Anal Bioanal Chem (2015) 407:7935–7943
DOI 10.1007/s00216-015-8960-3

RESEARCH PAPER

Fourier-transform-infrared-spectroscopy based spectral-biomarker selection towards optimum diagnostic differentiation of oral leukoplakia and cancer

Satarupa Banerjee[1] · Mousumi Pal[2] · Jitamanyu Chakrabarty[3] · Cyril Petibois[4] · Ranjan Rashmi Paul[2] · Amita Giri[5] · Jyotirmoy Chatterjee[1]

Received: 31 May 2015 / Revised: 24 July 2015 / Accepted: 4 August 2015 / Published online: 5 September 2015

Abstract In search of specific label-free biomarkers for differentiation of two oral lesions, namely oral leukoplakia (OLK) and oral squamous-cell carcinoma (OSCC), Fourier-transform infrared (FTIR) spectroscopy was performed on paraffin-embedded tissue sections from 47 human subjects (eight normal (NOM), 16 OLK, and 23 OSCC). Difference between mean spectra (DBMS), Mann–Whitney's *U* test, and forward feature selection (FFS) techniques were used for optimising spectral-marker selection. Classification of diseases was performed with linear and quadratic support vector machine (SVM) at 10-fold cross-validation, using different combinations of spectral features. It was observed that six features obtained through FFS enabled differentiation of NOM and OSCC tissue (1782, 1713, 1665, 1545, 1409, and 1161 cm^{-1}) and were most significant, able to classify OLK and OSCC with 81.3 % sensitivity, 95.7 % specificity, and 89.7 % overall accuracy. The 43 spectral markers extracted through Mann–Whitney's *U* Test were the least significant when quadratic SVM was used. Considering the high sensitivity and specificity of the FFS technique, extracting only six spectral biomarkers was thus most useful for diagnosis of OLK and OSCC, and to overcome inter and intra-observer variability experienced in diagnostic best-practice histopathological procedure. By considering the biochemical assignment of these six spectral signatures, this work also revealed altered glycogen and keratin content in histological sections which could able to discriminate OLK and OSCC. The method was validated through spectral selection by the DBMS technique. Thus this method has potential for diagnostic cost minimisation for oral lesions by label-free biomarker identification.

Keywords FTIR · Oral leukoplakia · Oral squamous-cell carcinoma · Forward feature selection · Support vector machine

Electronic supplementary material The online version of this article (doi:10.1007/s00216-015-8960-3) contains supplementary material, which is available to authorized users.

✉ Satarupa Banerjee
satarupa@smst.iitkgp.ernet.in

[1] School of Medical Science and Technology, Indian Institute of Technology, Kharagpur 721302, India

[2] Department of Oral and Maxillofacial Pathology, Guru Nanak Institute of Dental Science and Research, 157/F Nilganj Road, Panihati, Kolkata 700 114, India

[3] Department of Chemistry, National Institute of Technology, Durgapur 713209, India

[4] University of Bordeaux – Inserm U1029 LAMC – Biophysics of Vascular Plasticity, 33608 Pessac, France

[5] Department of Pathology, North Bengal Medical College and Hospital, Darjeeling 734012, India

Introduction

Oral carcinogenesis is a complex multistep phenomenon. Its progression starts from benign hyperplasia and evolves through dysplasia, carcinoma in situ, and finally to oral squamous-cell carcinoma (OSCC) [1]. Oral pre-malignant disorders (PMDs) are believed to be an intermediate step of oral cancer development. Among many PMDs, oral leukoplakia (OLK) is the most prevalent, and histopathological evaluation of biopsies is the diagnostic method of choice [2]. The definition of OLK, as amended in the workshop of the WHO Collaborating Centre for Oral Cancer and Pre-cancer in 2005 by the working group, is: "The term leukoplakia should be used to recognise white plaques of questionable risk having excluded (other) known diseases or disorders that carry no

increased risk for cancer" [2]. Despite substantial development in molecular pathology, there is a lack of specific molecular markers predicting malignant potential. OLK is often associated with different grades of dysplastic changes, assessment of which is too difficult because of inter or intra-observer variability regarding the biopsy specimen. Often, clinically indicated OLK is diagnosed as OSCC on histopathology. Current research thus focuses on more selective and specific OLK diagnostic-marker identification. Recent studies used Raman spectroscopy for oral cancer and PMD diagnosis or to diagnose recurrence using oral ex-vivo tissues [3], but high data-acquisition time and low efficiency of inelastic light scattering are major limitations of the procedure regarding translation into the clinic [4].

As well as histology, a few molecular markers can be studied by molecular pathology techniques, but definite expression study by this method has not been proved sufficient to predict malignancy [5]. Hence, recent studies used vibrational-spectroscopy tools to document label-free markers to overcome limitations of molecular pathology, especially to record many types of molecular and/or sub-molecular information from the same tissue section. In this study, we analyzed the efficacy of Fourier-transform-infrared-spectroscopy (FTIR)-based optimum spectral-biomarker identification and diagnostic segregation of OLK and OSCC for minimising inter or intra-observer variability and increasing diagnostic sensitivity and specificity.

