Theor Chem Acc (2011) 130:1261–1273
DOI 10.1007/s00214-011-1063-0

REGULAR ARTICLE

Overview of the use of theory to understand infrared and Raman spectra and images of biomolecules: colorectal cancer as an example

J. A. A. C. Piva · J. L. R. Silva · L. Raniero · A. A. Martin · H. G. Bohr · K. J. Jalkanen

Received: 30 June 2011 / Accepted: 6 October 2011 / Published online: 1 November 2011

Abstract In this work, we present the state of the art in the use of theory (first principles, molecular dynamics, and statistical methods) for interpreting and understanding the infrared (vibrational) absorption and Raman scattering spectra. It is discussed how they can be used in combination with purely experimental studies to generate infrared and Raman images of biomolecules in biologically relevant solutions, including fluids, cells, and both healthy and diseased tissue. The species and conformers of the individual biomolecules are in many cases not stable structures, species, or conformers in the isolated state or in non-polar non-strongly interacting solvents. Hence, it is better to think of the collective behavior of the system. The collective interaction is not the simple sum of the individual parts. Here, we will show that this is also not true for the infrared and Raman spectra and images and that the models used must take this into account. Hence, the use of statistical methods to interpret and understand the infrared and Raman spectra and images from biological tissues, cells, parts of cells, fluids, and even whole organism should change accordingly. As the species, conformers and structures of biomolecules are very sensitive to their environment and aggregation state, the combined use of infrared and Raman spectroscopy and imaging and theoretical simulations are clearly fields, which can benefit from their joint and mutual development.

Keywords Infrared · Raman · First principles · Molecular mechanics · Statistical methods · Principal component analysis · Linear discriminant analysis · Cluster analysis · Infrared imaging · Raman imaging · Image generation · Colorectal cancer diagnosis

Abbreviations

KS-DFT	Kohn–Sham density functional theory
PCA	Principal component analysis
LDA	Linear discriminant analysis
APT	Atomic polar tensor
VA	Vibrational absorption
IR	Infrared
RS	Raman scattering

Dedicated to Professor Akira Imamura on the occasion of his 77th birthday and published as part of the Imamura Festschrift Issue.

J. A. A. C. Piva · J. L. R. Silva · L. Raniero · A. A. Martin (✉) · K. J. Jalkanen
Laboratory of Biomedical Vibrational Spectroscopy, Institute of Research and Development, Universidade do Vale do Paraíba, UniVaP, Avenida Shishima Hifumi, 2911, Urbanova, São José dos Campos, São Paulo 1244-000, Brazil
e-mail: amartin@univap.br
URL: http://www.ipd.univap.br/levb

J. A. A. C. Piva
e-mail: juliana.piva@yahoo.com.br

J. L. R. Silva
e-mail: jlucasrangel@hotmail.com

L. Raniero
e-mail: lraniero@univap.br

H. G. Bohr · K. J. Jalkanen (✉)
Department of Physics, Quantum Protein Center, QuP, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
e-mail: karljalkanen@gmail.com
URL: http://www.researcherid.com/rid/A-2456-2008

H. G. Bohr
e-mail: hbohr@fysik.dtu.dk

K. J. Jalkanen
Division of Functional Genome Analysis, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany
e-mail: k.jalkanen@dkfz.de

1 Introduction

The first IR (infrared) spectra were measured using dispersive instruments, glow bar sources, and MCT (*Mercury Cadmium Telluride*) detectors [1]. On the other hand, early Raman measurements used conventional light sources, and the technique was not really widely used until the development of lasers. Due to the wavelength dependence of the Raman scattering, one would in many cases like to use the shortest wavelength possible. Early Raman instruments were dispersive and used visible lasers, for example, the Ar ion laser at 488 nm and the Nd:YAG laser at 532 nm. For many samples, these sources are fine, but for many biological and organic samples, there is a large fluorescence when one uses these lasers sources, so one can either go to even shorter wavelengths, into the UV or vacuum UV, or go to higher wavelength, for example, to use the 785 and 1,064 nm laser sources. In addition to the use of sources that do not give fluorescence, Fourier transform (FT) technology has been used.

With the development of more sensitive detectors and the reduced cost, one has been able to develop infrared and Raman imaging instruments. Here, one can measure the infrared and Raman spectra for large areas, for example, surfaces or heterogenous samples. As the technology has developed, so has the breadth of the applications. Initially, infrared and Raman instruments were only found in physics and physical and analytical chemistry laboratories. Spectroscopists needed optics, electrical engineering, and also quantum mechanics to be able to use the instruments and understand the first principles theory required to understand the bands, both frequencies and intensities that one measured in the laboratory. Initially, the harmonic approximation was used, both for the frequency and the infrared absorption intensities. But when the experimental spectra showed more bands than predicted within the harmonic approximation, terms like non-harmonic and Fermi resonance effects were used. Very few experimental vibrational spectroscopists understood the mathematics and theory for these so-called anharmonic effects. This was in large part due to the complexity of the mathematics and physics and the way the theory developed. In the early days of vibrational spectroscopy, one did not have computers and computer programs as developed as one has now. Also, the computer groups gave numbers that were less reliable than those one could "derive" from the experimental data. This is similar to what can now be done with many complex experimental data, where the ad hoc and empirical theories are quite well in describing the so-called underlying effects of the principal components.

Early in the developments of these spectroscopic methods, researchers collected a wealth of information. But due to the complex nature of the mathematics, it has been very hard to extract the information from the spectra directly. This is due to the fact the experimental observables/measurables are not the simple quantities of interest initially to the experimental scientists and chemist, i.e., bond lengths, valence angles, dihedral angles, specific atomic intermolecular distances, binding energies and even relative energies between the large number of possible chemical species with the same chemical formula, for example, C_2H_6O: ethanol and dimethylether. In the case of both infrared and Raman spectroscopy, one can distinguish between these two chemical entities (if one knows one has either one of them, that is, one can make the binary decision: 0 ethanol or 1 dimethylether) by measuring either the "vibrational absorption (VA) or Raman scattering" spectra for the "alcohol" or "ether" functional groups. Hence, there is large body of literature in both the VA and Raman literature on frequencies for various functional groups. Indeed in many papers on infrared/VA and Raman spectroscopy of cells and tissues one sees these tables reproduced to document and "assign" the bands "identified" by principle component analysis as the determiners (also called biomarkers) of various diseased states: cancer, malignant or healthy cells. But this is a very simplistic and not very useful interpretation and understanding as the functional groups are in many cases very generic and do not relate specifically to the biological function or dysfunction which is more important.

To get the level of understanding that is required for the field of molecular medicine to fully develop and interpret, one will need to be able to be a "bit" more specific. For example, to which specific functional group in which specific molecule or molecular complex are the bands in the spectra that have been identified by statistical methods due (to). And then, why? Is it due to a genetic mutation or is "simply" due to the exposure of the organism and/or molecule to (1) damaging radiation, or (2) a denaturing chemical (so-called carcinogen) which may then induce changes that though not at the DNA level, do affect biological processes, including DNA replication and gene transcription and translation. Finally, in many cases, changes may be induced by an invading organism, bacteria, the symptoms of which are similar to other "natural" occurring diseases of sometimes unknown origin. For example, Lyme disease has symptoms that are confused with arthritis and sometimes are failed to be diagnosed. In the meantime, the patient is not given the right treatment. The list can be very long.

