Development of an Expert System for the Identification of Bacteria by Focal Plane Array Fourier Transform Infrared Spectroscopy

By

Andrew Ghetler

Department of Food Science and Agricultural Chemistry

McGill University, Montreal

August, 2009

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

Short title:

EXPERT SYSTEM FOR BACTERIA IDENTIFICATION

ABSTRACT

This study presents new techniques for the analysis of data acquired by focal plane array Fourier transform infrared (FPA-FTIR) spectroscopy. FPA-FTIR spectrometers are capable of acquiring several orders of magnitude more data than conventional FTIR spectrometers, necessitating the use of novel data analysis techniques to exploit the information-rich nature of these infrared imaging systems. The techniques investigated in this study are demonstrated in the context of bacteria identification by FPA-FTIR spectroscopy. Initially, an examination is made of the image fidelity of three FPA-FTIR instruments and demonstrates the high degree of within-image and between-image variability that is encountered with this technology. This is followed by a description of the development of pixel filtration routines that allow for the extraction of the most representative data from the infrared images of non-uniform samples. A genetic algorithm (GA) approach is introduced for determining the relevancy of spectral features in relation to bacterial classification and is compared to other forms of classifier optimizations. A proof-of-concept study demonstrating the potential use of infrared imaging to detect bacterial samples originating from a mixed culture is then presented. Finally, an overall methodology involving the combination of these data analysis techniques and including additional approaches towards the development, maintenance, and validation of databases based on infrared imaging data is described. This methodology has been developed with an emphasis on accessibility by implementing the elements of an expert system which allows for this technology to be employed by a non-technical user.

RÉSUMÉ

Cette étude présente une nouvelle approche d'analyse de données spectrales résultant de l'utilisation de la spectroscopie infrarouge à transformée de Fourier couplée à un détecteur type «matrice à plan focal» (FPA-FTIR) à balayage rapide. Les spectromètres FPA-FTIR ont une capacité de capture de données de plusieurs ordres de grandeur supérieurs aux spectromètres traditionnels et nécessitent donc des techniques avancées d'analyse de données pour exploiter cette mine d'information que représente l'imagerie infrarouge. La spectroscopie FPA-FTIR a été utilisée dans cette étude pour l'identification des bactéries. L'étape initiale, celle de la comparaison de trois spectromètres FPA-FTIR sur les points de vue fidélité de l'image, tant image-image qu'entre images, a révélé de grandes variabilités qui sont propres à cette technologie. Cette étape est suivie du développement de routines de filtration de pixels permettant d'extraire les données caractéristiques de l'imagerie infrarouge des échantillons non-uniformes. Un algorithme génétique (GA) est ensuite introduit pour déterminer la pertinence des caractéristiques spectrales sur le plan de la classification bactérienne et a été comparé à d'autres formes de classification optimisée. Une étude de démonstration de la capacité de la technologie d'imagerie infrarouge pour la détection des échantillons de bactéries provenant de cultures mixtes s'en est suivie. Pour terminer, une méthodologie globale combinant ces techniques d'analyse de données et incluant d'autres étapes telles le développement, la mise à niveau et la validation des bases de données d'imagerie infrarouge a été présentée. Cette méthodologie met l'emphase sur le développement et l'implantation d'un système expert accessible d'utilisation à de non-experts.

ACKNOWLEDGMENTS

I wish to give my thanks and gratitude to a number of people without whom this work would not have been possible. It has truly been a group effort.

First and foremost, I would like to thank my supervisor Dr. Ashraf A. Ismail. He has been the guiding force in helping me to develop my ideas as they have evolved over the past few years. His unwavering support has ensured my success in what is to be one of my life's greatest achievements. I would like to thank Dr. Jacqueline Sedman, without her incredible logic, knowledge, and masterful editing this could not have been accomplished. I would like to thank Dr. Frederik R. Van de Voort for his encouragement and advice, for introducing me to food science, and for including me in various parts of his research. I would also like to thank my (former) colleague and good friend Dr. Jonah Kirkwood who has taught me so much and provided countless hours of help in achieving this goal. Finally, I would like to say thank you to all my graduate colleagues, each of whom have played a part in making this work a reality.

I would like to express my deepest gratitude to Brad, Dave, and Josh Pinchuk at Thermal-Lube Inc. for their moral and financial support throughout the years, and for the incredible experiences I've gained through working together. I'd also like to thank, Dr. Emanuel Akochi-Koble for the support and mentoring he has given.

I would like to thank Varian Inc. for supplying the instrumentation and for their continued support in this research, and the National Sciences and Engineering Research Council of Canada for awarding me with four years of an IPS scholarship which allowed me to focus on my graduate studies.

I wish to give a huge thank you to my parents for always believing in me and for providing me with the encouragement and support that I needed. I'd like to thank my brother for his motivation, and my sisters for their support. I would like to express my appreciation to all my friends who supported me through this endeavor. Last but far from least, a special thank you goes to my wife Meghan who has been supportive, encouraging, and understanding beyond belief even when things were taking a little longer than they should.

CONTRIBUTION OF AUTHORS

Chapter 3-6 are the text of papers being prepared for publication. The author of this thesis is responsible for developing the ideas and concepts presented, and for the analysis of all data in this study. Dr. Ashraf A. Ismail, the thesis supervisor, provided support and guidance throughout the course of this research. Dr. Jacqueline Sedman provided valuable advice and insight into all aspects of the research including the chemometrics and spectroscopy. Mr. Alexander Enfield provided assistance in creating the samples and acquiring the spectral data used in Chapter 3. Dr. Jonah Prevost Kirkwood provided the infrared image data used for Chapters 4-5 and his expertise in bacteria identification by FPA-FTIR spectroscopy. Mrs. Laura Carranza (a doctoral student under the supervision of Dr. Ismail) provided the data for Chapter 6.

Chapter 3
Ghetler, A., A. Enfield, J. Sedman, and A.A. Ismail, *Evaluation of Image Fidelity in Focal Plane Array Fourier Transform Infrared Spectroscopy.*

Chater 4
Ghetler, A., J.P. Kirkwood, J. Sedman, and A.A. Ismail. *Pixel Filtration for the Extraction of Representative Data from Infrared Images.*

Chapter 5
Ghetler, A., J.P. Kirkwood, J. Sedman, and A.A. Ismail. *Continuous Genetic Algorithms for Evaluating Feature Relevancy in Infrared Spectra.*

Chapter 6
Ghetler, A., L. Carranza, J. Sedman, and A.A. Ismail. *Identifying Hyperspectral Images Originating From a Heterogeneous Sample.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial neural networks
BDV	Between-deposit variability
BIV	Between-image variability
CCD	Charge-coupled device
CGA	Cumulative genetic algorithms
CPU	Central processing unit
CSD	Calibration standard deviation
CVA	Canonical variate analysis
DFA	Discrminant function analysis
DNA	Deoxyribonucleic acid
DTGS	Deuterated triglycine sulfate
FO	Full overlap
FOV	Field of view
FPA-FTIR	Focal plane array Fourier transform infrared
FTIR	Fourier transform infrared
FWHM	Full-width at half-maximum
GA	Genetic algorithm
GT	Grown together
HCA	Hierarchical cluster analysis
InSb	Indium antinomide
IR	Infrared
k-NN	K-nearest neighbor
LOOCV	Leave-one-out cross-validation
MCT	Mercury cadmium telluride
MD	Mix and deposit
PCA	Principal component analysis
PLS	Partial least squares
PO	Partial overlap
PRESS	Predictive error sum of squares
RMS	Root mean square
SD	Standard deviation
SIMCA	Soft independent modeling of class analogy
SNR	Signal-to-noise ratio
SVM	Support vector machines
WIV	Within-image variability
ZnSe	Zinc selenide

Chapter 1: Introduction

1.1 GENERAL INTRODUCTION

Advances in scientific instrumentation in the last 15 years have seen the development of a variety of information-rich technologies which have pushed the boundaries of conventional methods of data acquisition and analysis. These types of technologies acquire massive amounts of data in a short period of time, necessitating the use of computational methods for effective interpretation of the data. One such technology is focal plane array Fourier transform infrared (FPA-FTIR) spectroscopy, a technology for infrared imaging first introduced in 1995 [1]. The infrared images recorded by FPA-FTIR spectrometers consist of hundreds to thousands of simultaneously collected, spatially resolved infrared spectra, allowing the chemical composition of samples to be probed with a spatial resolution on the micron scale. Given that FPA-FTIR spectrometers acquire several orders of magnitude more data than conventional FTIR spectrometers are capable of acquiring in the same amount of time, the development of methods and tools for the analysis and interpretation of this data is key to making full use of the capabilities of this technology. This thesis looks to provide new insight into ways in which the information-rich data acquired by FPA-FTIR spectroscopy may be exploited.

Most of the applications that FPA-FTIR spectroscopy has found to date utilize the high spatial resolution and information content provided by this technology to visualize the spatial relationships within and between chemical components of heterogeneous samples. The present study examines a less explored facet of FPA-FTIR spectroscopy, which involves exploiting the high data-redundancy contained within the infrared images of ostensibly homogeneous samples together with the capability to evaluate sample uniformity and quality. By virtue of the small sampling area and the large number of spatially resolved spectra acquired from the sample by an FPA-FTIR spectrometer, certain advantages, including enhanced reliability with respect to

conventional FTIR spectroscopy, are attained when FPA-FTIR spectroscopy is used in this manner.

In this context, the target application of the present study is the use of FPA-FTIR spectroscopy as a tool for the identification of foodborne bacteria. In today's society food safety is of paramount importance. Infections and illnesses due to food contamination cost the Canadian health care system millions of dollars and occupy valuable resources which could otherwise be applied toward other health concerns. Current methodologies for bacteria identification do not meet the needs and requirements of agencies to rapidly monitor the possibility of microbial contamination of the food supply. For this reason, the McGill IR Group has undertaken research to develop a technology for the robust and reliable identification of bacteria based on the use of FPA-FTIR spectroscopy. The approach to bacteria identification being developed has the potential to assist government and regulatory agencies in the detection of microbial contamination of foods for the prevention of food-related illnesses by offering a high-throughput method for bacteria identification. The implications extend beyond food safety and into the domain of food surveillance; the monitoring of the degree of antimicrobial resistance found in foodborne pathogens. The availability of a high-throughput method could potentially reduce analysis times from days to hours, allowing regulatory agencies to take a more active stance on food safety through routine and regular analysis of the food supply.

The research presented in this thesis is an integral part of this initiative. The aim of this research was to provide a complete methodology for bacteria identification from spectral acquisition through to data processing, the development of databases, and finally the identification of unknowns. It was the intention to address all aspects of the bacteria identification process from the point of spectral acquisition to output of the results and to implement the developed methods in an accessible manner. Following a review of the relevant literature, this thesis presents the research from this study in five chapters corresponding to the various stages of the bacteria identification process. First, the image fidelity provided by FPA-FTIR spectroscopy is examined through an

investigation of the quality of the spectral data acquired on multiple FPA-FTIR instruments. This is followed by a description of the pixel filtration methods used to extract the most relevant data from the infrared images of non-uniform samples. Following this, an approach to defining the relevancy of spectral features for bacteria classification is presented and compared to other feature selection methods employed in this study. A first of its kind study on the potential for infrared imaging to detect bacterial samples originating from mixed cultures is then described. Finally, an overall methodology integrating all these components and its implementation in an expert system are presented.

1.2 RATIONALE AND OBJECTIVES OF THE RESEARCH

1.2.1 Overview of the research project

The overall objective of the research presented in this thesis is to develop new routines for data analysis and database development for spectral data acquired by FPA-FTIR spectroscopy and to incorporate these routines into an expert system for the identification of bacteria. This expert system is intended to address all the elements required for the implementation of FPA-FTIR spectroscopy as a tool for the routine, high-throughput analysis of foodborne bacteria.

1.2.2 Specific objectives of the research

The specific objectives that will be addressed are:

(i) An examination of the performance characteristics of present-day FPA-FTIR spectrometers through within-instrument and instrument-to-instrument spectral comparisons.

(ii) The development of data-mining and filtration routines to extract relevant information from infrared images of samples exhibiting a high degree of sample non-uniformity.

(iii) Determination of the most appropriate classification and identification methods for building reliable and robust classifiers for the identification of bacteria.

(iv) An examination into classifier optimization (features selection) techniques for use as part of a methodology of bacteria identification and as a means of extracting additional chemical information regarding the samples under analysis.

(v) An evaluation of the potential for FPA-FTIR spectroscopy to detect bacterial samples originating from mixed cultures.

(vi) The development of a complete methodology for the reliable identification of bacteria by FPA-FTIR spectroscopy.

(vii) The implementation of this methodology in the form of an expert system which provides informative feedback regarding the identification process and provides the necessary accessibility of the technology to the intended user.

Chapter 2: Literature Review

2.1 INTRODUCTION

Infrared spectroscopy is a tried-and-true technology for the chemical characterization of samples and has been applied in a large variety of scientific domains for purposes of qualitative and, to a lesser extent, quantitative analysis. Its applications range from the identification of pure chemical compounds to the analysis of highly complex biological samples. It is widely used in both academia and industry, and it is routinely featured on popular TV shows such as CSI owing to its important role in forensics. From a commercial perspective, infrared spectroscopy has given rise to a large industry with combined sales of infrared spectrometers worldwide amounting to hundreds of millions of dollars annually. There is also a large accessories and disposables market, further increasing the use and potential applications of infrared spectroscopy.

In the long history of infrared spectroscopy, which is now approaching its 100th anniversary, one of the most significant advances was the development of Fourier transform infrared (FTIR) spectroscopy, beginning in the late 1960s. Over the last few decades, FTIR instrumentation has evolved from large and cumbersome units with complex interfaces to easy-to-use bench-top instruments [2] with a footprint no larger than a laptop computer. The last decade has seen substantial miniaturization of FTIR components to the extent that hand-held devices are now available [3]. However, the performance of these devices is not typically on par with that of their larger counterparts. For a review of technologies relating to infrared spectroscopy see [4].

One of the currently emerging technologies in infrared spectroscopy is infrared imaging, using either linear array or two-dimensional array detectors. The latter, termed focal plane array (FPA) detectors, originated as a military de-classified technology and, in fact, the first FPA detectors sold were those that were deemed unfit for military applications [5]. Because of international political events over the last few years, certain restrictions are in place regarding the distribution and development of this technology.

Both linear array and FPA detectors allow the user to derive not only chemical information about a sample but also spatial information. In addition, such infrared imaging detectors are typically coupled with infrared microscopes allowing for spectral data to be acquired on the micron scale. The spatial resolution at this scale is diffraction limited and is approximately equal to the wavelength [6], with the nominal spectral resolution quoted by manufacturers typically being an average spatial resolution across the measurable wavelength range.

An overview of infrared microscopy and imaging spectroscopy is provided below; however, for a very complete and technical description of these technologies, refer to the following sources [1, 4, 7, 8].

2.2 INFRARED IMAGING OVERVIEW

2.2.1 Infrared microscopy

Prior to the introduction of infrared microscopes, the approach used to examine small samples (1 mm in size) was to aperture down the beam of a conventional spectrometer in order to only expose a small area of the sample to infrared light. Such an approach has the drawback of significantly reducing the amount of source energy that reaches the detector. For example, in a spectrometer with a beam width of 6 mm which is apertured down to 1 mm, roughly 97% of the source energy will be lost, resulting in a spectrum with a very poor signal-to-noise ratio (SNR). The SNR was improved upon by condensing the beam using a beam condenser, thus focusing the full intensity of the source radiation onto a smaller area of the sample. Around 1990, infrared microscopes, which differ from optical microscopes through their use of all-reflective optics, became available to facilitate the acquisition of spectra from samples on the micron scale. In addition, since detector noise is proportional to the square root of the area of the detector [9], the detectors employed in infrared microspectroscopy were reduced in size to improve upon the SNR of the acquired spectra. Microscopes were fitted with motorized stages for precision positioning of the sample, and apertures were used to further limit the transmission of infrared energy to the desired location on

the sample. In this form, spectra with a sufficient SNR could be acquired from an area as small as 10 µm. Sequential acquisition of spectra from neighboring portions of the sample by incremental positioning of the motorized stage allowed for the piecewise construction of an infrared image of the sample. The process of acquiring an infrared image in this manner, termed mapping, would take on the order of hours.

2.2.2 Infrared Imaging

The first example of the coupling of an infrared imaging detector with an FTIR spectrometer appeared in 1995 [1]. The detectors initially employed were FPA detectors fabricated using indium antimonide (InSb), which were insensitive below ~1800 cm^{-1}. Infrared imaging in the information-rich mid-infrared region became possible when InSb detectors were replaced by mercury cadmium telluride (MCT) FPA detectors, although InSb detectors continue to be widely employed in near-infrared (NIR) imaging, for which they are well suited. As opposed to a photoconductive mode typical of single-element MCT detectors, FPA MCT detectors operate in a photovoltaic mode [5], resulting in a wavenumber cut-off of 900 cm^{-1}. Like single-element MCT detectors, FPA MCT detectors operate at ~70 degrees Kelvin and must be cooled using liquid nitrogen. At the time when FPA-FTIR spectrometers first became commercially available, the electronics did not exist to transfer the acquired data from the detector to the computer in real-time. The result was that these instruments needed to be operated in a step-scan [10] mode and image acquisition times were significantly longer than what they are today. By 1999 advancements were made that allowed for the acquisition of images in rapid-scan [10] mode, though only for arrays of a smaller size. To this day improvements are being made that allow for larger and larger infrared images to be acquired at increasing rates of acquisition.

In 2001 Perkin Elmer (Wellesley, MA) introduced an infrared imaging technique employing a linear array (LA) of 16 MCT detectors. This technique combines imaging across one dimension with mapping across the second dimension to incrementally construct an image. The photoconductive linear MCT array offers a slightly lower wavenumber cut-off of 700 cm^{-1} as compared to FPA detectors (900 cm^{-1}). LA detectors

have been reported to provide a substantially higher SNR than FPA detectors [5], but for imaging a given area of a sample within a given amount of time, this advantage is offset to some extent because a complete image is acquired in a single scan with an FPA detector whereas it has to be mapped out with an LA detector. For example, it requires 20 times longer [11] to acquire an image from a sample area of 2 mm x 2 mm with a LA detector as compared to an FPA detector. Thus, it is clear that if the FPA detector were to acquire for a similar length of time (by increasing the number of co-added scans) as is required by the LA detector to acquire an image of the same area, the SNR in the spectral images from the FPA detector would be substantially improved. Manufacturers often use metrics which most favorably illustrate the capabilities of the technology they provide and so the optimal choice of instrument is primarily dependent on the intended application.

Current manufacturers of FPA-FTIR spectrometers are Varian (Melbourne, Australia) and Bruker Optics (Billerica, MA). Infrared imaging spectrometers equipped with LA detectors are manufactured by Perkin Elmer (Waltham, MA) and Thermo-Scientific (Madison, WI), and at the time of writing it is expected that Shimadzu (Kyoto, Japan) will be unveiling a spectrometer of this type in the near future. The data gathered as part of this research has only been acquired on FPA-FTIR spectrometers by Varian.

2.2.3 FPA-FTIR spectrometer

An FPA-FTIR spectrometer consists of a conventional FTIR spectrometer attached to an infrared (IR) microscope which houses the FPA detector. Current microscopes can be configured with both a single-element detector and an FPA detector, though acquisition can only occur on one of the detectors at any given time. The FPA-FTIR instrumentation employed in the present study, manufactured by Varian, is illustrated in Figure 2.1.

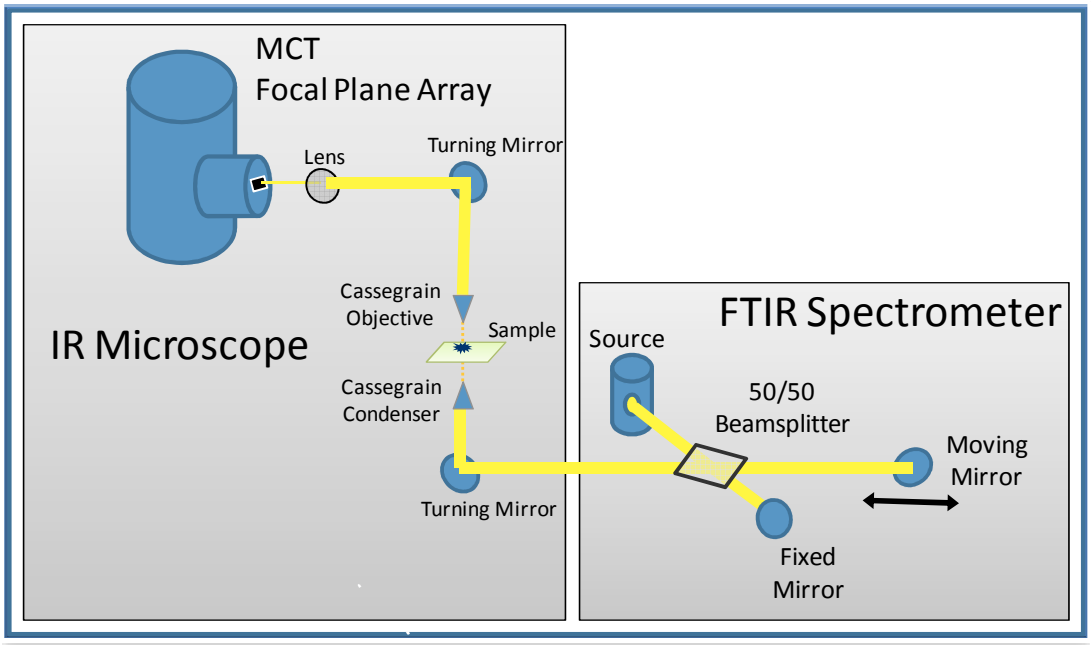


Figure 2.1: Pictorial diagram illustrating each of the components in the FTIR spectrometer and IR microscope.

Analogous to a digital camera, an FPA detector consists of a grid of pixels, usually square (Figure 2.2), where each pixel is an independent detector element. Currently, FPA MCT detectors are available in configurations of 16x16 (256 pixels), 32x32 (1024 pixels), 64x64 (4096 pixels) and 128x128 (16,384 pixels). Because MCT detectors operate at ~70 Kelvin, the detector housing is cooled with liquid nitrogen, which typically must be replenished by the user every 4 – 8 hours. For all array sizes, operation of the FTIR spectrometer in rapid-scan mode is possible, whereas a decade ago rapid-scan acquisition was not possible even with the smallest arrays.

With the instrument illustrated in Figure 2.1 and equipped with a 32x32 array, which was the configuration used throughout this study, the infrared radiation from the infrared source is focused onto an area of 188 µm x 188 µm, resulting in a nominal spatial resolution of 5.6 µm per pixel; however, as mentioned previously, the true spatial resolution is dependent on the diffraction limit of the infrared wavelength under examination. Current technology employed in the Varian FPA-FTIR spectrometer allows for an acquisition rate of 1.6 seconds per scan with a 32x32 array.

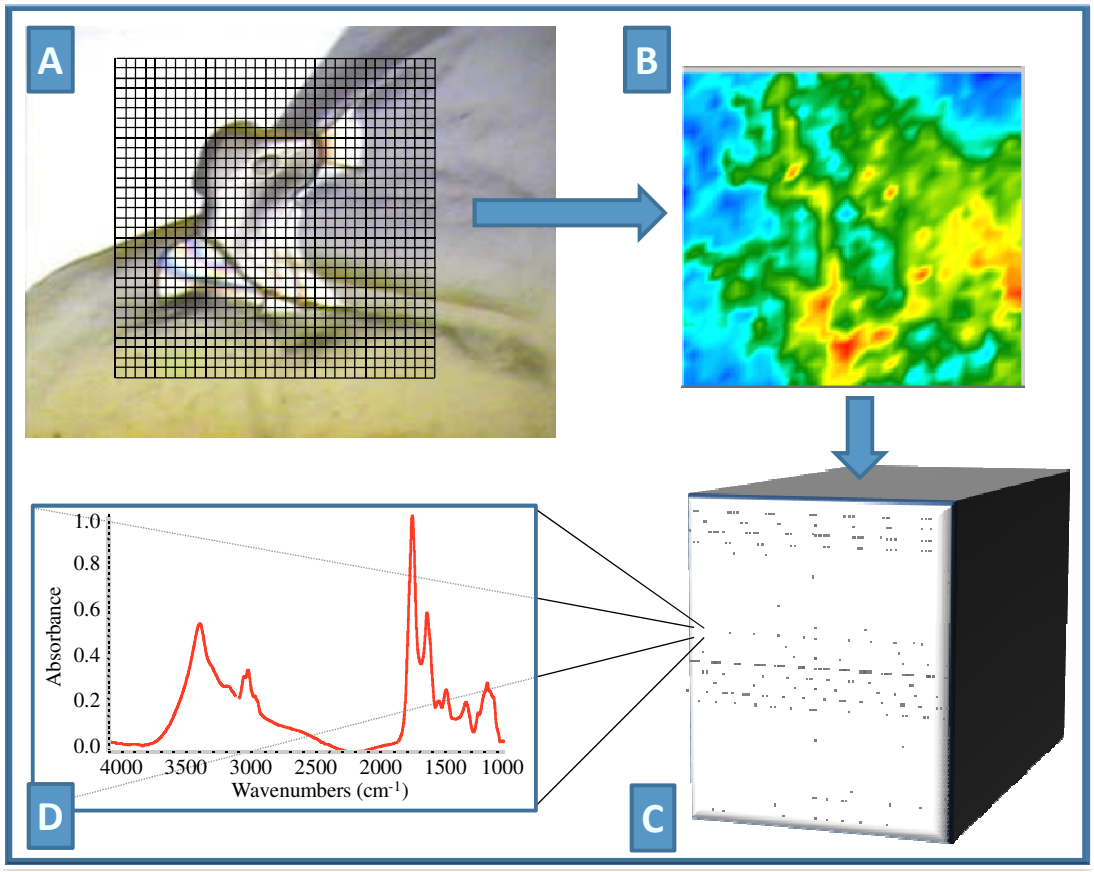


Figure 2.2: A depiction of the data acquired by a FPA-FTIR spectrometer. A) The bacterial smear on a zinc selenide slide as seen through the infrared microscope (in reflection mode). The visible area is approximately 250 µm x 250 µm, with the grid drawn over it representing the 188 µm x 188 µm area that is imaged by the FPA camera. B) The infrared image based on the total intensity of each spectrum as seen in the acquisition software (Resolutions Pro 4.0, Melbourne, Australia), where red and blue pixels represent spectra with bands of the greatest and lowest overall intensity, respectively. C) Representation of a 32x32 hyperspectral data cuboid illustrating that each pixel shares a third dimension (spectral wavelength) which in this example consists of 792 points from 4000 cm^{-1} to 950 cm^{-1}. D) An example of a spectrum from a single pixel.

Figure 2.2 provides an example of the resultant data of a single infrared image acquired from a 32x32 FPA detector. This information is often referred to as hyperspectral data, or a hypercube, although this is not entirely accurate as all three dimensions are not equivalent in size. The term hypercuboid would be more appropriate. FPA-FTIR spectroscopy has the capacity to generate a tremendous amount of data in a very short period of time. Figure 2.3 (presented at the end of the chapter)

provides an illustration of the data contained within a single infrared image. As is easily seen, the quantity of data associated with a single image (which can be acquired in seconds) is far more than can be analyzed visually by an individual. It is for this reason that data processing algorithms are a prerequisite to successfully exploit the information-rich data acquired using FPA-FTIR spectroscopy.

2.2.4 Image fidelity in FPA-FTIR spectroscopy

Several authors have discussed various issues that impact the fidelity of the infrared images acquired by FPA-FTIR spectroscopy [5, 12]. There are a number of non-sample sources of spectral variability that can degrade the quality of the images acquired by FPA detectors or introduce artifacts. In most types of infrared spectrometers, a major source of spectral variability is detector noise. Because detector noise is proportional to the square root of the detector size (area), the theoretical maximum SNR of FPA detectors has been estimated at 315,000 [12], many times higher than that of single-element detectors. However, this theoretical value was calculated on the assumption that the miniaturized MCT detectors constituting the FPA detector have the same detection characteristics as a single-element MCT detector, whereas, in practice, their performance is much poorer than that of a monolithic MCT detector, largely as a result of the processes involved in their fabrication [12] due to the incredibly small size of each detector. In addition, the circuitry that connects each detector element can result in additional sources of noise [8]. Earlier fabrication techniques produced a fixed pattern noise where certain pixels at regular intervals exhibited anomalous noise characteristics. This is supposedly no longer the case (personal correspondence) in the fabrication of newer detectors as a result of improved fabrication processes.

Occasional occurrences of “high-magnitude noise” may also contribute to spectral variability in FPA-FTIR spectroscopy [13]. Noise events whose magnitude exceeds the expected range are rare in the case of conventional single-element detectors, but as a result of the large number of detector elements present in FPA detectors, the likelihood of such noise events occurring during spectral acquisition is

drastically increased. High-magnitude noise is of particular importance when images are acquired in the step-scan mode owing to the longer overall acquisition times, leading to an increased likelihood of a high-noise event.

Another potential contributor to spectral variability in FPA-FTIR spectroscopy is the non-uniform intensity of radiation reaching the detector. When the total infrared response of the detector is examined, an aligned instrument (with no sample in the beam) will typically display a “hot spot” (a 2D Gaussian) near the center of the array. From our own experience, pixels at the periphery of the detector array may receive from 5% to 50% less energy than those at the center of the “hot-spot”, depending on the alignment of the instrument’s mirrors and the size of the array. Prior to acquiring spectral images, an open-beam calibration routine is conducted on every pixel so as to normalize the displayed intensity between each of the detectors of the array. This routine uses a two-point calibration based on a high and a low intensity value to estimate the response curve of each detector element. However, inherent to this procedure is the assumption of detector linearity between these two extremes, and thus image fidelity subsequent to calibration will be compromised by a non-linear response of any of the individual detector elements in the array.

Image fidelity can also be compromised when samples diffuse, scatter, or diffract the infrared light. The degree to which this can affect the acquired infrared images is dependent on the type and uniformity of the sample under analysis, making it very difficult to estimate the influence of this effect on real-world samples.

An examination into the quantitative capabilities of FPA-FTIR imaging was made by Snively and Koenig [12] in 1999 to evaluate the extent to which some of the above contributors to variability affect the quantitative capabilities of FPA-FTIR spectroscopy. Firstly, these authors introduced a diffuser into the beam path in order to have a more uniform distribution of light intensity across the detector. However, the overall effect was reported to be minimal. In addition, the effect of increasing the number of co-added scans was examined, and it was noted that the expected square-root increase in SNR was not obtained, although the trend was similar. It was also found that pixel-to-

pixel absorbances in the image recorded for a benzonitrile solution had a coefficient of variation of ~16% or higher but that increasing the number of co-added scans reduced that value to ~12%. The latter value still represents a fairly high standard deviation of 0.066 absorbance units, but the authors of the study did not comment on this result. Finally, the authors estimated a 9 mol % limit of detection and a >30 mol % limit of quantification for benzonitrile solutions. Again, no comment was made regarding these values, which indicate very poor analytical performance in relation to that typical of conventional FTIR spectrometers equipped with single-element detectors.

The above study, which appears to be the most detailed investigation to date of the quantitative performance of an FPA-FTIR spectrometer, was done in step-scan mode with a 64x64 FPA detector, and so the results are not entirely applicable to infrared images acquired in the rapid-scan mode. Nevertheless, generally speaking, it is evident from the literature that the performance of infrared imaging detectors is not comparable (on the pixel level) to that of conventional single-element detectors. In other words, at the present time it is unrealistic to expect that each pixel of an FPA detector can provide spectral performance that is comparable to that of a conventional single-element detector (although some may disagree given the cost of the technology). However, considering the great strides made in infrared imaging since its commercial availability, it is expected that the technology will improve even further in the years to come. For instance, some of the present limitations of this technology arise from the fact that FPA detectors are an addition to existing infrared spectroscopy and microscopy technologies. As results, the components used are not necessarily ideal for imaging purposes. One such example is the mirrors inside the spectrometer, which have been designed for maximum throughput but exhibit high spherical aberration, a phenomenon which is detrimental to image fidelity. Thus, together with improvements in detector fabrication, the design and construction of instruments exclusively for imaging should result in a substantial reduction in non-sample contributors to variability and hence enhanced image fidelity. This will open up the technology to a wider array of potential applications, resulting in an increased level of adoption and further reduction in

instrument costs. It is speculated (by the author) that in the next 20 years infrared imaging will reach a point of maturity and acceptance so as to be as familiar a technology as FTIR spectroscopy is today.

2.2.5 Applications of FPA-FTIR spectroscopy

Considering that FPA-FTIR spectroscopy has only been available to the scientific community for a short while, there are a limited number of examples of its use in the literature, dating back to 1995. To date, the majority of applications employing FPA-FTIR spectroscopy have been in the bio-sciences [7, 14-16] with some examples in the materials and chemical sciences [17-19]. The majority of biological applications have focused on studying and identifying cancerous tissues using infrared images [20-22].

Among the applications of FPA-FTIR spectroscopy that have been investigated to date, bacteria identification [23] is the specific focus of the research presented in this thesis. Accordingly, background information on this application is presented in the following section.

2.3 BACTERIA IDENTIFICATION

2.3.1 By FTIR spectroscopy

The identification of bacteria by infrared spectroscopy was investigated as early as the 1950s [24]. It was identified in the early work that the infrared spectrum of a microorganism provides information on the biochemical constituents of the microorganism. Bacteria identification by infrared spectroscopy is considered a method of “whole organism fingerprinting” due to the fact that the infrared spectrum is acquired from intact whole organisms and consists of a superposition of the absorption bands of all the biochemical components of the organism. As such, the spectrum of a microorganism effectively serves as its fingerprint, allowing one to build a database of ‘fingerprints’ which one can search to identify unknowns.

The key advantages of this approach are simplicity and speed. Once a pure culture of bacteria is available, the protocol merely entails its deposition onto an

infrared transparent slide (e.g. zinc selenide) and the acquisition of its infrared spectrum.

To date, there are hundreds of publications describing the use of FTIR spectroscopy for the discrimination and differentiation of microorganisms. Likewise, there have been numerous publications demonstrating the use of FTIR spectroscopy for the identification and analysis of bacteria. For a very complete look at the research that has been conducted in bacteria identification using FTIR spectroscopy see [23].

2.3.2 Classification techniques for bacteria identification

Nearly every chemometric, artificial intelligence, and statistical method has been explored in FTIR bacteria identification (see Table 2.1 for a complete list of methods, references describing these methods, and references providing examples of these methods in bacteria identification). In addition, an overview of the methods used up until 2001 can be found in [25]. A tabulation of the analysis methods used specifically for the clustering or classification of bacteria by FTIR spectroscopy in 50 publications dating from 1998 to the present day indicates that by far the most popular method (24/50 publications) for analyzing and displaying results is hierarchical cluster analysis (HCA). Although not typically used as a classification method, HCA does provide a clear visual representation of the relationship between the spectra from different bacteria. Additionally, some form of principal component analysis (PCA) was used in nearly every application either to plot principal components directly or to reduce the dimensionality of the data for input into another analysis method. In cases where principal component analysis was not used, a similar chemometric approach was applied such as canonical variate analysis (CVA - 6 publications) or discriminant function analysis (DFA - 6 publications). Soft independent modeling of class analogy (SIMCA) was used in 6 of the 50 publications examined. In general, all methods were deemed successful by the authors of the respective publications. Artificial neural networks (ANN) are a popular approach (8/50 publications) for training classifiers and in some cases were deemed to be essential to successful analysis at the strain level. Support vector machines (SVM) have been applied in very few publications on FTIR bacteria identification, which is likely

a result of the complexity and the limited number of applications of SVM to FTIR data as compared to the other techniques listed. Partial least squares (PLS) is a very popular regression technique that has seen only minimal application to bacteria identification; in the literature reviewed, it was examined by two groups in three publications. While classification is not a regression type problem, PLS has been demonstrated to be applicable to class-based analysis problems and has shown some promise in this area.

Table 2.1: List of chemometric methods with references describing the method and references providing examples in bacteria identification.

Analysis Method	Ref. for description	Examples in bacteria identification
Principal component analysis	[14]	[26-33]
Hierarchical cluster analysis	[34, 35]	[29, 33, 36-43]
k Nearest neighbors	[35, 44, 45]	[29, 45]
Artificial neural networks	[45]	[29, 37, 39, 46, 47]
Support vector machines	[48]	[49]
Partial least squares	[35, 50-52]	[29, 53, 54]
Soft independent modeling of class analogy	[55]	[26, 29, 31, 56]
Canonical variate analysis	[35, 57]	[27, 29, 32, 58]
Discriminant function analysis	[35, 57]	[33, 43, 55, 59]

In the vast majority of the publications examined, the regions of the infrared spectrum to be used for analysis (feature selection – see below) were identified by visual inspection. Additionally, in many cases either the entire spectrum or the fingerprint region (1800 - 950 cm^{-1}) was used. There are several examples in the literature of optimized regions being determined computationally [39, 54, 60], with genetic algorithms (GA) being one of the optimization methods used. While the various approaches to feature selection have not been widely used for the analysis of spectra from bacteria, there has been quite a substantial body of work conducted with infrared

spectral datasets from other applications. The most commonly described approach is genetic algorithms [61] as they are naturally suited to the selection of spectral features. Additionally, forward selection [62] and grid-greedy [63] methods have also been presented in the literature as viable methods for spectral feature selection. However, the success and implementation of these types of algorithms vary substantially depending on the application and so it is necessary to determine their applicability to the differentiation of bacteria at the various taxonomic levels.

Additional details regarding the principal data analysis techniques employed in this study are provided below.

2.4 DATA ANALYSIS

2.4.1 Spectral noise

2.4.1.1 Peak-to-peak noise

Peak-to-peak noise across a spectrum or a region of a spectrum is taken as the total difference in intensity between the maximum and minimum peak height. This measure should be taken across a portion of the spectrum where there are no absorbance peaks. This type of noise measure is sensitive to random fluctuations in the spectrum. In other words, one point of data that is significantly noisier than the rest will become the maximum or minimum taken in determining the peak-to-peak noise but may not be representative of the noise in the remainder of the spectrum.

2.4.1.2 Root-mean-squared (RMS) noise

The RMS noise is the square root of the mean of the squared values at each point across the portion of the spectrum examined. This type of noise measure is less susceptible to random sharp noise peaks because it computes the average across a wide range (the wider the range, the less the susceptibility). In addition, this measure can be used in conjunction with a linear least-squares fit across the points in order to minimize noise contributions as a result of baseline variations.

2.4.1 Spectrum as a Vector

When applying data-analysis and data-processing algorithms, a spectrum is routinely treated as a vector. A vector is an array of values, each value representing a position along a particular dimensional axis. For example, in its simplest form a vector with one value is represented by a point on a line. A vector with two values would correspond to a point on a familiar X vs. Y plot. If a series of spectra were to consist of only two values each, each value representing a different wavenumber, then these spectra could be mapped to a two-dimensional space on an X vs. Y plot, with each point representing a spectrum. However, current instrumentation will typically yield hundreds of points per spectrum. The result is large n-dimensional vectors (where n is the number of points in the spectrum) which are representative of a point in n-dimensional space. As such, standard vector operations can be applied between spectra (i.e addition, subtraction, etc.), and various mathematic, statistic (averaging), and chemometric (principal component analysis) models can be directly applied to a series of spectra. It is important to consider, however, that a precondition for treating a set of spectra in this manner is that the frequency (wavenumber) value of each point in all the spectra must near perfectly coincide (to the desired level of precision). Fortunately, the laser reference employed in FTIR spectrometers ensures excellent wavenumber reproducibility from one spectrum to the next. This condition may not be met when spectra have been recorded on more than one model of spectrometer, thus making it necessary to map the spectra to a common set of wavenumber values. Such interpolation can affect the resultant spectral intensities in ways which are hard to gauge and which are dependent on the spectral features. This is one of the difficulties encountered when evaluating instrument-to-instrument transferability between instruments of different manufacturers.

2.4.2 Vector Transforms

2.4.2.1 Derivative

Several different algorithms for numerical differentiation are applied in spectroscopy to eliminate baseline offsets or tilts and to mathematically enhance spectral resolution [4, 34]. The simplest is a first-difference derivative, which takes the derivative at some point x_i as x_i - x_{i-1}. Another approach referred to as a gap derivative takes the derivative at a point x_i as x_{i+g} - x_{i-g}. A consequence of taking the derivative of a spectrum is increased noise in the resultant derivatized spectrum. By taking a larger gap, the noise in the derivatized spectrum can be reduced but consequently so is the extent of resolution enhancement (with large gap values resulting in loss of spectral information through degradation of the resolution at which the spectrum was recorded). Another approach, referred to as a gap-segment derivative, applies a running average of a particular segment size to the spectrum prior to taking the gap derivative. The resultant derivative will exhibit less noise but, again, some spectral information may be lost. An improved version of the gap-segment derivative, referred to as Norris regression [64], attempts to optimize the size of the gap at each point along the spectrum. Another commonly employed derivatization technique is the Savitzky-Golay algorithm [65], which is equally used as a means of smoothing spectral data. It works by applying a regression of a specified order around the point of interest. For a point x_i and segment size of 5 (two points on either side of x_i), a regression will be determined on the points (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) and the value at x_i determined based on this regression. Given a regression of high enough order (for example, a second-order regression), the derivative is simultaneously computed and can be taken from the higher order coefficients of the regression.

2.4.2.2 Normalization

A variety of approaches exist to normalize a set of vectors [4, 34] but, in general, they all conduct the same operation, which is to divide each value in a vector by a normalization factor. The normalization factor can be derived as the height (absorbance

in a spectrum) of a particular peak which will result in the peak having a value of one after normalization. In spectroscopy it is also commonplace to use a normalization factor as the area under a peak (or set of peaks). Another approach is to use a vector's magnitude as the normalization factor (known as vector normalization). The type of normalization applied is typically dependent on the target application.

2.4.3 Vector distance metrics

One of the more common approaches to comparing two vectors is the Euclidean distance metric (Figure 2.4) [34]. This metric provides a measure of the distance between the two vectors in n-dimensional space by evaluating the straight-line distance between the points represented by the vectors (in the n-dimensional space). As applied to spectral data, this distance is then taken as a measure of spectral similarity. This type of metric is sensitive to changes in vector magnitude. Other metrics commonly used in spectroscopy are the cosine similarity metric, which is invariant to changes in magnitude of the vectors (scale-invariant), and the Pearson correlation coefficient, which examines the linear relationship between two vectors. Another distance metric well suited towards spectral data is the Mahalanobis distance metric [66]. Unlike the Euclidean distance metric, which assumes that points (in n-dimensional space) are spherically distributed about the center of a dataset, the Mahalanobis distance metric considers the case where points may have an ellipsoidal distribution. In other words, both the distance and direction in n-dimensional space are considered with respect to a reference dataset in determining a similarity value. The contribution to the overall distance for a given dimension is weighted by the range of values found in the reference dataset on that dimension. This metric requires computation of the covariance matrix of some reference dataset, which can be a computationally intensive process.

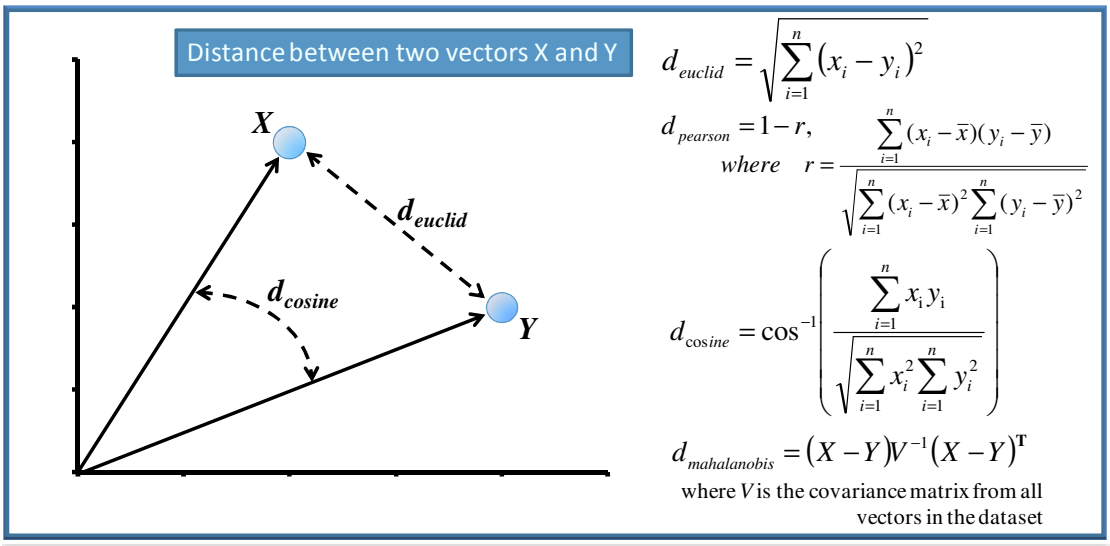


Figure 2.4: Description of the Euclidean, cosine, Pearson correlation coefficient, and Mahalanobis distance metrics where X and Y are two vectors and d the distance between them.

2.4.4 Construction of classifiers

2.4.4.1 Cross-validation

Ideally, when constructing a classifier using reference data, the data would be split into two groups, one used strictly for training the classifier and another for strictly validating that the classifier performs similarly on unseen data. It is often the case, though, that insufficient reference data is available to be split up into two groups. For this reason, cross-validation is often employed, one of the more popular techniques of which is a k-fold validation [45]. k-Fold validation involves splitting the reference data into k groups. A classifier is then built k times where on each iteration one of the groups is excluded from the training process. The excluded group is then used to validate the built classifier and the performance result (ex. number of correct predictions) stored. The expected performance of a classifier built using all of the reference data is taken as the average of the k classifiers previously built. Leave-one-out cross validation is a k-fold validation where k = n, the number of samples. In this scenario one sample is excluded at a time, a classifier built, the sample predicted and the result stored. This is repeated

for every sample in the dataset. For large datasets a leave-one-out cross validation is computationally expensive and will often overestimate the performance of the resultant classifier.

2.4.4.2 K-Nearest Neighbor

K-Nearest Neighbor (k-NN) is a simple classification method that is commonly used for generating classifiers. In its simplest form, k-NN is a lazy technique, meaning that it is not required to conduct a training phase whereby the classifier learns the reference data. Instead, the majority of computational work is conducted at the time of classification. The sole preparation of the data required is the storage of the spectral vectors. An unknown is identified through direct comparison of the unknown and the reference data. The method used for comparison must yield a value which allows for the reference vectors to be ranked whereby the top-ranked vectors are the most similar to the unknown. The unknown then receives an identification based on the *k* top-ranking reference vectors. If k = 1, then the unknown is simply given the same classification as the top-ranked reference vector. This is typically the case when the number of groups *g* is equivalent to the number of reference vectors *n.* If $g << n$ (for example, 20 reference vectors where 10 represent group A and the other 10 group B), then k > 1 may be used.

