Chemometrics Development using Multivariate Statistics and Vibrational Spectroscopy and its Application to Cancer Diagnosis

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Ran Li

Graduate Program in Chemistry

The Ohio State University

2015

Dissertation Committee:

Professor Heather C. Allen, Advisor

Professor James V. Coe, Co-Advisor

Professor Thomas J. Magliery

Professor Eylem Ekici

ABSTRACT

Cancer is one of the leading causes of mortality in both sexes worldwide. The project "cancer margin detection using vibrational spectroscopy" aims to develop methodologies capable of identifying and distinguishing cancer-bearing from non-cancer-bearing tissues, and to potentially complement standard histopathological tissue analysis for real-time cancer detection as it relates to the assessment of surgical resection margins and the completeness of surgical resection in both the operating room and the pathology department. To fulfill this goal, vibrational spectroscopy and multivariate statistics were incorporated into the methods development by integrating the knowledge of analytical chemistry, biochemistry, statistics and pathology.

In this study, a point detection technology using an ATR probe which provides real-time information on tissue groups (i.e. tumor and non-tumor) in light of a tissue discrimination model (TDM) is developed. The TDM is built on thin tissue sections of metastatic liver lesion from colorectal cancer using Fourier transform infrared (FTIR) spectroscopy imaging with multivariate statistics [*k*-means clustering and support vector machine (SVM)]. Biometrics values were further validated and the ones that have the greatest contribution to differentiate tumor and non-tumor tissues were selected to build the TDM. Subsequently, *k*-means clustering analysis using the selected biometrics were conducted on two groups (tumor and non-tumor groups). 10000 spectra from the FTIR

image data are chosen to build the training set along with their corresponding properties (i.e., tumor, and non-tumor). By studying the training set, SVM generates decision equations that are used subsequently to predict tumor in the attenuated total reflectance (ATR)-FTIR spectra obtained on the original resected liver tissues. The TDM was further validated on FTIR image data of other cases to test for the inherent variations between individuals. By comparing the prediction from the TDM with the results from *k*-means clustering analysis, the accuracy of five cases was in the range 95.4±5.7%. The statistical accuracy of the TDM as determined by Student's *t*-test was found to be 95.4 ± 5.4% ($P < 0.1$). Finally, this model was used in conjunction with an attenuated total reflection (ATR) probe as a point-detection method to differentiate cancer-bearing from normal tissue. This newly developed TDM positions ATR-FTIR spectroscopy a step closer toward its application as a real-time intraoperative tool for the objective identification of cancer-bearing tissues during surgery.

Distinct from previous methodology, the second approach investigates the alterations of protein secondary structures in tumor and non-tumor tissues through matrix multiplication of tissue spectra with calibrant IR spectra (calculated spectra dominant in α-helices and β-sheets extracted from protein standards of Dong, Carpenter, and Caughey) using the spectral range of 1500-1700 cm-^{1}, where the Amide I (1600−1700 cm^{-1}) and Amide II (1500−1600 cm^{-1}) bands are located. The plot of α-helix versus β-sheet scores differentiates the tumor and non-tumor spectra of rectal adenocarcinoma metastatic to

liver. Spectra obtained in the tumor region exhibit lower α-helix and β-sheet scores. The decrease is related to the reduced level of albumin in the tumor region.

DEDICATION

To the people I love; and to the things to love about life: piano, writing, hiking tennis and science.

谨以此文献给那些帮助过我的,我挚爱的人们。

ACKNOWLEDGMENTS

First and foremost, I would like to thank my research advisor, Professor Heather C. Allen for her constant support and encouragement through the completion of this project. I would like to extend my deepest appreciation to Professor James V. Coe, for taking me to the fantastic world of programming and for his advice through my Ph.D. study. I would also like to express my sincere gratitude to Dr. Dominique Verreault for his willingness to help, for those illuminating scientific discussions, and for providing valuable feedback on this thesis. Their guidance and integrity in science made me grow not only as a high level technician but also as an independent researcher and a professional scientist. I would also like to thank Professor Thomas J. Magliery, for accepting to be on in my committee and advising my final dissertation. Also, thanks to all Allen group members, past and present, in this exciting Ph.D. journey we all share.

Acknowledgement also goes to Drs. Charles L. Hitchcock, Edward W. Martin Jr. and Stephen P. Povoski. I would also like to express my gratitude to colleagues Zhaomin Chen, Barrie Miller, Ryan Butke, and Steven Nystrom. Also, special thanks to those people who helped me in various ways during my Ph.D. life.

I would also like to thank my previous mentors for stimulating my interest in science. I wish them all the best, and to all their dreams.

Last but not least, I would like to thank my parents for always believing in me and letting me be myself. Thanks to all my family members, especially my cousin for always answering my incessant questions. Thanks to those friends, who grew up with me. Their encouragement and enlightenment kept me sane during these years, and gave me the extra push to complete my Ph.D. All of them are my sources of happiness, and they made me who I am today.

VITA

2008.....B.S. Pharmaceutical Science, Hainan University

2010.....M.S. Chemistry, Eastern Michigan University

2015.....Ph.D. Chemistry, The Ohio State University

PUBLICATIONS

1. **Li, R.**, Verreault, D., Miller, B., Chen, Z., Hitchcock, C., Povoski, S., Martin Jr., E., Coe, J. and Allen, H. (2015) Development of a Tissue Discrimination Model Using Supervised Data Transformation with Support Vector Machines, *Anal. Bioanal. Chem. under review* (Manuscript No. ABC-01373-2015, Date: 31-July-2015, No. Pages Submitted: 31)
2. Coe, J., Nystrom, S., Chen, Z., **Li, R.**, Verreault, D., Hitchcock, C., Martin Jr., E., and Allen, H. (2015) Extracting Infrared Spectra of Protein Secondary Structures using a Library of Protein Spectra and the Ramachandran Plot, *J. Phys. Chem. B* 119(41), 13079-13092 (2015).
3. Coe, J., Chen, Z., **Li, R.**, Nystrom, S., Butke, B., Miller, B., Hitchcock, C., Allen, H., Povoski, S., Martin, Jr. E. (2015) Molecular Constituents of Colorectal Cancer Metastatic to the Liver by Imaging Infrared Spectroscopy, *Proc. SPIE* 8328, 93280R/1-93280R/7.
4. **Li, R.**, Verreault, D., Payne, A., Hitchcock, C., Povoski, S., Martin Jr., E., and Allen, H. (2014) Effects of Laser Excitation Wavelength and Optical Mode on Raman Spectra of Human Fresh Colon, Pancreas, and Prostate Tissues, *J. Raman Spectrosc*. 45(9): 773-780.
5. Coe, J., Chen, Z., **Li, R.**, Nystrom, S., Butke, B., Miller, B., Hitchcock, C., Allen, H., Povoski, S., Martin, Jr. E. (2014) Imaging Infrared Spectroscopy for Fixation-Free Liver Tumor Detection, *Proc. SPIE* 8947, 89470B/1-89470B/6.
6. **Li, R.**, Baker, S., DeRoo, C. S., Armitage, R. (2012) Characterization of the Binders and Pigments in the Rock Paintings of Cueva la Conga, Nicaragua Collaborative Endeavors in the Chemical Analysis of Art and Cultural Heritage Materials, Chapter

4, 75-89.

FIELDS OF STUDY

Major Field: Chemistry

TABLE OF CONTENTS

ABBREVIATIONS

ATR	attenuated total reflectance
EIT	electrical impedance tomography
FFT	fast Fourier transform
FTIR	Fourier transform infrared
IR	infrared
LDA	linear discriminant analysis
MCT	mercury-cadmium-telluride
MRI	magnetic resonance imaging
MSI	mass spectrometry imaging
OVA	one-versus-all
OVO	one-versus-one
PCA	principal component analysis
PLS	partial least squares
RBF	radial basis function
SVM	support vector machine
TDM	tissue discrimination model
WHO	World Health Organization

SYMBOLS

δ	phase difference
γ	kernel function parameter
λ	wavelength
λ_{IR}	IR wavelength
φ	dihedral angle, phase
$\bar{\nu}$	wavenumber
ψ	dihedral angle
θ	angle of incidence
$\sigma_{lineshape}$	standard deviation of experimental lineshape
$\sigma_{2nd\ derivative}$	standard deviation of second derivative
ω	frequency
a	semi-major axis width
a_i	coefficient of SVM decision function
A	absorbance
b	bias, semi-minor axis width
B	IR spectra of the secondary structure groups
c	penalty parameter
C	centroid
d_p	penetration depth

$\boldsymbol{E}$	total electric field
$\boldsymbol{E_0}$	electric field
$\boldsymbol{i, j}$	position indices
$\boldsymbol{k}$	stepping index
$K(\mathbf{x}_i, \mathbf{x}_j)$	kernel function
$\boldsymbol{n}$	refractive index
$\boldsymbol{n}_{\mathbf{crystal}}$	refractive index of ATR crystal
$\boldsymbol{n}_{\mathbf{tissue}}$	refractive index of biological tissue
$\boldsymbol{s}$	scaling factor
$\boldsymbol{S}$	matrix product
$\boldsymbol{x}$	fractions of amino acids secondary structures
$\boldsymbol{x}_{\mathbf{path}}$	path difference
$\mathbf{x}_i, \mathbf{x}_j$	real-valued n-dimensional training vectors
$\boldsymbol{xcal_{ij}}$	calibrant spectrum
$\boldsymbol{y}$	absorbance of each library protein
$\boldsymbol{z}$	distance from interface

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Motivation

The work presented in this dissertation focuses on chemometrics development with the aims of identifying and distinguishing cancer-bearing from non-cancer-bearing tissues using IR spectroscopy and multivariate statistics. In the long term, these strategies are to be applied in the operating room to help oncologists in surgical removal of malignant tumors. The strength of the chemometrics described in this dissertation is based on the combination of Fourier transform infrared (FTIR) spectroscopy with multivariate statistics. On one hand, FTIR spectroscopy has the advantage of being a non-destructive and label-free technique as it does not require the use of exogenous agents (fluorophores, antibodies, etc.) and it is sensitive to biochemical alterations in tissues rather than morphological features as is the case with current histopathological tissues assessment where tissue staining is required. On the other hand,

multivariate statistics and computational methodology considerably reduce the data processing time, allowing real-time time feedback, and enabling greater level of details to be obtained. Generally speaking, the approaches implemented in this study are combining FTIR mapping with ATR fiber optic probe, and applying multivariate statistics for data analysis.

Cancer is and remains one of the leading causes of mortality in both sexes worldwide. As reported by the World Health Organization (WHO), there were an estimated 14 million new cancer cases and 8 million cancer-related deaths that occurred in 2012. Likewise, it was estimated that there were 33 million people over the age of 15 years old who were alive with a cancer diagnosed within the prior five-year time frame.[1] Furthermore, current projections predict an increase in cancer worldwide with an estimated 21 million new cases and 13 million cancer-related deaths in 2030, respectively.[2]

Reliable diagnosis and resection of all cancer-bearing tissues at the time of surgery is therefore of critical significance for decreasing the future recurrence of the disease. Currently, histopathological tissue examination remains the standard-of-care method for the confirmation of a correct cancer diagnosis. However, major limitations include: (i) the

error (1–2%) associated with the routine histopathological tissue assessment,[3] (ii) inconsistent identification of tumor and assessment of surgical resection margins as a result of the subjective judgment of pathologists, and (iii) lack of real-time information, which is necessary for intraoperative decision making. The combination of these limitations, with a particular emphasis on the third one, brings about a situation that is detrimental to successful patient management and postoperative treatment planning.

To improve diagnostic accuracy, other techniques such as, for example, magnetic resonance imaging (MRI), mass spectrometry imaging (MSI), or electrical impedance tomography (EIT), are currently applied in diagnosing tumor occurrence, although each of them suffer from some disadvantages, e.g., introduction of exogenous agents, invasive to the surface of the sample, and sensitivity. Finally, regardless of their efficacy, these techniques require expensive instrumentation operated only by highly trained personal, thereby limiting their frequent application.[4]

This situation has prompted the development of alternative techniques relying on changes in chemical composition as a means to distinguish normal from diseased tissues, an approach often referred to as "molecular histopathology".[5] Among these newly developed techniques, FTIR spectroscopy has emerged as a promising technique that can

relate biochemical information contained in FTIR spectra to changes in cell or tissue composition. Although FTIR spectroscopy has certain limitations, e.g. more expensive and less automated compared to fluorescence imaging, it presents several advantages: (1) it requires only small amounts of tissue for analysis, (2) it is non-destructive and needs no labeling or staining, (3) it is inherently very sensitive to molecular-level biochemical changes. In addition, the use of an attenuated total reflectance (ATR) probe with FTIR allows measurements to be done on tissue sections or even in vivo. This combination has already been applied in the realm of cancer diagnosis to obtain spectroscopic information *in vivo*.[6-8]

FTIR spectroscopy examines biological tissues by probing spectral variations (peak intensity, position, and width) in vibrational modes arising from different functional groups in biomolecules, including proteins, lipids, nucleic acids, carbohydrates, etc. In light of previous studies, significant spectral differences between non cancer-bearing and cancer-bearing tissues have been observed in the mid-IR region (400–4000 cm^{-1}). Initially, cancer detection using FTIR focused on identifying cancer-bearing from non cancer-bearing tissues based on investigating differences in the spectral parameters (frequency, intensity, and shape) of the phosphate (P-O; 1082 and 1241 cm^{-1}) and

carbonyl (C-O; 1164 cm^{-1}) bands, as well as the C-H stretching region (2800–3000 cm^{-1}).[9-12] Band ratios were later introduced as an improved methodology for differentiation by eliminating the effect of sample thickness variation.[13] Nevertheless, two major problems were encountered in the treatment of spectral data: (1) under some circumstances spectra obtained from non cancer-bearing and cancer-bearing tissues exhibit subtle distinctions, difficult to identify, and (2) spectral differences not only come from cancerous/non-cancerous cells but are also related to different cell types.[14,15] Furthermore, the data processing is time-consuming, especially when dealing with high volume of data gained from FTIR mapping.

Introducing multivariate statistics such as partial least squares (PLS), linear discriminant analysis (LDA), principle component analysis (PCA), clustering analysis, support vector machines (SVM), genetic algorithms, neural networking, etc., into spectral data analysis and interpretation has been tremendously helpful, not only by allowing greater level of details to be examined but also by considerably reducing data processing time. *K*-means clustering analysis and SVM, in particular, have been applied quite successfully to interpret spectral data in cancer diagnosis.[16-20] *K*-means clustering analysis is able to identify different groups on tissue spectra mapping, and the groups can

simply be cancerous/non cancerous cells, or different cell types that can be incorporated into the *k*-means groups. SVM in tissue studies has been first applied to binary-class problems.[21-23] Widjaja *et al*. conducted a pioneering study on a multi-class classification of Raman spectra of colonic tissues using SVM, which has since then been widely used not only for tissue classification but also for prediction.[17,24] However, no matter which multivariate statistical analysis (LDA, PCA, *K*-means clustering, or SVM) was chosen for feature extraction, all previous studies transformed tissue spectral data by using eigenvectors.[17,21-24] Such an approach, albeit useful, suffers from some limitations. For example, eigenvectors are highly dependent on the dataset itself, meaning that each dataset will generate its own eigenvectors. Furthermore, the lack of universal standards in data transformation generally prohibits comparison of results between different studies. Furthermore, it remains difficult to relate the transformed datasets back to any biological or biochemical context, thus making their interpretation somewhat impractical.[25] Thus, in this work, we focus on the use of IR biometrics to circumvent this last issue.

The first approach in this dissertation describes a point-detection technology using ATR probe to assess YES vs NO on cancerous versus non cancerous human tissues.

"Supervised" biometrics are used to build the TDM from thin tissue sections of metastatic liver lesion from colorectal cancer, using FTIR mapping with *k*-means clustering and SVM. Besides reducing data analysis time and improving accuracy, the combination of SVM with IR biometrics has the merit that it enables a standardization of high-dimensional spectral data reduction using supervised data transformation and, more importantly, it gives a mean of correlating spectral (chemical) information with histopathological content. With the ultimate goal of building up a set of universal biometrics for cancer diagnosis, the biometrics are further evaluated and the ones most specific to cancerous/non-cancerous cells are used in the TDM. Furthermore, we demonstrate that the use of an ATR-FTIR probe in combination with this model as a point-detection technology enables the rapid differentiation of cancerous and non-cancerous tissues, which further supports its intraoperative application as a real-time diagnostic tool to assess tissues *in vivo* during cancer surgery.

The first approach qualitatively distinguishes cancer-bearing from non-cancer bearing tissues using TDM built upon SVM. In the second approach, changes in the protein secondary structures in the tumor and non-tumor areas are quantitatively identified. The methodology is established by matrix multiplication of tissue spectra with

calibrant IR spectra (spectra dominant in α-helices and β-sheets extracted from Ramachandran plot). Different spectral ranges, (e.g., 1200–1800 cm^{-1} and 1500–1700 cm^{-1}) and interpolation steps (2 and 10 cm^{-1}) were examined to figure out the optimum plot of α-helix versus β-sheet scores that enabled to distinctly differentiates the tumor and non-tumor spectra of rectal adenocarcinoma metastatic to liver. Spectra obtained in the tumor region exhibit lower α-helix and β-sheet scores. The decrease is related to the downgraded level of albumin in the tumor region.

1.2 Dissertation Highlights

Chapter 2 is made up of two sections dealing with instrumentation and multivariate statistics, respectively. The instrumentation section starts by providing the theoretical background and instrumentation of FTIR spectroscopy, including that of the ATR probe. The following section describe two multivariate statistics used in study for spectral data interpretation: *k*-means clustering and SVM. *K*-means clustering analysis is mainly used in analyzing FTIR mapping data, whereas SVM is necessary for the TDM development which is used to discriminate cancer-bearing from non-cancer bearing tissue spectra taken directly on the liver lesion with the ATR-FTIR probe. Chapter 3 describes a TDM

for discrimination between cancer-bearing and non-cancer-bearing tissues. This TDM is built upon the FTIR image data which takes 12 hours to obtain, and it is used to assess YES vs NO on cancerous versus non cancerous spectra obtained by an ATR probe. Discrimination of ATR-FTIR data only takes less than a minute to accomplish, providing the possibility of using the ATR probe combined with the TDM during the cancer surgery to give real-time feedback on surgical resection margins. Chapter 4 begins by discussing the Ramachandran plot and protein secondary structures, and the methods developed to extract IR spectra dominant in α-helix and β-sheet using linear least square analysis. The following sections describe a quantitative method to analyze protein secondary structures based on extracted α-helix and β-sheet spectra using matrix multiplication. Finally, Chapter 5 summarizes the methodologies for cancerous and pre-cancerous tissues identification presented in this dissertation, and a future outlook for work in this area is included.

1.3 Collaborations

The metastatic liver specimens used in this study were collected during cancer surgery performed by Drs Edward W. Martin Jr and Stephen P. Povoski in the

department of surgical oncology at the Ohio State University Hospital. Dr. Charles L. Hitchcock helps the team with histopathological tissue assessment as well as provides guidance on tissue handling procedures. Professors James Coe and Heather Allen have incorporated chemistry and statistics in comprehensively and reliably assessing cancerous and non-cancerous tissues in real time to assist oncologists in surgical removal of malignant tumors.

The author has developed a novel TDM utilizing*k*-means clustering algorithms (Prof. James Coe) and has incorporated SVM into the TDM. Different feature extraction methods on data pattern recognition have been explored by the author and the one with the best decision accuracy while able to provide change in chemical composition of tissues was selected. The author wrote the matlab program to facilitate the data processing and optimized the model with soft SVM and radial basis function kernel. All the initial work of this TDM development was conducted based upon the FTIR image data obtained from one metastatic liver case where the author performed the data collection. This model has been tested on other metastatic liver cases for validation yielding the overall statistic accuracy of $95.4 \pm 5.4\%$ ($P < 0.1$). The FTIR image data are from Ryan Butch, Zhaomin Chen and the author. The variations between individuals are

minimized by developing biometrics that have the greatest sensitivity to cancerous and non-cancerous tissues identification. At the same time, these biometrics eliminate the redundant information in the dataset and therefore reduce the data processing time. As a next step, the author collected the tissue spectra directly on the metastatic liver tissue sections with an ATR probe, and identified the properties of the ATR-FTIR spectra (i.e. tumor VS non-tumor) using this TDM. In summary, the chemometric TDM was developed by the author, and by coupling the probe with this chemometric model, real-time identification of cancerous and non-cancerous tissues has been achieved.

