ATR-FTIR spectroscopy for virus identification: A powerful alternative

Marfran C.D. Santos [a], Camilo L.M. Morais [b] and Kássio M.G. Lima [a],*

[a] *Química Biológica e Quimiometría, Instituto de Química, Universidade Federal do Rio Grande do Norte, Natal 59072-970, Brasil*
[b] *School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom*

Abstract. In pandemic times, like the one we are witnessing for COVID-19, the discussion about new efficient and rapid techniques for diagnosis of diseases is more evident. In this mini-review, we present to the virological scientific community the potential of attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy as a diagnosis technique. Herein, we explain the operation of this technique, as well as its advantages over standard methods. In addition, we also present the multivariate analysis tools that can be used to extract useful information from the data towards classification purposes. Tools such as Principal Component Analysis (PCA), Successive Projections Algorithm (SPA), Genetic Algorithm (GA) and Linear and Quadratic Discriminant Analysis (LDA and QDA) are covered, including examples of published studies. Finally, the advantages and disadvantages of ATR-FTIR spectroscopy are emphasized, as well as future prospects in this field of study that is only growing. One of the main aims of this paper is to encourage the scientific community to explore the potential of this spectroscopic tool to detect changes in biological samples such as those caused by the presence of viruses.

Keywords: ATR-FTIR spectroscopy, virus identification, multivariate classification

1. Introduction

Diseases caused by viruses are among the main public health problems. There is an incalculable number of viruses circulating in our environment, many already known by the scientific community and several still unknown. The Human Immunodeficiency Virus (HIV) [8] and arboviruses such as Dengue [44], Zika [38], Chikungunya [13] and Yellow Fever [26,27] are examples of well-known viruses that cause great damage to society, either because of their severity or because of their ability to change, giving rise to new serotypes. The Dengue virus, for example, is found in four different serotypes (DENV-1, -2, -3 and -4), so the same individual can contract the dengue virus up to four times [44]. In the case of new viruses, the most important case in the world is currently the new Coronavirus (COVID-19), responsible for the outbreak that began in December 2019 in Wuhan, China, and which is already in a pandemic situation, given the seriousness in which it occurs. As of March 11, 2020, more than 118,000 cases have been confirmed for COVID-19 in 114 countries, of which 4,291 people have died. [15,21,36,60,62,63]

Outbreaks such as COVID-19, currently in evidence, as well as other recent outbreaks, such as the Ebola Virus outbreak in Guinea in 2014 [3], the various outbreaks of arboviruses (Dengue, Zika, Chikungunya, Yellow Fever) in tropical countries at times of the year due to transmitting mosquitoes, and the

---

*Corresponding author. E-mail: kassiolima@gmail.com.

constant public health problem with the various influenza viruses [54], such as the H1N1 virus that emerged in Mexico and the United States of America in the first half of 2009, being also declared a pandemic state by the World Health Organization (WHO) [12], make evident the importance of reliable, accurate and fast diagnostic methods. It is evident that fast response (diagnosis) means fast treatment and less damage caused by the illnesses.

The most commonly available methods in diagnostic clinics or hospitals are serological methods. These methods are based on the detection of antibodies produced against the viruses. Once a certain antibody is detected, there is evidence of which virus is present. [5] This is the case, for example, of the widely used enzyme-linked immunosorbent assay (ELISA) method. [30] The big problem with using techniques based on the detection of antibodies is that, generally, for viruses of the same family as Dengue, Zika and Yellow fever, for example, cross reactions can occur. That is, the immune system produces antibodies to a virus different from the virus that is present in the body. Studies also show that the production of specific antibodies to Dengue can worsen the condition of patients infected with Zika virus. [9] This suggests that the same can occur for other viruses and from different families.

Spectroscopic techniques are based on the interaction between electromagnetic radiation with the sample. This interaction can provide valuable information from the sample compositional point of view. The technique of attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR) is one of the most well-known spectroscopic techniques, which works in the mid-infrared region. [24] This region of the electromagnetic spectrum comprises the 4000 to 400 cm^{-1} range. In biological samples, the range between 1800 to 900 cm^{-1} is known as the biofingerprint region because it has a high density of information regarding important biomolecules. The field of study where spectroscopic tools are used to analyze biological samples has become known as biospectroscopy [4,17], and has been widely used in chemometric approaches involving the identification of bacteria [22,23], viruses [42,43], cancer diagnosis [56], forensic entomology [2], among others.

The infrared spectra obtained using ATR-FTIR are an example of multivariate data. This is because in each spectrum there are several variables (wavenumbers) that inform the absorbance for a sample. To better interpret this type of data, multivariate analysis techniques can be used; this comprises statistical and mathematical tools capable of analyzing the data and provide reliable quantification or classification responses. [17,24]

In this mini-review, we will address the use of ATR-FTIR in conjunction with multivariate analysis techniques for classification and detection of viruses. We will highlight the operation of ATR-FTIR and some multivariate classification techniques, such as Principal Component Analysis (PCA), Successive Projections Algorithm (SPA), Genetic Algorithm (GA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). The potential of this approach, the limitations and future perspectives in the field are discussed in order to disseminate to the medical and virological community this methodology that has great potential in detecting viruses in a quick, accurate and reliable fashion.

2. General structure of viruses

It is important to know the biomolecular structure to be able to associate spectral information with the presence or absence of viruses. Viruses can vary in structure, size and even composition, however, in most cases there are similarities. Viruses are mainly made up of proteins. A general structure can be described as an outer membrane formed by a lipid bilayer, which is called a viral envelope (in the case of enveloped viruses); surface proteins responsible for the first contact of the virus with host cells are usually found encrusted in the viral envelope; and more internally, there is another protein membrane called
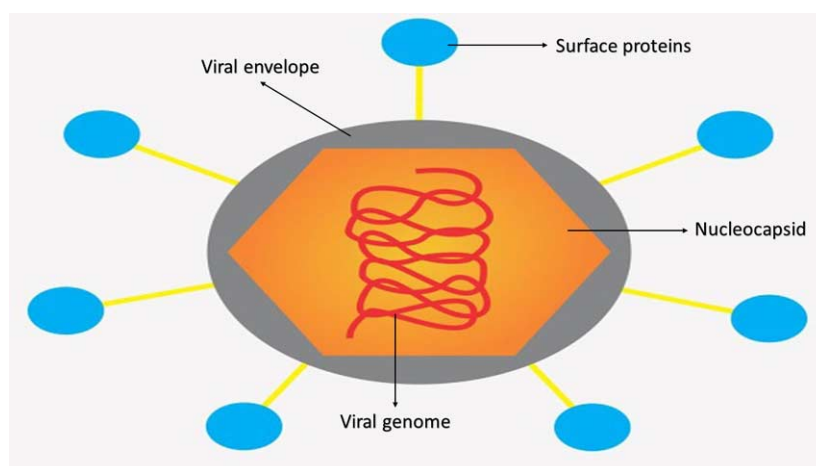
Fig. 1. General structure of a flavivirus with its identified parts. For viruses such as Dengue, Zika, Chikungunya and Yellow Fever, the viral genome is a positive single-stranded RNA according to the Baltimore Class IV classification.

a capsid (or nucleocapsid), which is responsible for protecting genetic information (RNA or DNA). [14,19,40] Fig. 1 shows a scheme for the general structure for viruses such as Dengue, Zika, Chikungunya, Yellow fever, among others. For these viruses, the viral genome is a positive single-stranded RNA according to the Baltimore Class IV classification.

