HYPERSPECTRAL MICROSCOPY FOR EARLY AND RAPID DETECTION OF

*Salmonella* serotypes

by

MATTHEW BRENT EADY

(Under the Direction of Bosoon Park)

ABSTRACT

Optical microbial detection methodologies have shown the potential as early and rapid pathogen detection methods. This proof-of-concept project explores the use of hyperspectral microscopy as a potential method for early and rapid classification of five *Salmonella* serotypes. Darkfield hyperspectral microscope images were collected at early incubation times of 6, 8, 10, and 12 hrs., then compared to 24 hrs. incubation. Hypercube data collected from cells were analyzed through multivariate data analysis (MVDA) methods to assess classification ability of early incubation times. A separate experiment was conducted to explore the ability of analyzing hyperspectral data collected from only informative spectral bands as opposed to the original 89 spectral bands. MVDA showed that incubation times as early as 8 hours had near identical spectra and classification abilities compared to 24 hours; while the original 89-point spectrum could be reduced to 3 selected spectral bands and maintain high serotype classification accuracy.

INDEX WORDS: *Salmonella*, hyperspectral microscopy, early detection, rapid detection, multivariate data
HYPERSPECTRAL MICROSCOPY FOR EARLY AND RAPID DETECTION OF

SALMONELLA SEROTYPES

by

MATTHEW BRENT EADY

BSA, The University of Georgia, 2011

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2014
HYPERSPECTRAL MICROSCOPY FOR EARLY AND RAPID DETECTION OF

SALMONELLA SEROTYPES

by

MATTHEW BRENT EADY

Major Professor: Bosoon Park
Committee: Mark Harrison
            Fanbin Kong

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
December 2014
DEDICATION

To my wife and children:

I dedicate this body of work in memory of my daughter, Amelie Eady, whose short time here taught me the true meaning of the word determination and inspired me to pursue the field of food safety, to my son, Judah Eady, who at the end of the day reminds me what all the effort is for, and finally, to my wife, Audrey Blackwell Eady, for her unshakable support of my education through the best of times, the worst of times, and everything in-between.
ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Bosoon Park, for the opportunity to be a part of his lab and his commitment to this project. I would also like to acknowledge my committee members, Drs. Mark Harrison and Fanbin Kong, for their time and input. I am grateful to Dr. Nasreen Bano for her expertise in microscopy and microbiological techniques. Thank you to Dr. Darlene Samuel for giving me my start with the USDA-ARS and the confidence to pursue this endeavor, and I am also grateful to Dr. Brian Bowker for his advice on the scientific writing and the peer review processes.

The past few years would not be possible without the support and encouragement of my wife, Audrey Blackwell Eady. Her constant support and sacrifices over the past few years have made this body of work possible. I would also like to thank my mother, Elizabeth Eady, and brother, Jake Eady, for their encouragement and support. Also acknowledged are Brenda and Larry Blackwell for making the many trips to Athens, GA over the years to ease the stress-filled days.

Sincerest appreciation,

Mathew Eady
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................vi
LIST OF TABLES ...................................................................................................................ix
LIST OF FIGURES ...............................................................................................................x

CHAPTER

1 INTRODUCTION ..............................................................................................................1
   References .......................................................................................................................6
   Figure .............................................................................................................................10

2 LITERATURE REVIEW ..............................................................................................11
   Optical detection of bacteria .....................................................................................11
   Hyperspectral imaging (HSI) and the hypercube .....................................................13
   Food quality applications of HSI .............................................................................14
   Food safety applications of HSI ...............................................................................14
   Hyperspectral microscope imaging (HMI) .............................................................15
   Data preprocessing methods .....................................................................................16
   HSI and chemometrics ...............................................................................................17
   Informative spectral band selection ..........................................................................20
   References ....................................................................................................................21
   Table .............................................................................................................................28
3 RAPID AND EARLY DETECTION OF *SALMONELLA* SEROTYPES 
WITH HYPERSPECTORAL MICROSCOPE AND MULTIVARIATE DATA

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>29</td>
</tr>
<tr>
<td>Introduction</td>
<td>30</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>31</td>
</tr>
<tr>
<td>Results</td>
<td>33</td>
</tr>
<tr>
<td>Discussion</td>
<td>38</td>
</tr>
<tr>
<td>References</td>
<td>40</td>
</tr>
<tr>
<td>Figures and Tables</td>
<td>44</td>
</tr>
</tbody>
</table>

4 VISIBLE NEAR-INFRARED HYPERSPECTORAL MICROSCOPY AND 
INFORMATIVE BAND SELECTION FOR CLASSIFICATION OF 
*SALMONELLA ENTERICA* SEROTYPES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>55</td>
</tr>
<tr>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>57</td>
</tr>
<tr>
<td>Results</td>
<td>60</td>
</tr>
<tr>
<td>Discussion</td>
<td>66</td>
</tr>
<tr>
<td>References</td>
<td>67</td>
</tr>
<tr>
<td>Figures and Tables</td>
<td>70</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>References</td>
<td>82</td>
</tr>
</tbody>
</table>

6 APPENDICES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>References</td>
<td>85</td>
</tr>
</tbody>
</table>
A Principal component analyses score plots from reduced variable selection

sets..............................................................................................................85
LIST OF TABLES

Table 2.1: Overview of HIS methods applied to food for quality and safety assessment ..................................................28
Table 3.1: Difference of serotype values from a one-way ANOVA Fisher test, to compare relative average spectral peak ratios (590 nm / 538 nm) at incubation times of 8 - 24 hrs .............................................................................................................52
Table 3.2: Mahalanobis distance (MD) values between serotype clusters calculated from PCA score plots at incubation times of 8 - 24 hrs ...........................................................................................................53
Table 3.3: Soft-independent modeling of class analogy (SIMCA) for all serotypes within each incubation period ........................................................................................................................................54
Table 4.1: Description of variable sets selected for analyses ..................................................................................................................80
Table 4.2: Confusion matrix results obtained from support vector machine classification (SVMC) ..............................................................................................................................................81
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The hyperspectral microscope setup</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Hyperspectral microscope images of <em>S</em>. Typhimurium cells at different incubation times: (a) 6 hrs., (b) 8 hrs., (c) 10 hrs., (d) 12 hrs., and (e) 24 hrs.</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>(a) Average spectra for <em>S</em>. Enteritidis normalized to 546nm, (b) average <em>S</em>. Enteritidis spectra with pretreatment algorithm applied.</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>PCA score plots with loading vectors for four incubation times: (a) 8 hrs., (b) 10 hrs., (c) 12 hrs., and (d) 24 hrs.</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>Sample collection and analysis flow chart</td>
<td>75</td>
</tr>
<tr>
<td>4.2</td>
<td>A mosaic of the 590 nm band with each variable selection set image (A-E) displaying the obtained HMI with a highlighted cell (top), with a respective 3D surface plot reproducing the highlighted pixels (middle), and accompanying spectra of the collected bands for the highlighted cell.</td>
<td>76</td>
</tr>
<tr>
<td>4.3</td>
<td>Spectral band variance, normalized to each band, (b) loading vectors for PCA score plots, and (c) average spectra for 5 <em>Salmonella</em> serotypes, vertical lines represent the range of bands selected for variable sets B-E (586 – 662 nm).</td>
<td>77</td>
</tr>
<tr>
<td>4.4</td>
<td>Example of how support vector machine (SVM) creates a hyperplane for classification of nonlinear data.</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of the 590-598 nm spectral range collected at each of the five variable selection sets, for <em>S</em>. Enteritidis.</td>
<td>79</td>
</tr>
</tbody>
</table>
Appendix A: Principal component analysis for reduced spectral band models A-E……..85
CHAPTER 1
INTRODUCTION

The detection of foodborne disease causative agents such as *Salmonella* is a global concern, highlighted by disease outbreaks in recent years. Market demand and advances in transportation logistics have brought foods from around the world to our doorsteps in a matter of hours (Käferstein et al., 1997). While this enhanced availability has placed fresher and more exotic products in consumer markets, it has also been accompanied by food safety concerns regarding traceability and accountability. The 2011 outbreak of Shiga toxin-producing *E. coli* (STEC) O104 in organic bean sprouts grown in the Lower Saxony region of Germany is an example of the potential for rapid global disease spread. Before positively identifying the sprouts as the contaminated product, Spanish cucumbers were incorrectly labeled as the pathogen-harboring product, resulting in an estimated loss equivalent to $200M US dollars per week for exporters (CDC, 2011). The sprouts were later positively identified as the source of the outbreak, tracing the problem back to Egyptian farms where fenugreek seeds were cultivated and contaminated (Bielaszewska et al., 2011; Buchholz et al., 2011). This outbreak highlights the spread of global agricultural commerce and the need for early and rapid foodborne disease detection methods.

The CDC estimated that in 2011 the United States had 47.8 million people affected by foodborne diseases, and approximately 3,000 deaths occurred as a result (CDC, 2014). *Salmonella*, a facultative intracellular pathogen of interest in many food
products, can result in gastroenteritis and, in some cases, death. Majowicz et al., 2010 estimated that there are 80.3 million global cases of foodborne nontyphoidal salmonellosis each year with 155,000 associated deaths.

Plating techniques involving nutrient-enriched growth media and immunological testing methods, such as enzyme-linked immunosorbent assay (ELISA), or molecular methods, such as polymerase chain reaction (PCR), have been the standard microbiological testing methods for foodborne bacteria for years (Barlen et al., 2007). While these methods are reliable, they have their disadvantages; confirmation through traditional plating media can take anywhere from several days to weeks due to incubation times and rounds of presumptive testing followed by confirmation plating (Ellis and Goodacre, 2006). Investigators are rushing to positively identify the contaminated products and can potentially name the wrong product, such as the previously-mentioned Spanish cucumbers. Molecular and immunological testing methods are less time-consuming; however, methods such as PCR require extensive training and can become expensive due to the high recurring cost of target-specific antibodies and pathogen-specific reagent kits, that are necessary for lysing the cell and extracting the DNA before amplification (Jay et al., 2005; Mullins et al., 1987).

