PAPER

Raman spectroscopy combined with principal component analysis and k nearest neighbour analysis for non-invasive detection of colon cancer

To cite this article: Xiaozhou Li *et al* 2016 *Laser Phys.* **26** 035702

IOP Publishing | Astro Ltd Laser Physics
Laser Phys. **26** (2016) 035702 (9pp) doi:10.1088/1054-660X/26/3/035702

Raman spectroscopy combined with principal component analysis and k nearest neighbour analysis for non-invasive detection of colon cancer

Xiaozhou Li[1,2], Tianyue Yang[1], Siqi Li[3], Deli Wang[1], Youtao Song[2] and Su Zhang[4]

[1] School of Science, Shenyang Ligong University, 110159, Shenyang, People's Republic of China
[2] College of Environmental Sciences, Liaoning University, 110036, Shenyang, People's Republic of China
[3] College of Medicine, Northeast Ohio Medical University, 44272, Rootstown, USA
[4] Section of Physiology and Biochemistry, Shenyang Sport University, 110102, Shenyang, People's Republic of China

E-mail: biophy@163.com (Xiaozhou Li) and ysong@lnu.edu.cn (Youtao Song)

Received 5 November 2015, revised 13 January 2016
Accepted for publication 15 January 2016
Published 15 February 2016

Abstract

This paper attempts to investigate the feasibility of using Raman spectroscopy for the diagnosis of colon cancer. Serum taken from 75 healthy volunteers, 65 colon cancer patients and 60 post-operation colon cancer patients was measured in this experiment. In the Raman spectra of all three groups, the Raman peaks at 750, 1083, 1165, 1321, 1629 and 1779 cm^{-1} assigned to nucleic acids, amino acids and chromophores were consistently observed. All of these six Raman peaks were observed to have statistically significant differences between groups. For quantitative analysis, the multivariate statistical techniques of principal component analysis (PCA) and k nearest neighbour analysis (KNN) were utilized to develop diagnostic algorithms for classification. In PCA, several peaks in the principal component (PC) loadings spectra were identified as the major contributors to the PC scores. Some of the peaks in the PC loadings spectra were also reported as characteristic peaks for colon tissues, which implies correlation between peaks in PC loadings spectra and those in the original Raman spectra. KNN was also performed on the obtained PCs, and a diagnostic accuracy of 91.0% and a specificity of 92.6% were achieved.

Keywords: Raman spectroscopy, serum, colon cancer, operation, diagnosis, principal component analysis, k nearest neighbour analysis

(Some figures may appear in colour only in the online journal)

Introduction

Colorectal cancer is the third and the second most common cancer in men and women respectively, and accounts for around 10% of global cancer incidence [1]. It has the highest incidence at ages 40 to 50. The 5-year survival rate for colon cancer at stages I-IIB (TNM Classification of Malignant Tumours) is 74 − 58%, and drops to 46 − 5.7% at stages III and IV. Stage IIIA was an exception—it has a 5-year survival rate as high as 73% [2]. Early diagnosis is the most effective way of thwarting cancers. However, the symptoms of early-stage cancers are rather unobvious and cannot be detected even with advanced equipment [3]. Some of the differences between normal cells and rudimentary cancerous cells are too slight to be detected by prevalently used morphological methods.

Conventional methods for colon cancer detection include colonoscopy, fecal occult blood test (FOBT) and sigmoidoscopy. Among those, colonoscopy is the gold standard for the assessment of colon cancer. But it is invasive, subjective and time-intensive [4, 5]. FOBT, which is recommended in many countries as the screening method [6], is much less sensitive than colonoscopy and may have a higher rate of false positives [7].

Serum contains a large amount of biomarkers. It is an informative testing object for cancer detection. Testing of serum is non-invasive and relatively economical compared to tissue imaging techniques and is thus a good candidate for cancer screening.

Contrary to conventional detection methods, optical spectroscopy on serum provides us a possibility to diagnose diseases non-invasively. Raman spectroscopy is based on inelastic vibrational scattering. Because of this mechanism, it can detect the secondary construction of molecules and is called fingerprint spectroscopy. The Raman techniques can detect the differences in cells and tissues on a molecular level [8]. Raman spectroscopy has been recognized as a potential method for cancer diagnosis, and has achieved promising results [9, 10]. Since water only has broad bands at the higher wavenumber region of 2500–4000 cm^{-1}[11], Raman is suitable for biofluid detection. In addition, several types of serum biomarkers for colon cancer have been verified, including tumor-associated antigens, cytokines, cell proliferation proteins and other DNA/RNA biomarkers [6]. Therefore, the combination of Raman technique and serum detection has foreseeable clinical applications.

Several types of Raman techniques have been used for the screening of colon cancers. Many studies on tissue detection using Raman have been conducted. Auto-fluorescence generated by endogenous porphyrins has been found to have significant differences between normal and colon adenocarcinoma samples [12]. Also, fourier transform infrared microscopy (FTIR-MSP) performed on colonic tissues can be used to detect the relapse of crypts [13]. However, these tissue detection methods are invasive and unsuitable for large-scale screening. Human serum has been measured using surface-enhanced Raman spectroscopy (SERS) for the diagnosis of colorectal cancer, and has achieved a sensitivity of 97% and specificity of 100% [14]. Although SERS can achieve a high sensitivity, the preparation of Raman enhancement substrates is hard to control and sometimes very complicated [15]. In this paper, we attempted to apply resonance Raman spectroscopy on human serum to observe Raman resonance peaks which were hard to detect with standard Raman [16].

