Theor Chem Acc (2011) 130:1249–1260
DOI 10.1007/s00214-011-0957-1

REGULAR ARTICLE

Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging

Martin Hedegaard · Christian Matthäus · Søren Hassing · Christoph Krafft · Max Diem · Jürgen Popp

Received: 17 November 2010 / Accepted: 7 May 2011 / Published online: 7 June 2011

Abstract A detailed comparison of six multivariate algorithms is presented to analyze and generate Raman microscopic images that consist of a large number of individual spectra. This includes the segmentation algorithms for hierarchical cluster analysis, fuzzy *C*-means cluster analysis, and *k*-means cluster analysis and the spectral unmixing techniques for principal component analysis and vertex component analysis (VCA). All algorithms are reviewed and compared. Furthermore, comparisons are made to the new approach N-FINDR. In contrast to the related VCA approach, the used implementation of N-FINDR searches for the original input spectrum from the non-dimension reduced input matrix and sets it as the endmember signature. The algorithms were applied to hyperspectral data from a Raman image of a single cell. This data set was acquired by collecting individual spectra in a raster pattern using a 0.5-µm step size via a commercial Raman microspectrometer. The results were also compared with a fluorescence staining of the cell including its mitochondrial distribution. The ability of each algorithm to extract chemical and spatial information of subcellular components in the cell is discussed together with advantages and disadvantages.

Keywords Chemometrics · Raman spectroscopy · Image processing · Hyperspectral data

1 Introduction

The analytical power of single-point Raman spectroscopy has been applied in biological and medical applications, and in the last decade, microscopic Raman imaging is becoming more and more popular [1]. The main inherent advantage of Raman spectroscopy lies in the label-free nature and high information content of the data. Recent technical progress has led to the development of sensitive and accurate microscopic systems for the collection of Raman images. Using high-magnification microscope objectives, the excitation laser can be focused to a diffraction limited spot of less than 1 µm diameter from which the Raman scattered light is collected. As most eukaryotic cells are larger and contain distinct subcellular compartments, a Raman image can be significantly more informative than a single Raman spectrum. Depending on the type and size of a given sample, image acquisition times range from minutes to hours. In addition, non-linear spectroscopic techniques such as coherent anti-Stokes Raman scattering and stimulated Raman scattering recently emerged to bring the acquisition times down to seconds per image [2, 3].

Raman microscopes collect a full spectrum for each image pixel, producing a hyperspectral data set with an enormous amount of chemical information; thus, the need for efficient ways for data analysis arises. The most widely

Dedicated to Professor Akira Imamura on the occasion of his 77th birthday and published as part of the Imamura Festschrift Issue.

M. Hedegaard · S. Hassing
Institute of Technology and Innovation,
University of Southern Denmark, Odense, Denmark

C. Matthäus · C. Krafft (✉) · J. Popp
Institute of Photonic Technology, Jena, Germany
e-mail: Christoph.Krafft@ipht-jena.de

M. Diem
Department of Chemistry and Chemical Biology,
Northeastern University, Boston, MA, USA

J. Popp
Institute of Physical Chemistry and Abbe Center of Photonics,
Friedrich Schiller University of Jena, Jena, Germany

accepted way and scope of the current paper are statistical methods to model the sample (here: a single cell) using the observed Raman spectra. However, it cannot be excluded that in the future first principles, semi-empirical, and empirical force field and molecular mechanics/molecular dynamics simulations of the spectra and their changes in cells, tissues, and the organism as a whole might contribute to a better understanding of the wealth of information. For further reading, a textbook is recommended as an introduction to vibrational spectroscopy with a focus on biomolecules, vibrational circular dichroism, and phase transitions (non-polar to polar solvents, solid to crystalline phase) [4].

A fundamental challenge is that the Raman spectrum of biological macromolecules depends on the structure and interaction of the monomeric units. For example, the spectra of proteins and nucleic acids constitute a complex superposition of contributions from amino acids, nucleotides, and backbone geometry. Furthermore, each Raman spectrum of a cell contains spectral contributions of numerous molecules within the laser focus. In the context of disease recognition by Raman spectroscopy, unsupervised and supervised algorithms have been described [5–7]. Rather than relying on single band intensities, diverse multivariate algorithms have been developed that use the entire spectrum for analysis. These methods can be roughly divided into factor methods that extract spectral information, cluster methods for partitioning a data set, and classification methods for modeling group differences. The large number of algorithms—implemented in commercial software as well as in user-written programs—for analyzing the hyperspectral data has motivated us to perform a comparison of advantages and disadvantages. Examples of hyperspectral Raman data analysis include cluster algorithms where the clusters and cluster memberships are defined based on similarity of the spectra. Different implementations are hierarchical cluster analysis (HCA), fuzzy *C*-means (FCM), and *k*-means cluster (KMC) analysis. The segmentation can be "crisp" (implying that a spectrum is assigned to one cluster only) or "soft", meaning that each spectrum can be assigned to more than one cluster with probability values between 1 and 0. False color images can be generated based on the cluster membership assigned to each pixel in the data set, or by plotting the probability values of each cluster. The biochemical content of each cluster is usually analyzed using the average cluster spectra.

Recently, spectral unmixing algorithms have been applied to hyperspectral Raman data sets [8–11]. These algorithms were first introduced to visible reflectance spectra from satellite detection systems [12–14]. Most of them are based on the following assumptions: (1) Pure pixels are available in the data set; their spectral signature is then defined to be an endmember. (2) All other pixels in the data set can be described as a linear combination of the pure pixel endmembers. The amount each endmember contributes to a pixel is described by the abundance fraction of that endmember. In general, a spectral unmixing algorithm finds the most extreme spectra and defines them as endmembers. If pure spectra are not present in the data set, mixed pixels can also be used as endmembers in algorithms like vertex component analysis (VCA) and N-FINDR [14]. Other algorithms such as iterated constrained endmembers (ICE) try to compensate for this limitation by combining spectral unmixing algorithms with multivariate curve resolution methods [15]. The abundance fraction is calculated by fitting the endmembers to each spectrum in the data set. When the abundance fraction is plotted versus the pixel coordinates, the contribution of each endmember can be visualized.

In this paper, we compare the chemical and spatial information that can be obtained from a Raman hyperspectral data set using segmentation algorithms and comparing them to spectral unmixing techniques. To investigate the properties of the algorithms, we have chosen a Raman data set of a cell grown in vitro. The resulting Raman spectral images were compared with a fluorescence image after specific staining of mitochondria. The Raman spectral signatures of subcellular features that were obtained from each algorithm were compared with each other and with those reported in reference [16]. Another summary of multivariate algorithms for the analysis of Raman hyperspectral data sets was recently published [8].

