Chapter 10

Multivariate Analysis for Fourier Transform Infrared Spectra of Complex Biological Systems and Processes

Diletta Ami, Paolo Mereghetti and Silvia Maria Doglia

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53850

1. Introduction

Fourier transform infrared (FTIR) spectroscopy is a label-free and non invasive technique that exerts an enormous attraction in biology and medicine, since it allows to obtain in a rapid way a biochemical fingerprint of the sample under investigation, giving information on its main biomolecule content. This spectroscopic tool is successfully applied not only to the study of the structural properties of isolated biomolecules, such as proteins, nucleic acids, lipids, and carbohydrates, but also to the characterization of complex biological systems, for instance intact cells, tissues, and whole model organisms.

In particular, FTIR microspectroscopy, obtained by the coupling of an infrared microscope to a FTIR spectrometer, makes it possible to collect the IR spectrum from a selected sample area down to ~ 20 microns x 20 microns when conventional IR source and detector are employed, and down to of a few micrometers when more specialized and sensitive detectors and the highly brilliant synchrotron light source are used. In this way, FTIR microspectroscopy provides detailed information on several biological processes in situ, among which stem cell differentiation [1-5], somatic cell reprogramming [6], cell maturation [7, 8], amyloid aggregation [9-12] and cancer onset and progression [13-15], making it possible to disclose the infrared response not only from single cells, but also from subcellular compartments [8, 16, 17].

The FTIR spectra of biological systems are very complex since they consist of the overlapping absorption of the main biomolecules; for this reason, to pull out the significant and non-redundant information contained in the spectra it is necessary to apply an appropriate multivariate analysis, able to process very high-dimensional data. This is even more crucial when time-dependent biological processes, such as cell maturation or

differentiation, are studied. Indeed, in this case it is fundamental to be able to extract from the spectral data the relevant information of the process you are investigating [18-21].

In Figure 1 we schematized the procedure that should be followed to successfully tackle the FTIR characterization of complex biological systems.

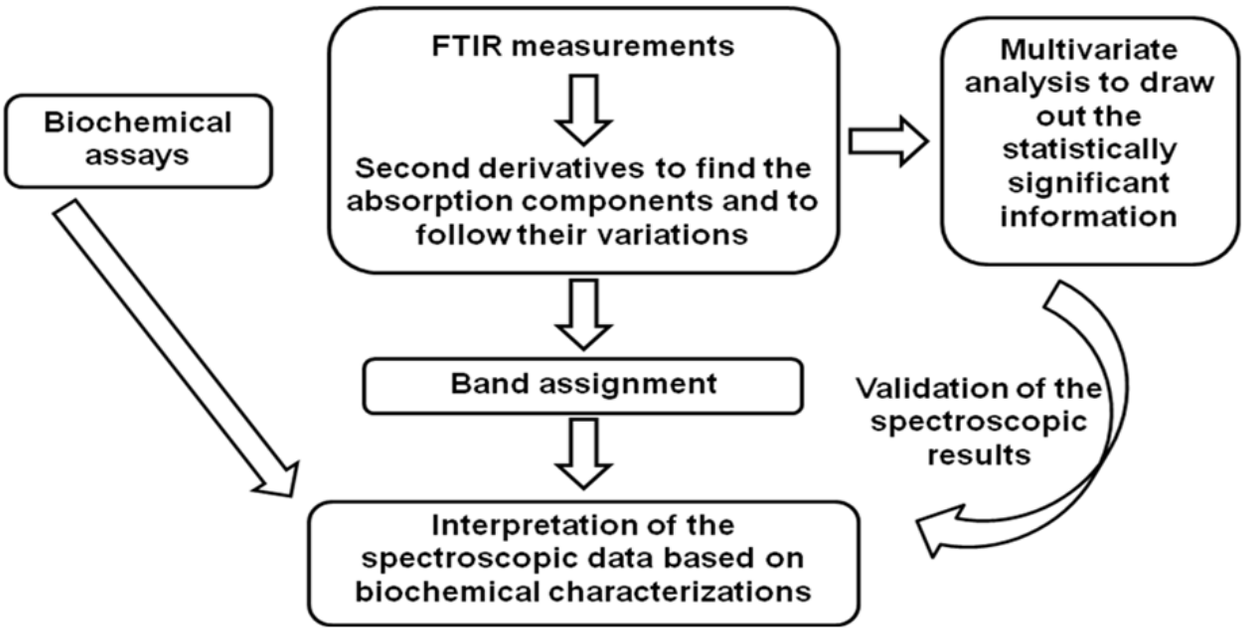

Figure 1. Scheme of the FTIR approach to study complex biological systems. The IR absorption spectra are analysed by resolution enhancement approaches (e.g. second derivatives) to resolve the overlapped absorption components and to monitor their variations during the process under investigation. The spectroscopic results are validated by an appropriate multivariate analysis approach, to identify firstly specific marker bands of the studied process. The interpretation of the spectroscopic data should be then confirmed by standard biochemical assays.

Several multivariate analysis approaches exist and for the scope of this book they can be divided into two main categories: regression and classification techniques. In the first category fall all methods that allow to derive a model describing the relationship between two sets of variables. The second category includes techniques to split observations into groups or classes.

In this chapter, we will firstly introduce the most widely used multivariate analysis approaches in the field of spectroscopy.

We will then illustrate the basic principles and experimental details for the application of principal component - linear discriminant analysis (PCA-LDA) to the analysis of FTIR spectral data of complex biological systems. The potential of these combined tools will be described on illustrative examples of cell biological process studies. In particular, we will discuss in details its application on our FTIR study of murine oocytes characterized by two different types of chromatin organisation around the nucleolus, strongly affecting their development after fertilization. In this case, PCA-LDA analysis made it possible to identify not only the maturation stage in which the fate separation between the two kinds of oocytes

occurred, but also to disclose the most significant cellular processes responsible for the different oocyte destiny, thus validating the visual inspection of the infrared spectra [7].

2. FTIR microspectroscopy of complex biological systems

Fourier transform infrared (FTIR) microspectroscopy is a powerful technique that allows to obtain a molecular fingerprint of the sample under investigation in a rapid and non-invasive way. In the case of complex biological systems it provides simultaneously, in a single measurement, information on the main biomolecules, such as lipids, proteins, nucleic acids, and carbohydrates, requiring also a very limited amount of sample. For these reasons, it became recently a very attracting tool for biomedical research [20, 22-24], being successfully employed for the study of several biological systems, from intact cells [6, 7, 25] to tissues [11, 26, 27] and whole model organisms (i.e. the nematode *Caenorhabditis elegans*) [9, 28].

As an example, in Figure 2 it is reported the FTIR absorption spectrum of a single intact murine oocyte. As shown, its IR response is very complex, being due to the absorption of the main biomolecules. In particular, between 3050 - 2800 cm^{-1} and 1500 - 1350 cm^{-1} the absorption of the lipid acyl chains occurs, while around 1740 cm^{-1} the ester carbonyl absorbs [29]. Moreover, the amide I and amide II bands - mainly due to the C=O stretching and the NH bending of the peptide bond respectively - give information on the protein secondary structure [30], while the spectral range between 1000 and 800 cm^{-1} is very informative on nucleic acid absorption, since it is due in particular to sugar vibrations sensitive to their conformation and to backbone vibrational modes [31, 32]. Finally, we should also mention the very complex spectral range between 1250 - 1000 cm^{-1}, mainly due to phosphodiester groups of nucleic acids and phospholipids and to the C-O absorption of glycogen and other carbohydrates [31, 33, 34].

Making it possible to obtain a sample biochemical fingerprint in a rapid and non destructive way, FTIR microspectroscopy is widely applied to the in situ characterization of cellular processes, such as cell maturation, differentiation, and reprogramming [3, 5-7, 25, 35], and to the detection of several diseases, as, for instance, cancer [13-15] and neurodegenerative disorders [10, 11], whose onset is accompanied by changes in the composition and structure of several biomolecules.

Since water has a strong absorption in the mid-infrared spectral range, samples have to be dried rapidly before IR measurements, in particular when working in transmission mode (see for details the following paragraph). The suitability of such “dry-fixing” has been proved by Raman spectroscopy, a vibrational tool complementary to FTIR, whose response is not affected by water. In particular, Raman measurements performed on differentiating human embryonic stem cells, hydrated and dry-fixed, demonstrated that the rapid desiccation didn’t affect the spectroscopic response of the main biomolecules. Indeed, in both cases the same temporal pattern of the differentiation marker bands - due to tryptophan, nucleic acid backbone and base vibrations - was observed during the biological process under investigation [36].

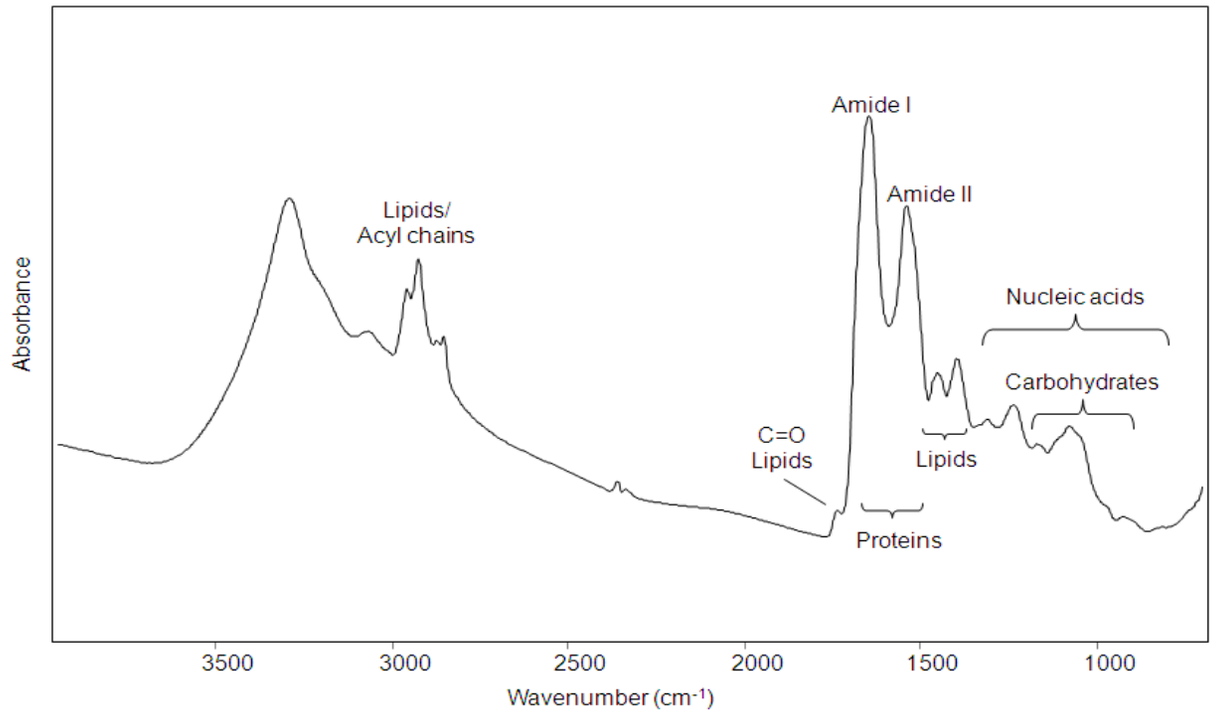

Figure 2. FTIR absorption spectrum of a single intact murine oocyte. The measured absorption spectrum of a single intact murine oocyte (surrounded nucleolus, MI 10 H) is reported without any corrections. The oocyte - deposited on a BaF_2 window - was measured in transmission by the IR microscope UMA 500, coupled to the FTIR spectrometer FTS 40A (both from Digilab), at a resolution of 2 cm^{-1}. The absorption regions of the main biomolecules are indicated.

