2011-11

Understanding the Molecular Information Contained in Principal Component Analysis of Vibrational Spectra of Biological Systems

Franck Bonnier
Technological University Dublin, Franck.Bonnier@tudublin.ie

Hugh Byrne
Technological University Dublin, hugh.byrne@tudublin.ie

Recommended Citation

Bonnier, F. & Byrne, H. (2012) Understanding the Molecular Information Contained in Principal Component Analysis of Vibrational Spectra of Biological Systems. *Analyst,* 2012, 137, 322. doi:10.1039/c1an15821j

Understanding the molecular information contained in Principal Component Analysis of Vibrational Spectra of Biological Systems

F. Bonnier, H.J. Byrne

Focas Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland

Abstract:

K-means clustering followed by Principal Component Analysis (PCA) is employed to analyse Raman spectroscopic maps of single biological cells. K-means clustering successfully identifies regions of cellular cytoplasm, nucleus and nucleoli, but the mean spectra do not differentiate their biochemical composition. The loadings of the principal components identified by PCA shed further light on the spectral basis for differentiation but they are complex and, as the number of spectra per cluster is imbalanced, particularly in the case of the nucleoli, the loadings under-represent the basis for differentiation of some cellular regions. Analysis of pure bio-molecules, both structurally and spectrally distinct, in the case of histone, ceramide and RNA, and similar in the case of the proteins albumin, collagen and histone, show the relative strong representation of spectrally sharp features in the spectral loadings, and the systematic variation of the loadings as one cluster becomes reduced in number. The more complex cellular environment is simulated by weighted sums of spectra, illustrating that although the loading become increasingly complex; their origin in a weighted sum of the constituent molecular components is still evident. Returning to the cellular analysis, the number of spectra per cluster is artificially balanced by increasing the weighting of the spectra of smaller number clusters. While it renders the PCA loading more complex for the three-way analysis, a pair wise analysis illustrates clear differences between the identified subcellular regions, and notably the

molecular differences between nuclear and nucleoli regions are elucidated. Overall, the study demonstrates how appropriate consideration of the data available can improve the understanding of the information delivered by PCA.

Keywords: Cellular imaging; Raman Spectroscopy; Multivariate analysis; K-mean Clustering Analysis; Principal Component Analysis;

* Corresponding author. *Tel.:* +353 1 4027917; *Fax*: +353 1 4027904;
E-mail address: fbonnier@dit.ie (F. Bonnier)

1. Introduction

The potential of Infrared and Raman spectroscopy as medical diagnostic tools has been well demonstrated [1-5]. The main advantages of these techniques are that they provide a non invasive, label free, molecular fingerprint of tissue and cells with a high specificity that can be used for the identification of different pathologies [6, 7] or variations in metabolism as a result of external agents [8, 9]. The two techniques are commonly described as complementary and are often performed in parallel for comparison purposes [10, 11] However, the different characteristics of the instrumentation can influence their applications. Although in the last few years, many further reaching applications have been described, including in radiobiology [12], toxicology and nanotoxicology [13, 14] and pharmacokinetics [15], Fourier Transform Infrared Spectroscopy (FTIR) remains widely used for analysis of tissue sections, mainly because the technique enables the analysis of large sample areas in a relatively short time and enables identification of pathological areas present in the tissue analysed [3, 16-18]. Nevertheless, due to the diffraction limit, the lateral resolution can be a major limitation in infrared spectroscopy. As it commonly employs optical or near infrared sources, Raman spectroscopy attains significantly higher lateral resolution enabling access to sub-cellular information [19, 20]. For this reason, it is often preferred for single cell analysis and imaging. Although the two techniques are based on different modes of interaction between the incident light and the sample, they suffer from similar concerns in terms of data processing and analysis. In both cases, the spectral background can be a combination of many signals contributing to and “contaminating” the spectra recorded. In FTIR microscopy, the physical origins of the broad undulating background [21] and the so-called dispersion artefact in transflection [22] and transmission [23] mode have recently been

elucidated and the Extended Multiplicative Scattering Correction (EMSC) has been further evolved to correct for both [24, 25]. In Raman spectroscopy, the most commonly reported artefact is the presence of a broad background which can often swamp the sample signal [26, 27]. Different algorithms can be applied to subtract this background from the spectra [28] or, alternatively it has been demonstrated recently that recording the spectra in immersion, or in collagen gels, greatly improves the signal to background ratio [29].

The improvements made in recent years through the development of new methods and approaches for sample analysis and data pre-processing have considerably increased the relevancy of the information contained in the data, facilitating the standardisation of methods for data analysis [30]. Different approaches such as PCA or K-means clustering are commonly used for the analysis of large amounts of data, allowing a discrimination of different samples or regions of a sample, according to differences in their biochemical content, and identification of the spectral features which manifest the highest degree of variability [1, 7, 31-33]. Although these methods are used routinely and are quite well developed, the question of the relevance and molecular specificity of the information contained in the data remains largely unaddressed. Notably, in an increasingly multidisciplinary field, the results and underlying processes are seldom scrutinised.

K-means clustering has been employed widely in tissue analysis, providing the possibility of grouping the spectra into different clusters based on spectral similarity and therefore identifying common biochemical signatures and their spatial distributions. The clustering is based on the molecular information contained in the individual spectra and the results are commonly displayed as false colour maps of the average spectra of each cluster. While the technique is valuable for visualisation of

the spectrally differentiated regions of the sample, the maps themselves do not show the original spectral data, and the mean spectrum representing each cluster can often under represent subtle differences between the different regions of the sample. PCA is a powerful approach for the analysis of large spectral data sets. It represents the spectra in data groupings of similar variability, allowing the identification and differentiation of different spectral groups. This approach is widely used to evaluate the possibility of discriminating different data sets (and therefore samples or regions of samples) using scatter plots [29]. However, the strength of this technique resides in the loadings, which give a representation of the spectral origin of the variations which differentiate the data grupings according to the wavenumbers [34-36]. A combination of K-means clustering followed by PCA constitutes a well adapted tool for analysis of spectral data [19]. Nevertheless, the analysis of the PCA loadings is not trivial and the relevance of the information contained in the plots often remains enigmatic.

In this study, Raman spectral mapping of single biological cells is presented and analysed using a combination of K-means cluster followed by PCA. In order to better understand the information obtained through the analysis, Raman spectra of individual biomolecular components are analysed using PCA. The observations made from the scatter plots are correlated with the different loadings calculated using simple models based on these samples, and the influence of spectral differences and number per dataset group are examined. The observations made are applied to the more complex data sets derived from sub-cellular maps recorded from single cells, demonstrating that a better understanding of the molecular contributions to the spectral variances can improve the data analysis process.

2. Materials and Methods

2.1. Sample preparation

A549 cells from a human lung adenocarcinoma with alveolar type II phenotype were obtained from ATTC (Manassas, VA, USA). Cells were cultured in DMEM (Sigma), penicillin and streptomycin (Gibco) and 10% foetal calf serum (FCS, Biochrom, Berlin) in a humidified atmosphere containing 5% CO_2 at 37ºC. Cells were loaded, at a concentration of 4×10^4 cells, onto CaF_2 and were incubated for 24h at 37°C, 95% CO_2. Before measurements, the cell samples were fixed using a 10% formalin solution for 10 mins. A number of studies on the effect of cell fixation using Raman spectroscopy and can be found in the literature [34, 37, 38]. By comparison between live cells and fixed cells using different fixation procedures, it has been demonstrated than the closest model for live cells is achieved using 10% formalin fixation. This is explained by the fact that while many fixation techniques require drying of the sample, formalin fixation keeps the cells in a hydrated state. Moreover, although the formalin could affect the protein, it also maintains the lipid content relatively intact compared to fixation methods employing alcohol. After fixation, the cells were washed 3 times using PBS and kept in this solution during the measurements.