Because only subtle variations exist in the Raman spectra of serum between different groups, it is essential to apply certain chemometric methods. Principal component analysis (PCA) and k nearest neighbours (KNN) were selected for the analysis of the obtained Raman spectra. Both of these two methods have been used successfully for the analysis of spectroscopy data [17–20]. PCA is a commonly used multivariate method for the dimension reduction of spectroscopy data. It can do this very effectively by projecting the original spectra into lower-dimension principal component subspaces while simultaneously retaining the key spectral information. KNN is a classical pattern recognition algorithm in machine learning. It classifies a point depending on the classification of the k nearest neighbouring points. KNN was used in this study to calculate the classification accuracy of principal components (PCs) obtained from the PCA model.

The study examines serum taken from pre-operative colon cancer patients (65 samples) and post-operative colon cancer patients (60 samples, a subset of pre-operative colon cancer patients) using Raman spectroscopy, and compares them with the controls (75 samples). Raman peaks in each group were assigned and compared; biomedical explanations of the peaks were discussed. The data reduction method of PCA and subsequent classification method of KNN were used consecutively on the Raman spectra to see the prediction accuracies.

Materials and methods

Sampling

A total of 65 pre-operative colon cancer patients, 60 post-operative colon cancer patients (a subset of pre-operative colon cancer patients) and 75 healthy volunteers participated in our experiment. All serum samples were obtained from the Tumor Hospital of Liaoning Province. For colon cancer patients, serum was collected both before and after the operation. Blood was phlebotomized before breakfast in the morning to avoid the interference of food. Then, the obtained venous blood was centrifuged at a speed of 3000 rot/min for 10 min. Afterwards, the serum layer was collected and stored in capped tubes at −4 °C. Samples were measured by Raman spectroscopy within three weeks.

Raman equipment

Figure 1 shows the main pieces of our Raman system: an Ar-ion laser (A240, former Nanjing State 772 Factory (presently Nanjing Sanle Electronic Information Industry Group Co., Ltd.)), a double spectrometer (HRD-1, Jobin Yvon Co., Ltd.), a PMT (R928, Hamamatsu Photonics Co., Ltd.) and a lock-in amplifier (SR830, Stanford Research Systems, Inc.). The sample chip (glass optical length: 0.735 mm) was excited directly by the laser beam (diameter: 2 mm), and then the Raman radiation of the sample was collected into the spectrometer. The focus length for the concave lens was −1000 mm and was 300 mm for the convex lens. Then, the scattering light was collected with a double spectrometer equipped with a PMT. Afterwards, the Raman signals were amplified with a lock-in amplifier. A personal computer was used to control the entire system, and to collect the output Raman spectra. The spectral range was 600–3000 cm^{-1}. The frequency of the chopper was set to 700 Hz for optimal signal-to-noise ratio, and the Ar-ion laser was operated at a wavelength of 514.5 nm at 100 mW.

Data preprocessing

All spectra data analysis including both the pre-processing and statistical analyses were done in the statistical environment R version 3.0.3 (www.R-project.org).

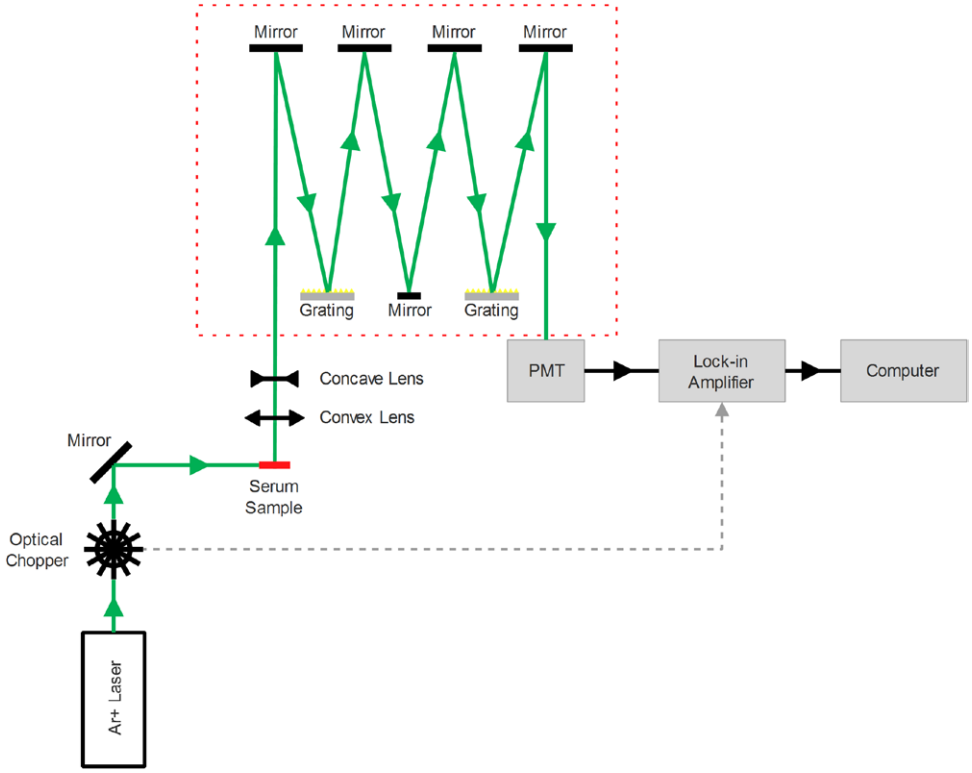

Figure 1. Experimental setup of the Raman spectroscopy system.

Spectroscopy data is inclined to be contaminated by factors such as fluorescence background, laser power fluctuation and sample concentration variances. Therefore, preprocessing was necessary to eliminate the interference. Smoothing, baseline correction and normalization were performed sequentially on the raw Raman spectra. This preprocessing can to a certain degree eliminate noise in the raw Raman spectra which may interfere with the following multi-variant analysis.

