JOURNAL OF RAMAN SPECTROSCOPY
J. Raman Spectrosc. 2002; **33**: 552–563
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jrs.877

JRS

Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis *in vivo*

Karen E. Shafer-Peltier,[1] Abigail S. Haka,[1] Maryann Fitzmaurice,[2] Joseph Crowe,[3] Jonathan Myles,[3] Ramachandra R. Dasari[1] and Michael S. Feld[1]*

[1] G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2] University Hospitals of Cleveland and Case Western Reserve University, Cleveland, Ohio, USA
[3] The Cleveland Clinic Foundation, Cleveland, Ohio, 44106, USA

Received 16 October 2001; Accepted 19 December 2001

Raman spectroscopy has the potential to provide real-time, *in situ* diagnosis of breast cancer during needle biopsy or surgery *via* an optical fiber probe. Understanding the chemical/morphological basis of the Raman spectrum of breast tissue is a necessary step in developing Raman spectroscopy as a tool for *in situ* breast cancer diagnosis. To understand the relationship between the Raman spectrum of a sample of breast tissue and its disease state, near-infrared Raman spectroscopic images of human breast tissue were acquired using a confocal microscope. These images were then compared with phase contrast and hematoxylin- and eosin-stained images to develop a chemical/morphological model of breast tissue Raman spectra. This model fits macroscopic tissue spectra with a linear combination of basis spectra derived from spectra of the cell cytoplasm, cell nucleus, fat, β-carotene, collagen, calcium hydroxyapatite, calcium oxalate dihydrate, cholesterol-like lipid deposits and water. Each basis spectrum represents data acquired from multiple patients and, when appropriate, from a variety of normal and diseased states. The model explains the spectral features of a range of normal and diseased breast tissue samples, including breast cancer. It can be used to relate the Raman spectrum of a breast tissue sample to diagnostic parameters used by pathologists.

INTRODUCTION

In 1999, ~176 000 new cases of breast cancer were diagnosed in the United States alone, 44 000 resulting in death.[1] In the last 20 years, there has been increasing interest in using optical techniques to diagnose breast cancer *in situ*.

Current methodologies, such as x-ray mammography and ultrasound, look for density changes in the breast. These techniques cannot reliably distinguish between benign and malignant tumors, and thus can only be used for detecting suspicious lesions and not for diagnosis. A tissue biopsy must be performed to determine whether or not a lesion is malignant, and 70–90% of breast biopsies are found to be benign upon pathological analysis.[2] However, instead of removing tissue for pathological analysis, it is possible to use optical techniques such as Raman spectroscopy to provide diagnostic information about a suspicious lesion *in situ*. Raman spectroscopy studies the spectral sidebands generated by the light scattered from a sample illuminated with monochromatic excitation light. Each chemical present has its own unique Raman spectral signature. By inserting a fiber-optic needle device into the breast it should be possible to collect Raman spectroscopic measurements from a lesion and extract chemical information almost instantaneously. Obtaining such information using a Raman needle device would result in more objective and faster (real-time) diagnosis and diminished trauma to the patient compared with biopsy techniques currently in use.

In addition to Raman spectroscopy, several other optical techniques are currently being explored. These include optical tomography, fiber-optic ductoscopy and fluorescence spectroscopy. Optical tomography uses visible or near-infrared light to illuminate a point on the surface of the breast, while a detector records the diffusely reflected or transmitted light at other points. In addition to providing information about the attenuation of the light signal as it traverses the breast, scattering and absorption information can also be extracted to measure quantitatively water, lipid and oxy-/deoxyhemoglobin concentrations. The use of this information to distinguish between benign and malignant tumors is under study.[3–6] Furthermore, an array of sources and detectors can be used

*Correspondence to: Michael S. Feld, Massachusetts Institute of Technology, 77 Massachusetts, Cambridge, Massachusetts 02139, USA. E-mail: msfeld@mit.edu
Contract/grant sponsor: NIH; Contract/grant number: P41-RR 02594.
Contract/grant sponsor: Pathology Associates of University Hospitals.

to form a measurement cup, allowing three-dimensional imaging.[7]

Fiber-optic ductoscopy adapts endoscopes developed to detect cancer in organs such as the colon, cervix and esophagus to the study of breast ducts. As most breast cancers and precancers start in the linings of the ducts and lobules, a very small fiberscope (<1 mm diameter) is introduced into the lactiferous duct through the nipple to look for intraductal abnormalities, primarily papillary lesions. The interior of the duct is illuminated and viewed via fiber-optics. The lactiferous duct, and its branches, can be observed using the device.[8]

Fluorescence spectroscopy has been used successfully to study cancerous lesions *in vivo* in the esophagus, colon, bladder and oral cavity.[9] Fluorescence spectroscopy of the breast has also been studied *ex vivo*, showing some promise for diagnosis, although as yet there is little understanding of the chemistry behind these results.[10,11] Fluorescence-based diagnosis is limited by the number of endogenous fluorophores present in breast tissue linked with cancer (primarily collagen and NADH). In comparison, there are many more Raman-active molecules present in tissue, which have been associated with cancer development (e.g. collagen, fibrinogen, DNA, calcium hydroxyapatite and various glycosaminoglycans).

Raman spectroscopy has been used for chemical analysis for many years, but only recently have researchers begun to apply it to biomedical problems. The ability to acquire Raman spectra in a clinical setting was made possible by the development of new technologies, such as compact diode lasers, CCD detectors and holographic notch filters. Each of these components contributes to the fabrication of compact, high-efficiency systems for medical diagnosis, previously unattainable.[12]

The first Raman spectroscopy measurements of human breast tissue used 1064 nm excitation coupled to a Fourier transform (FT) Raman system.[13] The investigators found that the relative intensities of the 1445 and 1651 cm^{-1} bands could be correlated with disease classification, although the biochemical origins of these spectral changes remained unexplained.

Redd and co-workers have since studied the Raman spectrum of breast tissue at a variety of excitation wavelengths, including 406, 488, 515, 647, 691, 784 and 830 nm.[14,15] Using 784 nm excitation to collect Raman spectra from normal, benign (fibrocystic disease) and malignant (infiltrating ductal carcinoma) breast tissue, they observed a shift of the 1439 cm^{-1} band in normal tissue to 1450 cm^{-1} in malignant tissue (due to changes in the chemical environment of the CH_2 bending mode). By using the area ratio of the 1654 (due to a combination of the C═C stretch and the amide I bands) and 1439 cm^{-1} bands, they were able to distinguish between malignant and normal tissue. They attributed this difference to increased protein concentrations in malignant samples. However, this test could not be used to distinguish benign from malignant lesions.

Each of these studies used excisional biopsy specimens, fixed in formalin. The fixation process chemically alters the tissue, primarily cross-linking the collagen proteins, and thus affects the Raman spectral signature of the tissue. As the ultimate goal is to use Raman spectroscopy to diagnosis tissue *in vivo*, we chose to study tissue that had been frozen and not fixed. Raman spectra of normal, benign and malignant breast tissue samples ($\sim$0.5 cm^3) recorded by our laboratory using 830 nm excitation have been reported previously.[16] Principal component analysis of this data permitted the differentiation of normal, benign and malignant tissue based on key spectroscopic features. However, principal component analysis does not allow the identification of the chemical or morphological origins of these spectroscopic signatures, and our data set at the time was too small for cross-validation (61 samples from 13 patients).

