-shot steps from different positions of the sample spot were accumulated and analyzed (automatic mode, default settings of MBT_AutoX method). The Bruker bacterial test standard (BTS; Bruker Daltonik GmbH) was used for calibration according to the instructions of the manufacturer.

MALDI-TOF MS spectra analysis was performed using two reference fungal libraries in a single run: (i) the Bruker Filamentous Fungi Library (Version 3.0) and (ii) the Charles River filamentous fungi MALDI library. A score between 0.00 and 3.00 was obtained, depending on the degree of similarity of a spectrum to the ones in the reference databases. The highest score of each triplicate was used. A score value $\geqslant 2$ indicates a highly reliable identification at the species level, values between 1.999 and 1.7 represent probable correct identification. Culture scraping failure resulted in spectra without peaks and required reanalysis.

After comparing the identification performance of SDA and IDFP, another 20 randomly selected Aspergillus spp. were subcultured on IDFP and underwent the same procedure for MALDI-TOF MS analysis as described above.

5.3.4. ATR-FTIR spectroscopy

Aspergillus spp. were grown on SDA at 25°C for 5 days. For each fungal strain, 3 independent agar plates were prepared to obtain 3 independent replicates. Before FTIR measurements, mycelium of each sample was transferred to a 1.5 mL Eppendorf tube with 160 µL sterile water. For deactivation of the filamentous mold, 940 µL of absolute ethanol was added to the mixture, and put aside for 2.5 h. The suspension was then centrifuged twice at 13,000 rpm for 2 min to discard the supernatant. The mycelial pellets were then spread thinly on an aluminum foil and allowed to dry overnight to form a film suitable for FTIR analysis.

The mid-infrared spectra were recorded using OMNIC™ Paradigm Software from Nicolet™ Summit (Thermo Fisher Scientific Waltham, US) FTIR spectrometer. Triplicate spectra were acquired from individual dried mycelium film by pressing the film against the ATR crystal of the FTIR spectrometer. All spectra were recorded in the region between 700 cm^{-1} and 4000 cm^{-1} in absorbance mode and Happ-Genzel apodization. For each FTIR spectrum, a total of 32 scans at a spectra resolution of 8 cm^{-1} were co-added and ratioed against a background spectrum collected from a clean ATR sampling surface.

FTIR spectra contain both biochemical information and information coming from physical effects due to light–matter interaction. The latter may introduce artifacts and large variabilities that can influence the classification model. Preprocessing of raw spectra is therefore an important step to extract important spectral information. All spectra were first visually examined to remove low quality spectra containing low water content (water absorbance <0.15). Spectra with artifacts caused by drifts in the baseline or atmospheric water vapor fluctuations were also removed. Mean spectra were calculated from the remaining spectra based on the triplicate measurements and the triplicate agar plate belonging to the same strain. Afterward, all spectra were pre-processed by to the first derivative, and vector normalized using either an in-house written software or commercially available spectral analysis software OMNIC (Thermo Scientific™, USA). Hierarchical cluster analysis (HCA) using Ward's algorithm was used to classify genus or species groups according to their proximity representing their similarity in a dendrogram (Figure 5.1).

Spectra were separated into a training and a validation set, and were exported from MATLAB (The MathWorks Inc., Natick, MA, US) as csv files using an in-house written software, and then imported directly into JMP Statistical Discovery software (Cary, NC), version 16 to perform data analysis. The relevant wavenumber ranges were narrowed down to 700-1800 cm^{-1} and 2800-3000 cm^{-1} in order to perform a forward region selection algorithm to discriminate among the species based on selected spectral features. The wavenumber combinations producing the most effective separation of classes were subsequently chosen.

The reference database for *Aspergillus* spp. was built on a multitier pairwise structure. Principal components (PCs) derived from PCA were used together with linear discriminant analysis (LDA) for validation of the database. For the validation sets, the analysis procedure was the same, except that the validation strains were labeled as "unknown" and were not employed in forward region selection or in building the PCA models. Pairwise identification at genus and species level of prediction sets was achieved by same multitier spectral database structure, and at each step identify the group in which the "unknown" spectra belong, until identification at species level is attained.

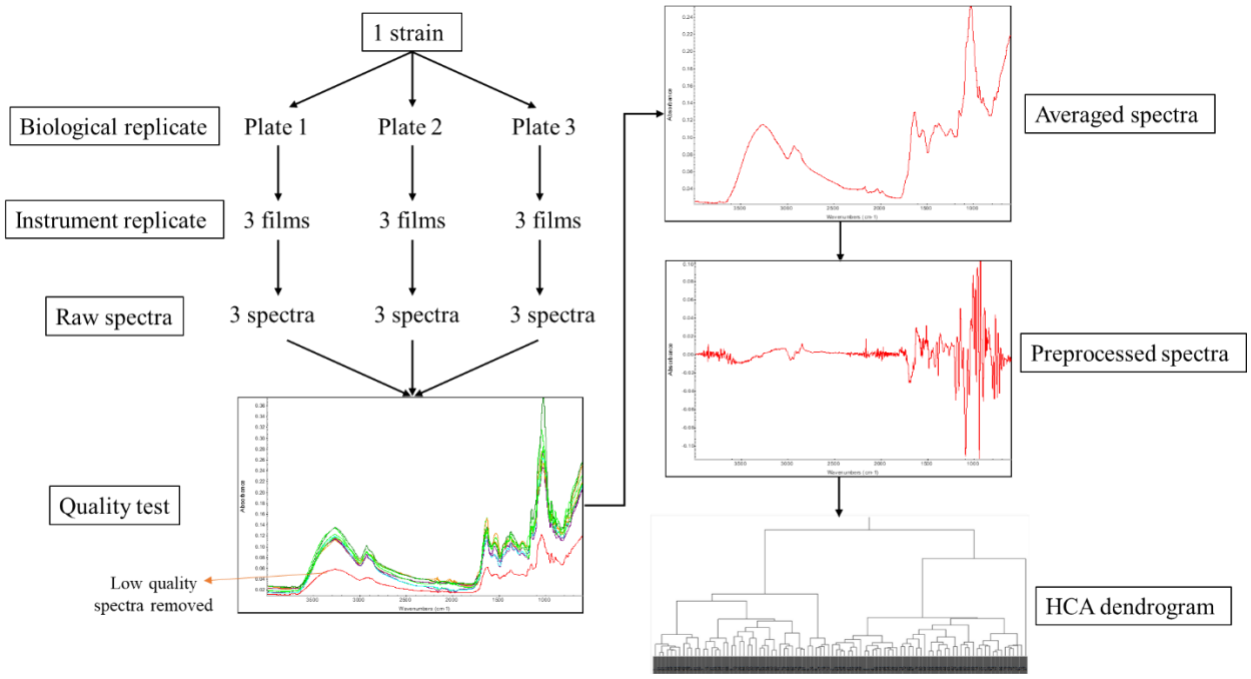


Figure 5.1. Schematic representation of sample preparation, sample analysis by FTIR spectroscopy, and spectral preprocessing.

5.3.5. Enlargement of the ATR-FTIR fungal reference database

The *Aspergillus* reference database was included 402 isolates, belonging to 13 fungal genera and 18 species (Table 5.3). The validation set, which is independent from the reference database strains, had a total of 80 isolates, including 2 *Aspergillus versicolor*, 15 *Cladosporium* spp., 11 *Geotrichum* spp., 21 *Penicillium* spp., 8 *Debaryomyces hansenii*, 5 *Issatchenkia orientalis*, 8 *Rhodotorula mucilaginosa*, 2 *Trichosporon asahii*, and 8 *Yarrowia lipolytica*. All isolates were obtained from the strain collection of Université de Laval (Québec, Canada) and Laboratoire de Santé Publique du Québec (LSPQ), and were stored in 10% glycerol at -80°C before being revived by growth on PDA. The identification of yeast and mold isolates was confirmed at LSPQ by MALDI-TOF MS and/or gene sequencing of the ribosomal DNA (rDNA) D1/D2 or internal transcribed spacer (ITS) regions (using NL1-NL4 or ITS1-ITS4 primers, respectively) and by comparing sequence similarity to that of reference sequences in GenBank, International Society of Human and Animal Mycology (ISHAM) ITS, and the Westerdjik Fungal Biodiversity Institute nucleotide databases. The identification of mold isolates was done by morphological analysis based on macroscopic and microscopic features. The identification was then confirmed by DNA sequencing. Briefly, genomic DNA was extracted using the 'FastDNA SPIN Kit' (MPBio, Illkirch, France) according to the manufacturer's instructions from mycelia grown on potato dextrose broth after 2-4 days at 25°C on a rotary shaker at 120 rpm. Depending on the fungal genus (ITS region including the 5.8S rRNA gene for all genera except *Fusarium* spp., partial β-tubulin gene for *Penicillium* and *Aspergillus* spp., and partial mcm7 and partial tsr1 genes for *Mucor* spp.), DNA amplification of 5 different regions was performed as described previously [15, 16]. After sequencing of the amplicons, the DNA sequences were compared to the GenBank database using the Basic Local Alignment Search Tool (BLAST) to determine the taxonomic assignment of fungal isolates.

Each sample was thawed and subcultured onto SDA. For non-filamentous mold or moist yeast culture, a loopful of fungal cells was isolated using a sterile disposable loop and simply deposited on the ATR diamond surface of the ATR accessory placed in the sample compartment of the FTIR spectrometer. Spectral preprocessing, statistical analysis and validation were performed as described in the previous section.

Table 5.3. List of the fungal isolates included the ATR-FTIR spectral database.

Fungal species		No. of isolates	Source
Aspergillus	*versicolor*	4	Food
Cladosporium	*cladosporioides*	5	Food
	cucumerinum	3	Food
	herbarum	2	Food
	sp	4	Food
	sphaerospermum	6	Food
Geotrichum	*candidum*	12	Food
	sp	3	Food
Mucor	*racemosus*	3	Food
Penicillium	*camemberti*	8	Food
	commune	3	Food
	roqueforti	7	Food
	sp	19	Food
Candida	*spp*	282	Clinical
Cryptococcus	*diffluens*	1	Clinical
Debaryomyces	*hansenii*	8	Food
Issatchenkia	*orientalis*	8	Food
Rhodotorula	*mucilaginosa*	9	Food and clinical
Saccharomyces	*cerevisiae*	2	Clinical
Trichosporon	*asahii*	3	Food and clinical
	cutaneum	1	Clinical
	sp	1	Clinical
Yarrowia	*lipolytica*	8	Food
	Total	**402**	-

5.4. Results & Discussion

5.4.1. Growth of *Aspergillus* spp.

After 5 days of incubation at 25°C , all *Aspergillus* isolates successfully grew on SDA and IDFP with different sizes, colors, shapes and textures, depending on the species. The phenotypic characteristics of mold varied among each other. For example, *Aspergillus flavus* (*A. flavus*) formed fluffy greyish-green colonies consisting predominantly of vegetative hyphae, while *Aspergillus niger* (*A. niger*) usually formed characteristic black colonies. *Aspergillus fumigatus* (*A. fumigatus*) was a rapid grower with typical velutinous, grey-blue-green colonies and uniseriate conidial heads. Aspergilli such as *A. flavus, A. niger,* and *Aspergillus terreus* (*A. terreus*) had similar growth rates to that of *A. fumigatus*. The rate of fungal growth was not significantly different between SDA and IDFP. In general, *Aspergillus* on IDFP grew faster, and harvesting of fungal material for the sample preparation was easier. However, the spectral quality of the inactivated thin film was not affected by the cultivation medium (Figure 5.2C). Once the filamentous mold was spread on the thin aluminum foil and dried, it can be stored safely for a long time. Figure 5.2 also shows the typical black powdery-like colony morphologies of an *A.niger* on (A) SDA and (B) IDFP plate.

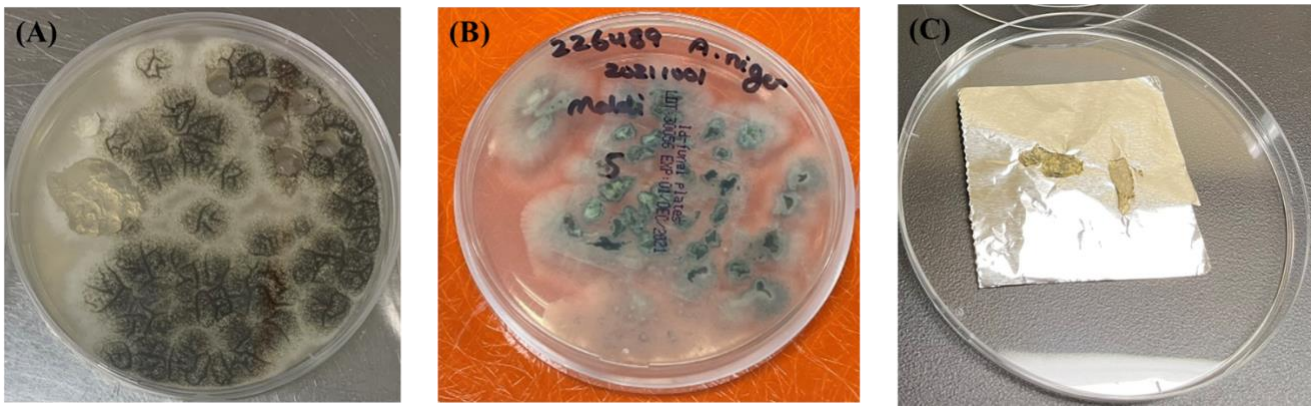


Figure 5.2. *Aspergillus niger* after 5-days of growth at 25℃ on (A) SDA and (B) IDFP, and (C) deactivated sample prepared for FTIR spectra acquisition.

5.4.2. Identification of *Aspergillus* spp. by Multiplex RT-qPCR

Multiplex real time qPCR have been widely applied for the quantification and identification of *Aspergillus* [12, 17-23]. As previously described, the PCR identification method was developed based on the *benA* gene for the four major *Aspergillus* sections (*Flavi, Fumigati, Nigri* and *Terrei*), as it reduces the diagnosis time without compromising the amplification [12].

The performance of multiplex real-time qPCR for 1167 fungal samples without monitoring the DNA concentration is reported in Table 5.4. The amplification efficiency of each probe was found to be 49.1% to 88.4%; The sensitivity and specificity were 46.5% to 100% and 41.7% to 83.3%, respectively. With a standardized DNA concentration of 10 ng/mL, the overall amplification results significantly improved, as shown in Table 5.5. The correct amplification percentage for 324 controlled fungal DNA concentration samples was 66.2% to 91.1%, whereas the sensitivity and specificity were 55.8% to 100% and 86.9% to 93.3% for *Flavi*, *Fumigati*, *Nigri* and *Terrei*, respectively. Overall, the correct amplification rate for all DNA assay was 63.2% (738/1167), and 71.3% (231/324) for standardized DNA samples. No contamination occurred in the negative control.

Table 5.4. Specific amplification of the designed PCR probes with DNA concentration not monitored (2.54-49.3 ng/µl) for the identification of all *Aspergillus* spp.

	***Aspergillus* section**			
Probe	***Fumigati***	***Nigri***	***Flavi***	***Terrei***
Ascomycetes	15/15 (100%)	102/113 (90.26%)	90/226 (39.8%)	28/30 (93.3%)
Fumi 1R2	15/15 (100%)	6/40 (15%)	36/76 (47.4%)	9/12 (75%)
Nig 1R26	5/5 (100%)	104/113 (92.0%)	34/76 (44.7%)	7/12 (58.3%)
Flavi 1F18	5/5 (100%)	7/40 (17.5%)	120/226 (53.1%)	5/12 (41.7%)
Terrei 1R29	5/5 (100%)	7/40 (17.5%)	34/76 (44.74%)	25/30 (83.3%)
Correct amplification	30/45 (66.67%)	306/346 (88.44%)	334/680 (49.1%)	68/96 (70.8%)
Incorrect amplification	15/45 (33.33%)	40/346 (11.56%)	346/680 (50.9%)	28/96 (29.2%)
Sensitivity	100%	91.15%	46.5%	88.3%
Specificity	N/A	83.33%	54.4%	41.7%

Table 5.5. Specific amplification of the designed probes with DNA concentration 10 ng/µl for *Aspergillus* spp.

	***Aspergillus* section**			
Probe	***Fumigati***	***Nigri***	***Flavi***	***Terrei***
Ascomycetes	15/15 (100%)	15/15 (100%)	30/69 (43.5%)	9/9 (100%)
Fumi 1R2	15/15 (100%)	1/5 (20%)	2/23 (8.7%)	1/3 (33.3%)
Nig 1R26	5/5 (100%)	12/15 (80%)	3/23 (13.0%)	0/3 (0%)
Flavi 1F18	5/5 (100%)	0/5 (0%)	47/69 (68.12%)	0/3 (0%)
Terrei 1R29	5/5 (100%)	0/5 (0%)	4/23 (17.4%)	6/9 (66.7%)
Correct amplification	30/45 (66.7%)	41/45 (91.1%)	137/207 (66.2%)	23/27 (85.2%)
Incorrect amplification	15/45 (33.3%)	4/45 (8.9%)	70/207 (33.8%)	4/27 (14.8%)
Sensitivity	100%	90%	55.8%	83.3%
Specificity	N/A	93.3%	86.9%	88.9%

A very interesting factor which influenced sensitivity in our study is the DNA concentration. Standardization of fungal DNA concentration yielded an amplification result improvement of 0% to 17.06% in general, with *Flavi* sections providing the highest improvement and *Fumigati* sections the lowest. Except for section *Flavi*, where a slight improvement (9.34%) in assay sensitivity (when DNA concentration was standardized to 10 ng/mL), *Nigri* and *Terrei* sections analysis showed a 1.2% and 5% reduction in sensitivity, respectively. On the other hand, specificity improved by 10% to 47.22% for all sections excluding *Fumigati*, where the qPCR results remained identical. There was also and improvement of Ascomycetes amplification for all section isolates, in which yielding 100% correct amplification at the genus level Ascomycetes

except for *Flavi*. For section *Flavi,* Ascomycetes consistently did not amplify well (<45%). Despite section *Flavi* showing the highest improvement in terms of correct amplification after the DNA adjustment, it still the lowest number of true positives compared to the other three sections. Amplification of isolates from section *Fumigati* happened for all section probes with low Ct numbers (<30), including Ascomycetes, regardless of DNA concentration. Although sensitivity was reduced, section *Nigri* and *Terrei* all had improved amplification percentage and specificity after DNA concentration standardization, especially for section *Terrei*.

Many publications on PCR-based identification for *Aspergillus* spp. were based on using different target DNA, such as 18S and 28S rRNA [17, 19], nuclear ribosomal internal transcribed spacer (ITS) regions [23], *benA* and *rodA* [12, 22]. Sequences in 18S or 28S rRNA regions are conserved across a wide range of fungi and have been used to detect fungal pathogens in clinical specimens [17], yet, it is difficult to design truly species-specific primers using these regions. The more variable ITS regions may be more useful for identification of fungal species as it is not only conserved, but is also present as multiple copies in the fungal genome, yielding sufficient taxonomic resolution for most fungi, and has the advantage that GenBank and other universal large genome database contain a large number of sequences from this locus, facilitating identification of the sequence from an unknown isolate. More recent studies suggest that probe-based RT-qPCR can be more sensitive, specific, and fast compared to molecular identification based on ITS sequencing [12]. Moreover, genetic analysis of *benA* is useful to distinguish cryptic species, which is not possible with ITS sequencing [18], hence the reason that *benA* gene was used to select a primer pair and probe specific to the major *Aspergillus* sections. In general, compared to the amplification results of this study, higher PCR identification performance for *Aspergillus* spp. has been reported using all the DNA regions aforementioned [12, 19-21, 23]. It is worth mentioning that the larger number of samples with false-negative or false-positive results in this study may be due to the large set of isolates used. It has been reported that more than 80% of fungal isolates can be isolated only once when tested at the molecular level [24], and that some cross-reactions exist when large number of samples are tested due to the greater diversity of fungi [25]. Some studies also encountered problems when using PCR for fungal identification, especially for species differentiation within section *Nigri* [26] and *A. fumigatus* [27] due to contamination. Another study has found that only 50-60% of the molecular identification results were concordant with phenotypic methods, and this lack of concordance could be attributed to the incompleteness of the

sequence database [5]. As a matter of fact, although independent meta-analysis showed comparable identification performance with regards to using different biomarkers and commercial assays, there is methodological diversity with respect to PCR identification of fungal isolates. In other words, there is currently no standard preparation method of *Aspergillus* spp. DNA isolates for multiplex RT-qPCR. Therefore, the comparability of PCR performance for *Aspergillus* spp. among studies remains questionable.

Design of primers and selection of probes are very important as they determine the specificity and sensitivity of PCR-based methods. A reliable method of DNA extraction is also crucial for PCR amplification, and several papers have reported the progress in improving the extraction step and the need to address inhibitors of PCR that could not be inactivated resulting in reduced performance of the PCR tests [23]. An optimal standard method for fungal PCR protocol may provide superior performance, as suggested in this work when fungal DNA concentration was standardized. Nevertheless, this study showed promising results for identifying section *Nigri* and *Terrei* using multiplex RT-qPCR.

5.4.3. Identification of *Aspergillus* spp. by MALDI-TOF MS

A total of 5 *Aspergillus* isolates, including 2 *A. flavus*, 2 *A. niger* and 1 *A. parasiticus* were grown on SDA and IDFP for the comparison of MALDI-TOF MS identification score (Table 5.6). Interpretation criteria for MALDI-TOF MS was applied with minor modification. Briefly, a score value > 1.400 was sufficient to consider as a probable correct identification, and score value $\leqslant$ 1.400 signifies no identification (failed identification). Among the 5 samples inoculated in SDA, three (60%) of them could not be analyzed as no peak was found, and one (20%) was not identified due to unreliable score. At the end, only a single isolate (20%) could be correctly identified on SDA regardless of Bruker or Charles River database with comparable MALDI score. With regards to the isolates grown on IDFA, unlike SDA with 3 isolates (60%) failed to have peaks resolved, all isolates obtained a prediction result from the two databases. The Bruker database correctly identified 1 isolate (20%), with 1 misidentified (20%) and 3 non-identified (60%); whereas the Charles River database had 4 isolates (80%) correctly identified and only 1 (20%) misidentified. With the preliminary results for the comparison of the two growth media, we noticed that IDFP clearly enhanced the performance of *Aspergillus* identification compared to SDA for both Bruker

and Charles River database. Therefore, IDFP was selected as the growth medium for the rest of MALDI-TOF MS identification.

As shown in Table 5.7, out of the 25 *Aspergillus* isolates, the Bruker and Charles River have obtained a correct identification of 24% (6/25) and 52% (13/25), a misidentification rate of 32% (8/25) and 24% (6/25), and a non-identification rate of 44% (11/25) and 24% (6/25), respectively. Species identification was able to achieve 42.89% (6/14) using Bruker and 68.42% (13/19) by Charles River database by excluding the non-identification rate. The number of non-identification was remarkably higher for Bruker (44%) database compared to Charles River (24%), as well as the number of misidentification (32% Bruker and 24% Charles River. Based on both databases, *A. fumigatus* obtained the highest identification rate (100%), despite the score value for all 3 *A. fumigatus* was significantly higher in Charles River database (> 2.000) than Bruker (1.57-1.88). For *A. flavus* (n = 7), the number of correct identifications were considerably larger for Charles River (85.71%) compared to Bruker (28.57%). Only a single *A. flavus* could not be identified (14.29%) in Charles River database, whereas there were three non-identified (42.86%) and two misidentified (28.57%) using the Bruker database. Correct identification rate of 16.67% was obtained for *A. niger* (n = 6) by both databases. However, a single isolate that could not be identified using the Bruker database, but correctly identified at genus level using the Charles River database. Unfortunately, it did not reach species identification by misidentifying *A. niger* as *A. tubingensis*. The one *A. oryzae* was misidentified by the Bruker database but was correctly identified by Charles River database with high score (1.724). For the case of *A. parasiticus* (n = 5), all of them were either misidentified (60%) or non-identified (40%) by the Bruker database. Interestingly, with the same number of misidentifications (60%), the two *A. parasiticus* that were not identified by the Bruker database were correctly identified (40%) by Charles River database. For *A. aculeatus* (n = 2) and *A. uvarum* (n = 1), both databases yielded the same results with similar MALDI-TOF scores. Table 5.9 contains the MALDI-TOF MS identification scores and results of each isolate based on the two databases for a clear comparison.

In general, Charles River database generated a higher percentage of correct identification and lower misidentification rate compared to Bruker. This result is not surprising as relatively fewer species of *Aspergillus* are represented in the Bruker database compared to Charles River database [28]. Furthermore, according to the official method for fungi identification by Bruker,

the sample preparation method was slightly modified in the present study, with the main adjustment being the use of IDFP instead of SDA [29]. Still, comparing *Aspergillus* identification on SDA and IDFP, the latter showed higher identification rate using both databases. Indeed, IDFP is the commercial plate manufactured by Charles River for MALDI-TOF MS identification of filamentous fungi.

Despite the relatively higher identification rate by Charles River database (68.42%), this number is far less than satisfactory. Identification of *Aspergillus* is rarely performed as routine tests in most microbiology laboratories, and even less common at species level. This may explain the considerably lower number of reference spectra for fungi (247 species) in commercial MALDI-TOF MS databases compared to bacteria (3893 species), which in turn reflect partially the low identification rate for *Aspergillus* spp. [30]. Additionally, although SDA is the suggested cultivation media by Centers for Disease Control and Prevention (CDC) for MALDI-TOF MS fungal identification, the Bruker filamentous fungi database library is created with mycelia from liquid broth culture. Noteworthy, Charles River database is built based on Bruker database but with additional to in-house reference spectra (Accugenix®). The relatively weak identification capacity of MALDI-TOF MS may be attributed to the different elements grown under the two culture methods, as short-term broth culture mainly yields mycelia while agar culture generates abundant conidia and hyphae. Since young colonies and mature colonies of the identical mold isolate could present some obvious differences in spectra, a shorter incubation time of about 2 days (instead of 5 days) and sampling of the front hyphea from the young colonies may be favorable for the identification of *Aspergillus* spp. grown on agar media by MALDI-TOF MS [31].

Results on identification of *Aspergillus* spp. by MALDI-TOF MS have large discrepancies in the literature. While some studies show promising identification capacity of 77.78% to 90.5% [32-34], many more have reported similar identification efficiency as the present study 54.2% to 69% [35-38]. These latter studies established in-house databases by expanding the commercial database with additional reference spectra and improved the identification performance of MALDI-TOF MS remarkably up to 93%. Other than changing the culture medium and optimizing the comprehensive databases, protein extraction method could also be employed in MALDI-TOF MS identification of filamentous fungi. One study applied bead grinding procedure with the extraction solution adding formic acid and acetonitrile in one step, and have shown to increase the

identification rates [39]. Overall, the use of commercially available filamentous fungal spectral libraries provides a low percentage of correct identification for *Aspergillus* spp. Expansion of the MALDI-TOF MS reference database will be necessary to improve the identification performance of mold genera as a whole.

Table 5.6. Comparison of growing medium SDA and IDFP for identification performance by MALDI-TOF MS.