As an example of diseases, colorectal cancer is a major public health problem, being the third most common cancer and the fourth leading cause of cancer deaths worldwide. Based on demographic trends of annual incidence, it is likely to increase approximately 80% (2.2 million new cases) over the next two decades, occurring especially in

less developed regions (62%) [2]. According to estimates from the American Cancer Society, 101,700 new cases of colon cancer and 39,510 cases of rectal cancer are expected to occur in 2011 [3, 4]. Thus, there is consensus that the control of colorectal cancer is crucial and of global significance, being based on a balance between prevention, diagnosis, and treatment [2].

Colorectal cancer is usually asymptomatic and is often diagnosed late, being detected after the occurrence of symptoms. Thus, a preliminary screening, which involves removal of polyps or tumors, is necessary in order to identify it early and thus significantly increase the chance of cure and hence survival [3, 4]. Three variables of colorectal lesions are known: non-neoplastic polyps, adenomatous polyps, and cancers or neoplastic polyps themselves. Polyp is defined as any mass that is projected onto the surface of normal mucosa. The non-neoplastic polyps are not generally considered precursors of cancer, but the polyps have great clinical importance, since they have a high probability of becoming malignant.

The procedures conventionally used for early detection of colorectal cancer are the following: examination of fecal occult blood test (FOBT), sigmoidoscopy, colonoscopy, double-contrast barium enema (DCBE), digital rectal examination (DRE, digital rectal) and biopsy surgery for histopathology, which is considered the gold standard for the diagnosis and classification of disease [5]. This analysis is an assessment of biopsy material using histological staining techniques that are traditional. However, the technique of traditional histology remains subjective, where significant problems that include missed lesions and unsatisfactory levels of inter- and intra-rater/observer agreement/reproducibility, are often found [6]. Thus, there is a need to develop diagnostic technologies that are simple, objective, and sensitive [7].

The optical techniques have been studied extensively as a proposal for the diagnosis of cancer, because instead of using an approach based on morphological changes, as currently occurs in histopathological studies [7], the analysis is automated and relies on the detection of biochemical changes that occur in tumor tissues [8]. One of the optical spectroscopy techniques that can effectively provide information concerning the structure and chemical composition of biological materials at the molecular level is Fourier transform infrared spectroscopy (FTIR).

Analysis by FTIR imaging has several advantages over conventional histology [6]. Among them, one is that it does not require the application or the use of dyes or other chemicals for obtaining the image, since the image is generated directly from the measured vibrational spectra of the unstained tissue samples. Here, the sample preparation procedures and possible changes in cellular constituents such as proteins, nucleic acids, and lipids are minimized. The biggest advantage of this technique is the high sensitivity combined with molecular spatial resolution of a few micrometers. One can analyze samples without pre-treatment, enabling identification [9], and the precise biochemical makeup/composition of both healthy areas and tumors [10]. Studies have demonstrated the potential of this technique particularly to differentiate normal from diseased tissues, completely removing the subjectivity and establishing itself as an invariant and reproducible technique [6]. Thus, combining the FTIR spectroscopy/imaging and histopathology is a way to strengthen the capabilities of both techniques. The advantages of an FTIR (and/or Raman) histopathological imaging method(s) would be an objective, easily applied, and observer independent method based on knowledge based image generation routines, similar to those which now exist for generating magnetic resonance imaging (MRI) and x-ray images [6].

Given the need to apply the technique of FTIR imaging to large data sets, it is important to obtain criteria for differentiation that are reasonable and flexible [11]. This can be achieved by the use of multivariate supervised classification strategies, for example, multilayer perceptron artifical neural networks (MLPANNs). These are the techniques of choice for developing robust and effective classifiers of infrared data which depends on and reflects the biochemical structure and composition of the tissue. These techniques can be efficiently monitored and optimized by pre-selecting the appropriate features from the spectral data. In their studies, Lasch et al. [11] applied the techniques of artificial neural networks as supervised techniques for obtaining FTIR images and demonstrated the applicability of this method for generating FTIR images from the histological data and the arrays of FTIR spectra. The FTIR microspectrometer was used to measure and store sets of hyperspectral data from human colorectal adenocarcinomas and build a database of spatially resolved spectral data. This database was composed of spectra data from 28 patients and 12 samples of different histological structures. The spectral information contained in the database was used to train and validate models of MLP-ANN. These classification models were used for data analysis and to produce color images of tissue from full FTIR spectral maps. An important aspect of this study was to demonstrate the sensitivity and specificity can be optimized in particular. The definition of topology of artificial neural network (ANN) was crucial to achieve a high degree of correspondence between the gold standard of histopathology and infrared spectroscopy. In particular, a hierarchical scheme of classification ANN proved to be superior to the classification of tissue spectra, as was concluded from the analysis of data, although the unsupervised clustering methods, specifically hierarchical cluster analysis (HCA), were useful in the early stages of model generation. They concluded that better results of classification can be

achieved if the class definitions of ANN are performed considering the classification information provided by cluster analysis [11–13].

Krafft et al. [12] made a comparative study of Raman and FTIR imaging on colon tissue. The goal of the combined application of FTIR and Raman imaging was to compare, complement, and to confirm the results. Four main groups of tissues were analyzed: muscle, connective, epithelial, and nerve cells. The Raman and FTIR microscopic images covered mucus, mucosa, submucosa and longitudinal and circular muscle layers. Raman images of three nuclei belonging to the mesenteric plexus were also analyzed. The results were discussed with respect to lateral resolution, spectral resolution, acquisition time and sensitivity of both modalities. The separation of Raman and FTIR microscopic images in clusters were very coincidental, despite fundamental differences in both modalities. The authors claim that this coincidence may be due to similar spectral changes that were found in some spectra. However, the divergence that occurred at the subcellular level and in the transitions between tissues was due to the lower lateral resolution inherent in infrared spectroscopy. The connective tissue, smooth muscle, epithelial tissue, glands, and mucous glands were distinguished by both infrared and Raman spectra. From these results, the authors concluded that the main advantages of FTIR imaging were lower acquisition times and better spectral quality, whereas Raman gives higher lateral and spectral resolutions.

Given the advantages of optical spectroscopy and the possible role that this technique will play in the near future in the medical clinic, this study aimed to correlate the different morphological structures with the corresponding biochemical images of normal colorectal tissue, adenomatous and adenocarcinoma obtained by infrared imaging.

2 Methods

2.1 Sample preparation

In this study, 12 samples of human colorectal tissue of patients undergoing colonoscopy or surgical resection in Gastrocentro and Surgical Center, Hospital of the University of Campinas, SP, were used. Of the 12 samples, 4 were normal, 1 tubular adenoma, 3 hyperplastic adenomas, and 4 moderately differentiated adenocarcinomas. This study followed the guidelines and rules for research involving human subjects according to Resolution 196/96 of the National Health Council and was approved by the Ethics Committee of the School of Medical Sciences, State University of Campinas Protocol H083/CEP/2009.

The sectioning of colorectal tissue samples was performed in Cryostat (Leica CM 1100) at −23 °C. Tissue Freezing Medium (Leica Instruments GmbH, Nussloch, Germany) was applied to the samples in order to fix the same on the cutting table. The samples were then placed, and cuts of 12 μm were performed.

Several serial sections were obtained for each sample. The first section of the pair, with 12 micrometers, was positioned over a window of 5 mm (mm) thickness of Calcium Fluoride (CaF_2) for FTIR spectroscopic analysis and the subsequent section, of the corresponding pair, was placed on slides for conventional histological characterization by hematoxylin and eosin (HE) staining. The slides were stained using the standard protocol by HE staining and then analyzed with the aim of identifying and classifying the structures present in tissue samples from normal, adenomatous and adenocarcinoma.