Smoothing was carried out using the smooth.Pspline function in the pspline package. This function will fit a polynomial smoothing spline of the original Raman spectra. An order of 9 was set as the order of the spline, where the order means the order of the derivative that is penalized. Baseline correction was completed by spc.fit.poly.below in the hyperSpec package. This function finds the least-square-fits of a polynomial with automatic support point determination. The polynomial number was set as 9. At this stage, the fluorescence was considered as the background radiation and was subtracted. Area normalization was performed through the sweep function in the hyperSpec package.

Statistical analysis

PCA is a non-parametric method that can compress large raw spectral data into several PCs. These PCs will contain most of the variance of the original spectra. In brief, PCA generates PCs through orthogonal transformation in the direction containing the most variation, and tries to rebuild the original Raman spectroscopy by PCs. The mathematical expression of such transposition can be written as

$$\mathrm{X}(\lambda) = u_1 p_1(\lambda) + u_2 p_2(\lambda) + \ldots + u_\mathrm{A} p_n(\lambda) = \sum_{i=1}^{A} u_i p_n$$

where $X(\lambda)$ is the original data (Raman spectroscopy), $p_1(\lambda), p_2(\lambda), \ldots, p_n(\lambda)$ are the new orthogonal vectors containing the most variations (loadings spectroscopy), and $u_1, u_2, \ldots, u_A$ are the values (latent variables or scores, also known as principal components) on the new orthogonal space (created by axes $p_1(\lambda), p_2(\lambda), \ldots, p_n(\lambda)$). Thus, $X(\lambda)$ can be represented by one point in the new space with the coordinates $(u_1, u_2, \ldots, u_A)$. This dimension reduction method is beneficial as variable selection is particularly difficult for spectroscopy data. However, one shortcoming of this data reduction method is its lack of physical meaning of the original data.

Several classification methods have been applied on spectroscopy, such as artificial neural network (ANN), decision tree (DT), support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbour (KNN). As the scatter dots drawn by PC1 and PC2 in this research has shown clear separated groups, classification methods based on distances is preferred. Since the interfaces between groups in the scatter plot may not be linear, non-linear methods such as SVM and KNN were better for our data.

KNN can separate groups via the irregular line and isolated cloud methods, and the closest k samples are used for the group estimation of a new sample [21]. It weighs the neighbours according to distances using certain kernel functions. The Minkowski distance function is used in this paper.

PCA was performed using the PCA related functions in the ChemSpec package, and the KNN was carried out using the kknn function in the kknn package.

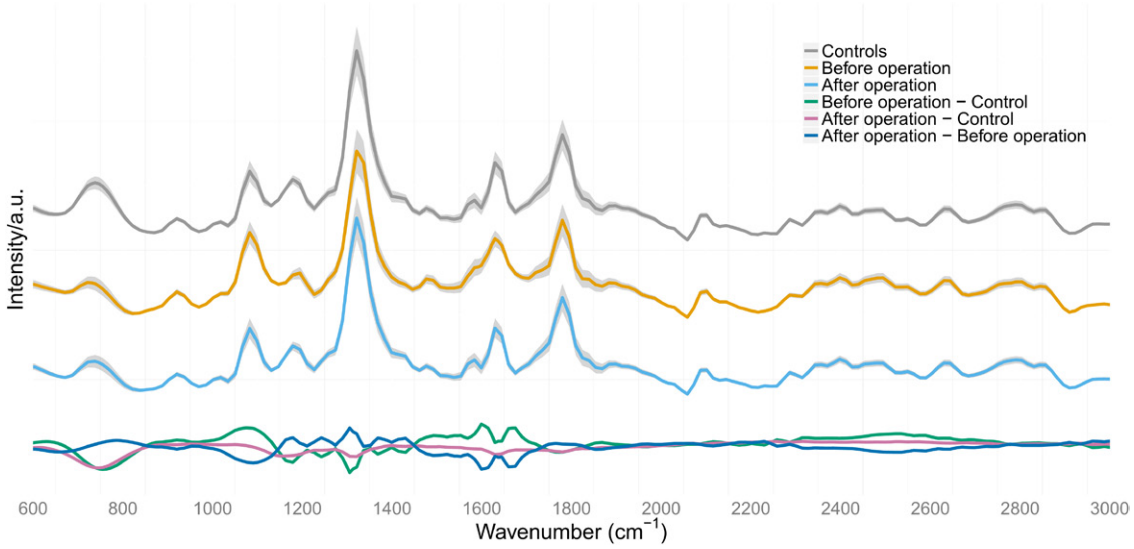

Figure 2. Mean (colored line) ± SD (gray shadow) spectra of the control group and the pre- and post-operative colon cancer group.

Table 1. Demographics of study population.

Groups		Controls ($n = 75$)	Colon cancer before operation ($n = 65$)	Colon cancer after operation ($n = 60$)
Age (years, s.d.)		67 (14)	59 (9)	59 (9)
Gender (% F)		51%	49%	45%
TNM Stages	I	NA	4	4
	II	NA	21	21
	III	NA	26	26
	IV	NA	14	9

Table 2. Peak assignments and p value.

Wavenumber (cm^{-1})	Vibrational modes	Molecules	Trends for cancer	Chromophore?
750	CH_2 rocking, symmetric breathing	Tryptophan [22]	Decrease ($P < 0.001$)	Yes
1083	CC or PO_2 stretching	Phospholipids in nucleic acids [23, 24]	Increase ($P < 0.001$)	No
1165	CC stretching	β-carotene [25]	Decrease ($P < 0.001$)	Yes
1321	Ring breathing	Nucleic acids adenine base [26]	Decrease ($P < 0.001$)	No
	CH_2CH_3 twisting	Collagen, tryptophan [27]		
1629	CC asymmetric stretching	Porphyrin moiety of hemoglobin [28]	Increase ($P < 0.001$)	Yes
1779	—	—	Decrease ($P = 0.003$)	NA

Results

Raman spectroscopy

Raman spectroscopy of serum taken from 65 pre-operative colon cancer patients, 60 post-operative colon cancer patients and 75 healthy volunteers was measured. For each serum sample, three measurements were performed and the averaged Raman spectra were recorded. The Raman spectra were preprocessed by smoothing, baseline correction and area normalization. Figure 2 shows the mean ± standard deviation (SD) spectra of the three groups. Six notable Raman peaks existed and can be consistently observed in all three groups. Peak intensity changes are evident at certain wavenumber regions as can be seen in the difference spectra in figure 2 (three lines at the bottom).