Apart from the human health effects caused by foodborne disease outbreaks, there is also an economic impact. Production of broiler chickens and peanuts are integral agricultural industries to the State of Georgia’s economy. Other large agricultural industries in the State of Georgia include blueberries, onions, peaches, and oysters, all of which can be susceptible to a variety of bacterial contaminations. In 2009, a *Salmonella* Typhimurium outbreak occurred at a peanut processing facility in Blakely, Georgia
resulting in 471 reported illnesses and nine deaths, as well as the shutdown of a company and the loss of approximately two hundred jobs in the state (CDC, 2012). The outbreak caused a dip in peanut sales nationwide for several months before returning to normal (Wittenberger, 2010). The CDC’s PulseNet is a national disease tracking network of public health agencies and food regulatory operations, which recognize a pattern of unique DNA fingerprints for *Salmonella* and other pathogenic bacteria through pulsed-field gel electrophoresis (PFGE) (Murase et al., 1995). Because it typically takes two to three weeks from ingestion of a contaminated food source until a positive diagnosis of disease from an outbreak, it is critical to reduce the days or weeks that may be required to then correctly match the disease to a strain isolated from a food source to minimize the spread of an outbreak.

In February of 2013, the Food Safety and Inspection Service (FSIS), of the United States Department of Agriculture (USDA), implemented more stringent allowable *Salmonella* limits for poultry processors. A method of rapid and early detection could potentially assist production facilities in meeting federal regulatory standards, while reducing the chance of production slowdowns and yield-loss due to extended quality control checks that rely on time-consuming incubation periods for microbial enumeration, and an early and rapid detection methodology could be implemented in-house to assist in meeting federal standards with less of a strain on production yields.

Optical detection has shown potential in detecting and classifying bacteria through spectroscopic methods that identify bacteria based on a spectral signature unique to the organism (Fan et al., 2008). Previously, spectroscopic instruments such as Raman, Fourier transform infrared (FTIR), laser-induced breakdown spectroscopy (LIBS), and
surface plasmon resonance (SPR) have been utilized for obtaining unique spectral signatures (Perkins and Squirrell et al., 2000; Rehse et al., 2010; Sundaram et al., 2013; Yang et al., 2014). Once the sample’s spectra are collected, multivariate data analysis (MVDA) methods are typically applied to classify samples. The resulting spectra can exhibit signs of collinearity with small differences in spectra. MVDA treats each spectral band as an independent variable; therefore, small differences in specific peaks can be identified as unique contributing factors to a specific bacterial species or serotypes. Other MVDA methods, such as the principal component analysis (PCA), soft independent modelling by class analogy (SIMCA), or partial least squares regression (PLSR), are typically applied to the selected spectra as a means of differentiating and classifying samples.

Hyperspectral imaging (HSI) is a technique that combines imaging with spectroscopy through the collection of data in a three-dimensional matrix. HSI has been most commonly associated with military astronomy, medicine and pharmaceuticals (Gowen et al., 2008). The ability to collect both spatial and spectral data from a sample offers an advantage over other spectroscopic techniques. Previously, HSI has been applied to a variety of agricultural commodities for quality measurements. Food safety measures have also been assessed with HSI on commodities such as broiler chickens (Park et al., 2006), with recent applications focusing on predicting the bacterial loads on chickens and differentiating non-O157:H7 Shiga-toxin producing E.coli. (STEC) (Feng and Sun, 2013; and Windham et al., 2013). The focus of this project is investigating the potential for hyperspectral microscopy (HMI) as a means of early and rapid classification
for pathogenic bacteria, as little research has been published on HMI as a means of foodborne bacteria detection. The HMI setup can be viewed in Figure 1.1.

This project investigates HMI data collected for five *Salmonella* serotypes (*S.* Enteritidis, *S.* Heidelberg, *S.* Infantis, *S.* Kentucky, and *S.* Typhimurium). The objectives of this research are: 1) to determine if HMI data collected for serotypes of the same bacterial species can be differentiated and if they can be differentiated at early incubation times of 6, 8, 10, and 12 hours as well as how the spectra compares to 24 hours incubation at 35-37°C, and 2) can the large data processing and storage requirement associated with hyperspectral imaging be reduced to only the necessary spectral bands while achieving a high classification accuracy in an effort to reduce the data processing and storage requirement.
References


Figure 1.1: The hyperspectral microscope setup
CHAPTER 2

LITERATURE REVIEW

Optical detection of bacteria

The exploration of optical detection methods for pathogenic bacteria is based on identifying from spectral signatures unique to the organism. Optical detection is typically viewed as a non-destructive rapid identification method that can potentially counter disadvantages of traditional detection methods such as nutrient-enriched growth media, immunological testing such as antigen/antibody interactions, and molecular methods such as polymerase chain reaction (PCR) (Jay et al., 2005). The use of nutrient-enriched growth media can take up to a week for results (Rose, 1998). While immunological testing and PCR require much less time, but there is also a high recurring cost due to cell lysing reagents and pathogen specific antibodies, with extensive training required for equipment associated with these methods (Barlen et al., 2007). Various spectroscopic approaches have been applied to bacterium identification, such as laser-induced breakdown spectroscopy (LIBS), Fourier-transform infrared (FTIR) spectroscopy, surface-enhanced Raman spectroscopy (SERS), and surface plasmon resonance (SPR) (Perkins and Squirrel, 2000; Rehse et al., 2010; Sundaram et al., 2013; Yang et al., 2014), with the objective of creating a reference library of spectral signatures that can be used when comparing unknown bacteria obtained from a suspect food product.

LIBS has seen wide remote applications in fields ranging from chemical analysis, military defense, and geology (Gottfried and Harmon, 2009). Use of the LIBS
system requires little sample preparation and a small sample size (Barnett et al., 2011). Rehse et al., (2010) stated that collection time of a bacterial sample with a LIBS system required only a few nanoseconds to obtain the spectra. A review of the literature shows that, for the research area of bacterial detection and classification, there is much variation in collection techniques with laser strength, type, and harmonizing levels, as well as the platform for observing the bacteria. Variations in platform include a nutrient-free agar, nutrient-enriched media, water suspensions, and creation of concentrated freeze-dried pellets of bacteria (Barnett et al., 2011; Baudelet et al., 2006; Gamble, 2014; Multari et al., 2010).

FT-IR collects data over a broad spectral range and requires the addition of a Fourier transformation algorithm applied to the raw data. Yang et al., 2014 used the FT-IR when searching for a rapid discrimination method for *Shigella sonnei* and *Escherichia coli* O157:H7 and collected absorption spectra between 4000-600 cm\(^{-1}\). The study found that focusing the spectra on a range of 1800-900 cm\(^{-1}\) highlighted the differences between the two species, which could then be differentiated by multivariate data analysis (MVDA).

SERS is another common method for the optical detection of bacteria. Typically, a substrate is prepared that enhances the Rayleigh scattering of light (Xu et al., 2013). Substrates are either roughened metallic substances, such as gold or silver, or are composed of nanoparticles. Compared to FT-IR, SERS has seen more bacterial identification applications, in part because FT-IR signals arise from asymmetrical molecular vibrations as opposed to SERS, which yields spectral bands arising from
symmetrical vibrations (Ellis and Goodacre, 2006). Water is also a weak Raman
scattering substance, which has a greater effect on FT-IR (Jarvis and Goodacre, 2008).

SPR-based immunoassays have also previously been developed for detecting
bacteria (Bokken et al., 2003; Mazundar et al., 2007; Perkins and Squirrell, 2000; Waswa
et al., 2007), which is done by binding bacterial cells to antibodies, creating an enhanced
signal for detection of SPR spectra with a biosensor (Barlen et al., 2007). While rapid
cell extraction and imaging can be achieved with this technology, there is once again the
high recurring cost of purchasing the antibodies in addition to the equipment training
requirements.

A review of the literature shows both advantages and disadvantages to these
optical detection methodologies. Some methods such as the LIBS system require
minimal sample preparation and a data acquisition time of only a few nanoseconds.
There is debate over an appropriate substrate for the bacteria sampling in addition to the
time requirement for fine-tuning the instrument settings. Other methods such as SERS,
SPR, or FT-IR may need a concentrated cell sample from the same strain, requiring
enrichment.

**Hyperspectral imaging (HSI) and hypercube**

There is an array of remote HSI applications ranging from astronomy,
pharmaceuticals, and medical applications (Gowen et al., 2008), and one advantage of
HSI over the previously mentioned spectroscopic techniques is the ability to collect data
in a three-dimensional hypercube. Two dimensions represent spatial data (x and y
coordinates) with a third spectral dimension (λ). Data are collected at each point over the
spectrum and at each pixel to form a band of the cube. Each band is combined together
and forms a three-dimensional cube. Because the first and second spatial dimensions are constant throughout the overlaying of images, the third dimension will yield a spectrum across all collected bands. Also, data from the hypercube can be analyzed on a band-by-band basis because each collected band produces a separate image.

**Food quality applications of HSI**

HSI has been implemented for analysis of quality attributes in a variety of agricultural commodities ranging from meats, dairy, fruits, and vegetables with an overview of examples from previously conducted research in this field listed in Table 2.1. This list is not all-inclusive, but is rather meant to provide an understanding of the variety in quality issues that have been previously investigated. This table also shows us that there has been an influx in research investigating HSI and food within the last 10 years. Projects have employed hyperspectral imaging to predict quality parameters within the commodity, such as predicting the drip-loss of salmon or assessing apples for bruising (He et al., 2014; Elmasry et al., 2008). Chao et al., (2007) demonstrated that hyperspectral imaging could be applied in a high-throughput setting to maintain production goals in a food processing facility.

**Food safety applications of HSI**

There has been far less published research applying HSI to food safety matters. Previously, work has been published in the area of applying HSI to detecting fecal matter on broiler chicken carcasses (Heitschmidt et al., 2007; Lawrence et al., 2007; Park et al., 2006; Windham et al., 2003). In general, an in-line hyperspectral imaging system was established for detection of fecal material present on broiler chicken carcasses during processing (Park et al., 2006). Images were collected and analyzed through a variety of
MVDA methods such as spectral angle mapper (SAM), Mahalanobis distances, support vector machine (SVM) classification, and principal component analysis (PCA). Once contaminated birds were identified, they could then be reprocessed through further decontamination steps.

