8-2009

Spectroscopic Imaging for the Detection and Identification of Bacterial Contaminations

Michael Gilbert
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Michael Gilbert entitled "Spectroscopic Imaging for the Detection and Identification of Bacterial Contaminations." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Chemistry.

Frank Vogt, Major Professor

We have read this dissertation and recommend its acceptance:

Michael Sepaniak, Michael Best, Nicole Labbé

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Michael K. Gilbert entitled "Spectroscopic Imaging for the Detection and Identification of Bacterial Contaminations". I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Chemistry.

Frank Vogt, Major Professor

We have read this dissertation and
recommend its acceptance:

Michael Sepaniak

Michael Best

Nicole Labbé

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Spectroscopic Imaging for the Detection and Identification of Bacterial Contaminations

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Michael K. Gilbert

August 2009

Abstract

There exist many types of harmful bacteria that can contaminate foods and cause serious illness, which often have their own best courses of treatment. This requires the classification of different types of bacteria. Traditional methods of bacterial identification, while able to provide accurate classification, are often very time-consuming processes. In the case of a potentially fatal bacterial infection, time is often of the essence. FTIR spectroscopy is a faster, more practical alternative that can discriminate different strains of bacteria, based on their spectral signatures, with high confidence.

Bacterial contaminations on food exist as small, localized colonies that must be found and identified. As FTIR spectroscopy is a fast technique for one sample location, it is advantageous to move towards spectroscopic imaging to do FTIR spectroscopy at many locations in parallel. The amount of information obtained at one time from this technique is large, thus it can be helpful to visualize this information by means of color images. A multivariate image analysis will produce such color images depicting the different chemical properties of a material as different colors. This enables the visual discrimination of a bacterial colony from its substrate, as they will have different chemical properties. Colonies can be found in a larger sample area by merging data cubes as presented here. Once a bacterial colony is found in an extended area, it must be identified. A method of bacterial identification has been developed here that can identify bacterial colonies based on chemical information obtained from spectroscopic imaging.

While spectroscopic imaging is useful for acquiring chemical information in two spatial dimensions, it inherently loses height information, which is important for investigating chemical reactions proceeding in three dimensions. This information is of key importance in the growth monitoring of bacterial colonies. Thus, novel illumination optics have been coupled with a spectroscopic imaging setup to probe the surface topography of a 3D sample while being able to acquire spectroscopic information simultaneously.

Preface

Further information on chemometric methods used in this dissertation can be found in the supplemental material of a first-author publication in the Journal of Chemical Education [1]. The spectroscopic topics discussed in Chapters 2 - 4 are adapted from first-author manuscripts in Applied Spectroscopy and Analytical Chemistry, respectively [2], [3]. All alterations are implemented in order to reflect the appropriate style of this dissertation.

Table of Contents

List of Figures

List of Tables

Table Page

Chapter 1

Introduction: Spectroscopic Imaging for Detection and Identification of Bacterial Contaminations

Each year, around 76 million cases of food borne diseases occur in the United States, as reported by the World Health Organization (WHO) [4]. While many of these cases are not fatal, an estimated five thousand people die as a result of these infections. If the infection is caught and treated early enough, these deaths can often be prevented. In order to treat effectively, the source of infection must be identified both correctly and quickly. If the source is not identified *correctly*, a wrong course of treatment could be prescribed that may not be helpful or could lead to antibiotic resistance. Traditional serological methods of bacterial identification, such as polymerase chain reaction [5], [6], are able to identify a source correctly, but are lacking when it comes to speed. Thus, there is a need for methods that can accomplish this task by meeting both requirements.

Although different strains of bacteria are composed of similar chemical components, there are small chemical differences between strains that can be used to discriminate and identify them. In recent years, new detection and identification strategies based on analytical instrumentation have been developed, focusing on utilizing these chemical differences. These methods have the advantage of being fast, typically providing information in hours rather than days. Methods involving electrochemistry, mass spectrometry, and spectroscopy have been shown to be useful for the detection and identification of various types of bacteria. In this dissertation, a developed method using Fourier transform infrared (FTIR) spectroscopy and correlation analysis will be presented for the discrimination of three spectroscopically similar strains of *E. coli* bacteria.

Karasinski et al. [7] have used dissolved oxygen (DOX) sensors to monitor the respiratory activities of different bacteria via oxygen consumption. By investigating five different types of bacteria (*E. coli, Escherichia adecarboxylata, Comamonas acidovorans, Corynebacterium glutamicum,* and *Staphylococcus epidermis*), they have

tested a method of differentiation. By adding non-lethal concentrations of broad-spectrum antibiotics to the growth medium for each type of bacteria, bacterial growth and oxygen consumption was partially inhibited. When compared to oxygen consumptions in antibiotic-free media, the effect of antibiotic addition was shown to be unique for the each type of bacteria investigated, giving rise to unique oxygen consumption fingerprints. By building a library of oxygen consumption fingerprints, unknown oxygen-consuming bacteria can be identified, provided they are contained in the library. Operating conditions such as growth medium and temperature must be controlled and consistent, as unknown bacteria grown under conditions that are not calibrated for could lead to incorrect identification.

In order to selectively detect a certain type of bacteria among a mixed bacterial population, Neufeld et al. [8] have incorporated a bacteriophage (phage). A phage acts as a parasite or virus that will infect certain bacterial species, causing the release of specific intracellular enzymes that can be monitored by amperometric measurements. This combination of phage used *and* enzyme released is strain-specific. In this study, bacteriophage lambda (λ) has been used to infect *E. coli K12*, causing the release of β-D-galactosidase during cell lysis. This enzyme is measured amperometrically by monitoring its oxidation at a carbon anode. The authors have shown that, in a mixture of *E. coli K12* and *Klebsiella pneumoniae*, the combination of phage used and enzyme released is specific for *E. coli K12*, as *K. pneumoniae* showed no enzymatic activity. By changing the combination of phage used and enzyme monitored, the detection of a wide variety of bacteria can be accomplished. Filtering and preincubating the bacterial strains before infection with the phage allowed the detection of concentrations of *E. coli K12* as low as 1 colony-forming unit (cfu)/mL within six to eight hours. The concentration of bacteria detected is two orders of magnitude lower than that detected by a method involving a fluorescent bacteriophage assay (100 cfu/mL) [9]. This method has the advantage of being able to detect certain species of bacteria within a mixture and is not limited to oxygen-consuming bacteria, as in [7].

Matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF-MS) has been used in the discrimination of environmental strains of *E. coli* as well as classifying them according to source [10], identifying clinical isolates of

Staphylococcus aureus by comparison with reference strains [11], and for the identification of gram-negative bacteria isolated from cystic fibrosis patients [12]. MALDI-TOF-MS methods of bacterial identification allow a high throughput in a short amount of time with minimal sample preparation, which is enticing for clinical applications. However, like electrochemical methods, MALDI-TOF-MS is destructive to the sample, rendering the sample unfit for further analyses. This can pose a problem if only a small amount of sample is available for analysis. Siegrist et al. [10] have focused on using MALDI-TOF-MS as a method of bacterial source tracking (BST) in order to identify possible sources of surface water contamination by *E. coli* from fecal input. As the source of *E. coli* is directly related to the health risks associated, the authors have used a model system consisting of *E. coli* isolates from both human and animal sources, looking for potential biomarkers that are unique to particular source groups. The animal source groups consisted of avian, bovine, and canine. Analysis of the mass spectra led to 73% of isolates being correctly grouped by their source, with 100% of the avian, bovine, and canine isolates grouped correctly. Of particular interest was the identification of five potential avian-specific biomarkers based on mass spectra. Rajakaruna et al. [11] have performed MALDI-TOF-MS on intact *Staphylococcus aureus* cells from 134 field isolates. The authors' aim was to collect these field isolates and, with minimal sample preparation, identify them correctly when compared to a database comprised of mass spectra from reference strains. By taking the entire mass spectrum into account, rather than focusing on specific mass ions, 130 isolates were successfully identified as *S. aureus* by comparison with the reference mass spectra. The spectral patterns produced by species that are closely related to *S. aureus* were also shown to have a high degree of similarity, though genus-specific biomarkers were found. Degand et al. [12] have built a reference database consisting of mass spectra from 58 species of gram-negative bacteria that lead to infection in cystic fibrosis patients. This database has been used to identify 559 clinical isolates of different types of bacteria. By determining the percentage of common mass ion peaks between the clinical isolate and the mass spectra contained in the database, 549 isolates were identified correctly. In most cases the percentage of common peaks averaged 80% or higher. MALDI-TOF-MS techniques, while fast, are destructive and require mixing with a protective matrix, thus increasing the amount of sample handling needed.

Spectroscopic methods such as laser-induced breakdown spectroscopy (LIBS) [13] and Raman Spectroscopy [14], [15] have also been used for bacterial identification. LIBS is a type of atomic emission spectroscopy that uses a laser pulse as an excitation source, ablating a very small amount of material, typically in the range of nanograms or picograms. Because such a small amount of material is consumed, the process is considered essentially non-destructive. Diedrich et al. [13] have used LIBS to discriminate between enterohemorrhagic *E. coli* O157:H7 (EHEC), which is pathogenic to humans, and three nonpathogenic strains of *E. coli* (two strains of K12 and one strain of C). These bacteria have specific outer membrane compositions containing divalent cations, such as Mg^{2+} and Ca^{2+}. The dominating spectral features arise from the ionized and neutral forms of Mg and Ca. Emission lines from other trace inorganics such as iron, potassium, sodium, manganese and phosphorus, etc. will also be present. The differences detected by LIBS analysis are not genetic differences, such as the elements comprising DNA or proteins. Rather, the differences between the chemical compositions of the outer membranes, which vary between different species of bacteria, are detected. By combining LIBS measurements with a discriminant function analysis (DFA), the authors have shown that significant differences exist between EHEC and one of the nonpathogenic strains (strain C), allowing the discrimination of the two strains. The authors have also reported the effects of changing the nutrient media used to culture bacteria on the LIBS spectra. The same type of bacteria grown in two different media had a higher similarity than did different types grown in the same media, showing that LIBS is a practical technique for identifying bacteria grown in a variety of environments.

Xie et al. [14] have used a combination of laser tweezers and confocal Raman spectroscopy (LTRS) to optically trap and identify individual bacteria in an aqueous environment. Confocal Raman spectroscopy has the advantage of collecting the Raman signal from only the focal region of the objective while any out-of-focus signals, such as background signals arising from the environment, are largely suppressed. This provides a high spatial resolution, enabling the detection of spectral features from single cells. In order to investigate a single cell within an aqueous environment, the cell must be immobilized. The authors have used optical trapping, which captures an individual moving cell in the focus of the laser beam and levitates it above the substrate, thereby maximizing the collection of Raman scattering and minimizing unwanted background

signal. This provides a high signal-to-noise ratio. Using six different species of bacteria, the authors have investigated the capability of confocal LTRS for the discrimination and identification of various species during different growth phases. Using a principal component analysis (PCA), cell cultures that were synchronized at the log phase could be well-discriminated from one another. Bacterial species taken from unsynchronized cell cultures at the stationary phase could also be discriminated using PCA and hierarchical cluster analysis (HCA). A general discrimination analysis (GDA) was used to classify bacterial species from unsynchronized cell cultures from random growth phases. Although the single-cell spectra were shown to fluctuate to some degree among the individual cells, the cells from different species during different growth phases could be discriminated.

Patel et al. [15] have used surface enhanced Raman spectroscopy (SERS) to identify closely-related members of the *Bacillus cereus* group of bacteria, including *Bacillus anthracis* (Sterne and Ames), *thuringiensis,* and *cereus*. When a molecule is in close proximity to a nanostructured metal surface, an enhancement in Raman scattering is observed. This signal enhancement permits the investigation of single bacterial cells on a very short time scale without the need for sample growth or labeling. PCA and HCA were performed on four sets of data: 1) normalized SERS spectra themselves, 2) first derivatives of the spectra, 3) second derivative spectra, and 4) binary barcodes based on the sign (+/-) of the second derivative spectra. Using the normalized and first derivative spectra, the *B. anthracis* Sterne and *B. cereus* strains formed distinct clusters from one another. However, the *B. anthracis* Ames and *B. thuringiensis* clusters overlapped and could not be differentiated. When using the second derivative spectra, better cluster separation is achieved, though the intragroup distances are large, meaning that they are somewhat spread out. The best clustering results were achieved using the second derivative-based barcoding approach. Here, barcodes are assigned based on the sign of the second derivative (+1 for positive second derivatives, 0 for negative second derivatives).The clusters corresponding to the four groups are well defined and separated, with smaller intragroup distances and larger intergroup separations. Once it was found that these groups could be discriminated well from one another, the authors tested the barcoding method for its ability to positively identify in-class unknowns and to minimize false positive identification of out-of-class unknowns. Using *B. anthracis* Sterne

and *B. anthracis* Ames as in-class unknowns, each falls into the correct strain cluster, thus providing positive identification. *E. coli, B. licheniformis,* and a different strain of *B. anthracis* Ames were used as out-of-class unknowns to test for false positive identification. In each case, the unknown does not find a match with any of the known clusters.

For the work presented in this dissertation, FTIR spectroscopy will be utilized for detecting and identifying small bacterial colonies within extended sample areas in which they grow. FTIR spectroscopy has distinct advantages for the field of bioanalysis: unlike electrochemical or mass spectrometric techniques, FTIR is nondestructive in nature and there is no need to alter the samples, such as mixing with a matrix, before analysis. Like other spectroscopic techniques, acquisition times are short, allowing the analysis of many samples in a short period of time. FTIR is easily introduced into clinical settings, as little sample handling is required, and the use of expensive reagents and extensive personnel training is not needed. For these reasons, FTIR spectroscopy has become a useful tool within the field of bacteriology, for which it has been shown to work well [16]-[20]. This dissertation will assess the capability of spectroscopic imaging to enable the detection and identification of bacterial contaminations.

Discrimination or identification of bacterial colonies can be achieved using FTIR spectroscopy if the location of such colonies is known. As bacterial colonies are often small and localized within a larger sample, it can be a challenge to find them. To accomplish this, extended sample areas must be probed with a high spatial resolution. FTIR can quickly probe one sample location, so it would be advantageous to utilize FTIR spectroscopy in conjunction with sampling many locations at one time. To this end, spectroscopic imaging [2], [3], [21], [22] has emerged, replacing a conventional detector with an infrared camera. Each pixel of this camera acts as a detector, providing both spatial and spectral information, and is capable of performing many FTIR measurements at one time with a high spatial resolution. This enables the use of sophisticated chemometric methods, potentially *in situ*, to obtain the spatial distributions of different chemical components. This spatial distribution of chemical information can also be represented visually by means of a multivariate image analysis (MIA) [23], [24], serving as a way to reinforce the chemometrics. A MIA is a visualization technique that displays

the spectroscopic signatures of different materials as different colors. This will enable the visual discrimination of *E. coli* bacterial colonies from a substrate by means of visibly different colors. This method can be applied to the localization of bacterial colonies within an extended area by merging spectroscopic data cubes.

E. coli is one among many types of bacteria that can contaminate food. To broaden the scope, three strains of *Bacillus* have been included in the investigation, leading to the need to find and identify more than one type of bacteria. Localization was shown to work well using spectroscopic imaging in the first part of chapter three, thus the second part focuses on identifying bacterial colonies. A method of identification has been developed that builds on the correlation analysis presented in chapter two by adding a principal component analysis [25], [26]. By acquiring data cubes from prepared bacterial samples, the identification method will be assessed as to how well it can identify bacterial colonies of known types.

Because chemical reactions can progress in three dimensions, both chemical information in an X-Y plane and topographic information (Z-dimension) are needed for complete analysis. Spectroscopic imaging, however, inherently loses this Z-dimension information, which is important for the growth monitoring of bacterial colonies. Chapter four will present an augmentation to a spectroscopic imaging setup in the form of external illumination optics, allowing samples having complex surface topographies to be investigated while at the same time acquiring chemical information.

Chapter 2

Differentiating Strains of Bacteria Using FTIR Spectroscopy

2.1 Introduction

Food contaminations by *Escherichia coli* bacteria are a major concern for public health [27], [28]. Of the different strains of *E. coli*, there are several harmful strains, some of which cause internal hemorrhaging [29]. Infections caused by each of these strains have their own optimum medical treatment. Thus, these strains must be able to be detected and told apart in order to provide proper treatment, as the wrong treatment may lead to more negative effects than the original infection. Methods of bacterial identification currently in place [5], [6] can easily discriminate these different strains, but are very time-consuming. They require expensive equipment, trained personnel, and ultimately render the sample unfit for further analyses. Presented here is a method utilizing FTIR spectroscopy, as it is fast, equipment is cheaper, requires no extensive personnel training, and is a non-destructive technique. Bacteria contain infrared-active components [30]-[35] that provide very characteristic peaks, good signal-to-noise, and low limits of detection.

The challenge to using FTIR to discriminate bacterial strains comes in the form of selectivity. Because bacteria all contain similar basic components, the spectra of different strains will look similar to one another, as can be seen in Figure 1. However, because the organization of these components is slightly different for each species, the hypothesis presented is here is two-fold: 1) bacteria of the same strain will have the same spectroscopic signatures, leading to high similarity between their spectra; 2) bacteria belonging to different strains will have different spectroscopic signatures and thus there will be a lower similarity between their spectra.

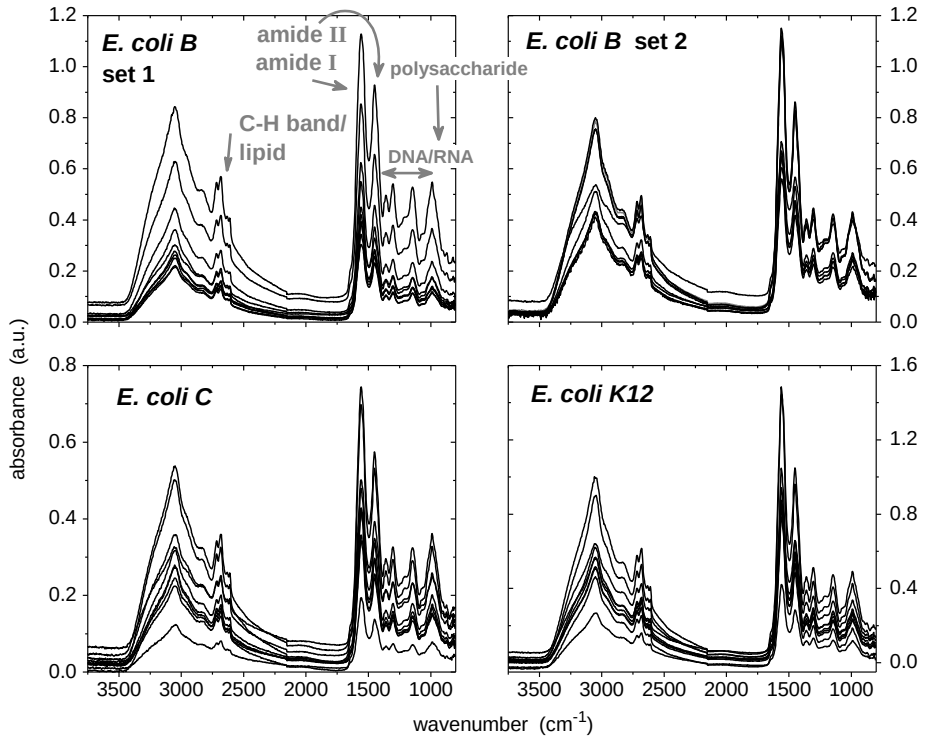

Figure 1: (top row) Comparing spectra acquired from 10 different locations of two different E. coli B samples; from these spectra it had to be assessed whether samples from the same strain are classified as the same. (bottom row) Spectra acquired from two additional strains (C and K12); a successful classification needs to determine that all three different strains have significantly different signatures despite their very similar spectroscopic signatures.

One way to measure the degree of similarity between two spectra is through the use of correlation coefficients [36]. In order to be able to discriminate different strains of bacteria, the correlation coefficient between the spectra of two different strains must be *significantly* different (lower) than between two spectra of the same strain. To indicate whether the correlation coefficients are significantly different, a *t*-test [37] is used. It will be shown here that, though their spectra are similar, three different strains of bacteria can be discriminated.

2.2 Discrimination of Similar Strains of Bacteria by FTIR Spectroscopy

In order to provide bacterial samples for analysis, three different strains of *E. coli* bacteria have been used here (*B*, *C*, and *K12*). For safety purposes, these strains are non-pathogenic to humans and are thus used as a proof-of-principle to show that different strains of the same type of bacteria can be discriminated. Spectra have been acquired from ten locations on two independently prepared samples of each strain to use for analysis.