Identifying distinct spectral features extracted from FTIR spectra to indicate a characteristic disease is very difficult, because human tissue is composed of widely different molecular structures where overlapping of individual spectral peaks leads to formation of broader ones [6]. Hence, dimensionality reduction is an important challenge in these studies. With the advance of computational techniques, many methods including wrapper, filter, and embedded methods are also used for feature or spectral-biomarker selection [7]. Other efficient methods include peak-picking from second-derivative spectra, curve-fitting methods, multivariate classifications, etc. In this study, the efficiency of rather uncommon computational methods, namely the feature forward selection (FFS) approach and Mann–Whitney's *U* test, was compared with difference between mean spectra (DBMS), the most commonly used technique for feature extraction. The selected spectral markers were then subjected to support vector machine (SVM) and its variants for OSCC and OLK classification to assess their classification performance. FFS, a greedy wrapper algorithm for feature subset selection, uses a classifier to guide the addition of new features. Starting with an empty feature set, the feature that gives best classification is chosen. FFS next chooses the feature that gives best classification together with the feature already chosen, and the process continues for a specified number of times. A random 90 % of the data is used for training, and testing is performed using the remaining 10 % of the dataset [8]. SVM was used as the classifier for disease classification from the selected spectral markers because it obtained excellent classification accuracy in previous disease-classification studies [9]. Because a recent study of the potential of formalin-fixed paraffin-embedded (FFPE) tissue obtained excellent IR spectra when optimised analytical procedures to differentiate normal oral tissues with OSCC were used [10], we tried to differentiate clinically more closely related oral lesions in this study. The major challenge associated with FTIR in transmission mode is the scattering resulting from tissue surface inhomogeneity, which was minimised through spectral pre-processing [11].

Materials and method

Tissue collection

Transmission FTIR was performed on FFPE tissue sections from 47 patients (eight normal (NOM), 16 OLK, and 23 OSCC). Tissues were collected from the repository of Guru Nanak Institute of Dental Science and Research, Kolkata, India under ethical clearance of the institution ethical committee (GNIDSR/IEC/15-1 dt. 05/01/2015) after histopathological confirmation of the biopsy samples by oncopathologists using haematoxylin and eosin (H&E) staining. Normal oral mucosae were collected from the distobuccal aspect of the third molar teeth. The excess amount of mucoperiosteal buccal flaps left after transalveolar surgery was excised and used as normal. Therefore, collection of normal samples from subjects devoid of any disease was limited for ethical reasons [12].

FTIR study

The study was performed using a Nicolet 6700 spectrometer, Thermo Fisher, USA. For spectral data acquisition in transmission mode, acetone-treated dried deparaffinised unstained 4 μm thick tissue sections were used. The FFPE tissue sections were first deparaffinised using 10 min xylene treatment, and then dried using 5 min acetone treatment. Because after drying the tissue became completely dried and brittle, it was treated as powder during KBr pellet preparation, which was used as substrate. All the FTIR spectra were obtained in the range 400–4000 cm^{-1} at a resolution of 4 cm^{-1} with 32 scans. An 8 mm aperture diameter and DTGS detector were used during data acquisition.

Histological confirmation and histochemical staining

Staining of the samples was performed in 4 μm thick paraffin sections placed on four albumin-coated glass slides. The slides were then deparaffinised using xylene for 10 min. The section was stained with Harris' Hematoxylin and counter-stained

with eosin (H&E). Stained tissue sections representative of each disease condition are shown in Fig. 1a–c. Only Type II OLK or OLK with moderate dysplasia and well-differentiated OSCC were considered in this study, along with their normal counterpart, to avoid subtype-associated variation, because previous studies suggested that Type II OLK has the maximum chance of malignant transformation [13, 14]. The observed difference in the mean spectra of the diseases in the range 1200–1400 cm^{-1} was caused by glycogen and carbohydrate [13]. Therefore, the variation was validated by PAS (periodic acid–Schiff)-stained representative tissue sections. Deparaffinised tissue sections were first oxidised with 0.5 % periodic acid for 10 min, then stained with Schiff reagent for 5 min, and finally counter-stained with Harris' Hematoxylin. Stained tissue sections representative of each disease condition are shown in Fig. 1d–f.

Spectral-biomarker identification and disease classification

The 900–1800 cm^{-1} spectral interval, being regarded as the "fingerprint region", was chosen to reveal the difference between mean spectra [14]. Clinically it was found that OLK and OSCC are highly correlated diseases. To guard against bias in favour of atypical features, heterogeneity between groups of samples was validated through classification of the full spectral interval 4000–400 cm^{-1}, where no classification was found when classified using Ward's algorithm, point-by-point Euclidian distances, vector normalisation, and no compression or derivative spectra. The result is shown in the Electronic Supplementary Material (ESM) Fig. S1a. Again, when three regions for the best fit, 3000–2800 cm^{-1}, 1600–1480 cm^{-1}, and 1145–1081 cm^{-1}, were considered for classification in the same manner, only one normal sample was out of the main cluster (ESM Fig. S1b). This result suggested that simple classification was not accurate enough to separate healthy from pathological samples. This classification also revealed that statistical differences exist at the global level, and thus any further treatment, for example multivariate analysis, was supposed to reveal more subtle differences. Therefore, rubberband-like baseline correction (RBBC) preprocessing followed by maximum vector normalisation and mean centring of spectra in the range 900–1800 cm^{-1} followed by principal component analysis (PCA) was performed; when linear discriminant analysis (LDA) was performed, LDA scores plotted with a confidence ellipse representing a 70 % confidence interval had significant overlapping in OLK and OSCC cases, although NOM and OSCC were completely non-overlapping (ESM Fig. S2). Therefore the optimum spectral biomarker was sought with two other techniques, namely forward feature selection (FFS) as a multivariate wrapper method and Mann–Whitney's U test as a univariate filter method. During FFS-based feature extraction, the spectral interval of 900–1800 cm^{-1} was first standardised. First a feature histogram was generated through repeated FFS by randomly splitting the dataset 100 times into 90 %:10 % training–test sets, selecting five variables (spectral bands selected as wavenumbers) each time. Finally, six variables were suggested on the basis of the maximum number of hits of the variables [8]. FFS was also performed for two-class problems separately, (i) NOM vs. OLK, (ii) NOM vs. OSCC, and (iii) OLK vs. OSCC, and six spectral biomarkers were selected for each set. For Mann–Whitney's U test, $-\log_{10}(p\text{-value})$ was considered, rather than the p-value itself. This form is