Foodborne pathogen detection on a colony scale has also recently been approached with HSI. Windham et al., (2013) and Yoon et al., (2013) were able to classify Shiga toxin-producing \textit{E. coli} (STEC) bacterial colonies growing on a rainbow agar plate. These studies observe bacterial colonies, six non-O157:H7 serogroups (O26, O45, O103, O111, O126, and O145) of STEC, on agar plates. A five-pixel crosshair is used to select pixels within the colonies, followed by MVDA. Windham et al., (2013) found that using HSI to identify overall classification of the six serogroups through a k-nearest neighbor (kNN) cluster analysis approach was accurate 99% of the time, with approximately 4% of O103 colonies misclassified as O26 colonies.

Feng et al., (2013) applied near-infrared hyperspectral imaging with a spectral range of 930-1450nm to predict the total \textit{Enterobacteriaceae} load on chicken breast fillets. This family of bacteria represents organisms commonly found in poultry, such as \textit{Salmonella}, \textit{E. coli}, and \textit{Enterobacter}. The results of this standoff imaging method showed that through a combination of hyperspectral image collection and partial least squares regression (PLSR) the bacterial loads on fillets was correctly predicted with $R^2 \geq 81\%$ accuracy and a root mean squared error (RMSE) of $\leq 0.47 \log_{10} \text{CFU g}^{-1}$.

**Hyperspectral microscope imaging (HMI)**

HMI applies the concepts of hyperspectral imaging to microscopic images. Hypercube data are generated for the cells captured within a darkfield image. In theory,
this will create a sensitive detection method that will require only a few cells for
classification. Little published research exists on the topic of bacterial classification
through HMI. Previously, Anderson et al, (2008) investigated the use of HMI to
differentiate between live and dead endospores. This microscope setup used a tunable
filter, for filtering the tungsten halogen light into its hyperspectral state and collected 32
bands over the range of 400 – 720nm. *Bacillus* strains were imaged under immersion oil
and 1000x magnification. The study found that live-viable cells varied from peroxide
killed cells, but not chlorine killed cells. It was theorized that this was due to the
peroxide killed cells having extensive damage to their outer layers, while the chlorine did
not have the same effect on the endospores.

**Hyperspectral data preprocessing methods**

The purpose of preprocessing methods is to highlight the intrinsic differences in a
sample’s unique spectral signature of samples and minimizing spectral noise that can be
attributed to background spectra, object scattering, limb-darkening, or differences in
particle size (Amigo et al., 2013; Park et al., 2006). Spectroscopic data have a tendency
to be collinear (Mark and Workman, 1991). A data pretreatment algorithm can reduce
noise while focusing on differences between samples. A variety of data pretreatment
algorithms have been applied to preprocess hyperspectral image data. Perhaps the most
commonly used is the Savitzky-Golay 1st or 2nd derivative-based algorithm (Savitzky et
al., 1964). In addition to Savitzky-Golay, other preprocessing measures can be used such
as moving average smoothing treatments, Norris gap derivatives, baseline offsets, or
segments-size derivatives. These preprocessing steps share a common goal of enhancing
the intrinsic spectral differences between samples and reducing spectral noise.
**HSI and chemometrics**

The term “chemometrics” originated in the early 1970s; although the origins of the MVDA approach is said to have dated back decades prior (Esbensen, 1990). Because the hypercube generates a full spectrum for each pixel in an image, chemometrics is often implemented for handling the large amount of data generated from hyperspectral imaging, that can also tend to be collinear (Amigo, 2010). Chemometrics can be used to extract meaningful data from hyperspectral images while reducing dimensionality (Amigo et al., 2013). Factorial analyses reduce large amounts of hyperspectral data into easy-to-read plots that relate the common underlying principles of the data trends between samples. A review of the literature shows many MVDA methods are applied, such as; principal component analyses (PCA), partial-least squares regression (PLSR), support vector machine (SVM) classification, multivariate curve resolution (MCR), multivariate linear regression or (MLR) to name a few (Elmasry et al., 2012; Feng et al., 2012; Sun 2010). Multiple MVDA approaches are also listed in Table 2.1.

PCA is a commonly used analytical method. PCA can be useful in simplifying the analysis, and reducing the number of variables by decomposing the spectral X matrix:

\[ X = TP^T + E \]

where \( X \) is the data matrix, \( T \) represents the score matrix, \( P \) represents the loading matrix, and \( E \) represents the residuals or spectral noise (Wold et al., 1987). The primary tools of a PCA analysis are the scores plot and x-loading vector plot. The PCA scores plot shows two PCA component vectors plotted against each other representing the underlying footprint of the spectra, while the x-loadings plot can be viewed as a “map of variables” that explains the influence each observed band has on the spectra as a whole (Esbensen,
PCA is particularly useful in dealing with HSI samples due to the ability to handle a large amount of data. The PCA score plot is a clear way to quickly visualize the cluster relationship among samples, and a quantitative representation amongst the PCA clusters can help in assessing cluster separation. The Mahalanobis distance (MD) can be utilized in the principal component space and takes into account the correlation in data by using the inverse of the variance-covariance matrix from the data set, and measuring from the center of one cluster to the center of another cluster (De Maesschalck et al., 2000). The MD can be computed as:

$$
MD^2 = (x_i - x_j)^T S^{-1} (x_i - x_j)
$$

Where $x_i$ and $x_j$ are the averages of groups $i$ and $j$, $S$ is the covariance matrix, superscript $T$ is the transverse of the matrix, and $MD^2$ represents the Mahalanobis distance.

SVM classification is a method of kernel based nonlinear classification. When the data is not necessarily linear and multidimensional clustering of sample classes is present, a kernel classification is implemented for clustering based on the three dimensional space (Alonso et al., 2011). A kernel based recognition algorithm is applied for reclassification of the data classes based on the separation capabilities of the created hyperplane, that serves as an intermediate step:

$$
h = w^T x + b
$$

where $h$ is the hyperplane data, $x$ is the HMI input pixel vector, $w$ is the adaptable weight vector, $b$ represents the bias, and $T$ is the transverse operator (Sun, 2010). Samples from classes are then reclassified based on their separation with this hyperplane.

PLSR is a generalization of MLR, that is utilized when assessing highly collinear spectra with many variables (Wold et al., 2001). PLSR compares the similarities between
two data matrices (Hair et al., 2010). Because we are observing spectra of serotypes within the same bacterial species the Pearson correlation values tend to be highly collinear. In this research project we are using PLSR as a method of assessing the goodness-of-fit for a global data pretreatment model. The second data matrix is the collection of untreated raw spectral data obtained through another repetition of our experiment, while the same global data pre-treatment model is applied, and the similarity of PCA results is observed. A high $R^2$ and low root-mean squared error of calibration (RMSEC) or validation (RMSEV) is desired.

The soft-independent model classification analogy (SIMCA) is another tool that has been previously used for validating the results of a PCA (Sundaram et al., 2013). Instead of addressing the goodness-of-fit such as PLSR, SIMCA applies the previously applied pre-treatment algorithm to a data training set to predict the classification abilities of the validation repetition, based on the results of a previously completed PCA (Dunn III et al., 1978).

Table 2.1 shows a summary of HSI data collected from agricultural commodities including processing technique, spectral bands analyzed, and the chemometric analysis that was applied to the study. The table shows a variety of approaches dependent on the scope of the study, type of image, and the nature of the data produced from the microscope hypercube data. For this project, the Unscrambler V10.3 software program (CAMO, Oslo, Norway) was applied. This program is widely used throughout the literature for spectral processing.
Informative spectral band selection

One of the downsides to this methodology is processing and storing the large amount of data generated from the hypercube. It is common practice in dealing with HSI projects to select the minimal number of bands necessary to reduce the spectral noise from unnecessary bands, while still maintaining the naturally occurring variance (Sun 2010). Several methods for band selection exist such as analyzing the loading vectors of PCAs for the bands explaining the largest amount of variance (Windham et al., 2013). Another method of band selection focuses on a Monte Carlo approach for variable elimination by using a robust statistical model applied with PLSR (Gowen et al., 2012). PLSR has also been applied for informative band selection (Lawrence et al., 2004). Gowen et al., (2008) applied MLR for selection of optimal bands. There is a large amount of variance in MVDA and informative band selection methods. The nature of the biological material, amount of spectral bands, type of filtering apparatus, and lighting source are all play factors in selecting an appropriate MVDA and informative band selection method.
References:


Table: 2.1 Overview of HSI methods applied to food for quality and safety assessment.

<table>
<thead>
<tr>
<th>Product</th>
<th>Quality Attribute</th>
<th>Spectra</th>
<th>Band Range (nm)</th>
<th>MVDA Method</th>
<th># of Optimal Bands Selected</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>Surface defects</td>
<td>Reflectance</td>
<td>450-940</td>
<td>ANN</td>
<td>N/A</td>
<td>Ariana et al., 2006</td>
</tr>
<tr>
<td>Beef</td>
<td>Tenderness</td>
<td>Reflectance</td>
<td>400-1000</td>
<td>PCA, CDA</td>
<td>8</td>
<td>Naganathan et al., 2008</td>
</tr>
<tr>
<td>Cheese</td>
<td>Composition</td>
<td>Absorbance</td>
<td>1100-2498</td>
<td>PCA, PLS</td>
<td>N/A</td>
<td>Burger and Geladi 2006</td>
</tr>
<tr>
<td>Chicken</td>
<td>Fecal detection</td>
<td>Reflectance</td>
<td>430-900</td>
<td>Par, Min, Mah, Max, Sam, Bin, PLSR</td>
<td>N/A</td>
<td>Park et al., 2006</td>
</tr>
<tr>
<td>Chicken</td>
<td>Enterobacteriacea</td>
<td>Reflectance</td>
<td>930-1660</td>
<td>Par, Min, Mah, Max, Sam, Bin, PLSR</td>
<td>3</td>
<td>Feng et al., 2013</td>
</tr>
<tr>
<td>Non O157:H7 STEC E. coli</td>
<td>Colony differentiation</td>
<td>Absorbance</td>
<td>400-1000</td>
<td>PCA, kNN</td>
<td>N/A</td>
<td>Windham et al., 2013</td>
</tr>
<tr>
<td>Grain</td>
<td>Insect damage</td>
<td>Reflectance</td>
<td>1000-1600</td>
<td>PCA, Bin</td>
<td>2</td>
<td>Singh et al., 2008</td>
</tr>
<tr>
<td>Mushroom</td>
<td>Deterioration</td>
<td>Reflectance</td>
<td>400-1000</td>
<td>PCR, MLR</td>
<td>20</td>
<td>Gown et al., 2008</td>
</tr>
<tr>
<td>Spinach</td>
<td>E. coli</td>
<td>Reflectance</td>
<td>400-1000</td>
<td>PCA, ANN</td>
<td>N/A</td>
<td>Siripatrawan et al., 2011</td>
</tr>
<tr>
<td>Salmon</td>
<td>WHC</td>
<td>Absorbance</td>
<td>400-1753</td>
<td>PLSR, LS-SVM</td>
<td>9, 12, 13</td>
<td>Wu and Sun 2013</td>
</tr>
</tbody>
</table>

