Chemometrics and Intelligent Laboratory Systems 155 (2016) 1–6

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Automatization of spike correction in Raman spectra of biological samples

Oleg Ryabchykov [a,b], Thomas Bocklitz [a,b,*], Anuradha Ramoji [a,c], Ute Neugebauer [a,b,c], Martin Foerster [d], Claus Kroegel [d], Michael Bauer [c], Michael Kiehntopf [e], Juergen Popp [a,b,c]

[a] Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany
[b] Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany
[c] Center for Sepsis Control and Care (CSCC), Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany
[d] Clinic for Internal Medicine I, Department of Pneumology and Allergy/Immunology, Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany
[e] Institute for Clinical Chemistry and Laboratory Diagnostics, Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany

ARTICLE INFO

Article history:
Received 22 January 2016
Received in revised form 17 March 2016
Accepted 19 March 2016
Available online 31 March 2016

Keywords:
Cosmic ray noise
Raman spectroscopy
Pre-processing
Automation
Laplacian

ABSTRACT

Raman spectroscopy as a technique has high potential for biological applications, e.g. cell and tissue analysis. In these applications, large data sets are normally recorded which require automated analysis. Unfortunately, a lot of disturbing external influences exist, which negatively affect the analysis of Raman spectra. A problematic corrupting effect in big data sets is cosmic ray noise, which usually produces intense spikes within the Raman spectra. In order to exploit Raman spectroscopy in real world applications, detection and removing of spikes should be stable, data-independent and performed without manual control. Herein, an automatic algorithm for cosmic ray noise correction is presented. The algorithm distinguishes spikes from spectra based on their response to a Laplacian, e.g. their sharpness. Manual rating of the spike presence was used as a benchmark for algorithm validation. The algorithm's sensitivity was estimated to be above 99%.

1. Introduction

With Raman spectra a label-free molecular characterization of biological samples, such as prokaryotic and eukaryotic cells or tissue specimen, can be carried out [1–3]. For example, the analysis of leukocytes by means of Raman spectroscopy offers a high potential for future application as a Raman spectroscopic hemogram [4]. Such spectroscopic hemogram can be utilized along with conventional hemogram for the diagnosis of infections [5] and for routine medical examination [6]. However, the analysis of Raman spectra of biological specimens requires sophisticated statistical data analysis methods as the biochemical changes occurring are subtle.

Prior to an application of these statistical methods, it is important to pretreat the Raman data [7]. This includes standardization and correction procedures for dealing with corrupting effects, which might mask the useful Raman information. The pre-processing should always start with a quality control of the Raman spectra, in order to verify that they contain useful information [8]. Thereafter, correction procedures, such as wavenumber correction, noise reduction and background removal, have to be performed. For each task, specialized and adapted correction algorithms have to be applied for a stable and reliable analysis [9].

Spikes are usually sharp and intense features within a measured Raman spectrum originating from cosmic ray noise on a CCD camera. They result from high energy particles which constantly bombard the earth and hit the CCD detector. When hitting the CCD detector, the particle generates a large number of electrons. If the amount of generated electrons is much larger than the charge packet of the CCD, then blooming and smear effects can appear and bright pixels will be observed in two or more consecutive frames or pixels until all electrons are transferred or leaked [10,11].

In Raman spectra, cosmic ray noise is represented as spikes — high intensity sharp peaks. The intensities, position along the wavenumber axis and the frequency of spikes occurrence are random. Only in the case of a blooming effect, spikes can be found on the same position with decreasing intensities in a few consecutive Raman spectra or in neighboring pixels [12]. Cosmic ray noise influences the analysis of large data negatively, because it affects the outcome of normalization procedure and analysis methods. Therefore, spike correction has to be performed during the pre-processing before the normalization is carried out. Here, we present a methodology for a cosmic-ray spikes detection, removal and data interpolation, where no manual optimization by an operator is required. There are some published methods for cosmic ray spikes correction and these methods differ in respect to different

* Corresponding author.
E-mail address: thomas.bocklitz@uni-jena.de (T. Bocklitz).

http://dx.doi.org/10.1016/j.chemolab.2016.03.024
0169-7439/

tasks. For some biological issues an algorithm could be used, which compares each data point with its nearest neighbors within the Raman spectra to determine presence of cosmic ray noise with a bandwidth of only 1–2 pixels [13]. Wavelet transform [14,15], polynomial filters or other smoothing methods [16] are also used for spikes correction along with noise reduction. Smoothed spectra can be used for further analysis or the difference between smoothed and original Raman spectra can be used for spikes detection [17]. The possibility to use these methods on a single Raman spectrum is an advantage of these methods. However, all of them require preset parameters, and the efficiency of the methods depends strongly on the parameter selection. For Raman spectral scans of biological samples this becomes a particular challenge, because of the big number of Raman spectra, which are often in the order of hundreds or thousands of spectra per scan. There are some algorithms that allow for a correction based on the similarity of spectra in a data set. Such methods as the Upper-Bound spectrum data matrices, which is a combination of the Upper-Bound Spectrum method with PCA [18], or the Nearest Neighbor Comparison method for Raman maps, which is based on calculation of correlation coefficients of the spectrum and each of its neighbors [19], are proven to be effective. However, these algorithms also require extensive processing time and adaptation of parameters to a specific data set.

Another important step for a spike correction method is the interpolation of the data points after removing of the spikes. Excluding data without interpolation is also possible, but can produce problems in further steps of analysis, due to missing data. The most reliable way for real-time systems is to repeat the measurement [13]. Different polynomials and linear approximations are most often used, but for the consecutive Raman spectra the average of neighboring Raman spectra can be also used.