We should add that to obtain reliable results on the studied process it is crucial to standardize firstly the sample preparation, since - for instance - metabolic changes due to cell aging could result in significant spectral changes that could, in turn, hide the IR response specifically due to the process of interest, as it has been recently reported in the literature [37]. For these reasons, it is fundamental to check accurately the stage of cell growth in culture before performing spectroscopic measurements.

We should also briefly mention that, before spectral analyses, the measured IR spectra could require some corrections due to artifacts that can interfere with the spectroscopic response. For instance, single cells, or subcellular compartments, or particles of the size of the same order of that of the incident infrared light (≅3-10 microns) could give rise to Mie scattering, that significantly distorts the measured spectrum, causing misinterpretation of the results. For this reason, before further analyses, it is strongly recommended to correct the measured spectra with opportune algorithms specifically developed to this aim [38].

Since the IR spectra of complex biological systems are due to the overlapping spectral features of multiple components, their analysis requires often the employment of resolution enhancement procedures to better resolve their absorption bands, an essential prerequisite for the identification of peak positions and their assignment to the vibrational modes of the different molecules. Among these, second derivative analysis is widely applied, as described in [39]. Since second derivative band intensity is inversely proportional to the square of the

original band half-width, this procedure introduces an enhancement of sharp lines, as those due to vapour and noise. For this reason, this analysis requires spectral data free of vapour absorption and with excellent signal to noise ratio.

Furthermore, due to the intrinsic complexity of biological systems, their spectral analysis requires the support of appropriate multivariate analysis approaches able to tackle the study of high-dimensional data, to verify firstly the reproducibility of the results and then to extract the most significant spectral information [18-21] (see for details paragraph 4).

3. FTIR microspectroscopy: Technical considerations

FTIR microspectroscopy is realized coupling to a FTIR spectrometer an infrared microscope characterized by an all reflecting optics, since typical lenses and condensers of visible microscopy - being made of glass, not transparent to the IR radiation - cannot be employed.

The main advantage of FTIR microspectroscopy is that it offers the possibility to study selected areas of the sample under investigation, resulting particularly useful in the case of systems characterized by an intrinsic heterogeneity, such as biological systems.

Two main types of IR microscopy exist, depending on the detector employed, and both equipped with an IR thermal source (globar), whose spatial resolution is diffraction-limited.

The first, conventional, generally equipped with a nitrogen cooled mercury cadmium telluride (MCT) detector, makes it possible to measure IR absorption spectra from a microvolume within the sample, selected by a variable aperture of the microscope, whose side can be adjusted down to a few tens of microns.

The second type of IR microscope, more advanced, is equipped with a focal plane array (FPA), consisting of an array of infrared detector elements, that enables not only to collect the IR absorption spectrum of the sample, but also an IR chemical imaging, where the image contrast is given by the response of selected sample regions to particular IR wavenumbers. Depending mainly on the detection array, the spatial resolution in this kind of microscopy is approximately between 20 and 5 microns, making it possible to reach, therefore, a resolution near to the diffraction limit.

We should, however, add that the use of a synchrotron IR light source, with a brightness of at least two orders of magnitude higher than that of a conventional thermal source, makes it possible to achieve diffraction-limited spatial resolution with enhanced signal-to-noise ratio. In this way, synchrotron light could allow to explore the IR spectra at the subcellular level.

A final remark should be done concerning the spectral acquisition mode. Indeed, infrared measurements can be mainly performed in transmission, reflectance or attenuated total reflection (ATR) mode. Typically, measurements on complex biological systems are performed in transmission mode, using appropriate IR transparent supports for the deposition of the sample, such as BaF_2, CaF_2, ZnSe. In this case, the IR beam goes through the sample, that - depending mainly on its molar extinction coefficient - should have a uniform thickness, not exceeding 15-20 microns.

Moreover, in reflectance mode - where the sample is placed onto proper reflective slides - the IR beam passes the sample, is reflected by the slide, and passes the sample again. In particular, the sample slides reflect mid-infrared radiation almost completely and usually are also transparent to visible light, allowing sample inspection by a conventional light microscope. This approach is, for instance, useful for tissue characterizations.

Finally, in the ATR approach, where the sample is placed into contact with a higher refractive index and an IR transparent element (mainly germanium and diamond), samples with higher thickness than in transmission can be processed. In particular, the IR beam reaches the interface between the ATR support and the sample at an angle larger than that corresponding to the total reflection. In this way the beam is totally reflected by the interface and penetrates into the sample as an evanescent wave, where it can be absorbed. The beam penetration depth is of the order of the IR wavelength (a few micrometers) and depends on the wavelength, the incident angle, as well as on the refractive indices of the sample and of the ATR element. Furthermore, it should be noted that this kind of approach makes it possible to measure also samples not necessarily deposited onto an IR transparent support, as in ATR measurements it is only required that the sample be in close contact with the ATR element.

For a review of the technical aspects of FTIR microspectroscopy, see [40-42].

4. Multivariate analyses

4.1. Introduction to multivariate analysis

Several phenomena can only be described or explained by taking into account several variables at the same time. These cases represent the realm of the Multivariate statistical analysis (MVA).

We now define the structure of our data that will be kept throughout the text for all described techniques. For a given phenomenon we perform a certain measurement and store the value in a uni- or multivariate variable called $\mathbf{y} = \left(y_1, y_2, ..., y_m\right)^T$, where m is the number of independent variables. The same measurement can be repeated several times on the same sample or on different samples. We then define a group as a collection of two or more replica of the same experiment and we also define the term instance or observation to refer to a specific experiment within one group.

Each instance associated to the variable $\mathbf{y}$ is stored in a matrix $\mathbf{Y}$ composed of n rows (the observations) and m columns (the independent variables).

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n1} & \cdots & y_{nm} \end{pmatrix} = \left(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\right)^T \quad (1)$$

Each element of matrix $\mathbf{Y}$ can be indicated as y_{ij} where i indicate the observation and j is an independent variable. In some cases we want to find or explain the relationship between the

independent variables **Y** and another set of uni- or multivariate variables **Z**. Similarly to the **Y** matrix, the matrix **Z** has *n* rows, one for each observation and *m* columns, the dependent variables.

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1\mathrm{m}} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n1} & \dots & z_{nm} \end{pmatrix} = \left(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\right)^T \quad (2)$$

The matrix **Y** (composed of the independent variables y) represents the only input for several multivariate techniques described here; in some other cases the matrices **Y** and **Z** (composed of the dependent variables z) are both required.

In the following part, we will make a distinction between regression and classification techniques. However, it should be clear that the separation between these two domains is not always sharp and the same technique can be either used for regression or for classification purposes.

4.2. Multivariate regression techniques

4.2.1. Linear Multivariate Regression (LMVR)

LMVR (or MLR) can be used to model linear relationships between one or more **z** (dependent variable) and one or more **y** (independent variable). In the most general case, we have *n* independent multivariate variables **y** represented by the matrix **Y** and the corresponding response multivariate variable *z*, stored in the matrix **Z**.

The LMVR is based, as many other statistical techniques, on the generalized linear model: $\mathbf{Z} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta}$ is a matrix containing the parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a matrix which models the errors or noise. The coefficients $\boldsymbol{\beta}$ are usually estimated using the ordinary least square, which consists of minimizing the sum of the square differences of the *n* observed ***y***'s from their modeled values. Mathematically, the optimal values of $\boldsymbol{\beta}$ are obtained by $\boldsymbol{\beta} = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{Z}$. To apply the least square method we must have *n - 1 > m* (e.g. the number of observation must be larger than the number of variables, which is often not the case), otherwise the matrix $\mathbf{Y}^T\mathbf{Y}$ is singular and cannot be inverted. Another common problem is the correlation between variables; more specifically, none of the independent variables must be a linear combination of any other. This phenomenon is called "multicollinearity" [43, 44] and it will be explained in more details in section 4.2.3.

4.2.2. Non-Linear Multivariate Regression (NLMVR)

In some cases linear models cannot be used and one could try to apply non-linear models.

Common models which frequently apply to natural phenomena are the exponentials (which, indeed, is a transformed linear model. A linear model can be applied upon on the

logarithm of the data), logistic models or power law models.The regressed model has the general form of $\mathbf{Z} = \boldsymbol{\beta} f\left(\mathbf{Y}\right) + \boldsymbol{\varepsilon}$, where *f*(**Y**) can be any non-linear function.

The optimal values for the coefficients $\boldsymbol{\beta}$ can be obtained using deterministic optimization algorithm such as the conjugate gradients [45] or the Levenberg-Marquard method [46, 47], or stochastic algorithm such as genetic algorithms [48].

4.2.3. The multicollinearity problem

When the number of observations is smaller than the number of variables (as it often happens for spectral data), the matrix $\mathbf{Y}^T\mathbf{Y}$ is singular and is not invertible. This rules out the possibility of using standard linear multivariate techniques (LMVR) based on the least square criterion, as the solution will not be unique.

Increasing the number of observations (above the number of variables) will not always solve the problem. This is due to the so-called near-multicollinearity which means that some variables can be written approximately as linear functions of other variables. This problem is often found among spectral measurements. Even if the solution will be mathematically unique, it may be unstable and lead to poor prediction performances.

Linearly correlated or quasi-linearly correlated variables have to be removed prior to apply a regression method. In the following sections, we will describe two methods that are frequently used to remove correlations among variables, namely principal component analysis (PCA) and partial least squares (PLS).

4.2.3.1. Principal Component Analysis (PCA)

We should first recall the structure of the data. Suppose that we have *n* observations, each one defined by a vector $\mathbf{y}_i$ composed of *m* variables, where *i=1,2,...,n* stands for the i-th observation. The matrix of the original data **Y** is then composed by *n* rows (the observations) and *m* columns (the variables).

By using PCA, our intent is to develop a smaller number of uncorrelated artificial variables, called principal components (PC), that will account for most of the variance in the observed variables. The new uncorrelated variables are obtained as linear combination of the original data as $\mathbf{T} = \mathbf{AY}$. Correlation among variables can be measured using the covariance matrix.

Given the sample mean of the m-dimensional vector $\mathbf{y}_i$, $\langle \mathbf{y} \rangle = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$, an unbiased estimator of the sample covariance matrix is $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{y}_i - \langle \mathbf{y} \rangle\right)\left(\mathbf{y}_i - \langle \mathbf{y} \rangle\right)^T$.

For uncorrelated variables, the off-diagonal values of the sample covariance matrix are zero, that is, **S** is diagonal. The covariance of linearly transformed variables $\mathbf{T} = \mathbf{AY}$ is equal to $\mathbf{S}_T = \mathbf{ASA}^T$, where **S** is the sample covariance of the original data **Y** [49].

Thus, we want to find the matrix **A** such that the covariance matrix of the transformed data, $\mathbf{S}_T$, is diagonal, which corresponds to find the eigenvectors of the covariance matrix and the corresponding eigenvalues.

The eigenvalues, which coincide with the matrix $\mathbf{S}_T$, are the sample variance of the principal components **T** and are ranked according to their magnitude. The first principal component is then the linear combination with maximal variance (the largest eigenvalue). The second principal component is the linear combination with the maximal variance along a direction orthogonal to the first component, and so on [44].

The number of eigenvalues is equal to the number of original variables; however, since the eigenvalues are equal to the variance of the principal components and they are sorted in a decreasing order, the first *k* eigenvalues can account for a large portion of the variance of the data.