As an associated study, alterations in protein secondary structures in cancerous and non-cancerous tissues have been investigated by the author. A chemometric method capable of differentiating tumor and non-tumor spectra based on the analysis of α-helix and β-sheet scores of spectra using matrix multiplication was also developed by the author. IR spectra dominated in protein secondary structures that have been used as calibrant spectra in matrix multiplication were extracted (Prof. James Coe) using linear least square analysis. The author facilitated the development of the corresponding algorithms. In addition, the author conducted data mining on FTIR images with principal component analysis, *k*-means clustering analysis and hierarchical clustering analysis to

evaluate the optimal strategies for hyper-dimensional data classification, and aided in the testing of the matlab programs. Additionally, the author was responsible for tissue acquisition and institutional review board (IRB) preparation.

Chapter 2: Fundamentals of FTIR and Multivariate Statistics

2.1 Theory and Instrumentation of FTIR Spectroscopy

2.1.1 FTIR Spectroscopy

Utilization of light in the IR region for studying matter started in 1905 by Coblentz, and the first IR spectrometer became available in the 1930s.[26] FTIR spectroscopy was under slow development because of the tedious calculations needed to transform the interferometric signal into a spectrum until the discovery of the fast Fourier transform (FFT) algorithm by Cooley and Tukey in 1964.[27] The next significant breakthrough happened in 1969 with the development of commercialized FTIR spectrometers by Digilab. Since then, together with the rapid development in computer technology, FTIR spectroscopy widened its application to different scientific fields.

FTIR spectroscopy makes use of the Michelson interferometer which advantageously combines high spectroscopic resolution with fast acquisition time. A schematic diagram of a Michelson interferometer is illustrated in Figure 2.1. The beam

from an IR source is split into two secondary beams by a beam splitter, which are later recombined and directed towards a detector. The path difference of these two beams is achieved by a moving mirror and measured as a function of the intensity. For an ideal beam splitter, each beam which can be represented by a time-harmonic electric field of amplitude E_0 and frequency ω are recombined at the detector and the total electric field (E) is given as

$$\begin{aligned} E(t) &= E_0 \sin(\omega t - \varphi_1) + E_0 \sin(\omega t - \varphi_2) \\ &= E_0[\sin(\omega t - \varphi_1) + \sin(\omega t - \varphi_2)] \end{aligned} \tag{2.1}$$

Squaring Eq.(2.1) gives

$$\begin{aligned} E^2(t) &= E_0^2[\sin(\omega t - \varphi_1) + \sin(\omega t - \varphi_2)]^2 \\ &= E_0^2 \left[1 + \cos(\varphi_1 - \varphi_2) - \frac{1}{2}(\cos 2(\omega t - \varphi_1) - \cos 2(\omega t - \varphi_2)) \right. \\ &\qquad \left. - \cos\big(2\omega t - (\varphi_1 - \varphi_2)\big)\right] \end{aligned} \tag{2.2}$$

Defining the phase difference $\delta = \varphi_1 - \varphi_2$ between the two beams and inserting into Eq.(2.2) yields

$$\begin{aligned} E^2(t) \; &= E_0^2 \left[1 + \cos\delta - \frac{1}{2}(\cos 2(\omega t - \varphi_1) - \cos 2(\omega t - \varphi_2)) \right. \\ &\qquad \left. - \cos(2\omega t - \delta)\right] \end{aligned} \tag{2.3}$$

The detected intensity is defined as the time average of Eq. (2.3)

$$I \; = \varepsilon_0 c \langle E^2(t) \rangle_t \tag{2.4}$$

Inserting Eq.(2.3) into Eq.(2.4) gives

$$I = \varepsilon_0 c E_0^2 \left[1 + \cos\delta - \frac{1}{2}(\langle \cos 2(\omega t - \varphi_1)\rangle_t - \langle \cos 2(\omega t - \varphi_2)\rangle_t) - \langle \cos(2\omega t - \delta)\rangle_t\right] \quad (2.5)$$

Because

$$\langle \cos(2\omega t - \varphi_1)\rangle_t = \frac{1}{\Delta t}\int_t^{t+\Delta t} cos2(\omega t' + \varphi_k)dt' = \frac{1}{2\omega\Delta t}\left[\frac{sin2(\omega t' + \varphi_k)}{2}\right]\Big|_t^{t+\Delta t} = \frac{1}{4\omega\Delta t}[sin2(\omega(t + \Delta t) + \varphi_k) - sin2(\omega t + \varphi_k)] \approx 0 \quad (2.6)$$

$$\langle \cos(2\omega t - \varphi_2)\rangle_t \approx 0 \quad (2.7)$$

$$\langle \cos(2\omega t - \delta)\rangle_t \approx 0 \quad (2.8)$$

then Eq.(2.5) reduces to

$$I = \varepsilon_0 c E_0^2 (1 + \cos\delta) = \varepsilon_0 c I_0 (1 + \cos\delta) \quad (2.9)$$

with $\delta = 2\pi x_{path}\nu$ where x_{path} is the path difference between the two beams. This is the equation of the interferogram.

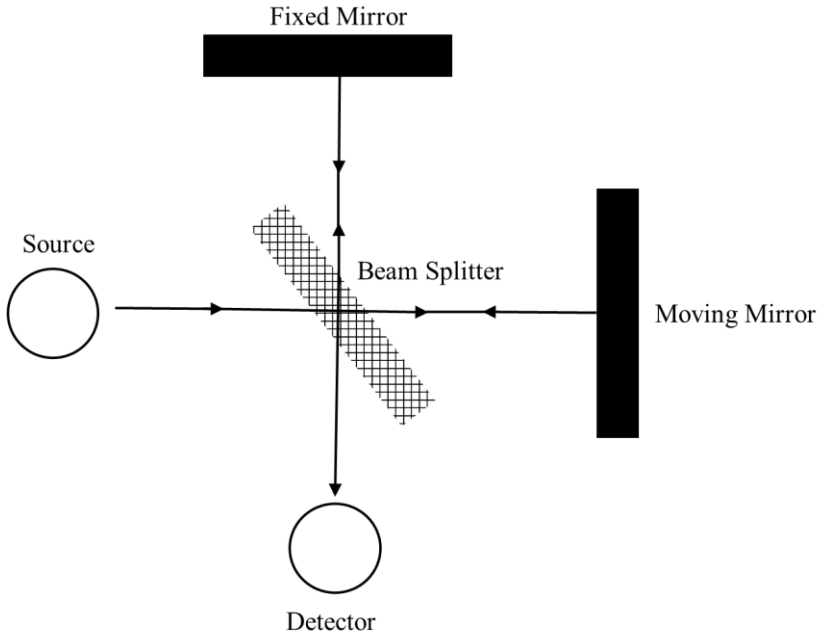

Figure 2.1 Schematic diagram of a Michelson interferometer.

To this point, the interferogram is generated by the Michelson interferometer, which is a function of the intensity versus the path difference between the two beams of the interferometer. The final IR spectrum is obtained by carrying out the Fourier transform of the interferogram (Eq.(2.9)):

$$S(\nu) = \int_{-\infty}^{\infty} I(x) \cos(2\pi x \nu)\, dx \tag{2.10}$$

2.1.2 FTIR Microscopy

In 1983, Muggli first coupled a microscope to a FTIR spectrometer. This was followed two years later by the development of the first commercialized FTIR microscope by Spectra-Tech. Since then, the FTIR microscope has become a widely used instrument in biological studies.[28] In an FTIR microscope, the IR radiation is generated by the source and then modulated by the interferometer of a FTIR spectrometer. Before the radiation passes through the sample, it is focused by lower Cassegrain optics. Subsequently, the transmitted radiation is collected by an upper Cassegrain before being directed to a mercury-cadmium-telluride (MCT) detector. The optical path of the IR beam for collecting an image in transmittance mode of a Perkin Elmer Spectrum Spotlight 300 FTIR microscope is illustrated in Figure 2.2.

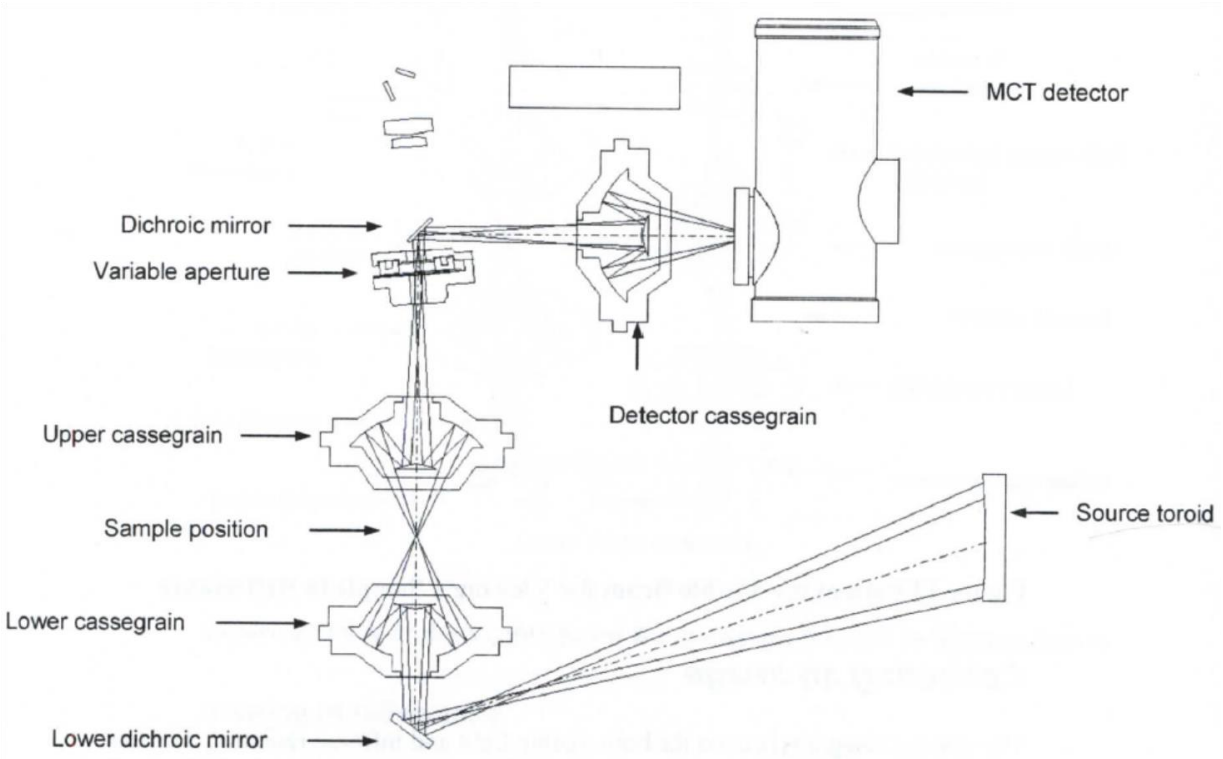

Figure 2.2 Schematic diagram of the IR optical path in a Perkin Elmer Spectrum Spotlight 300 FTIR microscope.[29]

2.1.3 ATR Probe

Application of FTIR spectroscopy to a real world problem requires the coupling to an ATR probe capable of direct examination of samples without further preparation. A typical ATR probe is made up of a shaft containing fiber optics for emission and collection of input and output IR beams and an ATR crystal, usually diamond, sealed in

the tip. Except the ATR crystal, the probe is usually covered and protected by a silicone covered metal sleeve. The React IR 15 (Mettler-Toledo) ATR probe used in this study for taking spectra on tissues is an example of a commercial ATR probe and is shown in Figure 2.3.

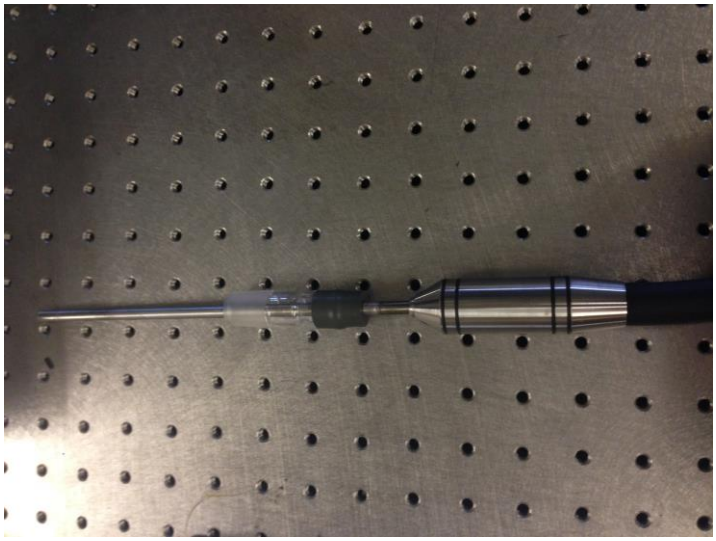

Figure 2.3 ReactIR 15 ATR probe from Mettler-Toledo.

As shown in Figure 2.4, when IR radiation propagates at or beyond a specific angle (called the critical angle) from an optically dense medium of refractive index n_1 (e.g., ATR crystal) to an adjacent medium of lower optical density ($n_2 < n_1$) (e.g., a biological

tissue) it will undergo total internal reflection at the interface, i.e., all radiation will be reflected back to the denser medium. Even though there is no net transmission of energy across the interface, the electromagnetic boundary conditions require the presence of an electric field along the interface. The wave generated from this field is known as an evanescent wave. The electric field of the evanescent wave decays exponentially as the distance increase from the interface:

$$E = E_0 e^{\frac{-2\pi}{\lambda_{\mathrm{IR}}}\left(\sin^2\theta - \left(\frac{n_{\mathrm{tissue}}}{n_{\mathrm{crystal}}}\right)^2\right)^{1/2} z} = E_0 e^{-z/d_p} \tag{2.11}$$

with the penetration depth d_p given by

$$d_p = \frac{\lambda_{\mathrm{IR}}}{2\pi\left[\sin^2\theta - \left(\frac{n_{\mathrm{tissue}}}{n_{\mathrm{crystal}}}\right)^2\right]^{1/2}} \tag{2.12}$$

and where E_0 is the electric field amplitude, λ_{IR} is the wavelength of the IR radiation in the denser medium, and z is the distance from the surface. For an IR wavelength in the range from 2.5 to 25 μm, and refractive indices n_{crystal} = 2.4 and $n_{\mathrm{tissue}} \approx 1.4$, the penetration depth of the evanescent wave is typically between 0.5 and 2 μm.

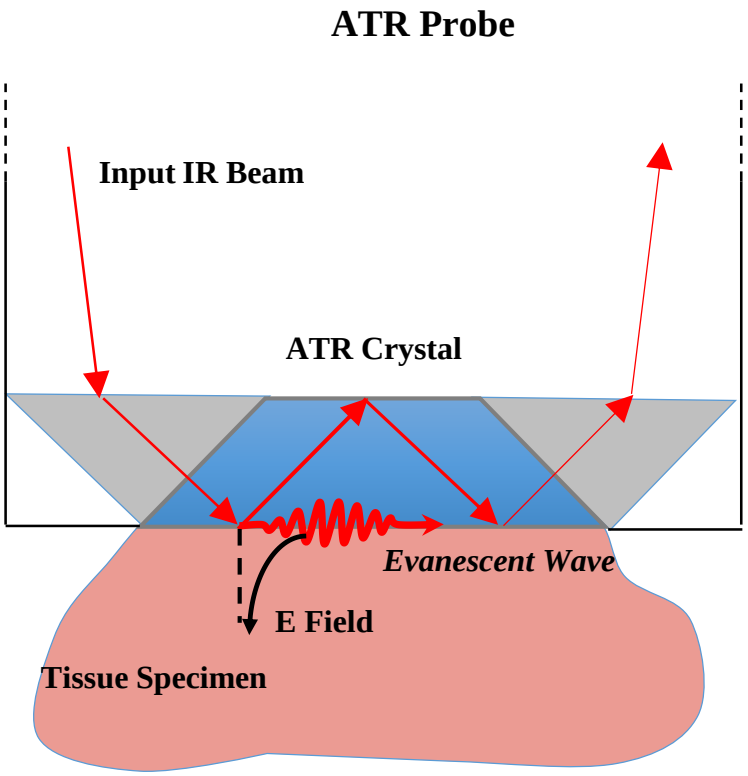

Figure 2.4 Schematic illustration of the ATR principle.

2.2 Theory of Multivariate Statistics

2.2.1 K-means Clustering

K-means clustering analysis partitions *X* observations into *k* clusters, with the restriction of minimizing the sum of the distances between the centroid *C* of each cluster and the location of each observation. To apply *k*-means clustering analysis to FTIR mappings, the FTIR imaging data is typically transformed through the process illustrated in Figure 2.5. The Matlab programs of *k*-means clustering analysis have been written by Professor James Coe et al. The detailed procedures of application of *k*-means clustering algorithm to analyze FTIR image data obtained from metastiatic liver tumor have been described in the paper "Infrared metrics for fixation-free liver tumor detection". [20]

The original FTIR data is stored in a three-dimensional matrix, data*(i, j, k)*, where *i* and *j* are the position indices, and *k* is the stepping index through the FTIR spectrum. Subsequently, the data is further reduced by ratioing two specific wavenumbers, i.e., by constructing biometrics. At this point, a new dataset*(i, j, m)* with *m* representing the index over the biometrics is generated. This three-dimensional dataset is finally reduced to a two-dimensional dataset *X(n, m)* by replacing *i* and *j* with *n*. The relationship between *i*, *j* and *n* is expressed as

$$n = (j-1)i_{max} + i \tag{2.12}$$

The process of *k*-means starts by randomly choosing the centroids, denoted as *C(k, m)*, then calculating the distance of all observations *X(n, m)* to the centroids. The sum of all the distances between the centroid of each cluster and each observation is given as

$$\sum_{\mathrm{k}} \mathrm{d}_{n\in G_k,k} = \sum_{k} \sum_{n\in G_k} \sqrt{\sum_{m} [X(n,m) - C(k,m)]^2} \tag{2.13}$$

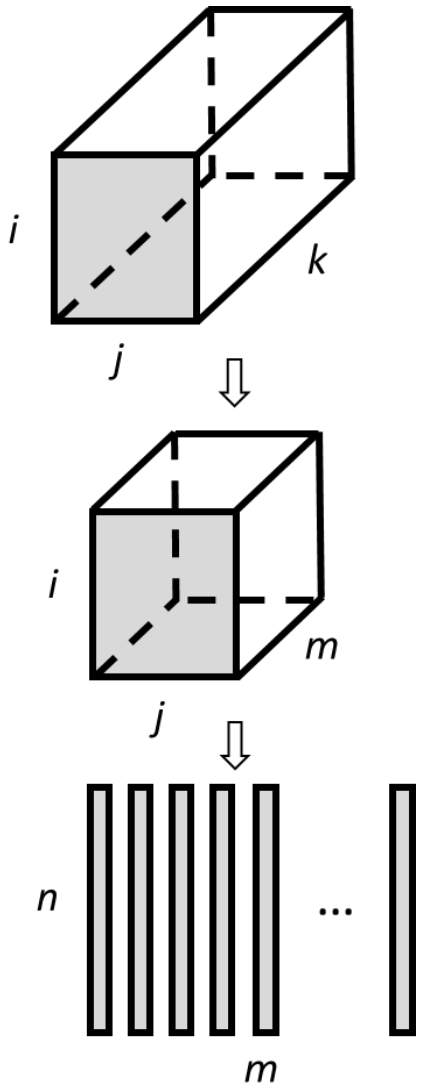

Figure 2.5 Process of FTIR imaging data transformation

Observations are assigned to the clusters based on the criterion of the shortest distance. After this, new centroids are determined by averaging the observations in each cluster. The process iterates until no observation has changed its cluster (Figure 2.6). Examples of *k*-means clustering analysis of metastatic liver cases are shown in Figure 2.7.

