Raman spectral post-processing for oral tissue discrimination – a step for an automatized diagnostic system

LUIZ FELIPE C. S. CARVALHO,1,4 MARCELO SAITO NOGUEIRA,2,5 LÁZARO P. M. NETO,1 TANMOY T. BHATTACHARJEE,1 AIRTON A. MARTIN3,6

1Univap/Instituto de Pesquisa e Desenvolvimento, Laboratório de Espectroscopia Vibracional Biomédica, Avenida Shishima Hifumi, 2911, São José dos Campos/SP, CEP: 12244-000, Brazil
2Universidade de São Paulo/ São Carlos Institute of Physics, Optics Group, Biophotonics Division, Avenida Trabalhador São Carlense, 400, São Carlos/SP, CEP: 13566-590, Brazil
3Biomedical Engineering Innovation Center - Biomedical Vibrational Spectroscopy Group, Universidade Brasil - UnBr - Rua Carolina Fonseca, 235 - 08230-030 – Itaquera, São Paulo/SP/
Visiting Professor Universidade Federal do Piauí - UFPI - Campus Ministro Petrônio Portella Departamento de Física – CCN Bairro Ininga Teresina, PI, CEP: 64049-550, Brazil
4luisfelipecarvalho@hotmail.com
5marcelosaitonogueira@gmail.com
6amartin@pq.cnpq.br

Abstract: Most oral injuries are diagnosed by histopathological analysis of a biopsy, which is an invasive procedure and does not give immediate results. On the other hand, Raman spectroscopy is a real time and minimally invasive analytical tool with potential for the diagnosis of diseases. The potential for diagnostics can be improved by data post-processing. Hence, this study aims to evaluate the performance of preprocessing steps and multivariate analysis methods for the classification of normal tissues and pathological oral lesion spectra. A total of 80 spectra acquired from normal and abnormal tissues using optical fiber Raman-based spectroscopy (OFRS) were subjected to PCA preprocessing in the z-scored data set, and the KNN (K-nearest neighbors), J48 (unpruned C4.5 decision tree), RBF (radial basis function), RF (random forest), and MLP (multilayer perceptron) classifiers at WEKA software (Waikato environment for knowledge analysis), after area normalization or maximum intensity normalization. Our results suggest the best classification was achieved by using maximum intensity normalization followed by MLP. Based on these results, software for automated analysis can be generated and validated using larger data sets. This would aid quick comprehension of spectroscopic data and easy diagnosis by medical practitioners in clinical settings.

© 2017 Optical Society of America

OCIS codes: (120.3890) Medical optics instrumentation; (120.6200) Spectrometers and spectroscopic instrumentation; (170.1610) Clinical applications; (200.4560) Optical data processing; (300.6330) Spectroscopy, inelastic scattering including Raman.

References and links
1. P. Holmstrup, P. Vedtofte, J. Reibel, and K. Stoltze, “Oral premalignant lesions: is a biopsy reliable?” J. Oral Pathol. Med. 36(5), 262–266 (2007).
2. P. N. Prasad, Introduction to Biophotonics (John Wiley & Sons, 2004).
3. L. F. C. Carvalho, F. Bonnier, K. O’Callaghan, J. O’Sullivan, S. Flint, H. J. Byrne, and F. M. Lyng, “Raman micro-spectroscopy for rapid screening of oral squamous cell carcinoma,” Exp. Mol. Pathol. 98(3), 502–509 (2015).
4. L. F. Carvalho, T. Dreyer, J. Almeida, and H. Martinho, “Diagnosis of inflammatory lesions by high-wavenumber FT-IR spectroscopy,” in Latin America Optics and Photonics Conference,(Optical Society of America, 2012), LT4A. 3.
5. A. Cosci, M. S. Nogueira, S. Pratavieira, A. Takahama, Jr., R. S. Azevedo, and C. Kurachi, “Time-resolved fluorescence spectroscopy for clinical diagnosis of actinic cheilitis,” Biomed. Opt. Express 7(10), 4210–4219 (2016).
1. Introduction