2.2 FTIR data collection

The biochemical images were obtained in transmittance mode by using the FTIR imaging microscope (Spotlight 400—Perkin-Elmer) equipped with MCT detector (Mercury Cadmium Telluride), operating at liquid nitrogen temperature and coupled to a FTIR spectrometer (Spectrum 400—Perkin-Elmer). The images were obtained in the range of 4,000–900 cm^{-1} with 32 scans per pixel (6.25 μm^2) and a resolution of 4 cm^{-1}.

After the spectra for the tissues samples were measured and stored, the spectral data files were imported into the program Cytospec® for spectral analysis and generation of the infrared images.

2.3 Data processing

The average spectra of each region were obtained using the software Cytospec® (version 1.4.02). A spectral quality test was performed to remove the spectra recorded in areas where there was no tissue, or which had a low signal to noise ratio. All spectra were submitted to a thickness test, and the maximum/minimum intensity in the spectral range from 1,700 to 1600 cm^{-1} was also used as the criterion. We excluded regions where the maximum absorption was greater than 1 (indicating a very thick sample), and where the minimum absorption was less than 0.2 (indicating a very thin sample).

An additional important problem/item which needs to be controlled and/or taken into account is the effect due to the absorption of atmospheric water (H_2O) and carbon dioxide (CO_2). One can purge the chamber with either dry nitrogen or dry air, or be very careful in using the regions where both water vapor and carbon dioxide absorb strongly in the infrared: 3,950 to 3,350 cm^{-1} and 1,900 to 1,300 cm^{-1} (H_2O) and 2,400 and 2,300 cm^{-1} (CO_2). One normally can also run a background spectrum under the same conditions

as one measures the spectrum of the sample, and then adjust for atmospheric water and carbon dioxide. The spectra which underwent the thickness tests and adjustments for water vapor and carbon dioxide subsequently underwent two further procedures: 9 point Savitsky Golay smoothing and subsequent first derivative calculation followed by vector normalization in the region from 4000 to 900 cm^{-1}. The first procedure produces a better identification minimizing the variability of the baseline, while the latter reduces the influence of intensity variations caused by differences in cell density (which indicates that the cellular constituents are more or less compact) and tissue thickness, therefore, to exclude differences in thickness of the sample.

To perform the calculation of distance matrix, the data were statistically analyzed using cluster analysis where two spectral regions were selected: 3,000–2,800 and 1,450–950 cm^{-1}. This calculation was applied in order to graphically view the proximity between the biochemical samples. The application of Pearson's correlation coefficient was then performed as a first step for statistical analysis.

After applying the Pearson's correlation coefficient, the scaling first range and Ward's algorithm were applied. This algorithm minimizes the heterogeneity between the elements of each group, thus building more homogeneous groups, so they can minimize the variance within these groups.

After this processing, the data were subjected to hierarchical cluster analysis (HCA) in the spectral range of 950–1,750 cm^{-1} for the classification of groups of spectra that are specific to each type of tissue. This is possible through direct correlation of spectral images, constructed from HCA analysis, with the morphological analysis of histological slides. The spectra were divided into classes that reproduce the histology of the tissue using this method.

Spectral data were analyzed by supervised test (Artificial Neural Networks-ANN) through routine contained called CytoSpec NeuroDeveloper 2.5 (Synthon GmbH, Heidelberg, Germany). While CytoSpec is a software package designed specifically for the generation of infrared images from the mapping of large amounts of infrared data, the software combines NeuroDeveloper modules for selection of spectral characteristics as the development model ANN classification model [11]. Based on neural networks, the interface can be used to re-mount images from an original data set of Cytospec. After this step of pre-processing of data, images were correlated with images of histological slides.

2.4 Neural network techniques

The special ability of the ANN methods [14] is to correlate and classify input data with output data through a carefully arranged training that can either be supervised or unsupervised [15]. The ANN used for the present type of problem are multi-layered feed forward perceptron neural networks [16]. In the following, the most common architecture of these networks is reviewed.

The basic elements of the ANN, the neurons, are processing units that produce output from a characteristic, non-linear function (often a sigmoidal function) of a weighted sum of input data. The ANN network consists of such processing units that can communicate with each other through "synaptic" interconnections between the neurons. The neuron elements are arranged in layers that are connected vertically through these synaptic wires. The first layer is receiving the input data, while the last processing layer is producing output. In between are so-called hidden layers of processing units with no direct communication to the outside.

The training process consists of presenting a set of selected, non-homologous input data while adjusting the variable interconnecting synaptic weights such that the output neurons are producing the desired output values. The network will through a set of training sessions, or training cycles in which the training data are presented to the network, gradually acquire a global information processing where input data leads, through back propagation error-correcting algorithms, to the output with minimal errors compared to known data and corresponding input data. After minimizing the error of the network output compared to known output values, the network will eventually be able to generate new output from new input. In our case, new infrared images are to be generated or chosen from combinations of infrared spectra as input. Before the construction of the training set, the data to be used were, as explained before, subjected to hierarchical cluster analysis (HCA) in a certain spectral range. This is done for making a more sharp distinction in the classification of spectra that are corresponding to specific types of tissue that are homogeneous within one type.

Next, we give some formulas for the network architecture and processing. If we denote a set of inputs by $\{x_i\}$ and the corresponding set of outputs is denoted by $\{y_i\}$, the processing of each neuron i in the network can be described as

$$y_i = f\left(\sum_j W_{ij}x_j + \eta_i\right),$$

where W_{ij} are the adjustable weights of the synaptic connections leading to the neuron i from the neuron j of the proceeding layer, η_i are the thresholds, and f is the non-linear function for the neuron i. In the case of hidden neuron layers the intermediate outputs y_i are propagated further to other layers through the above formula to eventually become real outputs z_i. The training process consist, as explained earlier, of adjusting the weights W_{ij} and the thresholds η_i so as to obtain final outputs z_i with

minimal errors from given inputs x_i by a gradient decent procedure [15]. The cost function, CF, is simply the squared sum of errors formed from the difference between the correct target values t_i and the actual values from the output neurons z_i such that

$$\mathrm{CF} = 1/2 \sum_{a,i} \left(t_i^a - z_i^a\right)^2.$$

It is important to do an assessment of the statistical errors or the precision of the classification of the data used. In the input data of absorption spectra, the precision is roughly 10% corresponding to 1 over the number of peak assignments, while in the output data of images, the accuracy is given as 1 divided by the number of classes of clearly different images which is again estimated to be around 10%. In Figs 1, 2, and 3 the processing of FTIR graphical data is shown in Figs. 1c, 2c, and 3c which should be input to the ANN, and shown in Figs. 1d, 2d, and 3d as output from the network, ANN. The goal of the network processing is seen here as a generalization of graphical features. However, ANNs also have the ability to distinguish useless noise from useful sparse data. The usefulness is determined by what graphics (image) produces the best classification during the training.

