Title: Spectrochemical analysis of liquid biopsy harnessed to multivariate analysis towards breast cancer screening

Type: Article

URL: https://clok.uclan.ac.uk/34228/

DOI: ##doi##

Date: 2020

Citation: Freitas, Daniel L. D., Câmara, Ingrid M., Silva, Priscila P., Wanderley, Nathália R. S., Alves, Maria B. C., Medeiros-De-morais, Camilo De Ielis orcid iconORCID: 0000-0003-2573-787X, Martin, Francis L orcid iconORCID: 0000-0001-8562-4944, Lajus, Tirzah B. P. and Lima, Kassio M. G. (2020) Spectrochemical analysis of liquid biopsy harnessed to multivariate analysis towards breast cancer screening. Scientific Reports, 10 (1).

Creators: Freitas, Daniel L. D., Câmara, Ingrid M., Silva, Priscila P., Wanderley, Nathália R. S., Alves, Maria B. C., Medeiros-De-morais, Camilo De Ielis, Martin, Francis L, Lajus, Tirzah B. P. and Lima, Kassio M. G.

It is advisable to refer to the publisher’s version if you intend to cite from the work. ##doi##

For information about Research at UCLan please go to http://www.uclan.ac.uk/research/

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the http://clok.uclan.ac.uk/policies/
Spectrochemical analysis of liquid biopsy harnessed to multivariate analysis towards breast cancer screening

Daniel L. D. Freitas¹, Ingrid M. Câmara¹, Priscila P. Silva¹, Nathália R. S. Wanderley², Maria B. C. Alves³,⁴, Camilo L. M. Morais⁵,⁶, Francis L. Martin⁵,⁶, Tirzah B. P. Lajus²,³,⁴ & Kassio M. G. Lima¹*

Mortality due to breast cancer could be reduced via screening programs where preliminary clinical tests employed in an asymptomatic well-population with the objective of identifying cancer biomarkers could allow earlier referral of women with altered results for deeper clinical analysis and treatment. The introduction of well-population screening using new and less-invasive technologies as a strategy for earlier detection of breast cancer is thus highly desirable. Herein, spectrochemical analyses harnessed to multivariate classification techniques are used as a bio-analytical tool for a Breast Cancer Screening Program using liquid biopsy in the form of blood plasma samples collected from 476 patients recruited over a 2-year period. This methodology is based on acquiring and analysing the spectrochemical fingerprint of plasma samples by attenuated total reflection Fourier-transform infrared spectroscopy; derived spectra reflect intrinsic biochemical composition, generating information on nucleic acids, carbohydrates, lipids and proteins. Excellent results in terms of sensitivity (94%) and specificity (91%) were obtained using this method in comparison with traditional mammography (88–93% and 85–94%, respectively). Additional advantages such as better disease prognosis thus allowing a more effective treatment, lower associated morbidity, fewer false-positive and false-negative results, lower-cost, and higher analytical frequency make this method attractive for translation to the clinical setting.

Breast cancer is the second most common and the leading cause of cancer-related death amongst women¹. According to the Brazilian Mortality Information System, 14,206 women died in 2013 due to this disease¹. In 2014, the estimation was about 49,240 cases, and in 2018 it was expected to reach 59,700 new cases of breast cancer in Brazil alone¹. This neoplasm is relatively rare in women <35 years old, and increases progressively above this age, especially after age 50 years³. Therefore, breast cancer is a major public health problem taking into consideration the detection and treatment costs⁴. The control of breast cancer has been a priority and is present in the Brazilian Strategic Action Plan for Confronting Non-transmissible Chronic Diseases since 2011⁵.

Only one in three cases of breast cancer can be cured if discovered at an early stage⁶ and there are no effective ways of reducing the incidence of this disease⁸. The best alternative approach to tackle breast cancer is the concept that the earlier the disease is detected, the more effective is the treatment. Early detection through screening is the only method that has proven to be effective in reducing mortality¹. Screening programs are an important health policy practice where the asymptomatic phase of disease is long enough to allow direct or indirect detection of the disease before the occurrence of symptoms.

