2003

Fourier-transform infrared spectroscopic imaging of prostate histopathology

Daniel Celestino Fernandez
University of South Florida

Scholar Commons Citation
http://scholarcommons.usf.edu/etd/1366

Follow this and additional works at: http://scholarcommons.usf.edu/etd
Part of the American Studies Commons
Fourier-Transform Infrared Spectroscopic Imaging of Prostate Histopathology

by

Daniel Celestino Fernandez

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Pathology and Laboratory Medicine
College of Medicine
University of South Florida

Co-Major Professor: Santo V. Nicosia, M.D.
Co-Major Professor: Ira W. Levin, Ph.D.
Wenlong Bai, Ph.D.
Luis H. Garcia-Rubio, Ph.D.
Maria Kallergi, Ph.D.
Patricia A. Kruk, Ph.D.

Date of Approval:
May 20, 2003

Keywords: FT-IR, adenocarcinoma, vibrational, spectroscopy, classification

© Copyright 2004, Daniel Celestino Fernandez
Dedication

To my parents, for their love and support,

and

my wife, for always believing in me.
Acknowledgements

- Howard Hughes Medical Institute – National Institutes of Health Research Scholars Program
- National Institutes of Health Graduate Partnership Program
- University of South Florida – College of Medicine – Department of Pathology and Laboratory Medicine
- National Institute of Diabetes, Digestive and Kidney Diseases
- Ira W. Levin, Ph.D.
- Santo V. Nicosia, M.D.
- Stephen M. Hewitt, M.D., Ph.D.
- Rohit Bhargava, Ph.D.
- Michael D. Schaeberle, Ph.D.
- Scott W. Huffman, Ph.D.
- Patricia McCarthy, Ph.D.
- Jamie Winderbaum Fernandez, M.D.
Table of Contents

List of Tables ........................................................................................................................................ iv

List of Figures .........................................................................................................................................v

Abstract .................................................................................................................................................. vii

Chapter One - Introduction ..................................................................................................................1
  1.1 Electromagnetic Spectrum ............................................................................................................. 1
    1.1.1 Interactions of Electromagnetic Radiation with Matter ......................................................... 4
  1.2 Basis of Infrared Absorption ........................................................................................................... 5
    1.2.1 Requirements for IR Absorption .............................................................................................. 6
    1.2.2 Number of Vibrational Modes ................................................................................................ 8
    1.2.3 Group Frequencies ..................................................................................................................... 9
  1.3 IR Spectral Feature of Tissues .........................................................................................................10
    1.3.1 Proteins .....................................................................................................................................10
    1.3.2 Carbohydrates ............................................................................................................................15
    1.3.3 Lipids .........................................................................................................................................15
    1.3.4 Nucleic Acids .............................................................................................................................17
  1.4 FTIR Spectroscopy Background ......................................................................................................17
    1.4.1 FTIR Spectrometers ..................................................................................................................18
    1.4.2 Infrared Microscopy ..................................................................................................................20
    1.4.3 Mapping with Single-Point Detectors .........................................................................................22
    1.4.4 Raster-scan Imaging Using Multichannel Detectors .................................................................24
    1.4.5 Global FTIR Spectroscopic Imaging ..........................................................................................25
  1.5 Spectroscopic Imaging: Data Structure and Applications ..............................................................27
    1.5.1 Image Classification Methods ................................................................................................30
  1.6 Prostate Background .......................................................................................................................31
    1.6.1 Anatomy and Histology .............................................................................................................31
    1.6.2 Prostate Pathology ....................................................................................................................33

Chapter Two - Methods ..........................................................................................................................44
  2.1 Tissue microarrays ..........................................................................................................................44
    2.1.1 Construction of Prostate Tissue Microarrays ............................................................................44
    2.1.2 Array P-16 Design ....................................................................................................................45
    2.1.3 Array P-40 Design ....................................................................................................................45
    2.1.4 Array P-80 Design ....................................................................................................................46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>Tissue Array Section preparation</td>
<td>47</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Optical Substrates for Tissue Array Sections</td>
<td>47</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Deparaffinization</td>
<td>48</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Optical imaging of H&amp;E sections</td>
<td>48</td>
</tr>
<tr>
<td>2.3</td>
<td>Spectroscopic Imaging Instrumentation</td>
<td>49</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Tissue Array FT-IR Data Collection Parameters</td>
<td>51</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Modifications and Environmental Considerations</td>
<td>52</td>
</tr>
<tr>
<td>2.4</td>
<td>Data Handling and Computational Considerations</td>
<td>53</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Data Pre-Processing</td>
<td>53</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Spectral Baseline Correction</td>
<td>54</td>
</tr>
<tr>
<td>3.1</td>
<td>Visualization of Spectral Images and Verification of Histologic Features</td>
<td>56</td>
</tr>
<tr>
<td>3.2</td>
<td>Creation of Ground Truth Data Regions of Interest</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Spectral analysis of histologic features and metric selection</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>Construction of a Supervised Classification Model for Prostate Histology</td>
<td>64</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Spectral Data Reduction</td>
<td>64</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Image Classification</td>
<td>66</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Array P-16, 20-metric, GML Self-classification results</td>
<td>68</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Leave-one-out metric evaluation</td>
<td>70</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Array P-16, 18-metric GML Classification Results</td>
<td>72</td>
</tr>
<tr>
<td>3.5</td>
<td>Validation of Prostate Histology Classification Model</td>
<td>75</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Cross-Array Validation</td>
<td>76</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusions and Further Directions</td>
<td>79</td>
</tr>
<tr>
<td>4.1</td>
<td>Classification strategy</td>
<td>81</td>
</tr>
<tr>
<td>4.2</td>
<td>Array P-80 H&amp;E Stained Section Pathology Analysis</td>
<td>82</td>
</tr>
<tr>
<td>4.3</td>
<td>Array P-80 Histology Classification Results</td>
<td>83</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Spatial Filtering of Histology Classification Results</td>
<td>84</td>
</tr>
<tr>
<td>4.4</td>
<td>Construction of a Supervised Classification Model for Prostate Pathology</td>
<td>88</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Creation of pathology ground truth ROIs</td>
<td>88</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Pathology Spectral Data Reduction</td>
<td>89</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Histogram analysis of Spectral Metric Data</td>
<td>91</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Mean-centering of epithelial metric data</td>
<td>92</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Metric Statistical Analysis</td>
<td>92</td>
</tr>
<tr>
<td>4.4.6</td>
<td>GML Pathology Classification of Array P-80</td>
<td>94</td>
</tr>
<tr>
<td>4.5</td>
<td>Individual Patient Evaluation of P-80 Pathology Classification</td>
<td>96</td>
</tr>
<tr>
<td>4.6</td>
<td>Cross-Array Validation</td>
<td>97</td>
</tr>
</tbody>
</table>
4.7 Conclusions and Further Directions ................................................................. 98
References .................................................................................................................. 101
About the Author ...................................................................................................... End Page
List of Tables

Table 1.1 - Spectroscopic techniques utilizing different regions of the electromagnetic spectrum ................................................................. 5
Table 1.2 - Staging of primary tumor (T) ................................................................................................................................. 42
Table 1.3 - Staging of regional lymph node involvement (N) ................................................................................................. 43
Table 2.1 - Spectral frequencies used for spectroscopic baseline correction ........................................................................... 55
Table 3.1 - Histologic class population data ......................................................................................................................... 61
Table 3.2 - Histology Spectral Metric Definitions .................................................................................................................. 66
Table 3.3 - Error Matrix of supervised GML Classification results using 20 spectroscopic metrics ......................................................... 69
Table 3.4 - Confusion matrix of supervised GML Classification attempt using 18 spectroscopic metrics ............................................................. 72
Table 3.5 - Revised 6-class histology ground truth ROIs for Array P-16 and Array P-40 ...................................................................................... 77
Table 3.6 - Error Matrix for 6-Class, GML Classification Results ...................................................................................................... 78
Table 4.1 - Pathology spectral metric parameters ........................................................................................................... 90
Table 4.2 - Results of t-test on mean adenocarcinoma metric values from population of 25 patients on array P-80 for 54 candidate pathology metrics ........................................................................................................ 93
Table 4.3 - Error matrix for 20-metric pathology GML classification of epithelial tissue on array P-80 ......................................................... 96
List of Figures

Figure 1.1 - The electromagnetic spectrum ................................................................. 2
Figure 1.2 - The infrared region of the electromagnetic spectrum ............................. 4
Figure 1.3 - Vibrational modes and IR activity of water vapor (A) and carbon dioxide (B) molecules ................................................................................... 8
Figure 1.4 - Vibrational modes of methylene group .................................................. 9
Figure 1.5 - Structure of a typical amino acid ............................................................ 11
Figure 1.6 - Basic polypeptide structure .................................................................... 12
Figure 1.7 - Common Protein Secondary Structures: α-helix and β-sheet ................. 13
Figure 1.8 - Michelson Interferometer ....................................................................... 19
Figure 1.9 - Three Instrumental Approaches for collection of spatially resolved FTIR spectroscopic data ................................................................. 22
Figure 1.10 - Schematic representation of the image cube ......................................... 28
Figure 1.11 - Zonal Anatomy of the Prostate .............................................................. 32
Figure 2.1 - Array P-80 Layout .................................................................................... 47
Figure 2.2 - Spectrum Spotlight 300 Microscope Optical Configuration ................. 51
Figure 3.1 - A) Baseline-corrected N-H stretching (3290cm-1) absorbance intensity image of four tissue array spots from a single patient on Array P-16 B) Optical images of corresponding H&E stained section ................................................................. 57
Figure 3.2 - Absorbance Band Ratio Images of tissue array spots from Array P-16 ...... 60
Figure 3.3 - Histologic class mean spectra ................................................................. 63
Figure 3.4 - Histograms of metric value class frequency distribution for the three most populated classes (epithelium, mixed stroma, & fibrous stroma) for: A) Metric 02 (band ratio 1080/1544cm-1), and B) Metric 11 (band ratio 1400/1390 cm-1) ................................................................................. 67
Figure 3.5 - Graphical Representation of results of the leave-one-out analysis ............ 71
Figure 3.6 - Classification results for 2 tissue array spots from the same patient .......... 73
Figure 4.1 - Array P-80 histology classification results ............................................ 83
Figure 4.2 - Optical images of H&E stained section of Array P-80 ............................ 84
Figure 4.3 - Spatial filtering techniques for classified image results .......................... 86
Figure 4.4 - Sieve operation spatial filtering of histology classification results for patient 2 from array P-80 ................................................................. 88
Figure 4.5 - Array P-80 pathology ground truth ROIs ............................................ 89
Figure 4.6 - Patient-to-patient metric variation ....................................................... 91
Figure 4.7 - Array P-80 pathology classification results ........................................... 95
Figure 4.8 - Individual patient analysis of 20-metric GML pathology classification ...... 97
Fourier-Transform Infrared Spectroscopic Imaging of Prostate Histopathology

Daniel Celestino Fernandez

ABSTRACT

Vibrational spectroscopic imaging techniques have emerged as powerful methods of obtaining sensitive spatially resolved molecular information from microscopic samples. The data obtained from such techniques reflect the intrinsic molecular chemistry of the sample and in particular yield a wealth of information regarding functional groups which comprise the majority of important molecules found in cells and tissue. These spectroscopic imaging techniques also have the advantage of acquisition of large numbers of spectral measurements which allow statistical analysis of spectral features which are characteristic of the normal histological state as well as different pathologic disease states. Databases of large numbers of samples can be acquired and used to build model systems that can be used to predict spatial properties of unknown samples.

The successful construction and application of such a model system relies on the ability to compile high-quality spectral database information on a large number of samples with minimal sample-to-sample preparation artifact. Tissue microarrays provide a consistent sample preparation for high-throughput infrared spectroscopic profiling of histologic specimens. Tissue arrays consisting of representative normal healthy prostate tissue as well as pathologic entities including prostatitis, benign prostatic hypertrophy,
and prostatic adenocarcinoma were constructed and used as sample populations for infrared spectroscopic imaging at high spatial and spectral resolutions.

Histological and pathological features of the imaged tissue were correlated with consecutive tissue sections stained with standard histologic stains and visualized via traditional optical microscopy and reviewed with a trained pathologist. Spectral analysis of histologic class mean spectra and subsequent cross-sample statistical validation were used to classify reliable spectral metrics for class discrimination. Multivariate Gaussian maximum likelihood classification algorithms were used to reliably classify all pixels in an image scene to one of six different histologic subclasses: epithelium, smooth muscular stroma, fibrous stroma, corpora amylacea, lymphocytic infiltration, and blood. The developed database-dependent classification methods were used as a tool to investigate subsequent microarrays designed with both normal epithelial tissue as well as adenocarcinoma from a large population of patients. Such investigation led to the identification of spectral features that proved useful in the preliminary discrimination of benign and malignant prostatic epithelial tissue.
Chapter One - Introduction

Spectroscopy deals with the interaction of various forms of electromagnetic (EM) radiation with matter. Vibrational spectroscopy provides information regarding the molecular composition and structure of a wide range of materials including biological tissues. Recent technological advances have led to powerful vibrational imaging approaches involving both near and mid-infrared, as well as Raman-based platforms providing spatially-resolved chemical information on a microscopic scale\[1\]. Infrared spectroscopic imaging microscopy, in particular, benefits from many decades of instrumentation advances and database compilations. A brief background into the theory and techniques of infrared spectroscopy follows.

1.1 Electromagnetic Spectrum

The wave nature of electromagnetic (EM) radiation treats the radiation in terms of oscillating electric and magnetic fields perpendicular to one another and to the direction of wave propagation traveling with the velocity of light. Certain continuous regions of the EM spectrum have been designated and appear in Figure 1.1\[2\]. Vibrational absorption spectra result from the interaction of oscillating dipole moments, which occur during molecular vibrations, with the electric field of the radiation, resulting in an energy exchange between the radiation and the molecular system.

Electromagnetic radiation is characterized by its wavelength $\lambda$. The specific units typically used to express wavelength vary across the spectrum from angstroms (Å) in the
gamma ray region to meters in the radio wave region or $\sim 10^{-10}$ to $10^2$ cm, respectively.

The units of µm are practical for describing radiation in the mid-infrared spectral region.

In the near-infrared (NIR) region the unit nm typically employed just as it is in the visible (VIS) and ultraviolet (UV) spectral regions.

Electromagnetic radiation can also be characterized by its frequency $\nu$, defined as the number of oscillations of the magnetic or electric field radiation vector per unit of time[2]. The frequency unit is $s^{-1}$ (oscillations per second), often specified in Hertz (Hz). The energy ($E$) of EM radiation is directly related to its frequency ($\nu$) by the equation

$$E = h\nu$$  \hspace{1cm} (1.1)

where $h$ is Planck’s constant with a value $h = 6.63 \cdot 10^{-34}$ J s.

The frequency and wavelength ($\lambda$) of EM radiation are related by the proportionality constant $c$ (the speed of light) according to the equation

$$\nu = \frac{c}{\lambda}$$  \hspace{1cm} (1.2)
where c has a value of \( \approx 2.99793 \times 10^{10} \text{ cm s}^{-1} \) (in a vacuum).

Infrared spectroscopists have adopted the convention of expressing frequency in terms of wavenumber with the units of cm\(^{-1}\)[3]. A simple expression for wavenumber is given by

\[
\bar{\nu} = \frac{1}{\lambda}
\]  

(1.3)

The units of wavenumber provide a convenient scale for IR spectroscopy, especially the mid-infrared region that spans 200-4000 cm\(^{-1}\). The units of wavenumber are also desirable for IR spectroscopists because they are directly proportional to the energy of radiation, which varies inversely with wavelength as described by equation 1.4[4].

\[
E = \frac{hc}{\lambda}
\]  

(1.4)

The relationships between energy, frequency, and wavelength and the various regions of the electromagnetic spectrum are detailed in figure 1.1. The infrared region of the electromagnetic spectrum is subdivided into three contiguous regions; the near, mid and far infrared regions. The nomenclature of these prefixes refers to the individual sub-region’s position relative to the visible region. Figure 1.2 shows these three regions of the infrared spectrum and the ranges they occupy on the wavelength, frequency and wavenumber scales.
1.1.1 Interactions of Electromagnetic Radiation with Matter

All forms of spectroscopy deal with the interaction of radiation and matter. Numerous possible types of interactions exist and many involve transitions between specific molecular energy states. The monitoring of the absorption and emission of radiation from different regions of electromagnetic spectrum provides information regarding these molecular transitions and consequently gives information regarding the atomic and molecular composition of samples[5].

Quantum mechanical treatments describe both the wave and particle nature of electromagnetic radiation[5, 6]. As seen in figure 1.1, the electromagnetic spectrum spans an extremely wide range of frequencies, and therefore, energies. There are a variety of energy levels that molecules can occupy leading to the possibility of many transitions between states. These energy transitions are of varying magnitudes with corresponding frequencies depending upon the specific regions of the spectrum in which they occur. Radiation from different regions of the electromagnetic spectrum are used as
the basis of the many spectroscopic techniques that exist, for which each technique provides molecular information regarding the sample[2].

Table 1.1 contains examples of different types of spectroscopy based on specific regions of the electromagnetic spectrum and the type of chemical information probed.