Optical biopsy refers to techniques where the light-tissue interaction is analyzed and information regarding the pathological state of the tissue is obtained, either in vivo or ex vivo. Optical spectroscopy techniques such as infrared absorption, fluorescence, optical coherence tomography, diffuse reflectance spectroscopy and Raman scattering can be employed. Holmstrup, et. al. [1] have described many molecular interactions and features in cells and tissues that cannot be assessed by conventional histopathology, but can be probed by optical techniques. Unlike the conventional histopathology, which is based on morphological changes; optical spectroscopy use biochemical information and hence can be used to obtain early and differential diagnosis of multiple lesions [2–9].

Currently, many researches are exploring the use of Raman spectroscopy as an in vivo tool [10–13], to help pathologists get an early and reliable diagnosis for therapeutic decision making. This technique can also be used as a tool for monitoring the type and stages of pathological processes. Another extremely important clinical application is guiding the...
regions to be biopsied. This procedure will assist pathologists in the diagnosis, optimizing the surgical procedure by the involved health professional [14–17].

In order to translate Raman spectroscopy into clinics, it is vital to develop easy to use and reliable tools that will help clinicians comprehend the results and make a diagnosis or decide on a therapeutic intervention. However, spectroscopy yields information in form of peaks. The intensity changes, shifts and other subtle variations can be best understood by spectroscopists and statisticians, not clinicians. This gap can be bridged by softwares that can automatically process and analyse the data to give output easily gauged by the clinicians. At the moment, experts in the field of Raman spectroscopy use a plethora of preprocessing tools and analysis methodologies [18]. Each routine followed have their own advantage and disadvantage in terms of sensitivity, specificity and application. A standard routine of spectra preprocessing followed by specific multivariate analysis may help write a software that can be used in clinics.

Therefore, in this study, we have used varied preprocessing and multivariate analysis methodologies on clinical samples; to help identify the best routine for classification between normal and abnormal oral samples. In order for the routine to be thoroughly clinic oriented, spectra were acquired using a fibre-optic probe that can easily be used in clinics. Results of the study are discussed in the manuscript.

2. Materials and methods

2.1 Research ethics

This work (number 1132237-2015) was approved by the Research Ethics Committee from Universidade do Vale do Paraíba (UNIVAP) after submission via Plataforma Brasil website.

2.2 Samples

Eight oral tissue samples (size of 1x1mm) – four normal and four abnormal mucosa, were obtained from a dental clinic. An expert pathologist was invited and confirmed the status of each tissue after histopathological evaluation. The abnormal samples were found to be inflammatory fibrous hyperplasia, malignant tissue, inflammatory minor salivary gland process and oral lichen planus. The normal tissue samples were taken from the region around the lesions when it was possible. After surgery, a fragment was excised from the surgical specimen, washed with saline solution 0.9% (NaCl) Aster and placed in previously identified Nalgene tube. For transportation of samples to the Laboratory Vibrational Spectroscopy Biomedical the UNIVAP, the Nalgene held fragment tubes were placed in a liquid nitrogen filled container. They were stored in a −70 °C freezer (Thermo Scientific LTD.)

2.3 Raman instrument

Spectra were acquired using a Raman spectrometer - Kaiser Optical Systems imaging spectrograph Holospec, f/ 1.8i-NIR and a 785nm. In order for the routine to be thoroughly clinic oriented, spectra were acquired using a fibre-optic probe (EMVision) that can easily be used in clinics. The Raman scattered light was collected by the same fiber, through a gold dichroic mirror and finally focused on the entrance aperture of the spectrometer through a holographic notch filter. The Raman scattered signal was then collected by a CCD detector (Andor - IDUs 420 Series) whose quantum efficiency is around 95%. Two 20 second acquisitions were averaged for each spectrum, power used being 60mW. Considering the small size of the tissue, 10 spectra were acquired per sample. This helps take tissue heterogeneity into consideration. The acquisition time was made with an interaction of 40 seconds for each spectrum (20x2seconds).