2.5 Theoretical

To calculate the infrared or Raman spectra of a molecule requires one first to determine the so-called optimized geometry of the isolated molecule. Depending on the environment and/or medium for which one is measuring one may or may not have to take the environment into account. The environment can normally be neglected to first order if the infrared and/or Raman measurements are made for a molecule at low concentration in a non-polar solvent, in an inert gas matrix (N_2 or Ar) at low temperature, or in the gas phase. For non-hydrogen bonding solvents and other solvent that do not strongly interact with the solute molecule, one can use one of many so-called continuum solvent models, the simplest of which is the Onsager continuum model [17]. Here, one places the molecule in a sphere with a dielectric constant. Improvements to this model include the polarized continuum model (PCM), the CPCM, and finally the COSMO model [18–20]. The latter takes the shape of the molecule into account and includes higher order effects between the molecular and the solvent. The Onsager model only includes the interaction of the molecules electric dipole moment with the solvent [17]. For molecules with no electric dipole moment, hence there is no solvent/environmental effect. For a complete review of continuum solvent models, the readers are referred to the three Chemical Review, two by the Tomasi group in Italy [19, 20], and the third by the Cramer group in Minnesota [21]. For strongly interacting and/or hydrogen bonding solvents like water, which can not only interact with the solute, but actually stabilize species which are not stable in non-polar solvents (the zwitterionic form of amino acids), or conformers that are also not stable in non-polar solvents [the P_{II} conformer of the alanine dipeptide, *N*-acetyl L-alanine *N*′-methylamide (NALANMA)], it is imperative that one include minimally the solvent molecules responsible for these fundamental "phase transitions". For the L-alanine zwitterion (LAZ), it has been shown that one should minimally include 4 water molecules [22], and then one needs 20 to fully encapsulate the LAZ [23]. For NALANMA, minimally 4 water molecules are necessary [24], and for the L-histidine zwitterion, 6 water molecules are necessary [25]. For all other amino acids and peptides, a good starting point is to mutate the methyl side chain of either the LAZ or the P_{II} conformer of NALANMA to get the species of interest, and then hydrate/solvate the amino acid residue, similar to the work of Deplazes et al. [25] for L-histidine and of Jalkanen et al. for *N*-acetyl L-histidine *N*′-methylamide (NALHNMA) [26, 27].

3 Results

Figures 1, 2, and 3 show the results of the analysis of three tissue samples of normal colon, adenoma, and adenocarcinoma, respectively. Figure 1a corresponds to the slide stained by HE the sample of normal mucosa. There is clearly a regular structure of the glands (also called glandular crypts) next to each other and occupying most of the volume of the mucosa. These glands are composed of columnar cells and goblet cells, indicating differentiation into two cell types morphologically and functionally distinct. The glands consist of goblet cells with regular nuclei in the shape, size, and number. In this figure, it is also be observed epithelial tissue (surrounding the glands). Figs. 1b and c show the sample in CaF_2 window and image biochemistry FTIR, respectively. Figure 1d refers to the classification obtained by ANN. Figure 1e shows the processed image obtained by HCA for the normal tissue and Fig. 1f shows the average spectrum of each different region discriminated in the image processed by the software Cystospec obtained by HCA, showing the vibrational bands of the chemical bonds within all the biochemical components of cells (proteins, nucleic acids, carbohydrates, and lipids). In the Table 1, we present the legend for the color for (d). The image was colored in six colors, determined by the software Cytospec. Each of these colors represents a structure that was identified and classified according to their proportion of the amount (given in percentage) and is part of a database. Each color of the

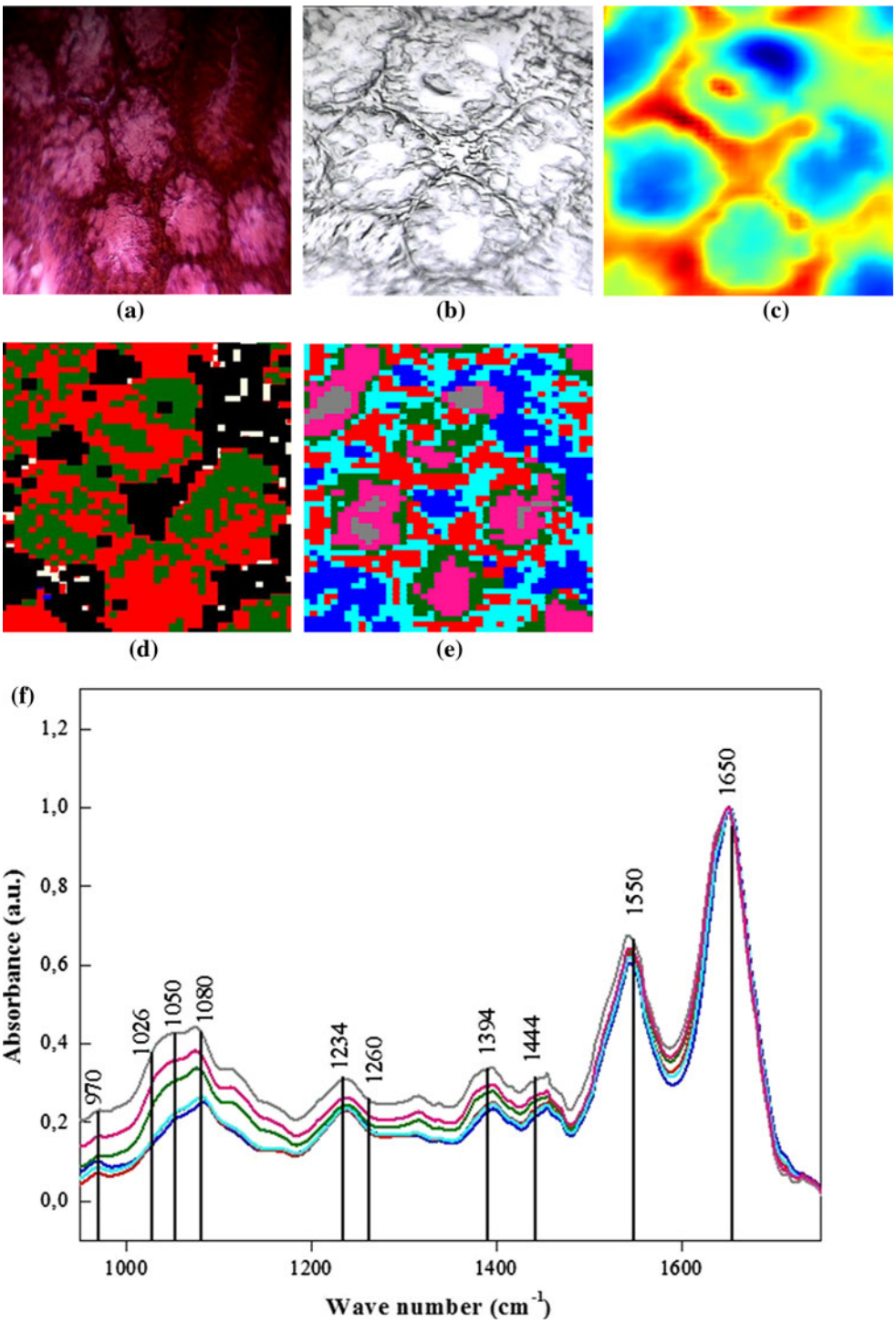

Fig. 1 Comparison of the infrared images for normal colorectal tissue: **a** the Blade of normal tissue stained by HE, **b** sample in CaF2 window, **c** biochemical image obtained by FTIR, **d** processed image obtained by artificial neural network (ANN), **e** processed image obtained by HCA, and **f** the average spectra obtained by HCA

average FTIR in Fig. 1f represents a different structure type as determined by HCA and compared with the histopathological analysis. In Fig. 1f, the light blue and dark green refers to the lamina propria of the mucosa and fibrovascular. The crypts are colored by red and magenta represent the central lumen of the crypts. The outer layers of cells of the crypts appear as gray and dark blue regions.