A clinical measurement of breast tissue using an optical fiber Raman needle probe will sample a region of tissue typically 1 mm^3 in volume. Cancer-related changes in the breast involve subtle alterations in the biochemical and morphological composition of the tissue. These changes occur at the microscopic level. Consequently, in order to develop a diagnostic algorithm that provides insight into the microscopic state of the tissue, it is important to characterize the Raman spectral features of the individual morphological components. A model employing these microscopic spectral features as building blocks to describe the macroscopic spectrum can then be used to extract information about the composition of the tissue at the microscopic level. By identifying the specific contributors to the Raman spectrum, a robust diagnostic algorithm can be developed.

In previous studies, Raman spectroscopy was used for quantitative biochemical analysis of atherosclerotic lesions in aorta and coronary artery tissue *in vitro*.[17] In these studies, the Raman spectrum of the tissue was modeled using a linear combination of Raman basis spectra collected from the major biochemicals present in arterial tissue. A related approach was instead to base the model on the Raman spectra of individual morphological features commonly found in artery, and to use these as the basis spectra for modeling.[18,19] The success of this approach has prompted the investigation of a similar morphological model for breast cancer diagnosis.

Why use morphologically derived basis spectra instead of pure chemicals? Primarily because the determination of which chemicals should be used to represent a morphological feature can be very difficult. For example, identifying every chemical in a complex mixture such as that found in a cell or tissue may not be possible. More importantly, those components that can be identified, such as collagen, may be present in human tissue in many different forms, each one having a slightly different Raman spectrum. The collagen found in breast tissue is, in fact, a combination

of several different types of collagen, but if each type of collagen were individually included in the model, this could lead to over-fitting. By using a single, morphologically derived collagen spectrum, one then obtains a picture of that chemical component in its microenvironment within normal or diseased tissue. Finally, chemicals purified in the laboratory or bought from commercial sources are not in their natural state. For instance, proteins such as collagen may have been exposed to caustic acids or other organic solvents. We can avoid all of these problems by using Raman spectra obtained from breast tissue itself. However, when necessary, synthesized or commercially available chemicals can be used.

In this paper, we present a morphological model of human breast tissue developed using a Raman confocal micro-imaging system. This model can characterize all of the spectroscopic features observed in macroscopic samples of breast tissue, both normal and diseased. It identifies the morphological components present in breast tissue through their unique Raman spectra, and uses them as building blocks to describe the morphological features of macroscopic samples.

EXPERIMENTAL

Breast tissue samples

Samples of breast tissue were obtained from surgical biopsy specimens. The samples were snap frozen in liquid nitrogen and stored at −85 °C until spectroscopic examination. Samples were then mounted on a cryostat chuck using Histoprep (Fisher Diagnostics, Orangeburg, NY, USA) and sliced into 6–8 μm thick sections using a microtome (International Equipment, Needham Heights, MA, USA). These sections were subsequently mounted on MgF_2 flats (Moose Hill Enterprises, Sperryville, VA, USA), selected because of their small Raman background signal, and kept moist with phosphate buffered saline (pH 7.4).

Raman confocal micro-imaging

Raman spectral images, produced using confocal Raman microspectroscopy, were collected from the unstained tissue sections and correlated with phase contrast images of the same section and serial hematoxylin and eosin-stained sections. The images were overlapped for comparison. When possible, examples of each morphological element were identified from a variety of patients and disease states. Spectra were then classified according to their morphological origin, i.e. as collagen fiber or epithelial cell, and the disease classification of the tissue sample. For example, initially extracellular matrix spectra from normal and malignant samples were kept separate. Once a library of spectra for each morphological element had been acquired, usually 60–80 spectra from 5–6 patients, they were analyzed for their degree of variation. If the spectra of a morphological element did not vary greatly or consistently, the spectra were averaged to create the morphologically derived basis spectrum used in our model. If consistent differences were observed, as was the case for the cellular components, the number of independently varying contributors was identified and used to extract independent basis spectra. In cases where single spectra had additional Raman bands when compared with other spectra in that morphological category, those spectra were removed from that category and analyzed independently to ensure that the additional spectral features could be explained by other elements in the model. If the spectral features could not be explained by the other elements of the model, a new basis spectrum was added to the model and the database of Raman micro-images was searched for similar spectral signatures. The phase contrast images and serial stained sections of all micro-images containing this new spectrum were reviewed. This methodology enabled new morphological features to be identified.

Raman spectroscopy of breast tissue and chemical components

Raman spectra were also collected from macroscopic breast tissue samples and various chemicals either synthesized in the laboratory or obtained from commercial sources. Raman spectra were obtained from the following commercially available chemicals (Sigma, St. Louis, MO, USA) for use in model development and image analysis: actin (chicken gizzard), β-carotene, calcium hydroxyapatite, cholesterol, cholesterol linoleate, collagen (bovine achilles tendon, type I), deoxyribonucleic acid (calf thymus), ribonucleic acid (calf liver), phosphatidylcholine and triolein, and also calcium oxalate, which was synthesized[20] in our laboratory.

Spectroscopy and micro-imaging system

A schematic diagram of the experimental setup is shown in Fig. 1. The same system was used for both macroscopic tissue samples and micro-imaging. The Raman excitation light (830 nm), provided by an argon ion laser-pumped Ti : sapphire laser (Coherent Innova 90/Spectra-Physics 3900S, Coherent, Santa Clara, CA, USA), traversed a band-pass filter (Kaiser Optical Systems, Ann Arbor, MI, USA) and was launched into either an aluminum holder for macroscopic tissue samples via a prism or into an epi-illuminated microscope (Zeiss Axioskop 50, Zeiss, Thornwood, NY, USA; axial resolution ~1 μm) for Raman micro-imaging using two mirrors. The microscope objective both focused the excitation and collected the Raman scattered light in a backscattering geometry. A dichroic beamsplitter and mirror combination redirected the Raman-scattered light from the microscope through a confocal pinhole of variable diameter to increase axial resolution. If the macroscopic assembly was used, a camera lens collected the Raman scattered light. The diameter of the light spot on a macroscopic tissue sample was ~1 mm, and the tissue volume sampled was typically 1 mm^3. For both configurations, the light passed through a holographic notch filter (Kaiser Optical Systems) and was then focused into a

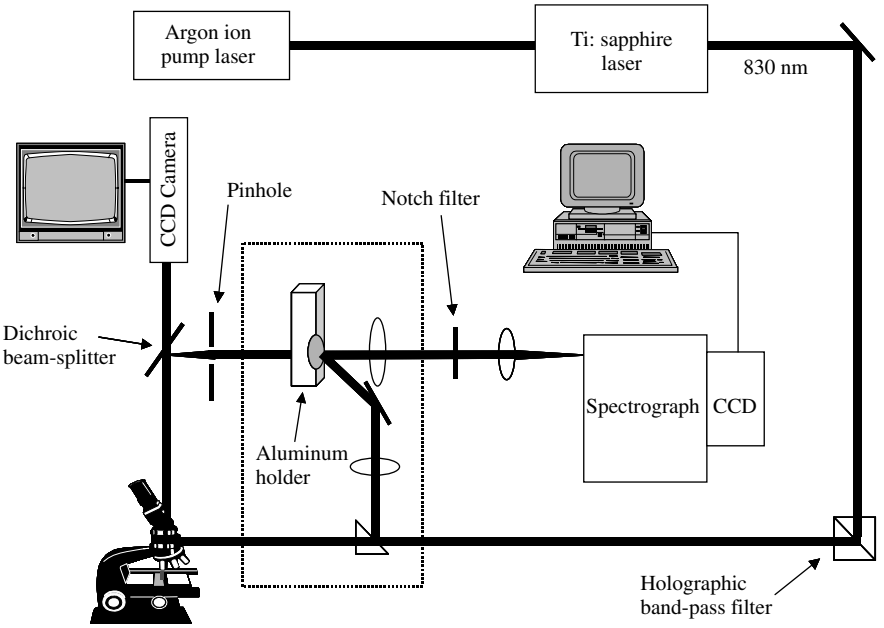

Figure 1. Schematic diagram of the Raman microscope system. The components for collecting macroscopic measurements, shown in dotted box, are displaced during microscopic data acquisition.