3. WHO methods for detecting Dengue, Zika and Chikungunya

In the literature, several research articles and review articles that address the main methods of virus diagnosis can be found for Influenza [59], Ebola [7,11], Hepatitis B [33], Dengue [5,60], Zika [16,37,49], Chikungunya [1,32], among others. In view of the large number of existing viruses, here we will focus on the arboviruses Dengue, Zika and Chikungunya, since they are viruses that cause very similar symptoms and coexist in the same geographic regions. Viruses that coexist and exhibit similar symptoms are very common worldwide, and since the ELISA and PCR diagnostic methods are used for the vast majority, the discussion in this topic can be used for other viruses as well.

In 2017, the WHO published a document entitled *"Tool for the diagnosis and care of patients with suspected arboviral diseases"* where methods are provided that can be used in differential diagnosis by laboratories with installed capacity to identify Dengue, Chikungunya and Zika viruses antigenically, molecularly and serologically. [57] In Fig. 2, we can see the methods for the differential diagnosis using RT-PCR (Fig. 2a), and for differential diagnosis using ELISA (Fig. 2b), according to WHO. As demonstrated in the figure, it is necessary to repeat the analysis several times for both RT-PCR and ELISA method. This is because each analysis is specific to each virus. Thus, it is necessary to rule out co-infections. Still, for the ELISA technique, the diagnosis is not accurate. At best, infection with a particular virus can only be presumed.

4. ATR-FTIR spectroscopy

ATR-FTIR spectroscopy can be an advantageous alternative both from the point of view of time consumed for analysis and financial, since the use of reagents is not required. When interrogating biological
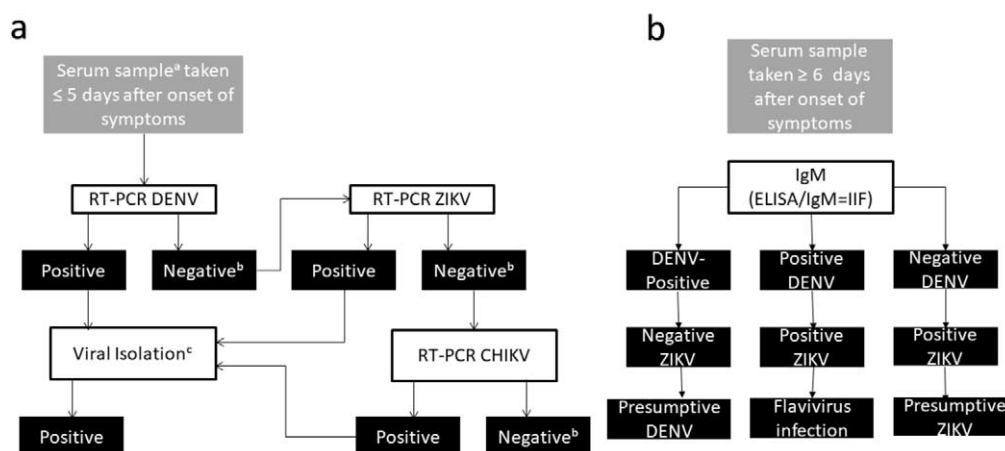
Fig. 2. Algorithm for laboratory diagnosis for (a) RT-PCA: suspected cases of arbovirus: acute phase; (b) ELISA: suspected cases of arbovirus: convalescent phase. Inspired by reference [57]. [a]A urine sample is also recommended for PCR ZIKV. [b]Consider dengue NS1 antigen for determining DENV infection. [c]Isolation is not required in order to confirm infection; it is considered complementary information for identifying serotypes, genotypes, and strains of the arbovirus in question.

samples with radiation in the mid-infrared region (4000–400 cm^{-1}), the region of greatest interest is that between 1800–900 cm^{-1}. This region is called the "biofingerprint" because it provides a large amount of information and representation of the composition of interrogated samples based on the vibrational modes of chemical bonds. Figure 3 shows a typical biological spectrum showing biomolecular peak assignments from 1,800 to 900 cm^{-1}, where biochemical assignments are shown. [17] Since nucleic acids are specific to each virus, differences in nucleic acid content are expected to be a key factor for an accurate assessment of ATR-FTIR. In addition, the complexity and dynamics of cellular and viral epitranscriptomes must be taken into account during infection. This has been studied in depth by McIntyre *et al.* [25], where the diversity of the viral epitranscriptome was first observed, which far exceeds the few modifications reported so far. All these possible changes that occur during infection by the virus +ssRNA must be taken into account, and may even be potential markers to identify the stage of infection. A characteristic IR spectrum can be derived by ATR-FTIR spectroscopy, FTIR microspectroscopy, or photothermal microspectroscopy. [24] Here, we will focus on the ATR-FTIR technique.

When acquiring spectra using an ATR-FTIR, a crystal (*e.g.*, diamond size 250 μm × 250 μm) is used to cause total reflection. The sample (*e.g.*, 10 μL) is then placed on this crystal and the infrared beam is passed through the crystal and fully reflected internally (without refraction). Samples can measured in either liquid or dry state. Preferably, samples should be allowed to dry before measurement to reduce the water content, hence, reducing OH interfering in the IR spectrum. Nevertheless, ATR also allows samples to be measured as liquid; in this case, one must be careful with OH absorptions at 3650–3600 cm^{-1} (free O-H stretching) and 3400–3300 cm^{-1} (hydrogen-bonded O-H stretching). [29] In liquid state, the bands are broaden, which might mask the lipids region at around 3000 cm^{-1}; therefore, the user must focus on analyzing the spectra on the fingerprint region (1800–900 cm^{-1}) only. In this region, the Amide I and Amide II bands tend to merge, where the Amide II band becomes an arm in the right-hand side of the Amide I peak at around 1650 cm^{-1}. The protein phosphorylation band at around 900 cm^{-1} also tends to be wider for liquid samples.