2.2.1 Analyzing the Similarity of Spectra

Discrimination must be based on spectroscopic signatures, the general shapes of the spectra. This is accomplished by measuring the degree of similarity between spectra of the same strain and comparing it to the degree of similarity between the spectra of different strains. Thus, the degree of similarity between the spectra of *E. coli B* should be significantly higher than the degree of similarity between the spectra of *E. coli B* and *E. coli C*. In order to measure the degree of similarity between two spectra $\mathbf{s}_1$ and $\mathbf{s}_2$ correlation coefficients. are utilized with $\bar{s}_{1,2}$ being the mean values of the two spectra respectively. If a correlation coefficient (1) is close to one, the two spectra are very similar. If it is considerably smaller than one, the spectra have different signatures.

$$r = \mathrm{corr}(\mathbf{s}_1, \mathbf{s}_2) = \frac{\sum_i \left[s_1(v_i) - \bar{s}_1\right] \cdot \left[s_2(v_i) - \bar{s}_2\right]}{\sqrt{\sum_i \left[s_1(v_i) - \bar{s}_1\right]^2 \cdot \sum_i \left[s_2(v_i) - \bar{s}_2\right]^2}}$$

(1)

The correlation coefficients calculated between samples of the same strain will be compared to those calculated between samples of different strains. In order to discriminate two different strains, the correlation coefficient between two spectra acquired from different strains must be *significantly* smaller than the correlation coefficient between two spectra from the same strain.

First, the 'equivalency' of spectra from the same strain must be determined. Because measurement imperfections are unavoidable, correlating spectra from the same strain of bacteria will result in correlation coefficients less than one. Thus, this 'equivalency' establishes a threshold of what is considered a high enough correlation coefficient to indicate that both spectra are from the same strain of *E. coli*. Second, the 'difference' between the spectra from different strains must be determined. Due to the high degree of similarity between such spectra, the correlation coefficients calculated will be close to one. Therefore, the 'difference' will establish a threshold of what is considered a low enough correlation coefficient to indicate two spectra are from different strains. The thresholds determined for both 'equivalency' and 'difference' are based on experimental data obtained.

For growing *E. coli* samples, tryptic soy agar (TSA) plates are prepared (Carolina Biological Supply Company). All three strains were purchased from the same vendor. Plates are streaked with bacteria and incubated at 37 °C for a minimum of 24 hours but no longer than 48 hours. Using an inoculating loop, a small amount of bacteria is removed from a plate and spread thinly on an infrared-transparent calcium fluoride (CaF_2) slide. Because TSA is highly infrared absorbing, care is taken not to remove any TSA from the plate along with the bacteria. Visible images of streaked bacteria can be seen in Figure 2. The left column shows images of bacteria streaks acquired at 4X

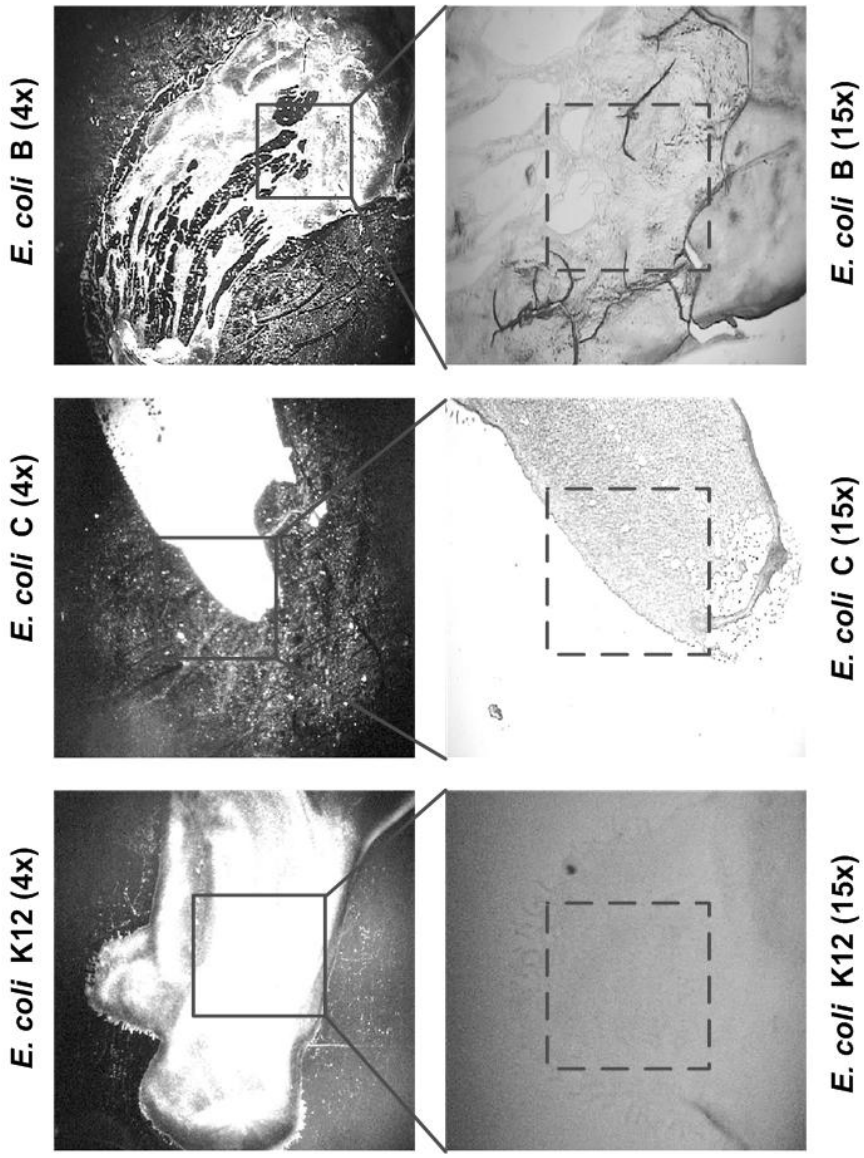

Figure 2: Images of three *E. coli* strains acquired in the visible at 4x and 15x magnification; the squares in the left columns depicts in the 4x images what is seen in the 15x image to its right; the dashed squares in the right image column depict the approximate areas sampled by the IR detector. For all IR measurements a resolution of 4 cm^{-1} is chosen and 32 scans are averaged. Background spectra are acquired from a blank spot on the CaF_2 slide.

magnification. The dashed squares in the right column of images depict the approximate areas that are analyzed at 15X magnification.

After growing and streaking, the slide containing the sample is placed on the stage of a Bruker Optics Hyperion 1000 VIS/IR microscope that is attached to a Vertex 70 FTIR spectrometer. This microscope is equipped with camera for the visible and a single pixel MCT detector. Mid-infrared (MIR) transmission spectra are measured from ten different locations on two samples of each of the three strains, resulting in a total of sixty spectra. In all spectra the wavenumber window from 2400 - 2250cm^{-1} is excluded in order to prevent changes in atmospheric CO_2 from affecting the similarity of the spectra.

Determining 'equivalency' of spectra from the *same* strain: To determine the degree to which the correlation coefficients (1) indicate 'equivalency', correlation coefficients of spectra acquired from the same *E. coli* sample are calculated. This involves calculating the correlation coefficients for all combinations of the ten spectra per sample, i.e. spectrum #1 is correlated to the other nine spectra in the set. Thus, there are a total of 45 different combinations per set, producing 45 correlation coefficients:

$$\begin{aligned} &\mathrm{corr}(\mathbf{s}_1,\mathbf{s}_2)\ ,\ \mathrm{corr}(\mathbf{s}_1,\mathbf{s}_3)\ ,\ \mathrm{corr}(\mathbf{s}_1,\mathbf{s}_4)\ ,\ \dots\ ,\ \mathrm{corr}(\mathbf{s}_1,\mathbf{s}_{10}) && \textit{9 combinations}\\ &\mathrm{corr}(\mathbf{s}_2,\mathbf{s}_3)\ ,\ \mathrm{corr}(\mathbf{s}_2,\mathbf{s}_4)\ ,\ \dots\ ,\ \mathrm{corr}(\mathbf{s}_2,\mathbf{s}_{10}) && \textit{8 combinations}\\ &\qquad\qquad\dots\ , && \\ &\mathrm{corr}(\mathbf{s}_9,\mathbf{s}_{10}) && \textit{1 combination} \end{aligned} \tag{2}$$

Certain combinations are excluded, seeing as they are either redundant, such as in the case of $\mathrm{corr}(\mathbf{s}_i,\mathbf{s}_k) = \mathrm{corr}(\mathbf{s}_k,\mathbf{s}_l)$, or irrelevant, $\mathrm{corr}(\mathbf{s}_i,\mathbf{s}_l) = 1$. Next, it is investigated whether two spectra from two independently prepared samples of the same strain result

in correlation coefficients close to the ones obtained by means of (2), as well as assessing the reproducibility of bacterial preparation. For example, the spectra of *E. coli B* contained in set #1, $\mathbf{s}_{1,(B1)}$, and set #2, $\mathbf{s}_{1,(B2)}$, are correlated, resulting in a total of 100 correlation coefficients:

$$
\begin{array}{ll}
corr\,\mathbf{s}_{1,(B1)},\mathbf{s}_{1,(B2)}\ ,\ corr\,\mathbf{s}_{1,(B1)},\mathbf{s}_{2,(B2)}\ ,\ \ldots\ ,\ corr\,\mathbf{s}_{1,(B1)},\mathbf{s}_{10,(B2)} & \textit{10 combinations} \\
corr\,\mathbf{s}_{2,(B1)},\mathbf{s}_{1,(B2)}\ ,\ corr\,\mathbf{s}_{2,(B1)},\mathbf{s}_{2,(B2)}\ ,\ \ldots\ ,\ corr\,\mathbf{s}_{2,(B1)},\mathbf{s}_{10,(B2)} & \textit{10 combinations} \\
\ldots & \\
corr\,\mathbf{s}_{10,(B1)},\mathbf{s}_{1,(B2)}\ ,\ corr\,\mathbf{s}_{10,(B1)},\mathbf{s}_{2,(B2)}\ ,\ \ldots\ ,\ corr\,\mathbf{s}_{10,(B1)},\mathbf{s}_{10,(B2)} & \textit{10 combinations}
\end{array}
\qquad (3)
$$

The procedures (2), (3) are performed for all three strains *E. coli B, C* and *K12*. For the remainder of this chapter, these sets of correlation coefficients (3) will be referred to as (B1, B2), (C1, C2) and (K12_1, K12_2), respectively.

Determining 'difference' of spectra from two samples of *different* strains: In the next step, a similar procedure correlates spectra from different strains, providing the threshold of 'low enough' correlation coefficients. If these 'inter-strain' correlation coefficients are found to be significantly different (lower) from the ones obtained from spectra from within one strain ('intra-strain'), then it can be concluded that the spectral signatures of the strains are different, thus establishing that they are from different strains. For example, correlating *E. coli B* set #1 and *E. coli C* set #1 (containing ten spectra each) derives the following 100 correlation coefficients, indicated by (B1, C1):

$$corr\left(s_{1,(B1)}, s_{1,(C1)}\right), corr\left(s_{1,(B1)}, s_{2,(C1)}\right), \ldots, corr\left(s_{1,(B1)}, s_{10,(C1)}\right) \quad \textit{10 combinations}$$

$$corr\left(s_{2,(B1)}, s_{1,(C1)}\right), corr\left(s_{2,(B1)}, s_{2,(C1)}\right), \ldots, corr\left(s_{2,(B1)}, s_{10,(C1)}\right) \quad \textit{10 combinations}$$

$$\ldots,$$

$$corr\left(s_{10,(B1)}, s_{1,(C1)}\right), corr\left(s_{10,(B1)}, s_{2,(C1)}\right), \ldots, corr\left(s_{10,(B1)}, s_{10,(C1)}\right) \quad \textit{10 combinations}$$

(4)

This (4) is done for all strain combinations, i.e.:

(B1, C1), (B2, C1), (B1, C2), (B2, C2), (C1, K12_1), (C2, K12_1), (C1, K12_2), (C2, K12_2), (B1, K12_1), (B2, K12_1), (B1, K12_2) and (B2, K12_2

(5)

Here B1, C1, K12_1 are the spectra from set #1 of *E. coli B*, *C*, and *K12*, respectively, and B2, C2, and K12_2 are the spectra from set #2 of each strain.

If the mean correlation coefficients of the spectra from different strains are significantly different (lower) than from spectra of the same strain, the strains can be discriminated based on their spectroscopic signatures. Comparing mean values is done by a *t*-test [37]. However, a *t*-test requires that the data be normally distributed. Correlation coefficients themselves are not normally distributed; the range of correlation coefficients r is restricted to the range from -1 to +1. In some cases, the mean plus one (or more) standard deviations reaches beyond a correlation coefficient of one and are thus meaningless. For this reason, a Fisher's z-transformation [36], [38] is applied to the correlation coefficients r prior to *t*-testing in order to make them more normally distributed. Each correlation coefficient r is z-transformed via:

$$z = \frac{1}{2} \ln\left[\frac{1+r}{1-r}\right]$$

(6)

These means are then *t*-tested instead of the correlation coefficients r themselves. For such a *t*-test, a certain probability of 'acceptable error' must be chosen. This is a probability that a wrong assessment will be made, i.e. stating two means are equivalent when they are not, or stating two means are not equivalent when they are. For this study, a confidence level of 95% had been chosen. When *t*-testing the mean *z*-values, a 'significance value' will be calculated. Using the confidence level of 95%, a significance value greater than 5% will indicate a high probability that both mean *z*-values are equivalent.

2.2.2 Results of Testing for Equivalency or Difference

In determining the 'equivalency' of spectra from the same strain, it is expected that the correlation coefficients produced will be very close to one, indicating that the spectroscopic signatures are the same. The threshold of 'high enough' correlation coefficients can be determined by calculating the mean correlation coefficient among the spectra of one set. For spectra obtained from the same strain, the mean z-values are not expected to be significantly different. In these cases, *t*-testing the mean *z*-values should produce a high significance value (> 5%), indicating a high probability that both mean z-values are equivalent. *T*-testing the mean *z*-values obtained from each of the two sets of *E. coli B, C* or *K12*, i.e. B vs. B, C vs. C, and K12 vs. K12, produced significance values above 5% for each case, indicating no significant differences between the two. Thus, reproducible spectra did result from the repeated sample preparation. The first row of Table 1 summarizes the mean correlation coefficients ± one standard deviation acquired from spectra of strain set #1 among themselves; the second row gives the same information computed from the strain set #2. Here, the 'high enough'

Table 1: (top part) means and standard deviations (sd) of the correlation (1) of spectra acquired from the same strains; (middle part) each correlation coefficient was transformed into z-values (6) by means of a Fisher's z-transform in order to obtain more normally distributed values; (bottom part) *t*-test results analyzing the mean z-values among different samples of the same strain; expected is a high significance (>5% significance level) demonstrating that the mean correlation coefficients are not significantly different.

correlation coeffs.	mean ± sd		mean ± sd		mean ± sd
B1 (intern)	0.9967 ± 0.0027	C1 (intern)	0.9972 ± 0.0022	K12_1 (intern)	0.9970 ± 0.0035
B2 (intern)	0.9947 ± 0.0086	C2 (intern)	0.9975 ± 0.0017	K12_2 (intern)	0.9967 ± 0.0030
B1 vs B2 = (B1,B2)	0.9956 ± 0.0068	C1 vs C2 = (C1,C2)	0.9961 ± 0.0034	K12_1 vs K12_2 = (K1,K2)	0.9965 ± 0.0031
Fisher's z values					
B1 (intern)	3.3732 ± 0.4433	C1 (intern)	3.4792 ± 0.4853	K12_1 (intern)	3.6556 ± 0.7351
B2 (intern)	3.5614 ± 0.7537	C2 (intern)	3.4832 ± 0.4538	K12_2 (intern)	3.4548 ± 0.5734
	significance		significance		significance
B1 vs B2	15.2%	C1 vs C2	96.8%	K12_1 vs K12_2	15.1%

correlation coefficients are at least 0.99 (first and second rows of Table 1), which does indicate a high degree of similarity among the spectra of one set of one strain as expected. The third row in Table 1 shows that the mean correlation coefficients obtained by correlating the spectra in set #1 with the spectra in set #2 are also high and close to the ones obtained by correlating the spectra of set #1 and set #2 internally. The fourth and fifth row give the mean *z*-values for each set internally, and the sixth row provides the significance value as calculated by *t*-testing the mean *z*-values.

In determining the 'difference' of spectra from different strains, it is expected that, while the correlation coefficients are close to one (Figure 1), they will be significantly lower than those obtained between spectra of the same strain. The mean correlation coefficients will determine the threshold of 'low enough' correlation coefficients so as to say that two spectra are from different strains. The mean correlation coefficients ± one standard deviation, for all combinations of different strains (5) are concluded in the top parts of Table 2 and Table 3. Here it can be seen that, while the mean correlation coefficients obtained by correlating the spectra of different strains are high, in most cases they are lower than the means obtained from the correlation of spectra from the same strain. Thus, it must be determined whether or not two different strains can be discriminated. It is expected that there will be a significant difference between the spectra of different strains. For this purpose, the 'intra-strain' *z*-values from spectra of a certain strain (e.g. B1 or B2 (intern), middle of Table 1) are *t*-tested against all *z*-values from 'inter-strain' correlations (4), (5) (middle of Table 2 and Table 3). For example, the *z*-value from internal B2 is *t*-tested against the 'inter-strain' correlation B2 vs. C2. This is done to investigate whether the correlation coefficients obtained by correlating the spectra of different strains are significantly different (lower) from those obtained by correlating the spectra from the same strain. In these cases, the significance is expected to be less than 5%. This has been confirmed for all possible combinations except for B1 versus (B1, C2), showing that the spectra of different strains of *E. coli*, though similar, can be discriminated based on their spectral features.

Table 2: (top part) The correlation coefficients ± one standard deviation were obtained from correlating spectra of different strains (compare (4) and (5)); (middle part) each correlation coefficient was transformed into z-values (6) by means of a Fisher's z-transform; (bottom part) *t*-tests comparing the mean correlation coefficients obtained from the samples: B1 (intern), C1 (intern) and K12 (intern) (see Table 1) and from different strains. If spectra from different strains show a lower similarity than spectra obtained from one sample, here, a significance value <5% is expected; in only one case this hypothesis was not supported.

correlation coeffs.	mean ± sd		mean ± sd		mean ± sd
(B1,C1)	0.9949 ± 0.0038	(B1,C1)	0.9949 ± 0.0038	(B1,K12_1)	0.9951 ± 0.0038
(B1,C2)	0.9964 ± 0.0023	(B2,C1)	0.9949 ± 0.0046	(B2,K12_1)	0.9927 ± 0.0100
(B1,K12_1)	0.9951 ± 0.0038	(C1,K12_1)	0.9901 ± 0.0058	(C1,K12_1)	0.9901 ± 0.0058
(B1,K12_2)	0.9955 ± 0.0037	(C1,K12_2)	0.9906 ± 0.0058	(C2,K12_1)	0.9947 ± 0.0031
Fisher's z values					
(B1,C1)	3.0883 ± 0.3119	(B1,C1)	3.0883 ± 0.3119	(B1,K12_1)	3.1413 ± 0.3742
(B1,C2)	3.2609 ± 0.3343	(B2,C1)	3.1186 ± 0.3527	(B2,K12_1)	3.0204 ± 0.4125
(B1,K12_1)	3.1413 ± 0.3742	(C1,K12_1)	2.7172 ± 0.2542	(C1,K12_1)	2.7172 ± 0.2542
(B1,K12_2)	3.1902 ± 0.3800	(C1,K12_2)	2.7558 ± 0.2826	(C2,K12_1)	3.0503 ± 0.3077
	significance		significance		significance
B1 vs (B1,C1)	0.0%	C1 vs (B1,C1)	0.0%	K12_1 vs (B1,K12_1)	0.0%
B1 vs (B1,C2)	*9.4%*	C1 vs (B2,C1)	0.0%	K12_1 vs (B1,K12_1)	0.0%
B1 vs (B1,K12_1)	0.1%	C1 vs (C1,K12_1)	0.0%	K12_1 vs (C1,K12_1)	0.0%

Table 3: Similar results as shown in Table 2 - here correlation coefficients derived from sample sets #2 are compared to all combinations of ‘inter-strain’.

correlation coeff.	mean ± sd		mean ± sd		mean ± sd
(B2,C1)	0.9949 ± 0.0046	(B1,C2)	0.9964 ± 0.0023	(B1,K12_2)	0.9955 ± 0.0037
(B2,C2)	0.9955 ± 0.0065	(B2,C2)	0.9955 ± 0.0065	(B2,K12_2)	0.9933 ± 0.0099
(B2,K12_1)	0.9927 ± 0.0100	(C2,K12_1)	0.9947 ± 0.0031	(C1,K12_2)	0.9906 ± 0.0058
(B2,K12_2)	0.9933 ± 0.0099	(C2,K12_2)	0.9947 ± 0.0032	(C2,K12_2)	0.9947 ± 0.0032
Fisher's z value					
(B2,C1)	3.1186 ± 0.3527	(B1,C2)	3.2609 ± 0.3343	(B1,K12_2)	3.1902 ± 0.3800
(B2,C2)	3.2992. ± 0.4424	(B2,C2)	3.2992 ± 0.4424	(B2,K12_2)	3.1286 ± 0.4678
(B2,K12_1)	3.0204 ± 0.4125	(C2,K12_1)	3.0503 ± 0.3077	(C1,K12_2)	2.7558 ± 0.2826
(B2,K12_2)	3.1286 ± 0.4678	(C2,K12_2)	3.0582 ± 0.3226	(C2,K12_2)	3.0582 ± 0.2256
	significance		significance		significance
B2 vs (B2,C1)	0.0%	C2 vs (B1,C2)	0.1%	K12_2 vs (B1,K12_2)	0.1%
B2 vs (B2,C2)	0.1%	C2 vs (B2,C2)	2.3%	K12_2 vs (B2,K12_2)	0.0%
B2 vs (B2,K12_1)	0.0%	C2 vs (C2,K12_1)	0.0%	K12_2 vs (C1,K12_2)	0.0%
B2 vs (B2,K12_2)	0.0%	C2 vs (C2, K12_2)	0.0%	K12_2 vs (C2,K12_2)	0.0%

2.3 Conclusions

Out of many strains of *E. coli* bacteria, there are several that are pathogenic to humans, each having their own best course of medical treatment. In order to treat a bacterial infection effectively, *E. coli* contaminations of food must be identified quickly and accurately. Traditional methods of bacterial identification, while able to identify contaminations accurately, are severely lacking when it comes to time. Thus, FTIR spectroscopy has been used in this study, as it is a faster technique. However, there is a challenge to FTIR spectroscopy, as the spectra of different strains of bacteria look very similar to one another. Using a correlation analysis, it has been shown here that there are statistically significant differences in the spectroscopic signatures of different strains of bacteria that allow them to be discriminated.