Peaks at the wavenumbers of 750, 1083, 1165, 1321, 1629 and 1779 cm^{-1} were assigned to nucleic acids, amino acids and choromophores such as β-carotene and porphyrin in accordance to previous research (table 2). To evaluate whether the peak intensity variations between groups have statistical significance, one-way ANOVA was performed on the Raman intensities of the six peaks. The ANOVA results show that all six peaks are statistically different among the three groups. For all six peaks, the peak intensities of the post-operative group tend to be closer to the ones of the healthy group compared to those of the pre-operative group. These variations indicate corresponding changes in serum substance between groups.

Principal component analysis (PCA)

PCA was utilized on the original spectra to extract the key information in the spectra. The input variables were the pre-treated Raman spectra. After the PCA analysis, scree plots of

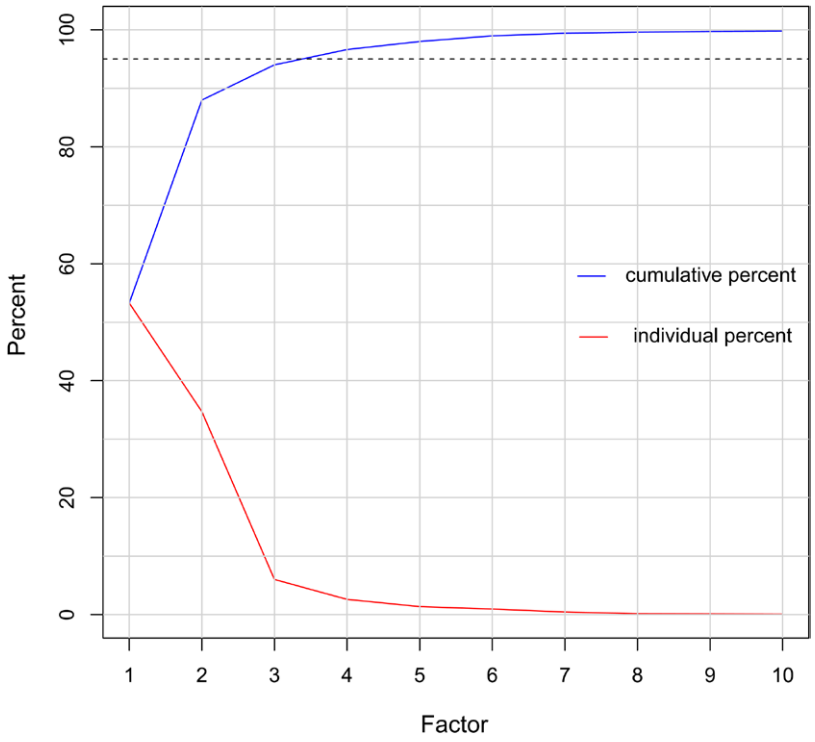

Figure 3. Scree plot of cumulative and individual percentages of the first 10 PCA scores (factors).

the cumulative percent of variance of the first 10 PCs were reviewed to see the contribution of the PCs to the total variation (figure 3). From the figure we can see the total variation of the first three PCs accounted for 94% of the total variation (PC1: 53%; PC2: 35%; PC3: 6%). Because the number of PCs retained should guarantee that the PCs describe no less than 90% of the total variance [29], the first three PCs were maintained for further KNN analysis.

The distribution of the three PCs was visualized by plotting them in a principal component space using two components as Cartesian axes for each plot. The majority of the spots (figure 4) were inside the 95% confidence ellipses. The scatter plot using PC1 and PC2 has the best discrimination performance: the confidence ellipses only have a small proportion of overlap (figure 4(a)). This demonstrated that samples can be discriminated well by PC1 and PC2. For the PC2-PC3 plot (figure 4(b)), the ellipses of the post-operative group (brown) and the normal group (gray) intersect in the middle but have little overlap with the blue ellipse that represents the pre-operative group. The three confidence eclipses in the third plot (figure 4(c)) have the largest overlap. This suggests that the sole use of PC1 and PC3 is not enough for discrimination.

Loading profiles of the first three PCs are displayed in figure 5. Peaks with differences between groups in the original spectra can also be found in the loadings spectra [30]. For the loadings plot of PC1 (which accounts for 53% of the total variance), peaks at the wavenumbers of 1178, 1305 and 1629 cm^{-1} have the highest intensity. For PC2 (35%), the peaks existed at 1211 and 1492 cm^{-1}. And for PC3, peaks at 1083, 1321 and 1779 cm^{-1} have the highest intensity. These results indicate that the peak positions at the wavenumbers of 1083, 1178, 1211, 1305, 1321, 1492, 1629 and 1779 cm^{-1} have the highest weights for the PCA discrimination of different groups.

For the evaluation of the diagnostic ability of the PCA model, an orthogonal distance plot for all the samples was drawn and marked with the horizontal critical boundary (figure 6). The orthogonal distance is the distance in the orthogonal direction of the PCA space and can be seen as a lack of fit [21]. By using the first three PCs which accounted for more than 90% of the total variance, only 7 spots (2 gray, 3 brown and 2 blue) were found above the horizontal dotted line and were marked as outliers.

