**OPEN ACCESS DOCUMENT**

Information of the Journal in which the present paper is published:

* Elsevier, Trends in Analytical Chemistry, 2017, 94, 130-140
* doi: 10.1016/j.trac.2017.07.004

**RELEVANT ASPECTS OF UNMIXING/RESOLUTION ANALYSIS FOR THE INTERPRETATION OF BIOLOGICAL VIBRATIONAL HYPERSPECTRAL IMAGES.**

Víctor Olmos1, Laura Benítez1, Mónica Marro2, Pablo Loza-Alvarez2, Benjamí Piña3, Romà Tauler3, Anna de Juan1, (\*).

1. Department of Chemical Engineering and Analytical Chemistry, University of Barcelona. Diagonal 645, 08028 Barcelona, Spain
2. ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, Carl Friedrich Gauss 3, 08860 Castelldefels (Barcelona), Spain
3. Department of Environmental Chemistry, Institute of Environmental Assessment and Water Diagnostic (IDAEA-CSIC), Jordi Girona 18, 08034 Barcelona, Spain

**ABSTRACT**

Multivariate unmixing/resolution analysis of biological vibrational hyperspectral images is crucial to characterize the morphology and spectral signatures of the different biological tissues. This work provides a general data analysis protocol to interpret Raman and FT-IR hyperspectral images of biological samples. To do so, dedicated preprocessing tools and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) are the tools proposed. A main focus is on the use of MCR-ALS to analyze image multiset structures that may include all the necessary variability to define appropriately a biological population. Multiset MCR-ALS analysis may help to solve problems, such as the modelling of mixed non-biological signal contributions or the elucidation and active use of information related to the presence/absence of biological contributions in different images (samples). Additional aspects, such as the only use of the FT-IR fingerprint region vs. using extended spectral ranges in FT-IR to improve the differentiation among biological contributions, are also considered.

**Keywords:** Raman hyperspectral images (HSI), FT-IR HSI, biological images, image unmixing, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS).

(\*) Corresponding author: [anna.dejuan@ub.edu](mailto:anna.dejuan@ub.edu).

**1. INTRODUCTION**

The use of hyperspectral imaging coupled with multivariate resolution analysis techniques is a powerful approach to study biological and biomedical samples. Hyperspectral images combine the spatial information of optical microscopy and the biochemical information provided by a spectroscopic technique. In the last years, the use of vibrational techniques like Raman or Fourier transform infrared (FT-IR) in hyperspectral imaging has increased[1–5]. Vibrational spectroscopic techniques provide sample information at molecular level. In the case of biological samples, spectra usually contain contributions of many different molecules that define a higher level structure, e.g. a cell compartment or a tissue. Besides, these techniques do not need the addition of external staining agents for the measurement. Raman and FT-IR imaging can help in biomedical and biological applications as non-invasive methods for disease diagnosis[2,6–8], improvement of image-guided surgery[9–14], study of cells and tissue dynamics and metabolites[15–19], etc.

Multivariate resolution image analysis allows the interpretation of the often small spectral variations within a biological image and links this variation to the characterization of biological components, e.g., tissues or cell compartments, depending on the spatial resolution of the technique used. The main goal of image resolution/unmixing methods is to recover the pure spectra and distribution maps of image constituents over the sample. In a biological context, an image constituent is hardly ever a single chemical compound. Instead, the definition of image constituent/contribution must be associated with a consistent mixture of biological molecules that define totally, or in combination with other contributions, a particular tissue or biological element. Therefore, the one-to-one association between tissue and image constituent is not general either and heavily depends on the within-tissue compositional variability.

The aim of this work is proposing a general data analysis protocol to handle Raman and FT-IR hyperspectral images of biological samples, taking into account preprocessing issues and particular aspects of image resolution, such as the use of information on the biological samples and the way to incorporate this information when analyzing biological populations in a multiset mode.

As a case study, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[20–22] has been applied to analyze a set of zebrafish (*Danio rerio*) tissue images by Raman and FT-IR imaging. The choice of zebrafish tissues is relevant because this organism is often taken as a model in human development studies[23–26]. Zebrafish genome is well described and presents numerous connections to human genome and, as for other vertebrates, analogous diseases in both organisms can be found. The easiness to breed zebrafish, its fast development and small size coupled with its high permeability to small molecules found in the environment, makes it very suitable in environmental –omics studies[27–31].

Zebrafish have been cryosectioned in order to obtain a suitable thickness for Raman and transmission FT-IR image acquisition. Transversal cryosections of zebrafish eyes on its larva growth stage at 120 hours post-fertilization have been obtained as an example for this study. Zebrafish eye is formed by many tissue layers of different composition, the most important being the crystalline lens, which contains a high amount of proteins, the retina and the cornea[32,33]. In addition to the histopathological importance of the eye area, the presence of many pigments and photoreceptors, which can be analyzed by spectroscopic techniques, makes the area also interesting in environmental studies related to contaminant stress.

Raman and FT-IR hyperspectral images of zebrafish cryosections have been preprocessed according to the nature and artifacts present in the different spectroscopic techniques. Hyperspectral images obtained with each spectroscopic technique have been analyzed separately. First, images have been analyzed individually by MCR-ALS and a first description of each cryosection has been obtained. The results of single image analysis help to obtain preliminary information to be incorporated in the simultaneous analysis of multiple images coming from the same spectroscopic technique (multiset analysis). Multiset analysis provides more reliable spectral signatures and distribution maps and helps in a better characterization of the samples analyzed together, particularly if all of them belong to the same population or were in the same biological conditions, e.g., healthy vs. diseased samples.. Pure spectral signatures and morphology and location of the biological contributions in the distribution maps can be used and compared afterwards with information in the literature in order to find out the biological meaning of the profiles obtained.

**2. EXPERIMENTAL**

2.1 Zebrafish breeding

Zebrafish (Danio rerio) were obtained by natural breeding under controlled conditions. Zebrafish embryos were raised at 28.5 ºC [25,34] with a 12L:12D photoperiod in embryo water (90 mg/mL of Instant Ocean (Aquarium Systems, Sarrebourg, France); 0.58 mM CaSO4.2H2O, dissolved in reverse osmosis purified water) till zebrafish became mature. During this period of time, no feeding was carried out in order to avoid modifying the embryo natural conditions. All individuals of the zebrafish population were genetically identical.

2.2 Sample Treatment

The 5-days zebrafish embryos were frozen using the flash-freezing method with the aim of preserving the sample in suitable conditions to perform cryosections to be used in Raman and FT-IR images. Thus, the zebrafish embryos were introduced in cryomolds and embedded in Tissue-Tek optimal cutting temperature (OCT) compound (Sakura Finetek, Flemingweg, The Netherlands) to proceed with the flash-freezing by immersion in an acetone - dry ice bath at -78 ºC.

20 µm thick zebrafish cryosections were obtained by a Leica CM30505 cryostat microtome located at Centre Científics i Tecnològics de la Universitat de Barcelona (CCiTUB) at -23 ºC. This thickness allows working with Raman imaging and with FT-IR in transmission mode. Cryosections were laid on CaF2 optical windows, a support suitable for Raman and FT-IR image acquisition.

2.3 Image acquisition

Raman hyperspectral images were acquired in Institut de Ciències Fotòniques (ICFO) in Parc Tecnològic de Castelldefels by an inVia Raman Microscope spectrometer manufactured by Renishaw. A 532 nm laser was used as light source and Raman spectra were recorded with a 10 s acquisition time and two accumulations using point scanning mode with a 50x objective (numerical aperture, NA: 0.75). The spectral range studied was the fingerprint region (850 cm-1 to 1700 cm-1), the spectral resolution was 3 cm-1 and the pixel size 5x5 µm2.

FT-IR images in transmission mode were recorded with a Nicolet iN10 MX microscope spectrometer manufactured by Thermo Scientific (located in CCiTUB). The images were acquired in the spectral range going from 916 cm-1 to 3715 cm-1 with 4 cm-1 spectral resolution and 25x25 µm2 pixel size. The acquisition mode was line scanning with sixteen scans and 5 s acquisition time for each line.

**3. DATA TREATMENT**

The data treatment consists of two steps described in the following sections:

* Preprocessing of the Raman and FT-IR images
* Application of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) to decompose cryosection images into spectral signatures and distribution maps of pure biological contributions.

Data preprocessing as well as MCR-ALS analysis have been performed using home-made routines and GUIs developed in MATLAB environment (MathWorks, Natick, MA, USA).

