Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Spectral model for diagnosis of acute leukemias in whole blood and plasma through Raman spectroscopy

Adriano Moraes da Silva
Fernanda Sant Ana de Siqueira e Oliveira
Pedro Luiz de Brito
Landulfo Silveira Jr.

SPIE.

Adriano Moraes da Silva, Fernanda Sant Ana de Siqueira e Oliveira, Pedro Luiz de Brito, Landulfo Silveira Jr. "Spectral model for diagnosis of acute leukemias in whole blood and plasma through Raman spectroscopy," *J. Biomed. Opt.* **23**(10), 107002 (2018), doi: 10.1117/1.JBO.23.10.107002.

Journal of Biomedical Optics 23(10), 107002 (October 2018)

Spectral model for diagnosis of acute leukemias in whole blood and plasma through Raman spectroscopy

Adriano Moraes da Silva,[a] Fernanda Sant Ana de Siqueira e Oliveira,[a] Pedro Luiz de Brito,[b] and Landulfo Silveira Jr.[c,*]
[a]Universidade Paulista—UNIP, Institute of Health Sciences, São José dos Campos, São Paulo, Brazil
[b]Grupo de Assistência à Criança com Câncer—GACC, São José dos Campos, São Paulo, Brazil
[c]Universidade Anhembi Morumbi—UAM, Center for Innovation, Techonology and Education—CITE, Parque Tecnológico de São José dos Campos, São José dos Campos, São Paulo, Brazil

Abstract. Acute leukemias are oncohematological diseases that compromise the bone marrow and have a complex diagnostic definition, leading to a high mortality when diagnosed late. This study proposed to determine the spectral differences between whole blood and plasma samples of healthy and leukemic subjects based on Raman spectroscopy (RS), correlating these differences with their resulting biochemical alterations and performing discriminant analysis of the samples ($n = 38$ whole blood and $n = 40$ plasma samples). Raman spectra were obtained using a dispersive Raman spectrometer (830-nm wavelength, 280-mW laser power, 30-s exposure time) with a Raman probe. The exploratory analysis based on principal component analysis (PCA) of the blood and plasma sample's spectra showed loading vectors with peaks related to amino acids, proteins, carbohydrates, lipids, and carotenoids, being the spectral differences related to amino acids and proteins for whole blood samples, and mainly carotenoids for plasma samples. Discriminant models based on partial least squares (PLS) and PCA were developed and classified the spectra as healthy or leukemic, with sensitivity of 91.9% (PLS) and 83.9% (PCA), specificity of 100% (both PLS and PCA), and overall accuracy of 96.5% (PLS) and 93.0% (PCA) for the whole blood spectra. In plasma, the sensitivity was 95.7% (PLS) and 11.6% (PCA), specificity of 98% (PLS) and 100% (PCA), and overall accuracy of 97.1% (PLS) and 64.1% (PCA). The study demonstrated that RS is a technique with potential to be applied in the diagnosis of acute leukemias in whole blood samples. [DOI: 10.1117/1.JBO.23.10.107002]

Keywords: Raman spectroscopy; acute leukemia; diagnosis; whole blood; plasma; discriminant analysis.

Paper 180357R received Jun. 14, 2018; accepted for publication Sep. 21, 2018; published online Oct. 22, 2018.

1 Introduction

Acute leukemias are defined as disorders where primitive hematopoietic cells (named blasts), especially white blood cells, suffer malignant transformations whereby they are produced in a clonal, uncontrolled, and autonomous way, with their functions, morphologies, and the maturation sequence generally altered, keeping them immature and inefficient.[1–3] Leukemias are oncological diseases of not fully known causes and present high incidence in the population; they are among the 11 most common cancer diseases in the world, reaching the mark of 257,000 new cases in 2017 in the world population;[4] and the National Cancer Institute estimates 10,800 new cases in Brazil in 2018.[5]

Leukemias are initially classified following the nomenclature of the cell's lineage, being the lymphoid and myeloid, depending on the type of blood cells involved (lymphoblasts or myeloblasts), and the time of development for each disorder, being considered acute leukemias those of fast and more aggressive development, and chronic leukemias those of slow and less aggressive development.[6]

The diagnosis of leukemia is based on a set of tests, starting with the blood count screening, which is able to evaluate (qualitatively and quantitatively) the blood cells, followed by a more invasive myelogram examination, which analyzes the intramedullary cells related to the alterations of the cell's morphologies and stages of maturation.[6–9] However, these morphological tests are not sufficient for an accurate diagnosis, and for this reason, clinicians request more complex laboratory tests capable of conclusive diagnosis. The cytogenetics and fluorescence *in situ* hybridization are aimed to identify specific genes and chromosomal alterations (deletions, translocations, and cryptic rearrangements) implicated in the process of leukemogenesis.[10] Immunophenotyping by flow cytometry and molecular biology are also part of the range of exams currently available for the elaboration of a definitive diagnosis.[11] Only after confirmation will the leukemia therapy be started and the prognosis known, but the average time to release the results of these tests varies from 3 to 5 days, impacting negatively the development and control of the disease.[12] The techniques based on myelogram and immunophenotyping present high sensitivity and specificity for the diagnosis of acute leukemia, and the accuracy to detect abnormal myeloid cells using phenotyping is shown to be around 94%,[13] whereas the sensitivity and specificity for the characterization of myeloid markers by immuniphenotyping are dependent on the antigen detected, with minimum sensitivity and specificity values of 97% and 88%, respectively.[14]

*Address all correspondence to: Landulfo Silveira, Jr., E-mail: landulfo.silveira@gmail.com; lsjunior@anhembi.br

1083-3668/2018/$25.00

The treatment occurs by administration of chemotherapy and depends not only on some factors inherent to the disease itself, but also on the clinical conditions of the patients. Despite records of a decrease in mortality, it is estimated that one-third of the patient present recurrence and the overall five-year survival rate is ~35%.[15] These numbers motivate the search for new methods and techniques that can reduce the time interval between the hypothesis and the diagnostic conclusion in favor of better prognoses for the patients.[15]

Raman spectroscopy (RS) is an optical technique capable of identifying biochemical changes in biological tissues and fluids related to pathological conditions,[16–18] diagnosing Alzheimer's disease in blood serum[19,20] as well as quantifying human blood and serum components *in vitro*.[21] RS has been applied to identify spectral information referring to blood cells and biochemical elements present in the blood, such as amino acids, proteins, lipids, nucleic acids, and carotenoids,[21] and the technique has been gaining importance in the field of diagnostics of different types of leukemia using cell lines,[22] blood smears,[23] and blood serum samples.[24] The Raman effect, which is the basis of RS, is based on the inelastic scattering of an incident laser light by polarizable molecules and has a great ability to provide detailed information on the vibrational energy levels of different materials.[25,26] When applied to the diagnosis of blood diseases, RS has advantages, such as potential for *in vivo* use, rapidness, no need for reagents or dyes to reveal the tissue biochemical information, and the possibility of obtaining the diagnosis using small amount of sample nondestructively, thus preserving its integrity.[26–29]

The objective of this study was to use the RS (830-nm excitation) to identify the spectral differences in whole blood and plasma samples from healthy and acute leukemia subjects and use these differences to discriminate both statuses. The spectral differences related to the biochemicals presented in each sample type will be statistically evaluated by the student's *t*-test, and then these peaks will be assigned to their corresponding chemical compositions already described in the scientific literature.[24,25,30] Then, the spectral dataset will be submitted to a classification model by discriminant analysis (DA) employing partial least squares (PLS) and principal component analysis (PCA) for the spectral differentiation of healthy from leukemic samples, through their most significant peaks.