0.25 m $f/4$ imaging spectrograph (Model 250IS/SM spectrograph monochromator, Chromex, Albuquerque, NM, USA) attached to a liquid nitrogen cooled CCD detector (Princeton Instruments, Princeton, NJ, USA). At the smallest confocal aperture diameter (~100 μm) the spatial resolution of the microscope system was ~2 μm^3.[19]

The spectrograph itself had an adjustable slit and a turret, which held three gratings (Chromex) for a range of measurements. For the Raman studies, a 600 groove mm^{-1} grating blazed at 1 μm was used along with the 140 μm spectrograph entrance slit setting, providing ~8 cm^{-1} resolution. As most biological samples do not exhibit Raman bandwidths narrower than 10 cm^{-1}, a spectrograph entrance slit 140 μm wide was generally used, providing maximized optical throughput (sometimes a 70 μm entrance slit was used for macroscopic measurements.)

A CCD camera (Sony, Cambridge, MA, USA) atop the microscope allowed for registration of the focused laser spot with a white light transilluminated image and recording of the image on a videotape. The microscope itself was equipped with a range of objectives, both normal and phase contrast. Typically, for Raman studies a 63× infinity-corrected water immersion objective (Zeiss Achroplan, numerical aperture 0.9) was used. Both the detector and the microscope translation stage were computer controlled. A complete Raman spectrum was collected at each tissue location, and spectral micro-images of the tissue were then created by moving the translation stage (Prior Scientific Instruments, Cambridge, MA, USA) in a raster-scan pattern under the microscope objective.[21] This method produced an image of typically $50 \times 50\ \mu m^2$, each pixel of which contained an entire Raman spectrum from 400 to 1850 cm^{-1}. The step size in both the x and y directions was typically 2 μm, consistent with the spatial resolution of the confocal microscope. Spectra were usually collected for 20 s at each pixel location at a power between 50 and 100 mW, hence an entire image required ~3.5 h. A droplet of phosphate-buffered saline kept the tissue section moist during data collection. Experiments were also performed to test for photochemical damage to the tissue. At 220 mW, the intensity of the fluorescence signal was observed to decrease with increased exposure time (over a period of minutes), whereas the Raman signal remained unaffected.[22] At this power photobleaching occurred, but its effect on the Raman signal was negligible. This photobleaching effect was not observed with lower excitation powers. As a result, the power was kept below 100 mW for 20–30 s exposures to reduce the effect of photobleaching during the collection of spectroscopic data.

Data processing and model fitting

The Raman spectra acquired underwent processing to ensure reproducibility of the data from day to day. First, they were corrected for the spectral response of the system using a tungsten light source. Then data were frequency calibrated using the known Raman lines of toluene. The MgF_2 background spectrum was then subtracted and the broad, slowly varying fluorescence background was removed by fitting the spectrum to a fifth-order polynomial (in wavenumber), and then subtracting this polynomial from the spectrum. Also, contributions from cosmic rays were removed, if necessary, using a derivative filter.

Micro-imaging or macroscopic tissue spectroscopic data were fitted simultaneously with the model basis spectra using MATLAB's non-negative least-squares fitting algorithm (MathWorks, Natick, MA, USA). For an estimate of the number of independently varying components, principal component analysis was used (MATLAB). In order to use either least-squares fitting techniques or principal component analysis, each Raman spectrum was represented as a vector of intensity values corresponding to each wavelength. A comprehensive discussion of ordinary least-squares, principal component analysis and other multivariate analysis techniques can be found in Hanlon *et al*.[12]

Another key issue when using our linear model was the orthogonality of the basis spectra. The degree of orthogonality of the elements was tested using the equation

$$\frac{x^{\mathrm{T}}y}{(x^{\mathrm{T}}x)(y^{\mathrm{T}}y)} \tag{1}$$

where x and y represent basis spectra of two morphological components, arranged as Raman intensities at each wavelength (x^{T} is the transpose of x). A value of zero indicates that the vectors are orthogonal and a value of one means that they are identical.

The level of error in the morphological model is determined by the signal-to-noise ratio of the spectra being used. Provided that the model basis spectra are not identical within the limits of the noise (i.e. they are more orthogonal than identical spectra 'altered' by noise), the ordinary least squares method can be used to separate them. Since the basis spectra are the average of many data points, collected for as long as necessary, they are virtually noiseless. Therefore, the limiting source of error in the model is due to the noise in the data being fitted. The error in the fit contribution of a particular basis spectrum is proportional to the noise in the spectrum being fitted:

$$E = NB \tag{2}$$

where $B = P^{\mathrm{T}}(PP^{\mathrm{T}})^{-1}$ is the calibration vector for the morphological basis spectrum P and N is the noise in the sample.[12,23]

RESULTS

Normal breast and breast cancer morphology and chemistry

Raman spectroscopy was used to extract information about the morphological and chemical components present in relatively large abundance in breast tissue, reviewed here. The breast contains two types of tissue: glandular and stromal. The glandular elements consist of lobules and ducts. The lobules are dense clusters of epithelial cells, which produce and secrete milk into a system of ducts that transport the milk to the nipple. The ducts consist of an inner layer of epithelial cells surrounded by a layer of myoepithelial cells.[24] Both layers are enclosed by a basement membrane, made primarily of collagen. The stromal elements provide the supportive network for these glandular units and include the extracellular matrix, fibroblasts, fat and blood vessels.[25] Whereas the glandular elements of the breast are mostly cellular, there are only a small number of cells in the stroma. Most of these cells are fibroblasts, responsible for producing the extracellular matrix, a supportive network of structural proteins and carbohydrates, mainly collagen and glycosaminoglycans. Fat is the only other major morphological structure present and makes up the bulk of normal breast tissue. Sometimes crystalline deposits of β-carotene, a lipophilic precursor to vitamin A, are also present.

Many of the morphological structures in benign and malignant breast lesions are similar to those in normal breast tissue. For example, fibrosis occurs in both benign and malignant breast lesions and involves a proliferation of the stroma. Fibrotic tissue is mainly collagen in composition, like most of the extracellular matrix, with an increase in the presence of proteins such as fibrinogen and fibronectin.

However, some of the morphological features of diseased breast are different from those in normal breast tissue. For example, breast cancer most commonly originates in the lobules and ducts as a rapid proliferation of epithelial cells, associated with nuclear enlargement, pleomorphism (variation in size and shape) and hyperchromatism (darker staining), atypical mitoses and DNA aneuploidy (gain or loss of a chromosome).[24] These morphological changes are not associated with a large-scale production of new chemicals, but rather a change in the relative concentrations of chemicals that are already present in the breast. For example, the above morphological changes are associated with a change in the nucleus-to-cytoplasm ratio, a qualitative indicator of malignancy used by pathologists.