¹Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil. ²Departamento de Biologia Celular e Genética – Serviço de Aconselhamento Genético, Centro de Oncologia Avançado/CECAN, Universidade Federal do Rio Grande do Norte – Hospital Liga Contra o Câncer, Natal, Brazil. ³Departament of Genetics and Cell Biology, Centro de Biociências, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil. ⁴Department of Pharmacy, Centro de Ciências da Saúde, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil. ⁵Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital, Fulwood, Preston PR2 9HT, UK. ⁶School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, UK. *email: kassiolima@gmail.com
pre-cancerous lesions. A significant degree of transformation in such lesions found in this phase would allow determination of their clinical significance and implementation of effective treatment to improve the patient’s prognosis. Such a screening test that diagnoses early disease needs to be acceptable to patients and available at a reasonable cost.

Mammography is the recommended method for routine screening of breast cancer worldwide. This technique performed with an x-ray machine is described as a radiological examination for evaluation of the breasts. It can be used for checking breast cancer-like lesions in apparently healthy women by finding nodules or calcifications. Exposure to this radiation rarely causes cancer, unless performed with a high periodic frequency whereby risk will increase. Besides being considered painful, relatively expensive, and a source of much discomfort and even embarrassment to patients, its sensitivity varies from 88 to 93%, while its specificity varies from 85 to 94%. Such statistical metrics demonstrate the proportion of women with breast cancer who will present a positive mammogram signalling disease presence, and the rate of women without breast cancer who will have a normal mammography, respectively. Some breast cancer screening tests also include breast self-examination (BSE), clinical examination of breasts (CBE), nuclear magnetic resonance (NMR), and ultrasonography. However, the time from initial patient examination until diagnosis can be too lengthy; about 70% of breast cancer cases lead to complete removal of the breast(s). Many examinations are required to identify the presence of neoplasm: mammogram, breast exam, biopsy, magnetic resonance imaging (MRI) and ultrasound.

Infrared (IR) spectroscopy is a vibrational technique capable of analysing biomolecules, such as nucleic acids (asymmetric PO$_2^-$ in DNA and RNA at $\sim$1,225 cm$^{-1}$), carbohydrates (C–O stretching at $\sim$1,155 cm$^{-1}$), proteins (amide II at $\sim$1,550 cm$^{-1}$ and amide I at $\sim$1,660 cm$^{-1}$) and lipids (C=C stretching at $\sim$1,750 cm$^{-1}$), that exhibit characteristic features in the IR region. Attenuated total reflection Fourier-transform IR (ATR-FTIR) spectroscopy has been used to analyse several biofluids due to its fast spectral acquisition, minimum sample preparation and sample volume, and its non-destructive nature to the sample. Recent research is progressing gradually in which excellent diagnostic results compared to traditional methods have been obtained in various types of cancer such as ovarian, cervical, and prostate; additionally, to diagnosis neurodegenerative diseases such as Alzheimer’s. Herein, we present the results of using ATR-FTIR spectroscopy together with chemometrics for classification of patients with breast cancer in a large-scale screening program using blood biopsies.

Results

The FTIR spectral data in the fingerprint region (900–1,800 cm$^{-1}$) were pre-processed by Savitzky–Golay smoothing (window of 7 points, 2nd order polynomial fitting) followed by AWLS baseline correction and normalization to the Amide I peak (1,650 cm$^{-1}$). The raw and pre-processed spectral data are shown in Fig. 1, where visual overlaps between breast cancer and healthy control spectra are present throughout the whole spectral region indicating the need of chemometric techniques to distinguish samples in such complex matrices. The pre-processed spectral data underwent chemometric analysis by several classification techniques (Table 1). Amongst the classification techniques tested, SPA-SVM presented the best classification performance with accuracy of 92.9% (94% sensitivity and 91% specificity) to detect breast cancer samples based on an external test set (15% of samples, $n=71$ patients). ~70% of samples ($n=334$ patients) were used for model construction and another 15% for internal validation ($n=71$ patients). Overall classification performance represented by the F-Score and G-Score values was good (93%), indicating equal performance with or without considering imbalanced data. Figure 2 shows the receiver operating characteristic (ROC) curve for all models. The best ROC curve (area under the curve [AUC] = 0.929) was found for SPA-SVM, indicating an excellent predictive performance. PCA-SVM (AUC = 0.886) and GA-SVM (AUC = 0.871) were, respectively, the second and third best classification algorithms, demonstrating a good classification performance.

The spectral variables selected by the best classification model (SPA-SVM) are shown in Fig. 3. In total, 16 wavenumbers (901, 959, 980, 999, 1,018, 1,277, 1,364, 1,402, 1,464, 1,489, 1,582, 1,311, 1,626, 1,643, 1,661, and...
1742 cm⁻¹) were responsible for class differentiation using SPA-SVM. The tentative biochemical assignments of these variables based on Movasaghi et al.¹³ are shown in Table 2.