Chapter 3

Localization & Identification of Bacterial Contaminations Using Spectroscopic Imaging

3.1 Introduction

It has been shown that different strains of *E. coli* bacteria, while having spectra that are similar to one another, can be discriminated based on their spectral features using FTIR spectroscopy. Performing such analyses requires knowing the location of such bacteria. On food samples, however, bacterial colonies are small and localized within a larger sample area. The challenge then becomes finding these colonies. Thus an entire sample must be investigated with a high spatial resolution. Probing one sample location using FTIR spectroscopy yields chemical information quickly. Therefore, a method combining the advantages of FTIR spectroscopy and the ability to sample many locations at one time is needed. For this reason, spectroscopic imaging [2], [3], [21], [22] has emerged, combining spectroscopy with imaging. This technique can be utilized to both find and identify bacterial colonies within an extended area. The first part of this chapter presents the localization of bacterial colonies, while the second part focuses on the identification of found colonies.

3.2 Localization of Bacterial Colonies

Spectroscopic imaging uses an infrared camera instead of a conventional detector. Each pixel of this camera acts as a detector, performing four thousand spectroscopic measurements at one time, making the technique fast. Spectroscopic imaging also enables the localization of samples, thereby providing information as to where an analyte is located within a sample. Because spectroscopic imaging supplies a wealth of information at one time, it can become useful to visualize this information. One method

of visualization is the depiction of different materials as different colors by relating the chemical properties of a material to a certain color. To achieve this relation, a characteristic signature within the chemical properties of a material must be found. This can be accomplished by utilizing a principal component analysis (PCA) [25], which can extract the spectroscopic features from a set of materials and provide a representation of the chemical properties; called principal components (PCs). A PCA will also provide information as to how much each PC contributes in a given material, called scores. It is important to note that the PCs and scores have no direct physical or chemical meaning. Instead, they are linear combinations of the physical/chemical properties of the set. As mentioned in chapter two, different strains of bacteria are composed of similar chemical components, i.e. functional groups, with slightly different mixtures of these components. These different mixtures will be represented differently by different linear combinations of the PCs.

Once the characteristic signature within the chemical properties of a material has been found, the next step is relating the chemical properties of the material to a certain color, thus combining PCA and imaging. One way to achieve this combination is through the use of a multivariate image analysis (MIA) [23], [24]. A MIA utilizes the scores produced by a PCA, compiling them into images, called score images, which represent the values of the scores at each pixel of the infrared camera. By color-coding these values, a color image is produced that represents the different chemical properties of a material as different colors, allowing the visualization of the information obtained by spectroscopic imaging.

Spectroscopic imaging makes use of a 2D detector array, called a focal plane array or FPA [21], containing many detectors, called ‘pixels’. Because each pixel acquires a spectrum simultaneously, the collection of chemical information from different sample locations is enabled. This chemical information is arranged into the form of a ‘data cube’ of information. This data cube has two spatial dimensions (X- and Y-dimensions), which provide information about location (Figure 3). The third dimension of this cube is a spectral dimension (Z-dimension), which contains chemical information. Thus, chemical information is acquired at many sample locations. For the setup used in this study, the FPA contains 4,096 individual pixels, having a 64X64 arrangement, with an infrared

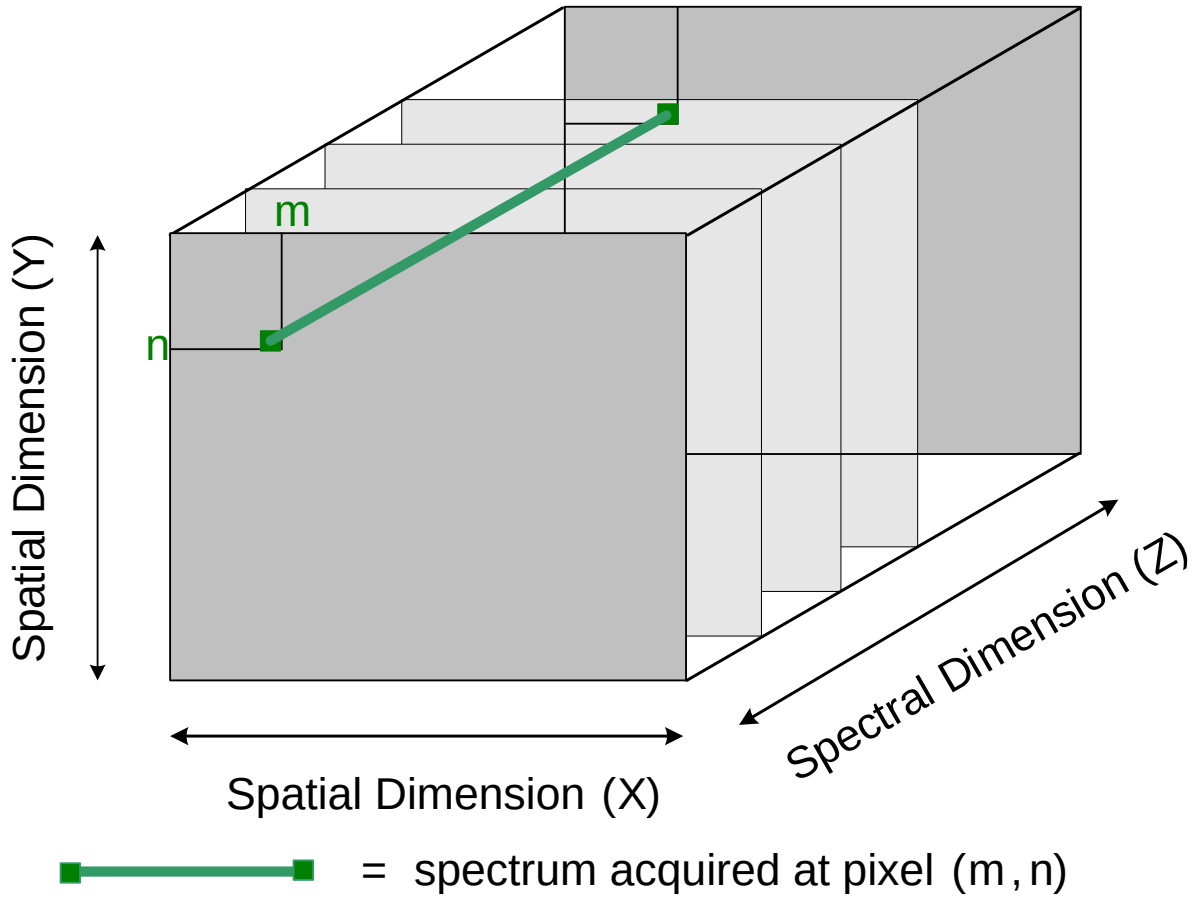

Figure 3: Diagram of a data cube produced using spectroscopic imaging. The X- and Y-dimensions are spatial dimensions, providing information about the location of a sample. The Z-dimension is a spectral dimension, providing a spectrum at each pixel. The transparent boxes indicate data acquired at different wavelength positions.

spectrum acquired at each pixel. A diagram of a data cube can be seen in Figure 3. Here, the pixel highlighted in green represents a spectrum acquired at a certain location. The transparent boxes shown represent wavelength positions contained in a spectrum. In this study, the same three strains of *E. coli* as used in chapter two were used to show the feasibility of using spectroscopic imaging to localize bacterial colonies. Bacterial streaks were prepared on CaF_2 in the same way. Spectroscopic imaging measurements were acquired from three *E. coli* colonies in transmission mode. To show that localization of bacterial colonies within an extended sample area is possible, the data cubes produced from these three measurements were merged together. It is expected that, after performing a MIA on this merged cube, the bacterial colonies will be clearly distinguishable from the substrate. Because three different strains have been used, the colors of the colonies will be different, though the color should be homogeneous within one colony. Finally, it is expected that the substrate will be a uniform color, as it is the same in each of the merged data cubes. The top of Figure 4 shows visible images of the three colonies investigated. The dashed boxes in each image depict the approximate area that is sampled by the infrared camera. The bottom shows these areas after MIA has been performed on the merged cube. It can be seen that each bacterial colony is distinguishable from the substrate and that each colony is mostly homogeneous in color within itself. The colony of *E. coli B* has been encoded in a yellow-green color, *E. coli C* has been encoded in a dark pink, and *E. coli K12* has been encoded in a lighter pink. The CaF_2 substrate in each of the three merged images is a uniform green color, which is expected.

3.3 Conclusions of Localization

Because bacterial colonies are often small and localized within a larger sample area, they must be found in order to be identified. Spectroscopic imaging has been shown here to be applicable to the localization of these bacterial colonies by visualizing the chemical information acquired. Visualization requires the relation of the chemical properties of a material to a certain color.

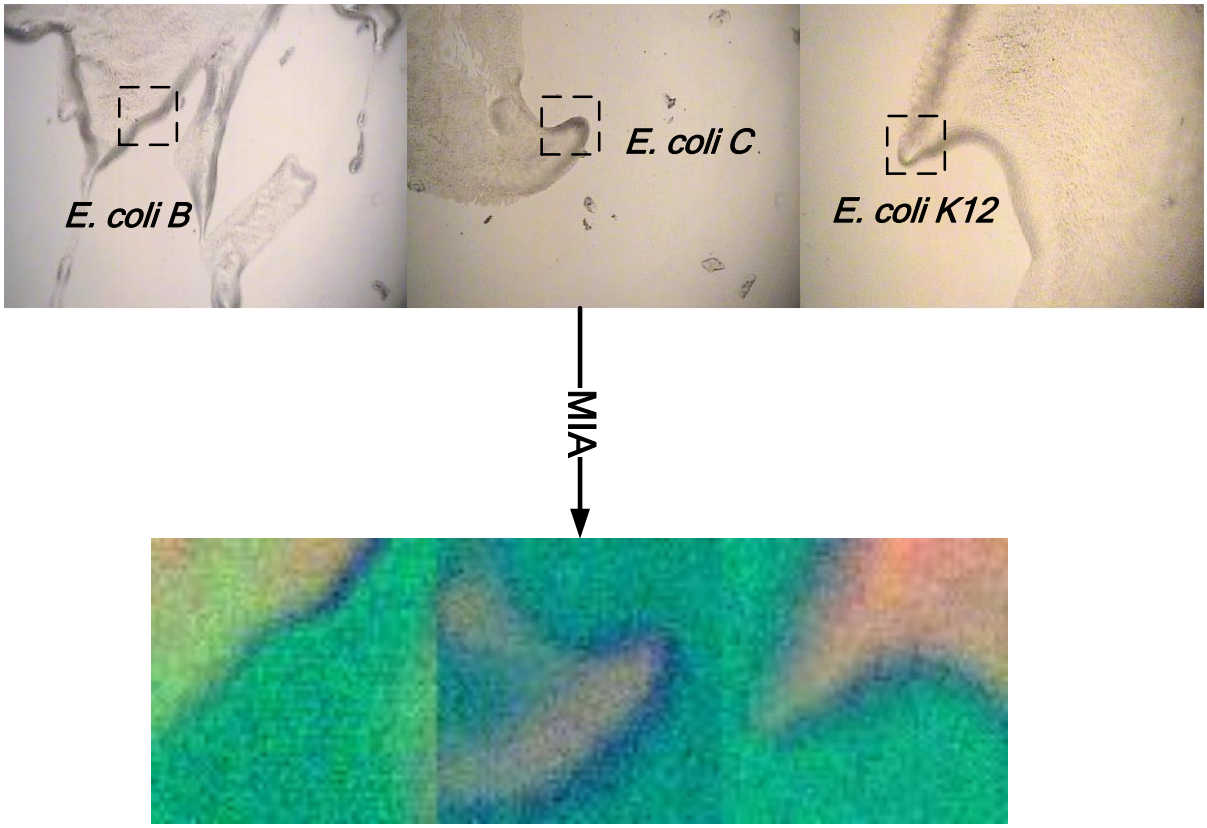

Figure 4: (top) Visible images from three *E. coli* colonies. The dashed boxes indicate the approximate area that the infrared camera samples in the acquisition of a data cube (bottom) The corresponding color images after a MIA has been performed on data cubes acquired from (top). It can be seen that the bacterial colonies in each color image can be clearly distinguished from their substrates, and that each is mostly homogeneous in color among itself. It is also important to note that the CaF_2 substrate in each image is a uniform green color, indicating a good reproducibility.

This relation necessitates finding a characteristic signature within the chemical properties, and can be accomplished by means of a principal component analysis. Once the characteristic signature has been found, the chemical properties of a material can be related to a certain color by means of a multivariate image analysis. This allows the visualization of chemical information by means of a color image. It has been shown here that, by performing a MIA on three data cubes of *E. coli*, bacterial colonies could be localized within a larger sample area and were clearly distinguishable from their common substrate.

3.4 Identification of Bacterial Colonies

It has been shown that different strains of *E. coli* can be discriminated from one another based on their spectroscopic signatures using FTIR spectroscopy. *E. coli* bacterial colonies can also be localized within a larger sample area through the use of spectroscopic imaging and multivariate image analysis. While *E. coli* is a very common food contamination, there are others, such as strains of *Bacillus*, that are equally as harmful [39]. It is important to be able to discriminate these different types of bacteria, as well as identify them. To demonstrate that spectroscopic imaging can be used to identify a variety of bacterial colonies, three strains of *Bacillus* are used here along with the same three strains of *E. coli* used previously.

By including *Bacillus* in the investigation, there is no longer only one species to account for. There are now two species of bacteria, each of which have their own different strains that need to be found and identified. A visible image of colonies from each strain can be seen in Figure 5. Because localization has been discussed earlier in this chapter, this section will focus on how well found colonies can be identified. This will be accomplished by acquiring several data cubes from prepared bacterial samples, meaning that the species and strain is known. An identification algorithm that builds on the correlation analysis presented in chapter two will be applied and how well the method identifies these known data cubes will be assessed. Before applying the identification algorithm, one thing must be addressed. This is the issue of drifting baselines in the spectra, as

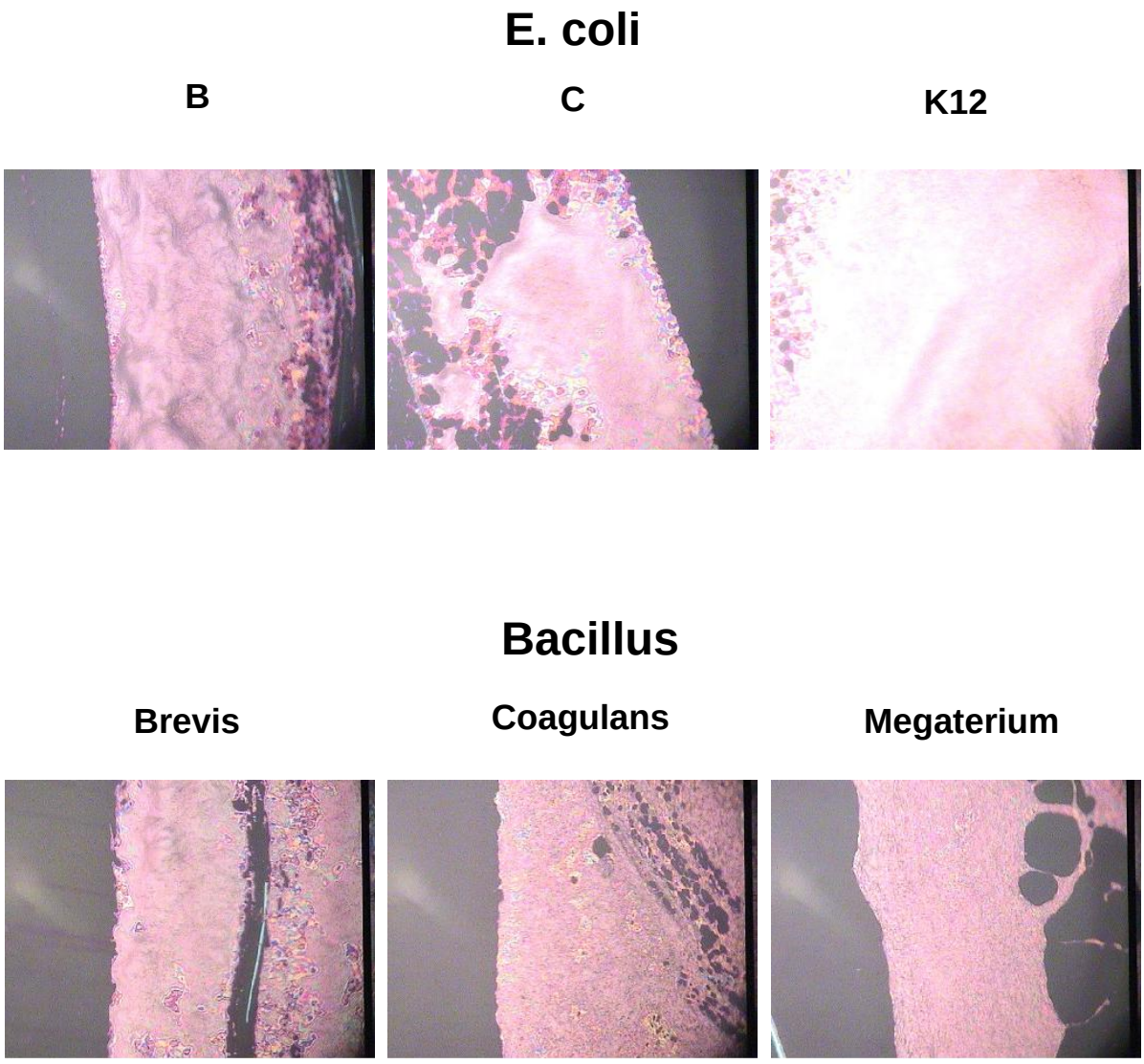

Figure 5: Visible images of bacterial streaks on an IR reflecting slide. The bacterial streaks are seen as pink areas. The pink color here is due to the reflective coating of the slide.

these could affect the outcome of identification. Thus, a method of baseline correction is applied to all measured spectra.

3.5 Experimental

Bacteria samples for this study are prepared in much the same way as in Chapter 2. Here, six strains of bacteria (*E. coli B*, *C*, and *K12* and *Bacillus brevis*, *coagulans*, and *megaterium*) are prepared and spread thinly on an infrared-reflecting slide (Kevley Technologies). Visible images of these colonies can be seen in Figure 5, having a pink color as opposed to the white color in chapter 2. This pink color is due to the infrared-reflective coating on the slide.