K nearest neighbour analysis (KNN)

KNN was used on the PCs obtained from PCA for classification. In the construction of KNN model, 1/3 of the samples were randomly selected as the training set, and the remaining 2/3 were used as the test set. For the selection of the k value, leave-one-out cross validation of all k values under 25 were tested, and the k values having the best prediction ability was retained. Based on the results, the k value was set as 1. And the kernel function was set as rectangular for its best performance. The test set was then used for the verification of the constructed KNN model.

Figure 7 shows the classification results of KNN. Three PCs were set as the parameters in the pairs plot. The correctly classified samples were printed in dark green, and the incorrect ones in red. The diagnostic results can be found in table 3: the accuracy was 91.0%, the specificity was 92.6%, the sensitivity was 87.0% for pre-operative colon cancer samples and was 94.1% for post-operative samples.

$$\text{Accuracy} = \frac{\text{number of correctly identified cases}}{\text{total cases}} \times 100\%$$

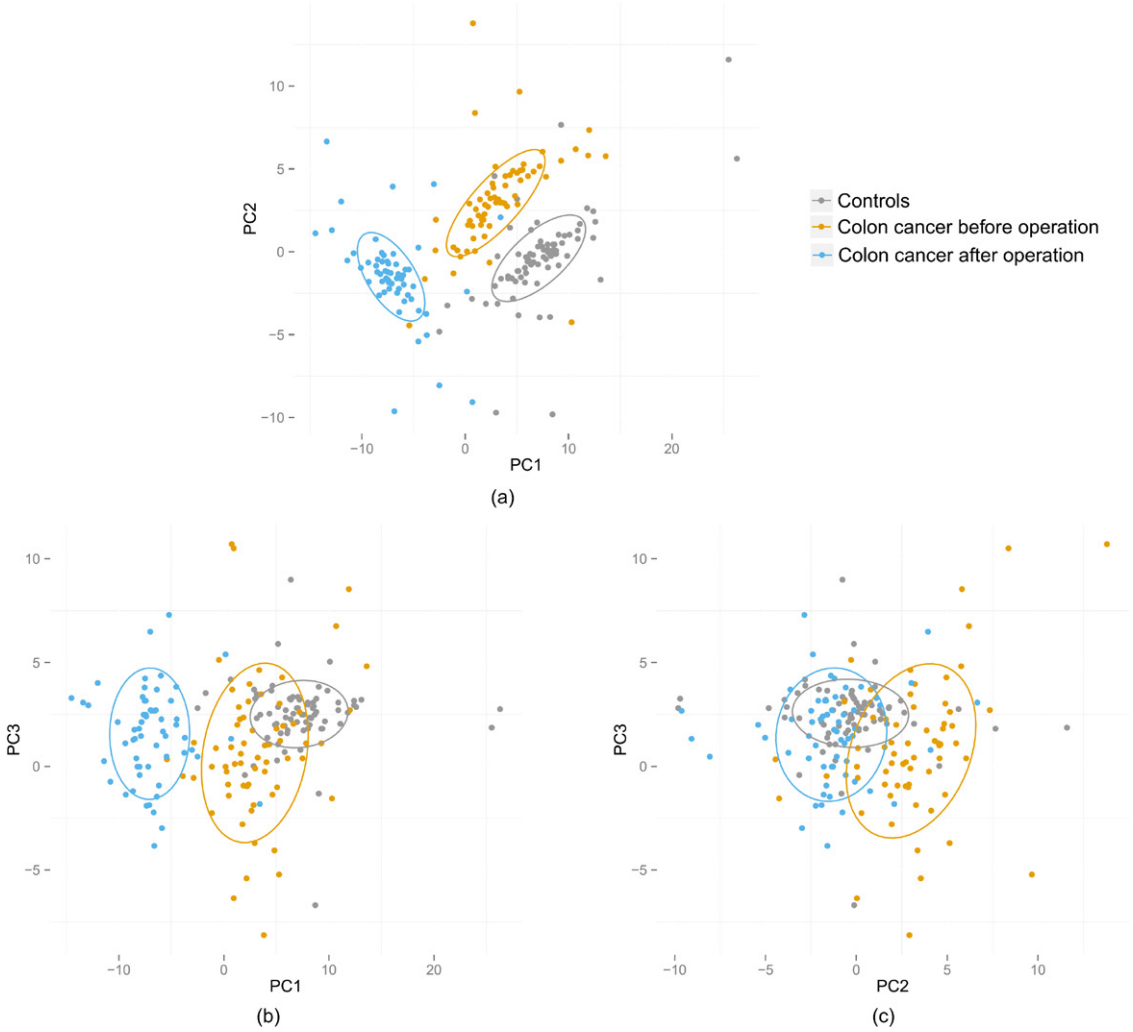

Figure 4. Scatter plot of three pairs of PCs: (a) PC1 versus PC2; (b) PC2 versus PC3; (c) PC1 versus PC3.