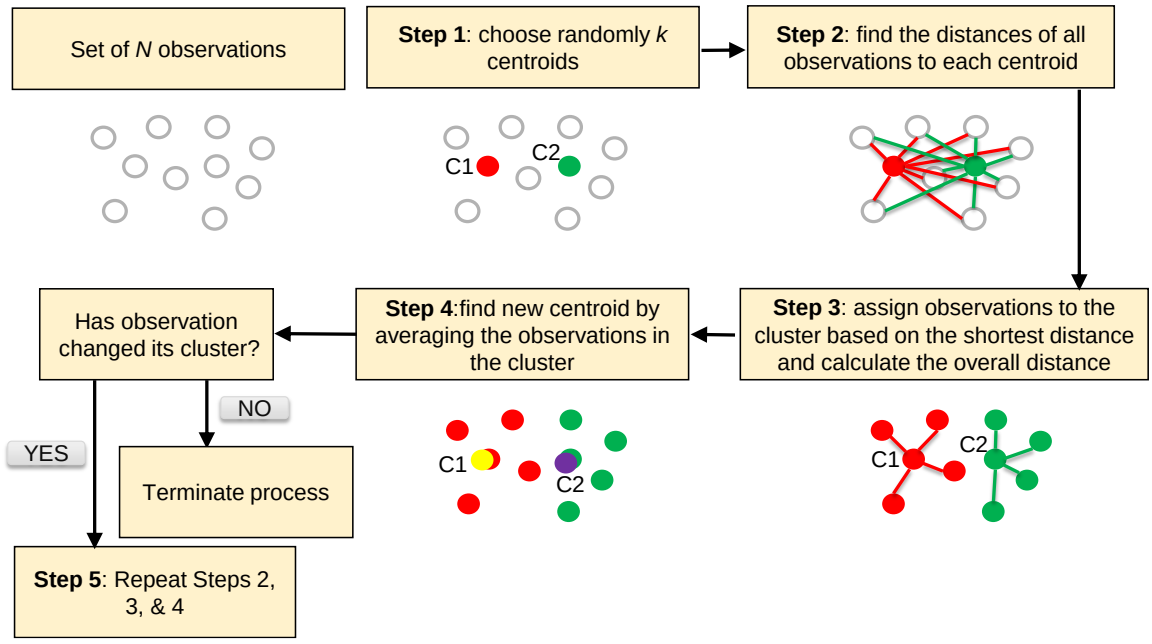

Figure 2.6 Flowchart of *k*-means clustering analysis.

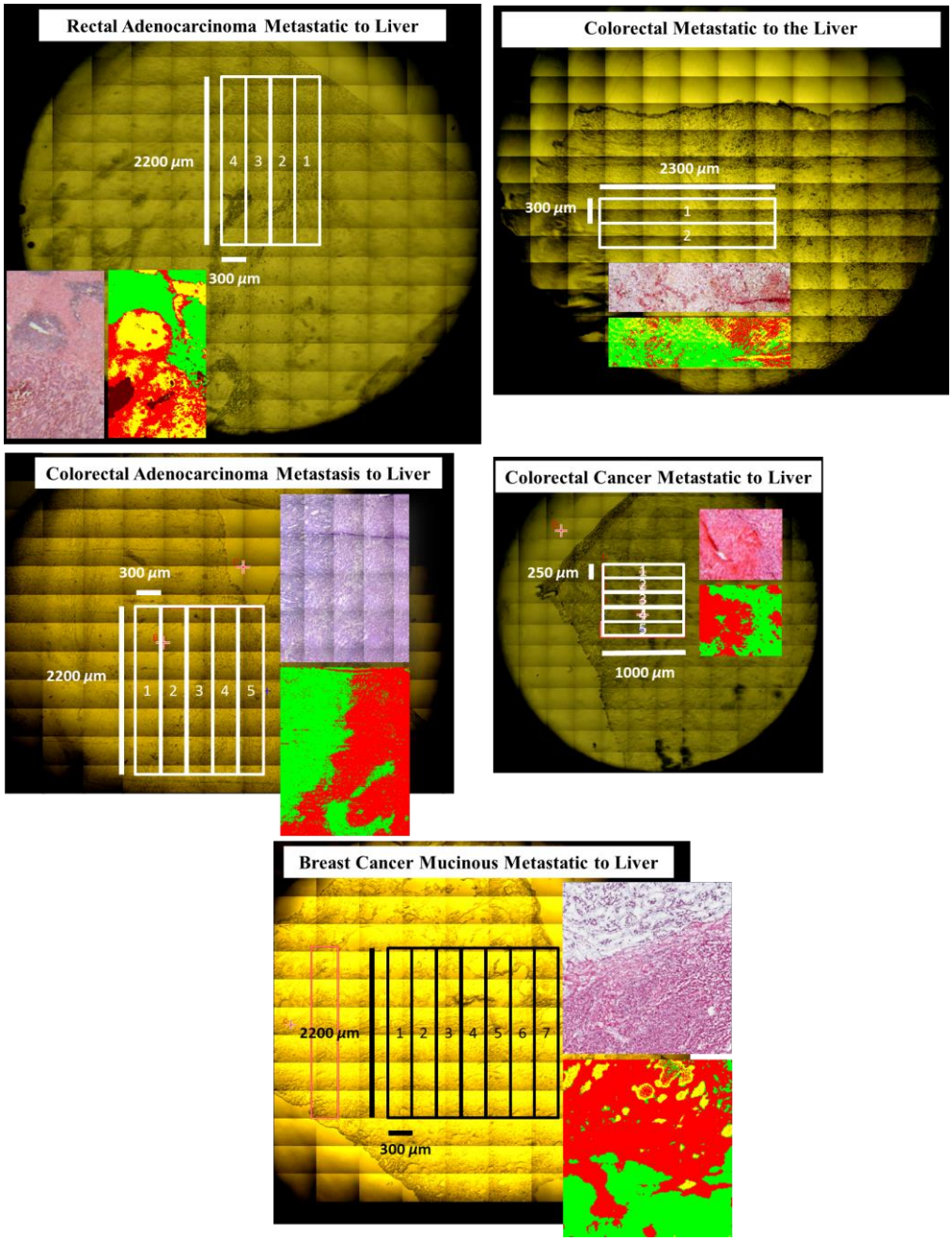

Figure 2.7 *K*-means clustering analysis of each metastatic liver case using a set of 64 IR biometrics, along with optical microscopic image and H&E-stained liver tissue specimen transferred onto ZnSe windows.

2.2.2 Support Vector Machine

SVM is a well-known supervised statistical learning method used for data analysis and pattern recognition. Details about the method can be found elsewhere.[30,31] Generally speaking, SVM uses support vectors which lies near the decision boundaries to construct a set of hyperplanes in high-dimensional space for classification or regression. An example of a linear binary-class classification using SVM is illustrated in Figure 2.8.

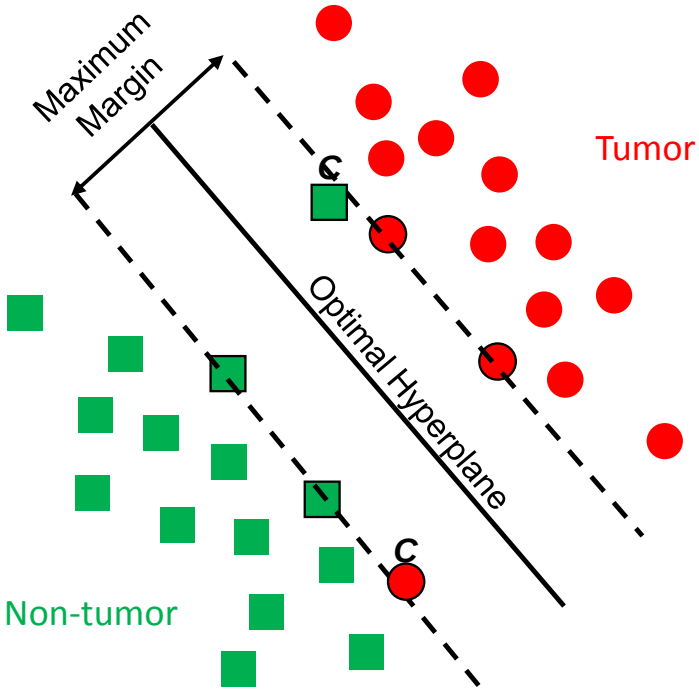

Figure 2.8 An example of a linear binary-class classification using soft SVM, i.e., a classifier that separates a set of observations into their respective groups (*green squares* (non-tumor) and *red circles* (tumor)) with a hyperplane (*full line*). Support vectors (*dashed lines*) are those lying near the margin. Penalty parameter (*c*) is incorporated in the development of SVM to allow misclassification of noise.

Assignment of unknown vectors or test set, is achieved by the decision equations build from the known data set or training set. For classification, SVM analysis of the training set generates decision functions of the form

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i\, K(\mathbf{x}_i, \mathbf{x}_j) + b\right) \tag{2.14}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function, and a_i and b are the coefficients and bias for each decision function, respectively. Whether the vector is assigned to the group or not is determined by the sign (*sgn*) of the decision equation as follows:

$$\text{if } f(x) \begin{cases} \geq 0, & \text{in group} \\ < 0, & \text{not in group} \end{cases} \tag{2.15}$$

In a multiple-class classification problem, several decision equations will be generated based on either of the two criteria: one-versus-one (OVO) or one-versus-all (OVA). In OVO approach, groups will be tested against each other and the number of decision equations generated is the combination of the group numbers, whereas in OVA, each group will be evaluated against all the other groups and the number of decision equations is the groups numbers.

Chapter 3: Development of a Tissue Discrimination Model using Supervised Data Transformation with Support Vector Machines

3.1 Overview

The proposed methodology is shown in the block diagram of Figure 3.1. The "supervised" biometrics (Table 3.1) are applied to build a TDM using FTIR imaging with *k*-means clustering and SVM, with the ultimate goal of building up a set of universal biometrics for cancer diagnosis. Besides reducing data analysis time and improving accuracy, the combination of SVM with IR biometrics has the merit that it enables a standardization of high-dimensional spectral data reduction using supervised data transformation and, more importantly, it gives a mean of correlating spectral (chemical) information with histopathological content. With the ultimate goal of building up a set of universal biometrics for cancer diagnosis, the used biometrics are further evaluated and the ones most specific to cancerous/non-cancerous cells are used in the TDM. Furthermore, we demonstrate that the use of an ATR-FTIR probe in combination with

this model as a point-detection technology enables the rapid differentiation of cancerous and non-cancerous tissues, which further supports its intraoperative application as a real-time diagnostic tool to assess tissues in vivo during cancer surgery.

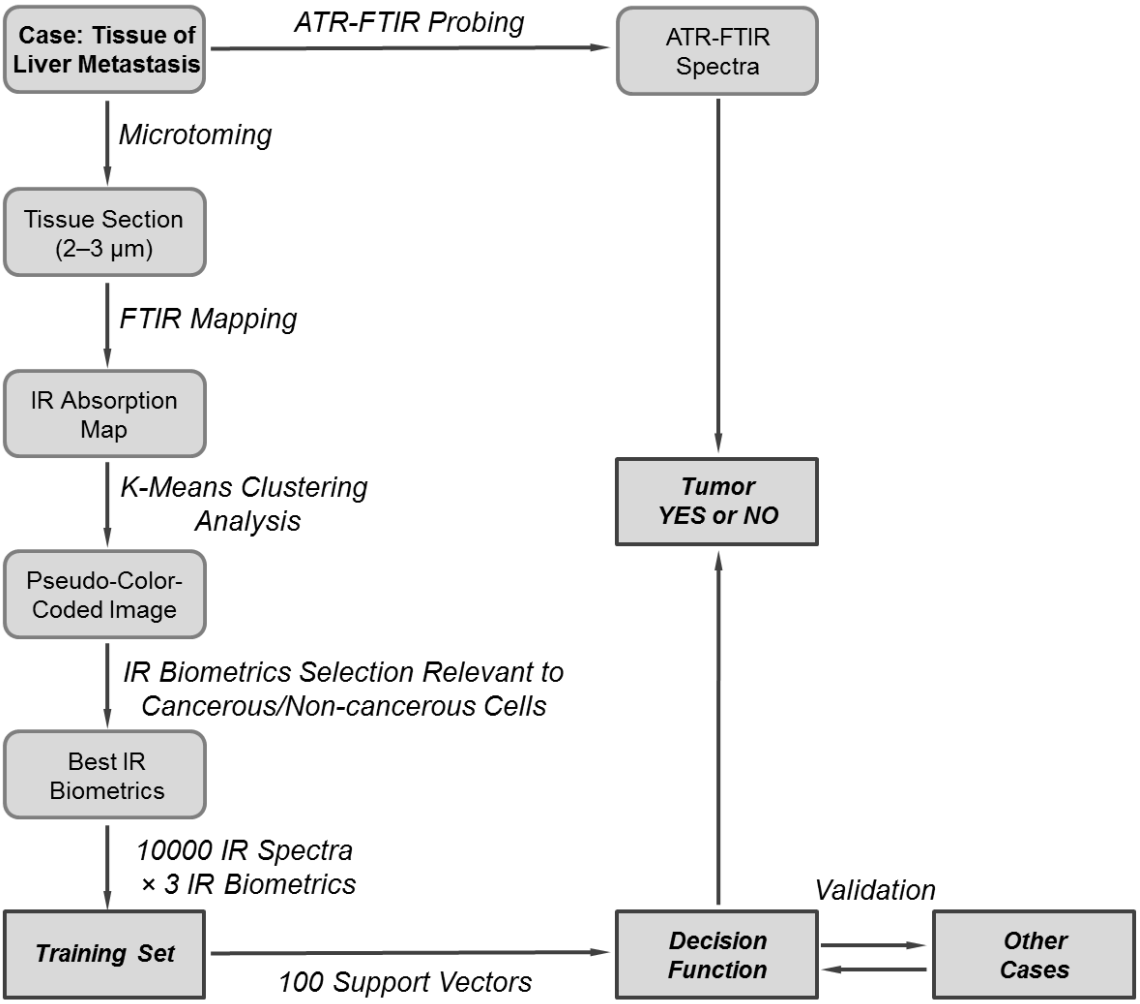

Figure 3.1 Block diagram of the proposed methodology.

3.2 Materials and Methods

3.2.1 Sample Preparation

Excised remnant liver tissues (~2.5 cm × ~2.5 cm × ~0.5 cm in size) containing both normal and cancer-bearing tissues were obtained from six different patients with metastatic colorectal cancer by the Department of Pathology at The Ohio State University (Columbus, OH) at the time of the patient's planned surgical procedure. Tissue acquisition and utilization was approved by the Institutional Review Board (No. 2011C0085). Immediately after collection, the tissue specimens were snap-frozen in liquid nitrogen (77 K) to preserve their structural integrity for further analysis. For the first five cases (cases 1–5), a thin (2–3 μm) tissue slice was obtained by cryo-sectioning and transferred onto an IR-transparent ZnSe window (Crystran Ltd., Poole, United Kingdom; 8.0 mm (diameter) × 1.0 mm (thickness)) mounted on a custom-built sample holder for FTIR mapping. After mapping, the slices were H&E-stained for comparison. Cases 1 and 6 were used for ATR-FTIR study.

3.2.2 FTIR Mapping

FTIR mapping was performed on an FTIR microscope (Spotlight 300, Perkin Elmer, Waltham, MA) used in transmission mode and equipped with a liquid nitrogen-cooled linear array of 16 HgCdTe (MCT) detectors. The mapped area of case 1 from which the TDM model was built is shown in the first image of Fig. 2.7. This area was divided into four rectangular regions. Each region was 2200 μm (length) × 300 μm (width) in size (i.e., 352 pixels × 48 pixels with 6.25 μm × 6.25 μm for each image pixel) and took about 3 hrs to scan. For cases 2–6, the sizes of the mapped areas are given in Table 3.2. For all mapped areas, sixteen spectra were averaged per pixel and each spectrum ranged from 750 to 4000 cm^{-1} with a spectral resolution of 4 cm^{-1}.

3.2.3 FTIR Imaging Data Processing

The data processing, including merging spectral data from all mapped windows into one data set, baseline correction, IR band ratioing, and *k*-means clustering analysis, was performed using custom-built routines written in Matlab (R2014, MathWorks, USA). The detailed procedures have been described in Section 2.2.1. All spectra were normalized using vector normalization

$$\bar{x}_{ij} = \sum_j x_{ij} \Big/ \sqrt{\sum_j x_{ij}^2}, \tag{3.1}$$

where the indices *i* and *j* refer to the *i*-th IR spectrum and *j*-th wavenumber in the spectrum. This normalization is an important and necessary step that accounts for variations in tissue thickness between sample slices and baseline shifts due to scattering.

3.2.4 TDM Construction

The TDM was built upon the FTIR image dataset where 10000 IR spectra obtained from the *k*-means clustering analysis of two groups (tumor and non-tumor; 5000 spectra/group) were chosen to build the training set. The original 1626-dimensional FTIR tissue spectra were reduced to 3 dimensions using a selected set of IR biometrics (Table 3.1). The selection of a biometric was based on its ability to differentiate specifically cancerous and non-cancerous cells.

Table 3.1 List of IR biometrics used in *k*-means clustering analysis. The biometrics selected for SVM analysis are boldfaced.

Biometric	Peak ratio	Biometric	Peak ratio
b1	1206/1544	b15	1516/1582
b2	1236/1544	b16	1588/1548
b3	1278/1544	b17	1520/1548
b4	1502/1544	b18	1600/1548
b5	1516/1544	b19	1620/1548
b6	1536/1544	b20	1632/1548
b7	1588/1544	b21	1556/1548
b8	1654/1544	b22	1252/1544
b9	1400/1390	b23	1100/1544
b10	1426/1450	b24	1070/1544
b11	1450/1544	**b25**	**1742/1256**
b12	1516/1236	b26	1556/1548
b13	**1744/1244**	b27	1252/1544
b14	**1744/1548**	b28	1472/1256

The training dataset was arranged as a 10000 (FTIR spectra) × 3 (IR biometrics) matrix and imported into SVM with soft margins classification using the LIBSVM 3.18 toolbox.[32] Soft margin SVM uses a penalty parameter *c*, which correlates inversely with the margin between groups, to indicate the degree of misclassification. For classification, SVM analysis of the training set generated a decision function of the form [30]

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i\, K(\mathbf{x}_i, \mathbf{x}_j) + b\right) \tag{3.2}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function, a_i and b are the coefficients and bias for each decision function, respectively. Hence, rather than using all 10000 spectra to generate the decision functions, the SVM analysis picks up 100 points lying near the boundary and uses them as support vectors (SV). The number of SV for tumor and non-tumor groups are 51 and 49, respectively. The generated decision functions will be used subsequently for test set prediction.

A Gaussian radial basis function (RBF) of the form [32,33]

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma|\mathbf{x}_i - \mathbf{x}_j|^2\right) \quad (\gamma > 0), \tag{3.3}$$

was selected as the nonlinear classifier, where $\mathbf{x}_i$, $\mathbf{x}_j$ are real-valued *n*-dimensional input training vectors and γ is the kernel function parameter. As pointed out by Widjaja et al., the RBF gives the best classification results and yields higher diagnostic accuracy compared to other nonlinear classifiers.[24] In order to find the *c* and γ values giving the highest accuracy in tissue discrimination, five-fold cross-validation was used. In five-fold cross-validation, the training set was first divided into five equal subsets and then one

subset was tested against the trained function inferred from the other four subsets. A grid search was employed in cross-validation in which various (c,γ) pairs were tried and the one with the best cross-validation accuracy was used in the subsequent analysis. In this work, five-fold cross validation yielded the best values c = 194.0117 and γ = 1.7411 with a best accuracy of 99.94%. These values were used in the final TDM. The model is further tested on other cases for validation.