3.1 Data preprocessing

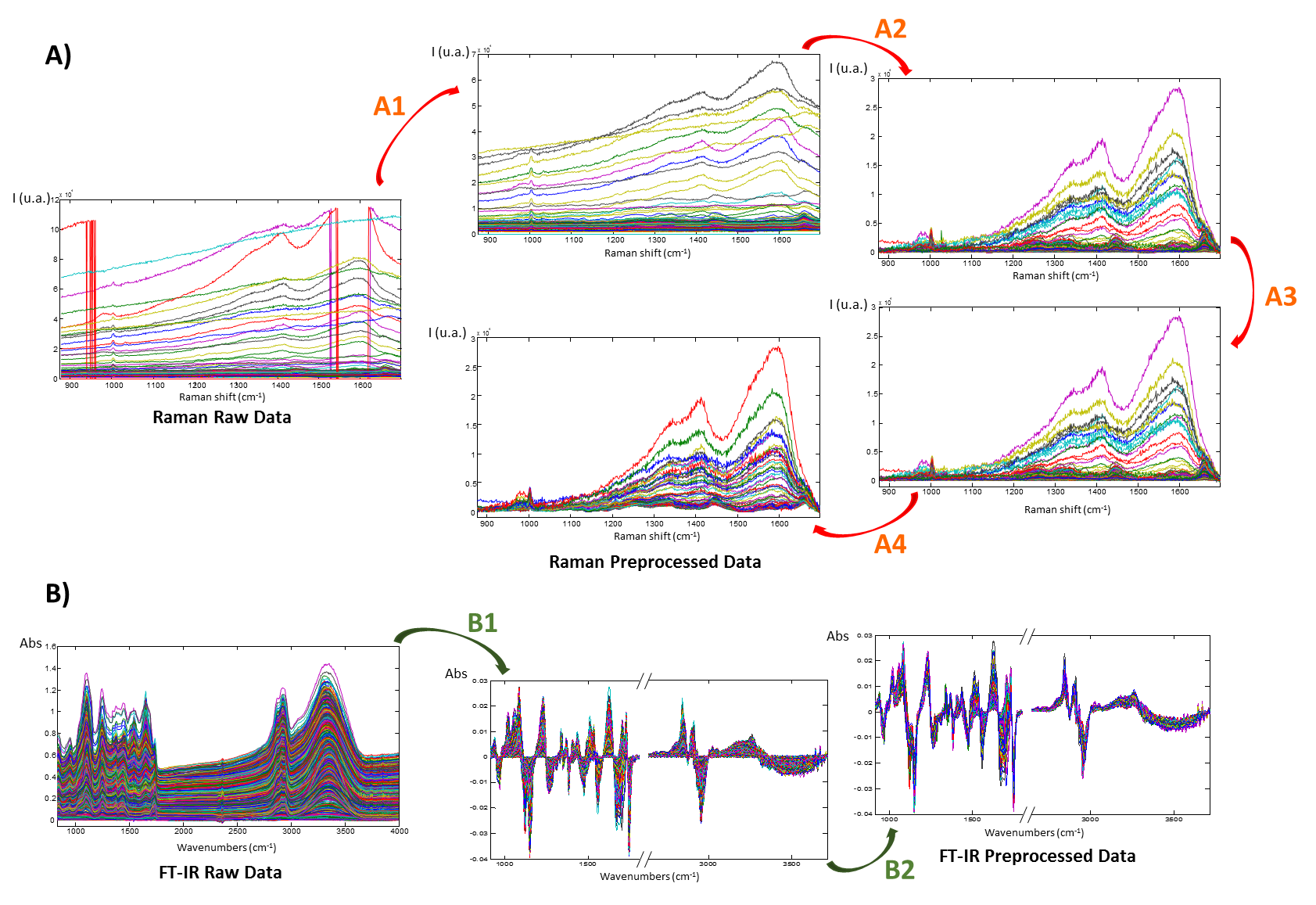
Hyperspectral images are structured in a data cube with two dimensions related to pixel coordinates on the sample surface (x and y) and the third linked to the spectral information (λ). In order to perform the sample data treatment, the image data cube has been unfolded into a data matrix (**D**), sized (total nr. pixels x nr. of spectral channels), which contains all pixel spectra one under the other.

Both Raman and FT-IR images have been preprocessed due to the presence of signal distortions created by changes in sample thickness, scattering, fluorescence and instrumental factors.

Raman image preprocessing includes the subsequent steps (see Figure 1A):

1. Elimination of anomalous pixel spectra because of too high or too low signal intensity (saturated and dead pixels). To avoid voids in the image, the spectra of these pixels were replaced by interpolation based on the immediate normal neighbouring pixel spectra.
2. Baseline correction by Asymmetric Least Squares (AsLS) [35]. This method is based on a recursive local fitting of the whole spectrum using a baseline obtained with a Whittaker smoother. The baseline fitting is controlled by two parameters, one associated with the smoothness of the fit (λ) and the other with the penalty imposed on the spectral readings providing residuals above the fitted baseline, i.e., the spectral readings associated with a spectral chemical feature (*p*).
3. Detection and elimination of cosmic peaks. Cosmic peaks in a spectrum were replaced by interpolation based on normal Raman intensity readings from small ranges of neighbouring spectral channels within the same spectrum.
4. Removal of pixel spectra related to the sample support. These pixels have null or very small signal. They are removed applying a threshold based on the total intensity of spectra. Pixels below that threshold are not considered for further analysis.

FT-IR images do not need as many preprocessing steps as Raman images (see Figure 1B). First of all, a characteristic band of CO2 variation during measurements (among 2000-2500 cm-1) has been suppressed by interpolation of the readings of the neighbouring spectral channels. This has been done because no information was present in this spectral range. After that, a preprocessing workflow similar to the used for Raman images could also have been performed due to the versatility of the AsLS baseline correction method. However, in the case of this set of FT-IR images, the baseline shape is fairly linear and does not require the use of AsLS. Instead, a first derivative by Savitzky-Golay method has been carried out (with a second polynomial grade and a window size of nine spectral channels, suitable parameters for the spectra of the data set analyzed)[36]. FT-IR images present little differences among spectra and the use of the first derivative preprocessing method corrects the baseline of the spectra and, at the same time, enhances these small spectral differences. Finally, as in Raman image preprocessing, the pixels of the sample support (with null signal) have been removed for further analysis.



**Figure 1:** (A) Raman image preprocessing: A1) elimination of anomalous pixel spectra (saturated or dead pixels); A2) baseline correction by AsLS; A3) Detection and elimination of cosmic peaks; A4) elimination of support related pixels; and (B) FT-IR image preprocessing: B1) first derivative of spectra using Savitzky-Golay method; B2) elimination of sample support related pixels.

3.2 Image resolution by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

The objective of image resolution is the decomposition of the images into the distribution maps (**C**) and the pure spectra (**ST**) of the constituents present in the sample. MCR-ALS, as any other resolution method, is based on decomposing the original mixed raw data into the chemically meaningful bilinear model formed by the contribution of the different pure components shown in Eq. (1), where **D** is the raw spectra matrix, **C** contains the concentration profiles of the pure components, **ST** the related pure spectra and **E** the experimental error associated with the raw measurement. For Raman and FT-IR images, the MCR model matches the bilinear model of the spectroscopic measurements defined by Equation 1.

Eq. 1

MCR-ALS performs the decomposition of the raw data into the bilinear model of pure contributions by an iterative algorithm. Some constraints can be applied to the profiles during the iterative optimization in order to maintain its chemical meaning. These constraints allow applying MCR-ALS to a high variety of data sets. In this study, MCR-ALS has been used to analyze hyperspectral images individually or in a multiset mode [20,21,37–39].. The essential steps to perform MCR-ALS are the following:

1. Determination of the number of chemical contributions in the raw data (**D**)
2. Generation of initial estimates of the **ST** matrix using a method to select the purest spectra based on SIMPLISMA[40].
3. Calculation of **C** and **ST** iteratively by alternating least squares under constraints until convergence is achieved.