In Fig. 2, we present (a) Blade adenomatous tissue stained by HE, (b) Sample in CaF2 window, (c) Biochemical FTIR Image (d) Processed image obtained by artificial neural network (ANN). Figure 2e represents the processed image obtained by HCA and Fig. 2f the average spectra obtained by HCA for colorectal adenoma tissues. Table 2 shows the identification and classification of the type of

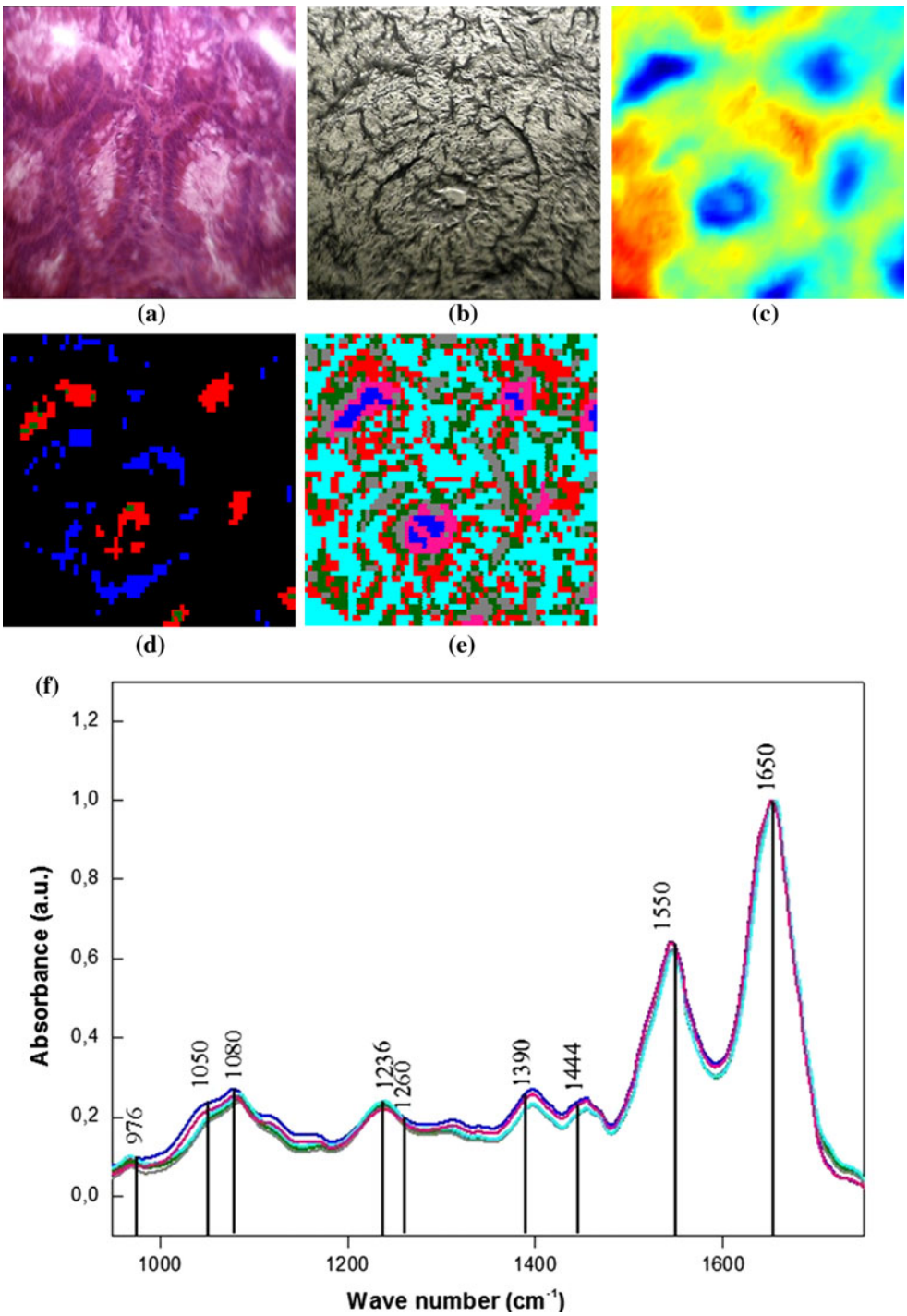

Fig. 2 Comparison of the infrared images for **a** the Blade of adenoma tissue stained by HE, **b** sample in CaF2 window, **c** image obtained by biochemical FTIR, **d** processed image obtained by artificial neutral network (ANN), **e** processed image obtained by HCA for colorectal adenoma tissues, and **f** the average spectra obtained by HCA

structure in the image obtained by ANN processing of the FTIR spectra for colorectal adenoma tissue.

In Fig. 3a, we present the images for Blade adenocarcinoma tissue stained by HE, showing a disorganized structure, atypical glands and with tumor infiltrating the submucosa. There is decrease in intracellular mucin with loss of differentiation into columnar cells. Figure 3b and c show the sample in CaF_2 window and biochemical FTIR image, respectively. In Fig. 3d and Table 3 one sees the results according to the classification obtained by ANN. Figure 3e shows the processed image obtained by HCA for adenocarcinoma tissues, and Fig. 3f shows the average spectra obtained for HCA for colorectal adenocarcinoma tissues.