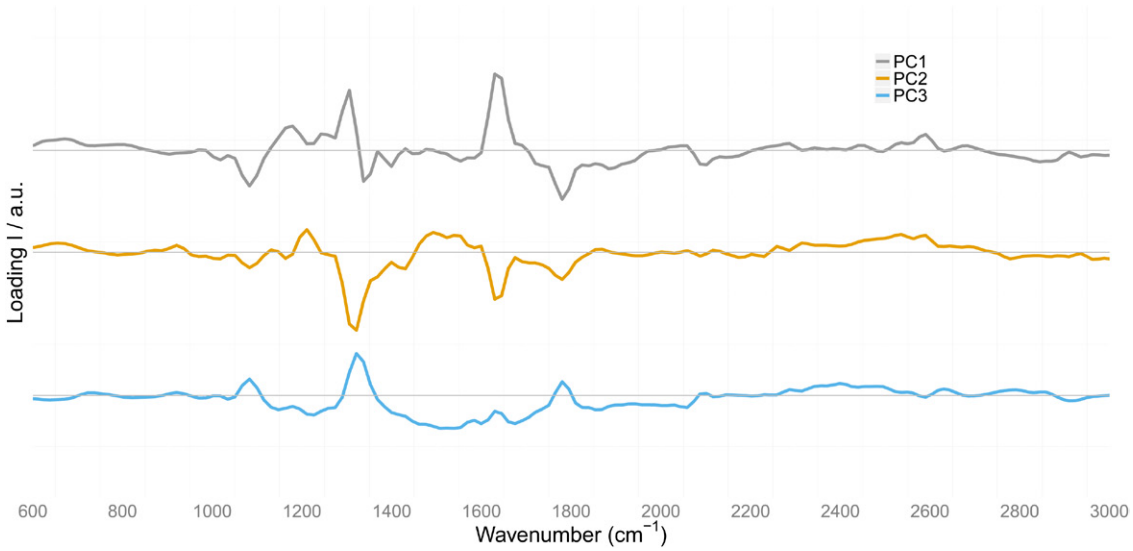

Figure 5. Loadings plot of PC1, PC2 and PC3. Peaks with higher intensity are the principal differences between groups.

Figure 6. Diagnostic plots using orthogonal distance (OD). The spots above the dotted line are outliers (grey: controls; brown: pre-operative; blue: post-operative).

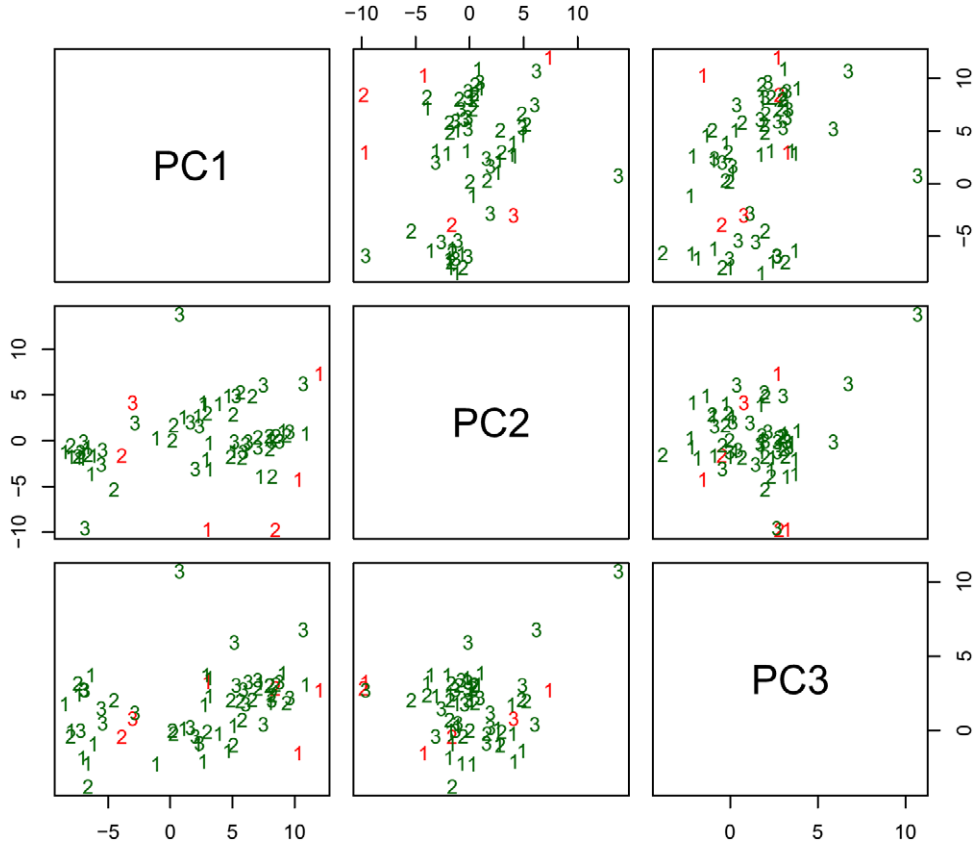

Figure 7. KNN classification results for three groups of samples. The correctly classified samples are shown in dark green and incorrectly classified samples are in red.

Table 3. Diagnostic results for PCA-KNN.

Group	Prediction			Accuracy (%)	Sensitivity (%)	Specificity (%)
	Control	Pre-operative	Post-operative			
Control	25	0	2	91.0	NULL	92.6
Pre-operative	2	20	1		87.0	
Post-operative	0	1	16		94.1	

$$\text{Sensitivity} = \frac{\text{number of correctly identified cases as colon cancer cancer}}{\text{total cases of colon cancer cancer}} \times 100\%$$

$$\text{Specificity} = \frac{\text{number of correctly identified cases as healthy}}{\text{total cases of healthy}} \times 100\%$$

Discussion

Raman spectroscopy techniques have been used for the detection of many diseases [31, 32] and various parts of colorectal tissues. The applications of Raman techniques on the colon include but are not limited to the detection of colonic polyps lesion [5], colon cancer relapse [13] and colon cancer screening [33]. However, few studies have applied Raman spectroscopy on the analysis of serum. Using serum spectroscopy for cancer screening is more convenient than tissue imaging techniques and fecal tests. Through the analysis of certain Raman peaks, content changes of bio-macromolecules and biomarkers for diagnosing colon cancer can be detected.