3.2.5 Spectral Curve Fitting of Amide I and II Bands

The lineshapes of the amide I and II bands and their second derivatives were fitted simultaneously by a set of Lorentzian functions.[34] These new spectral curve fitting algorithms are developed by Professor James Coe et al. For each amide sub-band, three Lorentzian parameters were defined: the intensity ($p_{3(j-1)+1}$), the position ($p_{3(j-1)+2}$), and the half-width-at-half-maximum ($p_{3(j-1)+3}$), where j = 1, 2, …, n is the peak label. The fitted lineshape is given by

$$f(\tilde{\nu}) = \sum_{j=1}^{n} \frac{p_{3(j-1)+1}}{1 + \left(\frac{\tilde{\nu} - p_{3(j-1)+2}}{p_{3(j-1)+3}}\right)^2} \tag{3.4}$$

The second derivative at a given IR wavenumber $\tilde{\nu}_i$ was calculated with the nine-point central difference formula given by [35]

$$\delta_i^{(2)} = \left.\frac{d^2 A_i(\tilde{\nu})}{dx^2}\right|_{\tilde{\nu}=\tilde{\nu}_i} \approx \frac{-9A_{i+4} + 128A_{i+3} - 1008A_{i+2} + 8064A_{i+1} - 14350A_i + 8064A_{i-1} - 1008A_{i-2} + 128A_{i-3} - 9A_{i-4}}{5040(\Delta\tilde{\nu})^2} \quad (3.5)$$

where $\Delta\tilde{\nu} = \tilde{\nu}_2 - \tilde{\nu}_1$ is the uniform grid spacing and $A_{i \pm k}$ (k = 0, 1, 2, 3, 4) is the absorbance value at the IR wavenumber $\tilde{\nu}_i \pm k\Delta\tilde{\nu}$.

The standard deviations of the experimental lineshape ($\sigma_{\text{lineshape}}$) and second derivative ($\sigma_{\text{2nd derivative}}$) relative to their fitted counterparts were bundled up in a calculated cost function:

$$\sigma = \sigma_{\text{lineshape}} + |s|\sigma_{\text{2nd derivative}}, \quad (3.6)$$

where s is a scaling factor making the lineshape and second-derivative terms contribute equally to the fitting. For an ideal fitting, the cost function equals zero such that Eq. (3.6) reduces to

$$s = \left|\frac{\sigma_{\text{lineshape}}}{\sigma_{\text{2nd derivative}}}\right|, \quad (3.7)$$

Initially, s was set to a certain value. The initial guesses of Lorentzian parameters were optimized by the "fminsearch" function in Matlab and a new set of values for the Lorentzian and s parameters was obtained. This process was iterated until s remains unchanged, after which the optimal fit was achieved.

3.2.6 ATR-FTIR Probing and Tissue Discrimination

A diamond-tipped (~1 mm in diameter) ATR fiber probe (0.6 cm (diameter) × 150 cm (length)) coupled to a portable FTIR spectrometer (ReactIR^{TM} 15, Mettler Toledo, USA) equipped with a liquid nitrogen-cooled MCT detector was used to collect data on two excised liver tissue sections. For case 1, sets of 19 and 18 spectra were randomly taken on the non-tumor (dark red) and tumor (light red) areas, respectively. For case 6, ten spectra were taken on both tumor and non-tumor areas. Each spectrum was recorded for 64 accumulations and ranged from 900−1800 cm^{-1} with a resolution of 4 cm^{-1}. All ATR-FTIR spectra were corrected using Perkin Elmer's proprietary software to compensate for the frequency-dependent variation of the IR beam penetration depth into the tissue.

The SVM decision functions for ATR-FTIR spectra discrimination were optimized and tested as in the model construction. Similar to the training set, the original ATR-FTIR spectra were reduced to 3 dimensions using the same IR biometrics. Subsequently, the TDM was applied to assign the group to the test data set through decision functions, where the spectra were classified into tumor and non-tumor groups.

3.3 Results and Discussion

3.3.1 Four groups k-Means Clustering Analysis with 28 IR biometrics

Fig. 3.2a shows the optical microscopic image of an H&E-stained liver tissue specimen after FTIR mapping. Following histopathological examination, four distinct regions were identified: tumor, non-tumor, lymphocytes, and red blood cells. As a comparison, Fig. 3.2b shows the pseudo-color image of the same specimen obtained from *k*-means clustering analysis of four groups using the 28 IR biometrics (Table 3.1). Similarly to the histopathological assessment, *k*-means clustering analysis delineates four regions: a non-tumor region comprising 21,882 spectra, a tumor region comprising 25,191 spectra, a lymphocyte-rich region accounting for 16,324 spectra, and a region rich in red blood cells involving 4187 spectra. By comparing Fig. 3.2a and b, the boundaries between tumor and non-tumor regions are consistent with each other.

The FTIR spectra extracted from the four groups *k*-means clustering analysis are shown in Fig. 3.3. Primary IR bands at 1082, 1241, 1395, 1445, 1545, 1655, 1744, 2855, 2924, and 2958 cm^{-1} were consistently observed in all groups with the strongest absorption found at 1545 and 1655 cm^{-1}. Bands at 1082 and 1241 cm^{-1} correspond to the symmetric and asymmetric phosphate stretching modes ($\nu_{s/a}$(O-P-O)), respectively.

Spectra from red blood cells showed a less intense ($\nu_{s/a}$(O-P-O) bands. These bands originate from the phosphodiester backbone of cellular nucleic acids and phosphate groups of membrane lipids.[36,37] The small C-O stretching band at around 1165 cm^{-1} arises from C-O groups in glycogen's carbohydrate residues.[37] In the non-tumor spectrum, this band changes dramatically in both shape and position, red-shifting from 1165 cm^{-1} in the tumor and lymphocyte spectra to 1154 cm^{-1} in the non-tumor spectrum.

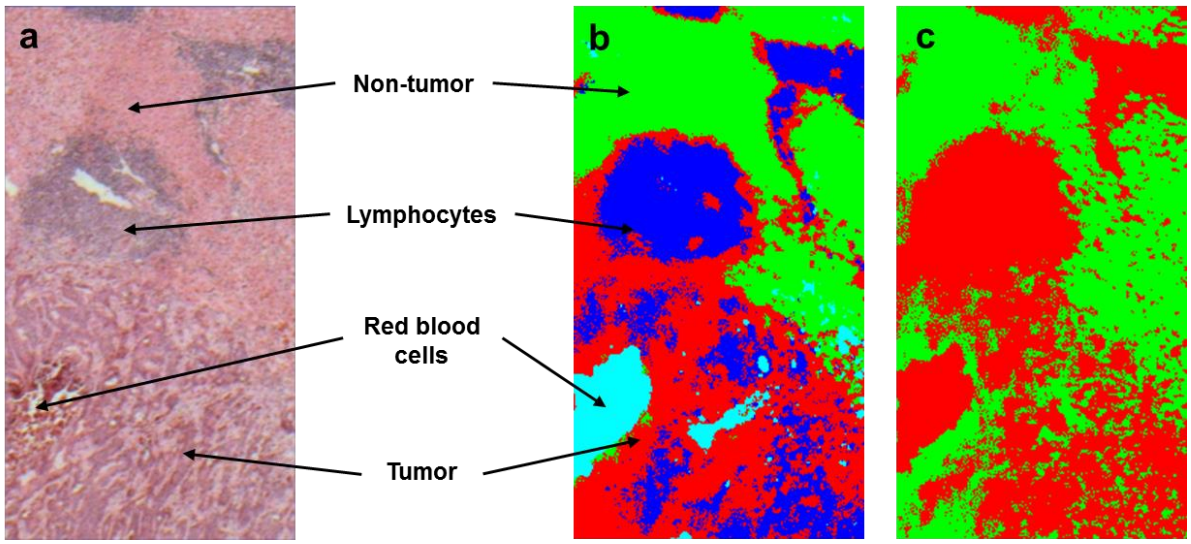

Figure 3.2 a Optical microscopic image of H&E-stained liver tissue specimen transferred onto a ZnSe window. **b** Four groups *k*-means clustering analysis with 28 IR biometrics: non-tumor (*green*), tumor (*red*), lymphocytes (*blue*), and red blood cells (*cyan*). **c** Two groups *k*-means clustering analysis with 3 IR biometrics: non-tumor (*green*) and tumor (*red*).

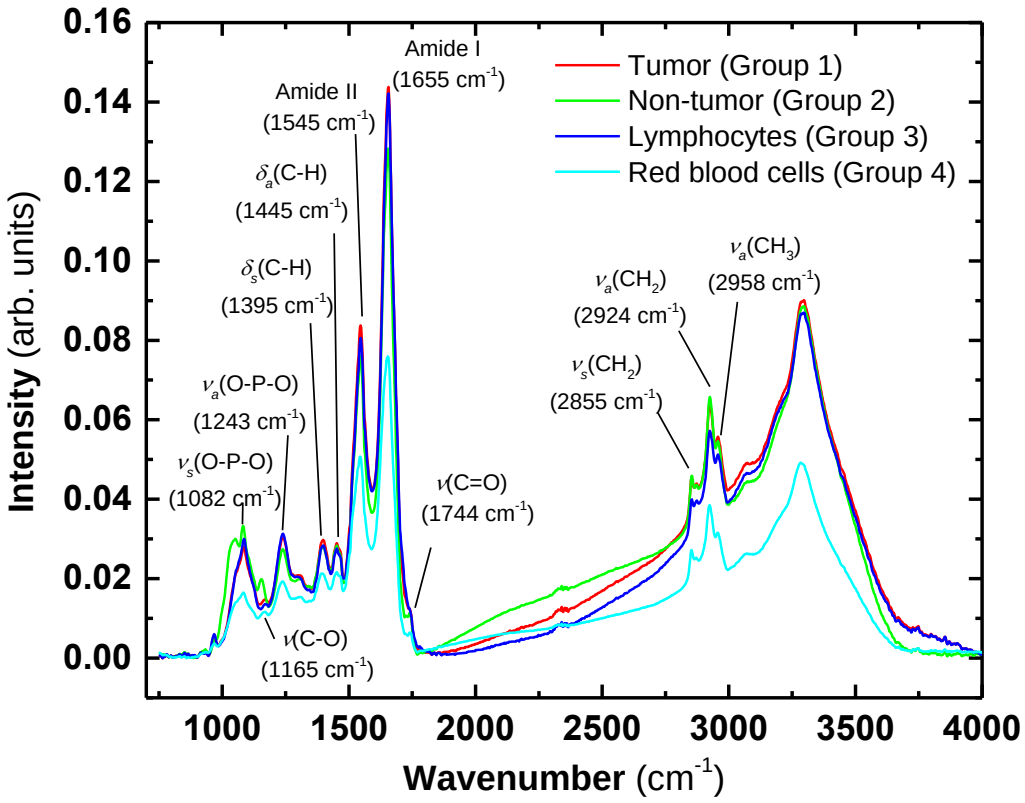

Figure 3.3 FTIR spectra of each of the four groups from the *k*-means clustering analysis of Fig. 3.2 The same color coding as in Fig. 3.2 was used.

The bands at 1395 and 1445 cm^{-1} are due to the symmetric and asymmetric C-H bending modes of the methylene (CH_2) groups ($\delta_{s/a}(CH_2)$), respectively. The red blood cells show the lowest peak intensity for these two modes. The fact that CH_2 groups are found abundantly in the acyl chains of fatty acids and lipids (e.g., phosphatidylcholines) commonly found in biological membranes suggests that the region rich in red blood cells has a lower fat content than other regions. The symmetric and asymmetric C-H stretches

of the CH_2 groups ($\nu_{s/a}(CH_2)$) are found at 2855 and 2924 cm^{-1}, respectively; the band at 2958 cm^{-1} is assigned to the asymmetric C-H stretching mode of methyl (CH_3) groups ($\nu_a(CH_3)$). The increase in the CH_2 and CH_3 bands intensity again indicates that more fat is present in the non-tumor. Finally, the band at 1744 cm^{-1} arising from the ester-linked C=O stretch has the highest intensity in the spectrum of non-tumor group. Because this band is insensitive to protein vibrational modes [36,38,39], it confirms that healthy tissue has typically a higher fat content than tumor. Overall, the spectrum of the lymphocytes shows some similarity with the spectra of tumor. Generally speaking red blood cells has lower molecular content compared to non-tumor, tumor and lymphocytes.

In addition, Fig. 3.3 reveals that the broad amide I (1600–1700 cm^{-1}) and amide II (1500–1600 cm^{-1}) bands are the most significant bands in the IR spectra of the liver tissue specimen. The amide I band is related to the protein backbone structures with primary contribution from C=O stretching and minor contribution from C-N stretching, whereas the amide II band is caused by N-H bending and C-N stretching vibrations. These bands involve unresolved protein sub-bands arising from the various vibrational modes of amino acids, which are sensitive to the local environment.[40,41]

In summary, non-tumor tissues have higher lipid content compared to tumor tissues, as the lipid bands at 1082, 1744, 2866, 2924 cm^{-1} are higher than those in tumor tissues. This is consistent with the results from other studies.[42-44]

3.3.2 Two groups k-means clustering analysis with 3 IR biometrics

In order to minimize the variations between different individuals, IR biometrics are evaluated to select the ones which have the greatest sensitivity to differentiate cancerous/non-cancerous cells. Biometrics sensitive to other features, e.g. red blood cells, are discarded. Fig. 3.4 shows the gray scale image of the 28 IR biometrics. Biometrics sets b4, b5, b6, b9, b17 and b12, b15, b28 are the ones sensitive to red blood cells and lymphocytes, respectively. However, among all biometrics, b13, b14, and b25 give the most distinctive contrast between tumor and non-tumor regions and are therefore used in the TDM development. Biometric b13 is the peak ratio of ester linked C=O stretch in lipids to O-P-O asymmetric stretch originating mainly from the phosphodiester backbone of cellular nucleic acids and phosphate groups of membrane lipids. b14 is the peak ratio of ester linked C=O stretch in lipids to amide II bands in protein. Similar to b13, b25 is the ratio of the same peaks while the peak position may shift in the tumor/non-tumor

regions. The results of *k*-means clustering analysis of two groups, tumor and non-tumor, using b13, b14, and b25 is shown in Fig 3.2c. The tumor region comprises 36,112 spectra, while non-tumor region involves 31,472 spectra.

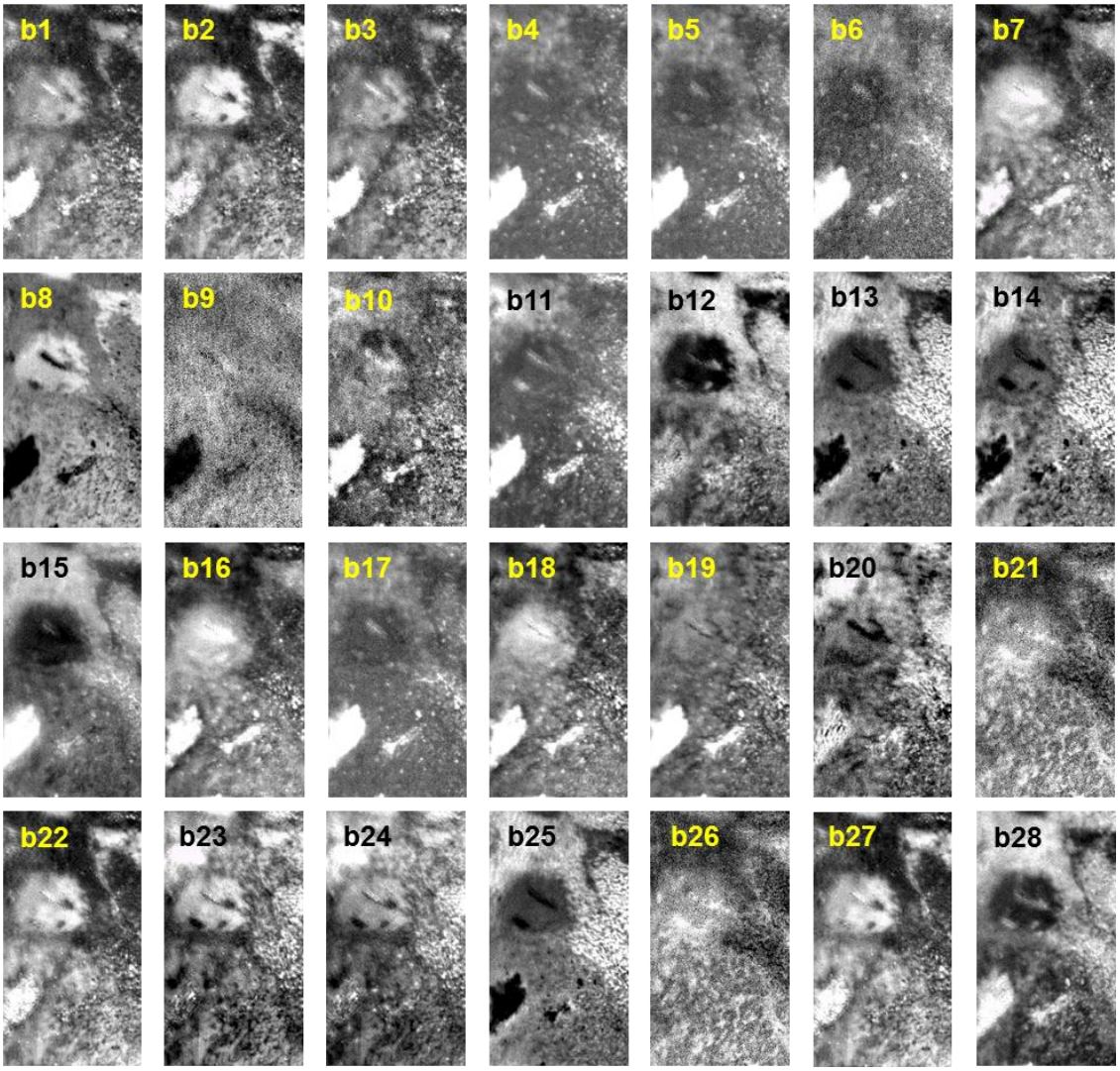

Figure 3.4 Gray-scale images of liver tissue specimen for each of the 28 IR biometrics found in Table 3.1.

3.3.3 TDM validation and predictions on ATR-FTIR spectra

The TDM is developed on a metastatic liver tumor specimen from one colorectal cancer case (case 1). Therefore, prior to applying the TDM to predict spectra obtained using the ATR probe, it is first tested on FTIR image data of four other cases (cases 2–5). All these cases are metastatic liver tumor also originating from the colon. By comparing the prediction from the TDM with the results from *k*-means clustering analysis, the accuracy found for four (cases 1, 2, 3, 5) of the five cases is greater than 94% (Table 3.2). The accuracy calculated on case 4 was slightly lower most likely because the tissue section contained more of the transition and tumor regions. However, using Student's *t*-test, the statistical accuracy was found to be 95.4 ± 5.4% ($P < 0.1$). This value compares well with results from other methods reported in the literature. For instance, using PCA on the IR spectra in the diagnosis of cervical cancer yielded overall 79% accuracy.[45] An accuracy of 88.6% was found by applying FTIR spectroscopy to the diagnosis of gastric cancer using 10 IR absorption bands as markers; this value could be increased to 92.2% with the help of SVM analysis.[46,16] Finally by combing PCA with SVM, even higher accuracies have been attained (99.8% and 96.4%, respectively) in the diagnosis of gastrointestinal malignancies and metastatic brain tumor.[24,17] The accuracy of TDM

could be further improved by using more functional biometrics. e.g., biometrics enabling the differentiation of proteins and lipids, proteins and polysaccharides, lipids and polysaccharides, etc.

Table 3.2 List of cases used for TDM validation.

Case	Mapped area (μm × μm)	Total number of spectra	Accuracy (%)
1	2200 × 1200	67584	99.89
2	2200 × 1500	84480	94.54
3	3300 × 1200	101376	99.16
4	2200 × 900	50688	85.95
5	1250 × 1000	48000	97.54

Fig. 3.5 shows FTIR spectra in the fingerprint and amide regions of tumor and non-tumor extracted from the TDM using biometrics b13, b14 and b25 and averaged ($P < 0.1$) over cases 1–5. While the amide I and II bands are nearly identical between tumor and non-tumor tissues, the ester-linked C=O stretch band at 1744 cm^{-1} is more intense in the non-tumor tissue. One major source of ester linkage in human tissues comes from phospholipids in cell membranes. In cancerous tissues, aberrant activation of sterol regulatory element-binding proteins (SREBPs) results in the reduced levels of membrane phospholipids.[47] This band could serve as a good marker for identifying cancerous and

non-cancerous tissues in colorectal metastatic liver tumors. A study using nuclear magnetic resonance (NMR) spectroscopy has shown that cholesterol, cholesterol esters, and phospholipids levels in human glioblastoma multiforme samples can be used as indicators in the tumor diagnosis.[48]

ATR-FTIR spectra were subsequently obtained on the remnant metastatic liver tissue section, specifically on two distinct areas respectively associated with non-tumor (dark red area) and tumor (light red area) tissues. By plotting biometrics against each other, e.g., b1 vs. b2 or b1 vs. b3, some classification of the spectra belonging to the dark and light red areas can be obtained (Fig. 3.6). However, even the use of biometrics does not completely separate spectra belonging to darker and lighter red areas. Therefore, to facilitate spectra differentiation SVM was used instead. The labels of the SVM predictions of two cases (cases 1 and 6) for the light and dark red areas are given in Table 3. If the decision value is positive, the spectra will be assigned as non-tumor and vice versa. For case 1, there are three and one discrepancies, respectively, in the light and dark red areas, with the accuracy of 89.19%. For case 6, there are four discrepancies in the light red area but only one in the dark red area, yielding 75.00% accuracy. Generally speaking, TDM gives more accurate diagnosis on the non-tumor regions, only one

discrepancies for both cases (94.73% for case 1, and 90.00% for case 6). The tumor regions actually contains both tumor and transition regions, and some parts in the transition regions might still be considered as non-tumor, resulting in some discrepancies in the tumor regions.