The process of total reflection generates an evanescent wave that can penetrate for a few micrometers in the sample. The sample then absorbs part of this radiation, attenuating it, so that the instrument
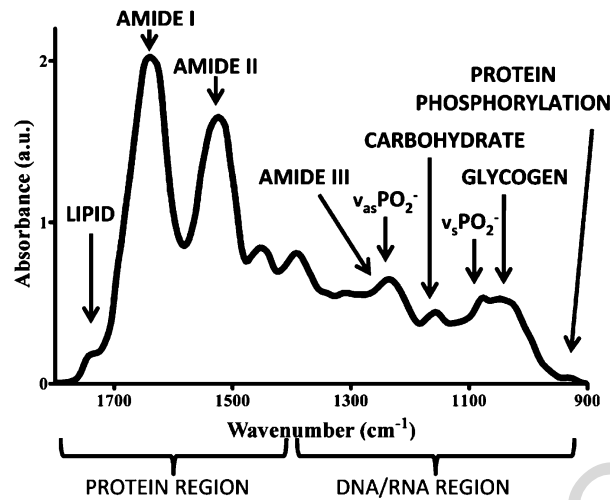
Fig. 3. Example of a biological sample spectrum in the biofingerprint region of the mid infrared range. Reprinted (adapted) with permission from reference [17] (J.G. Kelly, et al., Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers, *Journal of Proteome Research* **10** (2011), 1437–1448). Copyright (2011) American Chemical Society.
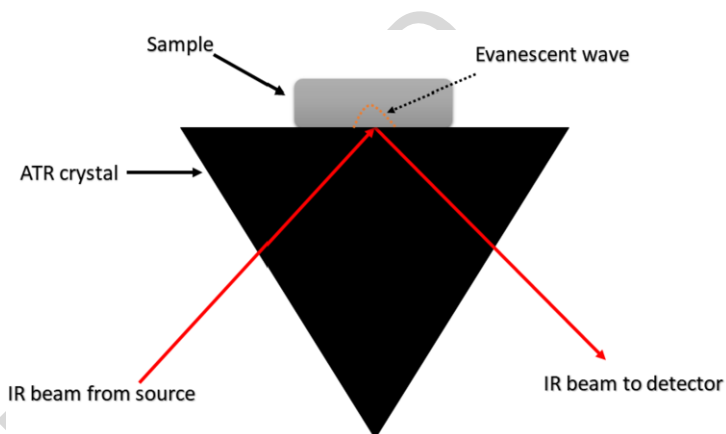

Fig. 4. Illustration of the operation of the ATR device in spectral acquisition. An evanescent wave is generated by the total reflection of the incident radiation. This evanescent wave can penetrate through a few micrometres in the sample, which absorbs part of the radiation. This absorption can be detected by the instrument, generating the spectrum. Inspired by reference [35].

detects when it has been absorbed (or transmitted) and provides a spectrum. [4,17,24,35] Fig. 4 shows the process taking place on the ATR crystal. It is important to note that, as absorption is related to vibrations, the spectra provide compositional information for the analyzed sample in terms of vibration signals. This means that the presence or absence of viruses can translate into spectral variations. However, these variations are subtle, and generally cannot be detected just by viewing the spectra. For this reason, it is necessary to use mathematical tools capable of finding spectral features that best differentiate one class from another (*e.g.*, infected *vs*. uninfected).

5. Viral contribution to the FTIR spectra during cell infection

Infections caused by viruses involve complex processes, with changes in the structures of biomolecules and consequent spectral variations. These changes, of course, could make it difficult to detect these viral infections by ATR-FTIR; however, if properly investigated, they may provide specific information on the infection stage. As an example, we can analyze what happens to viral RNA during the infection phase. As mentioned here earlier, recent studies have demonstrated the dimensions and diversity of the viral epitranscriptome, demonstrating that they are much larger than formerly believed. It was observed that the Zika, Dengue, hepatitis C, poliovirus and human immunodeficiency type 1 viruses significantly altered the global PTM landscape. Direct comparison of viral epitranscriptomes identified specific and MTPs common to all viruses. This suggests that the study of MTPs, that is, the study of changes involving the genetic material of the virus during infection, may be an important way to detect viral infections by ATR-FTIR. Specific modifications of dimethylcytosine were present only in the total RNA of virus-infected cells and in the intracellular RNA of HCV and viral RNA of Zika and HCV virions. On the other hand, ZIKV and DENV encode a methyltransferase responsible for introducing this modification into viral RNA, which helps to guarantee the efficient translation of viral proteins and camouflages the viral RNA from cellular defense mechanisms against foreign RNA. In contrast, HCV and PV have no methyltransferase activity and are known to achieve similar goals by different means. [25] These are just one of several processes that are involved in the complex dynamics of a viral infection. The characteristics of each virus can be a differential for spectral discrimination between one virus and another.

Although ATR-FTIR is a non-target technique greatly affected by the sample environment and the virus type, spectral changes after viral infection will primarily occur due to protein, cell DNA, and RNA changes (Fig. 3). [46] Blood compounds such as total cholesterol, high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol, triglycerides, albumin and total protein has been reported to decrease in concentration in patients infected with hepatitis B and C, while immunoglobulin G (IgG), A (IgA) and M (IgM) increase in concentration in infected patients when compared to controls. [41] Moreover, for hepatitis, several spectral features change for infected patients. [41] Table 1 summarizes these main spectral biomarkers that change after viral infection.

6. Computational procedure

After spectral acquisition, the spectra are imported into some suitable software in order to carry out pre-processing and multivariate analysis. Among the available software options, MATLAB (MathWorks, Inc., United States) stands out. However, other interesting options are The Unscrambler (Camo Analytics, Norway), Pirouette (InfoMetrix, Inc., United States), and freely available software such as GNU Octave (https://www.gnu.org/software/octave/) and R (https://www.r-project.org/). We will discuss the computational procedure from pre-processing techniques that can be used to extract information with high potential to discriminate classes.

6.1. Preprocessing

One of the main steps after spectral acquisition is data pre-processing. Pre-processing can be applied to improve the signal-to-noise ratio of the data, adjust the baseline, among other reasons related to the physical nature of the samples, instruments and environment. In short, pre-processing works by eliminating physical interferences and highlighting the signal of interest.