<table>
<thead>
<tr>
<th>spectral region</th>
<th>wavelength range ((\lambda))</th>
<th>spectroscopy</th>
<th>information</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma) - rays</td>
<td>&lt; 0.05 Å</td>
<td>(\gamma) - ray spectroscopy</td>
<td>nuclear decay emission</td>
</tr>
<tr>
<td>X-rays</td>
<td>0.05 Å to 10 nm</td>
<td>x-ray spectroscopy</td>
<td>electronic structure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x-ray crystallography</td>
<td>molecular structure</td>
</tr>
<tr>
<td>Ultraviolet (UV)</td>
<td>10 nm to 350 nm</td>
<td>UV-VIS spectroscopy</td>
<td>electronic transitions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fluorescence spectroscopy</td>
<td>fluorescence emission</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Raman spectroscopy</td>
<td>vibrational transitions</td>
</tr>
<tr>
<td>Visible (VIS)</td>
<td>350 nm to 770 nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Near infrared</td>
<td>770 nm to 2.5 (\mu)m</td>
<td>IR Absorption spectroscopy</td>
<td>vibrational transitions</td>
</tr>
<tr>
<td>Mid Infrared</td>
<td>2.5 (\mu)m to 50 (\mu)m</td>
<td>IR Reflection spectroscopy</td>
<td>thermal emission</td>
</tr>
<tr>
<td>Far Infrared</td>
<td>50 (\mu)m to 1 mm</td>
<td>IR emission spectroscopy</td>
<td></td>
</tr>
<tr>
<td>Microwave</td>
<td>1 mm to 300 mm</td>
<td>microwave spectroscopy</td>
<td>rotational transitions</td>
</tr>
<tr>
<td>Radio Waves</td>
<td>&gt; 300 mm</td>
<td>NMR Spectroscopy</td>
<td>nuclear spin transitions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(in magnetic field)</td>
</tr>
</tbody>
</table>

Table 1.1 - Spectroscopic techniques utilizing different regions of the electromagnetic spectrum

1.2 Basis of Infrared Absorption

Photons in the infrared spectral region have energies representative of transitions between molecular vibrational energy levels. While spectroscopic techniques exist which make use of the reflection and emission of infrared radiation, we are most concerned with the absorption of infrared radiation. Nearly all molecules exhibit an infrared spectrum, the noted exceptions being homonuclear diatomics, such as the common gases \(\text{N}_2\), \(\text{O}_2\), and \(\text{H}_2\)[5].
Various interactions can occur between radiation and matter that result in the transfer of energy. Quantum mechanical principles require that molecules exist in quantized energy states and thus the absorption of energy results in bands that characterize an infrared spectrum.

1.2.1 Requirements for IR Absorption

The wave nature of quantum mechanics is most simply represented by the time independent Schrödinger equation

\[ H\psi = E\psi \]  \hspace{1cm} (1.5)

where \( \psi \) is the wavefunction of the system, \( H \) is the Hamiltonian operator, and \( E \) is the energy of a state characterized by \( \psi \)[6]. The wavefunction can be used to calculate the transition moment \( R \) as shown in the equation

\[ R = \int \psi_i^* \mu \psi_j \, d\tau \]  \hspace{1cm} (1.6)

for a transition between states \( i \) and \( j \), where \( \mu \) is the electric dipole moment operator (\( \mu = er \), \( e \) is the electronic charge, \( r \) is the distance between the charges), and \( d\tau \) indicates the integration over all space. For vibrational motions, the electric dipole moment \( \mu \) is expressed as

\[ \mu = \mu_0 + (r - r_e) \left( \frac{\partial \mu}{\partial r} \right)_0 + \frac{1}{2} (r - r_e)^2 \left( \frac{\partial^2 \mu}{\partial r^2} \right)_0 + \ldots \]  \hspace{1cm} (1.7)

where \( \mu_0 \) is the permanent dipole moment, \( r \) is the internuclear distance and \( r_e \) is the equilibrium bond distance[5]. If we consider only the first two terms in equation 1.7 and substitute for \( \mu \) in equation 1.6 we obtain
\[ R = \int \psi_i^* \left[ \mu_0 + (r - r_e) \frac{\partial \mu}{\partial r} \right] \psi_j \, \partial \tau \]  

which reduces to

\[ R = \int \psi_i^* \left[ (r - r_e) \frac{\partial \mu}{\partial r} \right] \psi_j \, \partial \tau \]  

since \( \mu_0 \) is a constant and \( \int \psi_i^* \psi_j \, \partial \tau = 0 \) because of the orthogonality of the wavefunctions[2].

From equation 1.8 it is clear that there must be a change in dipole moment during the vibration in order for a molecule to absorb infrared radiation. The selection rules predict that the fundamental absorption will occur with vibrational quantum number \( \Delta \nu = \pm 1 \) for a harmonic oscillator, with much weaker overtone absorption corresponding to \( \Delta \nu = \pm 2 \) etc. for anharmonic conditions[6].

All molecules that are more complex than diatomics have multiple vibrational modes. These vibrational modes each have associated energies that correspond to the particular frequency or wavenumber of infrared radiation. The number, type, and energies of these vibrations are dictated by the molecular structure of the system in terms of the bonds, geometry, atomic masses, and force fields and are thus representative of specific molecules[2].

Vibrational modes that produce a change in dipole moment result in the absorption of IR radiation and are termed infrared-active. Vibrational modes that do not induce in a change in dipole moment are termed infrared-inactive. The requirement for a change in dipole moment during a molecular vibration explains why, for example, homonuclear diatomic molecules do not absorb infrared radiation[4].
1.2.2 Number of Vibrational Modes

While diatomic molecules can vibrate only in one dimension or mode, more complicated molecular structures present other possible vibrational modes. Linear molecules with \(N\) atoms exhibit \(3N-5\) vibrational modes, while nonlinear molecules have \(3N-6\) vibrational modes[5]. Water (a nonlinear triatomic) and carbon dioxide (a linear triatomic) are illustrative examples. As seen in figure 1.4, the carbon dioxide molecule’s additional symmetry provides it with four possible vibrational modes while the water molecule has only three. Note also that the symmetric stretch of the carbon dioxide molecule produces no net change in dipole moment and is thus infrared-inactive[4].

<table>
<thead>
<tr>
<th>A</th>
<th>water</th>
<th>B</th>
<th>carbon dioxide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vibration</td>
<td>(v_n)</td>
<td>Band position</td>
<td>Infrared activity</td>
</tr>
<tr>
<td>Symmetric Stretch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asymmetric Stretch</td>
<td>(v_2)</td>
<td>3756 cm(^{-1})</td>
<td>IR-active</td>
</tr>
<tr>
<td>Bend</td>
<td>(v_3)</td>
<td>1696 cm(^{-1})</td>
<td>IR-active</td>
</tr>
<tr>
<td>Bending (out-of-plane)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from [2]

Figure 1.3 - Vibrational modes and IR activity of water vapor (A) and carbon dioxide (B) molecules

As molecular structural complexity increases, other types of vibrational modes become possible. The methylene group, for example is capable of six different vibrational modes as illustrated in figure 1.4.
1.2.3 Group Frequencies

Various chemical functional groups exhibit specific infrared frequencies representative of their structures. Frequencies such as these are known as characteristic or group frequencies[4]. Many of the most common functional groups with characteristic group frequencies are familiar organic groups. Functional group frequencies allow the spectroscopist to use IR spectra to qualitatively identify structural elements in samples. Since vibrational frequency absorption profiles parallel functional group structure, the spectroscopist investigating biological material using vibrational techniques often depends upon existing databases and extensive compilations of spectral information.
1.3 IR Spectral Feature of Tissues

Modern approaches to histology categorize cells into different types based on their primary physiological function[7]. In such a system cells belong to one or more of the following groups: epithelial cells, support cells, contractile cells, nerve cells, germ cells, blood cells, immune cells, or hormone-secreting cells.

From a molecular point of view, all of these various types of specialized cells encountered in biological tissue are predominately comprised of four major types of biomolecules or their subunits: proteins, carbohydrates, lipids, and nucleic acids. Additionally, all four of these types of molecules each have a great deal of structural redundancy. That is, they tend to form polymeric molecules based on subunits that while different, reflect structural similarity. For example, thousands of different proteins exist in a typical cell, and while the individual structure of each protein is different, they are all made from the same set of amino acids, and share a common backbone structure.

1.3.1 Proteins

Protein molecules play many fundamental roles in the life of every cell in addition to serving various important extracellular functions in many tissues. The significance of proteins to biological organisms cannot be understated and their utility is evident in the many functions they perform including: enzymatic catalysis, transport and storage, coordinated motion, mechanical support, immune protection, generation and transmission of nerve impulses, and control of growth and differentiation[8].

All proteins are formed as linear chains of amino acid building blocks that can form various secondary and tertiary structures. Eukaryotic proteins are typically assembled
from a set of 20 different $\alpha$-amino acids that share a common template and are distinguished by unique side chain structures\[9\]. Figure 1.5 shows the molecular structure of a typical amino acid.

![Figure 1.5 - Structure of a typical amino acid](image)

All amino acids share a common structure that includes a central or $\alpha$-carbon atom bonded to a carboxyl group, an amino group and a hydrogen atom. At physiologic pH the amino group is protonated (NH$_3^+$) and the carboxyl group exists as the carboxylate ion (COO$^-$)[9], displayed in figure 1.5. Each different amino acid contains a distinctive structure at the side chain position designated as R in figure 1.5.

The primary protein or polypeptide structure is formed by linking these amino acid subunits together in a linear chain via a condensation reaction between the amino and carboxyl groups of adjacent amino acids in a linear chain[10]. The linkage that is formed between these amino acid subunits is known as a peptide bond and polypeptide chains that result form a repeating backbone structure that is the same for all proteins. Figure 1.6 shows the basic protein primary structure and the locations of these peptide bonds.
The polypeptide backbone structure consists of several functional groups, including a C-N group, a C-H group, an NH₂ group, and a carbonyl group (C=O). Since these functional groups repeat for every amino acid in a protein regardless of the protein’s identity or higher-order structure, the absorbance bands resulting from these structures dominate the IR spectra of most proteins. The most prominent of these absorbances include; the Amide I absorption near 1650 cm⁻¹ arising from C=O stretching vibrations (80%) weakly coupled to C-N stretching vibrations (20%), the Amide II absorption near 1545 cm⁻¹ arising from N-H bending vibrations (60%) coupled to C-N stretching vibrations (40%), the Amide III absorption near 1236 cm⁻¹ arising from C-N stretching vibrations, and the Amide A absorbance near 3290 cm⁻¹ arising from N-H stretching vibrations[11].

In their native states, most proteins do not exist as simple linear polypeptide structures, but instead form complex secondary and tertiary structures that impart a distinct three-dimensionality to a particular protein. The most common protein
secondary structures are the \( \alpha \)-helix and \( \beta \)-pleated sheet configurations depicted in figure 1.7.

![Figure 1.7](image)

Figure 1.7 - Common Protein Secondary Structures: \( \alpha \)-helix and \( \beta \)-sheet
\( \beta \)-pleated sheet structures can form between parallel polypeptide chains, or between strands with antiparallel orientation, as shown in the figure. The dotted lines indicate hydrogen bonds.

Both of these recurrent secondary structures involve hydrogen bonding between the oxygen atoms of backbone carbonyl groups and the hydrogen atoms of backbone N-H groups indicated in the figure as dotted lines. These structural arrangements change bond
angles and other structural parameters, causing frequency shifts of absorbance bands arising from backbone vibrations. As a result, the relationship between IR band positions of protein backbone absorbances, most notably the Amide I absorbance near 1650 cm\(^{-1}\), and protein structure has been the subject of much work over the past decade[12-16]. For example, several studies have examined the amide I bands of polypeptides and proteins whose structures are known to be dominated by one of the common secondary structure motifs, such as \(\alpha\)-helix, \(\beta\)-sheet, or unordered structures[17-19]. Such studies have led to the development of some empirical rules for the correlation of amide I band features and common secondary structural motifs.

On the basis of these empirical rules, IR bands in the 1660-1650 cm\(^{-1}\) spectral region are assigned to \(\alpha\)-helices, 1640-1620 cm\(^{-1}\) to \(\beta\)-sheets, 1695-1660 cm\(^{-1}\) to \(\beta\)-sheets and \(\beta\)-turns, and 1650-1640 cm\(^{-1}\) to unordered structures[20]. Such empirical rules are useful guidelines for obtaining structural information from vibrational spectroscopic information, however, many studies show that such rules are not free from shortcomings[19]. For instance, IR studies of proteins such as myoglobin and hemoglobin, for which x-ray crystallographic data suggests highly helical-structures with no \(\beta\)-sheets, have shown Amide I absorbances in the 1640-1620 cm\(^{-1}\) region[21, 22]. While no conclusive evidence exists to explain the presence of such lower-frequency \(\alpha\)-helix amide I bands, some have suggested that strong hydrogen bonding of peptide groups with solvent molecules and distortion of helix structures may contribute to such findings[23, 24].
1.3.2 Carbohydrates

Carbohydrates are aldehyde or ketone compounds with multiple hydroxyl groups. These important biomolecules play three central roles in all organisms: First, they serve as energy stores and metabolic intermediates. Stored glycogen can be readily broken down into glucose, a preferred metabolic fuel. Glucose is broken down to yield adenosine triphosphate (ATP), a phosphorylated sugar derivative and universal currency of energy in the organism. The second important role of carbohydrates is as basic structural components of nucleic acids. Ribose and deoxyribose sugars are structural units of all nucleotides and ribonucleotides whose sequence in nucleic acids is responsible for the storage and expression of genetic information. A third important role of carbohydrates in organisms is that they are often linked to proteins and lipids on cell membranes, many playing critical roles in cell signaling and recognition[25, 26].

Common cellular carbohydrates have many vibrational spectral features in the fingerprint region of the mid-IR spectrum due to various vibrational modes of C-O, C-C, and carboxylate groups. Infrared spectroscopy has been used extensively to help characterize biologically important polysaccharide cell-surface components, including glycolipids like diacyl sugars[27], cerebrosides[28, 29], gangliosides[30, 31], lipopolysaccharides[32-34], and mucopolysaccharides[35].

1.3.3 Lipids

Lipids form another important class of biomolecules found in tissue that play many important roles. Like carbohydrates, lipids provide an important source of energy for metabolism. The hydrophobic nature of lipids contributes significantly to their central
role in cellular membrane function, providing barriers which partition cells and
subcellular organelles. Additionally, lipids perform a variety of other important
functions, from the coenzyme roles of fat-soluble vitamins to the regulatory roles of
prostaglandins and steroid hormones to structural and functional roles in the nervous
system.

Lipids all share the characteristic of having non-polar, hydrophobic domains. In
many cases, long chain fatty acids are responsible for this hydrophobicity, and such lipids
have many vibrational modes associated with C-H groups across the fingerprint region of
the mid-IR. The spectral frequency region between 3000-2800 cm\(^{-1}\) also contains four
prominent absorbance bands common to many lipids: the methyl antisymmetric stretch
\((\text{as} \nu \text{CH}_3)\) at 2962 cm\(^{-1}\), the methyl symmetric stretch \((\text{s} \nu \text{CH}_3)\) at 2872 cm\(^{-1}\), the
antisymmetric CH\(_2\) stretch \((\text{as} \nu \text{CH}_2)\) between 2936-2916 cm\(^{-1}\), and the symmetric CH\(_2\)
stretch \((\text{s} \nu \text{CH}_2)\) between 2863-2843 cm\(^{-1}\)[36].

Unfortunately, most standard methods for the preparation of sectioned tissue
involve the use of one or more nonpolar solvents such as ethanol or xylenes that remove
lipids from the tissue section[37, 38]. As a tissue source for FT-IR spectroscopic studies,
formalin-fixed paraffin-embedded tissue offers some advantages over frozen tissue
including higher-quality preservation and access to large libraries of preserved tissue,
however, paraffin exhibits many of these common lipid absorbances, and therefore must
be removed from tissue sections intended for spectroscopic analysis. Effective paraffin
removal requires the use of strong nonpolar solvents such as hexane for several hours at
temperatures of 40°C further contributing to the extraction of physiologic lipids from
paraffin-embedded tissue.
1.3.4 Nucleic Acids

Nucleic Acids have been studied extensively in both purified state as well via model compounds[39]. The most prominent absorbances reported are due to vibrations of several functional groups on the repeating backbone structure of nucleic acids. These include absorbances near 1080 cm\(^{-1}\) and 1240 cm\(^{-1}\) attributed respectively to the symmetric and asymmetric stretch of phosphodiester (PO\(_2\)) moieties[40]. However, the ability of IR spectroscopy to attain vibrational information from quiescent nuclear DNA from cell preparations or tissue sections has recently been called into question and some theoretical analyses of chromatin density and packing used to support the idea that nuclear DNA is too dense to produce appreciable absorbances in transmission IR spectroscopic experiments[41].

1.4 FTIR Spectroscopy Background

Modern instrumental approaches to the collection of spatially-resolved infrared spectroscopic data share many characteristics and all benefit from the extensive advances made in the field of Fourier transform infrared (FTIR) spectroscopy over the past three decades. Several excellent books[4, 42, 43] have been written on the subject of FT-IR spectroscopy and contain comprehensive information on the technology that has been implemented for years in commercial FT-IR spectroscopy systems.

Infrared microspectroscopic imaging systems share many common features. Most consist of a research-grade FT-IR spectrometer that provides an output beam of modulated infrared radiation used as a source for an infrared microscope equipped with infrared detectors[44]. Modern approaches to the collection of spatially-resolved spectral
data are best differentiated in terms of the type of infrared detection employed. The
following sections discuss instrumental aspects of spectrometers and infrared
microscopes, as well as strategies for collecting FT-IR spectroscopic imaging data with
three different types of infrared detection: single-point mapping, raster scanning with
linear multichannel detectors, and global FT-IR imaging with Focal Plane Array (FPA)
detectors.

1.4.1 FTIR Spectrometers

The majority of commercial research-grade FTIR spectrometers incorporate a
broadband infrared source, Michelson interferometer, sample compartment, and infrared
detection with either deuterated triglycine sulfate (DTGS) or mercury cadmium telluride
(MCT) single-point detectors. Many commercial FTIR instruments exist for dedicated
analyses typically implemented in industrial settings for process assessment and quality
control analyses. Such spectrometers are typically designed to be lower in cost than
research-grade spectrometers, which offer more flexibility in the types of measurements
that are possible as well as increased sensitivity and higher spectral signal-to-noise ratios
(SNRs).

