A novel method for single bacteria identification by Raman spectroscopy

E. Schultz, Anne-Catherine Simon, S.A. Strola, R. Perenon, Isabelle Espagnon, C. Allier, P. Claustre, D. Jary, J.-M. Dinten

A novel method for single bacteria identification by Raman spectroscopy

Emmanuelle Schultz [a], Anne-Catherine Simon [b], Samy Andrea Strola [a], Rémi Perenon [a], Isabelle Espagnon [b], Cédric Allier [a], Patricia Claustre [a], Dorothée Jary [a], Jean-Marc Dinten [a]
[a] CEA-LETI, Minatec Campus, 17, avenue des Martyrs, 38054 Grenoble Cedex 9, FRANCE
[b] CEA, LIST, Gif-sur-Yvette, F-91191, FRANCE

ABSTRACT

In this paper we present results on single bacteria rapid identification obtained with a low-cost and compact Raman spectrometer. At present, we demonstrate that a 1 minute procedure, including the localization of single bacterium, is sufficient to acquire comprehensive Raman spectrum in the range of 600 to 3300 cm^{-1}. Localization and detection of single bacteria is performed by means of lensfree imaging over a large field of view of 24 mm^2. An excitation source of 532 nm and 30 mW illuminates single bacteria to collect Raman signal into a Tornado Spectral Systems prototype spectrometer (HTVS technology). The acquisition time to record a single bacterium spectrum is as low as 10 s owing to the high light throughput of this spectrometer. The spectra processing features different steps for cosmic spikes removal, background subtraction, and gain normalization to correct the residual inducted fluorescence and substrate fluctuations. This allows obtaining a fine chemical fingerprint analysis. We have recorded a total of 1200 spectra over 7 bacterial species (*E. coli*, *Bacillus* species, *S. epidermis*, *M. luteus*, *S. marcescens*). The analysis of this database results in a high classification score of almost 90 %. Hence we can conclude that our setup enables automatic recognition of bacteria species among 7 different species. The speed and the sensitivity (<30 minutes for localization and spectra collection of 30 single bacteria) of our Raman spectrometer pave the way for high-throughput and non-destructive real-time bacteria identification assays. This compact and low-cost technology can benefit biomedical, clinical diagnostic and environmental applications.

Keywords: Raman spectroscopy; Lensfree imaging; Bacteria identification; Bacteria chemical fingerprint.

1. INTRODUCTION

There is presently a continuing need for bacteria analysis using rapid, portable, reliable and user-friendly systems[1]. The addressed application domains are agro-food safety, clinical microbiology, or fight against bio-terrorism. All these domains would benefit from early detection and identification of microbial contamination or infection, in order to undertake as early as possible to specific decontamination processes or narrow spectrum antibiotic prescription that will limit the proliferation of drug resistant antibiotics.

Performing analysis on single cell bacteria rather than on micro- or macro-colonies is a way to drastically increase the rapidity of analysis since it avoids the time-consuming (24 h-48 h) and sometimes not possible cultivation step. Raman spectroscopy, by providing specific information about the chemical (molecular) composition of the sample, has already been efficiently used in microbiology to identify many medically or warfare relevant bacteria (*Escherichia coli* or *Bacillus anthracis*[2,3]). For instance, Hamasha *et* al.[4] identified with a good confidence a particular *E. coli* strain among a set of closely related *E. coli* strains using spontaneous Raman scattering associated with pre-processing and chemometric techniques. Palchaudhuri et al.[5] reported a study about the metabolism of Gram positive and Gram negative. These studies were conducted on colonies and dense pellets.

The ability of Raman spectroscopy to probe even single bacterial cell has been demonstrated[7-11]: for instance, Huang *et* al.[7] were able to differentiate between growth phases of a single species, and Stoëckel et al.[9] used Surface-Enhanced Raman Spectroscopy combined with chemometric approaches to identify *Bacillus anthracis* among 27 strains of *Bacillus*. Also, spore germination dynamics at the single cell level has been very recently observed[10,11] using a novel Raman imaging scheme.

However, regarding the instruments used in these studies, we can notice that they all are laboratory systems, expensive complex and bulky. These systems are obviously not immediately compatible with routine analysis, and even less with field applications. For instance Hamasha and Palchaudhuri[4,5] use a Jobin-Yvon Horiba TRIAX 550 spectrometer combined with a liquid-nitrogen CCD camera and mounted on a modified Olympus microscope; Huang et al uses a LABRAM 300 confocal Raman microscope (Jobin-Yvon Ltd). The reasons are first the need for high spatial and axial resolution (< 1 μm^3 in volume) due to the minute size of bacteria and, second, the weak intensity of Raman signals (1 photon per 1 million incidents). The former reason explains the use of confocal microscope, and the latter the use of performing spectrometers equipped with cooled CCD to enable long acquisition times. Even though laboratory micro-spectrometers are used, typical acquisition times remain quite long (30 s) for routine analysis. Advanced Raman spectroscopy techniques have been investigated to enhance Raman signals and thus shortening the exposure time to 1-6 s: SERS allows the use of less complex instrumentation, for instance, the BioParticleExplorer coupled to a TE-cooled HE532 Jobin-Yvon spectrometer[9]. However, we preferred not to investigate SERS because it suffers from some drawbacks, such as poor reproducibility, and the need for particular substrates or sample preparation, that make it hardly compatible with biological analysis[2].

In this paper, we first describe a novel system and associated method aimed at overcoming these two traditional challenges of Raman scattering, which confined up to now this technique to laboratory and research domains. Our system first combines lensfree imaging to Raman spectroscopy. This modality is used both as a tool to detect single bacteria and to align the Raman probe on it, and as a complementary optical modality to Raman spectroscopy. Indeed, lensfree imaging yields morphological information of the single cell, which adds to the chemical fingerprint provided by Raman analysis. This scheme avoids the use of standard microscope, in particular, the use of multiple microscopic objectives of increasing magnification generally required during the alignment step. The idea of multimodal architecture has already been investigated in order to increase the instrument overall performances, but results in by far more complex instrumentation[10,12]. On the contrary, our goal is to take benefit from the additional modality to simplify the system and accelerate the operation flow. Our system also integrates a HTVS-spectrometer prototype developed by Tornado Spectral Systems, which combines high throughput in the region of interest (500-1800 cm^{-1}), acceptable spectral resolution (7 cm^{-1}), middle price and especially a very good compactness compared to laboratory spectrometers.

The present study focuses on the Raman scattering modality, which is assessed in terms of discrimination and classification of bacteria at the specie level. Spectral data obtained from single bacterial cell are pre-processed and analyzed with a data classification approach, the so-called support vector machine (SVM) technique. A dataset of 1205 Raman spectra obtained from single bacteria of seven different species has been recorded using shorter acquisition time (10 s) than those usually employed in spontaneous Raman spectroscopy (30 s)[4], or in the range of SERS (6 s)[9]. The obtained classification score of 89 % demonstrates the ability of our system to perform single bacteria analysis, and more precisely to identify bacteria at the specie level. The study thus suggests that an alternative to complex and expensive laboratory systems is made possible to rapidly detect and identify bacterial cells.

2. EXPERIMENTAL SECTION

2.1 Lensfree imaging sample holder

The sample, typically a 5 µl droplet containing bacteria, is deposited and let evaporated for 2 minutes on a quartz cover slip (TedPella Inc. 19x19x0.5 mm) that was prior placed on a 8-bit 2592x1944 CMOS sensor (MT9P031, Aptina Imaging). This original sample holder (illustrated in Figure 1B) thus implements the lensfree on-chip technique reported in Allier *et al.*[14]. Briefly and as illustrated in Figure 1A, the image formed in transmission onto the sensor results from the interference between the light coming directly from the illumination source (here a laser beam), and the light scattered by the bacterial cell(s).