STEC = Shiga toxin producing E. coli, WHC = water holding capacity, ANN = artificial neural network, MLR = multiple linear regression, PCR = principal component regression, Par = parallelepiped, Min = minimum distance, Mah = Mahalanobis distance, Max = maximum likelihood, Sam = spectral angle mapper, Bin = binary, PCA = principal component analysis, CDA = canonical discriminate analysis, PLSR = partial least squares regression, LS-SVM = least-squares support vector machine, kNN = k nearest neighbor
CHAPTER 3

RAPID AND EARLY DETECTION OF SALMONELLA SEROTYPES WITH
HYPERSPECTRAL MICROSCOPE AND MULTIVARIATE DATA ANALYSIS

\(^1\)Eady, M.B., Park, B., and Choi, S. Submitted to “The Journal of Food Protection” on 08/04/14
Abstract

This study was designed to evaluate hyperspectral microscope images for early and rapid detection of *Salmonella* serotypes: *S. Enteritidis*, *S. Heidelberg*, *S. Infantis*, *S. Kentucky*, and *S. Typhimurium* at incubation times of 6, 8, 10, 12, and 24 h. Images were collected by an acousto-optic tunable filter (AOTF) hyperspectral microscope imaging (HMI) system, with a metal halide light source measuring 89 contiguous bands every 4 nm between 450 – 800 nm. Pearson correlation values were calculated for incubation times of 8, 10, and 12 hrs., compared to 24 hrs. to evaluate the change in spectral signatures from bacterial cells over time. Regions of interest (ROIs) were analyzed at 30% of the pixels in an average cell size per image. Preprocessing of the spectral data was performed by applying a global data transformation algorithm, and followed with a principal component analysis (PCA). The Mahalanobis Distance (MD) was calculated from PCA score plots for analyzing serotype cluster separation. Partial least squares regression (PLSR) was applied for calibration and validation of the model, while soft independent modeling of class analogy (SIMCA) was utilized to classify serotype clusters of training set. Pearson correlation values indicate very similar to spectral patterns for varying incubation times ranging from 0.987 to 0.999. PCA score plots show cluster separation at all incubation times, with MD values for incubation times ranging from 2.146 to 27.071. PLSR had a maximum RMSEC value of 0.0025 and RMSEV value of 0.0030. SIMCA correctly classified values at 8 h = 98.32%, 10 h = 96.67%, 12 h = 88.33%, and 24 h = 98.67% with the optimal number of principle components (four or five). The results of this study suggest that *Salmonella* serotypes
can be classified by applying a PCA to HMI data from samples at the earliest of 8 h incubation time.

**Introduction**

Identifying the causative agents and contaminated products of a foodborne disease outbreak can be a time consuming task. The use of nutrient enriched growth media has been the standard methodology for many years. This approach can prove to be lengthy with a best case scenario requiring several days (4). Methods such as PCR can reduce the necessary time by lysing cells and amplifying DNA sequences, while enzyme-linked immunosorbant assays (ELISAs) can reduce the required time through the binding of pathogen specific surface antigens to antibodies (12,19). These methods require extensive training with a high recurring cost of reagent kits and antibodies (13). These obstacles can be difficult to overcome when tracking foodborne disease outbreaks, due to the expedited measures of transportation logistics and rapid access to imported foods (14).

Optical detection offers a nondestructive rapid methodology for real-time recognition of pathogenic bacteria (20). Spectroscopic methods are utilized to produce a spectral signature that is unique to the organism. Multivariate data analysis (MVDA) methods are typically applied to categorize organisms based on the intrinsic differences in spectral signatures. Hyperspectral imaging (HSI) observes light in more bands than the three colors (red, green, and blue) recognizable by the human eye, and has the capability to compile contiguous wavebands for each pixel in an image (22). Hyperspectral microscope imaging (HMI) is a technique that collects microscopic data in a three dimensional hypercube matrix with two dimensions representing spatial data (x
and \( x \) and \( y \) coordinates), and a third representing spectral data (\( \lambda \)) that can be used as a tool in identifying the unique optical properties of bacteria \( (22) \). Machine learning algorithms could then potentially be applied to classify bacterial cells, resulting in a sensitive detection method.

In addition to the need for an early and rapid foodborne pathogen detection methodology for disease tracking purposes, there are potential economic benefits for food industry applications. A rapid nondestructive optical detection system can provide a means for maintaining the high throughput required for food production facilities, while streamlining quality assurance protocols. Early detection can dramatically reduce the amount of time that food products are kept in storage before being released to consumers. Because of the possibility of a foodborne disease outbreak companies have been known to implement a “test and hold” time period for sampling the microbial safety of new products, prior to release in consumer markets. Reducing the “test and hold” period to the shortest amount of time could have promising impacts on product safety, transportation logistics, and minimizing final product quality degradation. Clinical microbiology is another field that could benefit from the capabilities of early optical detection of pathogenic bacteria. By quickly diagnosing bacterial infections, healthcare professionals can implement specialized medical care sooner, providing the patient with a better chance for recovery. In addition to early detection, HMI offers the capabilities for rapid detection by eliminating some of the downsides to previously mentioned methodologies.

Previously, HSI systems have been applied to agricultural commodities studied as in-line fecal detection mechanisms for poultry processing facilities \( (11, 15, 17, 21, 23, \ldots) \).
pathogen detection in vegetables (26), safety characteristics of apples (16), and quality attributes of the beef industry (24). In 2008, live and dead Bacillus anthracis endospores were differentiated by HMI with fluorescent tags (1). Hyperspectral imaging has also been used in food safety applications on a macro-scale by predicting the Enterobacteriaceae loads, total viable bacteria count, and total Pseudomonas loads on raw broiler chicken breast meat (6, 7 8). Recently HMI was implemented as a way of differentiating between Shiga-toxin producing E. coli (STEC) serogroups on the colony scale (29, 32). HMI could be an effective tool for differentiating between various serotypes of a bacterial species at the cellular level, and comparing classification abilities at early detection times with micro-colonies inoculated less than 24 hrs. The objectives of this research were to 1) determine if the spectral signatures at 24 hrs. incubation time can be compared to the spectra of earlier time frames consisting of 6, 8, 10, and 12 hrs. of incubation, and 2) determine if the five Salmonella serotypes (S. Enteritidis, S. Heidelberg, S. Infantis, S. Kentucky, and S. Typhimurium) could be differentiated at the cellular level by their spectral signatures.

Materials and methods

Preparation of isolates and microscope slides

Glycerol stock cultures of S. Enteritidis (SE), S. Heidelberg (SH), S. Infantis (SI), S. Kentucky (SK), and S. Typhimurium (ST) were obtained from poultry carcasses rinses at the Poultry Processing and Swine Physiology Research Unit at the Russell Research Center in Athens, GA and stored at -80°C until needed. Samples were prepared for short term storage by inoculating agar slants and storing at 4°C. Fresh cultures were prepared by inoculating a few agar slant colonies into 10 ml of tryptic soy broth (TSB), incubating
samples at 35 to 37 °C for 18 to 24 h. The TSB cultures were serially diluted in peptone buffer up to $10^{-5}$ dilution, and 100 µl of $10^{-5}$ dilution was spread onto brilliant green sulfa (BGS) agar media plates, with a final plate dilution of $10^{-6}$. Samples were incubated at 35 to 37°C for a predetermined amount of time (6, 8, 10, 12, and 24 hrs.). From each plate 1 to 3 colonies were picked using a stereo-microscope (Olympus, SZK12, Center Valley, PA) (except 24 h incubated agar plates), and then resuspended in 10 µl of sterile water. To prepare the HMI microscope slides 3 µl of bacteria suspended in sterile water was placed in the center of a microscope slide, which were allowed to air dry in a biosafety cabinet (BSC, Baker, Sanford, ME) for 15 min. After drying, 0.8 µl of sterile water was added with a cover slide on top, and a drop of 50 cc immersion oil (Nikon, Melville, NY) on top of the cover slide.

**Hyperspectral microscope imaging (HMI) system**

The HMI system (HSI-400, Gooch & Housego, Orlando, FL) was used for acquiring images from the glass slides containing foodborne bacterial samples. The HMI system consists of a Nikon upright microscope (Nikon, Eclipse e80i, Lewisville, TX), acousto-optical tunable filter (AOTF) (Goouch & Housego, HSI-400, Ilminster, England), a high performance cooled electron multiplying charge coupled device (EMCCD), 16-bit camera (iXon, Andor Technology, Belfast, Northern Ireland), and dark-field illumination lighting sources (CytoViva 150 Unit, 24W metal halide, Auburn, AL). The AOTF used for this research has a high-speed, high-throughput, random-access solid state optical filter with an adjustable optical pass-band and exceptionally high rejected light levels. AOTF delivers diffraction-limited image quality with variable bandwidth resolution within 2 nm.
**HMI acquisitions**

HMI-400 software (ChromaDynamics, Lakewood, NJ), was used for acquiring HMIs and converted data from the individual spectral images to the contiguous hypercube format. The hypercube stack images obtained at all 89 bands, and overlays each pixel into a three dimensional cube. Once image acquisition software was executed, all required parameters including exposure, electronically magnified (EM) gain, binning, and bandwidth were selected for high quality images accordingly. In this experiment, the hyperspectral microscope imagery was collected from 450-800 nm at 2 nm bandwidth, 4 nm spectral intervals with a scanning exposure time of 250 ms, and a gain of 9. All images were acquired by spectral sweep mode that enabled us to collect a total of 89 contiguous spectral images by sweeping the entire image in one scan direction, and returning to capture the next spectral band, with an acquisition time of approximately 1.98 s per single spectral image.