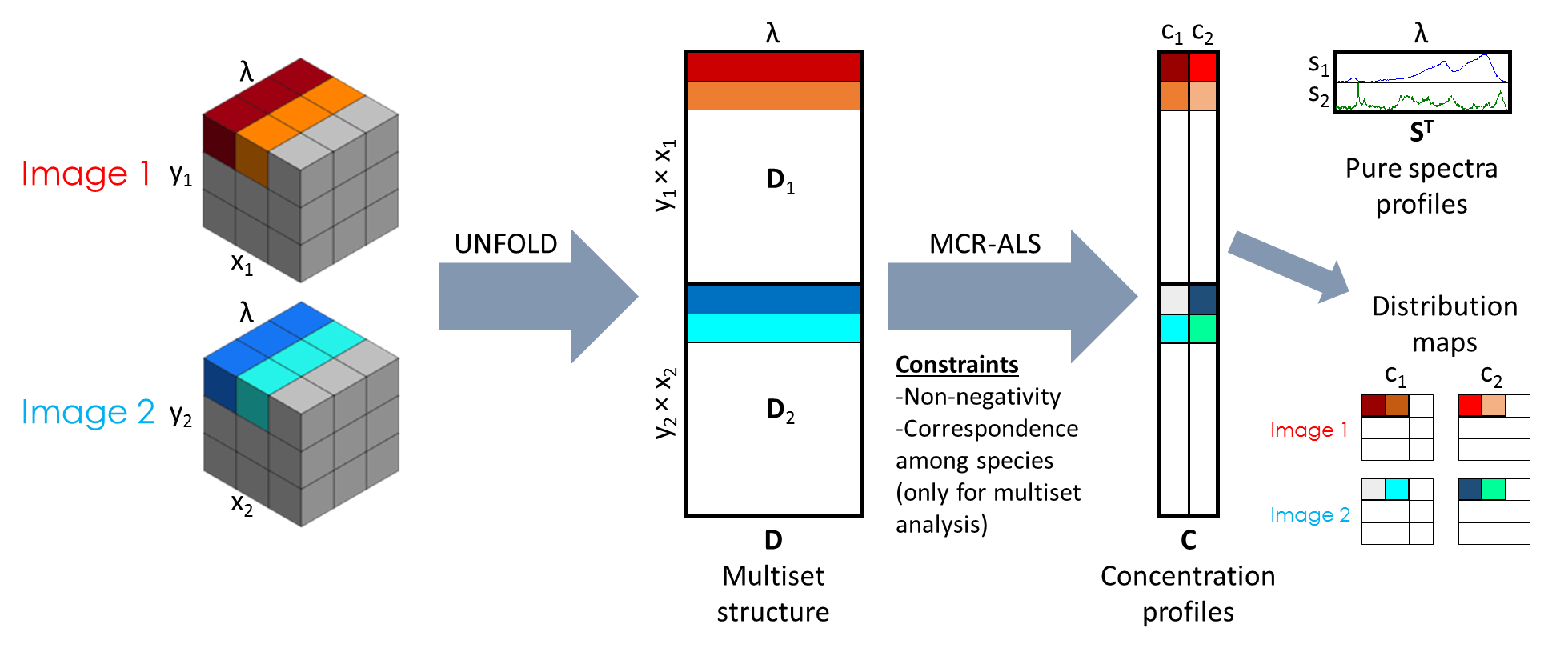
First of all, the number of chemical contribution has been estimated using singular value decomposition (SVD)[41]. Then, the initial estimates of the **ST** matrix have been generated by a SIMPLISMA-based approach[40]. Finally, the least squares algorithm is applied and involves the operations and alternatingly in each iterative cycle. The convergence criterion is achieved when the original data is well reproduced by the bilinear model of the concentration profiles and pure spectra and there is no significant variation in the fit of the original data among consecutive iterative cycles. The parameters used to assess the quality of the model are the percentage of lack of fit (see Eq. 2) and the variance explained, r2 (see Eq. 3).

Eq. 2

Eq. 3

where dij is the element of the original data matrix in row *i* and column *j* and eij is the residual obtained from the difference between the element dij of the original data and the analogous element obtained from the MCR-ALS model.

Very often, resolution of a single image can be difficult because of the similarities within **C** and/or **ST** profiles within a single sample. This problem can be solved by the addition of new information to the system. To do so, multiset analysis can be performed. In a multiset, augmented data matrices **D** are built containing different submatrices **Di**, which are related to different individual images sharing contributions in common. In this work, multiset structures are column-wise augmented matrices that contain spectra of several images acquired with the same spectroscopic technique one under the other. The bilinear decomposition (see Eq. 1) provides a single matrix **ST** that is common for all the images analyzed and a **C** matrix that has as many submatrices **Ci** as images in the data set. Every column of each **C** submatrix can be refolded to obtain the 2D distribution map of each constituent in the different images of the multiset (see Figure 2).



**Figure 2.** Data analysis workflow linked to multiset analysis of hyperspectral images by MCR-ALS.

As explained before, some constraints can be applied to the profiles during MCR-ALS optimization. These constraints help to provide chemically meaningful profiles and to decrease the ambiguity in the final results obtained. Non-negativity constraint is the most used to analyze hyperspectral images. In this case, non-negativity in the concentration direction has been applied to Raman and FT-IR images, whereas **ST** profiles have only been constrained to be positive in Raman images, since derivative spectra in FT-IR images can naturally contain negative values[38,39]. Normalization of spectra profiles in **ST** is also used to avoid scaling fluctuations in the profiles during the optimization.

Multiset analysis with MCR-ALS allow for the use of an additional constraint, called correspondence among species, which permits setting the presence or absence of any constituent in the different samples (images). This information is introduced through a binary matrix formed by as many columns as number of constituents in the data set and by as many rows as the number of images in the data set. When a certain constituent is absent in an image, the related concentration profile is constrained to be null. In biological samples, this information is usually not known in advance and single image MCR-ALS analysis is used to identify the presence or absence of constituents in the different samples. When the correspondence among species matrix is well defined, an improvement of the definition of spectral signatures and distribution maps is seen[21,42].

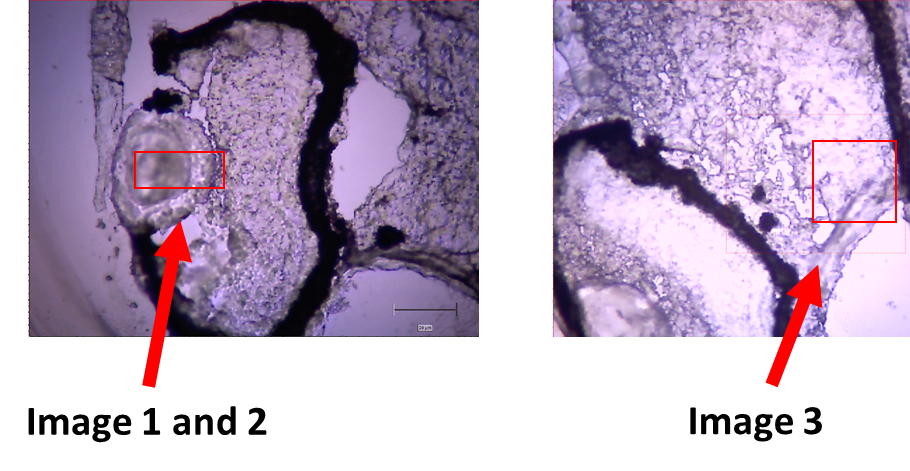
**4. RESULTS AND DISCUSSION**

The presented methodology has been applied to FT-IR and Raman hyperspectral images of cryosections of zebrafish eye tissue. This example is used to provide a data analysis workflow to study biological sample images by Multivariate Curve Resolution- Alternating Least Squares (MCR-ALS). As shown in section 3.2, this chemometric method provides the pure spectral signatures (**ST**) and the distribution maps of the image constituents (**C**). In the case of biological samples, the contributions resolved with MCR-ALS are not usually formed by only one chemical compound. Instead, each MCR contribution consists of a homogeneous mixture of chemical compounds that represents by itself or in combination with others, a certain biological structure, such as a specific kind of tissue.

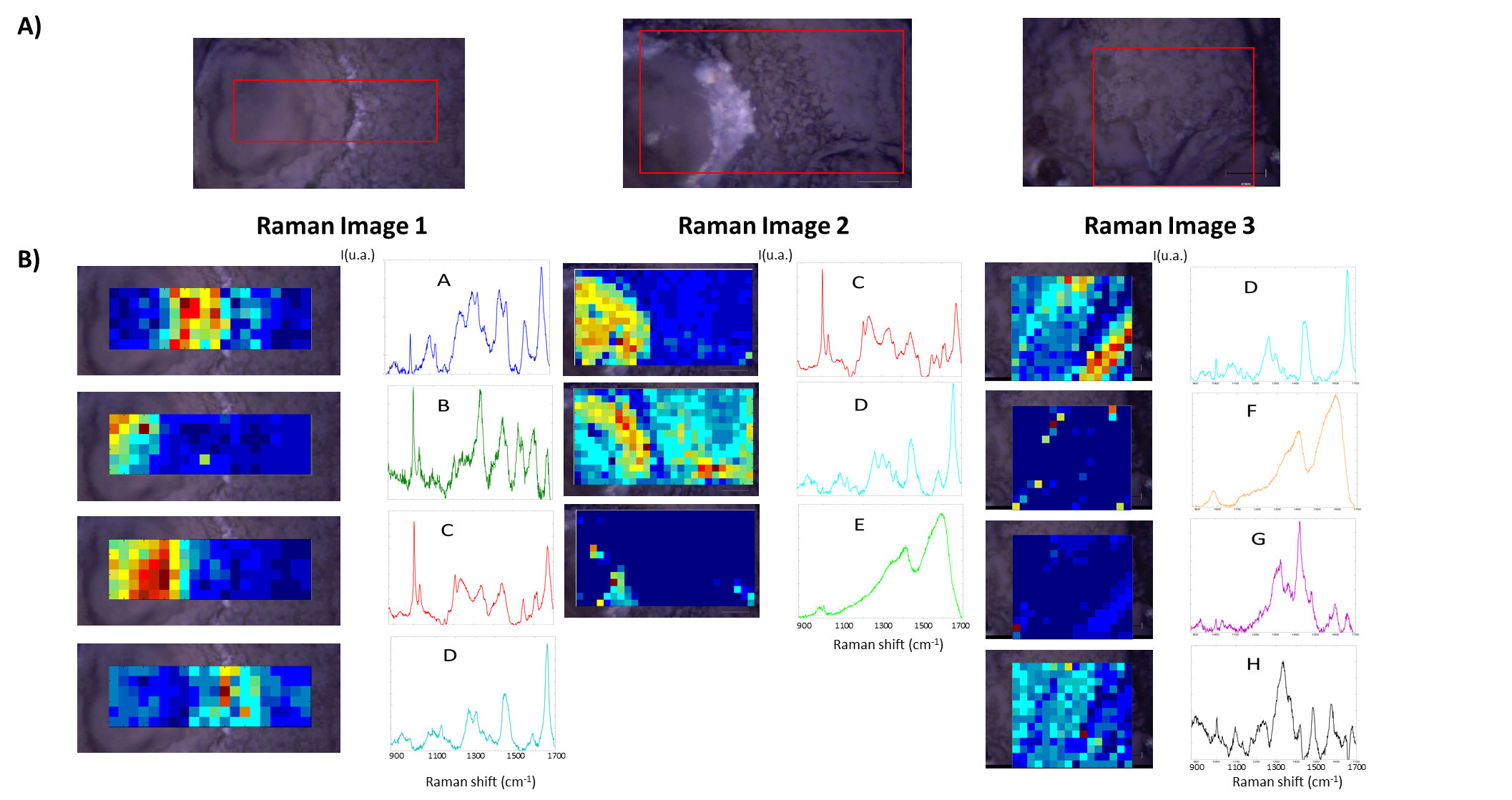
In the next sections, the application of MCR-ALS and the results obtained on Raman and FT-IR biological images are exposed and described in detail.