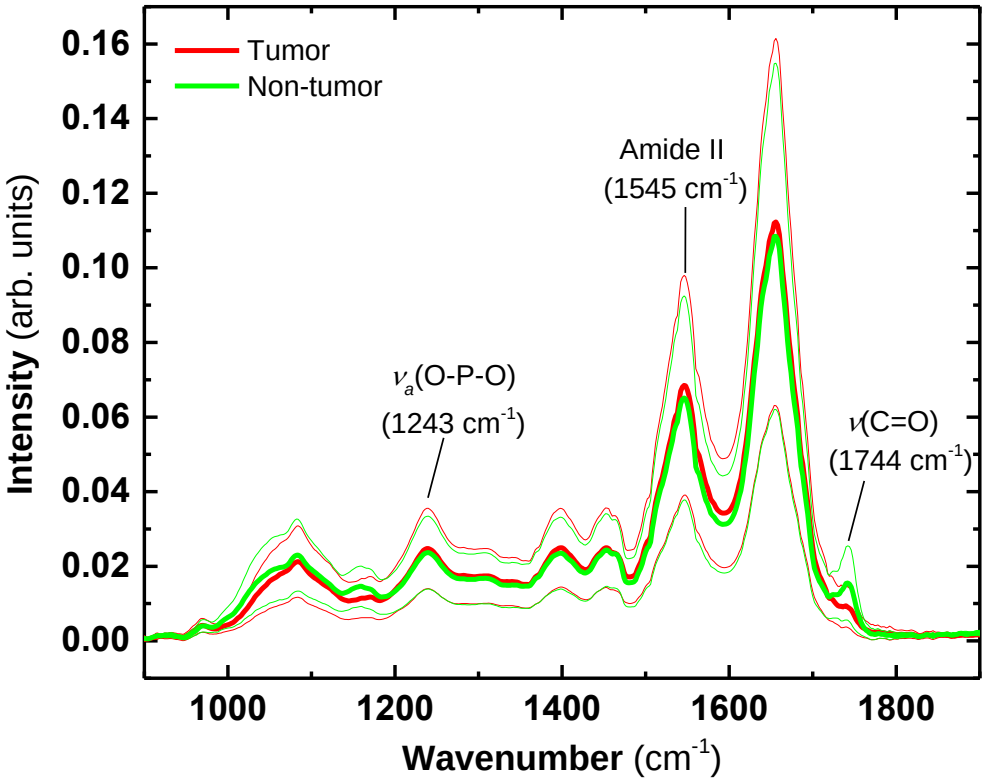

Figure 3.5 Averaged FTIR spectra ($P < 0.1$) in the fingerprint and amide spectral regions of tumor and non-tumor from each case extracted from the TDM using the 3 selected biometrics.

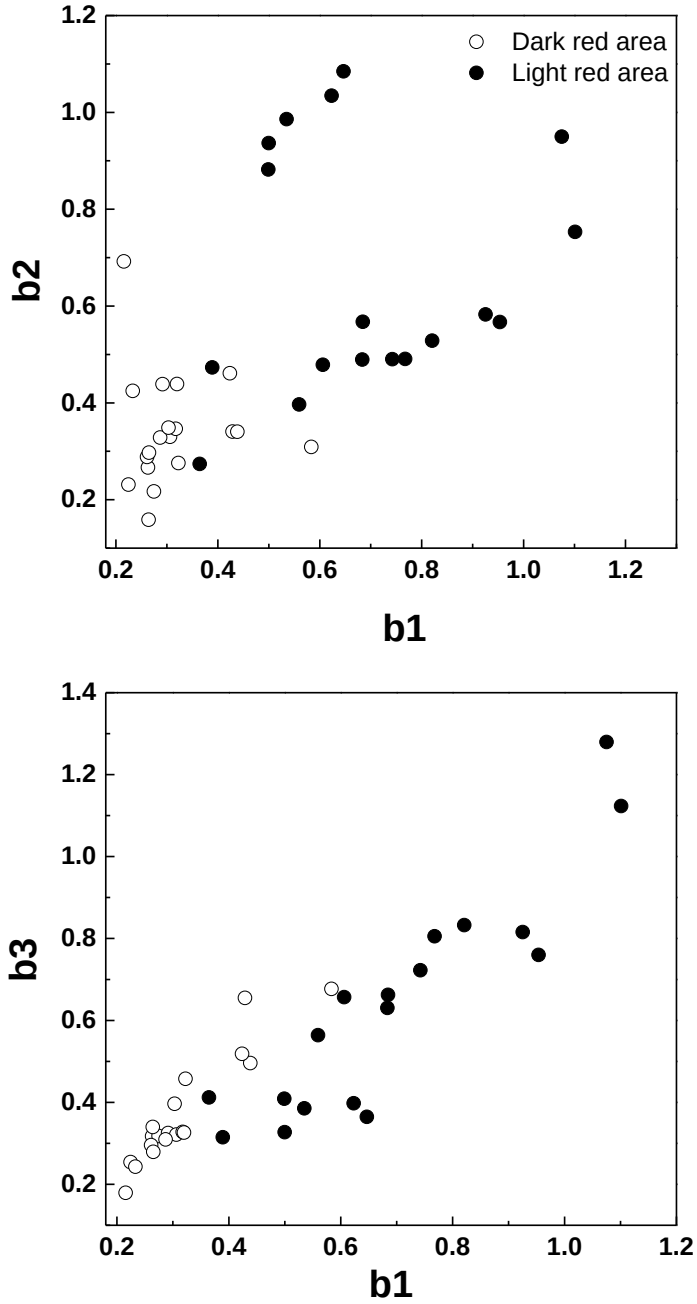

Figure 3.6 Examples of biometrics discrimination plots of ATR-FTIR spectra from the dark and light red areas of the remnant metastatic liver tissue section. **a** b1 versus b2. **b** b1 versus b3.

Table 3.3 TDM predictions from ATR-FTIR spectra obtained on cases 1 and 6. Group labels: (1) non-tumor, (2) tumor.

	Spectrum																		
Case 1																			
LRA	2	2	2	2	2	2	2	2	2	2	2	2	**1**	2	2	**1**	**1**	2	
DRA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**2**	1	1	1	1
Case 6																			
LRA	**1**	**1**	2	2	2	**1**	**1**	2	2	2									
DRA	1	1	1	1	1	**2**	1	1	1	1									

3.3.4 Simultaneous Fitting of Lineshape and Second Derivative of Amide I and II Bands

An example of simultaneous fits of lineshape and second derivative of a non-tumor group is shown in Figure 3.7a. The fitting process started with s = -40 and ended with s = -55. Fits in this specific group involve 40 peaks. The final values of both $\sigma_{\text{lineshape}}$ and $\sigma_{\text{2nd derivative}}$ are approximately equal to zero, indicating a good fit. This is also illustrated in Figure 3.7a where both the fitted lineshape and second derivative match well with the experimental data ($\sigma < 10^{-6}$). The integrated band intensities are plotted against sub-band positions for three pairs of *k*-means groups, i.e., tumor vs. non-tumor, tumor vs. lymphocytes, and non-tumor vs. lymphocytes. (Fig. 3.7b–d).

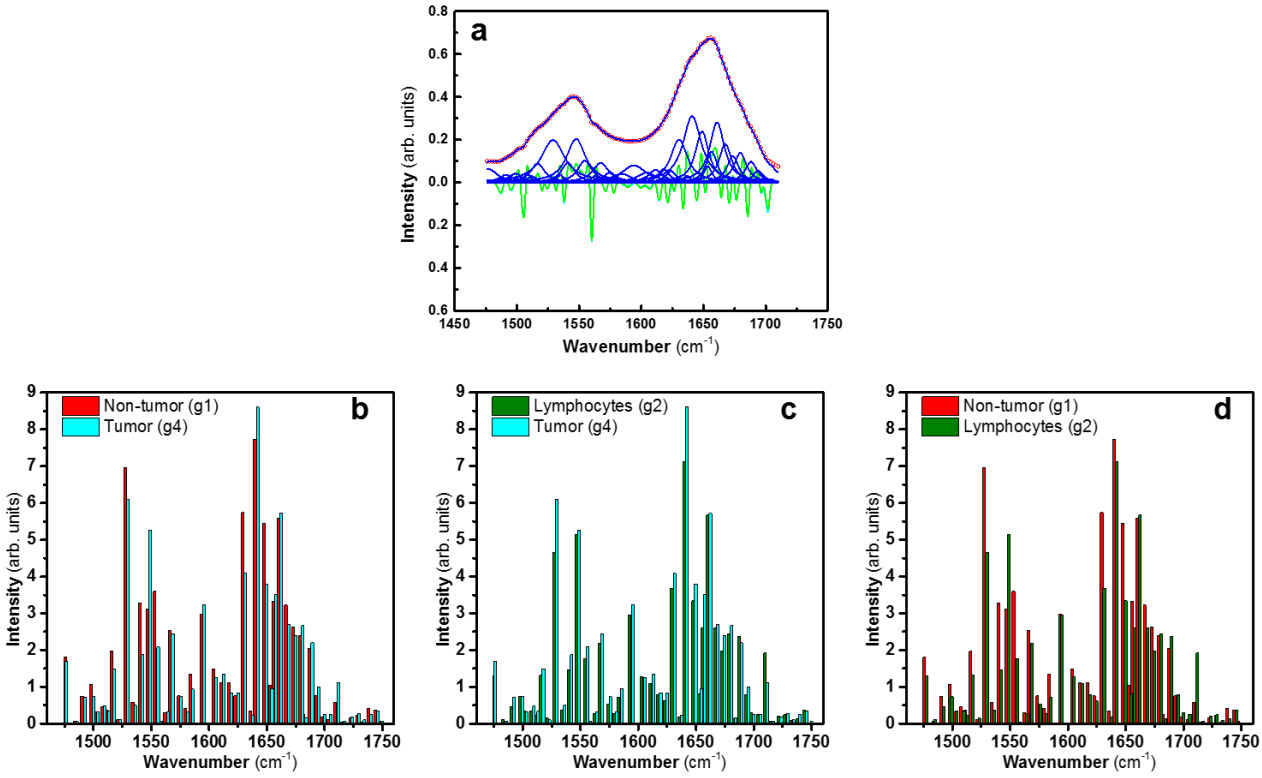

Figure 3.7 a Simultaneous fits of lineshape (raw data (*solid red*) and fit (*dashed blue*)) and second derivatives (raw data (*solid orange*) and fit (*dashed cyan*)) of amide I and II bands. The Lorentzian sub-peaks that sum to the fitted lineshape are also shown in blue. **b–d** Integrated band intensities against frequencies for three pairs of *k*-means groups, i.e., tumor (*cyan*) vs. non-tumor (*red*), tumor vs. lymphocytes (*green*), and non-tumor vs. lymphocytes.

In Figure 3.7b, most of the dominant protein sub-bands in the non-tumor area have greater intensities than those in the tumor area, except at 1550 and 1642 cm^{-1}. Comparing protein sub-bands of tumor to lymphocytes (Fig. 3.7c), the intensities in the tumor-area are higher than those in the lymphocyte-rich region. Similarly, protein sub-bands in the

non-tumor area have also greater intensities than those from the lymphocyte-rich region, except for the band at 1550 cm^{-1} (Fig. 3.7d). The most intense sub-bands of these groups are located at 1528, 1546, 1630, 1640, 1648, and 1660 cm^{-1}. These bands are attributed to protein secondary structures, particularly α-helices (1546 and 1648 cm^{-1}), β-sheets (1630 and 1640 cm^{-1}), and β-turns (1528 and 1660 cm^{-1}).

3.4 Conclusions

The combination of ATR-FTIR probing with a TDM is a promising diagnostic tool in cancer detection. With the FTIR microscopic imaging, relevant information on histopathology can be obtained. The tumor, non-tumor, and other regions are identified and classified by *k*-means clustering analysis. Biometrics are evaluated and a TDM was built to further discriminate the data taken with ATR probe. This TDM has the advantage that it minimizes the variations among individuals by using the biometrics only related to cancerous/non-cancerous cells. A logical extension of the methodology developed here could be improved by using more functional biometrics, e.g., biometrics enabling the differentiation of proteins and lipids, proteins and polysaccharides, lipids and polysaccharides, etc., with the ultimate goal of establishing a set of universal biometrics that can be used for data transformation. In addition, the entire procedure is label-free and objective as it does not rely on the judgment of pathologists. In future work, more specimens will be collected to further optimize the training set and to validate the model with the ultimate goal of using it in intraoperative applications.

Chapter 4: Detecting Metastatic Liver Tumors using Alpha-Helix and Beta-Sheet Scoring

4.1 Overview

The IR spectra of proteins are sensitive to their secondary structures; proteins dominated by α-helices (e.g., myoglobin) have Amide I (1600−1700 cm^{-1}) and II (1500−1600 cm^{-1}) bands altered more significantly than those that are abundant in β-sheets (e.g., concanavalin A).[49-52] Amide I and II bands involve a number of unresolved sub-bands arising from various amino acid vibrational modes. The Amide I band is related to the protein backbone structures with primary contribution from C=O stretching and minor contribution from C−N stretching, whereas the Amide II band is associated with N−H bending and C−N stretching vibrations. The correlation between a protein secondary structure fractions and its IR spectrum has been a topic intensely studied since 1950.[53,54,52] Most of these studies focused on qualitative analysis, including investigating shape and intensity alterations in Amide I and II regions as

proteins change their secondary structure, assigning sub-bands to different secondary structures, comparing IR spectra of proteins to standards, and developing strategies for deconvolution curve fitting. Quantitative analyses have been limited to area determination of sub-bands from curve fitting.

In this study, the prediction of tissue IR spectra obtained on metastatic liver lesion was achieved through analysis of α-helix and β-sheet scores using matrix multiplication with calibrants extracted from Ramachandran plot. By plotting α-helix against β-sheet scores, ATR-FTIR spectra obtained in the tumor region of colorectal cancer metastatic to liver were clearly separated from those from the non-tumor region. Reducing the number of IR metrics by only using a point every 10 cm^{-1} between 1500 and 1700 cm^{-1} does not deteriorate the separation of tumor and non-tumor and opens the possibility of applying quantum cascade IR lasers instead of a broadband IR source in future work. The results demonstrated that the use of an ATR-FTIR probe in combination with this approach allow high prediction accuracy for cancer-bearing tissue identification, which further supports its future intraoperative application as a real-time diagnostic tool to assess tissues *in vivo* during cancer surgery. To the authors' knowledge, this is the first time that

a methodology based on the quantitative analysis of protein secondary structures has been applied to differentiate tumor from non-tumor tissues.

4.2 Materials and Methods

4.2.1 Calculated IR Spectra of Protein Secondary Structures

The calculated IR spectra dominated in protein secondary structures were extracted using database protein standards of Dong, Carpenter, and Caughey. A Matlab routine written by Professor James Coe et al facilitated the extraction process of calibrant spectra dominated in protein secondary structures.[55] The database consists of 40 short-chain proteins (55–757 amino acids) whose identities are given in Table 4.1. These IR spectra were obtained from ~5 mg protein/ml protein solutions in a 10 mM phosphate buffer solution at pH 7.3. These spectra were recorded with a 6 μm pathlength over the range of 1200–2000 cm^{-1} at 4 cm^{-1} resolution. Phosphate buffer solution were taken as background and subtracted from the original spectra.

The α-helix and β-sheet dominated spectra are obtained using linear least squares to correlate the absorbance of each library proteins at a specific wavenumber and the fractions of amino acids in each secondary structures (i.e. α-helix, β-sheet and others). Using 1200 cm^{-1} as an example, the linear least squares relation is

$$\begin{bmatrix} y_{1,1200\ cm^{-1}} \\ y_{2,1200\ cm^{-1}} \\ \vdots \\ y_{m_s,1200\ cm^{-1}} \end{bmatrix} = \begin{bmatrix} x_{\alpha,1} & x_{\beta,1} & x_{O,1} \\ x_{\alpha,2} & x_{\beta,2} & x_{O,2} \\ \vdots & \vdots & \vdots \\ x_{\alpha,m_s} & x_{\beta,m_s} & x_{O,m_s} \end{bmatrix} \begin{bmatrix} b_{\alpha,1200\ cm^{-1}} \\ b_{\beta,1200\ cm^{-1}} \\ b_{O,1200\ cm^{-1}} \end{bmatrix}, \quad (4.1)$$

where the left-hand column of *y* values contains the absorbance of each library protein at the selected wavelength, the *x* values are the fractions of amino acids in each secondary structure group for each protein, and the *b* values are the IR spectra of the secondary structure groups at the selected wavelength. The method models each library protein's spectrum with a linear combination of the fractions and the secondary group spectra, $y_i = x_{\alpha,i}b_\alpha + x_{\beta,i}b_\beta + x_{O,i}b_O$. Upon extending the ordinary least squares procedure to all wavenumbers, then the relation becomes a multivariate regression which in matrix form is

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{B} \quad , \tag{4.2}$$

where the matrix **B** contains the IR spectra of the groups of protein secondary structures as rows, just as the **Y** matrix (defined earlier) contains the library protein IR spectra as rows. The number of rows in both **X** and **Y** is the number of library proteins ($m_s = 40$), while the number of columns in **Y** and **B** is the number of steps in the IR spectra ($n = 301$) in this work. There exist a variety of multivariate statistical analyses[56-60] for extracting information from IR spectra, however the strength of this work arises from its connection to the Ramachandran plot, not the mathematics. Its validity follows from three stages of error analysis, including calculations without weights, using weights from

the baselines of the library input spectra, and using covariance between the input library spectra which are highly correlated. The general least squares solution to equation (4) in matrix form is

$$\widehat{\mathbf{B}} = (\mathbf{X}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^{\mathrm{T}} \cdot \mathbf{Y} \quad , \tag{4.3}$$

where the 'hat' indicates a fitted result and **W** is the weighting matrix which is a square matrix of dimension m_s x m_s. The matrix **W** equals the identity matrix for unweighted least squares (**W=I**) and it has the reciprocal of each library spectrum's variance for weighted least squares

$$\mathbf{W} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \ldots & 0 \\ 0 & \frac{1}{\sigma_1^2} \ldots & 0 \\ \vdots & \vdots \quad \ldots & \vdots \\ 0 & 0 \ldots & \frac{1}{\sigma_{m_s}^2} \end{bmatrix} \quad . \tag{4.4}$$

These weights were chosen for the library spectra by calculating the standard deviation of the baseline noise, σ_i, of each library spectrum in a baseline region from 1840-1920 cm^{-1} of the normalized spectrum. The values of σ_i range from 0.00003-0.00029 in absorbance units of the normalized library spectra. These can be compared to the normalized average absorbance at 1654 cm^{-1} of 0.16 normalized absorbance units (see Figure 2a) giving errors of ~0.09% for amide I band of the input library spectra. The most general least squares approach takes account of the significant correlation between the input

library spectra. The correlations between different pairs of input library spectra varies from 0.802 to 0.998. This case is called a general least square problem or a least squares fit with covariance. In such a case, **W** is a nondiagonal matrix with correlation coefficients between each pair of library protein spectra in the off-diagonal positions. The general least squares problem is solved formally by decomposing the **W** matrix into two matrices by QR factorization, which in turn are used to reweight the **X** and **Y** matrices in such a way that the whole problem can be rewritten as a simple least squares (Stang and Graybill, MATLAB). Once the results of equation (4) are calculated for any of the three options with **W**, then the library protein spectra are calculated with $\widehat{\mathbf{Y}} = \mathbf{X} \cdot \widehat{\mathbf{B}}$, where the "hats" indicate fitted values. Since both **X** and **Y** are normalized quantities, it can be presumed that the output group spectra $\widehat{\mathbf{B}}$ are also normalized. In fact, the use of a group with small fractions does produce a raw solution with high absorbance. The raw solutions and their errors have been multiplied by the amino acid weighted fractions of the corresponding secondary structure groups to compensate for this effect.