**Region of interest (ROI) and preprocessing**

HMIs were analyzed through the environment for visualizing images (ENVI) software (Exelis, McLean, VA). The region of interest (ROI) extraction by threshold selection was implemented to select the pixels within the bacterial cells, removing the background. ROI pixels were then randomized to insure data subsets were a true representation of the image, as much as possible. Data subsets were calculated based on 30% of the pixels in an average cell for each image, in an effort to reduce the large data processing and storage requirement associated with HSI (approximately 500 pixels per cell). Six hour incubation did not produce enough cells per image with any of the five serotypes for an adequate comparison to other incubation times; therefore, 6 hrs. images
were not included for further analyses. Each sample subset size is thirty with the exception of SH (n=18) and ST (n=21) both at 8 hrs. and SK at 12 hrs. (n=28). ROI subsets were averaged, spectra was normalized to the excitation band at 546 nm, which has the maximum intensity for all serotypes, placing the spectra on a scale of 0 to 1. Pearson correlation values were calculated for each serotype to compare the average spectrum of shortened incubation times to that of 24 h.

Data preprocessing was carried out on the subsets through the Unscrambler software version 9.8 (Camo, Oslo, Norway). A first derivative Savitzky-Golay data transformation algorithm (25) with a quadratic polynomial, and 11-point convolution interval was applied to all data subsets to create more spectral band separation between serotypes, followed by a moving average smoothing treatment with a 5-point convolution interval (27).

A two-band relative ratio was calculated for each data subset. The band at 590 nm was selected based on the correlation of the peak to the original spectral peak at 590 nm (figure 2), and the 538 nm band was selected because this represents the second highest peak, which upon visual inspection also depicts variance in each serotype’s spectra.

\[
\text{Relative Ratio} = \frac{SI_{\lambda_1}}{SI_{\lambda_2}}
\]

Where \(SI_{\lambda_1}\) is the scattering intensity at 590 nm, and \(SI_{\lambda_2}\) is the scattering intensity at 538 nm. OriginPro9 software (OriginLab, Northampton, MA) was used to calculate one way ANOVA on the relative ratio values with a p-value of 0.05, and a Fisher’s least significant difference means comparison test (9). The relative ratios were calculated at 8
h, 10 h, 12 h, and 24 h to determine if the two-band relative ratios can be differentiated at each incubation time between serotypes.

*Principal component analysis (PCA)*

PCA was performed on ROI subsets at each incubation period. In general PCA is a factorial analysis used to reduce a large amount of multivariate data into manageable plots based on underlying data trends. Esbensen (5) presents additional information on PCA. For this experiment, PCAs were calculated with full leave-one-out cross-validation and Martene’s uncertainty principle was applied (18). PCA can be calculated by:

\[
X = TP^T + E
\]

where \(X\) is the data matrix, \(T\) represents the score matrix, \(P\) represents the loading matrix, and \(E\) represents the residuals or spectral noise (30). PCA score plots and loading plots are generated through the analysis. Score plots show sample clustering based on the underlying data trends, and loading plots were used to determine spectral bands contributing meaningful influence on the score plots. The Mahalanobis Distance (MD) values were calculated to quantify the distances between the center points of two PCA clusters. De Maesschalck et al. (2) and Hair et al. (10) provide further information on MD. The MD can be calculated by:

\[
MD^2 = (x_i - x_j)^T S^{-1} (x_i - x_j)
\]

where \(x_i\) and \(x_j\) are the averages of groups \(i\) and \(j\), \(T\) represents the transverse of the matrix, and \(S\) is the covariance matrix.

*Experiment repetition and validation*

The experiment was repeated with new cultures of the same five *Salmonella* serotypes grown and plated on BGS. Previously mentioned image collection, ROI
extraction, and spectral transformations were repeated. Partial least squares regression (PLSR) with Martens’s uncertainty principle and full leave-one-out cross validation were also performed on the validation data. PLSR was used as a method for validating the performance of the global data transformation model, by assessing the fit of the data collected from new images. Soft independent modeling of class analogy (SIMCA) was implemented as a supervised pattern recognition method. In general, SIMCA classifies the validation data based on results from the previously generated PCA. Dunn III et al. (3) provides additional information on SIMCA.

**Results**

*BGS plate growth and spectral comparisons*

*S. Infantis* showed no signs of colony growth at 6 hrs. of incubation time during the first repetition, and no or little growth at 6, 8, 10, or 12 hrs. during the second repetition of the experiment. However, the other four serotypes had noticeable colony growth at 6 – 24 hrs. Figure 3.1 shows hyperspectral microscopic composite images from *S. Typhimurium* cells grown at different incubation times, and Figure 3.2(a) shows a typical untreated spectral pattern for *S. Enteritidis* at the four incubation times, while Figure 3.2(b) represents the spectra with the global pretreatment algorithm applied. A sharp excitation peak occurs at 546 nm, with a second more gradual peak beginning at 590 nm in the raw spectra. Pearson correlation values were used to compare the spectral signatures of the shorter incubation times to the standard 24 h. Correlation values ranged from 0.987 to 0.999 for all serotypes and incubation times, representing highly correlated spectra between samples and justifying the need for MVDA. The results of the two-band relative ratios analyzed through ANOVA for incubation times of 8, 10, 12, and 24 hrs.
can be seen in Table 3.1, and shows that there is a significant difference in values for each serotype comparison at, with the exception of one (S. Heidelberg and S. Typhimurium at 8 h). This shows that the relationship of the transformed spectral peaks at 538 and 590 nm are responsible for a portion of the serotype difference explained in the PCA score plots.

*Principal component analysis (PCA)*

Figure 3.3 represents the PCA score plots overlaying the loading vectors for the first and second principle components (PCs). Visual serotype cluster separation is noticeable. Serotype clustering appears to be consistent in size and separation with the exception of ST at 12 h, which has higher inter-cluster variance. Each of the plots lists the amount of variance explained at the first and second PCs. The MD values listed in Table 2 represent a way to quantification of the cluster distances from the PCA scores plot by measuring the distance from center-point to center-point.

*Experimental repetition and validation*

PLSR shows a root mean squared error of calibration (RMSEC) < 0.0025 and $R^2 > 0.9844$, while the root mean square error of validation (RMSEV) is < 0.0030, with $R^2 > 0.9954$. SIMCA results are listed in Table 3.3. The SIMCA results are expressed as a percentage of correctly classified data subsets from the second experimental run based on the results of the first run’s PCA. The suggested optimal numbers of PCs for each incubation time are based on obtaining a desirable percentage of explained variance that was either four or five PCs.
Discussion

*Spectral differences between incubation times*

The lack of micro-colony growth at early incubation times for *S*. *Infantis* could signify that the serotype metabolizes nutrients differently than the other serotypes, possibly indicating a longer lag phase. Pearson correlation values are above 0.9870, describing a strong similarity between the spectral signatures. Figure 3.1 represents a typical image in which one can see *S*. *Typhimurium* cells of various life cycle stages, denoted by the various cell shapes. The strong correlation values indicate that incubation times are not having an effect on the spectral signatures, suggesting that times earlier than 8 h would have very similar spectral patterns that could be used for identification and classification purposes, if enough cells per image are obtained.

The two-band relative ratio difference of average values displayed in Table 3.1 shows that only *S*. *Heidelberg* and *S*. *Typhimurium* at 8 hrs. are not significantly different. The lack of significant difference between these two serotypes at 8 hrs. could be accounted for by the PCA only explaining 60% of the variance at the first PC. The first PC shows strong peaks at the 538 and 590 nm bands, while 8 h would be affected by the second PC more than the other incubation times, which has prominent peaks at 550, 598, 630, 666, and 726 nm.

*Principal component analysis (PCA)*

While the high correlation of the spectral signatures between 8 h and 24 h is advantageous for describing the early detection possibilities, tight grouping makes for difficult classification. Plotting the resulting score matrix gives a visual of dominate “object patterns” within the data matrix, whereas plotting the loading matrix shows the
complimentary “variable patterns” (30). Score plots display the relationships of principle components PCs to the central axis in the PC matrix, and loading plots are responsible for conveying information about the direction of each principle component relative to the original co-ordinate system. While the serotypes can be classified and validated at each incubation time, PCA score plots show different loading vector bands are influencing cluster scattering at the four incubation times. This could be due in part to the majority of the cells imaged at each incubation time being in different life cycle stages, affecting light scattering. The PCA cluster for ST at 12 hrs. contains a higher level of intracluster variance. This is due to the fact that the collected image contained a high amount of extra cellular particles, which were included within the ROI threshold limits. Those extracellular particles could be removed with further object-based image processing algorithms to minimize variation. These particles appear to be background material from the BGS agar and are accounting for the variation in the PCA score plot. The percentage of variance that is explained by each PC aids in determining the optimal number of PCs. It is desirable to have the least amount of PCs necessary, to reduce the data processing and storage requirements. The 12 hrs. plot has a lower percentage of variance described by the first PC due to the variance within ST. In Table 3.2 we can see that MD values range from 2.146 to 27.071, with the higher values equating to a higher level of classification. At 8 hrs. of incubation time the average inter-cluster distance values were the highest of reduced times, suggesting that the intrinsic differences in each serotype have a more noticeable effect on the sample’s spectra, leading to potentially higher classification results, which indicates MD algorithm could be a good candidate for early classification of Salmonella serotypes.
Validation with MVDA

The same ROI subset approach was applied to the validation data set and used to assess the fit of the data pretreatment model. The precision of the model parameters improves with the increasing number of relevant variables and observations \((31)\). Using the full 89 band variable set of data proved to be a close fit for the regression analysis with R\(^2\) values for both calibration and validation > 0.9821.

SIMCA is based on the idea that objects in one class or group show similar but not necessarily identical patterns \((3)\). This form of classification applies pattern recognition to the second set and uses the same global data pretreatment algorithm applied to the PCA, with the objective of predicting the serotype’s group classification. The samples at 12 hrs. of incubation had a lower percentage of correctly classified subsets as seen in Table 3.3, likely due to the extracellular particles in the ST image. Both 8 hrs. and 24 hrs. yielded classification percentages above 98% of the training data set, suggesting repeat experiments showed early identification is possible at high classification efficiencies.