Pure samples of biomolecular components were analysed by Raman spectroscopy to simulate biologically variable systems. In order to achieve the best homogeneity, samples were dissolved in appropriate solvents before deposition on CaF_2 substrates. Lyophilised ribonucleic acid, albumine and histones (Sigma-Aldrich, Ireland) were dispersed in water. Lyophilised deoxyribonucleic acid from calf thymus, ceramide and L-phosphatidyl-ethanolamine (Sigma-Aldrich, Ireland) were dispersed in chloroform. Collagen solution in acetic acid at a concentration of 5mg/mL (Gibco, Ireland) was deposited from acetic acid solution. For all biomolecular samples, the Raman signals were stable over prolonged measurement periods (~1hr). Signal

variations were largely due to inhomogeneity in the deposited sample and variation in the background present in the spectra collected.

The spectra were also utilised help identify the biochemical origins of the features of PC loading spectra. Isolation of the different biomolecular components from cells would present the best references for the characterisation of the cellular content. However, comparisons are qualitative only, and moreover, the similarity existing between the pure bio-molecules tested and the spectra recorded from the cells is satisfactory and gives a good representation of the information delivered by the PCA.

2.2. Raman spectroscopic measurements

A Horiba Jobin-Yvon LabRAM HR800 spectrometer was used throughout this work. For the measurements, either a x100 objective (MPlanN, Olympus), for the recording of spectra from the pure compounds, or a x100 immersion objective (LUMPlanF1, Olympus), for cell mapping, was employed, each providing a spot diameter of ~1μm at the sample. The confocal hole was set at 100μm for all measurements, the specified setting for confocal operation. The 785 nm laser line was used for this study, delivering a power of ~40 mW at the sample. The system was spectrally calibrated to the 520.7 cm^{-1} spectral line of silicon.

The LabRAM system is a confocal spectrometer that contains two interchangeable gratings (300 and 600 lines/mm respectively). In the following experiments, the 300 lines/mm grating was used, providing a spectral dispersion of approximately 1.5 cm^{-1} per pixel. The backscattered Raman signal was integrated for 10 second intervals over the spectral range from 400 to 1800 cm^{-1}. The detector used was a 16-bit dynamic range Peltier cooled CCD detector. 100 spectra were collected for each different bio-molecular component tested. The spectra recorded from single cells have been obtained using the mapping function of the Labspec software using a 1 μm step size.

The Raman cellular map presented in this paper has been selected for illustration purposes. It is a typical example and matches observations made in a large cell screening study as presented in previous work [19].

2.3 Data preprocessing

Data analysis was performed using Matlab (Mathworks, USA). Before statistical analysis, a Savitsky-Golay filter (5th order, 7 points) was applied to smooth the spectra and the reference spectrum constituting the background signal was subtracted. The data sets have also been corrected for baseline and vector normalized to facilitate comparison.

2.4 Data Analysis

K-means clustering analysis is one of the simplest unsupervised learning algorithms used for spectral image analysis. It groups the spectra according to their similarity, forming clusters, each one representing regions of the image with identical molecular properties. The distribution of chemical similarity can then be visualised across the sample image. The number of clusters (k) has to be determined *a priori* by the operator before initiation of the classification of the data set. K centroids are defined, ideally as far as possible from each other, and then each point belonging to a data set is associated to the nearest centroid. When all the points have been associated with a centroid, the initial grouping is done. The second step consists of the calculation of new centroids as barycentres of the clusters resulting from the previous step. A new grouping is implemented between the same data points and the new centroids. These operations are repeated until convergence is reached and there is no further movement

of the centroids. Finally, k clusters are determined, each containing the most similar spectra from the image, and represented by the mean of all spectra of that cluster. False color maps can then be constructed to visualise the organisation of the clusters in the original image.

PCA is a method of multivariate analysis broadly used with datasets of multiple dimensions [30]. It allows the reduction of the number of variables in a multidimensional dataset, although it retains most of the variation within the dataset. The order of the principal components (PCs) denotes their importance to the dataset. PC1 describes the highest amount of variation, PC2 the second highest, and so on. Therefore, var (PC1) ≥ var (PC2) ≥ var (PCp), where var (PCi) represents the variance of PCi in the considered data set. Generally, the first three PCs represent the highest variance present in the data sets, up to 99%, giving the best visualisation of the differentiation of the different clusters [39, 40]. However, when recording Raman data from single cells, the noise present in the spectra can increase the intra-group variability, thus reducing the specificity of the PCA. In such cases, typically the 10 first PCs can be taken into account for specific analysis [41]. Nevertheless, the PCs contribute less in decreasing order, meaning that the first PCs contain the most information [42]. In order to simplify interpretation of experimental observations, this study is aimed at understanding the molecular information contained in the first two PCs. It is considered that the observations made in this work can be applied for all the different PCs calculated using PCA.

PCA was employed for this study to highlight the variability existing in the spectral data set recording during the different experiments. The other advantage of this method is the observation of loadings which represent the variance for each variable (wavenumber) for a given PC. Analysing the loadings of a PC can give information

about the source of the variability inside a dataset, derived from variations in the molecular components contributing to the spectra.

3. Results and discussion

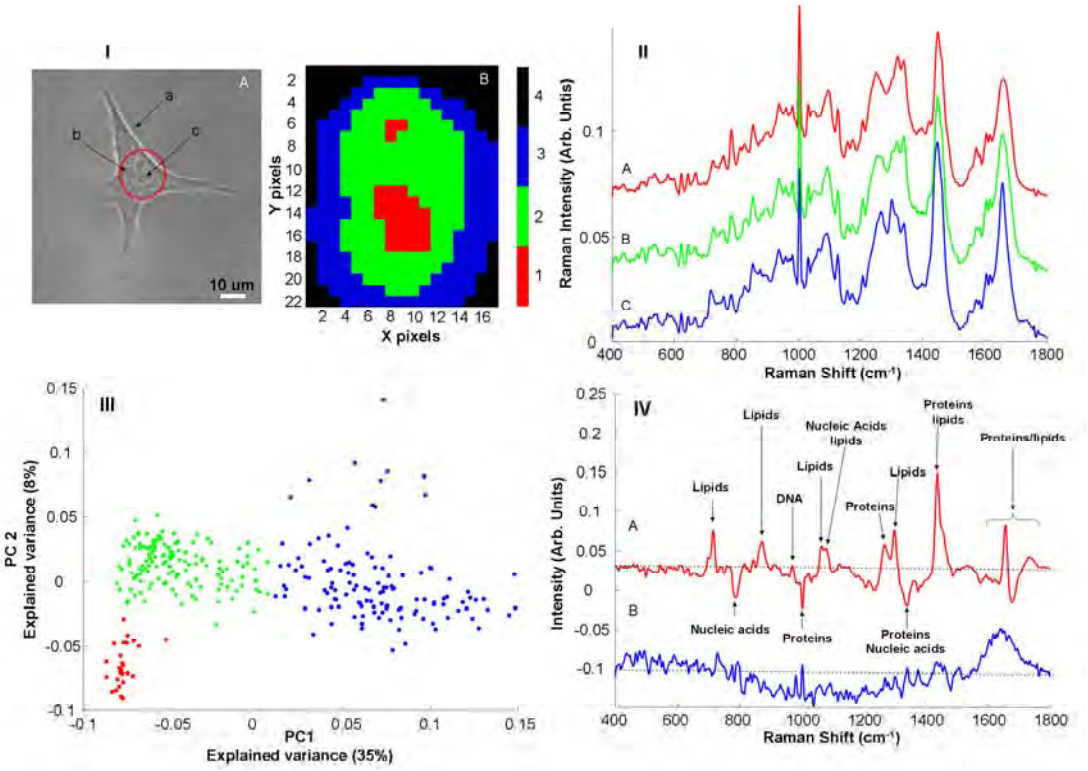

Figure 1. ***I:*** *(A), Typical confocal microscope image of an A549 cell. The different structures such as membrane (a), cytoplasm (b) and nucleus (c) are clearly identifiable. The nucleolus present inside the nucleus can also be seen. The area delineated by red indicates a "typical area" selected for Raman mapping (B), Example of K-means reconstructed image from a Raman map recorded on the nuclear area of an A549. In both X and Y directions, 1 pixel corresponds to a mapping step of 1 µm.* ***II:*** *Mean spectrum calculated for the different clusters obtained after K-means clustering analysis corresponding to the nucleoli (A), nucleus (B) and cytoplasm (C).* ***III:*** *Scores plot of the first two principal components after PCA performed on Raman spectra recorded from A549 cells. The individual data points have been colour coded according to the results of K-means cluster analysis; nucleus (green), nucleolus (blue) and cytoplasm (red).* ***IV:*** *Plot of the loadings of PC1 (A) and PC2 (B). Different features corresponding to the lipids, proteins and nucleic acids can be identified.*