Two additional morphological features that can be observed in breast cancer are calcifications and necrosis. Calcifications are important since they are radiodense, can be detected mammographically and are often seen in association with cancer. There are two major types that have similar morphological characteristics on mammograms. Type I calcifications are calcium oxalate dihydrate crystals, whereas type II calcifications are mainly calcium hydroxyapatite but contain other calcium phosphates and possibly also calcium carbonate.[26] Calcium oxalate crystals are more often found in benign than in malignant lesions and are thought to be the product of cellular secretions, whereas calcium hydroxyapatite deposits are found in both benign and malignant lesions and are thought to be the result of cellular degradation or necrosis (death).

With this basic knowledge of breast chemistry and architecture, and the changes induced by disease progression, it is possible to explain all of the major Raman spectral features of normal and diseased breast tissue.

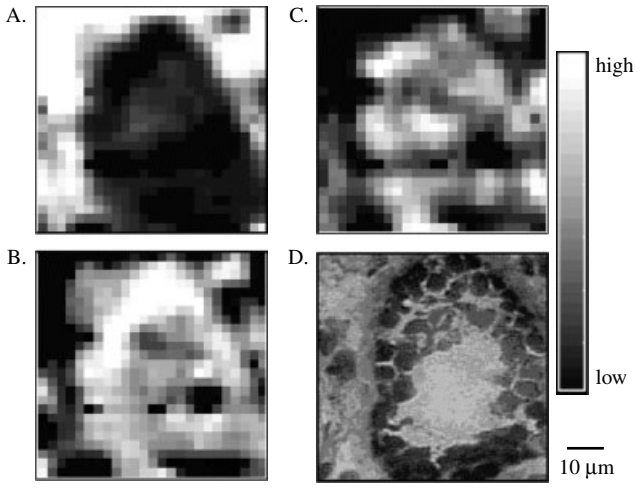

Figure 2. Raman images of normal breast duct [(A)–(C)] with corresponding serial stained section (D). Each image represents the contribution of a specific morphological element to the region being studied. (A) Collagen; (B) cell cytoplasm; (C) cell nucleus.

Raman imaging

We collected more than 60 Raman images from samples of normal, benign and malignant breast tissue. Raman images of a normal breast duct are shown in Fig. 2. Micro-images of collagen, cell cytoplasm and cell nucleus are produced by ordinary least-squares fitting of each data point in the image with these basis spectra. The serial stained section is shown for comparison. It is evident that the structures observed in the Raman images correlate well with the tissue architecture.

From the micro-imaging data, nine key basis spectra were identified: cell cytoplasm, cell nucleus, collagen, fat, cholesterol-like, β-carotene, calcium hydroxyapatite, calcium oxalate dihydrate and water. Some features were identified using Raman imaging, such as the cell membrane, but were not included in the model because they are not present in large quantities and have small Raman cross-sections, and therefore do not contribute significantly to macroscopic tissue spectra. Others were found to have virtually the same chemical composition as elements already in the model, and therefore could not be included as separate morphological features, as was the case for the basement membrane, which is composed mostly of collagen like the extracellular matrix. The number of spectra used to determine a model component spectrum depended on that morphological element's abundance, and also signal-to-noise ratio issues. For example, fat has an extremely strong Raman cross-section. As a result, very few fat spectra were needed from each patient to produce a clean spectrum. Both the extracellular matrix and the cellular components discussed required more spectra to increase the signal-to-noise ratio. The basis spectra used for the complete model of breast tissue are shown in Fig. 3.

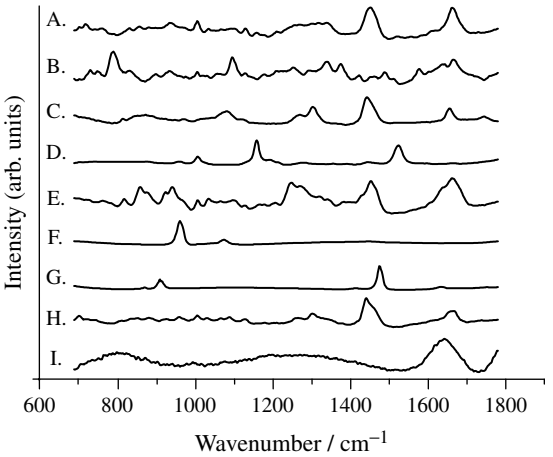

Figure 3. Basis spectra used in the morphological model of the breast. (A) Cell cytoplasm; (B) cell nucleus; (C) fat; (D) β-carotene; (E) collagen; (F) calcium hydroxyapatite; (G) calcium oxalate; (H) cholesterol-like; (I) water.

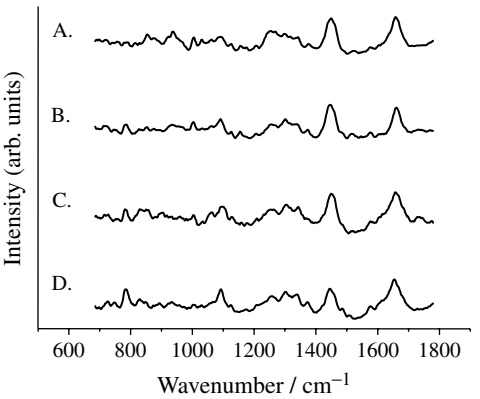

Figure 4. Raman spectra of four types of cells observed in normal or diseased human breast tissue. (A) Fibroblast (normal stroma); (B) epithelial cell (fibrocystic disease); (C) epithelial cell (normal duct); (D) malignant cell.

Cell nucleus and cell cytoplasm

In Fig. 4, spectra of a fibroblast and epithelial cells taken from normal, fibrocystic and malignant ducts are compared. Statistical analysis indicated that there are two major independently varying components, originating from the cell cytoplasm and cell nucleus. The spectrum of DNA [Fig. 5(A)] was very similar to that of the cell nucleus [Fig. 5(B)], although the cell nucleus spectrum also contained minor features related to RNA and histones. Similarly, the spectrum of actin [Fig. 5(C)] was the major contributor to the cell cytoplasm spectrum [Fig. 5(D)]. The cell cytoplasm spectrum also included minor features related to other elements found in the cytoplasm.