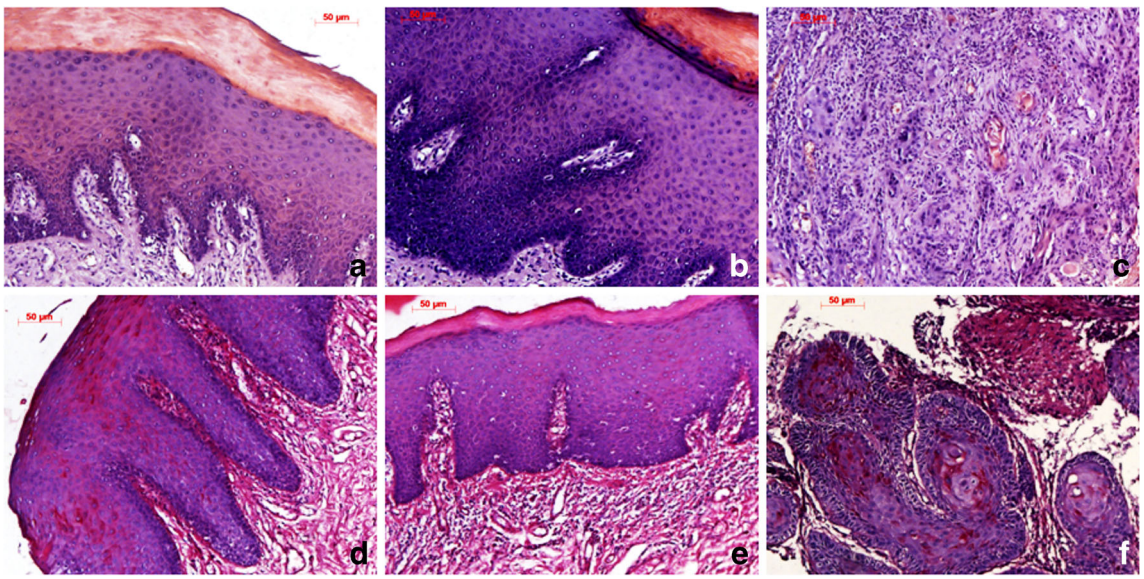

Fig. 1 H&E-stained tissue sections representative of each study condition (**a**) Keratinised normal tissue (NOM), (**b**) keratinised dysplastic epithelial oral tissue clinically featuring OLK, (**c**) keratinised dysplastic epithelial oral tissue clinically featuring OSCC, and PAS-stained tissue section of (**d**) NOM, (**e**) keratinised OLK, and (**f**) OSCC at 10× magnification

convenient because it transforms the *p*-value into a significance measure. The *U*-test, being non-parametric, is theoretically more appropriate than the *t*-test, because the probability distributions of the data variables may be skewed and/or bimodal, and the *t*-test assumes normal distribution. Finally, features selected by these two methods were classified using linear support vector machine (SVM) and quadratic SVM at 10-fold cross validation, followed by comparison of the results. Feature selection by the DBMS technique was performed using OMNIC9 software with 50 % sensitivity during peak finding. Spectral biomarker extraction by Mann–Whitney's *U* test and FFS was performed using "irootlab", a