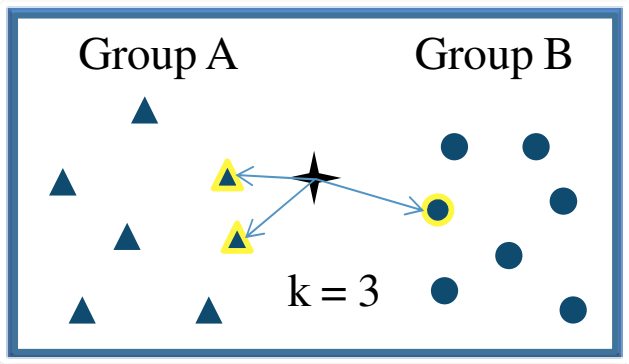


Figure 2.5: Example of a k-NN classification with k = 3. In this example the unknown (star) will be classified as group A based on a 2 out of 3 vote.

If the k top-ranked reference vectors consist of members from different groups, then a majority vote can be applied (Figure 2.5). In addition, the contribution of a top-

ranked reference vector to the vote can be weighted by that reference vector's similarity to the unknown. For example, consider a similarity measure using the Euclidean distance between two vectors. A distance of zero implies that two vectors are identical. If k = 3 and the top-ranked reference vector has a distance value of 2 and belongs to group A whereas the 2nd and 3rd place reference vectors belong to group B and have a distance value of 100 and 101, respectively, then it is clear that the votes of the top-ranked members must be weighted to take into account the discrepancy in similarity. One common approach is to use the measured distance as the contribution value to the vote and apply a classification based on the group with the lowest contribution (equating to the shortest distance). An important aspect in applying the k-NN algorithm is to recognize that this approach to classification will always provide a result. In other words, there are always nearest neighbors, it is how near these neighbors are that is relevant. For example, consider the situation where several cities in several countries around the world are the reference data. An unknown (say a city on Mars) is classified to a particular country using the k-NN algorithm based on the cities which are closest to the city on Mars. Of course, in this scenario the distances between the reference data (cities on Earth) is on a completely different scale than the distances determined in classifying the unknown and so the result should be rejected. There are multiple approaches to determining when a result should be rejected. The largest distance between any two reference spectra can be used as a cut-off, or the average and standard deviation of the distances between all reference data members ($n - (n - 1)$ distances) determined and used to evaluate the distance of the unknown relative to those in the reference dataset. The approach used in this study is to compare the average and standard deviation of reference data on the group level. The majority of data analyzed in this study is that of large datasets which are broken down into a much smaller number of groups ($n >> g$). For each group, the average and standard deviation of the distance to the mean vector of the group is determined. Upon classification to a particular group, the distance of the unknown to the mean vector is evaluated in relation to the previously determined average and standard deviation. This distance

provides a mechanism by which a confidence in the classification can be assigned. For example, a lower than average distance to the mean vector would imply a high confidence; conversely, a distance 1 or 2 standard deviations greater than the average would imply a low confidence. The primary metric used in this study is the Euclidean distance metric, a drawback of which is its invariance to direction. In this case the distance with respect to the mean vectors may be in a different direction than the reference vectors that make up the mean. The Mahalanobis metric is commonly used in spectroscopy for this very reason; however, it is a computationally expensive approach and as a result is currently not amenable to infrared imaging.

The benefits of the k-NN algorithm for classification are simplicity and transparency. It provides an easily understandable approach by which an unknown is identified and because it uses the reference data for classification there is little room for misinterpretation.

Generally speaking, k-NN is an unsupervised classifier in that there is no manipulation of the classification parameters in order to conform to the known designation of the reference data. In other words, the classification routine does not need to know which reference data corresponds to what. It will simply assign the nearest neighbors and return a result. Often, though, it is desirable to omit certain portions of a spectrum as part of the classification process. As noted above (2.3.2), it is unanimously accepted that the isolation of certain regions of the overall spectrum is ideal and often necessary for the construction of classifiers by FPA-FTIR spectroscopy. A feature is defined in this context as a value in a vector, or in the context of spectroscopy, the spectral intensity at a particular wavenumber. The process of feature selection involves a search through the state-space of feature combinations in order to find the combination of features which result in the best possible classifier. By applying a feature selection process prior to building and using k-NN classifiers, k-NN effectively becomes a supervised eager method. In other words, a training phase is employed in order to maximize the classification performance of a k-NN classifier. Used inappropriately, such routines are likely to overfit the reference data, resulting in classifiers that provide

exceptional performance when cross-validated on the reference data but much poorer predictive ability on previously unseen data. In this manner, standard optimization guidelines should be applied. For example, considering the substantive data width of spectroscopic data (each spectrum can consist of hundreds or more points), it is not unlikely that a single point or small number of points can be found which near-perfectly separate the reference data. There is no steadfast rule, but the inclusion of 20 to 30 data-points per spectrum should provide sufficient confidence that the found set of features is not the result of pure chance but is indicative of real (chemical) spectral differences.

2.4.4.3 Hierarchical cluster analysis

Hierarchical cluster analysis (HCA) can be described as the sequential grouping of data elements (in a dataset) into ever larger clusters (agglomerative clustering). The results of a hierarchical cluster analysis are typically visualized through the use of dendrograms. Similarly to principal component analysis (see below), HCA is often used for its ability to visualize the clustering or relationships between n-dimensional vectors in a 2-dimensional space. HCA is a two-stage process, the first of which is to compute the spectral distance (ex. Euclidean distance) between all vectors in the dataset, a total of [n x n – 1]/2 distance computations. The second stage uses a specified linkage method [34] to sequentially group vectors in clusters. The linkage method defines how clusters are linked together, the simple example of which is the ‘single’ linkage method which groups the two clusters which are closest together. Figure 2.6 provides details regarding four different types of linkage methods.

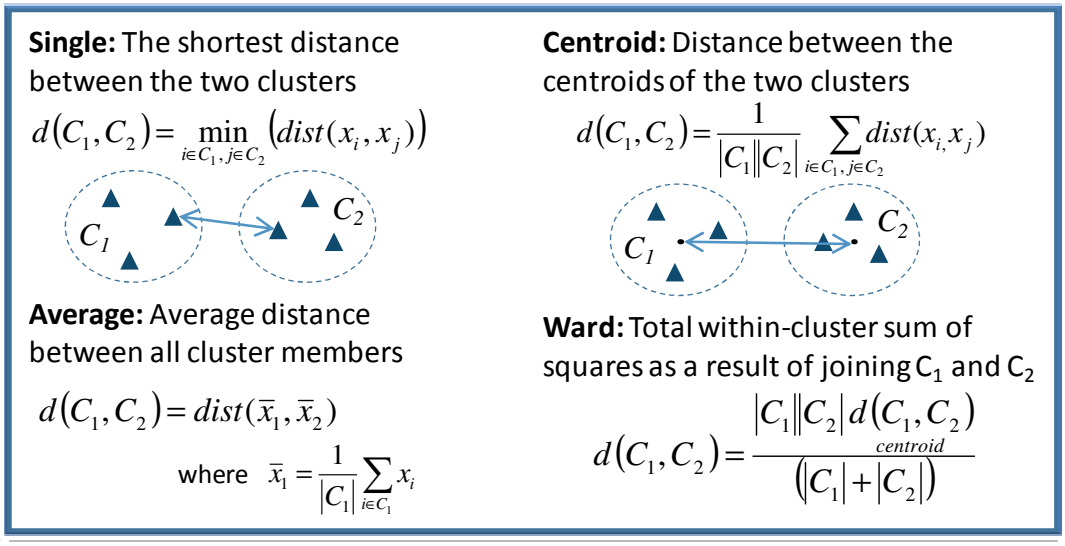


Figure 2.6: Description of four different linkage methods used to join clusters together in a hierarchical cluster analysis.

2.4.5 Features selection

2.4.5.1 Grid-greedy

The grid-greedy approach was first introduced for the selection of spectral features in infrared images in [63], although such an approach is commonly used in chemometrics and artificial intelligence as an optimization or search routine [45]. The grid search is simply a search which examines all possible combinations of features. It is a computationally expensive approach; however, it will necessarily find the optimal result for a given set of parameters. In the context of spectroscopy, a grid search may only be able to find the top two or three regions due to computational constraints, but it is certain these regions are the most optimal. A greedy search examines the combination of adjacent features and follows the path of greatest improvement. In this way it is greedy in that it only considers this path of greatest improvement. A greedy search requires a starting point which is provided by the result of the grid search. In addition, it allows for the examination of regions smaller than those examined by the grid search, and is, in essence, a fine tuning of the grid search result.

2.4.5.2 Forward search

Forward search is another commonly applied technique for search and optimization in chemometrics and artificial intelligence. It is a greedy search where the starting point is an empty set of features. The first feature selected is that which results in the best separation of the reference data. Afterwards, features are successively selected based on those that improve or maintain the highest level of separation. The routine will stop when adding additional features results in a poorer classifier and a minimum number of features have been selected. An interesting variation of the forward search can be seen in [62] which employs a mathematical property of the cosine distance metric to offer substantial improvements in search speeds.

2.4.5.2 Genetic algorithms

A genetic algorithm is a non-deterministic optimization routine which uses the principle of natural selection in an attempt to evolve an optimal result [45]. It allows for simultaneous examination of a large portion of the state space (the permutations of possible features to choose) and through mutation and cross-linking, analogous to their natural counterparts, evolves better performing feature sets. As with most optimization algorithms there is the possibility of arriving at a local maximum. A GA consists of a population of individuals with each individual having a unique set of genes. In this approach each gene is representative of a single feature (point at single wavenumber in the spectrum) which can have one of two states, 1 or 0, where 1 implies the inclusion of a feature and 0 the exclusion. Each individual of the population is randomly initialized and then evaluated by means of a fitness function. The fitness function is a measure of how well a set of features separates the classes. The better performing individuals are mated by cross-linking their genes (Figure 2.7) to create new population members. The selection of individuals to mate together is done semi-randomly by biasing the selection towards individuals that are better performing. In addition, a percentage of the top members may be included in the following generation. After mating completes, the new individuals are then evaluated and the process repeated. The overall population size is kept constant from one iteration to the next. The process continues until there is little

or no change in the classification score of the top-performing individual or a fixed number of iterations has been reached. In addition, at random intervals a random mutation will be introduced into an individual to stimulate variety. This random mutation serves as a means of avoiding local maxima.

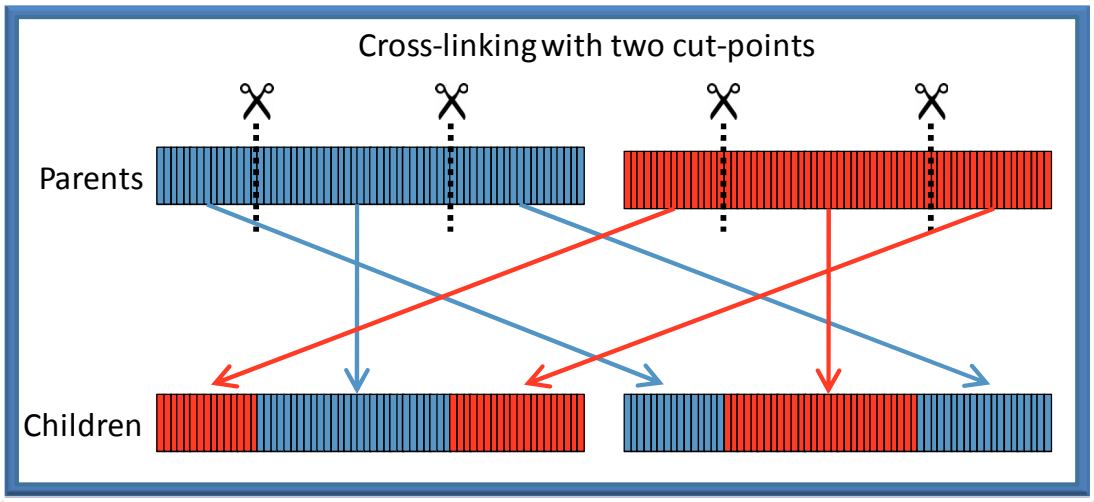


Figure 2.7: Description of the mating process between two individuals of a population. Cut-points are randomly chosen and the series of genes of each individual mixed to form the children.

There are a multitude of variations on the general GA concept. These include more complex approaches to selecting the individuals for mating and the actual mating procedure.

2.4.6 Principal Component Analysis

Principal component analysis is a commonly used approach for data-space reduction, noise elimination, and identifying features which most contribute towards variability in a dataset. Consider a dataset of n-dimensional vectors, as described previously, where each vector represents a sample and each value of a vector represents the position along a particular dimension or axis. The underlying concept of principal component analysis is a rotation of these axes to a new set of axes where the first axis (the first value in the transformed vectors) describes the largest contribution to variability, the second axis the second largest, and so on (Figure 2.8).

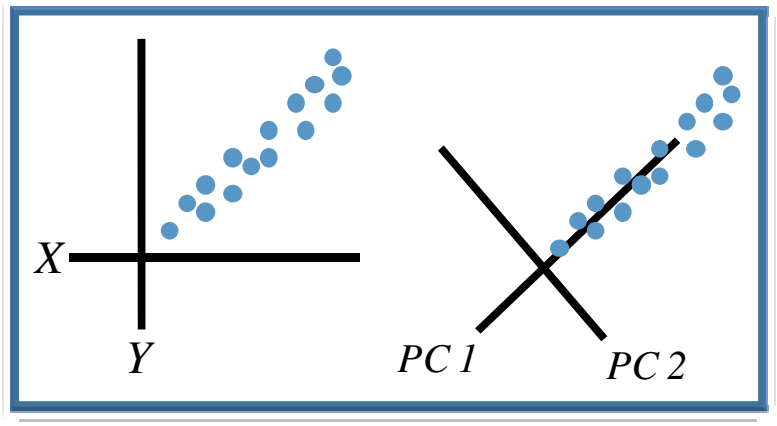


Figure 2.8: A 2D representation of PCA. By rotating the axis (X & Y) the majority of information can be described by a single axis (PC 1), whereas the remainder of information (PC 2) can be primarily attributed to noise.

In this manner the majority of information is often provided by the first few values (known as principal components) in the transformed vectors and can in many cases describe a dataset of spectra, each spectrum containing hundreds of points, with a dataset of vectors (principal components) of only several values and with minimal loss of information. The principal components beyond the first few will typically represent a very minor component of the original spectral information (< 1%). By discarding these values it is found that PCA can be used as an effective means of data noise reduction. As such, PCA can be used for data-space reduction, analysis of variability, and noise reduction. Furthermore, the data-space reduction and concentration of spectral information in the first few values of the transformed vectors will often allow for the separation of data groups in a 2D or 3D space. This is one of the major reasons for the use of PCA, as it allows for the separation and visualization of spectral data which normally resides in an n-dimensional space into an easily interpretable and plot-able 2D or 3D space.

2.4.7 Partial Least Squares Regression

Partial least squares (PLS) regression is a technique for multiple linear regression which tries to specify a linear relationship between the measured values (the spectra) and, in the case of infrared spectroscopy, concentration. It shares similar attributes to

PCA in that it attempts to find a set of latent variables (analogous to principal components); however, in this case the data matrix rotation determines the components in the data matrix that are also relevant to the concentration values. In other words, the first latent variable will describe the most variability in both the data matrix and concentration values simultaneously, while the second latent variable the second most, and so on. A regression is then performed on the rotated data matrix in order to build the calibration model. Examine [50] for a detailed description and technical derivation of PLS.

PLS regression is a “whole spectrum” technique whereby the calibration model is built using the entire spectrum (or rather the complete vector representing the spectrum). This is in contrast to a Beer’s law [67] type calibration, where only a single value, such as a peak height or peak area, is used in construction of the calibration model. One of the more significant advantages of PLS regression is that it can build calibrations with spectra having multiple overlapping components. A successful calibration can often be built from spectra consisting of multiple components where the concentration of only one of the components is known. This is unlike classical least squares [68], a multiple linear regression whereby all contributing components of the spectrum must be accounted for in order to build a valid model. In addition, the pure component, the spectrum consisting of 100% concentration of one component, can often be extracted using the resultant PLS model. A non-trivial and important requirement for employing a PLS regression is the selection of the number of latent variables to incorporate in the PLS model. A variety of statistical approaches have been suggested (see [50]), the majority of which rely upon the predictive error sum of squares (PRESS) value. The simplest approach is to plot the PRESS values for an increasing number of latent variables and take the point where the reduction of the PRESS value becomes minimal (i.e. where the plot flattens). Depending on the calibration data, differences in the selection of latent variables can have an effect on the predictive results of the PLS model and the pure components generated.

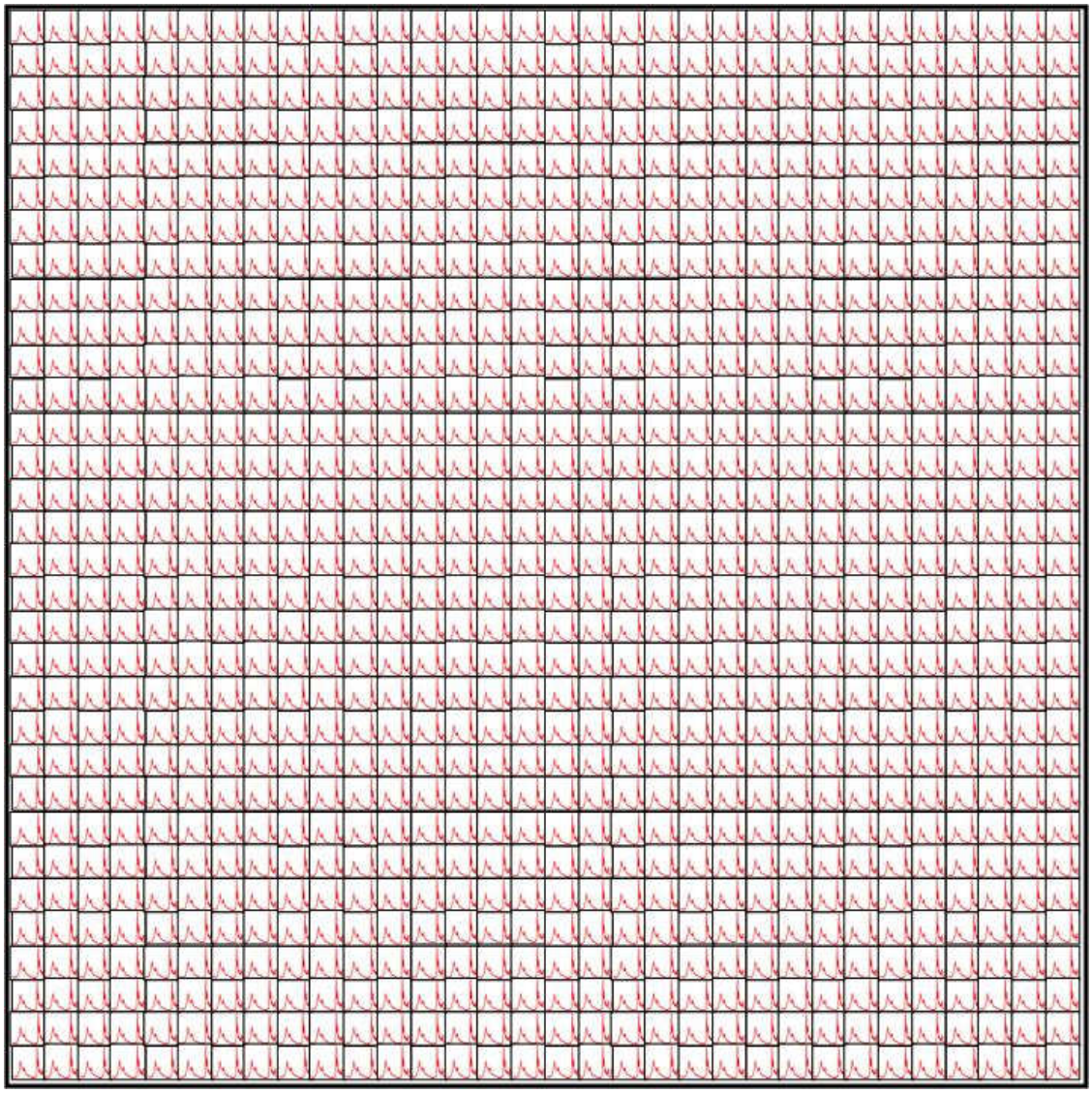


Figure 2.3: A visualization of 1024 spectra from a single 32x32 infrared image acquired in 2 minutes on a FPA-FTIR spectrometer. Each spectrum in each box consists of 100's to 1000's of points. As can be seen, this type of technology generates a tremendous amount of data in a very short period of time. Such quantities of data cannot be examined visually and require spectral processing algorithms for interpretation.

CONNECTING STATEMENT

In accordance with the aim of developing a complete methodology for the identification of bacteria by focal plane array Fourier transform infrared spectroscopy, the research described in the following chapters proceeds through the stages necessary to process infrared image data, apply optimization techniques, ensure sample homogeneity and construct classifiers capable of identifying unknown samples.

Prior to developing routines to filter and analyze infrared image data of bacteria, it is necessary to develop an understanding of the performance expectations and capabilities of a FPA-FTIR spectrometer. The literature has identified a number of contributors to variability and their potential influence on the spectra acquired by infrared imaging. Chapter 3 examines the degree to which non-sample sources of variability affect the spectrum acquired by each pixel in an infrared image by examining a variety of scenarios. In addition, our unique position of possessing multiple FPA-FTIR spectrometers allowed for a direct comparison of the levels and distribution of non-sample variability from one instrument to the next.

Chapter 3:

Evaluation of Image Fidelity in FPA-FTIR Spectroscopy

3.1 INTRODUCTION

Focal plane array Fourier transform infrared (FPA-FTIR) spectroscopy is a relatively new technology that has opened new avenues of scientific investigation and new applications of infrared spectroscopy in the bio-sciences [7, 14-16] as well as in chemistry and materials science [17-19]. The majority of studies to date have employed FPA-FTIR spectroscopy as an imaging technique, whereby the chemical composition of heterogeneous samples is probed on the micron scale. For example, a large number of studies have focused on examining and identifying cancerous tissues using infrared images [20-22]. There have also been a number of time-resolved studies exploiting the spatial relationship of pixels in conjunction with micro-fluidic mixers [69]. However, the data-rich nature of FPA-FTIR spectroscopy, which acquires thousands of times the amount of data that was previously possible in a comparable amount of time, opens up new applications which can benefit from high data-redundancy and spatial resolution. One such example is bacteria identification [23, 63], where the use of FPA-FTIR spectroscopy has been shown to enable a high-speed and robust approach to classifying bacteria at the species level. Other possible exploitations of infrared imaging would be for use in quantitative applications. Applications of this sort could include the use of multi-channel microfluidic systems for the simultaneous monitoring and analysis of fluids. This type of system could be used for quality control, process monitoring, or time-resolved studies, and depending on the configuration, a single FPA-FTIR spectrometer could replace multiple single element instruments. However, these types of applications necessitate high levels of reproducibility on a pixel-to-pixel and a detector-to-detector basis. Without a sufficient degree of image fidelity (the degree of which is dictated by

the application), the possibility of employing infrared imaging for quantitative and possibly certain qualitative applications is severely diminished. This study assesses the performance, response, and reproducibility characteristics of FPA detectors on a pixel-to-pixel and instrument-to-instrument basis to consider its further potential for use in quantitative and qualitative applications which specifically employ FPA-FITR spectroscopy for its data-rich, data-redundant, high-throughput, and high spatial resolution capabilities.

An FPA-FTIR spectrometer consists of a standard spectrometer attached to an infrared microscope. The infrared imaging detector is located on top of the microscope. The source and interferometer are located inside the spectrometer with an automated mirror to redirect the infrared beam into the microscope. The focusing optics incorporated into the microscope allows each pixel of the FPA detector to acquire a single spectrum from an average area of 5.6 µm x 5.6 µm. The average is taken as the true spatial resolution is related to the diffraction limit of a particular wavelength [6]. The pixels of the detector acquire data concurrently, resulting in the generation of thousands of spectra in a matter of seconds to minutes, depending on the number of spectral co-additions. Considering that each pixel is a unique detector capable of generating a complete spectrum, it is necessary to evaluate detector performance on a pixel-to-pixel basis.

The majority of FTIR spectrometers currently in use are detector noise limited, whereby the limitation of their sensitivity and random variability in the spectrum is a result of the noise of the detector. FPA-FTIR spectroscopy is, however, substantially more complex and there are several known sources of detector variability present in FPA-FTIR spectrometers that are not present in their single-element counterparts (see Chapter 2.2.4). These sources include high-magnitude noise as a result of the multiplexed nature of FPA detectors. Such noise is statistically small for a single-element detector but is substantially increased as a result of the hundreds to thousands of detector elements in an FPA detector. Its effect, though, is most prominent when images are acquired in step-scan mode (this study deals exclusively with operation of

the spectrometer in rapid-scan mode) due to the higher acquisition times. Another important source is the non-uniform distribution of energy that reaches the detector. Thus, when the total response of the detector is examined with an unobstructed beam, a “hot-spot” (2D Gaussian) across the array is noted. Prior to acquiring a background or sample image, a two-point linear calibration at either extremities of response is performed in order to estimate the response curve of each detector element. This process is adequate when dealing with uniform, homogeneous, isotropic samples but with real-world samples this is not typically the case. The result is that the infrared light is diffused, diffracted, and scattered by the sample, and thus the amount of energy reaching the detector is different from that with which it was calibrated. Finally, there is spatial pattern noise, also known as fixed pattern noise (with regard to the position on the array), which is a result of the complex and intricate manufacturing process of the FPA detector [5, 8]. The result is a non-random spatial pattern of noisy pixels in the infrared image, the effect of which is clearly illustrated in the results below.

An evaluation of the extent to which these sources of variability affect the quantitative capabilities of FPA-FTIR spectroscopy was conducted in 1999 by Snively and Koenig [12]. Results of note from that study included the lack of a square-root improvement in noise in relation to the number of co-additions and variability in pixel-to-pixel absorbance on a measured peak corresponding to a coefficient of variation of between 12% and 16%, the degree of which was affected by the number of co-added scans. As the last ten years have seen substantial improvements in instrument electronics and detector fabrication, a fresh examination into FPA detector performance is warranted. Pixel-to-pixel variability is examined in this study in the context of root-mean-squared (RMS) spectral noise, within a framework of per-pixel calibrations (Figure 3.1), and by comparing peak height ratios on a pixel-to-pixel basis. By building a complete calibration with each pixel of an infrared image, the calibration parameters (slope and intercept) and error can be evaluated on a per-pixel basis. In addition, the study was extended to multiple instruments to ensure that the results generated were representative of FPA detectors in general. Pixel-to-pixel and instrument-to-instrument

differences were evaluated based on the measured noise characteristics of a 100% transmission line, by evaluating pixel response from peak measurements in a simple Beer's law [67] calibration, and by comparing spectral peak height ratios.

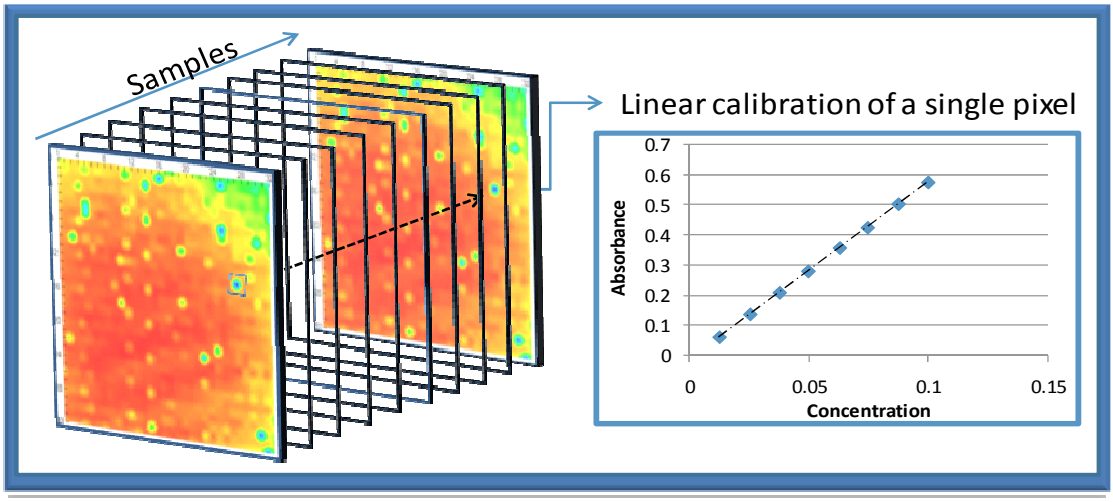


Figure 3.1: Each pixel is a unique detector capable of acquiring a complete spectrum (not shown). Calibrations can be built on a single pixel, and the calibrations between pixels compared for pixel-to-pixel reproducibility.

3.2 MATERIAL AND METHODS

3.2.1 Instruments

Five detectors (3 FPA detectors and 2 single element detectors) as part of four instruments were used in this study. Three instruments consisted of an Excalibur 3000 FTIR spectrometer (Varian, Melbourne, Australia) equipped with a deuterated triglycine sulphate (DTGS) detector. These spectrometers were coupled to a UMA 600 infrared microscope (Varian, Melbourne, Australia) which houses the 32x32 mercury cadmium telluride (MCT) focal-plane-array detectors. For the remainder of this report the FPA detectors are labeled as FPA 1, FPA 2, and FPA 3, the order in which the detectors were available for use. To our knowledge, FPA 2 and FPA 3 were fabricated by a newer process than that of FPA 1, though the details and differences of the fabrication process are not known. One Excalibur 3000 spectrometer in addition to an MB-Series FTIR spectrometer (ABB Bomem, Quebec, Qc) equipped with a DTGS detector was used for

comparison between the FPA detectors and the single element DTGS detectors. The FPA detectors acquisition parameters were set to a value which maximizes the SNR at a small cost of increased acquisition time. A constant flow of dry air was used to purge the spectrometer and the microscope, limiting spectral contributions from carbon dioxide and atmospheric water vapor. Each image was collected from a field of view of 188 x 188 μm^2 with a spatial resolution of 5.6 x 5.6 μm^2 per pixel.

3.2.2 Datasets and Sample Acquisition

Several datasets were used as part of this study. The first consisted of 100% transmission spectra acquired at 8 cm^{-1} resolution by co-addition of 256 scans (unless otherwise noted). The second consisted of methyl myristate ($CH_3(CH_2)_{12}COOCH_3$) added in increasing concentration to heavy mineral oil. Five samples were prepared ranging in concentration from 4.06 x 10^{-2} moles/L to 2.90 x 10^{-1} moles/L. Samples were manually injected into a 100-µm pathlength flow-through transmission cell using disposable syringes. The changes in the concentration of methyl myristate were monitored in the infrared spectrum by tracking the changes in the intensity of the ν(C=O) band at 1747 cm^{-1} with a full-width half-maximum (FWHM) of 15 cm^{-1}. These samples were analyzed on FPA 2 and FPA 3 (FPA 1 was no longer available at this time) at 8 cm^{-1} resolution with co-addition of 128 scans. The third dataset consisted of sodium hydrogen cyanamide (NaNHCN) added to ethanol in increasing concentrations ranging from 7.5 x 10^{-3} moles/L to 6.25 x 10^{-2} moles/L. The same 100-µm cell was used, and the concentration of NaNHCN measured using the band at 2110 cm^{-1} (assigned to ν(C≡N)) with a FWHM of 39 cm^{-1}. These samples were analyzed on FPA 2 at 4 cm^{-1} resolution and 128 co-additions (80 second acquisition) and the Bomem MB-series spectrometer at 4 cm^{-1} resolution and 32 co-additions (53 second acquisition). In addition, multiple images were acquired from an 85-µm thick polystyrene reference standard.

The acquisition process consisted of the following steps: An initial adjustment of the microscope sub-stage is required to maximize energy throughput to the FPA detector. The FPA detector is open-beam calibrated and a background taken. Examining the background ensures the dry-air purge is maintaining low moisture levels. A hundred

percent line open-beam spectrum is acquired to verify instrument noise levels and that the temperature of the detector (which is liquid nitrogen cooled) has stabilized. The transmission cell is positioned in the beam path on the microscope stage and the optics adjusted to focus within the fluid portion of the cell. The transmission cell is secured in place to ensure that throughout the entire experiment, all images are acquired from the identical area of the transmission cell. This was a very important step in an attempt to minimize variability which may occur as a result of removing the transmission cell from the microscope stage between samples. Placing the transmission cell into the beam path requires that the sub-stage be readjusted to re-establish maximum energy flux to the FPA detector. For this reason it is necessary to utilize a zero concentration sample as the background spectrum against which to ratio the sample spectra. The transmission cell is loaded with the pure solvent (zero concentration) and a background spectrum taken. Each of the samples is sequentially loaded into the transmission cell and the images collected are ratioed against the zero concentration background image.

3.2.3 Data analysis

3.2.3.1 Noise analysis

Noise values were measured by calculating the RMS noise (Equation 3.1) with respect to a baseline determined using a least-squares fit [70].

$$RMS = \sqrt{\frac{\sum (y_i - \bar{y})^2 - \frac{\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum (x_i - \bar{x})^2}}{n-2}} \quad (3.1)$$

y = intensity, x = wavenumber, n = # of points, $\bar{x}$ & $\bar{y}$ indicates the average of x and y respectively

With modern-day FTIR spectrometers it is typically found that the noise (y) improves as a function of the square root (root-factor r = 0.5) of the number of scans (x) that are acquired (Equation 3.2)

$$y = n_c x^{-r} \quad (3.2)$$

where n_c is the noise of a single scan (x = 1). However, when comparing the noise between instruments it is of greater interest to describe the relationship in terms of time (*t)* as each instrument may require significantly different amounts of time to perform the same number of scans. As such, the time-per-scan T (in seconds) can be used to define a new relationship (Equation 3.3) where *t* is the scan time, and n_t is the noise performance relative to scan time.

$$y = n_c \left(\frac{t}{T}\right)^{-r} = n_t t^{-r} \quad (3.3)$$

Similarly to co-adding repetitive passes of the interferometer, infrared imaging presents the possibility of averaging pixels in order to attain a spectrum with a lower signal-to-noise ratio than the individual pixels (Equation 3.4, a = the # of pixels averaged). Such a relationship requires that the noise characteristics from one pixel to the next be similar, otherwise the assumption fails and the benefits achieved will be diminished (root-factor *r* will be less than 0.5).

$$y = n_a a^{-r} \quad (3.4)$$

3.2.3.2 Beer's law calibration

The Beer-Lambert law (often referred to as Beer's law) is a concept that is often applied to building linear calibrations for quantitative use in infrared spectroscopy. With regards to infrared spectroscopy, it states that the absorbance of a species is directly proportional to its concentration. Thus, by measuring spectral properties (such as the height or area under a peak) from spectra across a range of concentration values, a calibration can be made relating absorbance to concentration. When a linear model is made relating concentration to absorbance for prediction of concentration from absorbance, the calibration is referred to as an inverse Beer's law calibration.

3.2.4 Data processing

After acquisition, the infrared image data was analyzed and the results generated using software developed in-house on the Visual Studio .Net 2005/2008 platform (Microsoft, Redmond, CA). This included all spectral processing (excluding fast Fourier transforms and background ratioing, which were performed by the instrument software), noise analysis, generation of calibrations, and images shown in this study. The processing routines were validated (where possible) by visual and computational comparison of outputs from Matlab 6.5 (Mathworks, Natick, MA) and Omnic 6/7 (Thermo Nicolet, Madison, WI). The processing was completed on a quad core Intel (Santa Clara, California) processor operating at 3.2 gigahertz with 4 gigabytes of random-access-memory running Windows Vista Business edition. The software developed as part of this study was optimized for parallel processing on multi-core central processing units (CPUs) allowing for significantly greater computational output than most other software available at the time.

3.3 RESULTS AND DISCUSSIONS

3.3.1 Analysis of instrument noise

The spectral noise and signal-to-noise levels in spectra acquired using an FPA-FTIR spectrometer were examined initially to assess the performance of individual pixels and compare it with that of spectrometers equipped with single-element DTGS detectors. One hundred percent transmission lines were acquired by co-addition of 256 scans and the RMS noise was calculated for each pixel. Figure 3.2A provides the RMS noise distribution profile for three FPA detectors. It is clearly seen that there are substantial differences between the noise distribution profiles of the detectors in terms of the range of pixel noise values and their average. While FPA 1 appears to have a more normal-like distribution, that of FPA 2 and FPA 3 appears highly skewed. As a consequence, the RMS noise of the pixels from FPA 2 and FPA 3 falls over a wider range of values, and these two detectors can generally be regarded as poorer performing.

Figure 3.2B1-2-3 spatially illustrates the pixel-to-pixel RMS noise, where red indicates higher noise and blue lower (note: the image for each FPA is on a different scale).

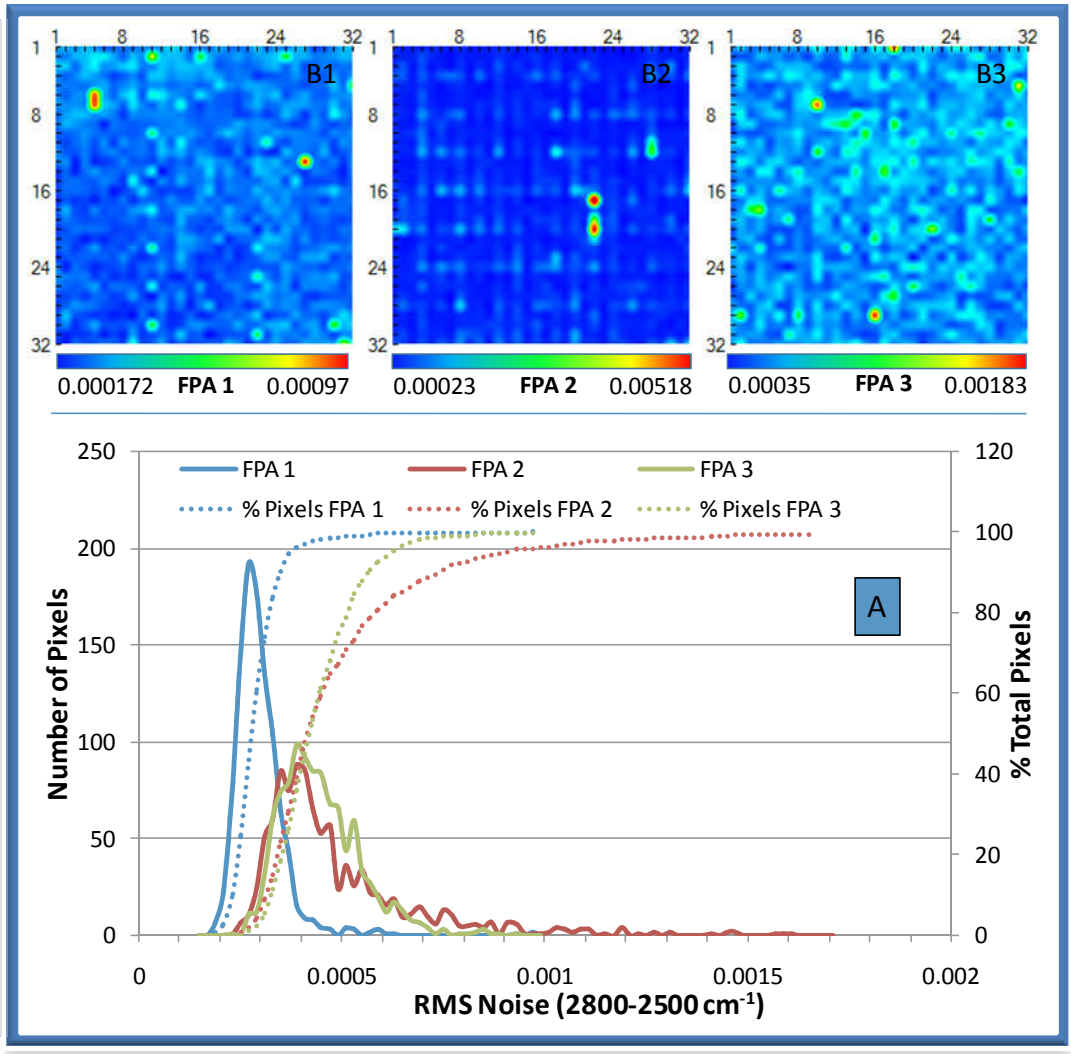


Figure 3.2: Statistical (A) and spatial (B) distribution of pixel RMS noise values. In B1-2-3 a blue value indicates lower noise, red higher, and each image is on a different scale. Aside from demonstrating the different distribution of noise values, spatial pattern noise is visible in the distribution of the pixels, most pronounced in B2.

From the images acquired by FPA 1 and FPA 2 it is apparent that there are spatially observable trends which are indicative of non-random factors affecting the noise performance of the pixels. In the image acquired by FPA 1, there are several columns (one very clear) where every 4th pixel exhibits a significantly greater than average RMS

noise. Similarly, but to a much greater extent, every 4th row and nearly every second column of FPA 2 contains pixels with higher levels of noise. The presence of these pixels explains the extended tail in the noise distribution profile of FPA 2, where nearly 20% of pixels have noise values substantially greater than the distribution peak. In the case of FPA 3, the spatial distribution of noisier pixels is not nearly as obvious; however, a minimal striping (columns of noisier pixels) is visible. From these distributions it appears that the information regarding the manufacturing process of the three FPA detectors is correct, namely, that FPA 2 and FPA 3 have been constructed by a newer/different fabrication process than that of FPA 1. The similar spatially non-random noise distribution profiles of FPA 2 and FPA 3 as compared to FPA 1 are most likely the result of spatial pattern noise, a property of the fabrication process of the FPA detector. According to the manufacturer and the manufacturer's software, this phenomenon, as seen in Figure 3.2, is considered within acceptable tolerances, and all three FPA detectors are performing within approved specifications (according to the acquisition software). The consequences of the pixel noise variability and techniques for coping with it are discussed below.

In order to place the measured FPA detector noise values in context, they were compared to those attained using two single element DTGS detectors. Table 2.1 lists the reduction in noise as a function of acquisition time (as opposed to the number of co-additions acquired) for each instrument. Contrary to the result of Snively and Koenig [12], all three FPA detectors exhibited a square-root reduction in noise as a function of scan time. This can probably be attributed to improvements in the technology (detector and electronics), and the operation of the spectrometer in rapid-scan mode as compared to step-scan mode. The latter results in a reduction in scan time which in turn reduces the occurrence of high-magnitude noise, a significant contributor to variability when the spectrometer is operated in step-scan mode. Likewise, the single-element instruments similarly produced the expected square-root reduction in noise in relation to the scan time.

From Table 3.1, it can be noted that the performance of an individual pixel (taken at the peak of the noise distribution profile in Figure 3.2A) from FPA 1 is 47.4 times poorer than that of the identical spectrometer (Excalibur) equipped with a single-element detector. The results for FPA 2 and FPA 3 reveal an even greater gap between the performance of the individual pixels and the single-element detectors, especially when considering the skewed noise distribution profiles of FPA 2 and FPA 3 and the significant number of pixels which have substantially higher noise values than the distribution peak.

Table 3.1: Actual and relative RMS noise for each detector normalized to 1 second scan time. *DP is the noise taken at the distribution peak visible in Figure 3.2A.

Instrument	RMS Noise	Relative
Bomem	2.05×10^{-4}	1.00
Excalibur	7.16×10^{-5}	0.35
FPA 1 Pixel DP	3.4×10^{-3}*	16.6
FPA 1 Avg.	3.12×10^{-4}	1.5
FPA 2 Pixel DP	5.05×10^{-3}*	24.6
FPA 2 Avg.	1.02×10^{-3}	5.0
FPA 3 Pixel DP	5.1×10^{-3}*	24.9
FPA 3 Avg.	1.35×10^{-3}	6.60

The poor noise performance of individual pixels can be somewhat improved by exploiting the multi-channel nature of FPA detectors, albeit with a consequent loss of spatial resolution. The high number of spectra acquired in a single image affords the option of averaging pixels in order to attain spectra with lower noise without having to increase the scan time. One must consider, though, that to achieve a similar noise level as the DTGS detectors for the identical scan time may require more pixels than are available on an FPA detector. For example, for FPA 2 or FPA 3 to achieve the equivalent noise level of the Excalibur spectrometer equipped with a single-element detector would require the co-addition of approximately 5,000 pixels, 5 times more pixels than is available on a 32 x 32 detector. This is assuming that the noise characteristics are

consistent from pixel to pixel, thus yielding a square-root improvement in noise as a function of the number of co-added pixels. However, as was seen in Figure 3.2, the noise distribution profile is not uniform and, as a result it would require significantly more pixels than are available with current detectors. For example, a 0.4 as opposed to 0.5 root factor, r in Equation 3.3, would require 100,000 pixels for a 100-fold improvement instead of 10,000 pixels. Figure 3.3 shows the improvement in noise as a function of increased number of pixels averaged for each detector. This was accomplished by randomly averaging a selected number of pixels (2, 4, 8, 16,…, 1024) and repeating until all pixels were used once (only factors of 1024 were used so that no pixels were discarded). The noise values were computed for each average spectrum and the average of the noise values reported. Due to the different noise properties from pixel to pixel, it was expected that the improvement in noise as a function of number of pixels averaged would be less than the square-root factor seen when increasing the number of co-added scans.

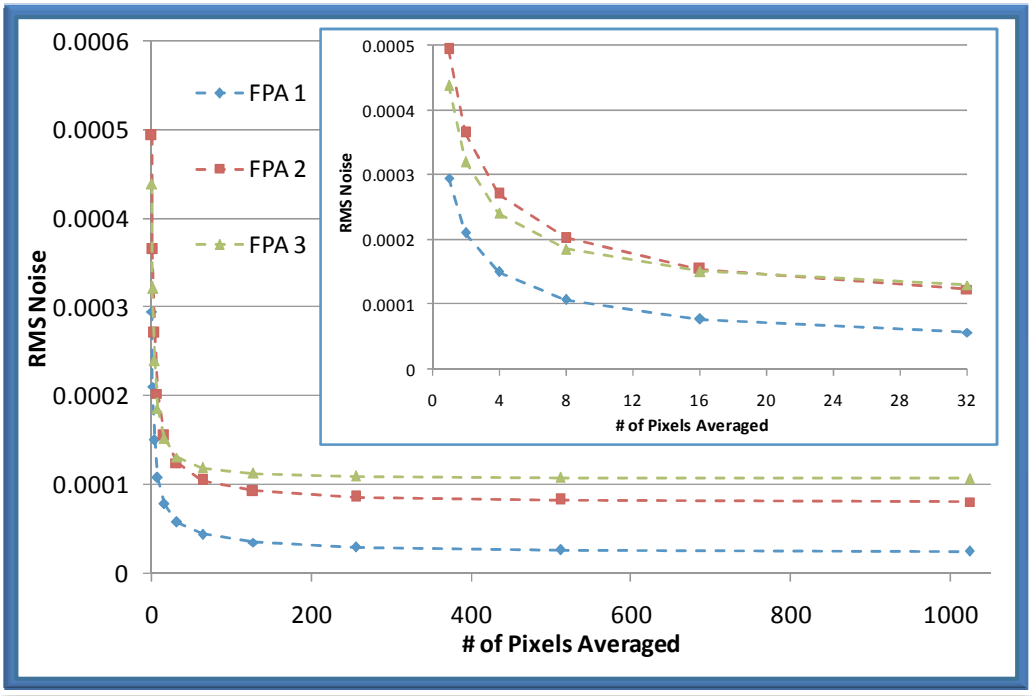


Figure 3.3: RMS noise (2800-2500 cm^{-1}) as a function of the number of pixels are averaged together. The dotted lines are simply a connection of the points in order to provide a visual cue of the power-like trend. It does not represent an actual power fit trendline.