Because the ability to collect pure spectra from the cell cytoplasm and the cell nucleus was limited by the collection volume of the Raman confocal microscope, the two basis

JRS

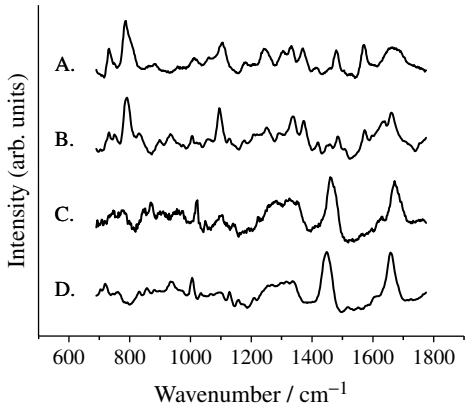

Figure 5. Comparison of commercially available [(A), (C)] and morphologically derived [(B), (D)] Raman spectra observed in cells. (A) DNA (Sigma); (B) cell nucleus (breast tissue); (C) actin (Sigma); (D) cell cytoplasm (breast tissue).

spectra were separated mathematically. To separate the two components, spectra of hundreds of cells (all types) from eight patients were collected using the Raman imaging system. Initially the spectra were fitted with two basis spectra, one taken from a cellular region with low nuclear content (determined by looking at the Raman signal) and one from purified DNA. Spectra with especially high DNA fit coefficients (DNA-rich), corresponding to spectra taken from the nuclear regions, were then separated from those spectra with little to no DNA (DNA-poor), collected from regions in the cell cytoplasm. The mean DNA-rich spectrum was then scaled and subtracted from the mean DNA-poor spectrum to produce a new cytoplasm-only spectrum with no DNA content. This new cytoplasm-only spectrum was then subtracted from the mean DNA-rich spectrum to remove all cytoplasm features, leaving a spectrum representative of only the nuclear material. The original data (both DNA-rich and DNA-poor) were then fitted with these two modified basis spectra. The procedure was repeated, using the two modified basis spectra rather than the purified DNA and the low nuclear content spectra, to produce the final cell cytoplasm and cell nucleus spectra. By using this iterative process, artifacts due to the inability of the purified DNA spectrum to model the nucleus (which contains DNA, RNA, histones and more) were minimized. These two basis spectra can be used to extract key diagnostic information about the cells, such as the nuclear-to-cytoplasm ratio.

Collagen (extracellular matrix and basement membrane)

Both the extracellular matrix and the basement membrane are composed primarily of collagen. Other structural proteins, such as fibrinogen and fibronectin, and proteoglycans are also present, but in such minute quantities and with such small Raman cross-sections that they did not contribute significantly to the overall Raman spectrum. Figure 6 compares

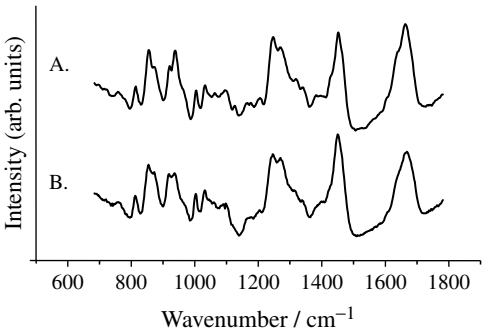

Figure 6. Comparison of (A) purified collagen and (B) morphologically derived collagen.

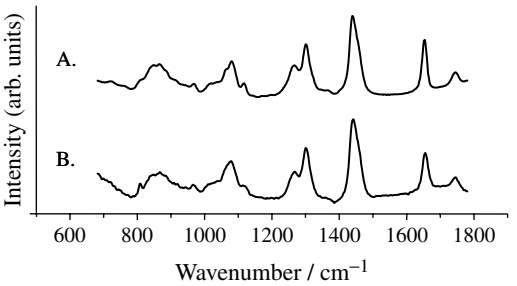

Figure 7. Comparison of (A) purified triolein and (B) morphologically derived fat.

the spectra of morphologically derived collagen (mostly type I, but some types III, IV and V were also present[27]) and that of purified collagen (type I). They are very similar, although a few minor differences can be observed in the region between 800 and 1200 cm^{-1}. The morphologically derived collagen spectrum was the mean of 215 spectra taken from seven patients, mostly from regions of extracellular matrix.

Fat

Fat is one of the strongest contributors to the Raman spectrum of normal breast tissue. It is present in large quantities and has a strong Raman cross-section. Its storage in humans primarily takes the form of triglycerides, especially triolein. Figure 7 compares the Raman spectrum of fat acquired from breast tissue with that of triolein, showing that, as expected, triolein was the major contributor to the spectrum. The fat spectrum included in our model and shown in Fig. 7 was the average of 28 spectra collected using data from five patients.

Cholesterol-like (necrosis)

Necrosis within the lumen of a malignant duct or the center of a malignant tumor is essentially the product of cellular degradation. Consequently, its composition varied significantly from location to location within even a single duct. Analysis of Raman spectra from three patients indicated that the necrotic material contained fat, collagen, calcification (calcium hydroxyapatite), free cholesterol and cholesterol ester (linoleate), in addition to cellular material (both cell

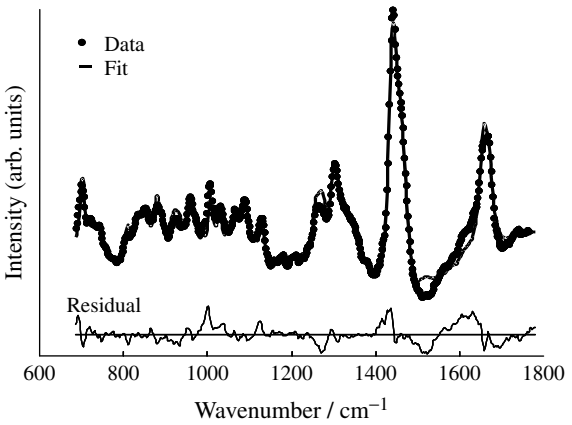

Figure 8. Spectrum of necrotic core ('cholesterol-like') fitted with a cell cytoplasm, cell nucleus, fat, cholesterol linoleate and cholesterol.

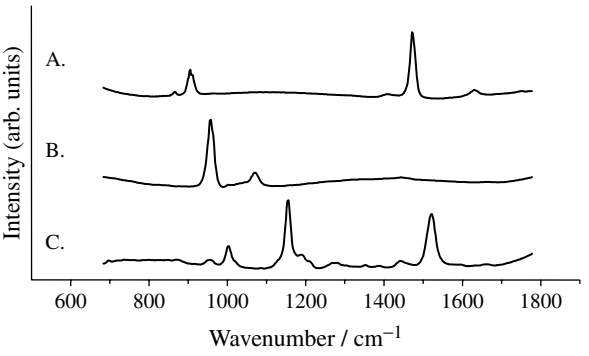

Figure 9. Spectra of breast deposits. (A) Calcium oxalate dihydrate; (B) calcium hydroxyapatite; (C) β-carotene.

cytoplasm and cell nucleus). As the ratios of these elements could vary significantly, the spectrum included in our model ('cholesterol-like') represents the common elements of these spectra not represented elsewhere in the model, mainly the cholesterol components, collected from a single patient. Chemical modeling tells us that the 'cholesterol-like' spectrum has major contributions from cholesterol and cholesterol linoleate, with minor contributions from cellular material (cell cytoplasm and cell nucleus) and fat. Figure 8 shows an ordinary least-squares fit to the data using these five elements. Pure chemical spectra of cholesterol and cholesterol linoleate were not used, because there is more than one type of cholesterol ester present that cannot be individually determined. As was the case for collagen, it is best to model the tissue using a biologically derived mixture than with one or two pure components. The necrotic material was not the only element in breast tissue containing cholesterol and cholesterol esters. Cell membranes also contain both of these chemicals, although they also include other chemicals, such as phospholipids. Thus, the 'cholesterol-like' basis spectrum was found to be present in small quantities in all tissue spectra (not just malignant specimens).