Figure 1.8 shows the schematic design of the Michelson interferometer, which is
the optical portion of the spectrometer that is used to modulate the radiation. The
interferometer is composed of two perpendicular beam paths often referred to as separate
arms of the interferometer. These beampaths intersect at the beamsplitter, an optical
component that when placed at 45-degree angle to the normal both reflects and transmits
exactly 50% of incident radiation. In the mid-IR region, beamsplitters are typically
constructed from potassium bromide (KBr) with a thin coating of germanium (Ge) or silicon (Si), and many commercial instruments allow beamsplitters to be changed to other materials for coverage of specific spectral regions[42].

Figure 1.8 - Michelson Interferometer

As depicted in Figure 1.8, polychromatic radiation from an infrared source, typically a ceramic globar, is passed through an aperture to form a beam. This beam strikes the beamsplitter at a 45° angle, dividing the beam in half. Half of the beam is directed at a fixed mirror, while the other half is diverted to a mirror whose displacement can be varied along the axis of the incident beam. After striking these mirrors, the beams in the two arms of the interferometer are sent back to the beamsplitter, where they recombine and interfere with each other. The beamsplitter divides the recombined beam in half again, sending half back toward the source, while the other half is used for spectroscopy and is directed through sample and subsequently detected[4].
When the moving mirror occupies a displacement where the pathlengths in the two arms of the interferometer are equal, then the recombining beams are precisely in-phase and only interfere constructively. This mirror position produces the most intense beam for every frequency of radiation. As the mirror moves from this position, a pathlength difference is created in the two arms of the interferometer that causes specific interference patterns for different mirror displacements. If the mirror is continuously scanned, then the intensity of the recombined beam will vary with respect to time in a frequency or wavelength dependent manner[2].

The function of the spectrometer is to encode a modulation on the polychromatic IR source radiation such that detection of the intensity of the encoded radiation with respect to time or in the “time domain” yields spectral information in the “frequency domain”. The “Fourier transform” part of the technique’s name refers to the mathematical operation that is required to transform the raw data collected by the instrument in the time domain, known as the interferogram, into a intensity profile in the frequency domain, otherwise know as an infrared spectrum.

1.4.2 Infrared Microscopy

Infrared microspectroscopic imaging systems typically couple the modulated output beam of a FTIR spectrometer to an infrared microscope for use as source radiation for obtaining spectroscopic information from microscopic regions of a sample. Infrared microscopes perform similarly to conventional optical microscopes and are typically set up to image with visible light along the same optical path. However, they have many structural differences that stem from some fundamental properties of infrared radiation.
One major limitation of infrared spectroscopy is related to its exceptional molecular sensitivity. As mentioned in section 1.2, all covalently bonded molecules, with the exception of homonuclear diatomics, absorb infrared radiation. Optical components used in conventional microscopes are composed almost exclusively of borosilicate glass or quartz, both of which have broad absorbances over much of the infrared spectrum. For this reason, infrared microscopes are designed to use reflective optics wherever possible, and refractive optics have to be manufactured from alternative materials, such as halide salts, which are transparent over the spectral regions of interest[42].

Most Infrared microscopes use Cassegrain condenser and objective lenses and can be operated in either transmission or reflectance modes. In reflectance mode, one side of the Cassegrain objective primary mirror is typically used to direct the radiation onto the sample while the opposite portion of the primary mirror is used to collect the reflected radiation. Infrared microscopes are often outfitted with automated high-precision motorized mapping stages, which permit the sample to be positioned precisely in the plane perpendicular to the optical path. Most microscopes incorporate a visible light source and detection system, typically a video camera. Adjustable mirrors are used to switch between visible and infrared modes and some models incorporate a beamsplitter to allow for simultaneous imaging in both spectral regions[45].

The different strategies that can be employed to collect spatially-resolved infrared microspectroscopic data depend on the types of infrared detection systems available of the microscope[44]. Panels A-C of Figure 1.9 depict three different approaches based respectively on single-point, linear-array, and focal plane array (FPA) detection. A discussion of each approach follows.
Figure 1.9 - Three Instrumental Approaches for collection of spatially resolved FTIR spectroscopic data
A) Point-mapping using single element detection; B) Raster-Scan imaging using linear multichannel detection; and C) Global FT-IR imaging using 2-D focal plane

1.4.3 Mapping with Single-Point Detectors

In single element microspectroscopic instrumentation, spectral information from a small, specified area of the sample is obtained by restricting the area illuminated by the infrared beam using opaque apertures of controlled size. The collected radiation is then diverted to a sensitive detector. To identify the area to be examined, however, a corresponding white light optical image is also required. Clearly, focusing the infrared
beam for maximal throughput and minimal dispersion in the sample plane requires the optical and infrared paths be parfocal and collinear[45].

By restricting the infrared beam to a small spatial area of the sample, and sequentially moving to different regularly-spaced sample locations with a high precision microscope stage, spatially-resolved spectroscopic data from large sample areas can be mapped out point by point. This strategy, often referred to as point-mapping, suffers from several limitations.

The cross-sectional diameter of the beams used in such infrared microscopes must be large enough to fully illuminate the area passed by the largest aperture setting that may be employed, for example a 100x100 um square. There is a tradeoff between the spatial resolution of mapping data that can be acquired and corresponding throughput due to the need to block out more and more of the available radiation. Aperture use decreases the instrumental throughput due to diffraction when the aperture is of the same dimension as the wavelength of light (~3-14um), thus limiting the highest achievable data spatial resolution. Apertures also permit the passage of some diffracted light from outside the apertured region. The use of a second set of apertures in tandem to reject stray radiation can improve spatial fidelity, unfortunately at the cost of additional throughput loss.

Throughput is important because it directly affects the spectral signal to noise ratio (SNR), and losses in throughput require larger acquisition times for signal recovery[42].

Data acquisition time is the major drawback to single-point mapping approaches. Spectral information is acquired for each spatial location in the final map one-by-one and there is significant time overhead for moving the sample to each new sampling location.
1.4.4  Raster-scan Imaging Using Multichannel Detectors

While single element microspectroscopy provides the capability to obtain spectra from small spatial regions, poor SNR characteristics, diffraction effects and stray light issues resulting from the use of apertures limit the applicability of this point mapping approach. A multichannel detection approach to circumvent some of these issues has recently been implemented[46] with a linear array detector employed to image an area corresponding to a rectangular spatial area on the sample. The sample stage is moved precisely to sequentially image a selected spatial area on the sample. This data collection strategy is referred to as *push-broom* mapping or *raster scanning*. The process is conceptually similar to point-by-point mapping but takes advantage of the multiple channels of detection. Hence, imaging a large sample area is faster by a factor of \( n \), for a linear array detector containing \( n \) elements. The instrument is schematically displayed in Figure 1.9B.

Point mapping detectors are typically 100 – 250 µm in size; in contrast, an individual detection element in a linear array detector is of the order of tens of micrometers. Employing a linear array eliminates the need for apertures, as small detector elements directly image different sample spatial regions. For example, a detector element 25 µm in size can be operated at 1:1 magnification or 4:1 magnification to provide a 25 µm or a 6.25 µm effective pixel size with available, relatively aberration-free infrared optics. This approach circumvents the debilitating diffraction effects resulting from the use of small apertures in single channel detection systems and provides higher quality data when desired spatial resolutions approach the wavelengths of light being used. In addition, the spatial resolution, data quality, and time for data acquisition
are no longer coupled as in point mapping methods. The data acquisition time depends solely on the size of the image and quality of data desired, and is correlated less with the spatial resolution, which is determined by the employed optics.

A high-precision, motorized stage that reproducibly steps in small increments is used and the interferometer is operated in a continuous scan mode. In combination with high performance multichannel detectors, this mode combines high performance multichannel detectors with the most desirable properties of rapid-scan interferometry to yield high quality spectroscopic imaging data.

1.4.5 Global FTIR Spectroscopic Imaging

The state of the art in FTIR microspectroscopic imaging instrumentation is the combination of an infrared microscope equipped with a focal plane array (FPA) detector and an FTIR spectrometer[47, 48], as shown in Figure 9C. FPA detectors are constructed of thousands of individual detection elements laid out in a two-dimensional grid pattern. An FPA matched to the characteristics of the optical system is capable of imaging the entire field of view afforded by the optics and of utilizing a large fraction of the infrared radiation spot size at the plane of the sample. The increase in the number of individual detectors with respect to a linear array provides a correspondingly larger multichannel advantage. For example, an FPA with pixel dimensions $p \times p$, provides a $p^2$ time savings relative to a single element detector and a $p^2/n$ time savings compared to a linear array detector containing $n$ elements. For a 128 x 128 element FPA detector relative to the single element case, the advantage is a factor of 16,384, while compared to a 16-element linear array detector; the multichannel advantage is a factor of 2048. FPA
detectors are also capable of imaging large spatial areas simultaneously without inherent inefficiencies of moving the sample or re-setting the interferometer to scan a different area. The considerable reduction in data acquisition times allows for imaging large areas, as well as the examination of dynamic processes in a single field of view[49].

The first and, to date, most popular approach to FTIR micro-imaging spectrometers incorporates a step-scan interferometer[50]. While continuous or rapid-scan spectrometry involves scanning the moving mirror at a constant velocity, a step-scan interferometer is capable of stepping the moving mirror to discrete, evenly-spaced intervals and maintaining individual mirror positions with very little displacement error. A constant retardation over an extended time period allows suitable time for signal averaging and for data readout and storage. Short time delays prior to data acquisition are necessary for mirror stabilization at the onset of the step. Detector signal is integrated for only a fraction of the total time required for collection of each frame. The integration time, number of frames co-added, and number of interferometer retardation steps (a function of desired spectral resolution) determine the total time required for the experiment. Since the integration time determines the data quality, efforts have been made to increase the ratio of the integration time to the total data acquisition time[51].

Imaging configurations that utilize a rapid scan interferometer have been proposed for small arrays[52]. Slow data readout and storage rates for many FPA detectors preclude conventional rapid-scan mirror velocities, thus approaches must make use of so-called slow-scan mirror velocities of \( \leq 0.01 \text{ cm/s} \). A generalized data acquisition scheme that permits true rapid scan data acquisition for FPA detectors has been proposed[53], where the integration time of individual frames collected by the FPA detector is
negligible with respect to the complete interferogram acquisition. For most FPA
detectors available today, the motion of the moving mirror does not allow co-addition of
frames at individual retardations in the continuous scanning mode, but successive single-
frame acquisitions can be averaged to increase data SNRs. Compared to step-scan data
acquisition, rapid scan data collection (mirror velocity > 0.025 cm/s) allows for fast
interferogram capture as no time is spent on mirror stabilization. The error arising from
the deviation in mirror position during frame collection is hypothesized to be the next
largest contributor of noise compared to the dominant contribution from random detector
noise[50]. At present, the advantages of continuous-scan relative to step-scan approaches
are a decreased cost of instrumentation and an increased data collection efficiency.

1.5 Spectroscopic Imaging: Data Structure and Applications

Spectroscopic imaging data, regardless of its method of collection, can be
conceptualized as an image cube with two dimensions corresponding to the spatial axes
of the sample and the third dimension to the spectral frequency or wavelength. Digital
image data is represented as a collection of rectangular picture elements or pixels, each
with an associated brightness value or magnitude. Spectroscopic image data can be
thought of as a collection of super-imposable and spectrally consecutive image planes,
whose pixel values consist of the spatially independent absorbance at the spectral
frequency or wavelength specified by the image plane. Alternatively, the data structure
can be conceptualized to consist of individual spatial locations or pixels each with an
associated absorbance spectrum. The concept of the image cube is represented
schematically in figure 1.10.
Figure 1.10 - Schematic representation of the image cube

These alternative views of the data structure influence the type of information that can be extracted from the data. For example, we can specify distinct spatial locations in a spectroscopic image, and display the associated spectra for simultaneous comparison of absorption features across the full spectral region collected. Alternatively we can specify a particular absorption feature of interest and display the associated spectral image plane. The brightness values of pixels in such an image will correspond to the sample’s spatial distribution of the species responsible for the absorption at the associated spectral frequency.

FTIR imaging of biological systems has demonstrated a potential to complement other imaging approaches. For biomedical applications, the technique may be used to
examine chemical changes due to pathological abnormalities and to follow histological alterations with high accuracy. Non-destructive morphological visualization of chemical composition rapidly provides structural and spatial information at an unprecedented level. Specifically, thousands of spectra routinely acquired in an imaging experiment may be employed for statistically meaningful data analyses, which in the example of biological tissue samples may prove ultimately useful in medical diagnoses. Since the visualization contrast is dictated by inherent chemical and molecular properties, no sample treatments, such as histopathological staining techniques required for optical microscopy, are necessary.

A typical example of the type of tissue information that can be retrieved was demonstrated by examining monkey cerebellum sections[54]. Distributions of lipid relative to protein allowed easy differentiation of white and gray matter areas. Purkinje cells in rat cerebella, which strongly influence motor coordination and memory processes, were visualized using FTIR imaging techniques[55, 56]. Neuropathologic effects of a genetic lipid storage disease, Niemann-Pick type C (NPC)[57], were distinguishable on the basis of spectral data without the use of external histological staining. Statistical analysis provided a numerical confirmation of these determinations consistent with a significant demyelination within the cerebellum of the NPC mouse. IR spectroscopy has been used for a number of years to characterize mineralized structures in living organisms (notably, bone). FTIR imaging spectroscopy[58, 59] of bone allows spatial variations of a number of chemical components to be non-destructively monitored. Correlations in bone between FTIR imaging and optical microscopy involving chemical composition, regional morphologies and the developmental processes have been made,
and an index of crystallinity/bone maturity could be determined providing structural information in a non-destructive manner[60].

1.5.1 Image Classification Methods

One of the most useful approaches to extracting data from such data structures is the process of image classification. Image classification algorithms automatically assign each pixel in an image scene to a specific class or group based on its spectral properties or pattern. *Unsupervised Classification* refers to the automatic partitioning of pixels into classes of spectral similarity without the use of any class training data. *Supervised Classification* is the process of classifying pixels into specific classes based on their spectral similarity to user-supplied training data for each class.

Unsupervised classification methods have the advantage that no extensive prior knowledge of the image scene is necessary and the potential for human error is far less than with supervised methods. Additionally, they are useful for finding natural spectral patterns and groups in spectral images. However, they are limited in their usefulness by the need to identify the resulting classes after the classification is performed[61]. For this reason, such unsupervised methods are of little usefulness for diagnostic implementation.

Supervised classification methods have several advantages relative to unsupervised strategies. First, the analyst has control over the specific number and identity of class categories and can tailor them for specific tasks. Supervised classification is tied to areas of known identity, determined through the process of selecting training regions. Additionally, regions of training data can be used during the process of classifier development to evaluate classifier performance. While inaccurate classification of
training data indicates serious classification problems and/or problems with training data selection, accurate classification of training data does not always assure accurate classification of other image data[62].

Supervised image classification methods have several disadvantages and limitations as well. By creating classes and assigning training populations, the analyst imposes a classification structure on the data. If the user-defined class structure does not match the natural class structure within the data, the classes may not be distinct or well defined in multidimensional space. Training populations that do not accurately represent the natural distribution of values within a class may result in severe classification error[63]. Finally, classes unknown to the analyst and not included in the training data may also be misclassified and thereby remain undiscovered.

1.6 Prostate Background

1.6.1 Anatomy and Histology

In men, the prostate is a retroperitoneal gland located just below the bladder that surrounds the urethra. The gland is divided into four zones: peripheral, central, transitional, and periurethral as shown in Figure 3.1. Distinctions between these zones are important because proliferative lesions vary according to the zone in which they occur. For instance, nodular hyperplasia, also known as benign prostatic hypertrophy or hyperplasia (BPH), occurs predominantly in the central zone, whereas most adenocarcinomas occur in the peripheral zone[64].
Histologically, the prostate is a compound tubuloalveolar gland in which glandular spaces are lined by epithelium. Specifically, the gland is lined by a layer of low cuboidal epithelium at the basal surface, which is covered by a layer of columnar mucus-secreting cells. The glands contain a discrete basement membrane and are separated by abundant fibromuscular stroma. Some ducts in the gland are lined by tall columnar epithelium, but as they approach the urethra, the epithelium changes to more cuboidal and eventually into the transitional epithelium that lines the urethra and urinary bladder[65].
While prostatic epithelial tissue and fibromuscular stroma make up the bulk of the gland, there are several other important histological features seen in the prostate. Numerous blood vessels run throughout the prostate, as well as peripheral nervous tissue innervating the gland. Prostates from older men frequently contain small, spherical corpora amylacea composed primarily of condensed glycoprotein in the glandular lumina[7].

1.6.2 Prostate Pathology

1.6.2.1 Incidence

Prostatic carcinoma is the most common form of cancer in men and it is estimated that 221,000 new cases will be diagnosed in the United States in 2003[66]. The incidence of newly diagnosed cases of prostate cancer in the US was 100,000 in 1988, and has risen steadily since then to just under 200,000 in 1994[67]. Mortality in the US due to prostate cancer rose from 28,000 to 36,000 during the same time period, however recent evidence suggests that mortality has peaked and may be falling[68]. The estimated mortality for US men in 2003 is 29,000[66]. This decline has been attributed to increased screening efforts and active treatment of localized disease by radiation and radical prostatectomy[69].

1.6.2.2 ‘Latent’ Prostate Cancer

In 1954, Franks observed an extraordinarily high prevalence of microscopic foci of what he termed ‘latent’ prostate cancer during autopsy of men who died from other diseases[70]. His observations have been corroborated by several investigators[71, 72]
and the occurrence of these incidental cancers has been shown to increase with age affecting approximately 20% of men in their 20’s, 30% of men in their 50’s, and 70% of men in their 80’s[73]. The lifetime chance that a man will develop clinically apparent prostate cancer is less than 10%[74], thus the majority of these tiny cancers detected at autopsy are clinically insignificant. While it is clear that early diagnosis and treatment of prostate adenocarcinoma leads to an improved mortality and morbidity, these findings point out the importance of being able to differentiate potentially dangerous cancers from the very small, well-differentiated, slow-growing lesions which are unlikely to present clinically during the patient’s natural lifespan.