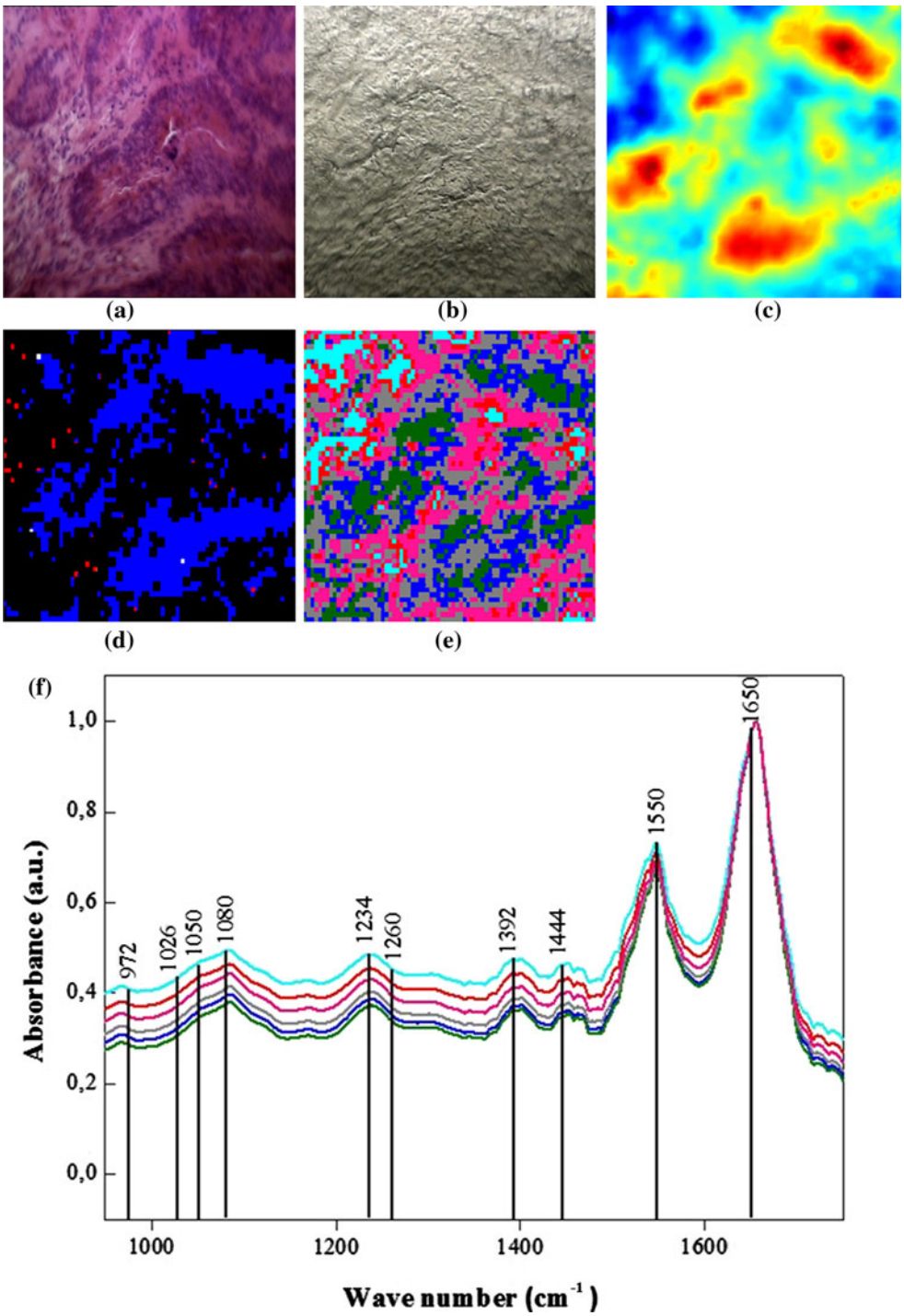

Fig. 3 Comparison of the infrared images for **a** the Blade of adenocarcinoma tissue stained by HE, **b** sample in CaF2 window, **c** image obtained by biochemical FTIR, **d** processed image obtained by artificial neutral network (ANN), **e** processed image obtained by HCA for colorectal adenocarcinoma tissues, and **f** the average spectra obtained by HCA

4 Discussion

4.1 Histological architecture and histopathologic colorectal

We begin the discussion with a brief review of the histopathological characteristics of colon tissue. Subsequently, clustering methods will be introduced, and the images obtained by applying the method of grouping the data set are compared to images of stained tissue. The cell adenocarcinomas are subsets of colorectal adenocarcinomas and malignant colonocytes. These tumors mostly originate from the reabsorption of epithelial cells, located mainly in the mucosal surface and the upper third of the tubular glands, also called crypts. Colorectal mucosal epithelial cells form a layer of columnar epithelium with

Table 1 Identification and classification of the type of structure in the image obtained by ANN processing of the FTIR spectra for normal colorectal tissue

Color	Type of structure color	Percent
Blue	Adenocarcinoma	0.4
Red	Crypts	45.19
Dark green	Mucin	29.51
Light yellow	Propria	9.79
Salmon	Necrosis	0.04
Black	Unclassified spectra	15.05
Total		99.98

Table 2 Identification and classification of the type of structure in the image obtained by ANN processing of the FTIR spectra for colorectal adenoma tissue

Color	Type of structure color	Percent
Blue	Adenocarcinoma	23.99
Red	Crypts	9.1
Dark green	Mucin	0.51
Black	Unclassified spectra	66.4
Total		100

Table 3 Identification and classification of the type of structure in the image obtained by ANN processing of the FTIR spectra for adenocarcinoma

Color	Type of structure color	Percent
Blue	Adenocarcinoma	59.54
Red	Crypts	5.02
Gray	Muscularis mucosae	0.04
Yellow light	Propria	0.13
Salmon	Necrosis	0.06
Blue light	Connective tissue	0.02
Black	Unclassified spectra	35.19
Total		100

well-preserved architecture. Crypts, these cells are arranged in groups with other types of epithelial cells (goblet cells, stem cells or stem cells and various types of endocrine cells functionally differentiated). The propria of mucosae, the crypts and muscularis form the membrane lining the colon [10, 28]. The normal mucosa is shown in Fig. 1a, where you can observe the cross-sectioned crypts with epithelial cells (colonocytes), goblet cells, and propria of mucosa. In the severe dysplastic high grade changes observed in Fig. 2a, the cell polarity is not possible to identify, the nuclei are dramatically expanded and acquire either a rounded or ovaloid shape, and irregular complex tubules are present [29]. The colorectal adenocarcinomas originated from epithelial cells and are able to infiltrate the underlying layers (submucosae, muscularis), colon and rectum [10]. This is observed in Fig. 3a. Furthermore, adenocarcinoma shows the typical morphological signs of malignancy such as atypical histoarchitecture with several layers of cells exhibiting pleomorphism and infiltration of the submucosa. Figures 1a, 2a, and 3a clearly show the complexity of histoarchitecture colorectal tissue.

As noted in Fig. 1d, Table 1 and Fig. 2d, Table 2 and Fig. 3d, Table 3, the processing of images by ANN showed regions of adenocarcinoma in the percentage of 0.4, 23.99 and 59.54 in normal tissue, adenoma and adenocarcinoma, respectively. In terms of statistical classification criteria, the sensitivity of diagnosis of adenocarcinoma is high, because all spectra were classified as adenocarcinoma and the number of false negatives is almost null. Moreover, there are some false positives, since many spectra of other spectral classes were classified as adenocarcinoma, as observed in Fig. 1d, Table 1 and Fig. 2d, Table 2. Thus, the image of Fig. 3d illustrates an example with high sensitivity and low specificity of the diagnosis, because this sample showed a higher percentage of tissues not identified, about 66.4%. However, generally there is a correlation between the relatively high gold standard of histopathology and diagnostic spectra observed in all samples.

The Figs. 1f, 2f, and 3f show the average spectra of typical infrared absorption of a normal human colorectal tissue, adenomatous, and adenocarcinoma, respectively, in the spectral range from 950 to 1,750 cm^{-1}.

The average spectra show bands at $\sim$970 cm^{-1} corresponding to phosphatidylcholine; 1,026 cm^{-1} to glycogen; $\sim$1,050 cm^{-1} glycolipid stretching vibration [ν (COH)] and cholesterol; 1,080 cm^{-1} to phospholipids symmetric stretching vibration [ν_s (PO_2^-)] and glycogen; 1,234 cm^{-1} to phospholipids asymmetric stretching vibration [ν_{as} (PO_2^-)]; $\sim$1,260–1,390 cm^{-1} to protein symmetric stretching vibration [ν_s (COO^-)] and cholesterol; 1,444 $cm^{-1}$to protein deformation vibration [δ (CH_3)] and collagen; 1,550 cm^{-1} to protein (amide II); and 1,650 cm^{-1} to protein (amide I). The spectrum is dominated by two bands 1,650 and 1,550 cm^{-1}, which are amide I and amide II, respectively. The amide I stretching vibration arises from the hydrogen bonded C=O stretch mode and the amide II vibration arises from a combination of the C–N stretching and the N–H bending modes. The intensity differences from normal tissue and cancerous polyp for the amide II modes were not significant in all three cases. Important spectral features were found in DNA and RNA and are associated with protein phosphorylation and nucleic acids ($\sim$970 cm^{-1}), ν_s PO_2^- ($\sim$1,080 cm^{-1}), ν_{as} PO_2^- ($\sim$1,234 cm^{-1}), amide I ($\sim$1,650 cm^{-1}), and amide II ($\sim$1,550 cm^{-1}) [30]. The weakest side chain band of amino acids, peptides and proteins, at 1,444 cm^{-1}, is with the scissoring and bending vibrations of the CH_2

and CH_3 groups. The absorption peaks of 1,080 and 1,234 cm^{-1} are due to vibrations of symmetric and asymmetric stretching of PO_2^-, respectively. The absorption from the normal tissue was higher than that of polyps and cancer in the whole region of the spectrum and in three situations. The bands of 1,026 and 1,050 cm^{-1} infrared spectrum are responsible for the stretching vibration modes of the CH_2OH groups and CO groups coupled with CO bending and C–OH of the carbohydrates [31]. Different spectral patterns were found between normal, polyps, and cancer in the region of 1,100–950 cm^{-1} (ν_S PO_2^- nucleic acids and carbohydrates) with maximum 1,080 and 1052 cm^{-1}, respectively [32].