Figure 6a shows ten MCT spectra from *E. coli C*. Here, it can be seen that there are drifting baselines, which cause the spectra to change inconsistently. Therefore, a method of baseline correction has been developed in-house [40]-[42] and has been utilized here. The hypothesis of the baseline correction method is that the absorbance bands in a spectrum are more localized than the baseline drifts and the correction method involves two steps: the first step is an initialization; the spectrum is fit by a polynomial. However, using a polynomial as the best estimate of the baseline e.g. can result in drift over-compensation. Thus, a second optimization step is needed. This optimization will correct wavenumber regions where the absorptions distract the fit from the true baseline, making the baseline estimate more similar to the spectrum, while avoiding that the baseline models the absorbance features. The optimization is an iterative approach with two antagonizing criteria: 1) at each wavelength position, the difference between the baseline estimation and the drifting spectrum is partially reduced, iteratively making the estimation more similar to the spectrum; 2) the estimation is made straighter piecewise in order to avoid modeling the absorbance features. This combination of criteria will ensure that the baseline estimation follows the general trend of the drifting spectrum without modeling the absorbance features. The result of baseline correction on the spectra in Figure 6a can be seen in Figure 6b. It is evident that the

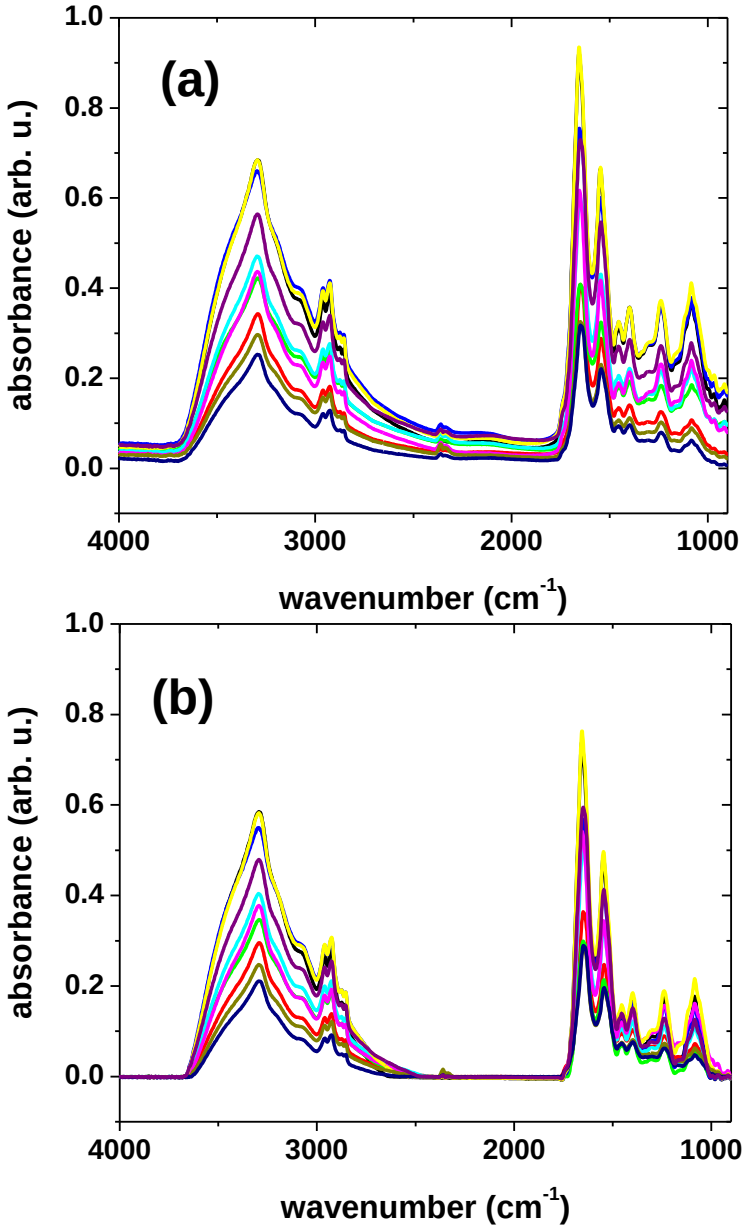

Figure 6: (a) 10 spectra (out of 100) of *E. coli C* are shown here, acquired in reflection mode. It can be seen that all 10 spectra have similar shapes, though they have drifting baselines, causing dissimilarity among spectra of the same strain. (b) The same 10 spectra after applying baseline correction. Here, the baseline for each spectrum is centered at zero absorbance units.

baseline drifts have been corrected, with the baseline for each spectrum centered at zero.

Representative MCT spectra from each of the six strains after baseline correction can be seen in Figure 7. It can be seen in (a) that the spectra from the three strains of *E. coli* look similar to one another, as is expected from the discussion in chapter 2. (b) shows spectra from the three strains of *Bacillus*. There are prominent differences between the spectra of these strains, most notably in the region of 1750cm^{-1}, where *megaterium* has a strong absorbance that is not as evident in *brevis* or *coagulans*. There are also clear spectral differences in the 'fingerprint region' of these three strains. (c) shows spectra from the three strains of *E. coli* and a spectrum from *Bacillus brevis*. It is evident that the spectrum of *brevis* is similar to the spectra from the strains of *E. coli*.

Looking closer at the MCT spectra (Figure 8), it can be seen that, within the spectra of one strain, there are different overall shapes, most notably in the 1650cm^{-1} region (Amide I/protein region [31]). The black and green spectra have a broad shape, while the shape clearly changes when looking at the red, blue, and purple spectra. These shape differences are hypothesized to arise from differing contributions from a number of possible secondary protein structures [43-45], including parallel and anti-parallel β-sheets, β-turns, α-helix, and random coils. Each type of structure will have a slightly different carbonyl stretching frequency, due to the fact that each structure has a distinct molecular geometry and hydrogen-bonding pattern. The shape changes in the blue and purple spectra are most likely occurring due to a larger contribution from the β-turns at 1675cm^{-1}. The overall shape of the Amide I peak can be described as a combination of these different contributions and can be thought of as overlapping absorbances. These differing shapes occur in every set of spectra acquired, meaning that there is no one spectrum that can be assigned to each of the different strains. Instead, the set of spectra for each strain contain these inherent differences, which must be taken into account. For this reason, a principal component analysis (PCA) [25] is utilized in the identification method, as it is able to model these differences and incorporate them into a calibration model for each species/strain, using several PCs to model these individual changes.

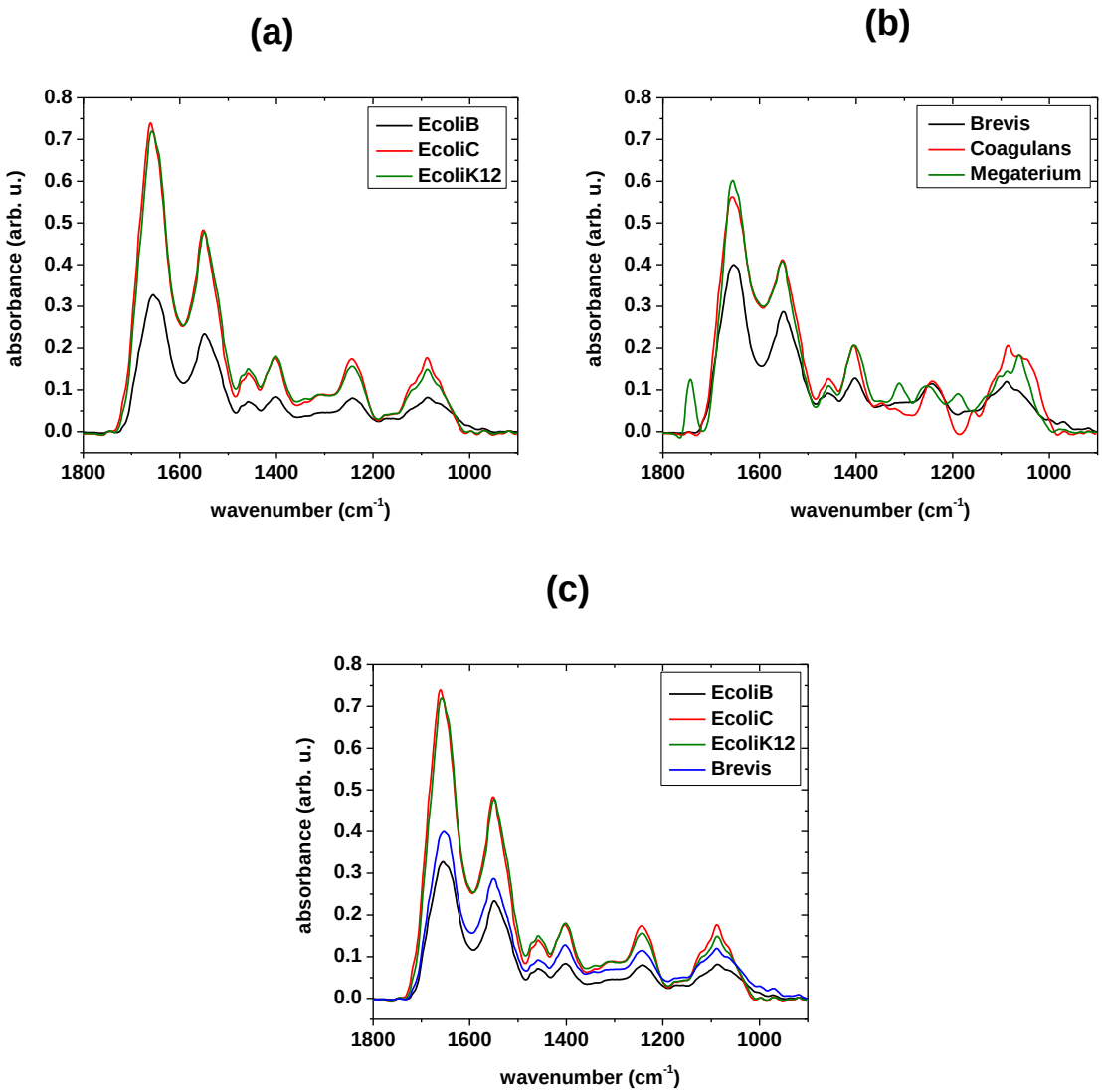

Figure 7: (a) representative MCT spectra of *E. coli* strains *B*, *C*, and *K12* having similarity, (b) representative MCT spectra of *Bacillus* strains *brevis*, *coagulans*, and *megaterium* having noticeable differences among themselves, (c) spectra from the three *E. coli* strains and *Bacillus brevis*, which are similar to one another.

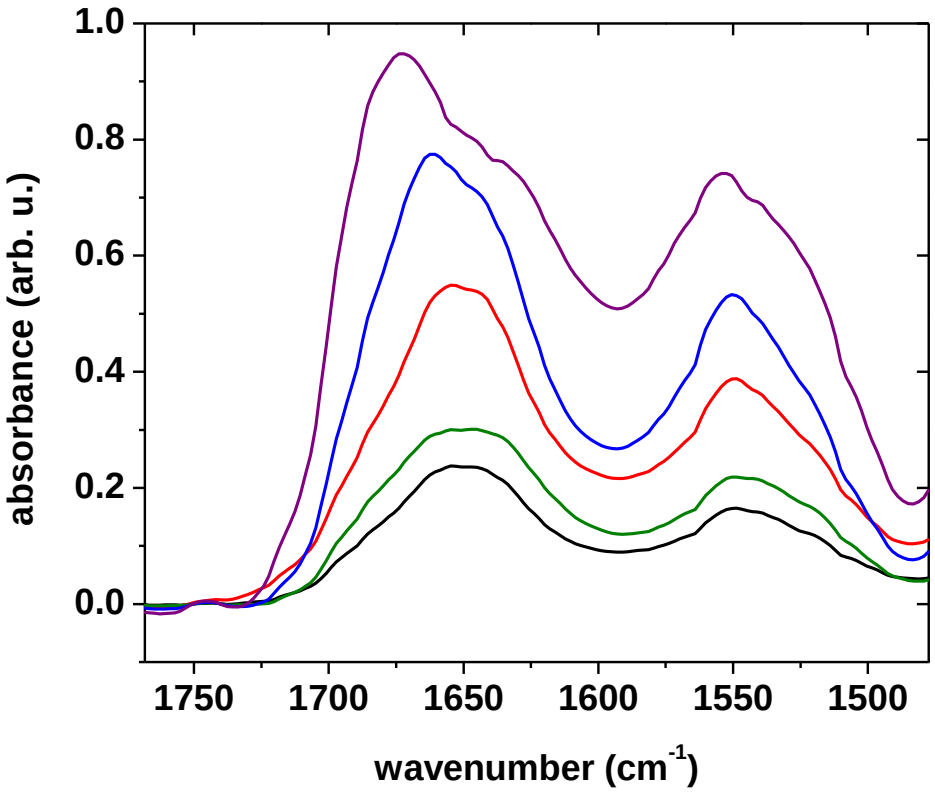

Figure 8: Spectral differences in the Amide I region (1650 cm^{-1}) leading to different shapes. It is hypothesized that these differences arise from varying contributions from protein secondary structures. Since the shape differences are present in all sets of spectra acquired, there is no single characteristic spectrum that can be assigned to each strain.

3.6 Identification Algorithm

A flowchart of the method for the identification of bacterial colonies can be seen in Figure 9 and will guide the discussion. The identification method is based on the hypothesis that the spectra within a data cube of a certain type of bacteria will have spectral features closer to those of a spectrum of the same type than to a spectrum of a different type. In order to have a 'gold standard' to which spectra in a data cube can be compared, a set of MCT spectra are used as a calibration set. This calibration set contains spectra from *known* locations of *known* types. Applying a correlation analysis between spectra from a data cube and spectra such as those in Figure 8 at this point in the method will yield low correlation coefficients, even if the spectra are of the same type of bacteria. For this reason, extra steps must be taken. Performing a PCA on this calibration set will determine the representations of chemical properties within that set, which, due to the chemical changes (Figure 8) within one set, will require more than one PC. Because different types of bacteria have slightly different arrangements of chemical components, a PCA will represent them differently; hence the PCs from different types of bacteria will be slightly different. Using this calibration set to evaluate the spectra within a data cube will produce a set of scores for the data cube, which will describe how much each PC from the calibration set contributes to the spectra. It is expected that the PCs from a calibration set will model the spectra contained in a data cube best if the calibration set and data cube are of the same type of bacteria. In order to determine how well a spectrum is modeled, the PCs obtained from the calibration set and the scores from that spectrum can be used to construct a second version of this spectrum. This second version is referred to as a 'reconstructed' spectrum. Here a correlation analysis can be used to compare the reconstructed spectrum and the original spectrum in the data cube. The correlation coefficient produced will indicate how similar the spectrum in the data cube is to the spectra contained in the calibration set. It is expected that the correlation coefficient between an original spectrum and the reconstructed spectrum will be higher if the original spectrum and the spectra in the calibration set are of the same type of bacteria than if they are of different types.

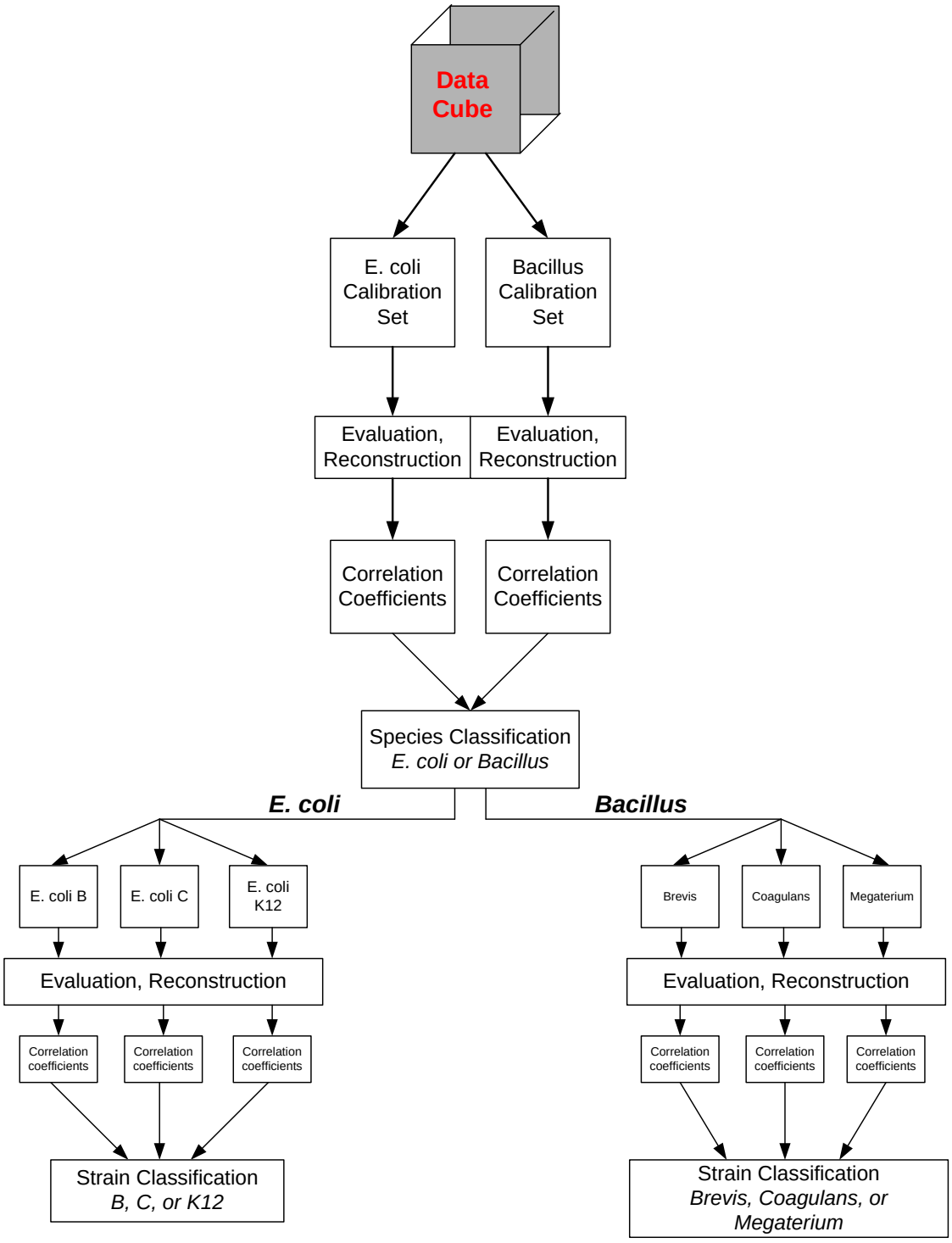

Figure 9: Flow chart of the data analysis method as described in section 3.6.

In some cases, a spectrum within the data cube is not reconstructed well, such as in the case of a bad pixel on the FPA or in the case of no sample present. In these situations, the correlation coefficient produced is low compared to the others. Thus, a threshold for the correlation coefficients is set. Here, the threshold is set at 0.9 based on experimental data, where the correlation coefficients largely fall between 0.93 and 0.98 and the correlation coefficients produced by a bad pixel, noise, or spectra containing a good deal of noise generally falling between 0.3 and 0.8. If the correlation coefficient is below this threshold, the spectrum will be considered ‘non-classifiable’.

3.7 Results and Discussion

3.7.1 Selection of Optimum Number of Principal Components

The number of principal components used to evaluate an FPA data cube is an important factor, as these are used to represent the chemical properties of different strains of bacteria. If too few PCs are used, small chemical differences that exist between different strains may not be taken into account and can result in an incorrect identification., However, it can become necessary to restrict the number of PCs used so as to sufficiently model relevant chemical information without over-fitting the data [46]. Over-fitting the data can occur when a model has too many degrees of freedom, resulting in poor prediction ability. In this case the noise contained within the spectra of the calibration set may be interpreted as an important chemical feature and can result in incorrect identifications. However, determining the proper number of PCs to use is often difficult. Cross-validation (CV), a technique for assessing the predictive capability of a model, is commonly used to achieve this goal, though it requires information about sample concentrations. Thus, it cannot be applied when data evaluation is based only on a PCA, as is the case here.

An alternative has been developed in reference [47], and selects the optimum number of PCs for each individual spectrum from an unknown sample. This optimum number of PCs is found by an iterative procedure. Starting with a ‘reduced calibration model’

containing only one PC, each iteration increases stepwise the number of PCs used and compares the reduced model to the 'full model', which contains all PCs. If the reduced model and the full model are not significantly different from one another, the reduced model contains the minimum number of PCs needed to describe the spectral features equally as well as the full model. To determine whether the reduced model and the full model have equivalent capabilities, a F-test on the residual variances is performed. The residual variances are obtained by: 1) reconstructing the unknown spectrum using both the reduced model and full model; 2) subtracting the reconstructions from the unknown spectrum, resulting in residual spectra; 3) determining the variances from the elements of the two residual spectra. The two residual variances are F-tested for significant differences. If both models are equivalent, no significant difference will be found. An F-test will determine a probability that the variances of two sets of residuals are significantly different; the smaller the probability, the more likely they are significantly different. A threshold for the probability must be defined, above which it is assumed that the models are not significantly different. The method developed in [47] required small thresholds (e.g. 10e-5) for experimental data.