The optical fiber used in the study was developed by Emvision – Advanced Optical designs and structurally composed of 1 100 micron fiber to drive energy from the laser to the tissue and six fibers (100 microns) to capture scattered light. It would be single fiber
excitation with 7 collection fibers that surround the excitation fiber. Band Pass filter at the probe tip on the excitation fiber, and a Long Pass filter at the probe tip on the 7 collection fibers. The fibers are separated at a breakout (Y style) to have the laser fiber separated for the collection fibers. The collection fibers are oriented in a line at the connector.

2.4 Data processing

The background subtraction was performed by fitting the data using an asymmetric quadratic polynomial function and a homemade routine in MATLAB software (R2012a version, Mathworks, Natick, Massachusetts, USA). The spectra smoothing was performed by a Savitsky-Golay filter (5th order, frame size 7). The spectral data were normalized by the area or by the intensity maximum of each spectrum. After this normalization, each of 80 spectra (40 from pathological tissue and 40 from normal tissue) was analysed by multivariate data analysis. In order to show the main differences between these spectra, the average spectrum of each group was obtained and plotted using the software OriginPro 8 (OriginLab Corporation, Northampton, Massachusetts, USA).

2.5 Data analysis

The multivariate data analysis was performed by using the PCA preprocessing in the z-scored data set, and the KNN (K-Nearest Neighbors), J48 (unpruned C4.5 decision tree), RBF (Radial Basis Function), RF (Random Forest), and MLP (Multilayer Perceptron) classifiers at WEKA software (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand). The classification was based on 10-fold cross-validation. The spectral curves after each normalization were compared by using the entire spectra or PCA components from PC2 to PC7, which contributed to about 80% of the total variance. The plot of PCs loadings were analyzed using other MATLAB routine. After the classification, each combination of classifier and set of parameters were compared by its accuracy and Area Under the Receiver Operating Characteristic curve (AUROC).

3. Results and discussion

The main idea of the study is to search for a preprocessing and analysis routine that can analyze spectra acquired from human samples in a clinic and provide fast, reliable, easily comprehensible output for the clinicians to aid diagnosis/therapeutic decision making. To this end, spectra from normal and abnormal oral tissues acquired in clinics were analyzed by different combinations of preprocessing and multivariate analysis (Table 1). Full spectra or PCA loading factors after area normalization or highest intensity normalization were used as input for KNN, J48, RBF, MLP, or RF.

| Normalization                  | Parameters used for classification | Classifiers            |
|--------------------------------|-----------------------------------|------------------------|
| Area normalization (RSNAS)     | Full spectra                      | KNN, J48, RBF, MLP, RF |
| Area normalization (RSNAS)     | PCA loading factors               | KNN, J48, RBF, MLP, RF |
| Highest peak normalization     |                                   |                        |
| (RSNIM)                        | Full spectra                      | KNN, J48, RBF, MLP, RF |
| Highest peak normalization     |                                   |                        |
| (RSNIM)                        | PCA loading factors               | KNN, J48, RBF, MLP, RF |

Table 1. Types of normalization, parameters and classifiers used for classification, where RSNAS is Raman Spectra Normalized by the Area of Spectrum and RSNIM is Raman Spectra Normalized by the Intensity Maximum.
Figure 1 shows mean spectra of normal and abnormal tissue after intensity (A) and area (B) normalization. Each spectra shows the Amide I, CH$_2$ bending, Amide III and the 1002 cm$^{-1}$ phenylalanine band. Changes in Amide III and 1000-1200 cm$^{-1}$ regions with respect to control are clearly visible in the pathological tissue spectra.