There are error statistics to consider for the fitting of the spectra of both the library proteins and the protein secondary structure groups. The error statistics for the

library spectra involve the rows of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ and the variances for each library spectrum are

$$\sigma_{Y,i}^2 = \frac{[\mathbf{Y}(i,:)-\widehat{\mathbf{Y}}(i,:)]\cdot[\mathbf{Y}(i,:)-\widehat{\mathbf{Y}}(i,:)]^T}{m_s-n_g} \quad , \tag{4.5}$$

where $i = 1,2,\ldots m_s$ is an index over the library spectra. The notation $(i,:)$ means all of the elements across row i, so this amounts to a sum of the errors squared across the IR spectrum for each library protein. The error statistics for the fitted group spectra $\widehat{\mathbf{B}}$ of protein secondary structures involve the columns of $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ and are given as a mean square of errors at each wavenumber in the spectrum as

$$mse_j = \frac{\mathbf{Y}(:,j)^T\cdot\mathbf{W}\cdot\mathbf{Y}(:,j)-\mathbf{Y}(:,j)^T\cdot\mathbf{W}\cdot\widehat{\mathbf{Y}}(:,j)}{n-n_g} \quad , \tag{4.6}$$

where $j = 1,2\ldots n$ is an index for the wavenumbers in the spectrum. The notation $(:,j)$ means all of the elements down the column j, so this is an assessment across the library proteins at each wavenumber. The variance-covariance matrix for the $\widehat{\mathbf{B}}$ parameters is calculated at each wavenumber step of the spectrum as

$$\widehat{\mathbf{V}}_j = (\mathbf{X}^T\cdot\mathbf{W}\cdot\mathbf{X})^{-1}mse_j \quad , \tag{4.7}$$

where again $j = 1,2\ldots n$ which steps through the wavenumbers. The estimated standard deviations of the fitted spectra of protein secondary structures are obtained from the square root of the diagonal elements of $\widehat{\mathbf{V}}_j$ at each wavenumber (index j steps through

wavenumbers). To summarize, the inputs are **X, Y**, **W** and the outputs are $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{Y}}$ and their errors.

Table 4.1 Secondary structure fractions for the 40 proteins of the database of Dong, Carpenter, and Coughey.

[a] Name of the IR spectrum from the library of Dong, Carpenter, and Caughey. [b] RSCB Protein Data Bank file number. [c] Number of amino acids for which there are specific dihedral angle values (not the actual protein chain length). [d] The fractions of α-helix and β-sheet calculated using linear least squares.

name[a]	protein (source)	pdb[b]	#aa[c]	α-fractio	β-fractio
a1pi	α1-Proteinase Inhibitor (human)	1KCT	375	0.0880	0.0800
bsa	Albumin (bovine serum, A-0281 Sigma)	4F5S	583	0.7204	0.0000
albumnhu	Albumin (human serum)	1E7I	582	0.7096	0.0000
alcdehho	Alcohol Dehydrogenase (equine liver)	6ADH	374	0.1738	0.2219
alcdehye	Alcohol Dehydrogenase (Baker's yeast)	2HCY	347	0.2767	0.2911
apoferit	Apoferritin (equine spleen)	4DE6	168	0.7738	0.0000
bfgf	Basic Fibroblast Growth Factor (recombinant; human)	1BFG	126	0.0000	0.4127
carbanhy	Carbonic Anhydrase (bovine erythrocytes)	1V9E	259	0.0734	0.3050
concanv	Concanavalin A (jack bean)	3CNA	237	0.0000	0.4304
chymbov	α-Chymotrypsin (bovine pancreas)	1YPH	131	0.0000	0.3511
cytreho4	Cytochrome c (reduced; equine heart)	2GIW	104	0.4038	0.0000
cytoxho4	Cytochrome c (oxidized; equine heart)	1AKK	104	0.3942	0.0385
cytoxtun	Cytochrome c (oxidized; tuna heart)	3CYT	103	0.4563	0.0388
cytoxiso	Cytochrome c (oxidized; Baker's yeast)	2LIR	108	0.3611	0.0000
ccobov	Cytochrome c Oxidase (oxidized; bovine heart)	10CC	512	0.2720	0.2960
dnase1	Deoxyribonuclease I (bovine pancreas)	1DNK	250	0.0667	0.3458
elastspo	Elastase (porcine pancreas)	2V35	240	0.3899	0.1743
enolase	Enolase (Baker's yeast)	3ENL	436	0.1260	0.4072
rfxiii	Factor XIII (recombinant; homodimer; human)	1F13	722	0.0000	0.4127
fibrgnhu	Fibrinogen (human plasma)	3GHG	401	0.3541	0.1970
hbcohu	Hemoglobin (carboxy; human)	1K0Y	141	0.7163	0.0000
hbmethor	Hemoglobin (aquomet; equine)	1NS6	141	0.7589	0.0000
iggbov	Immunoglobulin G (bovine)	1GB1	56	0.2500	0.4107
interfhu	Interferon-gamma (recombinant; human)	1EKU	252	0.7024	0.0000
lalbnca	α-Lactalbumin (Ca-bound; bovine milk)	1F6S	122	0.3443	0.0820
ldhrab	Lactic Dehydrogenase (rabbit muscle)	3H3F	331	0.4109	0.2145
blgabov	β-Lactoglobulin A (bovine milk)	1CJ5	162	0.0556	0.3457
blgbbov	β-Lactoglobulin B (bovine milk)	4IBA	157	0.1146	0.4140
len	Light-chain LEN (recombinant; human)	2LVE	113	0.0000	0.5133
lysozyme	Lysozyme (chicken egg white)	1AZF	129	0.3333	0.0620
ovalbum	Ovalbumin (chicken egg)	2FRF	152	0.7632	0.0000
papain	Papain (papaya latex)	9PAP	211	0.2322	0.1801
rnasea	RNase A (bovine pancreas)	2QCA	124	0.1935	0.3306
subtilis	Subtilisin Carlsberg (*Bacillus licheniformis*)	1SBC	274	0.3139	0.1642
sodoxbov	Cu,Zn-Superoxide Dismutase (oxidized; bovine liver)	1CB4	151	0.0397	0.4040
sodrebov	Cu,Zn-Superoxide Dismutase (reduced; bovine liver)	1SXN	151	0.0331	0.4172
staphnuc	Staphylococcal Nuclease (recombinant)	1NUC	135	0.2741	0.3111
tim	Triosephosphate Isomerase (rabbit muscle)	1R2S	247	0.4372	0.1579
trypsnb	Trypsin (bovine pancreas)	4I8L	223	0.0807	0.3363
trypgenb	Trypsinogen (bovine pancreas)	1TGN	222	0.0811	0.3468
sti	Trypsin Inhibitor (soybean)	1BA7	169	0.0000	0.4260

4.2.2 Matrix Product of IR Spectra with Protein Secondary Structure Calibrants

Calibrant spectra dominated in α-helix and β-sheet are shown in Figure 4.1. The IR spectra considered herein each consist of a sequence of absorbances at equally interpolated wavenumbers that can be represented by a vector with elements x_{ij}, where the indices i and j refer to the i^{th} IR spectrum and j^{th} wavenumber in the spectrum, respectively. The set of all IR vectors constitute the IR spectra matrix. Similarly, the calibrants (i.e., the α-helix and β-sheet IR spectra) can also be defined by vectors $xcal_{ij}$ with elements having the same length as x_{ij}.

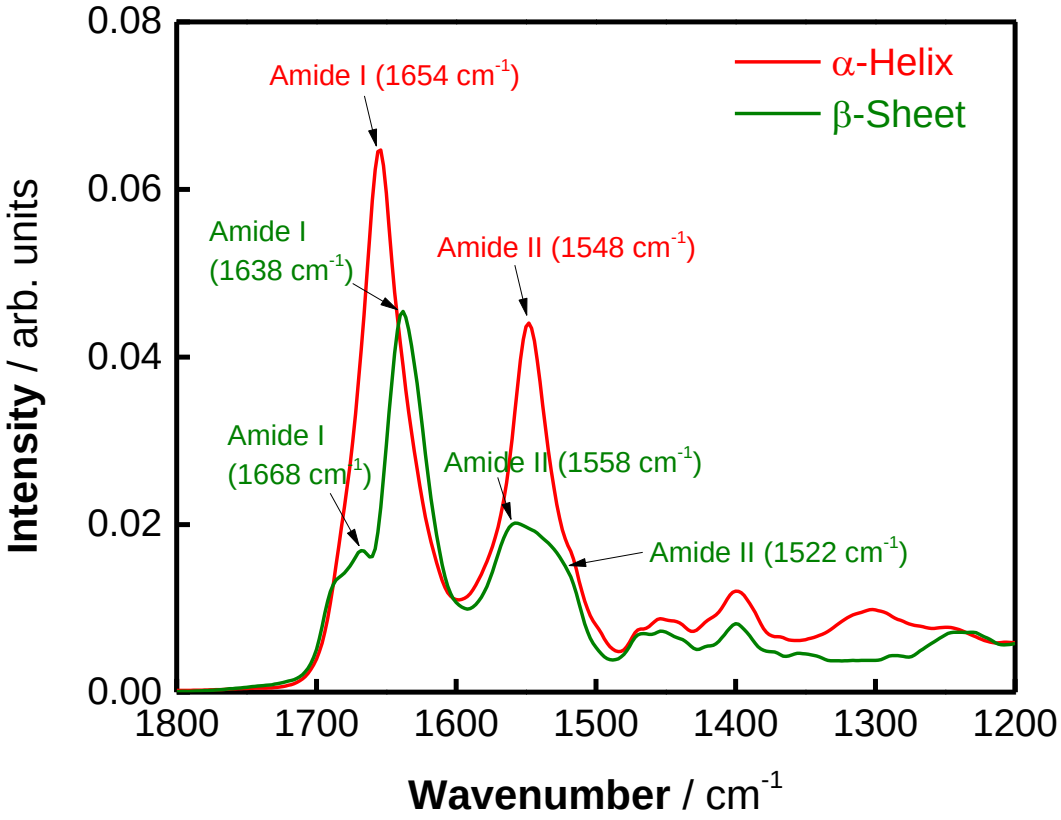

Figure 4.1 Spectra dominated in α-helix and β-sheet. These spectra are matrix multiplied with FTIR imaging/ATR-FTIR spectra to get α-helix and β-sheet scoring plots.

The set of calibrant vectors form the calibrant matrix. Prior to performing the product of these matrices, all vectors (spectra) were normalized to unity using

$$\bar{x}_{ij} = \sum_j x_{ij} / \sqrt{\sum_j x_{ij}^2}, \tag{4.4}$$

and

$$\bar{x}cal_{ij} = \sum_j xcal_{ij} / \sqrt{\sum_j xcal_{ij}^2}, \tag{4.5}$$

where x_{ij} and $xcal_{ij}$ are the element of the IR spectra and calibrant matrices, respectively. This normalization is an important and necessary step that accounts for variations in tissue thickness between sample slices and baseline shifts due to scattering.

The matrix product between the IR spectra (either from FTIR imaging or ATR-FTIR) and the calibrants given by

$$S_i = \sum_j \bar{x}_{ij}\, \bar{x}cal_{ij}{}^{\dagger} \tag{4.6}$$

can be used to generate the matrix of α-helix and β-sheet scores.

4.3 Results and Discussion

4.3.1 Identification of Distinct Tissular Regions with Protein Secondary Structure Score Plots from FTIR Imaging of Rectal Adenocarcinoma Metastatic to Liver Lesion

Figure 4.2 show the histograms of α-helix and β-sheet scores of rectal adenocarcinoma metastatic to liver obtained using the spectral range of 1200–1800 cm^{-1}, 1500–1700 cm^{-1}, and reduced 1500–1700 cm^{-1}. Scores from α-helix located in the range of 0.970 to 0.985, while those from β-sheet are in the lower range, between 0.920 and

0.940. Generally speaking, α-helix calibrant spectrum gives more separation of tumor and non-tumor than β-sheet one.

Figure 4.3 shows the contour plots of α-helix versus β-sheet scores of rectal adenocarcinoma metastatic to liver. Four areas are revealed in the plot that can be compared with the original histopathological examination of the H&E stain (Figure 3.3a). The non-tumor area has α-helix scores around 0.980, and β-sheet scores around 0.935. In the tumor area, α-helix and β-sheet scores are found in the ranges around 0.975 and 0.930, respectively. A small area adjacent to the tumor area, identified as a region rich in lymphocytes, has α-helix and β-sheet scores around 0.970 and 0.923, respectively. In contrast, red blood cells have a more extensive range of α-helix and β-sheet scores ranging from 0.940 to 0.980, and from 0.940 to 0.950, respectively. In light of Figures 4.2 and 4.3, reducing the spectral range from 1200–1800 cm^{-1} to 1500–1700 cm^{-1} and interpolation from 2 to 10 cm^{-1} does not deteriorate the separation of tumor and non-tumor.

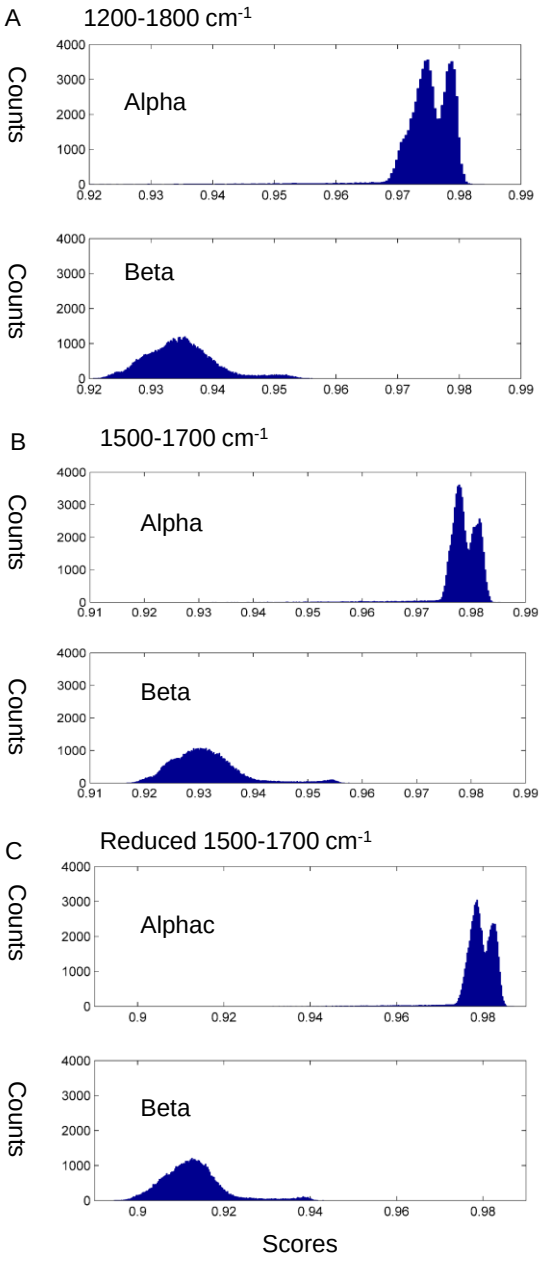

Figure 4.2 Histogram of α-helix and β-sheet scores obtained from FTIR imaging of rectal adenocarcinoma metastatic to liver. (A) 1200−1800 cm^{-1} (301 wavelengths) and (B) 1500−1700 cm^{-1} (101 wavelengths), both with 2 cm^{-1} interpolation, as well as (C) 1500−1700 cm^{-1} with 10 cm^{-1} interpolation (only 11 wavelengths).

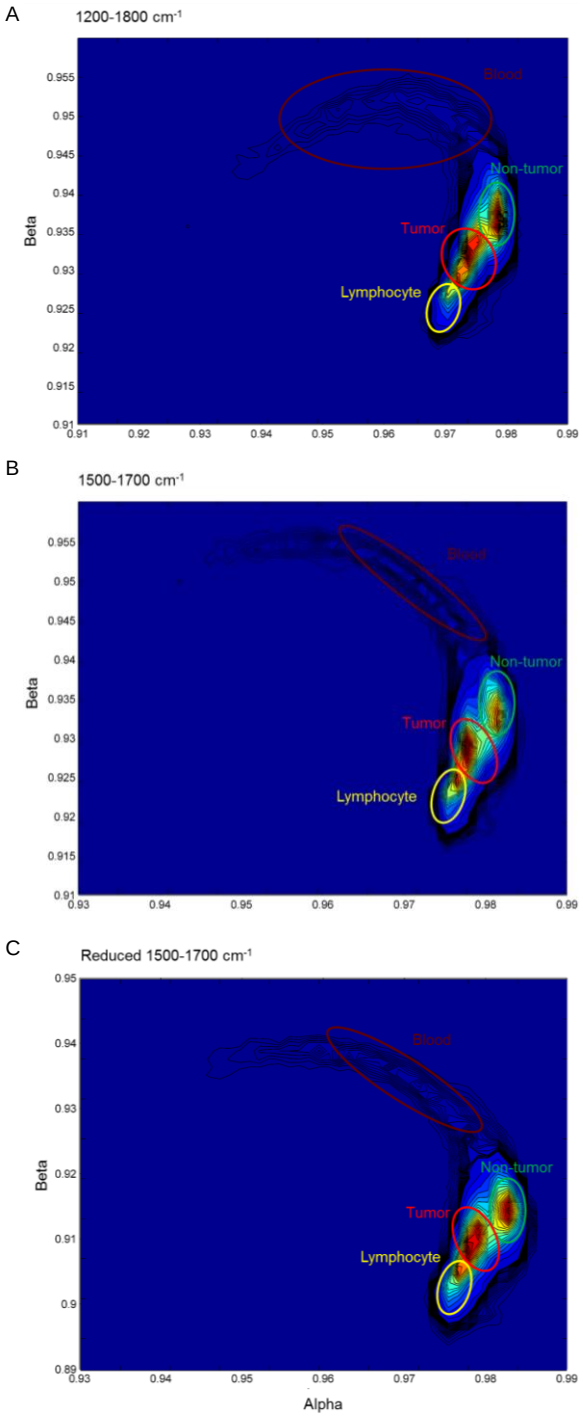

Figure 4.3 Contour plot of α-helix versus β-sheet scores obtained from FTIR imaging of rectal adenocarcinoma metastatic to liver. (A) 1200−1800 cm^{-1} (301 wavelengths) and (B) 1500−1700 cm^{-1} (101 wavelengths), both with 2 cm^{-1} interpolation, as well as (C) 1500−1700 cm^{-1} with 10 cm^{-1} interpolation (only 11 wavelengths).

4.3.2 Identification of Distinct Tissular Regions with Protein Secondary Structure Score Plots from ATR-FTIR Spectra of Rectal Adenocarcinoma Metastatic to Liver Lesion

ATR-FTIR spectra from the tumor and non-tumor regions of the original excised remnant tissue from the rectal adenocarcinoma metastatic to liver are shown in Figure 3.5b. Spectra from the tumor area show some distinct spectral features compared to those from the non-tumor area. For example, tumor spectra show less intense Amide I and II bands as well as alterations in band shapes. Although some of the spectral differences among tumor and non-tumor groups can be revealed from simple inspection, the differentiation remains inaccurate and time-consuming. Figure 4.4a shows the α-helix versus β-sheet scores plot of ATR-FTIR spectra obtained from the tumor and non-tumor areas (dark and light red in Figure 3.5a) of the liver tissue sample in the 1200−1800 cm^{-1} spectral range (for a 2 cm^{-1} interpolation, this range consists of 301 wavelengths). This spectral range provides a fair separation of tumor and non-tumor spectra with several discrepant points.