1.1 Data handling and preprocessing

Raman microscopic imaging is performed using a Raman spectrometer connected to a confocal microscope with a motorized stage or a scanning mirror or by global illumination methods. In all cases, a three-dimensional data set of dimension $N_x \times N_y \times N_R$ is collected, where $N_x \times N_y$ denotes the number of pixels in the x and y direction and N_R the number of data points in each spectrum. Modern confocal microscopes allow measuring Raman signals from a diffraction limited volume in the ≈ 0.1 cubic micrometer range and, consequently, achieving a very high spatial resolution at the sample.

As all the data analysis algorithms described in the following sections are applied to two-dimensional data matrices and the measured data are three-dimensional hyperspectral Raman data sets, the data set needs to be reshaped into the two-dimensional matrix of dimension $N_xN_y \times N_R$ with N_xN_y samples and N_R data points in each spectrum. All methods give one matrix in the case of univariate methods and crisp clustering or more matrices in the case of FCM and spectral unmixing.

Before applying any of the imaging algorithms, a quality test was performed to remove spectra from the data set with signal intensities below a given threshold. The intensity of CH valence vibrations from 2,800 to 3,000 cm^{-1} was used as criterion. The average of the removed spectra was then subtracted from the spectra remaining in the data set, effectively removing constant signals originating from substrate, medium, and optical elements in the light path and causing an offset. The spectral contributions of the cell are better identified without this constant offset. However, this step is not essential and, in principle, the algorithms will also work without this quality test.

In general, the statistical data evaluations described here compare each individual spectrum with all other spectra within a given data set. The spectra are then sorted according to their similarities or dissimilarities, without any external input. For complex data sets, this unsupervised approach is advantageous, since the spectral features of the constituents are not always known. Clustering methods relate the spectra by their variations and group them together based on similarity. Unmixing algorithms decompose a given data set into a new basis set based on the greatest variance or dissimilarity. Each spectrum can then be reconstructed as a linear combination of the new basis vectors. The individual unmixing methods basically differ by the type of constraints introduced to find the new basis. All data analyses were performed in MatLab (The Mathworks, Natick, MA) using algorithms written in house and implemented as described in the following sections. A selection of similar tools is available in a commercial software package Cytospec (http://www.cytospec.com) and in an open-source software package hyperSpec for programming language R (http://hyperspec.r-forge.r-project.org).

1.2 Hierarchical cluster analysis

Hierarchical cluster analysis (HCA) is a frequently used algorithm for generating false color images from hyperspectral data sets. HCA calculates the symmetrical distance matrix of size $n \times n$ between n spectra as a measure of their pairwise similarity. The algorithm then searches for the minimum distance, collects the two most similar spectra into a first cluster, and recalculates spectral differences between all remaining spectra and the first cluster. In the next step, the algorithm performs a new search for the most similar objects which now can be spectra or clusters. This iterative process is repeated $n - 1$ times until all spectra have been merged into one cluster. The most widely used implementation of HCA applies Euclidian distances and Ward's algorithm for clustering. The algorithm is agglomerative, employing a bottom up approach, grouping spectra based on the intra-spectral Euclidian distance, and linking them using the Ward's criterion of minimizing loss of information associated to each group [17]. The distances between the spectra can be visualized using tree-like, two-dimensional dendrograms in which one axis refers to the reduction in clusters with increasing number of iterations and the other axis to the respective spectral distances. This algorithm enables one to examine the clustering arrangements with different numbers of groups and to select a scheme that may be more easily interpretable. The cluster membership map at a defined distance is a vector of dimension $N_xN_y \times 1$ where each spectrum is assigned a number from 1 to p. False color images can be constructed assigning each cluster a color and the result refolded, thereby creating an image matrix of dimension $N_x \times N_y \times 1$. One inherent disadvantage of HCA is that it is computationally demanding compared with other clustering algorithms. This can however be greatly improved by a parallel implementation of HCA, e.g. using modern graphics cards [7].

1.3 *k*-means cluster analysis

The *k*-means cluster (KMC) analysis belongs to the partitioning methods, and the basic principles were explained by MacQueen [18]. In general, clustering is the partitioning of a data set into clusters so that the differences between the data within each cluster are minimized and the differences between clusters are maximized according to some defined distance measure. Similar to the HCA algorithm, different distance measures can be used in KMC. The Euclidian distance usually works satisfactorily for the Raman data sets. The algorithm first selects K random spectra as starting centroids. Centroids denote the center or mean of the clusters. Then, the distances are calculated between every spectrum and these centroids. Subsequently, each spectrum is assigned to a cluster whose centroid is nearest. When all spectra have been assigned to the K centroids, a new set of centroids is calculated based on the mean of the spectra associated with each centroid. This process is then reiterated until the assignment does not change and the incremental improvement is below a given threshold. An alternative stop criterion is the maximum number of iterations. Mathematically, the KMC algorithm relates the N_xN_y spectra with running sample index i into K clusters by minimizing the cluster variance with respect to the means $\{m_1,\ldots,m_K\}$. The cluster variance is then minimized by assigning each spectrum to the nearest cluster mean m_k, by calculating:

$$C(i) = \underset{1<k<K}{\arg\min} ||x_i - m_k||.$$

The result of this algorithm is a vector of dimension $N_xN_y \times 1$ where each spectrum is assigned a number from 1 to k. The membership from KMC analysis can be plotted as false color images of dimension $N_x \times N_y \times 1$ where each cluster is represented by a color. KMC analysis can be improved by seeding such as trying out multiple starting points and choosing the clustering with lowest cost or using starting points derived by another method such as HCA.

1.4 Fuzzy *C*-means clustering

The principle of fuzzy C-means clustering (FCM) is closely related to KMC as it also assigns each sample to C centroids. However, instead of crisp partitioning as in KMC, the output of FCM is a membership function describing how similar the sample is to that particular centroid based on distance to the centroids. The degree of membership can vary between 1 and 0 with 1 being identical to the cluster center and 0 being no class membership. FCM cluster imaging uses a fuzzy iterative algorithm that was introduced by Bedzek et al. [19, 20] to calculate the membership degree for each spectrum resulting in a vector of dimension $N_xN_y \times C$ since each spectrum has C membership values. The coefficients of the membership matrix U are defined by:

$$u_{iN_S} = \frac{1}{\sum_{c=1}^{C}\left(\frac{d_{iN_S}}{d_{cN_S}}\right)^{2/(m-1)}},$$

where u_{iN_S} is the membership of sample N_s in one cluster, d_{iN_S} and d_{cN_S} are the distances to the i and cth cluster centers, and m is the fuzziness factor between 1 and ∞. If m approaches 1, the algorithm resembles the KMC algorithm, and if m approaches infinity, every cluster belongs equally to all centroids. The membership functions are normalized in a way that all sample membership functions add up to one. When the matrix **U** is found, it can be reshaped into C images of dimensions $N_x \times N_y \times 1$. The chemical content of the clusters can be analyzed by investigating the C centroid spectra [21].