As can be seen in Figure 3.3, averaging pixels together does provide a power trend and there are reasonable improvements when averaging less than 64 pixels together. However, as more pixels are averaged together, the noise level approaches a plateau, to the extent that there is very little improvement between the average RMS noise of four spectra consisting of 256 averaged pixels and the RMS noise of a single spectrum consisting of all pixels in the image (1024) averaged together. Table 3.2 provides details regarding the improvement in the RMS noise relative to the average RMS noise of individual pixels.

Table 3.2: Tabulation of the power fit parameters taken relative to the average noise of pixels (no averaging). The values are generated by determining the power trend line ($y = cx^r$) from the noise values up to the indicated amount of averaging. For example, the values indicated for averaging 16 pixels together consists of the power fit of no averaging and averaging 2, 4, 8, and 16 pixels together. The absolute value of the root factors (RF) are shown.

FPA 1				FPA 2				FPA 3			
# of Pix. Avg.	RF	Coeff.	R^2	# of Pix. Avg.	RF	Coeff.	R^2	# of Pix. Avg.	RF	Coeff.	R^2
4	0.48	0.999	1	**4**	0.43	0.999	1	**4**	0.43	0.996	0.999
8	0.48	0.999	1	**8**	0.42	0.997	0.999	**8**	0.41	0.986	0.998
16	0.48	0.996	0.999	**16**	0.41	0.99	0.999	**16**	0.38	0.967	0.993
32	0.47	0.989	0.999	**32**	0.4	0.975	0.997	**32**	0.35	0.939	0.984
64	0.46	0.976	0.998	**64**	0.37	0.949	0.992	**64**	0.31	0.902	0.968
128	0.44	956	0.996	**128**	0.35	0.916	0.983	**128**	0.28	0.859	0.946
256	0.42	0.924	0.992	**256**	0.32	0.875	0.969	**256**	0.25	0.816	0.918
512	0.4	0.882	0.983	**512**	0.29	0.829	0.949	**512**	0.22	0.773	0.886
1024	0.37	0.833	0.969	**1024**	0.26	0.782	0.924	**1024**	0.19	0.733	0.853

The ideal theoretical improvement relative to no averaging at all would consist of a root-factor of 0.5 and a coefficient of 1.0 with an R^2 correlation of 1. FPA 1, which has a tighter pixel noise distribution than FPA 2 and FPA 3, achieves a root-factor of 0.46 when every 64 pixels are averaged together, whereas the root factors for FPA 2 and FPA 3 are 0.37 and 0.31 respectively. As such, FPA 1 not only has less average per-pixel noise than FPA 2 and FPA 3 but also shows greater improvements upon averaging of 64 pixels. As more pixels are averaged together, the root-factor decreases substantially, as does the coefficient and the quality of the fit, which is indicated by R^2. In addition, as the

number of pixels averaged together is increased, the power fit is in fact overestimating the improvement; In other words, the benefit of averaging diminishes substantially as more pixels are averaged together. This is further illustrated for clarity in Table 3.3, which considers the relative improvement in RMS noise with an increasing number of pixels averaged together.

Table 3.3: The relative improvement in RMS noise as a function of increased averaging. Each FPA is assessed independently with the average noise of individual pixels assigned a value of 1.

# of Pix. Avg.	FPA 1	FPA 2	FPA 3
4	1.96	1.82	1.83
8	2.74	2.44	2.36
16	3.79	3.19	2.90
32	5.13	3.98	3.36
64	6.74	4.71	3.70
128	8.49	5.34	3.90
256	10.04	5.77	4.02
512	11.24	6.00	4.08
1024	11.97	6.14	4.11

As seen, combining multiple pixels with substantially differing noise characteristics can significantly diminish the benefit of averaging. For this reason, the benefits of applying a noise filter to eliminate pixels with noise values beyond an acceptable range was examined. The intention of this filter is to eliminate all pixels in an image that do not meet a predefined RMS noise cut-off value. The result of applying a successively lower noise cut-off and the resulting noise of the average spectrum from all remaining pixels after filtration is presented in Figure 3.4.

Due to the tighter noise distribution of FPA 1, there is no improvement in the RMS noise of the average spectrum as the RMS noise cut-off is decreased. The reason for this is that although noisier pixels are being eliminated, there are also fewer pixels to average together, effectively cancelling out the benefit of removing the noisier pixels. The results are very different for FPA 2 and FPA 3. Owing to the wide distribution of pixel noise and the significant number of pixels with much greater than average noise,

eliminating the noisier pixels reduces the RMS noise in the average spectrum. In fact, the lowest noise is obtained with 23% of pixels remaining in the case of FPA 2 and with only 5% of pixels remaining in the case of FPA 3, suggesting that FPA 3 contains a small number of low-noise pixels. The concept of applying a noise filter will be discussed in the following sections as part of developing quantitative calibrations using FPA-FTIR spectroscopy.

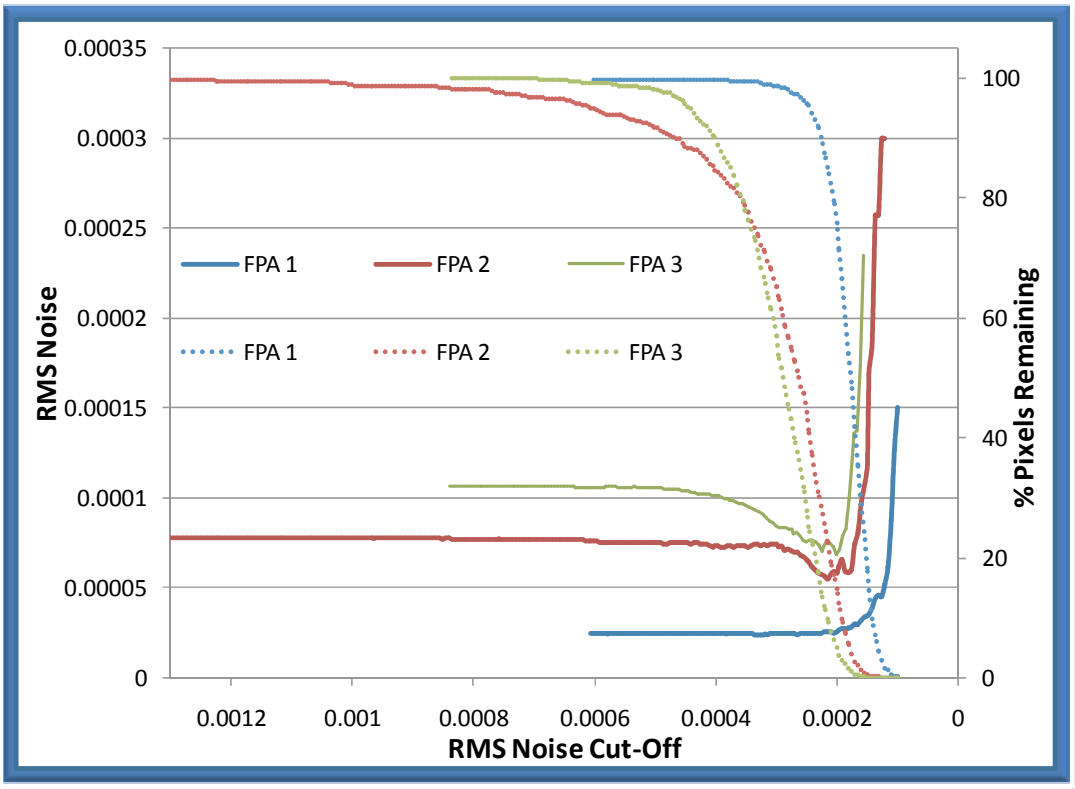


Figure 3.4: RMS noise of the average spectrum as the RMS noise cut-off is reduced, hence eliminating more and more pixels.

The application of a noise cut-off to eliminate pixels is also examined in the context of averaging pixels (Figure 3.5). The results are shown for FPA 2 because this detector had the greatest variability in the distribution of pixel RMS noise. Three aspects can be noted from Figure 3.5: (A) eliminating the noisier pixels did not have much effect on the reductions in noise (root-factor) achieved by averaging; (B) the fluctuations visible on the curves in Figure 3.5 originate from the random averaging of the pixels

remaining after the noise filter has been applied, resulting in slightly different average spectra for each iteration of the process; (C) when few pixels are averaged, reductions in noise are achieved as the RMS noise cut-off is decreased until there are very few pixels left (<2%); however, when a larger number of pixels is averaged, a low noise is achieved at a relatively high RMS noise cut-off.

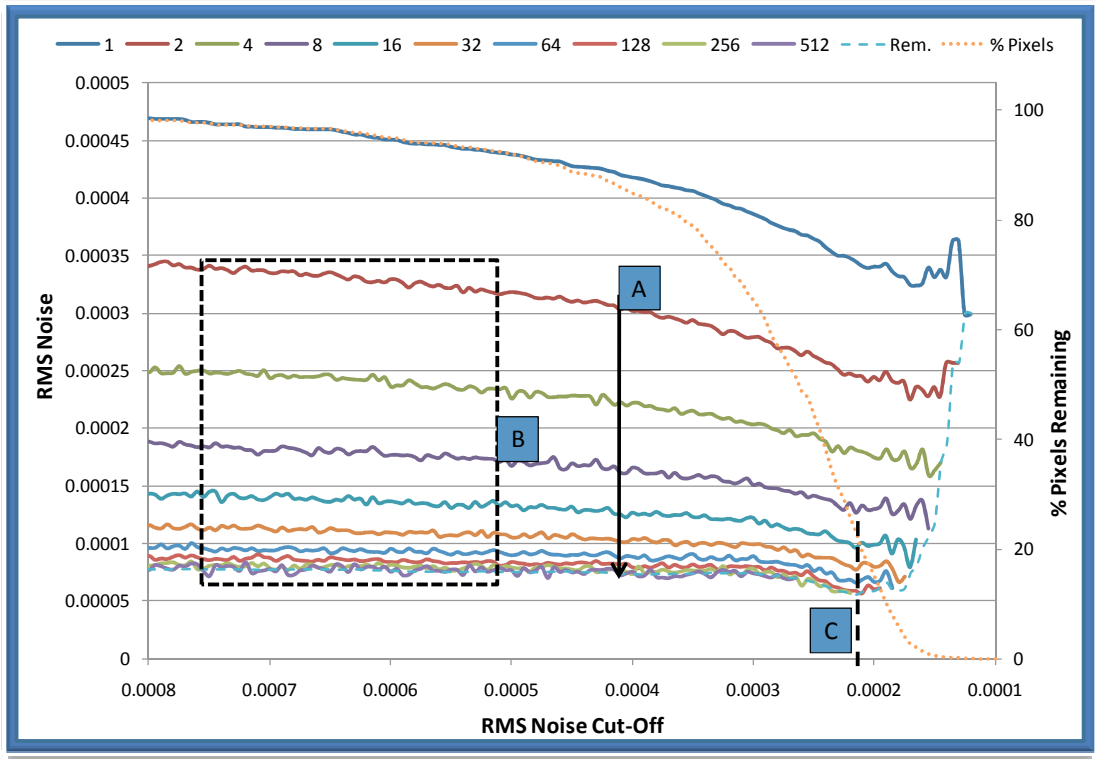


Figure 3.5: RMS noise of spectra consisting of an increasing number of pixels averaged together and for a decreasing RMS noise cut-off. RMS noise was measured in the region 2800-2500 cm^{-1}.

3.3.2 Analysis of quantitative performance

3.3.2.1 Methyl myristate in heavy mineral oil

To assess the quantitative accuracy of the FPA detectors, calibration standards prepared by gravimetric addition of methyl myristate to heavy mineral oil were scanned in a fixed-pathlength transmission cell and calibration equations relating the concentration of methyl myristate to the intensity (in absorbance units) of its ν(C=O) band at 1747 cm^{-1} were obtained by simple linear regression. Initially, the absorbance

values were measured after averaging all the spectra within each recorded image. Subsequently, to further investigate the noise issues discussed in the previous section, the effect of varying the noise cut-off, and hence the pixels contributing to the calibration, was investigated. Figure 3.6 illustrates the changes in the standard deviation of the calibration (CSD) as a function of both the RMS noise cut-off used in filtering each of the infrared images and the number of pixels averaged together.

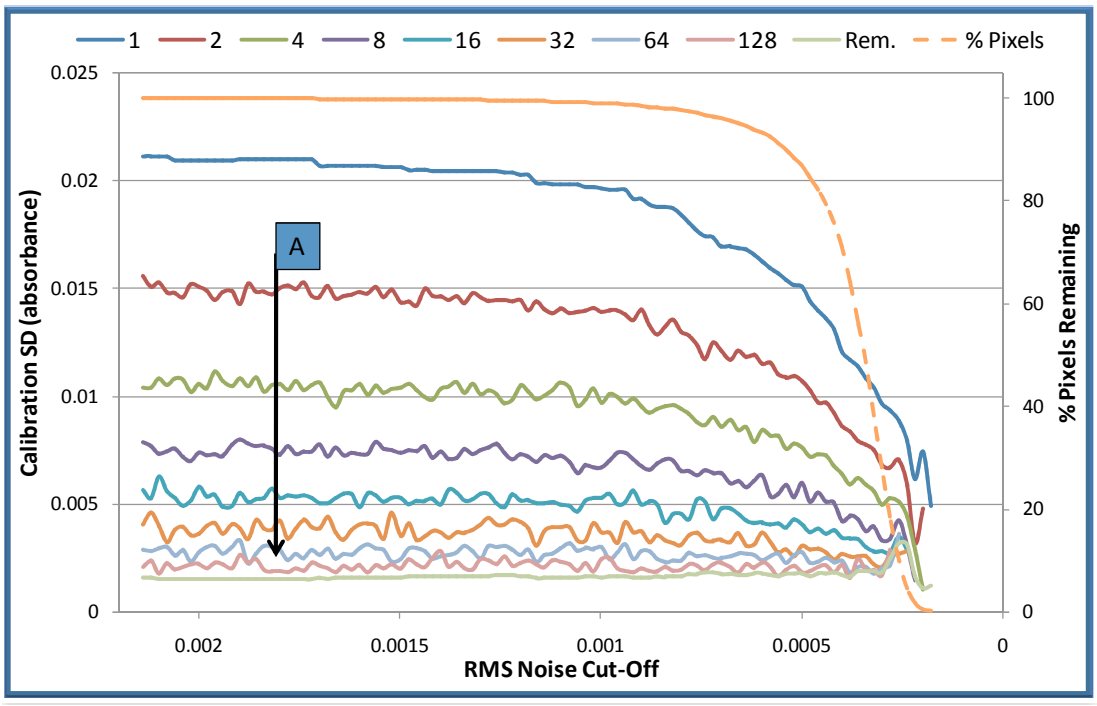


Figure 3.6: Reduction of the calibration SD (CSD) as pixels are eliminated by an RMS noise cut-off (measured in the region 2800-2500 cm^{-1}). The effect is shown for an increasing number of pixels averaged together (B).

After applying the noise filter, pixels were randomly averaged together up to the desired amount (2, 4, 8, etc.). Excess pixels that were insufficient in number to form an average of the desired amount were discarded. It can be seen that when averaging only a few pixels together, reasonable gains can be made by employing an RMS noise cut-off. For example, using a cut-off which leaves 40% of pixels remaining can reduce the CSD by approximately one-half. Contrary to what might be expected, eliminating the noisiest pixels had little effect on the CSD at all levels of averaging. In addition, when a large number of pixels were averaged together, there was almost no change to the CSD.

Analogous to the RMS noise, the fluctuations in the CSD curves are a result of the process of randomly selecting pixels to include in each average spectrum. This illustrates how the differences in pixel-to-pixel noise characteristics in conjunction with a poor combination of randomly selected pixels averaged together can yield noticeably worse results. Despite the non-deterministic nature of the averaging process, a more functional methodology for combining pixels has yet to be determined. The results suggest that if the intention is to average many pixels in order to improve the SNR, then a noise filter offers little benefit. However, if the desire is to maximize the number of spectra while maintaining as low as possible a level of noise, then applying a noise filter can indeed offer benefits in this regard. Lows in the CSD are achieved with 20% to 40% of pixels remaining depending on the number of pixels averaged together. The final drop in the CSD at the lowest noise cut-off levels is a result of entire images being eliminated (because none of the pixels meet the noise cut-off) and so having fewer points in the calibration. An additional and very interesting result of this investigation appears when the improvement in CSD as a function of the number of pixels averaged together is examined (Figure 3.6A). Averaging the root-factor attained by fitting the improvement in CSD at each increment of the RMS noise cut-off (except at the strictest cut-off values where few pixels remain) yielded a value of 0.5 with an average R^2 of 0.99. As such, the square root improvement in calibration SD as a function of the number of pixels averaged together is achieved. An explanation for these results becomes evident when examining the calibrations at the pixel level.

Figure 3.7A plots the baseline corrected absorbance at each of the concentration values for the individual pixels and the average of each image. The most noticeable (and surprising) aspect of this plot is the increasing spread in pixel absorbance values and the scale of the spread. For example, at the highest concentration the measured absorbance levels vary from 0.6 to 1.06 absorbance units, a spread of 0.46 on a mean absorbance of 0.88 absorbance units. As can be seen in Figure 3.7B1-B2, the pixel absorbance values are normally distributed about the mean.

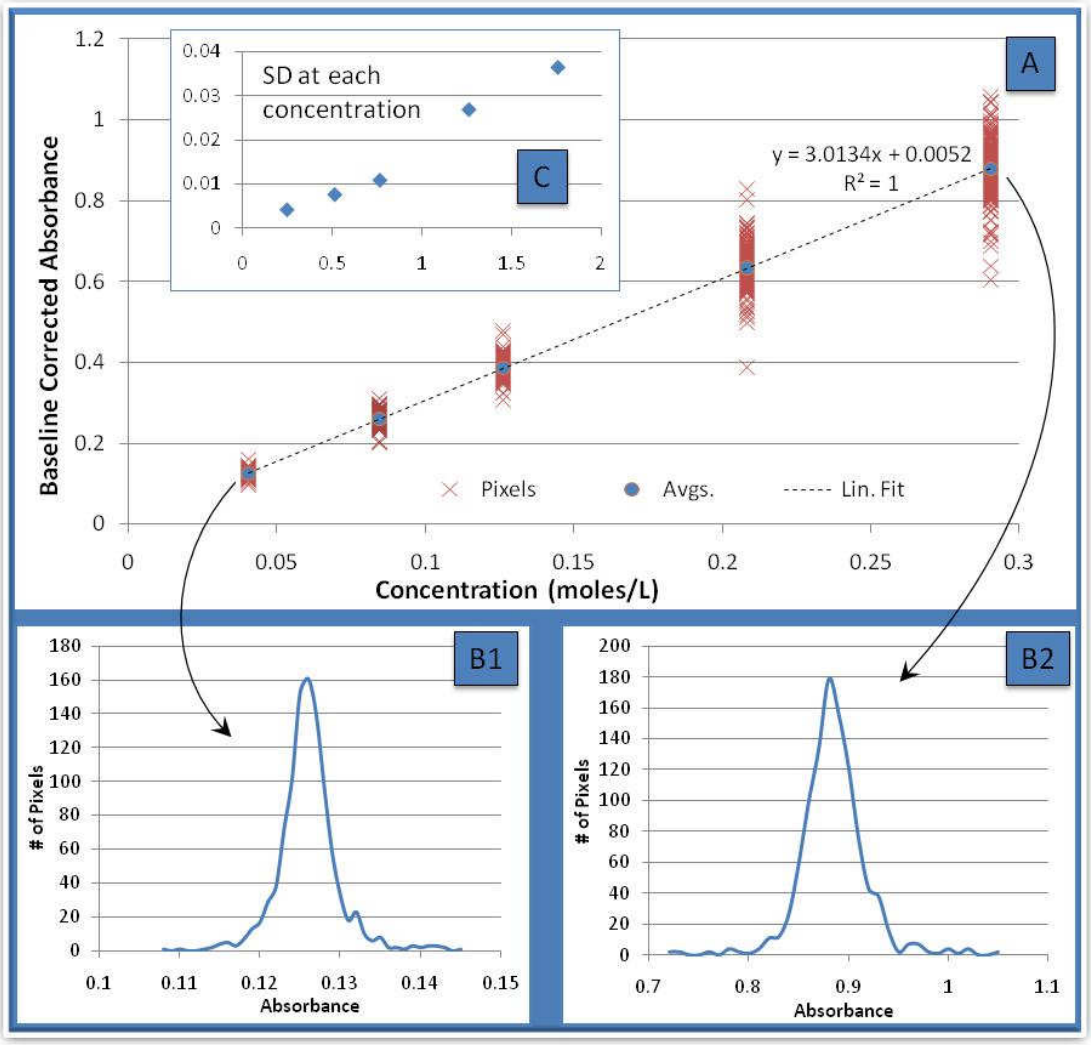


Figure 3.7: Calibration plot of methyl myristate and the measured absorbance (peak height @ 1747 cm^{-1} with a single point baseline taken as the average absorbance in the range 2100-1900 cm^{-1}) (A) and the absorbance distribution of the pixels at the lowest (B1) and highest concentrations (B2). In addition, the standard deviation (C) of the absorbance distributions is provided in the upper inset.

At the highest concentration value the standard deviation is 0.036, which implies that approximately 95% of pixels fall within the range of 0.81 - 0.96. The variation across each of the five concentrations ranged from 11% to 17% of the measured value at two standard deviations. This suggests that the variations in absorbance witnessed here are not a result of the same property as was described previously for the RMS noise. This also explains why a square root improvement in the CSD was noticed when examining the reduction in CSD as a function of the number of pixels averaged together.

A closer examination was made by examining the linear-least-squares models generated by building calibrations for every pixel (Figure 3.15). In this manner the performance can be evaluated on a pixel-to-pixel basis. Figure 3.15 (FPA 2 Run 1) illustrates the performance of each pixel in terms of its CSD and the absolute deviation from the average slope, and the average intercept. An example of the calibration achieved for a number of pixels at different performance levels is provided in Figure 3.17. As was seen with the images of the RMS noise, the spatial pattern of pixels exhibiting poor performance characteristics is clearly present. Although this might suggest that the pixels with poorer calibrations are as such due to their higher noise properties, this is not the case. Examining the calibrations of these pixels it is apparent that the deviation from the linear fit is well in excess of what would be caused by detector noise. In addition, it is not likely the case that it is simply a matter of having different than average response characteristics as this would only imply a different calibration slope or intercept and not necessarily a poor CSD (which is the case). Finally, the orderly layout of the poor performing pixels and the similarities between the results of the RMS noise analysis and the calibration analysis imply that this is not a sample/physical phenomenon.

3.3.2.2 Instrument-to-Instrument

In order to identify whether there is consistency in the pixel output, the identical samples were ran a second time one day later (FPA 2 Run 2). Plotting the calibration built using an average of each image results in almost the identical calibration (Figure 3.9). The spatial distribution of each pixel's calibration parameters (slope, intercept, CSD) is shown in Figure 3.15. Examining the images reveals that the pixels exhibiting spatial pattern noise have different response characteristics; however, plotting the slope, intercept, and CSD from FPA 2 Run 1 and FPA 2 Run 2 shows absolutely no correlation (Figure 3.8), implying that these pixels exhibit higher levels of variability (not just a different response characteristic). This is in-line with the generally higher level of noise seen previously with the spatial pattern noise pixels. Overall, though, the

distributions of pixel slopes, intercepts, and CSDs for the two runs are very similar, as is visible in Figure 3.16.

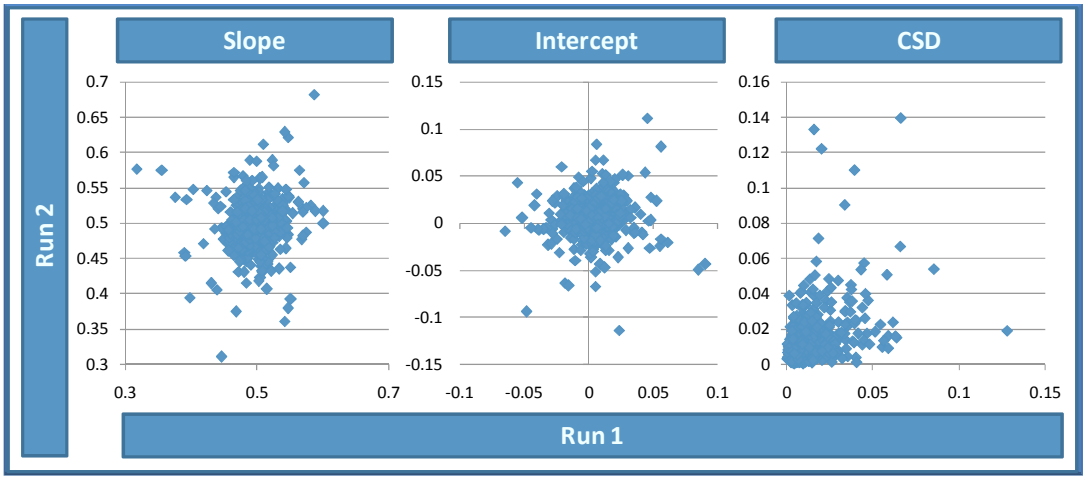


Figure 3.8: Plot of each pixel's slope, intercept and CSD for Run 1 vs. Run 2.

The run-to-run standard deviation, the SD between absorbance values measured from each run at a given pixel, is provided for all pixels in Figure 3.9. The results indicate that 60% of pixels have a run-to-run SD of less than 20 milli-absorbance units, while at the other end, nearly 20% of pixels have a run-to-run SD of 35 milli-absorbance units or much higher. These 20% of pixels constitute the pixels contributing to the spatial noise pattern seen previously.

To verify that the results from FPA 2 are representative, the methyl myristate calibration was repeated on FPA 3. The result of plotting the average for each image in the datasets is shown in Figure 3.10. While the differences between Run 1 and Run 2 on FPA 2 appear to be primarily noise related, the results suggest that the overall response of FPA 3 as a function of concentration is slightly lower than that of FPA 2. This is indicated by the smoothly increasing differential between FPA 2 and FPA 3 as the absorbance increases. Further examination at the pixel level is provided in Figures 3.15 and 3.16. The wider and slightly lower distribution of pixel slopes of FPA 3 (seen in Figure 3.16) explains the lower slope displayed by the averages. Considering that noisier pixels also have poorer calibrations, this was verified by increasing the noise filter cut-off, which resulted in the average slopes for FPA 2 and FPA 3 becoming closer together.

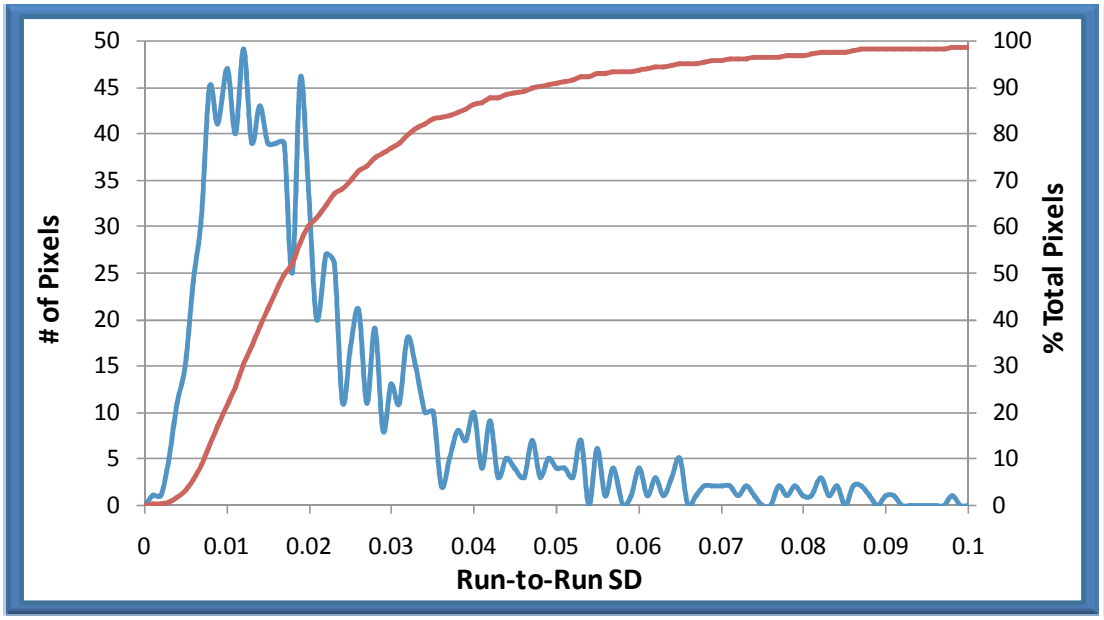


Figure 3.9: The run-to-run standard deviation is the SD between the absorbance values of the two runs. In other words, at every pixel the SD is determined between the different absorbance values of Run 1 and Run 2. The distribution of run-to-run SD values for all pixels is plotted above.

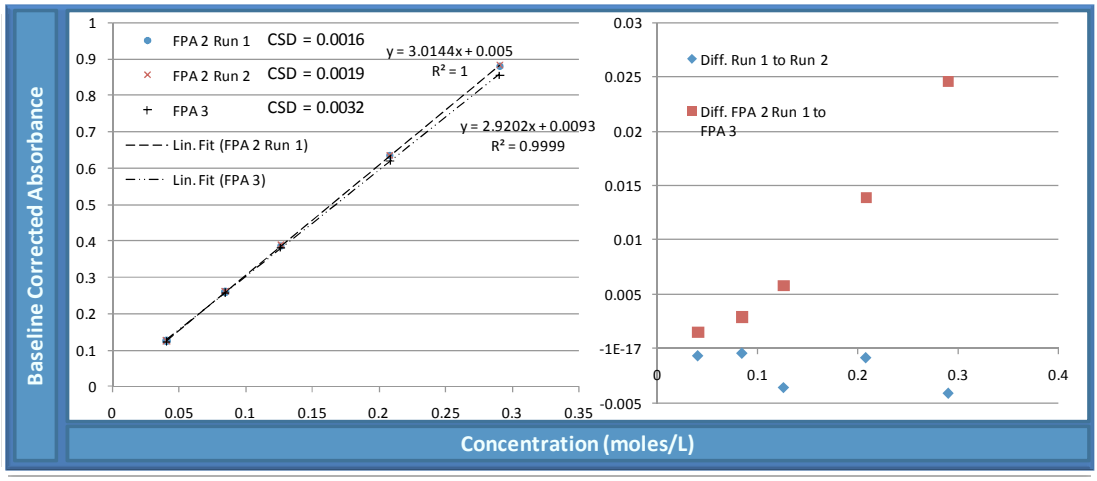


Figure 3.10: Plot of the calibrations of the average spectrum at each concentration for FPA 2 Run 1, FPA 2 Run 2, and FPA 3. The differences between the averages are shown relative to FPA 2 Run 1.

FPA 3 also exhibited a wide range of response levels from one pixel to the next (Figure 3.11), analogous to what is seen with FPA 2. However, the extremes are less severe and the distribution skewed slightly more towards the low absorbance side. Despite this, the standard deviation of absorbance values at each concentration is very similar between FPA 2 and FPA 3. Finally, unlike FPA 2, there is only a minimal presence of the spatial pattern noise pixels as was similarly found when examining the RMS noise values.

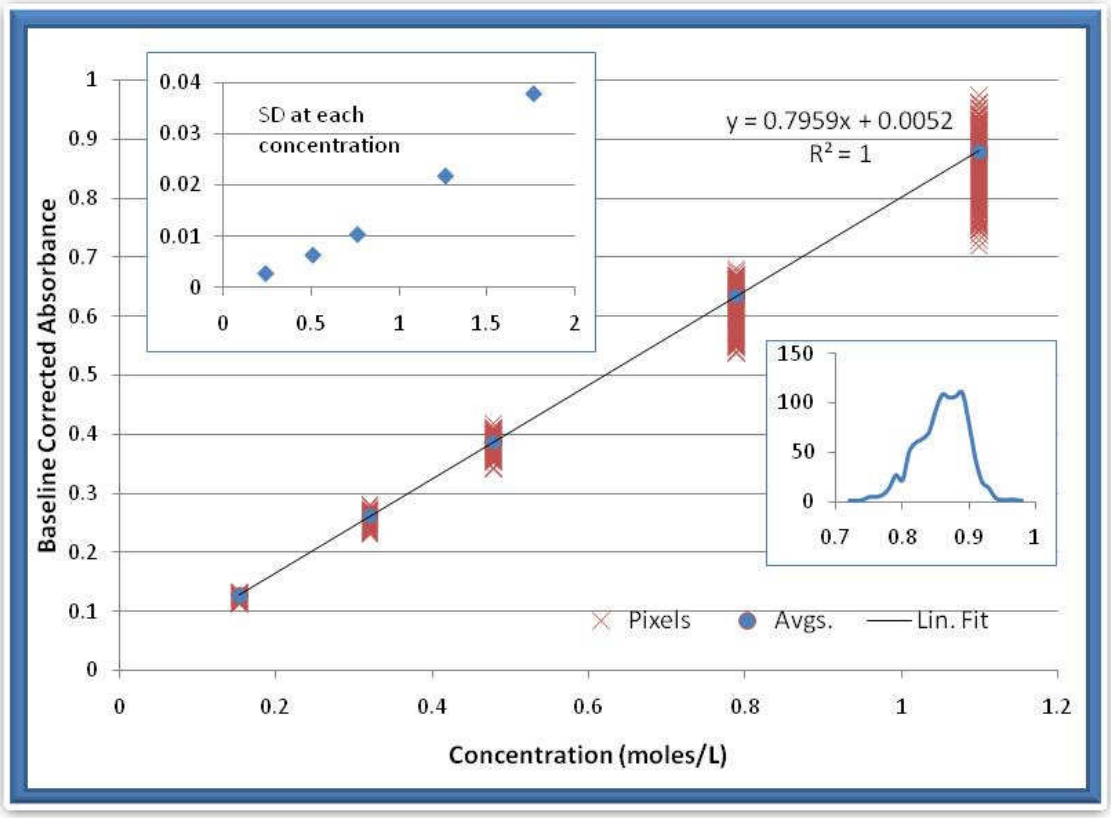


Figure 3.11: Plot of the measured absorbance (peak height @ 1747 cm^{-1} and a single point baseline taken as the average from 2100-1900 cm^{-1}) and the absorbance distribution of the pixels at the highest concentration (lower inset). The standard deviation of the absorbance distributions is shown in the upper inset.

3.3.2.3 Comparison to Single Element

In order to place the overall FPA performance into context, a comparison was made between the calibrations developed on FPA 2 and the Bomem MB-Series

spectrometer equipped with a DTGS detector. These calibrations were developed using the cyanamide dataset which provides an easily measurable peak at 2110 cm^{-1}. Despite the wider bandwidth of the cyanamide peak (as compared to that of methyl myristate) and the higher resolution (4 cm^{-1} vs. 8 cm^{-1}), the pixel-to-pixel spread on the measured peak heights was between 14% and 15% at two standard deviations. This is consistent with the results from the methyl myristate calibration and implies that the variability that was witnessed was not due to a combination of bandwidth and resolution of the measured peak. Despite a longer scan time (80 seconds vs. 53 seconds) for the FPA detector, a comparable calibration CSD comparable to that obtained with the single element detector could not be attained even when all pixels were averaged.

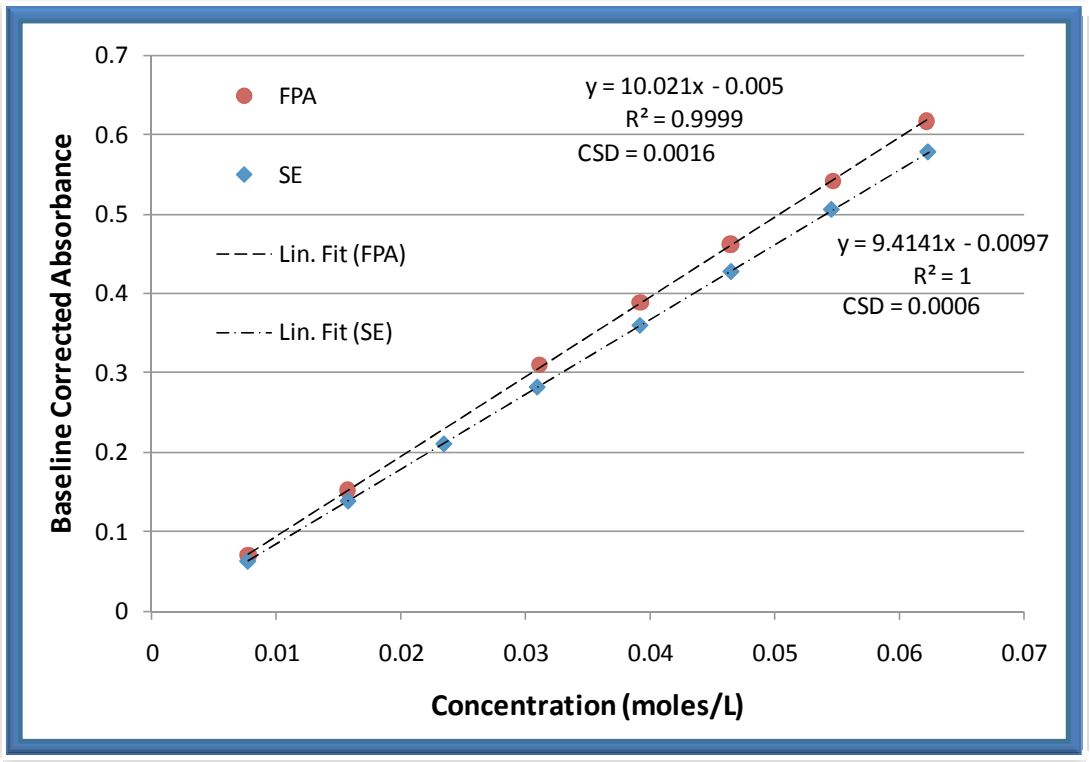


Figure 3.12: Comparison of the calibrations of NaHNCN (peak height @ 2110, two point baseline between 2189 –1989 cm^{-1}) made by FPA 2 and the Bomem MB-Series DTGS detector.

3.3.2.4 Comparing Spectral Profiles

It is evident by the pixel calibration images shown in Figure 3.15 that neighboring pixels can exhibit radically different properties. It has been shown from the calibrations that a given pixel can have a different erratic response from one image to the next. In addition, the scale of the anomalous readings makes it unlikely that the variable absorbances are simply a result of detector noise, but the pixels which exhibit this behavior also exhibit higher RMS noise in an open-beam spectrum. To further investigate whether this phenomenon is wavelength dependent, several images of a polystyrene reference standard were acquired (128 co-added scans at 4 cm^{-1} resolution). The absorbance values of two peaks within the spectrum were then plotted against each other with the expectation that if a pixel has a higher than average (compared to the other pixels) absorbance at one peak in the spectrum, it should at the other. The process was repeated for several orientations of the polystyrene to ensure that the results obtained were not due to a physical property of the polystyrene, and several images were acquired to verify consistency of the results.

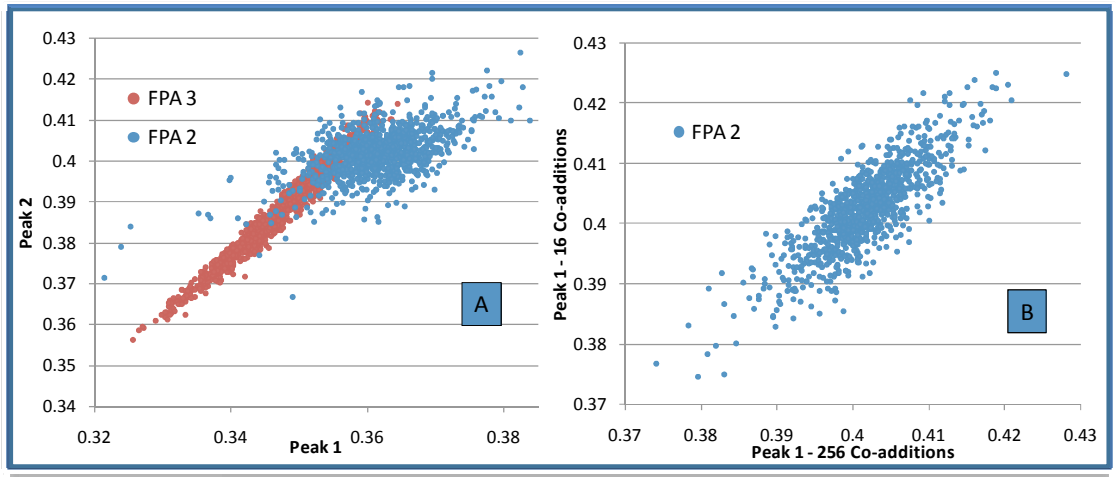


Figure 3.13: A) Comparison of peak heights for each pixel in two images of polystyrene, one for each FPA. Peak 1 is measured at 1069 cm^{-1} with a baseline at 1126 cm^{-1}, Peak 2 at 1373 cm^{-1} with a baseline point taken at 1413 cm^{-1}. B) Peak 1 measured for each pixel of FPA 2 from two images, one consisting of 256 co-additions, the second consisting of 16 co-additions.

From Figure 3.13A it can be seen that there is a correlation between the absorbance values of the two peaks that were measured. While FPA 3 appears to show a better correlation, the absorbance values span a much wider range. In addition, the response of FPA 3 is lower than that of FPA 2, as was noted with the calibration data. The correlation in the case of FPA 2 appears worse due to the larger number of pixels known to perform significantly worse than the average. Figure 3.13B plots the measured absorbances for peak 1 from one image consisting of 16 co-additions and another consisting of 256 co-additions. What is clearly illustrated is that the variation in measured absorbance is not dependent on detector noise, or rather, noise which is improved by co-adding multiple passes of the interferometer. Despite higher noise values being visible in low absorbance areas of the spectrum in the 16 co-additions image, there is no indication of this in Figure 3.13B.

When considering the use of FPA-FTIR spectroscopy for bacteria identification the consistency in pixel-to-pixel peak heights is not of concern (as the data is normalized) [63]. However, if applications such as these are to succeed it is crucial that spectral profiles are preserved from one pixel to the next. This was examined with a polystyrene image from FPA 2. As the intention was to compare spectral profiles, the polystyrene spectra were baseline corrected, vector normalized in the region where all peaks are on scale, and the first derivative taken. The first derivative also reduced baseline effects which can artificially lower the variation near the baseline points. The average and variance spectrum of all pixels is shown in Figure 3.14.

Except for several sharp peaks, such as the one at 1150 cm^{-1} and the off-scale bands, there are few noticeable effects on the variance spectrum from peaks at 2000 – 1650 cm^{-1} and those at 1400 – 1250 cm^{-1}. The magnitude of any of the noticeable variations, such as the small variance peak at 1950 cm^{-1}, is not sufficient to explain the variation in measured absorbances witnessed on the calibration and polystyrene data.

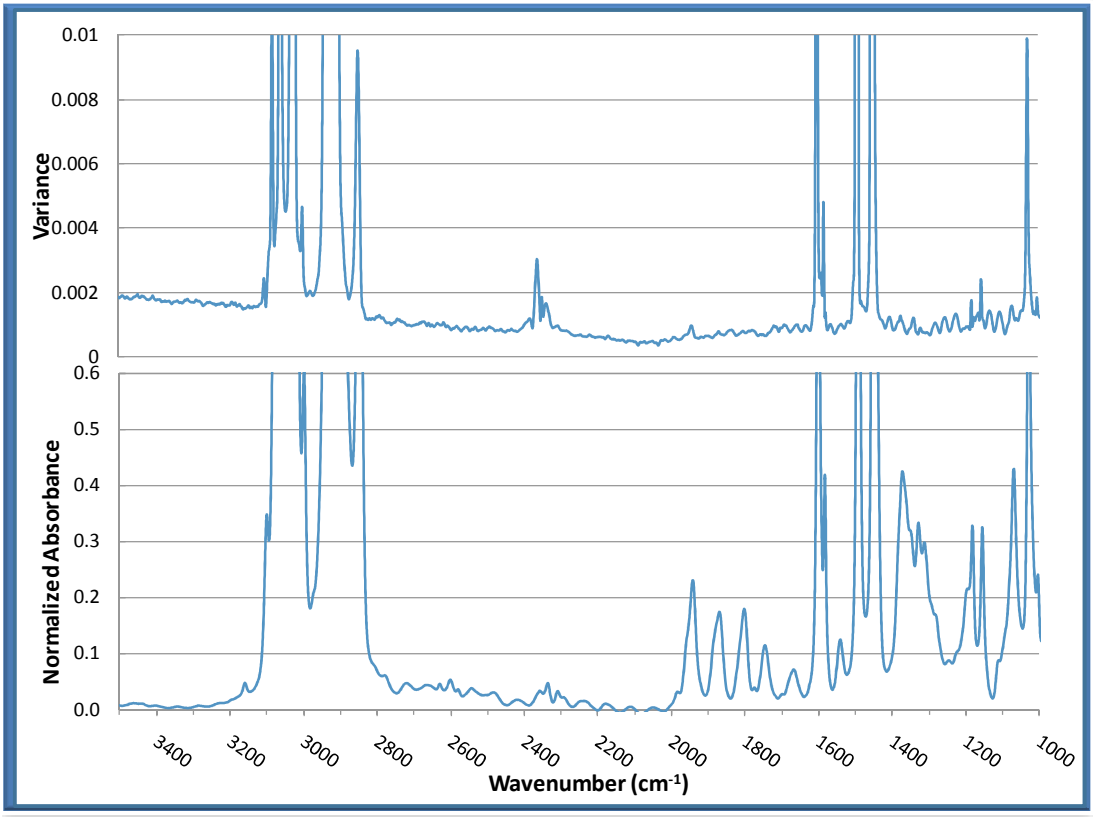


Figure 3.14: Baseline corrected and normalized average absorbance and variance spectrum (based on the first derivative spectra) for all 1024 pixels of an image of polystyrene. The pixels were baseline corrected (average between 2100 – 2032 cm^{-1}) and vector normalized using the region 1045 – 1413 cm^{-1}.