2 Materials and Methods

2.1 *Whole Blood and Plasma Samples*

The study was approved by the Research Ethics Committee of Universidade Paulista—UNIP (Process CAAE No. 67895617.5.0000.5512). Human blood samples were obtained from the laboratory of clinical analyses of a reference hospital for oncohematological diseases in São José dos Campos. The diagnosis of leukemia in the samples enrolled in the study was done with qualitative and quantitative analysis of the peripheral blood cells through hemogram, which results in a suggestive hematological disease, progressing to detailed exams, such as myelogram, immunophenotyping, cytogenetics, and molecular biology[31] when needed.

The blood samples were collected from peripheral veins of each subject by vacuum-closed method in tubes containing K3EDTA (7.2 mg) as anticoagulant (Sarstedt AG & Co., Nümbrecht, Germany). An aliquot of each blood sample was transferred to a tube for mechanical centrifugation at 3500 RPM for 10 min to obtain the plasma (model Combate, CELM Ltd., Barueri, SP, Brazil); the remnant blood was maintained in the original tube.

It was evaluated 25 samples of whole blood from healthy subjects and 17 samples from whole blood from acute leukemic subjects identified after conventional diagnostic techniques. The samples of whole blood and plasma were properly conditioned in thermal boxes (2°C to 8°C) in order to avoid changes in the biochemical constitution due to temperature. At the time of spectroscopy, these samples were separated into the two groups named healthy group and leukemic group.

2.2 *Raman Spectroscopy*

Raman spectra were obtained in whole blood and plasma samples without any preparation, by pipetting an amount of 80 μL in an aluminum sample holder with holes, using a single-channel micropipette of variable volume (model P200, Bio-Rad, Hercules, California). The spectra were obtained in a near-infrared Raman spectrometer (model Dimension P1, Lambda Solutions Inc., Massachusetts), which uses a diode laser at 830 nm coupled to a Raman probe fiber optic cable for sample's excitation, obtaining 280 mW of laser power at the excitation output of the probe. The scattering of the sample was collected by the Raman probe and coupled to the spectrometer for dispersion. The spectrometer disperses the scattered light onto a back thinned, deep-depleted charge-coupled device camera (1340 × 100 pixels, cooled to −75°C) in the spectral range between 400 and 1800 cm^{-1}, providing ~2 cm^{-1} of spectral resolution. The exposure time for obtaining each spectrum was 3 s with 10 accumulations (30 s of total exposure time), and each sample was analyzed between three and five replicates in order to increase the number of spectra in the discrimination model. Neither noticeable damage to the blood or serum nor spectral change was observed during the spectrum acquisition.

The collected Raman spectra were subjected to preprocessing to remove the Raman background (mainly fluorescence from blood components and cells) by fitting and subtracting a seventh-order polynomial over the entire spectral range of 400 to 1800 cm^{-1}. Spikes from cosmic rays were removed manually and then the spectra were normalized by the "area under the curve" (one-norm).[32] The mean spectra of each group were calculated for visual comparison and statistics. Table 1 presents an overview of the number of spectra collected in each of the groups. Four blood samples from the healthy group and two plasma samples from the leukemic group were excluded from the study due to the presence of hemolysis and low Raman signal-to-noise ratio identified after preprocessing.

Table 1 Number of samples and number of spectra in each group.

Sample	Type	Number of samples	Total collected spectra
Whole blood	Healthy	21	80
	Leukemic	17	62
Plasma	Healthy	25	101
	Leukemic	15	69

The mean spectra were plotted in order to identify visual differences in the intensities of the Raman peaks between the healthy and leukemic groups. The most intense peaks in both spectra were labeled and the Raman band positions were tabulated to determine the composition of whole blood and plasma based on the Raman features. Student's t-test with significance level of 5% ($p < 0.05$) was applied to the intensities of the peaks between healthy and leukemic groups to determine the peaks with significant differences, in a way to determine the differences in the biochemical composition between healthy and leukemic blood and plasma and to identify the differences in the biochemical profile of both groups. The t-test is used to compare the means when the samples follow normal distribution but with unknown variance, and the p-value is used to accept (p-value >0.05) or reject (p-value <0.05) the null hypothesis (equality in the mean between the two populations), being accepted the alternative hypothesis that the means come from different populations. In this particular study using the intensities of the Raman peaks, as the Raman peaks may present intensities higher or lower when comparing healthy and leukemic groups, the two-tailed t-test was used. Thus, the t-test can be used to make decisions based on statistical calculations with a high degree of confidence.[33]

2.3 *Exploratory Analysis and Discrimination*

2.3.1 *Exploratory analysis by principal component analysis*

The PCA is used to analyze data of a multivariate nature. It is a statistical tool that allows transforming a set of variables of a database (the Raman spectra) into its principal components (PCs) based on the variance of the data in the group. The PCA extracts the most significant information (based on the variance) from an original dataset, generating two new variables, called principal components loading vectors (PCs) and scores (SCs), where each PC loading vector, which resemble Raman spectra, presents a "weight," the SC, which indicates the intensity of each loading that is present in the original data.[29,32,34,35] From these two variables, the similarities and differences in the groups can be identified. The largest spectral variation is stored in PC1, and the extraction of the variations follows successively (PC2, PC3, etc.) until the lowest variance component, being the loadings extracted in a way that each PC is orthogonal to each other (no redundant variation is represented in each PC).

In the exploratory analysis, the aim is to identify which spectral variables (PC loadings) present significant differences between the groups, evaluated by the t-test applied to their scores, and to correlate these loadings (spectral differences or variances) with the biochemical differences between the healthy and leukemic groups. Then the loadings with significant differences were compared to the vibrational peaks assigned to the known biochemical components of the blood obtained from the published literature. The t-test was applied to the scores of both healthy and leukemic groups in order to identify which loading has statistically significant differences between the two groups ($p < 0.05$).

The MATLAB software (version 2007a, The MathWorks Inc., Natick, Massachusetts) was used to perform exploratory analysis based on PCA.

2.3.2 *Discriminant analysis by partial least squares and PCA*

Multivariate regression based on PLS is an important statistical tool applied to establish linear relationship models between multivariate measures.[36] Authors have applied PLS to classify and categorize some types of cancers by using the unique biochemical information present in the Raman spectra.[37] Since it is known that PLS is related to canonical correlation analysis (CCA) and that CCA is, in turn, related to linear discriminant analysis (LDA), PLS has similarities to LDA[38,39] and thus can be used to discriminate samples in groups based on a training dataset. By using PLS regression, the model finds the "Fisher's among-groups sum-of-squares and cross-product matrix," which means that any correlation between the predicted and predictor variables in the training set are estimated and maximized, and therefore used to model the output. This means that the "within-groups" variations are distinguished from the "among-groups" variations, and the discrimination is achieved by focusing on the "among groups" variations,[38–40] resulting in the identification of the most relevant differences in whole blood and plasma samples and using these differences to discriminate the spectra of leukemic from the healthy samples.

PCA was also applied to perform discrimination between healthy and leukemic spectra using the statistically significant scores ($p < 0.05$), meaning that these components bring the most significant differences between the healthy and leukemic groups for both whole blood and plasma, and that can be used as diagnostic or discrimination (classification) parameters.

The Chemoface software by Nunes et al.[41] was used to model the discrimination problem based on the Raman spectra by applying the cross-validation methodology of "leave-one-out" (withdrawing a sample, modeling with $n - 1$ samples, and validating the withdrawn sample), for both PLS-DA and PCA-DA, respectively. With the results obtained from the discrimination of whole blood and plasma, a confusion matrix and its plot was created, comparing the rate of discrimination of whole blood and plasma samples using the Raman spectral models (PLS-DA and PCA-DA) compared to the correct diagnostics of the conventional tests.