This technique succeeded in monitoring and counting cell, single bacteria, or viruses using a led as the illumination beam and a thin wetting film enabling enhancement of the micron sized particles by creating a micro-lenses effect on top of them[14-16]. In this study, no wetting film is used since it would be detected by Raman spectroscopy rather than the specific signal arising from the bacteria. Yet, single bacteria are efficiently detected thanks to the high quality wave front of the laser beam. Depending on the laser spot size impinging the sample, the field of view can be varied from 24 mm^2 to 100 μm^2. More precisely, these large and small field of views are obtained when the laser spot size is about the sensor size and when it is less than the size of the bacterium to probe (1 µm), respectively. In this latter configuration, the image formed onto the CMOS sensor is due to forward elastic scattering only, and thus reveals both the laser spot and the

bacterium patterns (Figure 2). The operator is then able to monitor accurately the alignment of the probe onto the sample.

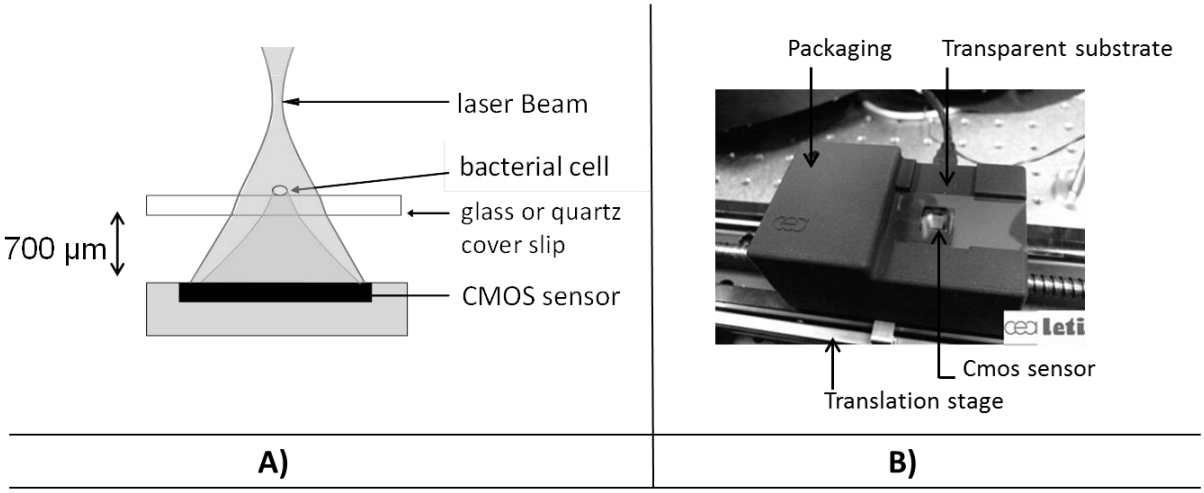

Figure 1. Lensfree Imaging Module: A) schematic illustrating the principle of lensfree image formation, B) Picture of the developed module used as the sample holder.

To conclude, the entire droplet as well as a zoomed view of a single bacterium can be easily observed, and the accurate lateral alignment of the laser probe can be achieved thanks to this so-called lensfree (or lensless) based scheme. Moreover, a forward scattering pattern, the so-called lensfree image, can be collected for each probed bacterium in order to extract its morphological characteristics. In practice, the spot size is easily adjusted by translating the sample along the laser beam using a vertical translation stage mounted on the optical bench, and the XY laser probe alignment is achieved using a double translation stages (PI micos VT-75) mounted below the lensfree module, as illustrated in Figure 3 and described in Strola *et al.*[17,18].

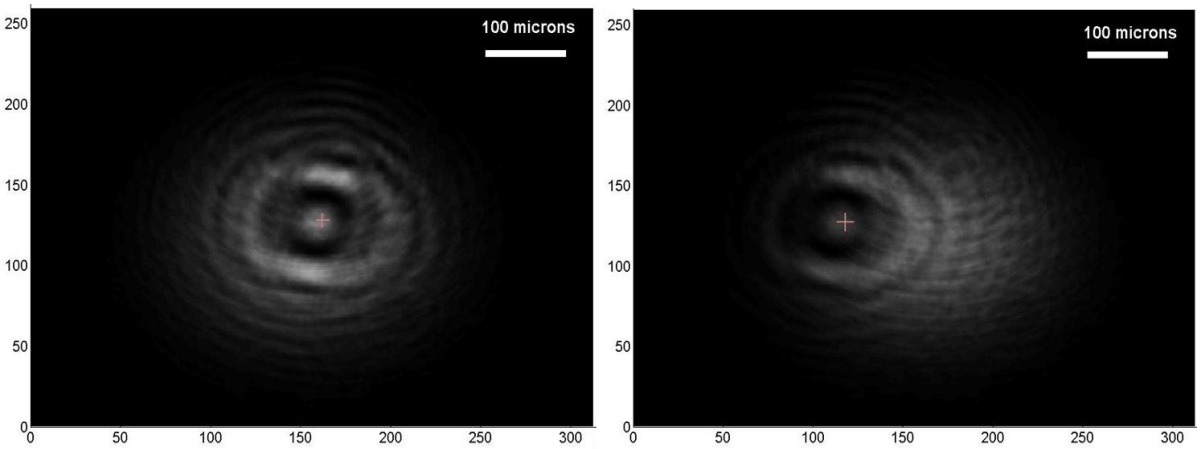

Figure 2. Lensfree images of single *B. cereus* illustrating lateral alignment process. On the left, laser probe is correctly aligned onto the bacteria. On the right image, the laser probe is about 0.8 µ right misaligned. Central cross marks the position of bacteria.