In this contribution an automated optimization of the spike correction is achieved, which allows the researcher to work with the large data sets necessary for Raman spectroscopic investigations of complex biological samples. Appropriate criteria, obtained from the feedback of the correction procedure, were used for the purpose of automation.

2. Materials and methods

2.1. Sample preparation

White blood cells (WBCs) were isolated from five healthy volunteers' blood with informed consent according to the Ethics Committee of the Jena University Hospital (Ethic vote 4004–02/14). Briefly, ~100 μl blood from fingertip was obtained using lancet and collected in ethylenediaminetetraacetic acid (EDTA) capillary tube. Red blood cell lysis was carried out by mixing the blood with an ammonium chloride solution such that the total volume of the diluted blood is 1 ml. After 5 min of incubation at room temperature (RT) the mixture is centrifuged for 10 min at 400 g at RT. The WBCs pellet at the bottom of the Eppendorf tube was collected by discarding the supernatant and suspended in phosphate buffer saline solution (PBS, Biochrom AG, Berlin, Germany). The WBCs were chemically fixed with 4% formaldehyde (Carl Roth GmbH & Co. KG, Karlsruhe, Germany) for 10 min, followed by washing the cells successively with PBS and 0.9% NaCl (Carl Roth GmbH & Co. KG, Karlsruhe, Germany). For Raman spectroscopy the WBCs (~1×10^6 cells) were suspended in 100 μL of 0.9% NaCl prepared in distilled water and coated onto CaF_2 slides (Crystal GmbH, Berlin, Germany) by means of cytospin (Shandon Cytospin3 Cytocentrifuge, ThermoScientific, Waltham, USA, 6 min, 300 g). To ensure immobilization of the WBCs the CaF_2 slides were precoated for 10 min at RT, with 0.2% gelatin (Sigma-Aldrich, Darmstadt, Germany) solution prepared using distilled water and sterilized by heating up to 121 °C.

2.2. Raman spectra acquisition

Raman spectra of WBCs were measured with an upright micro-Raman setup (CRM 300, WITec GmbH, Germany) equipped with a 300 g/mm grating (spectral resolution about 7 cm^{-1}) and a Deep Depletion CCD camera (DU401 BR-DD, ANDOR, 1024 × 127 pixels) cooled to −75 °C. An excitation wavelength of 785 nm (diode laser, Toptica Photonics, Germany) was utilized. The laser was focused through a Zeiss 100× objective (NA 0.9) onto the cells giving 75 mW of power in the object plane. Raman images of leukocytes were recorded in the scanning mode with a step size of 0.3 μm and integration time of 1 s per spectrum. The investigated data set, in total 30 Raman spectral scans of cells from 5 donors, featured 53.235 Raman spectra and 1024 wavenumbers positions ranging from 249 cm^{-1} to 3452 cm^{-1}.

3. Calculations

3.1. Computer system

All calculations were carried out using R (version 3.0.2) [20] running on a Windows 7 Professional 64-Bit system. (Intel® Core™ i5–4570 CPU @ 3.20 GHz 2.70 GHz with 8 GB RAM). The used packages were 'Peaks' [21] and 'e1071' [22]. All Raman scans, were imported into R and arranged into a matrix with Raman spectra in the same order as they were measured. Therefore, this matrix can be considered as a 2-dimensional data set with wavenumbers in one dimension and time in the other. Prior to spike correction, fluorescence background was removed by the SNIP algorithm [23]. The wavenumbers region between 249 cm^{-1} and 415 cm^{-1} was excluded from analysis, because of the presence of a CaF_2 band that originated from the substrate.

3.2. Algorithm

There are three important steps within a correction procedure for spikes removal: identification of spikes, choosing a threshold to separate spikes from other peaks and the interpolation of data after spike removal. Several basic characteristics of spikes may be helpful for their identification: spikes usually feature a high intensity, sharpness and random position in the Raman spectrum. A high number of spikes have intensities much higher than the intensity of Raman spectral bands, but some of them are comparable with Raman peaks and cannot be recognized by their intensity alone. Therefore, a marker has to be developed to better differentiate between spikes and Raman spectral bands. This marker should take the different spike characteristics into account. Spikes are usually represented by one or few pixels within the Raman spectra that are more intense than previous and subsequent points. However, in a real Raman spectrum of biological samples the change of intensity from point to point is not as abrupt as for spikes. A mathematical formula of this idea is the discrete Laplace operator D_x^2 (Eq. (1)), which corresponds to the sharpness of spectral features.

$$D_x^2 = [-1 \;\; 2 \;\; -1] \tag{1}$$

The discrete Laplace operator response of a sharp thin spectral feature, like a spike is much higher compared to the response of a wider peak with same maximal intensity. On each side of the response to a sharp spectral feature two minima are occurring, which exhibit high absolute values. As it is shown in Fig. 1, the usage of a Laplacian operator enhances the separation between Raman peaks and spikes. In this way not only the intensity is used for separation but also the Spikes sharpness.

Another typical characteristic of spikes, but not for Raman bands, is their random position. In the case of biological Raman scans, the change between Raman spectra of nearby scanning positions are usually small. If a spike appears in one of the Raman spectra, then the intensity change at the spike position from previous and subsequent Raman spectra is significantly higher than in other spectral regions.

This feature can be used to further enhance the Laplacian response for spikes. If the matrix with Raman spectra consists of consecutive spectra, a similar method as for sharpness can be used to achieve a high response for unexpected deviations of intensity within the current