Calcifications

Calcium hydroxyapatite and calcium oxalate dihydrate both have very strong Raman spectra [Fig. 9(A) and (B)]. However, they were not commonly found in our frozen breast tissue specimens, because calcifications are important for medical diagnosis and therefore tissue containing calcifications is generally not released for scientific study. Although calcifications were found in occasional frozen specimens, they were often punctate calcifications and difficult to study. For these reasons, we expanded our study to include spectra obtained from 6 μm thick deparaffinized sections of breast tissue fixed in formalin, in which calcifications were larger and more numerous. The fixation process altered the tissue proteins, but did not affect the relatively inert mineral deposits in the calcifications. Therefore, deparaffinized sections could be used to analyze a larger number of calcifications, identified for us by an experienced pathologist, from a range of patients and disease states.

Calcium hydroxyapatite was identified in frozen sections from three patients and from deparaffinized tissue sections in an additional 11 patients. The spectra from the frozen and deparaffinized samples were the same. The calcium hydroxyapatite basis spectrum used was acquired from a combination of these spectra. Calcium oxalate was only observed in one deparaffinized tissue section, owing to its rarity. Although its presence in breast tissue is well documented, calcium oxalate dihydrate is significantly less common than calcium hydroxyapatite in breast tissue.[26] Therefore, calcium oxalate dihydrate was synthesized in our laboratory for incorporation into our model. Both calcification spectra were consistent with previously published spectra.[28,29]

β-Carotene

β-Carotene is resonance enhanced when excited with 830 nm radiation. As a result, it has an extremely strong Raman signal. Although its peaks stand out, it is often found in conjunction with fat throughout the breast. To eliminate the need for extracting the fat content from the morphologically derived β-carotene spectra (of which we collected hundreds), we used the spectrum acquired from commercially available β-carotene [Fig. 9(C)]. Using our morphologically derived Raman spectra of β-carotene, we were able to confirm that the commercially available sample was an accurate representation of the β-carotene found in tissue.

Water

Although water is a weak Raman scatterer, it contributed to the spectrum through sheer volume. Water constitutes ~80% by weight of human tissue and is present in the phosphate-buffered saline used to keep the tissue moist. Previously, in studies of artery, it was determined that water

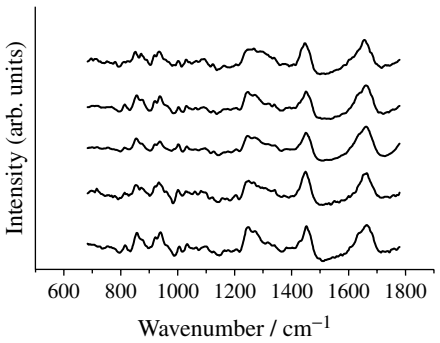

Figure 10. Raman spectra collected from the extracellular matrix of five patients.

did not contribute significantly to the Raman spectrum of human tissue.[30] However, in these studies of breast tissue its inclusion was found to be essential for fitting the data properly. Water has a single, relatively broad Raman band centered at 1650 cm^{-1}. If water was not included in the model, fitting of this band by the other morphological components was incomplete.

Model evaluation

One of the key requirements for successful morphological modeling was that there be very little inter-patient variation in the Raman spectra of a given morphological structure. By developing a model through averaging several spectra from many patients, we could ensure that our model includes the common elements of all morphological features. But were all of the extracellular matrix spectra really that similar? The answer is yes. The extracellular matrix spectrum is primarily collagen, regardless of the patient. In Fig. 10, the extracellular matrix spectra from five patients are shown. The interpatient variability is similar for all morphological features. The minor differences observed were due to the close proximity of other morphological features, i.e. a small fat droplet close to a collagen fiber being studied might result in small amounts of fat being observed in addition to collagen. We found that in the development of a Raman model of breast tissue, the lineshape variability unexplained by other basis spectra in the model was not significant.

When analyzing the orthogonality of the model components, four components were found to have values >0.5 when compared with each other: cell cytoplasm, fat, collagen and cholesterol-like (Table 1). This makes sense, as these four elements have many of the same functional groups (CH_2 bends, C–C stretches, etc.). Still, they were sufficiently orthogonal to be differentiated amongst when using ordinary least-squares fitting. Water and cell nucleus also overlapped considerably. Nonetheless, the key to successful fitting was to use as few elements as possible, while retaining relevant spectral information in order to avoid over-determining the spectrum. Despite having the highest degree of overlap (0.89), the differences between the fat and cholesterol-like spectra are greater than the noise component of the data fit with the model in Figs 11 and 12. Hence their incorporation in the model is reasonable. As discussed earlier, in this situation, the predictive value of the model is dependent on the signal-to-noise ratio of the data being fitted.

Using the morphological model developed here, the spectral features of a range of macroscopic tissue samples can be explained in terms of each sample's morphological composition. In Figs 11 and 12, Raman spectra from normal, fibrosis, adenosis, fibrosis/cysts, fibroadenoma and infiltrating ductal carcinoma tissue samples were fitted to a linear combination of the basis spectra of the morphological model. The fit coefficients given by the model (also shown in Figs 11 and 12), normalized to sum to one, represent percentage contributions of the normalized chemical and morphological basis spectra to the bulk tissue spectrum (excluding water, which varies independently). For example, the fibroadenoma and malignant samples shown in Fig. 12 both have a large cell cytoplasm content (31 and 34%, respectively) whereas the normal sample shown here has none. This observation reflects the greater cellularity of infiltrating carcinoma and fibroadenoma as compared with normal tissue or even the other benign lesions, which was confirmed by subsequent microscopic analysis of the samples by an

Table 1. Degree of orthogonality of the morphological model components

	Cell cytoplasm	Cell nucleus	Collagen	Fat	β-Carotene	Cholesterol-like	Calcium hydroxy apatite	Calcium oxalate	Water
Cell cytoplasm	1								
Cell nucleus	0.22	1							
Collagen	0.83	0.29	1						
Fat	0.73	0.08	0.58	1					
β-carotene	0.27	0.36	0.35	0.29	1				
Cholesterol-like	0.88	0.07	0.68	0.89	0.28	1			
Calcium hydroxy apatite	0.11	0.10	0.06	0.06	0.07	0.13	1		
Calcium oxalate	0.11	0.12	0.06	0.10	0.10	0.13	0.00	1	
Water	0.26	0.61	0.46	0.01	0.16	0.14	0.20	0.17	1