3 Results and Discussion

3.1 *Whole Blood and Plasma Spectra*

In this study, 312 spectra were collected from 42 individuals (25 healthy and 17 leukemic), being 142 spectra originated from whole blood (being 80 classified as healthy and 62 as leukemic) and 170 spectra originated from plasma (being 101 classified as healthy and 69 as leukemic), as shown in Table 1.

Figure 1 presents the mean Raman spectra of whole blood from healthy and leukemic groups. The spectrum from whole blood of the healthy group shows peaks in the positions of the blood constituents: cellular components (mainly leukocytes, erythrocytes, and platelets) and noncellular components (mainly plasma). Table 2 presents the positions of the main peaks of whole blood accompanied by their respective assignments as described in the literature[24,25,30,42,43] and organized by biochemical groups. By visual inspection, some peaks showed small but perceptible differences in intensity and bandwidth between healthy and leukemic, suggesting some relevant variation in these tissue components among the samples. The t-test ($p < 0.05$) applied to identify peaks with statistically significant differences in the healthy versus leukemic groups is presented in

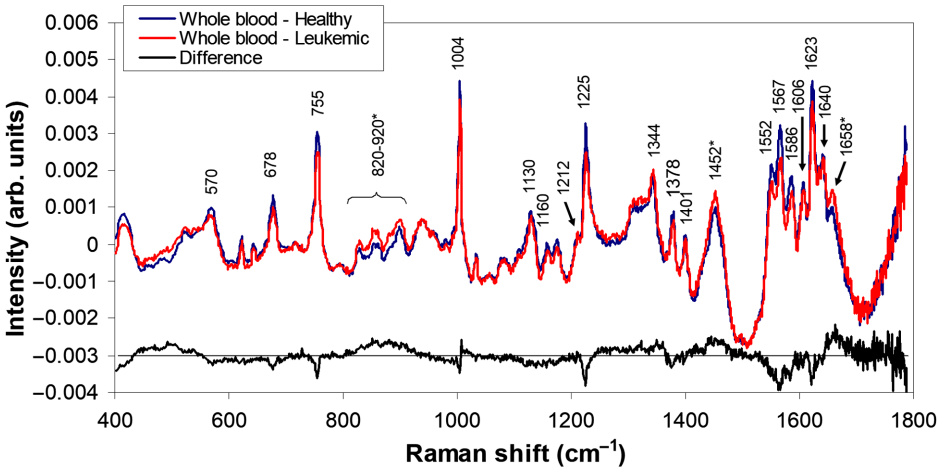

Fig. 1 Mean Raman spectra of whole blood from subjects in the healthy and leukemic groups and the spectrum of the difference between leukemic and healthy. The symbol * represents the peaks that are more intense in the spectra of leukemic than in the spectra of healthy.

Table 2. The higher differences were observed in the peaks at the positions of 570, 678, 755 cm^{-1}, peaks in the range between 820 and 920 $*$ cm^{-1}, and peaks at 1004, 1130, 1160, 1212, 1225, 1344, 1378, 1401, 1452*, 1552, 1567, 1586, 1623, 1640, and 1658 $*$ cm^{-1} (the symbol * denotes peaks in which the leukemic group shows greater intensity than the healthy group). Peaks at 1212, 1401, and 1640 cm^{-1} did not present statistically significant differences between the two groups ($p > 0.05$), suggesting that they do not aid in the differentiation between healthy and leukemic. As noted in Fig. 1, all the peaks in the spectrum of the healthy group are more intense when compared to the leukemic group, with the exception of the peaks at 820 to 920, 1452, and 1658 cm^{-1}, which present greater intensity in the leukemic group.

Figure 2 presents the mean Raman spectra of blood plasma from healthy and leukemic groups. The spectrum from plasma of the healthy group shows peaks in the positions of plasma constituents, mainly proteins (albumin, globulins, and amino acids), lipid fractions, carbohydrates (glucose), carotenoids, and metabolites. Table 2 presents the positions of the main peaks of plasma peaks accompanied by their respective assignments, as described in the literature[24,25,30,42,43] and organized by biochemical groups. By visual inspection, there were no relevant differences in the intensities of the peaks of the healthy versus leukemic groups; however, significant differences were observed in the peaks at 510, 721, 760, 837, 947, 1004, 1132, 1160, 1210, 1269, 1334, 1344, 1407, 1448, 1455, 1525, 1630, 1659, and 1666 cm^{-1} ($p < 0.05$). Due to the lower composition complexity of the plasma compared to the whole blood, the amount of peaks with statistically significant differences is lower in plasma than in whole blood, since in this specimen, the peaks referring to leukocytes, erythrocytes, and platelets are absent.

3.2 *Exploratory Analysis*

In the exploratory analysis, the PCA technique has been used to identify the spectral features that presented differences between the groups, through the PCs and SCs. Interpretation consists of identifying the peaks present in the first PCs, and correlating these peaks with the biochemical compounds present in each group according to the literature. The scores are then used to "quantify" these compounds in each of the healthy and leukemic groups. Positive peaks with positive scores, as well as negative peaks with negative scores, show that the specific biochemical is at a high concentration in that particular group, while positive peaks with negative scores and vice-versa show that the biochemical components attributed to that PC are presented in lower concentrations.

Figure 3 shows the plot of the PCs and SCs of the whole blood samples. The PC1 has characteristic peaks of whole blood, but the leukemic group presents a lower score of this loading (SC1) ($p < 0.001$), suggesting that despite the same constitution in terms of Raman features, the leukemic group presents these constituents in lower concentrations. PC2 shows peaks with negative intensities in the leukemic group and positive peaks in the healthy group, with a statistically significant difference in the SC2 ($p < 0.01$). This loading has positive peaks at 752, 1213, 1524, 1547, and 1619 cm^{-1} and negative peaks at 1007, 1381, 1403, and 1643 cm^{-1}, being the positive peaks assigned to proteins, amino acids and carotenoids, and the negative peaks related to amino acids, carbohydrates, and lipids. The positive features suggest the highest protein/amino acid concentration for the healthy group, mainly by observing peaks at 752, 1213, and 1547 cm^{-1}, which coincide with the peaks of the whole blood, especially erythrocytes.[27] Also, the presence of highest carotenoid concentration in the healthy has been evidenced. Still, PC3 presents peaks with negative intensities in the leukemic group and positive peaks in the healthy group, with significant difference for SC3 ($p < 0.001$). This component has positive peaks at 570, 678, 756, 1142, 1226, 1381, 1404, 1501, 1569, 1624, and 1643 cm^{-1} and negative peaks at 1003, 1212, 1451, 1544, and 1664 cm^{-1}, being the positive peaks assigned to proteins, amino acids, and carbohydrates, and negative peaks referred to amino acids and lipids. The SC3 indicates that the healthy group has higher concentrations of proteins (amide bands I and III and glutathione), amino acids (Trp and Tyr), and glucosamine, assigned to the biochemical components of whole blood. The leukemic group, with negative SC3, shows negative peaks related to amino acids (Phe and Trp) and lipids (cellular phospholipids), which indicates the presence of these compounds in higher concentrations in leukemia and suggests hypercellularity caused by blasts (white cells, particularly granulocytes).[30] Although PC4 presented

Table 2 Grouping of the main Raman peaks of whole blood and plasma according to the biochemical constitution, assignments according to the published literature,[24,25,30,42,43] and statistically significance (p-value) of the peak's intensities between healthy and leukemic.