1.5 Principal component analysis

Principal component analysis (PCA) decomposes a data set into a bilinear model of linear independent variables, the so-called principal components. Pearson initially proposed the idea in 1901 [22] to explain the variation in a data set with only a subset of the variables of the original data set. In the version of PCA used here, the algorithm starts with finding the vector best describing most of the variation of the data set, referred to as the first loading vector, and projecting each spectrum onto the loading vector. The score is defined as the projection of the sample vector onto the loading vector. The product of loading vector and score is defined as the first principal component. The next principal components are the vectors that describe the next largest variation not accounted for by previous components until all variations are explained. Mathematically, this can be expressed as decomposition by:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E},$$

with $\mathbf{X}$ having $N_xN_y \times N_R$ dimensions. The scores $\mathbf{T}$ are of $N_xN_y \times p$ dimensions with p as the number of principal components. The loadings $\mathbf{P}$ have the dimension $N_R \times p$ and the residual matrix E that contains the variation not explained by the principal components is of $N_xN_y \times N_R$ dimensions. The matrix $\mathbf{T}$ contains p $N_xN_y \times 1$ vectors. To show the spatial distribution of the scores, each vector can be reshaped into p images of dimension $N_x \times N_y \times 1$. These are usually referred to as abundance images as they represent the abundance of each loading vector for each pixel. The abundance images are then plotted in a color-scaled image. By assigning each abundance image an individual color channel, several images can be combined by overlaying them.

1.6 Vertex component analysis

Vertex component analysis (VCA) is an unsupervised spectral unmixing algorithm. VCA is based on the assumption that each spectrum can be seen as an n-dimensional vector spanning an n-dimensional Euclidian space, where each measured wavenumber is assigned to a coordinate within that space. When all spectra are represented in the n-dimensional space, they span a simplex. A simplex is defined as the n-dimensional generalization of a triangle or tetrahedron to an arbitrary dimension, i.e. the n-dimensional simplex has $n + 1$ corners. VCA utilizes the fact that each spectral vector belongs to a simplex and that the vertices of this simplex correspond to the most extreme spectra in the data set and thereby representing the purest spectra. Mathematically, VCA decomposes each spectrum in the data matrix into a linear combination of endmember spectra and a set of abundances for each spectrum describing how similar the spectrum is to the endmember. This can be written as

$$\mathbf{X} = \alpha\mathbf{M}^{\mathrm{T}} + \mathbf{E}.$$

$\mathbf{X}$ still has the dimension $N_xN_y \times N_R$. The abundance matrix $\boldsymbol{\alpha}$ is of the dimension $N_xN_y \times p$ for a model with p endmembers, and the endmember spectra $\mathbf{M}$ have the dimension $N_R \times p$. The variance not explained by the abundance matrix and endmember signatures is called $\mathbf{E}$ and has the dimension $N_xN_y \times N_R$. The VCA algorithm defines the vertex vectors k by finding the spectral vectors with the largest length:

$$k = \arg \max_{j=1\ldots N_X N_Y} ||y(j)||,$$

with ||y|| being the norm of the spectral vector. The next endmembers are found by iteratively calculating the orthogonal subspace to the endmembers already determined using orthogonal subspace projection. This process is continued until the desired number of endmembers is extracted.

Two different implementations of VCA have been described. The first one uses the complete data set to span the simplex [23]. The other one applies a dimension reduction, in most cases a PCA, and uses p principal components to describe the variation and thereby to reduce the problem from dimension N_R to dimension N_p [12]. The dimension reduction greatly reduces the computational complexity of the problem. The PCA approach also has the advantage that the endmembers have an improved signal-to-noise ratio because higher principal components that are dominated by noise are omitted. To visualize the abundance of each endmember spectrum, each abundance vector of dimension $N_x N_y \times 1$ in the matrix $\boldsymbol{\alpha}$ is reshaped into $N_x \times N_y \times 1$ images similar to the PCA images.

1.7 N-FINDR

Similar to VCA, N-FINDR also tries to find endmembers corresponding to pure spectra. The N-FINDR algorithm was proposed by Winter [14] to map minerals in geological images. However, the approach differs as it maximizes the volume of the *n*-simplex spanned by the spectra. The volume is determined by defining a matrix $\mathbf{M}_V$ containing the endmember spectra and augmented with a row of ones. Using the fact that the volume V is proportional to the determinant of $\mathbf{M}_V$,

$$V(\mathbf{M}_v) = \frac{1}{(N_R - 1)!} \mathrm{abs}(|\mathbf{M}_v|).$$

The algorithm finds the spectral vectors spanning the largest volume. When the positions of the endmember signature spectra have been determined, the used implementation of N-FINDR searches for the original input spectrum from the non-dimension reduced input matrix and sets the original spectrum as the endmember signature. Each pixel is then described as a linear combination of these endmember spectra using non-negativity constrained least squares fitting. The output format of the N-FINDR is analogous to the VCA algorithm and also results in the matrix being decomposed as:

$$\mathbf{X} = \alpha \mathbf{M}^{\mathrm{T}} + \mathbf{E}.$$

Therefore, the visualization of the data is identical.

2 Materials an methods

2.1 Cell culture

Human HeLa cells (cell line CCL-2, ATCC, Manassas, VA) were grown in 75-cm^3 culture flasks (Fisher Scientific, Loughborough, Leicestershire, UK) with 7 mL of Dulbecco's modified Eagle's medium (ATCC) and 10% fetal bovine serum (ATCC) at 37 °C and 5% CO_2. Cells were seeded onto and allowed to attach to polished calcium fluoride windows (Wilmad LabGlass, Buena, NJ), which were chosen to avoid background scattering that is observed from regular glass windows. The windows were removed from the culture medium after 12–24 h, and the cells were fixed in a 10% phosphate-buffered formalin solution (Sigma–Aldrich, St. Louis, MO) and washed in phosphate-buffered saline. For Raman and fluorescence measurements, the windows with the attached and fixed HeLa cells were submerged in buffer solution during the measurement. The fixed cells formed a subconfluent layer and were completely immobilized on the substrates [16].

2.2 Raman data acquisition

Raman spectra were acquired using a WITec (Ulm, Germany) Model CRM 2000 confocal Raman microscope. Excitation at 488 nm (~5 mW at the sample) was provided by an air-cooled Ar-ion laser (Melles Griot, Carlsbad, CA; Model 532). The exciting laser radiation was coupled into a Zeiss (Jena, Germany) microscope through a wavelength-specific single-mode optical fiber. The incident laser beam was collimated via an achromatic lens and passed through a holographic band-pass filter before it was focused onto the sample through the microscope objective. A Nikon (Japan) Fluor (60×/NA 1.00, working distance 2.0 mm) water immersion objective was used. The diameter of the laser focus was calculated as 621 nm according to the formula $4 \cdot \lambda/\pi \cdot NA$, which is above the Abbe's limit of the lateral resolution of $0.621 \cdot \lambda/NA = 303\,\mathrm{nm}$.