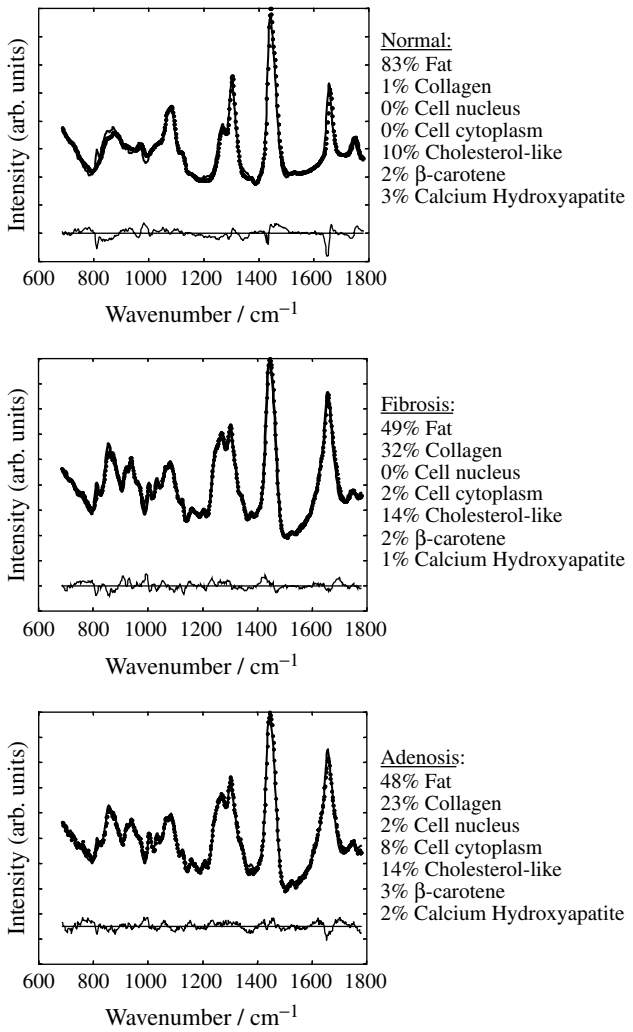

Figure 11. Demonstration of the quality of our model's fit to macroscopic breast tissue samples: normal (●) fit with model (—), fibrosis and adenosis. Below each spectrum is plotted the residual of the fit (with the zero line drawn). The percentages given at the side represent the fit coefficients of the basis spectra, normalized to sum to one (fit coefficient of water is not included in summation).

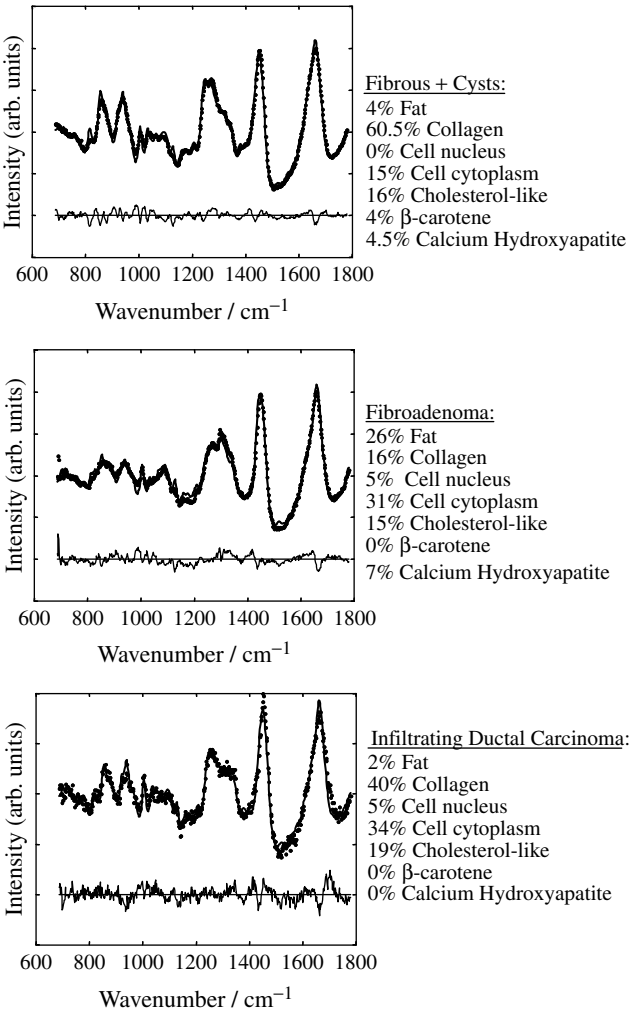

Figure 12. Further demonstration of the quality of our model's fit to macroscopic breast tissue samples: fibrosis + cysts (●) fit with model (—), fibroadenoma and infiltrating ductal carcinoma (residual plotted below).

experienced pathologist. The strong correlation between the model fit coefficients and the morphological changes known to accompany disease attests to the accuracy of the model. The small residuals observed in both Figs 11 and 12 indicate that all of the major spectroscopic features are explained by the model. Similarly, small residuals were observed when 101 macroscopic tissue spectra, collected from 37 patients representing a range of disease states, were fitted with the morphological model.

DISCUSSION

By comparing Raman images with phase contrast images, and also serial stained sections of the same tissue, it is possible to monitor spectral and thus chemical changes across a tissue surface. For example, not only can one compare spectra of ductal epithelial cells found in malignant tissue with those found in normal or benign tissue, but also progressive changes in these spectra can be monitored as the transition is made between a region of infiltrating carcinoma and one unaffected by the disease process within the same tissue section. Imaging also allows the identification of chemical/morphological differences that are not made visible by phase contrast or staining. With such information, a better understanding of the disease process and how it affects both the morphology and the chemistry of the tissue could be acquired, and a morphological model was developed.

Construction of a morphological model of breast tissue relied on three assumptions: first, that the Raman spectrum of a mixture was equal to the weighted linear sum of the individual components of the mixture; second, that biological

morphological features, such as cells, had the same Raman spectrum from one patient to another; and third that the basis spectra included in the model were sufficiently distinct to enable their differentiation based on their Raman spectrum. The first assumption was found to hold within experimental accuracy by Manoharan *et al*.[17] The other two were shown to be true by the observations reported here.

Although only some of the Raman micro-images collected were used to create the model presented, all of them were used to test the model's comprehensiveness. By using spectral data from a wide variety of patients with different pathologies, we ensured that our model explains all the major spectral features found in breast tissue including breast cancer. Excellent model fits also confirmed that the Raman spectrum of breast tissue is equal to the weighted linear sum of the spectra of the nine morphological/chemical elements included in the model. Each of the elements included had a strong spectroscopic signature, varied little from patient to patient and, except for calcium oxalate dihydrate, was present in large quantities. Some elements were not independently considered because their Raman spectrum overlapped too much with those of other elements. This overlap was an issue for the many cell types (epithelial, fibroblast, etc.), the basement membrane and the cell membrane (which contributes weakly to the tissue spectrum but was very similar to the necrotic material spectrum). Other chemicals present in breast tissue contributed so little to the aggregate Raman spectrum that they were insignificant. For example, glycosaminoglycans are present in the extracellular matrix in large quantities but have very weak Raman cross-sections, whereas matrix metalloproteinases are present in small quantities. Neither was observed in the breast tissue Raman spectrum.

The chemical composition of the morphological features identified by Raman micro-imaging was as expected. For example, the extracellular matrix was found to be mostly collagen, whereas fat droplets were primarily triolein. The cell types examined (fibroblasts, epithelial cells from a range of normal and diseased states and inflammatory cells) were all composed of the same basic components, cholesterol and cholesterol linoleate, actin and DNA. Each cell is enclosed by a cell membrane, mainly a lipid bilayer composed of phospholipids, cholesterols, triglycerides and some proteins. Making up the bulk of the cell is the cell cytoplasm, mostly the cytosol, an aqueous solution that fills the cell. Within the cytoplasm is the cytoskeleton, composed primarily of actin filaments, which allows controlled movement and organization within the cell; RNA and proteins involved in the machinery of the cell (mostly making and regulating the production of more proteins); and various organelles. The largest of these organelles is the cell nucleus. The nucleus is rich in DNA, RNA and histones (involved in helping DNA to form a compact structure).

Depending on the function of the cell, it will have varying amounts of each of these components and possibly a few additional ones. For example, fibroblasts are responsible for making and maintaining the extracellular matrix. In order to do so, they must produce collagen, fibrinogen and glycosaminoglycans within their cytoplasm and export them to the extracellular space. However, in terms of developing a Raman model of breast tissue, these components are already included in the spectrum of collagen, and therefore need not be considered independently.