4.1 Raman hyperspectral image analysis

Three hyperspectral images have been acquired using Raman spectroscopy on different cryosections (see pictures in figure 3). Raman images one and two relate to areas around the zebrafish eye, whereas image three includes an area enclosing the eye and part of the mouth of the sampled zebrafish. Red rectangles in Figure 3 shows the approximate physiological area imaged with respect to the full cryosection for a better understanding of the fish anatomy. The exact grey images of the sample areas analyzed are shown in Figure 4A.



**Figure 3.** Photographs of zebrafish cryosections showing approximately the Raman imaged areas in the context of the full cryosection. Images one and two contain a transversal section of a zebrafish eye, whereas image three includes a part of the eyes and the mouth of the zebrafish.



**Figure 4:** (A) Grey images related to Raman hyperspectral images (HSI). Red squared areas correspond to the areas imaged. (B) Right plots: Pure spectral signatures of the three Raman hyperspectral images resolved individually by MCR-ALS. Common signatures are represented in the same color and are assigned the same capital letter. Left plots: resolved distribution maps of biological contributions. Color scale in distribution maps ranges from blue (for low pixel concentrations) to red (for high pixel concentration).

The three Raman images were preprocessed as explained in section 3.1. Baseline correction is a critical step in the analysis of Raman images, even more evident in biological samples, because of the frequent presence of an intense and irregular fluorescence contribution in the spectra. AsLS[35] method has been used to correct the baseline of Raman spectra. This method is very suitable to fit irregularly shaped baselines and can cope with within-image changes of baseline shape and intensity. The only limitation is that the spectral features and the baseline contribution should have very different frequencies, e.g., a broad fluorescent contribution vs. very narrow and sharp Raman spectral features.

Raman images have been first analyzed individually by MCR-ALS to have a preliminary understanding of the composition of the samples (see section 3.2). Non-negativity constraint has been applied to pure spectra and distribution maps in all MCR analyses. Table 1 shows the number of components resolved, the lack of fit and the variance explained in the individual image analysis. In order to observe the differences of composition, the pure spectral signatures and the distribution maps obtained from the three images have been compared (see Figure 4B).

**Table 1:** Information and quality parameters of the MCR-ALS individual resolutions performed with Raman hyperspectral images.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Number of Components | Lack of fit (%) | R2 (%) |
| Image 1 | 4 | 10.08 | 98.98 |
| Image 2 | 3 | 6.75 | 99.54 |
| Image 3 | 4 | 10.34 | 98.93 |
| Multiset | 7 | 8.12 | 99.34 |

To check the quality of an MCR analysis, the lack of fit, the pure spectra and the distribution maps should be examined. For example, the individual resolution of Raman image one has a lack of fit of 10.8%, which corresponds to a variance explained of 98.98 %, satisfactory for Raman image data set. The morphology of the maps, highlighting different concentric layers from left to right on the image matches the morphology of tissues that can be visually observed in the related grey image. It can also be seen that distinct spectral signatures are observed, depending on the part of the related tissue, which are also in agreement with spectral features present in the raw image spectra. Besides, additional knowledge, such as an important presence of a band at 1000 cm-1, related to proteins, in the contributions closer to the center of the crystalline lens (C contribution) can be observed.

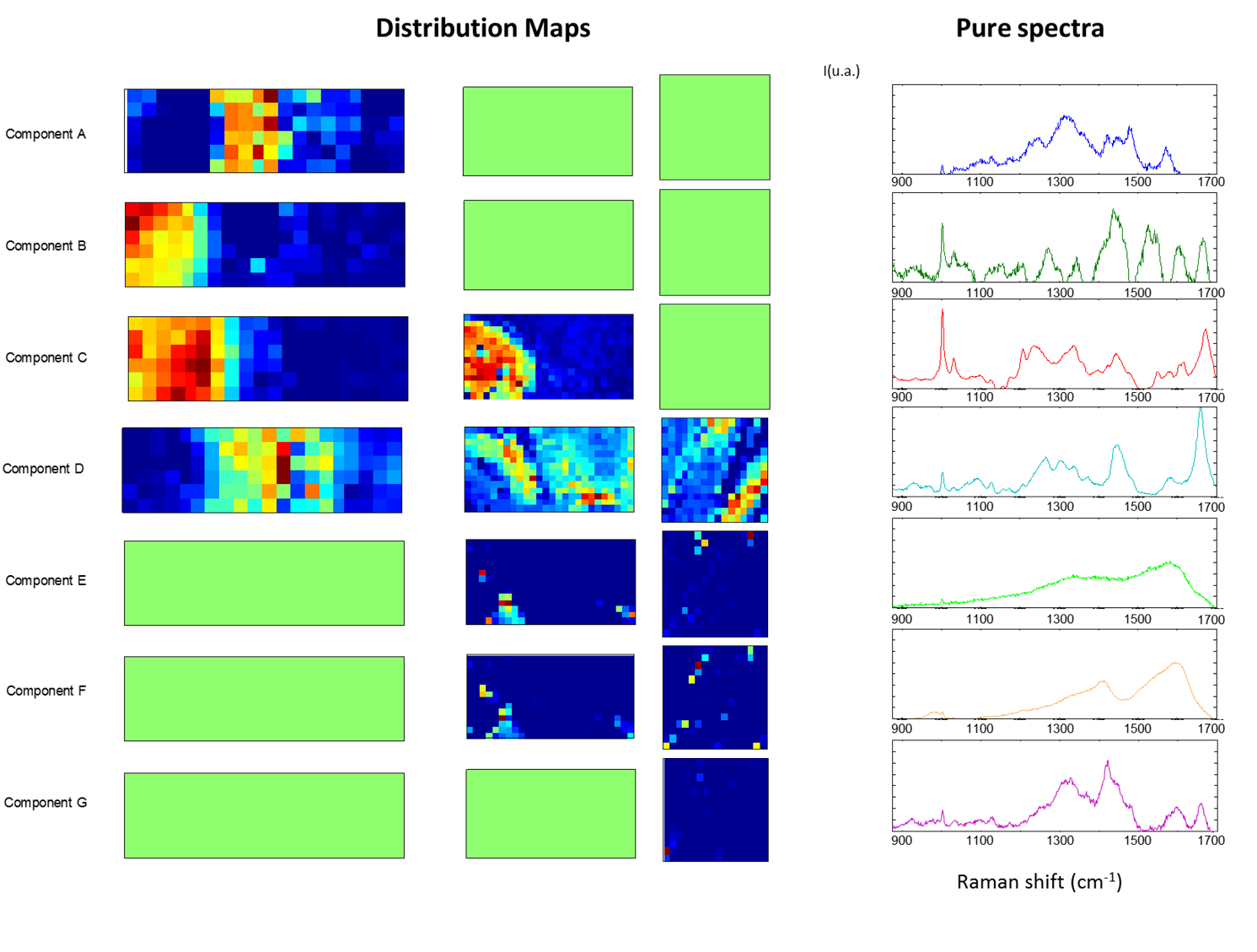
As it is shown in Figure 4B, the number and identity of image contributions resolved for each imaged cryosection is not the same. This may happen because there may be variability in the tissues scanned in each image and because of the variability among zebrafish individuals (biological variability). However, by visually comparing spectral signatures and morphology of the related maps, some common spectral signatures and contributions can be identified among images. To facilitate this interpretation, contributions are labeled with capital letters (A-H). The knowledge of the common and uncommon contributions among images is an information of great value when a multiset analysis is performed, as detailed below.