With the aforementioned results in mind, it is clear that present-day FPA-FTIR spectroscopy is not suitable in circumstances that require very high SNR or for high-throughput applications which require short scan times (as long scan times will most likely be necessary to achieve the desired noise levels). However, in circumstances where an individual pixel (or the average of multiple pixels) meets the noise requirements or where other variability inducing factors are present to a greater degree than the instrument noise, FPA-FTIR spectroscopy can offer a tremendous advantage by providing data redundancy, small sampling area, and multi-channel analysis. The results of this study do not negate the possibility of using FPA-FTIR spectroscopy for quantitative analysis; however, care must be taken in the selection of pixels used. It is entirely feasible to build a calibration and identify those pixels which meet the

necessary performance criteria, either directly based on the calibration results or against a validation dataset. These pixels could then be used for predictions in further analysis. If the signal-to-noise ratio requirements for prediction are low enough (not greater than 100) then by filtering incoming FPA imaging data a process could be developed which would allow for analysis of samples with minor contaminants. By taking advantage of the small sampling area of FPA-FTIR spectrometers quantifications could be made where it would be otherwise impossible using a single element detector (without a microscope). However, it should be considered that while FPA-FTIR spectroscopy has a high (5.6μm) spatial resolution, it is in fact limited by the quality of signal necessary in the target application. In other words, if it is necessary to average multiple pixels to achieve the desired level of spectral signal quality, then the resolution that is being sampled is effectively reduced. In the majority of applications where infrared imaging is used, the information of interest is generally that of examining gradients across the image which are indicative of substantial changes in the infrared absorbance. High pixel-to-pixel fidelity is not a primary concern as changes and trends in the infrared absorbance (if of sufficient magnitude) will be preserved. However, limitations occur if one wishes to infer slight changes in the properties of a sample at the pixel level. Due to the poor noise performance as compared to more traditional detectors and the variation on a pixel-to-pixel basis, the expectations of what can be interpreted from a single pixel are limited. Conversely, it should be considered that the expectations of what can be accomplished using FPA-FTIR spectrometers are exaggerated by the reliable, robust, and mature nature of single element DTGS based FTIR spectrometers.

3.4 CONCLUSION

This study has shown through the analysis of noise characteristics and quantitative calibrations that one cannot assume that FPA-FTIR spectroscopy can offer performance akin to what one has been accustomed to with single element detectors. It

can also be noted that the more recently fabricated FPA detectors (FPA 2 & FPA 3) exhibited an overall worse noise distribution than that of an older FPA detector (FPA 1). Perhaps this was a necessary requirement in reducing fabrications costs or in improving the longevity of the detector (against delamination). This, however, does not explain why a much higher level of spatial pattern noise was visible in FPA 2 as compared to FPA 3. If FPA-FTIR spectroscopy is going to be more readily adopted for quantitative analysis it will be necessary to improve the overall distribution profile of the FPA detector. Despite this, it has been shown that a linear calibration can be built with milli-absorbance accuracy, and performance improved through averaging and the use of noise cut-offs to eliminate noisier pixels. In addition, this is the first study to simultaneously compare multiple FPA-FTIR detectors. While one can envisage tremendous benefits and potential for this technology, the operation and maintenance of this type of instrumentation is currently more cumbersome and much less affordable than that of single element spectrometers which are widely available today. Nonetheless, this is to be expected considering the relatively new nature of this technology, and as with most analytical techniques, costs are expected to decrease and reliability to improve as the technology matures.

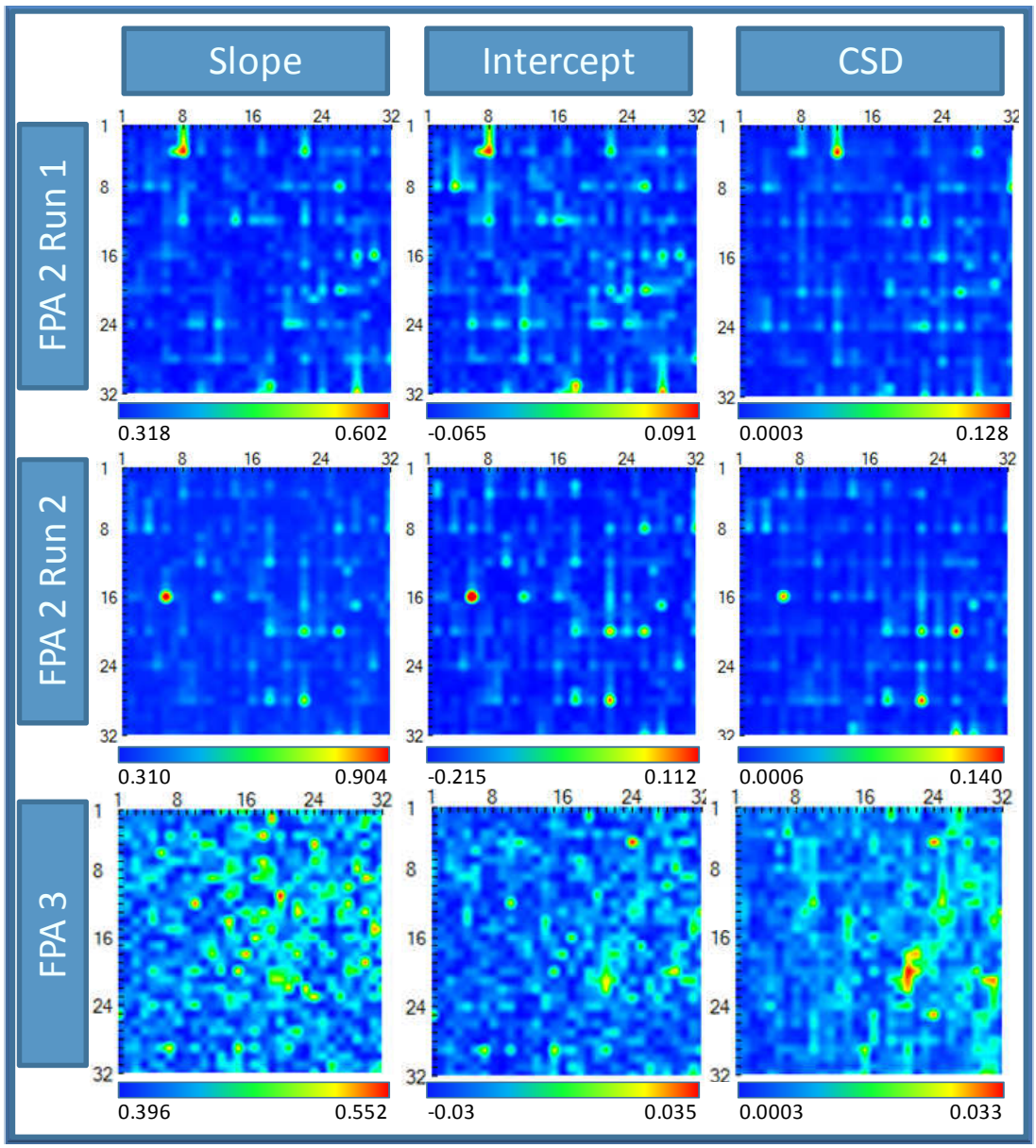
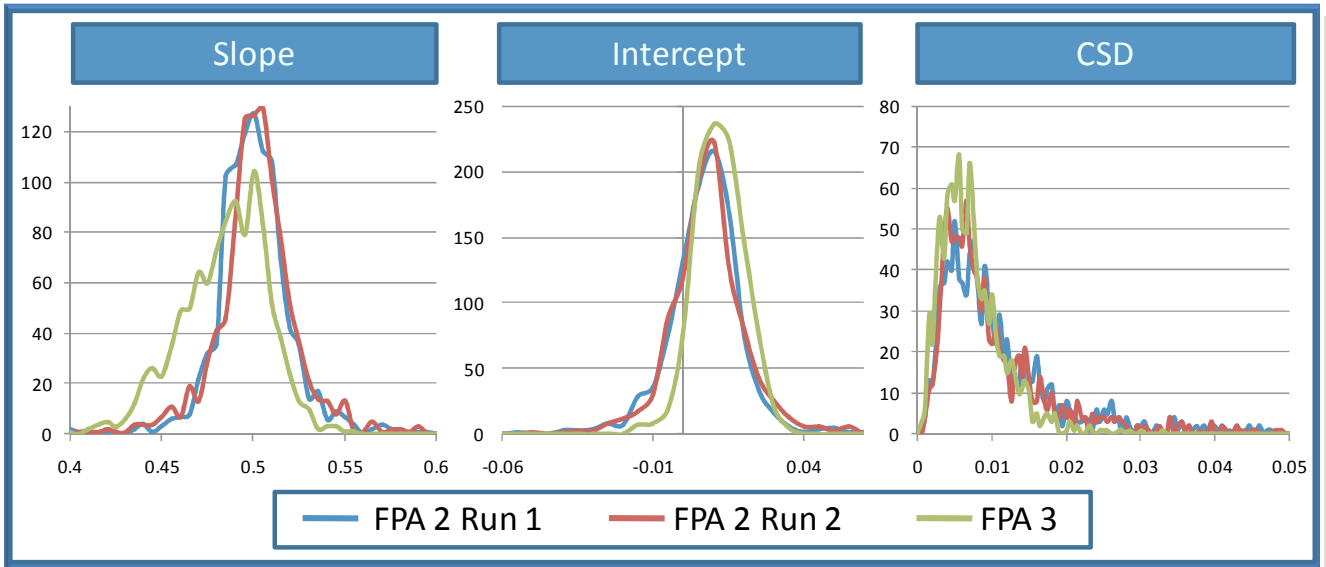


Figure 3.15: Spatial distribution of the calibration parameters slope, intercept, and calibration standard deviation for each pixel. The slope and intercept values are taken as the deviation relative to the average with blue being closer to average and red further.



	Slope Avg.	Intercept Avg.	CSD Avg.
FPA 2 Run 1	0.497 ± 0.023	0.005 ± 0.013	0.011 ± 0.010
FPA 2 Run 2	0.499 ± 0.029	0.005 ± 0.017	0.011 ± 0.011
FPA 3	0.482 ± 0.024	0.009 ± 0.008	0.007 ± 0.004

Figure 3.16: The pixel distributions of the slope, intercept and calibration standard deviation for FPA 2 run 1 & 2 and FPA 3. The average and standard deviation of the above distributions are provided in the table.

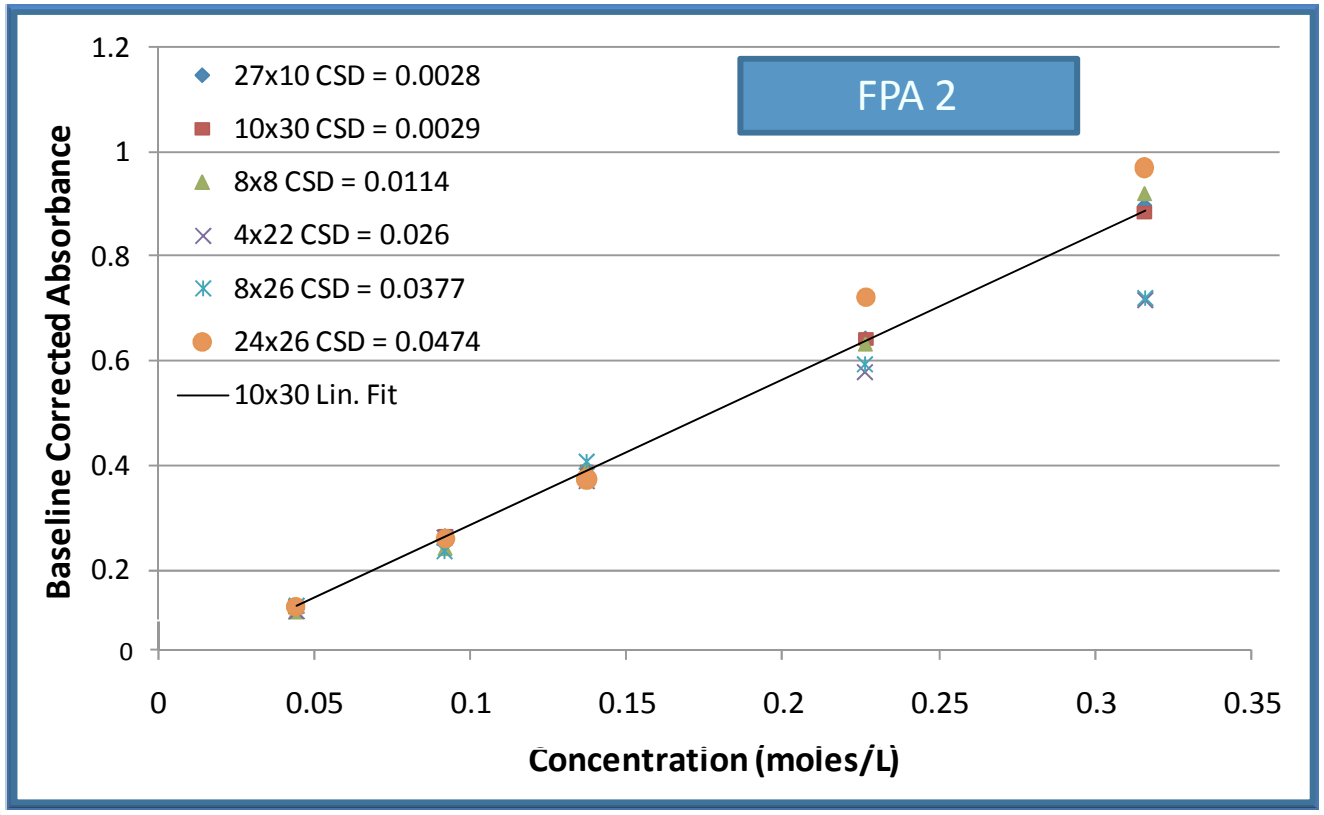


Figure 3.17: An example of the methyl myristate calibrations of several pixels of FPA 2 at a variety of performance levels.

CONNECTING STATEMENT

The first stage towards the analysis of infrared imaging data is the extraction of only those spectra that are representative of the sample. The previous chapter clearly demonstrated that variability from many sources exists in the infrared image. Such examples were taken from homogeneous samples with uniform thickness. In the case of bacteria identification, the sample, manually deposited on an infrared transparent slide, will have a non-uniform thickness which may further exacerbate the resultant variability. In addition, in any real-world application there is the risk of contamination or sample inhomogeneities that can occur. For this reason, the following study introduces an approach to extract only the most representative sample data by employing a pixel filtration process whereby pixels that do not conform to certain spectral characteristics are eliminated. This helps to ensure that only suitable data is used in the generation of spectral databases and during the classification of unknowns.

Chapter 4: Pixel Filtration for the Extraction of Representative Data from Infrared Images

4.1 INTRODUCTION

Prior research [23, 63, 71] has shown that significant advantages can be obtained by employing focal-plane-array Fourier transform infrared (FPA-FTIR) spectroscopy for the identification of bacteria. This method relies on a series of spectral transformations and data filtration techniques to extract reproducible spectra with predefined ideal spectral characteristics (based on criteria such as absorbance, noise, etc…). The following work describes the function and effects of each of the transformation and filtration methods in detail in relation to the discrimination of bacteria based on their infrared spectral profiles. The current sampling method used in conjunction with FPA-FTIR spectroscopy involves the manual deposition of bacterial films on an infrared transparent slide. Isolated bacteria colonies are lifted from an agar growth media and smeared directly onto a zinc selenide (ZnSe) slide using a sterile probe such as a wooden toothpick or a microbiological loop. This process results in a dry sample deposit which varies in thickness across its area as shown in Figure 3.1.

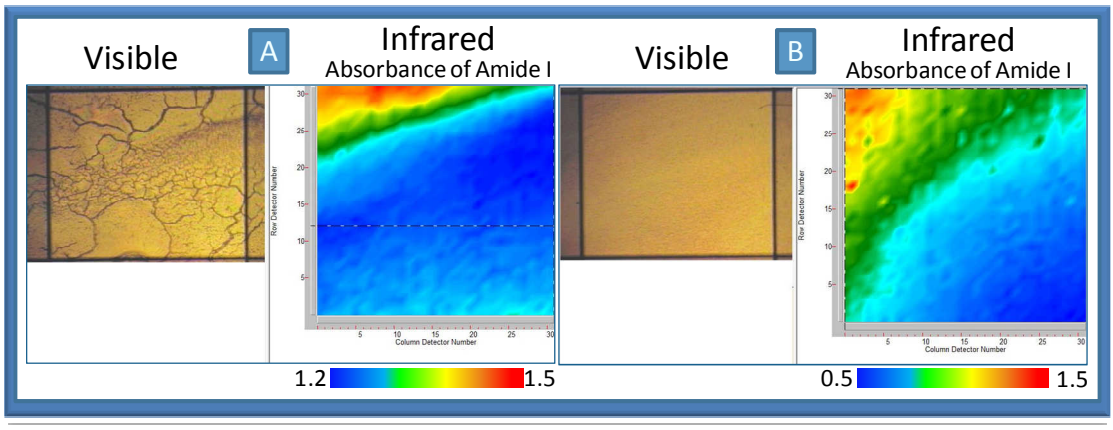


Figure 4.1: Two examples of a bacteria deposit on ZnSe in the visible and infrared. Although the visible image (4.1B) on the right appears more uniform, it in fact has a greater degree of variability in the infrared than the image on the left (4.1A).

Consequently, this deposition method can result in bacterial films with a substantial non-uniform pathlength/thickness across their area which can affect the measured spectral absorbances. It is not uncommon to find a single FPA-FTIR spectroscopic image which contains a large range of spectral absorbance values along with almost zero absorbance, on scale absorbances, and off-scale absorbances. With a little experience the operator can identify suitable portions of the sample in an attempt to minimize variability in the infrared using the visible light image from the microscope (directly or via a CCD camera). Nonetheless, nearly every spectral image of manually deposited bacteria will exhibit a level of variability which necessitates the elimination of pixels that do not meet a series of predefined spectral characteristics. The sample variability as a result of the deposition method can manifest itself in the infrared in a number forms. For example, sharp fluctuations in pathlength can skew spectral profiles while gaps between the sample and the ZnSe substrate upon which the bacteria are deposited can result in spectral fringes. In addition, other elements may contribute to infrared spectral variability including: sample contamination, changes in the operating environment, and instrument operating parameters. Contaminants such as dust or other particulates may cause light to scatter (skewing the spectral profile) or may introduce spectral artifacts in portions of the infrared image. Other variations that can affect an image as a whole include changes in the background spectrum (used to ratio against the sample spectrum), variability of the ZnSe crystal thickness, and warming of the mercury cadmium telluride (MCT) detector. The MCT FPA detector operates at a temperature of 70 degrees Kelvin and requires liquid nitrogen cooling. A consequence of the detector warming, which may not be readily indicated by the spectrometer's software, is increased levels of noise in the acquired spectra.

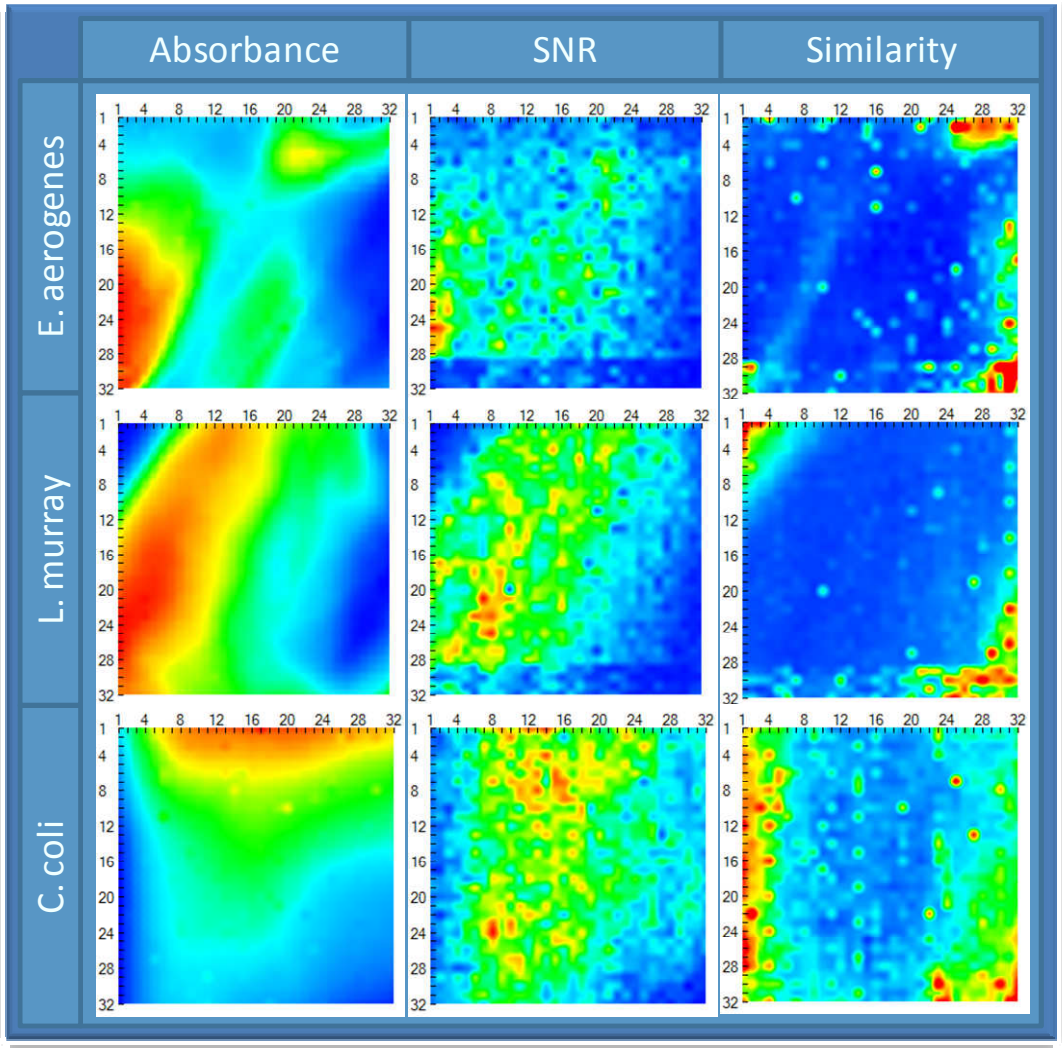


Figure 4.2: Three images of bacteria displayed in terms of absorbance, signal-to-noise ratio, and similarity. Absorbance was measured as the average absorbance on the region 1480-980 cm^{-1}. The signal-to-noise ratio was measured on 1st derivative spectra as the ratio of the peak-to-peak height in the region of 1480-980 cm^{-1} over the peak-to-peak height in the region of 2200-2000 cm^{-1}. Similarity is measured as the Euclidean distance to the mean spectrum of an image. In determining the mean spectrum, images were filtered to remove spectra with an average absorbance outside of 0.05 and 0.9 on the spectral region of 1480-980 cm^{-1}. This ensured that the average was not biased by pixels which consisted primarily of noise or off-scale absorbance. Euclidean distances were calculated on 1st derivative vector normalized spectra in the region of 1480-980 cm^{-1}. Scales for the values represented in this figure can be found by examining the per pixel plots found in Figure 4.3.

Several examples are shown in Figure 4.2 of how the infrared image of a deposit of bacteria on ZnSe varies in terms of its absorbance, signal-to-noise ratio (SNR), and within image pixel-to-pixel similarity. Similarity is measured as the Euclidean distance to the mean spectrum of an image. Even though commonalities are observed at the extremes of each of the measures, it is clearly seen that there are substantial differences in how the variability for each of the measures is spatially distributed. In addition, certain characteristics, such as the blue band visible on the bottom of the SNR and similarity images of *Enterobacter aerogenes* and *Listeria murray*, are not at all evident in the images of the absorbance measures. Further examination of the correlation between each of the measures is provided by plotting the measures against each other for each pixel (Figure 4.3). Figure 4.3 illustrates the range of values expressed in Figure 4.2 and underscores that despite obvious correlations (such as increased absorbance results in increased SNR) there is variability present in the spectral images which is on a comparable scale to the correlations. Contributors to spectral variability may include elements such as inconsistent pixel performance, sample irregularities, spectral scattering, and fringing, among others. As some of these factors vary from image to image, the extent to which the different measures correlate may also vary from one image to the next (as can be noted by Figure 4.3).

The aforementioned elements that contribute to spectral variability result in spectral images where a number of the pixels in each image are not suitable for use in further analysis. This motivated the development of pixel filtration routines, the goal of which is to minimize the contribution of variability that is not representative of a sample while achieving the best SNR possible all the while maintaining as many spectra as possible for further analysis by classification routines. The approach that has been developed in the following work consists of three spectral filters applied in the following order: an absorbance filter, a SNR filter, and a similarity filter. The absorbance and SNR filters eliminate spectra whose spectral quality do not meet per-defined criteria, while the similarity filter removes outliers, defines overall image homogeneity, and attempts to extract those pixels from an image which are most representative of a sample.

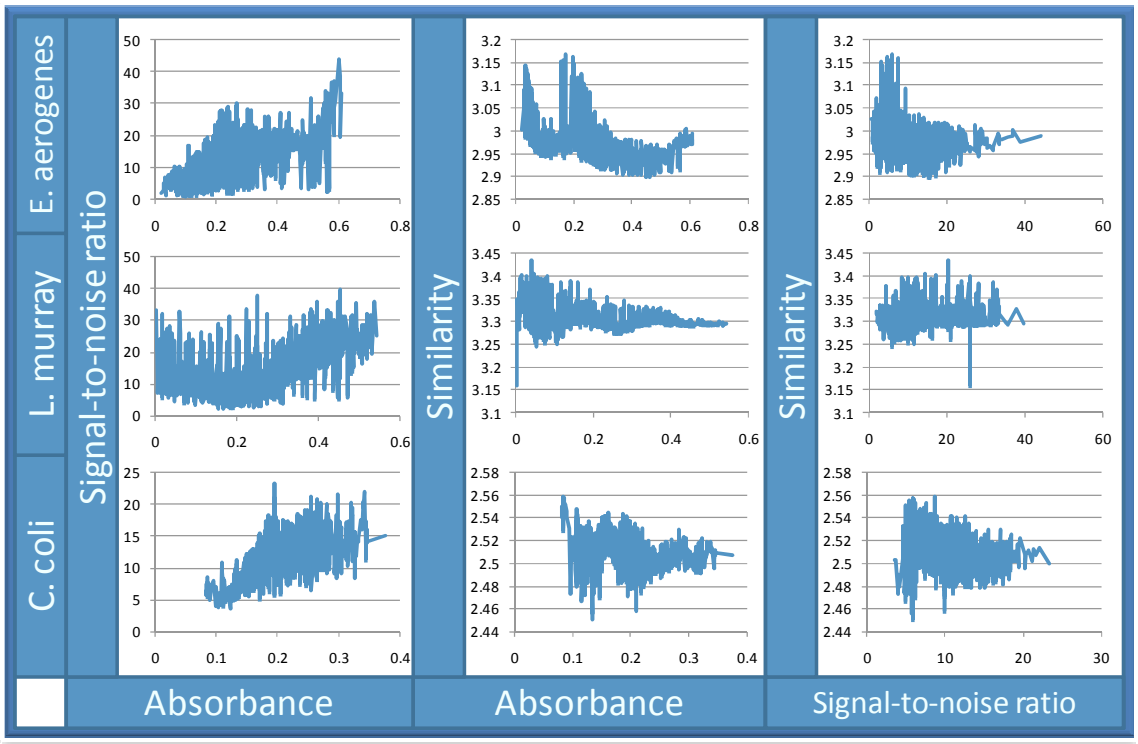


Figure 4.3: Examples of the correlation between the average absorbance, SNR, and similarity on a pixel basis. These plots represent the data from the images illustrated in Figure 4.2.

A consequence of having multiple factors which affect the spectral variability and a lack of a clear correlation between the measures of variability (absorbance, SNR, similarity) make it difficult to assess what are suitable limits and cut-off values to be applied to each of the three spectral filters. The implication is that one cannot define the cut-off values for each of the filters simply by observation (comparing correlations and expected values). For this reason a systematic approach has been taken to evaluate a wide range of potential filter cut-off values in order to determine the parameters which provide the most desirable result for a given set of data.

4.2 MATERIALS AND METHODS

4.2.1 Bacteria

Two reference datasets were used in this study. The first (referred to as GPN) consists of 191 images, 74 of which were Gram positive and 117 Gram negative bacteria. The second dataset (referred to as CCJ) consists of 294 images of *Campylobacter*, 77 of which were *Campylobacter coli* and 217 *Campylobacter jejuni*. These datasets were chosen as they provide examples at two different levels of taxonomic specification. The GPN and CCJ bacteria were prepared using standard methods and protocols as part of a previous study [23]. Specific details on the growth and preparation parameters of the bacteria are outside the scope of this report. Following incubation the bacteria were manually transferred to a zinc selenide infrared transparent slide using a sterile wooden probe.

4.2.2 Data acquisition

The spectral images were collected on a Varian Excalibur FTIR spectrometer operating under Varian Resolutions Pro 4.0 software (Varian, Melbourne, AU) and equipped with a UMA-600 infrared microscope and a 32 X 32 (1024 pixels) mercury cadmium telluride (MCT) focal-plane-array detector. A constant flow of dry air was used to purge the spectrometer and the microscope, limiting spectral contributions from carbon dioxide and atmospheric water vapor. Each spectral image was collected from a field of view of 188 X 188 μm^2 with a nominal spatial resolution of 5.6 µm per-pixel and a spectral resolution of 8 cm^{-1}. Each image consisted of 256 co-added scans which required approximately 5 minutes to acquire.

4.2.3 Data analysis

After spectral acquisition, the infrared image data was analyzed and the results generated using software developed in-house on the Visual Studio .Net 2005/2008 platform (Microsoft, Redmond, CA). This included all processing as part of the pixel filtration routines, hierarchical cluster analysis, and principal component analysis using

the non-linear iterative partial least squares algorithm. The processing routines were validated (where possible) by visual and computational comparison of outputs from Matlab 6.5 (Mathworks, Natick, MA) and Omnic 6/7 (Thermo Nicolet, Madison, WI). The processing was completed on a quad core Intel (Santa Clara, California) processor operating at 3.2 gigahertz with 4 gigabytes of random-access-memory running Windows Vista Business edition. The software developed as part of this study was optimized for parallel processing on multi-core central processing units (CPUs) allowing for significantly greater computational output than most other software available at the time.

4.2.4 Pixel filtration

The six stage process which has been developed to filter images is described in Figure 4.4. The process begins with an absorbance filter to identify the portions of the sample (within the image) which have a suitable pathlength through the selection of pixels which have an appropriate absorbance range. The remaining pixels are derivatized followed by a filter to eliminate pixels which exhibit a low SNR. The pixels remaining after the SNR filter are vector normalized and then a two-phase similarity filter is applied which eliminates outlier pixels, assesses image quality, and reduces overall spectral variability.

4.2.4.1 Absorbance filter

Large differences in sample absorbance from one pixel to the next can occur within an infrared image of bacteria as a result of the variations of the sample thickness across the field-of-view (FOV) of the detector. The absorbance filter will eliminate pixels whose overall absorbance level is too great or too little. Our initial work focused on the Amide I band (1700 – 1600 cm^{-1}) as an indicator of sample thickness due to its prominence in all bacteria under investigation. The absorbance filter was applied [63] by accepting only those spectra who's Amide I band fell in the range of 0.8-1.0 absorbance units. A tight range was seen as desirable to avoid differences that may occur as a result of detector non-linearity across a wider absorbance range.

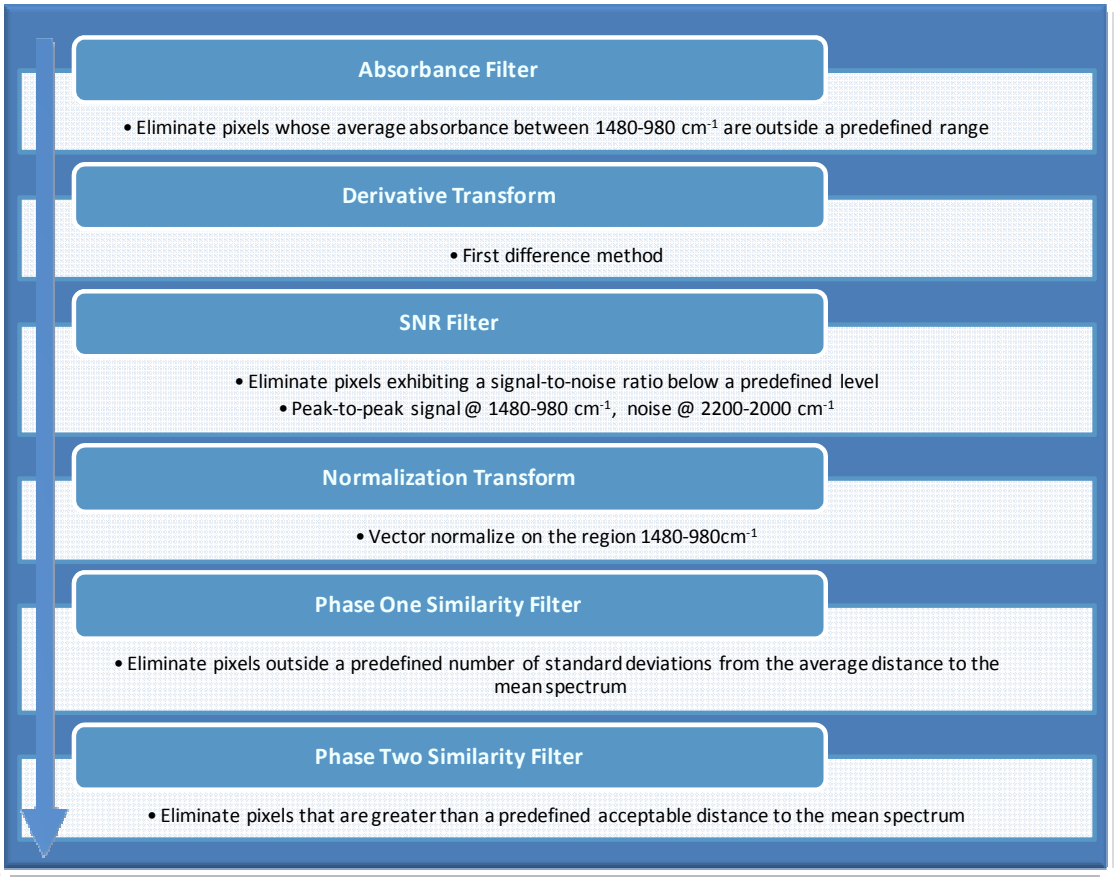


Figure 4.4: Overview of the pixel filtration process applied to extract spectra from infrared images of manually deposited bacteria that meet the pre-defined quality characteristics.

While this approach was successful, further work has provided insight into the effect of using various absorbance and wavelength ranges. It is important to note that the Amide I band is generally three times greater in magnitude than the remainder (Amide II aside) of the fingerprint region as shown Figure 4.5. For this reason, it is impossible to have the Amide I at an absorbance value that is within the linearity range of the detector (0.2 to 0.7 absorbance units) while having absorbance levels for the remainder of the fingerprint region (1480-980 cm^{-1}) that provide an adequate signal-to-noise ratio. A similar situation applies for the Amide II band but to a lesser extent. Considering that the 1480-980 cm^{-1} region has been shown in our research and others [56, 72, 73] to be highly effective for bacteria identification (more so than the Amide I/II alone), the measure of overall absorbance is now taken as the average absorbance from

1480-980 cm^{-1}. In order to maximize the signal-to-noise ratio and ensure that the majority of the 1480-980 cm^{-1} is on scale the average absorbance should be approximately in the range of 0.3-0.6. However, as part of this study a variety of ranges are tested to identify the optimal cut-off values and the consequences of using both wide and narrow ranges.

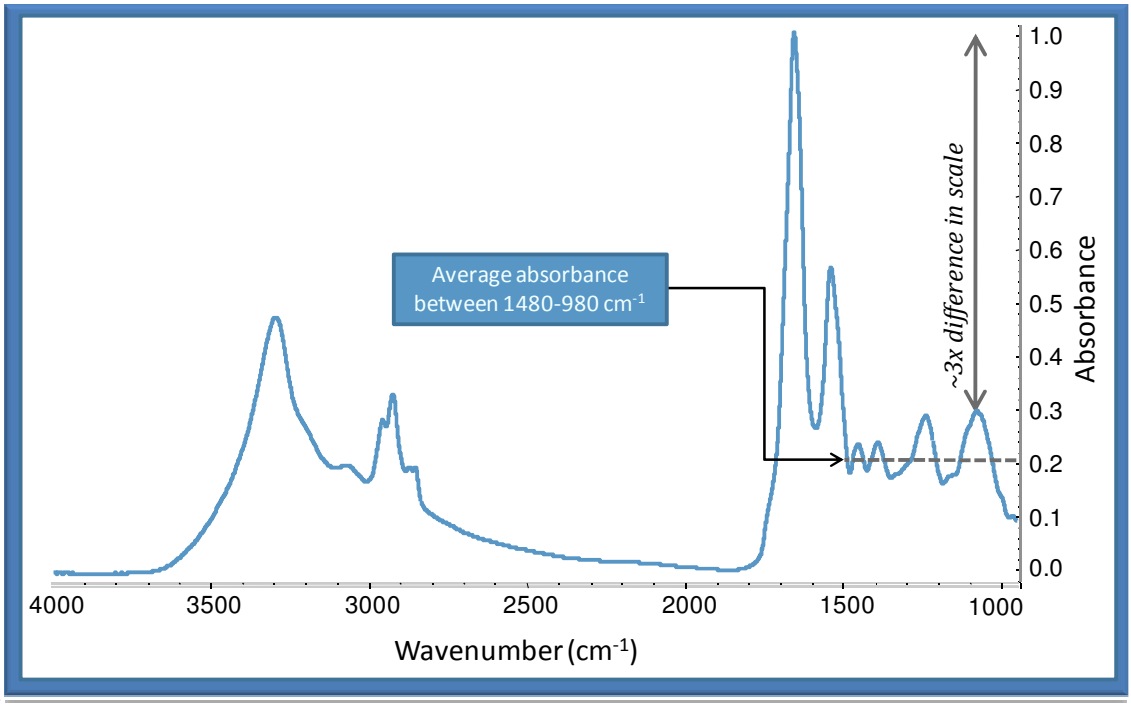


Figure 4.5: The absorbance measure used for filtration is based on the average absorbance in the region between 1480 – 980 cm^{-1}.

4.2.4.2 Signal-to-noise ratio filter

The second most defining characteristic of spectral quality (in the context of bacteria identification) is the SNR. It is important to quantify this characteristic in order to determine what the limitations of the acquired spectra are as discriminators for varying taxonomic levels of identification. If the degree of noise in a spectrum is comparable to the differences between the various bacteria under analysis then this will have a severe impact on the ability to develop a reliable and robust model. As such, the SNR filter eliminates pixels whose SNR is not adequate (regardless of the reason). Although, many pixels which exhibit a low SNR are those with low absorbance levels,

spectral images may often have pixels which exhibit a low SNR despite having adequate absorbance. In defining a SNR cut-off value it is important to realize that every pixel eliminated is a pixel which cannot be averaged with other pixels to provide an averaged spectrum with an even lower SNR. The SNR is measured in this study by taking the ratio of the peak-to-peak height in the region of 1480-980 cm^{-1} over the peak-to-peak height from 2200-1900 cm^{-1} on 1^{st} derivative spectra.

The SNR filter plays an important role as a result of the inconsistent noise characteristics of the FPA detector on a pixel-to-pixel basis (as described in chapter 2). Variability in an infrared image can result from a multitude of sources; as such, whether or not a pixel which exhibits a significantly greater noise (than the average) is eliminated, is dependent on the magnitude of variation introduced from other sources. In other words, if pixels whose noise levels are higher are present at a portion of the sample where absorbance levels are greater (than average), then the SNR of these pixels may be of a similar magnitude relative to those pixels with less noise that are measuring a portion of the sample with less absorbance.

A second function of the noise filter is to eliminate spectra which display a high degree of fringing (Figure 4.6). Spectra that exhibit fringing typically arise as a result of the bacteria deposit lifting off the ZnSe slide (as a result of the drying process or as the samples ages), creating a space between the sample and substrate where reflection occurs. This appears as a wide sinusoidal pattern across the length of the spectrum. If the sinusoidal pattern is sufficiently broad, then the slope of the fringes will be small enough to have almost no effect on the 1^{st} derivative spectrum (Figure 4.6A-B). Due to the nature of the SNR calculation, pixels which exhibit sufficient fringing to result in arching baselines on the 1^{st} derivative spectrum (Figure 4.6C) will have lower SNR values, and depending on the defined cut-off may result in the elimination of the pixel.

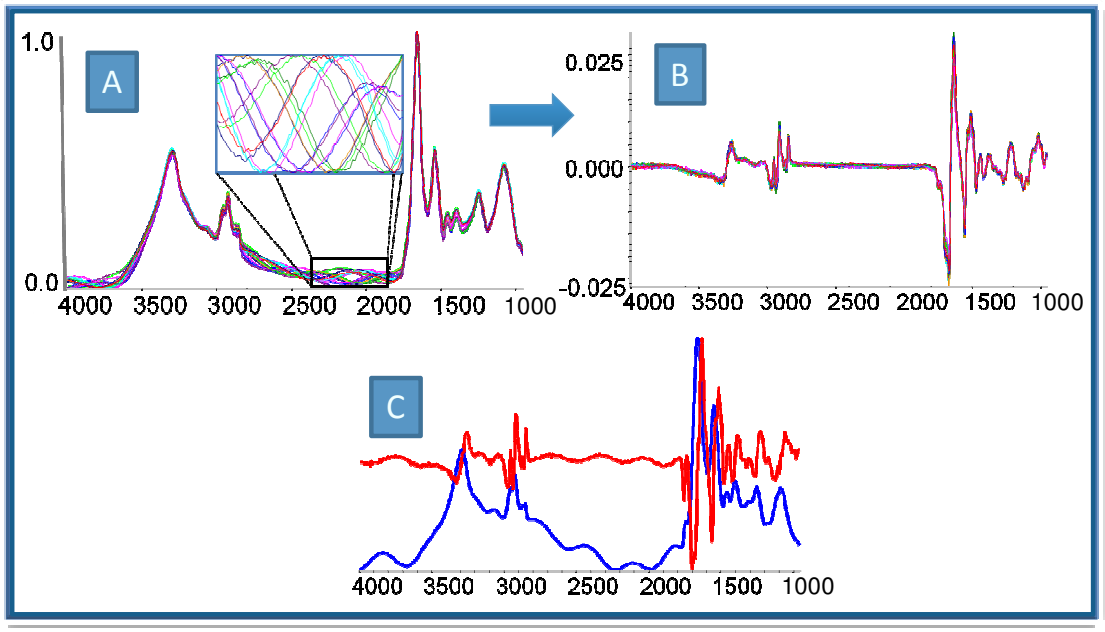


Figure 4.6: Depiction of the fringing effect that can occur when gaps form between the bacteria deposit and the infrared transparent substrate (ZnSe). Fringes (4.6A) that are sufficiently broad will have minimal effect after transformation to first derivative (4.6B). Excessive fringing (4.6C) will still be noticeable after transformation to first derivative

A tertiary benefit to this filter is the elimination of images which were acquired while the liquid nitrogen cooled MCT imaging detector was inadequately cooled. This typically occurs after several hours of instrument use as the liquid nitrogen in the detectors dewar becomes depleted and the MCT array detector begins to warm up. At the time of writing, the software which operates the instrumentation does not readily indicate the temperature status of the detector. If an infrared image is acquired while the detector is not adequately cooled then all pixels will exhibit a significantly reduced SNR. As such, this filter assists in removing such images from further analysis by rejecting all pixels in the image.

The SNR cut-off may be defined on the basis of *a priori* knowledge of the expected intensity differences between the types of samples to be classified. From our own research, individual pixels typically do not have an adequate SNR for use in the development of classifiers for bacteria identification. For this reason it is necessary to average multiple pixels in order to improve the SNR. Based on the number of pixels

remaining after the absorbance filter, a straight forward calculation can be applied to estimate the SNR necessary for a pixel so that the average of the remaining pixels will have a SNR adequate for separation of the bacteria samples. This approach had been used during the initial stages of this research, but proved infeasible when dealing with constantly evolving and expanding databases.

An alternative approach is to empirically determine the SNR cut-off based on the general performance characteristics of all the data in the reference data set. In this manner, one can generate realistic expectations of pixel performance and only eliminate those that lie outside of those expectations. This study examines a wide range of cut-off values in order to establish the optimal balance between eliminating low SNR pixels, while maintaining a sufficient number of pixels which can be averaged together into spectra of even greater SNR.

4.2.4.3 Similarity filter

The protocol for bacteria identification by FPA-FTIR spectroscopy requires that the sample under analysis be from a pure culture of bacteria. As such, the expectation is that each pixel should have a highly similar, if not identical spectrum. However, due to a combination of factors (related to both instrument and sample) this may not be the case. For this reason, it is appropriate to eliminate pixels which are not representative of the sample through the application of a similarity filter. This type of filter serves a dual role of eliminating outlying pixels and reducing overall image variability, ultimately providing more robust data (reduced within-group variability/increased between-group variability) for the development of a bacteria classifier. Prior to applying the similarity filter, all spectra are vector normalized on the region 1480 – 980 cm^{-1} (having been derivatized prior to the SNR filter). Due to the offset between the absorbance of the Amide I/II bands and the remainder of the fingerprint region, the Amide I/II peaks can exhibit higher noise and variability. For this reason the Amide I/II bands are omitted in the comparison of spectra on the basis of similarity.

Two approaches to filtering based on similarity are described and used in the following sections. The first method, employed in previous studies, defined the

representative spectrum as the average of all the pixels (after having applied the absorbance and SNR filters). The degree of similarity is defined as the spectral Euclidean distance to the mean spectrum. The distances to the mean spectrum are stored for each pixel and the average distance and standard deviation computed. The distribution of distances (to the mean spectrum) is approximated by a normal distribution and a cut-off defined in units of standard deviation. Pixels that fall outside the chosen number of standard deviations greater than the mean distance are eliminated. Figure 4.7A provides a graphical representation of this type of similarity filter with an example of a 2D projection of spectra about the mean spectrum. In order to reduce bias to the mean spectrum caused by extremely outlying pixels it is possible to iterate this filter until a desired number of pixels is reached or an acceptable level of variability about the mean spectrum is achieved. A weakness with this approach is that the level of accepted pixels is defined by the variability within an image. For example, if the pixels from 'image A' display much higher variability than those from 'image B' then the pixels that fall within the cut-off (defined by a number of standard deviations from the mean spectrum) of 'image A' will exhibit higher variability than those from 'image B'.

An alternative approach to defining a similarity filter is by examining the level of variability among all the acquired images in a given data set (assuming a sufficient number of images are available) and explicitly defining a permissible level of variability by establishing an acceptable distance to the mean spectrum. The elimination of pixels is accomplished by first computing the mean spectrum of all the pixels in an image (after having applied the absorbance and SNR filters) and then eliminating those pixels whose distance to the mean spectrum is greater than the defined cut-off value as illustrated in Figure 4.7B. This second approach offers several advantages. Pixels that diverge substantially from the mean spectrum, regardless of the cause, are eliminated. It allows for the evaluation of image quality by assessing the variability (average and standard deviation of the distances to the mean spectrum) with respect to the reference data set. It also ensures that the variability of a given image is below a defined threshold, with the potential elimination of entire images which exhibit much greater than average

variability. If it is found that a large number of consecutive images display below normal quality this can potentially be used as a warning or indicator that the instrument (or operation of) require review.