**Discussion**

Breast cancer accounts for approximately 15% of all female cancer deaths and has a 5-years survival rate ranging from approximately 40% in low-income countries to ≥ 80% in developing countries.¹⁴ Its incidence is continually increasing worldwide. This is partly due to a change in the distribution of risk factors: e.g., in developed countries such as the UK, there have been significant increases in women giving birth later in life and in the number of...
women childless by age 45 years. In addition, there has been an increasing adoption of Westernized lifestyles in developing countries\cite{14}, which may be a risk factor for breast cancer.

Mammography-based breast cancer screening is a common practice for early detection of breast cancers, where its efficiency has been demonstrated in randomized controlled trials and observational studies; hence, most organizations that issue recommendations endorse regular mammography as an important part of preventive care\cite{15}. However, although mammography-based breast cancer screening is associated with reduced morbidity and mortality, the majority of women who undergo screening will not develop breast cancer in their lifetime\cite{15}. In addition to the low risk of cumulative exposure to radiation over time and the great discomfort or shame associated with mammography-based screening, false positive results may lead to additional tests and investigations potentially causing psychological distress and anxiety. Conversely, negative results (i.e., where no signs of abnormality are found in the screening) may falsely reassure women when cancer is actually present\cite{14}. Moreover, mammography-based screening may also not benefit all women who are diagnosed with breast cancer, since it may lead to harm in women who undergo further biopsy for abnormalities that may not be breast cancer\cite{15}. For these reasons, less invasive and more accurate breast cancer screening strategies are urgently needed.

<table>
<thead>
<tr>
<th>Selected wavenumber (cm$^{-1}$)</th>
<th>Tentative assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>901</td>
<td>Phosphodiester (absorbances due to collagen and glycogen)</td>
</tr>
<tr>
<td>959</td>
<td>Symmetric stretching vibration of $\nu_2\text{PO}_4^-$</td>
</tr>
<tr>
<td>980</td>
<td>$\text{OCH}_3$ (polysaccharides)</td>
</tr>
<tr>
<td>999</td>
<td>Ring stretching vibrations mixed strongly with CH in plane bending</td>
</tr>
<tr>
<td>1,018</td>
<td>$\nu$(CO), $\nu$(CC), $\delta$(OCH), ring (polysaccharides, pectin)</td>
</tr>
<tr>
<td>1,277</td>
<td>Vibrational modes of collagen</td>
</tr>
<tr>
<td>1,311</td>
<td>Amide III band components of proteins</td>
</tr>
<tr>
<td>1,364</td>
<td>Stretching C–O, deformation C–H, deformation N–H</td>
</tr>
<tr>
<td>1,402</td>
<td>Symmetric CH$_3$ bending modes of the methyl groups of proteins</td>
</tr>
<tr>
<td>1,464</td>
<td>CH$_2$ scissoring mode of the acyl chain of lipid</td>
</tr>
<tr>
<td>1,489</td>
<td>In-plane CH bending vibration</td>
</tr>
<tr>
<td>1,582</td>
<td>Ring C–C stretch of phenyl</td>
</tr>
<tr>
<td>1,626</td>
<td>Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode</td>
</tr>
<tr>
<td>1,643</td>
<td>Amide I band (arises from C=O stretching vibrations)</td>
</tr>
<tr>
<td>1,661</td>
<td>$\delta$(C=C) cis in lipids and fatty acids</td>
</tr>
<tr>
<td>1,742</td>
<td>C=O stretching mode of lipids</td>
</tr>
</tbody>
</table>