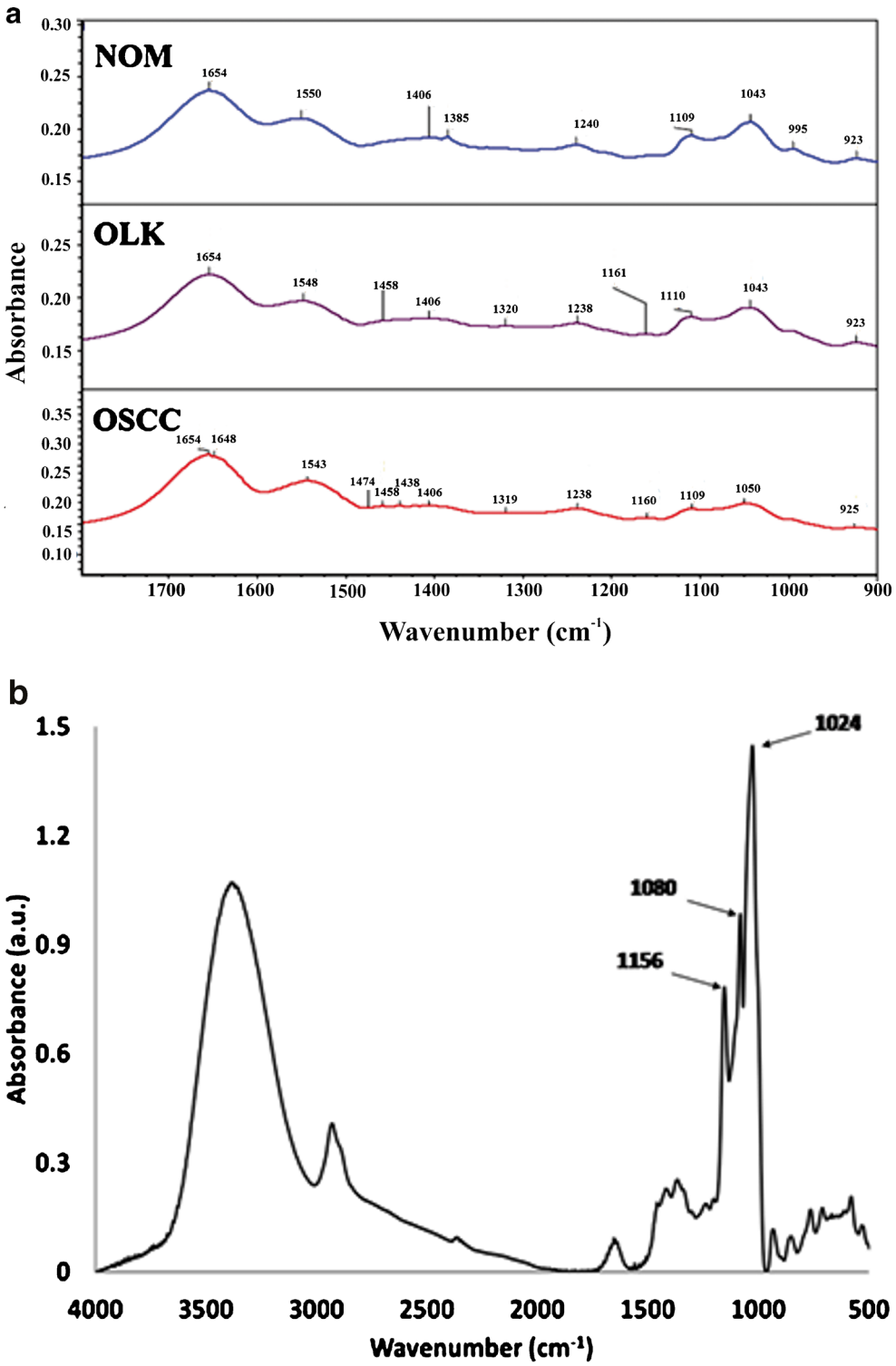

Fig. 2 (**a**) Difference between mean spectra (absorbance) in the range 900–1800 cm^{-1} for NOM, OLK, and OSCC tissues. (**b**) Typical spectra of glycogen (absorbance) in the range 400–4000 cm^{-1}

MATLAB toolbox used for vibrational spectroscopy in MATLAB R2012b (MathWorks, USA) [15]. Efficiency of disease classification was obtained using the "Classification Learner" app of MATLAB R2014b (MathWorks, USA). A schematic diagram of the method is provided in ESM Fig. S3.

Result and discussion

H&E-stained tissue sections representative of each disease condition, as shown in Fig. 1a–c, define the disease characteristics. In the NOM condition (Fig. 1a) finger-like rete pegs were present, whereas in the OLK condition (Fig. 1b) hyperplasia and moderate dysplasia were present. In the OSCC condition (Fig. 2c), epithelium and sub-epithelium could not be differentiated, and the presence of keratin pearl was evident. Selected spectral biomarkers, their corresponding feature-extraction technique, and commonly assigned biochemical components with their probable vibrational mode are presented in Table 1, and results of disease classification with sensitivity, specificity, and accuracy are presented in Table 2.

The difference between mean spectra is presented in Fig. 2a. The range 1000–1200 cm^{-1} is believed to be associated with polysaccharide and glycogen [13]. During feature selection for disease discrimination, four spectra extracted by DBMS and one extracted by FFS were found to be related to polysaccharide and glycogen content, as shown in Table 1.

Table 1 Spectral biomarkers extracted and their commonly assigned biochemical components

Wavenumber (cm^{-1})	Feature-extraction technique	Commonly assigned biochemical component	Vibrational mode	Ref.
1043	DBMS	Polysaccharides, glycogen and/or glucose, nucleic acid in absence of glycogen	γ (PO_2^-) in RNA, DNA	[13, 16, 17]
1050*	DBMS	Polysaccharides, glycogen	γs (C–O–O–C), γ (C–O) coupled with δ (C–O), of C–OH of carbohydrates	[13, 17]
1088*	FFS	Protein phosphorylation, DNA, phospholipids, phosphate I in B form DNA if not polysaccharide, glycogen	γs (PO_2^-)	[17–19]
1160, 1161	DBMS	Polysaccharides, RNA ribose	γ (C–C), δ (C–OH), γ (CO)	[13, 18, 19]
1161	FFS	Protein phosphorylation	γ (C–C), δ (C–OH), γ (CO)	[19]
1319, 1320	DBMS	Fatty acids and amino acids, collagen	–	[20, 21]
1385	FFS	Protein phosphorylation,	–	[15]
1385	DBMS	Phospholipid and fatty acyl chains triglyceride	δs (CH_3)	[15, 19]
1409	FFS	Protein phosphorylation	–	[15]
1438	DBMS	Keratin	–	[20]
1458	DBMS	Collagen	δas (CH_3)	[17, 19]
1474	DBMS	Keratin	δas (CH_2)	[20]
1543	DBMS	Amide II	60 % δ (N–H), 30 % γ (C–N), 10 % γ (C–C)	[13, 19]
1545	FFS	Amide II	δ (N–H), γ (C–N)	[15]
1548	DBMS	Amide II	60 % δ (N–H), 30 % γ (C–N), 10 % γ (C–C)	[13, 19]
1550	DBMS	Amide II	60 % δ (N–H), 30 % γ (C–N), 10 % γ (C–C)	[13, 19]
1628	FFS	Amide I	70–85 % γ (CO), 10–20 % (CN)	[22]
1638–1680	Mann–Whitney's *U* test	Amide I	70–85 % γ (CO), 10–20 % (CN)	[19, 22]
1648	DBMS	Amide I	γ (HOH), 70–85 % γ (CO), 10–20 % (CN)	[13, 19, 22]
1662	FFS	Amide I	70–85 % γ (CO), 10–20 % (CN)	[22]
1665	FFS	Amide I	70–85 % γ (CO), 10–20 % (CN)	[22]
1704	FFS	Fatty acid esters, lipid	–	[17]
1713	FFS	Fatty acid esters, lipid	–	[17]
1775	FFS	Unassigned	–	–
1782	FFS	Unassigned	–	–