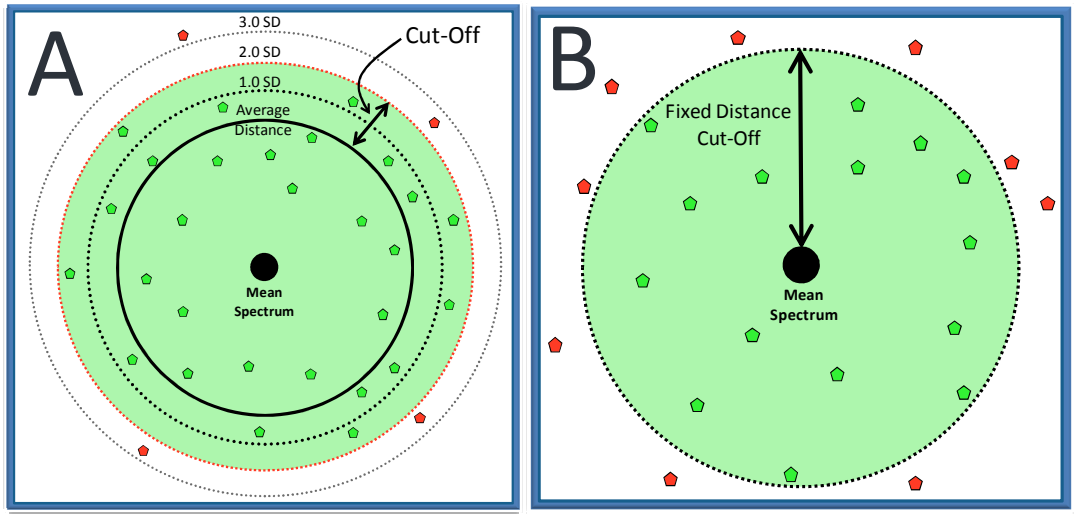


Figure 4.7: 2D projection of spectra about the mean spectrum illustrating the two approaches to filtering based on similarity.

4.7A: Similarity filter with a cut-off based on a number of standard deviations from the average distance to the mean spectrum. Any spectra lying outside the (cut-off) acceptable number of standard deviations (i.e. highly above average distance to the mean spectrum) are eliminated.

4.7B: Similarity filter based on a fixed distance cut-off. Spectra in all images are eliminated if they have a Euclidean distance to the mean spectrum that is greater than a predetermined value.

Spectral distances were computed using the conventional Euclidean distance measure on the region 1480-980 cm^{-1}. This type of metric may not be considered ideal due to the fact that it does not take into consideration the nature of the divergence between two spectra. In other words, as long as the magnitude of the differences is the same, the direction does not matter. For this reason, it is often preferable to use a metric which takes into account the nature of the divergence relative to the group. The Mahalanobis distance is such a metric; however, this metric requires computation of the covariance matrix and its inverse which in the case of multiple thousands of spectra can be a computationally expensive task. Multiple tests were carried out comparing the results of the Mahalanobis and Euclidean metric with the differences between them

being negligible in this particular application and not warranting the severe computational cost involved.

As described above, an individual pixel that is eliminated using a fixed distance cut-off is done so by comparing the spectral distance of the pixel to the average spectrum of that image. If there are significant outliers within the image then the average will be skewed and as a result pixels may be eliminated inappropriately. For this reason the approach that has been developed is to apply a two phase similarity filter. The first step defines a cut-off a certain number of standard deviations from the average based on the within-image variability, the goal of which is to eliminate extreme outliers. The mean spectrum of the image is then recomputed and pixels eliminated based on a fixed distance cut-off.

While the utility and function of the absorbance and noise filters can be expressed in terms of physical and measurable spectral qualities relating to sample thickness and instrument performance, the benefits accrued through the use of the similarity filter are not as clearly defined. On the one hand, the similarity filter is invaluable in eliminating obvious outliers (as a result of a contaminant or other foreign non-sample objects in the FOV), and indeed this aspect represents one of the significant advantages to applying imaging to this application. However, if the similarity filter is simply reducing pixel-to-pixel variability through the elimination of the less-similar pixels without considering the overall consequences, then the effect can be detrimental. A reduction in the number of available pixels results in a reduction in the number of pixels that can be averaged together, thereby reducing the potential for generating lower noise spectra (through increased averaging). In order to assess the optimal cut-off values, a variety of parameters have been examined for both phases of the similarity filter.

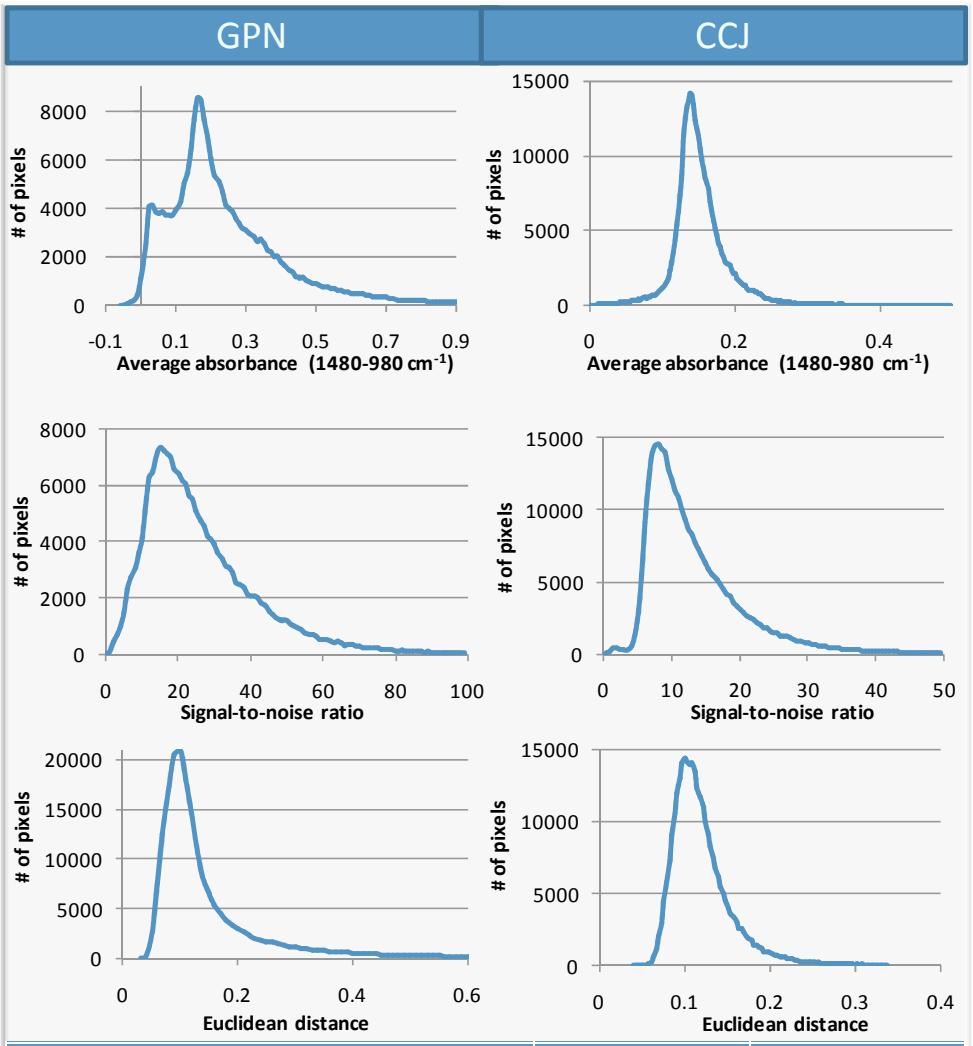
4.2.5 Parameter determination

In order to understand the effects caused by varying each of the parameter values for each of the different filters, a software routine was written to systematically test the filters across a wide range of values. To assess the ranges which should be

examined for each filter, the values for each measure (average absorbance, SNR, and similarity) were tabulated for all pixels in each dataset. Figure 4.8 plots the pixel values for each measure and provides details regarding the range of values tested.

Figure 4.8 highlights several differences between the GPN and CCJ datasets. The GPN dataset exhibits a broader cross-section of absorbance values with a peculiar trend on the low side of the GPN chart. Further examination of the GPN dataset showed that there appeared to be a section of pixels on the majority of images which had spectra lying primarily below the baseline. More details regarding these below-baseline spectra are provided in the results and discussions. Another notable difference between the two datasets is that the GPN dataset has an overall greater SNR. As will be seen in the results and discussions, these differences will have an influence on the cut-off parameters which are selected for each of the datasets.

The first stage of the parameter selection routine is to load the infrared images into the computer's memory. Each image is then filtered and the details stored according to the spectral measures listed in Table 4.1. These include the average Euclidean distance of each pixel from the mean spectrum, the SNR of the mean spectrum, and the average number of pixels remaining after filtration. For a given parameter combination, the average value for each of the previously mentioned measures (across all images) is stored. A wide range of parameter combinations were tested covering the limits described in Figure 4.8. The processing was optimized for parallel computation on multiple computer processor cores resulting in a computation time of 10-30 seconds for each parameter combination. Evaluation of the GPN and CCJ datasets required between 15-25 hours each to complete.



Parameter	Low	High
Average Absorbance Min	0.05	0.5
Average Absorbance Max	0.3	0.9
SNR cut-off	5	80
Phase-one similarity cut-off	3 SD	0.5 SD
Phase-two similarity cut-off	0.5	0.05

Figure 4.8: Statistics for all pixels (GPN = 195,584 pixels, CCJ = 301,056 pixels) in each dataset and the range of parameter values to test. The top chart indicates the average absorbance of pixels. The middle chart shows the SNR of pixels, and the bottom indicates the distance of pixels to the mean spectrum of their respective image. The ranges provided in the table were selected based on the observed trends. (SD = Standard deviation)

Table 4.1: The four measures that were stored for each parameter combination.

	Measure	Description
A_{snr}	The average of the SNR of the mean spectrum of the filtered pixels.	For each image and after all filters and transformations had been applied the mean spectrum was determined and the SNR taken.
N_{nz}	The number of images which had pixels remaining after filtering.	Excessively stringent filtration parameters may result in the elimination of all pixels. This value represent the number of images where this did not occur.
A_p	The average number of pixels remaining.	This is the average number of pixels remaining for all images after all filters and transformation had been applied.
An image does not contribute to the averages in the above measures if there are no pixels which remain after filtration.		
A_{rd}	The average Euclidean distance between the mean spectra of replicate images. A lower value is better.	Each bacteria type in the dataset consisted of a minimum of three replicate images. For a given set of replicates the mean spectrum was determined followed by the average Euclidean distance between each of the replicates. This measure is the average of these averages. A lower value represents replicate images whose mean spectra are more similar to each other. If there is one image or less that remains for a given set of replicates, then those replicates do not contribute to this measure.

A series of metrics detailed in Table 4.2 were used to evaluate the effectiveness of a given set of parameters based on the recorded values from Table 4.1. These metrics provide a numerical representation of the performance of a parameter combination based on a variety factors including: the similarity between replicate measurements, the average SNR, the average number of pixels remaining, and the number of images that did not have all pixels eliminated as a result of the parameters used.

The S_{rd} and S_{snr} are the A_{rd} and A_{snr} values linearly scaled between 0 and 1, where 0 represents the worst value and 1 the best. Multiplying these values (SS) provides the parameter combination which resulted in filtered spectral data that have the highest SNR and lowest average distance between replicates relative to the other parameter combinations tested. This metric does not take into consideration the number of images that remained after filtration nor the number of pixels filtered. The P metric multiplies SS by the ratio of the average number of pixels remaining to the total number of pixels (1024 pixels per image), whereas, the I metric considers the number of images which have pixels remaining after filtration relative to the total number of images in the

dataset. Finally, the T metric considers the combination of the SS metric, the number of pixels remaining, and the number of images remaining.

Table 4.2: The metrics used to evaluate the performance of a parameter combination.

	Description
S_{rd}	$1-(A_{rd}$ scaled between 0 and 1)*
S_{snr}	A_{snr} scaled between 0 and 1 *
SS	$S_{dm} \times S_{snr}$ provides parameter combinations with greatest mixture of a low average distance the mean spectrum and a high SNR of the mean spectrum
P	$SS \times A_p/N_{pi}$, N_{pi} = # of pixels per image
I	$SS \times N_{nz}/N_i$, N_i = # number of images
T	$SS \times N_{nz}/N_i \times A_p/N_{pi}$

*Values scaled by $x'_i = \frac{x_i - X_{min}}{X_{max} - X_{min}}$

4.3 RESULTS AND DISCUSSION

4.3.1 Determining filtration parameters

Which pixels are eliminated during the filtration process depends on the parameters used for each of the different filters. If overly stringent parameters are applied then very few pixels will remain per image which may reduce the capacity to average pixels together, or such parameters may result in the rejection of an unacceptable percentage of images (in their entirety). This, in turn, can make it overly difficult to acquire suitable data. Conversely, filtration parameters that are too relaxed may result in poorer performing and less robust models and classifiers. It is therefore necessary to find the optimal parameters which provide a balance between, maximizing SNR and the number of pixels maintained while also minimizing pixel-to-pixel differences and the number of images that are eliminated in their entirety.

The parameter combination evaluation routine was completed on both the GPN and CCJ datasets. Table 4.3 summarizes the results for the top performing parameter combination for each of the metrics. In instances where multiple combinations yielded identical results, the widest range was taken. For example, in many cases, extending the upper average absorbance cut-off to 0.9 had little effect due to the relatively small percentage of pixels that had an average absorbance near this value.

Table 4.3: Top performing parameter combinations for each of the metrics.

GPN	A_{min} - A_{max}	SNR	S_{P1}	S_{P2}	% pixels/image (A_{pi})	% of images remaining (N_{nz})	A_{snr}	A_{rd} (x 10^2) (lower is better)
S_{rd}	0.05 - 0.5	30	3	0.15	27 (277 ± 63)	6 (11)	164 ± 18	2.00 ± N/A
S_{snr}	0.05 - 0.7	30	1	0.1	25 (258 ± 39)	6 (11)	174 ± 20	2.88 ± 1.11
SS	0.05 - 0.7	30	1	0.1	25 (258 ± 39)	6 (11)	174 ± 20	2.88 ± 1.11
P	0.05 - 0.9	30	3	0.15	31 (315 ± 70)	9 (17)	167 ± 23	3.38 ± 1.00
I	0.1 - 0.9	15	2	0.15	41 (421 ± 162)	79 (150)	109 ± 25	5.30 ± 2.42
T	0.1 - 0.9	15	3	0.2	43 (443 ± 173)	81 (155)	107 ± 25	5.53 ± 2.39
CCJ	A_{min} - A_{max}	SNR	S_{P1}	S_{P2}	% pixels/image (A_{pi})	% of images remaining (N_{nz})	A_{snr}	A_{rd} (x 10^2) (lower is better)
S_{rd}	0.1 - 0.9	15	3	0.15	54 (553 ± 175)	80 (236)	96 ± 33	7.54 ± 2.82
S_{snr}	0.05 - 0.3	20	3	0.3	33 (340 ± 105)	39 (115)	116 ± 31	8.93 ± 3.83
SS	0.1 - 0.3	20	1	0.1	27 (274 ± 67)	22 (64)	109 ± 28	7.70 ± 3.57
P	0.05 - 0.9	15	3	0.3	54 (555 ± 180)	81 (239)	96 ± 33	7.57 ± 2.85
I	0.05 - 0.9	15	3	0.3	54 (555 ± 180)	81 (239)	96 ± 33	7.57 ± 2.85
T	0.05 - 0.9	15	3	0.3	54 (555 ± 180)	81 (239)	96 ± 33	7.57 ± 2.85

In the case of the GPN dataset, evaluating the parameter combinations on replicate similarity (S_{rd}), SNR (S_{snr}), or the combination (SS) resulted in a small but high-quality dataset (in terms of SNR and replicate distance). However, with 6% of the initial images remaining, these parameters are not feasible for general use. Taking into consideration the number of pixels and/or the number of images remaining (P, I, T) results in a relaxation of the parameter values. In particular, the reduction of the SNR cut-off from 30 to 15 results in an increase of the number of pixels per image to 41% (I) and 43% (T), however more importantly, the number of images remaining after filtration increases to 79% (I) and 81% (T). This result demonstrates the compromise that arises

when trying to maximize spectral quality spectra while preserving as much data as possible (a necessary component of the bacteria identification protocol that has been developed in conjunction with this work).

Examination of the CCJ dataset reveals a similar scenario where the P, I, and T metrics result in a relaxation of the cut-off parameters with a substantial gain in the percent pixels and images remaining. Interestingly, the parameter combination that resulted in the best result for the S_{rd} metric (which evaluates the differences in replicate images) yielded nearly identical results to the P, I, and T metrics. A possible explanation as to why this is the case with the CCJ dataset and not the GPN dataset may be due to the anomalous below-zero absorbance pixels present in the GPN dataset (see Figure 2.8). Overall, the T metric parameter combinations for GPN and CCJ resulted in nearly identical cut-off values. The higher A_{min} for the GPN dataset can be attributed to the below baseline pixels. The lower S_{P2} of the GPN dataset might also be explained in a similar manner or may be a due to the GPN dataset having a higher overall SNR.

4.3.2 Results of pixel filtration

The top performing parameter combinations for the T metric from the GPN and CCJ datasets were selected and evaluated below. Table 4.4 provides information regarding the percent of pixels and images that were eliminated at each stage of the pixel filtration routine. For both datasets it was found that the majority of pixels were eliminated as a result of the SNR filter. This exemplifies the importance of ensuring that pixels (and the mean spectrum of an image) are of as a great a SNR as is possible. Another important aspect to note is that while the phase 2 similarity filter removed only a few percent of the total number of pixels, it was responsible for eliminating nearly all the images that were eliminated by the filtration process. Although the phase 2 similarity filter eliminated 36 (GPN) and 51 (CCJ) images, the average pixel count of the eliminated images after the previous filters was 88 (GPN) and 131 (CCJ) respectively. These values are far below the averages listed in Table 4.3 indicating that these images contained sub-par spectral data. These results are a clear indicator of the usefulness of the similarity filter for identifying and eliminating images with excessive spectral

variability. The absorbance filter was responsible for eliminating 20.7% of the total number of pixels in the GPN dataset; the reason for which can be explained by the large number of below-zero absorbance pixels present. In general, these results suggest that it is more efficient to use a wider absorbance range and to ensure adequate spectral quality by eliminating pixels based on their SNR value. Considering that low absorbance spectra typically have a lower SNR, the detriment of a low average absorbance cut-off is minimized.

Table 4.4: Percentage of pixels and images eliminated by each of the filters for the GPN and CCJ datasets.

	GPN		CCJ	
	% of total pixels eliminated by filter (% of eliminated only)	% (#) of images eliminated by filter	% of total pixels eliminated by filter (% of eliminated only)	% (#) of images eliminated by filter
Absorbance	20.7 (32)	0	1.2 (2.2)	0.6 (2)
SNR	40.4 (62.3)	0.5 (1)	52 (93.8)	0.6 (2)
Phase 1 Sim.	1 (1.6)	0	0.7 (1.3)	0
Phase 2 Sim.	2.7 (4.1)	18.8 (36)	1.5 (2.7)	17.3 (51)
Total	64.8 (100)	19.3 (37)	55.5 (100)	18.7 (55)

During the filtration process, the number of times that a pixel was not filtered was recorded and the results shown by the images in Figure 4.9. This figure spatially illustrates the percent to which a pixel contributed to an image, with red indicating higher percent contribution and blue lower. As can be seen, there is a trend for pixels on the edges of the infrared image to be eliminated more often whereas the more contributing pixels are centrally located. A potential reason for this distribution is the bias that is introduced by the operator of the instrument due to preferentially positioning the desired portion of the sample at the center of the infrared image. In this way, it is more likely that portions of the sample that are either too thick or too thin will be found at the outskirts of the infrared image.

The blue band located at the bottom of the GPN dataset is also very apparent. Further examination showed that this band is the location of the below baseline spectra seen previously. The cause for this is most likely a result of delamination of the MCT

array and has been noted in other studies [12]. Delamination is a consequence of the ageing process of the detector and attributed to the repetitive cool-down and warm-up of the detector. The result is a region of pixels with anomalous behavior as can be seen in Figure 4.9. Consequentially, the fact that those pixels provided little to no contribution to the filtered data demonstrates the effectiveness of the filtration routines to extract functional data even if portions of the image are unusable. According to the manufacturer the likely hood of delamination occurring has been significantly reduced as a result of new fabrication processes allowing for 1000's of cool-down and heat-up cycles before it may occur.

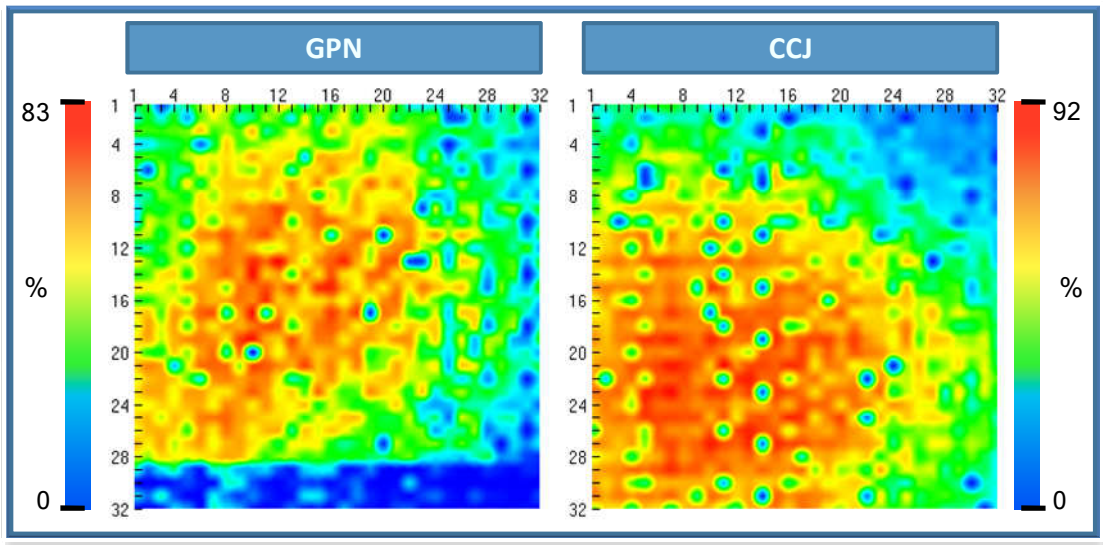


Figure 4.9: Images detailing the percentage that a pixel was <u>not</u> eliminated. A red point indicates that a pixel was rarely eliminated whereas a blue point indicates that the pixel was nearly always eliminated.

4.3.3 Benefits of pixel filtration

The primary intention of the pixel filtration routines is to filter out non-representative data. An example of the benefits which are gained by filtering out non-representative data is shown in Figure 4.10. This example consists of multiple replicates of several species and strains of *Listeria* that have been clustered using hierarchical cluster analysis. Differences in the spectra at this level of taxonomic specification are very small; as a result, minor differences between replicates will result in mixing of the

various types of *Listeria* in the dendrogram. As can be observed in Figure 4.10A, when replicate images are not filtered (except for an absorbance filter to eliminate extreme low and high absorbance spectra) it can be seen that the averages from the replicate images do not consistently cluster together. On other hand, when pixel filtering is applied the replicate images cluster together perfectly (Figure 4.10B). This demonstrates the importance of applying the pixel filtration routines on the infrared images in order to extract reproducible data that can be used to develop reliable models and to enhance the resolving capabilities of FPA-FTIR spectroscopy to the highest level of taxonomic specification possible.

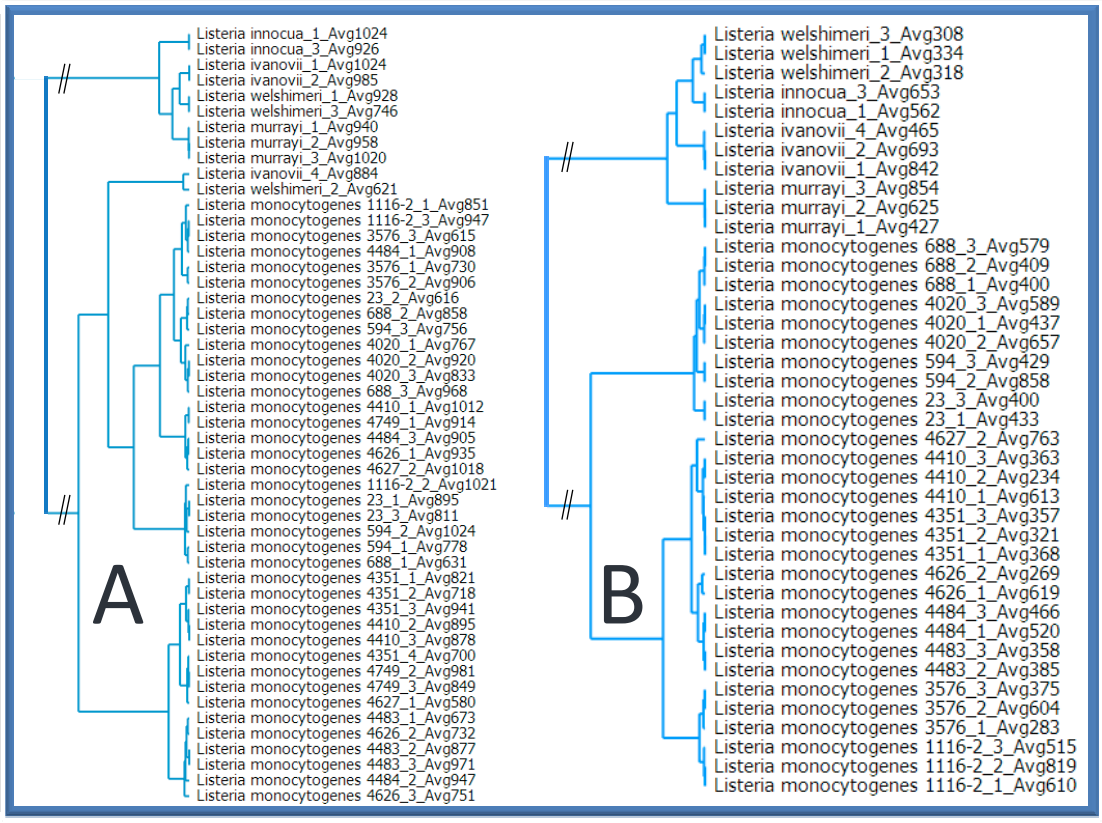


Figure 4.10: Dendrograms representing the averages from unfiltered (4.10A) (except absorbance) and filtered data (4.10B). Dendrograms were generated using the Euclidean distance measure and the Ward linkage method from spectral data in the region 1480 – 980cm^{-1}. Replicates are indicated by "_r_AvgNNN" where r indicates the replicate and NNN the number of pixels remaining after filtration that were then averaged together.

A second example of the effectiveness of the pixel filtration routines is shown in Figure 4.11. In this example, the GPN and CCJ datasets are clustered using principal component analysis on filtered and unfiltered data (except for an absorbance filter to eliminate extreme outlying pixels with either near-zero absorbance or off-scale absorbance levels). Pixel filtration results in a clear improvement in the separation of Gram positive/negative (GPN) and the separation of *Campylobacter coli/jejuni* (CCJ). In addition, non-representative variability has been reduced sufficiently that secondary groups (Figure 4.11 GPN-F) can be seen upon clustering. These groups represent the various genera and species making up the Gram positive and Gram negative data.

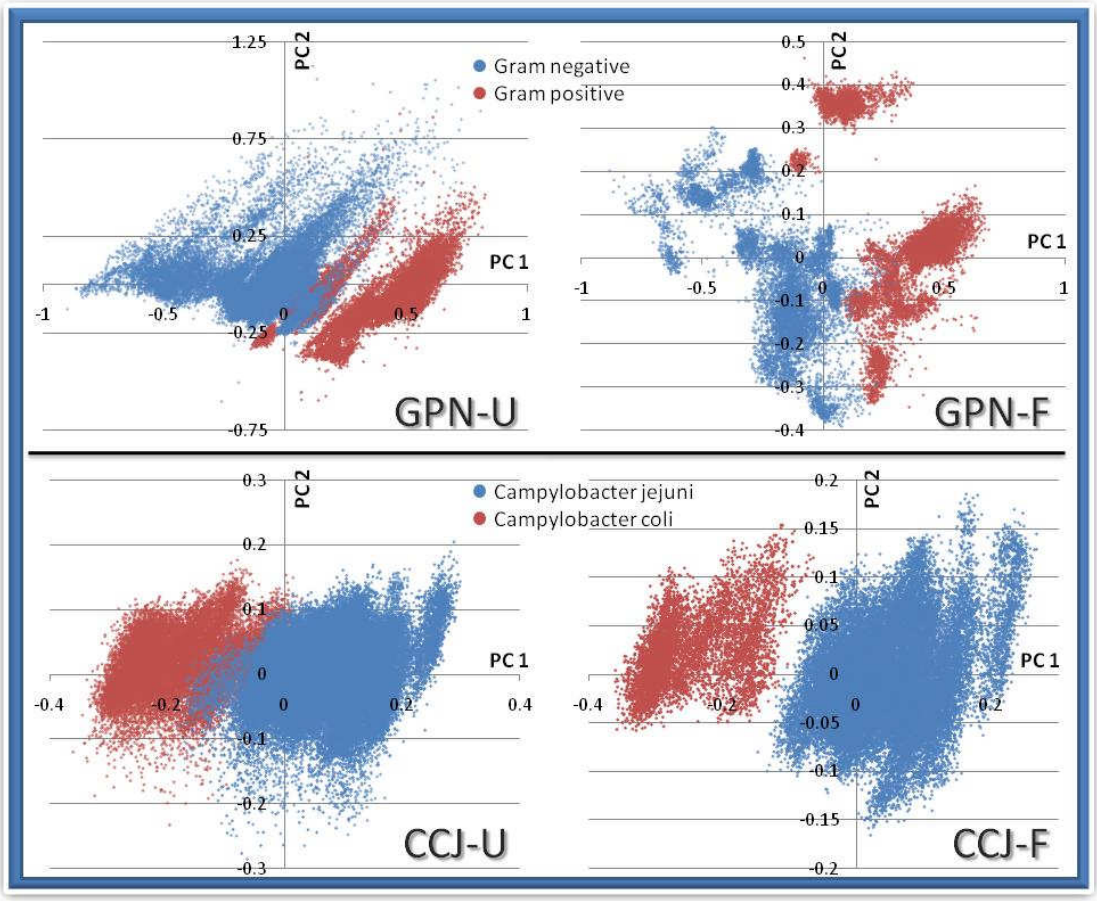


Figure 4.11: Principal component plots of the GPN and CCJ datasets without filtration (GPN-U,CCJ-U) and with filtration (GPN-F,CCJ-F). PCA was performed on spectral data in the region of 1480 – 980 cm^{-1}. Each spectrum used in the PCA consisted of the average of 4 pixels in order to facilitate the processing of principal components.

Pixel filtration has been shown to be beneficial for extracting the relevant data from infrared images of bacteria deposited onto zinc selenide. However, there is great opportunity to apply these techniques to a variety of applications. Each type of filter offers the potential for evaluating various characteristics of the sample under analysis and provides the possibility of extracting spectral data where it may not otherwise be feasible using traditional methods. Absorbance filters can be used to evaluate sample pathlength uniformity, or by targeting specific spectral features can be used to extract the desired chemical components for evaluation or classification when examining heterogeneous samples. The SNR filter can be used to maximize spectral quality by ensuring sufficient signal is present for further analysis. The similarity filter can equally be used in a variety of ways. Notably, it can provide the ability to extract specific components for classification from chemically heterogeneous samples, it can be used as a quality control mechanism by evaluating sample heterogeneity, and can potentially offer improved model development and enhanced classification for applications that deal with samples which contain frequent contaminations on the micron scale. Applying these filtration techniques allows for FPA-FTIR spectroscopy to be used as an enhanced approach to classification by FTIR (if SNR requirements are met) spectroscopy due to its ability to extract useful data from samples which exhibit physical and chemical heterogeneity. In addition, these techniques could potentially be applied for the optimization of sampling protocols, and the examination of instrument-to-instrument performance.

By applying FPA-FTIR spectroscopy as a tool for data-redundancy as opposed to being used as a technique for acquiring images for spatial chemical analysis a new domain of possible applications using FPA-FTIR spectroscopy has been created. Although a costly proposition, FPA-FTIR spectroscopy used in this manner can allow for databases to be generated, models to be built, and unknowns to be classified from types of samples that would make a similar endeavor using traditional single element detectors much more difficult and less reliable. In addition, the sheer number of spectra

acquired using FPA-FTIR spectroscopy allows for the use data analysis methods inaccessible to single element detectors.

4.4 CONCLUSION

This study introduces an approach to extract only the most representative sample data from infrared images by employing a pixel filtration process whereby pixels that do not conform to certain spectral quality characteristics are eliminated. In addition, it demonstrates a method for optimizing the filtration process and illustrated the benefits of this process on the classification of bacteria. The result is that infrared imaging can be applied to samples which may contain non-uniformities or inhomgeneities that would otherwise make the analysis of such samples impossible. With regards to bacteria identification, it ensures that only suitable data is used in the generation of spectral databases and during the classification of unknowns.

CONNECTING STATEMENT

Chapter 4 focused on the extraction of relevant data on the image level by employing pixel filtration routines which assist in eliminating data which is not representative of the sample. After extraction the next stage towards developing an expert system is the generation of classifiers. In the context of bacteria identification, many classifiers benefit from an optimization stage referred to as feature selection, which can help enhance their discriminatory power. This process involves the selection and omission of portions of the spectrum. Similarly to pixel filtration, feature selection is a form of data extraction; however, in this circumstance it is conducted on the spectrum level. This chapter introduces a weighted feature selection approach which attempts to improve upon the shortcomings of more commonly used techniques, all the while providing additional (as compared to other approaches) information regarding the chemical characteristics that are relevant with respect to the desired separation.

Chapter 5: Continuous Genetic Algorithms for Evaluating Feature Relevancy in Infrared Spectra

5. 1 INTRODUCTION

In many applications which employ infrared spectroscopy for class based discriminations it is necessary to identify the features (wavelengths) in the spectrum which supply relevant information. Identifying these regions (feature selection) assists in building reliable models and eliminating extraneous data which is not otherwise needed. In the majority of cases where a feature selection method is applied the resultant information indicates either the complete inclusion or exclusion of a feature for further use in developing models and classifiers. However, considering that in many cases the spectral bands under examination span multiple points (given adequate spectral resolution) there then exists a high correlation between neighboring spectral features. Additionally, in the case of complex data sets with a non-linear separation between the classes, features may only provide partial relevancy. This study demonstrates a technique to assess the relevancy of spectral features through the use of cumulative genetic algorithms (CGA). The goal of this technique is twofold; to use a weighted feature set in order to provide greater flexibility in ensuring that the classifier model is neither overfitted nor underfitted, and to extract information concerning the relative importance of the various regions of the spectrum towards the target classification. This information can be potentially used to infer generalized statements of chemical differences and similarities between the samples under analysis. In addition, this technique may be capable of providing an indication of spectral relevancy even when the development of an ideal classifier, where all groups separate completely, is not possible.

Genetic algorithms have been one of the more widely used methods of feature selection due to their ability to simultaneously examine large portions of the state space of possible features. This approach is independent of the number and size of the regions; however, due to the non-deterministic nature of genetic algorithms, repeated searches on the same data may yield slightly different results. It is this aspect of GA's that is exploited in this approach to defining feature relevancy.

While the majority of optimization techniques used for feature selection do not provide indications of relevancy, there are other approaches which may provide similar information as that which is attained through the use of CGA. The average spectrum for each class in a dataset can be compared to determine where they spectrally differentiate. The loadings which are computed by principal component analysis (PCA) can indicate which spectral features provide the greatest contribution to the first few principal components. However, this is assuming that the dataset can be adequately separated based on the first few principal components, which in the case of datasets of bacteria, the focus of this study, is often not the case. Due to the unsupervised nature of PCA, if the within class spectral differences are on the same order as between class differences then the variance identified may not necessarily be associated to the desired separation. Partial least squares regression and its ability to identify pure components can potentially be used to acquire information regarding the underlying differences between the classes. However, similar to PCA the non-linear nature of the differences between the two classes (two genera consisting of bacteria from multiple kinds of species) and the assumption of a pure component associated to a genus is overly optimistic. A non-linear PLS can be employed; however, this adds a layer of complexity, such as the selection of which kernel to use and its operational parameters, which would make the implementation and interpretation of such a technique more difficult than the proposed method in this study. Artificial neural networks is another approach which is capable of implicitly identifying the spectral features to optimize the target separation, however, extraction and interpretation of this information from the input weights of each of the layers of neurons is by no means a trivial task.

This study examines feature relevancy in the context of separating bacteria of various types and at different taxonomic levels. The spectrum acquired from bacteria consists of the superposition of all the chemical components which make up the bacteria. This includes the lipids, proteins, carbohydrates, DNA, and any another compounds which may be present in the bacteria. The result is a spectrum with broad bands where one cannot associate individual features to specific bacterial attributes (Figure 5.1).

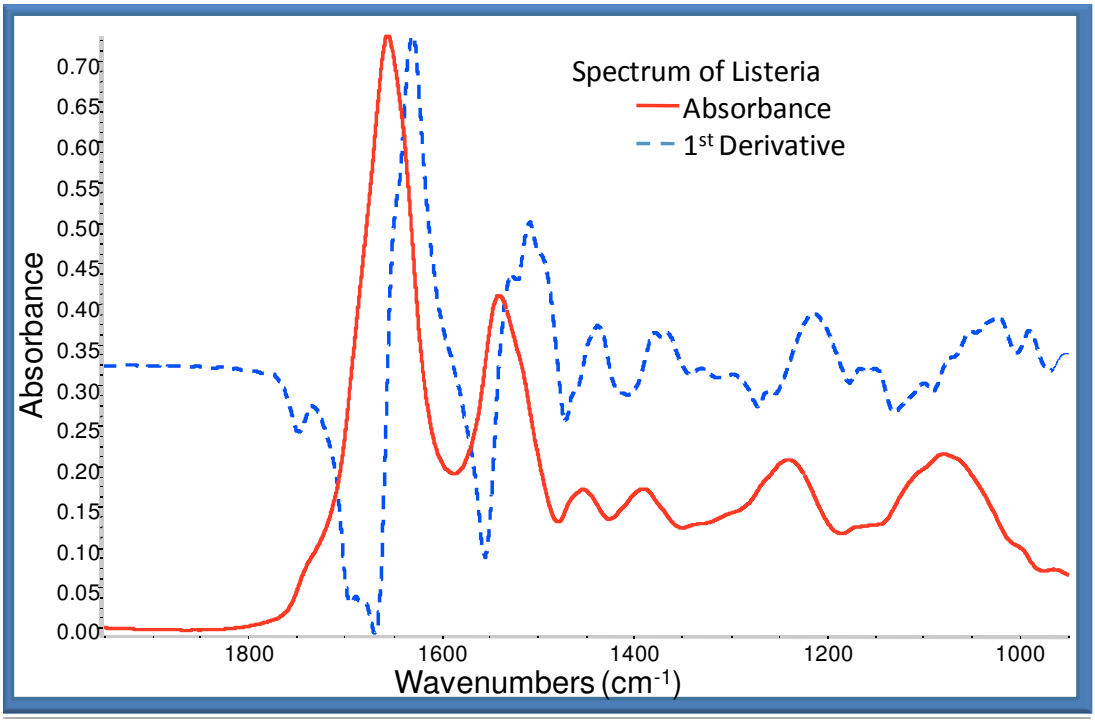


Figure 5.1: Example of a raw absorbance spectrum (red) and first derivative spectrum (blue) of *Listeria*.

In the case of bacteria identification, if one wishes to generate a model which can classify a bacteria as one genus or another where each genus consists of a variety of different species then it is typically the case that there is no single identifiable attribute which can be recognized as differentiating the two classes. In addition, it is often the case that the defining features for separation are not associated to spectral 'peaks' in the traditional sense. The features may be part of shoulders, valleys, or plateaus in the

raw absorbance spectrum. It is in these circumstances that a technique, which can provide an understanding of the relative importance of the various spectral features towards the separation of the two classes, can offer useful information of the types of chemical differences that exists between them.

5.1.1 Genetic algorithms

A genetic algorithm is a non-deterministic optimization routine which uses the principle of natural selection in an attempt to evolve an optimal result [45, 74]. It allows for simultaneous examination of a large portion of the state space (the permutations of possible features to choose) and through mutation and cross-linking, analogous to their natural counterparts, evolves better performing feature sets. As with most optimization algorithms there is the possibility of arriving at a local maximum. A GA consists of a population of individuals with each individual having a unique set of genes. In this approach each gene is representative of a single feature (point in the spectrum) which can have one of two states, 1 or 0, where 1 implies the inclusion of a feature and 0 the exclusion. Each individual of the population is randomly initialized and then evaluated by means of a fitness function. The fitness function is a measure of how well a set of features separates the classes, which in this study is defined by a classification score (see below). The better performing individuals are mated by cross-linking their genes (Figure 5.2) to create new population members. These new individuals are then evaluated and the process repeated. In this way the process continues until there is little or no change in the classification score of the top performing individual. In addition, at random intervals a random mutation will be introduced into an individual to stimulate variety. This random mutation serves as a means of avoiding local maxima.

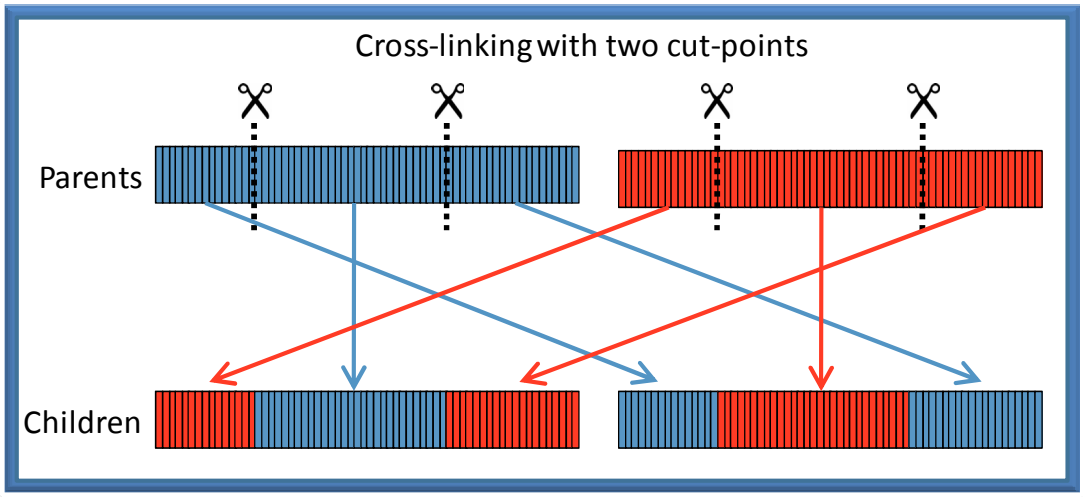


Figure 5.2: Description of the mating process between two individuals of a population. Cut-points are randomly chosen and the series of genes of each individual mixed to form the children.

5.1.2 Cumulative genetic algorithms

The approach developed in this study is termed cumulative genetic algorithms as it interprets the accumulated results from multiple independent runs of a genetic algorithm. A run is defined as the complete execution of a GA from initialization to completion by reaching one of the stopping criteria (no more improvement from one iteration to the next or the maximum number of iterations is reached). Due to the non-deterministic nature of GA's, a statistical representation of the importance of a feature towards the separation of different classes can be established by examining the number of times a feature is represented after multiple repetitions of the genetic algorithm feature selection routine. This approach is similar to bagging [75] (also known as ensemble classifiers), a technique used to improve the training of classifiers. Bagging involves training multiple classifiers on subsets of the training data. Predictions are then run on all of these classifiers and a voting scheme used to decide the final outcome. The CGA approach differs in that it uses all the data, not subsets, and relies on the non-deterministic nature of GA's to have varying output. Additionally, the ideas presented here were motivated by an approach where the initialization of one iteration of the GA was influenced by the features selected from the previous [76].

Due to the independent nature of each run of the genetic algorithm and given a sufficient number of repetitions the reproducibility of the results obtained using this method is very good. For example, comparison of the trace obtained from two runs (of 100 repetitions each) yielded a correlation coefficient of 0.991.

The procedure for defining feature relevancy using CGA is outlined in figure 5.3. The process begins by computing one complete run of a GA (evolving the population until no improvement is found). The genes of the top performing individual are extracted and multiplied by the fitness (classification score) of this individual. This serves as a weighting on the extracted genes (feature set) such that if the fitness of the top performing individual of a particular repetition is significantly worse or better then their contribution is scaled accordingly. Multiple individuals are not used at the completion of each run of a GA as typically the top performing members of a population will have a high degree of similarity to each other.

There are several operational parameters which must be defined as part of this algorithm. Namely, the initialization values for each individual of a population, the size of the population to use, and the number of repetitions of the GA. These parameters significantly affect the output of the CGA and must be defined appropriately. Details concerning the effect these parameters have on the output are provided in the results and discussion section.

With regards to the initialization values, it is often the case that the population of a GA is initialized randomly, however, this approach is contrary to the idea that there is a correlation between neighboring features. For this reason individuals are initialized by the approach defined in Equation 5.1. Two variables (C_N and C_S) can be adjusted in order to control the average number and average size of the regions in the initialized population (a region being defined as a consecutive set of genes with a value of 1).