Growth medium	SDA				IDFP			
MALDI-TOF MS database	**Bruker**		**Charles River**		**Bruker**		**Charles River**	
Species ID	Result	Score[1]	Result	Score	Result	Score	Result	Score
A. flavus (215373)	No peaks found	-	No peaks found	-	Myroides odoratus	1.36	*A. flavus*	1.886
A. flavus (215370)	*A. flavus*	1.76	*A. flavus*	1.756	*A. flavus*	1.79	*A. flavus*	1.793
A. niger (226489)	No peaks found	-	No peaks found	-	Lactobacillus paralimentarius	1.44	*A. niger*	1.783
A. niger (211079)	*Lacto-bacillus curvatus*	1.31	*Lacto-bacillus curvatus*	1.306	*Staphylo-coccus lutrae*	1.36	*Penicillium aurantioviolaceum*	1.684
A. parasiticus (239372)	No peaks found	-	No peaks found	-	Cryptococcus neoformans	1.33	*A. parasiticus*	1.592
Correct identification	**1/5 (20%)**		**1/5 (20%)**		**1/5 (20%)**		**4/5 (80%)**	
Incorrect identification[2]	**0 (0%)**		**0 (0%)**		**1/5 (20%)**		**1/5 (20%)**	
No identification[3]	**4/5 (80%)**		**4/5 (80%)**		**3/5 (60%)**		**0/5 (0%)**	

[1]For MALDI-TOF MS, a score value ≥2.000 indicates a reliable species identification, values between 1.999 and 1.700 represent probable correct identification
[2]Incorrect identification signifies that the MALDI-TOF MS results predicted wrongly, and the score was >1.400.
[3]No identification signifies that MALDI-TOF MS results were ≤1.400.

Table 5.7. MALDI-TOF MS Identification Results of *Aspergillus* spp. based on Charles River Database.

		Bruker database			Charles River database		
***Aspergillus* spp.**	**Number of isolates**	**Correctly identified**	**Mis-identified**	**Non-identified**	**Correctly identified**	**Mis-identified**	**Non-identified**
aculeatus	2	0/2 (0%)	0/2 (0%)	2/2 (100%)	0/2 (0%)	0/2 (0%)	2/2 (100%)
flavus	7	2/7 (28.57%)	2/7 (28.57%)	3/7 (42.86%)	6/7 (85.71%)	0/7 (0%)	1/7 (14.29%)
fumigatus	3	3/3 (100%)	0/3 (0%)	0/3 (0%)	3/3 (100%)	0/3 (0%)	0/3 (0%)
niger	6	1/6 (16.67%)	1/6 (16.67%)	4/6 (66.66%)	1/6 (16.67%)	2/6 (33.33%)	3/6 (50%)
oryzae	1	0/1 (0%)	1/1 (100%)	0/1 (0%)	1/1 (100%)	0/1 (0%)	0/1 (0%)
parasiticus	5	0/5 (0%)	3/5 (60%)	2/5 (40%)	2/5 (40%)	3/5 (60%)	0/5 (0%)
uvarum	1	0/1 (0%)	1/1 (100%)	0/1 (0%)	0/1 (0%)	1/1 (100%)	0/1 (0%)
Total	25	6/25 (24%)	8/25 (32%)	11/25 (44%)	13/25 (52%)	6/25 (24%)	6/25 (24%)
Correct percentage[1]		**6/14 (42.89%)**			**13/19 (68.42%)**		

[1]The correct identification percentage calculated did not account the non-identified isolates.

5.4.4. Identification if *Aspergillus* spp. by ATR-FTIR spectroscopy

ATR-FTIR spectral classification models for 54 *Aspergillus* spp. strains belonging to 10 species were generated using PCA-DA. The IR spectra were acquired in triplicate and subjected

to second derivative transformation and vector normalization prior to spectral analysis. A total of 54 averaged spectra were obtained after removing obvious outliers or low-quality spectra. Each of the averaged spectrum provide information of the similarities and differences in the biomolecular composition among isolates, and this can be visualized in peak shifts and relative intensity differences. Comparison of the 4 *Aspergillus* sections and their corresponding *Aspergillus* species are shown in Figure 5.3 and 5.4, respectively. Wavenumber regions 700-1800 cm^{-1} and 2800-3000 cm^{-1} were selected for the classification model development as these regions are known to be characteristic to biomolecules. The spectral regions are tentatively been assigned by others as the fingerprint region (700-900 cm^{-1}), polysaccharides (1200-900 cm^{-1}), proteins, lipids, and phosphate compounds (1500-1200 cm^{-1}), amides (1700-1500 cm^{-1}), and lipids (3000-2800 cm^{-1}) [7]. It is clear that regardless of the overall similarities between spectra of *Aspergillus* spp., unique spectral differences exist between species and sections. Based on the classification results obtained through the use of spectral search algorithms and HCA, each *Aspergillus* section was assigned into distinct groups to enable pairwise discrimination down to the species level.

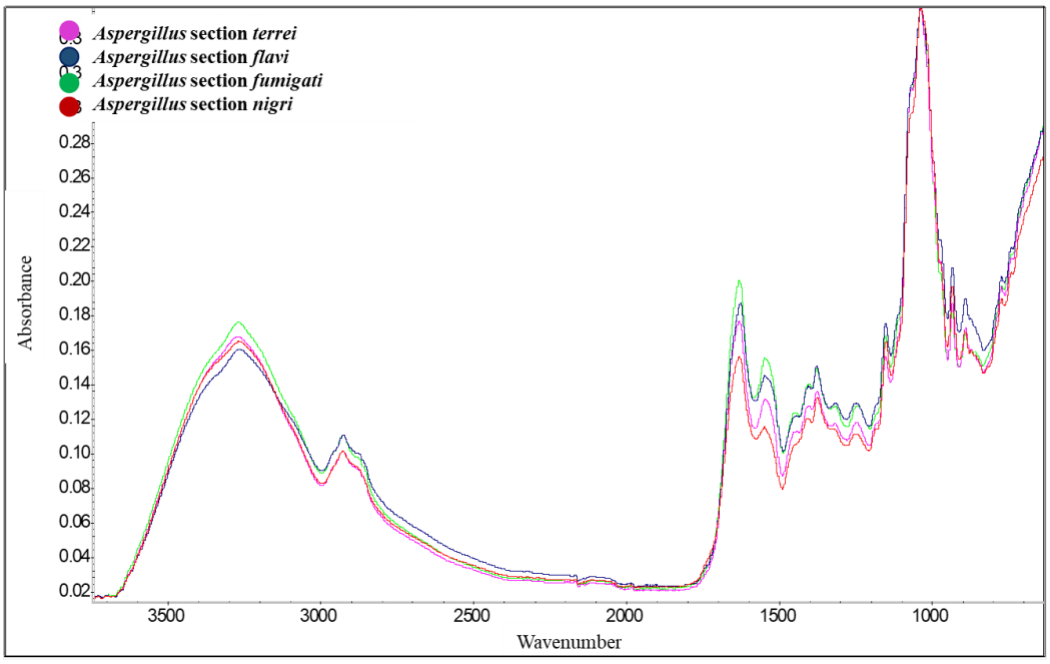


Figure 5.3. Superposition of averaged FTIR spectra of *Aspergillus* sections.

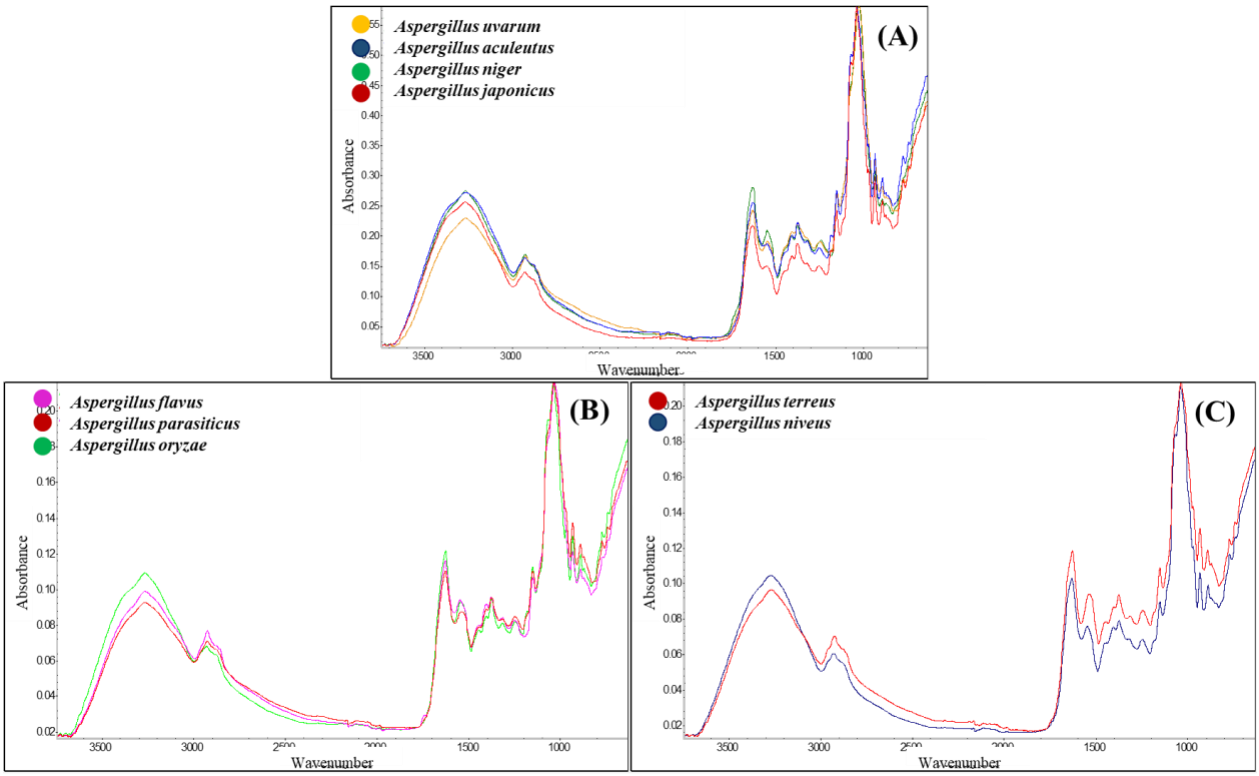


Figure 5.4. Superposition of averaged FTIR spectra of *Aspergillus* spp. in section (A) *Nigri*, (B) *Flavi*, and (C) *Terrei*. (Section *Fumigati* is not shown as it comprises only *A. fumigatus* in this study).

The database validation was carried out using an 'unknown' validation set composed of 39 *Aspergillus* strains belonging to 9 species, totalling 39 spectra after averaging the triplicates. To facilitate the comparison of identification performance with other methods involved in this study, the 39 validation strains included the 25 strains used for MALDI-TOF MS identification. Species of all these strains were represented in the database and were excluded when constructing the database. Identical spectral preprocessing procedure was used as for the reference spectra when building the database was performed, and pairwise identification down to species level was done for each of the validation spectra with a confidence percentage for each prediction result (Figure 5.5). To be considered as a reliable identification, the prediction confidence must be $\geqslant$0.8000 ($\geqslant$ 80%); if the percentage of prediction confidence is <0.8000 (< 80%), the prediction is reported as misidentified or unidentifiable. The identification result is reproted in Table 5.8. 36 strains out of 39 (92.3%) were correctly identified; only 2 strains (5.1%) were misidentified and a single strain (2.6%) as non-identified. Concerning the two misidentified strains (*A. japonicus* and *A. flavus)*, *A. japonicus* was identified as *A. aculeatus* with 98% confidence, and *A. flavus* as *A. parasiticus* with

high confidence (97.5%). Both isolates were misidentified within their same *Aspergillus* section. Based on the DNA sequence, *A. japonicus* and *A. aculeatus* were shown to be more closely related to each other than to other species within the section *Nigri* from DNA sequence [40]. With regards of *A. flavus* and *A. parasiticus*, both are capable to produce aflatoxin, which is implicated as acute toxicants and hepta-carcinogens in human [41]. Hence, identifying either one or the other will bring out comparable anti-fungal treatment. For the non-identified *A. parasiticus*, although it was predicted correctly to species level, the low confidence percentage (60.51%) caused it to be considered as a failed identification.

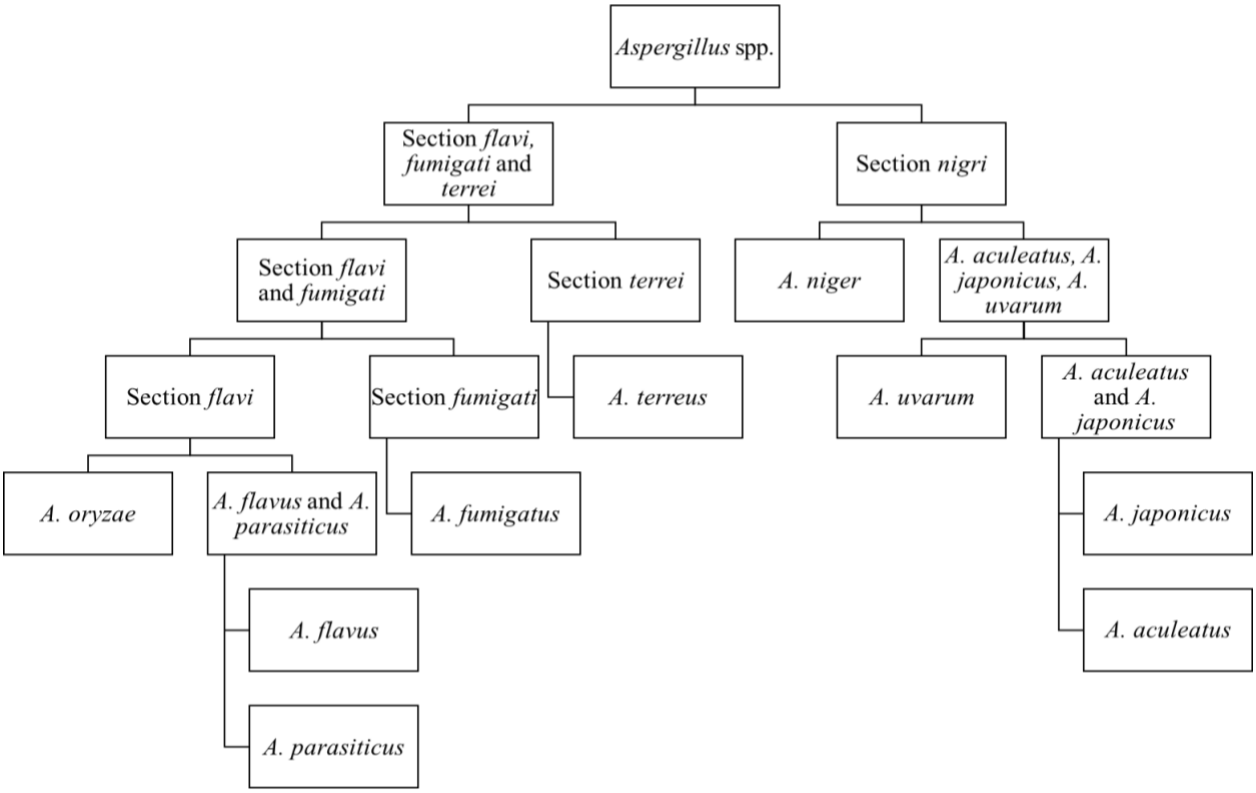


Figure 5.5. Identification flow chart of *Aspergillus* spp.

Table 5.8. ATR-FTIR-based identification results of *Aspergillus* spp.

***Aspergillus* species**	**In database**	**For prediction**	**Correct identification[1]**	**Misidentified[2]**	**Non-Identified[3]**
aculeatus	3	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
flavus	20	11	10/11 (90.9)	1/11 (9.1%)	0/11 (0%)
fumigatus	1	4	4/4 (100%)	0/4 (0%)	0/4 (0%)
japonicus	2	1	0/1 (0%)	1/1 (100%)	0/1 (0%)
niger	11	9	9/9 (100%)	0/9 (0%)	0/9 (0%)
niveus	1	0	-	-	-
oryzae	1	1	1/1 (100%)	0/1 (0%)	0/1 (0%)
parasiticus	7	7	6/7 (85.7%)	0/7 (0%)	1/7 (14.3%)
terreus	4	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
uvarum	4	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
Total	**54**	**39**	**36/39 (92.3%)**	**2/39 (5.1%)**	**1/39 (2.6%)**

[1]Correct identification signify that the percentage of correct prediction confidence must be ≥0.8000.
[2]Misidentification signify that the percentage of incorrect prediction confidence must be ≥0.8000.
[3]No identification signify that the percentage of prediction confidence result is <0.8000.

The correct classification of *Aspergillus* spp. at species level by ATR-FTIR spectroscopy is reported in this study. Classification models yielded a 92.3% identification down to the species level using an in-house built ATR-FTIR spectral database. Findings in this study were in agreement with transmission based FTIR spectroscopy studies for *Aspergillus* spp. identification. Lecellier et al. and Shapaval et al. developed their own filamentous fungi FTIR spectral database comprising *Aspergillus* spp., achieving a correct assignment of 94% to 99.17% and 92.3% to 98.77% at the genus and species level, respectively [42-44]. Another study has also revealed that FTIR spectroscopy was able to differentiate non-toxigenic and toxigenic *A. flavus* and *A. parasiticus* isolates with a correct classification rates of 75% and 100%, respectively [45]. Similar to the observation in this present study, Bhat R. analyzed *Aspergillus* FTIR spectra of *Aspergillus* species and found that a single and clearly distinguishable peak could be sufficient to differentiate between two different fungal species [8]. Overall, this study demonstrated good performance of ATR-FTIR spectroscopy for *Aspergillus* species identification using a spectral library of molds as compared to transmission-based FTIR methods.

5.4.5. Identification performance comparison of the three methods

Comparison of Multiplex RT-qPCR, MALDI-TOF MS (using both databases) and ATR-FTIR spectroscopy for *Aspergillus* spp. is summarized in Table 5.9. For ease of comparison, only the identification results of the 25 strains used for MALDI-TOF MS are shown for PCR and FTIR spectroscopy. In brief, PCR correctly amplified 68.7% (69/99), MADI-TOF MS attained 42. 9%

(6/14) by use of the Bruker database and 68.4% (13/19) by using the Charles River database, while FTIR spectroscopy yielded a 100% (24/24) correct identification when the non-identified strains are excluded. *A. aculeatus* (n = 2) could not be identified (score < 1.400) from both databases in MALDI-TOF MS, yet was amplified correctly in general by PCR. Same results were observed for *A. niger* (n = 6) and *A. uvarum* (n = 1), where MALDI-TOF MS only identified one *A. niger* isolate by Charles River database, whereas PCR correctly amplified all of them including replicate assays. Despite correct amplification of all 3 *A. fumigatus* isolates by PCR, MALDI-TOF MS may still provide as a reliable identification result as the PCR assays for all section probes (low specificity) as shown in Table 5.5. Similarly, MALDI-TOF MS also correctly identified *A. oryzae* (n = 1) and *A. flavus* (n = 7) with high score by Charles River compared to PCR. ATR-FTIR spectroscopy correctly identified all isolates aforementioned with high confidence. While PCR and MALDI-TOF MS had trouble amplifying or identifying all 5 *A. parasiticus* isolates, only one *A. parasiticus* was non-identified by FTIR spectroscopy due to low confidence (60.5%).

When comparing the identification of the three methods by *Aspergillus* section as illustrated in Table 5.10 for the 25 isolates, MALDI-TOF MS yielded the lowest performance of 52% (13/25), followed by Multiplex RT-qPCR of 68.7% (68/99), and the ATR-FTIR-based method attained the highest correct identification of 96% (24/25). MALDI-TOF MS could be reliable for identification of isolates belonging to section *Fumigati* (100%), and relatively poor for section *Flavi* (69.2%) and section *Nigri* (11.1%). The reason may be attributed to the need of improvement for the filamentous fungi infrared spectral library, or the method of cultivation as stated in previous section. In comparison, PCR performed better than MALDI-TOF MS for the identification of section *Nigri* with correct amplification of 96.6%; and for section *Flavi* (50.82%); and lower rate of identification for section *Fumigati* due to the low specificity despite 100% amplification of the *Fumigati* section probe. ATR-FTIR-based method performed the best for the identification of all *Aspergillus* spp. section compared to RT-qPCR, and MALDI TOF MS, with 100% correctness in section *Fumigati* and *Nigri*, and 92.3% for section *Flavi*.

To the best of our knowledge, although many studies have achieved comparable results with the identification methods used in this study, no publication has done the identification performance comparison of Multiplex RT-qPCR, MALDI-TOF MS and FTIR spectroscopy for *Aspergillus* spp. In this study, ATR-FTIR spectroscopy coupled with PCA-DA has demonstrated

species-specific discrimination and identification of filamentous fungi with high rate of correct identification compared to RT-qPCR and MALDI-TOF MS. Current morphological analysis for filamentous fungi identification could be difficult because of the very high phenotypic biodiversity. Furthermore, as shown in this study, the use of molecular approaches and MALDI-TOF MS identification of *Aspergillus* spp. is limited due to the cost of instruments, time, reagents, and the identification accuracy. ATR-FTIR spectroscopy represents a rapid, accurate and cost-effective *Aspergillus* species identification technique when using an appropriate well-represented spectral database.

Table 5.9. Identification performance comparison of Multiplex RT-qPCR, MALDI-TOF MS and FTIR spectroscopy of 25 *Aspergillus* isolates.

Species ID	Multiplex RT-qPCR	Bruker (MALDI-TOF MS)		Charles River (MALDI-TOF MS)		FTIR	
	Amplification result[1]	Result	Score	Result	Score	Result	Prediction Probability
***A. aculeatus* (188310)**	5/6 (83.33%)	*Lactobacillus paracasei spp paracasei*	1.34	*Lactobacillus paracasei*	1.338	*A. aculeatus*	93.92%
***A. aculeatus* (190937)**	3/3 (100%)	*Lactobacillus paracasei spp paracasei*	1.34	*Lactobacillus paracasei*	1.34	*A. aculeatus*	95.07%
***A. flavus* (147046)**	4/6 (66.67%)	*A. flavus*	1.66	*A. flavus*	1.846	*A. flavus*	100.00%
***A. flavus* (164836)**	3/6 (50%)	*Staphylococcus cohnii ssp cohnii*	1.43	*A. flavus*	1.646	*A. flavus*	85.65%
***A. flavus* (215370)**	3/3 (100%)	*A. flavus*	1.79	*A. flavus*	1.793	*A. flavus*	99.73%
***A. flavus* (215373)**	1/6 (16.67%)	*Myroides odoratus*	1.36	*A. flavus*	1.886	*A. flavus*	99.31%
***A. flavus* (225949)**	3/3 (100%)	*Lactobacillus plantarum*	1.38	*A. flavus*	1.5	*A. flavus*	91.91%
***A. flavus* (237618)**	1/3 (33.33%)	*Clostridium novyi*	1.39	*Clostridium novyi*	1.387	*A. flavus*	97.47%
***A. flavus* (239379)**	0/3 (0%)	*Kytococcus sedentarius*	1.42	*A. flavus*	1.594	*A. flavus*	99.16%
***A. fumigatus* (15734)**	3/3 (100%)	*A. fumigatus*	1.57	*A. fumigatus*	2.004	*A. fumigatus*	90.71%
***A. fumigatus* (215394)**	3/3 (100%)	*A. fumigatus*	1.84	*A. fumigatus*	2.06	*A. fumigatus*	88.32%
***A. fumigatus* (226481)**	3/3 (100%)	*A. fumigatus*	1.88	*A. fumigatus*	2.067	*A. fumigatus*	97.27%
***A. niger* (144018)**	3/3 (100%)	*A. niger*	1.53	*A. niger*	1.528	*A. niger*	100.00%
***A. niger* (160593)**	3/3 (100%)	*Rhizobium radiobacter*	1.26	*Agrobacterium radiobacter*	1.265	*A. niger*	100.00%
***A. niger* (191282)**	3/3 (100%)	*Clodostridium beijerinckii*	1.39	*Clodostridium beijerinckii*	1.387	*A. niger*	100.00%
***A. niger* (211079)**	3/3 (100%)	*Staphylococcus lutrae*	1.36	*Penicillium aurantioviolaceum*	1.684	*A. niger*	99.51%
***A. niger* (221143)**	2/2 (100%)	*Lactobacillus oligofermentans*	1.39	*Lactobacillus oligofermentans*	1.386	*A. niger*	100.00%
***A. niger* (226489)**	3/3 (100%)	*Lactobacillus paralimentarius*	1.44	*Aspergillus tubingensis*	1.783	*A. niger*	97.70%
***A. oryzae* (221144)**	1/3 (33.33%)	*Lactobacillus aviaries ssp aviarius*	1.41	*A. oryzae*	1.724	*A. oryzae*	93.70%

***A. parasiticus* (216343)**	5/7 (71.43%)	*Lactobacillus sakei*	1.51	*Lactobacillus sakei*	1.508	*A. parasiticus*	60.51%
***A. parasiticus* (221063)**	0/3 (0%)	*Actinocorallia libanotica*	1.6	*Actinocorallia libanotica*	1.595	*A. parasiticus*	100.00%
***A. parasiticus* (238957)**	3/6 (50%)	*Lactobacillus reuteri*	1.37	*A. parasiticus*	1.476	*A. parasiticus*	99.55%
***A. parasiticus* (239372)**	3/6 (50%)	*Cryptococcus neoformans*	1.33	*A. parasiticus*	1.592	*A. parasiticus*	99.97%
***A. parasiticus* (239391)**	4/6 (66.67%)	*Lactobacillus malefermentans*	1.41	*Aspergillus minisclerotigenes*	1.42	*A. parasiticus*	99.63%
***A. uvarum* (250032)**	3/3 (100%)	*Paeniclostridium sordellii*	1.4	*Paeniclostridium sordellii*	1.402	*A. uvarum*	90.23%
Correct identification[2]	**68/99 (68.69%)**		**6/14 (42.89%)**		**13/19 (68.42%)**		**24/24 (100%)**

[1]This category only shows the amplification results for the section-specific probe.
[2]The correct identification percentage calculated did not account the non-identified isolates.

Table 5.10. Identification performance comparison of 25 *Aspergillus* isolates by Multiplex RT-qPCR, MALDI-TOF MS and FTIR spectroscopy.

Section	Multiplex RT-qPCR	MALDI-TOF MS (Charles River)	FTIR
Flavi	31/61 (50.82%)	9/13 (69.23%)	12/13 (92.31%)
Fumigati	9/9 (100%)	3/3 (100%)	3/3 (100%)
Nigri	28/29 (96.55%)	1/9 (11.11%)	9/9 (100%)
Correct identification[1]	**68/99 (68.69%)**	**13/25 (52%)**	**24/25 (96%)**

[1]The correct identification percentage calculated accounted the non-identified isolates.

5.4.6. Expansion of the in-house built ATR-FTIR spectral database and its validation as a method for fungal identification

The *Aspergillus* ATR-FTIR spectral database was enlarged to include additional non-*Aspergillus* mold and yeast strains in order to evaluate the robustness of ATR-FTIR spectroscopy as a generalized method for fungal identification. Including the 54 *Aspergillus* reference strains, a total of 456 fungal isolates were included in the new database, comprising 29 species from 13 genera. The enlarged spectral database was used to generate by HCA and PCA-DA classification models. Discriminant spectral features were selected for each taxonomic rank (genus and species level) to allow stepwise classification of the strains to a specific genus to species. The performance of the classification models we assessed by using a validation set of 119 fungal strains from 23 species and 13 genera. The final identification flow for the validation set is shown in Figure 5.6. The identification rate of fungi and yeast spp. by FTIR spectroscopy were 96.6% (115/119), 2.52% (3/119) and 0.84% (1/119) for correct identification, misidentification, and non-identification, respectively (Table 5.11).

The prediction errors were associated with the mold strains and all yeast isolates were correctly identified (100% correct identification). The non-identified was *A. parasiticus* and two of the three misidentified spectra were *A. flavus* and *A. japonicus*, the same strains that were problematic in the previous *Aspergillus* database. Although the newly built database was enlarged with more fungal strains, no *Aspergillus* spp. except *Aspergillus versicolor* was added. And since identification of unknown was based on the stepwise classification model, the unchanging reference spectral group of *Aspergillus* species will deliver identical identification result. Adding new reference spectra to the database to include maximal intra-species diversity for identification

of troublesome species would be helpful to improve the identification performance of ATR-FTIR spectroscopy. In a previous study performed by our team member, identification rate was increased from 95.6% to 99.7% for *Candida* spp. when the database was enlarged with reference strains belonging to the same species as the previously misidentified ones [46]. Other than *Aspergillus*, only one *Cladosporium herbarum* was misidentified as *Geotrichum* spp.