Hence, to describe our original dataset we can use only the first *k* uncorrelated principal components, instead of the complete set of redundant *m* variables. In matrix notation this can be written as $\mathbf{T}_k = \mathbf{A}_k\mathbf{Y}$, where $\mathbf{A}_k$ is the eigenvector matrix truncated to the k-th eigenvector, and $\mathbf{T}_k$ is the matrix of the first *k* principal components, also called score matrix [50].

Choosing which and the number of principal components that should be retained in order to summarize our data is a task that can be solved using several strategies [43, 49]. For example, one way commonly used is to retain the first k principal components that explain a given total percentage of the variance, e.g. 90% [43, 44]. Another rule is to plot the eigenvalues in decreasing order. Moving from left to right, the eigenvalues usually have an initial steep drop followed by a slow decrease. All the components after the elbow between the steep and the flat part of the curve should be discarded. This test is called screen plot.

Alternatively, one can select the principal components that can be associated to a physical meaning related to the studied system. For example, following the differentiations of a cell line growing in different experimental conditions, one principal component may represents the different conditions, while another PC may describe the maturation stage of the cells. None of the above methods are better than the other; usually more than one test should be done and the results compared.

The principal component analysis allows to obtain uncorrelated variables and then to remove the multicollinearity problem.

4.2.3.2. Principal Component Regression (PCR): multivariate regression following PCA

Once a set of *k* principal components has been obtained using the PCA method, they can be used as input variables for a multivariate regression analysis instead of the original data. The regression equation $\mathbf{Z} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\varepsilon}$, shown in section 4.2.1, can be written as $\mathbf{Z} = \boldsymbol{\beta}\mathbf{T}_k + \boldsymbol{\varepsilon}$, where $\mathbf{T}_k$ is the matrix of the principal components (scores matrix) and the regression coefficients $\boldsymbol{\beta}$ can be estimated by least squares. When the number of principal components

is equal to the number of variables, this method becomes equivalent to the LMVR. By removing correlations in the original data, the PCR method allows to perform linear regression on a multicollinear dataset.

4.2.3.3. Partial Least Squares (PLS)

Another way to face the multicollinearity problem is to use PLS. The goal of PLS regression is to predict **Z** from **Y** and to describe their common structure [50].

In the PCR method described above, the principal components are selected based on their ability of explaining the variance of the **Y** matrix (the dependent variable matrix). By contrast, PLS regression finds components from **Y** that are also relevant for **Z**. Specifically, PLS regression searches for a set of components that performs a simultaneous decomposition of **Y** and **Z**, with the constraint that these components explain as much as possible the covariance between **Y** and **Z**. In this way, compared to the PCR, the principal components contain more information about the relationship between predictors and dependent variables [50]. For categorical dependent variables, the PLS method takes the name of partial least square discriminant analysis (PLS-DA) [43].

4.3. Multivariate classification techniques

Classification methods can be divided into two main categories, supervised and unsupervised. Supervised techniques require the knowledge of the group membership of the observations and can be used to understand the structure of the data, e.g. why certain observations belong to a given group. Moreover, once the classification model is calibrated on a "training" dataset, it can be used in a predictive way to group observations whose group membership is unknown.

On the other hand, unsupervised methods try to group the observations without any knowledge of the group membership.

In the following paragraph, we will describe the main multivariate classification approaches.

4.3.1. Discriminant Analysis (DA)

Discriminant analysis is mainly a supervised technique which was originally developed by Ronald Fisher as a way to subdivide a set of taxonomic observations into two groups based on some measured features [51]. Later, DA was extended to treat cases where there are more than two groups, the so-called "multiclass discriminant analysis" [49, 52, 53].

DA can have mainly two objectives. First, it can be used in a supervised way to describe and explain the differences among the groups. As we will see later, mathematically DA finds the optimal hyperplane that separates the groups among each other. Or, in other words, it finds the optimal linear combination of the original variables that maximizes the distance among the groups. The transformed observations are called discriminant functions.

The use of a linear combination implies that each original variable is weighted by a coefficient which can be used to study the relative importance of the variable in the separation among the groups. A second possible role of DA is to classify observations into groups. An observation, which has to be assigned to a group, is evaluated by a discriminant function (already calibrated on another dataset) and it is assigned to one of the groups at which most likely it belongs [43, 44, 49]; in this view DA is used as an unsupervised method.

When only linear transformations are applied to the variables used as DA input, the discriminant analysis is called linear discriminant analysis (LDA).

In some cases, LDA alone is not suitable and the original variables can be mapped to a new space via any non-linear function. Then, the LDA is applied in this non-linear space (which is equivalent to non-linear classification in the original space). This procedure can be seen under several names such as "non-linear DA" (NLDA) or "kernel Fisher discriminant analysis" (KFD) or "generalized discriminant analysis".

In the following sections we will focus on LDA, first describing the descriptive approach and subsequently the classification approach.

4.3.1.1. Linear DA (LDA) as a descriptive method

The initial dataset is an ensemble of multivariate observations partitioned into *G* distinct groups (e.g. different experimental treatments, times or conditions). Each of the *G* groups contains n_g observations, where *g* runs from *1* to *G* and refers to the g-th group. The multivariate observation vectors can be written as $\mathbf{y}_{gj}$ where *g* is the g-th group and *j* is the j-th observation. The vector has size *m*, which corresponds to the number of variables.

Our goal in LDA is to search for the linear combination that optimally separates our multivariate observation into *G* groups.

The linear transformation of $\mathbf{y}_{gj}$ is written as

$$z_{gj} = \mathbf{w}^T \mathbf{y}_{gj} \tag{3}$$

Since z_{gj} is a linear transformation of $\mathbf{y}_{gj}$, the mean of the group *g* of the transformed data can be written as

$$\left\langle z_g \right\rangle = \mathbf{w}^T \left\langle \mathbf{y}_g \right\rangle \tag{4}$$

where $\mathbf{y}_g$ is the mean, of the observations within a group, obtained as

$$\left\langle \mathbf{y}_g \right\rangle = \sum_{j=1}^{n_g} \mathbf{y}_{gj} / n_g$$

We now introduce the between groups sum of squares **B** in equation 5 (measure of the dispersion among the groups) and the within-group sum of squares **E** in equation 6

(measure of the dispersion within one group). First, we define them for the uni-dimensional case relatively to the untransformed data:

$$\mathbf{B}(y) = \sum_{g=1}^{G} \left(\langle y_g \rangle - \langle y \rangle \right)^2 \tag{5}$$

and

$$\mathbf{E}(y) = \sum_{g=1}^{G} \sum_{j=1}^{n_g} \left(y_{gj} - \langle y_g \rangle \right)^2 \tag{6}$$

where $\langle \mathbf{y} \rangle = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{n_g} \sum_{j=1}^{n_g} \mathbf{y}_{gj}$ is the total average of the data.

Analogously, in the multivariate case (where each observation is constituted by m variables) we have the two matrices:

$$\mathbf{B}(\mathbf{y}) = \sum_{g=1}^{G} n_g \left(\langle \mathbf{y}_g \rangle - \langle \mathbf{y} \rangle \right) \left(\langle \mathbf{y}_g \rangle - \langle \mathbf{y} \rangle \right)^T \tag{7}$$

and

$$\mathbf{E}(\mathbf{y}) = \sum_{g=1}^{G} n_g \sum_{j=1}^{n_g} \left(\langle \mathbf{y}_{gj} \rangle - \langle \mathbf{y}_g \rangle \right) \left(\langle \mathbf{y}_{gj} \rangle - \langle \mathbf{y}_g \rangle \right)^T \tag{8}$$

Finding the optimal linear combination that separates our multivariate observations into k groups means to find the vector $\boldsymbol{w}$ which maximizes the rate between the between-groups sum of squares over the within-groups sum of squares. Using the equations for the transformed data (equations 3 and 4) into the equations 7 and 8, we can write:

$$\lambda = \frac{\mathbf{w}^T \mathbf{B}(\mathbf{y}) \mathbf{w}}{\mathbf{w}^T \mathbf{E}(\mathbf{y}) \mathbf{w}} = \frac{\mathbf{B}(\mathbf{z})}{\mathbf{E}(\mathbf{z})} \tag{9}$$

We want to find $\boldsymbol{w}$ such that λ is maximized.

Equation 9 can be rewritten in the form $\mathbf{w}^T \left(\mathbf{B}\mathbf{w} - \lambda \mathbf{E}\mathbf{w} \right) = 0$; then we search for all the non trivial ($\mathbf{w}^T = 0$ is excluded) solutions of this equation and we choose the one which gives the maximum value of λ. This means to solve the eigenvalue problem $\mathbf{B}\mathbf{w} - \lambda \mathbf{E}\mathbf{w} = 0$ that can be written in the usual form:

$$\left(\mathbf{A} - \lambda \mathbf{I} \right) \mathbf{w} = 0 \tag{10}$$

where $\mathbf{A} = \mathbf{E}^{-1}\mathbf{B}$.

The solutions of equation 10 are the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_m$ associated to the eigenvectors $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w}_m$. The solutions are ranked for the eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_m$. Hence, the first eigenvalue λ_1 corresponds to the maximum value of equation 9.

The discriminant functions are then obtained considering only the first s positive eigenvalues and multiplying the original data by the eigenvectors $\mathbf{z}_1 = \mathbf{w}_1^T\mathbf{Y}, \mathbf{z}_2 = \mathbf{w}_2^T\mathbf{Y}, ..., \mathbf{z}_s = \mathbf{w}_s^T\mathbf{Y}$.

Discriminant functions are uncorrelated but not orthogonal since the matrix $\mathbf{A} = \mathbf{E}^{-1}\mathbf{B}$ is not symmetric.

In many cases the first two or three discriminant functions account for most of $\lambda_1 + \lambda_2 + ... + \lambda_s$. This allows to represent the multivariate observations as 2 or 3 dimensional points which can be plotted on a scatter plot. These plots are particularly helpful to visualize the separation of our observations into the different groups. Moreover, we can deduce, looking at the scatter plot, the meaning of a given discriminant function, i.e. we can associate the discriminant function to a given property of the analyzed system.

The weighting vectors $\mathbf{w_1}, \mathbf{w_2}, ... \mathbf{w}_s$ are called unstandardized discriminant function coefficients and give the weight associated to each variable on every discriminant function.

If the variables are on very different scales and with different variance, to assess the importance of each variable in the group separation the standardized discriminant functions can be used. The standardization is done by multiplying the unstandardized coefficients by the square root of the diagonal element of the within-group covariance matrix.

Another way to assess the variable importance is to look at the correlation between each variable and the discriminant function. These correlations are called structure or loading coefficients. However, it has been shown that these parameters are intrinsically univariate and they only show how a single variable contributes to the separation among groups, without taking into account the presence of the other variables [49].

4.3.1.2. Linear as a classification method

After a set of discriminant functions are calibrated as described in the previous section, the discriminant analysis can be applied to classify new observations into the most probable groups. From this point of view, the linear discriminant analysis becomes a predictive tool, since it is able to classify observations whose group membership is unknown [43, 49]. The discrimination ability of our LDA model can be tested by a procedure called "re-substitution" [49]. This method consists of producing an LDA model using our dataset (i.e. finding the optimal w). Then, each observation vector is re-submitted to the classification function ($\mathbf{z}_{gj} = \mathbf{w}^T\mathbf{y}_{gj}$) and assigned to a group. Since we know the group membership of the submitted vector, we can count the number of observations correctly classified and the number of observations misclassified. To measure the classification accuracy we can count

the number of observations correctly classified and the number of observations misclassified. Then, we can estimate the classification rate as the number of correctly classified observations over the total number of observations. In general, in evaluating the accuracy of a model, we have then to distinguish between two types of accuracy: the fitting accuracy and the prediction accuracy [43, 54].