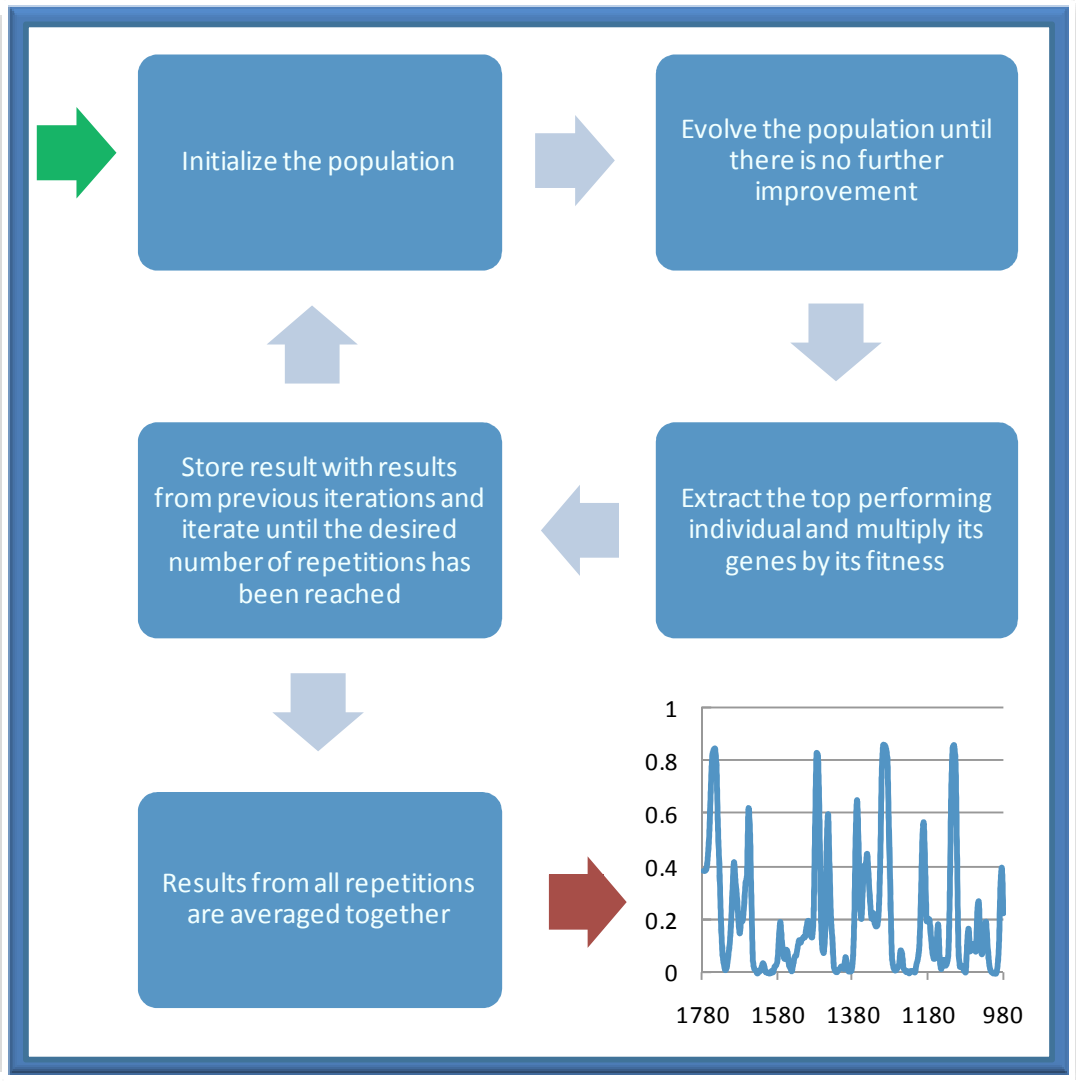


Figure 5.3: Description of the CGA routine. The process involves repetitions of a GA feature selection routine. After each repetitions the results are stored and the process repeated until the desired number of iterations is reached. The final CGA output is the cumulative (average) representation of each feature in all of the GA iterations. This value is represented as a contribution between 0 and 1, where 1 implies the feature was selected in all iterations of the GA's.

$$g_0 = \begin{cases} 1, & if\ rnd < C_N \end{cases}$$

$$g_i = \begin{cases} 0, & if\ (g_{i-1} = 0\ \&\ rnd >= C_N)\ or\ (g_{i-1} = 1\ \&\ rnd >= C_S) \\ 1, & if\ (g_{i-1} = 0\ \&\ rnd < C_N)\ or\ (g_{i-1} = 1\ \&\ rnd < C_S) \end{cases}$$

Equation 5.1: The initialization process used to define the genes of a new population. C_N affects the number of regions, whereas, C_S affects the size of the regions. *Rnd* is a random number between 0 and 1 inclusive which randomly changes as each gene is initialized.

As can be seen by equation 5.1, by adjusting the value of C_N, the likelihood that a region is initiated, the number of regions can be controlled. By adjusting the value of C_S, the likelihood that a region grows larger, the average size of the regions can be controlled.

5.1.3 Fitness function

Genetic algorithms are an optimization routine and, as such, need to be used in conjunction with an evaluation method (the fitness function). The fitness function provides a numerical score of the performance for a given set of features of an individual. The approach used in this study to evaluate a selection of features is that of a leave-one-out cross validation using k-nearest neighbor (LOOCV-KNN) [63].

A leave one out cross validation (LOOCV) [45] operates by incrementally removing each member of a data set and testing it against a model built using the remainder of the data set. In other words the idea is to remove a sample from the training set, build a model with the remaining samples and then test the one that has been removed. This process is repeated for every sample in the training set. In this approach the k-nearest neighbor (k-NN) is used to perform the evaluation at each step of the LOOCV where k is equal to one less than the number of members pertaining to the class of the removed member. For example, if the removed member belongs to a class composed of eight samples, the k-NN search will be conducted and return the seven (k=7 as one has been removed) nearest neighbors. In this model, the score for an individual (removed) member is determined as the number of samples out of seven that

belong to the class of the removed member. The final classification score used for a set of regions is taken as the average score of the individual scores (the score from being removed and reclassified). The score that results will be a value between 0 and 1 with 1 indicating that every nearest neighbor is a member of the same class for every member left out.

Some advantages to using this method for evaluating a set of features is that it provides a good indication of the separation between the classes, it is deterministic, and does not require training. However, the LOOCV-KNN approach can become computationally expensive as the number of reference samples increases. Specifically, the computational cost increases by the square of the number of reference samples. An upper limit can be set on the number of nearest neighbors to use, however, the effectiveness of the LOOCV-KNN algorithm would then be diminished as the reference set for each class grows larger while the number of neighbors used remains the same.

5.2 MATERIALS AND METHODS

5.2.1 Bacteria

Four datasets (Table 5.1) are used as inputs for the CGA routine in order to provide results from varying levels of taxonomic specification. All bacteria were prepared and deposited at Health Canada (Longeuil, Quebec, Canada) using standard methods and protocols [23]. After incubation the bacteria were transferred to a polished Zinc Selenide infrared transparent slide using a sterile wooden probe resulting in a bacterial film. Multiple infrared images were recorded from each deposited organism.

Table 5.1: Description of the four data sets used in this study

Taxonomic level	Abv.	Separation of	# of images
Gram	GPN	Gram positive vs. negative	74/117
Genus	ES	Escherichia vs. Salmonella	17/36
Species	SHG	Shigella flexneri vs. sonnei vs dysentarie	53/45/21
Species	CCJ	Campylobacter coli vs. jejuni	77/217

5.2.2 Spectral acquisition

The spectral images were collected on a Varian Excalibur FTIR spectrometer operating under Varian Resolutions Pro 4.0 software (Varian, Melbourne, AU) and equipped with a UMA-600 infrared microscope and a 32 X 32 (1024 pixels) mercury-cadmium-telluride focal-plane-array detector. A constant flow of dry air was used to purge the spectrometer and the microscope, limiting spectral contributions from carbon dioxide and atmospheric water vapor. Each spectral image was collected from a field of view of 188 X 188 μm^2 with a spatial resolution of 5.6 µm and a spectral resolution of 8 cm^{-1}. Each image consisted of 256 co-added scans which required approximately 5 minutes to acquire.

5.2.3 Data processing

All images were pre-processed applying pixel filtration routines in order to extract the most highly representative spectral information from each infrared image. Initially, all spectral data were truncated to the region of 1780-980 cm^{-1}. The pixel filtration steps involved an absorbance filter of 0.1 to 0.7 average absorbance units on the region 1480-980 cm^{-1} to remove spectra with too great or too little absorbance. All spectra were then derivatized and a noise filter applied to eliminate data with a measured signal-to-noise ratio lower than 15 by comparing the peak-to-peak value of 1480-980 cm^{-1} over 2200-2000 cm^{-1}. The spectra were vector normalized after which a two phase similarity filter was applied in order to eliminate outlier pixels and insure that the variability of each image was within acceptable norms. Finally, the average of the remaining pixels for each image was taken and this data used for the subsequent analysis. The average signal-to-noise ratio of the average spectrum from each image was within 85 to 100 for the datasets used. Prior to analyzing each of the datasets using the CGA routine, a preliminary analysis was conducted to identify and remove spectra which proved to be outliers during the classification process.

The pixel filtration, PLS regression, feature selection, and CGA was accomplished on software developed in-house on the Visual Studio .Net 2005/2008 platform (Microsoft, Redmond, CA). The processing routines were validated (where possible) by

visual and computational comparison of outputs from Matlab 6.5 (Mathworks, Natick, MA) and Omnic 6/7 (Thermo Nicolet, Madison, WI). The processing was completed on a quad core Intel (Santa Clara, California) processor operating at 3.2 gigahertz with 4 gigabytes of random-access-memory running Windows Vista Business edition. The software developed as part of this study was optimized for parallel processing on multi-core central processing units (CPUs) allowing for significantly greater computational output than most other software available at the time.

5.2.4 Feature selection

The grid-greedy feature selection was done with 3 regions of a minimum size of 20 wavenumbers (6 features) and a maximum size of 92 wavenumbers (24 features) per region. All possible combinations of such regions were evaluated between 1780-980 cm^{-1} and the region with the highest LOOCV-KNN classification score selected. The greedy portion of the algorithm examined combinations of adjacent features following the path of greatest improvement.

The forward selection began by evaluating the single feature with the highest classification score and followed by adding features one at a time which kept the score at a maximum. The routine stops when the classification score is no longer improved by adding features. The search would continue for a minimum of 21 features (10% of the total # of features) even if there was no further improvement in classification score in order to minimize overfitting of the training data.

5.2.5 Partial least squares (PLS) regression

Partial least squares (PLS) regression [77] is a well known technique for building "whole spectrum" linear calibrations. It offers excellent resolving capabilities in defining spectral pure components even when there are multiple un-accounted for contributions to the spectra. A PLS model of the bacteria data was constructed by assigning a component for each of the four types of bacteria under analysis. A particular spectrum of bacteria would be assigned (a concentration value of) 0 for all components except for the one it belonged to where it would receive a value of 1. The PLS models were built on

the region 1780-980 cm^{-1} using a K-fold cross-validation technique (K = 4). A Q^2 value [77] (cross-validated R^2) was determined in order to identify the appropriate number of latent variables to use.

5.2.6 Computational complexity

Evaluating feature relevancy in the manner described in this chapter can be a time consuming process. The computational time required depends on the size of the data set, the population size of the genetic algorithm, and the number of repetitions that are applied. Fortunately, the process scales linearly with the number of computer processors available. The software developed in this study can distribute the computational load evenly among all available processors; taking advantage of today's (and tomorrow's) multi-core CPUs. In terms of the operational parameters used, the process scales linearly with respect to the number of repetitions and the size of the population used. While it is known what affect the reference sample set size has on determining the classification score (using LOOCV-KNN), it is difficult to provide a generalized statement regarding the affect of the sample set size on the convergence rate of the GA. Any change in the sample set affects the complexity of the separation and as a result the average time to completion for each run of the GA. Likewise, it is difficult to compare the computational time of two data sets which consist of different spectral information. A dataset of 120 spectra with approximately 2-3 spectra per bacteria would require approximately 2 to 3 minutes to complete a single GA iteration. It would take approximately 4 hours to generate a relevancy plot consisting of 100 GA iterations. However, given the rapid rate of improvement in the computational capabilities of computers, the time taken for computing the CGA approach will likely be cut in half every two years [78] (assuming the latest computer technologies are applied).

5.3 RESULTS AND DISCUSSION

5.3.1 Initialization parameters

Table 5.2 provides statistics for the average number and size of regions as well as the total number of genes set to 1 after initialization of a population of 1000 individuals at several values of C_N and C_S.

Table 5.2: Population statistics for 1000 individuals after initialization by varying CN and CS values. Each individual consists of 204 genes corresponding to the 204 points between 1780-980 cm^{-1} at 8 cm^{-1} resolution.

C_N	C_S	Region Size	# Regions	Total Genes Set to 1
0.7	0.1	2.9 ± 0.8	15.8 ± 3	52.6 ± 12.7
0.7	0.15	2.9 ± 0.7	20.7 ± 3.1	69 ± 12.5
0.7	0.2	2.9 ± 0.6	24.9 ± 3.2	82.7 ± 11.9
0.7	0.25	2.9 ± 0.6	28.5 ± 3.4	94.3 ± 12.1
0.8	0.1	4.6 ± 1.3	13.9 ± 2.6	69.5 ± 15.6
0.8	0.15	4.5 ± 1.1	17.9 ± 2.7	88.2 ± 14.7
0.8	0.2	4.5 ± 1.1	20.9 ± 2.9	102.3 ± 14.7
0.8	0.25	4.4 ± 0.9	23.4 ± 3	114.4 ± 12.9
0.9	0.1	9.6 ± 3.4	10.6 ± 2.3	102.5 ± 22
0.9	0.15	9.5 ± 2.9	12.8 ± 2.5	122.8 ± 19.1
0.9	0.2	9.5 ± 2.7	14.3 ± 2.7	137.7 ± 15.7
0.9	0.25	9.5 ± 2.6	15.4 ± 2.9	147.7 ± 13.6

Figure 5.4 gives an example of the effect that occurs on the output of the CGA as a result of adjusting the initialization values. When initialized individuals contain many small regions (small sets of genes set to 1) we see that the resultant relevancy graph contains many sharp well defined peaks. This type of output is similar to that of conventional feature selection routines in that features are either totally included or excluded.

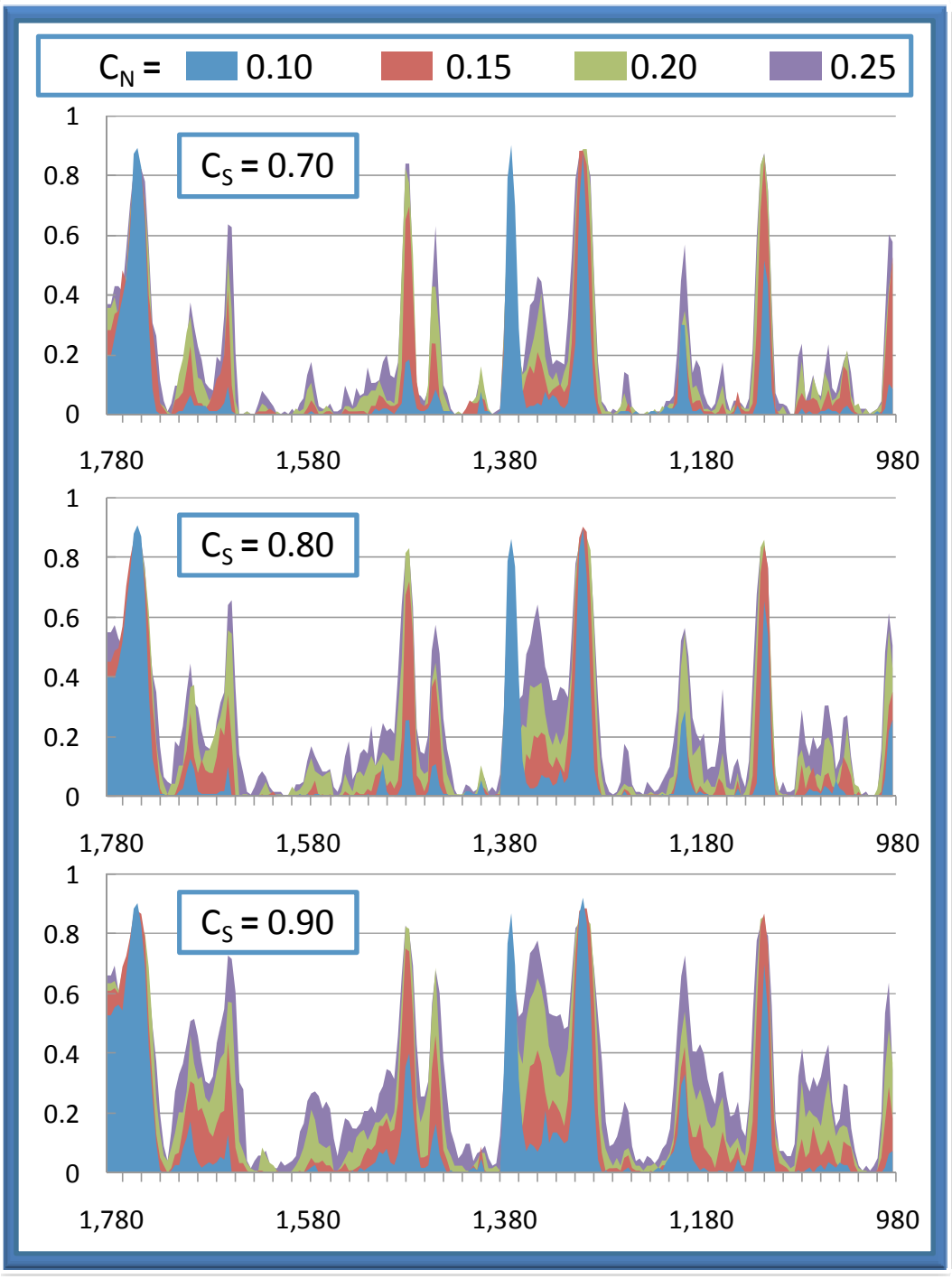


Figure 5.4: Example of how varying the initialization parameters affect the CGA output. The CGA routine was performed on the ES dataset with a population size of 1000 and 100 iterations.

Although, this identifies the most important spectral features (those that provide maximal separation), it does little to enhance the overall understanding of feature relevancy beyond what can be attained using conventional feature selection techniques. If the initialized population contains fewer but larger regions then the relevancy graph will consist of broad peaks with nearly all regions displaying some degree of relevancy. In this case, overall performance is degraded by having poor regions "piggy back" with the more performing features. In other words, the large size of the initialized regions makes it less likely that these (poor performing) features are eliminated during the evolutionary process. Table 5.2 outlines a range of initialization parameters and the number and type of regions that form as a result.

As can be seen by Figure 5.4, even if two different initialization parameters yield a similar number of features used, the difference in the number and size of the regions will result in a different relevancy plot (consider C_S=0.8,C_N=0.2 vs. C_S=0.9,C_N=0.1). The initialization parameters used in the remainder of this study consisted of C_N=0.2 and C_S=0.8 which results in ~50% of features being used after initialization of the population. These values were chosen as they offered a good compromise between near maximum classification score, a wide distribution of peaks at all level of relevancy, but with regions of zero contribution. Given sufficient computer resources, additional feature relevancy information could be attained by computing several relevancy plots with different initialization parameters.

5.3.2 Population and iteration parameters

The size of the population to be used by the genetic algorithm is another important parameter, the effect of which is shown by Figure 5.5. Three examples are given with both large and small populations and with high and low number of iterations. The number of iterations was kept inversely proportional to the population size in order to have comparable computational requirements. The effect of increasing the number of iterations is that the output of the CGA becomes smoother, as can be seen by comparing (B) and (C). However, computational restraints require that the population size be reduced if the number of iterations is increased. If the population is very small

then an iteration of a GA will often reach a sub-optimal result due to irrelevant or detrimental features finding their way into the top performing individuals. This can be seen by the fact that there are no features in which their percent contribution is zero for a population size of 10 (A). If the population is very large (C) then there is greater variability in the population and more children generated at each step of the GA. This increases the likelihood of achieving a highly optimized result as can be noted by the sharp bands and large number of features with zero contribution. As the population size is increased, the results of the CGA are analogous to those of typical inclusion/exclusion methods to feature selection (such as standard GA's, grid-greedy, or forward selection algorithms). Figure 5.5C may indeed be the optimal features to use in the separation of the classes, if comparison with a test set shows that the data has not been overfitted.

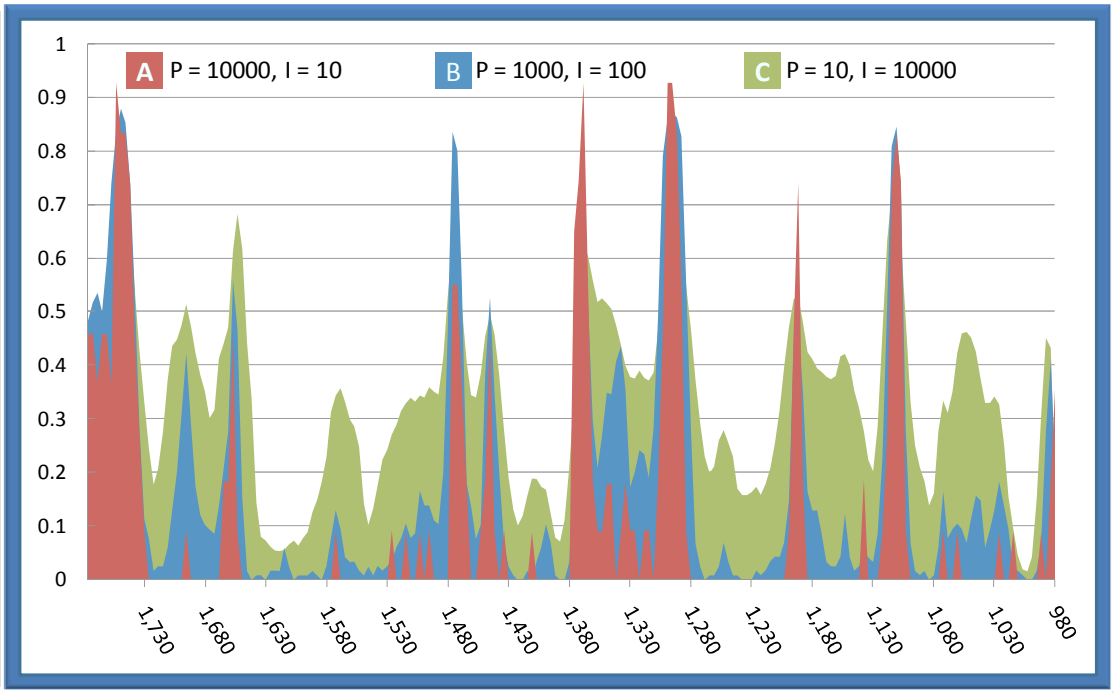


Figure 5.5: An example of how the GA population size (P) and number of iterations (I) affects the CGA output. The CGA routine was performed on the ES dataset with initialization parameters $C_N = 0.2$ and $C_S = 0.8$.

5.3.3 Comparison of CGA

5.3.3.1 Comparison of CGA and spectral variance

Figure 5.6 compares the output of the CGA to the variance between the average spectrum of each species in the Shigella dataset (recalling that the goal of this dataset is to separate three species of Shigella from each other). Although differences between averages may not be representative of the differences between the spectra which make up those averages, there is clearly more variation in the region of 1230-980 cm^{-1} which is the identical region that is indicated by the CGA as having features with the highest contribution. On the other hand, it can be seen that outside of the 1230-980 cm^{-1} region there are portions of high variance that offer zero contribution. In these regions the variation between the spectra is not representative of the different classes of bacteria and as such is detrimental to the classification.

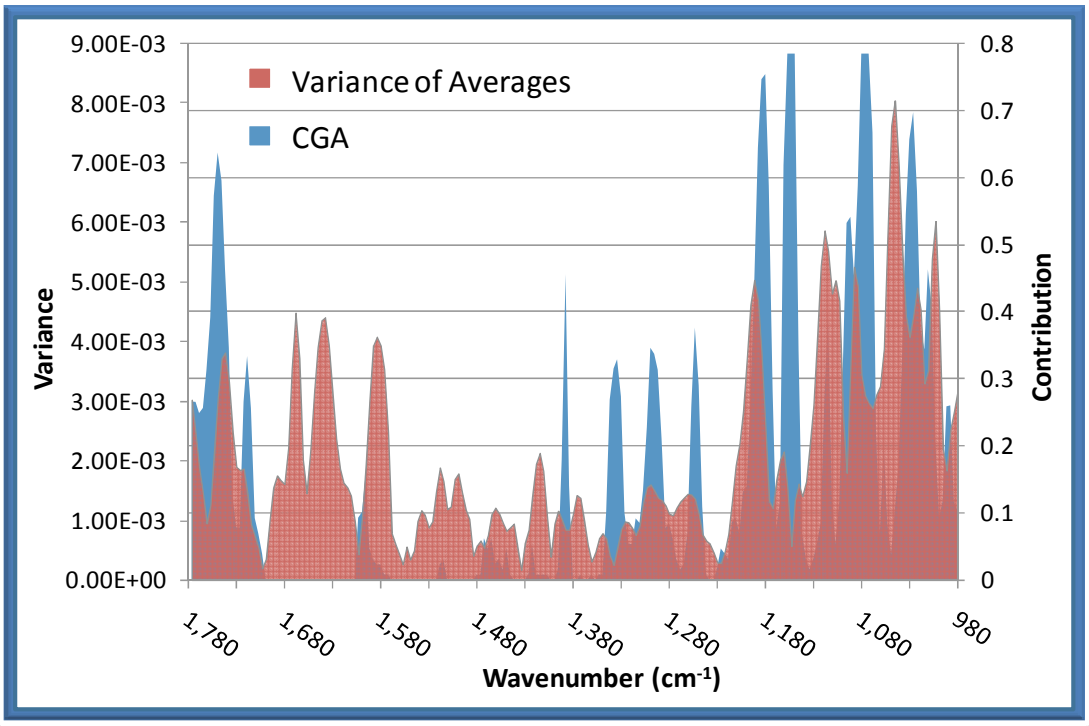


Figure 5.6: Comparison between the output of the CGA routine and the variance between the average spectrum of each species in the SHG dataset.

5.3.3.2 Comparison of CGA to PLS

The result of comparing the feature relevancy output of the CGA to the variance between pure components of the PLS model is illustrated in Figure 5.7. Again, similarities can be seen in the region from 1200-980 cm^{-1}. In addition, it appears that in nearly all places where there are larger peaks in the PLS output there are peaks in the CGA output. Considering that the outputs were generated using completely different techniques provides reassurance that the CGA relevancy plot peaks represent true differences between the types of bacteria being separated. There are still, however, a number of major peaks present in the CGA relevancy plot which are not indicated in the PLS plot. Despite statistical methods to determine the number of latent variables to use as part of the PLS regression, changes in the number of latent variables used would result in changes in the peaks present in the PLS plot, including those major peaks in the CGA plot not previously seen in the PLS plot. Additionally, it should be considered that the computational time of the PLS plot is on the order of minutes or less, whereas the feature relevancy plot required several hours to compute.

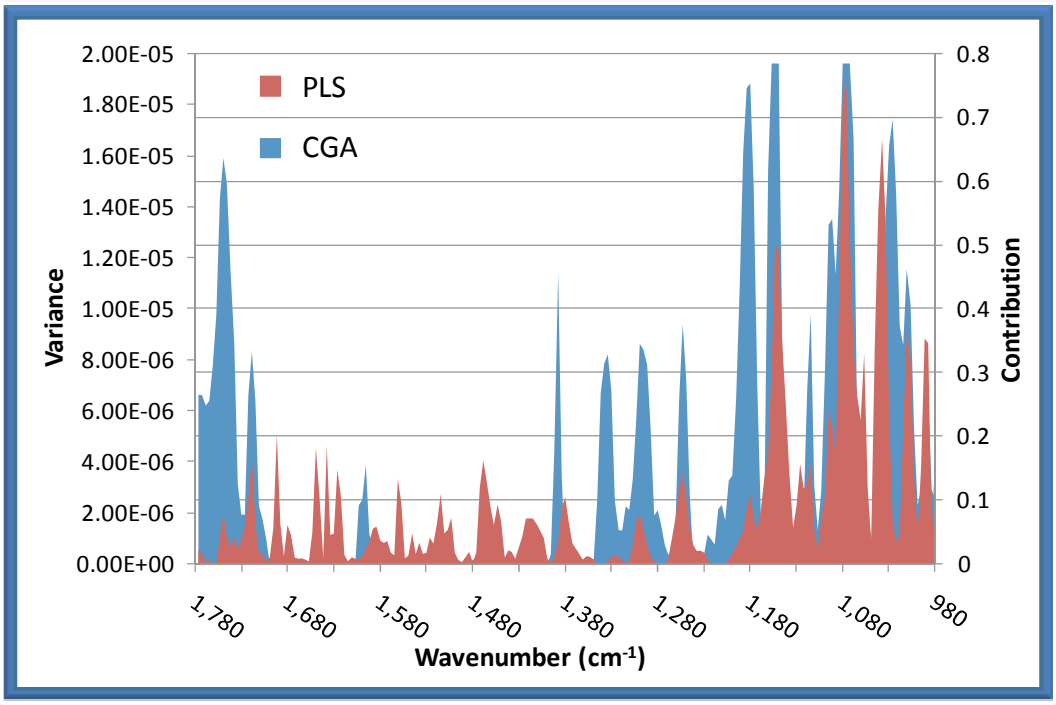
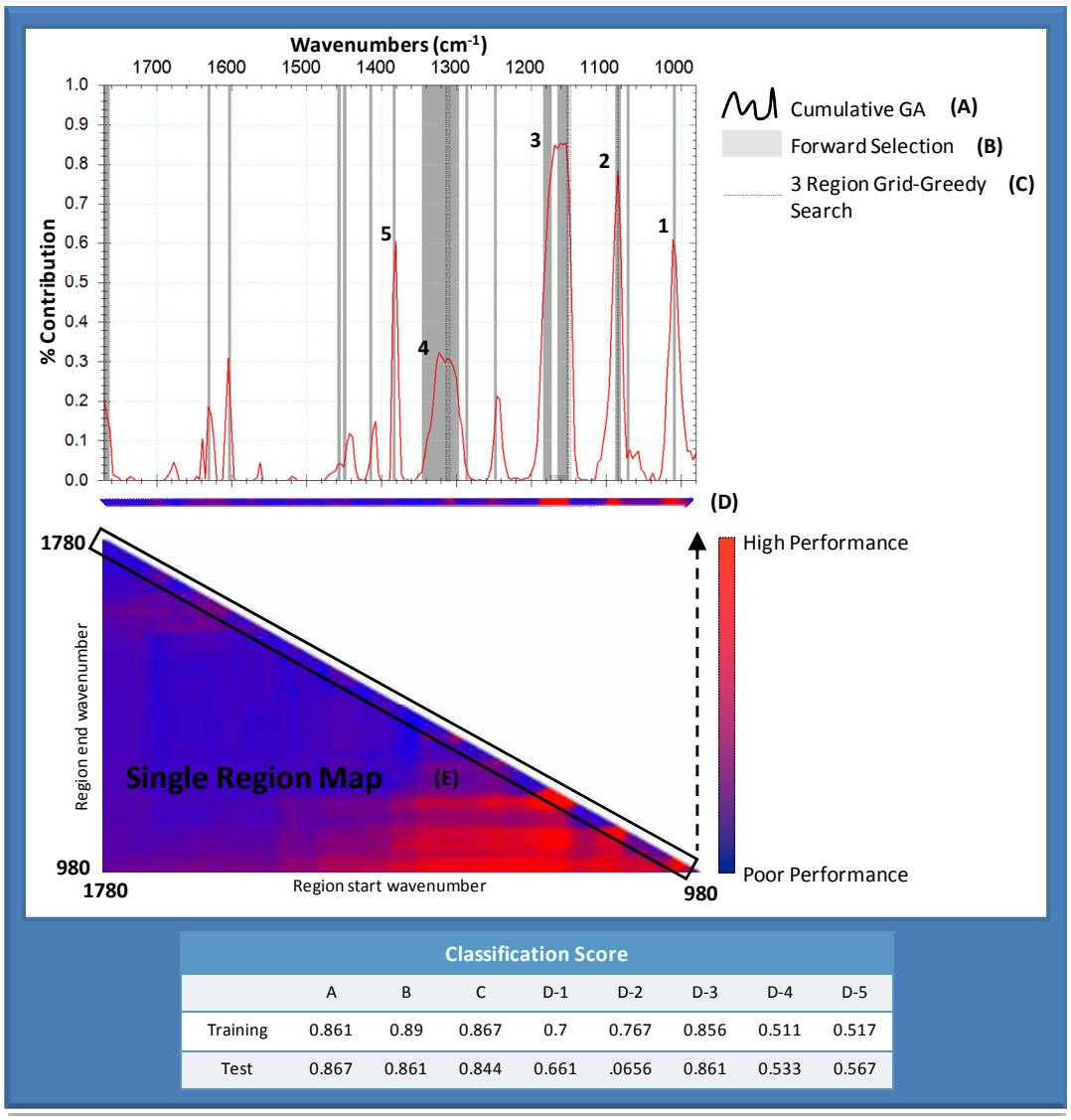


Figure 5.7: Comparison between the output of the CGA routine and the variance between pure components of a PLS model (10 latent variables) of the SHG dataset.

5.3.3.3 Comparison of CGA to other feature selection methods

The relevancy of the features identified by the CGA method was compared to the features determined by the forward selection and grid-greedy algorithms of feature selection. All techniques used the LOOCV-KNN method of defining a classification score for a given set of features. Figure 5.8 illustrates the regions defined by the cumulative genetic algorithm (A), forward selection (B), three-region grid-greedy search (C), and single region mapping (D and E). The regions defined by the forward selection routine parallel those of the cumulative genetic algorithm (CGA), however, it can be seen that there was a strong tendency towards a wide band which offered a minor contribution on the CGA and less emphasis on the stronger bands at 1 and 5 in the CGA profile. The three primary regions defined by the grid search correlated well with those of the forward selection and CGA. Additionally, sub-optimal regions (not shown) fell in line with the other primary peaks of the CGA. The single region map (E) provides information regarding the classification score if only a single region is used. It is interpreted by reading the start of a region on the x-axis and the end of a region on the y-axis. The point at the bottom left indicates the performance of using the region 1780-980 cm^{-1} whereas points along the diagonal consist of single feature regions (ex. 1540-1540 cm^{-1}, a point at a single wavenumber). In Figure 5.8, (D) represents a slice of the greater single region map. The small region map (D) clarifies an example of how the performance of regions consisting of 1-3 features parallels the weights defined by the CGA (bands 1, 2, and 3). It can also be seen that some regions defined as relevant to the forward selection and three-region grid-greedy algorithms show little relevance when evaluated individually (bands 4 and 5).



Classification Score								
	A	B	C	D-1	D-2	D-3	D-4	D-5
Training	0.861	0.89	0.867	0.7	0.767	0.856	0.511	0.517
Test	0.867	0.861	0.844	0.661	.0656	0.861	0.533	0.567

Figure 5.8: Comparison of CGA with forward selection and a 3 region grid-greedy search for the separation of Campylobacter coli and jejuni (CCJ). It can be seen that the primary selected regions from all three methods overlap. In addition, many of the smaller regions selected by the forward search are peaks in the CGA. From the single region map of small regions (D), it is interesting to note that features that do not perform well individually may still be relevant when used in combination with other features (such as peak 5). The classification score was computed using the LOOCV-KNN on the training data and a complete KNN on the test data.

As indicated by the table in figure 5.8, the forward selection algorithm provided the best result on the training set, followed by the grid-greedy search and finally the CGA. However, it is interesting to note that of the three techniques, only the CGA

yielded an improvement in classification score on the test set. The reason for which can be attributed to reduced overfitting of the data. Using a weighted feature set, as is provided by the CGA approach, can allow for partial contribution of features towards the classification. Whereas, more conventional techniques either completely include or exclude features which in certain circumstances can make it difficult to establish the boundaries between overfitting and underfitting. Similar results were found on a variety of datasets where the CGA typically demonstrated less of a tendency to overfit the training data.

5.3.4 Interpretation

Thus far it has been shown that the relevancy plot from the output of a CGA correlates well with the outputs from other feature selection routines, however, it is of interest to see how the relevancy can provide information regarding physical attributes of the sample that could not otherwise be obtained through more conventional means. Generally speaking, the bands of the CGA relevancy plot can be categorized as follows:

A. Bands near the maximum % contribution are highly relevant and are the primary bands used as part of the discrimination of the classes.
B. Bands that have a % contribution below the maximum and the maximum over c (c equal to the number of groups) infer a level of partial relevance.
C. Small bands near zero are representative of irrelevant features. Features that are only mildly detrimental or beneficial to the classification score.
D. Bands that are at zero represent features that are detrimental to the classification.
E. A final class of bands represent the portions of the spectrum where there is little absorbance among all spectra in the dataset.

Bands that are part of category (A) correlate well with the bands outputted using conventional feature selection routines. While bands in category B are not the primary separating features they are still relevant to the separation of the classes. These are bands which on their own may not result in a high classification score but will improve

the classification score when used in conjunction with the primary bands in category A (ex. Figure 5.8 band 4). In other words, they offer information beyond what is provided by the bands in category A. Bands in category C most likely represent portions of the spectrum where the classes are highly similar. As such, their contribution to the Euclidean distance measure used in the classification score would be very small resulting in a degree of indifference if they are included with better performing bands. However, on their own these bands would perform very poorly resulting in a small contribution value. Bands in category D represent those which negatively affect the classification score. These are portions of the spectrum where within-class differences are on a similar scale as between-class differences. Similar to bands in category C, bands in category E can occur when all spectra have near zero intensity and as such their contribution to the Euclidean distance is insignificant as compared to most other features. These bands can be easily disregarded by a quick examination of the reference spectra.

There have been a number of studies [79, 80] that have assigned specific infrared vibrations to biochemical constituents of the bacteria cell. However, having confidence in such assignments typically requires additional confirmatory approaches which lie outside the scope of this study. In addition, these assignments have typically been made along absorbance peaks, but as can be noted by the relevancy plots, often it is in shoulders, valleys, and plateaus of the raw spectrum that relevant features exist. The relevancy plots can be interpreted on a general level by examining combinations of bands in order to infer information regarding the nature of the cell components which are contributing to the separation of the bacteria. In order to associate differences to a particular component it is expected that all known bands of the component be present. It is not expected that the relevancy of these bands be identical as they are often superimposed by a multitude of other components. Table 5.3 indicates several cellular components and their expected positions in the spectrum.

Figure 5.9 illustrates the relevancy plots that have been generated for each of the datasets. From these plots, several aspects concerning the underlying components contributing to the separation of each of the classes can be noted.

Table 5.3: General regions of the spectrum known to contain spectral information pertaining to the primary cellular components. [79, 80]

Cellular component	Spectral regions
Lipids	1750-1730cm^{-1}, a variety of bands scattered between 1480-980 cm^{-1}
Polysaccharides	Many bands in the region 1200-1000 cm^{-1}
Proteins	Strong Amide I and II bands between 1700-1480 cm^{-1}, weak Amide III bands in the region of 1300-1230 cm^{-1}
DNA	Bands between 1260-1060 cm^{-1}

The bands located around 1110, 1300, 1470 and 1750 cm^{-1} of the relevancy plot separating Escherichia coli from Salmonella (ES) suggest that the differences in lipid content or fatty acid profile play an important role in differentiating the two genera. This is not surprising considering that it has been known for a long time that bacteria can be separated based on the analysis of their lipid profile by gas chromatography [81]. On the other hand, the primary factor in separating the three species of Shigella (SHG) is clearly related to differences in the polysaccharides seen by the large contribution of the bands between 1200-980 cm^{-1}. In addition, there are three bands located in the region of 1340-1230 cm^{-1} suggesting the contribution of other cellular components to their separation. This region contains spectral features relating to the Amide III protein bands but the lack of contribution from both the Amide I and II implies that these bands are not representative of protein. Shigella are known to be separated by biochemical tests and serology of their lipopolysaccharides [82]. The three bands from 1340-1230 cm^{-1} (which may be peaks relating to lipids) in conjunction with the stronger bands in the 1200-1000 cm^{-1} region may be indicative of this. The separation of Campylobacter coli and jejuni (see Figure 5.8) appears to be primarily a result of differences in carbohydrates as has been noted in other studies [83]. From the relevancy plot for the separation of Gram positive and Gram negative it is apparent that nearly all cellular components are contributing factors. Considering the diversity of the GPN dataset, this result is not unexpected.

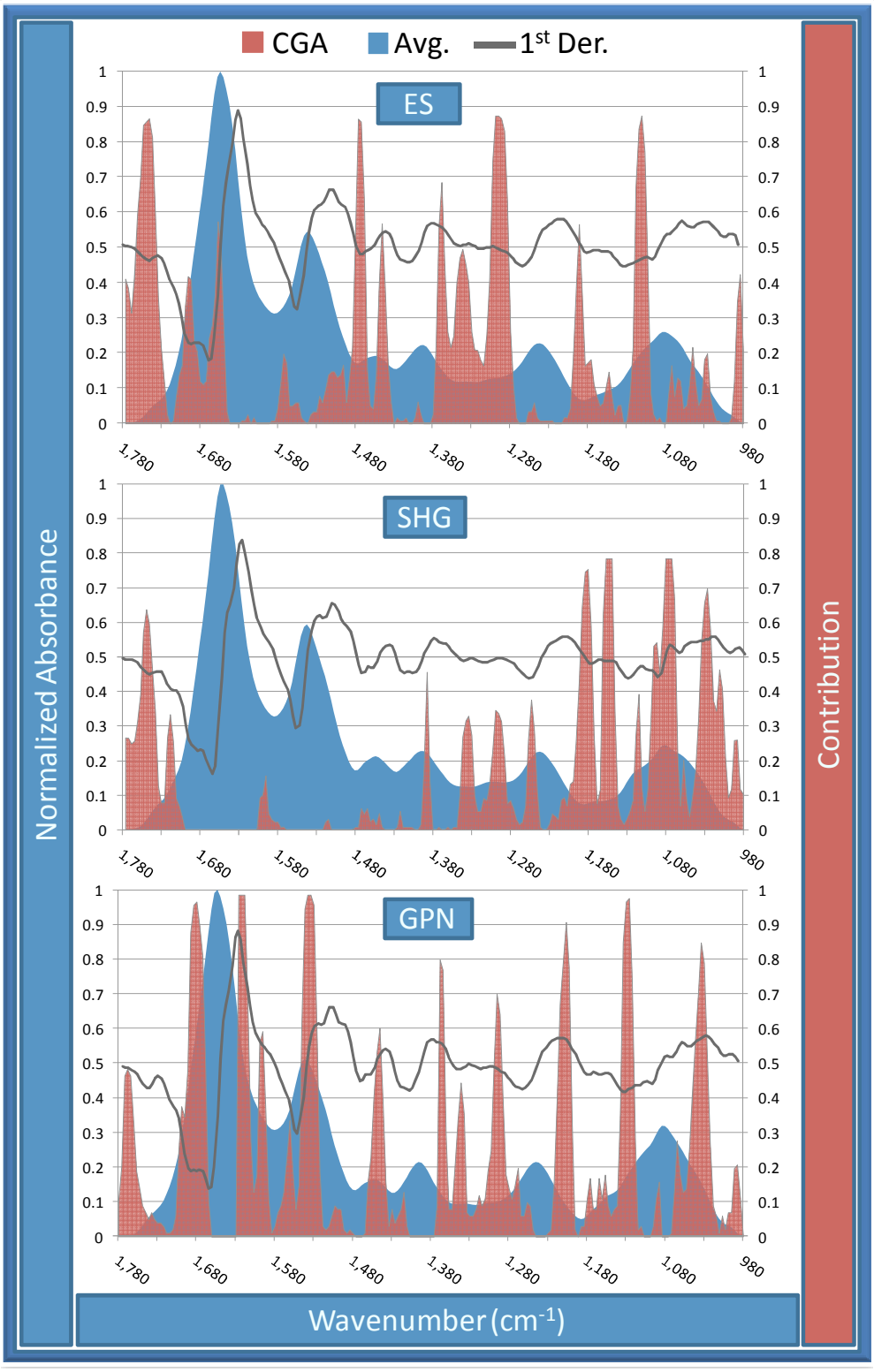


Figure 5.9: Relevancy plots (red) compared to the average spectrum (blue) and the 1st derivative of the average (scale not shown) of the ES, SHG, and GPN datasets.

CGA's have been shown to correlate well with other feature selection techniques and feature relevancy methods. There are a number of scenarios in which applying this approach can offer a benefit to the development and use of a classification model. In particular, the use of a weighted feature set provides more flexibility in defining a classification model which is less prone to overfitting or underfitting. However the advantage in this regard comes at a computational cost over feature selection methods such as a forward search. This makes the CGA approach only suitable in situations where models are not routinely updated. In these situations, the time required to compute the CGA output would make frequent changes to the database infeasible. In situations where time is not a primary concern, such as an infrequently changing commercial database or if adequate computational resources are available, CGA can indeed offer a benefit as a feature selection routine.

With regard to feature relevancy, although the CGA approach to defining feature relevancy is much more time consuming, it was found to provide visually appealing easily interpretable results as compared to other methods. In addition, any changes in the identified feature relevancy that occur as a result of adjustments of the initialization parameters are easily interpretable in the context of their value. This is unlike techniques such as PLS which can produce difficult to interpret outputs which often have changes that are not easily understood as a function of their defining parameters. For example, slight changes in the number of latent variables can result in changes to the underlying pure components. The nature of these changes may not be easily understood.

The smooth profile of the relevancy peaks highlights the fact that neighboring features are correlated and suggests that the use of a weighted feature set may provide advantages in developing robust classifiers.

5.4 CONCLUSION

This study has introduced the CGA approach for evaluating feature relevancy for the separation of spectra into their assigned groups. This approach can potentially offer a greater degree of information than inclusion/exclusion methods regarding the chemical components responsible for the target separation, particularly in situations where the separations are incomplete. As a result, applications of this technique may allow for interpretations which may otherwise not be possible. In addition, CGA's have been shown to be a viable (given sufficient computational time) method for determining a weighted feature set that can improve the reliability of classifiers by reducing the potential for overfitting.

CONNECTING STATEMENT

Previous work focused on the development of classifiers for bacteria identification where it is assumed and required that the bacteria samples are from a pure culture. Proper sample preparation generally ensures this is the case, but nonetheless the possibility of mixed cultures exists. The following chapter examines whether it is possible to take advantage of the spatial resolution and multi-channel nature of FPA-FITR spectrometers to identify images which may originate from a mixed-culture sample. Applying the techniques introduced in this chapter helps to build a robust expert system capable of dealing with the various scenarios that it may encounter in real-world use.

Chapter 6: Identifying Hyperspectral Images from a Heterogeneous Sample

6.1 INTRODUCTION

Research has shown focal-plane-array Fourier transform infrared (FPA-FTIR) spectroscopy to be a reliable and robust method for the rapid identification of bacteria. In conjunction with pixel filtration and data selection techniques, databases can be built and used for the identification of unknowns [23]. As is the case with the majority of high-throughput methods currently used for bacteria identification [84, 85], the protocol for identifying bacteria using FPA-FTIR spectroscopy necessitates that the bacteria under examination consists of a pure culture. Most methods of bacteria identification do not possess mechanisms to verify the possibility that the unknown sample consists of a mixed culture of bacteria nor the ability to identify the bacteria appropriately in the case of a mixed culture. Despite protocols to ensure that all analysis is conducted with a pure culture of bacteria there are multiple ways in which this may not be the case. This study will examine several of these possible scenarios and investigate techniques by which to identify that the infrared images acquired originate from a mixed culture. These scenarios include manual accidental mixing by either misplacement of the bacteria deposits or re-use of the sampling probe in addition to mixed cultures as a result of growing two types of bacteria on a single agar plate.