Biochemical group	Peak position (cm^{-1})	Assignments	p-value (respective to peak position)	Reference
Proteins and amino acids	510 (P)	Trp	<0.01	24
	755 (WB)	Protein, Trp	<0.0001	24, 25, 30
	760 (P)	Trp	Not significant	24
	820 to 920 (WB)	Tyr, Trp, glutathione	<0.0001	24, 30
	831 (P)	Tyr, Trp, glutathione	Not significant	24
	897 (P)	Tyr, Trp, glutathione	Not significant	24
	1004 (WB; P)	Phe	<0.001 (WB); <0.0001 (P)	24, 25, 30
	1130 (WB); 1132 (P)	Protein	<0.01 (WB); not significant (P)	24
	1210 (P); 1212 (WB)	Trp, Phe, Tyr, amide III	<0.01 (P); not significant (WB)	24, 30
	1225 (WB)	Protein, amide III	<0.0001	24
	1269 (P)	Protein, amide III	<0.01	24, 30
	1334 (P)	Trp	<0.001	24
	1344 (P)	Protein, Trp	<0.001	24, 25
	1401 (WB); 1407 (P)	Glutathione	Not significant (WB, P)	24
	1448 (P); 1452 (WB); 1455 (P)	Protein	<0.01 (P); not significant (WB); <0.0001 (P)	24, 25
	1552 (WB)	Trp, amide II	<0.01	24, 25, 30
	1586 (WB)	Protein, Tyr	<0.01	24, 25
	1606 (WB)	Protein, Tyr, Phe	<0.001	24, 25, 30
	1623 (WB)	Tyr, Trp	<0.0001	24
	1658 (WB); 1659 (P); 1666 (P)	Protein, amide I	<0.0001 (WB); not significant (P); not significant (P)	24, 25, 30, 43
Lipids	1130 (WB); 1132 (P)	Lipids, phospholipids	<0.01 (WB); not significant (P)	24, 30
	1225 (WB)	Lipids	<0.0001	30
	1269 (P)	Lipids, phospholipids	<0.01	30, 43
	1334 (WB)	Phospholipids	<0.01	24, 30, 43
	1344 (WB; P)	Phospholipids	<0.001 (WB; P)	24, 30, 43
	1448 (P); 1452 (WB); 1455 (P)	Lipids, phospholipids	<0.01 (P); <0.0001 (WB; P)	24, 25, 30, 43
	1658 (WB); 1659 (P)	Phospholipids	<0.0001 (WB); not significant (P)	24, 30, 42
	1666 (P)	Phospholipids	Not significant	30
Carbohydrates	721 (P)	Polysaccharides	<0.0001	24
	1378 (WB)	Glucosamine	<0.05	24, 30
Carotenoids	1004 (WB; P)	β-carotene	<0.001 (WB); <0.0001 (P)	30
	1160 (WB); 1160 (P)	β-carotene	<0.01 (WB); <0.001 (P)	24
	1525 (WB)	β-carotene	Not significant	24, 30

Note: Abbreviations: Phe: phenylalanine; Tyr: tyrosine; Trp: tryptophan. (WB): peaks referred only to whole blood and (P): peaks referred only to plasma.

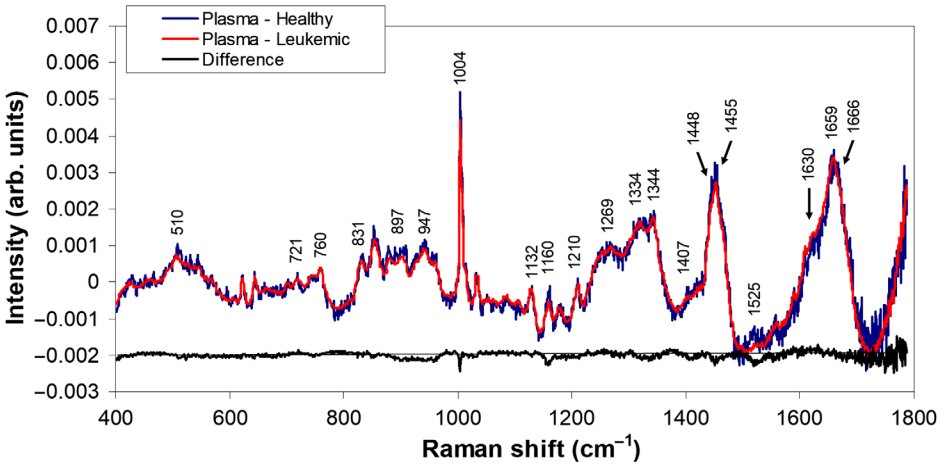

Fig. 2 Mean Raman spectra of plasma from subjects in the healthy and leukemic groups and the spectrum of the difference between leukemic and healthy.

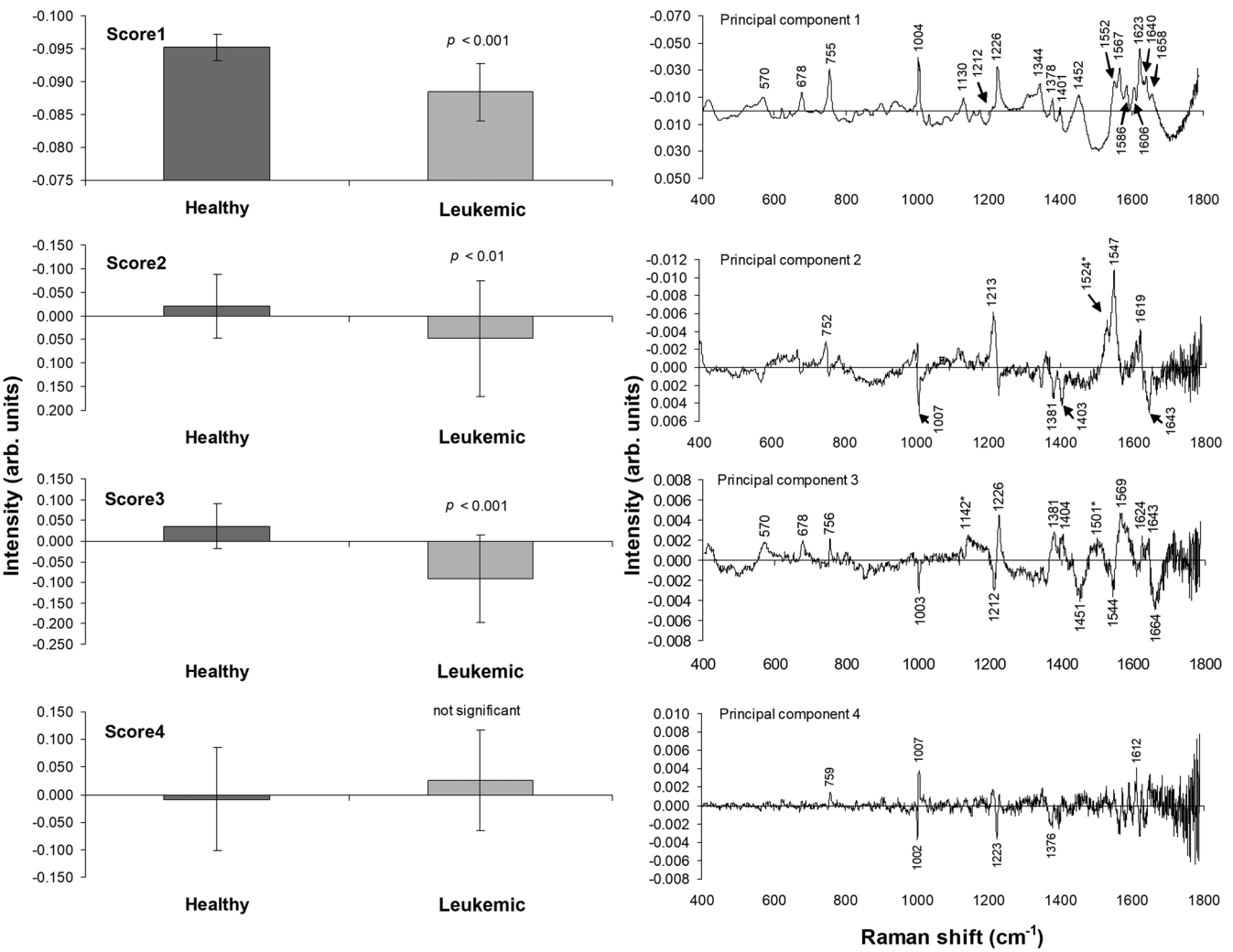

Fig. 3 Plots of the PCs and SCs calculated for the whole blood samples for exploratory analysis. The peaks marked with "*" represent peaks found in this study which do not present known assignment.

peaks in positions related to proteins and amino acids, SC4 did not present significant difference of healthy versus leukemic groups ($p > 0.05$). Therefore, PC2 and PC3 indicated the major differences, mainly higher concentration of red blood cells in healthy group and higher concentration of white cells in the leukemic group.