In order to search for spectral bands that lead to a better differentiation between the normal and pathological tissues, we analyzed the loadings of each PC for the Raman spectra with both normalizations. Since loadings were different for each type of normalization, the contributions of each PC may be lead to different results when performing classification methods. In Fig. 2 and 3, we show the results of PC2 and PC3 in the comparison between loadings of PCs for each type of normalized Raman spectra. Since higher differences may be noticed from 980 cm$^{-1}$ to 1090 cm$^{-1}$, from 1350 to 1600 cm$^{-1}$, and from 1730 to 1800 cm$^{-1}$ for PC2, and from 1000 to 1100 cm$^{-1}$, from 1200 to 1330 cm$^{-1}$, from 1440 to 1460 cm$^{-1}$, and 1580 to 1800 cm$^{-1}$ for PC3, these ranges may be related to the different results on sensitivity and specificity due to the contribution of these PCs. We could observe changes in both PC2 and PC3 associated to Amide III, and also the 1750 cm$^{-1}$ band in PC3 can be attributed to lipids, which are plentiful in normal oral mucosa. After doing the whole analysis using the full spectra or a particular spectral window, we observed that by using a narrow band in the spectra the sensitivity and specificity of the technique increase considerably. It was already shown in previous studies [3, 4, 9] that some misclassification occurs due to the similarity of
the tissues, i.e. inflammatory fibrous hyperplasia and normal tissue. In addition, it is important to understand that this “narrow Raman biomarker” could be useful in the next step of the work for *in vivo* analysis [9]. At the time of software development for spectra acquisition, the clinicians would be able to look for the biomarker in order to better discriminate the tissues instead to use the full spectra.

![PC2 loadings for Raman spectra](image1)

**Fig. 2.** Plot of PC2 loadings for Raman spectra normalized by the intensity maximum or by area. The behavior of loadings is mostly different from 900 cm\(^{-1}\) to 1600 cm\(^{-1}\) range.

![PC3 loadings for Raman spectra](image2)

**Fig. 3.** Plot of PC3 loadings for Raman spectra normalized by the intensity maximum or by area. The behavior of loadings is mostly different from 900 cm\(^{-1}\) to 1600 cm\(^{-1}\) range.

As it can be seen in Fig. 4, PC2 and PC5 separates normal from abnormal after area normalization (A) while PC2 and PC5 separate the same after highest peak normalization. It is however clear that there is scope of improvement in classification.

Although PC1 has a large contribution to variance, it may not always contain biologically relevant information for classification [3], due to the heterogeneity biological tissues. By excluding PC1 we eliminate features from the variability of our measurements, making our analysis less likely to overfitting and providing more chances of success for external validation. Thus, we excluded it and kept only the spectral components with attributes that can be associated with biological tissues for classification using PCA parameters.
Fig. 4: PCA Scores plot for Raman spectra A) normalized by the area under the spectrum (PC2 x PC5 x PC6) and B) normalized by the intensity maximum (PC2 x PC3 x PC5). Although we do not observe a good discrimination only by using three parameters, we achieved very good values of sensitivity and specificity when using classification methods for twenty one or more parameters.

Several statistical methods like neural network, principal component analysis, linear discriminant analysis and cluster analysis are being used in the discrimination of abnormal cases from the normal tissues. Different statistical approaches have been used to quantify the observed vibrational spectral differences between normal and diseased tissues, and also with respect to the different preprocessing methods, in order to evaluate the efficiency of these methods. Table 2 shows the results of the statistical analysis.