Multiset image analysis consists of the simultaneous analysis of several images. The three Raman images analyzed separately have also been analyzed together in a column-wise augmented multiset structure. Initial estimates of the **ST** matrix have been selected from the previous individual resolutions performed. MCR-ALS multiset analysis presents the possibility to apply the correspondence among species constraint. This constraint permits to indicate the presence or absence of the different MCR contributions in each one of the images of the multiset structure[21] (see section 3.2). The previous information obtained from the single image MCR analysis related to the common and uncommon contributions among images has been used to build the initial correspondence among species matrix (see Eq. 3), with as many columns as total number of MCR contributions in the three images analyzed simultaneously and as many rows as images in the multiset :

Eq. 4

It is important to say that the information on the correspondence among species in a multiset can be changed in view of the multiset analysis results. The first setting of this constraint, obtained from a visual comparison of resolved contributions in the individual image analysis is a good starting point to build the first correspondence among species matrix. After that, the multiset results obtained are checked and, in view of those, the correspondence among species can be modified. The use of the correspondence among species constraint allows providing the algorithm a better description of the data and converge in more reliable results, with less associated ambiguity[22,43].

The final multiset analysis of the three Raman images has been performed with seven MCR contributions, four of them common in two or more images and the other three present in only one image. In this case, the initial proposed resolution (with eight compounds) was decreased to a simpler model with seven contributions, which described with a very satisfactory fit the multiset and did not modify significantly the original matrix of correspondence among species. The results obtained from the multiset analysis of Raman images are shown in Figure 5. The lack of fit of the multiset resolution was 8.12 % (see Table 1), which is satisfactory for Raman image analysis. Distribution maps and spectral signatures of the multiset model are better defined than in the individual resolution due to the complementary information among the images added to the analysis and to the setting of presence/absence of contributions, which significantly helps to decrease the uncertainty linked to ambiguity in the final results. Interpretation of the resolved spectra is shown below.

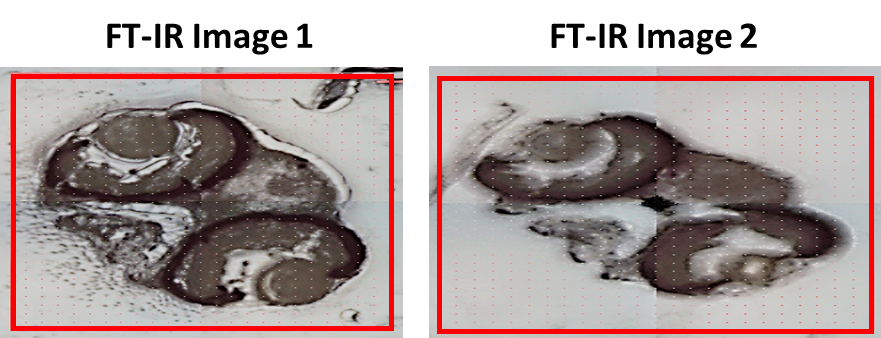


**Figure 5:** Multiset analysis of the three Raman hyperspectral images by MCR-ALS. Pure spectra and distribution maps of the 7 components resolved are shown. Distribution maps in light green refer to absent components in the image.

The identification of the resolved spectra is the last step in the resolution analysis. As explained before, spectral signatures of biological samples are not related to a single chemical compound and that is why the identification of the components is one of the most challenging goals after image resolution analysis of biological samples. Libraries of biological Raman spectra are usually very limited and present only pure compounds, which does not help when the resolved contributions are formed by compound mixtures. Raman bands interpretation can also help to identify the composition of the pure signatures resolved, as it is done in classical Raman spectroscopy[44], but previous knowledge is needed in order to suggest the identification of a tissue. Some of the compounds have been identified looking at Raman biological fingerprint spectra. Siebinga et al.[45] studied protein changes in human crystalline lens by Raman spectroscopy, and the component C resolved can be identified as the crystalline lens by the comparison of the fingerprint spectra and the more important bands. The bands corresponding to phenylalanine, amide III and amide I (1000, 1200 and 1600 cm-1 respectively) are shown in the crystalline lens spectra, which is a characteristic of a protein rich tissue. Components E and F have also been identified probably as two types of melanin pigments[46,47] located in the external pigmented part of the zebrafish eye. Given its relative position with the (putative) crystalline signal, component A may be related with the retina cells. Component B corresponds with a darker substructure in the lens tissue, probably reflecting a highly compacted zone, whereas component D co-localized with a highly birefringent material, of nature presently unknown.

4.2 FT-IR hyperspectral image analysis

Two hyperspectral images have been acquired using FT-IR spectroscopy. The grey images of the samples analyzed are shown in Figure 6. Both images have been obtained from transversal cryosections of zebrafish where the eyes and the mouth can be well located. FT-IR images have been analyzed in two ways, using the biological fingerprint (916-1800 cm-1) and also the whole spectral range acquired (916-3715 cm-1) to check if expanding the spectral range can improve the MCR‑ALS results obtained.



**Figure 6:** Grey images related to FT-IR hyperspectral images. Samples present transversal cryosections of the zebrafish head, including eyes, brain and mouth. Image area measured is squared in red.

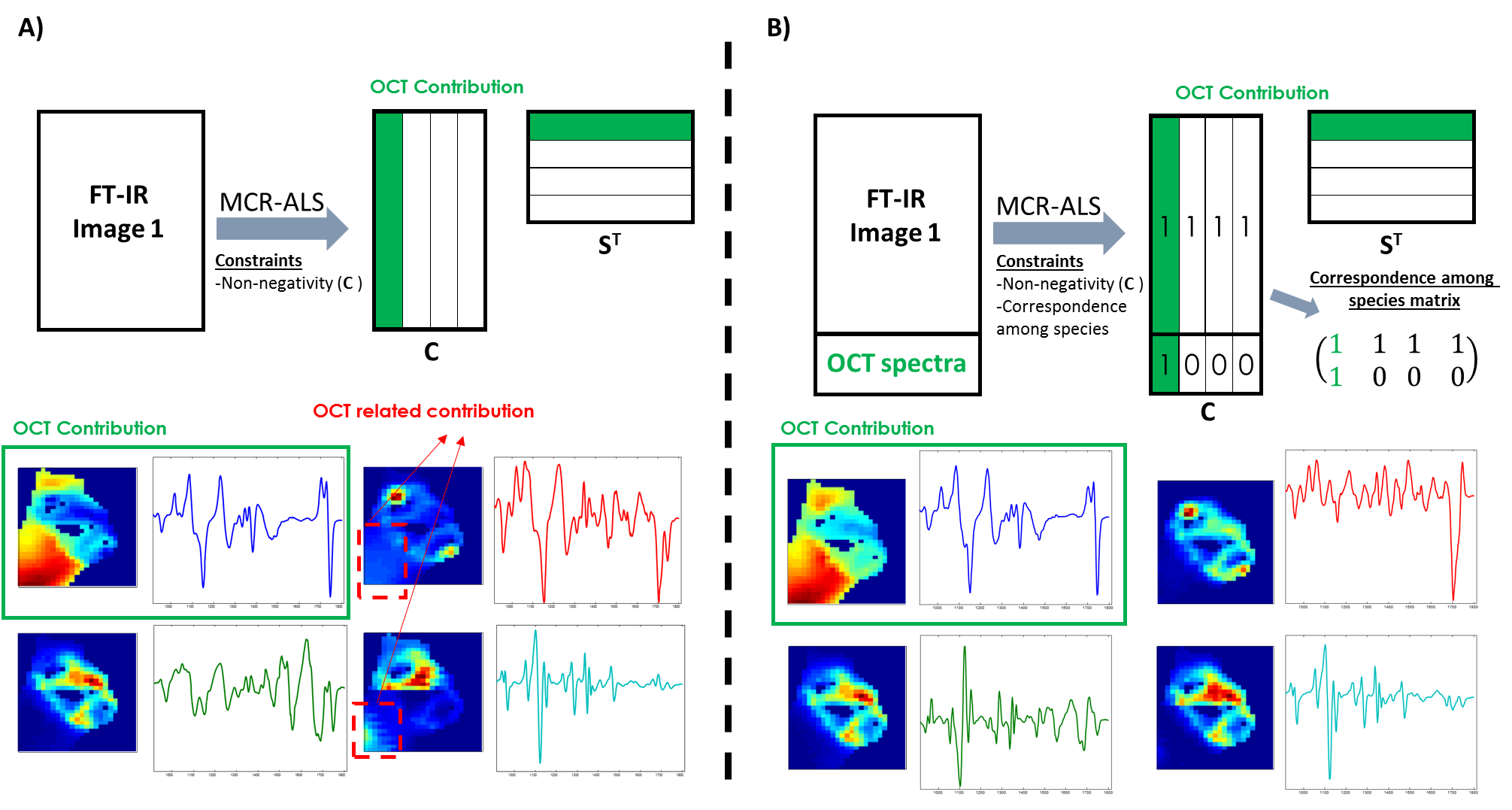
Image preprocessing has been performed as explained in section 3.1. A first derivative preprocessing using Savitzky-Golay method has been applied[36]. This method allows enhancing the small spectral differences among spectra shown in the FT-IR images and corrects satisfactorily the baseline of the spectra, since it is fairly linear in this data set. In other FT-IR spectra, where more irregular baselines due to more intense Mie scattering have been found, AsLS was successfully applied. The derivation of the data implies that some steps during the MCR-ALS analysis procedure have to be adapted to this kind of preprocessing. The first step modified is the generation of initial estimates by the SIMPLISMA[40]-based algorithm. This algorithm cannot perform adequately when negative data are present (as in derivative spectra). In this case, the SIMPLISMA-based algorithm has been applied on the squared raw dataset to detect the rows corresponding to the purest spectra. The spectra related to these rows in the derivative dataset have been afterwards used as initial estimates.