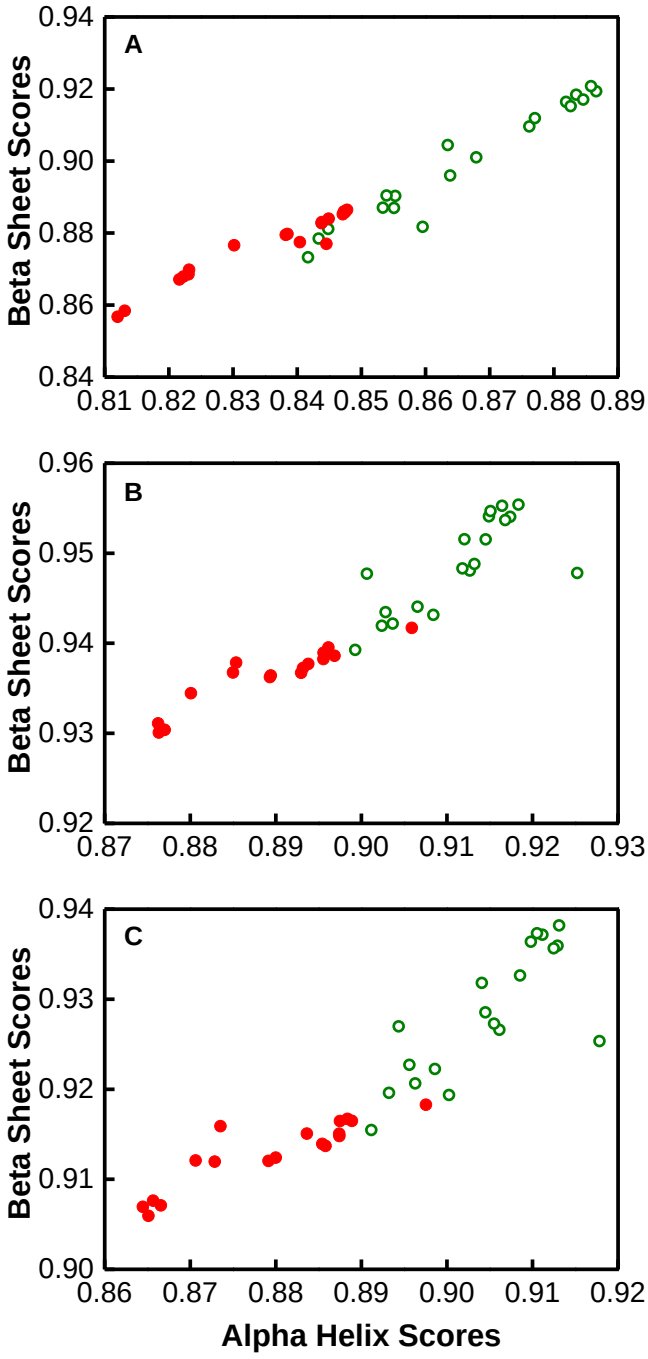

Figure 4.4 α-helix versus β-sheet scores using spectral data in the ranges (A) 1200−1800 cm^{-1} (301 wavelengths) and (B) 1500−1700 cm^{-1} (101 wavelengths), both with 2 cm^{-1} interpolation, as well as (C) 1500−1700 cm^{-1} with 10 cm^{-1} interpolation (only 11 wavelengths) for ATR-FTIR spectra from the tumor (*red*) and non-tumor (*green*) areas.

Figure 4.4b gives the α-helix versus β-sheet scores plot using a different spectral range, i.e., that of Amide I and II bands (1500−1700 cm^{-1}; for a 2 cm^{-1} interpolation, this range is now reduced to 101 wavelengths). Using this spectral range, the 18 and 19 spectra obtained from the tumor and non-tumor region, respectively, are well separated, except for one discrepancy. The non-tumor area has α-helix and β-sheet scores ranging from 0.900 to 0.930 and from 0.938 to 0.959, respectively. In comparison, the tumor area has slightly lower α-helix and β-sheet scores, ranging from 0.875 to 0.897 and from 0.930 to 0.940, respectively. The α-helix versus β-sheet scores plot for ATR-FTIR spectra follows a pattern identical to that observed from the score plot of FTIR imaging spectra. The number of IR metrics can be further reduced by using a 10 cm^{-1} interpolation of the same spectral range, the result of which are shown in Figure 4.4. This suggests that in principle one would require only 11 wavelengths between 1500−1700 cm^{-1} to differentiate tumor and non-tumor areas. Thus, this result also opens the possibility of using quantum cascade IR lasers instead of current broadband IR sources.

The spectra obtained by FTIR mapping and ATR-FTIR in the tumor region exhibit decreased intensity of Amide I and II bands. The results are in agreement with previous studies.[61-63] Based on previous research, more than 80% of cancer patients

demonstrate downgraded level of albumin.[64,65] Almost 80% of the protein in human liver is made up of albumin.[66] The reduced level of albumin gives rise to the lower intensity of Amide I and II bands arising from protein vibrations.

4.4 Conclusions

In summary, a method capable of differentiating tumor and non-tumor spectra based on the analysis of α-helix and β-sheet scores of spectra using matrix multiplication with calibrants extracted using linear least squre analysis was presented. The plot of α-helix versus β-sheet scores distinctly differentiates the tumor and non-tumor spectra of rectal adenocarcinoma metastatic to liver. Spectra obtained in the tumor region exhibit weaker Amide I and II bands as well as altered band shape. The decrease in intensity is related to protein misfolding in the tumor region. The alteration in band shape has something to do with the change in protein secondary structures between tumor and non-tumor regions. This particular result will be the object of future work. This approach, which can be applied without relying on training dataset development of SVM avoids the variations from different individuals and institutes. Reducing the number of IR metrics using larger interpolation does not affect the differentiation between tumor and non-tumor and could pave the way for the future application of quantum cascade IR lasers instead of current broadband IR sources as fewer wavelengths are required.

Chapter 5: Summary and Outlook

5.1 Summary

Work presented in this dissertation aimed at chemometrics development using FTIR spectroscopy and multivariate statistics for the accurate identification of cancer margin to complement surgical resection of cancer-bearing tissue, currently the most effective treatment of many forms of human cancer. The approach is to establish a TDM based on SVM using FTIR spectroscopic image data with *k*-means clustering analysis as the training set, and subsequently testing on the ATR-FTIR data. As an extension, alterations in protein secondary structures in tumor and non-tumor regions of colorectal cancer metastatic to the liver were analyzed using matrix multiplication based on the IR spectra extracted from linear least square analysis. To the author's knowledge, this is the first work quantifying the alterations in protein secondary structures associated with the tumors' progression.

The TDM is developed where the training set is based on FTIR imaging data with *k*-means clustering analysis. Subsequently, ATR-FTIR probe data is predicted by this

model. The results demonstrate the potential of using this approach as extremely important adjunct methodologies to that of standard histopathological tissue analysis for real-time cancer detection. Some details (e.g., spectral difference of tumor, non-tumor, lymphocytes, and red blood cells) were also given.

While chapter 3 provided a qualitative picture of differenting tumor and non-tumor tissues using a TDM, chapter 4 discussed the differentiation of tumor and non-tumor using a more quantitative approach. The strategies developed in this study are more straightforward compare to the prior use of curve deconvolution fitting and area calculation. The plot of α-helix versus β-sheet scores obtained from matrix multiplication distinctly differentiates the tumor and non-tumor spectra of rectal adenocarcinoma metastatic to liver. The results shows that the tumor has lower α-helix and β-sheet scores compared to the ones obtained from non-tumor. This may arise from the presence of protein misfolding in the tumor area.

To recap, the themes explored in this dissertation, thorough resection of cancerous tissue during surgical removal of malignant tumors is of critical significance. If the aims of the project are achieved, clinical surgical oncology practice may be greatly improved

upon dissemination of these pathology protocols, and this may affect survival rates for patients.

5.2 Outlook

TDM and dot products of protein secondary structures presented in this dissertation are the first attempts to identify cancer margins using FTIR spectroscopy combined with multivariate statistics. The approach is powerful and promising as it offers many possibilities for further development.

For example, the work conducted with FTIR/ATR-FTIR in this dissertation can also be tested with Raman spectroscopy, which is another vibrational spectroscopy complementary to FTIR. Recently, Raman spectroscopy has attracted much attention in cancer diagnosis because of its merit of being real-time, highly sensitive and non-invasive. The greatest advantage in applying Raman spectroscopy in cancer diagnosis is the little sample preparation due to the absence of any thickness requirement. Furthermore, because of the Raman selection rule, water is a weak scatterer,[67] which has for consequence that there is less interference in Raman spectra compared to IR spectra. However, one possible challenge of using Raman spectroscopy in cancer study

could come from autofluorescence. As such, tissue types are limited to breast, lung, prostate, colon, gastric mucosa, and stomach.

Additionally, the matrix multiplication methodology can also be expanded to investigate not only the alterations in protein secondary structures but also some other compounds in tumor and non-tumor areas, including albumin, triglycerides, glycogen and etc. Albumin is dominated in α-helices stuctures and is the most abundant protein in human liver.[68] Examining alterations of albumin in tumor and non-tumor areas could provide insightful molecular signatures of proteins in cancer. In a similar manner, lipids alterations going from tumor to non-tumor areas could be studied by looking at the changes in triglycerides.

Finally, as an extension, more IR spectra of protein standards should be added to the initial library of Dong, Carpenter, Caughey, especially those related to the metastatic liver cancer. Furthermore, not only proteins but also the lipid contents of cancer-bearing tissues should be investigated. As such, an IR spectra database of lipid standards should also be built.

REFERENCES

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2013) GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer. http://globocan.iarc.fr. Accessed accessed Aug 23 2014
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global Cancer Statistics. CA Cancer J Clin 61:69-90
3. Renshaw AA, Gould EW (2007) Measuring Errors in Surgical Pathology in Real-Life Practice. American Journal of Clinical Pathology 127:144-152
4. Nover AB, Jagtap S, Anjum W, Yegingil H, Shih WY, Shih W-H, Brooks AD (2009) Modern Breast Cancer Detection: a Technological Review. Journal of Biomedical Imaging 2009:1-14
5. Diem M, Mazur A, Lenau K, Schubert J, Bird B, Miljkovic M, Krafft C, Popp J (2013) Molecular Pathology via IR and Raman Spectral Imaging. J Biophotonics 6:855-886
6. Sun XL, Xu YZ, Wu JG, Zhang YF, Sun KL (2013) Detection of Lung Cancer Tissue by Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy — A Pilot Study of 60 Samples. J Surg Res 179:33-38
7. Gajjar K, Trevisan J, Owens G, Keating PJ, Wood NJ, Stringfellow HF, Martin-Hirsch PL, Martin FL (2013) Fourier-Transform Infrared Spectroscopy Coupled with A Classification Machine for the Analysis of Blood Plasma or Serum: A Novel Diagnostic Approach for Ovarian Cancer. Analyst 138:3917-3926
8. Mackanos MA, Contag CH (2010) Fiber-optic probes enable cancer detection with FTIR spectroscopy. Trends in Biotechnology 28:317-323
9. Rigas B, Wong PTT (1992) Human Colon Adenocarcinoma Cell Lines Display Infrared Spectroscopic Features of Malignant Colon Tissues. Cancer Res 52:84-88
10. Wong PTT, Wong RK, Caputo TA, Godwin TA, Rigas B (1991) Infrared Spectroscopy of Exfoliated Human Cervical Cells: Evidence of Extensive Structural Changes During Carcinogenesis. Proc Natl Acad Sci USA 88:10988-10992
11. Wong PTT, Goldstein SM, Grekin RC, Godwin TA, Pivik C, Rigas B (1993) Distinct Infrared Spectroscopic Patterns of Human Basal Cell Carcinoma of the Skin. Cancer Res 53:762-765
12. Rigas B, Morgello S, Goldman IS, Wong PTT (1990) Human Colorectal Cancers Display Abnormal Fourier-Transform Infrared Spectra. Proc Natl Acad Sci USA 87:8140-8144
13. Andrus PGL, Strickland RD (1998) Cancer Grading by Fourier Transform Infrared Spectroscopy. Biospectroscopy 4:37-46
14. Wood BR, Quinn MA, Burden FR, McNaughton D (1996) An Investigation into FTIR Spectroscopy as a Biodiagnostic Tool for Cervical Cancer. Biospectroscopy 2:143-153
15. Chiriboga L, Xie P, Yee H, Zarou D, Zakim D, Diem M (1998) Infrared Spectroscopy of Human Tissue. IV. Detection of Dysplastic and Neoplastic Changes of Human Cervical Tissue via Infrared Microscopy. Cell Mol Biol 44:219-229

16. Li QB, Wang W, Ling XF, Wu JG (2013) Detection of Gastric Cancer with Fourier Transform Infrared Spectroscopy and Support Vector Machine Classification. Biomed Res Int 942927:1-4
17. Bergner N, Romeike BFM, Reichart R, Kalff R, Krafft C, Popp JU (2013) Tumor Margin Identification and Prediction of the Primary Tumor from Brain Metastases using FTIR Imaging and Support Vector Machines. Analyst 138:3983-3990
18. Lee DW, Seo KW, Min BR (2007) Discrimination between Cancer and Normal Tissue using Near Infrared Spectroscopy. In: Kim SI, Suh TS (eds) World Congress on Medical Physics and Biomedical Engineering 2006, Vol. 14., vol 14. IFMBE Proceedings. Springer-Verlag, Berlin, Germany, pp 1341-1344
19. Derenne A, Van Hemelryck V, Lamoral-Theys D, Kiss R, Goormaghtigh E (2013) FTIR Spectroscopy: A New Valuable Tool to Classify the Effects of Polyphenolic Compounds on Cancer Cells. Biochim Biophys Acta-Mol Basis Dis 1832:46-56
20. Chen ZM, Butke R, Miller B, Hitchcock CL, Allen HC, Povoski SP, Martin EW, Coe JV (2013) Infrared Metrics for Fixation-Free Liver Tumor Detection. J Phys Chem B 117:12442-12450
21. Lin WM, Yuan X, Yuen P, Wei WI, Sham J, Shi PC, Qu J (2004) Classification of In Vivo Autofluorescence Spectra Using Support Vector Machines. J Biomed Opt 9:180-186
22. Palmer GM, Zhu CF, Breslin TM, Xu FS, Gilchrist KW, Ramanujam N (2003) Comparison of Multiexcitation Fluorescence and Diffuse Reflectance Spectroscopy for the Diagnosis of Breast Cancer IEEE Trans Biomed Eng 50:1233-1242
23. Majumder SK, Ghosh N, Gupta PK (2005) Support Vector Machine for Optical Diagnosis of Cancer. J Biomed Opt 10:024034
24. Widjaja E, Zheng W, Huang ZW (2008) Classification of Colonic Tissues Using Near-Infrared Raman Spectroscopy and Support Vector Machines. Int J Oncol 32:653-662
25. Härdle WK, Simar L (2007) Applied Multivariate Statistical Analysis. Springer, New York, NY
26. Coblentz WW (1908) Investigations of Infrared Spectra: Infra-Red Reflection Spectra. vol p. 5. Carnegie Inst. of Washington, Baltimore, MD
27. Cooley JW, Tukey JW (1965) An Algorithm for the Machine Calculation of Complex Fourier Series. Mathematics of Computation 19:297-301
28. Petibois C, Deleris G (2006) Chemical Mapping of Tumor Progression by FT-IR Imaging: Towards Molecular Histopathology. Trends Biotechnol 24:455-462
29. Guide PESU.
30. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK
31. Schölkopf B, Smola AJ (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA
32. Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2:1-27
33. Hsu CW, Chang CC, Lin CJ (2003) A Practical Guide to Support Vector Classification.

34. Coe JV, Chen ZM, Li R, Butke R, Miller B, Hitchcock CL, Allen HC, Povoski SP, Martin EW (2014) Imaging Infrared Spectroscopy for Fixation-Free Liver Tumor Detection. Proc SPIE 8947:89470B
35. Khan IR, Ohba R (1999) Closed-Form Expressions for the Finite Difference Approximations of First and Higher Derivatives Based on Taylor Series. J Comput Appl Math 107:179-193
36. Parker F (2012) Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine. Springer, New York, NY
37. Mostaco-Guidolin LB, Murakami LS, Batistuti MR, Nomizo A, Bachmann L (2010) Molecular and Chemical Characterization by Fourier Transform Infrared Spectroscopy of Human Breast Cancer Cells with Estrogen Receptor Expressed and Not Expressed. Spectroscopy 24:501-510
38. Movasaghi Z, Rehman S, Rehman IU (2008) Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. Appl Spectrosc Rev 43:134-179
39. Mantsch HH, Chapman D (1996) Infrared Spectroscopy of Biomolecules. John Wiley & Sons, Hoboken, NJ
40. Barth A (2007) Infrared Spectroscopy of Proteins. Biochim Biophys Acta-Bioenerg 1767:1073-1101
41. Tatulian SA (2013) Structural Characterization of Membrane Proteins and Peptides by FTIR and ATR-FTIR Spectroscopy. Methods in molecular biology (Clifton, NJ) 974:177-218
42. Amharref N, Beljebbar A, Dukic S, Venteo L, Schneider L, Pluot M, Manfait M (2007) Discriminating Healthy from Tumor and Necrosis Tissue in Rat Brain Tissue Samples by Raman Spectral Imaging. Biochimica et Biophysica Acta (BBA) - Biomembranes 1768:2605-2615
43. Sasic S, Ozaki Y (2011) Raman, Infrared, and Near-Infrared Chemical Imaging. Wiley,
44. Yates AJ, Thompson DK, Boesel CP, Albrightson C, Hart RW (1979) Lipid Composition of Human Neural Tumors. J Lipid Res 20:428-436
45. Cohenford MA, Godwin TA, Cahn F, Bhandare P, Caputo TA, Rigas B (1997) Infrared Spectroscopy of Normal and Abnormal Cervical Smears: Evaluation by Principal Component Analysis. Gynecologic oncology 66:59-65
46. Fujioka N, Morimoto Y, Arai T, Kikuchi M (2004) Discrimination between Normal and Malignant Human Gastric Tissues by Fourier Transform Infrared Spectroscopy. Cancer Detection and Prevention 28:32-36
47. Baenke F, Peck B, Miess H, Schulze A (2013) Hooked on Fat: the Role of Lipid Synthesis in Cancer Metabolism and Tumour Development. Disease models & mechanisms 6:1353-1363
48. Srivastava NK, Pradhan S, Gowda GA, Kumar R (2010) In vitro, High-Resolution 1H and 31P NMR Based Analysis of the Lipid Components in the Tissue, Serum, and CSF of the Patients with Primary Brain Tumors: One Possible Diagnostic View. NMR in biomedicine 23:113-122
49. Barth A (2007) Infrared spectroscopy of proteins. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1767:1073-1101