Conventional identification methods for fungi in routine are carried out mostly by morphological methods, that are mainly time-consuming and require highly trained personnel. Based on this study, ATR-FTIR spectroscopy combined with multivariate analysis methods is a promising technique to identify filamentous fungi. FTIR spectroscopy is a high throughput technique and is less expensive than traditional morphological and molecular methods, providing a good alternative for routine analysis. Limitation of using ATR-FTIR-spectroscopy as identification method exists. For instance, signal-to-noise ratio of the spectra, spectral reproducibility, sampling methods, variability in growth media and incubation can have varying impact on the identification of an unknown isolate. Furthermore, the in-house ATR-FTIR spectral database built is far from being sufficient with respect to the diversity of fungal species encountered in the general routine laboratory analysis, and no ATR-FTIR spectral library exists for fungal strain identification to date. A standardized protocol for ATR-FTIR spectroscopy for fungal identification is needed in order to be considered as a valuable diagnostic tool in clinical, food and agricultural settings for fungi and yeast.

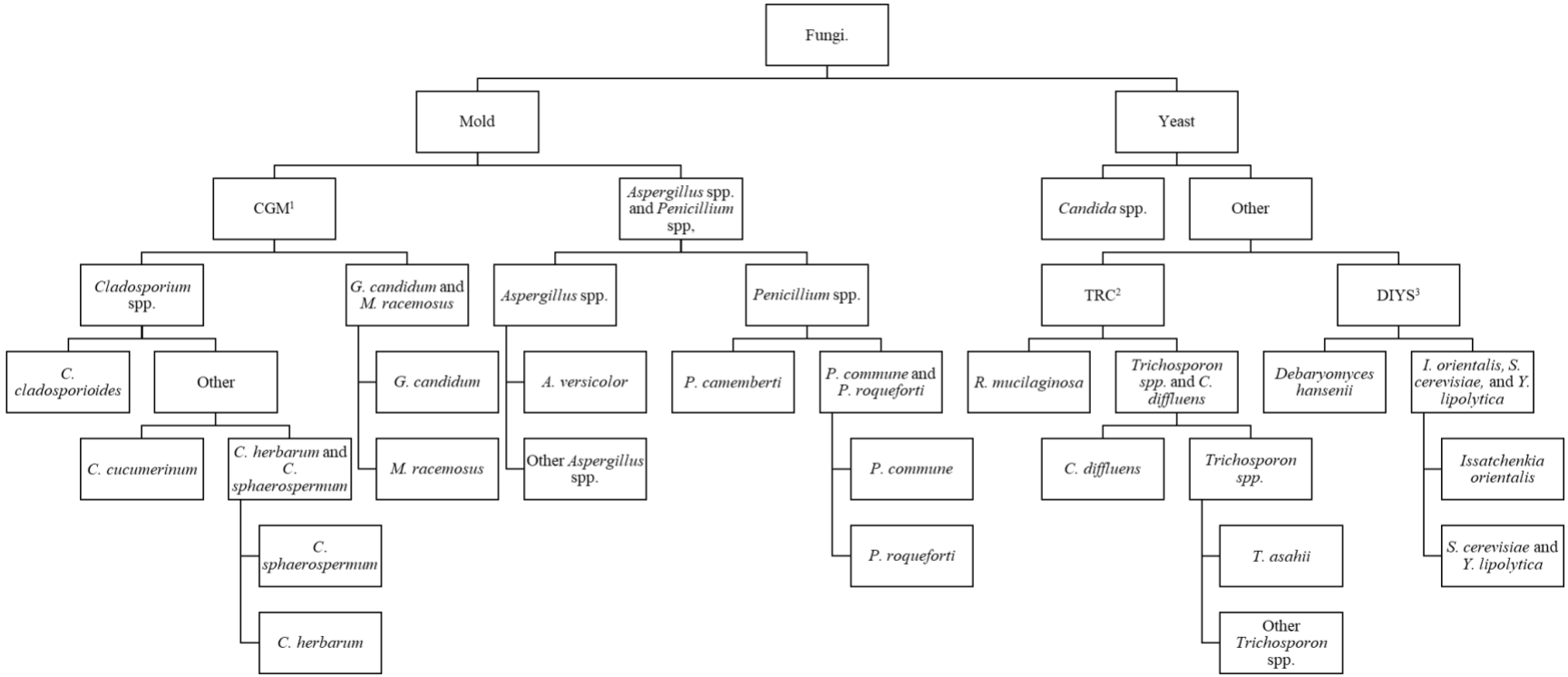


Figure 5.6. Identification flow chart of yeast and mold species.

[1]CGM is the abbreviation of *Cladosporium* spp., *Geotrichum candidum* and *Mucor racemosus.*

[2]TRC is the abbreviation of *Trichosporon* spp., *Rhodotorula mucilaginosa* and *Cryptococcus diffluens.*

[3]DIYS is the abbreviation of *Debaromyces hansenii, Issatchenkia orientalis, Yarrowia lipolytica* and *Saccaromyces cerevisiae.*

Table 5.11. FTIR identification using the in-house build fungi and yeast database.

	Genus	Species	In database	For prediction	Correct identification[1]	Misidentified[2]	Non-Identified[3]
Mold	*Aspergillus*	*versicolor*	4	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
		aculeatus	3	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
		flavus	20	11	10/11 (90.91%)	1/11 (9.09%)	0/11 (0%)
		fumigatus	1	4	4/4 (100%)	0/4 (0%)	0/4 (0%)
		japonicus	2	1	0/1 (0%)	1/1 (100%)	0/1 (0%)
		niger	11	9	9/9 (100%)	0/9 (0%)	0/9 (0%)
		niveus	1	0	-	-	-
		oryzae	1	1	1/1 (100%)	0/1 (0%)	0/1 (0%)
		parasiticus	7	7	6/7 (85.71%)	0/7 (0%)	1/7 (14.29%)
		terreus	4	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
		uvarum	4	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
	Cladosporium	*cladosporioides*	5	3	3/3 (100%)	0/3 (0%)	0/3 (0%)
		cucumerinum	3	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
		herbarum	2	1	0/1 (0%)	1/1 (100%)	0/1 (0%)
		spp.	4	3	3/3 (100%)	0/3 (0%)	0/3 (0%)
		sphaerospermum	6	6	6/6 (100%)	0/6 (0%)	0/6 (0%)
	Geotrichum	*candidum*	12	8	8/8 (100%)	0/8 (0%)	0/8 (0%)
		spp.	3	3	3/3 (100%)	0/3 (0%)	0/3 (0%)
	Mucor	*racemosus*	3	0	-	-	-
	Penicillium	*camemberti*	8	6	6/6 (100%)	0/6 (0%)	0/6 (0%)
		commune	3	1	1/1 (100%)	0/1 (0%)	0/1 (0%)
		roqueforti	7	4	4/4 (100%)	0/4 (0%)	0/4 (0%)
		spp.	19	10	10/10 (100%)	0/10 (0%)	0/10 (0%)
Yeast	*Candida*	*spp.*	282	0	-	-	-
	Cryptococcus	*diffluens*	1	0	-	-	-
	Debaryomyces	*hansenii*	8	8	8/8 (100%)	0/8 (0%)	0/8 (0%)
	Issatchenkia	*orientalis*	8	5	5/5 (100%)	0/5 (0%)	0/5 (0%)
	Rhodotorula	*mucilaginosa*	9	8	8/8 (100%)	0/8 (0%)	0/8 (0%)
	Saccaromyces	*cerevisiae*	2	0	-	-	-
	Trichosporon	*asahii*	3	2	2/2 (100%)	0/2 (0%)	0/2 (0%)
		cutaneum	1	0	-	-	-
		spp.	1	0	-	-	-
	Yarrowia	*lipolytica*	8	8	8/8 (100%)	0/8 (0%)	0/8 (0%)
	Total		**456**	**119**	**115/119 (96.64%)**	**3/119 (2.52%)**	**1/119 (0.84%)**

[1]Correct identification signify that the percentage of correct prediction must be ≥0.8000.
[2]Misidentification signify that the percentage of incorrect prediction must be ≥0.8000.
[3]No identification signify that the percentage of prediction result is <0.8000.

5.5. Conclusion

In this study, three different identification methods have been evaluated for the identification of *Aspergillus* strains. Multiplex RT-qPCR, MALDI-TOF MS and FTIR spectroscopy obtained 71.3%, 52% and 92.3% of correct identification, respectively. RT-qPCR may be more suitable for identification of *Aspergillus* strains in section *Nigri* and *Terrei* (*A. niger, A. japonicus, A. uvarum, A. aculeatus, A. terreus, A. niveus*); MALDI-TOF MS could be a good choice to use for identifying *A. fumigatus* and *A. flavus*; while FTIR spectroscopy provides the best means of identifying *Aspergillus* spp. The identification rate of 96.6% was obtained by FTIR spectroscopy following the enlargement of the database with additional fungal strains provides an effective means for the accurate identification of pathogenic fungal species for subsequent anti-fungal treatment cannot be overstated. Currently, phenotypic methods for the identification of the most common fungi remain useful in microbiology laboratories. Only a limited number of samples can be subjected to the identification analysis as such in a timely manner. The number of samples subjected to the identification analysis could be significantly increased by using ATR-FTIR spectroscopy without compromising the identification performance providing the added advantage of low cost and no reagent (enzyme or chemical) requirements as in the case for Multiplex RT-qPCR, MALDI-TOF MS analysis. On the basis of the results obtained in this study, ATR-FTIR spectroscopy can serve as a valuable technique in identifying different fungal species.

5.6. Reference

1. Geiser, D.M., et al., *The current status of species recognition and identification in Aspergillus.* Stud Mycol, 2007. **59**: p. 1-10.
2. Kujawa, M., *Some Naturally Occurring Substances: Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 56. Herausgegeben von der International Agency for Research on Cancer, World Health Organization. 599 Seiten, zahlr. Abb. und Tab. World Health Organization. Geneva 1993. Preis: 95, — Sw.fr; 95,50 US $.* Food / Nahrung, 1994. **38**(3): p. 351-351.
3. Pagano, L., et al., *Invasive Aspergillosis in patients with acute leukemia: update on morbidity and mortality--SEIFEM-C Report.* Clin Infect Dis, 2007. **44**(11): p. 1524-5.
4. Ragheb, S.M. and L. Jimenez, *Polymerase Chain Reaction/Rapid Methods Are Gaining a Foothold in Developing Countries.* PDA J Pharm Sci Technol, 2014. **68**(3): p. 239-255.
5. Hall, L., S. Wohlfiel, and G.D. Roberts, *Experience with the MicroSeq D2 large-subunit ribosomal DNA sequencing kit for identification of filamentous fungi encountered in the clinical laboratory.* J Clin Microbiol, 2004. **42**(2): p. 622-6.
6. Cassagne, C., et al., *Mould routine identification in the clinical laboratory by matrix-assisted laser desorption ionization time-of-flight mass spectrometry.* PLoS One, 2011. **6**(12): p. e28425.
7. Naumann, D., D. Helm, and H. Labischinski, *Microbiological characterizations by FT-IR spectroscopy.* Nature, 1991. **351**(6321): p. 81-2.
8. Bhat, R., *Potential Use of Fourier Transform Infrared Spectroscopy for Identification of Molds Capable of Producing Mycotoxins.* International Journal of Food Properties, 2013. **16**(8): p. 1819-1829.
9. Zhu, H., F. Qu, and L.H. Zhu, *Isolation of genomic DNAs from plants, fungi and bacteria using benzyl chloride.* Nucleic Acids Res, 1993. **21**(22): p. 5279-80.
10. Makimura, K., S.Y. Murayama, and H. Yamaguchi, *Detection of a wide range of medically important fungi by the polymerase chain reaction.* J Med Microbiol, 1994. **40**(5): p. 358-64.
11. Yang, Y., K. Zuzak, and J. Feng, *An improved simple method for DNA extraction from fungal mycelia.* Canadian Journal of Plant Pathology, 2016. **38**(4): p. 476-482.
12. Kim, W.B., et al., *Development of multiplex real-time PCR for rapid identification and quantitative analysis of Aspergillus species.* PLoS One, 2020. **15**(3): p. e0229561.
13. Bustin, S.A., et al., *The MIQE Guidelines: M inimum I nformation for Publication of Q uantitative Real-Time PCR E xperiments*. 2009, Oxford University Press.
14. Becker, P.T., et al., *Quality control in culture collections: Confirming identity of filamentous fungi by MALDI-TOF MS.* Mycoscience, 2015. **56**(3): p. 273-279.
15. Hermet, A., et al., *Molecular systematics in the genus Mucor with special regards to species encountered in cheese.* Fungal Biology, 2012. **116**(6): p. 692-705.
16. Schmitt, I., et al., *New primers for promising single-copy genes in fungal phylogenetics and systematics.* Persoonia-Molecular Phylogeny and Evolution of Fungi, 2009. **23**(1): p. 35-40.
17. Anand, A., et al., *Use of polymerase chain reaction in the diagnosis of fungal endophthalmitis.* Ophthalmology, 2001. **108**(2): p. 326-30.
18. Buil, J.B., et al., *Molecular Detection of Azole-Resistant Aspergillus fumigatus in Clinical Samples.* Front Microbiol, 2018. **9**: p. 515.

19. Embong, Z., et al., *Specific detection of fungal pathogens by 18S rRNA gene PCR in microbial keratitis.* BMC Ophthalmol, 2008. **8**: p. 7.
20. Mikulska, M., et al., *Use of Aspergillus fumigatus real-time PCR in bronchoalveolar lavage samples (BAL) for diagnosis of invasive aspergillosis, including azole-resistant cases, in high risk haematology patients: the need for a combined use with galactomannan.* Med Mycol, 2019. **57**(8): p. 987-996.
21. Sugita, C., et al., *PCR identification system for the genus Aspergillus and three major pathogenic species: Aspergillus fumigatusAspergillus flavus and Aspergillus niger.* Medical Mycology, 2004. **42**(5): p. 433-437.
22. Zarrin, M., et al., *Rapid Identification of Aspergillus Fumigatus Using Beta-Tubulin and RodletA Genes.* Open Access Maced J Med Sci, 2017. **5**(7): p. 848-851.
23. Zhao, J., et al., *Identification of Aspergillus fumigatus and related species by nested PCR targeting ribosomal DNA internal transcribed spacer regions.* J Clin Microbiol, 2001. **39**(6): p. 2261-6.
24. Chazalet, V., et al., *Molecular typing of environmental and patient isolates of Aspergillus fumigatus from various hospital settings.* J Clin Microbiol, 1998. **36**(6): p. 1494-500.
25. Einsele, H., et al., *Detection and identification of fungal pathogens in blood by using molecular probes.* J Clin Microbiol, 1997. **35**(6): p. 1353-60.
26. Palumbo, J.D. and T.L. O'Keeffe, *Detection and discrimination of four Aspergillus section Nigri species by PCR.* Letters in Applied Microbiology, 2015. **60**(2): p. 188-195.
27. Morton, C.O., et al., *RT-qPCR detection of Aspergillus fumigatus RNA in vitro and in a murine model of invasive aspergillosis utilizing the PAXgene® and Tempus™ RNA stabilization systems.* Medical Mycology, 2012. **50**(6): p. 661-666.
28. River, C. *Conidia IDFP for Filamentous Fungi Identifications*. 2020; Available from: https://criver.widen.net/s/yihnqhrxpb.
29. Prevention, C.f.D.C.a. *Identification of filamentous fungi using MALDI-ToF using the Bruker Biotyper*. SOPs 2022; Available from: https://www.cdc.gov/fungal/lab-professionals/identification_of_filamentous_fungi.html#processing.
30. Bruker. *MALDI Biotyper® for Food Microbiology*. 2021; Available from: https://www.bruker.com/en/applications/microbiology-and-diagnostics/food-beverage-microbiology/maldi-biotyper-for-food-microbiology.html.
31. Li, Y., et al., *Identification by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry and Antifungal Susceptibility Testing of Non-Aspergillus Molds.* Frontiers in Microbiology, 2020. **11**.
32. De Carolis, E., et al., *Species identification of Aspergillus, Fusarium and Mucorales with direct surface analysis by matrix-assisted laser desorption ionization time-of-flight mass spectrometry.* Clinical Microbiology and Infection, 2012. **18**(5): p. 475-484.
33. Becker, P., et al., *Identification of fungal isolates by MALDI-TOF mass spectrometry in veterinary practice: validation of a web application.* J Vet Diagn Invest, 2019. **31**(3): p. 471-474.
34. Vidal-Acuña, M.R., et al., *Identification of clinical isolates of Aspergillus, including cryptic species, by matrix assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS).* Med Mycol, 2018. **56**(7): p. 838-846.
35. Reeve, M.A., T.S. Caine, and A.G. Buddie, *Spectral Grouping of Nominally Aspergillus versicolor Microbial-Collection Deposits by MALDI-TOF MS.* Microorganisms, 2019.

7(8).
36. Schulthess, B., et al., *Use of the Bruker MALDI Biotyper for identification of molds in the clinical mycology laboratory.* J Clin Microbiol, 2014. **52**(8): p. 2797-803.
37. Riat, A., et al., *Confident identification of filamentous fungi by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry without subculture-based sample preparation.* International Journal of Infectious Diseases, 2015. **35**: p. 43-45.
38. Sleiman, S., et al., *Performance of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Identification of Aspergillus, Scedosporium, and Fusarium spp. in the Australian Clinical Setting.* J Clin Microbiol, 2016. **54**(8): p. 2182-6.
39. Luethy, P.M. and A.M. Zelazny, *Rapid one-step extraction method for the identification of molds using MALDI-TOF MS.* Diagn Microbiol Infect Dis, 2018. **91**(2): p. 130-135.
40. Parenicová, L., et al., *Combined molecular and biochemical approach identifies Aspergillus japonicus and Aspergillus aculeatus as two species.* Appl Environ Microbiol, 2001. **67**(2): p. 521-7.
41. Tamime, A.Y. and R.K. Robinson, *6 - Microbiology of yoghurt and related starter cultures*, in *Tamime and Robinson's Yoghurt (Third Edition)*, A.Y. Tamime and R.K. Robinson, Editors. 2007, Woodhead Publishing. p. 468-534.
42. Lecellier, A., et al., *Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds.* Food Microbiology, 2015. **45**: p. 126-134.
43. Shapaval, V., et al., *Characterization of food spoilage fungi by FTIR spectroscopy.* Journal of Applied Microbiology, 2013. **114**(3): p. 788-796.
44. Lecellier, A., et al., *Differentiation and identification of filamentous fungi by high-throughput FTIR spectroscopic analysis of mycelia.* International Journal of Food Microbiology, 2014. **168-169**: p. 32-41.
45. Garon, D., et al., *FT-IR Spectroscopy for Rapid Differentiation of Aspergillus flavus, Aspergillus fumigatus, Aspergillus parasiticus and Characterization of Aflatoxigenic Isolates Collected from Agricultural Environments.* Mycopathologia, 2010. **170**: p. 131-42.
46. Lam, L.M.T., et al., *Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy.* J Clin Microbiol, 2019. **57**(5).

Connecting Statement

The studies presented in the previous three chapters entailed implementation of fundamental principles of FTIR-based microbial identification in three fields of application, namely, veterinary microbiology, identification of foodborne pathogenic bacteria, and mycology. These fundamental principles include the need for culturing prior to spectral acquisition to obtain sufficient biomass and isolate pure colonies, the use of spectral databases consisting of spectra of reference strains (cultured under standardized conditions) in sufficient numbers to adequately represent the phenotypic diversity of the genera/species of interest, the application of multivariate classification or machine learning algorithms to discriminate among the genera/species represented in the database, and validation of the classification models to assess their predictive accuracy prior to implementing them in the identification of unknowns.

In the case of viral pathogens, similar principles may be applied with several notable exceptions. Although viruses are often referred to as microorganisms, they can only replicate in host cells and hence can only be cultured if an appropriate cell culture system exists. Furthermore, culturing of viruses does not produce sufficient viral mass to allow for FTIR spectral acquisition with a standard laboratory instrument, obviating any possibility of routine identification of viral pathogens by FTIR spectroscopy in the manner outlined above. On the other hand, for a given viral pathogen, the FTIR spectra of specimens of an appropriate biofluid collected from infected and uninfected individuals may exhibit differences serving as a basis for detection of the viral infection. Given that such differences are unlikely to be visually discernible, discriminatory features can be explored in accordance with the principles employed in discriminating among microbial pathogens on the basis of FTIR spectral differences. The Covid-19 pandemic provided the opportunity to conduct the proof-of-concept study described in the next chapter, utilizing saliva specimens collected from individuals undergoing routine PCR-based screening at a local hospital to develop an FTIR-based method for detection of SARS-CoV-2 infection.

Chapter 6. Case study for the application Transflection-FTIR spectroscopy for SARS-CoV-2 detection in heat-inactivated saliva fluids

6.1. Abstract

In December 2019, a novel coronavirus was implicated in a viral outbreak reported in Wuhan China; this virus and the disease it causes were subsequently were subsequently named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Covid-19, respectively [1]. As of April 2023 a total of 764,474,387 confirmed cases of COVID-19, including 6,915,286 deaths were reported by the World Health Organization (WHO) [2]. The most accurate test for the detection of the virus to date is the real time polymerase chain reaction (RT-qPCR) (from nasal swabs or saliva). RT-qPCR tests are expensive, labor-intensive, and have long turnaround time when considering the number of daily tests required to screen large population segments. Over the course of the pandemic, alternative rapid and cost-effective tests for Covid-19 diagnosis based on immunoassay were developed. These tests also required the use of reagents, and the sensitivity of the immunoassay-based methods were lower than RT-qPCR. Herein, we aimed to evaluate application of transflection Fourier-transform infrared (TR-FTIR) spectroscopy as a rapid diagnostic tool for Covid-19 using heat-inactivated saliva. This technique is reagent free, non-destructive, and rapid amenable to on-site testing. The results are obtained within minutes of acquiring a sample of saliva from the individual. In this feasibility study conducted in April 2021, a total of 940 raw saliva samples (418 tested positive for Covid-19 and 522 tested negatives for Covid-19 by RT-qPCR) were collected and heat inactivated at 85 ºC for 15 mins. Discrimination models were developed based on spectral differences between the two sample sets using k-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM) algorithms, and yielded 82.9%-85.4% sensitivity, 82.4%-86.3% specificity, 84.8%-85.9% accuracy, and 79.6-83.4% precision. A comparison between the mean TR-FTIR spectra of all PCR negative and PCR-positive saliva specimens revealed a stronger band at 1017-1031 cm^{-1} corresponding to sugar moieties of glycosylated proteins. In addition to this band, spectral regions between 870 and 890 cm^{-1}, 1711 and 1732 cm^{-1}, and 1761 and 1780 cm^{-1} were instrumental to the generation of effective discrimination models. In summary, TR-FTIR spectroscopy could serve as a rapid, reagent-free method capability of on-site screening of saliva samples for Covid-19.

6.2. Introduction

Following its initial detection at the end of 2019, the novel coronavirus SARS-CoV-2 spread rapidly across the globe, causing the Covid-19 pandemic declared on March 11th 2020 by the World Health Organization (WHO). Infection with this highly contagious virus produced symptoms ranging from mild disease to severe pneumonia and multi-organ failure, eventually leading to death, especially in older patients and immunocompromised individuals, This viral outbreak is undoubtedly viewed as a global catastrophic event, posing a tremendous threat to healthcare and the world economy and surpassing malaria as the world's most devastating infectious disease [2].

Among all the measures taken to prevent the spread of Covid-19, rapid detection and widespread testing were crucial, particularly prior to large-scale vaccine delivery. Reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) is considered the gold standard for Covid-19 diagnosis. However, this method is time consuming (up to 72 hours) and costly and requires special laboratory equipment and highly trained personnel to operate. Due to the high expense of RT-PCR and shortage of test kits, some countries were only able to provide testing for a limited number of individuals, forcing them to exclude infected patients with mild or no symptoms (both of whom can spread the contagion). Antigen detection tests are another alternative and are widely used for Covid-19 self-testing. Yet the current tests suffer from suboptimal sensitivity to rule out the disease and need to implement a clear guidance on correct interpretation for Covid-19 detection [3]. Serology tests, such as enzyme-linked immunosorbent assays (ELISA), identify antibodies in blood or saliva. Nonetheless, the antibody test may not detect acute infection in the first two weeks, when viral shedding and transmission is at the highest [4]. All that said, a rapid, efficient and cost-effective tool for the detection of Covid-19 along with clinically validated sensitivity and specificity is lacking. A possible candidate is Fourier transform infrared (FTIR) spectroscopy, which serves is a non-invasive, low-cost, reagent-free and rapid technique for the analysis of bio-fluids. To date, many studies have examined the reliability of FTIR spectroscopy for use in disease diagnostics, including but not limited to cancer, neurological disorders, respiratory disorders, gastrointestinal disorders, kidney and urinary tract disorders, gynecological disorders, hematological disorders, and dermatological disorders [5].

Nowadays, laboratory tests for Covid-19 diagnosis are performed mainly with the upper respiratory tract (nasopharyngeal swab) or lower respiratory specimens (saliva) [6]. The former may cause the patient to sneeze or cough increasing the risk of exposure of healthcare staff to the virus. Moreover, nasopharyngeal swab collection is uncomfortable for the patient, and bleeding may occur occasionally leading to further complications. On the other hand, saliva specimens have been reported to have similar or higher Covid-19 virus yield than nasopharyngeal samples with significant advantages [7]. Saliva specimen is stable for diagnostic purpose for 24 hours at room temperature and for a week at 4℃ without coagulation [8]. Additionally, saliva samples can be stored at -80℃ for months and remain useful for scientific investigation [9]. Saliva can also be easily self-collected by patients at home, without patient discomfort and minimizes viral exposure to healthcare staff. It can be considered as the best specimen for diagnosis for humans from an ethical point of view. Since it is non-painful and non-stressful, saliva collection can be used in large scale or epidemiological studies, allowing widespread testing. Mardani et al. showed that salivary biomarkers having the potential for Covid-19 diagnosis include angiotensin converting enzyme 2 (ACE2), antibodies (immunoglobulin A (IgA), IgG and IgM), alanine aminotransferase, C-reactive protein, neutrophil, lactate dehydrogenase, and serum urea [10].

Due to the consideration mentioned above, this study focuses on the use of transflection-FTIR (TR-FTIR) spectroscopy for Covid-19 diagnosis using heat-inactivated saliva samples. In contrast to many studies reported in the literature using raw saliva samples, we heated all saliva samples to inactivate the virus (if present) for safety consideration and ease of sample handling. Furthermore, the demographic features of our patients were studied and compared for prediction relevancy. The wavenumber regions associated with spectral differences between healthy and Covid-19 infected specimens, reflective of biomarkers, were identified by multivariate analysis. As effective diagnosis of viral infection from a spectrum that encompasses contributions from all biocomponents of saliva may be challenging, this paper will use k-nearest neighbor (KNN), support vector machine (SVM) and artificial neural network (ANN) algorithms to develop classification models allowing for Covid-19 screening from a limited number of regions in the spectra of the saliva specimen.

6.3. Materials and Methods

6.3.1. Participants and saliva collection

Between April 2021 and June 2021, a total of 940 patients aged 2-91 years (452 males, 486 females, and 2 unknowns) were randomly selected at Hôpital de la Cité-de-la-Santé located in Laval, QC, Canada. Saliva samples were self-collected by patients under supervision of healthcare provider in labelled sterile tubes. For convenience, saliva samples were stored at 4 ℃ after collection until TR-FTIR spectral acquisition (0-3 days) took place. Taxonomic composition of saliva was shown to be stable within 9 days by a previous study [11]. Nasopharyngeal cotton swabs were also collected at the same time for RT-qPCR analysis in the laboratory of Hôpital de la Cité-de-la-Santé. The RT-qPCR results from nasopharyngeal swab were used for assigning saliva samples as Covid-19 positive (PP) or Covid-19 negative (PN). For all participants, demographic data (age and gender) were collected and summarized in A.4. This study was examined and approved by the Ethics Committees from Hôpital de la Cité-de-la-Santé.