1.6.2.3 Etiology and Risk Factors

It has become clear that genetics play a significant role in the pathogenesis of prostate adenocarcinoma. Male relatives of men who have died from prostate cancer have a greater-than-expected incidence of the disease. An early study by Woolf of 228 men dying of prostate cancer found the relative nearly 3-fold increase in the relative risk of first-degree relatives compared to a control group[75]. Subsequent studies have confirmed this familial association[76-78], and demonstrated the importance of screening PSA values in asymptomatic men from families with 3 or more members affected by prostate cancer[74, 79].

Recent evidence supports the existence of a genuinely hereditary form of early onset prostate cancer exhibiting Mendelian autosomal dominant inheritance[80]. The exact gene defects have not been elucidated for these families but possible locations have been mapped to chromosome 1q24-25[81] as well as the X chromosome suggesting the
possibility of X-linked inheritance[82]. Recent evidence suggest that mutations in the tumor suppressor genes BRCA-1[83] and BRCA-2[84, 85] confer increased risk of developing prostatic adenocarcinoma, and attempts to screen for those at risk are currently being studied[86]. The most influential factor conferring risk of developing prostate cancer besides familial inheritance is age[87]. African-American men have roughly twice the lifetime risk of their white counterparts and higher PSA and tumor volume in a study adjusted for age, stage, pathologic stage, Gleason score, and volume of benign disease[88].

Other predisposing factors for clinical prostate cancer include the presence of testosterone and dihydrotestosterone (DHT), sexual history positive for early first sexual experience and multiple sexual partners[89], a diet high in saturated animal fat and low in yellow and green vegetables, and environmental or occupational exposure to several pollutants including cadmium[90] and the radioactive agents $^{51}$Cr, $^{59}$Fe, $^{60}$Co, and $^{65}$Zn[91]. Vasectomy has been suggested as a possible risk conferring event[92-94] though some studies failed to demonstrate a conclusive link[95, 96].

1.6.2.4 Diagnosis

1.6.2.4.1 Clinical Presentation

With the recent widespread increase of PSA testing in men at risk for prostate cancer, a large proportion of patients presenting with the disease are asymptomatic. Clinically apparent prostate cancer presents with a spectrum of symptoms related to the extent of disease progression. Urinary symptoms occur in localized as well as advanced disease states as well as in extremely common condition of benign prostatic hyperplasia.
(BPH). Symptoms related to bladder outflow obstruction, such as hesitancy, poor stream, and a sensation of incomplete voiding arise from urethral occlusion by the tumor or nodular mass. Urinary frequency and urgency are irritative symptoms that develop due to detrusor muscle instability secondary to outflow obstruction or directly by tumor invasion of the trigone of the bladder and pelvic nerves. Invasive cancer can produce other symptoms both locally and at distant sites. Local extension of prostate cancer can present with hematuria and/or hemospermia due to invasion of the prostatic urethra or seminal vesicles. Direct invasion of the distal urinary sphincter can cause urinary symptoms unrelated to outflow obstruction, while similar invasion of the neurovascular bundles posteriorly can lead to erectile dysfunction and pain. Significant posterior invasion of prostate cancer can produce lower bowel symptoms including rectal bleeding and constipation due to large intestine obstruction near the rectum. Symptoms that indicate local metastatic disease include bone pain, paraplegia due to cord compression, lymph node enlargement, lower limb lymphedema, and loin pain while lethargy, cachexia, and hemorrhage may indicate significant systemic metastases[97].

1.6.2.4.2 Digital Rectal Examination (DRE)

Digital rectal examination (DRE) is an inexpensive method of prostate cancer detection which has been the focus of many clinical studies[98-103]. One problem with the test is that it is subjective and consequently depends on the experience of the examiner. Another is that several other conditions can lead to a false-positive DRE finding, including BPH, prostatitis, prostatic calculi, ejaculatory duct anomaly, seminal vesicle anomaly, and rectal wall phlebolith or polyp/tumor. Early stages of prostate
cancer (T2a) are characterized by a firm peripheral nodule that does not distort the capsule, while more advanced cancers feel hard and more diffuse. T3 stage tumors often present an altered prostate contour while retaining movement of the gland as a whole contrasted with the fixed, immobile presentation of T4 stage tumors.

1.6.2.4.3 Prostate Specific Antigen (PSA)

Prostate-specific antigen is a 34 kD glycoprotein specifically found in prostate epithelium. It is a neutral serine protease designed to lyse seminal-vesicle protein. A small percentage of PSA normally escapes the prostatic ducts and enters the bloodstream where it exists bound mainly to the proteins alpha-1-antichymotrypsin (ACT) and alpha macroglobulin (αMG), leaving a small proportion of free PSA in the serum. Prostate-specific antigen has established utility for the immunohistochemical identification of metastatic disease of prostatic origin, for monitoring of “biochemical recurrence” after therapy and for assessment of disease status in men who are at high risk for biopsy complications.

Screening measures for serum PSA levels have increased the detection rate of early-stage prostate cancer and are thought to be in part responsible for the downward stage migration trend seen in the disease. Considerable variability exists in the world of PSA testing. The cutoff for normal total PSA is accepted to be 4.0 ng/mL though some evidence suggests lowering this cutoff in at risk populations. While most clinical assays measure total PSA (bound + free) a significant advantage is afforded when an additional test for free PSA is performed. Strong evidence exists that PSA complexed with ACT increases in prostatic carcinoma[104, 105] and the lack of availability of a test to
specifically measure serum ACT-complexed PSA led to the use of percent free-to-total PSA ratio to approximate complexed PSA[106]. Such ratios proved to be especially useful in the population of men with total PSA values in the ‘gray zone’ of 2.5 to 10 ng/ml[107]. Recent development of a reliable assay for ACT-PSA complex[108] looks promising and may outperform both total PSA and free-to-total PSA ratio as a more specific analyte for cancer[109]. Other methods to improve PSA performance that have been studied include PSA density[110, 111], transitional zone density[112], PSA velocity[113, 114], and age-specific PSA[115].

1.6.2.4.4 Diagnostic Imaging

Transrectal ultrasound imaging (TRUS) produces high-resolution images of the prostate which are useful for assessing extent of tumor involvement and extension as well for guiding needle biopsies to sample areas suspected of harboring tumor foci. Prostate cancers are frequently hypoechoic on TRUS, but can also be isoechoic and more rarely hyperechoic[116]. Characteristics of prostate cancer that can be evaluated by TRUS include asymmetry of prostate size, shape, indefinite differentiation between the central and peripheral zones, and bulging or disruption of the capsule. Advances in color Doppler TRUS allowing analysis of abnormal blood flow look promising for the identification of hypervascular regions in the peripheral zone[117]. Computed tomography (CT) scanning is useful in metastatic disease to identify the presence of lymphadenopathy in the pelvis and is suggested only when other factors identify risk of tumor spread (i.e. PSA>20ng/mL and Gleason grade > 7)[118]. Advances in Magnetic resonance (MR) imaging endorectal coil design[119] have allowed the acquisition of
high-resolution differentially weighted MR images of prostatic disease that are probably the most accurate technique currently available for assessing the extent of tumor involvement. Additionally, dynamic contrast enhanced MR imaging may provide tumor angiogenesis information[120].

1.6.2.5 Biopsy Interpretation and Grading of Prostatic Adenocarcinoma

The definitive diagnosis of prostatic adenocarcinoma involves the cytological and histological confirmation of the established criteria of malignancy. The diagnostic criteria for carcinomas in biopsies of the prostate involve both architectural and cytologic findings[121]. Low to medium power analysis of the arrangement of the glandular acini is useful and is the basis of the Gleason scale for grading prostatic adenocarcinoma, the predominant scoring system used in the United States[122]. Malignant acini are typically scattered haphazardly in the stroma either singly or in clusters. The acini in cancer are typically small to medium sized with contours that are less smooth than adjacent normal and hyperplastic acini. Cytologic abnormalities in adenocarcinoma include nuclear and nucleolar enlargement present in a majority of malignant cells. Nucleolar size greater than 1.5 mm suggests malignancy while identification of two or more nucleoli in a single cell is virtually diagnostic of malignancy[123].

1.6.2.5.1 Gleason Grading System

The Gleason Grading system is the most widely used system for grading prostatic adenocarcinoma. It relies heavily on the examination of low power architectural features of the arrangement of prostatic acini. The Gleason scale rates glandular patterns of proliferation on a scale of 1 (most differentiated) to 5 (least differentiated). Most prostate
cancers contain more than one of these patterns and thus the Gleason score for a biopsy interpretation is reported as the combination of the two most prominent patterns. Scores range from 2-10 and should be reported as the composite score and its component patterns with the most prevalent pattern listed first[124]. For example a biopsy sample with a predominant pattern of 3 and a secondary pattern of two would be reported as 3+2=5. In practice most cancers have at least one score of 3, and the score of 1 is rarely used.

Gleason grade 1 architecture is described as very well differentiated and is minimally distorted. Neoplastic glands are round, closely packed, single, separate, uniform in shape and diameter, and are sharply delineated from fibrovascular stroma. Hyperplastic glands also fulfill these criteria, therefore a classification as grade 1 adenocarcinoma also requires occasional enlarged nucleoli > 1mm in diameter. In practice a Gleason score of 1 is rarely used. Gleason grade 2 pattern (well differentiated) consists of glands which still exhibit a mild but definite stromal separation between glands with more variation in the shape and size of glands than is seen in grade 1, but less than that of grade 3. Grade 2 tumors remain circumscribed, and definite separation of the malignant glands exists at the tumor periphery suggesting ability to spread to the surrounding stroma. Tumor gland separation is usually less than one average gland diameter. Gleason grade 3 cancers exhibit more extreme variation in size, shape, and separation than grade 2 and are typically spaced more than one average gland diameter apart. The cytoplasm of grade 3 tumor cells tends to be more basophilic than lower grade cancers and nuclei are variable but still larger than lower grades and almost always contain prominent nucleoli. Gleason grade 4 cancers may exhibit any of 4 different
morphologic patterns. Glands with a cribiform pattern have large masses of tumor cells punctuated by sieve-like spaces. Such a pattern was classified as grade 3 by Gleason, however, subsequent reclassification to grade 4 was based on the conclusion that most, if not all examples of cribiform carcinoma are equivalent to grade 4 carcinoma growing within preexisting lumina[125]. The distinctive feature of grade 4 tumors is ragged and invading edges in contrast to the smooth edges of grade 3. Other architectural variants of grade 4 adenocarcinoma include solid, microacinar, and papillary. Gleason grade 5 tumors completely lack glandular differentiation. Such tumors can be arranged in solid masses, cords, trabeculae, sheets, or may appear as single cells infiltrating the stroma.

1.6.2.5.2 Importance of Histologic Grading

Cancer grade at time of diagnosis has been investigated extensively for correlations with other tumor characteristics and clinical behavior. Every measure of survival and recurrence is strongly correlated with cancer grade. These measures include crude survival, tumor-free survival after treatment, metastasis-free survival, and cause-specific survival. Such correlation has been described and validated in numerous studies[126-129]. Age-adjusted, fifteen-year, cancer-specific mortality rates for men with Gleason scores of 2-4, 5, 6, 7, and 8-10 are 4-7%, 6-11%, 18-30%, 42-70%, and 60-87% respectively[130]. Tumor volume has been correlated with histologic grade in both transurethral and radical prostatectomy specimens. A study by McNeal showed that in Gleason grade 4 and 5 tumors, 22 of 38 tumors >3.2 cm³ had tumor-positive nodes while positive nodes were present in only 1 out of 171 tumors <3.2 cm³. Two studies
independently confirmed that the strongest predictor of progression of poorly differentiated cancer is tumor volume[129, 131].

Other studies have found correlations between Gleason grade and PSA levels[132]. Gleason grade is also one of the strongest and most useful predictors of pathologic stage in many studies including the progression of capsular perforation, seminal vesicle invasion, and lymph node and bone metastases and can be correlated with expression levels of MIB-1 (Ki-57), a tissue marker for proliferation[133-136].

1.6.2.6 Staging of Prostatic Adenocarcinoma

Accurate assessment of the clinical stage of prostatic adenocarcinoma is important for the estimation of prognosis, selection of treatment, and evaluation of therapeutic results. The Tumor Node Metastasis (TNM) staging system is used to stage prostatic adenocarcinoma. The current TNM clinical staging is shown below in tables 1.2 and 1.3.

<table>
<thead>
<tr>
<th>TX</th>
<th>Primary tumor cannot be assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0</td>
<td>No evidence of primary tumor</td>
</tr>
<tr>
<td>T1</td>
<td>Clinically inapparent tumor not palpable or visible by imaging</td>
</tr>
<tr>
<td>T1a</td>
<td>Tumor incidental histological finding in 5% or less of tissue resected</td>
</tr>
<tr>
<td>T1b</td>
<td>Tumor incidental histological finding in more than 5% of tissue resected</td>
</tr>
<tr>
<td>T1c</td>
<td>Tumor identified by needle biopsy. Nonpalpable, not visible in imaging.</td>
</tr>
<tr>
<td>T2</td>
<td>Tumor confined within the prostate</td>
</tr>
<tr>
<td>T2a</td>
<td>Tumor involves one lobe</td>
</tr>
<tr>
<td>T2b</td>
<td>Tumor involves both lobes</td>
</tr>
<tr>
<td>T3</td>
<td>Tumor extends through the prostate capsules</td>
</tr>
<tr>
<td>T3a</td>
<td>Unilateral extracapsular extension</td>
</tr>
<tr>
<td>T3b</td>
<td>Bilateral extracapsular extension</td>
</tr>
<tr>
<td>T3c</td>
<td>Tumor invades the seminal vesicle(s)</td>
</tr>
<tr>
<td>T4</td>
<td>Tumor invades any of bladder neck, external sphincter, or rectum</td>
</tr>
<tr>
<td>T4a</td>
<td>Tumor invades any of bladder neck, external sphincter, or rectum</td>
</tr>
<tr>
<td>T4b</td>
<td>Tumor invades levator muscles and/or the pelvic wall</td>
</tr>
</tbody>
</table>

adapted from [69]

Table 1.2 - Staging of primary tumor (T)
<table>
<thead>
<tr>
<th>N</th>
<th>Staging of regional lymph node involvement (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NX</td>
<td>Regional lymph nodes cannot be assessed</td>
</tr>
<tr>
<td>N0</td>
<td>No regional lymph node metastasis</td>
</tr>
<tr>
<td>N1</td>
<td>Metastasis in regional node(s)</td>
</tr>
</tbody>
</table>

adapted from [69]
Chapter Two - Methods

2.1 Tissue microarrays

Tissue microarray technology provides a platform for the high throughput analysis of tissue specimens in research[137]. They are used for the target verification of cDNA microarray results[138], expression profiling of tumors and tissues[139], as well as epidemiology based investigations. Well-designed tissue arrays reduce the variability of experiments performed in a repetitive fashion on large populations, and provide consistent sample-to-sample preparation.

There are currently no reported studies applying vibrational spectroscopic imaging techniques to the analysis of tissue microarray specimens. The tissue microarray is an attractive sample platform for pathological spectroscopic imaging approaches for several reasons. First, tissue arrays can be constructed from archival material, allowing for large sample populations representative of normal tissue and disease processes to be examined. Second, tissue microarrays provide consistent sample preparation across a large sample population, minimizing sample-to-sample data variation. Finally, serial sections of tissue microarrays can be analyzed with other techniques to provide complementary information invaluable to the interpretation of spectroscopic imaging results.

2.1.1 Construction of Prostate Tissue Microarrays

Sections from three prostate tissue microarrays constructed in the Tissue Array Research Program Laboratory, Laboratory of Pathology, Center for Cancer Research, of
the National Cancer Institute by Dr. Stephen M. Hewitt were used as samples for the experiments in this study. The tissue array donor material was obtained from formalin-fixed paraffin-embedded blocks from radical prostatectomy specimens taken from cases of confirmed prostate adenocarcinoma from specimens obtained from the Cooperative Human Tissue Network (CHTN) with approval of the appropriate Institutional Review Boards or Office of Human Research Subjects. The tissue arrays were constructed with a 0.6 mm needles[139]. The arrays were constructed using a Beecher Instruments (Silver Spring, MD) ATA-27 Automated Tissue Arrayer.

For sake of clarity, the arrays will be referred to by the respective patient populations used for their construction. Specific details regarding the layout of Array P-16, Array P-40 and Array P-80 appear in the sections below.

2.1.2 Array P-16 Design

Array P-16 was constructed using donor tissue from a population of 16 patients with confirmed prostate adenocarcinoma. Eight unmapped 0.6 mm cores from each patient were used for a maximum spot number of 128 spots/section. Donor core locations were determined by examination of H&E stained sections of the donor blocks and were chosen to provide a representative sampling of both normal prostate histology and pathology from each patient.

2.1.3 Array P-40 Design

Array P-40 was constructed from donor tissue from a population of 40 patients that included the set 16 patients used in the construction of Array P-16. Five unmapped 0.6 mm cores from each of the forty patients were used for a maximum spot number of 200
spots/section. Donor core locations were chosen from locations representative of both adenocarcinoma and benign epithelium.

2.1.4 Array P-80 Design

Array P-80 was constructed of donor tissue from a population of 79 patients with confirmed adenocarcinoma. Two mapped 0.6 mm cores were used from each patient for a maximum spot number of 160 spots/section. H&E-stained sections of the donor tissue blocks were used as a guide to carefully select tissue from a region of adenocarcinoma for one core and benign epithelium for the corresponding core. Figure 2.1 below contains an image of an H&E stained section of Array P-80 and a corresponding schematic representation of the core layout.
### 2.2 Tissue Array Section preparation

#### 2.2.1 Optical Substrates for Tissue Array Sections

Standard optical materials, such as those found in microscope slides, are generally composed of glass, quartz or fused silica. These materials all absorb radiation in the infrared region at wavelengths longer than 2 µm. For this reason, transmission experiments in the mid-IR require the use of alternative optical materials. Several different halide salts are commonly used as optical materials for IR spectroscopy and each possess different optical and physical properties[42].

