Journal of Molecular Structure 1163 (2018) 472–479

Contents lists available at ScienceDirect

Journal of Molecular Structure

journal homepage: http://www.elsevier.com/locate/molstruc

A simple model for cell type recognition using 2D-correlation analysis of FTIR images from breast cancer tissue

Mohamed H. Ali [a], Fazle Rakib [b], Khalid Al-Saad [b], Rafif Al-Saady [c], Fiona M. Lyng [d], Erik Goormaghtigh [e, *]

[a] Qatar Biomedical Research Institute (QBRI), Hamad Bin Khalifa University (HBKU), Doha, Qatar
[b] Department of Chemistry and Earth Sciences, Qatar University, Doha, Qatar
[c] Pathology and Laboratory Medicine, Al Ahli Hospital, Doha, Qatar
[d] DIT Centre for Radiation and Environmental Science, FOCAS Research Institute, Dublin Institute of Technology, Ireland
[e] Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine CP206/02, Université Libre de Bruxelles CP206/2, B1050, Brussels, Belgium

ARTICLE INFO

Article history:
Available online 13 March 2018

Keywords:
Infrared imaging
Spectral histopathology
Infrared spectroscopy
2D correlation spectroscopy
Breast cancer

ABSTRACT

Breast cancer is the second most common cancer after lung cancer. So far, in clinical practice, most cancer parameters originating from histopathology rely on the visualization by a pathologist of microscopic structures observed in stained tissue sections, including immunohistochemistry markers. Fourier transform infrared spectroscopy (FTIR) spectroscopy provides a biochemical fingerprint of a biopsy sample and, together with advanced data analysis techniques, can accurately classify cell types. Yet, one of the challenges when dealing with FTIR imaging is the slow recording of the data. One cm^2 tissue section requires several hours of image recording. We show in the present paper that 2D covariance analysis singles out only a few wavenumbers where both variance and covariance are large. Simple models could be built using 4 wavenumbers to identify the 4 main cell types present in breast cancer tissue sections. Decision trees provide particularly simple models to reach discrimination between the 4 cell types. The robustness of these simple decision-tree models were challenged with FTIR spectral data obtained using different recording conditions. One test set was recorded by transflection on tissue sections in the presence of paraffin while the training set was obtained on dewaxed tissue sections by transmission. Furthermore, the test set was collected with a different brand of FTIR microscope and a different pixel size. Despite the different recording conditions, separating extracellular matrix (ECM) from carcinoma spectra was 100% successful, underlying the robustness of this univariate model and the utility of covariance analysis for revealing efficient wavenumbers. We suggest that 2D covariance maps using the full spectral range could be most useful to select the interesting wavenumbers and achieve very fast data acquisition on quantum cascade laser infrared imaging microscopes.

1. Introduction

Breast cancer is the second most common cancer after lung cancer. Its incidence is still growing (by 0.4% during the last 5 years) in the US [1] but increased by 40% among the Gulf Cooperation Council Countries women during the last 12-year period analyzed [2]. The precise identification of tumor molecular parameters and their integration in cancer management is crucial for improving cancer therapy as well as for moving towards individual therapy. So far, in clinical practice, most cancer parameters originating from histopathology rely on the visualization by a pathologist of microscopic structures observed in stained tissue sections, including immunohistochemistry (IHC) markers (HR, HER2, Ki-67and/or basal cell markers such as CK5/6 and EGFR) that are largely used in the clinic [4]. In practice, the classifications obtained using these markers help, statistically speaking, define a therapy and provide a prognosis. At individual level however, they are largely insufficient to deal with individual cases. The origin of this failure can be largely

* Corresponding author. Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine CP206/02, Université Libre de Bruxelles, Bld du Triomphe 2, CP206/2, B1050, Brussels, Belgium.
E-mail address: egoor@ulb.ac.be (E. Goormaghtigh).

https://doi.org/10.1016/j.molstruc.2018.03.044

found in the huge heterogeneity of the tumors. Tumor cells mutate at a high rate and a single tumor may contain many different clones [5,6] with different properties which are difficult to identify. A further complexity arises from the interplay between tumor cells and the microenvironment. Finally, the distinct molecular characteristics of the tumors that can have profound implications on clinical behaviors and long-term patient outcomes [7,8] are not fully considered. The result is that most of these parameters, e.g. the histological grade, are associated with inter- and intra-observer discrepancies, in addition to suffering from a quantification problem.

Addressing the heterogeneity of the tumor at microscopic level is not a simple problem. It requires both in depth analysis of the cells contained in the tissue section and a microscopic approach which can provide such an analysis at the cellular level. As far as in-depth analysis is concerned, the development of microarray-based gene expression profiling technologies has helped explain some of this heterogeneity, in particular by sub-classifying human breast carcinomas and establishing molecular-based prognostic and predictive signatures [4,9–12]. The complexity and cost of gene expression profiling limit its use as a routine hospital diagnostic tool. Furthermore, it cannot be applied systematically at the cellular level. This results in an incomplete picture of the heterogeneity of the tumor [13]. Reflecting this cellular heterogeneity therefore requires new approaches.

