IDENTIFICATION OF SPECTRAL MARKERS IN FOURIER TRANSFORM INFRARED SPECTRA OF BIOLOGICAL SAMPLES BY MEANS OF STATISTICAL ANALYSIS

Tautvydas Taraškevičius[1], Justinas Čeponkus[1]

[1]Vilnius University, Institute of Chemical Physics, Saulėtekio av. 3, 10257 Vilnius, Lithuania
tautvydas.taraskevicius@ff.stud.vu.lt

Fourier transform infrared (FTIR) spectroscopy is a versatile, non-destructive and non-specific vibrational spectroscopy method with great potential in biomedical research and diagnosis. FTIR spectroscopy has been successfully used in conjunction with statistical and machine learning methods for classifying different biological materials. In addition, many of those data analysis methods have the benefit of identifying the spectral regions associated with the biochemical differences between samples. However, the validity of these biomarkers is not well understood because of the complexity of vibrational spectra of biological samples, as well as the fact that usually there are no reference biomarkers to compare to. The aim of this study is to evaluate the validity of FTIR spectral biomarkers obtained by various statistical and machine learning methods and how they depend on spectral preprocessing.

For this study an experimental dataset was used containing FTIR spectra from two different types of microorganisms – bacteria and yeast – from four species: *Streptococcus pyogenes*, *Staphylococcus aureus* for bacteria and *Candida guilliermondii*, *Candida parapsilosis* for yeast. Three different methods of biomarker extraction were tested, each using one specific technique: univariate statistical hypothesis testing using the t-test; extracting loading vectors from classification models, in this case from a principal component analysis – linear discriminant classifier (PCA-LDC) model; multivariate feature selection using forward feature selection.

Based on the results, the following observations were made. First, while spectral preprocessing didn't appear to cause radical changes to the extracted biomarkers, it did qualitatively alter the interpretations of some of the spectral bands in the example spectra (see Fig. 1.). Furthermore, for the classification-based methods, these changes were observed even when the overall classification accuracy is optimal. Spectral markers obtained from the loadings of a classifier model had the advantage of indicating which spectral bands are correlated with the type of sample used. Multivariate feature selection is believed to be the most logically connected to spectral biomarker identification [1], but has the disadvantage of being computationally time consuming.

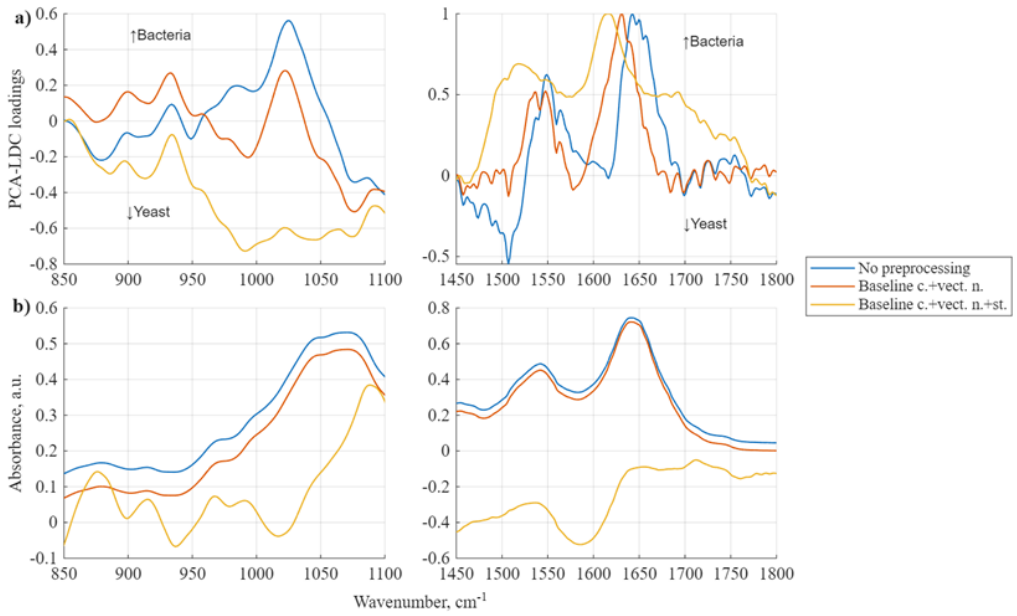

Fig. 1. a) Trained loadings (not to scale) of a PCA-LDC model classifying FTIR spectra of yeast and bacteria with different preprocessing: no preprocessing; linear baseline correction and vector normalization; linear baseline correction, vector normalization and standardization (Z-score normalization). Loadings with positive values are correlated with the spectrum belonging to bacteria and negative values – to yeast. b) Excerpts of a FTIR spectrum of bacteria that correspond to the PCA-LDC loadings in a).

[1] J. Trevisan et al., "Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy," J Biophotonics, vol. 7, no. 3–4, pp. 254–265, Apr. 2014, doi: 10.1002/jbio.201300190