The hypercellularity due to the presence of blasts in the leukemias causes an increase in the metabolic rate of the cells,

resulting in greater consumption of amino acids and proteins that participate in specific enzymatic actions, such as the tyrosine kinase. This enzyme has the function of phosphorylation of protein substrates, such as the myeloperoxidase, an enzyme present in some types of acute myeloid leukemias.[44–46] This may explain the fact that PC2 and PC3 exhibit some peaks related to amino acids with higher intensities in the healthy compared to leukemic groups, especially Tyr and Trp. The peaks of proteins and glutathione (linked to erythrocytes) are also found in lower intensities in the leukemic group, which was also observed by Sanches et al.,[47] who carried out a study comparing the biochemical profile between leukemic patients and healthy individuals. The peak assigned to the amino acid Phe was highlighted in the first three loadings. Although in low amounts, the literature describes that this amino acid is one of the activators of the BCR-ABL gene, which present in a class of chronic myeloid leukemia and other myeloproliferative disorders.[44] Its presence in the first loadings would suggest that Phe may play an important role in acute leukemias.

Figure 4 shows the plot of the PCs and SCs of plasma samples. The PC1 has characteristic peaks of plasma, but there was no statistically significant difference in SC1 for healthy versus leukemic groups ($p > 0.05$); therefore, such peaks are not relevant for the differentiation of the groups studied. PC2 shows negative peaks at 1630 cm^{-1} in the leukemic group and positive peaks at 947, 1004, 1159, 1348, 1407, 1453 and 1520 cm^{-1} in the healthy group, with a significant difference for SC2 ($p < 0.01$). These peaks have assignments related to proteins, amino acids, lipids, and carotenoids. The low intensity of SC2 suggests that the difference in the concentration of these constituents is small, but significant in the groups ($p < 0.01$), being that the healthy group has higher concentrations of proteins, Trp, Phe, glutathione, phospholipids (free), and carotenoids than the leukemic group. PC3 shows peaks with negative intensities in the leukemic group and positive in the healthy group, with a significant difference for SC3 ($p < 0.05$). This component has positive peaks at 1003 and 1433 cm^{-1} and negative at 1007, 1159, 1344, and 1527 cm^{-1}, assigned to proteins, amino acids, and carotenoids. The low intensity of SC3 suggests that the biochemical elements of these peaks are not useful for the differentiation between the healthy group and leukemic group. PC4 showed a statistically significant difference in the intensities of SC4 ($p < 0.001$), with intense negative peaks at 897, 959, 1004, 1160, 1447, and 1525 cm^{-1}, referring to the group of proteins, amino acids, and carotenoids, suggesting a higher concentration of these components in the healthy

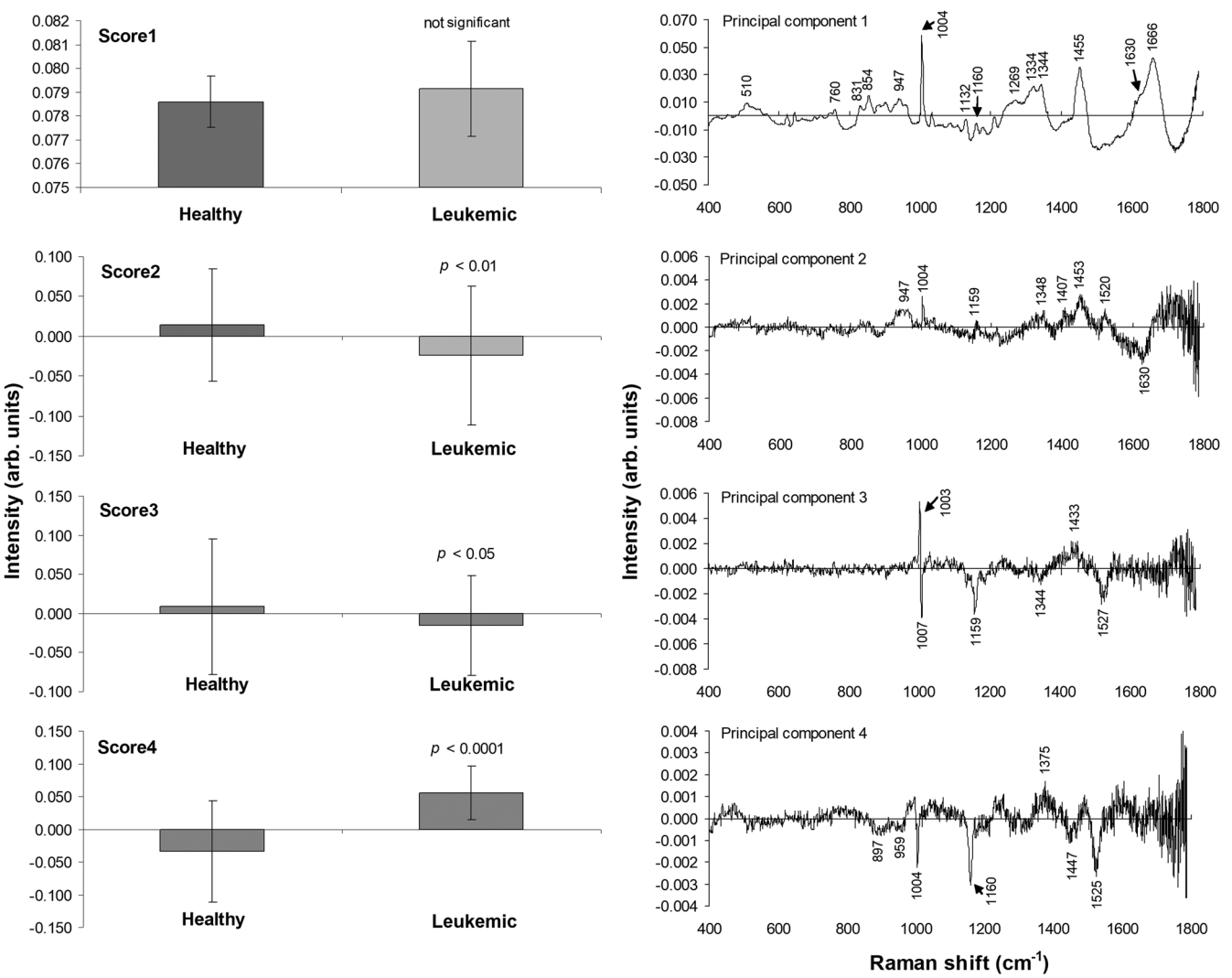

Fig. 4 Plots of the PCs and SCs calculated for the plasma samples for exploratory analysis.

group, and a positive peak at 1375 cm^{-1} attributed to glucosamine, which presents in a low concentration in the healthy group compared to the leukemic group.