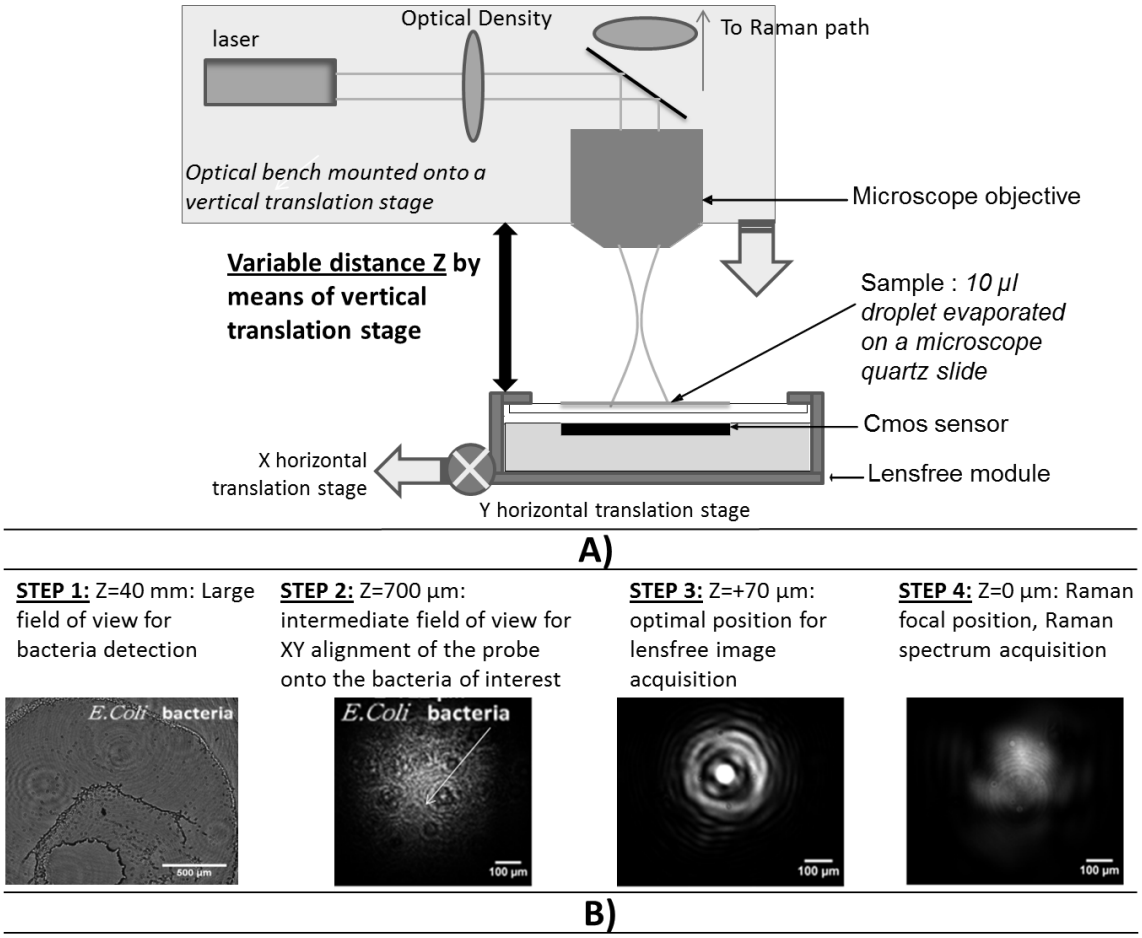

Figure 3. A) Schematic of the integration of lensfree imaging in the Raman optical bench. B) Illustration of the 4-steps operation flow: (1) using a high Z value, thus large spot size, the entire droplet is observed with diffraction patterns corresponding to bacteria; (2) using intermediate Z value enables the operator to choose the bacteria of interest and align in XY the laser probe onto it; (3) using a Z value 70 microns above the Raman focal position the interference pattern can be collected by the lensfree imaging module; (4) Raman focal position is found when the diffraction pattern blurred due to the laser spot size being smaller than the bacteria.

2.2 Setup

Figure 4 shows schematic of the setup (A) and a picture (B) of the lensfree imaging based Raman microspectrometer[14].
A 532 nm laser (Spectra Physics Excelsior 532-50-CDRH) is both the Raman excitation light and the alignment light. On the contrary, standard microscopes integrate separated light sources and paths, thus increasing the complexity and the risk of misalignments. The 50 mW optical output power is reduced by means of an optical density down to few nW during the alignment step, and to 34 mW for the Raman spectra collection. A razor-edge filter (Semrock LPD01-532RS-25) steers the laser beam at 45° into a 100x microscope objective (Olympus LMPLFLN) that illuminates the sample from above. The 100x magnification combined with the very high laser beam quality (M^2=1.02) enables a spot size of less than 1 µm in diameter at the beam waist together with a good wave front quality, an important specificity for large field lensfree imaging.
When the sample is in the beam waist position (Figure 1A), Raman scattering light is generated and collected by the microscope objective, transmitted through the razor-edge filter, filtered from elastic scattering by two notch filters (Semrock NF03-532E) and finally focused into the spectrometer optical fiber (Thorlabs M18L01, 0.22NA) using a 50 mm achromatic lens. The optical fiber diameter (105 µm) acts as the pinhole in confocal systems, and enables an acceptable axial resolution of 2 microns. The spectrometer, APEX-532, is a custom-built unit consisting in the best features from Tornado Spectral Systems' HyperFlux 532 spectrometer and Ocean Optics QE65000 detector. High

throughput virtual slit (HTVS) technology allows this unit to generate both broadband (spectral range from -4000 cm^{-1} to 4000 cm^{-1}) and high resolution (7 cm^{-1}) spectra while enhancing the optical throughput in the wavelength range of interest for bacteria analysis range (500-1500 cm^{-1}). APEX-532 is equipped with Hamamatsu detector thermoelectric cooled.
APEX-532 has been calibrated with Hg lamp calibration and polystyrene has been used as daily reference sample.
All these optical components, except for the spectrometer unit, are mounted on a single vertical translation stage (PI micos VT-75) with a resolution better than 0.4 µm that allows an accurate adjustment of the focal position. Its large course (50 mm) also enables the adjustment of the illumination spot size onto the sample according to the lensfree based method previously described.

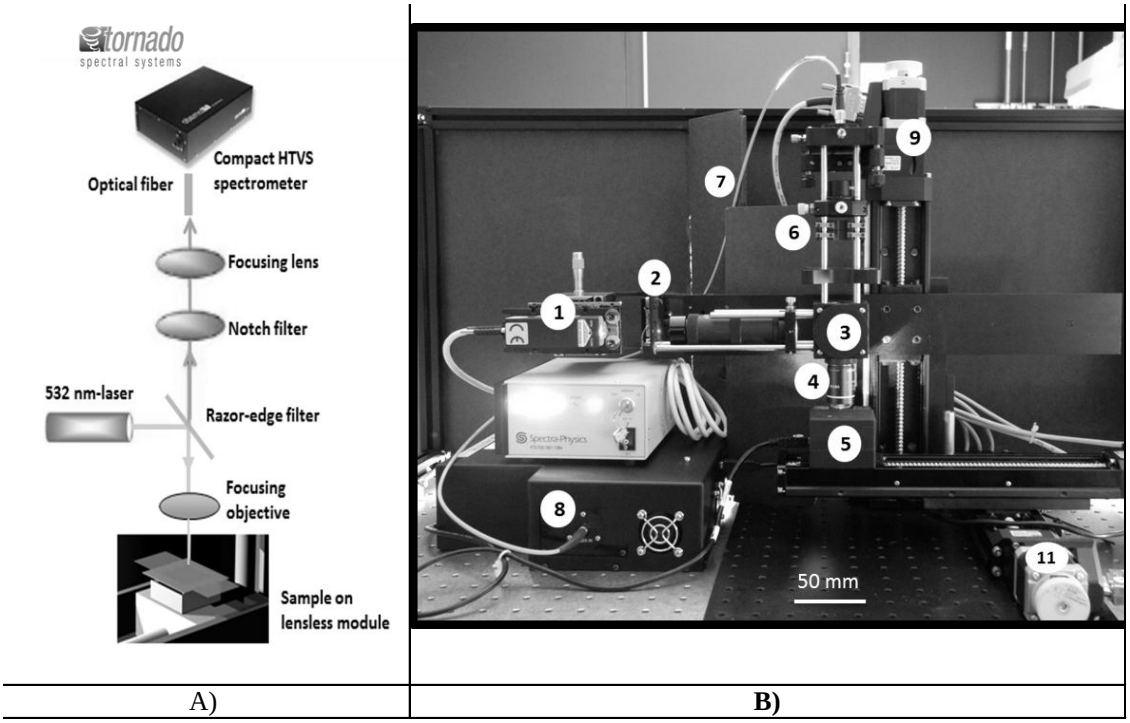

Figure 4. A) Schematic of the optical architecture. B) Picture of the apparatus consisting in 1) 532 nm CW TEM00 laser head, 2) optical density (0.3), 3) razor-edge 45° of incidence filter, 4) 100x, NA=0.8 microscope objective, 5) lensfree imaging module on which is placed the sample, 6) two notch filters and f=50 mm plane-concave focusing lens, 7) optical fiber 105 um NA=0.22, 8) Tornado Spectral systems prototype spectrometer, 9) vertical motorized translation stage, 10) and 11) translation stage in the plane of the optical table, XY.

The translation stages, spectrometer and CMOS sensor are controlled via a program developed under the software Labview® (version 2011). This program is a useful interface which allows a full control of the setup. Alignment protocol takes less than 1 minute from the moment the droplet has been evaporated until the Raman acquisition starts. The Z position provided by the lensfree based scheme is about 1-2 µm off the correct Raman focal position which is found by monitoring the appearance of the C-H band at 2925 cm^{-1} in the Raman spectra using minute translation steps (0.4 µm). An exposure time of 1 s is sufficient to obtain a Raman signal and guarantees a rapid alignment. Interference pattern collection is very fast, in the millisecond range, and the Raman exposure time has been decreased to 10 s thanks to the high spectrometer throughput. Typically, it takes 25 minutes to collect 30 spectra from 30 different single bacteria in a single droplet.
Scattering patterns and Raman spectra are analyzed off line using Matlab (R2013a) and RStudio programs, respectively. This paper focuses on the Raman scattering analysis (pre-processing and classification techniques) to demonstrate the scheme's feasibility to collect Raman comprehensive spectra, and to identify bacteria. Scattering patterns analysis will be reported in a future paper.