The PC selection method has been modified in this study. Rather than compare a reduced model to the full model, a reduced model containing Q PCs is compared to another reduced model containing $Q+1$ PCs. This will test whether adding one more PC significantly improves the model; if the addition of one more PC does not significantly improve the ability to model a certain spectrum, the variances between the residuals of the two models will not be significantly different. If this is the case, Q PCs are optimum; able to sufficiently model the spectroscopic features without over-fitting the data. Here, a threshold of 10% has been used and is a more reasonable threshold. E.g., if comparing a model using 5 PCs and a model using 6 PCs produces a probability value above 10%, it can be assumed that the two models are not significantly different, and the model using 5 PCs will be able to model a spectrum as well as the model using 6 PCs. The optimum number of PCs are determined for every single spectrum inside a FPA data cube and thus, the reconstruction of a spectrum is done with this optimum number of PCs.

3.7.2 Results of Identification

Eighteen data cubes from each species of bacteria have been evaluated by the identification algorithm. Each of these data cubes is of known species, and is used to assess the capability of the method. The data cubes will be classified by calculating a percentage. These percentages represent the number of spectra in the data cube that are classified as a certain species. The percentages are calculated based on the number of spectra classified as a certain species out of the number of 'classifiable spectra'. Recall that non-classifiable spectra are those having correlation coefficients less than the set threshold of 0.9. The number of non-classifiable spectra is subtracted from the total number of spectra in the data cube, resulting in the number of classifiable spectra. Of the six data cubes of *E. coli B* that were evaluated, two of the data cubes were classified as *E. coli*. Of the six data cubes of *E. coli C* evaluated, five were classified as *E. coli*. Five out of six data cubes of *E. coli K12* were classified as *E. coli*, with one classified as *Bacillus*. Four data cubes of *Bacillus brevis* were classified as *Bacillus*, while all six data cubes of *Bacillus coagulans* were classified as *Bacillus*. Five out the six data cubes of *Bacillus megaterium* were classified as *Bacillus*.

The left side of Table 4 shows the percentage classification as either *E. coli* or *Bacillus* for the evaluated *E. coli* data cubes. The first column indicates the data cube, where the letter indicates strain, i.e. *B0* is the first data cube of *E. coli B* acquired and evaluated. It is expected that all values in the second column are larger than those in the third column, as each data cube on the left is known to be *E. coli*. This expectation is fulfilled for twelve of the eighteen data cubes. The right side of Table 4 shows the percentage classification as either *E. coli* or *Bacillus* for the evaluated *Bacillus* data cubes. Here it is expected that all values in the last column will be higher than those in the fifth column. This expectation is fulfilled for fifteen of the eighteen data cubes. Misclassified data cubes are shown in *italics* on each side. It can be seen by looking at the percentages in Table 4 that ten of the twenty-seven data cubes classified as the correct species have a percentage of classification is less than 60%.

The left side of Table 5 shows the data cubes that were classified as *E. coli* in Table 4, while the right side shows the data cubes that were classified as *Bacillus*. Each side also

Table 4: Species classification based on percentage of classifiable spectra using *E. coli* calibration versus *Bacillus* calibration. Classification is based on majority of classifiable spectra. The higher percentages of correct classification are in **bold**. Incorrect species classifications are noted by *italics*. (B = *E. coli B*, C = *E. coli C*, K = *E. coli K12*, brev = *Bacillus brevis*, coag = *Bacillus coagulans*, mega = *Bacillus megaterium*)

Data Set	% E. coli Classification	% Bacillus Classification	Data Set	% E. coli Classification	% Bacillus Classification
B0	**66.19**	33.81	brev0	48.56	**51.44**
B1	*25.67*	*74.33*	brev1	41.5	**58.5**
B2	*47.84*	*52.16*	brev2	*60.92*	*39.08*
B3	*42.21*	*57.79*	brev3	47.86	**52.14**
B4	*34.78*	*65.22*	brev4	48.44	**51.56**
B5	**59.6**7	40.33	brev5	*50.1*	*49.9*
C0	**57.36**	42.64	coag0	22.78	**77.22**
C1	**54.49**	45.51	coag1	32.59	**67.41**
C2	*49.52*	*50.48*	coag2	37.67	**62.33**
C3	**57.47**	42.53	coag3	22.52	**77.48**
C4	**60.93**	39.07	coag4	24.03	**75.97**
C5	**64.47**	35.53	coag5	25.74	**74.26**
K0	**74.92**	25.08	mega0	1.06	**98.94**
K1	**68.84**	31.16	mega1	9.71	**90.29**
K2	*48.47*	*51.53*	mega2	12.81	**87.19**
K3	**53.81**	46.19	mega3	24.69	**75.31**
K4	**59.25**	40.75	mega4	*65.11*	*34.89*
K5	**78.85**	21.15	mega5	5.91	**94.09**

Table 5: Results of strain classification using MCT spectra as calibration sets and optimum PC selection algorithm with no adjustment. The higher percentages for correct classifications are in **bold**. Incorrect strain classifications are noted in *italics*. The data cubes classified as *E. coli* are shown on the left, while the data cubes classified as *Bacillus* are shown on the right.

Classified as E. coli

Data Set	% E coli B Classification	% E coli C Classification	% E coli K12 Classification
B0	**47.23**	26.84	25.93
B5	*27.28*	*31.25*	*41.47*
C0	41.78	**42.88**	15.34
C1	*35.65*	*31.13*	*33.22*
C3	28.15	**39.64**	32.21
C4	31.32	**40.77**	27.91
C5	*26.94*	*35.88*	*37.18*
K0	*23.08*	*43.85*	*33.07*
K1	*38.15*	*28.58*	*33.27*
K3	*26.7*	*46.72*	*26.58*
K4	4.33	34.99	**60.68**
K5	*4.76*	*60.09*	*35.15*
brev2	*26.22*	*32.54*	*41.24*
brev5	*24.28*	*38.59*	*37.13*
mega4	*44.86*	*35.9*	*19.24*

Classified as Bacillus

Data Set	% Brevis Classification	% Coagulans Classification	% Megaterium Classification
B1	*29.06*	*36.82*	*34.12*
B2	*78.98*	*15.73*	*5.29*
B3	*21.71*	*16.17*	*62.12*
B4	*2.11*	*97.66*	*0.23*
K2	*38.56*	*49.36*	*12.08*
brev0	**60.71**	17.5	21.79
brev1	**44.58**	23.97	31.45
brev3	**59.19**	17.76	23.05
brev4	**54.92**	20.4	24.68
coag0	13.17	**66.41**	20.42
coag1	12.26	**66.7**	21.04
coag2	24.93	**40.1**	34.97
coag3	9.32	**52.24**	38.44
coag4	12.98	**47.08**	39.94
coag5	13.59	**57.83**	28.58
mega0	0.84	6.11	**93.05**
mega1	2.97	4.74	**92.29**
mega2	17.53	5.26	**77.21**
mega3	4.29	19.34	**76.37**
mega5	12.76	12.14	**75.1**

shows the percentage of classification as *B*, *C*, or *K12* from *E. coli* or *brevis*, *coagulans*, or *megaterium* from *Bacillus*. Of the twelve *E. coli* data cubes correctly classified by species, five were correctly classified by strain. Of the fifteen data cubes correctly classified by species as *Bacillus*, all were correctly classified as their respective strains, shown towards the bottom of the right side. Overall, 75% of the data cubes correctly classified by species were also correctly classified by strain, though the percentages of classification were typically low, as was the case during species classification. These low percentages occur primarily in the *E. coli* and *Bacillus brevis* data cubes. This is likely due to the degree of similarity between the spectra of *E. coli* and the spectra of *Bacillus brevis* (Figure 7). The percentages are typically lower than 50%, except in the case of *K4*, where the percentage is just above 60%. It is hypothesized that this may also be due to the highly varying numbers of PCs selected by the optimum PC selection algorithm. Thus, it would be helpful to institute a degree of uniformity to the numbers of PCs used to evaluate a given data cube to see how the classification changes accordingly.

3.7.2.1 Instituting a Degree of Uniformity

In order to assess whether the highly varying number of optimum PCs used for each spectrum leads to the low percentages encountered, an adjustment was made to the optimum PC selection algorithm. This adjustment allows the algorithm to calculate the optimum number of PCs for every spectrum within a data cube and calculates the average number of PCs for that set. Utilizing the average number of PCs for every spectrum in the data cube will introduce a measure of uniformity while still allowing the spectroscopic features to be modeled with a minimum amount of over-fitting. It will also allow for only one variable, the spectrum, changing at a time.

The left and right sides of Table 6 show the same information as in Table 4, though here the average number of PCs has been used. Overall, of the twenty-seven correctly classified data cubes, the percentages of classification have increased, with all but one above 60%. It is interesting to note that, while the number of correctly classified data cubes has remained the same, some of the identities have changed. Previously, three of

Table 6: Species classification based percentage of classifiable spectra using the average number of PCs as calculated by the optimum PC selection algorithm. The higher percentages of correct classification are shown in **bold**. Incorrect species classifications are noted by *italics*. (B = *E. coli B*, C = *E. coli C*, K = *E. coli K12*, brev = *Bacillus brevis*, coag = *Bacillus coagulans*, mega = *Bacillus megaterium*)

Data Set	% E. coli Classification	% Bacillus Classification	Data Set	% E. coli Classification	% Bacillus Classification
B0	**95.68**	4.32	brev0	46.93	**53.07**
B1	*5.25*	*94.75*	brev1	*76.28*	*23.72*
B2	**100**	0	brev2	*67.72*	*32.28*
B3	**81.98**	18.02	brev3	*62.18*	*37.82*
B4	*0*	*100*	brev4	*66.58*	*33.42*
B5	**67.12**	32.88	brev5	*69.59*	*30.41*
C0	**96.01**	3.99	coag0	21.5	**78.5**
C1	**76.56**	23.44	coag1	21.62	**78.38**
C2	**67.47**	32.53	coag2	*52.38*	*47.62*
C3	**89.55**	10.45	coag3	19.71	**80.29**
C4	**93.6**	6.4	coag4	8.78	**91.22**
C5	**99.73**	0.27	coag5	23.64	**76.36**
K0	**98.93**	1.07	mega0	0	**100**
K1	**68.22**	31.78	mega1	0	**100**
K2	**95.22**	4.78	mega2	0	**100**
K3	**98.7**	1.3	mega3	0	**100**
K4	**99.8**	0.2	mega4	*86.9*	*13.1*
K5	**100**	0	mega5	0	**100**

the five *brevis* data cubes were correctly classified as *Bacillus*. In this section, only one of them has been correctly classified as *Bacillus*. Recall from Figure 7 that the spectra of *brevis* were similar to those of *E. coli* while being different from *coagulans* and *megaterium*. It is likely that the spectra contained in the *brevis* data cubes are being reconstructed better using the *E. coli* calibration set, to which they are more similar, than the *Bacillus* calibration set, to which they are different. Because the numbers of PCs used are restricted here, the higher PCs from the *Bacillus* calibration set which may model differences between the general shapes of *E. coli* and *Bacillus brevis* are not accessible and thus the cubes are misclassified.

The left side of Table 7 contains the data cubes classified as *E. coli* using the average numbers of PCs. The right side of this table shows the data cubes classified as *Bacillus*. Each side shows the percentage of classification by strain as well, with incorrect classifications shown in italics. Here, of the sixteen data cubes correctly classified as *E. coli*, five were correctly classified by strain. In this case, the adjustment to the optimum PC selection algorithm may be restricting the numbers of PCs used too much, not utilizing higher PCs that contain minor differences between the spectral shapes of the three *E. coli* strains. If the number of PCs used is too few, non-classifiable spectra will also result, as the correlation coefficients produced will be below the set threshold. On the right side of the table, the data cubes of *Bacillus* that were correctly classified by species have all been correctly classified by strain, except in the case of *coag4*. The percentages of classification have increased as well for each data cube. *Megaterium* again has large percentages, close to 100% for each data cube.

Overall, the number of data cubes that were correctly classified by strain did decrease, likely due to the adjustment to the optimum PC selection algorithm, causing too few PCs to be used and not allowing the modeling of small differences by the higher PCs. However, as was the case for species classification using this approach, the percentages of classification did show an overall increase. Misclassifications do remain, particularly among the *Bacillus brevis* data cubes. Thus, identifying a contamination of *Bacillus brevis* as such would prove difficult, as five of the six evaluated data cubes were classified as the wrong species. Thus, it would be advantageous to both increase the number of overall correct classifications as well as classify the different species equally

Table 7: Results of strain classification using MCT spectra as calibration sets and an adjusted optimum PC selection algorithm using average PCs. The higher percentages of correct classification are shown in **bold**. Incorrect strain classifications are noted in *italics*.

Classified as E. coli

Data Set	% E coli B Classification	% E coli C Classification	% E coli K12 Classification
B0	**84.63**	4.18	11.19
B2	*21.82*	*78.18*	*0*
B3	*1.77*	*97.92*	*0.31*
B5	*25.66*	*19.8*	*54.54*
C0	*87.29*	*12.54*	*0.17*
C1	23.36	**43.9**	32.74
C2	*0.25*	*0*	*99.75*
C3	5.22	**50.6**7	44.11
C4	*45.79*	*43.17*	*11.04*
C5	*30.25*	*28.68*	*41.07*
K0	*1.82*	*52.47*	*47.71*
K1	*6.92*	*77.79*	*15.29*
K2	0	0	**100**
K3	*0.22*	*99.04*	*0.74*
K4	0	0.02	**99.98**
K5	*0*	*99.14*	*0.86*
brev1	*20.24*	*31.91*	*47.85*
brev2	*13.1*	*31.5*	*55.4*
brev3	*27.81*	*35.19*	*37*
brev4	*7.89*	*37.87*	*54.24*
brev5	*8.35*	*37*	*54.65*
coag2	*54.18*	*45.82*	*0*
mega4	*17.33*	*82.67*	*0*

Classified as Bacillus

Data Set	% Brevis Classification	% Coagulans Classification	% Megaterium Classification
B1	*16.63*	*10.44*	*72.92*
B4	*0*	*100*	*0*
brev0	**97.54**	0.05	2.41
coag0	8.74	**90.1**	1.15
coag1	5.25	**84.35**	10.4
coag3	0.41	**65.32**	34.27
coag4	*2.42*	*35.16*	*62.42*
coag5	5.74	**66.39**	27.87
mega0	0.08	0.03	**99.89**
mega1	0	0	**100**
mega2	0	0	**100**
mega3	0	0	**100**
mega5	0	0	**100**

well, meaning that the data cubes belonging to one species are not correctly classified significantly better than the other, as in the case of *Bacillus brevis*. A second adjustment, using FPA data cubes rather than MCT spectra as calibration sets, is applied.

3.7.2.2 Using FPA Data Cubes as Calibration Sets

Figure 10 (top) depicts a MCT spectrum and a spectrum from a FPA data cube, both of *E. coli B*. It can be seen that, while the spectra have similar shapes, they do have inherent differences. This is evident when looking at the Amide I peak at ~1650cm^{-1}. Where the MCT spectrum has a more broad shape, the FPA spectrum does not. What looks like noise is also present in the FPA spectrum, though this may be masking other spectral features. The same FPA spectrum is shown with the first PCs from the *E. coli* and *Bacillus* calibration sets in the bottom of Figure 10. As MCT spectra from the different species are used in the calibration set, it can be seen that the PCs resemble these spectra; more importantly, they do not contain the differences seen in the FPA spectrum. The PCs produced from the MCT spectra will model the *overall shape* of the FPA spectrum, but will not be able to model these differences, which may lead to misclassifications and must be accounted for. For this reason, FPA data cubes have been used as calibration sets in a second adjustment to the classification method.

The data analysis method using FPA data cubes as calibration sets also uses the average numbers of PCs. Here, one data cube from each strain has been removed from evaluation, to be used for calibration. Again, for species classification, three data cubes from *E. coli* are combined into one calibration set and three data cubes from *Bacillus* are combined into a second calibration set. The left and right sides of Table 8 contain the same information as shown in Table 4 and Table 6, with some key differences. Here, cubes *B0*, *C3*, *K4*, *brev0*, *coag0*, and *mega0* have been used for calibration and thus are not evaluated. Data cubes *B2* and *B4* are marked with a “zero” in bold italics, as they were both deemed non-classifiable, meaning that no correlation coefficient produced was above the set threshold of 0.9. In fact, for cube *B4*, the highest correlation coefficient obtained from the *E. coli* calibration was 0.896, while the highest obtained from the *Bacillus* calibration set was 0.623.

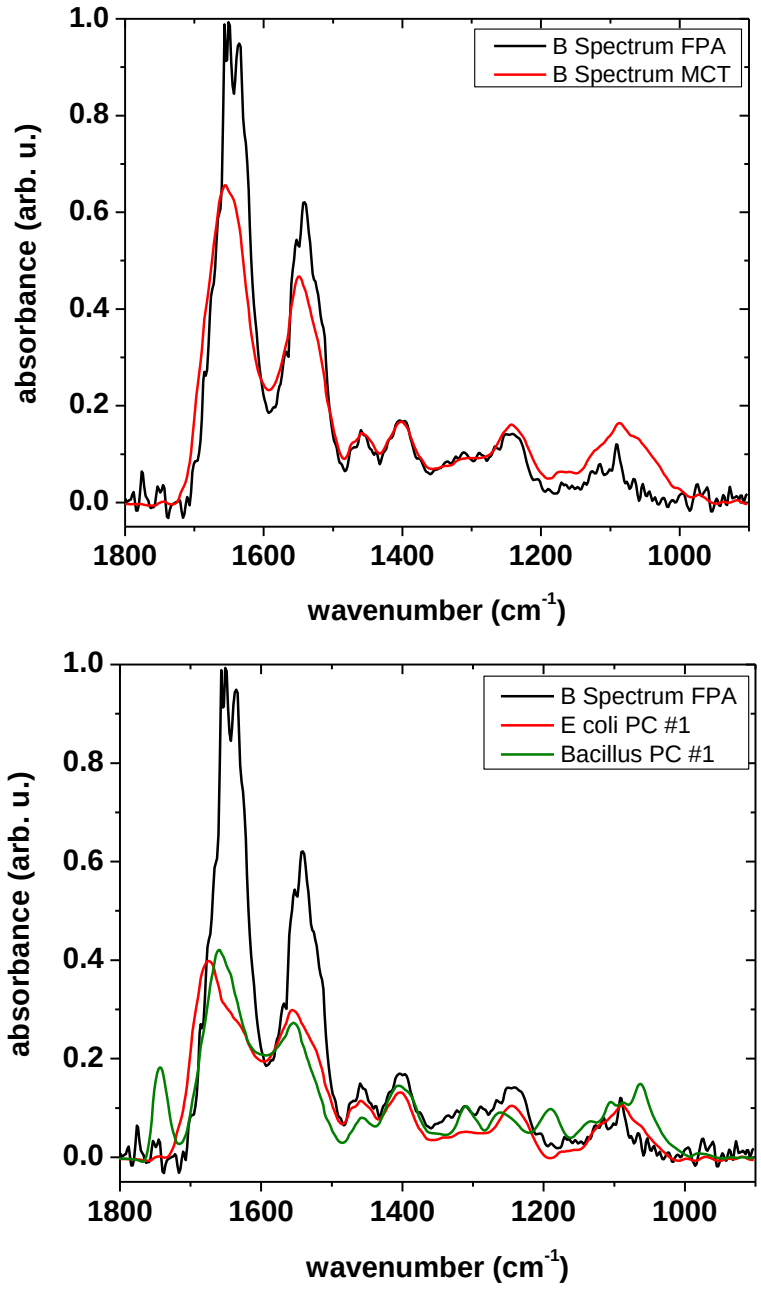

Figure 10: (top) comparison of MCT spectrum from *E. coli B* and a FPA spectrum from *E. coli B*. Though the spectra have a similar overall shape, there are inherent differences between the two. (bottom) comparison of FPA spectrum of *E. coli B* and first PCs from *E. coli* and *Bacillus* calibration sets. Here the first PCs resemble the MCT spectra they are produced from.

Table 8: Species classification results based on percentage of correct classification using FPA data cubes as calibration sets. The higher percentages of correct classification are shown in **bold**. Incorrect species classifications are noted by *italics*. Data cubes that are non-classifiable are noted by ***bold red italics***.