γ = stretching, δ = bending, s = symmetric, as = asymmetric, *1080 and 1156 cm^{-1} found on glycogen pure spectrum in Fig. 2b; the shifts result from the absorption from other saccharidic contents

The changes resulting from alteration in carbohydrate and glycogen content within the tissue sections were clearly visible in representative PAS sections (Fig. 1d–f). In NOM epithelium there was glycogen expression in the suprabasal region, which was not present in OLK. In OSCC, glycogen expression was visible in the focal region with a characteristic magenta colour. A typical spectrum of glycogen, revealing absorbance in the range 400–4000 cm^{-1}, is provided in Fig. 2b to validate the concept. The peak present at 1043 cm^{-1} in normal oral mucosa and OLK mean spectra was shifted to 1050 cm^{-1}, and again there was a peak at 1109 cm^{-1} which was found to reduce gradually in diseased cases. This result was in agreement with a previous study [20], and was validated by representative PAS-stained tissue sections (Fig. 2d–f). Previous studies proved that keratinisation is always preceded by glycogen accumulation, but interestingly glycogen content is increased in OLK when keratinisation is reduced [23]. This was correlated with an increase of specific absorptions of glycogen in spectra (bands centred at 1024, 1080, and 1156 cm^{-1} giving rise to global absorptions in the 1000–1200 cm^{-1} spectral interval of biological tissue spectra), as presented in Fig. 2b. Another study suggested that diminished glycogen concentration in OLK compared with normal buccal mucosa is suggestive of morphological changes [21]. This suggestion was validated histochemically in this study by the PAS-stained images.

The effect of glycogen in keratinisation is quite interesting. In one method glycogen was revealed to be the energy provider for keratinisation, whereas another study suggested glycogen to be the precursor of oleic acid, which is used for keratin synthesis [24]. Another study revealed that alteration of glycogen metabolism and phosphorylation of related proteins were probably the underlying causes of keratinisation in oral mucosa [25]. It was found that in non-keratinised mucosa there was a substantial amount of intercellular glycogen expression, but in OLK (Fig. 1e), with increase in the extent of keratinisation, intracellular glycogen content was reduced compared with that of the NOM (Fig. 1d), but the keratinised area remained positively stained. This might be caused by the glycosylated keratin filaments present in human keratinocyte [26]. Because the glycosylation is associated with keratan sulfate, increased expression of sulfur in PMDs can be associated substantially with keratinisation [27]. Another study revealed the presence of glycoproteins in keratin fibres [28], and the effect of glycogen and keratin was found to be interesting in OLK-mediated oral carcinogenesis, as illustrated by the spectral and histochemical results and by previous studies, because PAS positivity substantiates tissue glycogen and glycoprotein expression [29]. Increased glycogen was also found to be associated with increased cellular differentiation in carcinoma and an important property for disease classification [30]. The concept of such an association has been presented in and validated through histochemical and spectrochemical signatures towards label-free marker identification for glycogen. Therefore it can be deduced that during spectral differentiation of OLK and OSCC, the region of glycogen can be taken into account with the amide I region, as evident from Fig. 2.

During analysis of DBMS spectra, the highest intensity moved from 923 cm^{-1} to 925 cm^{-1}, revealing that the most

Table 2 Results of OLK and OSCC classification by linear and quadratic SVM showing sensitivity, specificity, and accuracy

Feature-extraction technique	Selected wave numbers	Type of SVM used	Sensitivity (%)	Specificity (%)	Accuracy (%)
DBMS	1043, 1050, 1050, 1061, 1319, 1320, 1385, 1438, 1474, 1543, 1548, 1550, and 1648 cm^{-1}	Linear	75	82.6	79.5
		Quadratic	**81.3**	**91.3**	**87.2**
FFS	All 18 features selected through FFS	Linear	68.8	82.6	76.9
		Quadratic	68.8	91.3	82.1
FFS	Features selected for NOM vs. OLK and OLK vs. OSCC delineation	Linear	68.8	78.3	74.4
		Quadratic	62.5	91.3	79.5
FFS	Features selected for NOM vs. OLK and NOM vs. OSCC delineation	Linear	75	78.3	76.9
		Quadratic	75	91.3	84.6
FFS	Features selected for OLK vs. OSCC and OLK vs. OSCC delineation	Linear	75	87	82.1
		Quadratic	68.8	91.3	82.1
FFS	Features selected for NOM vs. OLK delineation (1628, 1385, 1088, 1775, 1704, and 1662 cm^{-1})	Linear	68.8	82.6	76.9
		Quadratic	68.8	91.3	82.1
FFS	Features selected for OLK vs. OSCC delineation (1032, 956, 1707, 1639, 1606, and 1565 cm^{-1})	Linear	68.8	78.3	74.4
		Quadratic	68.8	91.3	82.1
FFS	Features selected for NOM vs. OSCC delineation (1782, 1713, 1665, 1545, 1409, and 1161 cm^{-1})	Linear	68.8	91.3	82.1
		Quadratic	**81.3**	**95.7**	**89.7**
Mann–Whitney's *U* test	1638–1680 cm^{-1}	Linear	50	65.2	59
		Quadratic	31.3	78.3	59

intense contributing peak had changed in the spectral range in OLK and OSCC, possibly caused by transformation into left-handed-helix DNA (Z form) through oxidative DNA damage [17] caused by damage of sugar and base moieties [31]. A previous study confirmed that conformation change into Z-DNA leads to genetic instability in such diseases as cancer [32]. In NOM, the peak at 995 cm^{-1} resulting from ring breathing [17] was found to be lost in cases of disease.