The sample was placed on a piezo-electrically driven microscope scanning stage with an *x*, *y* resolution of ~3 nm and a repeatability of 65 nm and z resolution of ~0.3 nm and repeatability of 62 nm. The sample was scanned through the laser focus in a raster pattern at a constant stage speed of fractions of a micrometer per second. Dissipation of laser-induced heat to the surrounding aqueous medium and continuous motion prevented sample degradation in the focal point of the laser beam. Spectra were collected at a 0.5-µm grid with a dwell time of 0.5 s. The excellent confocality of the Raman microscope was demonstrated in lateral and axial images of lipid droplets within single cells [24].

2.3 Fluorescence measurements

The green fluorescence of the Mitotracker® stain was detected using the confocal Raman setup described above using 488-nm excitation. To compare the mitochondrial distribution in Raman and fluorescence images, the Raman data were acquired first. Then, Mitotracker® stain was carefully added, and the fluorescence emission of the cell was rescanned at significantly lower laser power and a dwell time of 0.2 s per data point [9].

3 Results

To compare the different algorithms, a high-resolution Raman image of one single HeLa cell was analyzed. The data set has the dimensions of 120 × 120 × 1,024 pixels or data points and corresponds to a total area of 60 × 60 μm^2. A white light image of the cell is inserted in Fig. 1a. An intensity plot of the 2,935 cm^{-1} band is shown in Fig. 1b. Highest intensities are found for the nucleoli in the nucleus and for the perinuclear region. Immediately after the Raman measurement, the cell was stained with the Mitotracter® fluorescence stain and the image in Fig. 1c was obtained. The bright green fluorescent areas indicate high mitochondrial content. The highest fluorescence intensity was observed in the perinuclear region which correlates to cytoplasmic regions with high Raman intensities. The white light image of cell in Fig. 1a shows only weak contrast.

3.1 Hierarchical cluster analysis

To identify subcellular features, such as the nucleus, nucleoli, and cytoplasm, HCA was performed in the whole spectral region. The segmentation of the data set into eight clusters is displayed in Fig. 1d. The image clearly distinguishes between different regions in the cytoplasm (light green and green), regions that correlate with a high concentration of mitochondria (yellow, orange, red, and dark red) and the nucleus as well as nucleoli (light blue and blue). As the spectra are separated on the basis of Raman bands and their intensity variations, the average cluster spectra contain valuable information of the underlying biochemical differences. Mean spectra representing three clusters are shown in Fig. 2. These spectra correspond to the classes with the main spectral differences. Trace 2 represents the nuclear region (light blue in Fig. 1d), trace 3 the perinuclear regions (orange in Fig. 1d), and trace 4 the outer regions of the cytoplasm (green in Fig. 1d). In the perinuclear region, membraneous cell organelles such as the rough endoplasmic reticulum with ribosomes, the Golgi apparatus, and also mitochondria are found. All three regions show distinct protein bands at 1,655 cm^{-1} (amide I), the extended amide III region between 1,230 and 1,330 cm^{-1}, and the phenylalanine (all symmetrical ring breathing) band at 1,002 cm^{-1}. Apart from the contributions of proteins, differences in lipid contributions are evident. Spectral differences due to lipids are visible at the shoulder between 2,850 and 2,950 cm^{-1}, which are more pronounced in cytoplasm and perinuclear regions compared with the nucleus. Various organelles contribute with signals from cholesterol, phospholipids, and fatty acids from membrane-rich structures. Such organelles could be for instance the Golgi, lysosomes, mitochondria, intracellular vesicles, and endoplasmic reticulum that are, however, not completely resolved in the example here.

As the protein bands dominate in all Raman spectra of the cell, a difference spectrum (cytoplasm minus nucleus) was calculated to better visualize lipid and DNA bands as positive and negative differences, respectively. Spectral contributions from the phospholipids exhibit distinct features in several regions in the spectrum. The long aliphatic side-chains give rise to C–H stretches between 2,850 and 2,950 cm^{-1}. $CH_{2/3}$ deformations are found between 1,290 and 1,465 cm^{-1}, and the hydrophilic head groups between 700 and 900 cm^{-1}. A prominent band at 715 cm^{-1} is assigned to choline in lipids phosphatidylcholine and sphingomyelin. In general, it is observed that cytoplasm and perinuclear regions show relatively high concentrations of lipids opposed to the nucleus region that shows higher concentrations of DNA related vibrations at 785, 1,095, 1,335, and 1,678 cm^{-1}. Notable nucleic acids related vibrations in the cytoplasm regions are attributed to ribosomal, messenger, and transfer RNA.

3.2 *k*-means clustering

Similar to the HCA, the KMC algorithm using eight clusters was applied to the same data set for the whole spectral region. The resulting image is shown in Fig. 1e. The average cluster spectra from the KMC analysis show very similar spectral features compared with the HCA and therefore are not shown. The spatial distributions of the clusters are, however, slightly different. A separation of the nucleus from the perinuclear and cytoplasm regions is obvious. Furthermore, the algorithm recognizes a thin layer around the nucleus. The nuclear membrane may be too thin to actually be spatially resolved. However, it is possible that the lipids of the highly folded membrane contribute to the Raman signal from that region. The differences compared with HCA can be attributed to the differences in approaching the clustering problem. The centroid-based approach of KMC uses random starting points, compared with the agglomerative approach of HCA. The different starting centroids in KMC may result in slightly varying