The intensity of the band 1,050 cm^{-1} gives an estimate of the levels of carbohydrates. Carbohydrate levels between normal and cancer is a greater measure of disease progression. This may be because the absorption of carbohydrates (or your metabolism) is affected tissue with cancer especially in the later stages of the disease. In the early stages, the carbohydrate content calculated from the spectra was smaller than in the polyp and it was reversed in cancer. It was then possible to establish a strong correlation of spectral analysis between normal tissues and cancer. However, the difference between polyps and malignant tumors were not significant [25]. However, a band in 1,468 cm^{-1} concerning protein and carbohydrates, coupled with the asymmetric deformation CH_3 [DA (CH_3)] and symmetric deformation CH_2 [(Ds CH_2)], was seen in the spectrum of adenocarcinoma tissue.

The main spectral differences were found in the typical collagen bands at ~1,336 and ~1,452 cm^{-1}. Interestingly, the spectra obtained from the central part of the crypts have very characteristic infrared bands as well. These bands are different from those of collagen and can be found at 1,080 cm^{-1}. Most of these bands can be attributed to mucin, a glycoprotein rich in cysteine. Mucin is known to be present as a precursor in the secretory granules of goblet cells or mature, after their secretion into the lumen of the crypts. Due to the heterogeneity of the different types of mucin, the allocation of bandwidth is a little different in the literature. To give one example, the position of the peak more prominent mucin was found between 1,035 and 1,050 cm^{-1} [33]. Therefore, the spectral differences between the mucin-rich regions of the crypts (gray and pink spectrum of Figs. 1e and 2e, respectively, and stromal tissue (blue color spectrum of Fig. 3e) and the differences between the spectra from the neoplastic parenchyma (color pink, red, gray, dark green, and dark blue in Figs. 3e and 2e are rather small. The most relevant spectral changes were found in bands of PO_2^- at ~1,080 and ~1,234 cm^{-1}. It was postulated that the intensity of PO_2^- band may be related to the degree of activity of cell division. Given that cancer cells in the division rate is generally higher than in precursor cells, the average spectrum of tissue with cancer tissue must differ from "normal" (benign) due to a lower mitotic index. The experimental results of this study confirm these findings of data from transitional tissue, adenomas, the spectrum dark blue, light blue, red and green blend in Fig. 3f. Moreover, a progressive decrease in signal of mucin is also observed (spectrum light blue). This can be explained by the loss of the initial process arquiterura typical feature of the colon that starts in these structures.

5 Conclusions

The combination of spectroscopic imaging techniques and digital image analysis is a powerful new technique that can be used to re-assemble color images of histological sections. The results presented in this study demonstrated the potential use of FTIR imaging in the detection of morphological and biochemical changes that occur in tissues when they undergo from normal to diseased state. Thus, our results show the ability of this technique for future clinical use in histopathology.

There are two important new developments in this technique. One is to go from single type molecule spectroscopy to whole tissue spectroscopy where many different types of molecules in huge numbers are included and sampled over. The other development is that of using artificial neural networks, previously used for classifying molecular structures, instead to classify overall features of areas in tissue samples.

A few things are definitely to be improved. One is the number of examples in the training set that need to be enlarged. This problem of using too small sample sets can be seen in the uneven distribution or size of the percentage numbers in the three Tables 1, 2 and 3. In future, use of ANN for classifying tissue data, one should make use of the frequency distribution spectra for the analysis of chemical abundances, for example, how much is present in a certain tissue sample of the chemical compound of phosphor lipids or glycogen etc., based from an analysis of the height of peaks in the spectra. In addition, first principle calculations of the electronic absorption and fluorescence spectra in a variety of environments have now appeared [34, 35], including one in this issue of TCA by Pomogaev et al. [36]. In these works using the elongation method developed by Imamura et al. [37], and extensions [38], one is now able to simulate both the ground and excited electronic state structures and properties of large biomolecules, including their infrared, Raman, electronic absorption, and fluorescence spectra. The next great challenge is now to move from solution, to heterogeneous media like the cell, intracellular media, but in/under homeostatic conditions and in/under stress and denaturing condition, which in

many cases lead to diseases: cancer being the one which we have addressed in this work. Note that Professor Imamura started working in this field with semi-empirical methods and now has progressed to using the latest state of the art wave function theory (MP2 and CASPT2) and density functional theory (GGAs, hybrid, meta hybrid and finally LC- and LCgau-KS-DFT methods [39–42]) with already or in works in progress to treat large biomolecules, first in the isolated state, then in aqueous solution. The last big challenge will now to treat these molecules and their interactions and functions under biologically relevant conditions, at relatively high concentrations, at various pH values, ionic strengths, at body temperatures and atmospheric pressure.

Acknowledgements JAAC Piva, JLR Silva, L Raniero, AA Martin and KJ Jalkanen would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the São Paulo Research Foundation (FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the Universidade do Vale do Paraíba (UniVaP) for financial and infrastructure support. KJ Jalkanen and AA Martin would like to thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for FAPESP grant: 2009/16782-2 which allowed Dr. Jalkanen to visit the Laboratory of Biomedical Vibrational Spectroscopy, LEVB, at the Universidade do Vale do Paraíba (UniVaP) for the period from June 2010 to May 2011, during which time a majority of the experimental work was undertaken at the LEVB. KJ Jalkanen and HG Bohr would like to thank the Danish National Research Foundation, DG, for the funding to establish the Quantum Protein Center, QuP, in the Department of Physics at the Technical University of Denmark, DTU. Finally KJ Jalkanen would like to thank the German Cancer Research Center (DKFZ) for his stipendium to work in the Division of Functional Genome Analysis at the DKFZ in Heidelberg, and the Deutscher Akademischer Austausch Dienst (DAAD) for providing affordable Krankenversicherung (health insurance) for him during his Aufenthalt (stay) in Heidelberg, Germany at the DKFZ.