The fitting accuracy is the ability to reproduce the data, namely how the model is able to reproduce the data that were used to build the model (the training set). This corresponds to the apparent classification rate and it is obtained using the re-substitution procedure.

The prediction accuracy is the ability to predict the value or the class of an observation, that was not included in the construction of the model. This kind of accuracy is often referred to as the ability of the model to generalize. The data used to measure this accuracy are called "test set". The prediction accuracy can be called "actual classification rate". This is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. To have an estimation of the actual classification rate, two main procedures can be applied: the hold-out and cross-validation [43].

In the hold-out, the dataset is divided into two partitions: one partition is used to develop the model (e.g. the discriminant functions) and the second partition is given as input to the model. The first partition is usually called "training set" or "calibration set", while the second partition is the validation set [54].

When the number of observations is small, the cross-validation is usually preferred over the hold-out. The basic idea of the cross-validation procedure is to divide the entire dataset into L disjoint sets. L-1 sets are used to develop the model (i.e. the calibration set on which the discriminant functions are computed) and the omitted portion is used to test the model (i.e. the validation set given as input to the model). This is repeated for all the L sets and an average result is obtained.

Apparent or actual classification accuracies can be summarized in a confusion matrix. As an example, total N observations, n_1 , belong to the group 1 and n_2 belong to the group 2. C_{11} is the total number of observations correctly classified in group 1 and C_{12} is the total number of data misclassified in group 2. Similarly, C_{22} is the total number of observations correctly classified in group 2 and C_{21} is the number of misclassified in group 1.

The confusion matrix becomes then:

	Actual group	Predicted group
	1	2
1	C_{11}	C_{12}
2	C_{21}	C_{22}

and the accuracy (the actual or apparent classification rate (acr)) is computed as: $acr = \frac{C_{11} + C_{22}}{n_1 + n_2}$

4.3.2. PCA-LDA

A powerful analysis tool is the combination of the principal component analysis with the linear discriminant analysis [52]. This is particularly helpful when the number of variables is large. In particular, if the number of observations (N) is less than the number of variables (m) - specifically $N-1<m$ - the covariance matrix is singular and cannot be inverted (see section 4.2.3.). We then need to find a way to reduce the number of variables, for example using the PCA [49, 55]. This procedure has been widely used for several problems in different fields [35, 52, 56-60]. The condition $N-1<m$ almost always appears in spectroscopy, where the number of observations (N) is usually 10^2 and the number of variables (m) is typically within 10^2 to 10^3.

Let's take into account the same situation described for the many group linear discriminant analysis. The original dataset is an ensemble of multivariate observations which is partitioned into k distinct groups. Again, we want to find the discriminant functions which optimally separate our multivariate observation into the k groups. Then, the discriminant functions can be used to identify the most important variables in terms of ability of distinguishing among the groups. Thus, first the original dataset is submitted to the PCA to reduce the number of variables; subsequently, the reduced dataset is analyzed using the LDA.

Another way that can be used instead of PCA is to perform the PLS.

4.3.3. PLS-LDA

In a way analogous to the PCA-LDA procedure, here we first apply the PLS algorithm to the original data and then the LDA on the selected principal components [61].

Given that the PLS searches for a set of components that performs a simultaneous decomposition of the dependent and independent datasets, the main difference with PCA-LDA is that the principal components resulting as output of PLS better describe the relationship between independent and dependent variables. This does not necessarily mean that this method is better in general. Indeed, applying PCA or PLS on the same dataset often leads to similar results [62, 63] and the classification accuracy or the descriptive ability is mostly determined by the underlying structure of the data which can make one of the two methods more suitable than the other.

4.3.4. Cluster Analysis (CA)

The goal of cluster analysis is to find the best grouping of the multivariate observations such that the clusters are dissimilar to each other but the observations within a cluster are similar [44].

CA is an unsupervised technique, that is, the group membership of the observations (and often the number of groups) is not known in advance.

At first we have to define a measure of similarity or dissimilarity also called distance functions. The most common distance functions are: i) the Euclidean distance; ii) the Manatthan distance; iii) the Mahalanobis distance; iv) the maximum norm.

Based on the procedure they use, clustering algorithms can be divided into three main groups: hierarchical, partitional and density-based clustering. None of the following algorithms is better than the other. The choice of the clustering method strongly depends on the structure of the data and on which kind of results one would expect.

Hierarchical clustering algorithms can be again subdivided into agglomerative or divisive. The agglomerative clustering starts with all observations placed in different clusters and in each step an observation or a cluster of observations is merged into another cluster. The most commonly employed agglomerative clustering strategies are complete-linkage, average-linkage, single-linkage, centroid-linkage. The drawback of the agglomerative clustering algorithms is that observations cannot be moved among the clusters once a cluster is made.

The divisive method starts with one single cluster containing all observations and then it divides the cluster into two sub-clusters at each step. Divisive methods have the same drawback of the agglomerative clustering, that is, once a cluster is made, an observation cannot be moved to another cluster. Divisive methods are suited when large clusters are searched for.

The partitional algorithm assigns the observations to a set of clusters without using hierarchical approaches. One of the most used non-hierarchical approach is the k-means clustering.

The density-based clustering seeks to search for regions of high density without any assumption about the shape of the cluster.

4.4. Artificial Neural Networks (ANN)

The artificial neural networks are mathematical models that were developed in analogy to a network of biological neurons [64]. Mathematically, a neuron can be modeled as a switch that receives, as input, a series of values and produces an output consisting of a weighted sum of the input eventually transformed by a function f. Many neurons can be combined to create more complex networks. Depending on the type of neurons and on how the neurons are connected to each others, different kinds of neural networks can be created. The most common type of neural network is the feed-forward neural network, in which neurons are grouped into layers, each neuron of a layer is connected to all the neurons of the next layer and the information flows from the input to the output without loops. For a comprehensive description of neural networks and their applications see [54, 65].

5. Applications of multivariate analysis to spectroscopic data of complex biological systems

In the following, we will provide a few selected examples of the application of FTIR microspectroscopy coupled with multivariate analysis for biomedical relevant studies, with

the aim to highlight the importance of linking the two approaches to extract the most significant spectral information from highly informative systems.

In some cases, PCA alone represents a powerful method for the analysis of multidimensional FTIR spectra. Indeed, several interesting works are reported in the literature, in which this approach is employed to support the spectroscopic investigation of complex biological systems and processes. For instance, synchrotron based FTIR microspectroscopy coupled with PCA has been applied to the characterization of human corneal stem cells [27, 66], in cancer research for the screening of cervical cancer [14], as well as to disclose the effects induced by a surface glycoprotein in colon carcinoma cells [67].

For instance, Matthew German and colleagues [68] coupled high-resolution synchrotron radiation-based FTIR (SR-FTIR) microspectroscopy with PCA to investigate the characteristics of putative adult stem cell (SC), transiently amplified (TA) cell, and terminally differentiated (TD) cell populations of the corneal epithelium. Using PCA, each spectrum, composed by many variables (the wavenumbers), is reduced to a point in a low dimensional space. Then, each observation can be visualized in a two or three dimensional score plot. Choosing the appropriate principal components, the authors were able to clearly distinguish the three cell populations confirming the ability of SR-FTIR microspectroscopy to identify SC, TA cell, and TD cell populations.

PCA alone is extremely powerful to reduce the number of variables; however, it is not a clustering algorithm and the group into clusters must be done with other techniques.

For example, Tanthanuch and colleagues applied FTIR microspectroscopy-supported by PCA and unsupervised hierarchical cluster analysis (UHCA) to identify specific spectral markers of the differentiation of murine embryonic stem cell (mESCs) and to distinguish them into different neural cell types [25]. In particular, focal plane array (FPA) - FTIR and SR-FTIR microspectroscopy measurements - performed on cell clumps and single cells respectively - allowed to obtain a biochemical fingerprint of different mESC developmental stages, namely embryoid bodies (EBs), neural progenitor cells (NPCs) and embryonic stem-derived neural cells (ESNCs). Interestingly, it should be noted that the results obtained on cell clumps and on single cells were found to be comparable, corroborating the FPA-FTIR results on cell clumps. The analysis of second derivative spectra enabled to highlight important spectral changes occurring during ES cell differentiation, mainly in the lipid CH_2 and CH_3 stretching region and in the protein amide I band. Noteworthy, these results overall indicated that during neural differentiation the cell lipid content increased significantly, likely reflecting modifications in cell membranes, whose lipid content is known to have a key role in neural cell differentiation and signal transduction. Moreover, changes in the profile of amide I band, mainly involving the alpha-helix component around 1650-1652 cm^{-1}, indicated an increased expression of alpha-helix reach protein in ESNCs compared with their progenitor cells, a result that could reflect the expression of cytoskeleton protein, crucial for the establishment of neural structure and function. These results were then strongly supported by PCA, that made it possible to disclose regions of the IR spectrum which most contributed to the spectral variance, namely amide I band and C-H

stretching region. Furthermore, the application of UHCA allowed to successfully discriminate and classify each stage of ESNCs differentiation, again considering the spectra in the spectral range mainly due to acyl chain vibrations and the extended region between 1750 and 900 cm^{-1}.

As discussed previously, PCA is frequently used for preliminary dimensionality reduction before further analyses, as LDA [21]. Indeed, a limit of using PCA alone is that it does not allow to obtain an unambiguous grouping of the data into clusters, requiring therefore the application of another analysis step able to reduce the intra-category variation while maximizing that inter-category [69]. The coupling, for instance, of PCA with LDA is a well established procedure which enables not only to classify the observations into groups but to quantify the importance of the single variables for this group separation. In this view, the advantage of LDA is that it makes it possible to reveal clusters, identifying objectively also the most contributory wavenumbers responsible for spectra discrimination [21, 58]. In particular, the application of PCA-LDA to spectroscopic investigation of complex biological systems proved to be a useful tool for the identification of spectral biomarkers of the process under investigation [7, 35, 69, 70, 71].

One outstanding work, worth to mention here, was done by Kelly and colleagues [70], where the authors showed how infrared spectroscopy and multivariate techniques can be used as a novel diagnostic approach for endometrial cancer screening. They first demonstrated how SR-FTIR microspectroscopy with subsequent PCA-LDA allows the clear segregation of different subtypes of endometrial carcinoma. However, the requirement of a particle accelerator impairs the use of endometrial spectroscopy as practical diagnostic application.

Recently, Taylor and colleagues applied ATR-FTIR spectroscopy supported by PCA-LDA analysis to interrogate endometrial tissues, employing in particular a conventional IR radiation source [72], showing that this approach, that can be applied directly to liquid or solid samples without further preparation, could provide a useful and simple objective test for endometrial cancer diagnosis.

Furthermore, in the work of Walsh and colleagues [69], ATR microspectroscopy has been successfully applied to the characterization of samples of exfoliative cervical cytology of different categories, with increasing severity of atypia. The spectral analysis was supported by PCA, with or without subsequent LDA, to verify if it was possible to discriminate among normal, low grade and high grade of exfoliative cytology. Indeed, important differences were found in the spectral range between 1500 and 1000 cm^{-1}, mainly due to proteins, glycoproteins, phosphates and carbohydrates. Noteworthy, the authors stressed that only the employment of the combined PCA-LDA allowed to maximize the inter-category variance, whilst reducing that intra-category. In particular, they found that the glycogen content strongly influenced the intra-category variance, while that inter-category resulted to be mainly due to protein and DNA conformational changes. In this view, FTIR microspectroscopy coupled with PCA-LDA could allow for an objective classification approach to class cervical cytology.