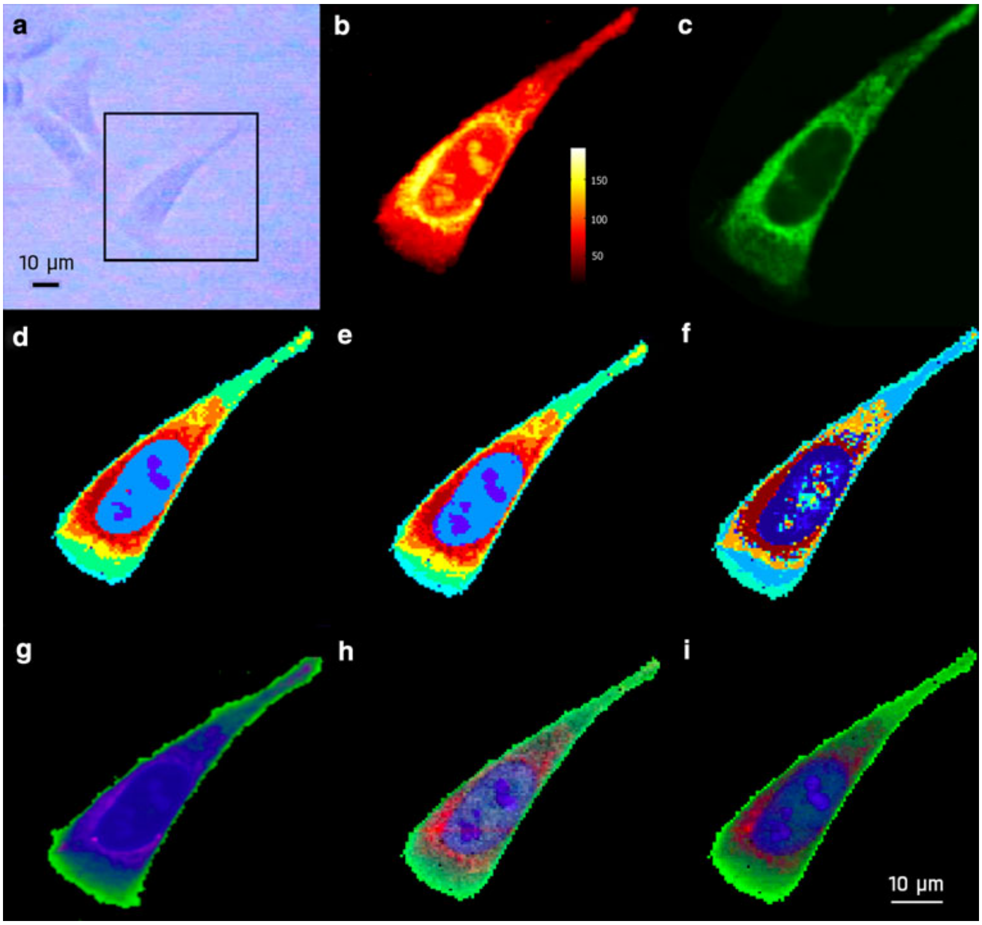

Fig. 1 **a** *White* light image of the cell. **b** Intensity plot of 2,935 cm^{-1} Raman band. **c** Fluorescence image tracking mitochondria. **d** HCA image using eight clusters. **e** KMC image using eight clusters. **f** Five-cluster FCM cluster image **g** PCA image. **h** VCA image. **i** N-FINDR image. Images **g**, **h**, and **i** were constructed using three components represented by *green*, *blue*, and *red*. The clusters in **d** and **e** are assigned to cytoplasm (*light green* and *green*), perinuclear region (*yellow*, *orange*, *red*, and *dark red*), nucleus (*light blue*), and nucleoli (*blue*)

cluster membership maps, as the assignment of clusters is based on distances from centroid to spectrum instead of the intra-spectral distances. In particular, the reproducibility of KMC is problematic for a large number of clusters. Here, KMC was repeated several times and gave stable and reproducible results for up to eight clusters. The results were less reproducible for more clusters.

3.3 Fuzzy *C*-means clustering

A five-cluster FCM model was fitted to the cell using a stop criterion of 0.0005 for the minimal amount of improvements and a fuzziness parameter of $m = 1.2$. Then, the denominator of the membership function becomes 10. Like in the other clustering methods, the nucleus is colored dark blue, perinuclear areas red, and cytoplasm cyan and light blue in Fig. 1f. Spectra from the different areas are mostly equivalent with the ones obtained by HCA and are therefore not shown. The image reveals some indications that further subcellular details are detected by the FCM model (orange) that were not resolved in the HCA and KMC images. A disadvantage of FCM is that the nucleoli inside the nucleus have not been resolved. Instead, clusters representing cytoplasmic and perinuclear region appear in the nucleus. An advantage of FCM is that a lower number of clusters are needed to visualize the main spectral differences present in the data set. Here, the main spectral features correspond to nucleus, cytoplasm with mitochondria, and cytoplasm without mitochondria as shown in Fig. 2. This may be because the transitions are much better represented by the membership functions compared with crisp clustering algorithms. However, FCM does suffer from similar problems as KMC regarding the reproducibility, and one also has to take into account the setup of the factor m for the fuzzy algorithm to obtain the best results, which means that the approach is not completely unsupervised, per se. However, m was not varied here.

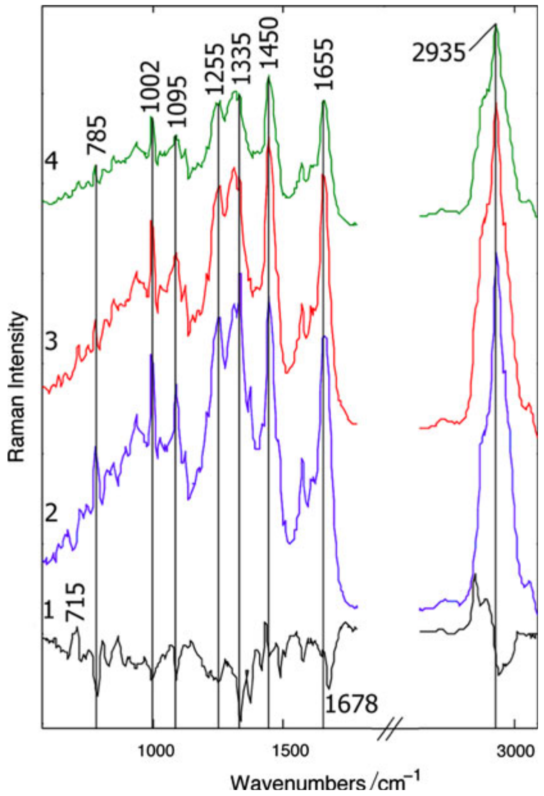

Fig. 2 Average spectra from HCA representing the nucleus (*trace 2*), perinuclear area (*trace 3*), cytoplasm (*trace 4*), and difference spectrum between perinuclear areas and nucleus (*trace 1*). The *lower* wave number regions are amplified by a factor of three

3.4 Principal component analysis

PCA decomposes the data set into a smaller set of linear independent vectors, called principal components (PCs). The score values of each PC become a measure of the contribution of the corresponding loading vector to each spectrum. Therefore, plotting the score values in an image gives the spatial distribution of that loading. Clustering approaches separate areas within the cell by assigning distinct clusters. This is not always possible in biological samples due to the nature of biological material, as there are nearly no crisp transitions between different regions or compartments. FCM can compensate for this limitation using soft cluster assignments. PCA imaging is a purely mathematical approach which is more reproducible and requires lower computation times. For illustration, PC1 to PC3 of a PCA was calculated for the Raman image of the single cell without mean-centering of the data set. In Fig. 1g, the distribution of the PC1 is plotted in green, PC2 in blue, and PC3 in red. The image is dominated by blue and green colors and reveals only low contrast. The corresponding loading vectors are displayed in Fig. 3, traces 1–3. PC1 exhibits high score values throughout the cell, since PCA finds a pseudo-average spectrum. In this particular case, the first loading is more similar to cytoplasm spectra as the number of spectra from the cytoplasm in the cell is larger compared with spectra from the perinuclear and nuclear regions. This is also evident from the spectral features of the loading that are similar to cytoplasm spectra obtained by cluster analysis. Negative features of the second loading describe the spectral contributions of the nucleus, and positive features spectral contributions of lipids. Interestingly, the PC2 loading resembles a pseudo-difference spectrum between nucleus and cytoplasm from the cluster analysis (see Fig. 2, trace 1). Consequently, the blue color of PC2 score in Fig. 1g is found in the nucleus and perinuclear region of the cytoplasm and overlaps with the green color of PC1 score. The PC3 loading is also associated with the perinuclear region and indicates higher contents of lipids visible in the region 2,850–2,900 cm^{-1} and in the $CH_{2/3}$ deformations found between 1,200 and 1,350 cm^{-1}. The overlay of the red color of the PC3 score with in the blue color of the PC2 score gives a violet color in Fig. 1g. The signal-to-noise ratios decrease from PC1 to PC3 because the spectral variances also decrease from PC1 to PC3, whereas the noise is distributed throughout the PCs. Consequently, the higher PCs are dominated by noise and were not displayed.