We expect most differences among cells, either within a type or between types, to be observed in the ratio of the cell cytoplasm to the cell nucleus. It is natural that there be some variation in this ratio, but it should be exaggerated greatly in malignant cells due to the occurrence of aneuploidy[31] and is used by pathologists to diagnose malignancy. Parameters such as the nuclear-to-cytoplasm ratio will be measurable in macroscopic tissue specimens in the future.

A number of non-cellular components were also found to be significant for modeling the Raman spectrum of breast tissue: collagen (extracellular matrix and basement membrane), fat, cholesterol-like (necrosis), calcium hydroxyapatite, calcium oxalate and β-carotene. Some of these, such as β-carotene, were significant only because they are strong Raman scatterers and therefore needed for good model fits. Others, such as 'cholesterol-like,' are also key features used by pathologists to diagnose malignancy.

We now know that the proteins that contribute the most to the Raman spectrum of breast tissue are collagen and actin. Collagen is representative of the extracellular matrix while actin is found in cells. As both are proteins, their Raman spectra are very similar, especially in the 1440–1660 cm^{-1} region, where researchers have previously looked for differences among normal, benign and malignant lesions. However, if one uses the information contained in these basis spectra to fit macroscopic tissue spectra in the model, it is possible to extract information about the relative quantities of cellular material (actin) and extracellular matrix (collagen) in a particular sample. This information has been used to develop an algorithm based on Raman spectroscopy to diagnose breast cancer, which will be presented in a future publication.

CONCLUSIONS

Raman spectroscopy has the potential to provide real-time, *in situ* diagnosis of breast cancer during needle biopsy or surgery via an optical fiber probe. Understanding the chemical/morphological basis of the Raman spectrum of breast tissue is a necessary step in developing Raman spectroscopy as a diagnostic tool. Near-infrared Raman micro-images of human breast tissue were acquired using a confocal microscope, and then used to develop a model of breast tissue Raman spectra. Nine morphological/chemical basis spectra were used in the model to explain the major Raman spectral features of a range of normal and diseased

human breast tissues, including breast cancer. It was possible to extract information about the relative quantities of various cellular and extracellular components in a particular breast tissue, by fitting its spectrum in the model using these basis spectra. This information was consistent with the known pathology of the tissue and is being used to develop an algorithm based on Raman spectroscopy to diagnose breast cancer.

Acknowledgments

The work was carried out at the MIT Laser Biomedical Research Center supported by NIH grant P41- RR 02594. We thank Pathology Associates of University Hospitals for funding part of this study. Tissue was provided in part by the Cooperative Human Tissue Network of the National Cancer Institute and by the Cleveland Clinic Foundation.

REFERENCES

1. Landis SH, Murray T, Bolden S, Wingo PA. *CA–Cancer J. Clin.* 1999; **49**: 8.
2. Johnson J, Dalton R, Wester S, Landercasper J, Lambert P. *Arch. Surg.* 1999; **134**: 712.
3. Quaresima V, Matcher SJ, Ferrari M. *Photochem. Photobiol.* 1998; **67**: 4.
4. Tromberg BJ, Coquoz O, Fishkin JB, Pham T, Anderson ER, Butler J, Cahn M, Gross JD, Venugopalan V, Pham D. *Philos. Trans. R. Soc. London, Ser. B* 1997; 661.
5. Fantini S, Walker SA, Franceschini MA, Kaschke M, Schlag PM, Moesta KT. *Appl. Opt.* 1998; **37**: 1982.
6. Ntziachristos V, Chance B. *Breast Cancer Res.* 2001; **3**: 41.
7. Hebden J, Veenstra H, Dehghani H, Hillman E, Schweiger M, Arridge S, Delpy D. *Appl. Opt.* 2001; **40**: 3278.
8. Shen K-W, Wu J, Lu J-S, Han Q-X, Shen Z-Z, Nguyen M, Shao Z-M, Barsky SH. *Cancer* 2000; **89**: 1512.
9. Zangaro RA, Silveria L, Manoharan R, Zonios G, Itzkan I, Dasari RR, Van Dam J, Feld MS. *Appl. Opt.* 1996; **35**: 5211.
10. Liu CH, Das BB, Glassman WLS, Tang GC, Yoo KM, Zhu HR, Akins DL, Lubicz SS, Cleary J, Prudente R, Celmer E, Caron A, Alfano RR. *J. Photochem. Photobiol. B* 1992; **16**: 187.
11. Gupta PK, Majumder SK, Uppal A. *Lasers Surg. Med.* 1997; **21**: 417.
12. Hanlon EB, Manoharan R, Koo T-W, Shafer KE, Motz JT, Fitzmaurice M, Kramer JR, Itzkan I, Dasari RR, Feld MS. *Phys. Med. Biol.* 2000; **45**: R1.
13. Alfano RR, Liu CH, Sha WL, Zhu HR, Akins DL, Cleary J, Prudente R, Cellmer E. *Lasers Life Sci.* 1991; **4**: 23.
14. Redd DCB, Feng ZC, Yue KT, Gansler TS. *Appl. Spectrosc.* 1993; **47**: 787.
15. Frank CJ, Redd DCB, Gansler TS, McCreery RL. R.L.M. *Anal. Chem.* 1994; **66**: 319.
16. Manoharan R, Shafer K, Perelman L, Wu J, Chen K, Deinum G, Fitzmaurice M, Myles J, Crowe J, Dasari R, Feld M. *Photochem. Photobiol.* 1998; **67**: 15.
17. Manoharan R, Baraga JJ, Feld MS, Rava RP. *J. Photochem. Photobiol. B: Biol.* 1992; **16**: 211.
18. Buschman HP, Motz JT, Deinum G, Romer TJ, Fitzmaurice M, Kramer JR, van der Laarse A, Bruschke AV, Feld MS. *Cardio. Pathol.* 2001; **10**: 59.
19. Buschman HPJ, Deinum G, Motz JT, van der Laarse A, Bruschke AVG, Feld MS. *Cardio. Pathol.* 2001; **10**: 69.
20. Kontoyannis CG, Bouropoulos NC, Koutsoukos PG. *Appl. Spectrosc.* 1997; **51**: 64.
21. Shafer-Peltier KE, Haka AS, Motz JT, Gardecki JA, Fitzmaurice M, Dasari RR, Feld MS. *J. Raman Spectrosc.* in preparation.
22. Shafer KE. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.
23. Sharma S. *Applied Multivariate Techniques*. John Wiley & Sons: New York, 1996.
24. Cotran RS, Kumar V, Collins T. *Robbins Pathologic Basis of Disease*. Saunders: Philadelphia, PA, 1999.
25. Sternberg SS. *Histopathology for Pathologists*. Lippincott-Raven: Philadelphia, PA, 1997.
26. Radi MJ. *Arch. Pathol. Lab. Med.* 1989; **113**: 1367.
27. Alini M, Losa GA. *Cancer Res.* 1991; **51**: 1443.
28. Kodati VR, Tomasi GE, Turumin JL, Tu AT. *Appl. Spectrosc.* 1990; **44**: 1408.
29. Kodati VR, Tomasi GE, Turumin JL, Tu AT. *Appl. Spectrosc.* 1991; **45**: 581.
30. Brennan JF. PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.
31. Cornelisse CJ, Kuipersdijkshoorn N, Vanvliet M, Hermans J, Devilee P. *Int. J. Cancer* 1992; **50**: 544.