We should note that a delicate point of PCA-LDA is the choice of the principal components to be used as LDA input and, as described in the previous section about PCA, several ways have been developed to perform this task. Alternatively, the PLS method can be used instead of PCA [6, 73, 74]. For instance, Sandt and colleagues, using synchrotron infrared microspectroscopy coupled with PLS-DA, were able to characterize the metabolic fingerprint of induced pluripotent stem cells (iPSCs). In particular, they found that iPSCs are characterized by a chemical composition that leads to a spectral signature indistinguishable from that of embryonic stem cells (ESCs), but entirely different from that of the original somatic cells [6].

5.1. FTIR microspectroscopy supported by PCA-LDA for the characterization of SN and NSN murine oocytes

Recently, we applied FTIR microspectroscopy supported by PCA-LDA to the study of murine oocytes characterized by two different types of chromatin organization, namely surrounded nucleolus (SN) oocytes in which the chromatin is highly condensed and forms a ring around the nucleolus, and the not surrounded nucleolus (NSN) type where chromatin is dispersed and less condensed around the nucleolus [7, 75]. Interestingly, only SN oocytes are capable to complete the embryonic development after fertilization, while the NSN type, if fertilized, arrests at the two cell stage. To try to get new insights on the mechanisms that drive the different chromatin organization in the two kinds of oocytes, crucial for their embryonic development after fertilization, we studied the infrared absorption of single intact cells at different maturation stages, namely antral germinal vesicle (GV), metaphase I (MI, matured for 10 hours in vitro), and metaphase II (MII, matured for 20 hours in vitro).

Indeed, as we will show in the following, the FTIR spectra of the oocytes taken at the different maturation stages are very complex, since they provide information on different processes that were taking place simultaneously within the cells. For this reason, beside a fundamental visual inspection of the data, enabling the identification and assignment of the different spectral bands, it was crucial the application of PCA-LDA that made it possible to draw out the most significant spectral information responsible for the different cell behavior. Moreover, PCA-LDA allowed to identify the stage at which the separation between the SN and NSN oocytes took place, leading to their well distinct cell destinies.

As we discussed in paragraph 2, since the FTIR spectrum of cells is due to the overlapping contributes of the main biomolecules (see Figure 2), we analysed the second derivative spectra to identify the band peak positions and to assign them to the different biomolecule vibrational modes. The spectral analysis, strongly supported by PCA-LDA, allowed us to disclose the most important spectral differences between the two types of oocytes, at each maturation stage, that were found to occur mainly in the lipid and nucleic acid absorption regions, as we will discuss below. For a full discussion of the results see [7].

5.1.1. Lipid analysis

5.1.1.1. NSN oocytes

The analysis between 3050 and 2800 cm^{-1}, mainly due to the lipid carbon-hydrogen stretching vibrations [29], disclosed significant variations in the lipid content of NSN oocytes during their maturation up to MII. Indeed, besides an increase of the CH_2 band intensity up to MII, respectively at 2922 cm^{-1} and 2852 cm^{-1}, important changes concerned mainly the unsaturated fatty acid composition, as indicated by variations of the band between 3020 and 3000 cm^{-1} due to the olefinic group absorption. Indeed, as shown in Figure 3A, a single peak around 3013 cm^{-1} was present at GV and MI stages, while a splitting in two components at ~ 3016 cm^{-1} and at ~ 3010 cm^{-1} characterized the MII stage (see the inset of Figure 3A). These results could reflect important changes in membrane fluidity, which in turn could confer to the oocyte a different division ability after fertilization [8].

5.1.1.2. SN oocytes

SN oocytes were found to be characterized - during maturation up to MII - by a significant increase of the 2937 cm^{-1} component that could be likely due to cholesterol and/or phospholipids (Figure 3B) [76, 77]. As discussed for NSN oocytes, the observed changes could reflect variations in the membrane properties, again highlighting the crucial role of lipids as markers of oocyte developmental competence [8, 78].

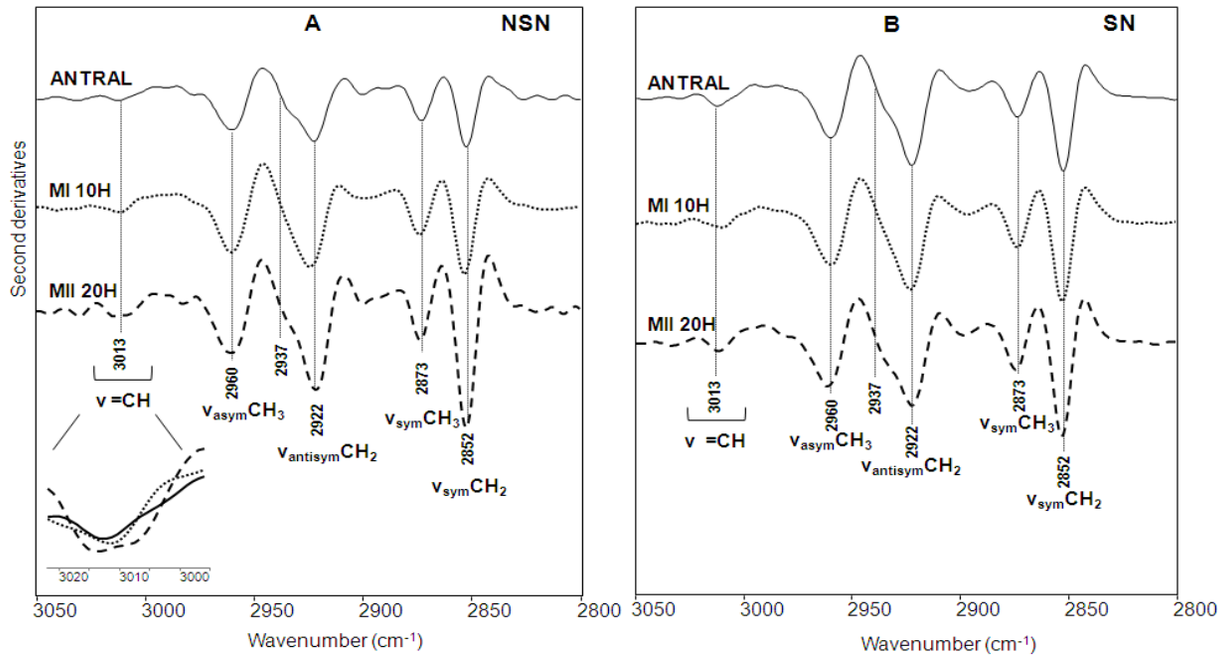

Figure 3. Second derivative absorption spectra of NSN (A) and SN (B) oocytes in the lipid absorption region. The second derivatives of the FTIR absorption spectra of single oocytes, measured at the antral (continuous line), MI 10 H (dotted line), and MII 20 H (dashed line) stages, are reported in the acyl chain absorption region, after normalization at the tyrosine peak (~1516 cm^{-1}). In the inset a magnification of the olefinic group band is shown.

5.1.1.3. PCA-LDA analysis

The results obtained by the direct inspection of second derivative spectra were confirmed by PCA-LDA analysis performed on raw spectra. Firstly, the analysis was made on each type of oocyte taken at the different maturation stages. For the SN oocytes, the component carrying the highest discrimination weight resulted that at 2938 cm^{-1}, likely due to cholesterol and / or phospholipids [76, 77], in agreement with what found by the direct inspection of the spectra.

Concerning the NSN oocytes, on the other hand, the wavenumbers with the highest discrimination weight were the 2922 cm^{-1}, due to the CH_2 stretching vibration, which increases up to MII, and the 3018 cm^{-1}, assigned to the olefinic group =CH of polyunsaturated fatty acids, whose absorption was observed to vary during the oocyte maturation.

We, then, compared the two types of oocyte at each maturation stage - as illustrated in Figure 4 - and we found that at the antral and MII stages the spectral components with the highest discrimination weight were those due to cholesterol and /or phospholipids, while at MI was that due to the olefinic group. Furthermore, to support the crucial role played by lipids in determining at some extent the oocyte developmental capacity, we should add that when we compared by PCA-LDA the spectra of the two oocyte types at the same maturation stage in the 1800-1500 cm^{-1} spectral range, dominated by the amide I and amide II absorption, the wavenumber with the highest discrimination weight was the 1739 cm^{-1}, due to the carbonyl stretching vibration of esters [7, 29].

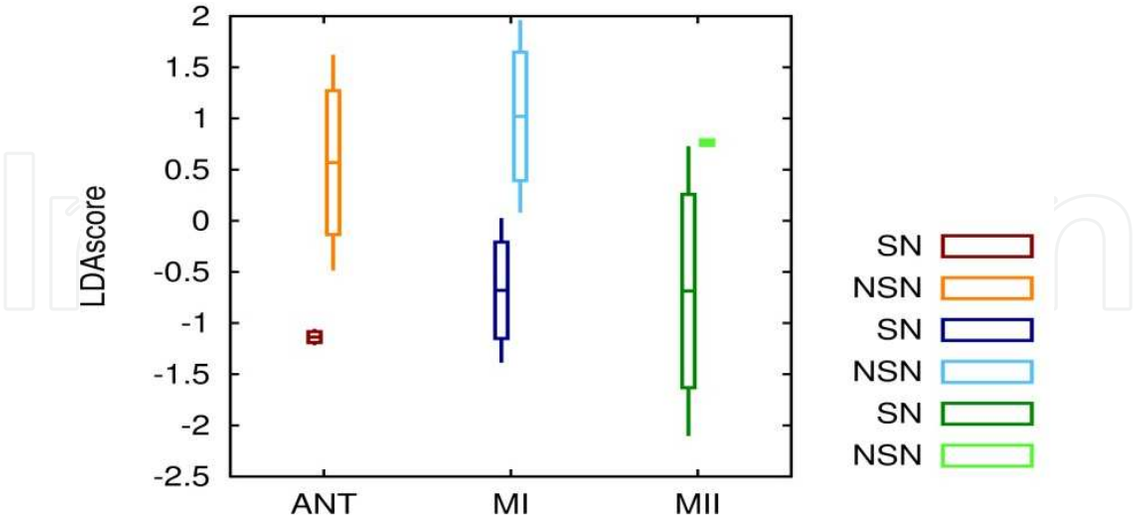

Figure 4. PCA-LDA analysis of SN and NSN oocytes in the lipid acyl chain absorption region (3050 – 2800 cm^{-1}). The separation between the two types of oocytes at each maturation stage is reported as average of PCA-LDA scores. The height of the boxes and the whiskers corresponds to 1 and 1.5 standard deviations from the mean values, respectively. The analysis has been performed on the measured spectra.

5.1.2. Nucleic acid analysis

5.1.2.1. NSN oocytes

We then analyzed the nucleic acid IR response of NSN and SN oocytes during their maturation, exploring the spectral region between 1000 and 800 cm^{-1}, where RNA and DNA vibrational modes mainly occur [31, 32].