Data Set	% E. coli Classification	% Bacillus Classification	Data Set	% E. coli Classification	% Bacillus Classification
B0	USED AS	CALIBRATION SET	brev0	USED AS	CALIBRATION SET
B1	**99.92**	0.08	brev1	*85.47*	*14.53*
B2	***0***	***0***	brev2	25.6	**74.4**
B3	**100**	0	brev3	0.33	**99.67**
B4	***0***	***0***	brev4	0.63	**99.37**
B5	**100**	0	brev5	0	**100**
C0	*2.64*	*97.36*	coag0	USED AS	CALIBRATION SET
C1	**100**	0	coag1	0	**100**
C2	**99.41**	0.59	coag2	0.21	**99.79**
C3	USED AS	CALIBRATION SET	coag3	0.25	**99.75**
C4	**98.06**	1.94	coag4	0	**100**
C5	*0.05*	*99.95*	coag5	0	**100**
K0	*21.23*	*78.77*	mega0	USED AS	CALIBRATION SET
K1	*0*	*100*	mega1	0	**100**
K2	**100**	0	mega2	0.02	**99.98**
K3	**78.93**	21.07	mega3	0	**100**
K4	USED AS	CALIBRATION SET	mega4	0.07	**99.93**
K5	**100**	0	mega5	0	**100**

The right side of the table shows the percentages of classification for each of the *Bacillus* cubes, again not counting the cubes used for calibration. Here, fourteen of the evaluated data cubes were correctly classified as *Bacillus* with at least 74% classification. Four of the *brevis* data cubes are now correctly classified as *Bacillus*, as are all of the evaluated *coagulans* and *megaterium* data cubes.

On the left side of Table 9, the data cubes that were classified as *E. coli* are shown. The right side contains the data cubes classified as *Bacillus*. The percentages of classification by strain are shown on each side in the grey/white rows, with incorrect classifications noted by italics. Note that the data cubes used as calibration sets are not shown here, nor are the two data cubes of *E. coli B* that were deemed non-classifiable. Nine *E. coli* data cubes, shown on the left, were correctly classified by species as *E. coli*. Of these, seven were correctly classified as their respective strains with percentages above 70%. The misclassification of the two *E. coli B* data cubes may again be due to restricting the numbers of PCs used. The *Bacillus* data cubes shown on the right side were also classified well by strain. All fourteen were correctly classified as their respective strains, with percentages of classification above 75% in all but *brev2*. The overall percentages of classification remained about the same as for the results presented in Table 7. The overall correct strain classification, however, improved a good deal. Here, 91% of the classifiable data cubes were correctly classified by strain.

3.8 Conclusions

Bacterial colonies are often small and localized within an extended sample area. Thus, they need to be found and identified. The first part of this chapter focused on finding such colonies. In order to investigate many locations at one time, spectroscopic imaging combines FTIR spectroscopy with an infrared camera, allowing the fast acquisition of data. As this produces a great deal of information, it can be useful to visualize this information by assigning colors to different materials, thereby relating a color to a chemical property. In order to accomplish this, a characteristic signature within the chemical properties of a material must be identified. A principal component analysis can

Table 9: Results of strain classification using FPA data cubes as calibration sets and a restricted optimum PC selection algorithm using average PCs. The higher percentages of correct classification are shown in **bold**. Incorrect classifications are noted in *italics*. Cubes B2 and B4 are not shown, as they were non-classifiable.

Classified as E. coli

Data Set	% E coli B Classification	% E coli C Classification	% E coli K12 Classification
B1	**87.19**	0.84	11.97
B3	*2.34*	*6.63*	*91.03*
B5	*3.62*	*63.73*	*32.65*
C1	7.57	**72.37**	20.06
C2	2.3	**78.49**	19.21
C4	3.92	**93.8**	2.28
K2	0	0	**100**
K3	0	0.02	**99.98**
K5	0	0.91	**99.09**
brev1	*0*	*0*	*100*

Classified as Bacillus

Data Set	% Brevis Classification	% Coagulans Classification	% Megaterium Classification
C0	*97.99*	*2.01*	*0*
C5	*97.79*	*2.21*	*0*
K0	*100*	*0*	*0*
K1	*100*	*0*	*0*
brev2	**50.56**	48.7	0.74
brev3	**98.61**	1.39	0
brev4	**97.47**	2.53	0
brev5	**92.38**	5.42	2.2
coag1	14.87	**83.22**	1.9
coag2	15.62	**80.37**	4.01
coag3	0.89	**98.96**	0.15
coag4	1.32	**98.07**	0.62
coag5	4.96	**92.69**	2.35
mega1	8.45	3.54	**88.01**
mega2	9.98	2.27	**87.75**
mega3	5.33	20.24	**74.43**
mega4	11.81	0.07	**88.11**
mega5	0.99	0.05	**98.96**

extract the spectroscopic features from a set of materials, providing a representation of the chemical properties. To combine PCA with imaging, a multivariate image analysis has been performed. A MIA produces color images in which different materials are displayed as different colors. This approach was applied to three colonies of *E. coli*. Data cubes obtained from these colonies were merged together to show that localization in an extended area was possible. It was shown that these colonies could be clearly discriminated from a substrate based on different spectroscopic signatures and that each of the strains could be discriminated from one another, with each of the colonies homogeneous in color within itself and different colonies encoded as different colors. The substrate was represented as a uniform green color throughout the merged image.

The second part of this chapter focused on the identification of found bacterial colonies using spectroscopic imaging. Building on the correlation analysis presented in chapter two, the identification method presented here was based on a similar hypothesis. The spectra contained within a data cube of a certain type should have spectral features that are more similar to the features of spectra of the same type than to those of a different type. MCT spectra of known type were used as a 'gold standard' to which the spectra in a data cube could be compared. A principal component analysis was applied to this gold standard in order to model different spectral shapes shown in Figure 8. This allowed the reconstruction of a spectrum within the data cube. A correlation analysis between the reconstructed spectrum and the original spectrum provided a measure of the similarity between a spectrum in a data cube and the gold standard. It was expected that the correlation coefficient would be higher if the spectrum in the data cube and the gold standard were of the same species. Using MCT spectra and FPA data cubes of known bacteria types, the identification method was assessed as to how well it was able to identify bacterial colonies.

With no adjustments, the identification method correctly classified twenty-seven out of thirty-six data cubes as the correct species, with twenty out of those twenty-seven identified as the correct strain. Two adjustments were made to the identification in order to maximize both the percentages of classification and the overall number of correct classifications. The first adjustment placed a restriction on the optimum PC selection algorithm, calculating the average number of optimum PCs for a given data cube. The

second adjustment utilized acquired FPA data cubes as calibration sets. Application of the first adjustment was found to raise the percentages of classification for both species and strain classifications, though the number of correct classifications was affected adversely. During species classification, the strains of *Bacillus brevis* were found to be classified as *E. coli*, to which they are spectrally similar. By restricting the numbers of PCs used, it is possible that higher PCs modeling the spectral differences between *Bacillus brevis* and *E. coli* were not utilized, resulting in misclassifications. This same situation was seen during strain classification, this time between the three strains of *E. coli*, resulting in an overall lower number of correct strain classifications than did the approach without using this adjustment.

Utilizing FPA data cubes as calibration sets was hypothesized to minimize the number of misclassifications due to using PCs from MCT spectra, which are able to model only the general shape of the FPA spectra and not inherent differences between the two. Applying this second adjustment in conjunction with the first did result in fewer misclassifications for both species and strain. Out of the twenty-eight classifiable data cubes, twenty-three were correctly classified by species. Of those, twenty-one were correctly classified by strain with high percentages of classification. Thus, using spectroscopic imaging for the identification of found bacterial colonies has been shown to work well.

Chapter 4

Augmenting Spectroscopic Imaging for Analyses of Samples with Complex Surface Topographies

4.1 Introduction

Spectroscopic imaging [48]-[61] has become widely used in the analysis of heterogeneous sample by incorporating focal plane array detectors into spectrometers. This allows the analyses of spatial distributions of chemical information in an X-Y plane at high time resolution. However, in the case of reactions proceeding in three spatial dimensions (X-Y-Z), the acquisition of spectroscopic information in an X-Y plane plus topographic information (Z-dimension) is needed. When capturing visible (2D) images from 3D samples, Z-dimension, or height information, is inherently lost. This information may become useful for monitoring the growth of bacterial colonies. Therefore, this chapter discusses an augmentation to a spectroscopic imaging setup that can acquire both types of data, i.e. spatially resolved spectroscopic information and topographic information at one time.

Custom-made illumination optics project a regularly shaped light pattern, i.e. areas of illumination and shadow, onto 3D samples (Figure 11, left). This light pattern is generated by placing a micromesh into the beam path of a visible light source. If the sample is flat or only substrate is present, the light pattern is preserved. If a non-flat structure is present, the light pattern is distorted accordingly, as can be seen in Figure 12. Since height information is encoded in these distortions, the topography can be determined by extracting the distortions by means of custom-designed computer software.

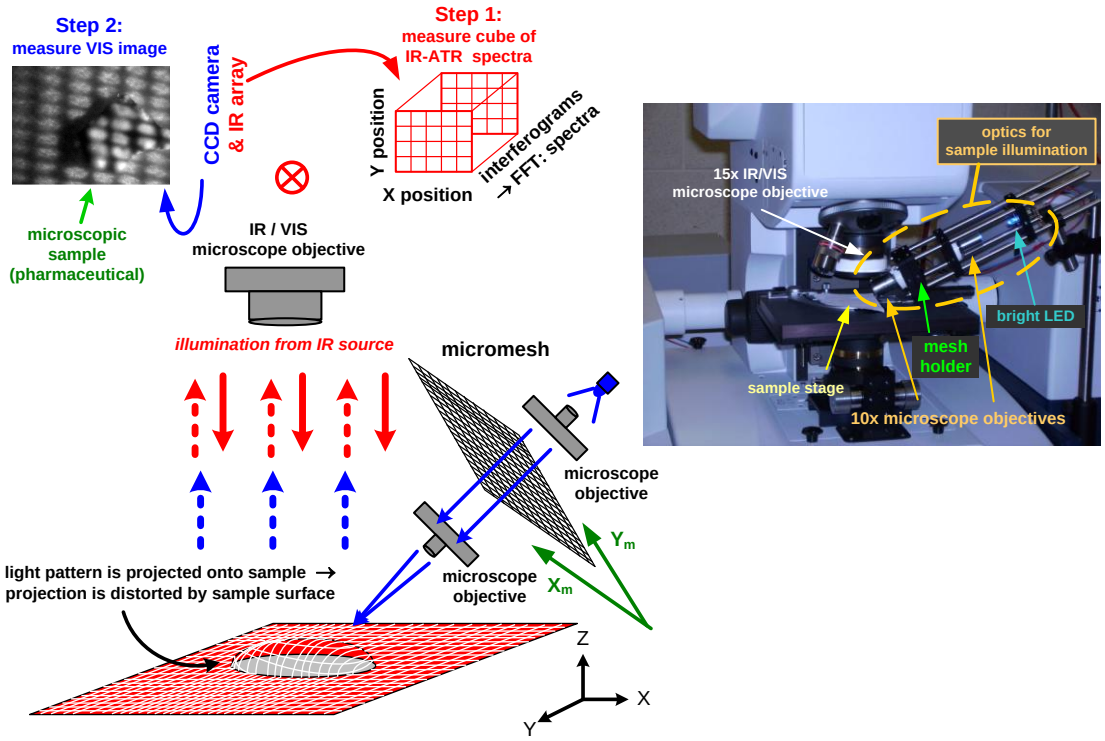

Figure 11: (left) Step 1: FTIR spectroscopic imaging in reflection, transmission or ATR mode probes samples at high spatial resolution. Step 2: Samples are illuminated with a light pattern in the visible wavelength range; this pattern gets distorted upon projection onto 3D samples and topographic information will be gained from these distortions. (right) Picture of the realized external illumination optics setup for generating and projecting a light pattern onto samples.

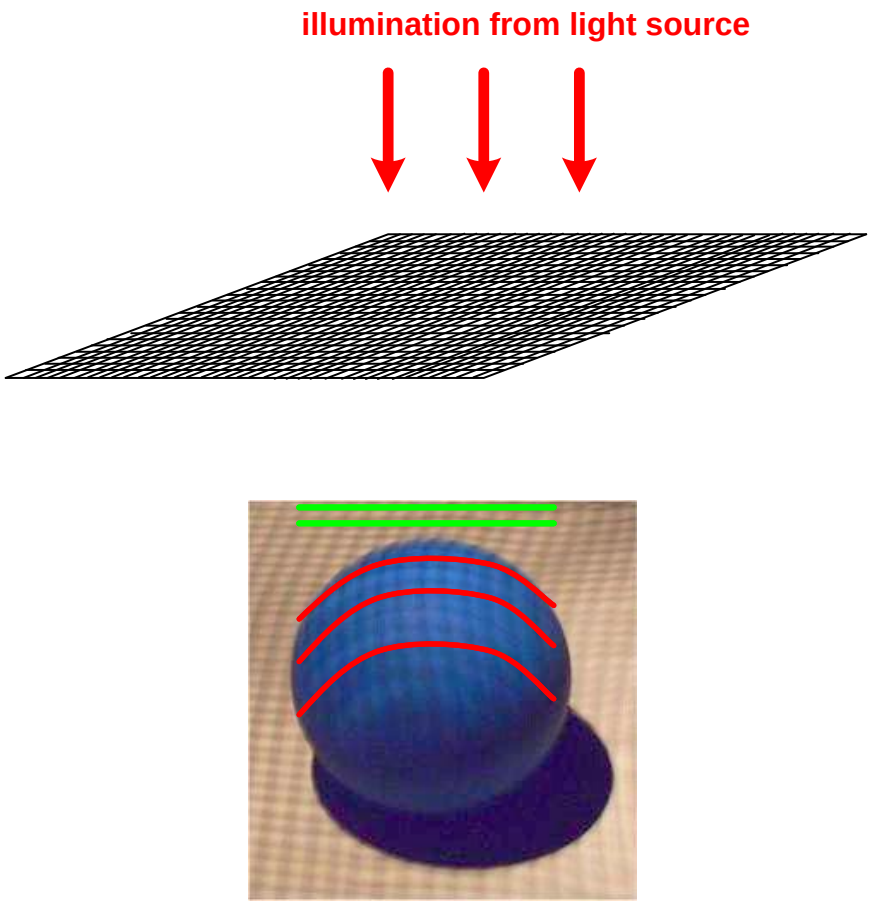

Figure 12: Placing a micromesh into the beam path of a visible light source projects a regular light pattern onto a sample. When projected onto a flat substrate the regularity of the pattern is retained, as shown by the green lines. When projected onto a sample have a complex surface topography, the pattern becomes distorted, as shown by the red curved lines.

Because topographic probing must not affect infrared measurements, different wavelength ranges are used; spectroscopic information is acquired in the mid-IR while the light pattern (topography probe) is generated in the visible. In order to relate distortions to physical height structures, the setup must be calibrated. To this end, calibration objects of known dimensions have been manufactured and are placed under the visible light source. Determining the distortions introduced by objects of known heights will derive a calibration between distortions and topographies. Due to mechanical restrictions the light pattern can only achieve a certain spatial resolution. In order to enhance the spatial resolution, scanning the light pattern in X and Y-directions is proposed.

4.2 Experimental Setup and Results

4.2.1 Experimental Setup

The setup used in this study is coupled to a commercial FTIR spectroscopic imager. This system features a 64 x 64 pixels focal plane array (FPA) and a CCD camera for the visible wavelength region (Sony Exwave HAD Digital Video Camera). An illumination unit projecting a light pattern onto samples has been developed in-house and is depicted in Figure 11 (right). The diverging light emitted from a blue LED is collimated into a parallel beam using a microscope objective, transmitted through a micromesh (Figure 13, left) and then focused onto the sample by a second microscope objective to provide sufficient sample illumination. This illumination unit operates in the visible wavelength range in order to prevent interference with infrared measurements. For this same reason, the mesh contained in the illumination unit must be kept outside of the IR beam path; otherwise artifacts would be contained in the spectroscopic data obtained. Another reason for choosing the visible wavelength range is that CCD cameras usually have larger FPAs (here: 370 x 280 pixels), which will facilitate a better resolution of the topography than an infrared-FPA.

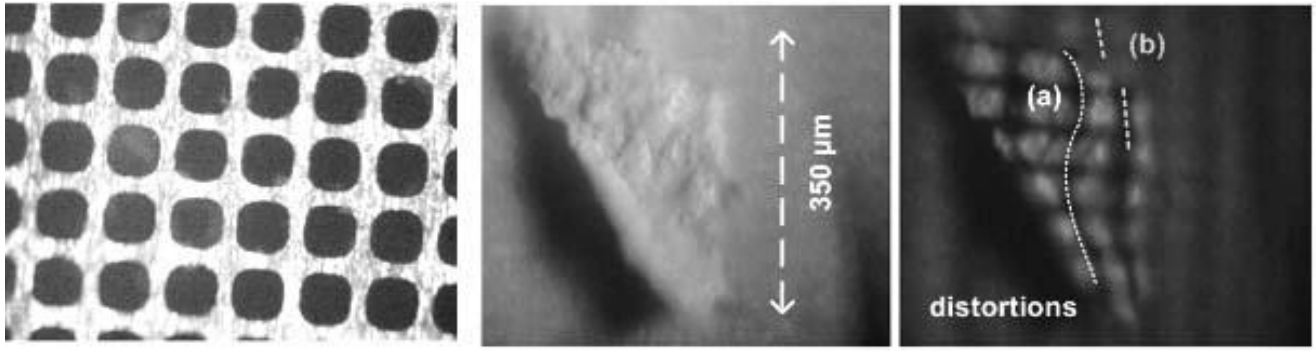

Figure 13: (left) Image of the micromesh used in these studies; the hole dimensions are 100µm x 100µm, wire thickness is 30µm. The transmission of such micromeshes typically ranges from 30% to 70%. (center) This image shows a piece of a pharmaceutical without mesh in the illumination path. (right) The mesh placed in the beam path reduces the overall light level but retains sufficient contrast. The white dotted line points out (a) distortions of the light pattern or (b, grey) a discontinuity in the pattern.

4.2.2 Sufficient Contrast

Since the micromesh (Figure 13, left) acts as a grey filter, it was first determined whether the remaining light level still facilitates a good contrast between shadow and illuminated areas. If there is insufficient contrast, the light pattern extraction will be difficult at best. In Figure 13 (center and right) images of a pharmaceutical sample are shown with and without the mesh present. It can be seen that the mesh somewhat reduces the light level but enough light is still transmitted through so as to ensure sufficient contrast. Beause the mesh is not placed in the IR beam path, the issue of sufficient contrast is of importance only to topographic analyses, as insufficient *visible* contrast will not affect the intensity of the IR radiation used for spectroscopic measurements. The micromesh used in these studies had a good compromise between light reduction and pronunciation of the light pattern. Using a micromesh having thinner wires and larger apertures may transmit more light but would also generate a less pronounced pattern. This would make the detection and extraction of the light pattern less reliable. Also the mesh shown has a limited ruggedness and thinner wires could cause an increased sensitivity to airflow. On the other hand thicker wires and smaller apertures would cover too much of the sample with shadows and thus would reduce contrast and the spatial resolution of the topography. Figure 13 (right) is an example demonstrating the distortions of the light pattern due to the sample's surface topography; the curved white line highlights the pattern of a selected wire.

4.2.3 Extraction of the Light Pattern

Once image acquisition is complete, the distorted light pattern must be extracted. The light pattern generated consists of areas of illumination and areas of shadow. Thus, the signal generated by this pattern will have areas of higher signal (maxima) and areas of lower signal (minima). Figure 14 (left) shows a typical visible image of the light pattern produced. Here, the pattern is projected onto a small piece of pharmaceutical tablet. A mesh extraction algorithm has been developed that searches this image for local minima, indicating that a shadow has been found. The mesh extraction operates as

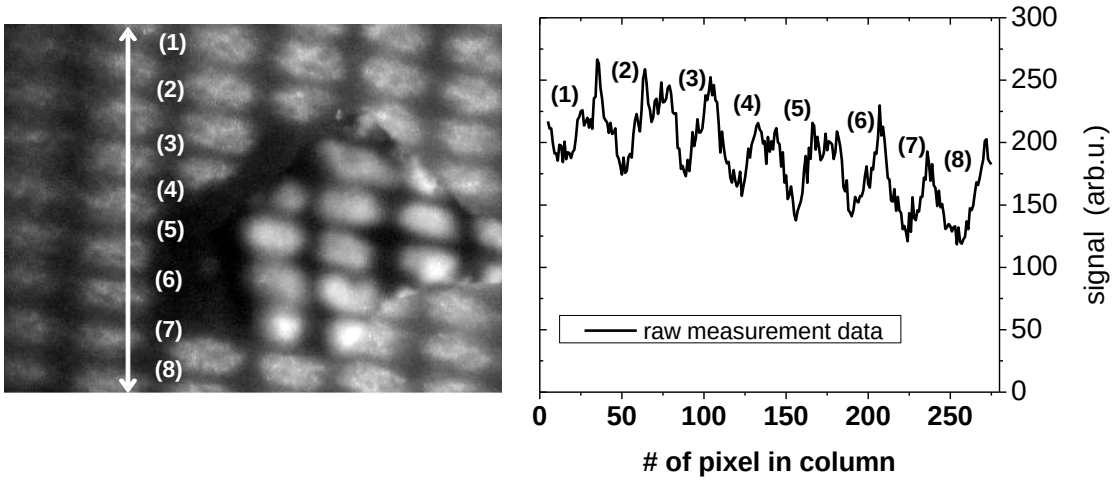

Figure 14: (left) Image of a microscopic sample acquired by the CCD camera (step 2 in Figure 11); (right) signal measured along the column indicated by a white arrow in the top left image; the locations marked by (1) - (8) correspond to local signal minima and shadow areas which receive less light.

such: the visible image shown in Figure 14 (left) in bitmap (BMP) form is loaded into a matrix, producing rows and columns of signal such as the one seen in Figure 14 (right). A window is moved stepwise over this signal, searching for a local minimum at each position. If the minimum is found *inside* the window and not at the edges, it is marked. This is done for every row and column of the matrix, thereby covering both the horizontal and vertical signals.