In the normal condition there was a peak at 1385 cm^{-1}, which moved to 1405 cm^{-1} for OLK and 1406 cm^{-1} for OSCC. Again in the normal condition, there was a peak at 1550 cm^{-1}, which moved to 1548 cm^{-1} for OLK and to 1543 cm^{-1} for OSCC. The peak at 1238 cm^{-1} in OSCC is in agreement with another study, which was also found to be present in OLK [10]. Two peaks at 1438 cm^{-1} and 1474 cm^{-1} in OSCC may be caused by alteration in keratin, as suggested by Fukuyama et al. [20]. Alteration in keratin expression is evident in both Figs. 1a and 2d–f; both H&E staining and PAS staining indicate keratin expression [29]. There was increased keratin deposition above the epithelium in OLK, and in OSCC distinct keratin pearls are observed in Fig. 2. Therefore, these spectra obtained through FFS can be regarded as a label-free biomarker for differential OLK and OSCC classification. When Mann–Whitney's *U* test was performed, 1638–1680 cm^{-1} was found to be significant with $-\log_{10}(p\text{-value})$ less than 1. This region is representative of the amide I band. Overlapping of broad underlying component bands in this region leads to difficulty in specific band assignment of proteins, other than distinctive absorbance maxima [33]. During FFS, 18 spectral wave numbers were extracted from the region 900–1800 cm^{-1}, six each to differentiate NOM vs. OLK (1628, 1385, 1088, 1775, 1704, and 1662 cm^{-1}), NOM vs. OSCC (1782, 1713, 1665, 1545, 1409, and 1161 cm^{-1}), and OLK vs. OSCC (1032, 956, 1707, 1639, 1606, and 1565 cm^{-1}). The chemistry behind the alteration of spectral shift or presence of exclusive spectral bands is depicted in Table 1. Five spectral markers were common in at least two feature-selection techniques. Peaks at 1161 and 1385 cm^{-1} were obtained with both DBMS and FFS, whereas peaks at 1648, 1662, and 1665 cm^{-1} were common in DBMS and Mann–Whitney's *U* test.