3.1 Single cell analysis

For the purpose of this study, rather than a full cellular analysis [18], the objective was to differentiate only three different types of spectra, and for this reason the acquisition was focused in the nucleus including some neighbouring cytoplasm. An optical image of a typical A549 cell is shown in Figure 1 IA. It has been previously demonstrated that Raman spectroscopy can effectively identify and discriminate the cytoplasm as well as nucleoli within the nucleus [18, 19]. Using K-means clustering analysis, three different clusters can thus be identified, corresponding to nucleoli, the nucleus and cytoplasm, as shown in the example of figure 1 IB. Note, the black region represents areas which were not sampled spectroscopically, rather than a fourth K-means cluster. Mean spectra corresponding to the different clusters were derived from the K-means clustering and are shown in figure 1 II. The three different spectra are rather similar and only small variations can be seen between the different structures, illustrating that although the analytical technique is very efficient at identifying sub-cellular regions, the mean spectra give little insight into the basis of differentiation on a molecular basis.

PCA can provide further insight into the source of the spectral variability and therefore differentiation of the different sub-cellular regions. In figure 1 III, the data are colour coded according to their classification by the K-means clustering analysis, but the PCA clusters are widely spread and not clearly differentiated. The spectra have been recorded from cells and the signal is relatively weak, resulting in relatively noisy spectra. Also, variations can be present due to the spatial non uniformity of the sample. Unavoidably, spectra of the nucleus will have varying contributions from the overlying cell membrane and cytoplasm, and the spatial separation of the sub-cellular regions is not necessarily distinct, within the measurement resolution and stepsize (1

μm), although this can be partially alleviated in confocal operation. Nevertheless, the three different groups are relatively well discriminated using this method. PC1, which accounts for 35% of the variance, discriminates the nuclear spectra from those of the cytoplasm, whereas PC2 accounts for 8% of the variance and allows discrimination of the spectra from within the nucleus.

Beyond differentiation and classification, the potential of PCA lies in the possibility to derive information regarding the basis for discrimination from the loadings corresponding to each PC. A clear representation of the spectral variability can be seen, and moreover these loadings can be compared to pristine Raman spectra for comparison [18].

The loadings of the principal components are shown in figure 1 IV. The plots are offset for clarity, the dotted line indicating the zero point in each case. PC1 has peaks which can be attributed to biochemical constituents such as nucleic acids (788, 1080, 1339 cm^{-1}), proteins (1003, 1268, 1339, 1437 cm^{-1}) and lipids (715, 872, 1066, 1080, 1299, 1437 cm^{-1}) (for detailed assignments, see for example [8, 10, 43, 44]). Their respective negative and positive loadings contribute substantially to the differentiation of the nuclear (negative) and cytoplasmic (positive) spectra in the scores plot of figure 1 III. Differentiation of nucleus and cytoplasm based on nucleic acid and lipidic content is somewhat trivial, however, and the loading is rich in features which further contribute to the differentiation, although a detailed analysis is complex. Furthermore, PC2 is rather noisy, and it is difficult to extract specific information relating to biomolecular constituents which differentiate the nuclear and nucleoli datasets in the scores plot.

It should be noted that PCA is an unsupervised technique, and does not differentiate between variability within the dataset groupings and variability between the

groupings. Thus, intra – group variability can contribute substantially to the variance and loadings which differentiate the groups [40, 41]. Nevertheless, in the case of the differentiation of cytoplasm and nucleus/nucleoli by PC1 and nucleus and nucleoli by PC2, there is a clear separation of the groups according to positive and negative scores, indicating that the loadings are representative of underlying biological differences.

PCA, in combination with K-means clustering, sheds further light on the basis for differentiation of the cellular regions. However, the loadings are complex or inconclusive, rendering interpretation of the underlying biology nontrivial. In order to further elucidate factors which govern the differentiation of spectral groups, a series of studies on pure biomolecular compounds was conducted.

3.2 Understanding the PCA

In an effort to better elucidate the process of differentiation by PCA, based on biochemical content, a comparison was made of the Raman spectra collected from 3 structurally distinct bio-molecules: RNA, histone (protein) and ceramide (lipids). 100 spectra were recorded for each sample. Figure 2 I presents mean spectra calculated for each of the samples recorded offset for clarity. These molecules are commonly found in biological samples and have been specifically selected for their high degree of dissimilarity in the spectra range 400-1800 cm^{-1}.

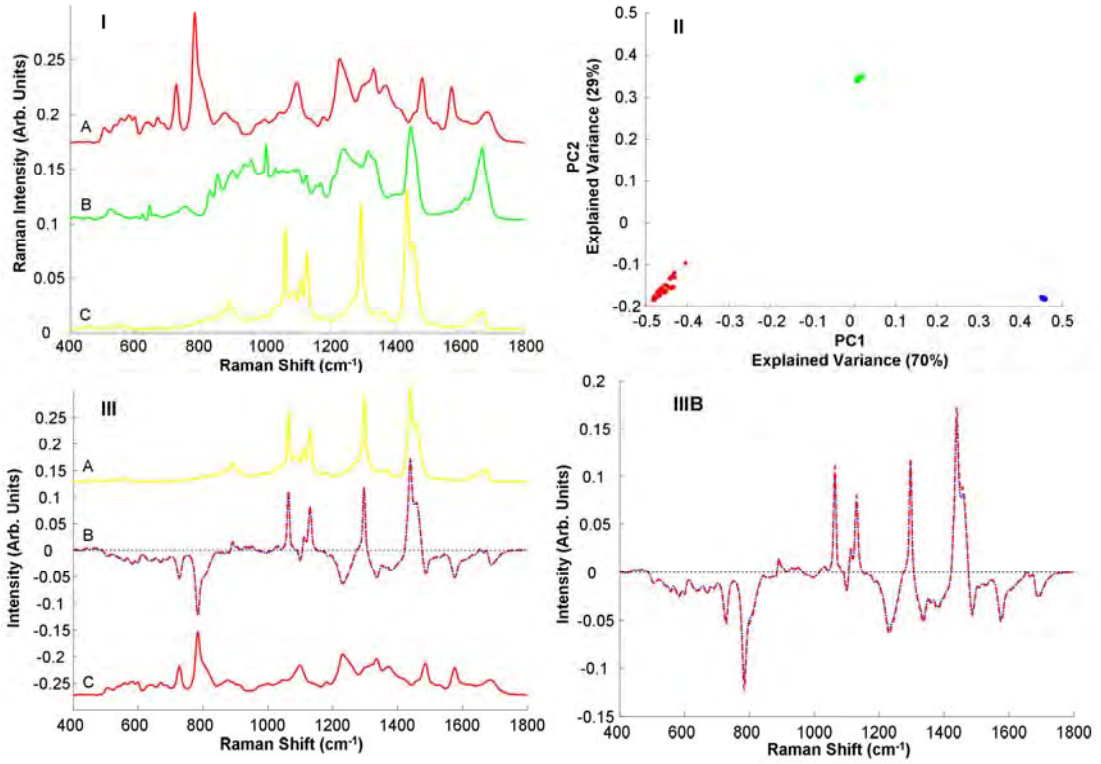

Figure 2. ***I:*** *Mean Raman spectra recorded from RNA (A), histone (B) and ceramide (C) on* CaF_2 *windows.* ***II:*** *Scores plot of the 2 first principal components after PCA performed on Raman spectra recorded from RNA, histone and ceramide.* ***III:*** *Plot of the loadings of PC1 (blue dot line) compared with the difference spectrum calculated from the mean spectrum of ceramide minus the mean spectrum of RNA (red dash line). The loadings are compared with spectra recorded from ceramide (A) and RNA (C), both offset for clarity.* ***IIIB:*** *Plot of the loadings of PC2 (blue dot line) compared with the difference between the mean spectrum of histone minus the average mean spectrum of RNA and ceramide (red dash line).*