3.5 Vertex component analysis

Vertex component analysis is a spectral unmixing algorithm that searches for a preselected number of endmember spectra and describes all the spectra in the data set on the basis of these endmembers by a fitting routine. The amount that each endmember contributes to a spectrum is the abundance value for that endmember.

After PCA for dimension reduction, a VCA model with five endmembers was calculated for the Raman image of the single cell. Three vertex components (VCs) were found to represent cytoplasm, perinuclear areas, and the nucleus. A plot of the distribution of the endmembers is shown in Fig. 1h with the cytoplasm colored in green, the nucleus in blue, and the perinuclear areas in red. The corresponding endmember spectra in Fig. 3 are assigned to cytoplasm, nucleus, and perinuclear areas, respectively. The other two endmember spectra represent spectral information that could not clearly be assigned to cellular features.

The endmembers and abundances of VCA have higher chemical relevance and can easier be interpreted as the loadings and scores of PCA that can contain both positive and negative values. The low contrast in Fig. 1g is consistent with overlapping spectral contributions in PC1–PC3. The contrast is improved in Fig. 1h because the spectral contributions in VC1–VC3 are better separated. In

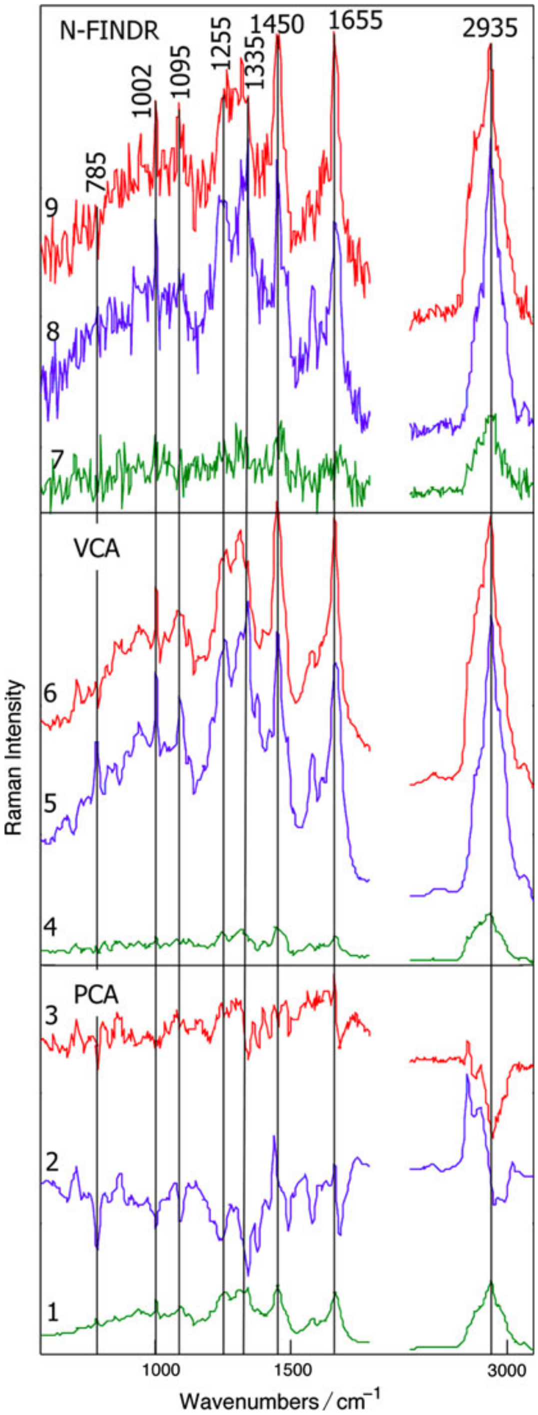

Fig. 3 Loading vectors PC1 to PC3 of a *PCA* representing a pseudo-average spectrum mostly correlated with cytoplasm (*trace 1*), the differences from cytoplasm to the nucleus (*trace 2*) and the differences between cytoplasm and perinuclear regions (*trace 3*). *VCA* endmember signatures representing cytoplasm (*trace 4*), nucleus (*trace 5*), and perinuclear regions (*trace 6*). *N-FINDR* endmember signatures for cytoplasm (*trace 7*), nucleus (*trace 8*), and perinuclear regions (*trace 9*). The *lower* wave number regions are amplified by a factor of three

particular, signatures of the nucleus and the mitochondria are well defined as evident in Fig. 3.

Compared with the results of cluster analysis, VCA gives nearly the same spatial information. The cytoplasm, perinuclear regions, nucleus, and nucleoli are clearly visible in Fig. 1h. The spectral signatures of each region can be analyzed using the endmember spectra, which, as expected, show very similar information to those obtained by HCA and KMC analysis. The advantage in VCA is, however, that transitions between regions are much better explained as combinations of the endmembers describe all the information in the data set.

3.6 N-FINDR

The N-FINDR algorithm calculated a five endmember model. Three endmembers were found to correspond to cytoplasm, perinuclear areas, and nucleus as in the VCA algorithm. The resulting image is shown in Fig. 1i with cytoplasm colored green, nucleus blue, and perinuclear areas red. The endmember signature spectra are shown in Fig. 3, traces 7–9 representing cytoplasm, nucleus, and perinuclear areas, respectively. The contrast and the spatial distribution of the components are similar in Fig. 1h, i obtained by the VCA and N-FINDR algorithm, respectively. The first difference is that the endmember spectra are not reconstructed using the PCA dimension reduction as in the VCA. Therefore, the signal quality of the endmember corresponds to the signal-to-noise ratio of the original spectra and not the PCA dimension reduced version. Second, the N-FINDR algorithm searches for the position of the spectrum which corresponds to the original spectrum from the data set and uses it for as an endmember for spectral unmixing. Due to the lower signal-to-noise ratio, a detailed assessment of the chemical information is impossible.