6.3.2. Sample preparation and TR-FTIR measurements

Collected saliva samples (1-mL) were transferred into 1.5 mL Eppendorf tubes, vortexed and incubated at 85℃ for 15 min to inactivate the Covid-19 virus [12]. Aliquots of 15-20 µl heat-inactivated saliva sample were deposited onto low e-glass slides and allowed to dry. Each e-glass slide was designed to have 10 spots for sample deposition, and each specimen was deposited onto 2 spots. The IR spectrum was measured once per spot, yielding two independent spectra per sample. A descriptive flow chart from saliva collection to spectral analysis is depicted in Figure 6.1. TR-FTIR spectra were acquired using a Nicolet^{TM} Summit (Thermo Fisher Scientific Waltham, US) FTIR spectrometer. All spectra were recorded in the region between 600 cm^{-1} and 4000 cm^{-1} in transflection mode. For each FTIR spectrum, 32 scans were co-added, averaged with a spectra resolution of 8 cm^{-1} and ratioed against a background spectrum collected from a clean e-glass slide surface.

Figure 6.1. Methodology flow chart of saliva specimen analysis by TR-FTIR.

6.3.3. Data pre-processing and Spectral analysis

Duplicate spectra were averaged for each sample, followed by baseline correction and normalization using an in-house written software and commercially available spectral analysis software OMNIC (Thermo Scientific™, USA). TR-FTIR spectra were separated into 3 sets: training set (60%), validation set (15%) and testing set (25%), and were exported from MATLAB (The MathWorks Inc., Natick, MA, US) as csv files using an in-house written algorithm, and imported directly into JMP Statistical Discovery software, version 16 to perform data analysis. Random forest algorithm is a decision-tree based approach that can analyze complex interactions and identify non-linearities of predictor effects. Random forest was applied to narrow down relevant wavenumber ranges out of the 3527 variables in the biological fingerprint region between 844 and 1780 cm^{-1} and 2800-3020 cm^{-1} with ascending order of contribution in terms of separation between Covid-19 negative and Covid-19 positive samples. The spectral combination contributing the most efficient separation of classes was subsequently chosen, and principal component analysis (PCA) was used to reduce dimensionality of the wavenumber and serve as inputs for the prediction models.

Developed database with analyzed classification models comprising all processed saliva spectra was then split into two databases for further analysis. Demographic information of each of

the three databases was evaluated and compared, a student's t-test and χ^2 test was performed to determine any significant difference among them. Fourier self-deconvolution was carried out in OMNIC (Thermo Scientific™, USA) using a bandwidth of 24 cm^{-1} and a resolution-enhancement factor of 2.0 for visual analysis of models' averaged spectra. Wavenumber range that was differentiating the models were being analyzed, and the original spectra are re-filtered accordingly to bring up ultimately a final database.

6.3.4. Prediction models

Three different classification models based on k-nearest neighbor (KNN), support vector machine (SVM) and artificial neural network (ANN) were developed using discrete spectral regions from the TR-FTIR spectra of saliva specimens. The three multivariate methods employed in this study have been widely reported in the literature for infrared spectral analysis [13-15]. Here, the three models, each developed using one of the above multivariate techniques were compared in terms of sensitivity and specificity for the discrimination between Covid-19-positive and negative samples by TR-FTIR spectroscopy.

KNN is a supervised machine learning classifier that predicts the correct class of test data (query point) by taking into account the distance between the test data and all the training points. The 'K' number of points which is closest to the test data is then selected using Euclidean distance metrics. Finally, the algorithm calculates the probability of the test data belonging to the 'K' training data class. The class with the highest probability will be selected as the prediction result of the test data [15]. KNN has been widely used in classification tasks for disease prediction due to its easiness to implement and interpret.

ANN is a self-training system and intelligently constructed to optimize the processing power of its own network. It works like the human nervous system, where each neuron receives a signal from neighboring neurons, later executes them to give the output signal. The number of neurons used may vary from ten to several thousands and are based on the training set. As more data is fed in, the algorithm gets smarter and more efficient at interpreting future inputs. One key aspect of ANN is that each neuron can be formulated to utilize a single algorithm that could be useful for certain datasets but poor to others [14]. And as weights are adjusted for each neuron, ANN learns by itself where to best analyze the data for having the highest confidence output and

continues to adjust neuron weights for more optimization of the network. Many studies showed that ANN models can be very useful for clinical data in terms of diagnosis [14].

SVM is a classification algorithm that aims to find hyperplanes that splits the class labels in the multidimensional space [13]. The best hyperplane that will be employed is the one that maximizes the margin among data points of different classes, that is, the hyperplane whose distance to the nearest data of each class is the largest. Test data is classified based on which side of the hyperplane it lies on. SVM uses complex kernel function to converts non-separable problems into separable problems when the data are non-linear by mapping the training data into higher-dimensional feature space [15]. The kernel function can be linear, polynomial or radial, and in our context, radial basis function is used for SVM.

6.3.5. Model performance evaluation

Model performance evaluation is an important step to construct effective machine learning model. Performance measure metrics are derived from confusion matrix, which includes true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

In this study, six performance metrics are adopted to assess the performance of the proposed models: accuracy, precision, sensitivity, specificity, F1 score values, and Mathews' Correlation Coefficient (MCC).

Accuracy is a widely used metric which measures the proportion of correctly predicted cases (TP and TN) over the whole dataset predictions. For unbalanced data, this metric can be misleading. It is measured by the equation given in as followed:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the percentage of correctly predicted Covid-19 positive instances out of the total predicted positive instances. That is, to find out what percentage the model is positive when it says it is positive. The equation of precision is given below:

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity is the percentage of positive instances that are correctly identified as positive out of the total actual positive instances. The following equation illustrates the sensitivity:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity is the proportion of predicted negative instances that was truly negative out of the total observed negatives. Mathematically represented as followed:

$$Specificity = \frac{TN}{TN + FP}$$

F1-score is a single-value metric by computing the harmonic mean of sensitivity and precision as shown in the following equation. F1-score is highly influenced by the positive class. In imbalanced data, F1-score may be unreliable if equal attention is needed for both positive and negative class. The range of F1-score is between 0 and 1, where 1 refers to a model that perfectly predicts each data to the correct class and 0 refers to a model that is unable to classify any data to the correct class.

$$F1\ score = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision}$$

Similar to F1-score, MCC is a single-value metric that summarizes the confusion matrix. However, as MCC accounts all four values from the confusion matrix, it is a more balanced assessment of classifiers, regardless which class is positive. MCC ranges between -1 and 1, where 1 refers to the best agreement between actuals and predictions and 0 refers to the predictions that are randomly assigned. The equation of MCC is represented as followed:

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

6.4. Results and Discussion

6.4.1. Development of Database 1, Database 2, and Database 3

The first TR-FTIR spectral database is comprised of a spectrum from a total of 940 individuals, in which 452 males (48.09%) and 486 females (51.7%) with an average age of 29.4 ± 18.5 years old. In the male group, 247 (54.6%) are healthy (PN) and 205 (45.4%) are Covid 19-positive (PP); in the female group, there were 275 (56.6%) PN and 211 (43.4%) PP. Through analysis by random forest and PCA, Database 1 was subsequently separated into 2 models to optimize prediction results: Database 2 and Database 3, each comprising 270 and 670 spectra, respectively of the original set.

Database 2 was integrated with 124 males (45.9%) and 146 females (54.1%) and had averaged age of 30.2 ± 18.0. On the other hand, average age of Database 3 is 28.9 ± 18.7 and comprised 328 males (49.0%) and 340 females (50.8%). Detailed demographic information on the study population of the 3 models is presented in Table 6.1.

Table 6.1. Patient characteristics.

Group	Gender	Age (Mean) years	Total patients by age groups	PCR result	
			Age range (years)	Positive (n=418)	Negative (n=522)
			0-30 (n=232)	84	148
			31-60 (n=230)	109	121
	Female (n=486)	30.51 ± 17.76	61-99 (n=24)	18	6
			Total	211	275
			0-30 (n=247)	86	161
			31-60 (n=181)	102	79
Database 1 (n=940)		28.14 ± 19.2	61-99 (n=24)	17	7
	Male (n=452)		**Total**	205	247
			Age range (years)	Positive (n=125)	Negative (n=145)
			0-30 (n=68)	32	36
			31-60 (n=71)	33	38
	Female (n=146)	31.54 ± 17.25	61-99 (n=7)	3	4
			Total	68	78
			0-30 (n=66)	35	31
			31-60 (n=53)	21	32
Database 2 (n=270)		28.67 ± 18.79	61-99 (n=5)	1	4
	Male (n=124)		**Total**	57	67

			Age range (years)	Positive (n=293)	Negative (n=377)	
			0-30 (n=164)	52	112	
			31-60 (n=78)	76	83	
	Female (n=340)	30.07 ± 17.98	61-99 (n=15)	15	2	
			Total	143	197	
			0-30 (n=181)	51	130	
			31-60 (n=128)	81	47	
Database 3 (n=670)		27.94 ± 19.38	61-99 (n=19)	16	3	
	Male (n=328)		**Total**	148	180	

Gender and age group distributions of the 3 databases are graphically presented in Figure 6.2, with PN represented as plain color, and PP with shaded color.

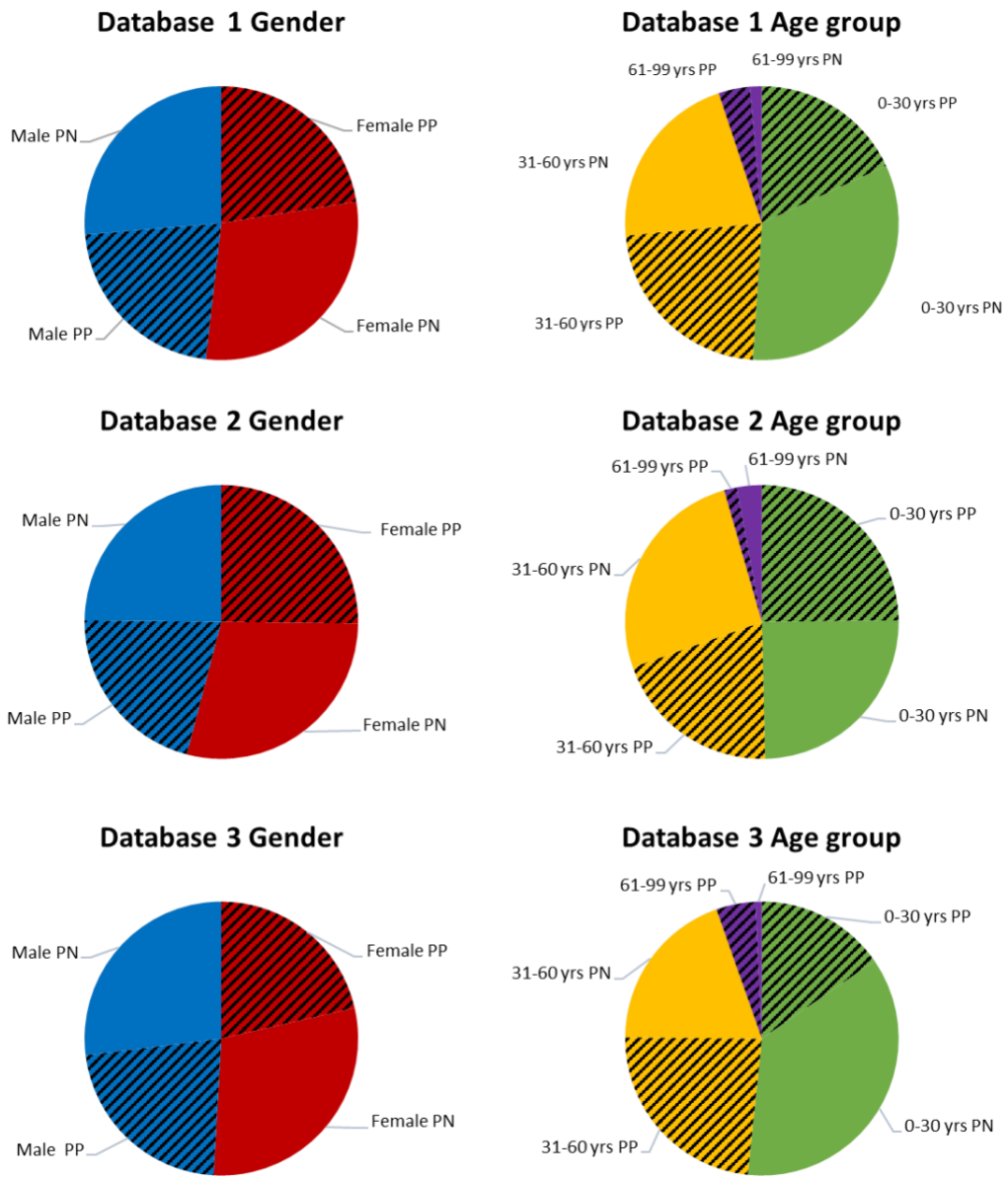


Figure 6.2. Gender and age groups statistics of patients.

6.4.1.1. Prediction results and Evaluation of Database 1, Database 2, and Database 3

Duplicate TR-FTIR spectra were averaged, baseline corrected, and area normalized as these pre-processing procedures delivers the best results for the classification models. Through random forest, wavenumbers with the highest contribution for the separation of PP and PN were selected for subsequent PCA, and the calculated principal components (PCs) were used as input for the development of classification models. Spectra were separated randomly into training (60%), validation (15%), and test (25%) set, stratified for PP and PN in each set. Similarly, training and

validation, test sets were used for optimisation and testing of the classification models respectively. A summary of the results is shown in Table 6.2. The classification model generated by SVM analysis of spectra from database 1 (comprising 940 averaged spectra from all the samples) had an average accuracy of 70%. Taking F1-score and MCC alone, their values are <0.7 and <0.5, respectively, suggesting that the algorithms are unreliable for COVID-19 screening. Other studies have demonstrated the capacity of FTIR spectroscopy to achieve sensitivity and specificity above 80% for Covid-19 diagnosis using saliva [16-18]. To improve on the performance of discrimination models, the original database (database 1) was split into two databases (database 2 and database 3) according to prediction error. Identical spectral analysis protocols were carried out for two additional spectra in Database 2 and 3. Prediction results and performance metrics for all 3 classification models were significantly improved (p=0.0001). All confusion matrices of Database 1, Database 2 and Database 3 can be found in the appendix section A.5. A summary of all the classification models of generated from the analysis of the spectra in the three spectral databases are disclosed in Table 6.2.

Table 6.2.. Prediction scores of Database 1, Database 2, and Database 3.

		Sample[1]			Quality parameters (%)[2]					
	Set	**Pos**	**Neg**	**Algorithm**	**Acc**	**Prec**	**Sens**	**Spec**	**F1**	**MCC**
				KNN	59.84	55.41	46	70.7	0.5027	0.1729
				ANN	69.84	66.3	64.4	74.1	0.6533	0.3867
	Training (n=630)	278	352	SVM	70.48	69.91	56.8	80.7	0.6268	0.3884
				KNN	68.12	66.67	59.6	75.2	0.6294	0.3528
				ANN	68.6	67.06	60.6	75.2	0.6367	0.3629
	Validation (n=207)	94	113	SVM	66.67	69.01	52.1	80.5	0.5938	0.3425
				KNN	61.17	56.25	58.7	63.2	0.5745	0.2178
				ANN	73.79	72.09	67.4	78.9	0.6967	0.4671
Database 1	Test (n=103)	46	57	SVM	76.7	73.81	67.4	80.7	0.7046	0.4865
				KNN	96.67	95.29	97.6	95.9	0.9643	0.9333
				ANN	100	100	100	100	1	1
	Training (n=180)	83	97	SVM	98.89	98.81	100	99	0.994	0.9889
				KNN	100	100	100	100	1	1
				ANN	100	100	100	100	1	1
	Validation (n=59)	27	32	SVM	98.31	96.3	96.3	96.9	0.963	0.9317
				KNN	93.55	93.33	93.3	93.8	0.9332	0.8708
Database 2	Test (n=31)	15	16	ANN	93.55	100	86.7	100	0.9288	0.8771

				SVM	93.55	100	86.7	100	0.9288	0.8777
				KNN	88.62	90.56	82.7	93.2	0.8645	0.7691
				ANN	95.09	93.97	94.9	95.2	0.9443	0.9005
	Training (n=448)	197	251	SVM	95.98	95.45	95.9	96.4	0.9568	0.923
				KNN	89.19	87.5	87.5	90.5	0.875	0.7798
				ANN	91.89	89.39	92.2	91.7	0.9078	0.8357
	Validation (n=148)	64	84	SVM	89.19	84.06	90.6	86.9	0.8721	0.7699
				KNN	87.84	81.08	93.8	83.3	0.8698	0.7638
				ANN	91.89	88.23	93.8	90.5	0.9093	0.8373
Database 3	Test (n=74)	32	42	SVM	89.19	85.29	90.6	88.1	0.8787	0.7826

[1]Pos = Covid-19 positive; Neg = Covid-19 negative.

[2]Acc = Accuracy; Prec = Precision; Sens = Sensitivity; Spec = Specificity; F1 = F1-score; MCC = Mathew's Correlation Coefficient.

6.4.1.2. Spectral Analysis of Database 2 and Database 3

Accuracy, precision, sensitivity and specificity were up to 100% and 96.4% for Database 2 and 3, respectively. F1-score and MCC value also increased to >0.85 and >0.75, respectively. The high scores indicate information lies within the FTIR spectra that differentiates between PP and PN. However, an important factor is misleading in the classification models, causing high misidentification rate once Database 2 and 3 are combined together into Database 1. Furthermore, this unknown factor overrules the difference between PP and PN in the spectra. No statistical difference (student's t-test, two-tailed equal variance, $\alpha<0.05$) is observed between Database 2 and 3 in terms of age (p = 0.3699). Through χ^2 test ($\alpha<0.05$), there were no sufficient statistical evidence to reject the null hypothesis, suggesting that gender (p = 0.4455) and PCR result (p = 0.3781) distribution were not dissimilar between the two Databases either. To resolve overlapped bands and improve information content of FTIR spectra, Fourier self-deconvolution was thus performed (bandwidth of 24 cm^{-1}, resolution-enhancement factor of 2.0) over the averaged raw PP and PN spectra of Database 2 and 3, as shown in Figure 6.3.

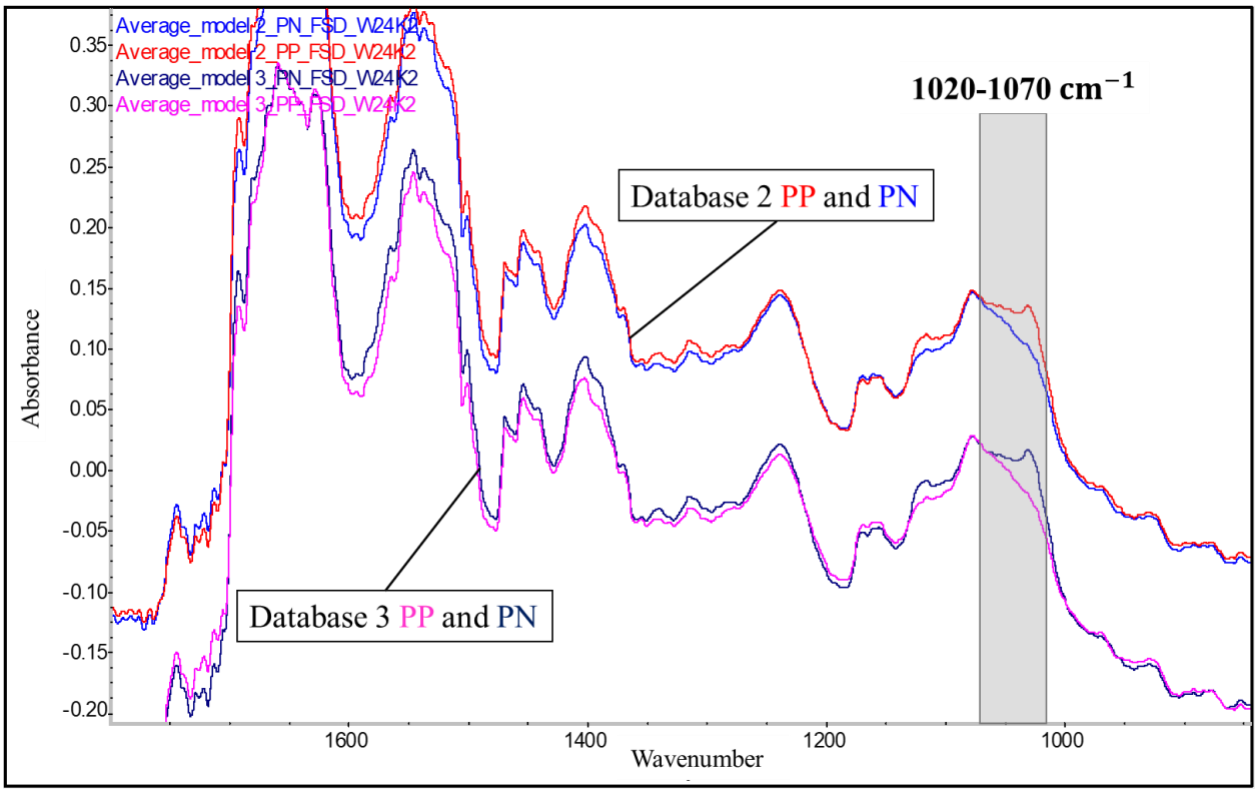


Figure 6.3. Mean PP and PN TR-FTIR spectra of Database 2 and Database 3.

An inverse band intensity of PP and PN at wavenumber region 1020-1070 cm^{-1} between the averaged spectra of Database 2 and Database 3 was observed (Figure 6.3). Briefly, averaged PP of Database 2 and averaged PN of Database 3 had an apparent peak at 1031 cm^{-1}, whereas their counterpart does not. This spectral region is associated to carbohydrates, and can be assigned to sugar moieties of glycosylated proteins in saliva [19]. High band absorbance at this region could be attributed to higher mucin content (a glycogen that serves as the main component of mucus) in some saliva samples [20]. It is not surprising to have sticky or thick saliva with high mucus level in some of the collected saliva samples. As a matter of fact, the spectral region of 1019-1027 cm^{-1} was used in the classification models to predict PP and PN in Database 1. And since some PP spectra and some PN spectra had inverted peaks over region 1020-1070 cm^{-1}, it is likely to have resulted in the low performance of the classification models from Database 1, as shown in Table 6.2. This could also explain the high prediction score when the spectra were split into two sets (Database 2 and Database 3).

6.4.2. Development of Database X

Re-filtration of the spectra was conducted based on the peak 1031 cm^{-1}. All averaged raw spectra were compiled into our in-house built software and baseline corrected. Peak height range of 1020-1070 cm^{-1} of all spectra is analyzed, and then ranked in ascending order. Spectra with the highest rank (peak absorbance abnormally high at 1020-1070 cm^{-1}) are being removed. Spectra with absorbance above 0.4 at peak 1031 cm^{-1} were considered as outliers. A final Database X was obtained with 828 spectra, in which 112 outlier spectra were removed.

The final Database X was integrated with 400 males (48.31%) and 426 females (51.45%) and had averaged age of 29.92 $\pm$ 18.39. In the male group, there were 214 (53.5%) PN and 186 (46.5%) PP, and 248 (58.2%) PN and 178 (41.8%) PP in the female group. Differently, the average age of outlier groups is 25.3 $\pm$ 18.8 and comprised 52 males (46.4%) and 60 females (53.6%). On the other hand, average age of Database 3 is 29.0 $\pm$ 18.7 and comprised 328 males (49.0%) and 340 females (50.8%). Detailed demographic information on the study population of Database X and outliers are shown in Table 6.3.

Table 6.3. Patient characteristics of Database X and Outliers.

Group	Gender	Age (Mean) years	Total patients by age groups	PCR result	
			Age range (years)	Positive (n=366)	Negative (n=462)
			0-30 (n=198)	68	130
			31-60 (n=206)	94	112
	Female (n=426)	31.00 ± 17.89	61-99 (n=22)	16	6
			Total	178	248
			0-30 (n=211)	77	134
			31-60 (n=169)	94	75
Database X (n=828)	Male (n=400)	28.77 ± 18.87	61-99 (n=20)	15	5
			Total	186	214
			Age range (years)	Positive (n=52)	Negative (n=60)
			0-30 (n=68)	16	18
			31-60 (n=71)	15	9
	Female (n=60)	27.03 ± 16.54	61-99 (n=7)	2	0
			Total	33	27
			0-30 (n=66)	9	27
			31-60 (n=53)	8	4
		23.31 ± 21.15	61-99 (n=5)	2	2
Outliers (n=112)	Male (n=52)		**Total**	18	33

Gender and age group distributions of the final Database X and outliers are graphically presented in Figure 6.4, with PN represented as plain color, and PP with shaded color.

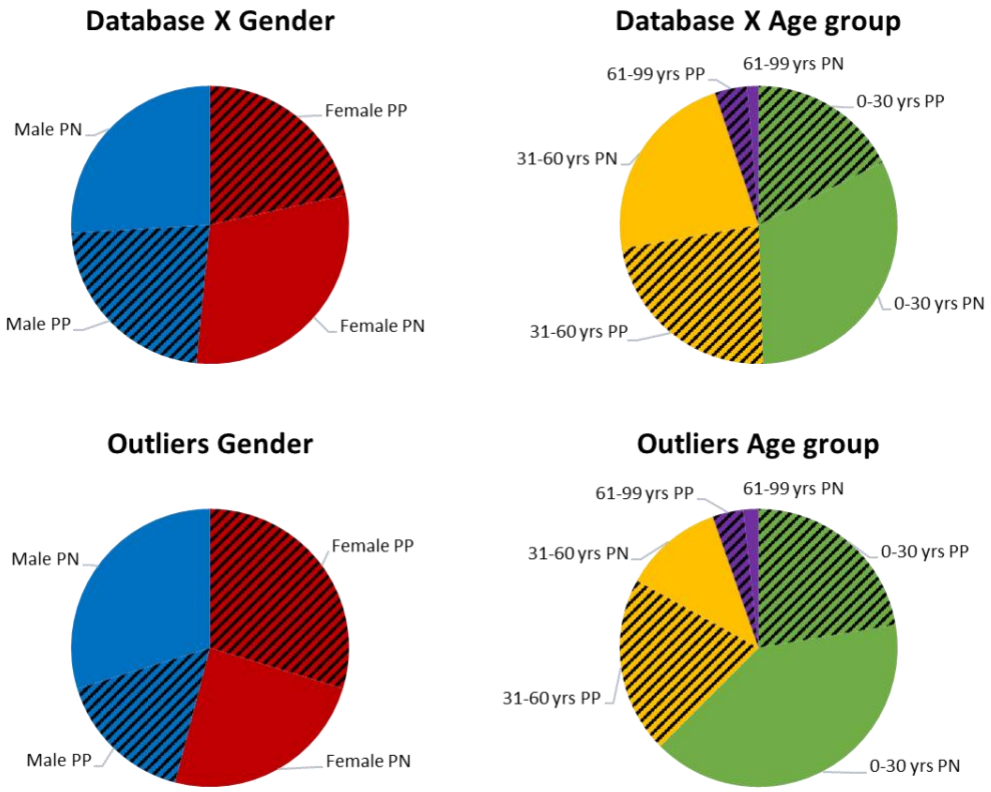


Figure 6.4. Gender and age groups statistics of Database X and Outliers.