The three data sets were loaded in Matlab and PCA was performed on the entire spectral window used for the acquisition. As expected, the three groups are well discriminated in the scores plot, as shown in Figure 2 II. PC1 represents 70% of the explained variance and allows the discrimination between the three groups. Notably, the histones are grouped at ~zero on the PC1 axis, while the RNA and Ceramide spectra are symmetrically grouped at the negative and positive extremes, respectively.

PC2 represents 29% of the variance and discriminates the histone spectra from the RNA and ceramide spectra. For this PC, little or no discrimination between RNA and ceramide is observed. In the example presented in the figure 2 II, the bio-molecules tested present highly different Raman signatures therefore the intra-group variability is very low and the different spectra for each group are closely grouped. The RNA sample was observed to be physically slightly more heterogeneous than the other materials after drying, resulting in significant differences in the intensity of the signal recorded. Thus, even after background and baseline correction, some variation persists. However, the groups are well defined and discriminated across the scores plots indicating that the PCs are a clear representation of inter-group variance.

Figure 2 III compares the loading of PC1 (figure 2 IIIB) with the spectra of pure ceramide (figure 2 IIIA) and RNA (figure 2 IIIC). The primary observation is that all the peaks corresponding to the ceramide appear as positive features in PC1, whereas those from the RNA correspond to negative features of PC1. In relation to the zero-line, the loadings are almost a perfect representation of each pure spectrum, the one of RNA being inverted. Notably, histone features are totally absent in this plot and have absolutely no influence on the information contained in PC1, as might be expected from figure 2 II. To illustrate this effect, the loading obtained from the PCA of the data sets of the RNA and ceramide alone was calculated and was found to overlap exactly with the loading 1 obtain using the 3 data sets (data not shown). However, this loading overlaps perfectly with figure 2 IIIB and so is not discernible. Notably, although all spectra have been normalised before analysis, PCA identifies the histone as having the lowest variability, perhaps because of the lack of sharp individual spectral features in comparison with the other species (Figure 2 I).

An interesting observation is the correlation existing between the values of the loading and the positions of the spectra in the PCA plot. The spectra of the RNA are negative with respect to the loading of PC1 whereas the spectra from the ceramide are positive. Thus, a link exist between the composition of the samples analysed and the profile of the loading of PC1, or, more precisely, the loading of PC1 in this case is a representation of the molecular composition of each sample. To understand the information contained in the loadings calculated from the PCA, a simple comparison is made in figure 2 III. The loading of PC1, which discriminates the RNA from the ceramide, is compared to the difference spectrum between the mean spectrum of ceramide and the mean spectrum of RNA. The two spectra overlap almost exactly and no major variations can be seen.

A similar relationship between the source molecular spectra and the PC loadings can also be demonstrated for PC2. In figure 2 IIIB, the loading of PC2 has been plotted, but in order to find a match, the mean spectra of RNA and ceramide have first been averaged before being subtracted from the mean spectrum of the histones. The resulting spectrum contains all the same peaks as the loading of PC2 and only slight variations in relative peaks intensities are observed.

Thus, PCA can provide information on the molecular composition or underlying biochemical differences of the data sets analysed and the results presented are comparable to the difference spectra that can be calculated by simple subtraction. A direct correlation exists between the position of the spectra in the scores plot and the value of the loadings. The negative scores are as informative as the positives ones, but correspond to two different directions in the scatter plot. Using the scale present for each PC, the negatives and positive peaks can be attributed to the different groups of

the plot and using different reference spectra these peaks can be then matched with specific features for the molecular characterisation of the samples.

3.3 Sensitivity of PCA

In the previous section, PCA was applied to three significantly different bio-molecules having strong dissimilarities in their Raman signatures. In this section, the histone spectra are compared to those of albumin and collagen. These three samples are proteins and therefore have more similar Raman profiles. Figure 3 I presents the mean spectra of albumin and collagen compared to the histone spectrum, offset for clarity. The three spectra contain many similar features and only the regions around 1300-1350 cm^{-1} and 800-950 cm^{-1} exhibit obvious differences between the three different molecules.

Despite the similarities of their spectra, PCA clearly discriminates the molecules as shown in the scores plot of Figure 3 II. As observed in the previous section, the intra-group variance is very low in comparison to the inter-group differentiation suggesting that the loading is s good representation of the spectral variation between the different bio-molecules tested. PC1 describes 71% of the observed variance, and PC2 26%. Albumin and collagen are largely discriminated by PC1, although they are not placed symmetrically about the origin, and they are also discriminated somewhat by PC2. Histone is differentiated from the other two proteins largely by PC2, although in this instance it does not sit at the origin of PC1 and so is somewhat discriminated by PC1. As in this case both PC1 and PC2 contribute to the differentiation of all three species, the loadings are more complex and are not as clearly derived from the individual molecular components as for the previous example. However, in comparison to the spectra of collagen and albumin, specific features for each of them can be identified in

the loadings. As expected, the loading of PC1 is dominated by these two components, as shown in figure 3 III. Indeed, the loading of the PC which differentiates a dataset of just collagen and albumin is almost identical to that which discriminates the three proteins (data not shown). As in the previous example, the loading of PC1 can be accurately reproduced by taking the difference of the mean spectra of albumin and collagen, as shown in figure 3 III.