2.3 Biological samples

Microorganisms *E. coli* (ATCC 9637, 35421, 8739, 11775, 25922), *B. subtilis* (ATCC 23857), *S. epidermidis* (ATCC 14990), *B. cereus* (ATCC 10702), *B. thuringiensis* (ATCC 33679), *M. luteus* (ATCC 4698), *S. marcescens* (ATCC 27137) were purchased from "American Type Culture Collection". All strains were cultivated in one broth culture in a liquid medium overnight. The preparation to obtain the final bacteria concentration was following the standard protocols (washing by centrifugation and counting with a microscope)[17]. An amount of 5 µL for each bacteria solution is pipetted on top of a quartz coverslip (TedPella Inc. 19 mm^2 and 0.5 mm thickness) previously rinsed with ethanol solution at 70 % (Sigma-Aldrich) and dried with nitrogen. Before starting the measurements, we let the liquid drop containing bacteria evaporate in air at room temperature. After each analysis, the quartz coverslip is carefully cleaned in ultrasonic bath (Novatec) for 10 minutes.

2.4 Raman data analysis

Data analysis (spectra pre-processing, calculation of indicators and classification) are performed using the R software environment, with existing functions or routines specifically developed for this use.

Data pre-processing is crucial in order to select the most pertinent information in the spectra and thus to prepare the Raman spectra to be analyzed in the classification algorithm. The first step of spectrum pre-processing consists in the cosmic spikes removal. The suppression of spikes is usually done by comparing several spectra successively acquired on the same bacterium. Another method, which presents the advantage to treat each spectrum individually without depending on the other spectra, is used in Espagnon *et al.*[19]. It is based on the specificity of the spikes, which are thinner and higher than the bacteria's peaks. So a peak search, carried out from the second derivative of the signal, is processed on each spectrum, and spikes are detected using optimized thresholds. Once detected, spikes are replaced by a background obtained by linear interpolation between both sides of the peak.

Then, several types of spectrum pre-processed can be carried out to form the input data of the classification method, from the simplest to the more complex ones: the raw spectrum, smoothed and normalized, the first derivative, and the normalized net spectrum after background subtraction were tried out. In all cases, the data are reduced to two regions of interest (ROI). The main region is from 650 cm^{-1} to 1800 cm^{-1} and contains most of the specific peaks of bacteria. An additional region of interest surrounds the CH stretching band and spreads from 2600 cm^{-1} to 3200 cm^{-1}.

More precisely, smoothed signal and first derivative are calculated by Savitzky-Golay polynomials filters[20] (degree 4, on 9 points), a powerful, simple and fast method to obtain the desired signal without loss of intensity.

Background subtraction is more complex and includes different steps. First of all, the aim is to eliminate the high quartz contribution in comparison with the bacteria signal. The first step consists in considering the mean quartz spectra, calculated from the different quartz spectra acquired at the same date as the bacterium spectrum and relative to the same strip, and in fitting it on each bacteria spectrum on the large peak spreading from 200 to 650 cm^{-1}, specific to quartz (Figure 5A). An approaching method is presented in Beier *et al.*[21], with a rather different spectrum topology.

A constraint on the relative level of quartz signal, which has to be smaller than the bacteria spectrum on the region up to 1700 cm^{-1}, is added to the fitting procedure. The fitted mean quartz spectrum is then removed from the considered bacteria spectrum. Afterwards, a peak stripping on the resulting spectrum is processed using the Clayton's algorithm (also used in the SNIP algorithm[22]), with a neighborhood window of 3 channels. This algorithm is iterative: when the number of iterations grows, the calculated background level drops. At the end of the process, the quartz contribution is extremely reduced and the bacteria peaks are thus emphasized, which facilitates their study. The resulting spectrum is called "specific net spectrum" (Figure 5B).

Normalization is the last step before cutting the spectrum to the ROI. It is obtained by dividing the signal by its mean value on a chosen region of interest. It enables to have all spectra at the same scale, independently of factors that may vary between two spectra or two experiments (laser power, etc.).

For 100 spectra, time of pre-processing is about 8 s to load the spectra and 7 s to remove cosmic spikes. Then the time to subtract the background depends on the iteration number. For example, time needed for 600 iterations is only 5 s, even though it is 30 s for 2500 iterations. These durations are given for a PC equipped with an Intel Core i5 at 2.40 GHz and 8 Go RAM. Once calculated and saved, the pre-processed spectra can be loaded as often as necessary for the following treatments.

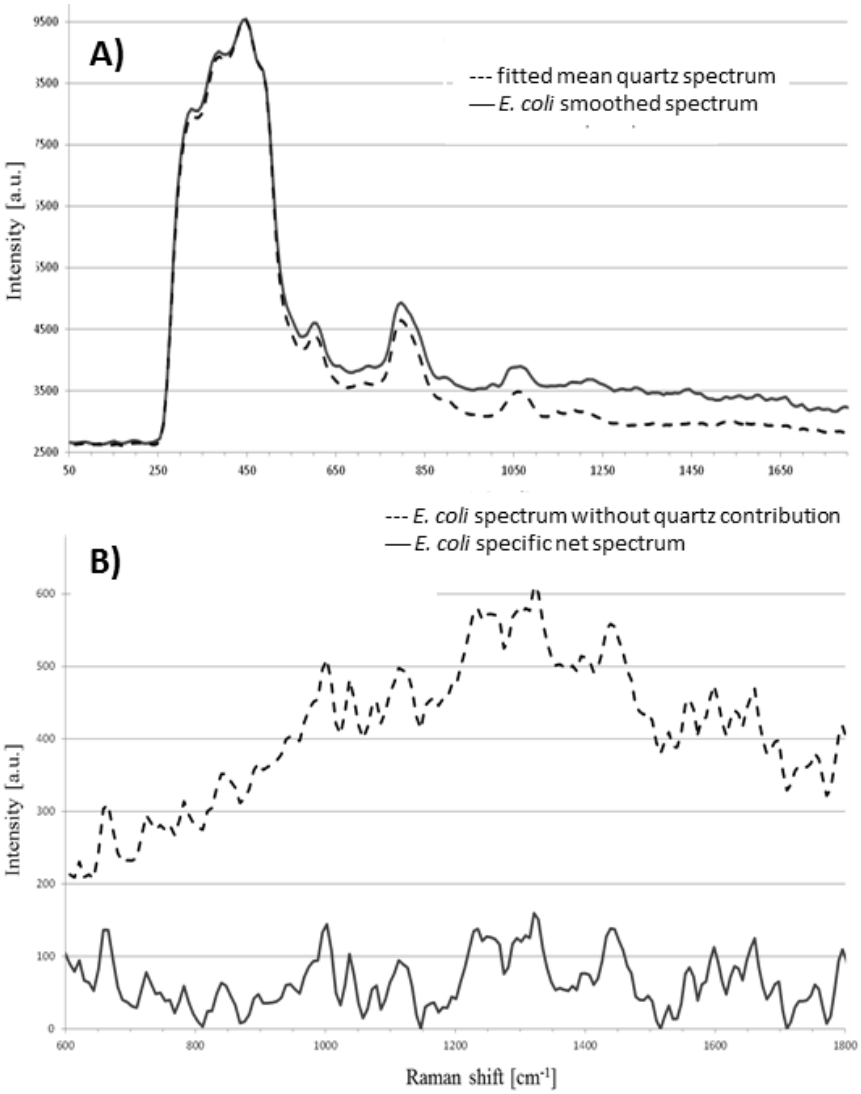

Figure 5. Examples of *E. coli* spectrum: A) raw spectrum and mean quartz spectrum after fitting on the main peak of quartz (both spectra were smoothed). The high contribution of the quartz in the bacterium spectrum can be observed. B) spectrum obtained after removing the quartz contribution (dotted line), then after Clayton's algorithm with 600 iterations (specific net spectrum).