6.4.2.1. Spectral Analysis of Database X

The application of χ^2 test ($\alpha<0.05$) showed that the gender (p = 0.6913) and PCR (p = 0.6374) between Database X and outlier group were not significantly different from each other. However, according to student's t-test, outlier group had significantly (p = 0.0131) younger age group distribution compared to Database X. As shown in previous studies, age-related changes in saliva composition are inevitable. For example, alpha-amylase was shown to have lower levels in the elderly [21]. However, other suggested no significant difference, or even decrease of this protein [22]. Different conclusion from literature could be due to different collection methods, as saliva is known to be simulated with stress. Within this context, higher α-amylase content may be observed in mostly younger population as no instruction was given prior saliva collection. Hence the youngster may have consumed confectionary increasing protein and carbohydrate content in

saliva. Mucin is another component worth mentioning. This protein decreases significantly with age, and this component may highly influence FTIR spectra [23, 24]. The higher content of proteins in outlier group, which comprises the younger group, can be visualized in TR-FTIR spectra in wavenumber region 1020-1130 cm^{-1} (Figure 6.5).

Fourier self-deconvolution was performed on the mean PP and PN spectra of Database X and outliers to investigate the possible protein content differences. As shown in Figure 6.5, the inverted peak shift observed in Figure 6.3 over wavenumber region 1020-1070 cm^{-1} was now untangled. Both averaged PN spectra of Database X and outlier group had a higher absorption intensity compared the PP spectra at peak located at ~1031 cm^{-1}. It is noteworthy that the outlier group had greater absorption at the entire 1020-1130 cm^{-1} region than Database X, regardless of PP and PN. This region encompasses carbohydrates, which include mucin, α-amylase, collagen and glycogen. The three dominant bands over region 1020-1130 cm^{-1} were 1031 cm^{-1} (sugar moieties of glycosylated proteins), 1075 cm^{-1} (DNA, RNA, phospholipids, and phosphorylated proteins), and 1130 cm^{-1} (carbohydrates, RNA, and phospholipids) [25, 26]. Furthermore, saliva samples were heated at 85℃ for 15 min before measurements for viral inactivation, and this procedure may contribute to spectral differences due to conformational changes of proteins. Nevertheless, the fact that PP and PN spectra in both Database X and outlier group can be visually discriminated indicates that FTIR spectroscopy is robust enough to overcome the minor saliva spectral differences induced by heat.

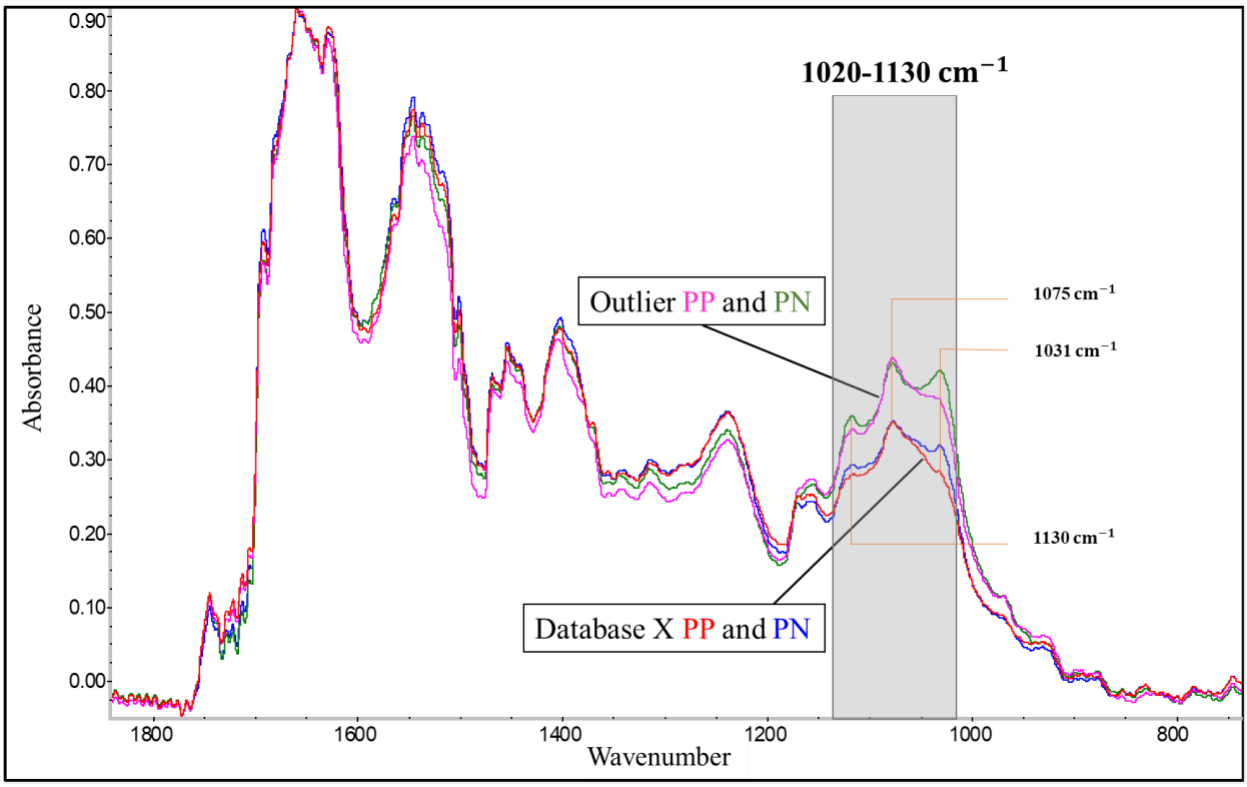


Figure 6.5. Mean PP and PN FTIR spectra of Database X and Outliers.

6.4.2.2. Prediction results and Evaluation of Database X

Classification models were developed and evaluated by dividing spectra into training (60%), validation (15%), and test set (25%). Same as previously done, the test set was not involved during model development. In brief, random forest was performed on averaged, baseline corrected, and area normalized spectra to detect wavenumbers that contributes the most for the separation of PP and PN in ascending order. PCA was then carried out to determine the corresponding PCs for model prediction evaluation. Similarly, the training and validation, test sets were used for all three classification models. Confusion matrices of Database X are shown In A.5. Prediction scores for Database X, comprising 828 spectra are shown in Table 6.4.

After removing outliers spectra with abnormal peaks within the wavenumber region 1020-1070 cm^{-1}, the test set accuracy was boosted to 84.8%, 84.8%, and 85.9% for KNN, ANN, and SVM respectively. Although SVM had the highest accuracy (85.9%), it had the lowest precision (79.6%), meaning that it had the lowest predicts of correct PP. In terms of ANN, it had comparable accuracy with KNN (84.8%), and had the highest precision (82.9%) and specificity (86.3%). However, its sensitivity (82.9%) is also the lowest compared to the other two algorithms. In the

Covid-19 context, higher chances of false-negative are more dangerous than high false-positives. Hence, FTIR spectra analysis by ANN may be more reliable in screening out PN rather than PP. On the other hand, KNN had comparable accuracy (84.8%) with ANN, moderate precision (81.4%), high sensitivity (85.4%) and moderate specificity (84.3%). Hence, KNN provides an overall better performance than the other two algorithms. Additionally, although all three models generated satisfactory F1-score (>0.8) and MCC value (>0.6), KNN also had the highest overall model performance parameters. Nonetheless, with larger dataset, ANN and SVM may outperform KNN due to their higher computational efficiency [27]. Still, it is always preferred to examine data with all three algorithms and to choose the one that has the most outstanding results. Most studies on Covid-19 diagnosis by saliva using FTIR spectroscopy were employed attenuated total reflectance (ATR) sampling method instead of transflection. Despite the fact that the sample size was limited (<240), they were all able to achieve high sensitivity (up to 95%) and high specificity (up to 89%) using discriminant analysis (linear discriminant analysis and partial least square discriminant analysis) [17, 18, 28]. It is worth mentioning that their methodologies are distinct from each other, especially the collection and treatment of saliva. Barauna et al. were using cotton swab of saliva and applying directly to the ATR crystal, while Nascimento et al. air-dried the raw saliva on aluminum foil and prior to spectral acquisition [17, 18]. Kazmer et al. inactivated the raw saliva as well with 70% ethanol [28]. One study used transflection-FTIR spectroscopy for Covid-19 screening and was able to achieve sensitivity of 93% and specificity of 82% [16]. However, their sample size was relatively small (n = 57).

Table 6.4. Prediction scores of Database X.

		Sample[1]			Quality parameters (%)[2]					
	Set	**Pos**	**Neg**	**Algorithm**	**Acc**	**Prec**	**Sens**	**Spec**	**F1**	**MCC**
				KNN	80.83	77.82	79.1	82.2	0.785	0.612
				ANN	84.99	83.13	82.8	86.7	0.83	0.6955
	Training (n=553)	244	309	SVM	86.26	83.27	85.7	86.4	0.845	0.7187
				KNN	80.87	79.49	76.5	84.3	0.78	0.6112
				ANN	87.43	86.25	85.2	89.2	0.857	0.745
	Validation (n=183)	81	102	SVM	81.97	81.82	77.8	86.3	0.798	0.6444
				KNN	84.78	81.4	85.4	84.3	0.834	0.6941
				ANN	84.78	82.93	82.9	86.3	0.829	0.692
Database X	Test (n=92)	41	51	SVM	85.87	79.55	85.4	82.4	0.824	0.6738

[1]Pos = Covid-19 positive; Neg = Covid-19 negative.

[2]Acc = Accuracy; Prec = Precision; Sens = Sensitivity; Spec = Specificity; F1 = F1-score; MCC = Mathew's Correlation Coefficient.

6.4.3. Spectral analysis of Databases

PP and PN differentiating bands determined by random forest of all four Databases are illustrated in Figure 6.6, with wavenumbers with highest contribution marked with a red star. In short, there were 84 wavenumbers, 86 wavenumbers, 46 wavenumbers, and 113 wavenumbers selected for Database 1, Database 2, Database 3, and Database X, respectively. In general, the discriminative wavenumbers lie within the region 800-1100 cm^{-1}, and 1500-1800 cm^{-1}. The former region can be associated with symmetric vibrations of $–PO^{2–}$ in phospholipid (~1080 $cm^{–1}$), nucleic acid for the vibrations of the sugar-phosphate backbone (780-1000 $cm^{–1}$), and carbohydrates in glycosylated proteins with the stretching vibrations of the C-O/C-O-C groups (800-1200 $cm^{–1}$) [29]. The amide I band at 1600-1700 $cm^{–1}$, and amide II band at 1500-1600 $cm^{–1}$, nucleic acid (1550-1780 $cm^{–1}$) assigned to in-plane vibrations of double bonds of the bases, and lipids at 1725-1745 $cm^{–1}$ symmetric stretching vibration of the ester carbonyl bond [29] also contribute to the development of the calibration models. Tentative band assignments of differentiating wavenumbers for all four Databases are listed in Table 6.5. Overall, PP and PN discriminating regions were mainly associated with the 844-890 $cm^{–1}$, 1015-1031 $cm^{–1}$, 1503-1570 $cm^{–1}$, 1620-1660 $cm^{–1}$, 1711-1735 $cm^{–1}$, 1752-1780 $cm^{–1}$, 2030-2033 $cm^{–1}$ spectral regions. Bands with highest contribution (red star in Figure 6.6) were ~860 $cm^{–1}$ (nucleic acid), ~1030 $cm^{–1}$ (glycosylated proteins), ~1650 $cm^{–1}$ (amide I), ~1725 $cm^{–1}$ (lipids), and ~1775 $cm^{–1}$ (lipids).

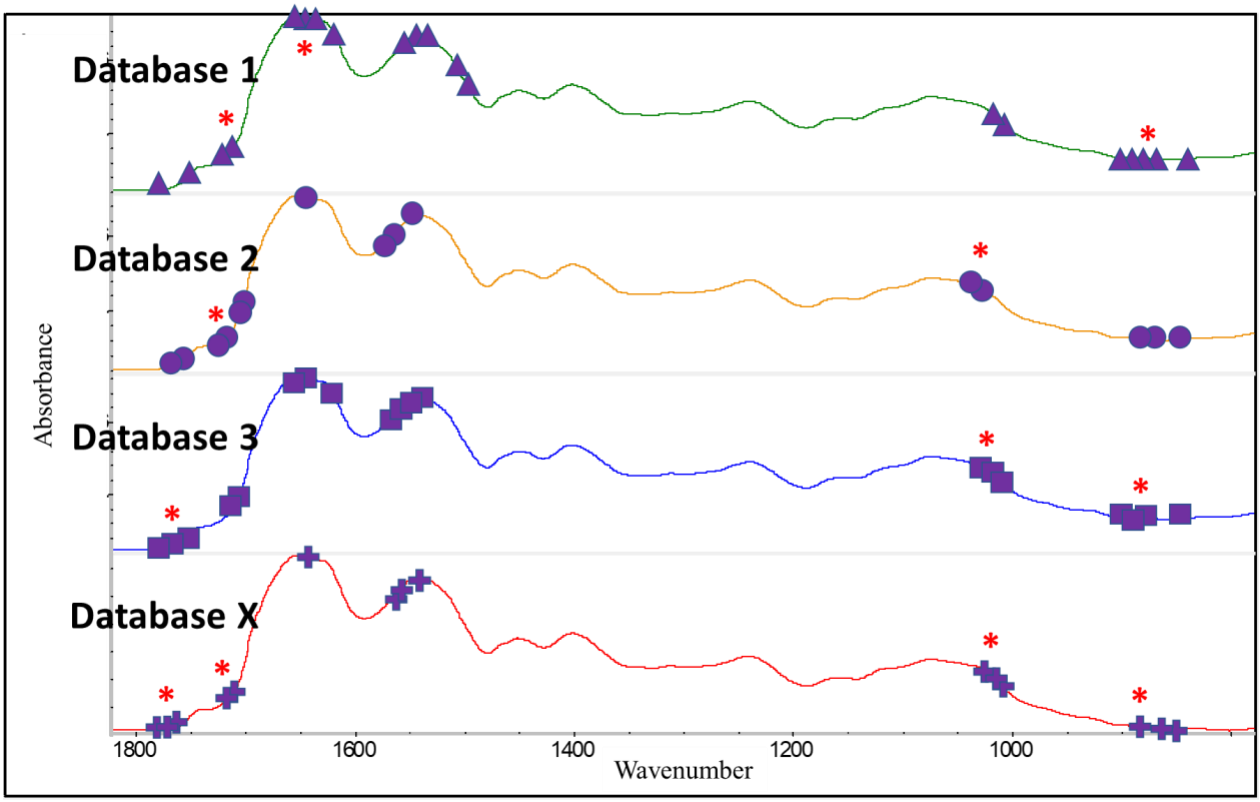


Figure 6.6. Figurative representation of wavenumbers contributing to PP and PN discrimination of all four Databases.

Table 6.5. Band assignment for Databases.

Band				Vibrational mode	Suggested biomolecule of saliva assignment	Reference
Database 1	**Database 2**	**Database 3**	**Database X**			
844	-	-	-	O-C-O bending, aromatics	CO_2	[30]
850-856*	850	850-852	850-853			
861	-	866-892*	865			
871-900	871	-	**870-890***	O-C-O bending, C-C, C-O, aromatics	CO_2, deoxyribose	[31, 32]
-	882-885	-	-	O-C-O bending	CO_2	[30]
1019-1027	1021-1032*	1015-1037*	**1017-1031***	CH_2OH groups, C-O stretching and C-OH groups bending; symmetric PO^{2-} stretching	Sugar moieties of glycosylated proteins (glycogen, alpha-amylase, collagen), nucleic acids (DNA & RNA), oligosaccharides, polysaccharides (glucose)	[19, 31-34]
1503	-	-	-	C-H bending	Phenyl ring	[35]
1510	-	-	-	CH_2 bending	Lipid, protein (Tyrosine)	[35]
1539-1561	1549	1547-1571	1553	N-H bending coupled to C-N stretching	Amide II (beta sheet)	[19, 35]
-	1559-1585	-	1559-1570			
-	-	1620	-	C=O stretching, C-N stretching, N-H bending	Amide I (parallel beta sheet) (peptide, protein)	[33, 35]
1636	-	-	-			
1644-1660*	1646	1644-1652	1651-1653	C=O stretching, C-N stretching, N-H bending	Amide I (alpha-helices, anti-parallel beta sheets) (peptide, protein)	[19, 35]
1712-1731*	1712	-	**1711-1732***	C=O stretching	Lipids (triacylglycerol, cholesterol esters, glycerophospholipids)	[35]
-	1723-1735*	-	-			
-	1752-1758	1756-1780*	-			
1760	-	-	**1761-1780***			
1778-1779	-	-	-			
-	-	-	2030-2033	Thiocyanate (SCN^-) anions	Product of detoxication of CN^- acquired with food	[33]

*Bands with a star (*) are the highest-ranking wavenumbers that contributes to the differentiation between PP and PN.

Changes in a diverse absorption bands reflecting changes in the biochemical composition of the saliva matrix associated with lipids, proteins, carbohydrates, and nucleic acids were identified through the use of region selection algorithm allowing for the effective differentiation between PP and PN (Table 6.5). The wavenumber ranges are shown in the overlapped spectra of PP and PN in the spectral region between 2150 and 750 cm^{-1} (Figure 6.7). PN showed higher absorption at 1031 cm^{-1} peak compared to PP. This could be associated to the higher α-amylase level in saliva of PN than PP [36]. Other studies did not observe pronounced differences in the 1031 cm^{-1} band between PP and PN samples [37]. The contradictory results may be due to different saliva collection procedures, for example, Martinez-Cuazitl et al., asked patients to fast at least 8h

prior saliva collection [37]. Furthermore, the heating process of saliva may alter the protein structure and cause irreversible conformational changes. Other discriminating bands were 850-853 cm^{-1}, 865 cm^{-1}, and 870-890 cm^{-1}, which could be associated with the stretching vibration C–C of DNA backbone and the vibration of ribose ring increased in PP due to the presence of the virus. An increase in band intensity of PP in these regions, were also reported by Kazmer et al. [28]. Another absorbance differences between PP and PN occurs in the spectral region between 1711 and 1732 cm^{-1} and 1761-1780 cm^{-1}, corresponding to lipids components. Song et al. also showed increase in lysophospholipids and sphingolipids with Covid-19 [38]. The protein amide I (1651-1653 cm^{-1}) and amide II (1553 cm^{-1} and 1559-1570 cm^{-1}) bands show minor differences (Figure 6.7). These wavenumber ranges were reported to be useful for the differentiation between PP and PN in other studies [17, 28]. The narrow band at 2030-2033 cm^{-1} corresponds to thiocyanate (SCN^{-}) anions [33] and is not usually present in biological samples other than saliva. Nevertheless, the minor variation in the band intensity in the 2030-2033 cm^{-1} region did not play a significant role in the PP/PN discrimination.

The immunoglobulin (Ig) spectral regions were assessed as well. It is known that saliva contains IgA, IgG, and IgG, with IgA and IgG being the principal antibody classes present [39]. As shown in Figure 6.7, PP group had higher absorption in IgA region (1187-1200 cm^{-1} and 1237-1285 cm^{-1}) [40]. We noticed that the absorption increase in IgA region correlates well with Covid-19 diagnosis, as reported in another study [37]. Nonetheless, the spectral changes in the region between 1464 and 1560 cm^{-1} attributed to IgG and the spectral region between 1028-1160 cm^{-1} and 1289-1420 cm^{-1} attributed to IgM did not show a higher absorption in PP samples Figure 6.7.

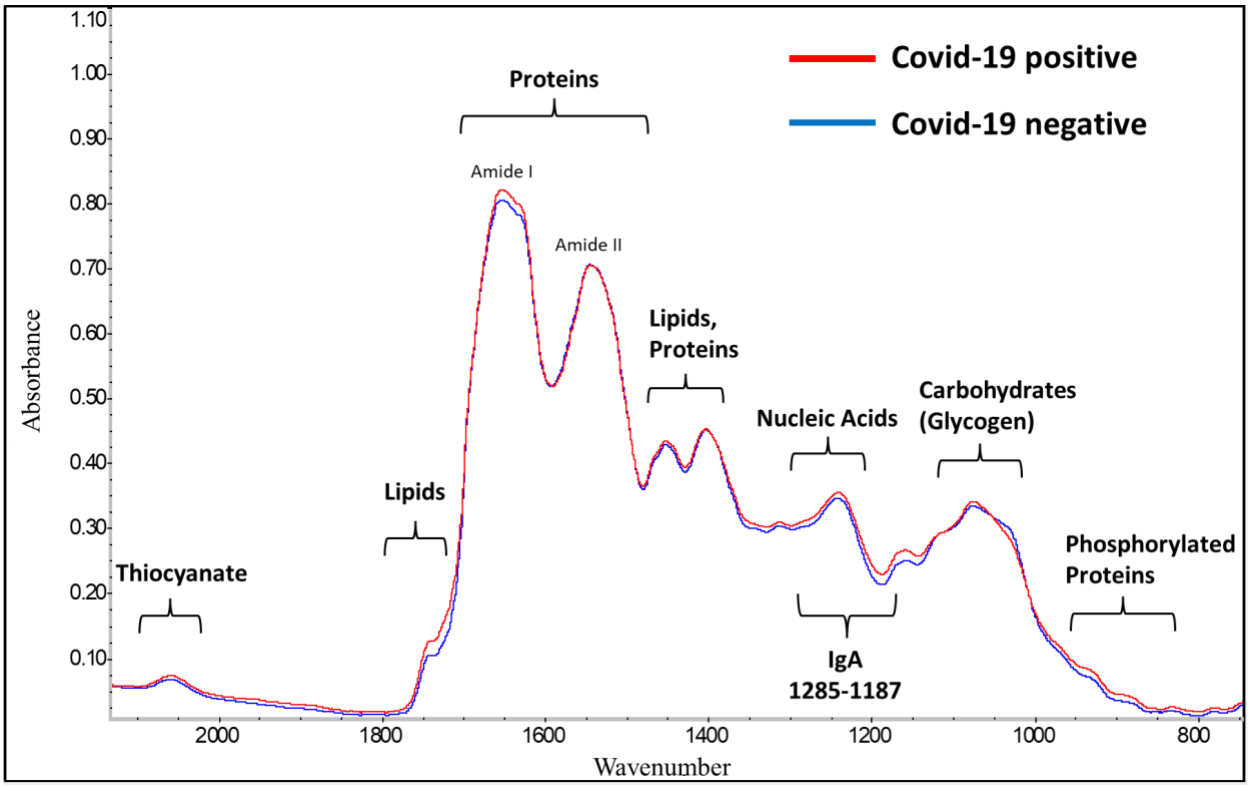


Figure 6.7. Mean raw spectra of PP and PN of Database X.

To the best of our knowledge, no previous study has investigated the use of heat-inactivated saliva for Covid-19 diagnosis by TR-FTIR spectroscopy. This study provided the possibility of using TR-FTIR spectroscopy as a fast and effective screening method for Covid-19 using different machine learning algorithms and showed both high sensitivity (85.4%) and specificity (86.3%). This technique also provided good accuracy (85.87%) and precision (82.93%). An important advantage of heat inactivation is related to the safety procedure of the technique. The overall F1-score (0.834) and MCC (0.6941) values suggest that KNN, ANN and SVM could be reliable classification models for TR-FTIR spectra analysis for use in Covid-19 diagnosis. Although most studies used nasopharyngeal swabs and blood plasma samples for FTIR spectroscopic analysis of Covid-19, self-collected saliva is the least invasive biofluid collection. In this study, we demonstrated that saliva samples could achieve comparable prediction results with nasopharyngeal (87% sensitivity and 66% specificity [41]; 97% sensitivity and 98.3% specificity [42]) and plasma samples (83.1% sensitivity and 98% specificity [43]; 94.1% sensitivity and 69.2% specificity [44]). Furthermore, saliva collection precludes the use of viral transport medium (VTM), which has shown to have spectral bands at 1578 cm^{-1}, 1408 cm^{-1} and 1078 cm^{-1} and may affect the spectral

quality [16]. The use of transflection instead of ATR-based FTIR spectral acquisition also provides several advantages. In transflectance, the sample is deposited on a reflective substrate and some of the IR beam passes through the surface layer, reflecting off the top layer of the substrate, and then passes through the sample a second time, doubling the pathlength. Therefore, transflection sampling mode results in greater absorbance and less noise compared to ATR. Another advantage is the ability to dry batch samples instead of depositing the sample one by one as in ATR. The use of e-glass slides with dried saliva spots can be stored for future re-analysis if needed. In addition, our team is developing an automated accessory for TR-FTIR spectrometers, that can acquire spectra automatically, allowing on-site testing and mass screening. This spectroscopic method for Covid-19 diagnosis could not replace the RT-qPCR, yet it could serve as a preliminary screening method to minimize the RT-qPCR reagents, which were in limited supply during the height of the pandemic.

6.4.4. Limitation and Future perspective

One limitation of our study is that saliva samples were all obtained from one hospital center, limiting the external validity of the prediction results. Another limitation is the unregulated saliva collection procedure. As no direction was previously given to patients, heterogenous content of saliva collected from different patients may greatly influence the spectra quality, hence the high number of outliers in this study. A further consideration would be to determine a specific cut-off absorbance at 1020-1070 cm^{-1} for outlier removal. Mucus may have played an important role. It has been shown that mucins, the major protein component of mucus, cause peaks at 1040 cm^{-1} due to C-O stretching [30]. A study simply skipped the peak associated with mucin for lung cancer diagnosis using FTIR spectroscopy [30]. Some saliva samples may contain more sputum than others, leading the heterogeneity in FTIR spectra relatable to mucus content in wavenumber range 1020-1070 cm^{-1}. Standardizing saliva collection procedure may dramatically decrease or eliminate the abnormal peak difference at this region.

Notwithstanding the limitations mentioned above, TR-FTIR spectroscopy coupled with the use of KNN, ANN, and SVM algorithms can provide a rapid, low-cost reagent free approach for COVID-19 screening. Up to now, no standard saliva collection method has been established, making FTIR spectroscopy studies with a similar objective complicate to compare. A standard protocol for saliva collection would be optimal to find out the best FTIR sampling method along

with multiple machine learning algorithms for the diagnosis of Covid-19. Besides, the greater the sample size is always better for developing a spectral database to encompass as much diversity as possible. In our study, elderly group were significantly smaller than youngster group. In the future, we may consider collecting more saliva samples from the elderly to alleviate this difference. Future studies could also consider developing models using FTIR spectroscopy for Covid-19 severeness and fatality prediction using different saliva and sputum components to delineate their contribution to the FTIR spectra.

6.5. Conclusion

In this study, we demonstrated the high efficiency (85.9% accuracy, 82.9% precision, 85.4% sensitivity, and 86.3% specificity) of TR-FTIR spectroscopy for the prediction of Covid-19 by KNN, ANN and SVM algorithms using self-collected saliva samples. Biomarker bands were also determined and investigated on raw spectra despite that saliva protein profile was disturbed by heat inactivation. Furthermore, heat inactivation of saliva prior TR-FTIR spectroscopic analysis help avoids any infection of the healthcare personnel performing the diagnosis. The use of saliva for Covid-19 detection also makes the biofluid sample collection easier for health workers. The possibility of automation for TR-FTIR spectroscopy enables workers with only basic skills to conduct the test, which could carry out for rapid on-site screening in public venues or airports. This cost-effective method would be likewise extremely cost-effective to implement in developing countries, as it does not require any reagents or sample preparation.