The primary focus of this study is to determine whether or not FPA-FTIR spectroscopy can identify if a deposit under analysis consists of a mixed culture. It is not the intention of this research to develop universal methods for identifying the constituents that make up all types of mixtures of bacteria. Although, it will be shown that in the simple two-mixture model system used in this study, FPA-FTIR spectroscopy can indeed identify the two bacteria in the mixed deposits. The challenges to creating classifiers capable of identifying deposits consisting of an unknown number and type of

bacteria will be discussed below and methods to try and overcome such difficulties will be proposed.

In theory, a spectrum acquired from a sample containing multiple types of bacteria consists of a linear combination of the spectra from the pure bacteria. Therefore, it could be possible to develop classifiers which can detect and identify multiple bacteria types simultaneously. In addition, by employing techniques such as partial least squares (PLS) regression it should be possible to identify the concentration of each bacteria type in a mixed sample. However, in practice this capability becomes infeasible as the number of types of bacteria increases. In particular, this process depends on defining bacteria types individually and not as groups (ie. genus, species) in order to maintain the linear relationship between spectral response and concentration. Due to the complex biochemical nature of bacteria, it is unlikely that successful PLS calibrations (that are linearly proportional to concentration) could be constructed for a taxonomic group (such as *Listeria*) as there is most definitely no spectral "pure component" which is representative of all species of *Listeria*. This is further exasperated by the fact that the phylogenetic labels given to bacteria are not defined based on spectral differences but as a result of other phenotypic traits which are not necessarily proportional to spectral differences.

The problem, however, can be simplified by not requiring the evaluation of concentration but strictly the presence of the bacteria in a sample. In such a circumstance alternative means of classification could be used, such as artificial neural networks (ANN) or k-nearest neighbors (k-NN). For example, a classification by ANN or k-NN which predicts with equivalent confidence multiple types of bacteria may suggest that the sample is not a pure culture. However, a difficulty is encountered when considering the evaluation of bacteria in this manner (and in the case of PLS calibrations but to a lesser degree). The equal mixture of two bacteria results in a spectrum that is "half-way" in between the original spectra of the bacteria. As such, it is difficult to ascertain whether or not such a spectrum is truly the result of a mixture or is in fact a new type of bacteria altogether. Processes can be used by which the classification is

broken down into a hierarchical set of identifications (Gram type, genus, species...) which could allow for the use of more traditional classification models at the more general levels of taxonomic specification and the PLS classifier at the most specific levels to assess sample purity. This would minimize the number of components in the PLS model, but difficulties would still arise if the sample consisted of bacteria of highly different phylogeny.

This study will attempt to circumvent the complexities of building and relying on such models by suggesting approaches which exploit the multi-channel nature and high spatial resolution of FPA-FTIR detectors to assess the spectral uniformity of a sample. FPA-FTIR spectrometers are capable of generating thousands of spectra from a field-of-view of 188 x 188 μm^2 and a nominal resolution of 5.6 µm. In this manner, chemical variability can be evaluated on the micron scale within a single image or on the mm scale by comparing multiple images. The expectation is that there will be chemical variability, either within-image or between-image, that would be encountered with a mixed deposit. This study attempts to identify the presence of a mixed culture by combining information regarding the measured spectral variability on several different scales (Table 6.1).

Table 6.1: The three different measures used to evaluate the variability of a sample.

Measure	Abv.	Description
Within-image variability	WIV	The average euclidean distance of all pixels in an image to the mean spectrum.
Between-image variability	BIV	The average Euclidean distance of all combination of pairs of mean spectra from each image of a deposit. For example, for three images the BIV is the average of three distances (1 – 2, 1 – 3, 2 – 3).
Between-deposit variability	BDV	The average Euclidean distance of all combination of pairs of mean spectra from each image of multiple deposits.

The first measure (referred to as WIV, within-image variability) exploits the high spatial resolution of FPA-FTIR spectroscopy with the expectation that at this scale an infrared image acquired from deposits that have been mixed (using the techniques

described below) may exhibit a greater degree of within-image chemical variability than those images which originate from a pure culture of bacteria. This is accomplished by examining the difference between each pixel and the mean spectrum of the image. The average and standard deviation of the Euclidean spectral distances to the mean spectrum of a mixed deposit are compared to the values that are obtained on images from pure deposits. However, it may be that there are only slight changes in the relative proportions of the two bacteria from pixel-to-pixel and so no indication from the within-image variability that the sample or deposit is mixed. The second measure (referred to as BIV, between-image variability) compares the differences in spectra from replicate images of a single deposit. Ideally, the protocol used for identifying bacteria involves three deposits for a given sample, and three images taken from each deposit (9 images total). In this manner the variability across the area of a deposit can be evaluated. The BIV takes the average distance between the mean spectrum of each image. In other words, for three images (1, 2, 3) the average Euclidean distance between 1 – 2, 1 – 3, and 2 – 3 is taken. This measure can be extended to examining the variability between deposits (referred to as BDV, between-deposit variability). The evaluation is done similarly to the BIV measure by comparing the mean spectrum of the images. The values provided by the WIV, BIV, and BDV measures are compared to those that are determined on a reference set of image data which is known to be from pure cultures. Measured variability that is significantly greater (> than 3 standard deviations) than what is typically found from spectral images of pure deposits would be indicative of increased chemical variability between replicate images. It does not guarantee that the deposit is from a mixed culture, only that there is spectral variability between replicates, one such explanation of which is that the deposit is mixed. The advantage of the above measures is that they do not rely on the development of classifiers and instead depend on statistically measured variation of mixed vs. non-mixed deposits. This can significantly simplify the approach of measuring sample purity as compared to trying to interpret misclassifications from classifiers which are dependent on suitable and sufficient reference data. However, it is clear that if the concentrations are too heavily

biased or spectral differences between the two types of bacteria are very small then these techniques will be limited in their capacity to identify that a deposit is mixed.

There are two primary factors that dictate the degree to which a mixed deposit will differentiate itself spectrally compared to a deposit from a pure culture. The first is the extent to which the two bacteria are mixed. If the mixture is heavily biased with one type of bacteria having a much higher concentration than the other than this will increase the difficulty in identifying that the culture is mixed. The second is dependent on the degree to which the spectra of the bacteria differ. For example, the spectra from bacteria which are from two different genera will typically differ to a greater extent than two types of bacteria from the same genera. The limit-of-detection in both these cases is proportional to the variability (both instrument and sample variability) that is encountered when acquiring spectra from deposits of bacteria of this type. Pixel filtration techniques (Chapter 4) are employed in this study to minimize the sample variability and to only use pixels which meet pre-defined spectral quality characteristics. When examining individual pixels, and not the average of multiple pixels, the signal-to-noise ratio of individual pixels (found to average around ~25 on the first derivative) is typically the limiting factor.

In order to assess the effectiveness of measuring the spectral variability to determine whether or not a deposit is mixed, the results are compared to the concentrations determined from a partial least squares (PLS) model. Due to the deposition process used it is impossible to know the true proportions of each type of bacteria in a deposit. As a result of the low SNR of individual pixels the values determined will not be highly precise but can still provide clear indications of the relative proportions of each of the bacteria constituents. The concentrations determined will be correlated to the values found for each of the measures (WIV, BIV, BDV).

6.2 MATERIAL AND METHODS

6.2.1 Sample preparation

Four types of bacteria including, *Staphylococcus aureus, Salmonella entrica typhimurium, Shigella flexneri,* and *Escherichia coli* were used in order to provide examples of samples with varying degrees of taxonomic separation. All bacteria were prepared and deposited at Health Canada (Longeuil, Quebec, Canada) using standard methods and protocols. After incubation all samples were deposited as outlined below.

6.2.2 Sample deposition

In order to recreate a number of different scenarios of how a deposit might become mixed four different deposition methods were used (Table 6.2). The first consisted of two types of bacteria that were deposited with a partial overlap. One type of bacteria was initially picked off the agar plate and deposited, the sampling probe discarded, and then a second type of bacteria picked from the agar plate and deposited adjacent to the first with a partial overlap. The majority of the two adjacent deposits consist of a pure culture with mixing only occurring around interface. The second approach is similar to the first except that the second sample is deposited directly on top of the first. This would be analogous to the technician accidentally depositing two different samples on top of each other. The third approach involved picking two different bacteria, one agar plate at a time and with the same sampling probe, and then depositing. This is analogous to accidentally swabbing a second bacteria without having deposited the first sample. The final approach involved the growth of two types of bacteria on the same agar plate. The bacteria were then indiscriminately picked up and deposited on the zinc selenide substrate. All deposits using the aforementioned approaches consisted of mixtures of two types of bacteria. Due to the samples being manually deposited the true concentrations are unknown. Figure 6.1 illustrates the layout of the four types of bacteria deposited samples on the zinc selenide infrared transparent slide.

Table 6.2: Types of depositions conducted to simulate different ways by which a deposit may consist of more than one type of bacteria.

Deposit Type	Abv.	Description
Partial Overlap	PO	Two types of bacteria grown separately then deposited one at a time. The second deposit is placed next to the first with a partial overlap.
Full Overlap	FO	Two types of bacteria grown separately then deposited one at a time. The second deposit is placed directly on top of the first.
Mix and Deposit	MD	Two types of bacteria grown separately. The first bacteria is picked up from the agar, then with the identical microbiological loop the second is picked up. The two are then deposited simultaneously on the infrared transparent slide.
Grown Together	GT	Two types of bacteria grown on a single agar plate. The bacteria are picked up indiscriminately from the agar and deposited as would be done with an agar plate with a single type of bacteria.

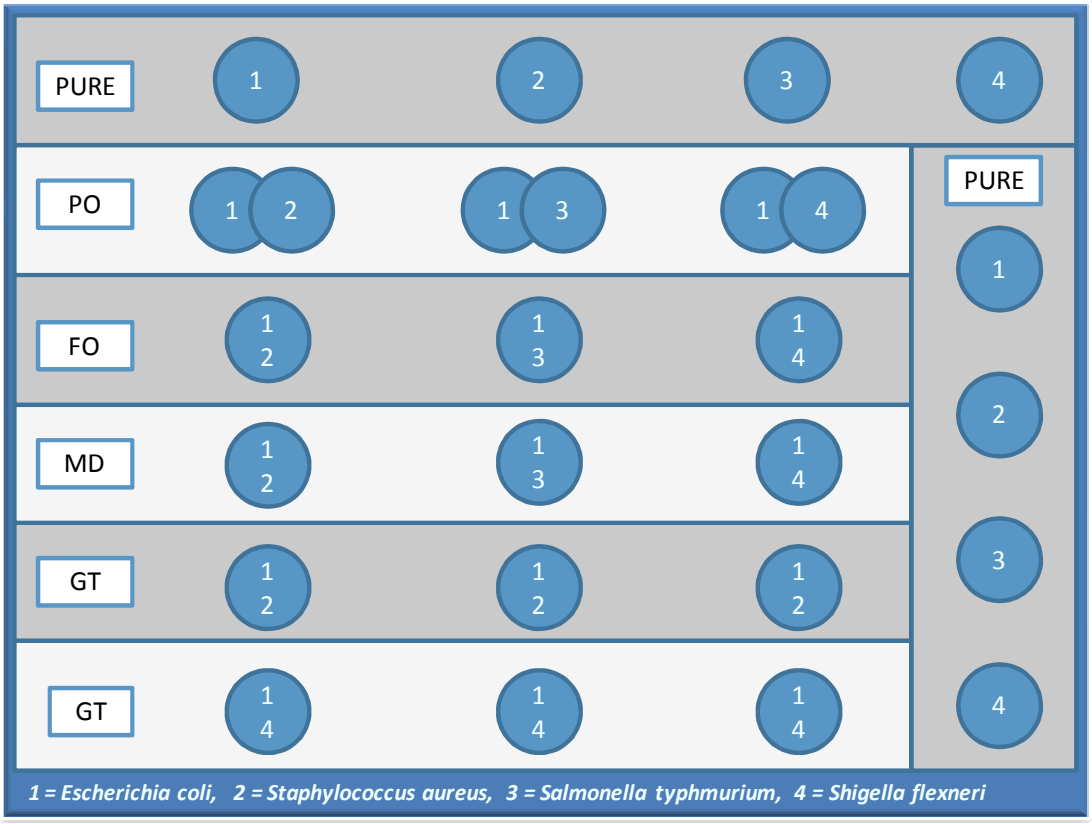


Figure 6.1: Illustration of the physical layout of the deposited samples on the zinc selenide infrared transparent slide. Deposits from a pure culture were deposited along the top and right-hand periphery. There is one row of samples each with partially overlapping (PO) and fully overlapping (FO) deposits, and one row of samples that were initially mixed then deposited (MD). There are two rows of samples consisting of samples that were grown together (GT).

The layout of the deposited samples on the zinc selenide slide consisted of two deposits of a pure culture for each type of bacteria (Figure 6.1 top and side). For the PO, FO, and MD deposit, Escherichia coli was mixed with each of the other three bacteria. In addition, there were three deposits of Escherichia coli and Staphylococcus aureus grown together (GT) and Escherichia coli and Shigella flexneri grown together (GT). Considering that the PO, FO, and MD depositions consist of errors during the deposition process, analysis of BDV does not apply in this context. On the other hand, the GT deposits consist of multiple direct depositions from a single agar, and so in this scenario one can verify spectral differences between deposits

6.2.3 Spectral acquisition

The spectral images were collected on a Varian Excalibur FTIR spectrometer operating under Varian Resolutions Pro 4.0 software (Varian, Melbourne, AU) and equipped with a UMA-600 infrared microscope and a 32 X 32 (1024 pixels) mercury-cadmium-telluride focal-plane-array detector. A constant flow of dry air was used to purge the spectrometer and the microscope, limiting spectral contributions from carbon dioxide and atmospheric water vapor. Each spectral image was collected from a field of view of 188 X 188 μm^2 with a spatial resolution of 5.6 µm and a spectral resolution of 8 cm^{-1}. Each image consisted of 256 co-added scans which required approximately 5 minutes to acquire.

6.2.4 Data analysis

Due to the pathlength non-uniformity of the sample deposits and to insure that spectra meet a minimum signal-to-noise ratio requirement, all images were initially filtered based on the appropriate spectral criteria (see Chapter 4). The pixel filtration process is composed of the following filters and transformations: An absorbance filter which eliminates any spectra whose average raw absorbance on the wavelength region 1480 – 980 cm^{-1} were outside the range of 0.1 – 0.9, calculation of spectral first derivatives using the first difference method, a noise filter to eliminate spectra with a 1st

derivative signal-to-noise ratio outside of 25 which is measured by taking the ratio of the peak-to-peak absorbance from 2200 – 2000 cm^{-1} to the peak-to-peak absorbance from 1480 – 980 cm^{-1}, and vector normalization of the data on the region from 1480 – 980 cm^{-1}. The WIV, BIV, and BDV measures were processed with the data remaining after the previous 4 stages as many of the images acquired from the mixed deposits examined in this study would be completely eliminated by the two-phase similarity filter that is applied subsequently. While this demonstrates the merit of the filters in ensuring spectral homogeneity it is not the focus of the study. Data used for building the PLS calibrations were further processed through the final stage of pixel filtration, the two phase spectral similarity filter. This filter is used to evaluate the pixel-to-pixel variability of the image (similarly to the WIV measure) and to reject pixels which appear to be outliers. All spectral distances were calculated using the Euclidean distance metric [34].

6.2.5 Partial least squares (PLS) regression

Partial least squares (PLS) regression [77] is a well known technique for building "whole spectrum" linear calibrations. It offers greater resolving capabilities in defining spectral pure components as compared to a classical least squares technique where there may be multiple un-accounted for contributions to the spectra. A PLS model of the bacteria data was constructed by assigning a component for each of the four types of bacteria under analysis. A particular spectrum of bacteria would be assigned (a concentration value of) 0 for all components except for the one it belonged to where it would receive a value of 1. The PLS models were built on the region 1480-980 cm^{-1} using a K-fold cross-validation technique (K = 3) and the number of latent variables (factors) used for the calibration was defined by picking the minimum in the cumulative (from the K = 3 groups) predicted residual sum of squares (PRESS) plot. The calibration developed was then used to predict the individual pixels of images acquired from deposits which consist of two types of bacteria. The data used to generate the model consisted of six images (three per deposit) for each of the four types of bacteria. All data were pre-filtered as described above and then spectral averages computed consisting of 64 pixels

randomly averaged together (i.e. not necessarily spatially adjacent pixels). For simplicity of the implementation, excess pixels insufficient in number (< 64) were discarded.

6.2.6 Data processing

All data processing as part of this study was accomplished on software developed in-house on the Visual Studio .Net 2005/2008 platform (Microsoft, Redmond, CA). The processing routines were validated (where possible) by visual and computational comparison of outputs from Matlab 6.5 (Mathworks, Natick, MA) and Omnic 6/7 (Thermo Nicolet, Madison, WI). The processing was completed on a quad core Intel (Santa Clara, California) processor operating at 3.2 gigahertz with 4 gigabytes of random-access-memory running Windows Vista Business edition. The software developed as part of this study was optimized for parallel processing on multi-core central processing units (CPUs) allowing for significantly greater computational output than most other software available at the time.

6.3 RESULTS AND DISCUSSION

6.3.1 PLS classifier

Before attempting to classify and predict concentrations of mixtures of bacteria it is necessary to ensure that a calibration can be successfully built that can separate the reference samples. Figure 6.2 illustrates the calibrations that have been made with the deposits from a pure culture of each of the four bacteria.

The calibration standard deviation ranged from 0.2% to 0.9% for the four calibrations. The calibrations were built with averaged spectra, each consisting of 64 pixels. It should be considered that further results are computed on individual pixels, so it is expected that the variability due to predictions made on individual pixels will be higher. This is not of significant concern, though, as even a moderately precise measure of concentration would be sufficient to evaluate the degree of mixture.

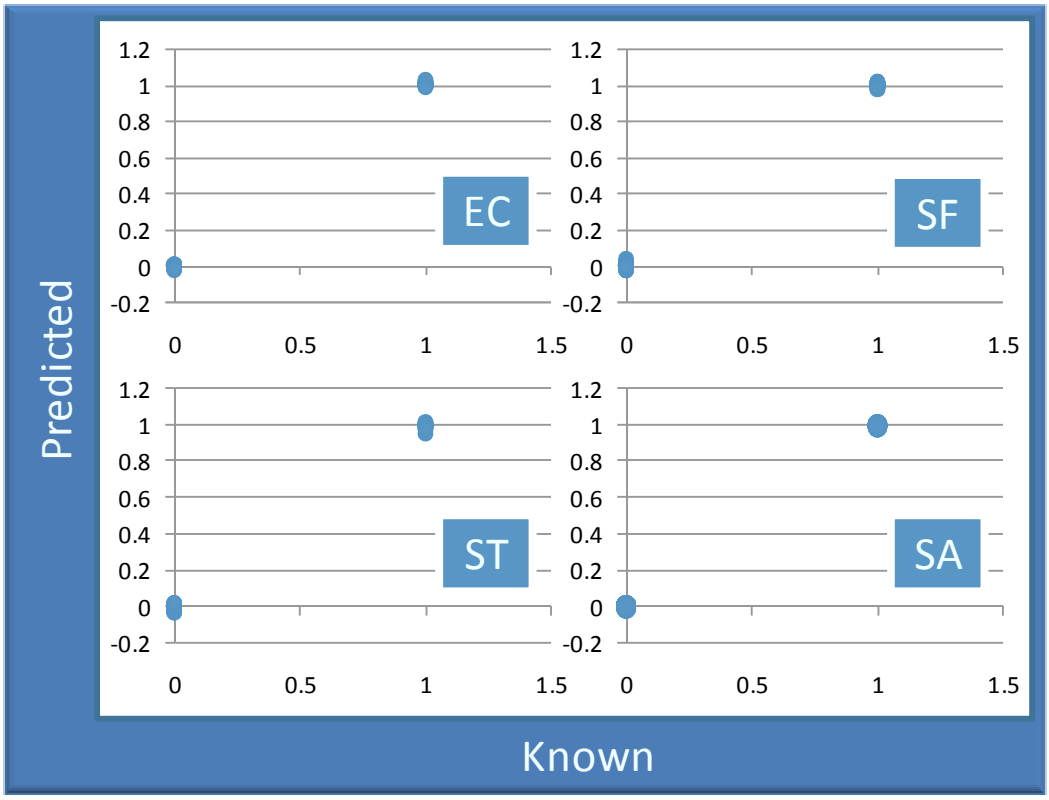


Figure 6.2: The PLS calibrations for each of the four components corresponding to each type of bacteria are shown. *EC = Escherichia coli, SA = Staphylococcus aureus, ST = Salmonella typhmurium, SF = Shigella flexneri*. Results shown are the predictions for each of the K-fold (K=3) validation groups. The standard deviations of the calibrations were less than 1 %.

The average WIV, BIV, and BDV measures for all spectral data acquired from pure cultures of bacteria are listed in Table 6.3. These values provide a baseline by which to compare the results found from the spectral data acquired from the mixed deposits. Measured values which are found to be 3 standard deviations or more outside of the values listed in Table 6.3 are indicative of spectral variability which cannot be accounted for.

Table 6.3: Average and standard deviation of the WIV, BIV, and BDV measures determined for pure deposits of each of the four types of bacteria used in this study. . EC = Escherichia coli, SA = Staphylococcus aureus, ST = Salmonella typhmurium, SF = Shigella flexneri.

Bacteria	Deposit Type	Pixels Rem.	WIV	BIV	BDV	
EC	PURE	823	5.2 ± 0.9	5.5 ± 0.9	5.2	2
SA	PURE	929	5.2 ± 1.1	5.4 ± 0.4	6.6	2
ST	PURE	902	5.1 ± 1.2	5.3 ± 0.5	6.8	2
SF	PURE	871	5.0 ± 1.1	5.2 ± 0.8	4.6	2

6.3.2 Samples deposited with a partial overlap

Samples that were deposited with only a partial overlap are discussed separately as unlike the other types of deposits discussed below; these have only a small area in which there is actual mixing of the two bacteria types. It was found that these types of deposits only showed mixing very close to the interface where the deposits overlapped. This is illustrated in Figure (6.3) where the pure bacteria are clearly seen (blue and red) at either sides of the interface (purple).

Within-image variability of images such as these is very high, whereas images taken at areas where the interface is outside the field-of-view of the detector had values similar to pure bacteria. The identification from multiple images taken across the area of a deposit such as this would yield a very high BIV value which would indicate to the operator that an error occurred during the deposition process. To our knowledge this is the first example of resolving multiple bacteria in a single infrared image.

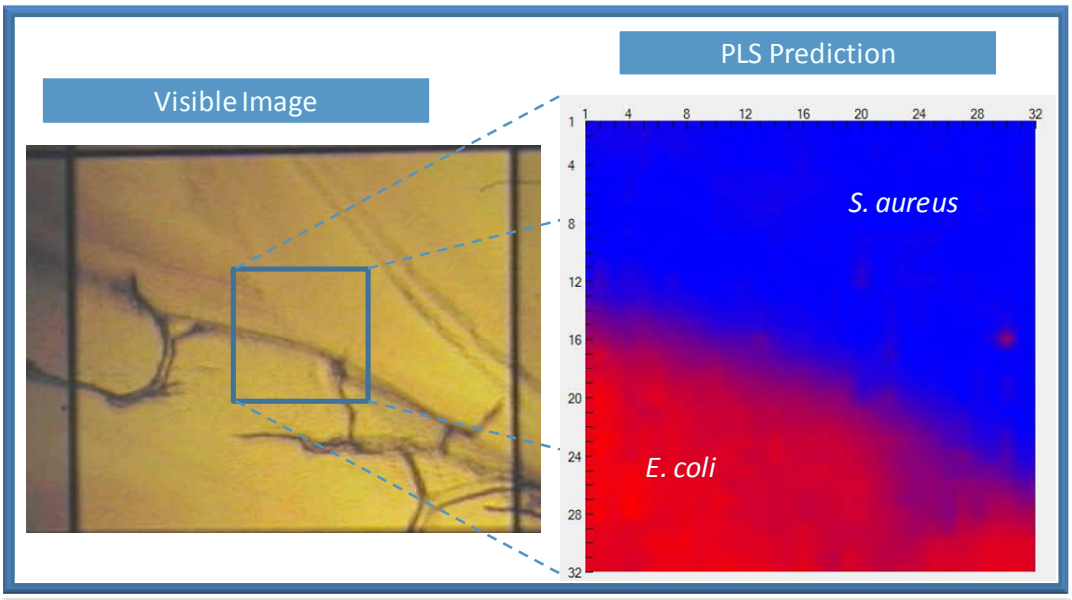


Figure 6.3: Example of the PLS prediction of pixels in an image acquired at the interface between two types of bacteria. Red indicates Escherichia coli, blue Staphylococcus aureus, whereas purple represents a 50/50 mixture.

6.3.3 Accidentally mixed deposits

The ability to detect mixed samples simply by examining spectral variation is assessed by comparing within-image and between-image variation to the concentrations determined by the PLS model. Table 6.4 displays the results for the WIV and BIV measures in conjunction with the concentrations determined by the PLS model for samples deposited of the type FO and MD. It can be seen that in cases where the deposited mixture is relatively uniform, as indicated by the PLS concentrations, the within-image and between-image variation does not give indication of the mixed nature of the deposit (Table 6.4 green highlight). In addition, it can be noted that higher standard deviations of the measured concentrations results in a higher within-image variation such as with the MD deposit of EC and ST (Table 6.4 orange highlight). On the other hand, between-image variance is minimally affected if the measured concentrations are similar. For example, despite having elevated within-image variation, the three images of the MD deposit of EC and ST have similar average concentrations, hence; the between-image variation is only slightly elevated. If the concentration changes highly across the area of the deposit then increased levels of variation can be

noted in both the within-image variation and the between-image variation (Table 6.4 red highlight).

Table 6.4: Results of the WIV and BIV measures for the FO and MD deposits, and the concentrations measured by the PLS mode. Areas highlighted in green indicate variability comparable to those found with the pure deposits. Orange indicates elevated variability, while the red indicates extreme variability as compared to the pure deposits.

Image	Deposit Type	Pixels Rem.	WIV	BIV	% E. coli	% S. flexneri	% S. typhmurium	% S. aureus	% Sum
EC & SA	1	FO	897	5.2 ± 1.6		34 ± 3	1 ± 2	0 ± 3	61 ± 2	96
EC & SA	2	FO	892	6.6 ± 1.6	5.8 ± 0.6	33 ± 3	3 ± 2	0 ± 1	65 ± 2	101
EC & SA	3	FO	875	6.8 ± 2.3		36 ± 3	3 ± 1	0 ± 3	62 ± 2	101
EC & ST	1	FO	849	14.6 ± 3.3		45 ± 16	2 ± 3	53 ± 14	-2 ± 2	99
EC & ST	2	FO	892	11.4 ± 2.8	18.1 ± 1.7	73 ± 4	3 ± 2	22 ± 7	-1 ± 1	98
EC & ST	3	FO	829	14.2 ± 3.0		60 ± 15	2 ± 4	35 ± 20	2 ± 1	100
EC & SF	1	FO	947	12.8 ± 3.2		39 ± 10	65 ± 15	-3 ± 2	0 ± 1	101
EC & SF	2	FO	875	8.1 ± 1.5	15.3 ± 2.4	88 ± 5	13 ± 8	1 ± 4	-1 ± 2	101
EC & SF	3	FO	909	9.7 ± 1.9		42 ± 7	56 ± 9	1 ± 1	-1 ± 2	98
EC & SA	1	MD	833	7.9 ± 2.3		45 ± 3	2 ± 2	0 ± 3	49 ± 2	97
EC & SA	2	MD	925	6.5 ± 1.5	6.3 ± 0.6	46 ± 3	3 ± 2	2 ± 2	48 ± 2	100
EC & SA	3	MD	839	6.6 ± 1.6		46 ± 3	1 ± 2	3 ± 3	47 ± 2	98
EC & ST	1	MD	793	10.2 ± 2.6		31 ± 5	-1 ± 1	72 ± 9	2 ± 3	104
EC & ST	2	MD	859	13.5 ± 4.0	7.1 ± 0.9	43 ± 9	3 ± 4	61 ± 13	-1 ± 4	106
EC & ST	3	MD	912	10.3 ± 2.3		42 ± 6	3 ± 3	59 ± 6	0 ± 4	104
EC & SF	1	MD	846	5.2 ± 1.4		15 ± 3	85 ± 3	-1 ± 1	2 ± 1	101
EC & SF	2	MD	960	5.5 ± 1.7	6.9 ± 0.7	22 ± 1	81 ± 6	-1 ± 2	2 ± 2	104
EC & SF	3	MD	882	6.4 ± 1.7		27 ± 2	72 ± 3	0 ± 1	2 ± 1	101

For those deposits that did not exhibit sufficient variability to be 'flagged' as erroneous (Table 6.4 green highlight) it is highly likely that there will be adequate differences between the other deposits (BDV) to provide indication that the deposit is not representative of the sample under analysis.

6.3.4 Mixed deposits from bacteria that are grown together

Table 6.5 displays the results for the WIV, BIV, and BDV measures in conjunction with the concentrations determined by the PLS model for samples that had two types of bacteria grown on a single agar plate and then deposited (GT). For these types of deposits there is only one indication (deposit 1 of EC and SF) from the WIV and BIV measures that the samples might be mixed (Table 6.5 orange highlight). Otherwise, the remaining images and deposits do not have sufficient variability to indicate that spectra are being acquired from samples with more than one type of bacteria. It is only when examining the between-deposit variation that it becomes apparent that the bacteria from the agar may not be a pure culture (Table 6.5 red highlight). The BDV of the two types of mixtures shown in Table 6.5 have spectral differences between deposits which are substantially greater than those seen with deposits of pure bacteria. This is mostly likely a result of sampling from different portions of the agar, resulting in a different mixture of bacteria for each deposition. This also highlights the benefits of acquiring spectral data from multiple deposits.

Another benefit of evaluating the WIV, BIV, and BDV measures on multiple spectral images from multiple deposits is that it can assist in identifying whether or not an unknown sample is the result of a mixed culture or a new type of bacteria not in the database. A pure culture of a new type of bacteria will exhibit variability levels comparable to what was found with the bacteria making up the reference database. The measures described above (WIV, BIV, BDV) identify increased variation in a sample deposit, or series of deposits, which may be indicative of a sample consisting of more than one type of bacteria. In practice, these types of measures would be used in conjunction with classification methods to better understand the nature of the classification result. For example, classifiers that encounter previously unseen spectral data (such as the infinitely possible combination of spectra from multiple bacteria types) will typically yield unpredictable difficult to interpret results. In this manner, conventional classifiers are greatly benefited when used in conjunction with the BIV, WIV, and BDV measures, due to the simplicity of interpretation which these measures

offer. In addition, the challenge of determining if a sample is in fact a mixed culture or a new type of bacteria not present in the database is facilitated by examining the BDV measure. A normal value of which would most likely imply that the sample is indeed a type of bacteria not present in the reference dataset suggesting further analysis is in order. In addition, these measures can act as an early warning system prior to classification that there may be a problem with the deposit(s) under analysis.

Table 6.5: Results of the WIV, BIV, BDV measures for the GT deposits, and the concentrations measured by the PLS model. Areas that are highlighted in red indicate increased variability as compared to that seen with the pure deposits. For the most part, bacteria grown together do not show increased levels of WIV of BIV compared to that of a pure culture. However, BDV is highly elevated as a result of differing concentrations between deposits.

Deposit	Image	Deposit Type	Pixels Rem.	WIV	BIV	BDV	% E. coli	% S. flexneri	% S. typhmurium	% S. aureus	% Sum
EC & SA	1	1	GT	913	5.2 ± 0.9	5.2 ± 0.2	12.7	9 ± 2	2 ± 1	-1 ± 3	90 ± 3	100
EC & SA		2	GT	920	5.3 ± 1.1			13 ± 3	2 ± 2	-2 ± 4	86 ± 2	99
EC & SA		3	GT	926	5.3 ± 1.1			15 ± 3	-1 ± 2	-1 ± 4	86 ± 2	99
EC & SA	2	1	GT	924	5.0 ± 1.2	5.1 ± 0.4		64 ± 4	0 ± 2	-1 ± 2	41 ± 2	104
EC & SA		2	GT	916	5.0 ± 0.9			61 ± 3	-2 ± 3	0 ± 2	40 ± 2	99
EC & SA		3	GT	929	5.1 ± 0.7			62 ± 3	-1 ± 2	1 ± 2	40 ± 2	102
EC & SA	3	1	GT	804	5.4 ± 1.0	5.7 ± 0.7		44 ± 1	0 ± 3	3 ± 2	48 ± 6	95
EC & SA		2	GT	854	5.2 ± 0.9			46 ± 1	-2 ± 2	3 ± 2	56 ± 7	103
EC & SA		3	GT	880	5.1 ± 0.9			41 ± 1	0 ± 1	2 ± 1	59 ± 6	102
EC & SF	1	1	GT	787	10.4 ± 1.9	10 ± 2.1	11.2	82 ± 7	16 ± 10	2 ± 1	2 ± 4	102
EC & SF		2	GT	903	12.2 ± 3.0			70 ± 8	25 ± 9	1 ± 2	3 ± 4	99
EC & SF		3	GT	859	11.7 ± 2.1			84 ± 8	17 ± 10	0 ± 2	2 ± 4	103
EC & SF	2	1	GT	833	7.1 ± 1.4	6.6 ± 1.8		55 ± 2	46 ± 2	-2 ± 2	2 ± 3	101
EC & SF		2	GT	843	6.2 ± 1.3			60 ± 2	37 ± 1	-2 ± 2	3 ± 3	98
EC & SF		3	GT	872	6.7 ± 1.5			59 ± 2	41 ± 1	-1 ± 2	0 ± 2	99
EC & SF	3	1	GT	874	5.4 ± 1.7	6.4 ± 1.4		67 ± 2	28 ± 3	4 ± 2	-1 ± 3	98
EC & SF		2	GT	866	5.6 ± 1.2			69 ± 1	28 ± 3	2 ± 1	-3 ± 3	96
EC & SF		3	GT	862	5.5 ± 1.2			65 ± 1	29 ± 2	3 ± 3	1 ± 3	98

EC = Escherichia coli, SA = Staphylococcus aureus, ST = Salmonella typhmurium, SF = Shigella flexneri

CONCLUSION

This study introduces several measures for enhancing the ability of FPA-FTIR spectroscopy to detect the possibility of a mixed cultural. By analyzing sample variability on multiple levels, this method is a simple and efficient means of assessing sample homogeneity without requiring a more complex approach using multiple classifiers. Furthermore, these methods used in conjunction with classification methods can allow for an understanding of the composition and source of variability in a sample, whether it is from a mixed culture or a new type of bacteria not present in the reference database, thereby increasing the explanatory capabilities of bacteria identification by FPA-FTIR spectroscopy. Future work in the area of mixed culture analysis along with improvements in sample preparation, computational methods, and instrument technologies may one day remove the requirement that a sample (for prediction) originate from a pure culture for an accurate and reliable analysis.

CONNECTING STATEMENT

The chapters up to this point have focused on individual aspects of the process of identifying bacteria by FPA-FTIR spectroscopy. These include pixel filtration for the extraction of representative data (Chapter 4), a comparison of methods for optimizing classifiers (Chapter 5), and an approach for evaluating the potential of identifying a non-homogenous deposit of bacteria (Chapter 6). The focus of the following chapter is to incorporate all these techniques into a complete methodology for the identification of bacteria by FPA-FTIR spectroscopy. This methodology is implemented in the form of an expert system which can build databases, analyze unknowns, and provide feedback regarding the properties of the samples under analysis.

Chapter 7: Expert System for the Identification of Bacteria by Focal Plane Array Fourier Transform Infrared Spectroscopy

7.1 INTRODUCTION

This chapter describes the methodology that has been developed for creating rapidly searchable and easily interpretable databases for the identification of bacteria from data mined from hyperspectral infrared images. Focal plane array Fourier transform Infrared spectroscopy is an information-rich technology that produces massive amounts of data in a short period of time. As a result, for the target application of this study (bacteria identification), visual examination of FPA-FTIR image data is inefficient, tedious, and, for the most part not even feasible. This necessitated the development of new techniques of data analysis and database management in order for this technology to be accessible for general use. The chapter outlines a methodology by which FPA-FTIR spectroscopy can be effectively and efficiently used for bacteria identification. Emphasis has been placed on developing techniques that are amenable to automation in order to have a complete, practical, and robust model accessible to the intended user. In addition, as a result of this, the development of searchable databases from infrared imaging data is facilitated, thereby increasing the applicability of FPA-FTIR spectroscopy as an analytical technique and thus extending the benefits gained by employing this information-rich spectroscopic method to a broader range of non-expert users. Finally, the overall goal of this work was the development of an expert system, which may be defined as the implementation of a computer program that incorporates the knowledge of an expert (or experts). In this context, an expert system is not only capable of analyzing samples but is able to provide intelligent and easily interpretable feedback to the user regarding the process and result of an analysis.

The infrared images considered in this study consist of a 32 x 32 matrix of spectra, each consisting of 792 points (4000 – 950 cm^{-1} at 8 cm^{-1} resolution). If we consider the physical storage space required strictly for the numerical data (omitting overhead such as titles, etc.), there are 811,008 points totaling 6,488,064 bytes of information per image. If we consider a scenario in which the data for each reference sample consists of triplicate images for each of three replicates, the raw data per reference sample will amount to approximately 50 megabytes (1 megabyte = 10^6 bytes). It is apparent that a modest-sized database would require a substantial amount of memory. Larger databases would require either high-end specialized computers to manage the increased memory requirements or long search times due to the retrieval of data from secondary storage devices (such as hard drives). Even in the best of circumstances, it can be seen that searching this quantity of data would result in long search times, and the often necessary step of spectral pre-processing (normalization, derivitization, etc.) would lead to even longer search times.

The situation is yet further exacerbated if we consider that when searching a database of infrared images for a match to the image of an unknown sample, we are in fact performing many repeated searches. For example, if we consider the 32 x 32 array, it would be necessary, in the worst case, to perform 1024 individual searches. Even after pixel filtering and averaging, tens to hundreds of iterations may still be required. Taking this into account, it is apparent that if a single search of a single spectrum from a database of images requires on the order of seconds, then the complete search of an unknown will take on the order of minutes to hours. The implication is that the infrared spectral data cannot be used as is in the generation of classifiers and thus additional steps such as data filtration, extraction, and averaging must be incorporated in database development to reduce the quantity and improve the quality of data used.

Considering the above, this study addresses several challenges that exist when trying to develop an expert system based on infrared imaging data, in particular:

- How to integrate infrared imaging data into a database in a practical manner.

- How to define a methodology by which a database consisting of infrared images can be created and manipulated in a similar fashion to one consisting of data acquired by conventional FTIR spectroscopy.
- How to make the benefits of information-rich infrared imaging technology accessible to the non-expert user.
- How to find a balance between the benefits of data-redundancy and the adverse effects of non-sample sources of spectral variability, given the FPA detector's pixel-to-pixel variability in terms of both detector response and SNR.

Each of these aspects is addressed by applying the techniques that have been presented in the previous chapters and through the development of custom software. The developed routines attempt to maximize success of the identification process by generating the highest quality data possible. In this context, quality is defined in terms of the extent to which spectral variability is a result of inherent chemical differences between the samples, given that the achievable sensitivity and specificity of the bacteria identification method are limited by non-sample contributions to spectral variability. Minimization of the latter contributions is accomplished through pixel filtration routines, spectral averaging, and feature selection for the optimization of the classifiers. An additional benefit of these operations is the streamlining of data into a more manageable and functional size.

During the development of the protocols introduced in this study a particular emphasis was placed on developing methods that are accessible and transparent. As the intention is for these protocols to be used commercially and by individuals with minimal understanding of infrared spectroscopy, infrared imaging, chemometrics, or data analysis/interpretation, it is necessary that optimization and classification techniques can be applied with minimal user input. In addition, the expert system should be able to provide intelligent feedback to the user regarding issues and results that are encountered during database development or the identification of unknowns.

7.2 METHODOLOGY FOR BACTERIA IDENITIFICATION

7.2.1 Overview

The overall methodology can be broken down into five stages, each of which constitutes a segment of the bacteria identification process. These stages are illustrated in Figure 7.1 and described below.

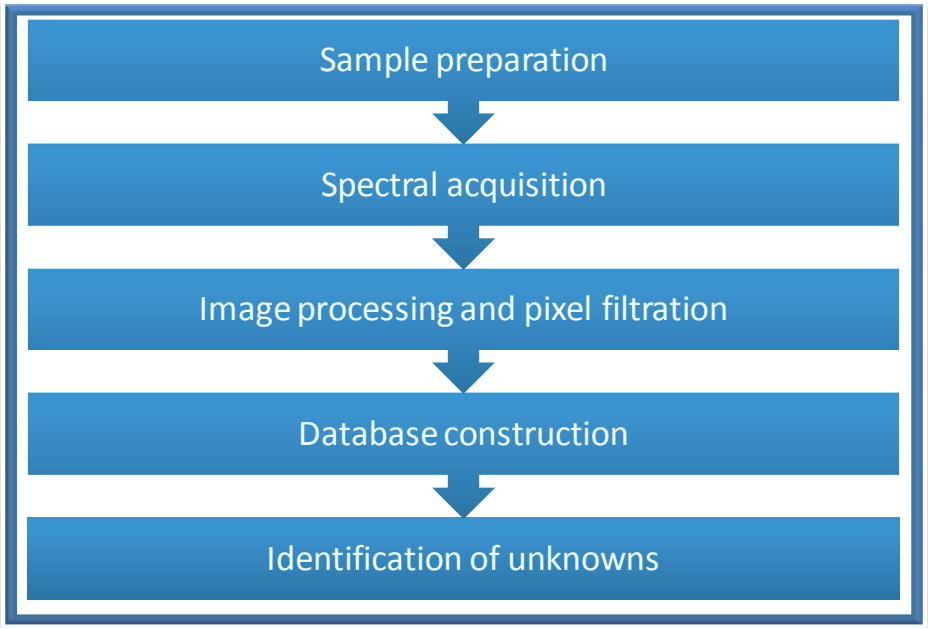


Figure 7.1: The five-stage methodology for bacteria identification by FPA-FTIR spectroscopy.

7.2.2 Sample preparation

Sample preparation is a crucial component of the overall FTIR bacteria identification process. However, specific details regarding the growth and deposition of the bacteria lie outside the scope of this study (information concerning these aspects of the bacteria identification process is provided in [23] and in upcoming publications by the McGill IR Group). Nonetheless, it is noted here that research conducted by the McGill IR Group as well as others has shown that consistent growth parameters, including growth media selection, growth times, and inactivation techniques, are a prerequisite to reliable and reproducible classification and identification of bacteria by FTIR spectroscopy. As such, if a database consists of samples that have been prepared

by different protocols, then information on the protocol that has been employed for each sample can be incorporated into the expert system as an initial separating factor prior to performing spectral comparisons.

Based on the work done to date by the McGill IR Group, the recommended protocol for database development involves the growth and deposition of three replicates of each reference sample and the acquisition of three images from each replicate for a total of 9 images per reference sample.

7.2.3 Spectral acquisition

The approach used to date to acquire images from a deposit of bacteria on a zinc selenide slide requires the operator to identify visually a portion of the deposit that is of appropriate thickness and would provide as uniform (in terms of sample pathlength) an image as possible. With experience, an operator can become quite proficient at identifying such locations, but there is still a degree of guesswork, and thus a preliminary image acquisition (typically, 8 co-added scans) is usually performed to ensure that the portion of the sample selected is suitable. The possibility of automating the examination of the visual image might be suggested as a means of eliminating this sometimes tedious step. However, while the visual image may provide cues regarding the uniformity of the sample's pathlength in the field-of-view, it does not give a reliable indication of the appropriateness of the pathlength.

In view of the above, an approach has been developed that involves the use of the total response of the FPA detector, a value (measured in counts) relating to the cumulative voltage response of the detector over time and indicative of the amount of infrared light (across the response range of the detector) impinging on the detector. Measurement of this value with a sample in the beam path provides an estimate of the sample's thickness. In addition, the pathlength uniformity can be easily evaluated by comparing total response values across the pixel array of the detector. In this manner, the area of a deposit can be quickly scanned and portions where the total response exhibits the appropriate value and uniformity can be used for acquisition. This approach requires an initial calibration stage where total response values are compared to

absorbance values. It is only necessary to determine the lower and upper limits of total response corresponding to the lower and upper acceptable limits of absorbance.

Images acquired after applying this technique showed significant improvements in image uniformity, with greater than 80% of pixels remaining after filtration as compared to 30%-50% when the visual approach is employed. While this represents an improvement over earlier approaches (using visible light to identify suitable portions of the sample), this aspect is not of primary relevance. The most important aspect of this approach is its amenability to automation. Through examination of the total response, the process of acquiring images becomes deterministic in nature and removes the guesswork (whether human or computer). A routine could be implemented where the computer automatically scans across the sample and acquires images at points of appropriate pathlength and uniformity. The calibration process can also be easily automated by having the computer initially scan multiple portions of a sample with differing total infrared response levels and correlating the response levels with the resulting absorbance values. Due to the inaccessibility of the instrument's operating software, the automated routine has not yet been implemented, but no hurdles are foreseen in implementing this routine should such functionality make itself available.

7.2.4 Image processing and pixel filtration

Regardless of whether an image is to be used as part of the reference dataset or to be identified as an unknown, the pre-processing and data extraction routines are the same. The process is described by Figure 7.2. First, individual pixels are processed as described in Chapter 4. Variability measured as part of the similarity filter provides an indication of the overall spectral variability within an acquired image. If the spectral variability within the image is found to be abnormally high, the process outlined in Chapter 6 can be conducted in order to evaluate whether the sample consists of more than one type of bacteria. Ultimately, though, such a process is only intended to provide information that may be beneficial to the user; irrespective of the results, the image is rejected as inappropriate for inclusion in the database or for use in the identification of the sample.