Indicators of spectra quality

In order to assess the overall quality of the spectra, we used two main types of indicators: the standard deviation of the means (SDM) and the signal to noise ratio (SNR).

The SDM of a set of spectra is the mean standard deviation of channels normalized to the standard deviation of the mean spectrum[9]. We calculate it on the normalized specific net spectra representative of a same strain. Low values of SDM indicate low variability by channel and high reproducibility of the spectra set, when high values, closed to 1, are the sign of spectra with high noise or disparities in the spectra set.

The SNR is a quality indicator of the individual spectra. It is calculated by the mean of the specific net signal in a region specific to bacteria, here the peak at 1445 cm^{-1}, divided by the standard deviation of the specific net signal in a region without bacteria signal, here the 2000-2500 cm^{-1} region.

In addition to these two quantitative indicators, more qualitative tools may be used: the dendrograms measure the distances between spectra and represent the latter under a tree according to the distance that separates them. Associated

to a classical principal component analysis (PCA) and to the calculation of the mean spectra plotted for each strain, they enable to bring out, not only simple outliers but also groups of spectra diverging from the others.

Classification method

The classification algorithm used here is the support vector machine (SVM), which is a supervised classifier (function "svm" of the R package "e1071", interfacing the" LIBSVM" library[23], with a linear kernel).

For cross-validation, we proceeded as follows. All species strains were represented in the reference spectra base. In order to avoid a too perfect match, called over fitting, a one-tenth of each strain, for all strains, was randomly chosen at each loop, and removed from the reference base to form the validation base. Thus, a 10-fold cross-validation enables to test all spectra. Moreover, this process was carried out ten times, so the mean and the standard deviation of the global correct identification rate for all strains, as well as a mean confusion matrix providing the results for each strain, could be given.

3. RESULTS

Prior staring the construction of a database, we had to ensure that the method enables to probe single bacteria and not aggregates. Indeed, the large field lensfree imaging patterns can be very similar for these two kinds of samples. We therefore added a direct imaging path attached to the razor-edge beamsplitter cube (Figure 6A). A mirror in the Raman path reflects the collected light which is then steered by the razor-edge filter back face to a CMOS sensor (Thorlabs) equipped with a Navitar objective. As samples are micron-sized, a pair of lens was used in front of the CMOS to magnify the image by a factor 40. The sample can be illuminated using a white light illuminator, or with the laser light. We are thus able to control that a single bacterium is well probed and that a good laser alignment quality is achieved using the lensfree based method. The scattering pattern and Raman spectra collected for that bacterium are displayed in Figure 6B and 6C respectively.

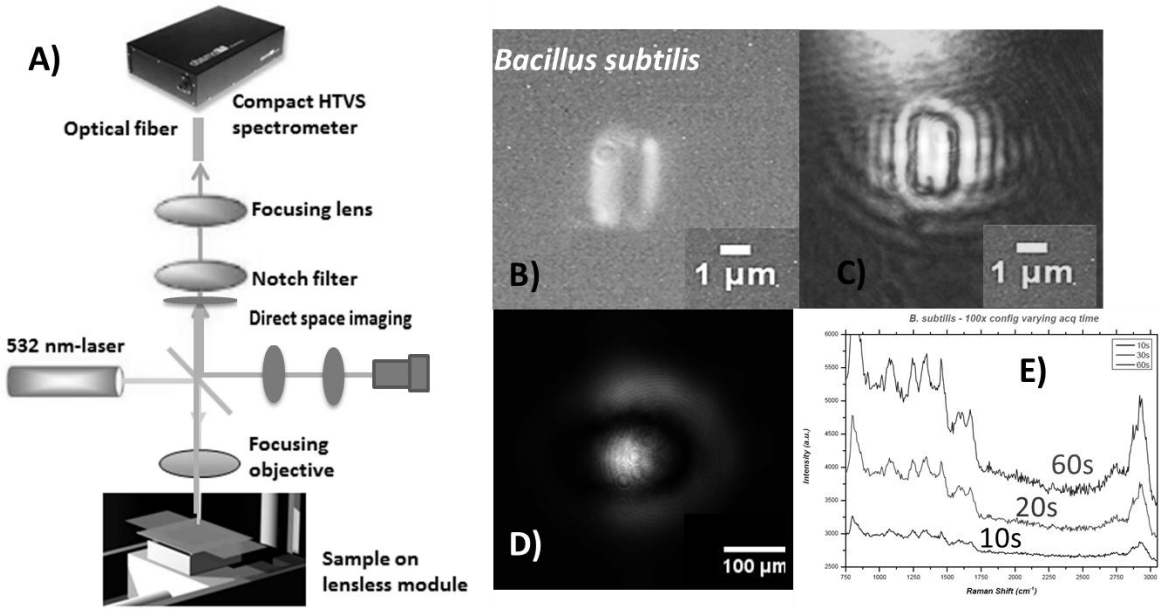

Figure 6. Validation of the lensfree based scheme ability to probe a single bacterial cell and collect scattering pattern and Raman spectrum. A) Schematic of the setup with a direct imaging modality added. B) Direct space white light image of the bacterial cell, C) Back-scattered image showing the laser probe is well aligned with the cell, D) forward scattered pattern collected using lensfree module, and E) Raman spectrum generated by the bacteria cell.

Figure 7 shows the average of processed Raman spectra acquired for *B. subtilis* and *E. coli* bacterial strains with 10 s exposure time. We can verify that the spectra consist in bands representing the cells contents: proteins, lipids, nucleic

acids. In particular, the peaks centered at 784 cm^{-1}, 1001 cm^{-1}, 1170 cm^{-1}, 1242 cm^{-1}, 1338 cm^{-1}, 1445 cm^{-1}, 1573 cm^{-1}, 1605 cm^{-1}, 1655 cm^{-1}, 2925 cm^{-1} can be detected for both strains. We assign the bands of Raman single bacteria spectra to cell wall and cytoplasm components since we have demonstrated that detected peaks cannot be attributed to quartz substrate after the pre-processing data treatment.

Typical signatures of cell components are CH stretching vibrations and are arising from signals reported at 2925 cm^{-1}. Contribution from various proteins is related to band centered at 1655 cm^{-1} where amide I, mainly associated with the C=O stretching vibration and directly related to the backbone conformation, is detected. Amide III, known as very complex band dependent on the details of the force field, the nature of side chains and hydrogen bonding, are revealed by 1242 cm^{-1} band. DNA bands arise in the region 1573 cm^{-1} via guanine and adenine nucleobases, but an overlapping with amide II contributions (C=C, N-H and C-N stretching) can be reported. CH_2 bending vibrations are reported by the band at 1445 cm^{-1} and give contribution from various lipids. The presence of bands arising from amino acid side chains must be recognized by 1605 cm^{-1} and 1001 cm^{-1} band: the signals can be both assigned to phenylalanine. Tyrosine and phenylalanine gives origin to the signal detected at 1170 cm^{-1}. The 1338 cm^{-1} region show fingerprint of cytoplasm fraction detected via DNA vibration. Band centered at 784 cm^{-1} is covered by cytosine stretching vibrations as part of DNA contribution.