In conclusion, early incubation times appear to have nearly identical spectra to 24 h and do not suggest any effect on the spectral signature of the bacteria, indicating that time is not a factor concerning early detection through HMI on the cellular level. The data from this study suggests that bacterial cells at any stage of their life cycle have a similar spectral signature unique to not only their species, but serotype as well. The data transformation model shows increased band separation and the classification of serotypes by HMI can be implemented as a rapid and early analysis tool for presumptive detection of bacterial pathogens. By applying a classification method such as SIMCA on the
second repetition to validate the results of the first data set collected, we showed that
classification through the MVDA steps is repeatable. Future experiments will focus on
reducing the number of bands to as few as possible while maintaining high classification
accuracy to reduce the data processing and storage requirements necessary for an
applicable implementation of this technology in the food industry.
References


determination and visualization of Pseudomonas loads in chicken fillets. *Talanta.* 1:74-83


Figure 3.1: Hyperspectral microscope images of S. Typhimurium cells at different incubation times: (a) 6 hrs., (b) 8 hrs., (c) 10 hrs., (d) 12 hrs., and (e) 24 hrs.
Figure 3.2: (a) Average spectra for S. Enteritidis normalized to 546 nm, (b) average S. Enteritidis spectra with pretreatment algorithm applied.
Figure 3.3: PCA score plots with loading vectors for four incubation times: (a) 8 hrs., (b) 10 hrs., (c) 12 hrs., and (d) 24 hrs.
Table 3.1: Difference of average serotype values from a one way ANOVA Fisher test, to compare relative average spectral peak ratios (590 nm / 538 nm) at incubation times 8 – 24 hrs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-0.11</td>
<td>n/a</td>
<td>0.07</td>
<td>-0.13</td>
<td>n/a</td>
<td>0.064</td>
<td>(0.00)</td>
<td>n/a</td>
<td>n/a</td>
<td>-0.07</td>
</tr>
<tr>
<td>10</td>
<td>-0.05</td>
<td>-0.10</td>
<td>-0.12</td>
<td>-0.08</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.026</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>12</td>
<td>-0.35</td>
<td>-0.42</td>
<td>-0.23</td>
<td>-0.10</td>
<td>-0.05</td>
<td>0.12</td>
<td>0.254</td>
<td>0.19</td>
<td>0.32</td>
<td>0.131</td>
</tr>
<tr>
<td>24</td>
<td>-0.20</td>
<td>-0.25</td>
<td>-0.62</td>
<td>-0.33</td>
<td>-0.05</td>
<td>0.14</td>
<td>-0.136</td>
<td>0.19</td>
<td>-0.08</td>
<td>-0.27</td>
</tr>
</tbody>
</table>

E = S. Enteritidis, H = S. Heidelberg, I = S. Infantis, K = S. Kentucky, T = S. Typhimurium

*p=0.05, ( ) represents a pair of average values that are not significantly different*
Table 3.2: Mahalanobis distance (MD) values between serotype clusters calculated from PCA score plots at incubation times of 8 – 24 hrs.

<table>
<thead>
<tr>
<th>Time (h)</th>
<th>Comparison between serotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>14.24</td>
</tr>
<tr>
<td>10</td>
<td>21.46</td>
</tr>
<tr>
<td>12</td>
<td>23.27</td>
</tr>
<tr>
<td>24</td>
<td>4.60</td>
</tr>
</tbody>
</table>

*E* = *S.* Enteritidis, *H* = *S.* Heidelberg, *I* = *S.* Infantis, *K* = *S.* Kentucky, *T* = *S.* Typhimurium
Table 3.3: Soft-independent modeling of class analogy (SIMCA) for all serotypes within each incubation period.

<table>
<thead>
<tr>
<th>Incubation time (h)</th>
<th>Optimal # of PCs</th>
<th>Total # of serotype subsets</th>
<th>% of subsets correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>119</td>
<td>98.32</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>120</td>
<td>96.67</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>120</td>
<td>88.33</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>150</td>
<td>98.67</td>
</tr>
</tbody>
</table>

*(Optimal # of PCs represents the # of principal components required to explain >96% of variance)*
CHAPTER 4

Visible Near-Infrared Hyperspectral Microscopy and Informative Band Selection for

Classification of *Salmonella Enterica* Serotypes\(^1\)

\(^{1}\)Eady, M.B., and Park, B. To be submitted to *The Journal of Microscopy*. 
Abstract

A novel method for classification of five Salmonella enterica serotypes (S. Enteritidis, S. Heidelberg, S. Infantis, S. Kentucky, and S. Typhimurium) through the analysis of images obtained from a hyperspectral microscope is proposed for a rapid detection method of pathogenic foodborne bacteria. Data was obtained from a hypercube that captured darkfield images of bacterial cells. A metal halide lighting source emits light through a glass slide containing a air dried bacterial suspension, and an acousto-optic tunable filter (AOTF) is applied with an electron-multiplying charge coupled device (EMCCD) to capture light scattering. Hyperspectral microscope images (HMI) were collected with 89 spectral bands, every 4 nm between 450 and 800 nm. Principal component analysis (PCA) was carried out on the images with informative variables selected based on PCA loading vectors and inter-band intensity variance. Reduced variable sets of 20, 12, 7, and 3 bands were chosen between the 586 and 662 nm range to compare. PCA score plots show cluster separation for the full 89 pt. spectrum, and at the reduced variable sets, however classification between S. Enteritidis and S. Typhimurium is less concise. A second set of HMIs were collected for the reduced number of bands. Support vector machine (SVM) classification was implemented as a tool for classifying the five serotypes into groups. The first repetition of HMIs saw SVM classification abilities drop to 84.2%, where as the HMIs collected at only the reduced variables had increased classification accuracies, with as few as three spectral bands selected (590, 594, and 598 nm), representing a 96.6% reduction in the data processing and storage requirement of HMI.
Introduction

Traditional foodborne detection methodologies such as the use of nutrient enriched growth media or polymerase chain reaction (PCR) have been widely implemented for bacterial detection and identification for some time, however these protocols come with disadvantages that can be limiting in assessing potential causative agents in foodborne disease outbreaks. Utilizing nutrient enriched growth media is time consuming with confirmation results requiring several days or longer. PCR analysis can be completed in considerably less time, although requires extensive training in addition to the high recurring cost of cell lysing reagents.

Optical detection methodologies have been approached as a rapid detection method for bacteria, focusing on the classification of microbes based on a spectral signature that is unique to that organism (Narsaiah et al., 2012). Spectroscopic techniques such as surface-enhanced Raman spectroscopy (SERS), Fourier-transform infrared (FT-IR), laser-induced breakdown spectroscopy (LIBS), and surface plasmon resonance (SPR) have previously been used as a means of nondestructive foodborne pathogen detection (Dudka et al., 2009; Rehse et al., 2010; Sundaram et al., 2013; Udelhoven et al., 2000). The proposed hyperspectral microscope image (HMI) analysis method has the potential for optical identification with the added benefit of sensitivity, requiring only a few cells captured within an image for classification.

Hyperspectral imaging (HSI) is a technique that generates both spatial and spectral information, representing a three dimensional data set or hypercube with two of the dimensions correlating to spatial data (x and y coordinates) and a third dimension yielding spectral data ($\lambda$) (Sun 2010). This imaging technique creates a data hypercube, which consists of capturing an image for each band selected. The hypercube then overlays each of these images, creating a
spectrum for each pixel within the image, while keeping the x and y coordinates constant. HSI has been implemented for medicine (Kellicut et al., 2005), pharmaceuticals (Roggo et al., 2005), and astronomy (Wood et al., 2002) applications. In recent years the exploration for applications of stand-off HSI in agricultural commodities has risen. This technology has been used as a nondestructive automated defect detection system in agricultural commodities such as broiler chickens, apples, mushrooms, and salmon (Gowen et al., 2008; Lawrence et al., 2003; Liu et al., 2007; Wu et al., 2013) to name a few. Park et al., (2012) showed that Salmonella and E. coli differ in HMI obtained spectra. This experiment proposes a novel method for detecting and classifying bacteria on a microscopic level through analysis of the HSI.

Salmonellosis is the disease accompanying a Salmonella infection, typically resulting in gastrointestinal symptoms. It is estimated that 93.8 million cases occur worldwide, resulting in approximately 155,000 deaths each year (Majowicz et al., 2010).

In addition to the risk of adverse human health issues, the presence of Salmonella can produce significant economic problems for the food industry, including food producers, processors, and distributors. This gram-negative pathogenic bacterium has caused foodborne disease outbreaks in poultry, eggs, milk, nuts, fruits, and vegetables (Cox et al., 2013). Costly product recalls can result in a negative consumer perception in addition to the profit loss. The growing acceptance of food safety programs such as the Hazard Analysis and Critical Control Points (HACCP) program put a focus on analyzing each processing step for potential chemical, physical, or biological hazards. The microbial analysis of a HACCP plan requires monitoring and validation of results. A rapid detection method for foodborne pathogens such as Salmonella
*enterica* serotypes can assist in automating this process, which typically focuses on microbial detection methods using enriched growth media or PCR.

A large amount of data is generated from HSI hypercubes due to each pixel yielding a spectrum generated from the overlapping images. It is important to reduce the amount of data collected to as little as possible while still maintaining a high level of classification accuracy for the purpose of applying this technology to a quality control facility, where a multitude of samples must be collected daily. Previously, various spectral reduction approaches have been undertaken to reduce the number of variables necessary with success in maintaining classification abilities. Windham *et al.*, 2013 used a principal component analysis (PCA) loading vector to find bands that explained the highest amount of variance. Partial least squares regression (PLSR) techniques have been applied to each band selecting the bands with the smallest root-mean squared error of calibration (RMSEC) or validation (RMSEV) (Wu *et al.*, 2013). Other techniques implemented have been an ensemble Monte Carlo selection method, discriminate analysis (DA), analysis of spectral difference (ASD), successive projections algorithm (SPA), and principle component regression (PCR) (Esquerre *et al.*, 2011; Gowen *et al.*, 2008; Liu *et al.*, 2007; Ye *et al.*, 2008). While reducing the number of bands necessary for identification purposes, caution must be utilized to not over reduce the spectra losing variables that explain the naturally occurring variance that drive spectral based classification abilities. In this study, we evaluated the necessity of each band based on two different selection techniques; loading vectors of principal component analysis, and the amount of inter-band variance created by normalizing the spectra to each band, then averaging and assessing the average variance at each band.
The objective of this experiment is to develop a hyperspectral microscopic method to definitively classify five serotypes of *Salmonella*, while selecting the fewest amount of informative variables or spectral bands necessary for retaining a high rate of classification compared to the full 89 band near-infrared spectrum between 450 and 800 nm.