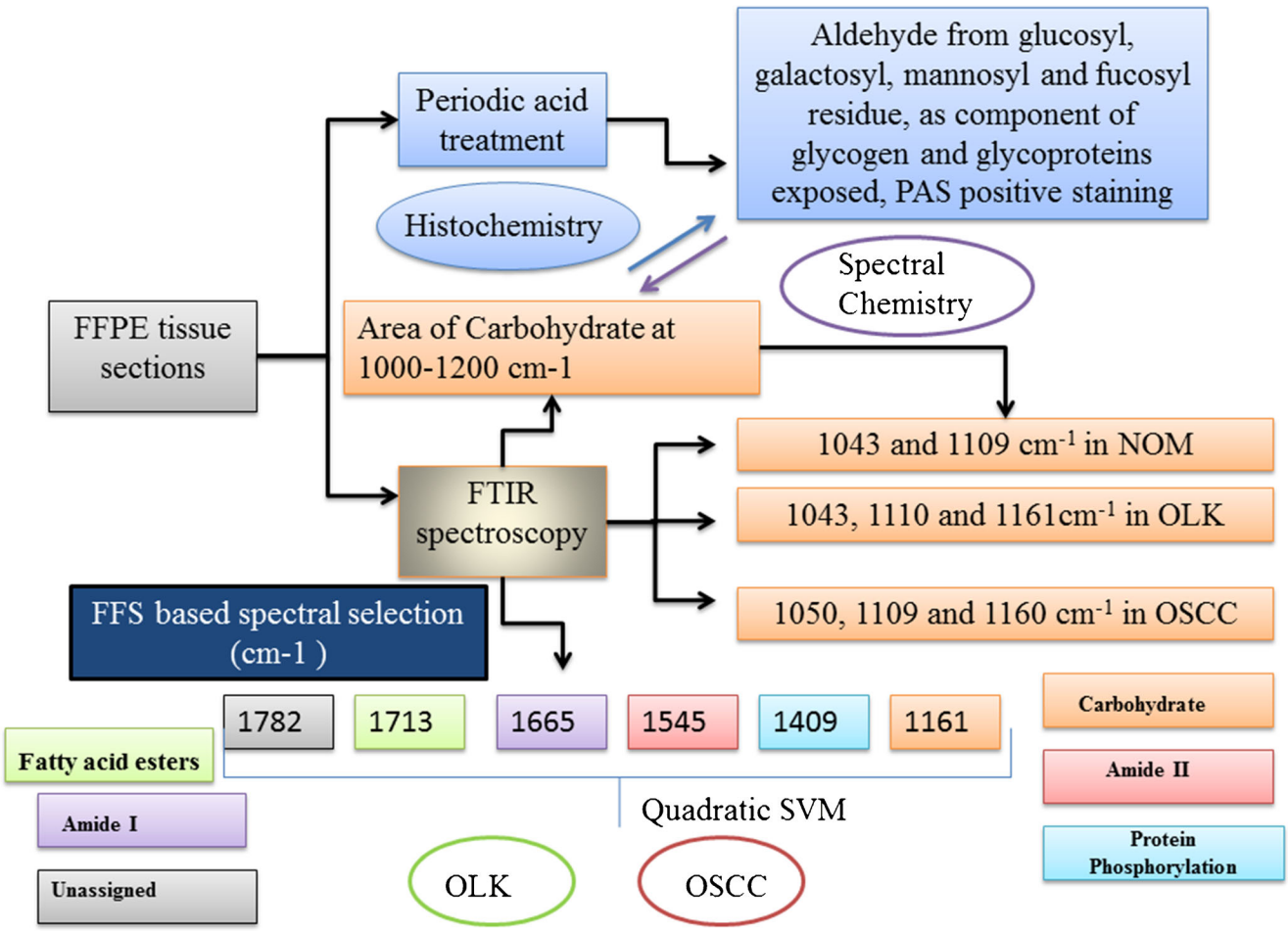

Fig. 3 Schematic diagram depicting connection between histochemical and spectrochemical signatures, with potential for label-free-marker identification and diagnostic differentiation of OLK and OSCC

The diseases were classified with linear and quadratic SVM using different combinations of spectra, and it is evident from Table 2 that quadratic SVM was more efficient than the linear SVM model. Six features obtained through FFS to differentiate NOM and OSCC were the most significant, with 81.3 % sensitivity, 95.7 % specificity, and 89.7 % overall accuracy, whereas 43 features extracted through Mann–Whitney's *U* Test were the least significant (50 % sensitivity and 65.2 % specificity) (Table 2). The extracted features obtained through DBMS were also found to be highly significant, with 81.3 % sensitivity, 91.3 % specificity, and 87.2 % overall accuracy to differentiate OLK and OSCC. Hence, features selected through FFS for NOM and OSCC differentiation and through DBMS could be regarded as important spectral biomarkers for classification of highly correlated OLK and OSCC, and had translational value and has been highlighted in Table 2. Because in the latter case the number of features was higher and sensitivity was lower, it can be concluded that the six spectral biomarkers 1782, 1713, 1665, 1545, 1409, and 1161 cm^{-1} can be used for objective diagnosis of OLK and OSCC, and for removing ambiguity associated with inter and intra-observer variability. Structural lipids remained unaltered during tissue processing, and were also found to be important in disease differentiation, along with the other proteins. Beyond 1750 cm^{-1} it is assumed that no more relevant absorption can be found, but, interestingly, a theoretical selection of spectra at 1782 cm^{-1} augmented classification efficiency. It can be observed from Fig. 3 that during DBMS spectra selection the region of carbohydrate and glycogen is be useful, whereas the FFS technique revealed the involvement of fatty acid esters, amide I and II regions, and protein phosphorylation in OLK and OSCC differentiation.

In the context of augmenting precise clinical decision making to differentiate moderately dysplastic OLK and OSCC conditions, this spectroscopic study aided with computational analytics was found to be successful for identification of discriminatory spectral markers for these diseases. These markers, being associated with important bio-molecular changes at qualitative and quantitative levels, were also found to be effective for classification of the diseases with high sensitivity and specificity. This work also suggests FTIR spectroscopy as a reagent-free and observer-independent method for glycogen and keratin assessment in OLK and OSCC histology. Thus this study not only supports FTIR-based label-free-marker development endeavours in cancer research, but also reveals clinical potential for disease classification and assessing malignant potential of oral lesions.

Acknowledgment The authors would like to acknowledge financial finding from MHRD, Government of India, New Delhi (IIT/SRIC/SMST/IAN/2013-14/222). The authors also wish to thank Mr B. Mohan Rao for his help during data acquisition and the anonymous reviewers for their suggestions.

Conflict of interest The authors declare no conflict of interest.

Statement on informed consent We would like to state that the work was performed under ethical clearance of the institution ethical committee of GNIDSR, Kolkata (GNIDSR/IEC/15-1 dt. 05/01/2015) and informed consent was obtained from all the subjects (both normal and diseased) recruited in the study.

References

1. AbdulMajeed AA, Farah CS (2013) Can immunohistochemistry serve as an alternative to subjective histopathological diagnosis of oral epithelial dysplasia? Biomark Cancer 5:49–60. doi:10.4137/BIC.S12951
2. van der Waal I (2009) Potentially malignant disorders of the oral and oropharyngeal mucosa; terminology, classification and present concepts of management. Oral Oncol 45(4–5):317–323. doi:10.1016/j.oraloncology.2008.05.016
3. Singh SP, Deshmukh A, Chaturvedi P, Murali Krishna C (2012) In vivo Raman spectroscopic identification of premalignant lesions in oral buccal mucosa. J Biomed Opt 17(10):105002. doi:10.1117/1.jbo.17.10.105002
4. Kong K, Kendall C, Stone N, Notingher I Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection. Adv Drug Deliv Rev. doi:10.1016/j.addr.2015.03.009
5. Reibel J (2003) Prognosis of oral pre-malignant lesions: significance of clinical, histopathological, and molecular biological characteristics. Crit Rev Oral Biol Med 14(1):47–62
6. Lee S, Kim K, Lee H, Jun CH, Chung H, Park JJ (2013) Improving the classification accuracy for IR spectroscopic diagnosis of stomach and colon malignancy using non-linear spectral feature extraction methods. Analyst 138(14):4076–4082. doi:10.1039/c3an00256j
7. Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL (2012) Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. Analyst 137(14):3202–3215
8. Trevisan J, Park J, Angelov PP, Ahmadzai AA, Gajjar K, Scott AD, Carmichael PL, Martin FL (2014) Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy. J Biophotonics 7(3–4):254–265. doi:10.1002/jbio.201300190
9. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 10(1):16
10. Pallua JD, Pezzei C, Zelger B, Schaefer G, Bittner LK, Huck-Pezzei VA, Schoenbichler SA, Hahn H, Kloss-Brandstaetter A, Kloss F, Bonn GK, Huck CW (2012) Fourier transform infrared imaging analysis in discrimination studies of squamous cell carcinoma. Analyst 137(17):3965–3974. doi:10.1039/c2an35483g
11. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, Fielden PR, Fogarty SW, Fullwood NJ, Heys KA, Hughes C, Lasch P, Martin-Hirsch PL, Obinaju B, Sockalingum GD, Sulé-Suso J, Strong RJ, Walsh MJ, Wood BR, Gardner P, Martin FL (2014) Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc 9(8):1771–1791. doi:10.1038/nprot.2014.110, **http://www.nature.com/nprot/journal/v9/n8/abs/nprot.2014.110.html#supplementary-information**
12. Anura A, Conjeti S, Das RK, Pal M, Paul RR, Bag S, Ray AK, Chatterjee J (2015) Computer-aided molecular pathology