Table 2. Confusion matrices and the Area Under the Receiver Operating Characteristic curve (AUROC) for the classification of the Raman Spectra Normalized by the Area of Spectrum (RSNAS) and the Raman Spectra Normalized by Intensity Maximum (RSNIM) using the PCA parameters or the full spectrum.

|               | KNN (K=9) | J48 | RBF | MLP | RF |
|---------------|-----------|-----|-----|-----|----|
|               | PT NT ROC Area | PT NT ROC Area | PT NT ROC Area | PT NT ROC Area | PT NT ROC Area |
| RNAS          | Pathological Tissue (PT) | 38 2 0.9 | 30 10 0.764 | 29 11 0.822 | 34 6 0.89 | 32 8 0.913 |
|               | Normal Tissue (NT) | 6 34 0.9 | 10 30 0.764 | 6 34 0.804 | 6 34 0.89 | 6 34 0.913 |
| RSNAS (PCA)   | Pathological Tissue (PT) | 34 6 0.817 | 32 8 0.675 | 29 11 0.776 | 31 9 0.747 | 34 6 0.858 |
|               | Normal Tissue (NT) | 14 26 0.817 | 21 19 0.675 | 10 30 0.763 | 12 28 0.727 | 8 32 0.858 |
| RSNIM         | Pathological Tissue (PT) | 38 2 0.925 | 26 14 0.668 | 34 6 0.886 | 33 7 0.953 | 32 8 0.917 |
|               | Normal Tissue (NT) | 4 36 0.925 | 11 29 0.668 | 6 34 0.868 | 4 36 0.953 | 7 33 0.917 |
| RSNIM (PCA)   | Pathological Tissue (PT) | 35 5 0.862 | 31 9 0.786 | 30 10 0.787 | 29 11 0.841 | 35 5 0.925 |
|               | Normal Tissue (NT) | 8 32 0.862 | 10 30 0.786 | 8 32 0.789 | 6 34 0.841 | 7 33 0.925 |

The confusion matrices show the accuracy for the analysis of the full spectrum normalized by area was higher compared to that normalized by the intensity maximum, except for the MLP method. The classification performed by the MLP, KNN, and RF achieved the best values of sensitivity, specificity, accuracy and AUROC.

Excellent AUROC values were achieved using the KNN, MLP, RF classifiers for the full spectra, and the RF classifier using PCA parameters of RSNIM, while good AUROC values were presented for the KNN and RF (RSNAS) classifiers using the PCA parameters, and RBF.
for full spectra. This result suggests that the accuracy of diagnostic test may be higher when using the correct combination between parameters and classifiers. In addition, the PCA preprocessing may also reduce possible data overfitting and the complexity of analyses concerning many parameters of the whole spectrum.

Comparing the classification using the full spectrum or the PCA parameters (Table 3), RBF and RF classifiers show similar values of accuracy, suggesting that these classifiers could be less affected by a huge amount of samples analyzed. On the other hand, KNN, J48 and MLP accuracies decrease between 11.25% to 15%, respectively, when using the RSNAS with PCA parameters. This decrease may be associated to the information lost when using six PCs rather than the full spectrum parameters, as these PCs may describe a limited variety spectral features, despite the 80% variance contribution. However, this simpler description allows a faster analysis.

Another interesting result is, when using PCA parameters instead of the full spectrum ones for RSNIM instead of RSNAS, we observed a lower decrease between the accuracy for KNN and MLP, and also an increase for both decision trees (J48 and RF). This effect can also be observed on the ROC curves (Fig. 5), since the AUROC decreases for all the classifiers but J48 and RF, making the J48’s ROC curve comparable to the RBF one, and the RF’s AUROC greater than the KNN’s one when using the full spectra to analyze the data. This shows the potential of PCA to describe spectral characteristics responsible to the detection of oral pathologies, and to simultaneously increase the diagnosis speed. In addition, a global analysis of AUROC and the percentage of correctly classified instances for all classifiers shows the normalization by the intensity maximum of each spectrum is most appropriate (Table 3).

Table 3. Overall accuracy of the different statistical analyses used in this study

|                  | KNN (K = 9) | J48 | RBF  | MLP  | RF   |
|------------------|-------------|-----|------|------|------|
| RSNAS            | 90%         | 75% | 78.75%| 85%  | 82.5%|
| RSNAS (PCA)      | 75%         | 63.75%| 73.75%| 73.75%| 82.5%|
| RSNIM            | 92.5%       | 68.75%| 85%  | 86.25%| 81.25%|
| RSNIM (PCA)      | 83.75%      | 76.25%| 77.5%| 78.75%| 85%  |

In fact, the microscope Raman-based system is better in general, as signal/noise ratio, resolution, etc. However, our objective for use a fiber optic Raman-based system is for *in vivo*
clinical applications, and show the limitations of the study using this type of system. If we aim to use an optical biopsy system for clinical applications the microscope Raman-based system is completely out of focus.