FT-IR images have also been analyzed by MCR-ALS, as explained in section 3.2. In this case, non-negativity constraint has been applied only to the distribution maps, since derivative spectra present naturally negative values. In this kind of images, we would like to address two issues: a) how to cope with the presence of non-biological contributions with a significant FT-IR signal, e.g., those related to reagents used to embed the samples for the cryosectioning process, and b) the effect of using a spectral region (from 916 to 1800 and 2650 to 3715) wider than the usual FT-IR fingerprint (from 916 to 1800) to enhance the capability to distinguish biological contributions.

1. *Modelling of non-biological contributions (embedding medium).*

Cryosections often come from samples embedded in reagents used to freeze and preserve adequately the samples before cryosection. Although these reagents do not penetrate in the biological sample tissues, some embedding material can remain in the surface of the cryosection analyzed. On the one hand, intensive rinsing can damage the sample and may not eliminate completely this material; on the other hand, when no rinsing is applied, remaining inert embedding reagents, such as OCT, may give FT-IR signal.

Therefore, it is important to propose chemometric strategies to distinguish the contribution to the signal of the embedding medium from the biological contributions. Figure 7A show the result of the individual resolution of FT-IR image one, where one of the resolved contributions (mainly in the outer part of the image) can be clearly associated with the presence of OCT compound, used as an embedding medium for cryosectioning (see section 2.2). In the distribution maps of the biological contributions, the mixed contribution of OCT can still be seen, which distorts both the distribution maps and, as a consequence, the spectral signatures recovered for contributions related to the zebrafish sample.



**Figure 7:** Strategy used for the modeling of embedding OCT compound in FT-IR image one. (A) MCR-ALS resolution of FT-IR image one. (B) MCR-ALS resolution of the same image by appending and using actively specific OCT information in the data analysis.

In order to improve the differentiation of this OCT contribution from the relevant biological contributions, spectra where only the embedding material is present need to be added to the data set analyzed. The procedure consists of adding a second matrix that contains several spectra of the embedding medium to create a multiset structure (see Figure 7B). The MCR-ALS resolution of this multiset is performed using the correspondence among species constraint to indicate that the 2nd matrix added has only one contribution, related to the OCT compound.

Figure 7B shows the effect of applying this procedure to remove the non-biological contribution of the OCT embedding material. As it can be seen, the distribution maps of the biological contributions related to zebrafish are better defined. In Table 2 the number of components, the lack of fit and the variance explained in the individual analysis of the images is shown. The distribution maps related to biological contributions neither show the shades in zones where only OCT was present nor the spectral features clearly associated with the embedding compound appear on the related spectral signatures resolved.

**Table 2:** Information and quality parameters of the MCR-ALS individual resolutions performed with FT-IR hyperspectral images.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Fingerprint and Complete Spectra | Number of Components | Lack of fit (%) | R2 (%) |
| Image 1 | Fingerprint | 4 | 8.99 | 99.19 |
| Image 2 | Fingerprint | 4 | 10.55 | 98.89 |
| Image 1 | Complete spectra | 4 | 8.79 | 99.23 |
| Image 2 | Complete spectra | 4 | 16.24 | 97.36 |
| Multiset | Fingerprint | 4 | 12.87 | 98.34 |
| Multiset | Complete spectra | 4 | 11.60 | 98.66 |

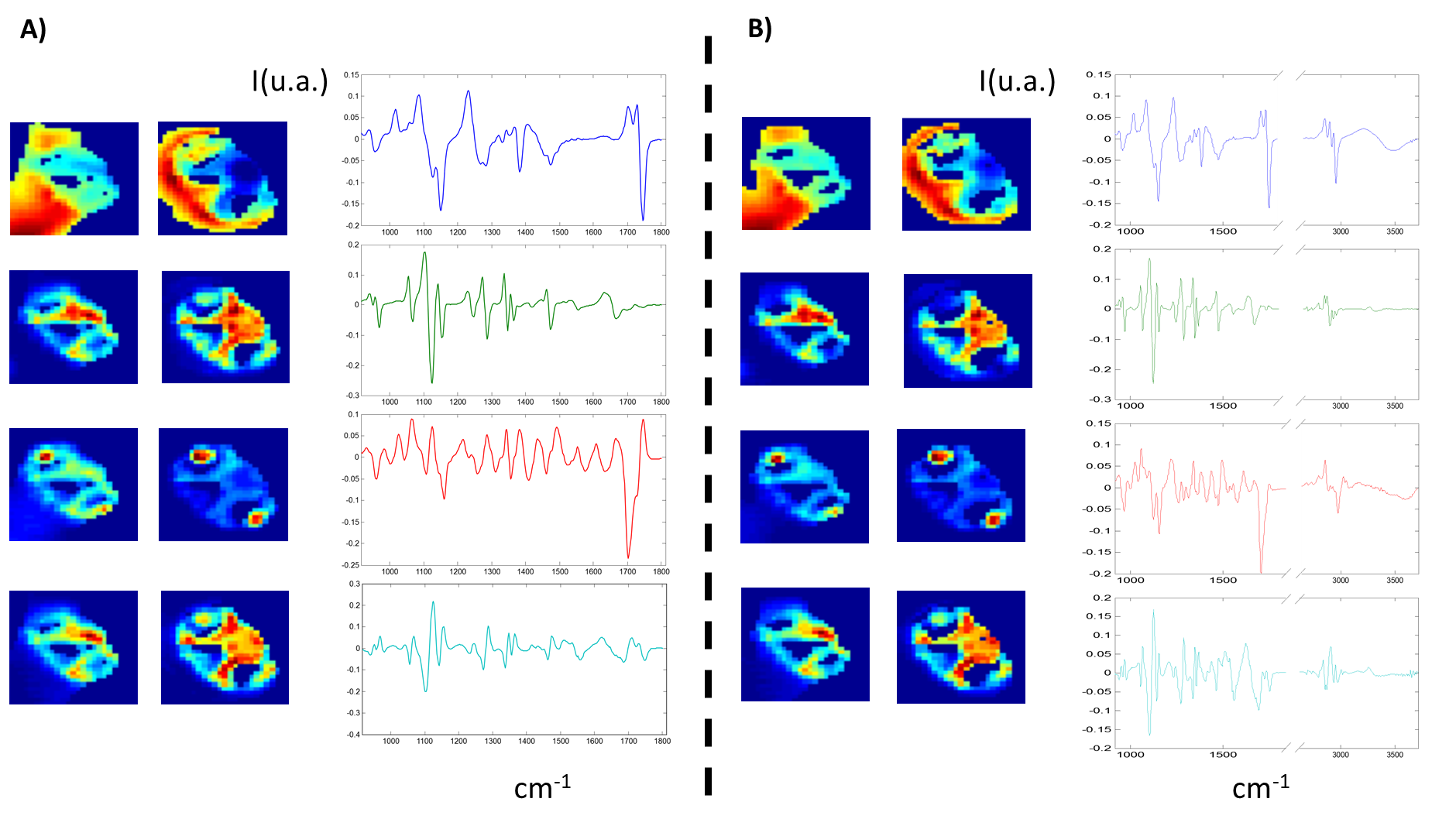
This strategy can be used, in general, when contributions related to non-biological material (reagents or sample supports) provide a significant signal in the images scanned. In these cases, obtaining pure spectra of these well identified contributions is easy and adding them into a multiset analysis improves the modelling of the biological contributions of interest.

1. *Use of extended spectral region (916 to 3715 cm-1) vs. using fingerprint region (916 to 1800 cm-1).*

In most FT-IR images of biological materials, the attention is focused on the fingerprint region because this spectral range provides better features for biological interpretation. However, the inclusion of a wider spectral range may improve the capability of differentiation among biological contributions. To check this fact, a multiset analysis was performed including the two FT-IR images in Figure 6 using only the fingerprint region and the extended spectral range.

The individual resolution of each of the images needed four contributions in the two images, one of them from the OCT compound and the other three from common biological contributions (independently of the spectral range analyzed). From these results, two multisets have been built that contain the two biological images and a third matrix containing the pure OCT spectra, as explained above. One multiset covered the extended spectral range and the other one only the fingerprint region.

Multiset analysis has been performed on the extended spectral range and the fingerprint region. Correspondence among species constraint has been implemented to distinguish better the contribution related to the embedding compound from the rest. Figure 8 shows the final results of the two multiset resolutions. Lack of fit is satisfactory for the multiset analysis of FT-IR spectra. All the contributions resolved are present in both multisets, but the resolution using the extended spectral range presents better discrimination among contributions. This can be seen in the best morphological definition of the third component (associated with the crystalline), which has less presence of surrounding tissues and on the distinction of spectral signatures, since the region between 2650 to 3715 cm-1 is clearly different among certain contributions (OCT and crystalline have a much dominant signal in the band located around 3500 cm-1, as compared with the other biological contributions). This is also seen in the map of the second resolved component, which now does not show significant concentration in the crystalline area. Therefore, it is important to note that whereas the fingerprint region may be more informative in order to identify biological structures, the improvement of the definition of the MCR contributions on the sample when the full spectra is used is clearly valuable.