In our experiment, six Raman peaks were found to have statistically significant changes between groups ($P < 0.05$). The decrease of tryptophan represented by peaks at 750 cm^{-1} may be the result of degradation of tryptophan caused by the enzyme indoleamin (2,3)-dioxygenase stimulated by Th1 immune response (IDO) in colorectal cancer patients [34]. The increased intensity of nucleic acids at 1083 cm^{-1} are due to higher nucleic acids bases in serum caused by the abnormal metabolism of DNA and RNA in cancer patients [24]. Higher levels of anti-oxidant β-carotene in serum were associated with a higher survival rate for colon cancer [35], this indicate that cancer may cause the consumption and subsequent decrease of β-carotene in serum represented by 1165 cm^{-1}. The peaks at 1321 cm^{-1}, contributed to by nucleic acids, collagen and tryptophan also decreased and reflected lower concentrations of corresponding substances in serum. This peak is also apparent in SERS of blood serum for colorectal cancer detection [14], but the differences in our experiment were more apparent. The Raman peaks at 1629 cm^{-1} were caused by the porphyrin moiety of hemoglobin. Hemoglobin is modulated by haptoglobin in response to inflammation and injury [36]. The increase of peak 1629 cm^{-1} may be caused by the response of human body to cancerous cells. Peaks at 1779 cm^{-1} have not been found in literature. It is possibly caused by the blue-shift of some of the resonance Raman peaks caused by laser power fluctuation. Since SERS may cause Raman peaks to shift and blood is the excretion of tissues, we can conclude that the two peaks at 1629 and 1779 cm^{-1} were produced by the same substances that caused 1655 cm^{-1} in the SERS experiment [14]. The decrease of those two peaks indicates that bond breakups have occurred and bio-functions were lost in the occurrence of colon cancer.

In the scatter plot drawn by PCs, most spots belonging to different groups have little overlap in the PC1-PC2 plot which indicates a good separation between groups. The orthogonal distance plot showed that a good diagnostic accuracy has been obtained by this PCA model. Also, the loadings spectra of PCs enabled us to identify the positions with the highest weights for discriminating groups, and several bonds were found to contribute a large degree to the PC scores. The PC loading peaks at 1492 cm^{-1} were reported as the characteristic peak for colon tissues [26]. This implies that PC loadings peaks have certain correlations with their original Raman peaks. However, as the PC scores are only the orthogonal transformation of the original data, they cannot be interpreted physically or chemically. KNN then was used on the obtained PCs for discriminant analysis. The sensitivities obtained were 87.0% and 94.1% for the pre- and post-operative groups respectively, and the specificity was 92.6%. These results are commensurate to that of colonoscopy (sensitivity: 83–95%; specificity: 83–99%) and higher than fecal occult blood test (FOBT) (sensitivity: 11–89%; specificity: 91–98%), flexible sigmoidoscopy (sensitivity: 60–70%; specificity: 60–70%) and CT colonography (sensitivity: 53–95%; specificity: 18–99%) [37, 38]. However, as a classification method, KNN has both merits and demerits. As a technique focusing on distances between objects, it does not rely on any assumption [39]. It is correlated strongly with the selection of the value of k, which may induce overfitting or underfitting. It is also time-consuming when manipulating large datasets. Thus, KNN is often used as a reference method [21]. All the above stated points complicate the application of KNN as a grouping method for clinical use.

Conclusion

In this study, serum taken from colon cancer patients both before and after operation was tested using Raman spectroscopy and was compared with serum from the controls. Six Raman peaks at the positions of 750, 1083, 1165, 1321, 1629 and 1779 cm^{-1} were assigned to nucleic acids, tryptophan and chromophores, and were found to have significant differences between the three groups. The concentration changes of chemical components in serum is a major factor for the spectral changes, thus the results indicate corresponding component changes that have taken place along with the development of colon cancer. The multivariate statistical techniques of PCA and KNN were

utilized to develop diagnostic algorithms for the classification. The scatter plot of PCs revealed dispersed spot distribution in the PC space between groups. The following KNN analysis further confirmed the diagnostic ability of those PCs, achieving a diagnostic accuracy of 91.0% and a specificity of 92.6%. This paper is an exploratory study for using Raman spectroscopy of serum for the diagnosis of colon cancer, and hopefully will be perfected in future research for clinical applications.

References

[1] International Agency for Research on Cancer 2014 *World Cancer Report 2014* (Geneva: World Health Organization)

[2] Edge S, Byrd D, Compton C, Fritz A, Greene F and Trotti A (eds) 2011 *AJCC Cancer Staging Manual* 7th edn (Berlin: Springer)

[3] Winawer S J 2007 Colorectal cancer screening *Best Pract. Res. Clin. Gastroenterol.* **21** 1031–48

[4] Davila R E *et al* 2006 ASGE guideline: colorectal cancer screening and surveillance *Gastrointest Endosc.* **63** 546–57

[5] Cappell M S 2007 From colonic polyps to colon cancer: pathophysiology, clinical presentation, screening and colonoscopic therapy *Minerva Gastroenterol. Dietol.* **53** 351–73

[6] Creeden D J, Junker F, Vogel-Ziebolz S and Rex D 2011 Serum tests for colorectal cancer screening *Mol. Diag. Ther.* **15** 129–41

[7] Cappell M S 2008 Reducing the incidence and mortality of colon cancer: mass screening and colonoscopic polypectomy *Gastroenterol. Clin. North Am.* **37** 129–60

[8] Schie I W and Huser T 2013 Methods and applications of Raman microspectroscopy to single-cell analysis *Appl. Spectrosc.* **67** 813–28

[9] Li X, Yang T and Li S 2012 Discrimination of serum Raman spectroscopy between normal and colorectal cancer using selected parameters and regression-discriminant analysis *Appl. Opt.* **51** 5038–43

[10] Li X, Yang T and Lin J 2012 Spectral analysis of human saliva for detection of lung cancer using surface-enhanced Raman spectroscopy (SERS) *J. Biomed. Opt.* **17** 037003