Table 1

Main spectral markers associated with hepatitis infection. [41]

Class	Assignment	Wavenumber (cm^{-1})
Infected	A-DNA backbone ν(C-C)	966
Infected	RNA, ν(C-O) ribose	1015, 1018
Infected	Proteins (Ig)	985, 988, 989, 1074
Infected	DNA, RNA $\nu_s(PO_2^-)$ or phospholipid $\nu_s(PO_2^-)$	1078
Infected	RNA, ν(C-O) ribose	1067
Infected	Valence vibration of CC and CO in glucose from polysaccharide	1093, 1096, 1100
Infected	RNA, ν(C=O) ribose	1119
Infected	RNA, ν(C=O) ribose	1037
Infected	ν_{as}(CO-O-C) of glycan and nucleic acids (DNA and RNA)	1145, 1148
Infected	A-DNA, $\nu_{as}(PO_2^-)$ and RNA, $\nu_{as}(PO_2^-)$ or phospholipid phosphodiester bond (lipid bilayer) or protein (IgG), Amide III	1223, 1226, 1234, 1249, 1286
Infected	Amide I, protein, β pleated sheet	1635, 1631
Infected	$\nu_a(C_2$=O) vibration in RNA	1691
Infected, Uninfected	Lipids, ν(C=O) ester carbonyl	1732, 1747
Uninfected	Ester C-O-C asymmetric stretching from phospholipid, triglyceride and cholesterol esters	1167, 1174, 1182, 1186
Uninfected	Proteins, lipids, symmetric $\delta(CH_3)$	1379
Uninfected	Proteins, Amide II	1501, 1542, 1591
Uninfected	Proteins, Amide III	1308
Uninfected	Amide I, protein, α-helical or DNA (C=O and C=N)	1650, 1654

The main pre-processing technique used to eliminate noise is Savitzky-Golay smoothing. However, attention must be taken and the correct parameters must be used since this technique can introduce distortions in the spectra and also "smooth" important information. As a result, parameters need to be chosen so that noise is reduced and peaks are valued. [48,58]

For removal of physical distortions in the data, differentiation techniques can be applied to correct baseline problems and solve band overlapping problems. Differentiation works by improving the differences between the spectral bands of interest and the existing baselines. First-order differentiation has been widely used in pre-processing steps, however, second-order differentiation provides symmetry which the absorption value. However, it is important to be careful when using second-order differentiation, as each order of differentiation greatly increase the noise. To reduce this effect, a Savitzky-Golay differentiation can be used which has the implicit smoothing. [48,58] Physical phenomena can also translate into distorted baselines. Rubberband baseline correction is one of the techniques capable of solving this problem. In this approach, a convex polygonal line whose edges are 'valleys' within the spectrum are found. Another possibility is to use manual point baseline correction. In this case, the user chooses the regions of wavenumbers of the polygonal line to be subtracted from the absorption spectra. [17,50] Manual baseline correction has been successfully applied by Zucchiatti et al. [64] In this study, the contribution of ribonucleic acid (RNA) to Fourier transform infrared (FTIR) spectra of eukaryotic cells was investigated. [64]

When analyzing biofluids such as plasma or serum, samples usually contain differences in the concentrations of some species. These concentration differences are one of the main sources of spectral variations between samples, thus affecting information of structural biochemical differences of interest

(*i.e.*, molecular structural changes caused by the presence or absence of viruses) in the spectrum. To minimize this problem, the spectra can be scaled through normalizations using some specific criteria. Two examples of widely used normalization are normalization by the peak of Amide I (1650 cm^{-1}) and normalization by the peak of amide II (1550 cm^{-1}) that can be applied when these peaks are present in all spectra dataset and are not a distinguishing criteria. However, it is not recommended to apply the two normalizations together. [4,58] Other common example of normalization is vector normalization, which is applied after 1st or 2nd derivatives pre-processing. [4,58] It is important to keep in mind that f pre-processing can greatly modify the signal, changing the spectral shape and even physical interpretation.

6.2. *Multivariate analysis*

After the pre-processing step, the mathematical tools capable of finding features in the data that differentiate samples can be applied. Since mid-infrared spectra are multivariate data (each spectrum is composed of many wavenumbers associated with their respective absorbance intensities), the analysis of these data can be done using multivariate analysis techniques. There are numerous techniques that can be used for this purpose, here we will describe some that have been used in studies involving virus analysis.

Principal component analysis (PCA) is an unsupervised analysis technique widely applied in chemometric studies. Undoubtedly, among unsupervised multivariate analysis techniques, PCA is the best known. This technique is called unsupervised because it does not work with training-labelled information, that is, no information is given about the spectra that are provided. Therefore, PCA is used mainly for exploratory analysis where similarities and dissimilarities can be observed between the samples in the principal components space. The principal components are those that have the greatest observed variance in the data after applying a linear transformation to the original data. This transformation provides new variables, which are the principal components (PCs). The PCs are composed of scores, representing the variance on sample direction, and loadings, representing the variance on wavenumber direction. The PCA scores are used to assess similarities/dissimilarities between the samples and the loadings to identify important wavenumbers responsible for the scores distribution pattern. PC1 is the component with the largest amount of information (largest observed variance), PC2 is the one that has the second largest observed variance, and so on. [35] Fig. 5 explains the structure of the PCA model in mathematical terms, where **X** is a $I \times J$ matrix with the rows (I) being the spectra and the columns (J) the variables (wavenumbers); **T** is the scores vector, which are numerical values for each sample present on the selected components; **P** is the loadings vector, which informs the most important variables for the respective component; and **E** is the residual matrix. [6]

Fig. 5. Matrixial representation of the PCA model, where *X* is a matrix with rows representing spectra and columns representing spectral variables (wavenumbers). *T* is a matrix containing the PCA scores, *P* is a matrix containing the PCA loadings, and *E* is a residual matrix. Superscript *T* represents the matrix-transpose operation.

PCA has been applied in the exploration of data for studies involving virus, where, for example, plasma samples with and without HIV virus have been discriminate by PCA [34]. However, this technique is not much efficient in distinguishing samples based on biofluids' spectra in the presence or absence of viruses due to the large sample complexity in some cases. [40,45–47,52] This probably happens due to the high similarity observed in the samples and, consequently, in the spectra. Viruses represent the smallest part of the sample, therefore, spectral differences are minimal. That is, the similarity of the data is high, so that, is not possible to identify differences only by exploring the data in a unsupervised way. For this type of problem, a supervised approach is often necessary.