Regarding the carotenoid peaks in plasma, the exploratory analysis by PCA showed that both PC2 and PC4 presented peaks of this component (1004, 1159/1160, and 1520/1525 cm^{-1}),[24,30] and SC2 and SC4 indicated significant differences, indicating the presence of these components in high concentrations in the healthy group, as demonstrated by González-Solís et al.,[24] when analyzing Raman spectra of blood in leukemic subjects. The literature has shown that healthy individuals have a high plasma carotenoid concentration and that carotenoids act as protectors against neoplasms (antioncogenic), such as acute leukemias.[48,49] Another study showed that carotenoid concentration increased in leukemic individuals in remission.[24]

3.3 *Discriminant Analysis (PLS-DA and PCA-DA)*

Discrimination techniques based on PLS and PCA (PLS-DA and PCA-DA, respectively) applied to spectral data have been used with relative success as predictors of discrimination or differentiation between healthy and diseased tissues, especially in oncology.[37,50,51] The DA using PLS and PCA was applied to the normalized spectra of the healthy and leukemic groups using the entire spectral range (400 to 1800 cm^{-1}) for both whole blood and plasma samples. The "leave-one-out" cross-validation method defined an initial condition of 10 latent variables (PLS-DA) and 10 PCs (PCA-DA) to be modeled.[39]

The results of the discrimination models using the spectra of whole blood and plasma from each sample group were tabulated (confusion table, Table 3), presenting sensitivity, specificity, and overall accuracy values.[52] For the whole blood, using the PLS-DA, the maximum accuracy occurred by using the first three latent variables, with a value of 96.5% success classification and 91.9% sensitivity, while the maximum accuracy using PCA-DA occurred by using the first four PCs, with a 93.0% success classification and 83.9% sensitivity. Both discriminant models reached 100% specificity, showing that healthy individuals presented a spectral profile capable of allocating them to the healthy group independently on the model used. For the plasma, using the PLS-DA, the maximum accuracy occurred using the first four latent variables, with a 97.1% of success classification, 95.7% sensitivity, and 98.0% specificity, while the accuracy using PCA-DA occurred using the first PC, with 64.1% of success classification, 11.6% sensitivity, and 100% specificity.

Figure 5 shows the confusion plot with the resulting PLS-DA and PCA-DA discrimination for whole blood and plasma, respectively, as shown in Table 3. The sensitivity, specificity, and accuracy values for the PLS-DA were effective in both types of samples, being the results for the correct classification for both whole blood and plasma samples close to each other (around 97%). In general, the results for both PLS-DA and PCA-DA models were more favorable to the whole blood samples (96.5% and 93.0%, respectively), defining this sample as a good option for the differentiation of the leukemic from the healthy group, with spectral variables related to the red blood cells and white cells which allowed to obtain a reduced set of predictors with greater potential for success in the discrimination.

In the PCA-DA, the results obtained in the whole blood were high, demonstrating that the presence of blood cells had relevance in the description of the differences in the groups by the PCs, while the lower biochemical constituents of the plasma reduced the sensitivity of the PCA model, resulting in an accuracy of 64.1%. Gonzáles-Solís et al.[24] used RS in the serum to monitor patients with acute leukemia under chemotherapeutic

Table 3 Confusion matrix with the results of sensitivity, specificity, and correct classification for the discrimination model using the Raman spectra of whole blood and plasma samples.

Diagnosis by conventional methods	Raman diagnostics/PLS-DA		Raman diagnostics/PCA-DA	
	Healthy	Leukemic	Healthy	Leukemic
	Whole blood			
Whole blood healthy ($n = 21$)	80	0	80	00
Whole blood leukemic ($n = 17$)	5	57	10	52
Sensitivity	91.9%		83.9%	
Specificity	100%		100%	
Correct classification (accuracy)	96.5%		93.0%	
	Plasma			
Healthy plasma ($n = 25$)	99	02	101	00
Leukemic plasma ($n = 15$)	03	66	61	08
Sensitivity	95.7%		11.6%	
Specificity	98.0%		100%	
Correct classification (accuracy)	97.1%		64.1%	

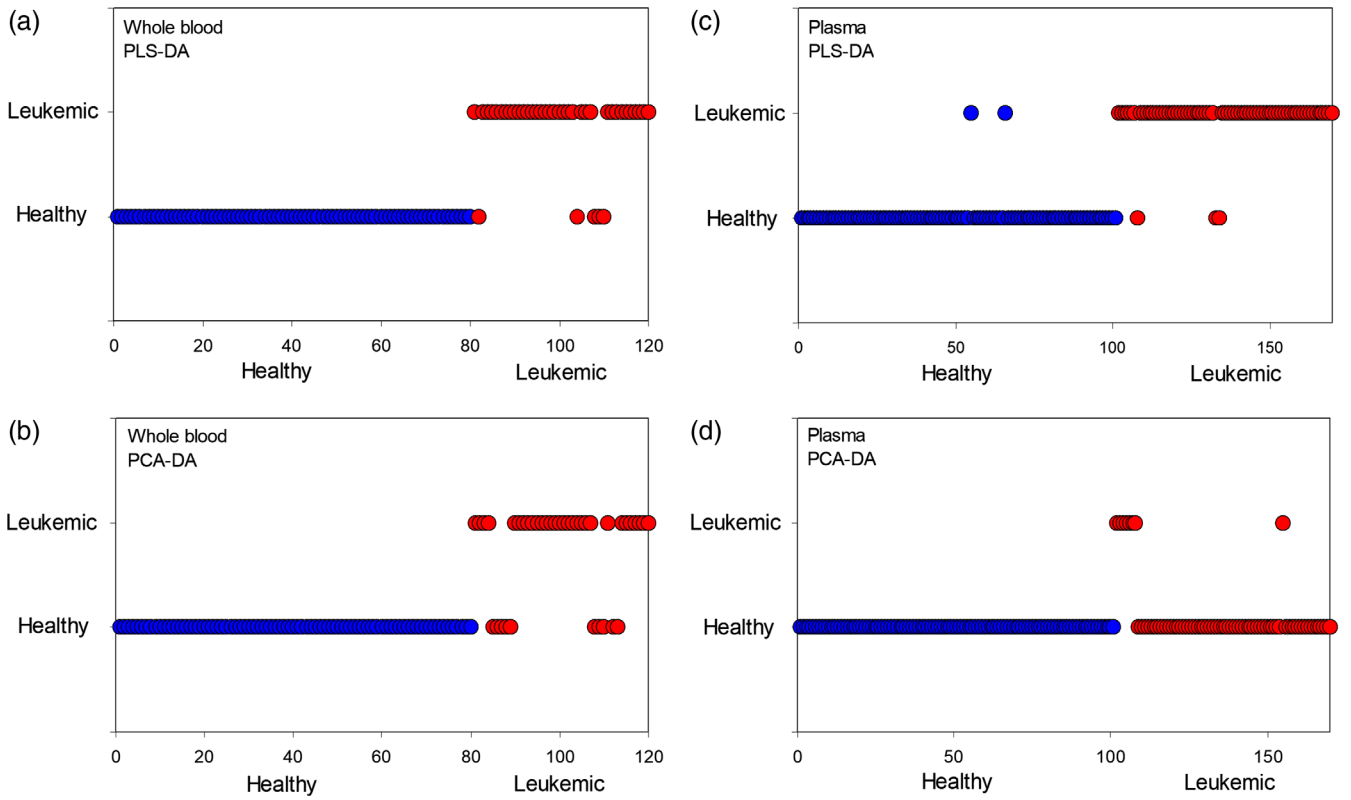

Fig. 5 Confusion plot with the classifications by healthy group and leukemic group, through the DA, being (a) PLS-DA (whole blood); (b) PCA-DA (whole blood); (c) PLS-DA (plasma); and (d) PCA-DA (plasma).

treatment, and through PCA and LDA, successfully identified sample groups belonging to healthy and leukemic individuals, with significant differences between biochemical components, such as proteins, amino acids, and carotenoids, which resemble the results of the research presented here.