Some emerging imaging technologies (e.g., mass spectrometry-based "omics" techniques [14,15]) reveal many more biomarkers than immunohistochemical approaches [16,17]. They are promising but not yet robust enough to ensure high-throughput analysis of histological sections in routine pathology. Presently, vibrational spectral histopathology (FTIR or Raman imaging) represents the only method that provides simultaneously, for each pixel of the acquired image, hundreds of robust markers related to the molecular content of the cells [18,19]. Fourier transform infrared (FTIR) spectroscopy has recently shown great potential for disease diagnosis in the field of breast cancer [20–22]. These techniques can provide a biochemical fingerprint of a biopsy sample and, together with advanced data analysis techniques, can accurately classify cell types. Each cellular component has a characteristic set of vibrational transitions resulting in a unique spectrum. It is now considered that vibrational spectroscopies provide as much information as DNA microarrays as far as diagnostic purposes are concerned. Importantly, they can also identify all molecule types as well as details of their chemical structure. For instance the length of lipid acyl chains, the degree of unsaturation of lipids [23], the lipid/protein ratio, DNA condensation state [24] and many other parameters can be obtained from vibrational spectra. In particular, information not only on the chemical nature of cell molecules but also their conformations can be obtained. They are, in particular very sensitive to protein secondary structure [25,26]. Altogether, the various contributions to the spectrum form a signature of the molecular composition of the cell that is unique. Multivariate analysis of spectral data results in remarkably sensitive segmentation of the cells present in a tissue section [27]. Recent progresses in FTIR imaging makes it possible to record images of tissue sections with a spatial resolution close to the cell size. Yet, one of the challenges when dealing with FTIR imaging is the still slow recording and handling of data [19]. When recording FTIR images with a pixel size of $2.7 \times 2.7\ \mu m^2$ as described in this work, a $1\ cm^2$ image contains 13.4 million spectra which are recorded between 4000 and 900 cm^{-1} for most Focal Plane Array (FPA) detectors, each one containing about 1500 data points. Recording such an image takes several hours, limiting the application of the technology. Yet, not all of this data is useful for identifying cell types. Recording FTIR images with quantum cascade lasers [28–30] and their analysis would be considerably faster if the useful spectral region were known. For this purpose, analysis of variance is a key issue. Spectral regions with no variance throughout the dataset are obviously of no interest and the co-varying wavenumbers could be reduced to a limited number of wavenumbers as they describe the same variations among cell types.

We propose in the present paper to use 2D correlation analysis to simplify the dataset and possibly identify the few wavenumbers that are of interest for the identification of the various cell types such as carcinoma cells, erythrocytes, lymphocytes and the extracellular matrix (ECM). In the near future, this first approach will have to be followed by an attempt to identify various types of tumor clones in a tissue section.

2. Materials and methods

Spectra of breast carcinoma cells, erythrocytes, lymphocytes and extracellular matrix were obtained from a database maintained in-house. A full description of the samples can be found in Benard et al. [27]. Briefly formalin-fixed paraffin-embedded (FFPE) breast cancer tissues from 66 patients were provided from the tumor tissue bank of the Jules Bordet Institute (Brussels, Belgium). To enrich the database on the immune and stromal responses, seven lymph nodes and three tonsils as well as seven scars from mastectomy biopsied tissue samples were also included. As described in the previous paper [27], more than 13,000 representative spectra were extracted from the images under the supervision of a trained pathologist.

As described previously [27], the FTIR data were collected using a Hyperion 3000 FTIR imaging system (Bruker Optics, Ettlingen, Germany), equipped with a 64*64 Mercury Cadmium Telluride (MCT) Focal Plane Array (FPA) detector. Data were collected in transmission mode from sample regions of 184*184 μm^2. Each pixel corresponds to an area of 2.7*2.7 μm^2. One FTIR image (unit image) resulted in 4096 spectra, each one being the average of 256 scans recorded in a spectral range from 3900 to 800 cm^{-1}. The spectral resolution was set to 8 cm^{-1} and data points encoded every 1 cm^{-1}. Spectral processing was performed as described [27]. Briefly, water vapor contribution was subtracted with 1956-1935 cm^{-1} as the reference peak. In order to eliminate any intensity variation caused by changes in the thickness of the tissue section or quantity of cellular material, the spectra were normalized for equal area between 1725 and 1481 cm^{-1}. An 11-point baseline correction was subtracted. For this purpose, straight lines were interpolated between the spectral points at 3620, 2995, 2800, 2395, 2247, 1765, 1724, 1480, 1355, 1144 and 950 cm^{-1} and subtracted from each spectrum. Pre-processed spectra were retained for further analyses when the Signal-to-Noise ratio (S/N) was greater than 300:1 for the Amide I and II region (from 1750 to 1480 cm^{-1}). This ratio was calculated considering Signal as the maximum absorbance within 1750-1480 cm^{-1} spectra range and Noise as the standard deviation within 2200-2100 cm^{-1} range.

For validation, a new set of samples was obtained from the Cancer Registry Office in Al-Amal Hospital, Hamad Medical Corporation (HMC), Doha, Qatar, cut as 5 μm thick sections and imaged without removing paraffin. FTIR data were collected using an Agilent 128×128 focal plane array (FPA) mid-IR imager. No binning was applied. Spectra were collected between 3950 and 900 cm^{-1} at a nominal resolution of 8 cm^{-1} and encoded every 2 cm^{-1}. Each spectrum was the mean of 64 scans. The microscope was equipped with a liquid nitrogen cooled 128×128 Mercury Cadmium Telluride (MCT) Focal Plane Array (FPA) detector and a $15\times$ objective (NA = 0.62). Each pixel corresponds to an area $5.5 \times 5.5\ \mu m^2$. The data were collected in transflexion mode from sample regions of $700 \times 700\ \mu m^2$. One FTIR image (unit image) resulted in 16,384