**Figure 8:** MCR-ALS resolution results of multiset analysis of FT-IR images using only the fingerprint spectral range (A) and the full spectral range from 916 to 3715 cm-1 (B).

The identification of the resolved spectra acquired by FT-IR is more difficult than in Raman without previous knowledge of the composition of the sample because the higher similarity among spectral signatures. The combination of the distribution maps and the fingerprint spectra can be used to suggest a biological tissue related to a resolved feature; e.g. the 3rd component resolved is mainly present in the crystalline lens of the eyes. In the case that a spectral library is available, the same preprocessing used to analyze the image, e.g., first derivative in this case, can be applied to the reference spectra, as explained by Piqueras et al.[48,49], and each spectrum of the library can be projected into the **ST** matrix to identify the possible components of the resolved spectra. Band interpretation is often used in FT-IR because identification due to some bond vibrations can, sometimes, be related to a specific family of biomolecules. Proteins (amide A, B, I-VII bands), lipids and nucleic acids are usually identified in the literature[5,50–52]. In spite of the first derivative applied to the spectra some bands can be assigned (component related to embedding OCT is avoided). All component present bands related to proteins in the amide I area (1600-1700 cm-1), with a clear dominance in the crystalline compound. The component related to the eye crystalline lens (the 3rd one) has a wide band around 1260 cm-1 which may be responsible for the presence of amide III. Component two and three show bands around 1100 and 1300 cm-1, related to DNA structures; other minor bands near 1000 cm-1 in the 2nd and 3rd components may be due to the presence of carbohydrates. Components 2 and 4 have a very similar distribution pattern, consistent with the distribution of cartilage from the tectum, trabecula and quadrate in zebra fish. The separation of signals from these structures into two components may be related to the binary composition of cartilage, formed by a mixture of collagen and peptidoglycan, which may not be homogeneous throughout this kind of tissue. This hypothesis is consistent with the protein signals present in both components and with the sugar signals present in component 4.

**5. CONCLUSIONS**

The potential of Raman and IR hyperspectral images in biological and biomedical samples has been clearly proven in recent research. The protocols explained provide possible solutions to handle Raman and IR images and to solve problems linked to spectroscopic and biological aspects. Selection of preprocessing methodologies supported on the nature of measurement artifacts and the degree of spectral similarity is described. The use of MCR-ALS as a chemometric tool to analyze biological images is described in detail as a powerful tool to unravel and characterize components related to biological structures present in the samples. Multiset analysis of several images is strongly recommended because the use of information coming from several samples or several individuals allows a more robust description of the resolved biological contributions. In the framework of multiset analysis, useful strategies are proposed: a) to facilitate the modeling of several images together by setting presence of common and specific sample contributions and b) to model non-biological contributions, such as tissue embedding compounds or, in general, spectroscopic contributions linked to any kind of sample support by adding specific information related to these spurious sample contributions. Use of spectral ranges going beyond the spectral fingerprint region is also suggested in FT-IR images to increase the discriminating power among compounds. All the steps proposed can be used as complete data analysis protocols or be combined adequately depending on the specific problem or dataset of interest.

**6. ACKNOWLEDGEMENTS**

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073 (CHEMAGEB project). The authors of this work belong to the network of recognized research groups by the Catalan government (2014 SGR 1106). L.B. acknowledges the support of the Spanish government through project CTQ2015-66254-C2-2-P.

ICFO would like to acknowledge financial support from Laserlab-Europe (EU-H2020 654148), the Spanish MINECO (Severo Ochoa grant SEV-2015-0522), Marató de TV3 (20142030), and the National Institute of Health (NIH, grant 5R21CA187890-02). The research conducted at ICFO’s Super Resolution Light Microscopy and Nanoscopy Facility has been partially supported by Fundació Cellex Barcelona.

**REFERENCES**

[1] C. Zhang, D. Zhang, J.-X. Cheng, Coherent Raman Scattering Microscopy in Biology and Medicine, 2015. doi:10.1146/annurev-bioeng-071114-040554.

[2] C.R. Flach, D.J. Moore, Infrared and Raman imaging spectroscopy of ex vivo skin, Int. J. Cosmet. Sci. 35 (2013) 125–135. doi:10.1111/ics.12020.

[3] A. a. Bunaciu, S. Fleschin, H.Y. Aboul-Enein, Biomedical Investigations Using Fourier Transform-Infrared Microspectroscopy, Crit. Rev. Anal. Chem. 44 (2014) 270–276. doi:10.1080/10408347.2013.829389.

[4] J.-X. Cheng, X.S. Xie, Vibrational spectroscopic imaging of living systems: An emerging platform for biology and medicine, Science (80-. ). 350 (2015) aaa8870-aaa8870. doi:10.1126/science.aaa8870.

[5] C. Petibois, K. Wehbe, K. Belbachir, R. Noreen, G. Deleris, Current Trends in the Development of FTIR Imaging for the Quantitative Analysis of Biological Samples, Acta Phys. Pol. A. 115 (2009) 507–512.

[6] M. Ogawa, Y. Harada, Y. Yamaoka, K. Fujita, H. Yaku, T. Takamatsu, Label-free biochemical imaging of heart tissue with high-speed spontaneous Raman microscopy, Biochem. Biophys. Res. Commun. 382 (2009) 370–374. doi:10.1016/j.bbrc.2009.03.028.

[7] K.R. Bambery, E. Schültke, B.R. Wood, S.T. Rigley MacDonald, K. Ataelmannan, R.W. Griebel, B.H.J. Juurlink, D. McNaughton, A Fourier transform infrared microspectroscopic imaging investigation into an animal model exhibiting glioblastoma multiforme., Biochim. Biophys. Acta. 1758 (2006) 900–7. doi:10.1016/j.bbamem.2006.05.004.

[8] S. Piqueras, C. Krafft, C. Beleites, K. Egodage, F. von Eggeling, O. Guntinas-Lichius, J. Popp, R. Tauler, a. de Juan, Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues, Anal. Chim. Acta. 881 (2015) 24–36. doi:10.1016/j.aca.2015.04.053.

[9] J.N. Bentley, M. Ji, X.S. Xie, D. a Orringer, Real-time image guidance for brain tumor surgery through stimulated Raman scattering microscopy., Expert Rev. Anticancer Ther. 14 (2014) 359–61. doi:10.1586/14737140.2013.877844.

[10] K. Kong, F. Zaabar, E. Rakha, I. Ellis, A. Koloydenko, I. Notingher, Towards intra-operative diagnosis of tumours during breast conserving surgery by selective-sampling Raman micro-spectroscopy., Phys. Med. Biol. 59 (2014) 6141–52. doi:10.1088/0031-9155/59/20/6141.

[11] B.F.M. Romeike, T. Meyer, R. Reichart, R. Kalff, I. Petersen, B. Dietzek, J. Popp, Coherent anti-Stokes Raman scattering and two photon excited fluorescence for neurosurgery, Clin. Neurol. Neurosurg. 131 (2015) 42–46. doi:10.1016/j.clineuro.2015.01.022.

[12] S. Takamori, K. Kong, S. Varma, I. Leach, H.C. Williams, I. Notingher, Optimization of multimodal spectral imaging for assessment of resection margins during Mohs micrographic surgery for basal cell carcinoma, Biomed. Opt. Express. 6 (2015) 98. doi:10.1364/BOE.6.000098.

[13] E. Locatelli, I. Monaco, M.C. Franchini, Hard and Soft Nanoparticles for Image Guided Surgery in Nanomedicine, J. Nanoparticle Res. 17 (2015) 1–17. doi:10.1007/s11051-015-3135-x.

[14] F.L. Cals, T.C. Bakker Schut, J.A. Hardillo, R.J. Baatenburg de Jong, S. Koljenović, G.J. Puppels, Investigation of the potential of Raman spectroscopy for oral cancer detection in surgical margins, Lab. Investig. 95 (2015) 1186–1196. doi:10.1038/labinvest.2015.85.

[15] S. Piqueras, L. Duponchel, M. Offroy, F. Jamme, R. Tauler, a. De Juan, Chemometric strategies to unmix information and increase the spatial description of hyperspectral images: A single-cell case study, Anal. Chem. 85 (2013) 6303–6311. doi:10.1021/ac4005265.

[16] K. Lau, A. Hobro, T. Smith, T. Thurston, B. Lendl, Label-free non-destructive in situ biochemical analysis of nematode Steinernema kraussei using FPA-FTIR and Raman spectroscopic imaging, Vib. Spectrosc. 60 (2012) 34–42. doi:10.1016/j.vibspec.2012.01.009.

[17] H.-J. van Manen, A. Lenferink, C. Otto, Noninvasive imaging of protein metabolic labeling in single human cells using stable isotopes and Raman microscopy., Anal. Chem. 80 (2008) 9576–9582. doi:10.1021/ac801841y.