The term "supervised" refers to the fact that sample label information is provided in the training stage, where a pattern classification or recognition model is trained to recognize and associate information with their respective classes, for later, be proven in a test step where new samples (MIR spectra, in this case) are provided "blindly". In the construction of this model, the samples can be separated into a training and test set randomly, or using sampling algorithms such as Kennard–Stone [18] or MLM [28]. Two of the most common supervised techniques are Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

LDA is one of the most used supervised techniques in chemometric and multivariate classification studies. LDA is based on a Mahalanobis distance calculation between the samples that fits a linear plane of separation perpendicular to the main direction of data variance. [55] In order to obtain the discriminant profile, the LDA calculates the classification score (L_{ik}) for a given class considering that the classes covariance matrices are the same for all classes (Eq. (1)):

$$L_{ik} = (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \sum_{\text{pooled}}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T - 2\log_e \pi_k \tag{1}$$

where $\mathbf{x}_i$ represents a measurement vector for a sample i; $\overline{\mathbf{x}}_k$ represents an average vector of class k; $\sum_{\text{pooled}}$ is the pooled covariance matrix; and π_k is the prior probability of class k. [51]

QDA works similar to LDA; however, in QDA the covariance matrices are not considered equal (if they are really equal, the decision limit will be linear and the QDA will be reduced to LDA). [51] The classification score for QDA is found based on Eq. (2):

$$Q_{ik} = (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \sum_k^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_k) + \log_e |\sum_k| - 2\log_e \pi_k \tag{2}$$

where $\sum_k$ is the variance-covariance matrix of class k; $\log_e |\sum_k|$ is the natural logarithm of the variance-covariance matrix $\sum_k$. The prior probability (π_k), pooled covariance matrix ($\sum_{\text{pooled}}$) and variance-covariance matrix ($\sum_k$) are calculated as follows:

$$\pi_k = \frac{N_k}{N} \tag{3}$$

$$\sum_{\text{pooled}} = \frac{1}{N} \sum_{k=1}^{k} N_k \sum_k \tag{4}$$

$$\sum_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \tag{5}$$

with N_k represents the number of objects of class k, N the total number of samples in the training set, and K the total number of classes. [10,61]

Based on the principle of using covariance matrices that are similar for all classes in LDA, and different in QDA, we can assume that, normally, for more complex data where the variance between the classes is highly different, QDA will obtain a better response. On the other hand, for simpler data sets with unique variance structure, LDA should get better results. Therefore, both LDA or QDA can be applied for supervised classification purposes. However, for big multivariate data such as spectra of biological samples, where the number of spectra variables is often larger than the number of training samples and a high degree of spectral overlapping features is present, a previous method is generally applied to reduce the sample size before applying LDA or QDA. This assists the LDA or QDA in the classification work, since the dataset is reduced and a large amount of redundant information is eliminated. A method widely used for dimensional reduction is precisely the PCA previously described. Therefore, PCA with LDA or QDA can be used together, which is called PCA-LDA or PCA-QDA, respectively.

Other techniques can be used instead of the PCA for feature extraction/selection. For example, techniques of variables selection, such as the Successive Projections Algorithm (SPA) and the Genetic Algorithm (GA). The SPA considers each variable present in the training set as a vector. These vectors are subjected to projection operations resulting in the creation of K chains of variables. A given chain starts with a variable and progressively increases with variables that have the least redundancy in relation to the previous ones (least collinearity) [46], and the collinearity is assessed based on the projections.

GA selects variables based on a computational approach following Darwin's natural selection process. In the process, GA creates an initial population formed by subsets of variables. Each variable is initially randomly assigned a value of 0 (variables not initially selected) or a value of 1 (variables initially selected). Each subset of variables is assigned a fitness value based on a fitness function. Based on this fitness criterium, the selection stage takes place where the subset of lesser fitness is eliminated, and those of greater fitness can be duplicated. In a second step, the mutation and crossover genetic operators can change selected variables to unselected (or the opposite), and cross two subsets of variables, respectively. This whole process is called generation. This process is repeated for a number of generations, and in the end, the subset of the best fitness is that of the selected variables. [53] Finally, it is also possible to associate SPA and GA with LDA or QDA for classification. In this case, we call these algorithms SPA-LDA, SPA-QDA, GA-LDA or GA-QDA.

7. Calculations for quality measurement

To measure the predictive capacity of the model, we rely on true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Based on these values, quality measures such as sensitivity and specificity can be calculated. Sensitivity (SENS) represents the portion of positive samples correctly classified and specificity (SPEC) represents the portion of negative samples correctly classified. Equations (6) and (7) show how to calculate sensitivity and specificity, respectively. Other figures of merit
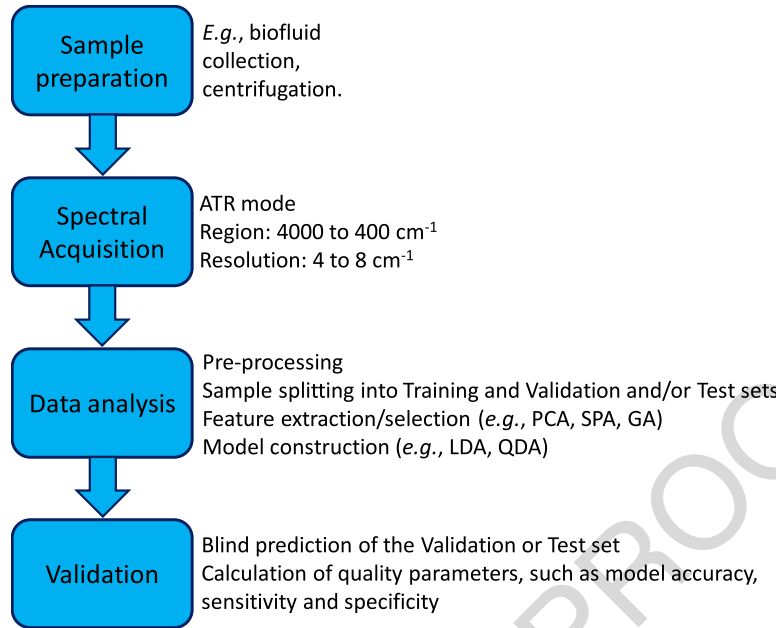
Fig. 6. Flowchart summarizing the fundamental steps of a multivariate classification study based on ATR-FTIR spectra.

can be found easily in the literature. [46]

$$SENS(\%) = \frac{TP}{TP + FN} \times 100 \tag{6}$$

$$SPEC(\%) = \frac{TN}{TN + FP} \times 100 \tag{7}$$

In Fig. 6 we can see a flow chart summarizing the fundamental steps for a study based on the fundamentals presented here.