Table 2. Selected wavenumbers by the SPA-SVM to distinguish healthy controls and breast cancer samples.
Herein, ATR-FTIR spectroscopy in conjunction with chemometric techniques was used to detect breast cancer in a total cohort of 476 patients recruited over 2 years for an early-stage breast cancer screening program in Natal, Brazil. Breast cancer detection among normal samples was successfully performed based on the blood plasma spectra with 93% accuracy (94% sensitivity, 91% specificity, AUC = 0.929) in an external (blind) cohort of 71 patients using the SPA-SVM algorithm. Sixteen spectral features were responsible for class differentiation in the fingerprint region (Table 2). These are predominantly associated with phosphodiester (P–O vibrations), polysaccharides (C–O stretching), proteins (CH₃ bending, Amide III, Amide I band), nucleic acids (C=O stretching and C=C ring breathing mode), and lipids (C=O stretching and C=C(unsaturated)). C–O vibrations in carbohydrates, P–O vibrations in phosphodiester, and proteins vibrations; these have been previously associated with breast cancer in serum. Serum applications for breast cancer detection have been performed using IR spectroscopy by Backhaus et al., where 98% sensitivity and 95% specificity (using cluster analysis) and 92% sensitivity and 100% specificity [using artificial neural networks (ANN)] was obtained in a study carried out with 196 patients. Likewise, Elmi et al. detected breast cancer in serum-based IR spectroscopy with 76% sensitivity and 72% specificity for breast cancer cases using principal component analysis linear discriminant analysis (PCA-LDA) in a study with 86 samples (43 breast cancer, 43 healthy controls). The results reported herein are higher taking into consideration the large number of patients, where the sensitivity and specificity are found to be > 90%; being comparable to results obtained by more sophisticated methods such as using quantum cascade laser IR imaging, where sensitivity and specificity has been reported at 94% and 86%, respectively, using a random forest classifier. However, there are no studies reporting breast cancer screening based on plasma samples using IR spectroscopy for a big cohort of samples. Herein, 476 patients were studied resulting in a diagnostic accuracy, sensitivity and specificity above 90% for cancer detection.

**Methods**

**Samples.** In this study, we evaluated two groups of women. The first, Breast Cancer (BC), refers to a group of women diagnosed with breast cancer, with or without neoadjuvant treatment, and were collected by professionals trained at the Liga Contra o Câncer Hospital (Natal/RN, Brazil), during a period of 2 years. The second, Healthy Controls (HC), refers to a group of women with no previous or current diagnosis of breast cancer, collected at the Prontoclínica Dr. Paulo Gurgel (Natal/RN, Brazil), during the same time period. In both groups, patients were > 18 years old, and family history related to some type of cancer was not taken into account. The Institutional Ethics Committee for Human Research of the Hospital Universitário Onofre Lopes (HUOL), of the Federal University of Rio Grande do Norte (UFRRN), Brazil, approved this study (Ethical Approval Number—44113115.1.1001.5292) and informed consent was obtained from all subjects. Also, all the methods carried out in this study were by the approved guidelines. Samples from both groups were obtained after the reading of a Free Informed Consent Form and signature of the patients. Vacutainer tubes BD with 5 mL EDTA were used with disposable vacuum syringes. Thereafter, they were centrifuged for 10 min, and frozen at approximately − 20 °C until the time of data analysis. A total of 476 samples were obtained.

**ATR-FTIR spectroscopy.** The samples were removed from the freezer 15 min before analysis to allow thawing. Samples were randomized and, to minimize temporal or instrumental effects, a similar number of samples from both groups were measured on each day. The absorption spectra were obtained using an attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectrometer model IRAffinity-IS (Shimadzu Corp., Kyoto, Japan). The spectra were obtained in the range between 600 and 4,000 cm⁻¹, with 32 co-added scans and 4 cm⁻¹ spectral resolution (2 cm⁻¹ data spacing). The ATR crystal was cleaned with alcohol (70% v/v) and acetone (P.A.) for each new sample and before setting the new background. A 10-μL staken performed. This procedure was repeated in triplicate. The measurement time for each sample was approximately 5 min.

Three spectra collected per sample were first averaged and the following pre-processing was applied to the dataset: truncation to the fingerprint region (900–1800 cm⁻¹ with 468 wavenumber data points), Savitzky–Golay (SG) smoothing to remove random noise (window = 15 points, 2nd order polynomial fitting), automatic weighted least squares baseline correction, and normalization to the Amide I peak (1,650 cm⁻¹).

**Data analysis.** The spectral data import, pre-processing and construction of multivariate classification models were performed using the MATLAB R2014b environment version 8.4 (MathWorks, Inc., Natick, USA) with the PLS-Toolbox version 7.9.3 (Eigenvector Research, Inc., Manson, USA) and laboratory-made routines. All spectra were organized into a data matrix, where samples were represented as rows and the wavenumbers as columns. The samples were divided into three different subsets by the Kennard–Stone (KS) sample selection algorithm; training (70%), validation (15%) and test (15%) sets. The training set was used to build the classification models, while the validation set to optimize and evaluate its internal performance. Finally, the test set was used to evaluate the model classification performance towards external samples.