interpretation in exploring prospective markers for oral submucous fibrosis progression. Head Neck. doi:10.1002/hed.23962
13. Krishna CM, Sockalingum GD, Bhat RA, Venteo L, Kushtagi P, Pluot M, Manfait M (2007) FTIR and Raman microspectroscopy of normal, benign, and malignant formalin-fixed ovarian tissues. Anal Bioanal Chem 387(5):1649–1656. doi:10.1007/s00216-006-0827-1
14. Gajjar K, Heppenstall LD, Pang W, Ashton KM, Trevisan J, Patel II, Llabjani V, Stringfellow HF, Martin-Hirsch PL, Dawson T, Martin FL (2012) Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis. Anal Methods Adv Methods Appl 5:89–102. doi:10.1039/C2AY25544H
15. Trevisan J, Angelov PP, Scott AD, Carmichael PL, Martin FL (2013) IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis. Bioinformatics (Oxford, England) 29(8):1095–1097. doi:10.1093/bioinformatics/btt084
16. Zohdi V, Whelan DR, Wood BR, Pearson JT, Bambery KR, Black MJ (2015) Importance of tissue preparation methods in FTIR micro-spectroscopical analysis of biological tissues: 'Traps for new users'. PLoS One 10(2), e0116491. doi:10.1371/journal.pone.0116491
17. Movasaghi Z, Rehman S, ur Rehman DI (2008) Fourier transform infrared (FTIR) spectroscopy of biological tissues. Appl Spectrosc Rev 43(2):134–179
18. Fujioka N, Morimoto Y, Arai T, Kikuchi M (2004) Discrimination between normal and malignant human gastric tissues by Fourier transform infrared spectroscopy. Cancer Detect Prev 28(1):32–36. doi:10.1016/j.cdp.2003.11.004
19. Bellisola G, Sorio C (2012) Infrared spectroscopy and microscopy in cancer research and diagnosis. Am J Cancer Res 2(1):1–21
20. Fukuyama Y, Yoshida S, Yanagisawa S, Shimizu M (1999) A study on the differences between oral squamous cell carcinomas and normal oral mucosas measured by Fourier transform infrared spectroscopy. Biospectroscopy 5(2):117–126. doi:10.1002/(sici)1520-6343(1999)5:2<117::aid-bspy5>3.0.co;2-k
21. Weisberger D, Fischer CJ (1960) Glycogen content of human normal buccal mucosa and buccal leukoplakia. Ann N Y Acad Sci 85(1):349–350
22. Sun QJ, Xiong L, Bu XH, Liu Y (2012) Study on mechanism of cross-linking of peanut protein isolate modified with transglutaminase. Adv Mater Res 550:1304–1308
23. Doyle JL, Manhold JH, Weisinger E (1968) Study of glycogen content and "basement membrane" in benign and malignant oral lesions. Oral Surg Oral Med Oral Pathol 26(5):667–673
24. Steiner K (1955) A histochemical study of epidermal glycogen in skin diseases. J Invest Dermatol 24(6):599–618
25. Goltz RW, Fusaro RM, Jarvis J (1958) Observations on glycogen in epithelial tumors1. J Invest Dermatol 31(6):331–341
26. Schafer IA, Sorrell JM (1993) Human keratinocytes contain keratin filaments that are glycosylated with keratan sulfate. Exp Cell Res 207(2):213–219. doi:10.1006/excr.1993.1185
27. Paul RR, Chatterjee J, Das AK, Cervera ML, de la Guardia M, Chaudhuri K (2002) Altered elemental profile as indicator of homeostatic imbalance in pathogenesis of oral submucous fibrosis. Biol Trace Elem Res 87(1–3):45–56
28. Allen AK, Ellis J, Rivett DE (1991) The presence of glycoproteins in the cell membrane complex of a variety of keratin fibres. Biochim Biophys Acta Gen Subj 1074(2):331–333. doi:10.1016/0304-4165(91)90172-D
29. Kiernan JA (1999) Histological and histochemical methods: theory and practice. Shock 12(6):479
30. Yano K, Ohoshima S, Shimizu Y, Moriguchi T, Katayama H (1996) Evaluation of glycogen level in human lung carcinoma tissues by an infrared spectroscopic method. Cancer Lett 110(1–2):29–34
31. Cooke MS, Evans MD, Dizdaroglu M, Lunec J (2003) Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J 17(10):1195–1214. doi:10.1096/fj.02-0752rev
32. May Promote Z-DNA, Damage DNA (2006) Cancer Biol Ther 5(3):253. doi:10.4161/cbt.5.3.2593
33. Kong J, Yu S (2007) Fourier transform infrared spectroscopic analysis of protein secondary structures. Acta Biochim Biophys Sin 39(8):549–559