8. Applications of ATR-FTIR spectroscopy in virus identification

Detection and quantification of poliovirus infections using FTIR spectroscopy and cell culture has been done by Lee-Montiel and collaborators. [20] In this study, the authors explain the problems found in current virus detection methods. They are complex and time-consuming, making detection at the point-of-care difficult. For this, they suggest the use of ATR-FTIR spectroscopy as a fast, sensitive and highly specific method to quantify potentially dangerous viral pathogens and to determine whether suspicious materials contain viable viral particles. Therefore, poliovirus (PV1) was used to evaluate the usefulness of FTIR spectroscopy with cell culture for rapid detection of infectious viral particles. For this, buffalo green monkey kidney cells (BGMK) infected with different virus titers were studied from 1 to 12 hours after infection (h.p.i.). It was concluded that this approach for the detection and quantification of poliovirus has the potential to be extended to other viruses, and can be adaptable to an automated scheme for use in water safety monitoring, medical diagnosis, among others. [20]

In another study, Santos *et. al.* [45] used ATR-FTIR spectroscopy in conjunction with multivariate analysis techniques to detect different viral loads of dengue serotype 3 (DENV-3) in blood and serum. In this study, 40 blood and serum samples were infected with four different viral loads of DENV-3 (10 samples for each viral load). Then, spectra in the 1800 to 900 cm^{-1} range were imported into MATLAB where they were pre-processed and separated into training, validation and test sets. The multivariate classification models used in this study were PCA-LDA, SPA-LDA and GA-LDA. Among the figures of merit used to assess the capacity of the models, sensitivity and specificity were used, where it was seen that the best results were obtained for blood samples, where all models obtained 100% sensitivity and specificity. Among the variables selected by SPA and GA, wavenumbers that are related to proteins and variables related to RNA were observed refereeing to important structures in the composition of the virus. This suggests that the technique was able to identify the virus in the blood. The authors conclude by stating that although the results were encouraging, it would be necessary to carry out more in-depth studies with a larger number of samples for greater reliability. However, the study shows that this field of research is worth exploring and, in the near future, we will be able to rely on portable spectroscopic instruments in clinical diagnostic environments. [45] Santos *et. al.* [47] also used ATR-FTIR spectroscopy in conjunction with PCA-LDA, SPA-LDA and GA-LDA to discriminate clinical samples from patients infected with dengue, zika and chikungunya and samples from healthy volunteers (without the presence of these diseases). The main objective of this study was to evaluate the specificity of the technique in discriminating blood samples based only on the presence or absence of these viruses and to compare the results with some standard diagnostic methods. The results of this study showed sensitivity and specificity values of 100% for the healthy, dengue and chikungunya classes and values close to 90% for zika, suggesting that this methodology has the potential to detect biochemical variations caused by the presence of the virus in the blood. [47]

Naseer *et. al.* [31] also studied ATR-FTIR spectroscopy in the diagnosis of dengue. In this study, they sought to identify biochemical marks differentially expressed in lyophilized human blood sera infected with dengue and healthy. They used a total of 77 samples, where 57 were infected with DENV and 20 were healthy. As multivariate analysis algorithms, they used principal component analysis (PCA), linear discriminant analysis (LDA) and partial least squares regression (PLSR) in the differentiation between classes based on the region between 1601 and 1501 cm^{-1}. The PCA-LDA model showed sensitivity and specificity of 89% and 95%, respectively. While the PLSR model has successfully found biochemical changes, with an $R^2 = 0.9980$. The authors conclude by emphasizing that ATR-FTIR spectroscopy in conjunction with multivariate classification algorithms can be used as an effective tool for diagnosing dengue. [31]

ATR-FTIR spectroscopy was also employed in the diagnosis of hepatitis C and B viruses. [41] In this study, the capacity of the technique was evaluated together with multivariate analysis techniques in the classification of human serum samples based on the presence of HBC and HCV infections. The classification technique used by them was the partial least squares discriminant analysis (PLS-DA), which is a variation of the PLS algorithm used in the discriminant analysis. The samples were separated into training (70%) and test sets (30%). The positive spectra for HBC and HCV showed an intense band observed at 1631 cm^{-1} attributed to the protein marker immunoglobulin (Ig). A 1093 cm^{-1} band was observed only in the spectra of samples infected with HBV, and was attributable to the C-C and CO- modes of the hepatitis B surface antigen (HBsAg) N-glycan polysaccharide. However, the authors should also consider vibrations that overlap with C-C and CO- stretching, such as a PO_2^- symmetrical stretching vibration that is related to the dsDNA of HBV. Finally, the authors conclude by stating that ATR-FTIR spectroscopy has great potential in the study of blood composition and in the identification

of possible disease markers. However, they emphasize that care must be taken so that modeling does not influence inflammation markers, which could confuse the diagnosis. [41]

9. Advantages and disadvantages of the technique

Among the advantages of ATR-FTIR spectroscopy we can highlight the high signal-to-noise ratio, reduced dispersion, good spatial resolution, non-destructiveness, no sample preparation (or minimal preparation), low relative cost, and automated analysis. Among the disadvantages, we can highlight that the analysis can be destructive if too much pressure is applied on the sample, CO_2 interference, and sample thickness problems that can lead to systematic spectral shifts. [4,46] A disadvantage in the use of ATR-FTIR in the detection of viruses is a possible overlapping of spectral bands between the virus (capsid protein, proteins released during infection, genetic material) and the host cell. In the region between 1800 to 900 cm^{-1} we will find peaks related to proteins, and genetic material, however, since these biochemical structures can be found in both cells and viruses, it is expected an overlapping of spectral bands. This can make discrimination difficult.

10. Future perspectives

As has been seen, viral diagnostic techniques that are used nowadays are a double-edged sword, where advantages and disadvantages are mutually observed. Direct methods are more specific, however, they take time and are more expensive. Indirect methods are faster and cheaper, but are less specific. Based on this assessment, ATR-FTIR spectroscopy emerges as a tool with the potential to solve the deficiencies found by standard techniques. ATR-FTIR spectroscopy is known to have a fast response and provide reliable information about the sample composition, and has been used in several virological applications for screening or diagnosis of viral infections.

Knowing that vibrational spectroscopy is fast, non-destructive and has a low-cost, we can imagine that, with new studies developed in this area, soon we can count with spectroscopic tools in clinics and hospitals, being used for routine diagnostic or acting as a reliable diagnostic aiding tool. For this, only a minimal amount of collected biofluid would be needed. Spectral acquisition of this biofluid would be done by an instrument coupled to a computer where this spectral information would be automatically imported into software that would perform all computational procedures in real time (pre-processing and multivariate classification), based on a previously constructed and optimized model.

11. Conclusions

It cannot be denied that ATR-FTIR spectroscopy has several advantages that attribute to this technique a great potential towards viral diagnostic routines. This technique cannot be overlooked, and must be considered an possible alternative for viral diagnosis especially when experiencing global problems such as the one caused by COVID-19. Several people die every day for lack of a quick, reliable and relatively inexpensive diagnosis. In conjunction with spectroscopic analysis, multivariate data analysis provides powerful support for interpretation and pattern recognition. Finally, spectroscopy approaches combined with multivariate analysis provide a powerful weapon in studies for development of rapid diagnostics, which can be extended to several viruses of different types and strains. With that, we encourage the scientific community to explore this field of study.