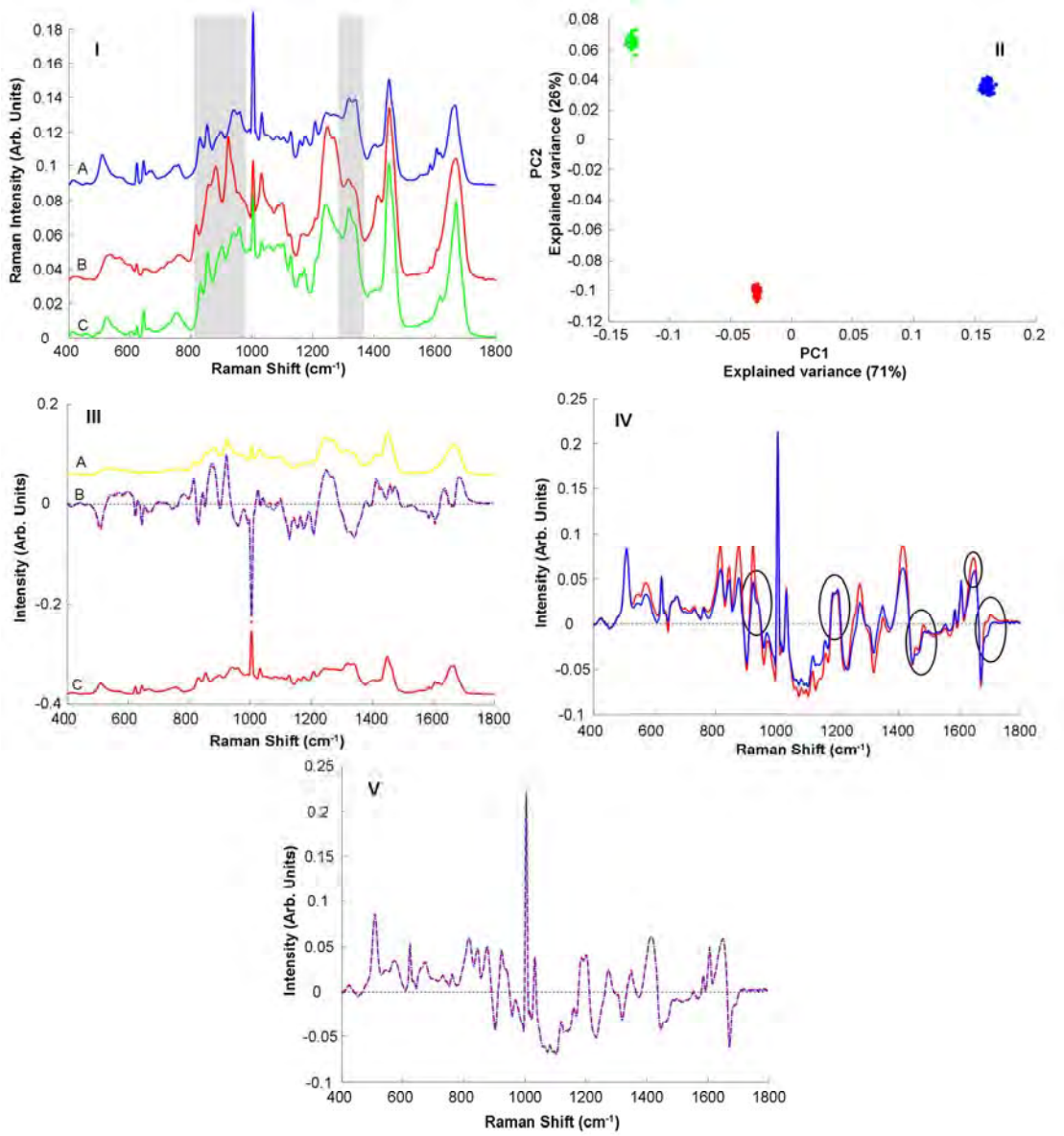

Figure 3. ***I:*** *Mean Raman spectra recorded from albumin (A) collagen (B) and histone (C) on CaF_2 windows.* ***II:*** *Scores plots of the first two principal components after PCA performed on Raman spectra recorded from albumin (green), histone (red)*

and collagen (blue). ***III:*** *Plot of the loadings of PC1 (blue dotted line) compared with the difference between the mean spectrum of collagen minus the mean spectrum of albumin (red dashed line). The loadings are compared with spectra recorded from collagen (A) and albumin (C), both offset for clarity.* ***IV:*** *Plot of the loadings of PC2 (blue line) compared with the difference between the mean spectrum of collagen minus the mean spectrum of albumin (red line).* ***V:*** *Plot of the loadings of PC2 (blue dotted line) compared with the simulated weighted sum of the different spectra according to their position on the scatter plot (red dashed line).*

In the case of PC2, however, the loading is not easily associated with the spectrum of any one of the individual components, although some of the stronger features can be identified, including the disulfide stretching at 509 cm^{-1}, the C–C twisting mode of Phe (proteins) at 623 cm^{-1}, the C-C stretching (proteins) at 816 cm^{-1}, the C–C aromatic ring stretching in Phe at 1005 cm^{-1} or the amide I band region around 1655 cm^{-1} [10, 45-47] (figure 3 IV). However, when summed according to their scores in the scores plot, (0.065 x (albumin spectrum) + 0.035 x (collagen spectrum)) – 0.1 x (histone spectrum), an excellent match to the loading of PC2 is achieved, as shown in figure 3 V. Notably, when weighted by the number of datapoints per group, (100), the respective sums of the negative and positive loadings each equals 1. Thus, although the principal components for more complex systems are not easily identifiable with the spectra of constituent biomolecular components of the sample, they are clearly determined by weighted sums of the spectra of those components.

3.4 Influence of imbalanced spectral datasets on PCA

As described in the previous section, the weighted sum of the spectra from the different components determines the profiles of the loadings when the number of spectra in each datasets is equal. However, when working on biological samples,

especially when the spectra are extracted from large maps, the numbers of spectra present in each group is usually determined by the relative sizes of the tissue or cellular regions, the laser spot size and sampling step size, and thus can be imbalanced. To illustrate the effect of such an imbalance in the datasets on the PCA, the numbers of the spectra composing the dataset recorded from the albumin was gradually reduced from 100 to 75, 50, 25 and finally 10 spectra, while the number of spectra of histone and collagen was kept constant. PCA was run for each dataset and the final plots are merged in figure 4 I, to visualise the evolution of the position of each cluster in the scatter plot.

The first observation is that although the number of spectra in only one group has been varied, the entire scores plot is affected and the positions of the data sets corresponding to the histone and collagen vary considerably. As the number of albumin spectra is systematically varied, they both move toward the zero position of PC2, approaching zero when the group corresponding to the albumin is reduced to 10 spectra. At this point they are predominantly discriminated by PC1 alone. Their positions also evolve according to PC1, their relative spacing increasing such that in the last data set of 10 albumin spectra, they are almost equidistant from the zero position of PC1. Simultaneously, the cluster corresponding to albumin is also significantly shifted when its number of spectra is reduced. When the groups have equal numbers, this cluster is positioned near the zero position of PC2, but gradually moves away from it as its constituent number is reduced. Although less pronounced, the cluster drifts to lower (negative) values of PC1, such that the position of the albumin data set and histone data sets is almost aligned with respect to PC1 when the number of spectra for albumin is reduced to 10 spectra.