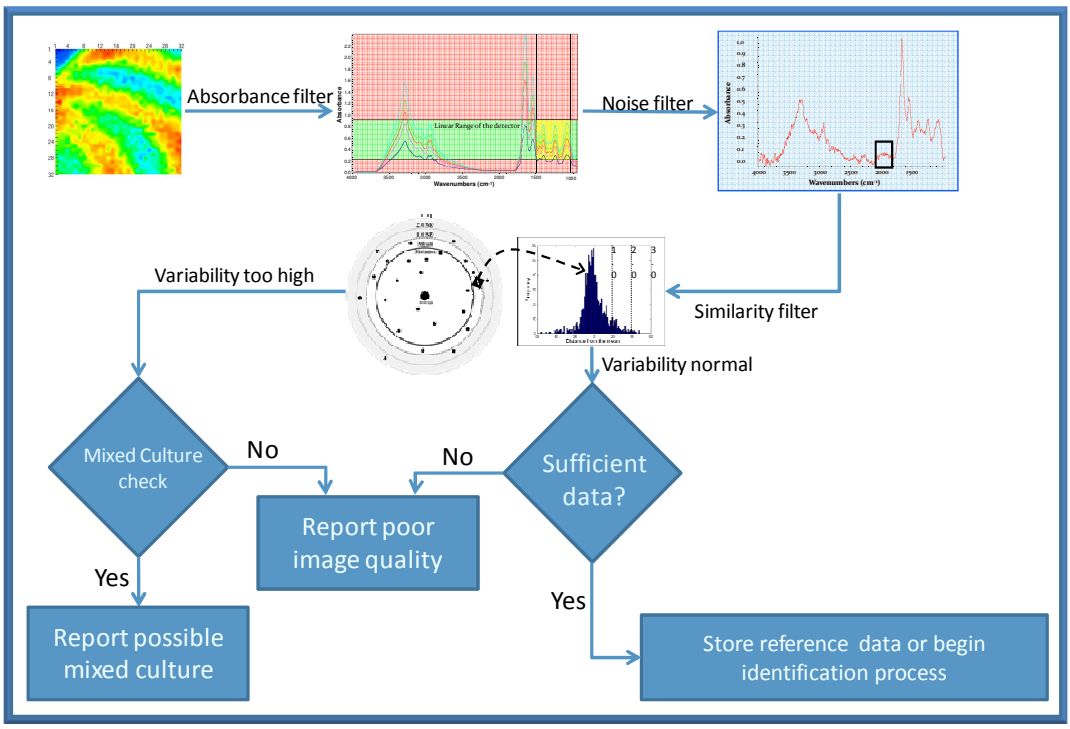


Figure 7.2: An overview of the image processing and filtration routines. The images are initially processed through an absorbance and a noise filter. The similarity filter is then applied and the measured variability used to determine the suitability of the image. If the image exhibits a high variability it can be further examined as a possible mixed culture. If variability is appropriate and sufficient data remains after applying the three filters, then the spectral data is stored for later use.

If a reasonable number of pixels remain after filtration, the data is stored to be used later for the construction of a spectral database or for identification. From the results of Chapter 3, it has been shown that averaging more than 64 pixels together yields negligible improvements in the overall SNR. For this reason, 64 pixels is considered the bare minimum number of acceptable pixels that must remain for an image not to be rejected in its entirety, although, when constructing a reference database, a minimum of 512 pixels (providing up to 8 spectra consisting of 64 averaged pixels each) would be more prudent. This value, however, is somewhat arbitrary and is primarily dependent on the experimental results obtained during the initial stages of database construction. Employing the automated approach described in Section 7.2.3 will ensure that the images acquired will exceed a predefined minimum number of acceptable pixels. However, for the majority of data acquired thus far and discussed in

this study, the spectral acquisition technique described above was not employed and so a minimum pixel cut-off is of relevance in these cases.

7.2.5 Database construction

7.2.5.1 Database development with Infrared Images

To our knowledge, there is no example in the literature of a database methodology using mid-infrared spectral image data. Although classifiers have been built from data acquired using FPA-FTIR technology, there is no precedent as to how one might go about constructing, optimizing, and maintaining a comprehensive database consisting of data acquired from infrared images. Additionally, there are no software solutions available which allow a user to easily interpret and manipulate infrared images directly, and to conveniently build, modify, and optimize a database consisting of infrared images. Such resources and tools were a necessity for the success of a FPA-FTIR spectroscopic approach to bacteria identification, and thus the necessary tools were developed as part of this study. In particular, software and data processing routines for efficient interpretation, manipulation, and management of a large number of images were created.

7.2.5.2 Multi-tiered database structure

A multi-tiered database structure was implemented as a technique for improving performance of the spectral database. Such an approach has been used before for bacteria identification [37, 47] and is a technique commonly used in chemometrics and artificial intelligence. The prerequisite to hierarchical structuring of the database is that the data must be divisible on multiple layers. This is precisely the case when developing a database of microbiological samples. Microbiological samples such as bacteria are categorized on different levels (taxonomy) based on their phenotypic and genotypic characteristics. These characteristics are the means by which microbiologists have categorized bacteria into several groups such as Gram-type (i.e. Gram-positive or Gram-negative), genus, species, strain, and so on.

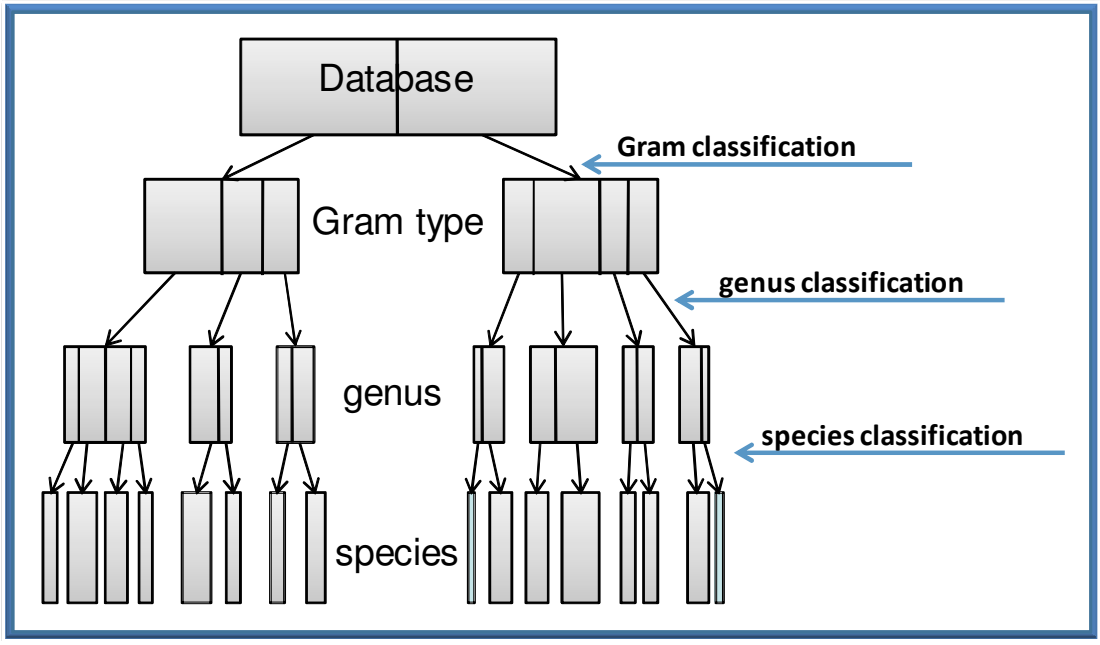


Figure 7.3: An illustration of the hierarchical database structure. Multiple classifiers are built (except for Gram separation) at each level of the hierarchical tree. An identification would start at the top by assigning a Gram classification and then move its way down the tree until the most specific level of taxonomic specification is reached.

This type of database structure offers several advantages, such as the potential for improvements in search speed, an improvement in classification performance, and the ability to offer different levels of classification. A search speed advantage can be gained by using less data for easier classifications (such as Gram-positive versus Gram-negative). In this context, this can be achieved by simply averaging the reference spectra for a particular class so that there are less reference spectra per class. With regards to a Gram classification, this averaging will have little to no effect on the classification accuracy and will reduce the overall search time. Furthermore, classification at this level eliminates a large percentage of the reference samples as candidates in identifying the unknown sample at later levels (genus, species) of classification. By extending this principle at each level of classification (Gram-type, genus, species, and strain), a sharp reduction in the overall number of comparisons can be achieved, hence improving the overall classification speed.

Another advantage to this approach is the generalization of the classification of an unknown. By defining and searching the database in a hierarchical fashion, a result can be given at each level of classification. If at a certain level a classification cannot be attained with the desired degree of confidence, then the result of a previous level can be provided. For example, a search may yield with high confidence the results that an unknown sample is Gram-positive and of the genus *Listeria*; however, it may be that a suitable match with a specific species of *Listeria* cannot be found, in which case the search result can offer as output the genus classification. This is of particular importance since bacteria are continually changing and evolving, making it highly likely that one would encounter on a regular basis samples that differ slightly from the reference strains contained within a database.

The multi-level approach to classification offers improved classifier performance by restricting difficult classifications to small concise subsets where local optimizations can be performed. For example, by assigning an unknown a classification of Gram-positive or Gram-negative a large portion of the database is eliminated from the succeeding levels of analysis. It would be very difficult to construct a single classifier to reliably identify an unknown from a collection of hundreds of different strains of bacteria due to the small differences seen between spectra at the strain level. On the other hand, identification from a concise list of several strains of a single species is much more likely to yield success. Additionally, classifiers can be optimized locally, further improving the odds of success. The primary cost of having a locally optimized multi-tier database structure is that there is an initial increase in the amount of optimization required. When a database is first constructed, each node at each of the different levels will require optimization of a classifier. However, the added computational cost is outweighed by the long-term benefit of handling changes and additions to the database locally. In other words, if a new strain of *Salmonella* is added to the database, it will not be necessary to re-optimize the classifier associated with any other genus.

7.2.5.3 Classifier development

There are a multitude of classification algorithms that could be used for the construction of spectral databases. In general, though, the degree to which a classifier can successfully be used with a particular set of data is dependent on the quality of the separations that exist in the data. If the data separates naturally and cleanly, then it will usually be found that regardless of the classification techniques used, the resultant classifier will perform well. Conversely, it is only when the data for different classes separate poorly that specialized techniques are employed. Many such classification techniques are specialized variations of their more familiar counterparts. Nonetheless, care must always be taken as even basic versions of techniques such as artificial neural networks and support vector machines have an incredible capacity to learn training data. However, if the data contains high levels of noise, the resultant classifiers will be of no practical use [86].

The approach used in this study for database development and the identification of unknowns is that of a K-nearest neighbor algorithm optimized using a forward search feature selection routine. The benefits of this approach are that it is deterministic, has few operational parameters, and to date has provided good results in our research.

The forward search feature selection technique was chosen over methods such as genetic algorithms [74] and grid-greedy [63] search for several reasons. Firstly, the search routine is deterministic and so it will yield the identical optimization result given identical data. This is important with regards to automation, as the user will typically not be able to judge the adequacy of an optimization result and so it is desirable that the output be repeatable and predictable. Secondly, the forward-search routine is the fastest of the optimization techniques used in this study. In particular, it has been found that as the level of taxonomic specification increases, the percentage of the spectrum that will be relevant for the classification decreases. Since a forward search begins with an empty set of features, it may only require several iterations of the search routine to find a small but effective set of features which best separates the data. The stopping conditions can also be explicitly set to ensure that a certain degree of optimization is

attempted. By setting the use of a minimum number of features as a condition, the potential for overfitting is diminished. If the performance of the k-NN classifier with the minimum number of features is unsatisfactory, this provides an indication to the expert system (and the user) to review the reference data. In addition, the process of determining a classification score (see below) allows for the identification of those spectra that are performing poorly in the classification model. Drawbacks of the forward-search approach include the somewhat arbitrary value assigned as a stopping condition, the proneness to local maxima, and the search times for classifiers where nearly all features are useable (such as in the case of Gram positive vs. Gram negative). Certain enhancements have been made to the implementation of the forward search to deal with these limitations. These include handling tied classification scores by branching the search in order to examine multiple pathways, and a simultaneous backward search in situations where the classification score is near perfect (= 1). In our studies this approach has been found to be the preferred approach to optimizing k-NN based classifiers.

The cumulative genetic algorithm (CGA) approach to feature selection could also be used for the optimization of the k-NN classifier. However, it was felt that for the initial applications of this technology, the databases being used will likely see frequent modification and updating. The computational time required by the CGA approach makes such a method impractical for the time being. However, if a definitive database which sees only minimal changes is used, then the CGA would be the preferred approach to optimizing classifiers. In addition, as the computational capabilities of computers improve, the optimization time required by the CGA will reduce to a point where it will be practical in most situations.

Of the variety of approaches examined (including artificial neural networks, support vector machines, decision trees, hierarchical cluster analysis, principal component analysis, and partial least squares regression) for use as a classifier, in the majority of cases, when one method was capable of building a suitable classifier, so were the others. However, some sacrifices are made when a k-NN algorithm is

employed instead of techniques such as ANN and SVM. In particular, the k-NN algorithm uses the reference data directly, at the time of identification, for establishing the identification of an unknown (referred to as a lazy learner). In contrast, an eager algorithm creates a classifier based on the training data, at which point only the classifier (and not the original data) is needed for further classification of unknowns. At first glance, it may seem that the eager approach is more suited to the type of scenario considered in this study since the time-consuming step of training could be accomplished in advance, making subsequent classification a much simpler and speedier affair. However, there are a number of limitations to eager learners which are not self-evident, in particular in relation to their amenability to automation, interpretation of output, reaction to a new type of sample (not in the database), and feasibility in a multi-class environment. While eager techniques (such as ANN and SVM) worked well in the lab, where operational parameters can be adjusted as necessary, it was found that these approaches were difficult to automate. In addition, the response of an eager learner to types of samples not seen during training is unpredictable and difficult to interpret. For these reasons and also due to its transparency and simplicity, the k-NN approach is used.

7.2.5.4 Classification score

The classification score used in this study is based on a leave one out cross validation (LOOCV) [45], which operates by incrementally removing each member of a dataset and testing it against a model built using the remainder of the dataset. In other words, the idea is to remove a sample from the training set, build a model with the remaining samples, and then test the one that has been removed. This process is repeated for every sample in the training set. In this approach the k-nearest neighbor (k-NN) algorithm is used to perform the evaluation at each step of the LOOCV, with k being set to one less than the number of members belonging to the class of the removed member. For example, if the removed member belongs to a class composed of eight samples, the k-NN search will be conducted and return the seven (k = 7 as one has been removed) nearest neighbors. In this model, the score for an individual (removed)

member is determined as the number of samples out of seven that belong to the same class as the removed member. The final classification score is taken as the average of the individual scores obtained in this manner for all the samples in the training set. The score that results will have a value between 0 and 1, with 1 indicating that for every iteration of the LOOCV, every nearest neighbor is a member of the same class as the member left out.

Among the advantages of using this method for evaluating a set of features are that it provides a good indication of the separation between the classes, it is deterministic, and it does not require training. However, the k-NN/LOOCV approach can become computationally expensive as the number of reference samples increases. Specifically, the computational cost increases by the square of the number of reference samples. An upper limit can be set on the number of nearest neighbors to use; however, the effectiveness of the k-NN/LOOCV approach would then be diminished as the reference set for each class grows larger while the value of k used remains the same.

7.2.5.5 Data compilation

To effectively use a database consisting of k-NN classifiers, the reference data must be compiled into one cohesive data package consisting of all the classifiers at each level of taxonomic specification in order to minimize computational processing times at the time of identification. Otherwise, the computational cost of conducting the retrieval and extraction of the necessary reference data at the time of identification would render the k-NN approach infeasible for larger databases. After the feature selection optimization has completed, the compilation process is carried out by extracting and storing the relevant portions of the spectrum (those portions chosen by the feature selection routine) in a database. This process is repeated for each classifier in the search tree. The reduction from the original data size to the reduced size is quite significant as is described in Figure 7.4.

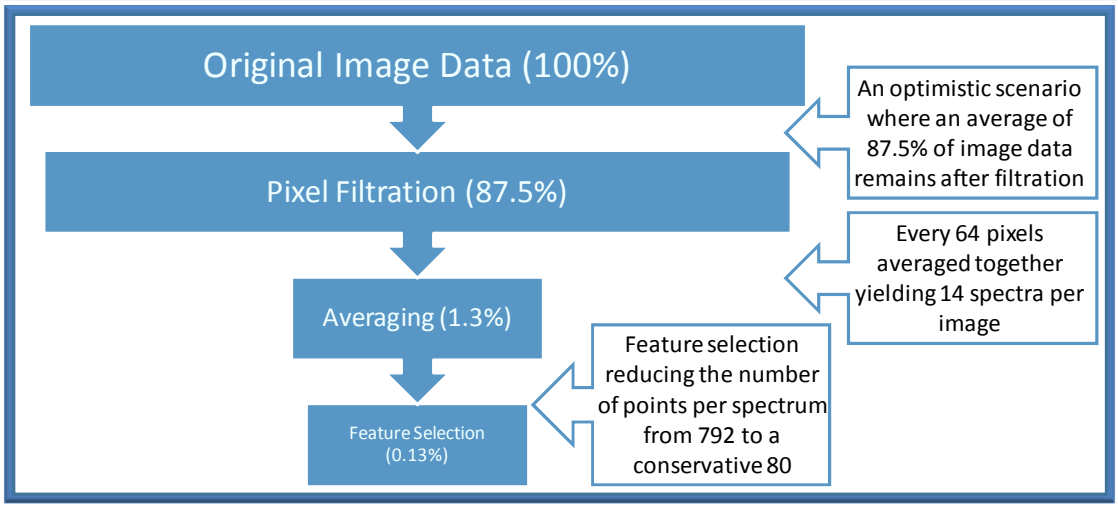


Figure 7.4: Conservative description of the reduction in data size as a result of each stage of image processing and optimization. As can be seen, these processes reduce the data to 0.13% of its original size. A database of 1000 images would normally occupy ~3 gigabytes ($3x10^9$ bytes) of storage space, a large quantity for a modern desktop computer. By applying these routines, the data is reduced to ~4 megabytes ($4x10^6$ bytes), an amount easily handled by today's computers.

As is illustrated by Figure 7.4, the resultant database file is reduced enough in size to allow for the database to be loaded in its entirety into computer memory, which in turn allows for an efficient search of the database to identify unknowns. It should be noted that the original spectral data is still necessary if the database is to be modified or augmented, as new classifiers might need to be built and additional optimization routines conducted.

An additional consideration is that the compilation step can be seen as a security enhancing feature since compilation of the database results in a masking of the original spectral images, so that the end user is unable to modify the database without the original data. This would allow the creator/developer of the database to provide search access without exposing the original data, and making it very difficult for the user to modify the contents of a database.

7.2.6 Identification

Prior to identification, unknown images are processed as outlined in Figure 7.2, with a minimum of 64 pixels being averaged to generate a spectrum of the unknown.

The spectra from multiple replicate images of the unknown are then put through the identification process simultaneously. Unknowns are identified using a weighted k-NN with k equal to the smallest number of spectra belonging to a particular group of a classifier. For example, if the desired separation is between *Escherichia* and *Salmonella* for which there are 25 and 10 reference spectra, respectively, then the k-NN algorithm will be conducted with k = 10. This is one of the reasons why it is preferable to have each classification group consist of a reasonable number of representatives. If the reference bacteria have been prepared under different conditions or using different protocols, then this information (provided by the user) is used to define which portion of the database to search. The classification then begins at the most general level of taxonomic specification, generates a classification, and then proceeds to the next appropriate classifier. A classification and confidence in the classification are generated at each level unless a result with extremely poor confidence (defined below) occurs at an upper level, at which point the search will not continue down the hierarchical tree of classifiers. A three-tiered confidence is used in this methodology which is determined from three difference aspects of the classification. Firstly, the distance to the mean spectrum of the identified group is compared to the distance of the members of the identified group to the mean spectrum (as described in 2.4.4). Secondly, the classification process produces a classification score based on the weighted percentage of the k nearest neighbors that belong to the identified group. These two confidence measures can be put into context by examining the average and standard deviation for the particular confidence measure. The approach used and implemented in this study is to assign a 'High', 'Medium', and 'Low' confidence output based on the confidence measure being less than one, two, and three standard deviations from the average of the identified group. However, these criteria may be adjusted depending on the intended commercial application. A third confidence measure is the combination of classification and confidence assignments of all spectra (each consisting of 64 pixels) belonging to the unknown images of a single sample. A best-case scenario where no pixels are filtered would result in 144 spectral classifications and confidence

assignments. Similarly to what is described in Chapter 6, there is much information that can be attained by comparing the results from multiple spectra from a single image and from multiple images of a single sample. A high confidence from multiple spectra from multiple images provides an even greater degree of confidence in the classification. Likewise, if the confidence or classification is found to vary within or between images, then this provides additional information regarding the sample that can be relayed to the user.

An additional benefit of employing the k-NN algorithm, in conjunction with the hierarchical database structure, is that it may allow for the interpretation of spectral data obtained from bacterial strains not represented in the reference database. Generally, although it is not always the case, a bacterial strain that is not found in the database will be spectrally quite similar to other bacterial strains of a similar taxonomic specification (e.g., different strains of *Listeria monocytogenes*). As a result, when the k-NN algorithm, encounters a new bacterial strain, it will often yield a result with low confidence at the lowest level (strain) but a higher confidence at an upper level (species or genus). This information can then be passed on to the user for further interpretation. In addition, if portions of the spectrum show larger differences than others, through spectral subtractions of the spectrum of the unknown from the closest matching reference spectra, then information regarding the biochemical origin of the difference between the new strain and those in the spectral database can be provided to the user.

7.3 IMPLEMENTATION

The methodology described in this study has been implemented in a user-friendly software. The software has been separated into two components, each targeting a different end user. The first is a database-building program which allows the user to construct new databases at multiple levels of taxonomic specification. The software provides feedback during the database building process regarding the quality of reference images and performance of the classifiers developed (as outlined above). The initial users that will most likely be targeted for the commercial introduction of

bacteria identification by FPA-FTIR spectroscopy are government or regulatory agencies whose use for this technology would be the construction of databases specific to their needs. The second component of the software is a program specifically for identification, with no functionality for building or modifying databases. In this scenario, a comprehensive database is supplied to the user and the FPA-FTIR instrumentation is used strictly as an analyzer for the identification of bacteria. This program aims to abstract the infrared imaging data into a form which is familiar to the user, who is most likely not a spectroscopist. The user will still be required to direct the acquisition of infrared images until a completely automated approach is implemented, but the user will not be required to examine, analyze, or interpret the imaging data and instead receives feedback regarding the suitability of the imaging data and the confidence of the identifications as was described above.

7.4 RESULTS AND DISCUSSION

Figure 7.5 provides a list of some of the databases that have been successfully constructed and using the strategies described in this chapter. Due to the nature of the expert system and the fact that these datasets were generated in a research setting, the evaluation of the predictive abilities of these datasets does not apply. In other words, the expert system is designed to optimize the databases for the target classifications by eliminating images that do not meet the defined spectral quality criteria or are identified as outliers during the construction of classifiers. As such, the databases are built so that they contain only the data which results in classifiers that (near) perfectly separate the data. If this is not the case then the poor separability is indicated to the user and the database is not used for classifying unknowns. The expert system successfully built classifiers that separated the datasets listed in Table 7.1; however, a proper validation of the predictive ability of the built classifiers (particularly at the strain level) would require the growth of additional bacterial samples for testing as unknowns. Regardless, the purpose is not to test the predictive abilities of the expert system as it is well established that given good data, the spectra corresponding to the different groups

of bacteria can be separated. The focus has been on evaluating and optimizing the feedback given by the expert system during the construction of databases and by intentionally predicting known images in order to assess the response. The software developed as the implementation of this methodology has been successfully used by multiple individuals as part of their own research allowing for the analysis, interpretation, and evaluation of databases of infrared images that would not have otherwise been possible.

Table 7.1: An overview of some of the datasets that have been analyzed by the methodology described in this study. These datasets were acquired over a period of several years as part of several different studies conducted by current and former students of the McGill IR-Group.

Taxonomic level	Separation of	# of images
Gram –Type	Gram-positive, Gram-negative	74/117
Genus	*Salmonella, Escherichia, Shigella*	17/36/15
	Listeria, Clostridium	51/24
Species/Group	*Shigella*: *flexneri, dysenteriae, sonnei*	53/45/21
	Salmonella : *branderberg, derby, infantis*, newport, *heidelberg, kentucky, typhmurium*	4/3/3/6/5/8/13
	Campylobacter: *jejuni, coli*	77/217
	Clostridium botulinum: group A & B*	55
Strain	*Campylobacter jejuni* strains	117 (from 35 different strains)
	Salmonella typhmurium strains	13 (4 different strains)

* Acquired on a 16x16 array

Currently in development is a comprehensive database containing a broad cross section of bacteria, prepared under an established protocol, for which images are being acquired using the improved acquisition procedure (7.2.3), and employing the most recent imaging technology available. This database represents the ideal use of the expert system and will allow for further validation of all components of this methodology. Nonetheless, each component of this methodology has been thoroughly studied and examined as is illustrated through the chapters and the multitude of

datasets presented in this thesis. The methods outlined in this chapter are also being validated for their applicability to other microorganisms such as molds, fungi, and plankton [87].

7.8 CONCLUSION

This study has introduced a methodology by which an expert system can be created for the identification of bacteria by FPA-FTIR spectroscopy. The methodology includes techniques for acquiring, processing, analyzing, interpreting, relaying, and managing large amounts of infrared imaging data by automating the necessary processing, optimization and classification routines and providing the user with important feedback and relevant information. With this approach, FPA-FTIR spectroscopy may become a versatile and easy means of rapid microbial identification.

Chapter 8:
Contributions to knowledge

This thesis introduces a number of novel ideas and methods for the analysis of data acquired by FPA-FTIR spectroscopy and incorporates these ideas into a complete methodology for the identification of bacteria. These include routines for improved spectral acquisition, the development and optimization of databases, and the identification of unknowns. The primary contributions to knowledge resulting from this research are summarized below:

1. The first simultaneous comparison of multiple FPA-FTIR spectrometers in terms of SNR and quantitative performance on a pixel-to-pixel and instrument-to-instrument basis.

While the performance characteristics of FPA detectors have been investigated in other studies, this study is the first to make direct instrument-to-instrument comparisons as well as direct comparisons of the performance of FPA-FTIR detectors with that of single-element detectors. Instruments were evaluated based on their SNR values, spectral response, and quantitative accuracy, as assessed by the calibration SD of Beer's law calibrations, at the pixel level, and the effects of pixel filtering and co-addition were investigated. The results of this study clearly illustrated the pixel-to-pixel variability that exists with present-day FPA-FTIR spectrometers, which is important information for any prospective application which looks to employ this technology.

2. Developed pixel filtration routines for the extraction of representative data from non-uniform samples.

The non-uniform nature of bacteria samples deposited on IR-transparent substrates for FTIR analysis and the pixel-to-pixel variability encountered with the FPA detectors necessitated the development of pixel filtration routines for the extraction of relevant data from the images acquired from bacteria. These pixel filtration routines eliminate pixels that do not meet pre-defined spectral quality criteria. In addition, a computational approach was introduced for optimizing the spectral quality criteria used for the filtration of pixels so as to maximize the quantity and quality of data extracted from the infrared images.

3. Introduced a novel method for determining feature relevancy in the form of cumulative genetic algorithms. This approach allows for a greater degree of interpretation regarding the importance of particular spectral features with respect to separation of the spectral data into their different groups.

There are many techniques for the selection of features in spectroscopic data, as part of classifier optimization. Few of these approaches, however, provide details regarding the relative relevancy of selected features. This study demonstrated such an approach through the use of a cumulative genetic algorithm. This approach provides the potential for extracting useful information regarding the chemical components that differentiate the classes of spectral data under examination. In addition, it also showed potential as a method for generating weighted feature sets which are less prone to overfitting.

4. Conducted a proof-of-concept study into the potential for identifying spectral images originating from a mixed culture of bacteria. This is the first study which attempts to identify the presence of multiple bacteria in a single sample using FTIR spectroscopy.

All methods for bacteria identification by FTIR spectroscopy have so far required that the sample under analysis be from a pure culture. This study introduced new concepts of analysis for identifying a sample originating from a mixed culture. It is also the first study to resolve and identify two different bacteria in a single infrared image.

5. Outlined a novel methodology for building databases from infrared images and the methods by which unknowns can be identified. These methods are incorporated into an expert system resulting in an efficient and accessible method for bacteria identification by FPA-FTIR spectroscopy.

The sheer quantity of data acquired by FPA-FTIR spectroscopy results in a number of challenges when trying to build databases that are to be used for the identification of unknowns. This study introduces novel methods for overcoming those challenges. These methods allow for optimized databases to be built from hundreds to thousands of images in an efficient manner and for unknowns to be identified with multiple measures of confidence. In addition, this study proposes a new approach to image acquisition, whereby the total infrared response of the FPA detector is measured. This approach provides improved uniformity in the spectral images acquired and a technique which is amenable to automation. Finally, these ideas are combined into a total methodology for the identification of bacteria in the form of an expert system, a necessary component for the effective use of FPA-FTIR spectroscopy for bacteria identification. While developed specifically for bacteria identification, the ideas presented in this study have general applicability to applications looking to build databases using infrared imaging data.

In conclusion, this study has accomplished several important goals. It has introduced new methods of data analysis, data processing, data interpretation, and database development for use in FPA-FTIR spectroscopy. This opens the door to a variety of new applications by exploiting the information-rich nature of FPA-FTIR images in an efficient and effective manner. Finally, all these components have been implemented in an expert system. In so doing, FPA-FTIR spectroscopy becomes an efficient and easy to use method for the high-throughput identification of bacteria. The availability of a high-throughput method for bacteria identification could potentially reduce analysis times from days to hours, allowing regulatory agencies to take a more active stance on food safety through routine and regular analysis of the food supply.

References

1. Lewis, E.N., P.J. Treado, R.C. Reeder, G.M. Story, A.E. Dowrey, C. Marcott, and I.W. Levin, *Fourier Transform Spectroscopic Imaging Using an Infrared Focal-Plane Array Detector.* Analytical Chemistry, 1995. **67**(19): p. 3377-3381.

2. *ALPHA FT-IR spectrometers - Bruker Optics*. [cited; Available from: http://www.brukeroptics.com/alpha.html.

3. *Ahura Scientific | TruDefender FT | Handheld FTIR Spectrometer*. [cited; Available from: http://www.ahurascientific.com/chemical-explosives-id/products/trudefenderft/index.php.

4. Griffiths, P., *Fourier transform infrared spectrometry*. 2007, Hoboken N.J.: Wiley-Interscience.

5. Griffiths, P.R. and J.A. de Haseth, *Microspectroscopy and Imaging*, in *Fourier Transform Infrared Spectrometry*. 2007, John Wiley & Sons, Inc.: Hoboken, NJ, USA. p. 303-320.

6. Sommer, A.J. and J.E. Katon, *Diffraction-Induced Stray Light in Infrared Microspectroscopy and Its Effect on Spatial Resolution.* Applied Spectroscopy, 1991. **45**(10): p. 1633-1640.

7. Bhargava, R. and I.W. Levin, *Fourier transform infrared imaging: a new spectroscopic tool for microscopic analyses of biological tissue.* Trends in Applied Spectroscopy, 2001: p. 3:57-71-3:57-71.

8. Scribner, D.A., M.R. Kruer, and J.M. Killiany, *Infrared focal plane array technology.* Proceedings of the IEEE, 1991. **79**(1): p. 66-85.

9. Griffiths, P.R. and J.A. de Haseth, *Signal-to-Noise Ratio*, in *Fourier Transform Infrared Spectrometry*. 2007, John Wiley & Sons, Inc.: Hoboken, NJ, USA. p. 161-175.

10. Griffiths, P.R., *Theoretical Background*, in *Fourier Transform Infrared Spectrometry*. 2007, John Wiley & Sons, Inc.: Hoboken, NJ, USA. p. 19-55.

11. *FT-IR Spectrochemical Imaging products from Varian, Inc*. [cited; Available from: http://www.varianinc.com/cgi-bin/nav?products/spectr/ftir/ftir_imaging/index&cid=KPLQIPNIFL.

12. Snively, C.M. and J.L. Koenig, *Characterizing the Performance of a Fast FT-IR Imaging Spectrometer.* Applied Spectroscopy, 1999. **53**(2): p. 170-177.

13. Bhargava, R. and I.W. Levin, *Effective Time Averaging of Multiplexed Measurements: A Critical Analysis.* Analytical Chemistry, 2002. **74**(6): p. 1429-1435.

14. Jolliffe, I., *Principal component analysis*. 2002, New York: Springer.

15. Lasch, P., *FT-IR spectroscopic investigations of single cells on the subcellular level.* Vibrational Spectroscopy, 2002. **28**(1): p. 147-157.

16. Wood, B.R. and D. McNaughton, *FPA Imaging and Spectroscopy for Monitoring Chemical Changes in Tissue*, in *Spectrochemical Analysis Using Infrared Multichannel Detectors*. 2005, Blackwell Publishing Ltd: Oxford, UK. p. p.204-233-p.204-233.

17. Bhargava, R., B.G. Wall, and J.L. Koenig, *Comparison of the FT-IR Mapping and Imaging Techniques Applied to Polymeric Systems.* Applied Spectroscopy, 2000. **54**(4): p. 470-479.

18. Gupper, A., K.L.A. Chan, and S.G. Kazarian, *FT-IR Imaging of Solvent-Induced Crystallization in Polymers.* Macromolecules, 2004. **37**(17): p. 6498-6503.

19. Kazarian, S., K. Kong, M. Bajomo, J. Vanderweerd, and K. Chan, *Spectroscopic Imaging Applied to Drug Release.* Food and Bioproducts Processing, 2005. **83**(2): p. 127-135.

20. Bambery, K.R., B.R. Wood, M.A. Quinn, and D. McNaughton, *Fourier Transform Infrared Imaging and Unsupervised Hierarchical Clustering Applied to Cervical Biopsies.* Australian Journal of Chemistry, 2004. **57**(12): p. 1139-1139.

21. Koller, D. and M. Sahami. *Toward Optimal Feature Selection*. in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*. 1996: Morgan Kaufmann Publishers.

22. Marcott, C., G.M. Story, and R.K. Dukor, *Infrared Spectral Imaging of H&E-Stained Breast Tissue Biopsies.* Microscopy and Microanalysis, 2004. **10**(S02).

23. Prevost Kirkwood, J., *Identification of bacteria by infrared imaging with the use of focal plane array Fourier transform infrared spectroscopy (Ph.D. Thesis)*, in *Department of Food Science and Agricultural Chemistry*. 2007, McGill University.

24. Stevenson, H.J.R. and O.E.A. Bolduan, *Infrared Spectrophotometry as a Means for Identification of Bacteria.* Science, 1952. **116**(3005): p. 111-113.

25. Mariey, L., *Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics.* Vibrational Spectroscopy, 2001. **26**(2): p. 151-159.

26. Alholy, M., M. Lin, A. Cavinato, and B. Rasco, *The use of Fourier transform infrared spectroscopy to differentiate O157:H7 from other bacteria inoculated into apple juice.* Food Microbiology, 2006. **23**(2): p. 162-168.

27. Irudayaraj, J., *Differentiation and detection of microorganisms using Fourier transform infrared photoacoustic spectroscopy.* Journal of Molecular Structure, 2002. **606**(1-3): p. 181-188.

28. Lin, M., M. Al-Holy, H. Al-Qadiri, D.-H. Kang, A.G. Cavinato, Y. Huang, and B.A. Rasco, *Discrimination of Intact and Injured Listeria monocytogenes by Fourier Transform Infrared Spectroscopy and Principal Component Analysis.* Journal of Agricultural and Food Chemistry, 2004. **52**(19): p. 5769-5772.

29. Oust, A., T. Møretrø, C. Kirschner, J.A. Narvhus, and A. Kohler, *FT-IR spectroscopy for identification of closely related lactobacilli.* Journal of Microbiological Methods, 2004. **59**(2): p. 149-162.

30. Perkins, D.L., C.R. Lovell, B.V. Bronk, B. Setlow, P. Setlow, and M.L. Myrick, *Effects of autoclaving on bacterial endospores studied by Fourier transform infrared microspectroscopy.* Applied Spectroscopy, 2004. **58**(6): p. 749-753.

31. Whittaker, P., *Identification of foodborne bacteria by infrared spectroscopy using cellular fatty acid methyl esters.* Journal of Microbiological Methods, 2003. **55**(3): p. 709-716.

32. Yang, H., J. Irudayaraj, and S. Sakhamuri, *Characterization of Edible Coatings and Microorganisms on Food Surfaces Using Fourier Transform Infrared Photoacoustic Spectroscopy.* Applied Spectroscopy, 2001. **55**(5): p. 571-583.

33. Zhao, H., Y. Kassama, M. Young, D.B. Kell, and R. Goodacre, *Differentiation of Micromonospora Isolates from a Coastal Sediment in Wales on the Basis of Fourier Transform Infrared Spectroscopy, 16S rRNA Sequence Analysis, and the Amplified Fragment Length Polymorphism Technique.* Applied and Environmental Microbiology, 2004. **70**(11): p. 6619-6627.

34. Adams, M.J., *Chemometrics in Analytical Spectroscopy*. 2004, Cambridge: Royal Society of Chemistry.

35. *StatSoft Inc. Cluster Analysis. 2006. Available at: http://www.statsoft.com/textbook/stcluan.html. Accessed 13/June/2007*.

36. Choo-Smith, L.P., K. Maquelin, T. van Vreeswijk, H.A. Bruining, G.J. Puppels, N.A.N. Thi, C. Kirschner, D. Naumann, D. Ami, A.M. Villa, F. Orsini, S.M. Doglia, H. Lamfarraj, G.D. Sockalingum, M. Manfait, P. Allouch, and H.P. Endtz, *Investigating Microbial (Micro)colony Heterogeneity by Vibrational Spectroscopy.* Applied and Environmental Microbiology, 2001. **67**(4): p. 1461-1469.

37. Maquelin, K., C. Kirschner, L.P. Choo-Smith, N.A. Ngo-Thi, T. van Vreeswijk, M. Stammler, H.P. Endtz, H.A. Bruining, D. Naumann, and G.J. Puppels, *Prospective Study of the Performance of Vibrational Spectroscopies for Rapid Identification of Bacterial and Fungal Pathogens Recovered from Blood Cultures.* Journal of Clinical Microbiology, 2003. **41**(1): p. 324-329.

38. Oberreuter, H., A. Brodbeck, S. von Stetten, S. Goerges, and S. Scherer, *Fourier-transform infrared (FT-IR) spectroscopy is a promising tool for monitoring the population dynamics of microorganisms in food stuff.* European food research & technology, 2003. **216**(5): p. 434-439.

39. Rebuffo, C.A., J. Schmitt, M. Wenning, F. von Stetten, and S. Scherer, *Reliable and Rapid Identification of Listeria monocytogenes and Listeria Species by Artificial Neural Network-Based Fourier Transform Infrared Spectroscopy.* Applied and Environmental Microbiology, 2006. **72**(2): p. 994-1000.

40. Wenning, M., V. Theilmann, and S. Scherer, *Rapid analysis of two food-borne microbial communities at the species level by Fourier-transform infrared microspectroscopy.* Environmental Microbiology, 2006. **8**(5): p. 848-857.

41. Wieser, M., E.B.M. Denner, P. Kämpfer, P. Schumann, B. Tindall, U. Steiner, D. Vybiral, W. Lubitz, A.M. Maszenan, B.K.C. Patel, R.J. Seviour, C. Radax, and H.-J. Busse, *Emended descriptions of the genus Micrococcus, Micrococcus luteus (Cohn 1872) and Micrococcus lylae (Kloos et al. 1974).* International Journal of Systematic and Evolutionary Microbiology, 2002. **52**(Pt 2): p. 629-637.

42. Winder, C.L., E. Carr, R. Goodacre, and R. Seviour, *The rapid identification of Acinetobacter species using Fourier transform infrared spectroscopy.* Journal of Applied Microbiology, 2004. **96**(2): p. 328-339.

43. Zhao, H., R.L. Parry, D.I. Ellis, G.W. Griffith, and R. Goodacre, *The rapid differentiation of Streptomyces isolates using Fourier transform infrared spectroscopy.* Vibrational Spectroscopy, 2006. **40**(2): p. 213-218.

44. Beebe, K., *Chemometrics : A practical guide*. 1998, New York: Wiley.

45. Russell, S., *Artificial intelligence : a modern approach*. 2003, Upper Saddle River N.J.: Prentice Hall/Pearson Education.

46. Mouwen, D., R. Capita, C. Alonsocalleja, J. Prietogomez, and M. Prieto, *Artificial neural network based identification of Campylobacter species by Fourier transform infrared spectroscopy.* Journal of Microbiological Methods, 2006. **67**(1): p. 131-140.

47. Udelhoven, T., D. Naumann, and J. Schmitt, *Development of a Hierarchical Classification System with Artificial Neural Networks and FT-IR Spectra for the Identification of Bacteria.* Applied Spectroscopy, 2000. **54**(10): p. 1471-1479.

48. Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. **2**(2): p. 121-167.

49. Harz, M., P. Rösch, K.D. Peschke, O. Ronneberger, H. Burkhardt, and J. Popp, *Micro-Raman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions.* The Analyst, 2005. **130**(11): p. 1543-1543.

50. Martens, H., *Multivariate calibration*. 1989, Chichester, England: Wiley.

51. Næs, T., *A user-friendly guide to multivariate calibration and classification*. 2004, Chichester UK: NIR Publications.

52. Höskuldsson, A., *PLS regression methods.* Journal of Chemometrics, 1988. **2**(3): p. 211-228.

53. Oust, A., T. Møretrø, C. Kirschner, J.A. Narvhus, and A. Kohler, *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.

54. Preisner, O., J.A. Lopes, R. Guiomar, J. Machado, and J.C. Menezes, *Fourier transform infrared (FT-IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination.* Analytical and Bioanalytical Chemistry, 2007. **387**(5): p. 1739-1748.

55. Wold, S. and M. SjÖStrÖM, *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy*, in *Chemometrics : theory and application : a symposium*. 1977, American Chemical Society: Washington. p. Chapter 12, pp 243–282-Chapter 12, pp 243–282.

56. Baldauf, N.A., L.A. Rodriguez-Romo, A.E. Yousef, and L.E. Rodriguez-Saona, *Differentiation of selected Salmonella enterica serovars by Fourier transform mid-infrared spectroscopy.* Applied Spectroscopy, 2006. **60**(6): p. 592-598.

57. Manly, B., *Multivariate statistical methods : a primer*. 2005, Boca Raton FL: Chapman & Hall/CRC Press.

58. Sivakesava, S., J. Irudayaraj, and C. DebRoy, *Differentiation of microorganisms by FTIR-ATR and NIR spectroscopy.* ASABE, 2004(47): p. 951-957.

59. Lai, S., *Whole-organism Fingerprinting of the Genus using Fourier Transform Infrared Spectroscopy (FT-IR).* Systematic and Applied Microbiology, 2004. **27**(2): p. 186-191.

60. Udelhoven, T., D. Naumann, and J. Schmitt, *Development of a Hierarchical Classification System with Artificial Neural Networks and FT-IR Spectra for the Identification of Bacteria.* Applied Spectroscopy, 2000. **Vol. 54**(Issue 10): p. 1471-1479.

61. Jarvis, R.M., *Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data.* Bioinformatics, 2004. **21**(7): p. 860-868.

62. Keshava, N., *Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries.* IEEE Transactions on Geoscience and Remote Sensing, 2004. **42**(7): p. 1552-1565.

63. Prevost Kirkwood, J., A. Ghetler, J. Sedman, D. Leclair, F. Pagotto, J.W. Austin, and A.A. Ismail, *Differentiation of group I and group II strains of Clostridium botulinum by focal plane array Fourier transform infrared spectroscopy.* Journal of Food Protection, 2006. **69**(10): p. 2377-2383.

64. Karl, H.N., *Applying Norris Derivatives. Understanding and correcting the factors which affect diffuse transmittance spectra.* NIR news, 2001. **12**(3): p. 6-9.

65. Savitzky, A. and M.J.E. Golay, *Smoothing and Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 1964. **36**(8): p. 1627-1639.

66. Mahalanobis, P.C. *On the generalised distance in statistics*. in *Proceedings National Institute of Science, India*. 1936.

67. Griffiths, P.R., *Quantitative Analysis*, in *Fourier Transform Infrared Spectrometry*. 2007, John Wiley & Sons, Inc.: Hoboken, NJ, USA. p. 197-224.

68. Kramer, R., *Chemometric techniques for quantitative analysis*. 1998, New York: Marcel Dekker.

69. Kaun, N., M.J. Vellekoop, and B. Lendl, *Time-resolved Fourier transform infrared spectroscopy of chemical reactions in solution using a focal plane array detector.* Applied Spectroscopy, 2006. **60**(11): p. 1273-1278.

70. *Omnic Version 7.3. Excerpted from: Help File, RMS Noise. Thermo Fisher Scientific, Inc. Waltham, MA*.

71. Prevost Kirkwood, J., S.F. Al-Khaldi, M.M. Mossoba, J. Sedman, and A.A. Ismail, *Fourier transform infrared bacteria identification with the use of a focal-plane-array detector and microarray printing.* Applied Spectroscopy, 2004. **58**(11): p. 1364-1368.

72. Ngo-Thi, N., *Characterization and identification of microorganisms by FT-IR microspectrometry.* Journal of Molecular Structure, 2003. **661-662**: p. 371-380.

73. Vandermei, H., D. Naumann, and H. Busscher, *Grouping of strains grown on different growth media by FT-IR.* Infrared Physics & Technology, 1996. **37**(4): p. 561-564.

74. Goldberg, D., *Genetic algorithms in search, optimization, and machine learning*. 1989, Reading Mass.: Addison-Wesley Pub. Co.

75. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**(2): p. 123-140.

76. Leardi, R., *Application of genetic algorithm-PLS for feature selection in spectral data sets.* Journal of Chemometrics, 2000. **14**(5-6): p. 643-655.

77. Martens, M., H. Martens, and J.R. Piggot, *Partial least squares regression*, in *Statistical procedures in food research*. 1986, Elsevier Applied Science: London, England.

78. Transcript, V., *Excerpts from a Conversation with Gordon Moore: Moore‟ s Law.* Intel Corporation, 2005.

79. Ngo-Thi, N., C. Kirschner, and D. Naumann, *Characterization and identification of microorganisms by FT-IR microspectrometry.* Journal of Molecular Structure, 2003. **661-662**: p. 371-380.

80. Maquelin, K., C. Kirschner, L.P. Choo-Smith, N. van den Braak, H.P. Endtz, D. Naumann, and G.J. Puppels, *Identification of medically relevant microorganisms by vibrational spectroscopy.* Journal of Microbiological Methods, 2002. **51**(3): p. 255-271.

81. Henis, Y., J.R. Gould, and M. Alexander, *Detection and Identification of Bacteria by Gas Chromatography.* Applied Microbiology, 1966(14): p. 513-524.

82. *Identification of Shigella species*. 2007 [cited; Available from: http://www.hpa-standardmethods.org.uk/documents/bsopid/pdf/bsopid20.pdf.

83. Mouwen, D.J.M., M.J.B.M. Weijtens, R. Capita, C. Alonso-Calleja, and M. Prieto, *Discrimination of Enterobacterial Repetitive Intergenic Consensus PCR Types of Campylobacter coli and Campylobacter jejuni by Fourier Transform Infrared Spectroscopy.* Applied and Environmental Microbiology, 2005. **71**(8): p. 4318-4324.

84. *Sherlock Microbial Identification System*. [cited; Available from: http://www.midi-inc.com/pages/microbial_id.html.

85. Pincus, D.H., *Microbial Identification Using The Biomérieux VITEK® 2 System*, in *Encyclopedia of Rapid Microbiological Methods*, M.J. Miller, Editor. 2006, PFA/DHI.

86. *Garbage In, Garbage Out from FOLDOC*. [cited; Available from: http://foldoc.org/Garbage+In,+Garbage+Out.

87. Pinchuk, O.R., *Focal plane array-Fourier transform-infrared (FPA-FTIR) spectroscopy as a tool in the simple and rapid classification of common environmental and food spoilage fungi*. McGill theses. 2008.