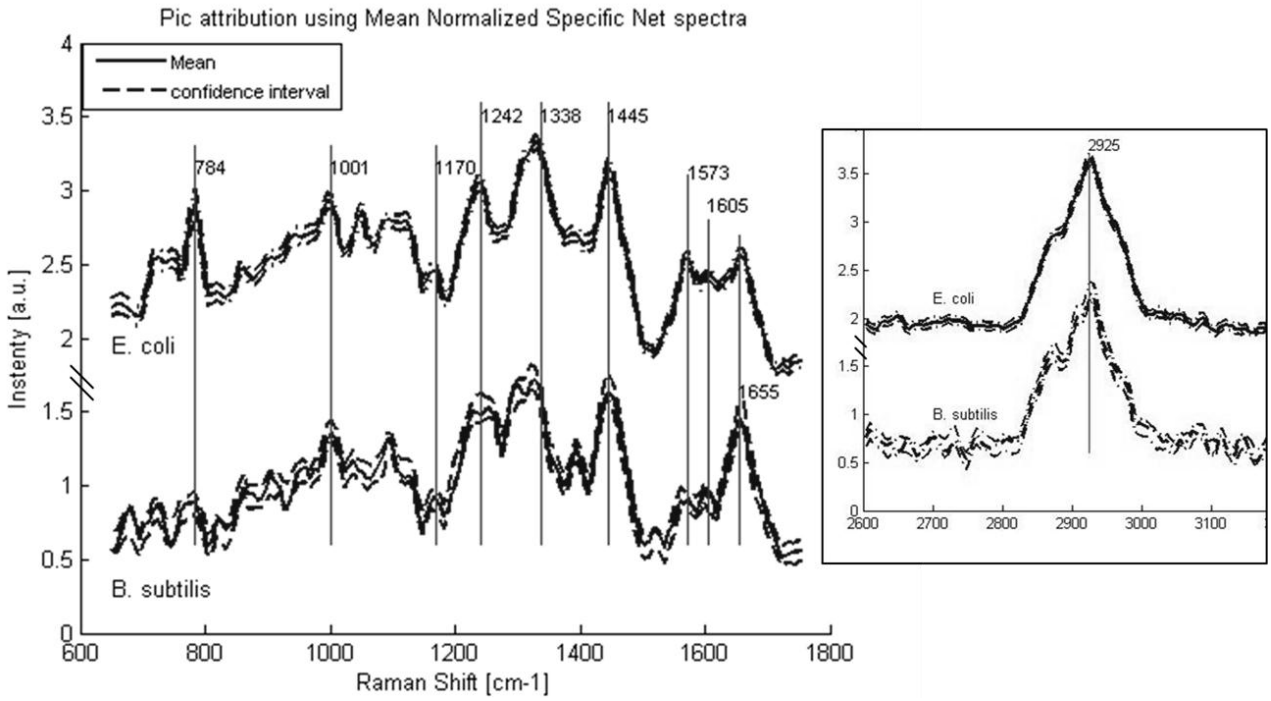

Figure 7. Mean normalized net spectrum for E. coli and B. subtilis. Dotted lines correspond to three times the mean standard deviation of the mean value, divided by the square root of the number of spectra (154 and 273 for E. coli and B. subtilis, respectively). Raman bands detected are indicated, and are found consistent with previous studies[25].

Database description

The database consisted of seven species, with one strain for each species: *B. cereus*, *B. subtilis*, *B. thuringiensis*, *E. coli*, *M. luteus*, *S. epidermidis*, and *S. marcescens*.

For each strain, acquisitions were carried out at four dates at the minimum (8 dates for *B. cereus* and *B. subtilis*). At least 30 spectra were acquired per day, which represented a total number of 1209 spectra, with a 10 seconds acquisition time.

The SDM and SNR for each strain were calculated (Table 1). SDM values are rather high, from 0.6 for *S. epidermidis* to values largely greater than 1, with 1.3 for *E. coli* and 1.7 for *S. marcescens*, mainly due to the low acquisition time. We find the same trend with the SNR: the mean SNR for each strain ranges from 3 to 6 corresponding to the noisiest series (*S. marcescens* and *E. coli*) and to the best series (*M. luteus* and *S. epidermidis*), respectively.

Table 1. SDM and SNR values for the different bacteria investigated.

Bacteria	SDM	Mean SNR
B.cereus	1.1	4
B. subtilis	1.0	5
B. thuringiensis	0.9	4
E. coli	1.3	3
M. luteus	0.8	6
S. epidermidis	0.6	6
S. marcescens	1.7	3

The dendrograms, PCA and mean spectra, plotted for each strain, brought out some aberrant spectra, probably due to a contamination of the quartz substrate. That may explain the higher value of SDM obtained for *M. luteus* (0.8), compared to *S. epidermidis*, although the SNR are similar.

Choice of the best pre-processing

The classification results were employed to choose the most adapted pre-processing for the identification aim. First of all, the iteration number for the Clayton algorithm was optimized. Indeed, if we can consider that 600 is a good choice to analyze bacteria peaks in the region of interest, it is not necessarily the most fitted one for the identification. So we used the global classification rate as a criterion to find the optimal iteration number, between 100 and 20000 iterations, from the normalized specific net spectra. In this part, we used a 2-fold cross-validation, giving lower classification rates but quicker than a 10-fold cross-validation. In a first phase, this search was done on both regions of interest independently. For each region, although the topology of regions is different, the same trend was observed (Figure 8): when the iteration number increases from 100 to 5000, the classification rate grows quickly.

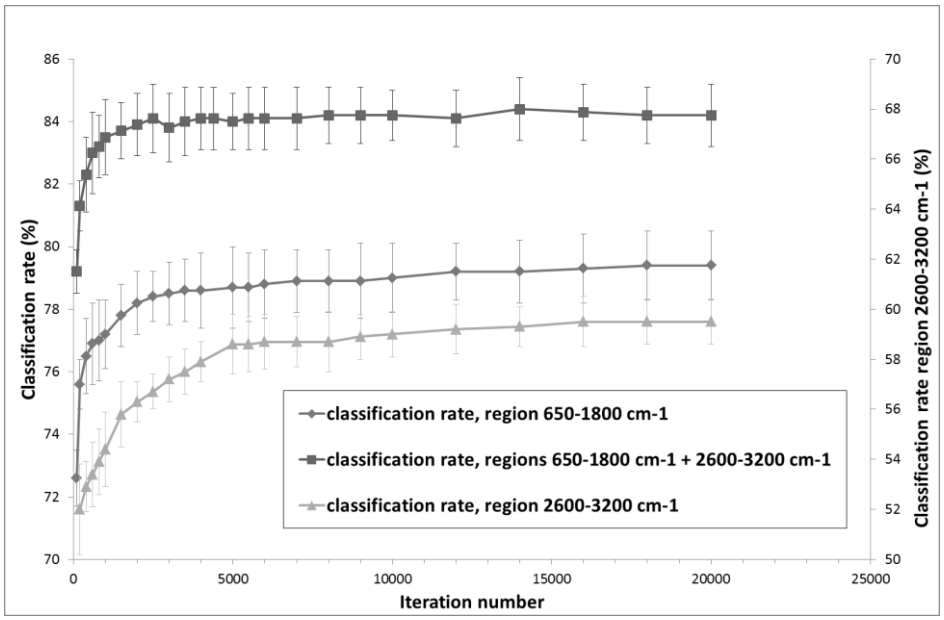

Figure 8. Classification rate according to the iteration number in the Clayton's algorithm, for each region of interest (2-fold cross-validation).

Beyond 5000 iterations, the classification rate follows an asymptotic tendency. So the iteration number can be chosen differently according to the priority given to the time execution or to the classification rate and its standard deviation. It can be noted that there is a great difference between the classification rates obtained by each region: lower than 60 % for the second region and almost 80 % for the first one. Then both regions were associated in the classification, with the same iteration number. Clearly better results were obtained in comparison to the use of the first region only, with a classification rate increasing more rapidly between 100 and 2500 iterations, then rather stable at 84 %. It was verified that a satisfying iteration number of 12000 for the first region, associated with 18000 iterations for the second region, gives a similar result as 14000 iterations for both regions. In conclusion, a number of 4000 iterations represent a good compromise between time execution and classification rate.