Acknowledgements

Conflict of interest

The author has no conflict of interest to report.

References

[1] M.E. Álvarez-Argüelles et al., Diagnosis and molecular characterization of Chikungunya virus infections, in: *Current Topics in Neglected Tropical Diseases*, IntechOpen, 2019, pp. 375–385.

[2] T.C. Baia et al., FTIR microspectroscopy coupled with variable selection methods for the identification of flunitrazepan in necrophageous flies, *Analytical Methods* **8** (2016), 968–972. doi:10.1039/C5AY02342D.

[3] S. Baize et al., Emergence of Zaire Ebola virus disease in Guinea, *The New England Journal of Medicine* **371** (2014), 1418–1425. doi:10.1056/NEJMoa1404505.

[4] M.J. Baker et al., Using Fourier transform IR spectroscopy to analyze biological materials, *Nature Protocols* **9** (2014), 1771–1791. doi:10.1038/nprot.2014.110.

[5] N. Boonham et al., Methods in virus diagnostics: From ELISA to next generation sequencing, *Virus Research* **186** (2014), 20–31. doi:10.1016/j.virusres.2013.12.007.

[6] R. Bro and A.K. Smilde, Principal component analysis, *Analytical Methods* **6** (2014), 2812–2931. doi:10.1039/C3AY41907J.

[7] M.J. Broadhurst, T.J.G. Brooks and N.R. Pollock, Diagnosis of Ebola virus disease: Past, present, and future, *Clinical Microbiology Reviews* **29** (2016), 773–793. doi:10.1128/CMR.00003-16.

[8] S.G. Deeks and H.I.V. Infection, Inflammation, immunosenescence, and aging, *Annu. Rev. Med.* **62** (2011), 141–155. doi:10.1146/annurev-med-042909-093756.

[9] W. Dejnirattisai et al., Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with Zika vírus, *Nature Immunology* **17** (2016), 1102–1108. doi:10.1038/ni.3515.

[10] S.J. Dixon and R.G. Brereton, Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure, *Chemometrics and Intelligent Laboratory Systems* **95** (2009), 1–17. doi:10.1016/j.chemolab.2008.07.010.

[11] C. Drosten et al., Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, Dengue virus, and Yellow Fever virus by real-time reverse transcription-PCR, *Journal of Clinical Microbiology* **40** (2002), 2323–2330. doi:10.1128/JCM.40.7.2323-2330.2002.

[12] M.P. Girard et al., The 2009 A (H1N1) influenza virus pandemic: A review, *Vaccine* **28** (2010), 4895–4902. doi:10.1016/j.vaccine.2010.05.031.

[13] B.A. Goupil and C.N. Mores, A review of Chikungunya virus-induced arthralgia: Clinical manifestations, therapeutics, and pathogenesis, *The Open Rheumatology Journal* **10** (2016), 129–140. doi:10.2174/1874312901610010129.

[14] F.X. Heinz and K. Stiasny, The antigenic structure of Zika virus and its relation to other flaviviruses: Implications for infection and immunoprophylaxis, *Microbiology and Molecular Biology Reviews* **81** (2017), 1–27.

[15] M.L. Holshue et al., First case of 2019 novel coronavirus in the United States, *The New England Journal of Medicine* **382** (2020), 929–936. doi:10.1056/NEJMoa2001191.

[16] A.J. Jaaskelainena et al., Validation of serological and molecular methods for diagnosis of Zika vírus infections, *Journal of Virological Methods* **263** (2019), 68–74. doi:10.1016/j.jviromet.2018.10.011.

[17] J.G. Kelly et al., Biospectroscopy to metabolically profile biomolecular structure: A multistage approach linking computational analysis with biomarkers, *Journal of Proteome Research* **10** (2011), 1437–1448. doi:10.1021/pr101067u.

[18] R.W. Kennard and L.A. Stone, Computer aided design of experiments, *Technometrics* **11** (1969), 137–148. doi:10.1080/00401706.1969.10490666.

[19] R.J. Kuhn et al., Structure of Dengue virus: Implications for flavivirus organization, maturation, and fusion, *Cell* **108** (2002), 717–725. doi:10.1016/S0092-8674(02)00660-8.

[20] F.T. Lee-Montiel, K.A. Reynolds and M.R. Riley, Detection and quantification of poliovirus infection using FTIR spectroscopy and cell culture, *Journal of Biological Engineering* (2011), 5–16.

[21] Q. Li et al., Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia, *The New England Journal of Medicine* **382** (2020), 1199–1207. doi:10.1056/NEJMoa2001316.

[22] A.S. Marques et al., Feature selection strategies for identification of Staphylococcus aureus recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: A case study, *Journal of Microbiological Methods* **98** (2014), 26–30. doi:10.1016/j.mimet.2013.12.015.

[23] A.S. Marques et al., Rapid discrimination of klebsiella pneumoniae carbapenemase 2 – producing and non-producing klebsiella pneumoniae strains using near-infrared spectroscopy (NIRS) and multivariate análisis, *Talanta* **134** (2015), 126–131. doi:10.1016/j.talanta.2014.11.006.

[24] F.L. Martin et al., Distinguishing cell types or populations based on the computational analysis of their infrared spectra, *Nature Protocols* **5** (2010), 1748–1760. doi:10.1038/nprot.2010.133.

[25] W. McIntyre et al., Positive-sense RNA viruses reveal the complexity and dynamics of the cellular and viral epitranscriptomes during infection, *Nucleic Acids Research* **46** (2018), 5776–5791. doi:10.1093/nar/gky029.

[26] T.P. Monath, Yellow fever: An update, *The Lancet Infectious Diseases* **1** (2011), 11–20. doi:10.1016/S1473-3099(01)00016-0.

[27] T.P. Monath and P.F.C. Vasconcelos, Yellow fever, *Journal of Clinical Virology* **64** (2015), 160–173. doi:10.1016/j.jcv.2014.08.030.

[28] C.L.M. Morais et al., Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard–Stone algorithm approach, *Bioinformatics* **35** (2019), 5257–5263. doi:10.1093/bioinformatics/btz421.

[29] C.L.M. Morais et al., Standardization of complex biologically derived spectrochemical datasets, *Nature Protocols* **14** (2019), 1546–1577. doi:10.1038/s41596-019-0150-x.

[30] D.A. Muller, A.C.I. Depelsenaire and P.R. Young, Clinical and laboratory diagnosis of Dengue virus infection, *The Journal of Infectious Diseases* **215** (2017), 89–95. doi:10.1093/infdis/jiw649.

[31] K. Naseer et al., FTIR spectroscopy of freeze-dried human sera as a novel approach for dengue diagnosis, *Infrared Physics and Technology* **102** (2019), 102998–103002. doi:10.1016/j.infrared.2019.102998.