We found that NSN oocytes maintain, in all the studied stages, an appreciable transcriptional activity as indicated mainly by the simultaneous presence of the RNA ribose component around 921 cm^{-1} and of the DNA deoxyribose between 895-898 cm^{-1} - indicative of a DNA/RNA hybrid - whose relative intensities were seen to vary during maturation (see Figure 5A). In particular, the intensity of these two components is higher at the antral stage, while it decreases at MI, to increase again up to MII. These results were also supported by the response of the complex band between 980-950 cm^{-1}, mainly due to the CC stretching vibration of DNA backbone. Indeed, the profile of this band varies depending on the DNA structure that, in turn, could reflect a different nucleic acid activity. In particular, for the NSN oocytes we found that at the antral stage DNA is mainly in A-form - with a triplet at 975 cm^{-1}, 966 cm^{-1} and 951 cm^{-1} - typical of the DNA/RNA hybrid during transcription. At MI, the reduction of the 975 cm^{-1} and 966 cm^{-1} bands and the appearance of that at 969 cm^{-1} indicate that DNA is mainly in the B-form, suggesting a sort of transcriptional "stand by state", further supported by the reduction extent of the DNA/RNA hybrid, as discussed above. From this "stand by state" NSN oocytes seem to resume their transcriptional activity at MII, where a coexistence of DNA A and B forms was observed, as indicated by the increase of the ~ 975 cm^{-1} band and again in agreement with the simultaneous increase of the ribose (921 cm^{-1}) and deoxyribose (898 cm^{-1}) components.

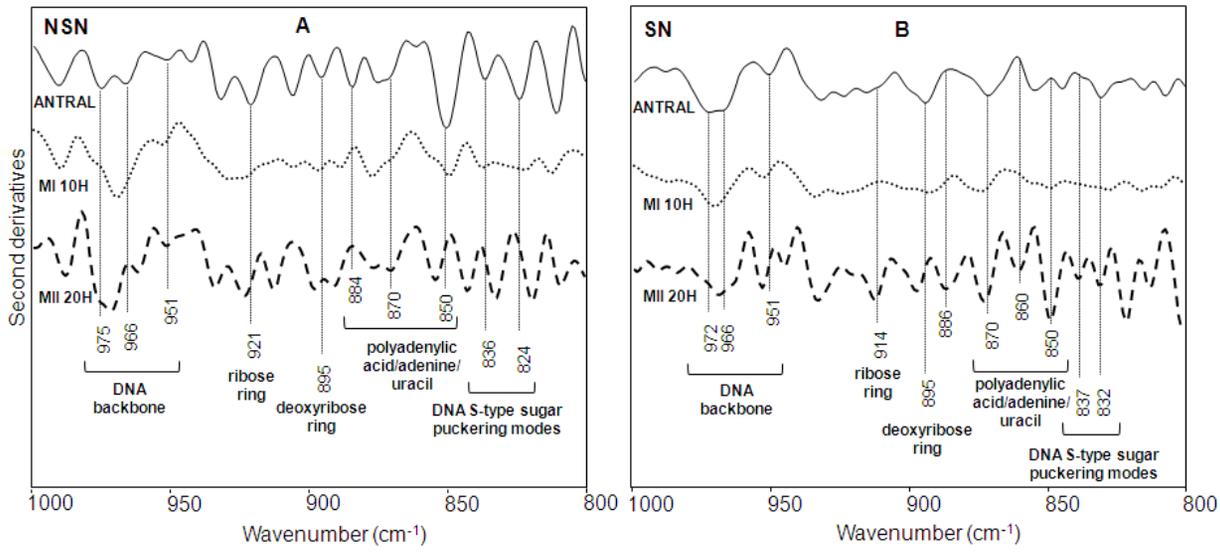

Figure 5. Second derivative absorption spectra of NSN (A) and SN (B) oocytes in the nucleic acid absorption region. The second derivatives of the FTIR absorption spectra of single oocytes, measured at the antral (continuous line), MI 10 H (dotted line), and MII 20 H (dashed line) stages, are reported in the 1000-800 cm^{-1} absorption region, after normalization at the tyrosine peak (~1516 cm^{-1}).

Furthermore, the analysis of the low frequency range, between 840-820 cm^{-1}, allowed us to obtain information on DNA methylation. In particular, in this spectral range, bands due to DNA S-type sugar puckering modes occur, which are sensitive to changes in the DNA sugar conformation induced by cytosine methylation [32]. The possibility to monitor changes in the profile of this spectral region in whole intact cells makes it possible, therefore, to obtain information on the variation of global DNA methylation in the CpG islands. In this way, we found that in the NSN oocytes DNA methylation was high at the antral stage, while it became very low, almost negligible at MII, in agreement with what found for the transcriptional activity pattern at the different maturation stages.

Finally, significant spectral differences were found between 890 and 850 cm^{-1}, where four different bands due to adenine and uracil vibrational modes occur (see Figure 5) [79]. Interestingly, the relative variation of these bands enables to monitor the mRNA polyadenylation extent, a crucial mechanism that regulates transcription. We found, in particular, that NSN oocytes were characterized during maturation by a low level of mRNA polyadenylation, being the polyadenylic acid band at 884 cm^{-1} absent at MII, while a new band at 854 cm^{-1} - likely due to adenine possibly not involved in polyA tail [80] – appeared. These results seem to suggest that an inadequate level of mRNA polyadenylation could preclude the possibility to resume meiosis, leaving the NSN oocytes in an unsuccessful transcriptional state.

5.1.2.2. SN oocytes

The analysis of SN oocytes (Figure 5B) in the spectral range between 1000 and 800 cm^{-1} led to very different results compared to NSN oocytes (see Figure 5A). Briefly, during all the studied maturation stages, the SN oocyte transcriptional activity was found to be maintained at lower levels than NSN oocytes, as revealed by the analysis of the CC stretching of the DNA backbone (980-950 cm^{-1}) and the monitoring of the ribose (~ 922 cm^{-1}) and deoxyribose (895-898 cm^{-1}) vibrations. These results were supported by the temporal evolution of the DNA methylation bands that suggested a partial CpG methylation at the antral and MI stages, which dramatically increased at MII, contrary to what observed for NSN oocytes.

Noteworthy, while no evidence of mRNA polyadenylation was observed for SN oocytes at the antral stage - as indicated by the absence of the two polyadenylic acid bands around 884 cm^{-1} and 860 cm^{-1} - starting from MI the adenine and uracile bands at 870 cm^{-1} and 850 cm^{-1} appeared, to then dramatically increase up to MII. These findings likely indicate that SN MII oocytes are characterized by an adequate level of maternal polyadenylated mRNAs, making them ready to sustain a proper embryo development, contrary to NSN oocytes.

5.1.2.3. PCA-LDA analysis

The above results overall indicate that the IR spectra of oocytes at different maturation stages are very informative in the nucleic acid absorption region, allowing to obtain information on several cell processes simultaneously, including transcriptional activity, DNA methylation, and RNA polyadenylation. For this reason, PCA-LDA analysis was

crucial to disclose the most significant spectral response, enabling to identify the marker bands able to discriminate between the two kinds of oocytes.

Firstly, we analyzed the different maturation stages of each kind of oocyte. In particular, NSN oocytes displayed a segregation into three separated clusters, each corresponding to a maturation stage, with a classification accuracy of about 80%. Noteworthy, the wavenumber with the highest weight (1.0) was that around 880 cm^{-1}, due to polyadenylic acid, that, as revealed by second derivative analysis, was present only at the antral stage and disappeared upon maturation up to MII.

On the other hand, PCA-LDA analysis of SN oocytes led to an excellent discrimination accuracy (97%), with the wavenumbers with the highest discrimination weight at 817 cm^{-1} (1.0) and 859 cm^{-1} (0.83). While this last component is due to polyadenylic acid, the assignment of the 817 cm^{-1} band is not unequivocal, being due to overlapping contributions of DNA and polyadenylic acid.

The above results were then confirmed by the PCA-LDA analysis performed between 1400-1000 cm^{-1}, where contributions due to nucleic acids, such as sugar-phosphate vibrations, also occur [31]. In particular, for the NSN oocytes the wavenumber with the highest discrimination weight (1.0) was the 1305 cm^{-1}, which is due to free adenine, possibly not involved in polyadenylation [79]. In agreement with the temporal pattern of the adenine band at 870 cm^{-1}, discussed previously, the 1305 cm^{-1} component displayed a higher intensity at MII, confirming that an inadequate mRNA polyadenylation could preclude NSN oocytes from a successful embryonic development (see Figure 6).

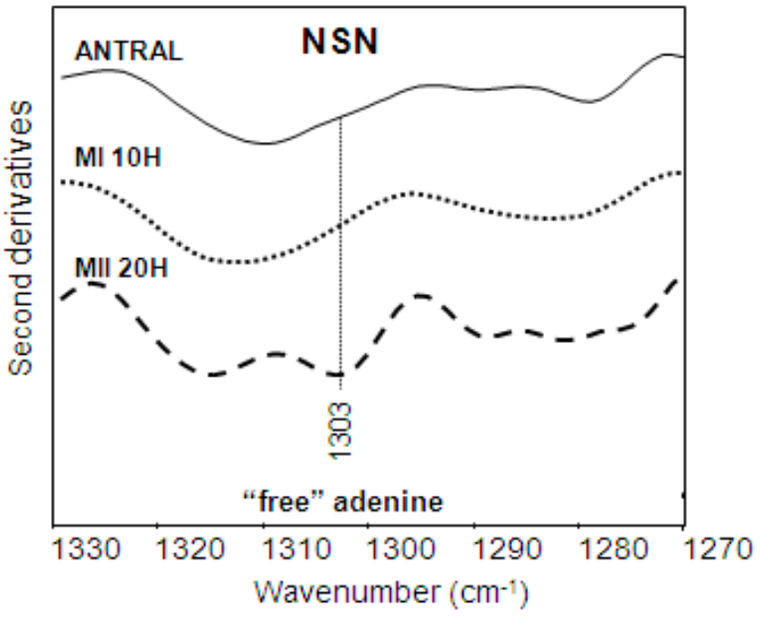

Figure 6. Second derivative absorption spectra of NSN oocytes in the absorption region of "free" adenine. The second derivatives of the FTIR absorption spectra of single NSN oocytes, measured at the antral (continuous line), MI 10 H (dotted line), and MII 20 H (dashed line) stages, are reported in the 1330-1270 cm^{-1} spectral range, where "free" adenine absorbs, after normalization at the tyrosine peak (~1516 cm^{-1}).

We then compared by PCA-LDA the two types of oocytes taken at the same maturation stage. As reported in Figure 7, we found the largest spectral distance at MI (92% classification accuracy), with the components carrying the highest discrimination weight due to A-DNA, likely reflecting differences in the transcriptional activity. In this view, MI stage could be considered a sort of crucial checkpoint, when some molecular rearrangements occur, deciding the oocyte fate.

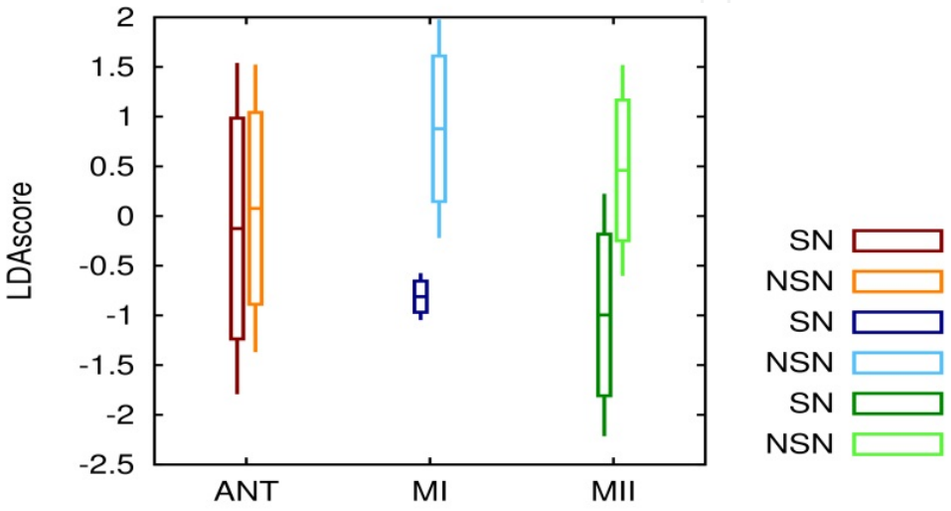

Figure 7. PCA-LDA analysis of SN and NSN oocytes in the nucleic acid absorption region (1000 - 800 cm^{-1}). The separation between the two types of oocytes at each maturation stage is reported as average of PCA-LDA scores. The height of the boxes and the whiskers corresponds to 1 and 1.5 standard deviations from the mean values, respectively. The analysis has been performed on the measured spectra.