As said before, for each bacterium spectrum, the mean quartz spectra used to subtract the quartz contribution is calculated from the quartz spectra acquired at the same date on the same strip. We verified that this solution gives better classification results (84.1 % ± 1 %) than the use of a mean quartz spectrum calculated from all available quartz spectra (71.3 % ± 1.1 %). So, even if the quartz spectra seem to be very similar to each other, they present differences, in particular in the slope of the background at about 650 cm^{-1}, that influence the classification very much.

Then, we searched if this pre-processing really brings some advantages to the classification algorithm. Indeed, the classifier may be able to exploit information in the spectrum that we don't consider as useful. So we compared the classification rate given by the normalized specific net spectrum with the results obtained with the raw spectrum, after smoothing and normalization, and with the first derivative, both on the main region and with the additional region.

Results are presented in Table 2 with a 10-fold cross-validation. They show that first derivatives give less good results than the other methods, both for one region of interest (classification rate equal to 76.4 % ± 0.6 %) or two regions (74.7 % ± 0.4 %). We can note that the use of the second region does not improve the classification rate, and even it degrades it. We have an opposite effect when the smoothed spectra are used as input in the classification algorithm: results are much improved with two regions (classification rate of 84.5 % ± 0.3 %) compared to one region (classification rate of 78.8 % ± 0.8 %). That means that the second region between 2600 cm^{-1} and 3200 cm^{-1} in the smoothed spectra contains information that can be exploited by the classification algorithm.

Table 2.Mean classification rates and standard deviation according to the type of pre-processing (10-fold cross-validation).

	Mean classification rate and standard deviation	
	ROI 650-1800 cm^{-1}	ROI 650-1800 cm^{-1} + 2600-3200 cm^{-1}
Normalized smoothed spectrum	78.8 % ± 0.8 %	84.5 % ± 0.3 %
First derivative	76.4 % ± 0.6 %	74.7 % ± 0.4 %
Normalized specific net spectrum	82.7 % ± 0.6 % (7000 iterations)	86.5 % ± 0.5 % (4000 iterations)

However, the best classification rate is obtained from the normalized specific net spectra, with a classification rate of 86.5 % ± 0.5% with 4000 iterations, so we can conclude that the pre-processing is useful to extract the right information, specific to the bacteria, to be provided to the classifier.

4. DISCUSSION AND CONCLUSIONS

The detailed classification rates for each strain are given under the form of a confusion matrix (Figure 9 left). It is interesting to note that the highest classification rates (higher than 90 %) are obtained for *M. luteus* and *S. epidermidis*, which present the best SDM and SNR. On the opposite, *S. marcescens*, which has the highest SDM and the lowest SNR, obtains a classification rate of only 79.9 %. The most frequent confusions are observed between *B. cereus* and the other *Bacillus* or between *B. cereus* and the two *Enterobacteriaceae*, and between the two *Enterobacteriaceae* themselves.

Simpler than bacteria strain identification, one can also aim at distinguishing the different families of bacteria, a classification according to the families was carried out. The resulting confusion matrix is displayed in Figure 9 right. A very satisfying classification rate was obtained for the bacillus family, with only 7.2 % of the *Bacillus* spectra classed as non-*Bacillus*. The lowest classification rate (83.3 %) was obtained for the *Enterobacteriaceae*.

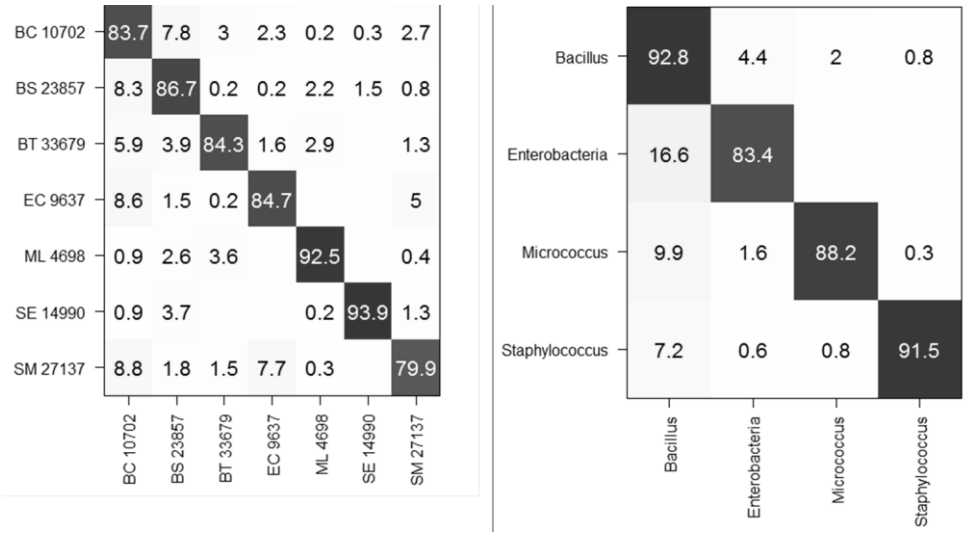

Figure 9. Mean confusion matrix using 10 s acquisition time and SVM classification technique. Left) species level classification (86.5 % ± 0.5 %) and Right) families level classification (89 % ± 0.5 %).

In order to measure the influence of the acquisition time on the spectrum quality and further on the classification results, experiments were carried out on E. coli and *S. marcescens* with a longer acquisition time, equal to 20 seconds. The same procedure was followed, with four days of experiments and at least 30 spectra a day. The new values of SDM are clearly lower, 0.7 for both strains, which represents a decreasing by 1.9 and 2.7 for *E. coli* and for *S. marcescens* respectively. For the SNR, a net improvement was also observed with higher SNR values (5 for *E. coli* and 6 for *S. marcescens* means a gain of 1.5 and 1.9). The new confusion matrix is presented in Figure 10 left. The classification rates for *E. coli* and *S. marcescens* are clearly better, as they grow respectively from 84.7 % to 91.1 % and from 79.9 % to 95.6 %. This has a positive influence on some of the other bacteria too, and the global classification becomes 89.8 % ± 0.5 %. The same effect can be noted with the classification according to the different families: the classification rate is improved by 10 % for the *Enterobacteriaceae*, and by a few percent for the *Bacillus* (Figure 10 right). These results clearly show that acquisition affects the classification results. Yet we show that acquisition as low as 10 seconds only gives consistent results.

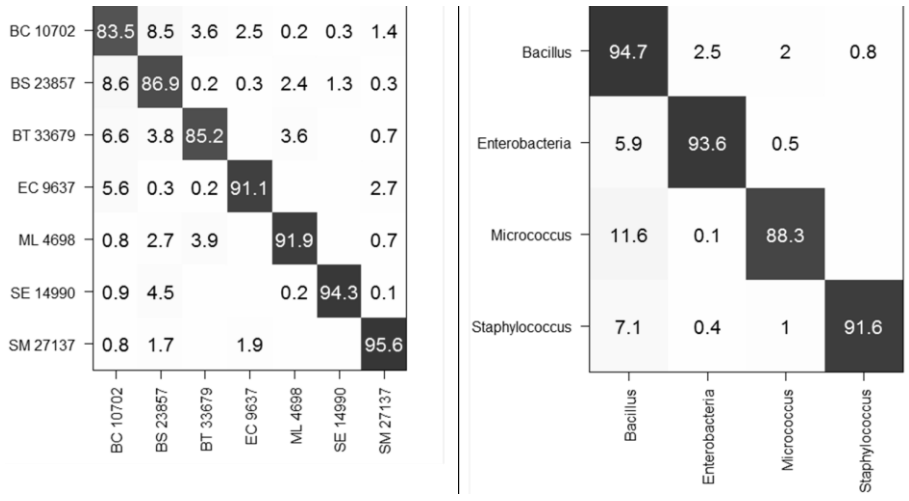

Figure 10. Mean confusion matrix obtained using longer exposure times (20 s instead of 10 s) for *E. coli* and *S. marcescens*. Left) classification at the strain level (89.8 % ± 0.5 %), and Right) at the family level (92 % ± 0.5 %).