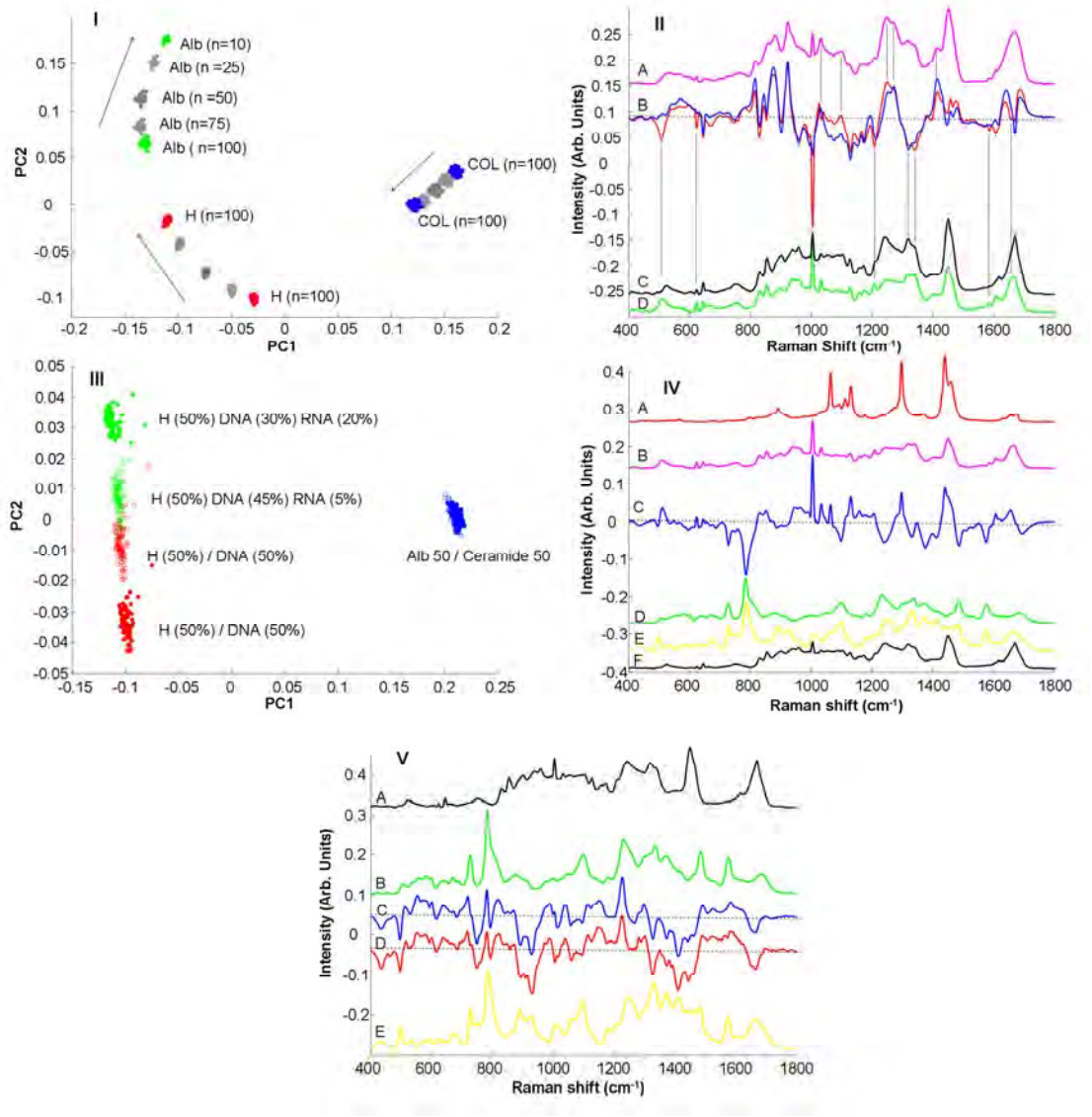

Figure 4. ***I:*** *Scores plot of the first two principal components after PCA performed on Raman spectra recorded from albumin (Alb - green), histone (H -red) and collagen (COL -blue). The data set corresponding to the albumin has been systematically reduced from 100 to 75, 50, 25 and 10 spectra.* ***II:*** *Plot of the loadings of PC1 (B) corresponding respectively to the first principal component resulting from the PCA analysis for Alb n = 100 (blue) and Alb n=10 (red). These loading have been compared with spectra recorded from collagen (A) and histone (C) and albumin (D).* ***III:*** *Scores plot of the first two principal components of PCA performed on Raman spectra recorded from protein mixtures. For all data sets, filled circles correspond to PCA with 50/30/20 histone/DNA/RNA, whereas unfilled squares correspond to PCA with 50/45/5 histone/DNA/RNA.* ***IV:*** *Plot of the loading of PC1 (C) of the PCA analysis for proteins mixture. This loading is compared with spectra recorded from different compounds such as ceramide (A) and albumin (B) and RNA (D), DNA (E) and histone (F).* ***V:*** *Plot of the loadings of PC2 of the PCA analysis for protein mixtures with the data set H 50% DNA 30% RNA 20% (C) and H 50% DNA 45% RNA 5% (D). The loading are compared to spectra recorded from different compounds such as histone (A) and RNA (B) and DNA (E).*

These modifications in the relative positions of the different groups reflect the variation in the relative contribution of each compound to the loadings corresponding to the different PCs. To illustrate these variations, the PC1 loadings calculated when the groups are balanced with the same number of spectra has been compared to the PC1 obtained when the albumin has been reduced to 10 spectra (figure 4 II).

As the group is located in the negative part of the scores plot with respect to PC1, the peaks related to the histones in the loadings are negative. The comparison between the two loadings clearly highlights the diminution of the specific features related to the albumin in the loading of PC1, when under-represented compared to the other groups. In quantitatively constructing the loadings from the spectra of the constituent spectra, a general formula can be applied:

$$PC = \Sigma_j S_j N_j \zeta_j(\nu) \qquad \text{Equation 1}$$

where j is the number of groups in the scatter plot, S_j is the average score of group j, N_j is the number of spectra contained in the group and ζ_j is the constituent spectrum as a function of frequency, ν. Thus, in the case where a group is represented by a lower number of constituent spectra, the average score is substantially reduced compared to the larger group, as is the contribution to the loading of the PC.

In the reduced dataset, however, specifics peaks corresponding to the histones are identifiable more easily, as they play a more significant role in the discrimination from the date set corresponding to the collagen. Moreover, the positive peaks in the loading of PC1 are also affected and specifics features corresponding to the collagen are also better defined when the groups are imbalanced. This is due to the increased differentiation of the two clusters according to PC1. Notably, therefore, in the reduced dataset, the molecular origin of the discrimination of the two majority clusters becomes clearer.

3.5 PCA of Complex Mixtures

The studies outlined above illustrate how PCA can differentiate structurally (and therefore spectrally) distinct and similar molecular species. In biological samples such as tissues or cells, however, these species are spatially mixed, and therefore within a typical sampling area are spectrally mixed. In order to simulate such mixed spectra, mixed data sets were constructed using weighted sums of the spectra of individual compounds. Initially, a dataset of the three groups was constructed: (i) 50/50 albumin/ceramide, (ii) 50/50 histone/DNA, and (iii) 50/30/20 histone/DNA/RNA. The mixtures thus mimic regions of equal protein content, but which are relatively rich in either lipidic (i) or nucleic acid (ii), or have differing nucleic acid (iii) content. In a simplistic approach, these can be proposed to represent (i) cytoplasmic/membrane, (ii) nuclear and (iii) nucleolar regions.

As expected, PCA effectively differentiates the three simulated spectral groups, as shown in figure 4 III (filled circles for all groups). PC1 represents 96% of the variance and largely differentiates the albumin/ceramide cluster from the histone/DNA and histone/DNA/RNA groups. The loading of PC1, shown in figure 4 IV, is dominated by albumin and ceramide spectral features in the positive sense, and by nucleic acid features in the negative sense. Again, the sharp features of the lipidic spectra contribute strongly loading. PC2 (figure 4 VC) differentiates the histone/DNA and histone/DNA/RNA groups and its loading is dominated by features of RNA (positive) and DNA (negative). As the relative contributions of DNA and RNA to group (iii) are varied from 50/30/20 to 50/45/5, the differentiation of groups (ii) and (iii) as described by PC2 is reduced considerably, although little change is seen with respect to the differentiation of clusters (i), (ii) and (iii), with respect to PC1 (Figure 4 III,

squares for all groups), consistent with the negligible contributions of nucleic acid peaks to the loading of PC1. The loading for PC2 is relatively unchanged, however, as shown in Figure 4 VD, due to the similarity of the spectra of DNA and RNA, although subtle differences can be seen.

3.6 Single cell analysis revisited

Analysis of the results of PCA on imbalanced datasets illustrates that the groups with smaller numbers are unrepresented in the loadings which differentiate them. In the case of the cellular analysis in section 3.1, the classifications as identified by K-cluster analysis result in three groupings corresponding to cytoplasm, nucleus and nucleoli. Within these, the number of spectra assigned to nucleus is the highest (n=143) followed by cytoplasm (n=126) and nucleoli (n=27). Thus, PC1, has strong contributions from lipids, and although some features of DNA are discernible, neither PC1 nor PC2 exhibit peaks which can unambiguously be ascribed to the nucleoli. The number of spectra per cluster cannot be determined *a priori* as it is dictated by the relative areas of the subcellular features. The imbalance can be overcome in a number of ways however: (a) reduction of the numbers of the larger clusters to the numbers of the smallest group (b) an increase in the number of the smaller clusters by for example duplication of randomly selected spectra, or means of a number of randomly selected spectra (c) increasing the relative strength of the spectra in the smaller data sets. For illustrative purposes, approach (b) has been chosen, and the numbers of spectra in the nucleoli (originally n=27) and cytoplasmic (originally n=120) groups was increased to the number of the nuclear group (n=143) by random selection and duplication of existent spectra.