The use of multivariate analysis with the objective of identifying and classifying sample groups in the most diverse areas of knowledge is well known.[24,34,35,37,42,51] The PCA aims to perform the segregation of groups based on the variances that, maximized, support the classification of samples according to the differences between the groups, and has shown to be an important tool when the main information capable of differentiating the sample groups is just the variability between groups, which needs to be greater than the intragroup variability.[38,39] However, the PLS-DA stands out from the PCA-DA because, in addition to the information of the differences between the groups, the variances obtained within each group are also recognized, and these variances are associated to the groups when modeling the regression curve;[38–40] therefore, the PLS-DA leads to better performance in the classification of samples when compared to PCA-DA.[38,39]

Compared to the techniques currently available for the diagnosis of acute leukemias based on cytomorphology and immunophenotyping,[9–11,13,14] RS model presented accuracy (97.1%) close to phenotyping (94%[13]) and sensitivity and specificity (95.7% and 98.0%) close or overmatch the flow cytometry technique (97% and 88%[14]), despite the difficulty in obtaining the sensitivity and specificity reference values for the standard diagnostic techniques of myelogram and phenotyping. Thus, Raman-based analysis could significantly differentiate between healthy and leukemic individuals based on differences in the spectral pattern related to the chemical composition (amino acids and carotenoids) of the blood plasma, and cellularity (red and white cells) and chemical composition (proteins) in the whole blood using a small volume of peripheral blood. This infers a minimally invasive character when obtaining the samples to be diagnosed.

RS does not require the use of reagents or special preparation of the sample, spectral analysis can be done in reduced time and mathematical and statistical algorithms can lead to a simplification in the analysis and interpretation of the results. The RS technique can be included in the list of diagnostic screening tests, since its high specificity makes it possible to exclude the suspicion of acute leukemia, allowing physicians to extend the diagnostic investigation to other diseases, defining the therapeutic conduction on their patients early, since the time between diagnosis and treatment is crucial for a good prognosis. The technique can benefit from compact Raman systems that resemble bedside instrumentation for the point-of-care use,[53,54] where blood and serum samples recently withdrawn can be evaluated rapidly and nondestructively, with advantages when compared to established screening techniques, such as blood count and myelogram. The Raman technique may establish a new technology for rapid and accurate diagnosis of complex diseases related to human blood including leukemias.

4 Conclusion

In this study, RS has been applied to samples of whole blood and plasma to identify spectral differences among healthy subjects and acute leukemic patients based on their biochemical components (proteins, amino acids, carbohydrates, lipids, and carotenoids). Exploratory analysis by PCA applied to the spectra of

whole blood revealed that the PC loadings 2 and 3 presented peaks attributed to proteins, amino acids, and carbohydrates, showing higher intensity in the healthy group, while in the leukemic group, particularly the loading 3, presented higher intensity of phospholipids constituents of the cell's membrane due to the characteristic hypercellularity of acute leukemias. In plasma, PCA identified major differences in PC loadings 2 and 4, where the healthy group showed peaks that indicate higher amount of proteins, amino acids, free phospholipids, and carotenoids compared to the leukemic group. The discrimination model based on PLS applied to whole blood spectra showed superiority in the discrimination of healthy from leukemic group in relation to plasma, with sensitivity of 91.9%, specificity of 100%, and accuracy of 96.5%; PLS also presented better results in the classification of the groups using plasma when compared to the discrimination by PCA, with sensitivity of 95.7%, specificity of 98%, and accuracy of 97.1%. RS has shown potential as a diagnostic tool for acute leukemia to increase the range of techniques currently available for screening and confirming the acute leukemia, quickly and minimally invasive with high sensitivity and specificity using both whole blood and plasma.

Disclosures

All authors declare no conflicts of interests.

Acknowledgments

L.S. acknowledges São Paulo Research Foundation—FAPESP (Grant No. 2009/01788-5) and National Council for Scientific and Technological Development— CNPq (Grant No. 306344/2017-3). A.M.S. and F.S.S.O. acknowledge Coordination of Superior Level Staff Improvement - CAPES-PROSUP for the doctorate fellowship and Universidade Anhembi Morumbi—UAM for the financial support.