---

**Figure 2.1 - Array P-80 Layout.**

The right panel contains a visible optical image of an H&E stained section of array P-80. A schematic representation of the core layout appears on the left with patient numbers.
Tissue array sections intended for IR imaging experiments were mounted on 3 mm-thick, polished, barium fluoride (BaF$_2$) optical windows. Barium fluoride is transparent from 0.15-12.5 µm, which covers the visible and the entire spectral range of the FT-IR instrument. Additionally, BaF$_2$ optical elements have the lowest solubility in water (0.17 gm/100 gm water at 23 °C) of materials with similar optical characteristics[2].

2.2.2 Deparaffinization

Histology grade, low melt (58-62 °C) paraffin was removed from tissue array sections by covering the tissue surface with hexane for 5 minutes. The samples were rinsed with hexane several times and deparaffinization was continued by immersion in hexane at 40°C with continuous stirring for 48 hours. Every 3-4 hours during the deparaffinization process, the immersion vessel was emptied, rinsed thoroughly with acetone followed by hexane. Once dry, the vessel was filled with fresh neat hexane to promote flow of embedded paraffin from the tissue. Thorough deparaffinization was assured by monitoring the disappearance of the paraffin band at 1462 cm$^{-1}$ at several sites on the tissue arrays.

2.2.3 Optical imaging of H&E sections

Tissue array sections contiguous with those used for IR imaging analysis were mounted on glass slides and stained with hematoxylin and eosin for traditional histopathological analysis. H&E stained tissue microarray sections were optically imaged using an Olympus BH-2 microscope equipped with a high resolution (3 megapixel), Peltier cooled, 10-bit, Q-imaging micropublisher digital camera. Tissue array spots were imaged individually through a 4x Olympus ∞-corrected microscope
objective and 10x camera eyepiece objective. The H&E sections were reviewed with a pathologist for diagnostic features on a two-headed teaching microscope.

2.3 Spectroscopic Imaging Instrumentation

FT-IR spectroscopic imaging was performed on a Perkin-Elmer (Shelton, CT) Spectrum Spotlight 300 imaging microspectrometer equipped with dual mode detection system. The imaging system is comprised of two main optical components: the spectrometer and microscope. The spectrometer houses a ceramic globar broadband infrared source, a continuous-scanning Michelson interferometer, and a macro sampling area for non-microscopic single point FT-IR measurements. The modulated infrared output beam of the spectrometer is coupled to an infrared microscope and focused onto the sample using Cassegrain optics.

In transmission mode, the infrared beam is focused onto the sample through a Cassegrain condenser. The condenser position can be varied along the beam axis to correct for optical effects caused by substrate thickness and refractive index. Transmitted radiation is collected by a Cassegrain objective and focused onto one of two mercury cadmium telluride (MCT) detectors. Figure 2.1 shows a detailed diagram of the microscope and the optical path in transmission mode.

A traditional single mercury cadmium telluride (MCT) detector is used in the instrument’s point mode to take single-point spectroscopic measurements. Seamless software control of variable-width, rotating knife-edge apertures, and a motorized mapping stage with a precision error < 1 µm allow flexible collection of high quality infrared spectroscopic mapping data in point detection mode. The microscope portion of
the instrument features a visible LED light source and video camera that are linked with control software for automated collection of visible-light images as shown in figure 2.1. Self-referenced stage position is dynamically linked with both captured visible images and spectroscopic imaging results. This feature allows the operator to choose sample areas for spectroscopic imaging experiments via a simple interface by selecting a rectangular area on the displayed optical visible image of the sample, and provides interactive registration of infrared spectroscopic imaging results with corresponding optical visible images.

The spotlight 300’s image mode utilizes a 16-element MCT linear array detector to build infrared spectroscopic images of any designated rectangular sample area in a line-mapping fashion. A fixed optical zoom allows the instrument to collect image data at two different spatial resolutions. The effective pixel size of these two resolution are 25 x 25 µm in low resolution and 6.25 x 6.25 µm at high resolution.
The 1 GB of RAM in the controlling computer limits the size of a single line-mapping image cube acquisition in the imaging mode. The maximum sample area size that can be collected is thus a function of several collection parameters including spatial resolution (high or low), spectral resolution, and spectral wavelength range. Practical considerations such as the liquid nitrogen dewar hold time of 7 hr can also limit the maximum size of image data collection in practice.

2.3.1 Tissue Array FT-IR Data Collection Parameters

IR Spectroscopic images of the tissue array spots were collected in transmission configuration in image mode at the high-resolution zoom setting (pixel size of 6.25µm). 1641 data points were collected across the spectral region from 4000-720 cm⁻¹ yielding
spectra with a resolution of 4 cm\(^{-1}\) (2 cm\(^{-1}\) data point interval). Four interferograms were co-added for each individual measurement to increase data signal-to-noise ratios (SNRs). Background spectra consisting of 190 coadded interferograms were collected from nearby locations on the BaF\(_2\) flats between the tissue spots.

Data collection with these parameters for a typical 600 \(\mu\)m tissue array spot results in a spectroscopic imaging data set with spatial dimensions of \(~115 \times 115\) pixels and a file size of approximately 85 MB. Acquisition time for a typical tissue array spot was approximately 35-40 min. The average SNR for a single pixel absorbance spectrum of tissue was \(>500:1\).

2.3.2 Modifications and Environmental Considerations

The microscope and spectrometer assemblies were enclosed in a Plexiglas housing to enable efficient purging with dry nitrogen gas to remove water vapor and to eliminate air currents. The computer controlling the system was situated outside the housing and the exhaust streams from the cooling fans of the spectrometer (source) and microscope (detector electronics) were vented out of the housing to maintain a stable room temperature atmosphere within the housing during data collection. Once the sample was placed on the stage, all positioning, focusing, and experimental control could be performed remotely by computer control without opening the housing to the atmosphere. After opening the housing for any reason, 20 minutes were allowed for atmospheric equilibration before spectroscopic measurements were resumed.
2.4 Data Handling and Computational Considerations

2.4.1 Data Pre-Processing

In its imaging mode, the Spectrum Spotlight 300 makes use of the dead time while the microscope stage is stepped to a new position to perform several computational tasks. The functions include interferogram apodization, fast Fourier transform of collected data to single beam spectra, and ratioing of sample spectra to background spectra to provide absorbance spectra. Spectroscopic imaging data of tissue array spots were collected individually or in small contiguous groups, checked for spectral quality (SNR, baseline fluctuations, etc.), and corrected for atmospheric water vapor and carbon dioxide using Perkin Elmer proprietary software.

The resulting, atmosphere-corrected, spectroscopic images were imported into ENVI (RSI Inc., Boulder, CO) using software written in IDL by Dr. Rohit Bhargava; all subsequent image processing was performed in this software environment. Some downstream statistical analyses and chart plotting were performed using Microsoft Excel and Origin. All processing was carried out computers equipped with 1.7 GHz Intel Pentium 4 processors and a minimum of 1 GB of RAM.

Individual tissue array spots were mosaicked into one large spectroscopic image dataset for each individual array section for further processing. For Array P-16, the final size of the whole-array spectroscopic image was ~ 500 x 3680 pixels (or ~1.8 million individual spectra) producing a file size of ~14 GB. Spectroscopic image datasets of the two sections of Array P-40 were ~ 4370 x 550 pixels or (or ~2.4 million individual
spectra) with a file size of ~17 GB. Array P-80 had a final size of 2160x1250 pixels (or ~2.7 million spectra) with a file size of ~18.5 GB.

2.4.2 Spectral Baseline Correction

Every infrared absorbance spectrum in the image scene was individually baseline corrected using custom-designed routines written in IDL by Dr. Rohit Bhargava. Regression is used to calculate the values that lie on the line-segment intersecting each pair of points. These values are subsequently subtracted from the spectral absorbance at the corresponding frequency, and the process is repeated for each spectrum in the image scene. Several hundred average spectra from different tissue regions on multiple spots of Array P-16 were compared and frequency positions observed to be consistent local minima were chosen as baseline points. A list of the frequency positions used as spectral baseline points appears in Table 2.1.
Table 2.1 - Spectral frequencies used for spectroscopic baseline correction

<table>
<thead>
<tr>
<th>spectral baseline points (cm⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3774</td>
</tr>
<tr>
<td>3682</td>
</tr>
<tr>
<td>3000</td>
</tr>
<tr>
<td>2708</td>
</tr>
<tr>
<td>2644</td>
</tr>
<tr>
<td>2542</td>
</tr>
<tr>
<td>2392</td>
</tr>
<tr>
<td>2282</td>
</tr>
<tr>
<td>1984</td>
</tr>
<tr>
<td>1764</td>
</tr>
<tr>
<td>1478</td>
</tr>
<tr>
<td>1352</td>
</tr>
<tr>
<td>1328</td>
</tr>
<tr>
<td>1296</td>
</tr>
<tr>
<td>1184</td>
</tr>
<tr>
<td>1144</td>
</tr>
<tr>
<td>982</td>
</tr>
<tr>
<td>948</td>
</tr>
</tbody>
</table>

The baseline-corrected absorbance intensity of the N-H stretching protein backbone vibration (or Amide A) at 3290 cm⁻¹ was used to differentiate tissue from empty space on the array. All pixels with an absorbance less than 0.08 at 3290 cm⁻¹ were masked to zero for all spectral data points and disregarded during any subsequent processing.
Chapter Three - Infrared Spectroscopic Histology of Prostate

3.1 Visualization of Spectral Images and Verification of Histologic Features

Infrared spectroscopic imaging datasets of prostate tissue microarray sections were initially visualized by plotting images of the baseline-corrected absorbance at 3290 cm\(^{-1}\). This wavenumber position corresponds to the N-H stretching absorbance band or Amide A absorbance, a backbone vibration found in all proteins. Since proteins are basic structural elements of all prostate tissue, Amide A absorbance images are useful for verifying the presence of spots and structural correlation of features with visible optical images of the corresponding H&E stained section. The baseline corrected Amide A absorbance images for 4 tissue array spots from a single patient are shown in fig 3.1A along with a corresponding H&E stained consecutive section in Fig 3.1B.
The tissue microarray sections used for IR spectroscopic imaging experiments are subject to harsh deparaffinization conditions of immersion in hexane at 40ºC for 4 hours. These conditions caused artifactual damage to a handful of spots in each array sections. Typical artifactual problems included partial or complete absence of spots, spots that folded over onto themselves, and spots which were partially detached from the surface of
the optical flat. N-H stretching absorbance images such as those seen in figure 3.1A were extremely useful for discovering spots that were subject to such damage so that they could be eliminated from further analysis.

3.2 Creation of Ground Truth Data Regions of Interest

In order to analyze spectra and to train and test classification models, ground truth data for different histological features or classes needed to be established. The name ground truth stems from remote sensing applications where field data from various sources on the ground are acquired and registered with image data to enable class training and/or evaluation of classification performance[61].

A pathologist examined the matching H&E stained tissue array sections microscopically and different histological features present in each spot were marked on optical images of the corresponding H&E stained sections. The region of interest (ROI) tool in ENVI allows the user to designate a collection of pixels as belonging to a set, or ROI. ROIs can be manually generated by selecting geometric areas on the spectroscopic images with drawing tools such as rectangles, ellipses, or polygons. Pixels may be added to or deleted from ROIs individually, allowing the user to carefully edit such groups. ROIs can also be generated from parameters of the data itself, which can be particularly useful. Once created, these ROIs can be used in a variety of image analysis operations from image subsetting and masking to statistical analyses and image classification.

In analyzing the spectroscopic datasets, specific images derived from various absorbance band ratios provided high contrast for discerning different histologic features in the tissue. Fig 3.2A shows the 1080 cm\(^{-1}\)/1544 cm\(^{-1}\) absorbance band ratio image of
four tissue array spots from a single patient on Array P16. The 1080 cm\(^{-1}\) band is attributed to a C-O stretching vibration of glycogen and the band at 1544 cm\(^{-1}\) to the Amide II vibration of the protein backbone. The 1080 cm\(^{-1}\)/1544 cm\(^{-1}\) image provides high contrast between prostate epithelium and stroma. Areas of higher ratio intensity in Fig 3.2A correspond to the basophilic-staining epithelial regions in the optical image of the corresponding H&E stained section in panel B. The eosinophilic stromal regions of the tissue correspond to lower intensity regions of the 1080 cm\(^{-1}\)/1544 cm\(^{-1}\) ratio image suggesting that glycogen/protein levels are higher in epithelial tissue than in stroma.

Another absorbance band ratio that produced useful images was 1206 cm\(^{-1}\)/1544 cm\(^{-1}\). At the spectral resolution used of 4 cm\(^{-1}\), the absorbance feature at 1206 cm\(^{-1}\) typically appears as a shoulder off the higher intensity combination band at 1236 cm\(^{-1}\) attributed to both Amide III vibrational mode of proteins and the asymmetric stretch of phosphodiester (PO\(_2^\sim\)) groups in phospholipids and nucleic acids. Fig 3.2C shows the 1206 cm\(^{-1}\)/1544 cm\(^{-1}\) absorbance band ratio image of 4 tissue array spots taken from a different patient on Array P-16. Comparison with the image of the matching H&E stained section (Fig 3.2D) reveals poor contrast between epithelial and stromal tissues, however, excellent contrast is seen between an area of lymphocytic infiltration, indicated by the highest intensity area in the upper spot, and the surrounding stromal and epithelial components.
Various absorbance band images and band ratio images were interactively overlaid and used to assist the ROI creation process. Using the pathologist-reviewed, marked optical images of the H&E stained sections as a guide, collections of pixels in the spectroscopic image of each tissue spot were assigned to one of the ten histological class ROIs listed in table 3.1. The epithelial class includes pixels from different histopathological states, including normal benign epithelium, benign prostatic hyperplasia (BPH), prostatic intraepithelial neoplasia (PIN), and prostatic adenocarcinoma (CaP).
Stromal histological features were separated into 3 subclasses: fibrous stroma, smooth muscular stroma, and mixed stroma based on the H&E section images and spectral differences noted between these three subclasses. Remaining classes included sites of lymphocytic infiltration, vessel endothelium and muscular coat, peripheral nerve tissue, ganglion cells, blood cells, and corpora amylacea. In making the component analysis, much care was taken to include only those pixels that were definitively representative of a particular class, and therefore pixels near edges or class borders were eliminated to insure that class spectral statistics remain uncontaminated.

<table>
<thead>
<tr>
<th>Histologic Class</th>
<th>number of spectra in class ROI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16 patient array</td>
</tr>
<tr>
<td>epithelial tissue</td>
<td>80293</td>
</tr>
<tr>
<td>fibrous stroma</td>
<td>11444</td>
</tr>
<tr>
<td>mixed stroma</td>
<td>74609</td>
</tr>
<tr>
<td>smooth muscle stroma</td>
<td>2751</td>
</tr>
<tr>
<td>lymphocytes</td>
<td>1976</td>
</tr>
<tr>
<td>endothelium</td>
<td>359</td>
</tr>
<tr>
<td>peripheral nerve</td>
<td>2362</td>
</tr>
<tr>
<td>ganglion cells</td>
<td>438</td>
</tr>
<tr>
<td>blood cells</td>
<td>628</td>
</tr>
<tr>
<td>corpora amylacea</td>
<td>1039</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>162956</strong></td>
</tr>
</tbody>
</table>

Table 3.1 - Histologic class population data

Class data were stored separately for each spot and histologic class as individual regions of interest (ROI) in ENVI and could be operated on individually at the spot level or merged to patient level or into a single ROI at the class level. This flexibility enables downstream comparisons to be made at the spot-spot and patient-patient level for each class and across classes.
3.3 Spectral analysis of histologic features and metric selection

The individual ROIs from each spot were merged together to form a single large ROI for each of the ten histologic classes for each array. The total number of pixels, where each pixel represents an individual spectrum, is shown for each histologic class in table 3.1. The spectra from each ROI were averaged to create a mean spectrum for each class, displayed in figure 3.3.
Figure 3.3 - Histologic class mean spectra

The spectra were calculated from baseline corrected spectra and were normalized to amide II absorbance at 1544 cm$^{-1}$. Panel A contains the full spectral window collected 720-4000 cm$^{-1}$. Panels B, C, and D contain enlargements of the corresponding boxes in panel A.
3.4 Construction of a Supervised Classification Model for Prostate Histology

3.4.1 Spectral Data Reduction

The mean spectra for each histologic class were compared and spectral features, frequencies, and band ratios could be identified for distinguishing the various classes from one another. A set of metrics was developed involving absorbance band ratios and peak centers of gravity for features across the entire spectral region. Metric values were computed using software routines written in the statistical language IDL by Dr. Rohit Bhargava and implemented in the remote sensing software environment ENVI (RSI, inc., Boulder, CO).

Histograms of each training class population were plotted and compared for each metric. Most distributions approximated a normal distribution and showed some variation in mean and standard deviation between classes. Metrics which did not approximate a normal distribution for most classes were discarded, since such data can lead to poor performance with parametric classification methods, particularly with Gaussian Maximum Likelihood classification algorithms[62] discussed below in section 3.4.2. Metrics that showed no significant variation between classes were also discarded, since their inclusion would likely add only noise to the classification. The spectroscopic imaging dataset was reduced from 1641 spectral bands (wavenumber positions) to a 20-band set of candidate spectral metrics, reducing the tissue array imaging dataset from 14 GB to a manageable 160 MB.