References

1. Spotlight 400 Series User's Guide (2008) PerkinElmer Ltd, Bucks HP9 2FX, United Kingdom
2. Karsa LV, Lignini TA, Patnick J, Lambert R, Sauvaget C (2010) The dimensions of the CRC problem. Best Pract Res Clin Gastroenterol 24:381–396
3. Cancer Facts & Figures 2010 (2010) American Cancer Society, Los Angeles, CA. http://www.cancer.org/Research/CancerFacts Figures/CancerFactsFigures/cancer-facts-figures-2010-rev, Accessed 28 May 2011
4. Colorectal Cancer (2011) American Cancer Society, Los Angeles, CA. http://www.cancer.org/Cancer/ColonandRectumCancer/DetailedGuide/colorectal-cancer-pdf12, Accessed 28 May 2011
5. Winawer SJ (2007) Colorectal cancer screening. Best Pract Res Clin Gastroenterol 21:1031–1048
6. Bird B, Miljkovic M, Romeo MJ, Smith J, Stone N, George MW, Diem M (2008) Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology. BMC Clin Pathol 8:8
7. Fernandez D, Bhargava R, Hewitt SM, Levin IW (2005) Infrared spectroscopic imaging for histopathologic recognition. Nat Biotechnol 23:469–474
8. Andrade PO, Bitar RA, Yassoyama K, Martinho H, Santo AM, Bruno PM, Martin AA (2007) Study of normal colorectal tissue by FT-Raman spectroscopy. Anal Bioanal Chem 387:1643–1648
9. Steiner G, Koch E (2009) Trends in Fourier transform infrared spectroscopic imaging. Anal Bioanal Chem 394:671–678
10. Petibois C, Déléris G (2006) Chemical mapping of tumor progression by FTIR imaging: towards molecular histopathology. Trends Biotechnol 24:455–462
11. Lasch P, Diem M, Hansch W, Naumann D (2006) Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. J Chemom 20:209–220
12. Krafft C, Codrich D, Pelizzo G, Sergo V (2008) Raman and FTIR microscopic imaging of colon tissue: a comparative study. J Biophotonics 1:154–169
13. Hedegaard M, Matthaus C, Hassing S, Krafft C, Diem M, Popp J (2011) Spectral imaging and clustering algorithms for assessment of single cells by Raman microscopic imaging. Theor Chem Acc. doi:10.1007/s00214-011-0957-1
14. Hertz JA, Krogh AS, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Redwood City
15. Rommelhardt DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
16. Bohr H, Bohr J, Brunak S, Cotterill RM, Lautrup B, Nørskov L, Olsen OH, Petersen SB (1988) Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. FEBS Lett 241:223–238
17. Onsager L (1936) Electric moments of molecules in liquids. J Am Chem Soc 58:1486–1493
18. Nemzkin VN, Hadt RG, Belosludov RV, Mizuseki H, Kawazoe Y (2007) Influence of molecular geometry, exchange-correlation functional, and solvent effects in the modeling of vertical excitation energies in phthalocyanines using time-dependent density functional theory (TDDFT) and polarized continuum model TDDFT methods: can modern computational chemistry methods explain experimental controversies? J Phys Chem A 111:12901–12913
19. Tomasi J, Pérsico M (1994) Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. Chem Rev 94:2027–2094
20. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chem Rev 105:2999–3093
21. Cramer C, Truhlar DG (1995) Continuum solvation models: classical and quantum mechanical implementations. In: Lipkowitz KB, Boyd DB (eds) Rev Comput Chem 6:1–72. doi: 10.1002/9780470125830.ch1
22. Tajkhorshid E, Jalkanen KJ, Suhai S (1998) Structure and vibrational spectra of the zwitterion L-alanine in the presence of explicit water molecules: a density functional analysis. J Phys Chem B 102:5899–5913
23. Jalkanen KJ, Degtyarenko IM, Nieminen RM, Cao X, Nafie LA, Zhu F, Barron LD (2008) Role of hydration in determining the structure and vibrational spectra of L-alanine and N-acetyl L-alanine N′-methylamide in aqueous solution: a combined theoretical and experimental approach. Theor Chem Acc 119:191–210
24. Jalkanen KJ, Suhai S (1996) N-acetyl-L-alanine N'-methylamide: a density functional analysis of the vibrational absorption and vibrational circular dichroism spectra. Chem Phys 208:81–116
25. Deplazes E, van Bronswijk W, Zhu F, Barron LD, Ma S, Nafie LA, Jalkanen KJ (2008) A combined theoretical and experimental study of the structure and vibrational absorption, vibrational circular dichroism, Raman and Raman optical activity spectra of the L-histidine zwitterion. Theor Chem Acc 119:155–176
26. Jalkanen KJ, Jürgensen VW, Claussen A, Jensen GM, Rahim A, Wade RC, Nardi F, Jung C, Nieminen RM, Degtyarenko IM, Herrmann F, Knapp-Mohammady M, Niehaus T, Frimand K, Suhai S (2006) The use of vibrational spectra to study protein and DNA structures, hydration, and binding of biomolecules: a

combined theoretical and experimental approach. Int J Quantum Chem 106:1160–1198
27. Jalkanen KJ, Knapp-Mohammady M, Jensen KBS, Ma S, Nafie LA, Zhu F, Barron LD, Aoki Y, Mannfors B, Bohr J, Nieminen RM, Rodarte A, Pomogaev V, Alves RS, Carvalho CS, JACC Piva, Martin AA (2011). (in preparation)
28. Talbot I, Price A, Tellez M (2007) Biopsy pathology in colorectal disease, 2nd edn. Oxford University Press, Oxford
29. Lasch P, Pacifico A, Diem M (2002) Spatially resolved IR microspectroscopy of single cells. Biopolym Biospectrosc 67:335–338
30. Walsh MJ, Hammiche A, Fellous TG, Nicholson JM, Cotte M, Susini J, Fullwood NJ, Martin-Hirsch PL, Alison MR, Martin FL (2009) Tracking the cell hierarchy in the human intestine using biochemical signatures derives by mid-infrared microespectroscopy. Stem Cell Res 3:15–27
31. Ramesh J, Salman A, Mordechai S, Argov S, Goldstein J, Sinelnikov I, Walfisch S, Guterman H (2001) FTIR microscopic studies on normal, polyp, and malignant human colonic tissues. Subsurfac Sens Tech Appl 2:99–117
32. Conti C, Ferraris P, Giorgini E, Rubini C, Sabbatini S, Tosi G, Anastassopoulou J, Arapantoni P, Boukaki E, Konstadoudakis S, Theophanidesc T, Valavanis C (2008) FT-IR microimaging spectroscopy: a comparison between healthy and neoplastic human colon tissues. J Mol Struct 881:46–51
33. Lasch P, Haensch W, Naumann D, Diem M (2004) Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim Biophys Acta 1688:176–186
34. Pomogaev V, Gu FL, Pomogaeva A, Aoki Y (2009) Elongation method for calculating excited states of aromatic molecules embedded in polymers. Int J Quantum Chem 109:1328–1340
35. Cusasi T, Granucci G, Persico M (2011) Photodynamics and time-resolved fluorescence of azobenzene in solution: a mixed quantum-classical simulation. J Am Chem Soc 133:5109–5123
36. Pomogaev V, Pomogaeva A, Avramov P, Jalkanen KJ, Kachin S (2011) Thermo-dynamical contours of electronic-vibrational spectra simulated using the statistical quantum mechanical methods. Theor Chem Acc. doi:10.1007/s00214-011-0936-6
37. Imamura A, Aoki Y, Maekawa K (1990) A theoretical synthesis of polymers by using uniform localization of molecular orbitals: proposal of an elongation method. J Chem Phys 95:5419–5431
38. Makowski M, Korchowiec J, Gu FL, Aoki Y (2006) Efficiency and accuracy of the elongation method as applied to electronic structures of large systems. J Comput Chem 27:1603–1619
39. Tawada Y, Tsuneda T, Yanagisawa S, Yanai T, Hirao K (2004) A long-range-corrected time-dependent density functional theory. J Chem Phys 120:8425–8433
40. Yanai T, Tew DP, Handy NC (2004) A new hybrid exchange-correlation functional using the coulomb-attenuating method (CAM-B3LYP). Chem Phys Lett 393:51–57
41. Peach MJG, Helgaker T, Sałek P, Keal TW, Lutnæ OB, Tozer DJ, Handy NC (2004) Assessment of a coulomb-attenuated exchange-correlation energy functional. Phys Chem Chem Phys 8:558–562
42. Song J-W, Tsuneda T, Sato T, Hirao K (2011) An examination of density functional theories on isomerization energy calculations of organic molecules. Theor Chem Acc. doi:10.1007/s00214-011-0997-6