[32] M.S. Natrajan, A. Rojas and J.J. Waggoner, Beyond fever and pain: Diagnostic methods for Chikungunya virus, *Journal of Clinical Microbiology*, **57**, (2019), e00350-19.

[33] M.H. Nguyen et al., Hepatitis B virus: Advances in prevention, diagnosis, and therapy, *Clinical Microbiology Reviews* **33** (2020), 1–38. doi:10.1128/CMR.00046-19.

[34] B. Otange et al., Estimation of HIV-1 viral load in plasma of HIV-1-infected people based on the associated Raman spectroscopic peaks, *Journal of Raman Spectroscopy* **50** (2019), 620–628. doi:10.1002/jrs.5557.

[35] M. Paraskevaidi, P.L. Martin-Hirsch and F.L. Martin, ATR-FTIR spectroscopy tools for medical diagnosis and disease investigation, in: *Nanotechnology Characterization Tools for Biosensing and Medical Diagnosis*, 2018, pp. 163–211. doi:10.1007/978-3-662-56333-5_4.

[36] S. Perlman, A. Decade and A. Coronavirus, *The New England Journal of Medicine* **382** (2020), 760–762. doi:10.1056/NEJMe2001126.

[37] R. Peters and M. Stevenson, Zika virus diagnosis: Challenges and solutions, *Clinical Microbiology and Infection* **25** (2019), 142–146. doi:10.1016/j.cmi.2018.12.002.

[38] A.R. Plourde and E.M. Bloch, A literature review of Zika virus, *Emerging Infectious Diseases* **22** (2016), 1185–1192. doi:10.3201/eid2207.151990.

[39] M. Ringnér, What is principal component analysis?, *Nature Biotechnology* **26** (2008), 303–304. doi:10.1038/nbt0308-303.

[40] M.G. Rossmann, Structure of viruses: A short history, *Quarterly Reviews of Biophysics* **46** (2013), 133–180. doi:10.1017/S0033583513000012.

[41] S. Roy et al., Spectroscopy goes viral: Diagnosis of hepatitis B and C virus infection from human sera using ATR-FTIR spectroscopy, *Clinical Spectroscopy* **1** (2019), 100001. doi:10.1016/j.clispe.2020.100001.

[42] J. Saade et al., Identification of hepatitis C in human blood serum by near-infrared Raman spectroscopy, *Spectroscopy* **22** (2008), 387–395. doi:10.1155/2008/419783.

[43] A. Sakudo et al., Diagnosis of HIV-1 infection by near-infrared spectroscopy: Analysis using molecular clones of various HIV-1 subtypes, *Clinica Chimica Acta* **413** (2012), 467–472. doi:10.1016/j.cca.2011.10.035.

[44] T.S. Salles et al., History, epidemiology and diagnostics of dengue in the American and Brazilian contexts: A review, *Parasites & Vectors* **11** (2018), 264–275. doi:10.1186/s13071-018-2830-8.

[45] M.C.D. Santos et al., ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in different concentrations in blood and serum: A new approach, *RSC Advances* **7** (2017), 25640–25649. doi:10.1039/C7RA03361C.

[46] M.C.D. Santos et al., Spectroscopy with computational analysis in virological studies: A decade (2006–2016), *Trends in Analytical Chemistry* **97** (2017), 244–256.

[47] M.C.D. Santos et al., ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dengue vs. chikungunya vs. Zika clinical samples, *Analytical Methods* **10** (2018), 1280–1285. doi:10.1039/C7AY02784B.

[48] A. Savitzky and M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry* **36** (1964), 1627–1639. doi:10.1021/ac60214a047.

[49] C. Shan et al., Zika virus: Diagnosis, therapeutics, and vaccine, *ACS Infectious Diseases* **2**, (2016), 170–172.

[50] X. Shen et al., Study on baseline correction methods for the Fourier transform infrared spectra with different signal-to-noise ratios, *Applied Optics* **57** (2018), 5794–5799. doi:10.1364/AO.57.005794.

[51] L.F.S. Siqueira et al., LDA vs. QDA for FT-MIR prostate cancer tissue classification, *Chemometrics and Intelligent Laboratory Systems* **162** (2017), 123–129. doi:10.1016/j.chemolab.2017.01.021.

[52] L. Sitole et al., Mid-ATR-FTIR spectroscopic profiling of HIV/AIDS sera for novel systems diagnostics in global health, *OMICS A Journal of Integrative Biology* **18** (2014), 513–523. doi:10.1089/omi.2013.0157.

[53] S.F.C. Soares et al., The successive projections algorithm, *Trends in Analytical Chemistry* **42** (2013), 84–98. doi:10.1016/j.trac.2012.09.006.

[54] R. Tellier, Review of aerosol transmission of Influenza A virus, *Emerging Infectious Diseases* **12** (2006), 1657–1662. doi:10.3201/eid1211.060426.

[55] A. Tharwat et al., Linear discriminant analysis: A detailed tutorial, *AI Communications* **30** (2017), 169–190. doi:10.3233/AIC-170729.

[56] G. Theophilou et al., ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: Classifying subtypes of human câncer, *Analyst* **141** (2016), 585–594. doi:10.1039/C5AN00939A.

[57] Tool for the diagnosis and care of patients with suspected arboviral diseases, Washington, D.C.: PAHO, 2017.

[58] J. Trevisan et al., Extracting biological information with computational analysis of Fourier transform infrared (FTIR) biospectroscopy datasets: Current practices to future perspectives, *Analyst* **137** (2012), 3202–3215. doi:10.1039/c2an16300d.

[59] S.V. Vemula et al., Current approaches for diagnosis of influenza virus infections in humans, *Viruses* **8** (2016), 96–110. doi:10.3390/v8040096.

[60] C. Wang et al., A novel coronavirus outbreak of global health concern, *The Lancet Comment* **395** (2020), 470–473. doi:10.1016/S0140-6736(20)30185-9.

[61] W. Wu et al., Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data, *Analytica Chimica Acta* **329** (1996), 257–265. doi:10.1016/0003-2670(96)00142-0.

[62] P. Zhou et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature* **579** (2020), 270–273. doi:10.1038/s41586-020-2012-7.

[63] N. Zhu et al., A novel coronavirus from patients with pneumonia in China, 2019, *The New England Journal of Medicine* **382** (2020), 727–733.

[64] Zucchiatti et al., Contribution of ribonucleic acid (RNA) to the Fourier transform infrared (FTIR) spectrum of eukaryotic cells, *Analytical Chemistry* **88** (2016), 12090–12098. doi:10.1021/acs.analchem.6b02744.