The computational analysis consisted of testing three algorithms for feature extraction and selection: principal component analysis (PCA), successive projections algorithm (SPA) or genetic algorithm (GA), followed by discriminant analysis classifiers: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) or support vector machines (SVM). These algorithms were coupled as feature extraction/selection and classification as: PCA-LDA, PCA-QDA, and PCA-SVM; SPA-LDA, SPA-QDA, and SPA-SVM; and GA-LDA, GA-QDA, and GA-SVM.

PCA is a feature extraction method widely used for data reduction. It decomposes the pre-processed spectral data into a small number of principal components (PCs) containing scores (variance on sample direction) and loadings (variance on wavenumber direction). The PCA scores are used to assess similarities/dissimilarities between the samples, while the PCA loadings to investigate potential spectral markers. SPA is a forward feature
The statistical parameters for the evaluation of the classification models were: accuracy (AC), sensitivity (SENS), specificity (SPEC), Youden's Index (YOU), positive predictive value (PPV), negative predictive value (NPV), F-Score and G-Score. AC is related to the percentage of correct classification achieved by the model. SENS measures the proportion of positive results that are correctly identified while SPEC measures the proportion of negative results that are correctly identified. In this study, when we have a case-control patients approach, sensitivity can be understood as the probability to find a positive result when the disease is present, while specificity can be understood as the probability to find a negative result when the disease is not present. Youden's index (YOU) evaluates the classifier's ability to avoid failure. The PPV measures the proportion
Table 3. Equations to calculate the figures of merit for model evaluation. FN stands for false negative, FP for false positive, TP for true positive, and TN for true negative.

<table>
<thead>
<tr>
<th>Parameter (%)</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (AC)</td>
<td>$\frac{TP + TN}{TP + TN + FP + FN} \times 100$</td>
</tr>
<tr>
<td>Sensitivity (SENS)</td>
<td>$\frac{TP}{TP + FN} \times 100$</td>
</tr>
<tr>
<td>Specificity (SPEC)</td>
<td>$\frac{TN}{FP + TN} \times 100$</td>
</tr>
<tr>
<td>Toudens index (YOU)</td>
<td>$\frac{SENS}{1 - (100 - SPEC)}$</td>
</tr>
<tr>
<td>Positive predictive value (PPV)</td>
<td>$\frac{TP}{TP + FP} \times 100$</td>
</tr>
<tr>
<td>Negative predictive value (NPV)</td>
<td>$\frac{TN}{TN + FN} \times 100$</td>
</tr>
<tr>
<td>F-score</td>
<td>$\frac{2 \times SENS \times SPEC}{SENS + SPEC}$</td>
</tr>
<tr>
<td>G-score</td>
<td>$\sqrt{SENS \times SPEC}$</td>
</tr>
</tbody>
</table>

of positives that are correctly assigned (its value varies between 0 and 1); the NPV measures the proportion of negatives that are correctly assigned (its value varies between 0 and 1); the F-score represents the weighted average of the precision and sensitivity; and the G-score accounts for the model precision and sensitivity without the influence of positive and negative class sizes. These parameters are calculated based on the equations shown in Table 3. In addition, a receiver operating characteristics (ROC) curve was generated to all models. The area under curve (AUC) value was calculated to evaluate how well the model can distinguish the samples between the different classes analysed.

Received: 22 March 2020; Accepted: 17 June 2020
Published online: 30 July 2020

References


**Acknowledgements**

D.L.D. Freitas would like to thank CAPES/PPGQ/UFRN for financial support. C.L.M. Morais would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil (Grant 88881.128982/2016-01), for his research grant. The authors greatly appreciate the contribution of all women involved in this study, and the collaboration of the Universidade Federal do Rio Grande do Norte (UFRN), Liga Contra o Câncer Hospital (Natal, Brazil), and the Prontoclinica Dr. Paulo Gurgel (Natal, Brazil) for the support during sample acquisition.

**Author contributions**

D.L.D.F was responsible for the construction of the chemometric models and multivariate analysis. I.M.C and D.L.D.F were responsible for acquiring the spectral data and writing the first draft of the manuscript. P.P.S, N.R.S.W and M.B.C.A were responsible for collecting the patients’ samples. C.L.M.M. and F.L.M. provided chemometric support and finalised the manuscript. T.B.P.L and K.M.G.L. supervised the project and revised the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Correspondence** and requests for materials should be addressed to K.M.G.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

© The Author(s) 2020