[18] M. Ishikawa, Y. Maruyama, J.Y. Ye, M. Futamata, Single-molecule imaging and spectroscopy of adenine and an analog of adenine using surface-enhanced Raman scattering and fluorescence, J. Lumin. 98 (2002) 81–89. doi:10.1016/S0022-2313(02)00255-7.

[19] Y.-Y. Ha, Y.-W. Park, H. Kweon, Y.-Y. Jo, S.-G. Kim, Comparison of the physical properties and in vivo bioactivities of silkworm-cocoon-derived silk membrane, collagen membrane, and polytetrafluoroethylene membrane for guided bone regeneration, Macromol. Res. 22 (2014) 1018–1023. doi:10.1007/s13233-014-2138-2.

[20] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176. doi:10.1080/10408340600970005.

[21] R. Tauler, M. Maeder, A. De Juan, S. Analyzed, D. Matrices, Q. Information, E. Variances, U. Different, R.L. Profiles, F. Bands, N. Propagation, Multiset Data Analysis : Extended Multivariate Curve Resolution, in: Compr. Chemom., 2009: pp. 473–505. doi:10.1016/B978-044452701-1.00055-7.

[22] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12. doi:10.1016/j.chemolab.2014.10.003.

[23] M.C. Fishman, Vertebrate, Science (80-. ). 294 (2001) 1290–1291.

[24] D.M. Parichy, M.R. Elizondo, M.G. Mills, T.N. Gordon, R.E. Engeszer, Normal table of postembryonic zebrafish development: staging by externally visible anatomy of the living fish., Dev. Dyn. 238 (2009) 2975–3015. doi:10.1002/dvdy.22113.

[25] C.B. Kimmel, W.W. Ballard, S.R. Kimmel, B. Ullmann, T.F. Schilling, Stages of Embryonic Development of the Zebrafish, Dev. Dyn. 203 (1995) 253–310.

[26] D.G. Howe, Y.M. Bradford, T. Conlin, A.E. Eagle, D. Fashena, K. Frazer, J. Knight, P. Mani, R. Martin, S.A.T. Moxon, H. Paddock, C. Pich, S. Ramachandran, B.J. Ruef, L. Ruzicka, K. Schaper, X. Shao, A. Singer, B. Sprunger, C.E. Van Slyke, M. Westerfield, ZFIN, the Zebrafish Model Organism Database: Increased support for mutants and transgenics, Nucleic Acids Res. 41 (2013). doi:10.1093/nar/gks938.

[27] A. V. Kalueff, D.J. Echevarria, S. Homechaudhuri, A.M. Stewart, A.D. Collier, A.A. Kaluyeva, S. Li, Y. Liu, P. Chen, J. Wang, L. Yang, A. Mitra, S. Pal, A. Chaudhuri, A. Roy, M. Biswas, D. Roy, A. Podder, M.K. Poudel, D.P. Katare, R.J. Mani, E.J. Kyzar, S. Gaikwad, M. Nguyen, C. Song, Zebrafish neurobehavioral phenomics for aquatic neuropharmacology and toxicology research, Aquat. Toxicol. 170 (2015) 297–309. doi:10.1016/j.aquatox.2015.08.007.

[28] U. Gündel, S. Kalkhof, D. Zitzkat, M. von Bergen, R. Altenburger, E. Küster, Concentration-response concept in ecotoxicoproteomics: Effects of different phenanthrene concentrations to the zebrafish (Danio rerio) embryo proteome, Ecotoxicol. Environ. Saf. 76 (2012) 11–22. doi:10.1016/j.ecoenv.2011.10.010.

[29] D. Raldúa, B. Piña, *In vivo* zebrafish assays for analyzing drug toxicity, Expert Opin. Drug Metab. Toxicol. 10 (2014) 685–697. doi:10.1517/17425255.2014.896339.

[30] S.R. Mesquita, J. Dachs, B.L. van Drooge, J. Castro-Jiménez, L. Navarro-Martín, C. Barata, N. Vieira, L. Guimarães, B. Piña, Toxicity assessment of atmospheric particulate matter in the Mediterranean and Black Seas open waters, Sci. Total Environ. 545–546 (2016) 163–170. doi:10.1016/j.scitotenv.2015.12.055.

[31] S.R. Mesquita, B.L. Van Drooge, C. Reche, L. Guimarães, J.O. Grimalt, C. Barata, B. Piña, Toxic assessment of urban atmospheric particle-bound PAHs: Relevance of composition and particle size in Barcelona (Spain), Environ. Pollut. 184 (2014) 555–562. doi:10.1016/j.envpol.2013.09.034.

[32] A.C. Price, C.J. Weadick, J. Shim, F.H. Rodd, P.E.T. Al, Pigments , Patterns , and Fish Behavior, Zebrafish. 5 (2008).

[33] P. a. Raymond, L.K. Barthel, G. a. Curran, Developmental patterning of rod and cone photoreceptors in embryonic zebrafish, J. Comp. Neurol. 359 (1995) 537–550. doi:10.1002/cne.903590403.

[34] E. Oliveira, M. Casado, M. Faria, A.M.V.M. Soares, J.M. Navas, C. Barata, B. Piña, Transcriptomic response of zebrafish embryos to polyaminoamine (PAMAM) dendrimers., Nanotoxicology. 8 Suppl 1 (2014) 92–9. doi:10.3109/17435390.2013.858376.

[35] P.H.C. Eilers, H.F.M. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, Life Sci. (2005) 1–26. doi:10.1021/ac034173t.

[36] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. doi:10.1021/ac60214a047.

[37] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146. doi:10.1016/0169-7439(95)00047-X.

[38] R. Tauler, A. Smilde, B. Kowalski, Selectivity, Local Rank, 3-Way Data-Analysis And Ambiguity In Multivariate Curve Resolution, J. Chemom. 9 (1995) 31–58.

[39] A. De Juan, S.C. Rutan, R. Tauler, Two-Way Data Analysis: Multivariate Curve Resolution - Iterative Resolution Methods, in: Compr. Chemom., 2010: pp. 325–344. doi:10.1016/B978-044452701-1.00050-8.

[40] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, Anal. Chem. 63 (1991) 1425–1432. doi:10.1021/ac00014a016.

[41] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420. doi:10.1007/BF02163027.

[42] R. Tauler, D. Barceló, Multivariate curve resolution applied to liquid chromatography-diode array detection, Trends Anal. Chem. 12 (1993) 319–327.

[43] R. Manne, On the resolution problem in hyphenated chromatography, Chemom. Intell. Lab. Syst. 27 (1995) 89–94. doi:10.1016/0169-7439(95)80009-X.

[44] A.C.S. Talari, Z. Movasaghi, S. Rehman, I.U. Rehman, Raman Spectroscopy of Biological Tissues, Appl. Spectrosc. Rev. 50 (2015) 46–111. doi:10.1080/05704928.2014.923902.

[45] I. Siebinga, G.F.J.M. Vrensen, K. Otto, G.J. Puppels, F.F.M. De Mul, J. Greve, Ageing and changes in protein conformation in the human lens: A Raman microspectroscopic study, Exp. Eye Res. 54 (1992) 759–767. doi:10.1016/0014-4835(92)90031-M.

[46] S. V Saenko, J. Teyssier, D. van der Marel, M.C. Milinkovitch, Precise colocalization of interacting structural and pigmentary elements generates extensive color pattern variation in Phelsuma lizards., BMC Biol. 11 (2013) 105. doi:10.1186/1741-7007-11-105.

[47] Z. Huang, H. Lui, X.K. Chen, A. Alajlan, D.I. McLean, H. Zeng, Raman spectroscopy of in vivo cutaneous melanin., J. Biomed. Opt. 9 (2004) 1198–205. doi:10.1117/1.1805553.

[48] S. Piqueras, L. Duponchel, R. Tauler, A. De Juan, Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares, Anal. Chim. Acta. 705 (2011) 182–192. doi:10.1016/j.aca.2011.05.020.

[49] R. de Juan, Anna; Piqueras, Sara; Maeder, Marcel; Hancewicz; Duponchel, Ludovic; Tauler, Chemometric Tools for Image Analysis, in: H. Salzer, Reiner; Siesler (Ed.), Infrared Raman Spectrosc. Imaging, Second, Co, Wiley-VCH, 2014: pp. 57–106.

[50] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, M. Konrad, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, (2014) 1–35. doi:10.1038/nprot.2014.110.

[51] J. Kong, S. Yu, Fourier Transform Infrared Spectroscopic Analysis of Protein Secondary Structures Protein FTIR Data Analysis and Band Assignment, Acta Biochim. Biophys. Sin. (Shanghai). 39 (2007) 549–559. doi:10.1186/1479-5876-10-117.

[52] J.G. Kelly, J. Trevisan, A.D. Scott, P.L. Carmichael, H.M. Pollock, P.L. Martin-Hirsch, F.L. Martin, Biospectroscopy to metabolically profile biomolecular structure: A multistage approach linking computational analysis with biomarkers, J. Proteome Res. 10 (2011) 1437–1448. doi:10.1021/pr101067u.