6.6. Reference

1. Acuti Martellucci, C., et al., *SARS-CoV-2 pandemic: An overview.* Advances in Biological Regulation, 2020. **77**: p. 100736.
2. Organization, W.H. *World Health Organisation Coronavirus Disease (COVID-19) Dashboard*. 2022 [cited 2022; Available from: https://covid19.who.int/.
3. Khurshid, Z., et al., *Role of Salivary Biomarkers in Oral Cancer Detection.* Adv Clin Chem, 2018. **86**: p. 23-70.
4. Abdul Rehman, S., et al., *Role of Salivary Biomarkers in Detection of Cardiovascular Diseases (CVD).* Proteomes, 2017. **5**(3).
5. De Bruyne, S., M.M. Speeckaert, and J.R. Delanghe, *Applications of mid-infrared spectroscopy in the clinical laboratory setting.* Crit Rev Clin Lab Sci, 2018. **55**(1): p. 1-20.
6. Lai, C.K.C. and W. Lam, *Laboratory testing for the diagnosis of COVID-19.* Biochem Biophys Res Commun, 2021. **538**: p. 226-230.
7. Wyllie, A.L., et al., *Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2.* N Engl J Med, 2020. **383**(13): p. 1283-1286.
8. Kaufman, E. and I.B. Lamster, *The Diagnostic Applications of Saliva—A Review.* Critical Reviews in Oral Biology & Medicine, 2002. **13**(2): p. 197-212.
9. Chiappin, S., et al., *Saliva specimen: a new laboratory tool for diagnostic and basic investigation.* Clin Chim Acta, 2007. **383**(1-2): p. 30-40.
10. Mardani, R., et al., *Laboratory Parameters in Detection of COVID-19 Patients with Positive RT-PCR; a Diagnostic Accuracy Study.* Arch Acad Emerg Med, 2020. **8**(1): p. e43.
11. Frediani, J.K., et al., *SARS-CoV-2 reliably detected in frozen saliva samples stored up to one year.* PLOS ONE, 2022. **17**(8): p. e0272971.
12. Pastorino, B., et al., *Heat Inactivation of Different Types of SARS-CoV-2 Samples: What Protocols for Biosafety, Molecular Detection and Serological Diagnostics?* Viruses, 2020. **12**(7).
13. Guhathakurata, S., et al., *A novel approach to predict COVID-19 using support vector machine*. Data Science for COVID-19. 2021:351-64. doi: 10.1016/B978-0-12-824536-1.00014-9. Epub 2021 May 21.
14. Shanbehzadeh, M., R. Nopour, and H. Kazemi-Arpanahi, *Developing an artificial neural network for detecting COVID-19 disease.* J Educ Health Promot, 2022. **11**: p. 2.
15. Theerthagiri, P., et al., *Prediction of COVID-19 Possibilities using KNN Classification Algorithm*. 2020.
16. Wood, B.R., et al., *Infrared Based Saliva Screening Test for COVID-19.* Angew Chem Int Ed Engl, 2021. **60**(31): p. 17102-17107.
17. Nascimento, M.H.C., et al., *Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis.* Anal Chem, 2022. **94**(5): p. 2425-2433.
18. Barauna, V.G., et al., *Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity.* Analytical Chemistry, 2021. **93**(5): p. 2950-2958.
19. Mistek-Morabito, E. and I.K. Lednev, *FT-IR spectroscopy for identification of biological stains for forensic purposes.* Spectroscopy (Santa Monica), 2018. **33**: p. 8-19.
20. Ma, J., B.K. Rubin, and J.A. Voynow, *Mucins, Mucus, and Goblet Cells.* Chest, 2018. **154**(1): p. 169-176.
21. Ben-Aryeh, H., et al., *The salivary flow rate and composition of whole and parotid resting and stimulated saliva in young and old healthy subjects.* Biochem Med Metab Biol, 1986.

36(2): p. 260-5.
22. Nassar, M., et al., *Age-related changes in salivary biomarkers.* Journal of Dental Sciences, 2014. **9**(1): p. 85-90.
23. Denny, P.C., et al., *Age-related changes in mucins from human whole saliva.* J Dent Res, 1991. **70**(10): p. 1320-7.
24. Mungia, R., et al., *Interaction of age and specific saliva component output on caries.* Aging Clin Exp Res, 2008. **20**(6): p. 503-8.
25. Ferreira, I.C.C., et al., *Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR) Spectroscopy Analysis of Saliva for Breast Cancer Diagnosis.* Journal of Oncology, 2020. **2020**: p. 4343590.
26. Bel'skaya, L.V., E.A. Sarf, and N.A. Makarova, *Use of Fourier Transform IR Spectroscopy for the Study of Saliva Composition.* Journal of Applied Spectroscopy, 2018. **85**(3): p. 445-451.
27. Li, Z., Q. Zhang, and X. Zhao, *Performance analysis of K-nearest neighbor, support vector machine, and artificial neural network classifiers for driver drowsiness detection with different road geometries.* International Journal of Distributed Sensor Networks, 2017. **13**(9): p. 1550147717733391.
28. Kazmer, S.T., et al., *Pathophysiological Response to SARS-CoV-2 Infection Detected by Infrared Spectroscopy Enables Rapid and Robust Saliva Screening for COVID-19.* Biomedicines, 2022. **10**(2).
29. Wang, R. and Y. Wang, *Fourier Transform Infrared Spectroscopy in Oral Cancer Diagnosis.* International Journal of Molecular Sciences, 2021. **22**(3): p. 1206.
30. Lewis, P.D., et al., *Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using sputum.* BMC Cancer, 2010. **10**(1): p. 640.
31. Lee, B.-J., et al., *Discrimination and prediction of the origin of Chinese and Korean soybeans using Fourier transform infrared spectrometry (FT-IR) with multivariate statistical analysis.* PLOS ONE, 2018. **13**: p. e0196315.
32. kumar j, K. and A.G.D. Prasad, *Identification and comparison of biomolecules in medicinal plants of Tephrosia tinctoria and Atylosia albicans by using FTIR.* Romanian J Biophys, 2011. **21**.
33. Pokrowiecki, R., et al., *Nanoparticles And Human Saliva: A Step Towards Drug Delivery Systems For Dental And Craniofacial Biomaterials.* Int J Nanomedicine, 2019. **14**: p. 9235-9257.
34. Wang, Q., et al., *UV–Vis and ATR–FTIR spectroscopic investigations of postmortem interval based on the changes in rabbit plasma.* PLOS ONE, 2017. **12**(7): p. e0182161.
35. Mateus, T., et al., *Fourier-Transform Infrared Spectroscopy as a Discriminatory Tool for Myotonic Dystrophy Type 1 Metabolism: A Pilot Study.* Int J Environ Res Public Health, 2021. **18**(7).
36. Muñoz-Prieto, A., et al., *Saliva changes in composition associated to COVID-19: a preliminary study.* Scientific Reports, 2022. **12**(1): p. 10879.
37. Martinez-Cuazitl, A., et al., *ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients.* Scientific Reports, 2021. **11**(1): p. 19980.
38. Song, J.W., et al., *Omics-Driven Systems Interrogation of Metabolic Dysregulation in COVID-19 Pathogenesis.* Cell Metab, 2020. **32**(2): p. 188-202.e5.
39. Brandtzaeg, P., *Secretory immunity with special reference to the oral cavity.* J Oral Microbiol, 2013. **5**.

40. Benezzeddine-Boussaidi, L., G. Cazorla, and A.-M. Melin, *Validation for quantification of immunoglobulins by Fourier transform infrared spectrometry.* Clinical Chemistry and Laboratory Medicine, 2009. **47**(1): p. 83-90.
41. Nogueira, M.S., et al., *Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning.* Scientific Reports, 2021. **11**(1): p. 15409.
42. Kitane, D.L., et al., *A simple and fast spectroscopy-based technique for Covid-19 diagnosis.* Scientific Reports, 2021. **11**(1): p. 16740.
43. Zhang, L., et al., *Fast Screening and Primary Diagnosis of COVID-19 by ATR–FT-IR.* Analytical Chemistry, 2021. **93**(4): p. 2191-2199.
44. Banerjee, A., et al., *Rapid Classification of COVID-19 Severity by ATR-FTIR Spectroscopy of Plasma Samples.* Analytical Chemistry, 2021. **93**(30): p. 10391-10396.

Chapter 7. General Discussion and Conclusion

Rapid and accurate identification of microorganisms is an utmost concern in microbiology and clinical laboratories. Conventional microbial identification relies mainly on traditional biochemical and serological methods. These methods have been used for over a century and are well established; however, interpretation of the results can be time consuming, subjective and may not always allow for identification of the microorganism at the species and strain levels. In clinical microbiology laboratories dealing with large numbers of samples, MALDI-TOF MS has been widely adopted in the past decade owing to its ability to identify microorganisms at species level in a rapid manner. However, closely related bacterial species may be misidentified, and subspecies-level discriminatory capability is generally lacking. For strain-specific or epidemiological analysis, molecular methods based on DNA or RNA fingerprints such as PCR or WGS are usually the first choice. Due to their high accuracy, these methods are considered the 'gold standard'. Nonetheless, many small regional microbiology laboratories do not use genotypically-based identification techniques, as they cannot afford the high cost of analysis. Moreover, by comparison with traditional methods of microbial identification, these methods are more technically challenging to perform and require more expensive equipment and supplies. Genotypic methods are also the gold standard for the detection of viral pathogens and provide much greater sensitivity than rapid immunoassay tests, thereby substantially reducing instances of false-negative results. In this regard, RT-qPCR is the gold standard for detection of SARS-CoV-2, the novel coronavirus causing Covid-19, and serves as the reference method against which all other methods are compared (or trained, as was the case in the development of the FTIR spectroscopic method in Chapter 6 of this thesis).

In the current context of available microbial identification methods, FTIR spectroscopy has been successfully applied for the identification and classification of microorganisms in numerous proof-of-concept studies by researchers worldwide. However, given the lack of public or commercial infrared spectral databases for this application of FTIR spectroscopy, its implementation requires that laboratories develop their own databases and is accordingly still largely restricted to research laboratories. This status is due, at least in part, to its perceived impracticality for routine use stem from the large number of methodological factors that would be subject to interlaboratory variability and could potentially affect the accuracy of the results. By

addressing this issue to a certain extent, the research work undertaken in this thesis aims to contribute toward eventual inclusion of FTIR spectroscopy among the currently accepted techniques, including sampling techniques, common cultivation medium used, and different manufacturers, briefly considered above. Also, with this aim in mind, this thesis focuses on ATR-FTIR and TR-FTIR spectroscopy (rather than the transmission FTIR-based approach employed in a majority of the research publications in this field) because of their respective practical advantages of ease of sample preparation and amenability to automation.

The focus of the work reported in Chapter 3 was the evaluation of identification accuracy impact due to differences in sampling techniques and cultivation medium by constructing a dedicated spectral database of bovine mastitis-related Gram-positive cocci. Infrared spectral features (serving as biomarker equivalents) were identified using spectral search algorithms and used as input for developing robust classification models for discrimination among the bacterial genera and species represented in the spectral database by FTIR spectroscopy. Bacteria were grown on TSA and CBA and their ATR-FTIR and TR-FTIR spectra acquired. Preliminary results revealed three mislabelled *S. aureus* strains that were actually CoNS, showing the strong differentiating competency of FTIR spectroscopy. The prediction results demonstrated that both ATR-FTIR and TR-FTIR are promising candidates for identification at bovine mastitis pathogens at both the genus and the species level. Identification of important CoNS species and *S. aureus* by FTIR spectroscopy yielded high rates of correct identification. Overall, growth of bacteria on CBA coupled with TR-FTIR yielded the highest identification accuracy at species level but performed less well than the combination of growth on TSA and ATR-FTIR for identification at genus level. This may be due to the longer effective pathlength of TR-FTIR across the full spectral range, as opposed to the increasingly short effective pathlength with increasing wavenumber in ATR-FTIR spectral acquisition. As a consequence, TR-FTIR spectra tend to have a generally higher absorbance and contain more biochemical information than ATR-FTIR spectra, making them very useful for species level identification. However, the extensive biochemical information that TR spectra encompass appears to have detrimental effects on the stepwise identification process for genus and *S. aureus* vs CoNS identification. While the amenability of TR-FTIR to automation and the possibility of archiving samples on the low-e glass slides used to acquire TR-FTIR spectra are advantageous for high sample throughputs, ATR-FTIR is simpler and faster as it does not require sample deposition or drying. This chapter also combined databases built using different culture

media. Since the sample preparation method, including the growth medium employed, could greatly affect bacteria growth and FTIR spectral profile, the spectrum of a bacterial strain may differ when using different sample preparation methodologies and grown on different media, which may affect its identification by FTIR spectroscopy. However, with a large enough spectral database comprising spectra of microbial strains grown on common culture media, accurate identification of *Staphylococcus* spp. and *Streptococcus* spp. will be achieved with either choice of agar medium. In general terms, it is always recommended to use the same growth medium as employed in developing the database for accurate identification, but identifying unknowns at genus level against a database containing the spectra of bacteria grown on different media may be a possible option if the end users of the methodology for which the database is designed are likely to resist changing their standard protocols to comply with an arbitrary growth protocol in order to adopt FTIR spectroscopy for routine microbial identification.

Many commercial FTIR instruments are available on the market, manufactured by different companies. Due to the availability and price, not all laboratories may have purchased the same brand. This raises the question of how well FTIR spectroscopy will perform on bacteria identification if unknowns were recorded by one instrument, while the database was created on another. For this reason, Chapter 4 evaluated two common FTIR instruments for the identification of common foodborne pathogens. Spectral features for identification of foodborne pathogens were selected to create a database on a specific FTIR instrument. HCA of FTIR spectra reflected the heterogeneity of species, which was concordant with the genomics. Then, two test sets were obtained, one test set from the same FTIR instrument as used for database creation, and one test set from a different instrument. Overall, FTIR spectroscopy maintained its efficacy in the identification of pathogenic foodborne bacteria at the genus level no matter the instrument used. Yet the change in FTIR instruments may make the species prediction results doubtful. Based on these preliminary findings with two instruments from different manufacturers, laboratories using different FTIR instruments would be able to share a common database if only genus level identification is needed, for instance, when FTIR spectroscopy is employed as an alternative to biochemical tests or potentially as a preliminary screening technique prior to the use of a more costly genotypic method for microbial identification.

Bacteria identification using FTIR spectroscopy has been evaluated previously by many researchers but the promising results that were obtained have not translated into practical applications in routine microbiology laboratories. The previous chapters contributed to the demonstration of the robustness of FTIR spectral databases, aiming at expansion of the applicability of FTIR spectroscopy in microbiology laboratories. In Chapter 5, the focus was on filamentous fungi. Traditional identification of filamentous fungi is based on macroscopic and microscopic methods. Genotypic methods, although considered as the gold standard for identification, are employed mainly for the characterization of fungi due to the tedious preparation procedure. FTIR spectroscopy was compared against multiplex RT-qPCR and MALDI-TOF MS for the identification of *Aspergillus* species. While promising results for identifying section *Nigri* and *Terrei* of *Aspergillus* spp. using multiplex RT-qPCR was demonstrated, the other sections investigated did not show optimal results. Identification using MALDI-TOF MS was also problematic, despite using a specially formulated agar plate for enhancing the identification capacity of filamentous fungi. Although gaining more attention, identification of molds such as *Aspergillus* spp. is rarely performed routinely in most microbiology laboratories, and even less often at species level. Hence, RT-qPCR sequence and MALDI-TOF MS databases are considerably less well developed for molds as compared to bacteria and yeasts, and this may explain the low identification rates obtained with these methods. On the other hand, with a relatively small database, FTIR spectroscopy was still capable to yield high identification accuracy at species level. The second part of this chapter enlarged the fungal database with more filamentous fungi and yeasts. In the test set, yeast isolates were correctly identified with 100% correct identification, with only several mold strains misidentified. The number of samples subjected to fungi identification analysis could be significantly improved by using FTIR spectroscopy without compromising the identification performance and being a cheaper method compared to genotypic methods.

In the past decade, increasing studies of disease diagnosis using FTIR spectroscopy have been reported. With the sudden emergence of Covid-19, Chapter 6 used this opportunity to evaluate the diagnosis of this novel disease by FTIR spectroscopy using saliva samples. As this virus is brand new, every step of the protocol was developed and constantly edited and improved as the research progressed. For safety purpose and for evaluating the robustness of FTIR spectroscopy, saliva samples were heat inactivated prior to spectral acquisition. The drawback of

the heating procedure is that it may inevitably alter protein composition of saliva. Nevertheless, FTIR spectroscopy coupled with efficient machine learning algorithms allowed high diagnostic accuracy, sensitivity, and specificity for Covid-19. Spectral regions discriminating Covid-19 negative from Covid-19 positive saliva were also found although they were divergent from those identified in other similar studies. Differences in terms of prediction rate is also observed from other papers. These discrepancies may be mainly due to the diverse methodologies employed. No standardized protocol exists yet for this novel virus. Additionally, other studies used either raw or ethanol-inactivated saliva, which causes great difference in saliva composition that are reflected in the spectra. All that said, FTIR spectroscopy has demonstrated its possibility to be employed as a diagnostic or screening tool for Covid-19.

This research demonstrated the applicability of FTIR spectroscopy for routine identification of bovine mastitis pathogens in veterinary microbiology and bacterial pathogens of relevance to food safety as well as the diagnosis of foodborne illnesses in clinical microbiology. FTIR spectra are rich in information, and hence providing all the necessary bands for correct identification to species and even strain level. However, this feature also makes using and interpreting their spectra challenging. Currently, pure microbial colony must be isolated in order to obtain a promising prediction result. Mixtures of different compounds cause absorption peaks to overlap, hence complicating the identification process. As seen in Chapter 6, the prediction of raw biospecimen (saliva) collected was complicated due to unanticipated peak, disturbing and mixing the PP and PN spectra. The identification value at the end was therefore less appealing than those from the previous chapters, where only pure colonies were used for spectra acquisition. The few algorithms investigated in Chapter 6 boosted the identification of saliva mixture up to an acceptable value. Other algorithms may possibly provide insights in the identification of specific features within a mixture using FTIR spectroscopy. Implementing algorithms within the database to generate identification results automatically would also be a huge advancement for FTIR spectroscopy in the application of microbial identification, and will eventually contribute in its wilder applicability. Another limitation of FTIR spectroscopy is its high detection limit. It requires a certain amount of biomass or thickness to be able to generate a high quality spectra. Water, for example, will absorb infrared light and may interfere with the analysis of wet samples. In the case of Chapter 3 and 4, some watery bacterial colony would be highly difficult to identify, as the peak absorbance were lower than our pre-set acceptance criteria. They will be considered as outliers

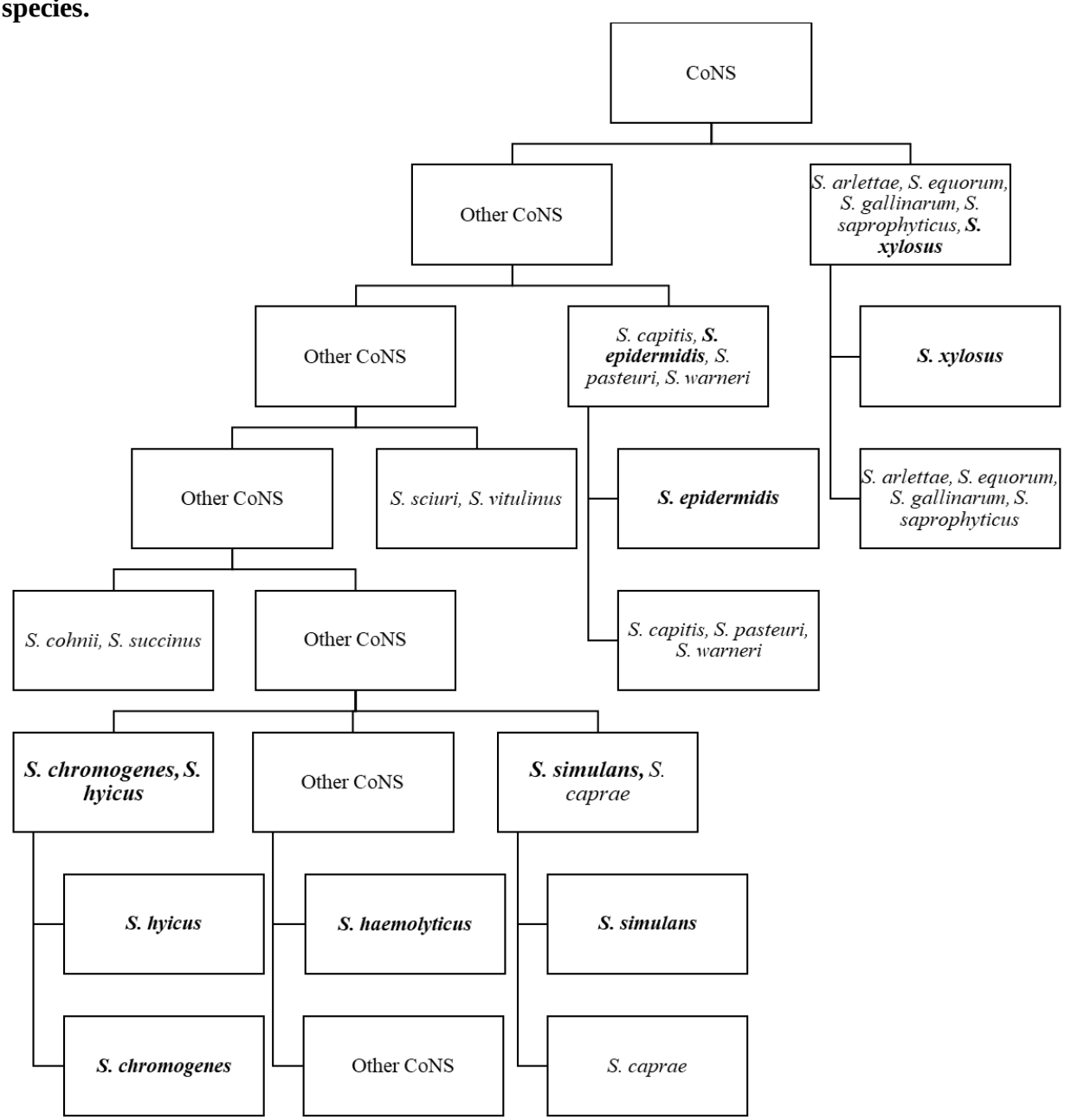
and rejected. Identification by TR-FTIR is less biased by water disturbance as the water is already dried out on the e-glass, leaving only the biomass. Identification of bacteria was simple using either ATR or TR sampling technique. Yet, due to the spore-forming characteristic of mold, identification of filamentous mold required few additional steps to ensure lab personnel safety during handling. And since FTIR spectrum records all biochemical constitution, the sample preparation steps may influenced the spectrum. There are different sampling techniques in the existing literature, ranging from incubation condition (time, medium, and temperature) to preparation of the mold pellet (distilled water and ethanol). Different sample preparation techniques for mold may have altered its identification results in MALDI-TOF MS and FTIR spectroscopy. The comparison results obtained in Chapter 5 may support the fact that FTIR spectroscopy is outperforming MALDI-TOF MS and multiplex RT-PCR for identification of *Aspergillus* species using the presented sample preparation method. Yet, other sampling techniques for mold are also worth experimenting for identification. Simplification of the methodology for mold identification using FTIR spectroscopy could be another possible future research direction, as it was quite more tedious in preparation for mold compared to bacteria. While Chapter 5 focused on *Aspergillus* spp., the identification of different mold species has gained attention. Other common mold genus, such as *Penicillium* spp., *Mucor* spp., and *Fusarium* spp. may also raise interest in evaluating the identification efficiency comparison of PCR, MALDI-TOF MS and FTIR spectroscopy. In the research for saliva, sample preparation was again a limitation that could be mitigated in future projects. As no prior instruction, collected saliva specimen were heterogenous in all manners. It was not easy to analyze data without a traceable pattern, and to figure out the noise peaks and the useful ones. But because that no instruction were given, we had the opportunity to study a real-world scenario of differentiating saliva mixture and evaluate the robustness of FTIR spectroscopy to distinguish PP and PN within complex biochemical mixture. Despite FTIR spectroscopy takes up all information from a spectrum, it was still able to find apparent biomarkers to predict infected and uninfected saliva. The heating process for saliva minimizes cross-infection risk for laboratory personnel. Yet, heating may cause conformational change of proteins in the saliva. Subsequent research can focus on the direct use of raw saliva onto reflective substrate for TR-FTIR spectral acquisition. Last but not least, enlarging the spectral database would ultimately increase the robustness of identification by FTIR spectroscopy. A database created using two common cultivation media and different

manufacturers could be possible to accommodate with demands of different laboratories. However, to ensure species level identification, the medium and FTIR instrument same as the database should be used. Since it works already perfectly with a small database and it is easy to create, FTIR spectroscopy is employed in many research laboratory. Nonetheless, as seen as Bruker's database and Charles River's database for MALDI-TOF MS, different databases can be created depending on the needs of different laboratories. Currently, no commercialized database exists for FTIR spectroscopy, therefore limiting its use as routine identification methods in many clinical and industrial laboratories. A standardized database for FTIR spectroscopy should be developed, validated, and officialized.

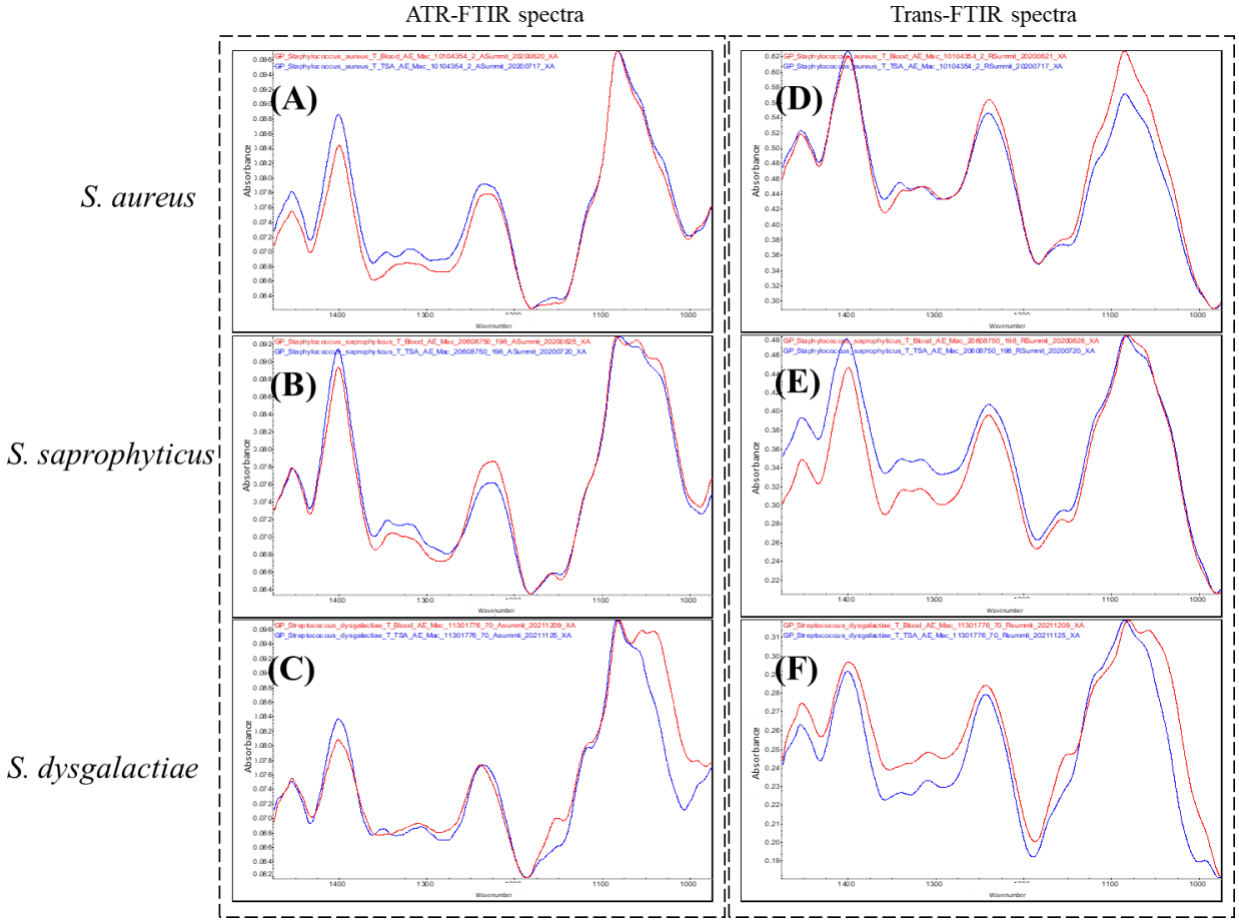
The results of experiments in this research performed to evaluate methodological variations in terms of culture medium, sampling method, and type of FTIR instrument showed that the FTIR methodology may be sufficiently robust to accommodate a certain amount of latitude, as would be required for interlaboratory use of a single database for microbial identification. FTIR spectroscopy was also compared with RT-qPCR and MALDI-TOF MS for the identification of filamentous fungi and yeasts and proved to have noticeable advantages over the other two methods. Lastly, FTIR spectroscopy showed promising sensitivity and specificity for Covid-19 diagnosis with the use of saliva samples, as would be suitable for on-site rapid screening in public venues or airports. Unlike most molecular methods, portable FTIR instruments can be easily accommodated in most laboratories due to their cost-effectiveness, ease of use, and compact design. This microbial identification technology would be likewise amenable to implementation in third-world countries, as it does not require costly reagents. However, as no universal FTIR spectral library exists for microorganisms, a commercial spectral database encompassing the diversity of species encountered in the general routine laboratory, together with a standard protocol for microbial identification with the use of the database, needs to be developed for this method to be widely applicable. Regular updates of the database to encompass spectral diversity resulting from mutations of the strains represented as well as to incorporate additional species would be straightforward if a stepwise identification process conducted in a pairwise manner, as demonstrated in this thesis, were adopted. All that said, this high-throughput and cost-effective technique for microbial identification has the potential to provide a good alternative to molecular methods for routine analysis.