**Materials and methods**

**Sample preparation**

Five *Salmonella* serotypes: *S*. Enteritidis (SE), *S*. Heidelberg (SH), *S*. Infantis (SI), *S*. Kentucky (SK), and *S*. Typhimurium (ST) were obtained from poultry carcasses rinses in the Poultry Processing and Swine Physiology Research Unit at the Russell Research Center in Athens, GA. Cultures were stored at -80°C. Fresh cultures were prepared from frozen cultures and incubated for 24 hrs. at 35-37°C. Several colonies were picked from the slants and inoculated into tubes containing triptc-soy broth (TSB), and incubated overnight at 35-37°C. Serial dilutions were prepared to $10^{-5}$ for each of the five serotypes, then 100 µL of $10^{-5}$ was pipetted onto brilliant green sulfide (BGS) plates with a final plate count dilution of $10^{-6}$. BGS plates were incubated for 24 hrs. at 35-37 °C.

**Hyperspectral microscope and image collection**

The HMI system can be viewed in Figure 1. A Nikon Eclipse 80i upright microscope (Nikon, Lewisville, TX) was equipped with an acousto-optic tunable filter (AOTF) (HIS-400, Gooch & Housego, Orlando, FL) and a high performance cooled electron multiplying charge coupled device (EMCCD), 16-bit camera (iXon) (Andor Technology, Belfast, Northern Ireland), and a darkfield illuminating metal halide lighting source (Cytoviva 150 unit, 24 W, Auburn, AL) that travels through a fiber optic cable illuminating from underneath the sample stage. The
AOTF filter used in this experiment has a high-speed, high-throughput, random access solid state optical filter with an adjustable optical pass-band. The variable bandwidth of the diffraction-limited image resolution was 2 nm.

Two colonies were picked from the BGS plates and suspended into 10 µL of autoclaved deionized water, followed by a brief vortex. Three µL of the bacterial suspensions were placed on a glass slide and allowed to air dry in a bio-safety cabinet (BSC, Baker, Sanford, ME) for 15 min. Sterile water (0.8 µL) was placed on top of the dried suspension and a glass cover slip was applied. Pressure was applied to the glass cover slip to immobilize cells and prevent movement during the multispectral sweep, collecting 89 contiguous images.

Initially, 89 contiguous bands were collected between 450 and 800 nm, every 4 nm for the first repetition of image collection. Once the informative variables were identified the second repetition focused on collecting those selected spectral bands only. The first repetition had an acquisition time of 1.98 s between each acquired image within the hypercube.

**Hypercube and pixel extraction**

The hypercube represents the overlapping of multiple images collected through the hyperspectral imaging process at various bands or spectral bands, and an experimental flow chart can be seen in Figure 4.1. These images are stacked upon each other creating the hypercube by combining both spatial and spectral data. The x and y planes are consistent as they represent the spatial coordinates of each pixel in the image. Lawrence et al., (2003) stated that this information can be used to describe physical or geometric observations within an image such as shape color and size. Figure 4.2 represents a mosaic of the same microscope image captured at
each of the five selected variable sets. The middle row shows a typical *Salmonella* cell and how intensity values from the 590 nm band are used to recreate a single band of the hypercube, in which physical features of a cell are defined.

Collected images were analyzed through the Environment for Visualizing Images (ENVI) software program (Exelis, McLean, VA). The region of interest (ROI) is a selection of pixels representing an object in the image. Pixels representing only the cells were extracted from the HSI images in order to remove the background. This was performed by setting minimum and maximum threshold values of each pixel to extract only cell data. After ROIs were extracted pixels were randomized to account for optical effects such as limb darkening of the cell boundaries. A thirty percent subset was calculated for each image based on the average number of pixels per cell and per image. Because the scope of this project focuses on an efficient data reduction method we also looked at extracted a minimal amount of data from each image. A typical rod shaped bacterial cell may contain approximately 500 pixels, however not all of these pixels are necessary for classification. The average number of pixels in a cell was calculated per image by dividing the number of pixels extracted from the threshold ROIs divided by the number of cells per image. Thirty percent of the average number of pixels were then randomized and averaged to reduce the amount of data required for classification. Based on the number of pixels obtainable per image, subsets of $n=120$ for each image were calculated for the first repetition of serotype images in this experiment. The randomization assures that subsets are not extracted multiple times from the same cell as a whole, and yields a true representation of the data set.
Principal component analysis (PCA) and informative band selection

Data was then imported into the Unscrambler software program version 10.3 (Oslo, Norway) for multivariant data analysis (MVDA). MVDA is typically necessary when analyzing HMI data due to the high correlation of the spectra obtained from samples (Hair et al., 2010). A first derivative Savitzky-Golay preprocessing step (Savitzky et al., 1964) with a quadratic polynomial for removing some of the baseline effect, and an 11-point convolution interval for spectral noise reduction of an appropriate segment size for the 89 contiguous bands collected or independent variables. A 5-point moving average smoothing treatment was then applied for additional smoothing of the sharp spectral peaks (Soderberg et al., 2000). This helped reduce noise and aided in selecting only independent bands.

PCAs were performed on the first repetition of HMI for each of the variable selection set models A-E. However, due to a reduction in spectral noise by eliminating the uninformative variables, data pretreatment was not needed for variable sets B-E. A PCA can be used to simplify an analysis, in which multiple variables are present as with spectroscopy data treating each band as a separate variable. This offers a scatter plot for simple visualization of the five serotype clusters. A PCA decomposes the spectral X matrix in which:

\[ X = TP^T + E \]

where \( X \) is the data matrix, \( T \) represents the score matrix, \( P \) represents the loading matrix, and \( E \) represents the residuals or spectral noise (Wold et al., 1987). PCAs in this experiment were calculated for each of the 600 ROI data subsets (n=120 per serotype x 5 serotypes) with full leave-one-out cross-validation, and Martene’s uncertainty principle was used for creating submodels and applying a cross-validation with a jack-knifing method of resampling for leaving
out each observation (band) in the data set and calculating the 89 submodel averages (Martenes et al., 1999).

The selection of informative bands for an abbreviated data processing and analysis protocol was defined by treating each of the 89 contiguous bands as independent variables. The PCA loading vectors represent the principal component (PC) direction relative to the original information and are considered to be a transformation matrix between the original variable space and the PCA matrix (Esbensen, 2010). Each PC within a PCA analysis has a loading vector associated with it. When the loading vectors for essential PCs (explaining a significant amount of variance) are plotted together, spectral bands with peak intensities can be identified.

In addition to analyzing the PCA loading vectors, the amount of variance associated with each band was assessed through normalization of the spectra at each band. The 89 spectral bands were normalized to each collected band in the spectrum, and the resulting average serotype variance plotted in Figure 4.3(a). Spectral bands were selected for the reduced variable sets by comparing the loading vectors Figure 4.3(b) to the spectral band variance observed through normalization. Figure 4.3(c) shows the range of variables selected as informative.

**Collection of validation hyperspectral microscope images (HMIs)**

Informative bands were identified with the first set of images, and the purpose of recollecting HMIs with a second repetition is to 1) validate classification results and 2) to determine if a reduction in the hypercube data will in fact be able to classify the five *Salmonella* serotypes at a high efficiency. A second set of HMIs was collected for the same five serotypes of *Salmonella*. The new bacterial cultures were grown following the same protocol as outlined
for the first repetition. Glass slides were prepared in the same manner. The collection of HMI differed. The second set of HMIs were collected at only the specified informative spectral bands. Variable set A was collected first, followed by changing the band selection range within the chemometric software to represent 586 to 663 nm (set B) obtaining the same microscopic image followed by sets C, D, and E collected at respective bands. The bandwidth for all repetition 2 images was 2.5 nm with spectral intervals of 4 nm, and a scanning exposure time of 250 ms. Theoretically classification results should be similar to that of the first HMI set. The purpose of this was to not just select bands that appear informative, but to carry out the hypercube data collection with the reduced bands and determine if a practical collection method would yield similar spectra and classification results to the first repetition that analyzes the bands based on the collection of the full 89 spectral bands.

Support vector machine (SVM) classification

Classification of serotypes for both repetitions and each of the five variable selection sets were performed with SVM classification algorithms. SVM is a form of supervised classification based on object recognition of nonlinear data (Mercier 2003). A kernel-based recognition algorithm creates a hyperplane or high-dimensional space that is then used to classify the serotypes through the input space with the equation:

\[ h = w^T x + b \]

where \( h \) is the hyperplane value, and \( x \) is the HMI input vector from pixel intensity, \( w \) is the adaptable weight vector, \( b \) represents the bias, and \( T \) is the transverse operator (Sun, 2010). Figure 4.4 shows a representation of how SVM classification uses a kernel based approach for
distinguishing objects in a high-dimensional space. While one class is visually noticeable in the original variable matrix sometimes a nonlinear classification algorithm is necessary. SVM creates a high-dimensional linear feature space or hyperplane to classify objects based on a linear separation (Sun, 2010). The original variable space is a low-dimensional space where a kernel is applied to create a hyperplane. A confusion matrix describes the classification results obtained from the SVM algorithms, by grouping serotype subsets into one of five classes.

**Results**

**Informative band selection**

Informative band selection was determined by comparing both PCA loading vectors with the average variance attributed to normalization at each band. The results can be seen in Figures 4.3(a) and 4.3(b). Figure 4.3(c) shows the original spectra with the area for informative band selection highlighted. All four of the abbreviated variable selection sets are within the range of 586 and 662 nm. Table 4.1 lists the five variable selection sets, the associated spectral ranges, and the percent reduction from the original 89 band spectra (set A).