The oral cancer is an aggressive disease, potentially fatal, originating from multiple factors and high rate of incidence. In the Brazilian context, it is the seventh among the most common malignancies. Around 90% of oral cancers instances refers to squamous carcinoma lesions [19]. However, there are some oral pathologies that precede the malignant neoplastic process of the mouth, which are called potentially precancerous lesions. Among them, we may highlight leukoplakia, erythroplakia, and actinic queilitis; all set out in the World Health Organization classification in 2005. Among the main factors related to the onset of the disease process of potentially precancerous lesions, tobacco stands out for oral intra oral lesions, and solar radiation, for extra oral lesions [20]. In general, the diagnosis is made by histopathology, which requires the execution of a biopsy - a surgical procedure, which comes with the full or partial removal of the lesion. The conventional histopathology technique provides morphological information of the tissue and may take days to complete a diagnosis. Recently, a range of new diagnosis tools based on photonic technologies are being developed. One of these tools is the optical or spectroscopic biopsy. Apart from speed and patient comfort advantages, optical biopsies may help guide surgeons during the excisions, decreasing the surgery time and providing conditions for a precise diagnosis.

The speed of diagnosis by Raman spectroscopy technique is an important factor to be considered over the currently used, for prognostic predictions. Furthermore, the applicability of the optical technique in real time in the near future will provide pathologists precise biochemical information’s about the analyzed material, thus ensuring greater accuracy in the diagnosis of many types of lesions. Recent publications involving Raman spectroscopy in vivo have been conducted with some promising results covering various anatomical areas, as for the diagnosis of cervical cancer [21], skin cancer [22] and also for the modelling thyroid cancer [23]. With the data to be obtained, it will be possible for educate patients in real-time risk of suffering a particular injury malignant transformation, or if it has inflammatory characteristics or even we face an early cancer. Depending on the spectral diagnosis professional can quickly take the patient, and in the short term lead to the same total cure of the disease. In the background, after the results we must extend our study to other types of cancers and disease processes and thus settle among health professionals (dentists, doctors, physiotherapists and nurses) in vivo Raman spectroscopy as a new and fast tool diagnostic. Considering all these advantages and potential, it becomes important to understand the shortcomings that has prevented its application in clinics. One of the reasons is clearly the lack of clear-cut diagnostic/ prognostic answers. The current study shows that a particular routine of preprocessin and analysis gives better results than others, making it possible to approach possibility of analysis standardization and automation.

4. Conclusions

The field of oral cancer diagnostics is open for new innovations. The large number of oral lesion types made the idea of real time, fast, label free and noninvasive techniques for diagnostics is a very attractive option. One of the obstacles in clinical translation is the complexity presented by spectral data. A routine automated software will help overcome this. Our study shows that different post-processing data methods could achieve distinct results, by using or not using PCA and two types of normalization can improve the sensitivity and specificity. Results of the study suggest that a combination between RSNAS parameters together MLP were better for discrimination of the tissues. A study with larger data set should be used to verify this. However, the paper presents the first step towards analysis automation.
Funding

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - 2014/05978-1 and FAPESP - 2014/16154-0); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – PNPD).

Acknowledgements

Luis Felipe CS Carvalho is funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - 2014/05978-1). Marcelo Saito Nogueira acknowledges the support provided by the Sao Paulo Research Foundation FAPESP: 2014/16154-0 (MS scholarship). Tanmoy T Bhattacharjee is funded by Programa Nacional de Pós-Doutorado (PNPD) Grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).