These findings have been strongly supported by the comparison of the SN and NSN oocytes at each maturation stage, altogether. A very good discrimination accuracy (89%) was again found analyzing the nucleic acid absorption region, between 1000 and 800 cm^{-1}, that led to a clear cut separation into two groups (see Figure 8): one containing only the MII SN oocytes, and the other containing all the other SN and NSN stages. In particular, the wavenumbers carrying the highest discrimination weight were found at 926 cm^{-1} (1.00), due to ribose vibration, and at 855 cm^{-1} (0.97), assigned to adenine vibration, indicating again that differences in the temporal evolution and extent of transcription and polyadenylation play a crucial role in determining the different oocyte fate: the MII SN oocytes, with their proper content of maternal mRNAs polyadenylated, ready to support successfully the embryonic development; on the other hand, the MII NSN oocytes, with their mRNA lacking the appropriate polyadenylation, are kept in an unsuccessful transcriptional state.

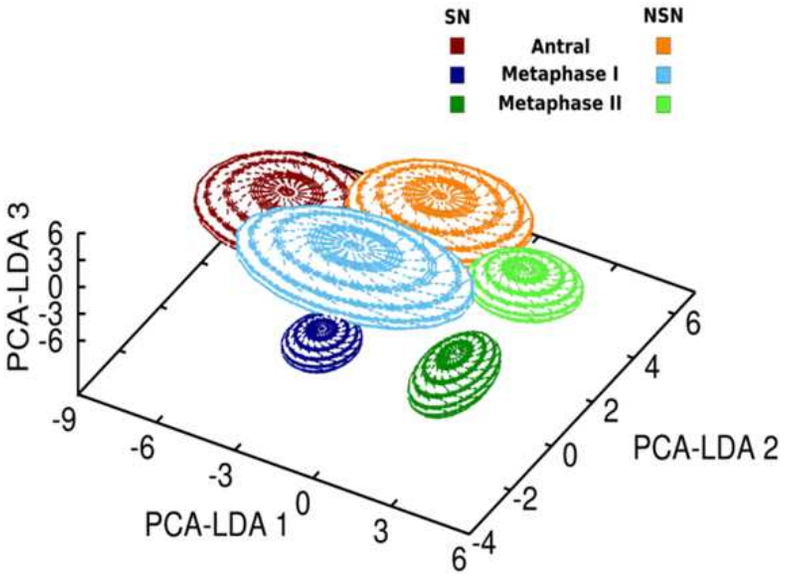

Figure 8. PCA–LDA analysis of SN and NSN oocytes in the nucleic acid absorption region. The PCA–LDA analysis has been carried out on measured FTIR absorption spectra obtained from SN and NSN oocytes at each maturation stage, between 1000 and 800 cm^{-1}. The semi-axes of ellipsoids in the 3D score plot correspond to two standard deviations of the data along each direction.

6. Conclusions

FTIR microspectroscopy has recently emerged as a powerful tool in biomedical research, thanks to the possibility of providing, in a non-invasive and rapid way, a chemical fingerprint of biological samples. In particular, being successfully applied to the study of complex biological systems, it makes it possible not only to characterize in situ biological processes, but also to provide a rapid diagnosis of several diseases, such as cancer and amyloid-based disorders.

We should, however, note that the intrinsic complexity of the IR response of biological systems - due to the overlapping absorption of the main biomolecules - requires the support of an appropriate multivariate analysis approach able to draw out the significant and non-redundant information contained in these highly dimensional data. Indeed, only a suitable combination of biospectroscopy and of multivariate analysis would provide robust and reliable results through the identification of specific biomarkers, an essential prerequisite for unbiased result interpretation [19, 20].

Author details

Diletta Ami* and Silvia Maria Doglia
Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milano, Italy
Consorzio Nazionale Interuniversitario per le Scienze Fisiche della Materia (CNISM), UdR Milano-Bicocca, Milan, Italy

* Corresponding Author

Paolo Mereghetti
Center for Nanotechnology Innovation @NEST, Italian Institute of Technology (IIT), Pisa, Italy

Acknowledgement

D. A. is indebted to the University of Milano-Bicocca (I) for the supporting postdoctoral fellowship. P. M. acknowledges a postdoctoral fellowship from Italian Institute of Technology. S.M. D. acknowledges the financial support of the FAR (Fondo di Ateneo per la Ricerca) of the University of Milano-Bicocca (I).

The authors wish to thank Carlo Alberto Redi and his collaborators at the University of Pavia (I) for the collaboration on murine oocyte maturation, and Antonino Natalello of the University of Milano-Bicocca (I) for helpful discussions.

7. References

[1] Aksoy C, Severcan F. Role of Vibrational Spectroscopy in Stem Cell Research. Spectroscopy: An International Journal 2012; 27(3) 167-184.

[2] Ami D, Mereghetti P, Natalello A, Doglia SM. Fourier transform infrared microspectroscopy as a tool for embryonic stem cell studies. In: Atwood CS. (ed.) Stem Cells in Clinic and Research. Rijeka: InTech, 2011. p. 193-218.

[3] Heraud P, Nga ES, Caine S, Yu QC, Hirst C, Mayberry R, Bruce A, Wood BR, McNaughton D, Stanley EG, Elefanty AG. Fourier transform infrared microspectroscopy identifies early lineage commitment in differentiating human embryonic stem cells. Stem Cell Research 2010; 4(2) 140-147.

[4] Chan JW, Lieu DK. Label-free biochemical characterization of stem cells using vibrational spectroscopy. Journal of Biophotonics 2009; 2(11) 656-668.

[5] Walsh M J, Hammiche A, Fellous TG, Nicholson JM, Cotte M, Susini J, Fullwood NJ, Martin-Hirsch PL, Alison MR, Martin FL. Tracking the cell hierarchy in the human intestine using biochemical signatures derived by mid-infrared microspectroscopy. Stem Cell Research 2009; 3(1) 15-27.

[6] Sandt C, Féraud O, Oudrhiri N, Bonnet ML, Meunier MC, Valogne Y, Bertrand A, Raphaël M, Griscelli F, Turhan AG, Dumas P, Bennaceur-Griscelli A. Identification of spectral modifications occurring during reprogramming of somatic cells. PLoS ONE 2012; 7(4) e30743.

[7] Ami D, Mereghetti P, Natalello A, Doglia SM, Zanoni M, Redi CA, Monti M. FTIR spectral signatures of mouse antral oocytes: molecular markers of oocyte maturation and developmental competence. Biochimica et Biophysica Acta, 2011; 1813(6) 1220–1229.

[8] Wood BR, Chernenko T, Matthäus C, Diem M, Chong C, Bernhard U, Jene C, Brandli AA, McNaughton D, Tobin MJ, Trounson A, Lacham-Kaplan O. Shedding New Light on the Molecular Architecture of Oocytes Using a Combination of Synchrotron Fourier

Transform-Infrared and Raman Spectroscopic Mapping. Analytical Chemistry, 2008; 80(23) 9065-9072.

[9] Diomede L, Cassata G, Fiordaliso F, Salio M, Ami D, Natalello A, Doglia SM, De Luigi A, Salmona M. Tetracycline and its analogues protect Caenorhabditis elegans from β amyloid-induced toxicity by targeting oligomers. Neurobiology of Disease, 2010; 40(2) 424-431.

[10] Kuzyk A, Kastyak M, Agrawal V, Gallant M, Sivakumar G, Rak M, Del Bigio MR, Westaway D, Julian R, Gough KM. Association among amyloid plaque, lipid, and creatine in hippocampus of TgCRND8 mouse model for Alzheimer disease. The Journal of Biological Chemistry, 2010; 285(41) 31202-31207.

[11] Choo LP, Wetzel DL, Halliday WC, Jackson M, LeVine SM, Mantsch HH. In situ characterization of beta-amyloid in Alzheimer's diseased tissue by synchrotron Fourier transform infrared microspectroscopy. Biophysical Journal, 1996; 71(4) 1672-1679.

[12] Kneipp J, Miller LM, Joncic M, Kittel M, Lasch P, Beekes M, Naumann D. In situ identification of protein structural changes in prion-infected tissue. Biochimica et Biophysica Acta, 2003; 639(3) 152-158.

[13] Bellisola G, Sorio C. Infrared spectroscopy and microscopy in cancer research and diagnosis. American Journal of Cancer Research, 2012; 2(1) 1-21.

[14] Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, Kyrgiou M, Stringfellow HF, Paraskevaidis E, Martin-Hirsch PL, Martin FL. IR microspectroscopy: potential applications in cervical cancer screening. Cancer Letters, 2007; 246(1-2) 1-11.

[15] Petibois C, Déléris G. Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology. Trends in Biotechnology, 2006; 24(10) 455-462.

[16] Kastyak-Ibrahim MZ, Nasse MJ, Rak M, Hirschmugl C, Del Bigio MR, Albensi BC, Gough KM. Biochemical label-free tissue imaging with subcellular-resolution synchrotron FTIR with focal plane array detector. Neuroimage, 2012; 60(1) 376-383.

[17] Miller LM, Dumas P. From structure to cellular mechanism with infrared microspectroscopy. Current Opinion in Structural Biology, 2010; 20(5) 649-656.

[18] Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. The Analyst, 2012; 137(14) 3202-3215.

[19] Kelly JG, Trevisan J, Scott AD, Carmichael PL, Pollock HM, Martin-Hirsch PL, Martin FL. Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers. Journal of Proteome Research, 2011; 10(4) 1437-1448.

[20] Wang L, Mizaikoff B. Application of multivariate data-analysis techniques to biomedical diagnostics based on mid-infrared spectroscopy. Analytical and Bioanalytical Chemistry, 2008; 391(5) 1641-1654.

[21] Martin FL, German MJ, Wit E, Fearn T, Ragavan N, Pollock HM. Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample. Journal of Computational Biology, 2007; 14(9) 1176-1184.