4 Discussion

Six different multivariate methods were compared to assess a Raman image of a single cell. Although the chemical image in Fig. 1b has a good contrast, the chemical information content of multivariate approaches is superior. Subcellular features including their spectral signatures were identified such as the nucleus, the nucleoli, mitochondria, and cytoplasm. The perinuclear distribution correlated well with the fluorescence image of the cell tracing the mitochondria. However, the fluorescent Mitotracker label did not reveal information on the nucleus or other parts of the cytoplasm. Using Raman imaging, more prominent features have been observed in other cells such as lipid droplets, apoptotic bodies, or the malaria pigment hemozoin as reported in the literature [1]. Among the main advantages of Raman imaging of cells is that no labels are required to obtain detailed information on the cell morphology and biochemistry that makes the approach particular interesting for live cell studies.

Clustering approaches have frequently been applied in the past. HCA, KMC, and FCM gave similar segmentations as displayed in Figs. 1c, d, and e. Differences were evident at transitions between subcellular features that demonstrate limitations of clustering approaches. In transition areas, spectra are found with mixed signals from each region potentially giving an extra cluster. A large number of clusters are often needed to describe all relevant regions properly, which complicates the interpretation. In this particular case, eight clusters of the crisp algorithms were required to differentiate the nucleoli from the nucleus, and several clusters differentiate both cytoplasm and perinuclear regions. This is due to small transitions and slight intensity differences between each cluster. The number of clusters was reduced to five for the soft clustering approach FCM because transitions between subcellular features could be described better.

The bottom up clustering of HCA is apparently very effective in resolving small differences between spectra. The approach, however, can run into problems in some cases resolving larger spectral differences. This limitation is avoided in KMC and FCM, as the centroid-based approach nearly always finds clusters to explain large differences. KMC and FCM, however, encounter problems with the reproducibility of the results, as both approaches use random starting points. The concept of FCM differs from HCA and KMC since the assignments are not crisp. The membership functions make it possible to better explain transitions than crisp clustering, but its also more computational intensive than KMC. HCA requires the highest computational power for large data sets as distance matrices are calculated $(n - 1)$ times for n spectra.

Principal component analysis performs a decomposition of the data set in a sense that the variations are described in lower dimensions. This results in a number of loading vectors with a corresponding set of score values. This allows clear imaging of transition regions as each pixel is composed of contributions from each loading. A high score value for one loading vector can be interpreted as a high contribution of the features in the loading vector, similar to the membership functions in FCM. However, scores and loadings are purely defined by a mathematical algorithm, and the chemical interpretation is not straightforward as positive and negative values exist, and spectral contributions of a chemical feature can be distributed over several principal components. This was observed in PC2 and PC3 for the perinuclear areas. For the Raman image of the single cell, the number of principal components was set to three and corresponds to the expected number of main spectral components. PC1 to PC3 describe 99.64% of the spectral information. That means nearly all variations are explained in the first three components. Therefore, three loading vectors contain almost all chemical differences between the regions as opposed to crisp clustering that needed eight clusters and FCM clustering that needed five clusters. Disadvantages include that the image often has a low contrast, and the loadings are not readily interpretable as they can be positive and negative and can consist of overlapping contributions.

The spectral unmixing algorithms VCA and N-FINDR address some of the problems of PCA. By calculating endmember spectra and describing the rest of the spectra as a linear combination of the endmembers, readily interpretable and chemically relevant spectral information is obtained. The endmembers were similar to the cluster centroids and could be assigned to cytoplasm, perinuclear region, and nucleus. Furthermore, abundance plots of the endmembers reveal high contrast images as seen in Fig. 1h and i. Analogous to the image reconstruction after PCA using three PCs, the abundance plots of three endmembers were considered for image reconstruction. In combination with the dimension reduction by PCA, VCA is faster and less computationally complex than N-FINDR.

Clustering and spectral unmixing algorithms depend on preprocessing procedures such as intensity normalization and baseline correction. Here, preprocessing was kept to a minimum. First, the constant spectral background was subtracted that contain spectral contributions of the substrate, buffer, and optical elements in the light path. Second, Raman spectra with intensities below a threshold were removed from the data set. A baseline correction was not required as single HeLa cells give only low fluorescence background under the experimental conditions. The cell was immersed in a buffer solution, and in this case, the absolute intensities of the spectra can be considered as density of biomolecules that is relevant information. This information would be lost after normalization, and we found that the nucleoli inside the nucleus could no longer be resolved by cluster analysis (Fig. 4). The results of vector normalization are shown as an example. Similar results were obtained using other normalization routines such as multiplicative signal correction (not shown). As mean-centering of the data set before PCA would also remove a significant amount of variation, it was not applied here. Systematic experiments are required to further study the effects of preprocessing on the algorithms.

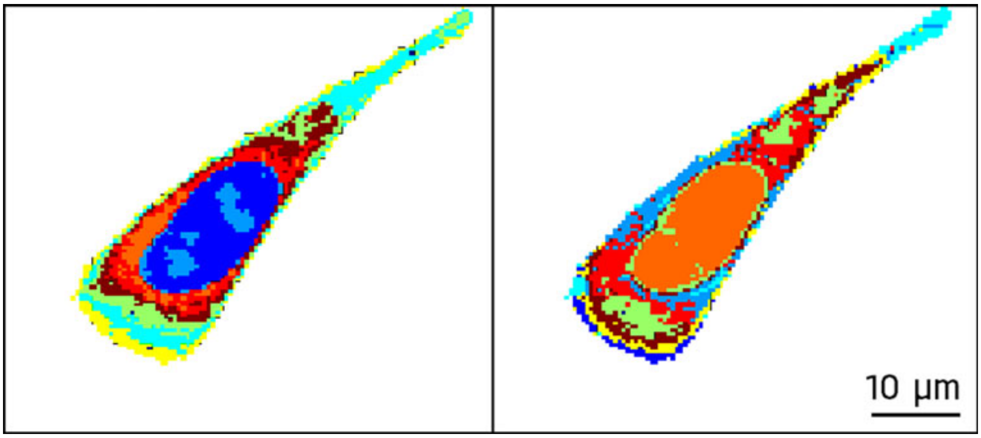

Fig. 4 KMC analysis of Raman image of the cell shown in Fig. 1. Comparison without normalization of the spectra on the left, and after vector normalization of the spectra on the right. *Colors* are not corrected to represent the same features in the figures on the left and right. See text for more details

5 Conclusions

Additional multivariate algorithms for analysis of Raman spectroscopic images may become useful in the future. Some of them are iterated constrained endmembers algorithm (ICE) presented by Berman et al. [15], self-modeling curve resolution presented by Awa et al. [25], and simplex identification via split augmented Lagrangian by Lopes et al. [26], which addresses the problems in spectral unmixing by calculating artificial endmembers. There are also people who are developing molecular level theories and models to better understand, interpret, and assign Raman modes that probe molecular vibrations. Molecular dynamic simulations can be implemented in classical, empirical, semi-empirical, and/or quantum chemical/mechanical first principles. Biomolecules were simulated making up cells and its various components proteins, lipids, carbohydrates, and nucleic acids. For example, the conformational structures of protonated polyalanine peptides were investigated in the gas phase using a combination of quantum chemical calculations and vibrational spectroscopy [27]. Concepts, simulations, and challenges were reviewed for coherent multidimensional vibrational spectroscopy of biomolecules [28]. Here, the basic principles of modern two-dimensional infrared spectroscopy are analogous to those of multidimensional NMR spectroscopy. The combination of these approaches and theories to model molecular biological systems with the algorithms described in this paper might offer advantages and opportunities to focus on Raman spectral details in local regions of interest.