A complementary test was undertaken: the spectra with a 10 s acquisition time for the two *Enterobacteriaceae* were put in the reference base, and the spectra acquired in 20 s were placed in the test base. The idea would be to adjust in real time the acquisition duration for each new unknown bacterium according to its measured SNR. But this didn't provide good results, because the spectra of different acquisition times, even normalized, cannot be compared (form of the peaks, relative heights between the peaks), due to an evolution of the Raman signal with time. However, if necessary, the acquisition time could be increased for all acquisitions in order to improve the identification.

As mentioned in Schutze[4], the type of cross-validation described in this work is not totally realistic because spectra of a same set (a same experiment date and a same type of bacterium) are employed both in the reference base and in the validation base, even if each spectrum is not in both bases at the same time. In order to perform an external validation, the influence of important factors like growing time, or nutrition, or matrix (sterile or real) effects need to be taken into account in the dataset. For instance, we already observed that our system can discriminate between different growth phases of *E. coli* and *B. subtilis* bacteria based on Raman analysis[24]. Next work will be then to build a reliable dataset comprising these different parameters, and to investigate how scattering patterns collected using the lensfree module can help the identification of bacteria strains.

To conclude, we described our novel Raman spectrometer based on a lensfree imaging scheme that overcomes some of the challenges encountered in Raman microscopy, namely compactness, rapidity and simplicity. The present study was focused on the Raman spectroscopy acquisition system. We demonstrated the ability of our system to probe single bacterial cell, and to differentiate bacterial species with a confidence interval of 86% via SVM classification. An improved classification rate is expected thanks to an enlarged Raman dataset, and by adding the morphological information yielded by the analysis of lensfree images. However, the obtained results can already be considered as comparable with those reported in literature but obtained with different techniques (i.e. SERS) or using research bulky and not user-friendly microspectrometer devices. This study thus paves the way for the development of the next-generation of compact and high performing spectroscopic devices designed for biomedical applications.

ACKNOWLEDGEMENTS

The authors thank the French trans-governmental CBRN-E R&D program for its financial support.

REFERENCES

[1] Premasiri W.R., Sauer-Budge A.F., Lee J.C., Klapperich C.M., Ziegler L.D., "Rapid Bacterial Diagnostics Via Surface-Enhanced Raman Microscopy", Spectroscopy, (2013)

[2] Das R.S., Agrawal Y.K., "Raman spectroscopy: recent advancements, techniques and applications", Vibrational Spectroscopy, 57, 163-176 (2011)

[3] Ashton L., Lau C., Winder C.L., Goodacre R., "Raman spectroscopy: lighting up the future of microbial identification", Future Microbiology 6(9), 991-997 (2011)

[4] Schütze K., "Summary Raman Litterature", CellTool GmbH, Bernried, Germany (2013)

[5] Hamasha K., Mohaidat Q.I., Putnam R. A, Woodman R.C., Palchaudhuri S., Rehse S. J, "Sensitive and specific discrimination of pathogenic and nonpathogenic Escherichia coli using Raman spectroscopy – a comparison of two multivariate analysis techniques", Biomedical Optics Express 4 (4), 481-489 (2013)

[6] Palchaudhuri S., Rehse S.J., Hamasha K., Syed T., Kurtovic E., Kurtovic E., Stenger J., "Raman Spectroscopy of Xylitol Uptake and Metabolism in Gram-Positive and Gram-negative Bacteria", Applied and Environmental Microbiology, 77, 131-137 (2011)

[7] Huang W.E., Griffiths R.I., Thompson I.P., Bailey M.J., Whiteley A.S., "Raman Microscopic Analysis of Single Microbial Cells", Anal. Chem. 76, 4452-4458 (2004)

[8] Rösch P., Harz M., Schmitt M., Peschke K-F, Ronneberger O., Burkhardt H., Motzkus H-W, Lankers M., Hofer S., Thiele H., Popp J., "Chemotaxonomic Identification of Single Bacteria by Micro-Raman Spectroscopy: Application to Clean-Room-Relevant Biological contaminations", Applied and Environmental Microbiology 71 (3), 1626-1637 (2005)

[9] Stöckel S., Meisel S., Elschner M., Rösch P., Popp J., "Identification of Bacillus anthracis via Raman Spectroscopy and Chemometric Approaches", Anal. Chem. 84, 9873-9880 (2012)

[10] Kong L., Zhang P., Wang G., Yu J., Setlow P., Li Y-Q, "Characterization of bacterial spore germination using phase-constrast and fluorescence microscopy, Raman spectroscopy and optical tweezers", Nature Protocols 6 (5), 625-639 (2011)
[11] Kong L., Setlow P., Li Y-q, "Observation of the dynamic germination of single bacterial spores using rapid Raman imaging", Journal of Biomedical Optics, 19 (1), 011003-1-6 (2014)
[12] Smith Z.J. and Berger A.J., "Construction of an integrated Raman- and angular-scattering microscope", Review of Scientific Instruments., 80, 044302 (2009)
[13] Schuster K. C., Reese I., Urlaub E., Gapes J. R. and Lendl B., Multidimensional information on the chemical composition of single bacterial cells by confocal Raman microspectroscopy, Anal. Chem., 72(22), 5529–5 534 (2000)
[14] Allier C.P., Hiernard G., Poher V., Dinten J.M, « Bacteria detection with thin film wetting film lensless imaging », Biomedical Optics Express, 1(3), 762-770 (2010)
[15] Hennequin Y., Allier C. P., McLeod E, Mudanyali O., Migliozzi D., Ozcan A., and. Dinten J.-M, "Optical Detection and Sizing of Single Nano-Particles Using Continuous Wetting Films.," ACS Nano, 7 (9), 7601–7609 (2013)
[16] Mudanyali O., Mcleod E., Luo W., Greenbaum A., Coskun A. F.,. Allier C. P, and Ozcan A., "Wide-field optical imaging of single nano-particles and viruses using computational on-chip microscopy and self-assembled liquid nano-lenses," Nature Photon.7, 247-254 (2013)
[17] Strola S.A., Schultz E., Allier C.P, DesRoches B., Lemonnier J., Dinten J-M., "Raman microspectrometer combined with scattering microscopy and lensless imaging for bacteria identification" Paper 8572-39, BIOS2013, Advanced Biomedical and Clinical Diagnostic Systems XI, Session 8: Vibrational Spectroscopy, 4th February 2013, San Francisco (2013)
[18] Schultz E., S. Strola, C.P; Allier, M. Dupoy, Patent FR13-50857 (2013)
[19] Espagnon I; and al., "Direct identification of clinically relevant bacterial and yeast microcolonies and macrocolonies on culture media by Raman spectroscopy", to be published
[20] Savitzky A., Golay M.J.E., "Smoothing and Differentiation of Data by Simplified Least Squares Procedures", Anal. Chem. 36(8), 1627–1639 (1964)
[21] Brooke D. Beier and al; "Method for automated background subtraction from Raman spectra containing known contaminants", Analyst, 134, 1198-1202 (2009)
[22] Ryan C.G., and al., "SNIP, a statistics-sensitive background treatment for the quantitative of PIXE spectra in geoscience applications", Nucl. Instr. and Meth. B34, 396-402 (1988)
[23] Chang C.C., Lin C.J., LIBSVM: a library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm (2013)
[24] Strola S.A., Marcoux P.R., Schultz E., Perenon R., Simon A.C., Espagnon I.., Allier C.P., Dinten J.M., "Differentiating the growth phases of single bacteria using Raman spectroscopy", paper 8939-4, upcoming BiOS Conference 2014 (2014)
[25] Maquelin K., Kirschner, C., Choo-smith L.P., Van den Braak N., Endtz H.P., Naumann D., Puppels G.J., "Identification of medically relevant microorganisms by vibrational microscopy", J. Microbiol. Meth. 51,255-271 (2002)