[22] Heraud P, Tobin MJ. The emergence of biospectroscopy in stem cell research. Stem Cell Research, 2009; 3(1) 12-14.
[23] Kazarian SG, Chan KL. Applications of ATR-FTIR spectroscopic imaging to biomedical samples. Biochimica et Biophysica Acta, 2006; 1758(7) 858-867.
[24] Holman H-YN, Martin MC, McKinney WR. Tracking chemical changes in a live cell: Biomedical applications of SR-FTIR spectromicroscopy, 2003; 17 (2-3) 139-159.
[25] Tanthanuch W, Thumanu K, Lorthongpanich C, Parnpai R, Heraud P. Neural differentiation of mouse embryonic stem cells studied by FTIR spectroscopy. Journal of Molecular Structure, 2010; 967(1–3) 189–195.
[26] Caine S, Heraud P, Tobin MJ, McNaughton D, Bernard CC. The application of Fourier transform infrared microspectroscopy for the study of diseased central nervous system tissue. Neuroimage, 2012; 59(4) 3624-3640.
[27] Nakamura T, Kelly JG, Trevisan J, Cooper LJ, Bentley AJ, Carmichael PL, Scott AD, Cotte M, Susini J, Martin-Hirsch PL, Kinoshita S, Fullwood NJ, Martin FL. Microspectroscopy of spectral biomarkers associated with human corneal stem cells. Molecular Vision, 2010; 16 359-368.
[28] Ami D, Natalello A, Zullini A, Doglia SM. Fourier transform infrared microspectroscopy as a new tool for nematode studies. FEBS Letters, 2004; 576(3) 297-300.
[29] Casal HL, Mantsch HH. Polymorphic phase behaviour of phospholipid membranes studied by infrared spectroscopy. Biochimica et Biophysica Acta, 1984; 779(4) 381-401.
[30] Barth A. Infrared spectroscopy of proteins. Biochimica et Biophysica Acta, 2007; 1767 (9) 1073–1101.
[31] Banyay M, Sarkar M, Gräslund A. A library of IR bands of nucleic acids in solution. Biophysical Chemistry, 2003; 104(2) 477-488.
[32] Banyay M, Gräslund A. Structural effects of cytosine methylation on DNA sugar pucker studied by FTIR. Journal of Molecular Biology, 2002; 324(4) 667-676.
[33] Kačuráková M, Mathlouthi M. FTIR and laser-Raman spectra of oligosaccharides in water: characterization of the glycosidic bond. Carbohydrate Research, 1996; 284 145-157.
[34] Wong PT, Wong RK, Caputo TA, Godwin TA, Rigas B. Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis. Proceedings of the National Academy of Sciences USA, 1991; 88(24) 10988-10992.
[35] Ami D, Neri T, Natalello A, Mereghetti P, Doglia SM, Zanoni M, Zuccotti M, Garagna S, Redi CA. Embryonic stem cell differentiation studied by FT-IR spectroscopy. Biochimica et Biophysica Acta, 2008; 1783(1) 98-106.
[36] Konorov SO, Schulze HG, Caron NJ, Piret JM, Blades MW, Turner RFB. Raman microspectroscopic evidence that dry-fixing preserves the temporal pattern of non-specific differentiation in live human embryonic stem cells. Journal of Raman Spectroscopy, 2011; 42(4) 576–579.

[37] Zhao R, Quaroni L, Casson AG. Fourier transform infrared (FTIR) spectromicroscopic characterisation of stem-like cell populations in human esophageal normal and adenocarcinoma cell lines. The Analyst, 2010; 135(1) 53–61.
[38] Bassan P, Kohler A, Martens H, Lee J, Byrne HJ, Dumas P, Gazi E, Brown M, Clarke N, Gardner P. Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. The Analyst, 2010; 135(2) 268-277.
[39] Susi H, Byler DM. Resolution-enhanced Fourier transform infrared spectroscopy of enzymes. Methods in Enzymology, 1986;130 290-311.
[40] Levin IW, Bhargava R. Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. Annual Review of Physical Chemistry, 2005; 56 429-474.
[41] Dumas P, Miller L. The use of synchrotron infrared microspectroscopy in biological and biomedical investigations. Vibrational Spectroscopy, 2003; 32 3–21.
[42] Katon JE. Infrared Microspectroscopy. A Review of Fundamentals and Applications. Micron, 1996; 27(5) 303-314.
[43] Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikstrom C, Wold S. Multivariate and Megavariate Data Analysis Basic Principles and Applications. San Jose: Umetrics Academy; 2006.
[44] Manly BFJ. Multivariate Statistical Methods. London: Chapman & Hall/CRC press; 2004.
[45] Hestenes MR, Stiefel E. Methods of Conjugate Gradients for Solving Linear Systems. Journal of Research of the National Bureau of Standards 1952;49: 409-436.
[46] Marquardt D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics 1963; 11(2) 431-441.
[47] Levenberg K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. Quarterly of Applied Mathematics 1944; 2 164-168.
[48] Goldberg D E. Genetic Algorithms in Search Optimization and Machine Learning. Boston: Addison-Wesley Longman Publishing; 1969.
[49] Rencher A C. Methods of Multivariate Analysis. Hoboken: Wiley; 2002.
[50] Nas T, Isaksson T, Fearn T, Davies T. Multivariate Calibration and Classification. Chichester: NIR Publications; 2004.
[51] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936; 7 179-188.
[52] Fearn, T. (2002). Discriminant analysis, In: Handbook of Vibrational Spectroscopy, Chalmers, J.M. & Griffiths, P.R. (eds.), New York: Wiley, p2086–2093.
[53] Fukunaga K. Introduction to Statistical Pattern Recognition. San Diego: Academic Press Professional Inc.; 1990.
[54] Bishop CM. Neural Networks for Pattern Recognition. New York: Oxford University Press; 1995.
[55] Jonathan P, McCarthy WV, Roberts MIA. Discriminant Analysis With Singular Covariance Matrices. A Method Incorporating Cross-Validation And Efficient Randomized Permutation Tests. Journal Of Chemometrics, 1996; 10(4) 189-213.

[56] Rezzi S, Giani I, Héberger K, Axelson DE, Moretti VM, Reniero F and Claude G. Classification of Gilthead Sea Bream (Sparus aurata) from 1H NMR Lipid Profiling Combined with Principal Component and Linear Discriminant Analysis. Journal of agricultural and food chemistry 2007; 55 9963-9968.

[57] Skrobot LS, Castro VRE, Pereira RCC, Pasa VMD and Fortes ICP. Use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Gas Chromatographic (GC) Data in the Investigation of Gasoline Adulteration. Energy & Fuels 2007; 21 3394-3400.

[58] Walsh MJ, Singh MN, Pollock HM, Cooper LJ, German MJ, Stringfellow HF, Fullwood NJ, Paraskevaidis E, Martin-Hirsch PL, Martin FL. ATR microspectroscopy with multivariate analysis segregates grades of exfoliative cervical cytology. Biochemical and Biophysical Research Communications, 2007; 352(1) 213-219.

[59] Pereira RCC, Skrobot VL, Castro EVR, Fortes ICP, Pasa VMD. Determination of Gasoline Adulteration by Principal Components Analysis-Linear Discriminant Analysis Applied to FTIR Spectra. Energy & Fuels 2006; 20: 1097-1102.

[60] Héberger K, Csomós E andLivia S. Principal Component and Linear Discriminant Analyses of Free Amino Acids and Biogenic Amines in Hungarian Wines. Journal of Agricultural and Food Chemistry 2003; 51 8055-8060.

[61] Indahl UG, Liland KH, Naes T. Canonical partial least squares - a unified PLS approach to classification and regression problems. Journal of Chemometrics 2009; 23 495–504.

[62] Rieppo L, Rieppo J, Jurvelin JS, Saarakkala S: Fourier transform infrared spectroscopic imaging and multivariate regression for prediction of proteoglycan content of articular cartilage. PloS ONE 2012, 7 e32344.

[63] Hemmateenejad B, Akhond M and Samari F. A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2007; 67 958-965.

[64] Krogh A. What are artificial neural networks? Nature Biotechnology, 2008; 25(2) 195-197

[65] Haykin S. Neural Networks: A Comprehensive Foundation. Englewood Cliffs: Prentice Hall Inc.; 1999.

[66] Bentley AJ, Nakamura T, Hammiche A, Pollock HM, Martin FL, Kinoshita S, Fullwood NJ. Characterization of human corneal stem cells by synchrotron infrared micro-spectroscopy. Molecular Vision, 2007; 13 237-242.

[67] Yousef I, Bréard J, SidAhmed-Adrar N, Maâmer-Azzabi A, Marchal C, Dumas P, Le Naour F. Infrared spectral signatures of CDCP1-induced effects in colon carcinoma cells. The Analyst, 2011; 136(24) 5162-5168.

[68] German MJ, Hammiche A, Ragavan N, Tobin MJ, Cooper LJ, Matanhelia SS, Hindley AC, Nicholson CM, Fullwood NJ, Pollock HM, Martin FL. Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. Biophysical Journal, 2006; 90(10) 3783-3795.

[69] Walsh MJ, Singh MN, Stringfellow HF, Pollock HM, Hammiche A, Grude O, Fullwood NJ, Pitt MA, Martin-Hirsch PL, Martin FL. FTIR Microspectroscopy Coupled with Two-

Class Discrimination Segregates Markers Responsible for Inter- and Intra-Category Variance in Exfoliative Cervical Cytology. Biomarker Insights, 2008; 3 179–189.

[70] Kelly JG, Singh MN, Stringfellow HF, Walsh MJ, Nicholson JM, Bahrami F, Ashton KM, Pitt MA, Martin-Hirsch PL, Martin FL. Derivation of a subtype-specific biochemical signature of endometrial carcinoma using synchrotron-based Fourier-transform infrared microspectroscopy. Cancer Letters, 2009; 274(2) 208-217.

[71] Walsh MJ, Fellous TG, Hammiche A, Lin WR, Fullwood NJ, Grude O, Bahrami F, Nicholson JM, Cotte M, Susini J, Pollock HM, Brittan M, Martin-Hirsch PL, Alison MR, Martin FL. Fourier transform infrared microspectroscopy identifies symmetric PO(2)(-) modifications as a marker of the putative stem cell region of human intestinal crypts. Stem Cells, 2008; 26(1) 108-118.

[72] Taylor SE, Cheung KT, Patel II, Trevisan J, Stringfellow HF, Ashton KM, Wood NJ, Keating PJ, Martin-Hirsch PL, Martin FL. Infrared spectroscopy with multivariate analysis to interrogate endometrial tissue: a novel and objective diagnostic approach. British Journal of Cancer, 2011; 104(5) 790-797.

[73] Chonanant C, Jearanaikoon N, Leelayuwat C, Limpaiboon T, Tobin MJ, Jearanaikoon P, Heraud P. Characterisation of chondrogenic differentiation of human mesenchymal stem cells using synchrotron FTIR microspectroscopy. The Analyst, 2011; 136(12) 2542-2551.

[74] Thumanu K, Tanthanuch W, Danna Y, Anawat S, Chanchao L, Rangsun P, Philip H. Spectroscopic signature of mouse embryonic stem cell-derived hepatocytes using synchrotron Fourier transform infrared microspectroscopy. Journal of Biomedical Optics 2011; 16(5) 057005.

[75] Debey P, Szöllösi MS, Szöllösi D, Vautier D, Girousse A, Besombes D. Competent mouse oocytes isolated from antral follicles exhibit different chromatin organization and follow different maturation dynamics. Molecular Reproduction and Development, 1993; 36(1) 59-74.

[76] Marty R, N'soukpoé-Kossi CN, Charbonneau DM, Kreplak L, Tajmir-Riahi HA. Structural characterization of cationic lipid-tRNA complexes. Nucleic Acid Research, 2009; 37(15) 5197-5207.

[77] Liu J, Conboy JC. Structure of a gel phase lipid bilayer prepared by the Langmuir-Blodgett/Langmuir-Schaefer method characterized by sum-frequency vibrational spectroscopy. Langmuir, 2005; 21(20) 9091-9097.

[78] Gentile L, Monti M, Sebastiano V, Merico V, Nicolai R, Calvani M, Garagna S, Redi CA, Zuccotti M. Single-cell quantitative RT-PCR analysis of Cpt1b and Cpt2 gene expression in mouse antral oocytes and in preimplantation embryos. Cytogenetic and Genome Research, 2004; 105(2-4) 215-21.

[79] Zhizhina GP, Oleinik EF. Infrared spectroscopy of nucleic acids. Russian Chemical Reviews, 1972; 41(3) 258 -280.

[80] Ten GN, Baranov VI. Manifestation of intramolecular proton transfer in imidazole in the electronic-vibrational spectrum. Journal of Applied Spectroscopy, 2008; 75(2) 168–173.