Appendix

A. 1. Structure of FTIR spectral databases for the identification of the six prevelant CoNS species.



A. 2. Comparison of CBA- (red) and TSA-grown (blue) spectra of same S. aureus (A, D), S. saprophyticus (B, E), and S. dysgalactiae (C, F) isolates. Spectra variability of identical isolate due to difference in growth medium over the broad wavenumber region 900-1500 cm^{-1} can be visualized in both ATR-FTIR spectra (A, B, C) and Trans-FTIR spectra (D, E, F).



A. 3. Ingredient list of TSA and CBA.

Growth medium	Ingredient	Amoung (g/L)
BD Difco Tryptic Soy agar	Pancreatic Digest of Casein	15
	Papaic Digest of Soybean Meal	5
	Sodium Chloride	5
	Agar	15
Oxoid Columbia Agar with 5% Sheep Blood	Pancreatic Digest of Casein	12
	Peptic Digest of Animal Tissue	5
	Yeast Extract	3
	Beef Extract	3
	Corn Starch	1
	Sodium Chloride	5
	Agar	13.5
	Sheep Blood, Defibrinated	5%

A. 4. List of patients and their corresponding characteristics and PCR results.

Database X										
	Patient characterisitcs		PCR results				Genetic analysis			Notes
PCR positive (PP) or Healthy (PN)	Gender	Age	Gene E	Gene N	Gene RDRP	Internal Control	N51Y	H69/V7	E484K	
PN	M	8				24.8				
PN	M	58				20.84				
PN	M	11				24.99				
PN	M	8				24.28				
PN	M	50				24.91				
PN	M	8				25.2				
PN	M	57				22.17				
PN	M	12				24.27				
PN	M	11				22.45				
PN	M	9				23.26				
PN	M	15				23.56				
PN	M	4				22.85				
PN	M	13				24.13				
PN	M	8				24.9				
PN	M	10				25.77				
PN	M	12				23.16				
PN	M	26				20.68				
PN	M	41				22.4				
PN	M	10				19.37				
PN	M	32				24.17				
PN	M	35				23.6				
PN	M	9				24.63				
PN	M	6				22.41				
PN	M	38				24.28				
PN	M	30				22.54				

PN	M	35	21.76	
PN	M	12	25.55	
PN	M	6	26.9	
PN	M	57	23.15	
PN	M	40	21.93	
PN	M	41	23.14	
PN	M	62	22.9	
PN	M	91	25.2	
PN	M	23	23.93	
PN	M	29	22.26	
PN	M	39	19.71	
PN	M	10	24.2	
PN	M	27	21.7	
PN	M	31	22.83	
PN	M	4	23.48	
PN	M	9	22.95	
PN	M	7	22.54	
PN	M	39	24.22	
PN	M	6	25.16	
PN	M	43	18.35	
PN	M	10	23.4	
PN	M	58	23.4	
PN	M	25	23.79	
PN	M	20	22.21	
PN	M	8	22.14	
PN	M	53	22.64	
PN	M	9	23.57	
PN	M	6	17.58	
PN	M	38	19.97	

PN	M	12	20.87	
PN	M	37	21.39	
PN	M	8	22.13	
PN	M	21	24.93	
PN	M	33	22.59	
PN	M	8	24.26	
PN	M	36	20.75	
PN	M	57	19.53	
PN	M	40	22.68	
PN	M	11	22.41	
PN	M	28	22.85	
PN	M	32	22.1	
PN	M	47	20.85	
PN	M	9	20.77	
PN	M	42	26.87	
PN	M	9	24.76	
PN	M	5	21.11	
PN	M	7	22.54	
PN	M	11	23.1	
PN	M	7	24.53	
PN	M	45	22.2	
PN	M	6	28.5	
PN	M	56	25.67	
PN	M	44	22.78	
PN	M	58	23.17	
PN	M	8	26.7	
PN	M	6	21.74	
PN	M	35	24.36	
PN	M	63	23.78	

PN	M	14	23.96		
PN	M	47	24.57		
PN	M	35	23.11		
PN	M	41	21.33		
PN	M	45	26.37		
PN	M	31	24.25		
PN	M	52	23.55		
PN	M	8	25.4		
PN	M	45	23.5		
PN	M	40	23.89		
PN	M	27	22.8		
PN	M	34	21.29		
PN	M	11	24.47		
PN	M	10	19.19		
PN	M	8	27.3		
PN	M	10	23.91		
PN	M	7	25		
PN	M	20	23.18		
PN	M	52	24.63		
PN	M	11	24.13		
PN	M	5	24.3		
PN	M	33	25.46		
PN	M	15	22.93		
PN	M	39	22.5		
PN	M	34	25.2		
PN	M	10	24.7		
PN	M	3	24.46		
PN	M	34	21.74		
PN	M	3	23.22		

PN	M	31	24.83	
PN	M	38	21.8	
PN	M	14	25.1	
PN	M	46	24.71	
PN	M	2	24.42	
PN	M	41	21.16	
PN	M	13	24.83	
PN	M	7	20.9	
PN	M	7	20.9	
PN	M	54	22.18	
PN	M	46	25.83	
PN	M	35	24.74	
PN	M	38	26.91	
PN	M	4	31.26	
PN	M	36	23.2	
PN	M	48	25.67	
PN	M	24	22.18	
PN	M	22	23.11	
PN	M	43	23.45	
PN	M	36	23.46	
PN	M	22	21.73	
PN	M	21	25.18	
PN	M	11	25.25	
PN	M	56	21.92	
PN	M	9	25.88	
PN	M	4	23.35	
PN	M	21	23.49	
PN	M	11	28.7	
PN	M	51	22.45	

PN	M	13	25.71	
PN	M	9	23.54	
PN	M	15	24.59	
PN	M	6	23.73	
PN	M	8	25.6	
PN	M	8	23.93	
PN	M	18	27.99	
PN	M	12	26.36	
PN	M	11	26.6	
PN	M	37	25.5	
PN	M	5	25.47	
PN	M	3	28.11	
PN	M	11	27.32	
PN	M	10	27.59	
PN	M	13	26.33	
PN	M	7	25.58	
PN	M	18	25.26	
PN	M	13	26.8	
PN	M	55	26.63	
PN	M	39	26	
PN	M	8	26.3	
PN	M	8	26.56	
PN	M	6	29.77	
PN	M	33	24.27	
PN	M	10	21.33	
PN	M	12	27.53	
PN	M	7	24.81	
PN	M	45	25.48	
PN	M	41	24.92	

PN	M	12	25.84	
PN	M	43	21.85	
PN	M	33	26.63	
PN	M	30	25.87	
PN	M	8	24.58	
PN	M	5	23.24	
PN	M	6	26.23	
PN	M	38	26.9	
PN	M	6	26.76	
PN	M	49	27.4	
PN	M	4	25.76	
PN	M	3	24.85	
PN	M	33	24.39	
PN	M	39	22.78	
PN	M	4	25.44	
PN	M	7	22.99	
PN	M	9	23.35	
PN	M	50	26.3	
PN	M	27	23.95	
PN	M	11	24.65	
PN	M	6	24.36	
PN	M	4	26.2	
PN	M	7	24.61	
PN	M	61	22.42	
PN	M	8	22.96	
PN	M	3	26.67	
PN	M	11	24.3	
PN	M	6	21.78	
PN	M	5	25.88	

PN	M	7				26.18				
PN	M	11				24.2				
PN	M	9				24.8				
PN	M	55				24.68				
PN	M	7				23.6				
PN	M	11				32.75				
PN	M	24				26.73				
PN	M	6				27.26				
PN	M	12				23.73				
PN	M	10				25.23				
PN	M	60				28.2				
PN	M	80				25.3				
PN	M	7				24.85				
PN	M	6				25.8				
PN	M	43				31.6				
PP	M	10	31.76	33.55	34.24	22.44	ND			
PP	M	22	23.7	25.55	26.69	21.59	ND			
PP	M	22	23.7	25.55	26.69	21.59	ND			
PP	M	37	16.41	19.5	19.44	21.48	D			
PP	M	10	31.74	33.35	33.17	24.27	ND			
PP	M	54	14.75	17.23	17.52	25.12	ND			
PP	M	34	11.33	13	15.18	25.65	D	D	ND	
PP	M	31	12.27	14.5	16.47	21.55	D	D	ND	B.1.1.7
PP	M	44	12.36	14.6	17.2	23.62	D	D	ND	B.1.1.7
PP	M	32	13.18	15.29	16.38	25.5	D	D	ND	B.1.1.7
PP	M	31	13.74	16	18.91	20.91	D	D	ND	
PP	M	36	14.24	15.55	17.23	22.94	D	D	ND	
PP	M	16	18.77	21.2	22.17	21.11	ND			B.1.1.7
PP	M	72	16.82	16.97	29.5	26.2	D	D		B.1.1.7

PP	M	34	15.6	16.57	19.8	21.4	D	D	ND	B.1.1.7
PP	M	58	15.9	17.4	21.38	20.64	D	D	ND	
PP	M	7	14.92	17.13	17.98	20.1	D	D	ND	
PP	M	20	13.92	16.53	17.32	23.32	D	D	ND	
PP	M	42	14.89	18.1	19.69	20.52	ND	ND		
PP	M	18	17.15	17.99	20.39	23.88	D	D	ND	B.1.1.7
PP	M	26	16.75	18.29	18.88	19.99	ND	ND	ND	B.1.1.7
PP	M	46	17.58	19.3	27.26	27.41	D	D	ND	
PP	M	58	18.31	18.51	29.93	20.48	D	D	ND	
PP	M	63	18.2	20.14	25.21	22.8	ND	ND		
PP	M	69	18.63	20.11	25.3	20.77	ND	ND	ND	B.1.1.7
PP	M	58	18.9	20.13	23.39	26.16	ND	ND	ND	B.1.1.7
PP	M	54	18.42	20.3	2.97	22.39	D	D	ND	
PP	M	26	17.94	20.7	22.21	24.4	D	D	ND	
PP	M	49	18.11	19.93	21.39	22.15	D	D	ND	
PP	M	34	18.44	20.32	22.9	25.19	D	D	ND	B.1.1.7
PP	M	23	18.95	20.42	22.23	23.95	D	D	ND	B.1.1.7
PP	M	49	18.41	20.9	21.7	25.49	D	D	ND	
PP	M	11	22.43	24.48	25.54	24.55	ND			
PP	M	47	19.2	21.34	22.35	24.72	ND	ND		
PP	M	11	19.42	20.6	23.4	22.39	ND		D	B.1.1.7
PP	M	45	19.17	20.53	23.37	22.9	D	D	ND	B.1.1.7
PP	M	5	19.45	21.4	21.64	22.77	ND	ND	ND	
PP	M	38	18.7	20.83	22.13	20.76	ND	ND	ND	
PP	M	24	19.35	20.53		22.99				
PP	M	38	20.18	21.34	23.3	23.95	D	D	ND	
PP	M	37	19.6	21.33	21.41	23.73	D	ND	D	
PP	M	43	21.63	21.26	34.98	23.25	D	D	ND	B.1.1.7
PP	M	7	24.15	26.18	26.64	21.88	D			

PP	M	65	20.71	22.9	25.88	21.73	D	D		
PP	M	50	20.41	21.69	24.19	25.7	D	D		
PP	M	56	19.54	21.94	24.58	21.39	D	D		
PP	M	30	19.71	21.64	23.74	23.9	ND	ND	ND	
PP	M	63	22.6	22.47	26.25	24.68	ND	ND	ND	
PP	M	44	19.8	21.87	23.26	23.8	D	D	ND	
PP	M	65	20.99	22.29	23.87	22.49	D	D	ND	
PP	M	17	19.66	21.65	22.86	24.4	ND	ND	ND	B.1.1.7
PP	M	13	20.85	22.44	23.6	24.67	ND	ND	ND	
PP	M	59	20.46	21.71	23.87	24.73	D	D	ND	
PP	M	65	21.1	23.23	25.35	22.82	ND	ND		B.1.1.7
PP	M	46	21.71	23.33	25.29	21.23	D	D		
PP	M	24	21.18	22.71	24.4	24.23	D	D		B.1.1.7
PP	M	14	20.67	22.83	23.6	24.44	ND	ND	ND	B.1.1.7
PP	M	35	20.85	22.84	23.39	24.88	D	ND	D	B.1.1.7
PP	M	40	21.67	22.94	27.38	19.45	D	D	ND	B.1.1.7
PP	M	30	21.61	23.34	26.73	21.58	D	D	ND	
PP	M	38	19.94	23.24	22.73	23.43	D	D	ND	
PP	M	38	21.34	22.58	23.81	25.9	D	D	ND	
PP	M	43	20.52	23.14	22.59	24.32	D	ND	D	
PP	M	39	24.31	26.67	27.39	25.38	ND			
PP	M	40	22.2	24.33	26.19	22.76	D	D		
PP	M	16	21.14	23.82	24.8	21.62	D	D		
PP	M	53	20.77	23.97	24.44	23.18	D	ND	D	
PP	M	60	21.47	23.8	25.5	21.91	D	D		
PP	M	66	22.37	23.77	28.89	23.41	D	D	ND	
PP	M	57	22.4	24.45	26.54	22.75	D	D	ND	B.1.1.7
PP	M	28	22.49	24.41	26.11	22.44	D	D	ND	
PP	M	46	22.17	23.94	25.92	21.31	D	D	ND	B.1.1.7

PP	M	20	22.48	25.7	25.82	24.15	D	D		
PP	M	35	24.55	25.39	29.62	21.3	ND	ND	ND	B.1.1.7
PP	M	34	23.71	25.42	26.69	22.88	D	ND	D	
PP	M	38	22.79	25.9	26.6	22.9	D	ND	D	B.1.1.7
PP	M	30	22.78	24.82	26.81	24.23	D	D	ND	B.1.1.7
PP	M	47	23.28	24.76	28.12	23.67	ND	ND	ND	B.1.1.7
PP	M	27	22.86	24.73	26.2	23.59	D	D	ND	B.1.1.7
PP	M	37	22.53	24.74	26.4	25.32	D	D	ND	
PP	M	16	22.98	24.79	26.53	23.92	D	D	ND	
PP	M	51	24.6	25.33	27.75	24.7	D	D	ND	
PP	M	50	23.97	25.49	27.19	22.59	D	D	ND	B.1.1.7
PP	M	24	21.74	24.56	27.92	24.78	D	D	ND	
PP	M	55	22.19	24.62	26.26	25.84	D	D	ND	
PP	M	61	23.88	25.81	27.81	22.41	ND	ND	ND	
PP	M	64	28.5	25.53		24.1	D	D		
PP	M	10	24.1	25.92	27.29	21.96	ND	ND	ND	B.1.1.7
PP	M	50	24.63	26.3	31.69	19.75	ND	ND	ND	
PP	M	18	23.88	26.26	26.45	24.26	ND	ND	ND	
PP	M	44	23.79	26.3	27.72	24.44	D	ND	D	B.1.1.7
PP	M	13	24.51	26.34	28.91	25.52	D	D	ND	
PP	M	38	24.63	26.31	29.8	20.88	D	D	ND	
PP	M	68	23.35	25.51	26.17	24.55	D	D	ND	
PP	M	38	24.93	26.17	27.72	22.41	ND	ND	ND	
PP	M	8	24.1	26.24	27.46	22.93	D	D	ND	
PP	M	22	24.5	25.94	26.99	26.5	D	D	ND	B.1.1.7
PP	M	11	24.55	26.44	28.2	26.2	D	D	ND	
PP	M	25	24.38	26.39	27.71	24.6	ND	ND	ND	
PP	M	23	24.36	25.64	27.22	21.74	D	D	ND	
PP	M	31	24.92	25.52	29.61	25.47	D	D	ND	

PP	M	39	20.28	22.4	24.7	25.2	D	D		
PP	M	49	25.4	27.39	28.71	23.35	ND	ND		B.1.1.7
PP	M	13	25.5	26.83	30.8	22.54	D	D		
PP	M	52	25.11	27.3	28.38	21.64	ND	ND	ND	B.1.1.7
PP	M	16	25.66	27.93	31.38	22.12	D	D		
PP	M	8	24.92	26.71	27.35	25.47	ND	ND	ND	
PP	M	41	24.55	26.55	26.61	25.59	ND	ND	ND	B.1.1.7
PP	M	47	25.5	26.54	29.39	23.63	D	D	ND	B.1.1.7
PP	M	55	25.68	26.82	30.17	25.15	D	D	ND	
PP	M	30	24.69	27.8	27.82	23.8	D	D	ND	
PP	M	10	24.61	26.54	28.97	26.93	D	D	ND	
PP	M	5	25.89	27.38	27.94	24.83	D	D	ND	
PP	M	31	24.7	26.94	26.72	27.82	D	D	ND	
PP	M	52	25.44	26.96	30.15	24.46	D	D	ND	
PP	M	60	24.25	26.77	27.72	25.64	D	D	ND	
PP	M	45	25	26.85	29.9	25.86	D	D	ND	B.1.1.7
PP	M	20	25.76	27.59	28.36	25.99	D	D	ND	
PP	M	15	26.49	27.97	29.85	23.84	ND	ND	ND	
PP	M	36	25.36	27.77	29.44	25.5	D	D	ND	
PP	M	51	26.86	28.45	30.85	24.94	D	D	ND	
PP	M	13	25.82	27.75	28.59	26.6	D	D	ND	
PP	M	20	25.76	28.3	27.85	25.9	ND	ND	ND	
PP	M	23	26.33	27.85	28.79	24.9	D	D	ND	B.1.1.7
PP	M	45	25.95	27.95	29.9	2.61	D	D	ND	
PP	M	30	27.79	29.15	31.2	24.31	D	D		
PP	M	64	27.13	28.66	33.12	28.15	D	D	ND	
PP	M	41	27.9	29.9	30.81	24.44	D	D	ND	B.1.1.7
PP	M	9	27.7	29.37	30.79	23.91	ND	ND	ND	
PP	M	57	26.97		30.77	20.3	ND	ND	ND	B.1.1.7

PP	M	57	27.41	28.8	29.93	26.4	ND	ND	ND	
PP	M	23	27.39	29.11	31.95	26.41	D	D	ND	B.1.1.7
PP	M	33	26.41	28.64	30.68	24.11	D	D	ND	
PP	M	44	27.65	29.12	30.45	21.3	ND	ND	ND	
PP	M	23	27.3	28.87	29.41	25.76	D	D	ND	
PP	M	61	29.51	29.27		21.2	D	D	ND	
PP	M	5	32.61	34.46	36.41	21.12	ND			
PP	M	19	28.23	30.32	31.52	27.63	D	D	ND	
PP	M	11	27.91	30.22	31.6	23.38	D	D	ND	
PP	M	42	27.69	30.18	3.83	25.52	D	D	ND	
PP	M	58	27.53	29.9		23.5	D	D	ND	B.1.1.7
PP	M	8	27.63	30.3	33.28	26.81	D	ND	ND	
PP	M	16	29.17	30.9	35.96	21.66	D	D	ND	
PP	M	31	30.1	31.43		25.47				
PP	M	43	27.22	30.68	30.95	24.8	D	D		
PP	M	11	30.86	31.13	39.1	17.33				B.1.1.7
PP	M	33	28.58	30.61	31.3	22.12	D	D	ND	
PP	M	21	28.84	31.38	32.36	24.37	D	D	ND	
PP	M	54	30.2	31.23	35.93	25.43	D	D	ND	
PP	M	35	21.33	23.6	25.48	25.56	ND			
PP	M	56	29.25	31.65	33.33	23.94				
PP	M	6	32.18	32.39	35.98	24.25				
PP	M	33	30.9	31.7	32.93	23.65	D	ND	ND	
PP	M	60	31.8	31.73		23.84	ND	ND	ND	
PP	M	34	30.17	32.37	33.39	24.6	ND	ND	ND	
PP	M	22	29.34	32.25	32.76	23.89				
PP	M	57	29.69	31.75	32.9	22.4				
PP	M	29	30.39	31.81	34.11	23.65	ND	ND	ND	
PP	M	18	30.72	32.47		28.3	ND	ND	ND	

PP	M	50	30.75	33.2		19.7				
PP	M	27	29.28	32.55	34.1	21.11				
PP	M	23	29.86	32.76	35.18	23.85	D	ND	ND	
PP	M	19	31.49	33.31	36.41	24.74				
PP	M	63	30.88	32.54	35.33	24.85				
PP	M	35	30.92	32.55	33.15	25.97	ND	ND	ND	
PP	M	31	31.76	33.85	34.82	27.16				
PP	M	30	32.2	33.55	34.85	27.86	D			B.1.1.7
PP	M	37		24.47		30.12				
PP	M	30	32.33	34.66	36.51	24.58				
PP	M	46	34.61	35.26	37.51	21.53				
PP	M	20	33.69	35.38		22.81				B.1.1.7
PP	M	66	34.9	34.86	38.43	22.83				B.1.1.7
PP	M	8	33.61	34.84	36.79	25.38				
PP	M	59	34.24	34.91		26.26				
PP	M	14	34.4	35.83	36.69	25.81				
PP	M	37	34.45	36.1		21.81				
PP	M	27		36.6		18.5				
PP	M	5		36.44		27.39				B.1.1.7
PP	M	11								
PP	M	58		36.64		20.28				
PP	M	44	37.1	37.4	37.91	22.8				
PP	M	21	36.4	37.13	38.51	22.77				
PP	M	36	37.72	37.76		22.36				
PP	M	42	36.53	37.51	35.81	25.6				B.1.1.7
PP	M	6		38.5		23.65				
PP	M	11	27.91	30.22	31.6	23.38	D	D	ND	
PN	F	33				25.18				
PN	F	8				18.32				

PN	F	31	21.62	
PN	F	55	22.53	
PN	F	43	24.61	
PN	F	48	23.31	
PN	F	38	24.39	
PN	F	43	21.9	
PN	F	9	21.7	
PN	F	15	23.43	
PN	F	36	23.5	
PN	F	31	22.72	
PN	F	17		
PN	F	14	24.16	
PN	F	14	21.97	
PN	F	15	22.92	
PN	F	15	22.74	
PN	F	33	25.21	
PN	F	33	19.94	
PN	F	16	26.72	
PN	F	22	23.4	
PN	F	23	23.44	
PN	F	6	25.13	
PN	F	44	24.2	
PN	F	52	26.52	
PN	F	49	20.68	
PN	F	29	25.12	
PN	F	52	22.2	
PN	F	31	21.5	
PN	F	5	23.17	
PN	F	10	25.17	

PN	F	89	24.61	
PN	F	32	24.16	
PN	F	46	22.73	
PN	F	7	24.3	
PN	F	38	21.11	
PN	F	7	22.8	
PN	F	42	21.8	
PN	F	37	21.79	
PN	F	7	23.82	
PN	F	31	22.61	
PN	F	43	24.32	
PN	F	5	22.47	
PN	F	11	22.39	
PN	F	6	23.19	
PN	F	36	24.14	
PN	F	34	21.19	
PN	F	6	24.73	
PN	F	10	21.29	
PN	F	37	22.61	
PN	F	14	21.68	
PN	F	42	21.56	
PN	F	11	23.6	
PN	F	8	22.52	
PN	F	60	23.27	
PN	F	9	26.29	
PN	F	55	23.42	
PN	F	25	21.68	
PN	F	44	20.75	
PN	F	8	21.24	

PN	F	36	18.27	
PN	F	50	21.85	
PN	F	10	21.53	
PN	F	7	22.65	
PN	F	31	24.54	
PN	F	10	23.3	
PN	F	57	21.2	
PN	F	7	23.45	
PN	F	4	22.65	
PN	F	50	23.9	
PN	F	8	23.83	
PN	F	30		
PN	F	38	24.35	
PN	F	89	25.9	
PN	F	7	24.89	
PN	F	5	17.87	
PN	F	41	25.95	
PN	F	17	26.16	
PN	F	39	20.91	
PN	F	13	23.11	
PN	F	54	22.71	
PN	F	41	20.64	
PN	F	14	24.51	
PN	F	30	24.4	
PN	F	14	23.24	
PN	F	26	22.44	
PN	F	57	23.92	
PN	F	7	25.13	
PN	F	31	23.69	

PN	F	39	23.8	
PN	F	37	24.57	
PN	F	45	24.67	
PN	F	11	24.42	
PN	F	7	27.28	
PN	F	12	26.42	
PN	F	44	23.99	
PN	F	53	22.94	
PN	F	24	24.62	
PN	F	33	24.81	
PN	F	18	21.68	
PN	F	26	22.71	
PN	F	34	23.6	
PN	F	32	22.86	
PN	F	23	24.66	
PN	F	25	21.19	
PN	F	39	26.11	
PN	F	26	25.66	
PN	F	7	22.54	
PN	F	27	19.28	
PN	F	25	22.43	
PN	F	33	20.43	
PN	F	31	23.5	
PN	F	29	24.82	
PN	F	41	24.12	
PN	F	6	21.2	
PN	F	7	23.43	
PN	F	42	20.82	
PN	F	8	24.39	

PN	F	34	24.28	
PN	F	4	24.52	
PN	F	40	20.45	
PN	F	14	26.21	
PN	F	36	24.29	
PN	F	37	23.83	
PN	F	33	21.62	
PN	F	8	24.2	
PN	F	15	24.9	
PN	F	8	27.3	
PN	F	8	24.78	
PN	F	7	24.91	
PN	F	42	25.1	
PN	F	44	25.9	
PN	F	14	25.32	
PN	F	13	23.83	
PN	F	39	23.26	
PN	F	37	24.2	
PN	F	40	27.86	
PN	F	57	22.4	
PN	F	53	26.63	
PN	F	9	26.71	
PN	F	36	22.61	
PN	F	45	25.4	
PN	F	5	24.75	
PN	F	42	21.61	
PN	F	40	22.64	
PN	F	18	22.59	
PN	F	5	25.12	

PN	F	30	25.62	
PN	F	13	27.32	
PN	F	38	23.9	
PN	F	38	22.47	
PN	F	30	21.59	
PN	F	29	27.91	
PN	F	9	27.65	
PN	F	8	23.23	
PN	F	2	26.66	
PN	F	24	25.83	
PN	F	14	21.84	
PN	F	7	22.91	
PN	F	9	24.5	
PN	F	9	24.6	
PN	F	42	24.5	
PN	F	51	23.23	
PN	F	6	25.37	
PN	F	38	24.67	
PN	F	49	24.42	
PN	F	33	23.32	
PN	F	45	22.66	
PN	F	4	23.88	
PN	F	59		
PN	F	10	27.17	
PN	F	10	26.38	
PN	F	40	25.48	
PN	F	49	28.19	
PN	F	39	24.61	
PN	F	36	27.11	