References

1. E. Matutes et al., "Definition of acute biphenotypic leukemia," *Haematologica* **82**, 64–66 (1997).
2. ASH—American Society of Hematology, "Leukemia," http://www.hematology.org/Patients/Cancers/Leukemia.aspx (10 August 2017).
3. M. A. Sekeres, "Treatment of older adults with acute myeloid leukemia: state of the art and current perspectives," *Haematologica* **93**, 1769–1772 (2008).
4. WHO—World Health Organization, "Incidence, mortality & survival databases," http://www.who.int/cancer/resources/incidences/en/ (30 May 2018)
5. INCA—National Cancer Institute José Alencar Gomes da Silva, "Cancer estimates for 2018, INCA," http://www.inca.gov.br/estimativa/2018/casos-taxas-brasil.asp (13 May 2018).
6. E. Stieglitz and M. L. Loh, "Genetic predisposition to childhood leukemia," *Ther. Adv. Hematol.* **4**(4), 270–290 (2013).
7. J. M. Bennett et al., "The morphological classification of acute lymphoblastic leukaemia: concordance among observers and clinical correlations," *Br. J. Hematol.* **47**(4), 553–561 (1981).
8. C. C. Chen et al., "Acute leukemia presenting with extramedullary diseases and completely normal hemogram: an extremely unusual manifestation unique Topre-B ALL," *Am. J. Hematol.* **85**(9), 729–731 (2010).
9. G. C. Silva et al., "Laboratory diagnosis of acute myeloid leukemias," *J. Bras. Patol. Med. Lab.* **42**(2), 77–84 (2006).
10. G. W. Dewald, "Cytogenetic and FISH studies in myelodisplasia, acute myeloid leukemia, cronic lymphocytic leukemia and lymphoma," *Int. J. Hematol.* **76**(Suppl. 2), 65–74 (2002).
11. D. A. Arber et al., "The 2016 revision to the World Health Organization classification of myeloid neoplasm and acute leukemia," *Blood* **127**(20), 2391–2405 (2016).
12. T. Haferlach et al., "Genetic classification of acute myeloid leukemia (AML)," *Ann. Hematol.* **83**(1), S97–S100 (2004).
13. A. Al-Mawali et al., "Incidence, sensitivity, and specificity of leukemia-associated phenotypes in acute myeloid leukemia using specific five-color multiparameter flow cytometry," *Am. J. Clin. Pathol.* **129**(6), 934–945 (2008).
14. R. Paredes-Aguilera et al., "Flow cytometric analysis of cell-surface and intracellular antigens in the diagnosis of acute leukemia," *Am. J. Hematol.* **68**(2), 69–74 (2001).
15. M. C. Lima et al., "Acute myeloid leukemia: analysis of epidemiological profile and survival rate," *J. Pediatr.* **92**(3), 283–289 (2002).
16. A. Mahadevan-Jansen and R. Richards-Kortum, "Raman spectroscopy for the detection of cancer and precancer," *J. Biomed. Opt.* **1**(1), 31–71 (1996).
17. L. Yongzeng et al., "Micro-Raman spectroscopy study of cancerous and normal nasopharyngeal tissues," *J. Biomed. Opt.* **18**(2), 027003 (2013)
18. A. Sahu et al., "Oral cancer screening: serum Raman spectroscopic approach," *J. Biomed. Opt.* **20**(11), 115006 (2015).
19. E. Ryzhikova et al., "Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia," *J. Biophotonics* **8**(7), 584–596 (2015).
20. N. Ralbovsky and I. K. Lednev, "Raman hyperspectroscopy shows promise for diagnosis of Alzheimer's," *Biophotonics* **4**(25), 33–37 (2018).
21. C. G. Atkins et al., "Raman spectroscopy of blood and blood components," *Appl. Spectrosc.* **71**(5), 767–793 (2017).
22. R. Vanna et al., "Label-free imaging and identification of typical cells of acute myeloid leukaemia and myelodysplastic syndrome by Raman microspectroscopy," *Analyst* **140**(4), 1054–1064 (2015).
23. T. Happillon et al., "Diagnosis approach of chronic lymphocytic leukemia on unstained blood smears using Raman microspectroscopy and supervised classification," *Analyst* **140**(13), 4465–4472 (2015).
24. J. L. González-Solís et al., "Monitoring of chemotherapy leukemia treatment using Raman spectroscopy and principal component analysis," *Lasers Med. Sci.* **29**(3), 1241–1249 (2014).
25. R. Petry, J. Popp, and M. Schimitt, "Raman spectroscopy - a prospective tool in the life sciences," *ChemPhysChem.* **4**(1), 14–30 (2003).
26. E. Cordero et al., "In-vivo Raman spectroscopy: from basics to applications," *J. Biomed. Opt.* **23**(7), 071210 (2018).
27. C. Matthäus et al., "Infrared and Raman microscopy in cell biology," *Method Cell. Biol.* **89**, 275–308 (2008).
28. W. E. Huang et al., "Raman microscopic analysis of single microbial cells," *Anal. Chem.* **76**(15), 4452–4458 (2004).
29. E. B. Hanlon et al., "Prospects for in vivo Raman spectroscopy," *Phys. Med. Biol.* **45**(2), R1–R59 (2000).
30. A. Bankapur et al., "Raman tweezers spectroscopy of live, single red and white blood cells," *PLoS One* **5**(4), e10427 (2010).
31. H. Dohner et al., "Diagnosis and management of acute myeloid leukemia in adults: recommendations from an International Expert Panel, on behalf of the European LeukemiaNet," *Blood* **115**(3), 453–474 (2009).
32. P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometr. Intell. Lab. Syst.* **117**, 100–114 (2012).
33. J. Neyman, "Outline of a theory of statistical estimation based on the classical theory of probability," *Philos. Trans. R. Soc. London A* **236**, 333–380 (1937).
34. F. S. de Siqueira e Oliveira, H. E. Giana, and L. Silveira, "Discrimination of selected species of pathogenic bacteria using near-infrared Raman spectroscopy and principal components analysis," *J. Biomed. Opt.* **17**(10), 107004 (2012).
35. F. L. Silveira et al., "Discrimination of non-melanoma skin lesions from non-tumor human skin tissues in vivo using Raman spectroscopy and multivariate statistics," *Lasers Surg. Med.* **47**(1), 6–16 (2015).
36. S. A. Morellato and C. A. R. Diniz, "PLS regression models with heteroscedastic errors," Dissertation, UFSCar—Universidade Federal de São Carlos, São Carlos (2010).
37. D. V. Nguyen and D. M. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics* **18**(9), 1216–1226 (2002).
38. M. Barker and W. Rayens, "Partial least squares for discrimination," *J. Chemometrics* **17**, 166–173 (2007).

39. Y. Liu and W. Rayens, "PLS and dimension reduction for classification," *Comput. Stat.* **22**(2), 189–208 (2007).
40. M. M. C. Ferreira et al., "Chemometrics I: multivariate calibration, a tutorial," *Quim. Nova* **22**(5), 724–731 (1999).
41. C. A. Nunes et al., "Chemoface: a novel free user-friendly interface for chemometrics," *J. Braz. Chem. Soc.* **23**(11), 2003–2010 (2012).
42. D. V. Nguyen and D. M. Rocke, "Tumor classifications by partial least squares using microarray gene expression data," *Bioinformatics* **18**(1), 39–50 (2002).
43. E. G. Santos, "Development of a rapid and low cost methodology for the diagnosis of sickle cell anemia," Thesis, USP—Universidade de São Paulo, São Paulo (2014).
44. R. L. Levine and D. G. Gilliland, "Myeloproliferative disorders," *Blood* **112**, 2190–2198 (2008).
45. B. H. Yip and C. W. So, "Mixed lineage leukemia protein in normal and leukemic stem cells," *Exp. Biol. Med.* **238**(3), 315–323 (2013).
46. M. Ohanian et al., "Tyrosine kinase inhibitors in acute and cronic leukemias," *Expert. Opin. Pharmacother.* **13**(7), 927–938 (2012).
47. F. L. Sanches et al., "Comparison of biochemical and immunological profile of pediatric patients with acute myeloid leukemia in relation to healthy individuals," *J. Pediatr.* **91**, 478–484 (2015).
48. G. F. Silva, "Carotenoids: a possible protection against the development of cancer," *Rev. Nutr.* **20**(5), 537–548 (2007).
49. H. Nishino et al., "Cancer prevention by natural carotenoids," *Biofactors* **13**(1–4), 89–94 (2000).
50. G. Fort and S. Lambert-Lacroix, "Classification using partial least squares with penalized logistic regression," *Bioinf. J.* **21**(7), 1104–1111 (2005).
51. F. L. Cals et al., "Investigation of the potential of Raman spectroscopy for oral cancer detection in surgical margins," *Lab. Invest.* **95**, 1186–1196 (2015).
52. R. Parikh et al., "Understanding and using sensitivity, specificity and predictive values," *Indian J. Ophthalmol.* **56**(1), 45–50 (2008).
53. I. Pence and A. Mahadevan-Jansen, "Clinical instrumentation and applications of Raman spectroscopy," *Chem. Soc. Rev.* **45**(7), 1958–1979 (2016).
54. J. Hutchings et al., "Rapid Raman microscopic imaging for potential histological screening," *Proc. SPIE* **6853**, 685305 (2008).

Adriano Moraes da Silva graduated in biomedicine from the University of Mogi das Cruzes and received his master's degree in biomedical engineering from the University of Vale do Paraíba. In 2018, he earned his PhD in biomedical engineering for defending his thesis for diagnosis of acute leukemias through Raman spectroscopy and partial least squares analysis. Currently, he is a professor of clinical hematology and coordinator of the biomedicine courses at Universidade Paulista.

Fernanda Sant Ana de Siqueira e Oliveira holds her bachelor's degree in biomedicine from the University of Mogi das Cruzes, and her master's degree in biomedical engineering from Camilo Castelo Branco University. In 2018, she received her PhD in biomedical engineering for her work on biochemical characterization of bacteria through Raman spectroscopy. Currently, she is a professor at the Institute of Health Sciences, Universidade Paulista, São José dos Campos, São Paulo, Brazil.

Pedro Luiz de Brito graduated in medicine from the Faculty of Medicine, University of São Paulo (FMUSP) in 1984, with specialization in pediatric surgery at the Clinical Hospital of FMUSP. He is a specialist in pediatric surgery, as recognized by the Brazilian Society of Pediatric Surgery, since 1990. He received his PhD in surgery and experimentation from the Federal University of São Paulo (UNIFESP) in 2011.

Landulfo Silveira Jr. received his bachelor's degree in electrical engineering and master's degree in bioengineering in 1994 and 1998, respectively. In 2001, he obtained his doctor of science degree from the FMUSP. Currently, he is a professor at the Universidade Anhembi Morumbi for the master's and doctorate programs of biomedical engineering, with the research interests in optical spectroscopy applied to diagnosis in tissues and fluids, and optical instrumentation.