50. Jackson M, Mantsch HH (1991) Protein secondary structure from FT-IR spectroscopy: correlation with dihedral angles from three-dimensional Ramachandran plots. Canadian Journal of Chemistry 69:1639-1642
51. Dousseau F, Pezolet M (1990) Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods. Biochemistry 29:8771-8779
52. Tatulian SA (2013) Structural characterization of membrane proteins and peptides by FTIR and ATR-FTIR spectroscopy. Methods Mol Biol 974:177-218
53. Adochitei A, Drochioiu G (2011) Rapid characterization of peptide secondary structure by FT-IR spectroscopy. Revue Roumaine De Chimie 56:783-791
54. Kong J, Yu S (2007) Fourier transform infrared spectroscopic analysis of protein secondary structures. Acta Biochimica Et Biophysica Sinica 39:549-559
55. Coe JV, Nystrom, S. V., Chen, Z., Li, R., Verreault, D., Hitchcock, C. L., Martin, Jr., E. W., Allen, H. C. (2015) The Ramachandran Plot and Infrared Spectra of Protein Secondary Structures: Alpha-Helix and Beta-Sheet. *submitted*
56. Navea S, Tauler R, de JA (2005) Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure. Analytical Biochemistry 336:231-242
57. Lee DC, Haris PI, Chapman D, Mitchell RC (1990) Determination of protein secondary structure using factor analysis of infrared spectra. Biochemistry 29:9185-9193
58. Cai S, Singh BR (2004) A distinct utility of the amide III infrared band for secondary structure estimation of aqueous protein solutions using partial least squares methods. Biochemistry 43:2541-2549
59. Rahmelow K, Hubner W (1996) Secondary structure determination of proteins in aqueous solution by infrared spectroscopy: a comparison of multivariate data analysis methods. Analytical Biochemistry 241:5-13
60. Shariati-Rad M, Hasani M (2009) Application of multivariate curve resolution-alternating least squares (MCR-ALS) for secondary structure resolving of proteins. Biochimie 91:850-856
61. Colagar AH, Chaichi MJ, Khadjvand T (2011) Fourier transform infrared microspectroscopy as a diagnostic tool for distinguishing between normal and malignant human gastric tissue. Journal of Biosciences 36:669-677
62. Kondepati VR, Keese M, Mueller R, Manegold BC, Backhaus J (2007) Application of near-infrared spectroscopy for the diagnosis of colorectal cancer in resected human tissue specimens. Vibrational Spectroscopy 44:236-242
63. Paluszkiewicz C, Kwiatek WM, Banaś A, Kisiel A, Marcelli A, Piccinini M (2007) SR-FTIR spectroscopic preliminary findings of non-cancerous, cancerous, and hyperplastic human prostate tissues. Vibrational Spectroscopy 43:237-242
64. Buchanan M (2010) You Can Prevent and Reverse Cancer. Xlibris US,
65. Seve P, Ray-Coquard I, Trillet-Lenoir V, Sawyer M, Hanson J, Broussolle C, Negrier S, Dumontet C, Mackey JR (2006) Low serum albumin levels and liver metastasis are powerful prognostic markers for survival in patients with carcinomas of unknown primary site. Cancer 107:2698-2705
66. First insight into the human liver proteome from PROTEOME(SKY)-LIVER(Hu) 1.0, a publicly available database (2010). Journal of proteome research 9:79-94

67. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, Hercogova J, Rossen K, Thomsen HK, Gniadecki R, Hansen LK, Wulf HC (2004) Melanoma Diagnosis by Raman Spectroscopy and Neural Networks: Structure Alterations in Proteins and Lipids in Intact Cancer Tissue. J Investig Dermatol 122:443-449
68. Consortium CHLPP (2010) First Insight into the Human Liver Proteome from PROTEOME(SKY)-LIVER(Hu) 1.0, a Publicly Available Database. Journal of proteome research 9:79-94

Appendix A Matlab Program for Matrix Multiplication

(from Professor James Coe et al.)

```
clear title xlabel ylabel
%just for case8, get dimensions of data
[np,nz]=size(X8full);

% calibrants for aichun dong's library have a smaller range of wavenumbers
% so we pick out the part of our spectra matching that range
istart=find(nu==1800); % find index of starting wavenumber
iend=find(nu==1200); % find index of ending wavenumber
X8=X8full(:,istart:iend);
[np,nz]=size(X8);
nu2=nu(istart:iend,1);

%**********calibrants**********
% input the calibrants
c1=dlmread('result_a1_2.txt');
c2=dlmread('result_b1_2.txt');
c3=dlmread('result_ot_2.txt');

% normalize calibrants, inner product with themselves is one
nc1=(c1(:,2)'*c1(:,2)); c1(:,2)=c1(:,2)/sqrt(nc1);
```

```
nc2=(c2(:,2)'*c2(:,2)); c2(:,2)=c2(:,2)/sqrt(nc2);
nc3=(c3(:,2)'*c3(:,2)); c3(:,2)=c3(:,2)/sqrt(nc3);

% put them into a matrix like the X file
Xc=cat(1,c1(:,2)',c2(:,2)',c3(:,2)');
[nb ns]=size(Xc); % get the number of calibrant metrics
nb
%*********************************************************************

% get intensity metric (norm) for each pixel spectra
disp('norms of X8')
for k=1:np
   XI(k,1)=sqrt(X8(k,:)*X8(k,:)');
end
disp('normalizing X8')
X8c=X8;
for k=1:np
   if k==1000; disp('1000'); end;
   if k==10000; disp('10000'); end;
   if k==20000; disp('20000'); end
   if k==40000; disp('40000'); end
   X8c(k,:)=X8(k,:)./XI(k,1);
end
```

```
%
**************************************************************************
% normalized X matrix dotted with normalized calibrants
disp('dotting normalized spectra with normalized calibrants')
Xrc=X8c*Xc';

% % get average of all IR spectra for case
sp_all=sum(X8,1);
sp_all=sp_all/nx8*ny8;
spout=cat(2,nu2,sp_all');
save wavg_case8.txt spout -ASCII;

figure(1)
% plot a histogram for each biomarker
%Xrcc=0:0.005:1;
for m=1:nb
   subplot(3,1,m);
   hist(Xrc(:,m),200)
   axis([.84 1 0 4000]);
end
xlabel('score')
ylabel('counts')
```

```
text(.85,13000,'alpha','FontSize',14)
text(.85,8000,'beta','FontSize',14)
text(.85,1*2000,'other','FontSize',14)
set(gcf,'PaperUnits','inches','PaperPosition',[0 0 6 7])
print(1,'-dtiff','-r600','calibrant_histograms.tif')

% ***************Dot Plot*************************************************
% This plot has the numerical values of the scores
figure(20)
plot(Xrc(:,1),Xrc(:,2),'.','markers',2)
xlabel('alpha')
ylabel('beta')
axis([0.86 0.94 0.875 0.93]);
%hold on
% % add an ellipse
% for i=1:101
%     t=(i/100)*2*pi;
%     xe(i)=Xc1+ae*cos(t)*cos(phip1)-be*sin(t)*sin(phip1);
%     ye(i)=Yc1+ae*cos(t)*sin(phip1)+be*sin(t)*cos(phip1);
% end
% plot(xe,ye,'w')
%set(gca,'Color',[0 0 0]);
print(20,'-dtiff','-r600','Cal_dot_1_2.tif')
```

```
% scale scores
Xrc_unscaled=Xrc; % keep the unscaled scores
% calculate average and std dev of each
Xbar=mean(Xrc); sX=std(Xrc);
% scale biomarkers
Xs=Xrc; % initialize Xs
for m=1:nb
   Xs(:,m)=0.5+(Xrc(:,m)-Xbar(m)).*(0.5/2)./sX(m);
end
Xrc=Xs;

% Make images of Case 8 from scaled scores
for i=1:nx8
   for j=1:ny8
      bp8(i,j,:)=Xrc((j-1)*nx8+i,:);
   end
end
for k=1:nb
   imwrite(bp8(:,:,k),['Case8_',num2str(k),'.bmp']);
end

Xrc=Xrc_unscaled;
```

```
%**********************ATR DATA*********************************
% input the ATR spectra
atr_n1=dlmread('nontumor_ATR_1800to1200.txt'); [nn1 nns1]=size(atr_n1);
atr_t1=dlmread('tumor_ATR_1800to1200.txt'); [nt1 nts1]=size(atr_t1);
nu_atr=atr_n1(1,:); % first row is wavenumbers
atrn=atr_n1(2:nn1,:); % remove first row
atrt=atr_t1(2:nt1,:); % remove first row

% normalize ATR groups, inner product with themselves is one
for i=1:nn1-1
   natrn(i,1)=sqrt(atrn(i,:)*atrn(i,:)');
end
for i=1:nn1-1
   atrnc(i,:)=atrn(i,:)./natrn(i,:);
end

for i=1:nt1-1
   natrt(i,1)=sqrt(atrt(i,:)*atrt(i,:)');
end
for i=1:nt1-1
   atrtc(i,:)=atrt(i,:)./natrt(i,:);
end
```

```
%*****************************************************************
disp('dotting normalized ATR spectra with normalized calibrants')
Xnc=atrnc*Xc';
Xtc=atrtc*Xc';

figure(21)
plot(Xnc(:,1),Xnc(:,2),'.g',Xtc(:,1),Xtc(:,2),'.r')
h=text(0.78,0.814,'alpha')
set(h,'rotation',0);
h=text(0.743,0.85,'beta')
set(h,'rotation',90);
```

Appendix B Matlab Program for Merging Cases into X Files

(from Professor James Coe et al.)

```
clc; clear all; close all;
% **************Merge all cases into one X file*****************************
% Case 4
WndRng=[6 12 2]; % Window#Range [start stop direction]
str=sprintf('acwindow%i.fsm',WndRng(1));
[d4,xAxis4,yAxis4,zAxis4,misc]=fsmload(str);
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
   str=sprintf('acwindow%i.fsm',i); % Enter file name
   [data, xAxis, yAxis, zAxis, misc] = fsmload(str);
   d4=cat(WndRng(3),d4,data);
   if WndRng(3)==2; xAxis4=cat(2,xAxis4,xAxis); end
   if WndRng(3)==1; yAxis4=cat(2,yAxis4,yAxis); end
end
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx4,ny4,nz4]=size(d4);
X4=zeros(nx4*ny4,nz4);
for j=1:ny4
      k=(j-1)*nx4+i;
      X4(k,:)=-log(d4(i,j,:)/100);
```

```
    end
end
X4=real(X4);
X=cat(1,X,X4);
Str=sprintf('Case 4 is loaded into the X matrix');
disp(Str)
%*************************************************************************
***
% Case 5
WndRng=[1 5 2]; % Window # Range: [start stop direction]
str=sprintf('ac_case5_%i.fsm',WndRng(1));
[d5,xAxis5,yAxis5,zAxis5,misc]=fsmload(str);
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
    str=sprintf('ac_case5_%i.fsm',i); % Enter file name
    [data, xAxis, yAxis, zAxis, misc] = fsmload(str)
```

```
 d5=cat(WndRng(3),d5,data);
    if WndRng(3)==2; xAxis5=cat(2,xAxis5,xAxis); end
    if WndRng(3)==1; yAxis5=cat(2,yAxis5,yAxis); end
end
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx5,ny5,nz5]=size(d5);
X5=zeros(nx5*ny5,nz5);
for i=1:nx5
    for j=1:ny5
        k=(j-1)*nx5+i;
        X5(k,:)=-log(d5(i,j,:)/100);
    end
end
X5=real(X5);
X=cat(1,X,X5);
Str=sprintf('Case 5 is loaded into the X matrix');
disp(Str)
```

```
%*************************************************************************
***
% Case 7 spectra have a different range, i.e. 4500 to 750 cm-1
WndRng=[1 4 2]; % Window # Range: [start stop direction]
str=sprintf('ac_case7_%i.fsm',WndRng(1));
[d7,xAxis7,yAxis7,zAxis7,misc]=fsmload(str);
zAxis7=zAxis7(251:1876); % need to select out 1st 250 of zAxis7
d7=d7(:,:,251:1876); % removes 4500-4002 cm-1
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
   str=sprintf('ac_case7_%i.fsm',i); % Enter file name
   [data, xAxis, yAxis, zAxis, misc] = fsmload(str);
   data=data(:,:,251:1876);
   d7=cat(WndRng(3),d7,data); % removes 4500-4002 cm-1
   if WndRng(3)==2; xAxis7=cat(2,xAxis7,xAxis); end
   if WndRng(3)==1; yAxis7=cat(2,yAxis7,yAxis); end
end
```

```
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx7,ny7,nz7]=size(d7);
X7=zeros(nx7*ny7,nz7);
for i=1:nx7
   for j=1:ny7
      k=(j-1)*nx7+i;
      X7(k,:)=-log(d7(i,j,:)/100);
   end
end
X7=real(X7);
X=cat(1,X,X7);
Str=sprintf('Case 7 is loaded into the X matrix');
disp(Str)
%*************************************************************************
***
% Case 8 renumbered *.fsm files to go from left to right
WndRng=[1 4 2]; % Window # Range: [start stop direction]
```

```
str=sprintf('ac_case8_%i.fsm',WndRng(1));
[d8,xAxis8,yAxis8,zAxis8,misc]=fsmload(str);
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
   str=sprintf('ac_case8_%i.fsm',i); % Enter file name
   [data, xAxis, yAxis, zAxis, misc] = fsmload(str);
   d8=cat(WndRng(3),d8,data);
   if WndRng(3)==2; xAxis8=cat(2,xAxis8,xAxis); end
   if WndRng(3)==1; yAxis8=cat(2,yAxis8,yAxis); end
end
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx8,ny8,nz8]=size(d8);
X8=zeros(nx8*ny8,nz8);
for i=1:nx8
   for j=1:ny8
      k=(j-1)*nx8+i;
      X8(k,:)=-log(d8(i,j,:)/100);
```

```
    end
end
X8=real(X8);
X=cat(1,X,X8);
Str=sprintf('Case 8 is loaded into the X matrix');
disp(Str)
%**************************************************************************
***
% Case 9
WndRng=[1 5 2]; % Window # Range: [start stop direction]
str=sprintf('ac_case9_%i.fsm',WndRng(1));
[d9,xAxis9,yAxis9,zAxis9,misc]=fsmload(str);
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
    str=sprintf('ac_case9_%i.fsm',i); % Enter file name
    [data, xAxis, yAxis, zAxis, misc] = fsmload(str);
    d9=cat(WndRng(3),d9,data);
```

```
    if WndRng(3)==2; xAxis9=cat(2,xAxis9,xAxis); end
    if WndRng(3)==1; yAxis9=cat(2,yAxis9,yAxis); end
end
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx9,ny9,nz9]=size(d9);
X9=zeros(nx9*ny9,nz9);
for i=1:nx9
    for j=1:ny9
        k=(j-1)*nx9+i;
        X9(k,:)=-log(d9(i,j,:)/100);
    end
end
X9=real(X9);
X=cat(1,X,X9);
Str=sprintf('Case 9 is loaded into the X matrix');
disp(Str)
```

```
%*************************************************************************
***
% Case 10 couldn't get 5th window to load into X?
WndRng=[1 4 1]; % Window # Range: [start stop direction]
str=sprintf('ac_case10_%i.fsm',WndRng(1));
[d10,xAxis10,yAxis10,zAxis10,misc]=fsmload(str);
% Loop through the first row (or column)
for i=WndRng(1)+1:WndRng(2)
   str=sprintf('ac_case10_%i.fsm',i); % Enter file name
   [data, xAxis, yAxis, zAxis, misc] = fsmload(str);
   d10=cat(WndRng(3),d10,data);
   if WndRng(3)==2; xAxis10=cat(2,xAxis10,xAxis); end
   if WndRng(3)==1; yAxis10=cat(2,yAxis10,yAxis); end
end
% rearrange as X matrix (each pixel has a row of its IR spectrum)
[nx10,ny10,nz10]=size(d10);
X10=zeros(nx10*ny10,nz10);
```

```
for i=1:nx10
   for j=1:ny10
      k=(j-1)*nx10+i;
      X10(k,:)=-log(d10(i,j,:)/100);
   end
end
X10=real(X10);
X=cat(1,X,X10);
Str=sprintf('Case 10 is loaded into the X matrix');
disp(Str)
Str=sprintf('All cases are merged into the X matrix');
disp(Str)
[np nc]=size(X); % get the number of pixels
disp('number of pixels or IR spectra')
np % report the number of pixels
```

Appendix C Matlab Program for *K*-Means Clustering Analysis

(from Professor James Coe et al.)

```
close all

% get the size of the data

[nx ny nz]=size(data)

% get the number of biomarkers

[nxny nb]=size(X)

[nx ny nb]=size(b)



% pick the number of clusters

nclusters=2;

% write the number of groups

nclusters
```

```
% do kmeans analysis

opts=statset('Display','final','MaxIter',200)

[idx,ctrs,sumd] = kmeans(X,nclusters,'distance','city','replicates',4,'Options',opts);

save centers.txt ctrs -ASCII
```

```
% count the number of members of each group
ic(1:nclusters)=0;
for k=1:nx*ny
   for m=1:nclusters
      if(idx(k,1)==m)
         ic(m)=ic(m)+1;
      end
   end
end
ic
save counts.txt ic -ASCII
icsum=sum(ic)


% get the spread of each group
for k=1:nclusters
   sigma(k)=sqrt(sumd(k)/(ic(k)-1));
end
```

```
sigma



% pick biomarkers for plotting

n1=12;n2=20;n3=7;

% pick the colormap

ColorOrder=[1 0 0;...

        0 1 0;...

        0 0 1;...

        0 1 1;...

        1 1 0;...

        1 0 1;...

        1 0.8 1;...

        1 0.549 0;...

        0 0 0.5;...

        0.5 0 0.5;...

        0.604 0.804 0.196;...

        0.721 0.525 0.043;...
```

```
        0 0.5 0.0;...

        0.5 0 0;...

        0.196 0.604 0.804;...

        0.5 0.5 0.5;...

        0.8 0.7 0.3;...

        0.3 0.8 0.7;...

        0.7 0.3 0.8;...

        0.75 0.75 0.75;...

        0.3 0.6 0.9;...

        0.9 0.3 0.6;...

        0.6 0.9 0.3;...

        0.2 0.2 0.2;...

        0.4 0.4 0.4];

set(0,'DefaultAxesColorOrder',ColorOrder)



figure(2)
```

```
for k=1:nclusters
    plot(X(idx==k,n1),X(idx==k,n2),'.','MarkerSize',3)
hold all
end
plot(ctrs(:,n1),ctrs(:,n2),'ko')
hold all
plot(ctrs(:,n1),ctrs(:,n2),'kX')
hold all


figure(3)
for k=1:nclusters
scatter3(X(idx==k,n1),X(idx==k,n2),X(idx==k,n3),'.')
hold on
end
scatter3(ctrs(:,n1),ctrs(:,n2),ctrs(:,n3),'k','o')
hold on
scatter3(ctrs(:,n1),ctrs(:,n2),ctrs(:,n3),'k','X')
```

```
hold on


% make a color image of all groups
% put idx back into an image plane
for i=1:nx
   for j=1:ny
      brg(i,j)=idx((j-1)*nx+i);
   end
end
% get a color for each group, construct red, green, and blue, and combine
red=brg;green=brg;blue=brg;
for i=1:nx
   for j=1:ny
      for k=1:nclusters
         if brg(i,j)==k
            red(i,j)=ColorOrder(k,1);
            green(i,j)=ColorOrder(k,2);
```

```
            blue(i,j)=ColorOrder(k,3);
          end
        end
    end
end

brg_color=cat(3,red,green,blue);
imwrite(brg_color,'kmeans_color.bmp');
figure(4)
imshow('kmeans_color.bmp');

% make a color image of each group
% make a image plane filter for each group
g=zeros(nx,ny,nb);
for i=1:nx
    for j=1:ny
        for k=1:nclusters
```

```
            if(brg(i,j)==k)
                g(i,j,k)=1;
            end
        end
    end
end
% write an image plane bitmap of each group
%red=brg;green=brg;blue=brg;
red=zeros(nx,ny);green=zeros(nx,ny);blue=zeros(nx,ny);
for k=1:nclusters
    x=k;
    for i=1:nx
        for j=1:ny
            if(brg(i,j)==k)
                red(i,j)=ColorOrder(k,1);
                green(i,j)=ColorOrder(k,2);
                blue(i,j)=ColorOrder(k,3);
```

```
        end
    end
end
brg_color=cat(3,red,green,blue);
imwrite(brg_color,['g',num2str(x),'.bmp']);
red=zeros(nx,ny);green=zeros(nx,ny);blue=zeros(nx,ny);
```

Appendix D Matlab Program for SVM

```
[Train,PS] = mapminmax(train');

Train = Train';

Test = mapminmax('apply',test',PS);

Test = Test';

[c,g] = meshgrid(-10:0.2:10,-10:0.2:10);

[m,n] = size(c);

cg = zeros(m,n);

eps = 10^(-4);

v = 5;

bestc = 1;

bestg = 0.1;
```

```
bestacc = 0;
for i = 1:m
    for j = 1:n
        cmd = ['-v ',num2str(v),' -t 2',' -c ',num2str(2^c(i,j)),' -g ',num2str(2^g(i,j))];
        cg(i,j) = svmtrain(train_label,Train,cmd);
        if cg(i,j) > bestacc
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
        if abs( cg(i,j)-bestacc )<=eps && bestc > 2^c(i,j)
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
    end
end
```

```
cmd = [' -t 2',' -c ',num2str(bestc),' -g ',num2str(bestg)];

model = svmtrain(train_label,Train,cmd);

[predicted_label_1,accuracy_1,
decision_values_1]=svmpredict(train_label,Train,model,'-b 0')
test_label=[0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0]
[predicted_label_2,accuracy_2, decision_values_2]=svmpredict(test_label,Test,model,'-b
0')
```