PN	F	5	27.84	
PN	F	31	22.72	
PN	F	52	24.19	
PN	F	32	28.2	
PN	F	7	26.79	
PN	F	28	25.67	
PN	F	7	24.7	
PN	F	7	28.8	
PN	F	10	27.88	
PN	F	16	24.42	
PN	F	42	24.77	
PN	F	30	25.55	
PN	F	15	22.42	
PN	F	10	25.22	
PN	F	40	25.13	
PN	F	11	26.36	
PN	F	32	25.42	
PN	F	62	23.65	
PN	F	5	26.85	
PN	F	36	22.58	
PN	F	23	23.38	
PN	F	45	25.6	
PN	F	61	24.95	
PN	F	59	26.56	
PN	F	62	22.45	
PN	F	28	24.98	
PN	F	6	23.44	
PN	F	13	22.37	
PN	F	13	24.43	

PN	F	14	26.21	
PN	F	18	24.28	
PN	F	4	21.17	
PN	F	19	24.38	
PN	F	43	23.47	
PN	F	7	26.62	
PN	F	38	24.11	
PN	F	6	25.4	
PN	F	30	24.33	
PN	F	20	25.4	
PN	F	47	23.14	
PN	F	30	24.81	
PN	F	12	24.42	
PN	F	5	23.8	
PN	F	37	24.24	
PN	F	5	24.6	
PN	F	53	23.78	
PN	F	41	25.76	
PN	F	14	24.76	
PN	F	27	24.64	
PN	F	5	24.25	
PN	F	6	22.15	
PN	F	45	22.58	
PN	F	46	22.83	
PN	F	34	26.93	
PN	F	39	23.56	
PN	F	65	27.19	
PN	F	35	24.48	
PN	F	5	24.62	

PN	F	22				22.73				
PN	F	36				23.58				
PN	F	32				26.9				
PN	F	57				24.84				
PN	F	54				26.48				
PN	F	9				25.6				
PN	F	41				25.77				
PN	F	8				23.8				
PN	F	36				25.12				
PN	F	40				24.28				
PN	F	13				22.93				
PN	F	44				26.28				
PN	F	6				24.76				
PN	F	3				25.32				
PP	F	7	26.6	28.9	28.95	22.5	D			B.1.525
PP	F	7	26.6	28.9	28.95	22.5	D			B.1.525
PP	F	43	27.21	29.46	30.95	20.58	ND			B.1.1.7
PP	F	43	27.21	29.46	30.95	20.58	ND			B.1.1.7
PP	F	32	22.99	25.8	25.91	25.32	D			
PP	F	47	23.14	24.71	25.47	22.99	D			B.1.1.7
PP	F	41	18.21	20.81	20.86	22.4	ND			
PP	F	37	14.22	16.53	18.11	20.66	D	D	ND	B.1.1.7
PP	F	33	16.83	18.46	20.98	33.48	D	D	ND	
PP	F	76	16.11	18.85	21.56	21.18	D	D		
PP	F	52	32.87	35.76		18.82				B.1.1.7
PP	F	39	18.18	19.47	22.96	18.5	ND	ND	ND	B.1.1.7
PP	F	75	17.66	18.68	35.24	24.24	D	D	ND	
PP	F	17	16.73	19.32	21.21	23.99	D	D	ND	
PP	F	34	17.9	19.9	20.48	21.58	ND	ND	ND	

PP	F	35	18.11	19.61	19.98	21.6	D	D	ND	B.1.1.7
PP	F	26	19.23	20.98	22.28	22.98	ND	ND	ND	
PP	F	44	19.37	21.32	22.42	22.78	D	D	ND	B.1.1.7
PP	F	51	18.89	20.92	21.72	23.27	ND	ND	ND	B.1.1.7
PP	F	38	19.64	21.39	22.47	24.1	D	D	ND	
PP	F	64	18.56	20.51	22.22	21.39	ND	ND	ND	
PP	F	30	18.29	20.63	21.2	23.57	D	D	ND	
PP	F	32	18.57	20.57	22.29	21.8	D	D	ND	B.1.1.7
PP	F	42	19.69	21.79	23.9	21.83	ND	ND		
PP	F	13	20.65	22.2	25.5	20.59	ND	ND	ND	B.1.1.7
PP	F	43	20.23	21.65	24.69	21.15	D	D		
PP	F	29	19.22	21.63	21.89	21.64	ND	ND	ND	
PP	F	51	19.77	22.6	22.78	23.98	D	D	ND	
PP	F	25	20.38	21.7	25.51	21.99	D	D	ND	B.1.1.7
PP	F	45	20.99	22.41	23.68	24.65	D	D	ND	
PP	F	12	19.97	22.23	23.2	24.38	D	D	ND	
PP	F	22	22.3	23.78	25.27	21.88	ND			B.1.1.7
PP	F	25	21.34	23.36	25.49	24.9	D	D		B.1.1.7
PP	F	17	21.33	22.65	25.98	20.34	D	D		B.1.1.7
PP	F	28	20.56	23	23.35	19.8	D	D	ND	B.1.1.7
PP	F	47	20.75	22.98	24.1	20.81	ND	ND	ND	B.1.1.7
PP	F	69	21.49	22.74	24.99	21.52	D	D	ND	B.1.1.7
PP	F	15	20.89	23.11	25.18	22.88	D	D	ND	B.1.1.7
PP	F	53	21.3	22.53	25.98	21.17	D	D	ND	B.1.1.7
PP	F	69	20.93	23.14	24.51	25.69	D	D	ND	B.1.1.7
PP	F	25	26.62	27.96	36.12	24.45	D	D	ND	B.1.1.7
PP	F	63	21.7	22.96	25.48	21.66	D	D	ND	B.1.1.7
PP	F	35	21	23.1	24.78	24.21	D	D	ND	
PP	F	35	21.28	22.83	24.45	23.7	D	D	ND	

PP	F	44	19.85	22.65	22.15	23.75	ND	ND	ND	
PP	F	30	25.5	28.35	28.68	22.25	ND			
PP	F	38	28.49	30.97	33.93	21.9	ND			B.1.1.7
PP	F	11	30.57	33.75	33.47	23.64	D			
PP	F	52	23.23	25.55	26.28	22.98	D			
PP	F	64	23.84	25.23	30.1	19.66	D	D		
PP	F	33	22.61	24.37	25.71	23.53	D	D	ND	
PP	F	28	23.31	24.19	25.17	24.39	ND	ND	ND	B.1.1.7
PP	F	46	22.25	23.9	25.42	25.39	D	D	ND	B.1.1.7
PP	F	38	22.35	24.35	26.2	2.16	D	ND	D	
PP	F	14	21.69	24.35	24.96	22.58	D	ND	D	
PP	F	24	22.65	23.72	25.57	23.2	D	D	ND	
PP	F	62	22.87	25.46	32.8	21.71	ND	ND		B.1.1.7
PP	F	48	23.5	25.41	27.2	24.3	ND	ND	D	B.1.1.7
PP	F	20	23.12	25.24	28.53	19.56	D	D		
PP	F	68	23.59	25.35	28.36	23.54	D	D	ND	B.1.1.7
PP	F	31	23.53	24.97	27.43	22.34	D	D	ND	B.1.1.7
PP	F	17	22.83	24.92	25.79	24.39	ND	ND	ND	
PP	F	50	23.1	24.98	26.23	23.45	D	D	ND	B.1.1.7
PP	F	62	22.6	24.61	25.62	21.3	ND	ND	ND	B.1.1.7
PP	F	38	23.66	25.38	28.75	22.16	D	D	ND	
PP	F	30	26.63	24.74		25.56	D	D	ND	
PP	F	45	23.18	25.1		24.3	D	D	ND	B.1.1.7
PP	F	28	23.15	25.17	27.8	25.38	D	D	ND	B.1.1.7
PP	F	23	22.95	25.6	25.76	24.37	ND	ND	ND	
PP	F	14	22.71	25.1	25.49	23.15	D	D	ND	B.1.1.7
PP	F	43	23.46	24.97	26.45	24.16	D	D	ND	
PP	F	38	23.58	25.1	25.81	23.83	D	D	ND	
PP	F	11	22.87	24.87	25.37	24.75	D	D	ND	B.1.1.7

PP	F	60	21.97	23.92	25.94	19.87	ND			B.1.1.7
PP	F	16	23.5		26.13	24.54				
PP	F	58	24.67	25.95	28.35	19.9	D	D	ND	
PP	F	42	24.9	25.77	27.2	23.21	D	D	ND	B.1.1.7
PP	F	44	23.66	25.63	26.14	23.81	ND	ND	D	
PP	F	12	26.18	26.47	33.37	31.8	D	D	ND	
PP	F	31	24.21	25.78	28.15	23.45	ND	ND	ND	B.1.1.7
PP	F	44	24.43	25.95	26.94	24.83	D	D	ND	B.1.1.7
PP	F	27	24.53	26.2	28.85	25.35	D	D	ND	
PP	F	19	24.3	25.95	26.23	24.79	D	D	ND	
PP	F	77	23.12	25.6	25.85	25.51	D	D	ND	
PP	F	25	24.34	26.51	28.79	24.91	D	D		
PP	F	19	24.84	26.84	30.5	20.81	ND	ND	D	B.1.1.7
PP	F	22	25.51	27.35	28.31	24.21	ND	ND	ND	
PP	F	44	25.93	26.86	35.6	23.41	ND	ND	ND	B.1.1.7
PP	F	11	24.82	26.91	27.78	25.64	ND	ND	ND	
PP	F	39	25.47	26.96	30.46	24.14	D	D	ND	
PP	F	27	23.91	26.78	26.44	24.36	ND	D	D	
PP	F	57	25.5	26.74	29.38	23.52	ND	ND	ND	
PP	F	58	25.98	27.43	28.38	26.45	D	D	ND	
PP	F	42	24.4	26.93	27.96	24.88	D	D	ND	
PP	F	46	25.57	27.4	28.13	25.95	D	D	ND	
PP	F	58	25.92	27.36	31.18	22.76	D	D	ND	
PP	F	53	25.41	27.27	29.85	19.91	D	D	ND	
PP	F	55	26.16	27.67	31.12	20.6	D	D		
PP	F	42	26.81	28.22	29.91	24.13	D	D	ND	
PP	F	51	26.83	28.23	30.55	24.8	D	D	ND	B.1.1.7
PP	F	47	26.73	28.38	31.6	23.7	ND	ND	D	
PP	F	48	26.2	27.67	30.79	23.27	ND	ND	ND	

PP	F	43	25.41	27.58	28.1	25.7	ND	ND	ND	
PP	F	13	26.53	28.6	29.67	23.25	D	D	ND	B.1.1.7
PP	F	25	25.33		28.19	24.9	D	D	ND	B.1.1.7
PP	F	31	25.72	28	30.8	21.3	D	D	ND	B.1.1.7
PP	F	32	25.13	27.73	31.88	19.95	D	ND	ND	
PP	F	65	25.73	28.24	31.15	24.48	D	D	ND	B.1.1.7
PP	F	29	26.19	28.34	28.83	24.29	ND	ND	ND	
PP	F	38	25.5	28.43	27.83	25.12	ND	ND	ND	
PP	F	17	26.64	28.68	29.62	24.87	D	D	ND	
PP	F	36	27.49	29.27	31.46	22.4	ND	ND	ND	
PP	F	23	27.36	29.3	29.81	27.36	ND	ND	ND	
PP	F	47	27.41	28.9	29.92	25.55	ND	ND	ND	B.1.1.7
PP	F	57	28.74	29.22		21.2	ND	ND	ND	
PP	F	31	26.91	29.53	31.24	24.19	ND	ND		
PP	F	48	28.56	29.89	34.58	20.79	D	D		
PP	F	17	27.72	30.2	32.99	21.11				B.1.1.7
PP	F	47	27.99	29.95	31.92	23.92	D	D		
PP	F	37	28.36	30.13	33.58	21.83				
PP	F	16	27.73	30.36	30.33	25.41	ND	ND	ND	
PP	F	24	28.94	30.31	31.54	25.71	ND	ND	ND	B.1.1.7
PP	F	7	26.32	29.53	31.42	21.21	D	D	ND	
PP	F	36	29.48	30.21	32.58	24.8				
PP	F	34	28.79	30.4	33.27	24.98	ND	ND	ND	
PP	F	52	27.82	29.59	34.2	26.69				B.1.1.7
PP	F	40	26.54	29.71	30.53	23.8	D	D		
PP	F	13	27.93	30.1		26.54	D	D	ND	
PP	F	12	32.37	34.68	35.64	23.25	ND			
PP	F	25	29	31.21	32.86	20.99	D	D		
PP	F	35	28.55	30.53	33.69	21.3	D	D		

PP	F	31	28.34	30.84	33.7	21.1	D	D		
PP	F	75	29.91	31.34		24.23	ND	ND	ND	B.1.1.7
PP	F	18	29.63	31.31	34.53	24.31				
PP	F	23	28.74	31.43	33.54	26.13	D	D	ND	
PP	F	12	28.89	31.37	32.78	23.63				
PP	F	23	28.84	30.53	31.8	25.85	ND	ND	ND	B.1.1.7
PP	F	10	30.18	32.2	34.35	25.92	D	D		B.1.1.7
PP	F	34	30.22	32.16	36.34	23.76	D	D		
PP	F	21	30.31		32.35	34.95				
PP	F	45	31.95	32.18	34.7	21.12				
PP	F	48	29.9	31.72	33.66	19.2	D	D		
PP	F	30	30.24	32.4	39.13	19.47				
PP	F	15	29.13	32.16		24.67	D	D	ND	
PP	F	44	29.69	31.71	36.4	23.59				B.1.1.7
PP	F	47	30.77	33.76	34.24	21.56	ND			
PP	F	9	30.95	32.57	34.29	25.29	D	ND	ND	B.1.1.7
PP	F	11	32.49	33.35	35.88	23.7				B.1.1.7
PP	F	63	30.96	32.82	35.59	21.46	D	ND	ND	
PP	F	10	30.95	32.58	33.12	23.41				
PP	F	51	32.96	33.33	36.9	22.49				
PP	F	12	32.6	33.81	36.47	24.72	D	D		
PP	F	41	32.83	34.48	38.17	23.22				
PP	F	55	34.75	34.34		23.33				
PP	F	39	38.89	33.54		30.66	D	D	ND	
PP	F	22	32.17	34.16	37.95	22.79				
PP	F	54	30.85	33.56	38.76	23.37		D		
PP	F	46	30.87	33.82		17.76	ND	ND	ND	
PP	F	43	33.48	34.34		25.28				
PP	F	67	34.42	35.2	39.95	24.81				B.1.1.7

PP	F	58	31.91	34.13	36.5	21.67	D	D		
PP	F	5	33.13	34.6	35.19	24.43				
PP	F	25	33.42	35.3	37	21.82				
PP	F	6	33.15		35.81	26.9				
PP	F	59		36.44		34.21				
PP	F	32		36.19		20.6				
PP	F	63	35.76	36.89		22.93				
PP	F	40	33.8	36.68	38.66	19.97				
PP	F	17		36.88		23.65				B.1.1.7
PP	F	54	33.2	37.15	38.6	24.89				
PP	F	44		38.17	39.42	22.39				
PP	F	36	35.42	37.89	38.97	23.2				B.1.1.7
PP	F	55	27.44	29.44	34.45	21.2	D			B.1.1.7
PP	F	49	35.42	38.87		28.96				B.1.1.7
PP	F	12		38.81		23.35				
PP	F	36		38.81		24.51				
PP	F	20	38.4	38.95		24.53				
PP	F	8	28.84	31.88	31.28	22.99	ND			
PP	-	-								
PP	-	-								

Outliers

	Patient characterisitcs		PCR results				Genetic analysis			Notes
PCR positive (PP) or Healthy (PN)	Gender	Age	Gene E	Gene N	Gene RDRP	Internal Control	N51Y	H69/V7	E484K	
PN	M	3				24.2				
PN	M	3				23.62				
PN	M	4				22.89				
PN	M	4				29.42				
PN	M	4				30.51				
PN	M	5				23.1				

PN	M	5				23.84		
PN	M	5				22.21		
PN	M	5				25.73		
PN	M	5				26.8		
PN	M	6				25.55		
PN	M	7				24.03		
PN	M	7				22.74		
PN	M	7				28.17		
PN	M	8				21.65		
PN	M	8				24.79		
PN	M	9				24.16		
PN	M	10				22.49		
PN	M	10				24.98		
PN	M	11				25.42		
PN	M	11				24.88		
PN	M	11				23		
PN	M	14				23.49		
PN	M	16				25.57		
PN	M	17				22.8		
PN	M	19				24.33		
PN	M	24				20.14		
PN	M	37				22.69		
PN	M	41				21.16		
PN	M	46				25.89		
PN	M	52				25.75		
PN	M	70				27.05		
PN	M	74				24.36		
PP	M	8	32.07	34.73	34.68	25.89		B.1.1.7
PP	M	10	25.15	27.28	27.71	27.1		

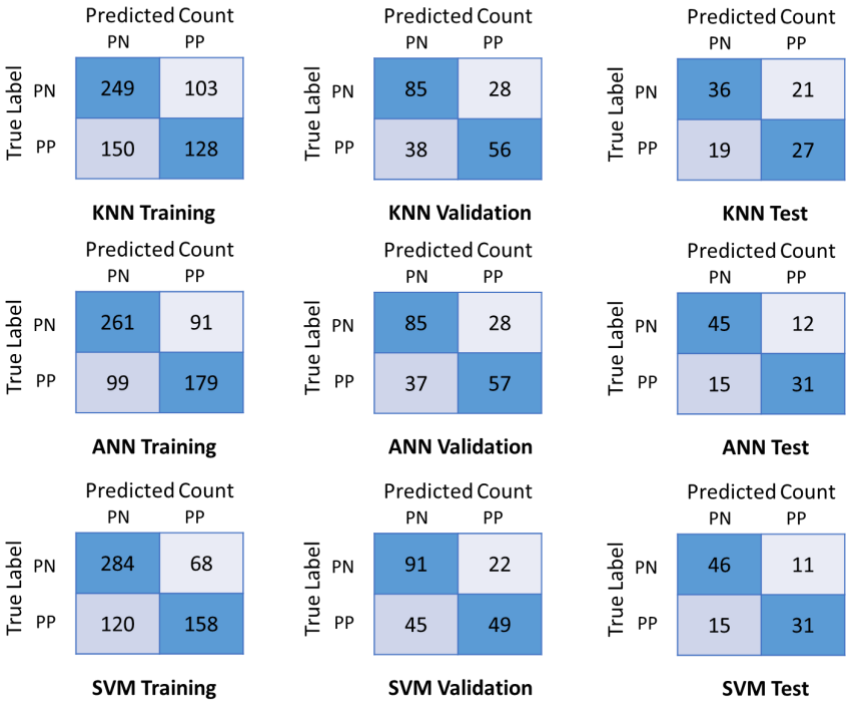
PP	M	12	28.8	31.11	31.56	24.86	ND			
PP	M	14	25.58	27.26	28.83	27.04	D	D	ND	B.1.1.7
PP	M	15	23.03	24.98	25.93	25.09	D	D	ND	B.1.1.7
PP	M	17	28.32	29.34	32.3	21.43	D	D	ND	
PP	M	21	26.27	27.3	28.97	23.63	D			
PP	M	26	32.2	34.33	37.02	23.17	ND	ND	ND	
PP	M	26	24.09		26.66	24.29	D	D	ND	
PP	M	35	12.43	15.18	18.54	21.78	ND	ND	ND	B.1.1.7
PP	M	38	28.03	29.01	32.1	22.97	D	D	ND	
PP	M	40	15.17	17.65	18.11	26.2	ND	ND	ND	
PP	M	41	17.39	19.01	20.19	21.89	D	D	ND	
PP	M	47	22.35	24.28	25.79	26.4	D	D	ND	
PP	M	48	24.69	26.64	29.51	23.27	ND	ND	ND	
PP	M	54	13.57	16.35	18.95	21.15	ND			
PP	M	56	17.3	18.91	20.6	22.85	D	D	ND	
PP	M	72	16.82	16.97	29.5	26.2	D	D		B.1.1.7
PP	M	74				24.36				
PN	F	5				26.34				
PN	F	5				24.5				
PN	F	5				25.41				
PN	F	6				24.38				
PN	F	6				28.45				
PN	F	6				25.62				
PN	F	6				23.61				
PN	F	7				24.25				
PN	F	8				24.91				
PN	F	8				26.46				
PN	F	8				23.29				
PN	F	10				25.99				

PN	F	16				21.78				
PN	F	17				26.23				
PN	F	21				22.92				
PN	F	25				22.87				
PN	F	26				24.7				
PN	F	27				22.47				
PN	F	31				22.75				
PN	F	34				25.04				
PN	F	35				22.1				
PN	F	36				22.11				
PN	F	40				21.53				
PN	F	46				24.21				
PN	F	47				22.4				
PN	F	49				26.41				
PN	F	51				24.68				
PP	F	4	28.86	30.44	31.15	26.92	D	D	ND	B.1.1.7
PP	F	5	35.38	37.28		27.08				
PP	F	11	23.7	25.19	26.8	23.91	D	D	ND	
PP	F	13	23.82	25.6	27.51	23.55	D	D		B.1.1.7
PP	F	14	30.05	32.13	33.69	23.82	D	D	ND	B.1.1.7
PP	F	15	29.86	32.25	32.86	22.12	D	ND	ND	B.1.1.7
PP	F	15	16.06	17.16		20.61	D	D	ND	B.1.1.7
PP	F	15	27.23	29.01	29.95	24.83	ND	ND	ND	
PP	F	19	31.3	31.71	34.48	23.62	ND	ND	ND	
PP	F	20	28.47	29.96	32.24	22.83	D	D		B.1.1.7
PP	F	25	17.36	20.19	20.54	23.42	ND	ND	ND	
PP	F	25	29.58	31.35	32.23	25.04				
PP	F	26	24.56	26.2	27.18	25.06	D	D	ND	
PP	F	26	24.48	26.56	28.32	23.45	D	D	ND	

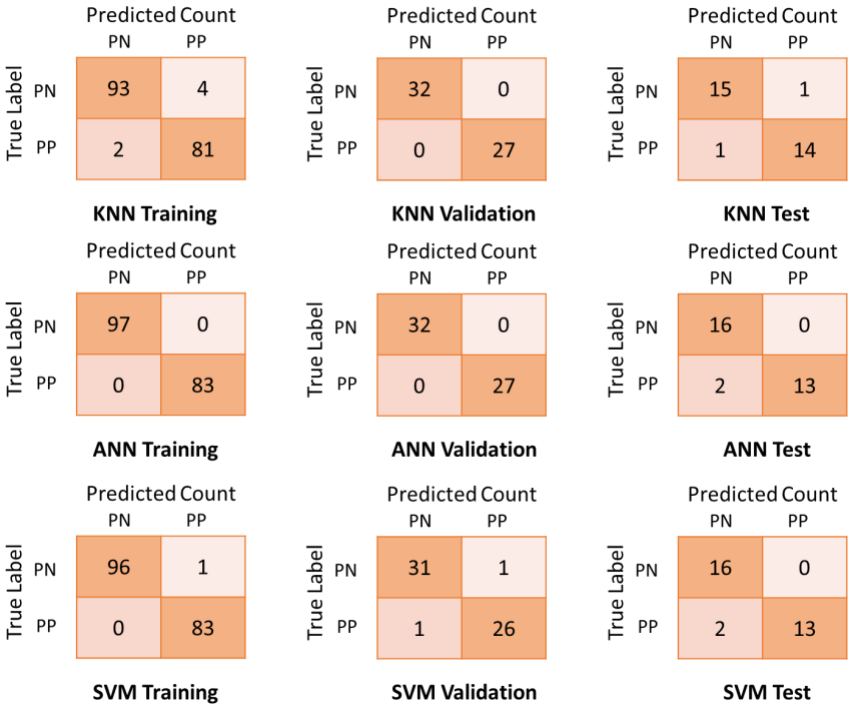
PP	F	27	22.66	24.73	25.51	21.9	D	D	ND	
PP	F	28	16.97	18.07	20.76	21.2	D	D	ND	
PP	F	33	29.31	31.08		27.05	D	D	ND	
PP	F	34	28.81	30.63	31.66	27.27	ND	ND	ND	
PP	F	35	27.84	29.8	31.8	20.29	ND			
PP	F	37	27.59	24.93	27.13	21.17	ND	ND	ND	
PP	F	37	27.64	29.55	32.31	22.6		D		
PP	F	39	19.93	21.63	23.12	21.65	D	D	ND	
PP	F	39	18.09	20.62	22.27	19.18	ND	ND	ND	B.1.1.7
PP	F	41	19.52	22.9	22.43	22.62	ND			
PP	F	41	23.44	25.51	25.96	23	D	D	ND	
PP	F	43	18.62	20.81	20.45	21.85	ND	ND	ND	
PP	F	44	23.62	26.24	25.61	25.2	D	D	ND	
PP	F	45	21.31	23.95	24.39	23.13	D	D	ND	B.1.1.7
PP	F	51	26.57	26.35	34.5	22.46	D	D	ND	
PP	F	53	28.4	30.55		23.98	D	D	ND	
PP	F	57	22.5	23.49		35.12	D	D	ND	B.1.1.7
PP	F	61	24.7	26.88	28.13	22.6	D	D	ND	
PP	F	63	21.62	22.82	24.09	22.13	D	D	ND	

A. 5. Confusion matrices of Database 1, Database 2, Database 3, and Database X.

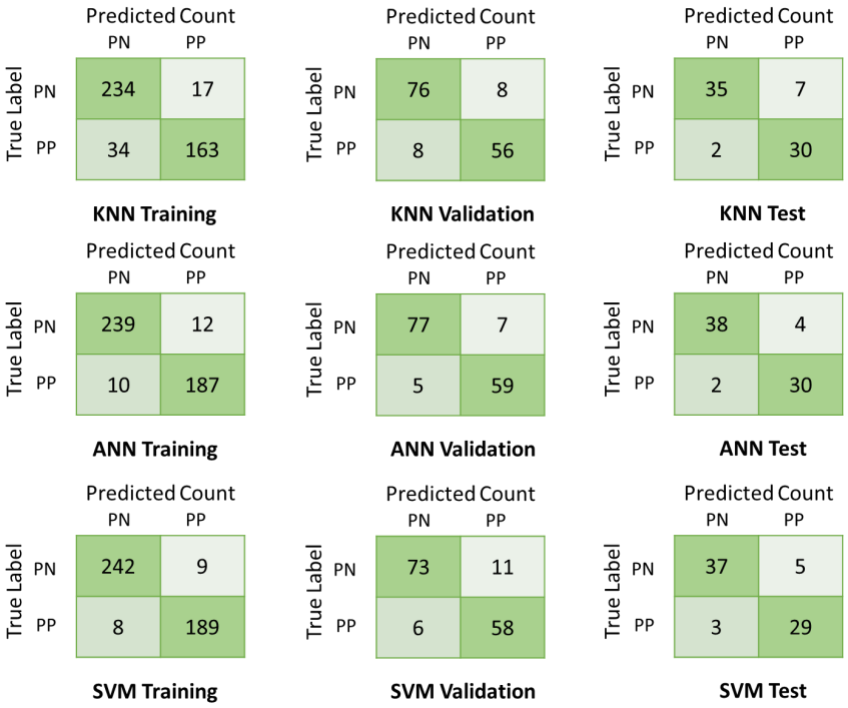
Database 1 Confusion Matrices



Database 2 Confusion Matrices



Database 3 Confusion Matrices



Database X Confusion Matrices