**Principal component analysis (PCA)**

The PCA score plots can be viewed in appendix a for the first repetition of images collected. The score plots give us a visual representation of the underlying trends in this factorial analysis. Cluster separation is noticeable, however with variable reduction typically comes less cluster separation. Appendix a depicts models A and D with clear cluster separation, while models B, C, and E display cluster separation although *S. Typhimurium* show some of the overlapping in cluster boundaries.
Support vector machine (SVM) classification

Figure 4.2 represents a mosaic of the HMI collected during repetition 2 of this experiment, where the same image was collected only at the specified spectral bands between 586 and 662 nm for each variable set. In the top row of the image, HMIs collected at 590 nm are visible, which appear very similar to each other. In the middle row, reconstruction of the same S. Enteritidis cell is repeated for each variable selection set. Comparing the middle row to the bottom row with corresponding spectra of the same pixel shows that while there is a slight change in spectral intensity, the spectral pattern is consistent across all five variable sets. For all of the pixels selected as a whole, Figure 4.5 describes the relationship of the 590-598 nm spectral range with each variable selection set. Variable set A is similar in pattern to variable set E. The change in spectral intensities can be attributed to several unknown factors at this time.

Table 4.2 shows that for the second repetition of the experiment SVM classification results were perfect (120 correctly identified, for n=120 per serotype) for much of the serotypes in variable sets A and B. Variable sets C, D, and E saw reductions in the classification abilities, dropping to 84.2% correctly identified. The confusion matrix also shows a 100% classification accuracy for all serotypes of variable sets of the second repetition.

Discussion

Spectral bands within the 450-800 nm range for darkfield images of bacterial cells showed that the bands between 586 nm and 662 nm appeared to be the most informative. Bands explaining the most variance in the PCA loading vector also had the lowest average band variance. The mosaic in Figure 4 shows that the 590 nm band, light scattering from a single cell, and pixel spectrum may change slightly in intensity over time, however the spectral pattern
remains similar. Figure 4.5 verifies that the average spectra of all extracted pixels for the 590-598 nm band range are similar with variable sets A and E being very similar.

The kernel-based classification algorithm succeeded with high classification accuracy because of the robust nature of SVM applied to nonlinear data. SVM results from repetition two confirm the PCA score plot results of the same images. Through both analyses there is clear classification for three serotypes, however in both analyses S. Enteritidis and S. Typhimurium are not separated as efficiently. By collecting only the 590, 594, and 598 nm spectral bands this shows high classification accuracy amongst all five Salmonella serotypes by reducing noise associated with uninformative variables.

Park et al., (2012) showed that spectral signatures from E. coli and Salmonella cells differed. This experiment suggests that it is possible to classify serotypes of the same bacterial species with as few as three spectral bands (590, 594, and 598 nm) through a kernel-based SVM classification analysis. The Hughes phenomenon is a situation that is common with hyperspectral or multispectral data sets. The Hughes phenomena states that with a fixed number of samples, the accuracy of the classification algorithm will decrease (Hughes et al., 1964). Alonso et al., (2011) showed that the SVM classification algorithm, being a non-parametric classifier was robust enough to negate the Hughes phenomena. From this, we can assume that the increase in correct SVM classification percentages during the second repetition where only the selected bands were collected is not influenced by this statistical method; however, possibly influenced by the reduction in spectral noise associated with the Salmonella serotypes.

In conclusion, this project shows the sensitive nature of hyperspectral imaging and that it can be applied to differentiate images of bacterial serotypes within the same species based on a
microscope image analysis. While selecting as few as three spectral bands (590, 594, and 598 nm) for classification will greatly reduce the amount of data stored and processed, a kernel-based SVM algorithm is useful for classification. This novel methodology could be implemented as a more cost efficient alternative to bacterial identification through PCR, while considerably less time consuming than traditional plating techniques. Further in-depth analysis on the hyperspectral signature of bacteria would be required prior to a creation of a spectral library followed by repeated sampling trials. Thus, an optical method with HMI will be a good candidate to detect and identify bacterial serotypes as well as species rapidly and easily compared to conventional plating and immune assay methods.
References


Figure 4.1: Sample collection and analysis flow chart

Salmonella Serotype Preparation on BGS Agar → Glass Slides Prepared for each Serotype → HMI Collected for 89 Bands (450-800nm) → Informative Bands Selected

HMI Collected for Selected Bands Only (#89, 20, 12, 7, and 3 bands) → New Salmonella Serotypes and Glass Slides Prepared

PLSR → SVM Classification

PCA
Figure 4.2: A mosaic of the 590 nm band with each variable selection set (A-E) displaying the obtained HMI with a highlighted cell (top), with a respective 3D surface plot reproducing the highlighted pixels (middle), and accompanying spectra of the collected bands for the highlighted cell (bottom).
Figure 4.3(a): Spectral band variance, normalized to each band, (b) loading vectors for PCA score plots (c) average spectra for 5 *Salmonella* serotypes, vertical lines represent the range of bands selected for variable sets B-E (586 – 662 nm).
Figure 4.4: Example of how support vector machine (SVM) classification creates a hyperplane, for classification of nonlinear data.
Figure 4.5: Comparison of the 590-598 nm spectral range collected at each of the five variable selection sets, for *S. Enteritidis.*
Table 4.1: Description of variable sets selected for analyses.

<table>
<thead>
<tr>
<th>Variable Set</th>
<th>Spectral band range (nm)</th>
<th>Number of Bands</th>
<th>Band Reduction Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>450-800</td>
<td>89</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>586-662</td>
<td>20</td>
<td>77.5</td>
</tr>
<tr>
<td>C</td>
<td>586-630</td>
<td>12</td>
<td>86.5</td>
</tr>
<tr>
<td>D</td>
<td>590-614</td>
<td>7</td>
<td>92.1</td>
</tr>
<tr>
<td>E</td>
<td>590-598</td>
<td>3</td>
<td>96.6</td>
</tr>
</tbody>
</table>
Table 4.2: Confusion matrix results obtained from support vector machine classification (SVMC).

<table>
<thead>
<tr>
<th>Serotype</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. Enteritidis</td>
<td>100(100)</td>
<td>95.0 (100)</td>
<td>93.3 (100)</td>
<td>85.8 (100)</td>
<td>85.8 (100)</td>
</tr>
<tr>
<td>S. Heidelberg</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>100 (100)</td>
</tr>
<tr>
<td>S. Infantis</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>98.3 (100)</td>
<td>98.3 (100)</td>
</tr>
<tr>
<td>S. Kentucky</td>
<td>100 (100)</td>
<td>100 (100)</td>
<td>99.2 (100)</td>
<td>89.2 (100)</td>
<td>89.2 (100)</td>
</tr>
<tr>
<td>S. Typhimurium</td>
<td>97.5 (100)</td>
<td>87.5 (100)</td>
<td>84.4 (100)</td>
<td>84.2 (100)</td>
<td>84.2 (100)</td>
</tr>
</tbody>
</table>

Values represent percentage of correctly classified samples.

( ) represents confusion matrix results for correctly classifying percentage of Salmonella serotype subsets, results of second repetition with variable set E (n=60)
CHAPTER 5

CONCLUSIONS

The primary objectives of this study were to first determine if five serotypes of the same *Salmonella enterica* subspecies could be classified through hyperspectral microscope imaging (HMI). The second objective was to analyze spectra obtained from early incubation times (6, 8, 10, 12, hrs.), and to compare the similarities in spectra and classification to the spectra obtained from 24 hrs. of incubation time at 35-37°C. The other objective of this study was to determine if the large data processing and analysis requirements associated with hyperspectral imaging can be reduced to only informative spectral bands, while still maintaining similar classification efficiencies to the full 89 spectra originally obtained from 450 – 800 nm.

This study shows that early classification of bacterial cells is possible through HMI. A comparison between the spectra of each serotype from incubation times of 6, 8, 10, and 12 hrs. compared to 24 hrs. yielded Pearson correlation values > 0.980, indicating strong similarities in samples grown at early incubation times. Application of the principal component analysis (PCA) showed visual cluster separation through a PCA scores plot, used as a factorial multivariate data analysis (MVDA) tool, with inter-cluster distances represented by Mahalanobis distance values to quantify separation between serotypes. Partial least squares regression (PLSR) showed that repetition of the experiment yielded $R^2$ calibration and validation values > 0.9821, describing spectral data that is consistent through repetition. Soft-independent classification analogy (SIMCA) was used in this experiment to classify the five serotypes into clusters, based on the
results of the PCAs obtained through the first repetition of sample culturing and image collection. This method verified that serotype classification accuracy for both 8 and 24 hrs. is greater than 98%.

Due to the large data processing and storage requirement of hyperspectral imaging (HSI) it is essential that analyses be completed with the only the informative spectral bands only. The support vector machine (SVM) applied a kernel-based classification technique to serotype cluster separation by introduction of a hyperplane for classification purposes. SVM classification of the full spectra provided cluster separation > 84.2%, however collecting data at only three spectral bands 590, 594, and 598 nm yielded SVM classification results of 100% for all five serotypes. The range between 590-598nm appears to represent key intrinsic differences in cell physiology between the five Salmonella serotypes.

The end goal for this area of study is to develop an early and rapid analytical system that counters drawbacks to traditional microbial testing procedures such as the high recurring cost and training requirements of PCR, or the extended time requirements of nutrient enriched growth media. While this study shows that detection and classification of pure culture Salmonella isolates is plausible at reduced incubation times, further experimentation on conditions that may affect cellular spectra need to be examined. It is unknown if environmental factors such as temperature or ionic variances, as well as morphological factors such as cellular evolutions or mutations influence spectral signatures. Correlating spectral peaks to physical characteristics of the cell is also needed, as is further understanding of cellular geometric orientation within an image, and the effects of limb-darkening on the cell’s overall spectral profile. This project was able to show that five Salmonella serotypes are capable of being classified with as little as 8 hrs.
of incubation time using a method requiring approximately 15 minutes to complete. It is also possible to carry out the analysis in as few as three spectral bands ranging between 590 and 598 nm. In conclusion, HMI for early and rapid bacterial identification is a promising novel microbial identification methodology. Future work is needed before practical application such as: exploring the effects of various food matrix compositions and environmental factors on the bacteria’s spectra.
Appendix A: Principal component analyses score plots from reduced variable selection sets.

(A.a): PCA scores plot for model A (89 bands).
(A.b): PCA scores plot for model B (20 bands).

(A.c): PCA scores plot for model C (12 bands).
(A.d): PCA scores plot for model D (7 bands).

(A.e): PCA scores plot for model E (3 bands).