The construction of successful classification model is by nature an interactive, process. Information is gained in small bits as individual problems are identified and
strategies are altered to adjust. A common problem encountered is the existence of classes which possess bimodal distributions in several spectral bands. Such observations typically indicate that the class is composed of two or more spectrally distinct subclasses. In such cases, classification accuracy can often be dramatically improved by splitting the training data for the suspect class into separate classes[63]. Similar histogram analysis performed on several absorbance band ratio images from early FT-IR imaging studies of non-array prostate tissue indicated that stromal tissue in the prostate was composed of spectrally distinct subclasses. These preliminary results formed the basis for splitting stroma into three separate subclasses: fibrous stroma, smooth muscular stroma, and mixed fibromuscular stroma.

A listing of the parameters for each of the 20 candidate spectral metrics appears below in Table 3.2.
Several different algorithms exist for the supervised classification of multispectral image data. Some of the more simplistic classification algorithms such as parallelepiped or minimum-distance approaches do not consider variation that may be present within spectral classes and do not perform well when frequency distributions from separate classes overlap[62].

Histogram analysis of individual metric value class distributions indicated that both significant intraclass variation in spectral metric values exist and that significant overlap between metric value frequency distributions of different classes was common. As examples, individual class histograms for the three most common or populated training
classes (epithelium, mixed stroma, and fibrous stroma) are displayed for metric 02 values (Fig. 3.4A) and for metric 11 values (Fig 3.4B).

A parametric approach to supervised classification that is particularly well suited to deal with such natural intraclass spectral variation and interclass overlap of metric frequency distributions is the Gaussian Maximum Likelihood (GML) Classifier[62]. An
$n$-dimensional probability surface for each class is generated from both class mean and variance statistics for training data consisting of $n$ spectral bands. As classification ensues, each pixel’s discrete spectrum can be used to calculate the corresponding conditional probability or likelihood that the pixel belongs to each class separately from the individual class $n$-dimensional probability surfaces[63]. The pixel is then assigned to the class with the highest conditional probability. Classification sensitivity can be adjusted by imposing minimum probability thresholds that cause pixels below a user-supplied minimum conditional probability to be relabeled as unclassified.

A supervised Gaussian Maximum Likelihood (GML) algorithm implemented in ENVI was used to classify the 20 metric dataset of the entire tissue array. The 10 different histologic class ROIs were used as input to train the classifier. No thresholding was imposed during the classification forcing each pixel in the tissue array image scene to be classified as one of the ten histologic subtypes. The 10 histologic training ROIs were used next as a preliminary validation set to evaluate the performance of the classification.

3.4.3 Array P-16, 20-metric, GML Self-classification results

An extremely useful tool for the evaluation of image classification results is the expression of classification accuracy in terms of an error matrix. Such error matrices are also commonly referred to (appropriately) as confusion matrices. Error matrices compare, on a class-by-class basis, the relationship between known reference data (ground truth class data) and the corresponding results of a classification attempt.
The same 10 histology class ROIs from Array P-16 that were used as training input for the 20-metric, Gaussian Maximum Likelihood (GML) classification of all tissue on Array P-16 were used as ground truth to calculate an error matrix for the same classification. These data appear below as Table 3.3.

<table>
<thead>
<tr>
<th>Ground Truth Class</th>
<th>Result of Classification</th>
<th>EPITHELIUM</th>
<th>MIXED STROMA</th>
<th>FIBROUS STROMA</th>
<th>SMOOTH MUSCLE</th>
<th>CORPORAAMYLACEA</th>
<th>ENDOTHELIUM</th>
<th>NERVE</th>
<th>GANGLION</th>
<th>BLOOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPITHELIUM</td>
<td>89.64</td>
<td>0.18</td>
<td>0.11</td>
<td>0.00</td>
<td>0.19</td>
<td>2.07</td>
<td>0.28</td>
<td>0.17</td>
<td>0.46</td>
<td>0.00</td>
</tr>
<tr>
<td>MIXED STROMA</td>
<td>0.00</td>
<td>70.42</td>
<td>0.13</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>2.88</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
</tr>
<tr>
<td>FIBROUS STROMA</td>
<td>0.25</td>
<td>4.82</td>
<td>94.19</td>
<td>0.69</td>
<td>0.00</td>
<td>0.00</td>
<td>3.56</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SMOOTH MUSCLE</td>
<td>0.00</td>
<td>23.18</td>
<td>0.73</td>
<td>96.36</td>
<td>0.00</td>
<td>0.00</td>
<td>3.90</td>
<td>0.04</td>
<td>0.00</td>
<td>0.64</td>
</tr>
<tr>
<td>CORPORA AMYLACEA</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>99.81</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LYMPHOCYTES</td>
<td>3.48</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>97.82</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ENDOTHELIUM</td>
<td>0.12</td>
<td>1.16</td>
<td>0.85</td>
<td>2.29</td>
<td>0.00</td>
<td>0.00</td>
<td>94.15</td>
<td>0.42</td>
<td>0.46</td>
<td>11.46</td>
</tr>
<tr>
<td>NERVE</td>
<td>0.79</td>
<td>0.03</td>
<td>3.92</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.56</td>
<td>94.12</td>
<td>1.37</td>
<td>0.00</td>
</tr>
<tr>
<td>GANGLION</td>
<td>5.68</td>
<td>0.05</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.28</td>
<td>1.69</td>
<td>97.72</td>
<td>0.00</td>
</tr>
<tr>
<td>BLOOD</td>
<td>0.03</td>
<td>0.14</td>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
<td>0.56</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>87.74</td>
</tr>
</tbody>
</table>

Table 3.3 - Error Matrix of supervised GML Classification results using 20 spectroscopic metrics. The classifier was implemented in ENVI and was trained on sets of reference spectra assigned to one of ten histologic classes. All matrix values are given in units of percent of ground truth class pixels.

The columns represent the ground truth or correct class designation and the rows represent the result class as assigned by the GML classifier. The numbers at each position are the percent of the number of total pixels in the column (ground truth class) that were classified as the class of the row. For example, if we examine the epithelium column, we see that 89.64% of epithelial pixels were correctly classified, 0.25% of epithelial pixels were misclassified as fibrous stroma, 3.48% of epithelial pixels were
misclassified as lymphocytes, etc. The values that occupy the diagonal of the confusion
matrix (shown in red in Table 3.3) are the classification accuracy for a given class. These
values show that this initial classification attempt performs above 94% for all classes
except for epithelium (89.6%), mixed stroma (70.4%), and blood (87.7%).

3.4.4 Leave-one-out metric evaluation

It was clear from the histogram analysis of the individual metrics in the original set of 20 that certain metrics were better for discriminating certain classes than others. In
light of the significant frequency distribution overlaps seen in many cases, a given
metric’s inclusion in the classification attempt might provide little to modest increase in
classification accuracy for single class or small number of classes while causing a
significant decrease in accuracy in the remaining classes. To test for the presence of such
contaminating metrics in the original set of 20, a leave-one-out analysis was performed.
The image scene was reclassified 20 separate times using a total of 19 spectral metrics
per attempt, leaving out a different metric for each successive trial. The accuracy change
for the 3 classes with the worst 20-metric classification accuracy (epithelium, mixed
stroma, and blood) with respect to the 20 metric classification was recorded for each
successive trial and is shown below in Figure 3.5.
Figure 3.5 - Graphical Representation of results of the leave-one-out analysis

The tissue array data was reclassified 20 separate times with a total of 19 metrics, sequentially leaving out a different metric. The percent change in classification accuracy for the three histologic classes which performed poorly in the 20 metric classification attempt (epithelium, mixed stroma, and blood) are plotted with the metric number left out varying along the x-axis.

While the results of the leave one out analysis were analyzed for every class, for the sake of clarity, Figure 3.5 contains only data from the three classes (epithelium, mixed stroma, and blood) which had the most classification error in the original 20-metric classification. These three classes stand to benefit most from the removal of a possible contaminating metric, and from the results in Fig 3.5 we see that two metrics clearly stood out as detrimental to classification accuracy. All three of the poorly classified classes (epithelium, mixed stroma, and blood) show a significant increase in accuracy when metric 9 and metric 18 are left out individually.
3.4.5 Array P-16, 18-metric GML Classification Results

Reclassification with the GML algorithm in the absence of both metric 9 and metric 18 produced the most promising self-evaluated results, which are represented as a error matrix in Table 3.4. The training ROIs were used as ground truth input to generate the error matrix results.

<table>
<thead>
<tr>
<th>Ground Truth Class</th>
<th>Result of Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPITHELIUM</td>
</tr>
<tr>
<td>EPITHELIUM</td>
<td>95.55</td>
</tr>
<tr>
<td>MIXED STROMA</td>
<td>0.00</td>
</tr>
<tr>
<td>FIBROUS STROMA</td>
<td>0.19</td>
</tr>
<tr>
<td>SMOOTH MUSCLE</td>
<td>0.00</td>
</tr>
<tr>
<td>CORPORA AMYLACEA</td>
<td>0.01</td>
</tr>
<tr>
<td>LYMPHOCYTES</td>
<td>0.92</td>
</tr>
<tr>
<td>ENDOTHELUM</td>
<td>0.07</td>
</tr>
<tr>
<td>NERVE</td>
<td>0.48</td>
</tr>
<tr>
<td>GANGLION</td>
<td>2.77</td>
</tr>
<tr>
<td>BLOOD</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3.4 - Confusion matrix of supervised GML Classification attempt using 18 spectroscopic metrics Metric 9 and metric 18 were left out of the original set of 20.

All 10 classes are classified at an accuracy above 92.5%. A color-coded classified result image of four tissue array spots from a single patient are shown in figure 3.6 with the corresponding H&E section in panel B for comparison. Classification correspondence with the histological features observed in the H&E section is outstanding.
Figure 3.6 - Classification results for 2 tissue array spots from the same patient
GML Classification was performed with a total of 18 metrics selected from the results of the leave-one-out analysis as shown in figure 3.5.

Epithelial pixels were classified correctly 95.6% of the time with the majority of misclassification as ganglion (2.8%) and lymphocytes (0.9%). Mixed stroma pixels were classified correctly 92.5% of the time with the majority of misclassification, not surprisingly, as smooth muscle stroma (5.5%) and fibrous stroma (1.15%).

Fibrous stroma pixels were classified correctly 93% of the time with the major misclassification predominately occurring as nerve (4%). Interestingly, nerve pixels were correctly classified with an accuracy of 93.7% with the majority of misclassification as fibrous stroma (3.7%). Upon close inspection, the mean spectra of the fibrous stroma and nerve training ROI’s proved to have many similarities, as seen in figure 3.3. Spectral similarities between nerve and fibrous stroma include absorbance peaks at 1034 cm$^{-1}$ and
1206 cm$^{-1}$, and a shoulder at 1280cm$^{-1}$. As a result of this spectral similarity, a substantial number of pixels at the stromal-epithelial interface were observed to be misclassified as nerve when they probably belong to the fibrous stroma or mixed stroma class.

Smooth muscle stroma pixels were classified with an accuracy of 94% with the majority of misclassification as mixed stroma (2.8%) and endothelium (2%). Again of note is that while endothelium was correctly classified 92.8% of the time, the majority of misclassification occurred as smooth muscle stroma. While fibrous stroma and nerve represent a pair of classes whose similarity seems likely based on a compositional similarity, the connection between endothelium and smooth muscle stroma is probably due to impurity in the endothelial training class. The endothelial training class by far had the fewest number of training spectra at 359. This reflects both the paucity of discernible endothelial tissue visible in prostate sections on H&E staining and the difficulty in correctly identifying it in corresponding IR spectroscopic images. Endothelial cells are typically very hard to identify as they are single-layered, and are contiguous with the smooth muscular media which is more pronounced in arterial vessels.[7] With a single pixel in the IR spectroscopic images representing 6.25 µm of tissue per edge, it seems highly likely some of the endothelial training pixels are contaminated with signal from smooth muscle tissue of the vessel media. Similarly, blood pixels were classified with an accuracy of 97.6% with the majority of misclassification as endothelium.

Lymphocyte pixels were classified with an accuracy of 96.2% with all of the misclassification as epithelial pixels (3.8%). A large proportion of pixels which were incorrectly classified as lymphocytes probably represent true spectral mixtures of different
class types, since lymphocytic infiltration necessarily overlays regions of stroma and epithelial tissue. Ganglion pixels were classified to an impressive 97.3% with the majority of misclassification as nerve. Corpora amylacea were classified to an accuracy of 99.8%. While this accuracy value seems aberrantly high compared with the other classes, examination of the class mean spectrum of corpora amylacea compared with the other class mean spectra (figure 3.3) reveals that it is quite extreme compared with every other spectrum which probably accounts for the high self-classification accuracy.

3.5 Validation of Prostate Histology Classification Model

These impressive results with a simple set of 20 metrics hint at the promise of this approach. One can be certain that many more metrics exist that if included would improve classification accuracy. One of the many advantages of this approach is that we can design our metrics to highlight the property of a spectral feature that is changing across classes, whether it be band height relative to another band or band center of gravity irrespective of height. Metrics which measure other spectral properties such as absorbance band widths are other obvious choices to be tested in the future, while data collection at higher spectral resolution and with higher single-pixel SNRs will uncover newly resolvable spectral features which can be harnessed as metrics to improve classification accuracy.

An important caveat mentioned prominently most remote sensing references [61-63, 141] is that accuracy estimates made using training data regions as ground truth do not necessarily indicate that similar results will be seen when classifying other regions of the image scene. The pixels in the ROI sets used for classifier training and evaluation
make up only a tiny fraction of the total number of tissue pixels in the full spectroscopic image of Array P-16. Several spots from Array P-16 were purposely avoided during the training ROI selection process so that they could be used for qualitative validation of promising classification results. Examination of these spots with respect to their matching H&E stained sections gave a qualitative sense that the 18-metric classification was performing quite well on tissue that was not included in the training sets. As an example, the lower spot in Figure 3.6 contains no pixels used in any of the 10 training ROIs, any the classification results agree well with the image of the matching H&E-stained section.

3.5.1 Cross-Array Validation

As noted in table 3.2, a set of histology ground truth ROIs was constructed for the spectroscopic imaging dataset of Array P-40 in the same manner as described in section 3.2 in reference to Array P-16.

In light of the observed classification trends seen in the 18 metric, P-16 training data error matrix in Table 3.3 and discussed in section 3.5, adjustments were made to the classification model class structure. The endothelial class was discarded due to insufficient ground truth ROI pixel populations on both Array P-16 and Array P-40. The extremely thin nature of this tissue structure on cross-section further adds to the difficulty in both establishing ground truth information for this potential class and evaluating results since pixels in the spectroscopic images have a size of 6.25 μm of tissue per pixel edge. Visual analysis of the H&E-stained section of Array P-40 revealed almost no contiguous areas of pure smooth muscle as seen frequently in Array P-16. Furthermore,
the 18-Metric self-classification results indicated that most of the misclassified mixed stroma pixels were incorrectly classified as smooth muscle stroma and vice versa.

Consequently, the ground truth data smooth muscle stroma class and mixed stroma classes were merged into a single mixed stroma class separately for both Array P-16 and Array P-40. The spectral similarity and commission errors seen between the fibrous stroma and nerve classes suggested they might also be better off combined as a single fibrous-stroma class. However, no appreciable nerve or ganglion tissue was found in any of the Array P-40 spots so both the P-16 nerve and ganglion training data were excluded from the cross-array classification attempt. These adjustments to the histology class structure result in a total of 6 classes. Table 3.5 contains the revised, 6-class, histology ground truth class ROI set population data for both Array P-16 and Array P-40.

<table>
<thead>
<tr>
<th>Histologic Class</th>
<th>number of spectra in class ROI</th>
<th>16 patient array</th>
<th>25 patient array</th>
</tr>
</thead>
<tbody>
<tr>
<td>epithelial tissue</td>
<td></td>
<td>80293</td>
<td>1134</td>
</tr>
<tr>
<td>mixed stroma</td>
<td></td>
<td>77360</td>
<td>30704</td>
</tr>
<tr>
<td>fibrous stroma</td>
<td></td>
<td>11444</td>
<td>19092</td>
</tr>
<tr>
<td>corpora amylacea</td>
<td></td>
<td>1039</td>
<td>828</td>
</tr>
<tr>
<td>lymphocytes</td>
<td></td>
<td>359</td>
<td>1134</td>
</tr>
<tr>
<td>blood cells</td>
<td></td>
<td>628</td>
<td>767</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>162956</td>
<td>153554</td>
</tr>
</tbody>
</table>

Table 3.5 - Revised 6-class histology ground truth ROIs for Array P-16 and Array P-40

The 6 ground truth ROIs for Array P-16 listed in Table 3.5 were used as training data for supervised classification of all tissue from Array P-40. The same 18 metrics used for classification of Array P-16 in section 3.5.5 were used for both the training data from Array P-16 and for P-40 image data to be classified by the GML algorithm.
All pixels in the P-40 image scene were classified and the 6 ground truth class ROIs from Array P-40 were used to construct an error matrix for the cross-array classification result which appears below in Table 3.6.

<table>
<thead>
<tr>
<th>Ground Truth Class</th>
<th>Result of Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPITHELIUM</td>
</tr>
<tr>
<td>EPITHELIUM</td>
<td>95.74</td>
</tr>
<tr>
<td>MIXED STROMA</td>
<td>0.58</td>
</tr>
<tr>
<td>FIBROUS STROMA</td>
<td>1.26</td>
</tr>
<tr>
<td>CORPORA AMYLACEA</td>
<td>0.10</td>
</tr>
<tr>
<td>LYMPHOCYTES</td>
<td>2.30</td>
</tr>
<tr>
<td>BLOOD</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 3.6 - Error Matrix for 6-Class, GML Classification Results

The classifier was trained on 6-class ground truth data from Array P-16 and applied to classify all tissue pixels in the image data from Array P-40. The same set of 18 spectral metrics used in section 3.5.5 were used for this classification.