The application of these algorithms is not restricted to the assessment of Raman spectroscopic images from single cells, but they also can be adapted to the analysis of Fourier transform infrared (FTIR) spectroscopy, vibrational circular dichroism, Raman optical activity, and related hyperspectral data from tissues. The confocal nature of Raman microscopy allows collecting depth profiles as shown for single cells [24] and skin [29], and the combination of the lateral information with the axial resolution even enables three-dimensional reconstruction of samples. Finally, Raman and FTIR imaging can be applied to any heterogeneous sample. Powerful multivariate algorithms will open new applications for label-free and non-destructive analyses.

Acknowledgments CK and JP acknowledge financial support of the European Union via the Europäischer Fonds für Regionale Entwicklung (EFRE) and the “Thüringer Ministerium für Bildung, Wissenschaft und Kultur” (Project: B714-07037).

References

1. Krafft C, Dietzek B, Popp J (2009) Raman and CARS microspectroscopy of cells and tissues. Analyst 134:1046–1057
2. Nan X, Potma E, Xie X (2006) Nonperturbative chemical imaging of organelle transport in living cells with coherent anti-stokes Raman scattering microscopy. Biophys J 91:728–735
3. Freudiger CW, Min W, Saar BG, Lu S, Holtom GR, He C, Tsai JC, Kang JX, Xie XS (2008) Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. Science 322:1857–1861
4. Diem M (1993) Introduction to modern vibrational spectroscopy. Wiley, Hoboken
5. Krafft C, Steiner G, Beleites C, Salzer R (2009) Disease recognition by infrared and Raman spectroscopy. J Biophotonics 2:13–28
6. Bocklitz T, Putsche M, Stüber C, Käs J, Niendorf A, Rösch P, Popp J (2009) A comprehensive study of classification methods for medical diagnosis. J Raman Spectrosc 40:1759–1765
7. Hedegaard M, Krafft C, Ditzel HJ, Johansen LE, Hassing S, Popp J (2009) Discriminating isogenic cancer cells and identifying altered unsaturated fatty acid content as associated with metastasis status, using k-means clustering and PLS-DA of Raman maps. Anal Chem 82:2797–2802
8. Miljkovic M, Chernenko T, Romeo MJ, Bird B, Matthäus C, Diem M (2010) Label-free imaging of human cells: algorithms for image reconstruction of Raman hyperspectral datasets. Analyst 135:2002–2013
9. Matthäus C, Chernenko T, Quintero L, Milane L, Kale A, Amiji M, Torchilin V, Diem M (2008) Raman microscopic imaging of

cells and applications monitoring the uptake of drug delivery systems. Proc SPIE 6991, 699106. doi:10.1117/12.800385

10. Chernenko T, Matthäus C, Milane L, Quintero L, Amiji M, Diem M (2009) Label-free Raman spectral imaging of intracellular delivery and degradation of polymeric nanoparticle systems. ACS Nano 3:3552–3559
11. Krafft C, Alipour Diderhoshan M, Recknagel P, Miljkovic M, Bauer M, Popp J (2011) Crisp and soft multivariate methods visualize individual cell nuclei in Raman images of liver tissue sections. Vib Spectrosc 55:90–100
12. Nascimento JMP, Bioucas-Dias JM (2005) Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans Geosci Remote Sens 43:898–910
13. Keshava N (2003) A survey of spectral unmixing algorithms. Lincoln Lab J 14:55–73. www.ll.mit.edu/publications/journal/pdf/vol14_no1/14_1survey.pdf
14. Winter ME (1999) N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. Proc SPIE 3753:266–275. doi:10.1117/12.366289
15. Berman M, Phatak A, Lagerstrom R, Wood BR (2009) ICE: a new method for the multivariate curve resolution of hyperspectral images. J Chemometrics 23:101–116
16. Matthäus C, Chernenko T, Newmark JA, Warner CM, Diem M (2007) Label-Free detection of mitochondrial distribution in cells by nonresonant Raman microspectroscopy. Biophys J 93:668–673
17. Ward JH (1963) Hierarchical grouping to optimize objective function. J Am Statistical Assoc 58:236–244
18. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab 1:287–297. http://www-m9.ma.tum.de/foswiki/pub/WS2010/CombOptSem/kMeans.pdf
19. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Kluwer, Norwell
20. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comp Geosci 10:191–203
21. Lasch P, Haensch W, Naumann D, Diem M (2004) Cluster analysis of colorectal adenocarcinoma imaging data: a FT-IR microspectroscopic study. Biochim Biophys Acta 1688:176–186
22. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 2(11):559–572
23. Nascimento JMP, Bioucas-Dias JM (2003) Vertex component analysis: a fast algorithm to extract endmembers spectra from hyperspectral data. Proc First IbPRIA, ser LNCS 2652:626–635. doi:10.1007/978-3-540-44871-6_73
24. Matthäus C, Kale A, Chernenko T, Torchilin V, Diem M (2008) New ways of imaging uptake and intracellular fate of liposomal drug carrier systems inside individual cells, based on Raman microscopy. Mol Pharm 5:287–293
25. Awa K, Okumura T, Shinzawa H, Otsuka M, Ozaki Y (2008) Self-modeling curve resolution (SMCR) analysis of near-infrared (NIR) imaging data of pharmaceutical tablets. Anal Chim Acta 619:81–86
26. Lopes MB, Wolff J, Bioucas-Dias JM, Figueiredo MAT (2010) Near-infrared hyperspectral unmixing based on a minimum volume criterion for fast and accurate chemometric characterization of counterfeit tablets. Anal Chem 82:1462–1469
27. Vaden TD, de Boer TS, Simons JP, Snoek LC, Suhai S, Paisz B (2008) Vibrational spectroscopy and conformational structure of protonated polyalanine peptides isolated in the gas phase. J Phys Chem A 112:4608–4616
28. Zhuang W, Hayashi T, Mukamel S (2009) Coherent multidimensional vibrational spectroscopy of biomolecules: concepts, simulations, and challenges. Angew Chem Int Ed Engl 48:3750–3781
29. Caspers PJ, Lucassen GW, Puppels GJ (2003) Combined in vivo confocal Raman spectroscopy and confocal microscopy of human skin. Biophys J 85:572–580