[11] Scherer J R, Go M K and Kint S 1974 Raman spectra and structure of water from − 10 to 90 deg *J Phys Chem.* **78** 1304–13

[12] Liu L *et al* 2013 *Characterizing Autofluorescence Generated from Endogenous Porphyrins in Cancerous tissue of Human Colon: Case Studies* (San Francisco: SPIE Press) p 857703–7

[13] Sahu R K, Argov S, Walfisch S, Bogomolny E, Moreh R and Mordechai S 2010 Prediction potential of IR-micro spectroscopy for colon cancer relapse *Analyst* **135** 538–44

[14] Lin D *et al* 2011 Colorectal cancer detection by gold nanoparticle based surface-enhanced Raman spectroscopy of blood serum and statistical analysis *Opt Express* **19** 13565–77

[15] Lin X-M, Cui Y, Xu Y-H, Ren B and Tian Z-Q 2009 Surface-enhanced Raman spectroscopy: substrate-related issues *Anal. Bioanal. Chem.* **394** 1729–45

[16] Robert B 2009 Resonance Raman spectroscopy *Photosynth. Res.* **101** 147–55

[17] Weakley A T, Warwick P C T, Bitterwolf T E and Aston D E 2012 Multivariate analysis of micro-raman spectra of thermoplastic polyurethane blends using principal component analysis and principal component regression *Appl. Spectrosc.* **66** 1269–78

[18] Yasser M, Shaikh R, Chilakapati M K and Teni T 2014 Raman spectroscopic study of radioresistant oral cancer sublines established by fractionated ionizing radiation *PLoS ONE* **9** e97777

[19] Kelly J 2010 Robust classification of low-grade cervical cytology following analysis with ATR-FTIR spectroscopy and subsequent application of self-learning classifier eClass *Anal. Bioanal. Chem.* **398** 2191–201

[20] Teh S K 2009 Spectroscopic diagnosis of laryngeal carcinoma using near-infrared Raman spectroscopy and random recursive partitioning ensemble techniques *Analyst* **134** 1232–9

[21] Varmuza K and Filzmoser P 2009 *Introduction to Multivariate Statistical Analysis in Chemometrics* (Boca Raton: CRC Press) p 336

[22] Liu C-H *et al* 2013 Resonance Raman and Raman spectroscopy for breast cancer detection *Technol. Cancer Res. Treat.* **12** 371–82

[23] Lin D *et al* 2012 *Diagnostic Potential for Gold Nanoparticle-Based Surface-Enhanced Raman Spectroscopy to Provide Colorectal Cancer Screening Using Blood Serum Sample* (Wuhan, China: SPIE Press) pp 83290L–10

[24] Chen Y *et al* 2012 Label-free serum ribonucleic acid analysis for colorectal cancer detection by surface-enhanced Raman spectroscopy and multivariate analysis *J. Biomed. Opt.* **17** 067003

[25] Tschirner N *et al* 2009 Resonance Raman spectra of β-carotene in solution and in photosystems revisited: an experimental and theoretical study *Phys. Chem Chem Phys.* **11** 11471–8

[26] Pinzaru S C, Andronie L M, Domsa I, Cozar O and Astilean S 2008 Bridging biomolecules with nanoparticles: surface-enhanced Raman scattering from colon carcinoma and normal tissue *J. Raman Spectrosc.* **39** 331–4

[27] Li S *et al* 2014 Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques *Opt. Express* **22** 25895–908

[28] Lemler P, Premasiri W R, DelMonaco A and Ziegler L D 2014 NIR Raman spectra of whole human blood: effects of laser-induced and *in vitro* hemoglobin denaturation *Anal. Bioanal. Chem.* **406** 193–200

[29] Lee D-S, Noh B-S, Bae S-Y and Kim K 1998 Characterization of fatty acids composition in vegetable oils by gas chromatography and chemometrics *Anal. Chim. Acta.* **358** 163–75

[30] Nogueira G V 2005 Raman spectroscopy study of atherosclerosis in human carotid artery *J. Biomed. Opt.* **10** 031117

[31] Li Z *et al* 2014 Surface-enhanced Raman spectroscopy for differentiation between benign and malignant thyroid tissues *Laser Phys. Lett.* **11** 045602

[32] Chen L *et al* 2013 Near-infrared confocal micro-Raman spectroscopy combined with PCA-LDA multivariate analysis for detection of esophageal cancer *Laser Phys.* **23** 065601

[33] Levin B *et al* 2008 Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American cancer society, the US multi-society task force on colorectal cancer, and the american college of radiology *CA Cancer J. Clin.* **58** 130–60

[34] Huang A, Fuchs D, Widner B, Glover C, Henderson D C and Allen-Mersh T G. 2002 Serum tryptophan decrease correlates with immune activation and impaired quality of life in colorectal cancer *Br. J. Cancer.* **86** 1691–6

[35] Moy K A, Weinstein S, Männistö S and Albanes D 2014 Serum alpha-tocopherol, beta carotene and cancer survival in the ATBC Study *Cancer Res.* **74** 2183–3

[36] Yu C, Xu C, Xu L, Yu J, Miao M and Li Y 2012 Serum proteomic analysis revealed diagnostic value of hemoglobin for nonalcoholic fatty liver disease *J. Hepatol.* **56** 241–7

[37] Martín-López J E, Beltrán-Calvo C, Rodríguez-López R and Molina-López T 2014 Comparison of the accuracy of CT colonography and colonoscopy in the diagnosis of colorectal cancer *Colorectal Dis.* **16** O82–9

[38] Bretthauer M 2011 Colorectal cancer screening *J. Intern. Med.* **270** 87–98

[39] Wehrens R 2011 *Chemometrics with R—Multivariate Data Analysis in the Natural Sciences and Life Sciences San Michele all'Adige* (Berlin: Springer)