Figure 5 I shows the resultant PCA scores plot, which is significantly different than that of the imbalanced groups (Figure 1 III). Although this should be a more accurate reflection of differentiation of the clusters, notably, the discrimination of the different clusters according to either PC1 or PC2 is no longer as clear. PC2 differentiates the nucleus from the nucleoli and cytoplasm clusters whereas in figure 1 III the nucleoli cluster was well discrimiated from the other two clusters. PC1 now does not provide a clear differentiation between clusters. Although the nucleoli seems to be well discriminated from the cytoplasm cluster, the spectra from the nucleus are unequally distributed about the origin, meaning these spectra will influence the loadings unevenly. By direct comparison, in figure 1 III, the nucleus and cytoplasm clusters are positioned on respectively the negative and positive sides of the scores plot according to PC1 which indicates a better differentiation.

The main limitation in this example is the presence of a high intra-group variability which can distort the information contained in the loadings. The nuclear spectra are distributed on the negative and positive side of PC1 and thus the contribution to the loading of PC1 is difficult to interpret. In terms of PC2, the intra-group variability of the cytoplasm spectra is higher than the difference existing between the nucleoli and nucleus clusters. Therefore the specificity of loading of PC2 is reduced. Therefore, as shown for the pure bio-molecules in section 3.2, a clearer interpretation of the loadings is provided by a direct comparison of individual clusters. Thus PCA of the balanced cytoplasmic and nuclear regions shows them to be exclusively differentiated by PC1 (data not shown).

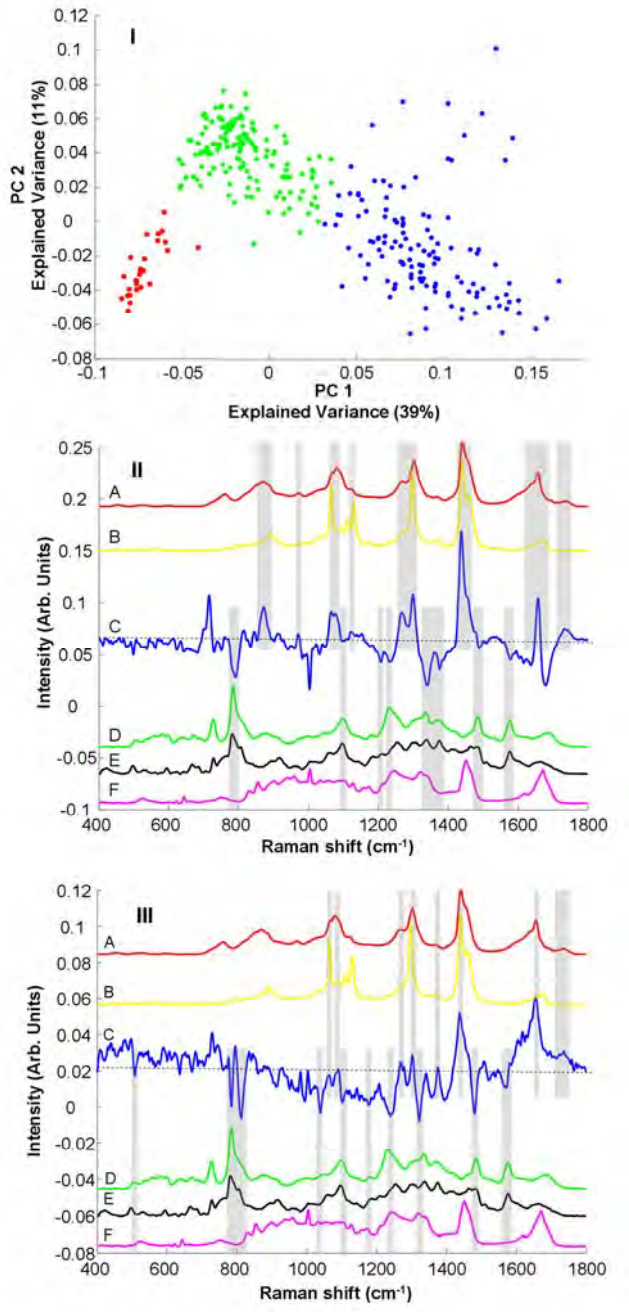

Figure 5. ***I:*** *Scores plot of the first two principal components of PCA performed on Raman spectra of A549 cells with numerically balanced groups. The spectra have been selected from the K-means clusters corresponding to the nucleus (green), nucleoli (red) and cytoplasm (blue).* ***II:*** *Plot of the loadings of PC1 of the PCA analysis based on 2 clusters (C). The loadings are compared with spectra recorded from different compounds such as L-phosphatidyl-ethanolamine (A), ceramide (B), RNA (D), DNA (E) and histone (F).* ***III:*** *Plot of the loadings of PC1 resulting from the PCA analysis based on 2 clusters (C). These loadings have been compared with spectra recorded from different compounds such as L-phosphatidyl-ethanolamine (A), ceramide (B), RNA (D), DNA (E) and histone (F).*

The loading is similar to that obtained for the imbalanced groups, which might be expected as the two original groups are not hugely different in size. Intragroup variability is still relatively large, but the fact that the number of spectra in each group is now balanced results in a balanced distribution of the two groups negatively or positively with respect to PC1, as seen with pure biomolecules, reducing the impact of the intra-group variability. Many different intense positive peaks can be identified in PC1 (figure 5 II), mostly related to the lipidic composition of the cytoplasm. The spectra of pure L-phosphatidyl-ethanolamine and ceramide have been plotted for comparison (figure 5 IIA and 5 IIB). The regions of interest have been highlighted in grey. Concerning the negative peaks, most of them can be found in the spectra of the RNA and DNA (figure 5 IID and 5 IIE). However, no match is found to the histone spectrum (figure 5 IIF) which is expected as this protein is specific to the nucleoli.

For the case of the nucleus and nucleoli only, PCA for the balanced groups also shows a clear differentiation according to PC1 (data not shown). The loading is now considerably less noisy and more detailed than that for the imbalanced groups (Figure 5 III). The positive peaks can be partially attributed to the presence of lipid in the structure of the nucleus (figure 5 IIIA and 5 IIIB). The negative peaks are related to the composition of the nucleolus and peaks specific for DNA and RNA can be found (figure 5 IIID and 5 IIIE). Most notably, specific features of histone are now discernable (figure 5 IIIF), such as the C-H in-plane Phe located at 1033 cm^{-1}, the ν C-C, ν C- N and ν C-O band at 1103 cm^{-1}, the C–O stretch and COH bend at 1173 cm^{-1}, the amide III (β-sheet, protein) at 1247 cm^{-1} and the C-H at 1321 cm^{-1} [43, 44, 48].

4. Conclusion

Although Raman spectroscopy is a powerful tool for the analysis of biochemical content at a sub-cellular level, analysis of the spatial distribution of spectral signatures requires complex multivariate analytical techniques. K-means clustering analysis provides an elegant map of spectrally differentiated regions, but little basis for the underlying biochemical differences between the identified regions. In conjunction with K-means cluster analysis, PCA loadings in principle provide this information, but they can be a complex mixture of many contributing molecular components. The loadings are strongly weighted by sharply varying features, such as those of lipids or nucleic acids. They are also weighted by the number of spectra within the individual clusters and thus molecular contributions to lower number clusters can be misrepresented. Balanced clusters provide the best representation of relative contributions to loadings, but ultimately, the best results are provided by a pair wise analysis of identified sub cellular regions, which yield a clearer representation of the underlying biochemical differences.

Acknowledgements

This research was supported by the National Biophotonics and Imaging Platform (NBIP) Ireland funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycle 4, co-funded by the Irish Government and the European Union Structural fund.