Looking at the white arrow in Figure 14 (left), it can be seen that, starting at the top of the image and moving to the bottom, there are eight areas of shadow, marked (1)-(8), produced by different mesh wires. These are local minima that the mesh extraction algorithm will search for. The signal along this arrow can be seen by the graph in Figure 14 (right). The numbers (1)-(8) in the right image correspond to the local minima marked (1)-(8) in the left image. In order to search for these local minima the mesh extraction algorithm moves a window covering N measurement points stepwise over the signal. At each window position, the minimal signal reading inside this window is determined.

In moving a window over such a signal as the one in Figure 14 (right), it can be seen that the minimum value inside the window can be found either at one of the borders of the window or inside the borders of the window (arrows marked a, b, and c in Figure 15): (a) If the minimum signal reading is found at the right border of the window, the window is covering a downhill portion of the signal and a local minimum has not yet been reached. (b) If the minimum is found at the left border of the window, the window covers an uphill portion of the signal and a local minimum has been passed. (c) If the minimum value is found inside the window and not at one of the borders, a local minimum is contained inside the window. In this case the pixel having the smallest value is marked. It is important to adapt the window length to the given data. If the window length is too long, it may cover more than one local minima and the detection is ambiguous. Conversely, if the window is too narrow, small noise spikes might be misinterpreted as mesh shadow. A window length of $N = 10$ measurement points was found to work well for this application.

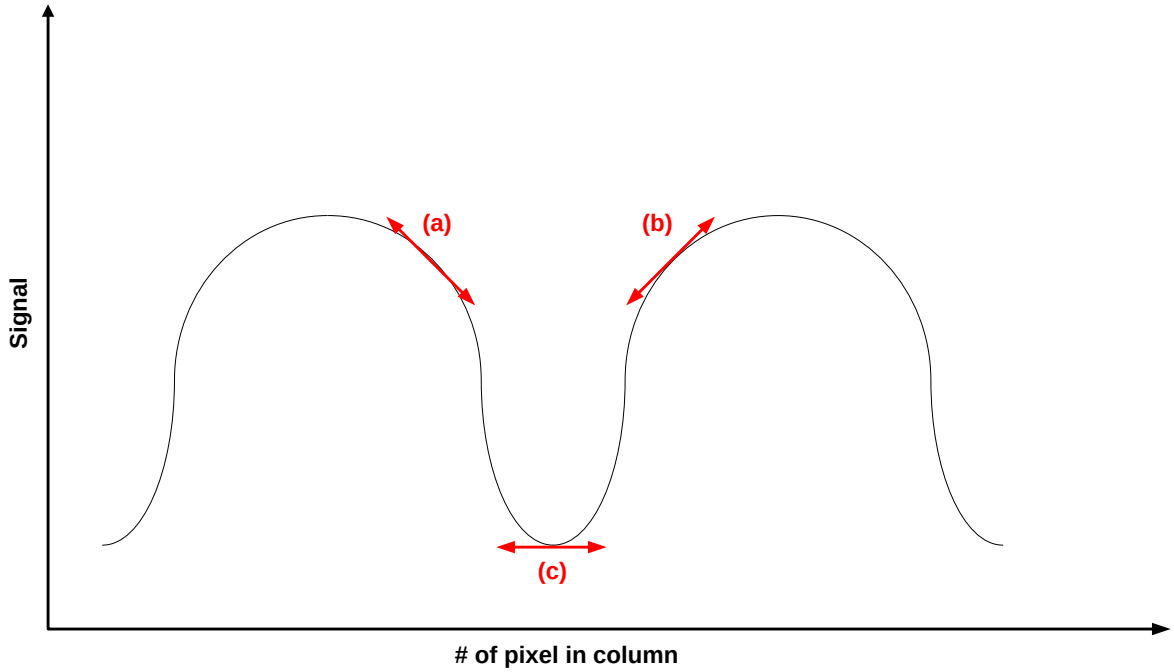

Figure 15: (a) If the minimum value inside the window is found at the right border, the window is on a downhill portion of the signal and has not yet reach a true local minimum. (b) if the minimum value is found at the left border of the window, the window is on an uphill portion of the signal and has passed a true local minimum. (c) if the minimum value is found inside the window, not at the either border, the window is covering a portion of the signal containing a true local minimum.

In some cases, such as using an airflow-sensitive mesh that produces noisy images, it is advantageous to average over several images in order to enhance the signal-to-noise ratio. Here, though the light pattern could be extracted well without averaging, twenty images have been averaged. As the camera has a frame rate of several tens of Hertz, this does not reduce the time resolution of the analysis significantly.

4.2.4 2D FFT Band Pass Filtering

It is evident that the signal in Figure 14 (right) is somewhat noisy. Applying the mesh extraction algorithm to such a noisy signal will result in several local minima being found in an area of low signal, as can be seen in Figure 16. The several local minima are marked by red circles. The results of the extraction in such a case are shown in Figure 17 (bottom, left). The local minima found by the mesh extraction algorithm have been marked with white dots. Obviously, the general shape of the light pattern has been derived. However, the white dots do not form a continuous line and many minima have been found outside the mesh's shadow. Due to the noise level in the signal, minima detection is unreliable and leads to discontinuity in the extracted pattern. Looking at the raw measurement data in Figure 14 (right), it can be seen that the signal is "sine-shaped". While this pattern must be preserved, it is advantageous to remove the noise, thus producing a smoother signal that does not have many local minima within one area of low signal.

In order to enhance the signal-to-noise, the raw images have been bandpass filtered in the Fourier domain. For this purpose the image has been 2D fast Fourier transformed (FFT) [37]. To the transformed image a bandpass filter (fine-tuned for the images acquired here) has been applied. This somewhat suppresses DC signal components, as DC does not contain information about the mesh shadow. The high frequencies, i.e. noise, have also been suppressed. In this way the bandpass filter suppresses signal components that have clearly different (spatial) frequencies than the more or less regular mesh shadow. This bandpass filter has been empirically fine-tuned to this optical setup.

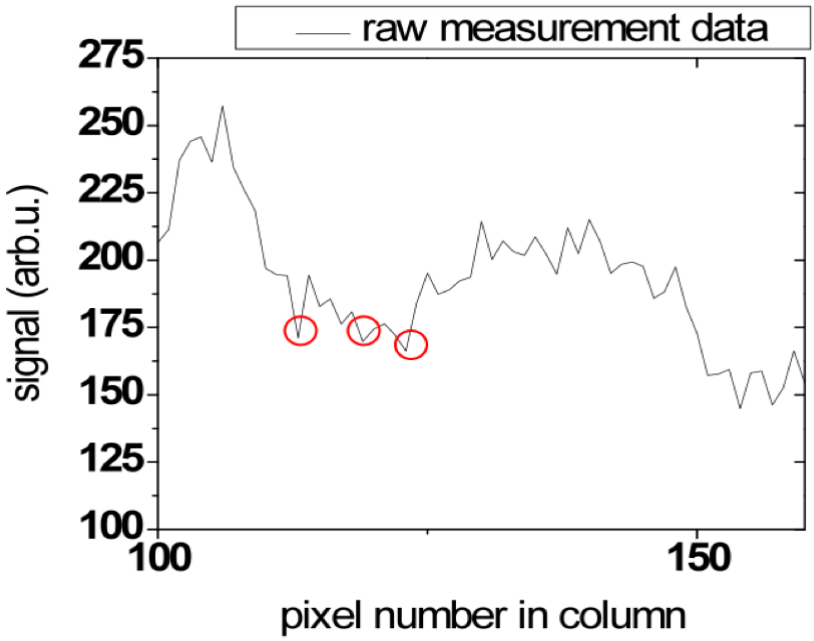

Figure 16: Expanded view of one area of low signal (shadow). It can be seen that, due to the noise level of the signal, several local minima are found within one area of low signal. This will cause unreliability and discontinuity in the mesh extraction.

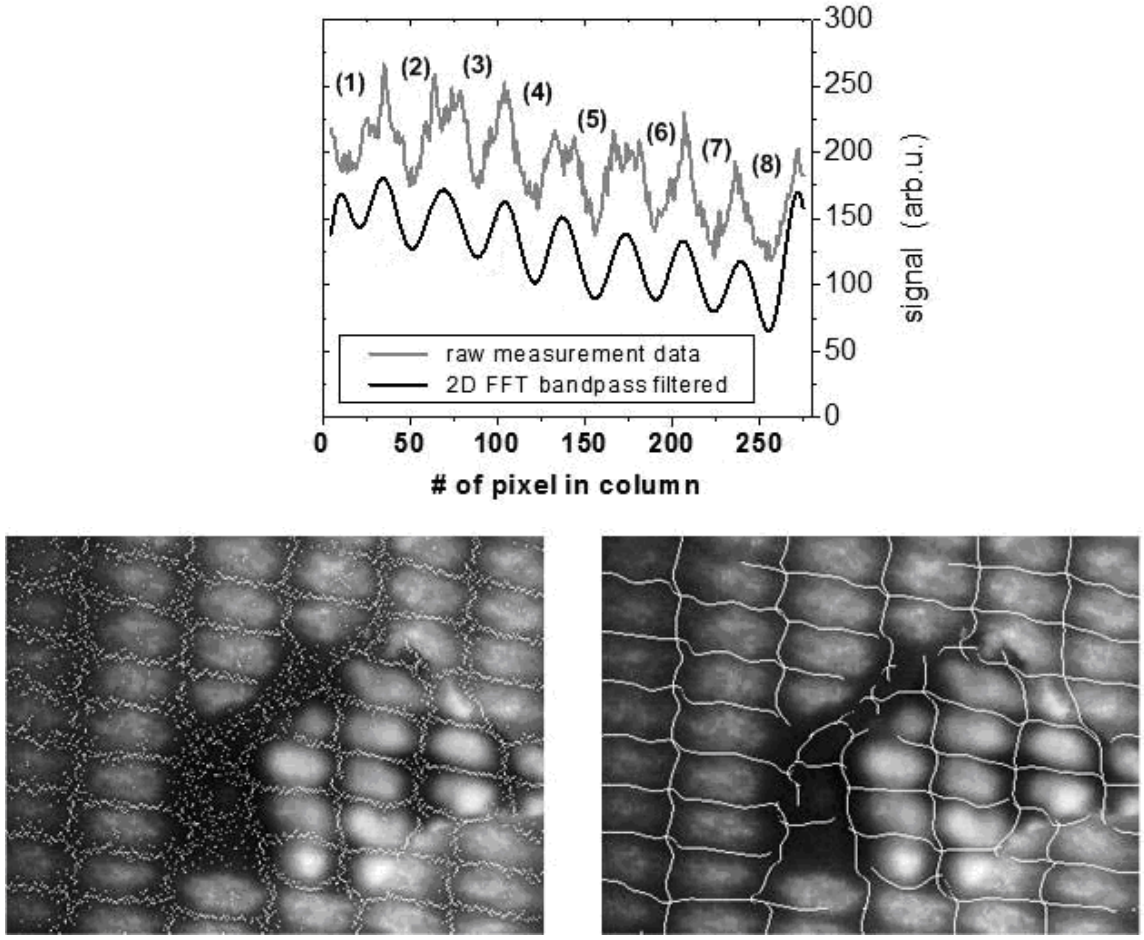

Figure 17: (top, grey graph) raw measurement signal from Figure 14. (top, black graph) measurement signal after apply 2D FFT band pass filtering. The noise level is considerably lower. (bottom, left) the extracted light pattern is marked by white dots overlaying the image; no 2D FFT band pass filtering had been applied; (bottom, right) extracted light pattern marked by white *lines* overlaying the image after utilizing a 2D FFT band pass filter.

In Figure 17 (top) the grey signal after bandpass filtering is shown in black. The enhanced signal-to-noise is clearly visible; the signal minima induced by the mesh's shadow also become better defined. The search for local minima becomes much less ambiguous and the extracted light pattern Figure 17 (bottom, right) contains much less noise. Furthermore, since the lines of the light pattern are now continuous the determination of the surface topography is enhanced.

4.2.5 Enhancing Spatial Resolution

The number of shadow lines projected onto a sample determines the spatial resolution with which the topography can be derived. As bacterial contaminations begin as small colonies, a good spatial resolution is needed; otherwise only a few mesh lines may cover the colony. In Figure 18, the distortion of the light pattern reveals a ridge running across the piece of pharmaceutical, however, the spatial resolution is limited because only a few lines cover the sample. In order to better probe the topography more shadow lines need be projected onto the samples. To achieve this, a finer mesh could be chosen, which would introduce mechanical challenges as mentioned before. Alternatively, one could choose a large magnification factor of the second microscope objective. The resulting, smaller focal depth, however, would require the second microscope objective being placed very close to the sample. Both objectives, i.e. IR microscope and illumination optics, have typical working distances of only a few millimeters. Bringing them close together is often not possible since there is simply not enough space for both objectives. Another more feasible solution is to scan the mesh. For this purpose, a U-shaped holder (Figure 19) of the micromesh is mounted onto an XY translation stage. While the mesh is scanned stepwise in X_m- and Y_m-direction (Figure 18) images are acquired at each position. From each of these images the (distorted) light pattern is extracted and combined to a final topographic image having an enhanced spatial resolution. In Figure 20, extracted distortions generated by shifting the mesh are shown. Dashed white lines are used as markers to visualize the shift of the light pattern to the right in the bottom image. The depicted shift has been obtained after five steps to the right. Thus, a spatial resolution of the topography on the micrometer scale is feasible.

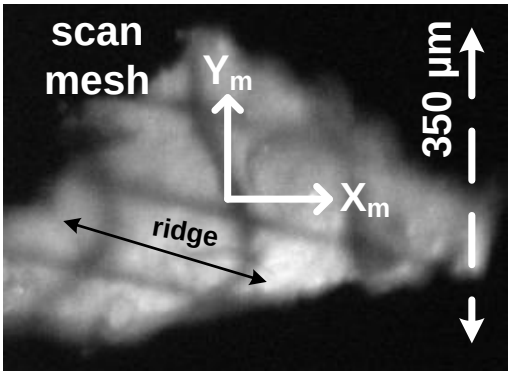

Figure 18: In order to enhance the spatial resolution of the surface structure, the micromesh is scanned stepwise over the sample in X_m and Y_m-direction; the subscript $_{m(esh)}$ refers to the X_m -Y_m plane of the mesh (Figure 11, center) which does not coincide with the X-Y plane shown in Figure 11 (bottom, center). After extracting the light pattern at each position all extracted patterns are combined. This enhances the spatial resolution of the sample's topography and allows a better assessment of (here) the sample's ridge structure.

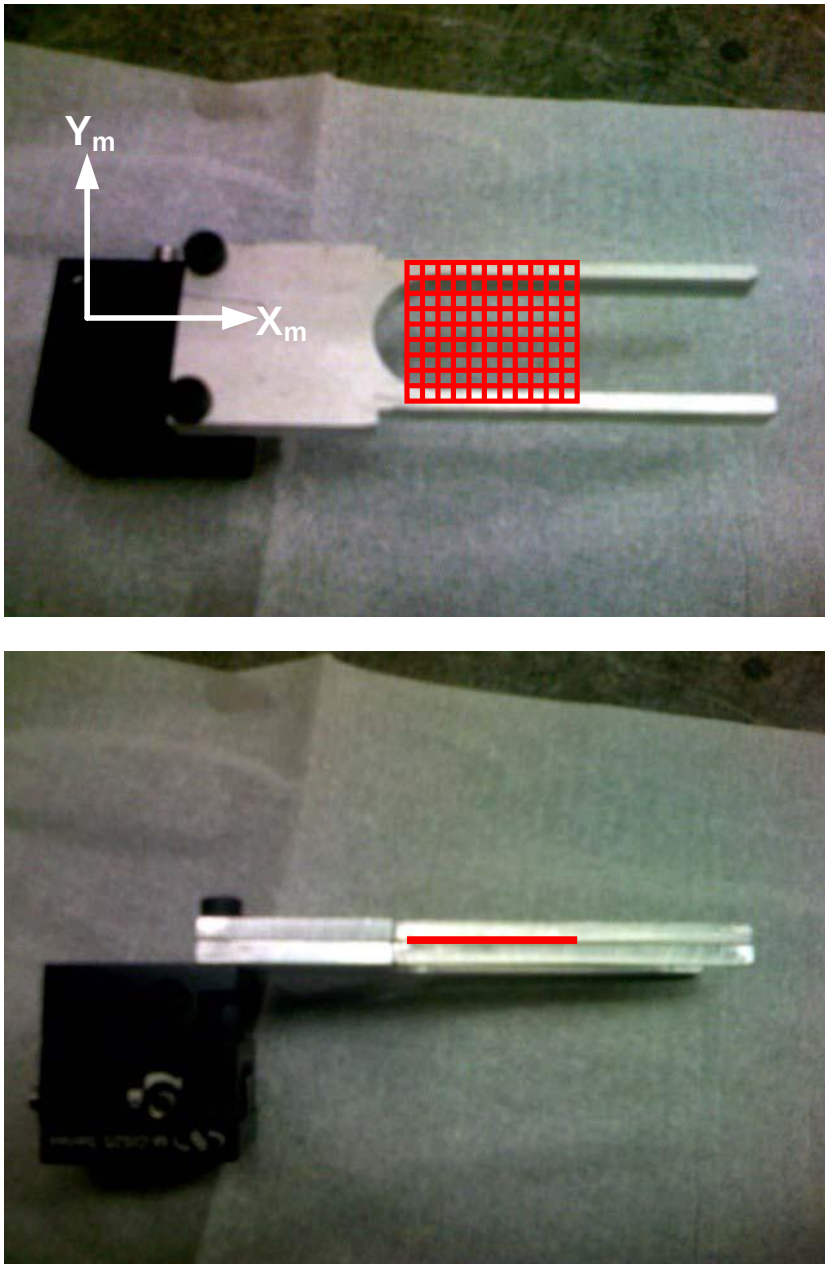

Figure 19: (top) Top and (bottom) side views of U-shaped mesh holder. The mesh holder is mounted to an X-Y translation stage that allows is to be scanned in both X- and Y-directions. An image is acquired at each position, then are combined into a final image having an enhanced spatial resolution. The mesh is simulated in red.

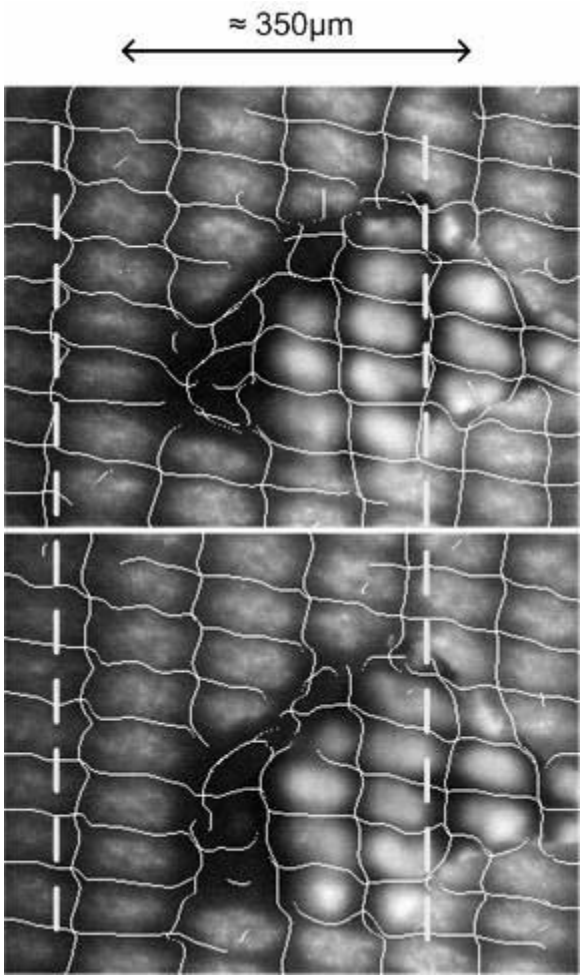

Figure 20: Demonstration of enhancing the spatial resolution of the topography probing by stepwise scanning the mesh in the X_m -Y_m plane. The mesh in bottom image is shifted to the right relative to the top image; the white dashed lines serve as markers. The arrow shown at the top gives an approximate length scale; spatial resolutions on the micrometer scale are feasible.