The error matrix results indicate that classification accuracy in 5 out of 6 classes exceeds 94.5%. Fibrous stroma was the class with the lowest classification accuracy at 91.5%, however, nearly all of such misclassified pixels were incorrectly classified as mixed stroma. This result likely speaks more to the heterogeneity of stroma in general than to any serious problems with the classification itself.
3.6 Conclusions and Further Directions

These results indicate that such a 6-class, supervised GML classification model can be used to successfully segment spectroscopic images of unstained sections of prostate tissue into useful histologic classes based on their spectral properties with respect to spectral class information from a database of previously imaged tissue from a number of patients. Histological class information obtained from such images is useful for image display, however standard staining procedures are far cheaper and provide similar information. FT-IR spectroscopic imaging data analyzed in this fashion can provide histological image information from unstained specimens. Standard staining techniques can interfere with other analytical techniques, such as immunohistochemistry and in situ hybridization, as well as, nucleic acid recovery from laser capture microdissected material[142].

The histological class information obtained could also be used to study morphological relationships, such as epithelial/stromal density ratios in various different states of normal prostate tissue, nodular hyperplasia (BPH), and varying grades of prostatic adenocarcinoma[143, 144]. The supervised classification methods for providing histological class information from IR spectroscopic imaging data developed in the above sections are well-suited for automation, providing a means for rapid evaluation necessary for high throughput analyses.

Furthermore, such histological classifications can be used as a tool for downstream analysis of spectral information from epithelial tissue in an effort to further study the infrared spectroscopic properties of benign prostate epithelial tissue and prostatic adenocarcinoma in many patients. If reliable spectral indicators of disease presence and
progression can be found, then FTIR microspectroscopic imaging techniques can be used as an objective tool to aid in the detection and diagnosis of prostatic adenocarcinoma.

The next section continues with some preliminary experiments using a third tissue array, P-80, designed to investigate some of these issues.
Chapter Four - Infrared Spectroscopic Histopathology of Prostate

4.1 Classification strategy

Array P-80 is the most logical choice as a starting point for the analysis of spectral features of populations of benign and malignant prostate epithelial tissue. Array P-80 was constructed from formalin-fixed, paraffin-embedded tissue blocks cut from radical prostatectomy specimens from population of 80 patients with confirmed prostatic adenocarcinoma. The array was constructed with 2 cores from each patient, one from a region of representative adenocarcinoma, and one from a region with only normal benign epithelium. The intention of the array design was to provide a large patient population and relatively even sampling of benign and malignant tissue for every patient.

The first step of the analysis will apply the histology classification developed in section 3, using the class statistics from the P-16-Array training populations to train the classifier. The histology classification results will be used along with the pathologist’s interpretation of the matching H&E-stained section to designate separate ROIs for benign and malignant epithelium for each patient. Mean spectra will be used to develop a large set of candidate spectral metrics for distinguishing between benign epithelium and adenocarcinoma. Spectral metrics that show a statistically significant difference between the benign and adenocarcinoma populations will then be used in attempts to self-classify Array-P-80 and cross-validate by classifying other arrays with training data from Array-P-80.
4.2 Array P-80 H&E Stained Section Pathology Analysis

The H&E stained, matching section of Array P-80 was carefully reviewed with a pathologist and each spot was evaluated for several histopathological parameters. Before the review process, the visible optical image of each spot was printed on a separate sheet of paper and used to record the pathologist’s comments during the review process. Each spot was assessed initially for tissue preservation and preparation. Spots that contained significant preparation artifact or no epithelium were removed from analysis. The pathologist carefully characterized the remaining spots and detailed records were kept for subsequent ROI creation and analysis. The pathological status of epithelial tissue in each spot was considered individually and any epithelial tissue for which the pathological or preparation status was at all questionable was marked on the optical images so that it would not be considered in later analyses.

All regions of confirmed prostatic adenocarcinoma were individually assigned a Gleason Grade of 1-5 indicating the predominant Gleason pattern seen in the spot. Once the pathology analysis was complete, the results were tabulated and it was found that a total 38 patients contained usable benign epithelial tissue 51 patients contained usable regions of prostatic adenocarcinoma. A total of 25 patients from array P-80 contained regions of both benign epithelium and confirmed prostatic carcinoma. Without the corresponding benign tissue from the same patient as a control, any analysis of spectral features of adenocarcinoma tissue would be questionable. For this reason, the full-spectrum spectroscopic imaging datasets for the tissue array spots for these 25 patients were mosaicked into a single spectroscopic image for faster processing during downstream analyses.
4.3 Array P-80 Histology Classification Results

The 18 histology metrics used in sections 3.5.5 and 3.6.1 were calculated for the new 25 patient image of array P-80 from the baseline-corrected full spectrum image data. The same histology classification performed in section 3.6.1 as cross-array validation, was applied instead to the 25-patient, 18-metric image of array P-80. The 6 histologic class ROIs from array P-16 listed in table 3.4 were used as class training data for the GML classification of the 18-metric image. The histology classification results for this 25-patient image are displayed in Figure 4.1. Figure 4.2 contains the corresponding optical images from the matching H&E stained section.

![Figure 4.1- Array P-80 histology classification results](image)
4.3.1 Spatial Filtering of Histology Classification Results

The collected datasets have an effective pixel size of 6.25 µm x 6.25 µm. The spectral data are collected over the wavelength range from 4000-720 cm\(^{-1}\), or 2.5-13.8 µm. A given pixel will therefore contain some spectral information from tissue locations represented in the image data by neighboring pixels. Since most cells also have a size within or near the spectral wavelength range of radiation, some misclassification can be attributed to spectral bleeding from neighboring pixels that contain a different class of tissue. As expected, this phenomenon is most prevalent along borders between different histologic classes. Additionally, though the histological GML classification performs quite accurately, it is after all, a model and like all models has an inherent error rate.
While many types of spatial filtering techniques exist for digital image processing, the nature of the classification results suggest a particular method is most applicable for removing randomly distributed misclassified pixels. Spectroscopic image data and spectral metric data both span continuous ranges of data values within a single image plane. Most commonly applied spatial image filtering techniques work well with such data and involve some type of spatially-dependent averaging of pixel values within a defined local neighborhood of pixels.

The GML classifier assigns each pixel in the image scene to one of 6 discrete classes. Each class is represented in the image results data by a unique integer values. In such a case, the data values associated with each pixel do not form any sort of continuous scale, thus any spatial filtering techniques that rely on averaging would produce meaningless results. Several useful spatial filtering techniques have been developed for image classification results. Some complex methods utilize the conditional probability statistics developed during the application of the GML classification algorithm to analyze individual pixels with respect to a defined local neighborhood of pixels.[62]

Two more simple operations, which also produce satisfactory spatial filtering results, are the sieve operation and majority analysis. A sieve operation considers the neighborhood of pixels around a center pixel of class X, and applies a group minimum threshold. If the number of pixels in the neighborhood classified as X is less than the group minimum, then the pixel is relabeled as unclassified. The process is repeated for every pixel in the image. An alternative spatial filtering technique that can be applied to improve the appearance of image classification results is a majority analysis. This technique also considers a kernel, or set of neighboring pixels, which is rastered across
the image pixel-by-pixel. As the kernel moves, the center pixel is changed to the class that occupies the majority of the kernel positions that do not contain unclassified pixels. The weight of the center pixel can be changed in integer increments to alter the amount of filtering applied. This technique provides effective filtering of randomly misclassified pixels and changes them to the class that dominates the neighborhood. For this reason, majority-filtered images appear smoother than sieve-filtered results, which contain more unclassified pixels. Figure 4.2 contains results of these two different filtering strategies on a small example region of a classified prostate histology image representing a typical border between epithelium and mixed stroma.

![Spatial filtering techniques for classified image results](image)

The majority analysis produces results that are extremely smooth and preferable for general classification image display, however, it is important to point out that the
majority analysis changes the class designation of pixels based solely on spatial information without any spectral information whatsoever. At this stage in the data analysis, the sieve method is much more appropriate precisely because it is subtractive. Since the histology classification results will be used to construct epithelial ROIs for downstream spectroscopic analyses, it is important that such populations be as spectrally pure as possible.

The histology classification result image displayed in figure 4.1 was spatially filtered using a sieve operation implemented in ENVI using a neighborhood of eight pixels, and a group minimum threshold of five. The results were qualitatively compared with the matching H&E section and found to provide satisfactory removal of randomly misclassified pixels, while also removing questionable pixels near class boundaries. Figure 4.3 contains images of the raw histology classification and post-sieve operation classification image for patient 2 from array P-80.
4.4 Construction of a Supervised Classification Model for Prostate Pathology

4.4.1 Creation of pathology ground truth ROIs

The sieved histology classification results produced in section 4.3 were used as the starting point for the designation of pathology ground truth ROIs for array P-80. First, the sieved histology classification result for a given spot was compared with the annotated optical image of the matching H&E stained section. The epithelial classification result pixels that corresponded to epithelial tissue selected for use in the
marked optical H&E stained imaged were grouped into separate ROIs for benign epithelium and prostatic adenocarcinoma for each patient for a total of 50 ROIs. An image of the pathology ground truth ROIs and the corresponding number of pixels in each ROI is shown in Figure 4.4.

<table>
<thead>
<tr>
<th>Patient 01</th>
<th>Patient 02</th>
<th>Patient 03</th>
<th>Patient 04</th>
<th>Patient 05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1334</td>
<td>1464</td>
<td>1113</td>
<td>690</td>
<td>2001</td>
</tr>
<tr>
<td>19</td>
<td>1479</td>
<td>532</td>
<td>1219</td>
<td>96</td>
</tr>
<tr>
<td>2781</td>
<td>394</td>
<td>1774</td>
<td>2464</td>
<td>2037</td>
</tr>
<tr>
<td>1832</td>
<td>1476</td>
<td>518</td>
<td>803</td>
<td>3126</td>
</tr>
<tr>
<td>1812</td>
<td>1434</td>
<td>633</td>
<td>3038</td>
<td>1143</td>
</tr>
<tr>
<td>406</td>
<td>623</td>
<td>350</td>
<td>404</td>
<td>510</td>
</tr>
<tr>
<td></td>
<td>2604</td>
<td>1549</td>
<td>302</td>
<td>662</td>
</tr>
<tr>
<td></td>
<td>378</td>
<td>985</td>
<td>222</td>
<td>698</td>
</tr>
<tr>
<td></td>
<td>2934</td>
<td>400</td>
<td>2248</td>
<td>3883</td>
</tr>
<tr>
<td></td>
<td>137</td>
<td>289</td>
<td>302</td>
<td>885</td>
</tr>
</tbody>
</table>

**Adenocarcinoma**
- Total = 42,239 Pixels

**Benign Epithelium**
- Total = 19,492 Pixels

Figure 4.5 - Array P-80 pathology ground truth ROIs

4.4.2 Pathology Spectral Data Reduction

The mean infrared absorbance spectrum for each ROI was created and normalized to Amide II protein backbone absorbance at 1544 cm\(^{-1}\). These mean spectra were compared and spectral features, frequencies, and band ratios could be identified for distinguishing benign epithelial tissue from prostatic adenocarcinoma. A set of 54 candidate metrics was developed involving absorbance band ratios and peak centers of gravity for features across the entire spectral region. A listing of the parameters for each of the 54 candidate spectral metrics appears below in table 4.1.
<table>
<thead>
<tr>
<th>Metric #</th>
<th>Type of Metric</th>
<th>Band Ratio Parameters</th>
<th>Center-of-Gravity Spectral Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Numerator Band (cm⁻¹)</td>
<td>Denominator Band (cm⁻¹)</td>
</tr>
<tr>
<td>1</td>
<td>Band Ratio</td>
<td>966</td>
<td>1544</td>
</tr>
<tr>
<td>2</td>
<td>Band Ratio</td>
<td>1012</td>
<td>1544</td>
</tr>
<tr>
<td>3</td>
<td>Band Ratio</td>
<td>1040</td>
<td>1544</td>
</tr>
<tr>
<td>4</td>
<td>Band Ratio</td>
<td>1062</td>
<td>1544</td>
</tr>
<tr>
<td>5</td>
<td>Band Ratio</td>
<td>1080</td>
<td>1544</td>
</tr>
<tr>
<td>6</td>
<td>Band Ratio</td>
<td>1116</td>
<td>1544</td>
</tr>
<tr>
<td>7</td>
<td>Band Ratio</td>
<td>1158</td>
<td>1544</td>
</tr>
<tr>
<td>8</td>
<td>Band Ratio</td>
<td>1170</td>
<td>1544</td>
</tr>
<tr>
<td>9</td>
<td>Band Ratio</td>
<td>1206</td>
<td>1544</td>
</tr>
<tr>
<td>10</td>
<td>Band Ratio</td>
<td>1236</td>
<td>1544</td>
</tr>
<tr>
<td>11</td>
<td>Band Ratio</td>
<td>1278</td>
<td>1544</td>
</tr>
<tr>
<td>12</td>
<td>Band Ratio</td>
<td>1312</td>
<td>1544</td>
</tr>
<tr>
<td>13</td>
<td>Band Ratio</td>
<td>1450</td>
<td>1544</td>
</tr>
<tr>
<td>14</td>
<td>Band Ratio</td>
<td>1502</td>
<td>1544</td>
</tr>
<tr>
<td>15</td>
<td>Band Ratio</td>
<td>1516</td>
<td>1544</td>
</tr>
<tr>
<td>16</td>
<td>Band Ratio</td>
<td>1536</td>
<td>1544</td>
</tr>
<tr>
<td>17</td>
<td>Band Ratio</td>
<td>1562</td>
<td>1544</td>
</tr>
<tr>
<td>18</td>
<td>Band Ratio</td>
<td>1588</td>
<td>1544</td>
</tr>
<tr>
<td>19</td>
<td>Band Ratio</td>
<td>1652</td>
<td>1544</td>
</tr>
<tr>
<td>20</td>
<td>Band Ratio</td>
<td>3082</td>
<td>1544</td>
</tr>
<tr>
<td>21</td>
<td>Band Ratio</td>
<td>3290</td>
<td>1544</td>
</tr>
<tr>
<td>22</td>
<td>Band Ratio</td>
<td>3450</td>
<td>1544</td>
</tr>
<tr>
<td>23</td>
<td>Band Ratio</td>
<td>1426</td>
<td>1450</td>
</tr>
<tr>
<td>24</td>
<td>Band Ratio</td>
<td>1390</td>
<td>1450</td>
</tr>
<tr>
<td>25</td>
<td>Band Ratio</td>
<td>1400</td>
<td>1450</td>
</tr>
<tr>
<td>26</td>
<td>Band Ratio</td>
<td>1032</td>
<td>1080</td>
</tr>
<tr>
<td>27</td>
<td>Band Ratio</td>
<td>1016</td>
<td>1080</td>
</tr>
<tr>
<td>28</td>
<td>Band Ratio</td>
<td>1208</td>
<td>1236</td>
</tr>
<tr>
<td>29</td>
<td>Band Ratio</td>
<td>1262</td>
<td>1236</td>
</tr>
<tr>
<td>30</td>
<td>Band Ratio</td>
<td>1278</td>
<td>1236</td>
</tr>
<tr>
<td>31</td>
<td>Band Ratio</td>
<td>1516</td>
<td>1236</td>
</tr>
<tr>
<td>32</td>
<td>Band Ratio</td>
<td>1080</td>
<td>1236</td>
</tr>
<tr>
<td>33</td>
<td>Band Ratio</td>
<td>1080</td>
<td>1450</td>
</tr>
<tr>
<td>34</td>
<td>Band Ratio</td>
<td>1080</td>
<td>1652</td>
</tr>
<tr>
<td>35</td>
<td>Band Ratio</td>
<td>1080</td>
<td>3290</td>
</tr>
<tr>
<td>36</td>
<td>Band Ratio</td>
<td>3044</td>
<td>3290</td>
</tr>
<tr>
<td>37</td>
<td>Band Ratio</td>
<td>3050</td>
<td>3290</td>
</tr>
<tr>
<td>38</td>
<td>Band Ratio</td>
<td>3060</td>
<td>3290</td>
</tr>
<tr>
<td>39</td>
<td>Band Ratio</td>
<td>3064</td>
<td>3290</td>
</tr>
<tr>
<td>40</td>
<td>Band Ratio</td>
<td>3078</td>
<td>3290</td>
</tr>
<tr>
<td>41</td>
<td>Band Ratio</td>
<td>3084</td>
<td>3290</td>
</tr>
<tr>
<td>42</td>
<td>Band Ratio</td>
<td>3180</td>
<td>3290</td>
</tr>
<tr>
<td>43</td>
<td>Band Ratio</td>
<td>3192</td>
<td>3290</td>
</tr>
<tr>
<td>44</td>
<td>Band Ratio</td>
<td>3202</td>
<td>3290</td>
</tr>
<tr>
<td>45</td>
<td>Band Ratio</td>
<td>3214</td>
<td>3290</td>
</tr>
<tr>
<td>46</td>
<td>Band Ratio</td>
<td>3226</td>
<td>3290</td>
</tr>
<tr>
<td>47</td>
<td>Band Ratio</td>
<td>3232</td>
<td>3290</td>
</tr>
<tr>
<td>48</td>
<td>Center of Gravity</td>
<td>1144</td>
<td>982</td>
</tr>
<tr>
<td>49</td>
<td>Center of Gravity</td>
<td>1182</td>
<td>1144</td>
</tr>
<tr>
<td>50</td>
<td>Center of Gravity</td>
<td>1296</td>
<td>1182</td>
</tr>
<tr>
<td>51</td>
<td>Center of Gravity</td>
<td>1426</td>
<td>1352</td>
</tr>
<tr>
<td>52</td>
<td>Center of Gravity</td>
<td>1578</td>
<td>1478</td>
</tr>
<tr>
<td>53</td>
<td>Center of Gravity</td>
<td>1718</td>
<td>1585</td>
</tr>
<tr>
<td>54</td>
<td>Center of Gravity</td>
<td>3682</td>
<td>3000</td>
</tr>
</tbody>
</table>

Table 4.1 - Pathology spectral metric parameters
4.4.3 Histogram analysis of Spectral Metric Data

Initial metric evaluation was conducted plotting histograms of different pathology ground truth ROIs for individual metrics. Histograms analyzed on a patient-to-patient basis revealed that for many metrics, a similar directional shift in the means of frequency distributions between benign and adenocarcinoma populations was present. For many of these metrics, while the direction of the shift was consistent from patient-to-patient, the absolute values of the respective distributions varied quite significantly among patients. This situation is depicted schematically in figure 4.5.