REFERENCES

1. Bonnier, F., et al., *Detection of pathological aortic tissues by infrared multispectral imaging and chemometrics.* Analyst, 2008. **133**(6): p. 784-90.
2. Lyng, F.M., et al., *Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool.* Exp Mol Pathol, 2007. **82**(2): p. 121-9.
3. Tfayli, A., et al., *Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy.* Biochim Biophys Acta, 2005. **1724**(3): p. 262-9.
4. Larraona-Puy, M., et al., *Development of Raman microspectroscopy for automated detection and imaging of basal cell carcinoma.* J Biomed Opt, 2009. **14**(5): p. 054031.
5. Moss, D., *Biomedical Applications of Synchrotron Infrared Microspectroscopy: A practical Approach*, ed. N.W. Barnett. 2011, Cambridge: RCS Publishing.
6. Sebiskveradze, D., et al., *Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections.* Lab Invest, 2011. **91**((5)): p. 799-811.
7. Wolthuis, R., et al., *IR spectral imaging for histopathological characterization of xenografted human colon carcinomas.* Anal Chem, 2008. **80**(22): p. 8461-9.
8. Nawaz, H., et al., *Evaluation of the potential of Raman microspectroscopy for prediction of chemotherapeutic response to cisplatin in lung adenocarcinoma.* Analyst, 2010. **135**(12): p. 3070-6.
9. Draux, F., et al., *IR spectroscopy reveals effect of non-cytotoxic doses of anti-tumour drug on cancer cells.* Anal Bioanal Chem, 2009. **395**(7): p. 2293-301.
10. Ostrowska, K.M., et al., *Investigation of the influence of high-risk human papillomavirus on the biochemical composition of cervical cancer cells using vibrational spectroscopy.* Analyst, 2010. **135**(12): p. 3087-93.
11. Krafft, C., et al., *Raman mapping and FTIR imaging of lung tissue: congenital cystic adenomatoid malformation.* Analyst, 2008. **133**(3): p. 361-71.
12. Meade, A.D., H.J. Byrne, and F.M. Lyng, *Spectroscopic and chemometric approaches to radiobiological analyses.* Mutat Res, 2010. **704**(1-3): p. 108-14.
13. Knief, P., et al., *Raman spectroscopy--a potential platform for the rapid measurement of carbon nanotube-induced cytotoxicity.* Analyst, 2009. **134**(6): p. 1182-91.
14. Barber, J.L., et al., *Low-dose treatment with polybrominated diphenyl ethers (PBDEs) induce altered characteristics in MCF-7 cells.* Mutagenesis, 2006. **21**(5): p. 351-60.
15. Ling, J., et al., *Direct Raman imaging techniques for study of the subcellular distribution of a drug.* Appl Opt, 2002. **41**(28): p. 6006-17.
16. Rubin, S., et al., *Analysis of structural changes in normal and aneurismal human aortic tissues using FTIR microscopy.* Biopolymers, 2008. **89**(2): p. 160-9.
17. Amharref, N., et al., *Discriminating healthy from tumor and necrosis tissue in rat brain tissue samples by Raman spectral imaging.* Biochim Biophys Acta, 2007. **1768**(10): p. 2605-15.

18. Krishna, C.M., et al., *FTIR and Raman microspectroscopy of normal, benign, and malignant formalin-fixed ovarian tissues.* Anal Bioanal Chem, 2007. **387**(5): p. 1649-56.
19. Bonnier, F., et al., *Imaging live cells grown on a three dimensional collagen matrix using Raman microspectroscopy.* Analyst, 2010. **135**(12): p. 3169-77.
20. Miljkovic, M., et al., *Label-free imaging of human cells: algorithms for image reconstruction of Raman hyperspectral datasets.* Analyst, 2010. **135**(8): p. 2002-13.
21. Mohlenhoff, B., et al., *Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy.* Biophys J, 2005. **88**(5): p. 3635-40.
22. Bassan, P., et al., *Resonant Mie scattering in infrared spectroscopy of biological materials--understanding the 'dispersion artefact'.* Analyst, 2009. **134**(8): p. 1586-93.
23. Romeo, M.J. and M. Diem, *Infrared spectral imaging of lymph nodes: Strategies for analysis and artifact reduction.* Vib Spectrosc, 2005. **38**(1-2): p. 115-119.
24. Bassan, P., et al., *Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples.* Analyst, 2010. **135**(2): p. 268-77.
25. Thennadil, S.N., H. Martens, and A. Kohler, *Physics-based multiplicative scatter correction approaches for improving the performance of calibration models.* Appl Spectrosc, 2006. **60**(3): p. 315-21.
26. Lieber, C.A. and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra.* Appl Spectrosc, 2003. **57**(11): p. 1363-7.
27. Hanlon, E.B., et al., *Prospects for in vivo Raman spectroscopy.* Phys Med Biol, 2000. **45**(2): p. R1-59.
28. Beier, B.D. and A.J. Berger, *Method for automated background subtraction from Raman spectra containing known contaminants.* Analyst, 2009. **134**(6): p. 1198-202.
29. Bonnier, F., et al., *In vitro analysis of immersed human tissues by Ramanmicrospectroscopy.* Journal of Raman spectroscopy, 2010. **DOI 10.1002/jrs.2825**.
30. Varmuza, K., *Introduction to multivariate statistical analysis in chemometrics.* Taylor & Francis Group ed. 2009, New york: CRC Press. ch.3, pp.59-102.
31. Ly, E., et al., *Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition.* Analyst, 2009. **134**(6): p. 1208-14.
32. Ly, E., et al., *Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies.* Analyst, 2008. **133**(2): p. 197-205.
33. Oshima, Y., et al., *Discrimination analysis of human lung cancer cells associated with histological type and malignancy using Raman spectroscopy.* J Biomed Opt. **15**(1): p. 017009.
34. Mariani, M.M., et al., *Impact of fixation on in vitro cell culture lines monitored with Raman spectroscopy.* Analyst, 2009. **134**(6): p. 1154-61.
35. Ostrowska, K.M., et al., *Correlation of p16(INK4A) expression and HPV copy number with cellular FTIR spectroscopic signatures of cervical cancer cells.* Analyst, 2011. **136**(7): p. 1365-73.

36. Taleb, A., et al., *Raman microscopy for the chemometric analysis of tumor cells.* J Phys Chem B, 2006. **110**(39): p. 19625-31.
37. Draux, F., et al., *Raman spectral imaging of single cancer cells: probing the impact of sample fixation methods.* Anal Bioanal Chem, 2010. **397**(7): p. 2727-37.
38. Meade, A.D., et al., *Studies of chemical fixation effects in human cell lines using Raman microspectroscopy.* Anal Bioanal Chem, 2010. **396**(5): p. 1781-91.
39. Korenius, T., J. Laurikkala, and M. Juhola, *On principal component analysis, cosine and Euclidean measures in information retrieval.* Information Sciences, 2007. **177**(22): p. 4893-4905
40. German, M.J., et al., *Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell.* Biophys J, 2006. **90**(10): p. 3783-95.
41. Martin, F.L., et al., *Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample.* J Comput Biol, 2007. **14**(9): p. 1176-84.
42. Kelly, J.G., et al., *Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers.* J Proteome Res, 2011. **10**(4): p. 1437-48.
43. Meade, A.D., et al., *Growth substrate induced functional changes elucidated by FTIR and Raman spectroscopy in in-vitro cultured human keratinocytes.* Anal Bioanal Chem, 2007. **387**(5): p. 1717-28.
44. Notingher, I., *Raman Spectroscopy Cell-based Biosensors.* sensors, 2007. **7**: p. 1343-1358.
45. Tfayli, A., et al., *Molecular characterization of reconstructed skin model by Raman microspectroscopy: comparison with excised human skin.* Biopolymers, 2007. **87**(4): p. 261-74.
46. Ivanov, A.I., et al., *INFRARED AND RAMAN SPECTROSCOPIC OF THE STRUCTURE OF HUMAN SERUM UNDER VARIOUS LIGAND LOADS.* Journal of Applied Spectroscopy, 1994. **60**(5-6).
47. Frushour, B.G. and J.L. Koenig, *Raman scattering of collagen, gelatin, and elastin.* Biopolymers, 1975. **14**(2): p. 379-91.
48. Maiti, N.C., et al., *Raman spectroscopic characterization of secondary structure in natively unfolded proteins: alpha-synuclein.* J Am Chem Soc, 2004. **126**(8): p. 2399-408.