4.2.6 Height Calibration

Once the light pattern and its distortions have been extracted, these distortions measured in pixels can be related to the samples' true physical height (i.e. topography) in micrometers. This can be useful if height information about a bacterial colony is of interest. For this purpose, 'calibration objects' of known dimensions are placed on the microscope stage and are illuminated with the light pattern. For these investigations, pillars of different heights (115±2.4 µm, 200±1.3 µm, 211±1.5 µm, 250±2.5 µm, and 310±1.4 µm) have been milled out of a Teflon slide (Figure 21, right). The footprint and the heights of these pillars have been determined under a microscope in topview and sideview, respectively. The extraction algorithm determines the induced distortions and derives a translation from measured pixel shifts to known heights (Figure 21, left). Using a flat substrate as a "baseline" will establish the undisturbed light pattern (no height); from this the amount of distortion caused by the calibration objects can be measured and related to calibration heights.

4.3 Conclusions

Determining the topography of samples is important for the chemical sensing of reactions that proceed in three spatial dimensions, such as the growth of bacterial contaminations, and requires the need to collect both spectroscopic information and topographic information from a sample. This topographic information is inherently lost when using spectroscopic imaging. A spectroscopic imaging setup has been augmented in order to gain topographic information from samples along with spectroscopic data. This setup is readily applicable to the detection and growth monitoring of small bacterial colonies.

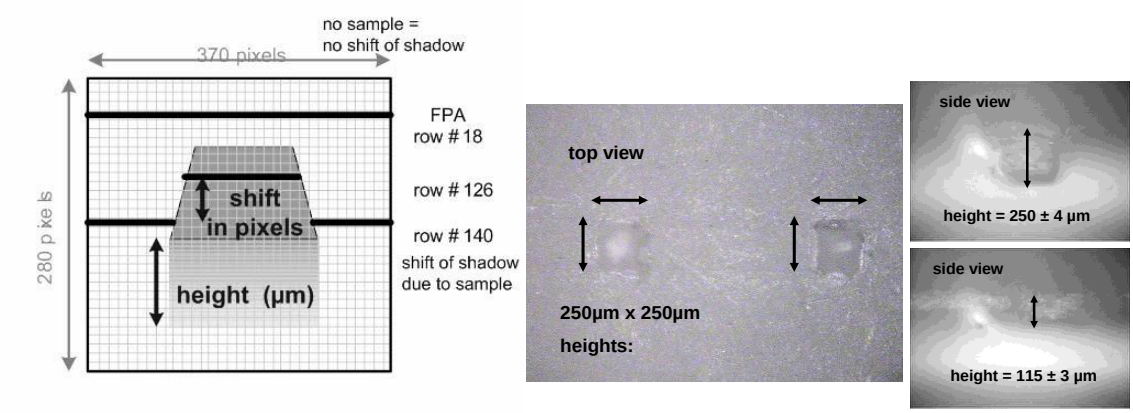

Figure 21: (left) This schematic depicts the use of calibration objects for relating their known heights in µm to distortions of the light patterns (compare Figure 14 - Figure 20) induced by them; (right) these microscopic calibration objects (only two are shown) have been milled out of a white Teflon slide. The footprint (approx. 250 µm x 250µm) and the heights of the pillars have been determined under a microscope. After projecting a light pattern onto one or several of these pillars the pixel shift-to-height relation can be determined.

For topographic investigation, external illumination optics for the visible wavelength range have been developed that project a regular light pattern onto the 3D samples. This light pattern is generated by placing a micromesh into the illumination beam path. If this light pattern is projected onto a flat substrate, it is preserved. If, however, the samples have a 3D surface topography, the regular light pattern is distorted. Information about surface structures is encoded into these distortions. By means of a novel algorithm the (distorted) light patterns are extracted from images acquired with a CCD camera. For translating these distortions into physical height information a calibration is required. This calibration is derived by means of calibration objects of known dimensions placed under the same illumination optics. Since these calibration objects also induce a distortion of the light pattern, their known heights can be related to induced shifts. Due to space limitations under the microscope objective and due to the fragility of micromeshes used, the light patterns are limited to length scales on the order of tens of micrometers. Because bacterial contaminations start out as colonies that are small, higher resolutions are required. This can be accomplished by scanning the mesh by means of a microstage, and has been shown to enhance the topographic spatial resolution.

Chapter 5

Summary and Conclusions

Methods of bacterial identification currently in place involve the collection of numerous possible sources of contamination, extraction techniques, culturing, and then finally identification. This can take a considerable amount of time, which is not conducive to the timely treatment of potentially fatal bacterial infections. Due to its speed, nondestructive nature, ease-of-use, and the wealth of information it provides, infrared spectroscopy has become a useful analysis technique in the field of bacteriology. FTIR is fast, requires little sample handling, no expensive dyes or labels, and yet, can detect infrared-absorbing components within bacterial samples. These components produce strong features with a good signal-to-noise. While the spectroscopic features produced by different strains of bacteria may not be visibly different, it has been shown here that, using a correlation analysis, three spectroscopically similar strains of *E. coli* bacteria could be differentiated from one another based on statistically significant differences in their spectra.

FTIR spectroscopy can quickly discriminate similar strains of bacteria if their location is known. However, bacterial colonies are often small and localized within a larger area. The first part of chapter three presented a method of finding such colonies by visualizing information obtained by spectroscopic imaging. This method utilized a multivariate image analysis to visualize the chemical information obtained from different materials as different colors. Due to the spectral differences between a bacterial colony and the substrate used here, three bacterial colonies were represented as noticeably different colors than the substrate. By merging data cubes, the detection of bacterial contaminations within an extended sample area has been shown to be feasible.

Once a bacterial colony has been found, it is important to correctly identify the contamination at the strain level in order to provide correct medical treatment. The second part of chapter three expanded the scope by introducing a second species of bacteria in order to assess the capability of an identification method. This method utilizes

a principal component analysis to model ‘gold standard’ spectra of known species and strains, evaluates a spectrum from a spectroscopic imaging data cube, and constructs a second version of that spectrum. A correlation analysis is performed, calculating the degree of similarity between the original spectrum and the reconstructed spectrum, providing a measure of how similar the original spectrum is to the ‘gold standard’. To maximize the number of correct classifications and the percentages of classification, two adjustments were made to the method. By making these two adjustments, the identification method succeeded in achieving both high percentages of classification as well as a large number of correct classifications.

While spectroscopic imaging is able to determine the spatial distributions of chemical information in an X-Y plane, the investigation of chemical reactions proceeding in three dimensions (X-Y-Z), such as growing bacterial contaminations, requires the acquisition of both chemical information *and* topographic information. Capturing visible images of a three-dimensional object will inherently lose height information. An external optical setup for probing the surface topographies of three-dimensional samples has been presented in chapter four. This setup uses a micromesh inserted into the beam path of a visible illumination unit to project a regular light pattern onto the surface of a sample. If only a flat substrate is present, the regularity of the light pattern is retained. However, if a sample having a complex surface topography is present, the light pattern becomes distorted. Encoded within these distortions is height information, which can be extracted by means of a novel method developed here, and related to actual height information using calibration objects of known dimensions. The spatial resolution at which a sample is probed can be enhanced, if needed, by scanning the micromesh in X and Y, extracting the distortions at each position, then recombining all extracted distortions into one image having a higher spatial resolution. Because this setup is coupled to a spectroscopic imaging setup, the growth monitoring of bacterial colonies can be achieved.

List of References

1 Gilbert, M.; Luttrell, R.; Stout, D.; Vogt, F. *J. Chem. Ed.* **2008**, *85*, 135.

2 Gilbert, M.; Frick, C.; Wodowski, A.; Vogt, F. *Appl. Spectrosc.* **2009**, *63*, 6.

3 Gilbert, M.; Vogt, F. *Anal. Chem.* **2007**, *79*, 5424-5428.

4 World Health Organization, "Food Safety and Foodborne Illness", http://www.who.int/mediacentre/factsheets/fs237/en/, accessed June 9, 2009.

5 Vaneechoutte, M.; Van Eldere, J. *J. Med. Microbio.* **1997**, *46*, 188.

6 Belgrader, P.; Benett, W.; Hadley, D.; Richards, J.; Stratton, P.; Mariella Jr., R.; Milanovich, F. *Science*, **1999**, *284*, 449.

7 Karasinski, J.; White, L.; Zhang, Y.; Wang, E.; Andreescu, S.; Sadik, O.; Lavine, B.; Vora, M. *Biosens. Bioelectron.* **2007**, *22*, 2643.

8 Neufeld, T.; Schwartz-Mittlemann, A.; Biran, D.; Ron, E.; Rishpon, J. *Anal. Chem.* **2003**, *75*, 580.

9 Goodridge, L.; Chen, J.; Griffiths, M. *Int. J. Food Microbiol.* **1999**, *47*, 43.

10 Siegrist, T.; Anderson, P.; Huen, W.; Kleinheinz, G.; McDermott, C.; Sandrin, T. *J. Microbiol. Methods* **2007**, *68*, 554.

11 Rajakaruna, L.; Hallas, G.; Molenaar, L.; Dare, D.; Sutton, H.; Encheva, V.; Culak, R.; Innes, I.; Ball, G.; Sefton, A.; Eydmann, M.; Kearns, A.; Shah, H. *Infect. Genet. Evol.* **2009**, *9*, 507.

12 Degand, N.; Carbonnelle, E.; Dauphin, B.; Beretti, J.; Bourgeois, M.; Sermet-Gaudelus, I.; Segonds, C.; Berche, P.; Nassif, X.; Ferroni, A. *J. Clin. Microbiol.* **2008**, *46*, 3361.

13 Diedrich, J.; Rehse, S.; Palchaudhuri, S. *J. Appl. Phys.* **2007**, *102*, 014702-1.

14 Xie, C.; Mace, J.; Dinno, M.; Li, Y.; Tang, W.; Newton, R.; Gemperline, P. *Anal. Chem.* **2005**, *77*, 4390.

15 Patel, I.; Premasiri, W.; Moir, D.; Ziegler, L. *J. Raman Spectrosc.* **2008**, *39*, 1660.

16 Sandt, C.; Madoulet, C.; Kohler, A.; Allouch, P.; De Champs, C.; Manfait, M.; Sockalingum, G.D. *J. Appl. Microbiol.* **2006**, *101*, 785.

17 Foster, N.; Thompson, S.; Valentine, N.; Amonette, J.; Johnson, T. *Appl. Spectrosc.* **2004**, *58*, 203.

18 Erukhimovitch, V.; Pavlov, V.; Talyshinsky, M.; Souprun, Y.; Huleihel, M. *J. Pharmaceut. Biomed. Anal.* **2005**, *37*, 1105.

19 Naumann, D.; Keller, S.; Helm, D.; Schultz, C.; Schrader, B. *J. Mol. Struct.* **1995**, *347*, 399.

20 Samuels, A.; Snyder, A.; Emge, D.; St. Amant, D.; Minter, J.; Campbell, M.; Tripathi, A. *Appl. Spectrosc.* **2009**, *63*, 14.

21 Lewis, E.N.; Treado, P.J.; Reeder, R.C.; Story, G.M.; Dowery, A.E.; Marcott, C.; Levin, I.W. *Anal. Chem.* **1995**, *67*, 3377.

22 de Juan A.; Tauler R.; Dyson R.; Marcolli C.; Rault M.; Maeder M. *Trend Anal. Chem.* **2004**, *23*, 70.

23 Geladi, P.; Grahn, H. *Multivariate Image Analysis*; John Wiley & Sons: Chichester; 1996.

24 Malinowski, E. *Factor Analysis in Chemistry*, 3rd ed.; John Wiley & Sons: New York; 2002.

25 Jolliffe, I. *Principal Component Analysis*, 2nd ed.; Springer: New York; 2002.

26 Martens, H.; Næs, T. *Multivariate Calibration*, 2nd ed.; John Wiley & Sons: New York; 1991.

27 Cody, S.; Glynn, M.; Farrar, J.; Cairns, K.; Griffin, P.; Kobayashi, J.; Fyfe, M.; Hoffman, R.; King, A.; Lewis, J.; Swaminathan, B.; Bryant, R.; Vugia, D. *Ann Intern Med.* **1999**, *130*, 202.

28 Herwaldt, B.; Beach, M. *Ann Intern Med.* **1999**, *130*, 210.

29 Todar, K. "Pathogenic E. coli", *Online Textbook of Bacteriology*. University of Wisconsin-Madison Department of Bacteriology, accessed June 2008.

30 Choo-Smith, L.; Maquelin, K.; Van Vreeswijk, T.; Brunning, H.; Puppels, G.; Thi, N.; Kirschner, C.; Naumann, D.; Ami, D.; Villa, A.; Orisini, F.; Doglia, S.; Lamfarraj, H.;

Sockalingum, G.; Manfait, M.; Allouch, P.; Endtz, H. *Appl. Environm. Microbio.* **2001**, *67*, 1461.

31 Maquelin, K.; Kirschner, C.; Choo-Smith, L.; van den Braak, N.; Endtz, H.; Naumann, D.; Puppels, G. *J. Microbiol. Methods* **2002**, *51*, 255.

32 Al-Holy, M.; Lin, M.; Cavinato, A.; Rasco, B. *Food Microbio.* **2006**, *23*, 162.

33 Al-Qadiri, H.; Lin, M.; Cavinato, A.; Rasco, B. *Int. J. Food Microbio.* **2006**, *111*, 73.

34 Kim, S.; Burgula, Y.; Ojanen-Reuhs, T.; Cousin, M.; Reuhs, B.; Mauer, L. *J. Food Sci.* **2006**, *71*, M57.

35 Nadtochenko, V.; Rincon, A.; Stanca, S.; Kiwi, J. *J. Photochem. Photobiol., A.* **2005**, *169*, 131.

36 http://www2.chass.ncsu.edu/garson/PA765/correl.htm, accessed October 2008

37 Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C,* 2nd ed.; Cambridge University Press: New York; 1992.

38 Draper, N.; Smith, H. *Applied Regression Analysis*, John Wiley & Sons, New York, 1998 (3rd ed.)

39 http://www.foodsafey.gov/~mow/chap12.html, accessed April 2009.

40 Vogt, F. unpublished results.

41 Gilbert, M.; Vogt, F. *Spectroscopic Imaging and Chemometric Methods for Detection and Discrimination of Different E. coli Strains*, Presentation at: PITTCON, Chicago, IL, March **2009**

42 Burke, R.; Gilbert, M.; Khandal, D.; Giordano, M.; Vogt, F. *Using Microalgae Biodiversity along with FTIR and Chemometric Methods to Detect Environmental Changes in Marine Ecosystems*, Presentation at: PITTCON, Chicago, IL, March **2009**

43 Susi, H.; Byler, M.; *Biochem. Biophys. Res. Comm.* **1983**, *115*, 391.

44 Susi, H.; Byler, M. *Arch. Biochem. Biophys.* **1987**, *258*, 465.

45 Kong, J.; Yu, S. *Acta Bioch. Bioph. Sin.* **2007**, *39*, 549.

46 Mandel, J. *Am. Statist.* **1982**, *36*, 15.

47 Vogt, F.; Mizaikoff, B. *J. Chemometrics* **2003**, *17*, 346.

48 Colarusso, P.; Kidder, L.; Levin, I.; Fraser, J.; Arens, J.; Levis, E. *Appl. Spectrosc.* **1998**, *52*, 106A.

49 Wold, J.; Kvaal, K. *Appl. Spectrosc.* **2000**, *54*, 900.

50 Shaw, G.; Manolakis, D. *IEEE Signal Process. Mag.* **2002**, *19*, 12.

51 Landgrebe, D. *IEEE Signal Process. Mag.* **2002**, *19*, 17.

52 Manolakis, D.; Shaw, G. *IEEE Signal Process. Mag.* **2002**, *19*, 29.

53 Lawrence, K.; Windham, W.; Park, B.; Buhr, R. *J. Near Infrared Spectrosc*. **2003**,*11*, 269.

54 Liu, Y.; Windham, W.; Lawrence, K.; Park, B. *Appl. Spectrosc.* **2003**, *57*, 1609.

55 Tran, C. *Appl. Spec. Rev*. **2003**, *38*, 133.

56 Yu, H.; Macgregor, *J. Chemom. Intell. Lab. Syst.* **2003**, *67*, 125.

57 Vogt, F.; Dable, B.; Cramer, J.; Booksh, K. *Analyst*. **2004**, *129,* 492.

58 Vogt, F.; Cramer, J.; Booksh, K. *J. Chemom.* **2005**, *19*, 510.

59 Beleites, C.; Steiner, G.; Sowa, M.; Baumgartner, R.; Sobottka, S.; Schackert, G.; Salzer, R. *Vib. Spectrosc.* **2005**, *38*, 143.

60 Fernandez, D.C.; Bhargava, R.; Hewitt, S.M.; Levin, I.W. *Nat. Biotechnol.* **2005**, *23*, 469.

61 Romeo, M.J.; Diem, M. *Vib. Spectrosc.* **2005**, *38*, 115.

Appendix

Summary of Graduate School Honors, Publications, and Presentations

Honors

1. Eastman Graduate Summer Fellowship, Eastman Chemical Company, Kingsport, TN, April **2008**
2. Gleb Mamantov Graduate Chemistry Scholar, Department of Chemistry, University of Tennessee, Knoxville, TN, April **2008**
3. C.W. Keenan Outstanding Graduate Teaching Award, Department of Chemistry, University of Tennessee, Knoxville, TN, April **2007**

Publications

1. M.K. Gilbert, C. Frick, A. Wodowski, F. Vogt

   Spectroscopic Imaging for Detection and Discrimination of Different *E. coli* Strains, *Applied Spectroscopy 63(1)*, 6-13.

2. M.K. Gilbert, R.D. Luttrell, D. Stout, F. Vogt

   Introducing Chemometrics to the Analytical Curriculum—Combining Theory and Lab Experience, *Journal of Chemical Education**, 2008**, 85(1),* 135-137 *(+45 pages of supplemental material)*

3. M.K. Gilbert, F. Vogt

Augmenting Spectroscopic Imaging for Analyses of Samples with Complex Surface Topographies, *Analytical Chemistry**, 2007**, 79(14),* 5424-5428

4. R.D. Luttrell, M.K. Gilbert, F. Vogt

   Composing Hybrid Wavelets for Optimum and Near-Optimum Representation and Accelerated Evaluation of N-way Data Sets, *Journal of Chemometrics, **2007**, 21(1-2),* 65-75

Presentations

1. M.K. Gilbert, F. Vogt
   Spectroscopic Imaging and Chemometric Methods for Detection and Discrimination of Different E. coli Strains, Presentation at: PITTCON, Chicago, IL, March **2009**

2. M.K. Gilbert, F. Vogt
   Spectroscopic Imaging and Chemometric Methods for Detection and Discrimination of Different E. coli Strains, Presentation at: Eastman Chemical Company, Kingsport, TN, February **2009**

3. M.K. Gilbert, C. Frick, A. Wodowski, F. Vogt
   Spectroscopic Imaging for Detection and Discrimination of Different E. Coli Strains*,* Presentation at: FACSS, Reno, NV, October **2008**

4. M.K. Gilbert, F. Vogt
   Augmenting Spectroscopic Imaging for Analyses of Samples with Complex Surface Topographies, Poster at: Pittcon, New Orleans, LA, March **2008**

5. M.K. Gilbert, F. Vogt
   Augmenting Spectroscopic Imaging for Analyses of Samples with Complex Surface Topographies, Presentation at: FACSS, Memphis, TN, October **2007**

Vita

Michael Gilbert graduated from Shelby County High School in May 2000. He went on to attend Bellarmine University in Louisville, KY and received a B.A. in Chemistry in May 2005. Michael entered the University of Tennessee as a graduate student in August 2005 and joined the research group of Dr. Frank Vogt in December 2005. He will receive a doctorate in Chemistry under Dr. Vogt's direction in summer 2009. After graduating, Michael will join the faculty at Texas Wesleyan University in Fort Worth, TX as an Assistant Professor of Analytical Chemistry.