![Graph showing patient-to-patient metric variation](image)

Figure 4.6 - Patient-to-patient metric variation

It was clear that many of these metrics were providing information regarding real spectral differences between benign and cancerous prostate tissue, however, the significant patient-to-patient variation rendered these metrics ineffective for use in parametric classification attempts.
4.4.4 Mean-centering of epithelial metric data.

The data were mean-centered in order to make use of the spectroscopic information contained in the metrics affected by significant patient-to-patient variation and to simplify the process of metric evaluation. The mean metric spectrum for each patient’s benign ground truth ROI was calculated by averaging the individual metric-spectra within each ROI. The discrete 54-metric spectrum of each individual epithelial pixel was divided by the mean benign metric spectrum from the corresponding patient. This calculation has the effect of normalizing the benign population distributions for all patients individually for each metric. Thus, all patient-patient variation among benign metric distributions is effectively collapsed such that recalculation of the benign 54-metric spectrum for any patient would yield a value of one at every metric.

4.4.5 Metric Statistical Analysis

A major advantage of mean-centering the metric data is that it simplifies the task of identifying which metrics provide statistically significant discrimination between benign and adenocarcinoma patient populations. The mean 54-metric spectrum for each patient’s adenocarcinoma ground truth ROI population was recalculated from the benign mean-centered metric data. A one population t-test was applied to each of 54 sets of 25 patient-mean metric values to determine if the 25 patient population was significantly different from the constant 1.0 at the 0.05 level. The results of the t-test for each metric and the associated p-values are listed in table 4.2.
<table>
<thead>
<tr>
<th>Metric #</th>
<th>Mean</th>
<th>Variance</th>
<th>t-value</th>
<th>p-value</th>
<th>Significantly different from 1.0?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.137</td>
<td>0.078</td>
<td>2.463</td>
<td>0.021</td>
<td>YES</td>
</tr>
<tr>
<td>2</td>
<td>0.865</td>
<td>0.017</td>
<td>-5.099</td>
<td>0.003</td>
<td>YES</td>
</tr>
<tr>
<td>3</td>
<td>0.994</td>
<td>0.011</td>
<td>-0.277</td>
<td>0.783</td>
<td>NO</td>
</tr>
<tr>
<td>4</td>
<td>1.029</td>
<td>0.015</td>
<td>1.175</td>
<td>0.251</td>
<td>NO</td>
</tr>
<tr>
<td>5</td>
<td>1.026</td>
<td>0.014</td>
<td>1.086</td>
<td>0.288</td>
<td>NO</td>
</tr>
<tr>
<td>6</td>
<td>1.056</td>
<td>0.035</td>
<td>1.504</td>
<td>0.146</td>
<td>NO</td>
</tr>
<tr>
<td>7</td>
<td>0.879</td>
<td>0.038</td>
<td>-3.092</td>
<td>0.005</td>
<td>YES</td>
</tr>
<tr>
<td>8</td>
<td>0.934</td>
<td>0.012</td>
<td>-3.023</td>
<td>0.006</td>
<td>YES</td>
</tr>
<tr>
<td>9</td>
<td>1.069</td>
<td>0.016</td>
<td>2.715</td>
<td>0.012</td>
<td>YES</td>
</tr>
<tr>
<td>10</td>
<td>1.052</td>
<td>0.007</td>
<td>3.099</td>
<td>0.005</td>
<td>YES</td>
</tr>
<tr>
<td>11</td>
<td>1.075</td>
<td>0.032</td>
<td>2.095</td>
<td>0.047</td>
<td>YES</td>
</tr>
<tr>
<td>12</td>
<td>0.961</td>
<td>0.029</td>
<td>-1.169</td>
<td>0.254</td>
<td>NO</td>
</tr>
<tr>
<td>13</td>
<td>0.988</td>
<td>0.007</td>
<td>-0.727</td>
<td>0.474</td>
<td>NO</td>
</tr>
<tr>
<td>14</td>
<td>0.960</td>
<td>0.006</td>
<td>-2.681</td>
<td>0.013</td>
<td>YES</td>
</tr>
<tr>
<td>15</td>
<td>0.969</td>
<td>0.002</td>
<td>-3.597</td>
<td>0.001</td>
<td>YES</td>
</tr>
<tr>
<td>16</td>
<td>0.991</td>
<td>1.8x10^-4</td>
<td>-3.179</td>
<td>0.004</td>
<td>YES</td>
</tr>
<tr>
<td>17</td>
<td>1.003</td>
<td>0.001</td>
<td>0.559</td>
<td>0.582</td>
<td>NO</td>
</tr>
<tr>
<td>18</td>
<td>0.983</td>
<td>4.4x10^-4</td>
<td>-4.087</td>
<td>4.2x10^-4</td>
<td>YES</td>
</tr>
<tr>
<td>19</td>
<td>1.000</td>
<td>0.001</td>
<td>-0.085</td>
<td>0.933</td>
<td>NO</td>
</tr>
<tr>
<td>20</td>
<td>0.991</td>
<td>0.003</td>
<td>-0.812</td>
<td>0.425</td>
<td>NO</td>
</tr>
<tr>
<td>21</td>
<td>0.977</td>
<td>0.002</td>
<td>-2.682</td>
<td>0.013</td>
<td>YES</td>
</tr>
<tr>
<td>22</td>
<td>0.978</td>
<td>0.017</td>
<td>-0.822</td>
<td>0.419</td>
<td>NO</td>
</tr>
<tr>
<td>23</td>
<td>1.073</td>
<td>0.009</td>
<td>3.873</td>
<td>0.001</td>
<td>YES</td>
</tr>
<tr>
<td>24</td>
<td>1.011</td>
<td>0.009</td>
<td>0.581</td>
<td>0.567</td>
<td>NO</td>
</tr>
<tr>
<td>25</td>
<td>1.020</td>
<td>0.008</td>
<td>1.098</td>
<td>0.283</td>
<td>NO</td>
</tr>
<tr>
<td>26</td>
<td>0.946</td>
<td>0.007</td>
<td>-3.264</td>
<td>0.003</td>
<td>YES</td>
</tr>
<tr>
<td>27</td>
<td>0.870</td>
<td>0.013</td>
<td>-5.578</td>
<td>9.7x10^-4</td>
<td>YES</td>
</tr>
<tr>
<td>28</td>
<td>1.008</td>
<td>0.004</td>
<td>0.671</td>
<td>0.509</td>
<td>NO</td>
</tr>
<tr>
<td>29</td>
<td>0.984</td>
<td>0.003</td>
<td>-1.548</td>
<td>0.135</td>
<td>NO</td>
</tr>
<tr>
<td>30</td>
<td>1.019</td>
<td>0.014</td>
<td>0.787</td>
<td>0.439</td>
<td>NO</td>
</tr>
<tr>
<td>31</td>
<td>0.928</td>
<td>0.010</td>
<td>-3.653</td>
<td>0.001</td>
<td>YES</td>
</tr>
<tr>
<td>32</td>
<td>0.978</td>
<td>0.013</td>
<td>-0.967</td>
<td>0.343</td>
<td>NO</td>
</tr>
<tr>
<td>33</td>
<td>1.049</td>
<td>0.027</td>
<td>1.505</td>
<td>0.145</td>
<td>NO</td>
</tr>
<tr>
<td>34</td>
<td>1.025</td>
<td>0.011</td>
<td>1.192</td>
<td>0.245</td>
<td>NO</td>
</tr>
<tr>
<td>35</td>
<td>1.047</td>
<td>0.011</td>
<td>2.298</td>
<td>0.031</td>
<td>YES</td>
</tr>
<tr>
<td>36</td>
<td>1.010</td>
<td>0.003</td>
<td>0.991</td>
<td>0.332</td>
<td>NO</td>
</tr>
<tr>
<td>37</td>
<td>1.003</td>
<td>0.003</td>
<td>0.246</td>
<td>0.808</td>
<td>NO</td>
</tr>
<tr>
<td>38</td>
<td>1.007</td>
<td>0.003</td>
<td>0.684</td>
<td>0.501</td>
<td>NO</td>
</tr>
<tr>
<td>39</td>
<td>1.014</td>
<td>0.002</td>
<td>1.468</td>
<td>0.155</td>
<td>NO</td>
</tr>
<tr>
<td>40</td>
<td>1.015</td>
<td>0.003</td>
<td>1.393</td>
<td>0.176</td>
<td>NO</td>
</tr>
<tr>
<td>41</td>
<td>1.014</td>
<td>0.003</td>
<td>1.388</td>
<td>0.178</td>
<td>NO</td>
</tr>
<tr>
<td>42</td>
<td>1.000</td>
<td>0.002</td>
<td>-0.027</td>
<td>0.979</td>
<td>NO</td>
</tr>
<tr>
<td>43</td>
<td>0.999</td>
<td>0.002</td>
<td>-0.119</td>
<td>0.907</td>
<td>NO</td>
</tr>
<tr>
<td>44</td>
<td>0.998</td>
<td>0.001</td>
<td>-0.210</td>
<td>0.635</td>
<td>NO</td>
</tr>
<tr>
<td>45</td>
<td>0.998</td>
<td>0.001</td>
<td>-0.289</td>
<td>0.775</td>
<td>NO</td>
</tr>
<tr>
<td>46</td>
<td>0.997</td>
<td>0.001</td>
<td>-0.421</td>
<td>0.678</td>
<td>NO</td>
</tr>
<tr>
<td>47</td>
<td>0.997</td>
<td>0.001</td>
<td>-0.482</td>
<td>0.634</td>
<td>NO</td>
</tr>
<tr>
<td>48</td>
<td>1.001</td>
<td>3.016x10^-4</td>
<td>3.806</td>
<td>0.001</td>
<td>YES</td>
</tr>
<tr>
<td>49</td>
<td>1.000</td>
<td>8.991x10^-4</td>
<td>2.630</td>
<td>0.015</td>
<td>YES</td>
</tr>
<tr>
<td>50</td>
<td>1.000</td>
<td>3.754x10^-4</td>
<td>-1.127</td>
<td>0.271</td>
<td>NO</td>
</tr>
<tr>
<td>51</td>
<td>1.000</td>
<td>2.437x10^-4</td>
<td>2.049</td>
<td>0.052</td>
<td>NO</td>
</tr>
<tr>
<td>52</td>
<td>1.000</td>
<td>1.663x10^-4</td>
<td>2.786</td>
<td>0.010</td>
<td>YES</td>
</tr>
<tr>
<td>53</td>
<td>1.000</td>
<td>3.885x10^-4</td>
<td>0.564</td>
<td>0.578</td>
<td>NO</td>
</tr>
<tr>
<td>54</td>
<td>1.000</td>
<td>3.350x10^-4</td>
<td>-0.444</td>
<td>0.661</td>
<td>NO</td>
</tr>
</tbody>
</table>

Table 4.2 - Results of t-test on mean adenocarcinoma metric values from population of 25 patients on array P-80 for 54 candidate pathology metrics
The t-test results indicate that 20 metrics from the candidate set of 54 pathology metrics show statistically significant deviation between their respective populations of adenocarcinoma and patient-matched benign epithelial pixels.

4.4.6 GML Pathology Classification of Array P-80

The set of 20 pathology metrics identified in section 4.2.8 for discriminating between benign and malignant prostate tissue were used as data for GML classification of all epithelial pixels from P-80 ground truth epithelial ROIs. The benign ground truth ROI sets for all 25 patients were merged into one large benign training ROI comprised of 19,492 total pixels. Likewise, the adenocarcinoma ground truth ROI sets for all 25 patients were merged into one large adenocarcinoma training ROI comprised of 42,239 total pixels. A supervised 2-class (benign epithelium & adenocarcinoma) classification was implemented in ENVI using the 20 metrics identified in section 4.2.8.

The classification image results for all 25 patients appear below in Figure 4.6.
The ground truth training ROIs were used to construct an error matrix to evaluate the classification results on a whole-array basis. The error matrix appears below in Figure 4.3.
These classification results give a sense that in general, the classifier is performing adequately for distinguishing benign epithelium from regions of adenocarcinoma. While very little misclassification of ground truth benign pixels is seen, there are a handful of adenocarcinoma spots in Figure 4.8 which seem to be classified with less certainty than the remainder of the patients.

4.5 Individual Patient Evaluation of P-80 Pathology Classification

The 20-metric pathology classification results were analyzed next on an individual patient basis. For each of the fifty pathology ground truth ROIs (25 benign + 25 adenocarcinoma), the percentage of ROI pixels classified as adenocarcinoma for each ROI was plotted as a bar chart in Figure 4.7 below.
The data reveals that the pixels from the benign ROIs of all 25 patients were classified with an accuracy > 80%. Imposing a minimum threshold for adenocarcinoma classification of 20% on the data in figure 4.7 provides 100% discrimination between foci benign and malignant epithelial tissue across the entire population of 25 patients.

4.6 Cross-Array Validation

Again it must be noted that such self-evaluation of training data ROIs represents the best possible scenario for producing accurate supervised results. To examine the cross-array performance of the pathology classification model, training data from array p-80 was used to classify mean-centered 20-metric data from other arrays.

Upon the pathologist’s review of the H&E stained section, arrays P-16 and P-40 were each found to contain 5 patients with usable regions of both benign epithelium. Initial cross-array classification attempts did not yield consistent results. While some
individual patients yielded results similar to those seen with Array P-80, the limited population sizes of five patients on each array made it impossible to draw any substantive conclusions regarding the cross-array performance of the developed pathology classification model.

4.7 Conclusions and Further Directions

These results indicate that spectral features from FTIR spectroscopic imaging data can be used to differentiate between regions of healthly benign prostate epithelial tissue and regions harboring prostatic adenocarcinoma. The results presented in this section represent an initial attempt to probe the infrared spectroscopic characteristics of prostate histopathology and serve to highlight the promise that such vibrational spectroscopic imaging techniques hold for the objective analysis of sectioned tissue.

Such methods provide simple, readily interpretable image-based results that convey histological and pathological information provided by referencing spectral database information. Some of the most useful of these preliminary results are those from section 4.8 from the t-test analyses of individual metrics.

The t-test results displayed in table 4.2 show the 20 metrics for which the patient populations of patient-mean adenocarcinoma metric values differed significantly from corresponding benign metric values. Close examination of the spectral parameters of each of these 20 metrics listed in table 4.1 reveals that most successful metrics involve spectral information from a handful of spectral regions corresponding to specific vibrational modes. Several metrics involve spectral information from the spectral region between 1200-1000 cm\(^{-1}\), a region with prominent absorbances due to vibrational modes
of glycogen, as well as symmetric stretching of phosphodiester (PO$_2^-$) groups of nucleic acids. The region between 1300-1200 cm$^{-1}$ also contributed to several metrics; this region contains spectral absorbances due to protein Amide III modes and antisymmetric stretching modes of nucleic acid PO$_2^-$ groups. Finally, many metrics involved features from the spectral region between 1590-1500 cm$^{-1}$, a region whose main absorbance is the Amide II mode of proteins arising from N-H bending modes coupled to C-N stretching on the protein backbone.

Future spectroscopic imaging of prostate tissue at higher spectral resolutions will allow more information to be extracted from these spectral regions. Alternative classification methods, such as spectral-angle mapping and hierarchical cluster analysis more readily make use of continuous spectral information, and can be employed using data from these isolated regions of the spectrum. If performed on substantially larger patient populations, it is likely that such approaches will lead to more specific information regarding, spectrally similar subgroups of related cancers and correlations with histologic grade and/or disease progression.

Clearly, studies conducted with larger tissue microarrays and patient populations will advance our understanding of the spectroscopic properties of prostate pathology. The technology utilized to collect vibrational spectroscopic imaging data is advancing at a rapid pace. Faster collection times, better SNRs, and higher data collection at higher spatial and spectral resolutions will all add to the power of this technique in the future.

Among the most promising future analytical approaches will be to create techniques to register spectroscopic image data with results from other analytical techniques, such as immunohistochemical staining and in-situ hybridization conducted after IR data
collection, or performed on serial tissue array section. Such combinatorial approaches should enable calibrations to be constructed that could tentatively predict staining patterns for multiple panels of antibodies or other probes via spectral pattern recognition from spectroscopic image data of unstained tissue.
References


46. Perkin-Elmer (Shelton, CT) Spectrum Spotlight 300 FTIR Imaging System


About the Author

Daniel Celestino Fernandez received his A.B. in Chemistry in 1997 from Amherst College in Amherst, MA. During his senior year he undertook an honors research project where he was exposed to vibrational spectroscopy for the first time and wrote a thesis entitled “Spectroscopic Characterization of the Iron-Sulfur Cluster of Pyruvate Formate-Lyase Activating Enzyme.” After college he returned home to Tampa, FL where he began medical school at the University of South Florida. After finishing his first two years of study toward an M.D. degree, he accepted a Howard Hughes Medical Institute Research Scholar Fellowship at the National Institutes of Health in Bethesda, MD. In the Section of Molecular Biophysics of the Laboratory of Chemical Physics of the National Institute of Diabetes, Digestive, and Kidney Diseases, he joined a team working on biological applications of spectroscopic imaging techniques with Dr. Ira W. Levin. After two years as an HHMI-NIH Research Scholar, with the help of the NIH Graduate Partnerships Program, he enrolled in the Medical Sciences Ph.D. program in the Department of Pathology and Laboratory Medicine at the College of Medicine of the University of South Florida and was able to stay at the NIH to complete two additional years of doctoral research. He has now transferred to the Mount Sinai School of Medicine in New York City where he is finishing his last year of study toward the M.D. degree. After graduation he plans to complete a residency in diagnostic radiology.