Direct identification of clinically relevant bacterial and yeast microcolonies and macrocolonies on solid culture media by Raman spectroscopy

Isabelle Espagnon, Denis Ostrovskii, Raphaël Mathey, Mathieu Dupoy, Pierre Joly, Armelle Novelli-Rousseau, Frédéric Pinston, Olivier Gal, Frédéric Mallard, Denis Leroux

Journal of Biomedical Optics 19(2), 027004 (February 2014)

Direct identification of clinically relevant bacterial and yeast microcolonies and macrocolonies on solid culture media by Raman spectroscopy

Isabelle Espagnon,[a] Denis Ostrovskii,[b] Raphaël Mathey,[b] Mathieu Dupoy,[c] Pierre L. Joly,[c] Armelle Novelli-Rousseau,[b] Frédéric Pinston,[b] Olivier Gal,[a] Frédéric Mallard,[b] and Denis F. Leroux[d,*]
[a]CEA, LIST, Département Métrologie, Instrumentation et Information, 91191 Gif-sur-Yvette, France
[b]bioMérieux, Technology Research Department, 5 rue des Berges, 38000 Grenoble, France
[c]CEA, LETI, MINATEC Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
[d]bioMérieux, Technology Research Department, Chemin de l'Orme, 69280 Marcy l'Etoile, France

Abstract. Decreasing turnaround time is a paramount objective in clinical diagnosis. We evaluated the discrimination power of Raman spectroscopy when analyzing colonies from 80 strains belonging to nine bacterial and one yeast species directly on solid culture medium after 24-h (macrocolonies) and 6-h (microcolonies) incubation. This approach, that minimizes sample preparation and culture time, would allow resuming culture after identification to perform downstream antibiotic susceptibility testing. Correct identification rates measured for macrocolonies and microcolonies reached 94.1% and 91.5%, respectively, in a leave-one-strain-out cross-validation mode without any correction for possible medium interference. Large spectral differences were observed between macrocolonies and microcolonies, that were attributed to true biological differences. Our results, conducted on a very diversified panel of species and strains, were obtained by using simple and robust sample preparation and preprocessing procedures, while still confirming published results obtained by using more complex elaborated protocols. Instrumentation is simplified by the use of 532-nm laser excitation yielding a Raman signal in the visible range. It is, to our knowledge, the first side-by-side full classification study of microorganisms in the exponential and stationary phases confirming the excellent performance of Raman spectroscopy for early species-level identification of microorganisms directly from an agar culture. [DOI: 10.1117/1.JBO.19.2.027004]

Keywords: Raman spectroscopy; rapid microbiology; species identification; bacteria colonies; culture medium; classification algorithms.

Paper 130538R received Jul. 31, 2013; revised manuscript received Nov. 27, 2013; accepted for publication Dec. 31, 2013; published online Feb. 12, 2014.

1 Introduction

In vitro microbiological diagnostics are still heavily relying on time-consuming cultivation of microorganisms to identify infectious agents and to prescribe therapeutic antibiotics against the diseases they cause. Because of long turnaround time (TAT), clinicians prescribe broad-spectrum antibiotics prior to the availability of a more precise diagnosis facilitating a more targeted therapy. In addition, pathogens accumulate multiple antibiotic resistance traits. This phenomenon added to the fact that fewer new antibiotics are being discovered constitutes a major public health problem. Reducing the TAT and the time needed for the microbial identification from 24 h or more to approximately 6 h would be a valuable step in the right direction.

Raman spectroscopy is a technology with strong assets for the use in *in vitro* diagnostic (IVD) applications as it is a sensitive technique amenable to automation,[1] nonintrusive and possibly nondestructive assuming a proper selection of acquisition parameters. In microbiology, the ultimate sensitivity has been demonstrated as good quality Raman spectra can be acquired from a single bacterium,[2] therefore even suggesting the possible elimination of culture as a whole. The fact that the technology may be nondestructive is another key point since resistance and susceptibility testing of pathogens is systematically conducted after identification. Another key element is the possibility to perform real-time analysis of pathogens directly on the culture medium. This would allow for further cultivation or immediate abrogation of the process in case no further characterization would be required. In addition to the clinical value, this would reduce the cost of running a clinical laboratory and it could increase the process traceability and robustness.

The questions raised by the direct measurement of microorganisms on solid culture media at different development stages were already addressed in earlier works. In 2000, Maquelin et al.[3] demonstrated for the first time the possibility to discriminate at the species level between four species by direct measurement of Raman spectra from 6-h old microcolonies on the solid culture medium. The possibility to cluster the data in definite classes based on microcolonies' Raman spectra was demonstrated through hierarchical clustering analysis (HCA). In 2002, Maquelin et al.[4] conducted a full classification study on 6-h old microcolonies of five yeasts (42 strains) belonging to the *Candida* genus, demonstrating the possibility to identify at the species level with a high prediction accuracy ranging from 97% to 100%. Classification was performed using a rather complex methodology based on the use of four linear discriminant analyses (LDA), each based on a separate model, requiring to

*Address all correspondence to: Denis Leroux, E-mail: denis.leroux@biomerieux.com

run *a posteriori* an HCA. An orthogonalization procedure had been used to correct for medium and water contributions. This step, deemed necessary by the author, requires 60 min of acquisition time on bare medium, presumably a one-time procedure for a given medium. A 97% average correct identification level was reported. In 2003, Maquelin et al.[5] conducted a more extensive study, building a larger reference database of bacteria and yeasts commonly detected in bloodstream infections, collected from positive hemocultures and grown 6 to 8 h on the solid culture medium and then directly analyzed by Raman spectroscopy. This study resulted in a correct classification level of 92.2%, after the analysis of 115 strains grouped in 11 classes (some of those classes included more than one species and 17 strains were excluded from the comparison because the phenotypic identification yielded a species not included in the database). Lowest identification rates were observed at 80% for *Enterobacter aerogenes* and *Enterobacter cloacae*. The total acquisition time reported per sample (50 spectra per sample for 10 replicates from five colonies) was 25 min not including the orthogonalization step. All data processing was conducted on first derivative spectra, as a way to remove the fluorescence background, and the classification analysis was conducted according to a leave-one-strain-out cross-validation method, what we call a "stringent" mode. In the three studies mentioned above, a procedure[6] based on vector algebra was used to subtract unwanted signals from the medium (although a highly confocal optical setup was being used to minimize signal contribution originating from the culture medium and variations in water content).

In all those works, an 830-nm near-infrared (NIR) laser was used to minimize sample auto-fluorescence. Unfortunately, classification performance obtained from spectra after fluorescence suppression but without prior correction for medium and water variation was not shown, preventing the quantification of the benefits of such an approach.

In 2001, Choo-Smith et al.[7] conducted a thorough study to compare compositional heterogeneities in microcolonies and macrocolonies cultured 6, 10, and 24 h. Taking advantage of the high spatial resolution provided by Raman spectroscopy, a precise depth-profile analysis of the colonies was conducted, showing that microcolonies are more homogeneous than macrocolonies and therefore presumably better suited for classification studies. Levels of RNA and glycogen were shown to differ depending on the growth stage, with young bacteria being characterized by a higher metabolism and therefore a higher RNA content compared to old bacteria. Old colonies, being constituted of a mix of young and old bacteria, appeared to be more heterogeneous in their biochemical composition. In this study, no compensation of the signal from water or medium was attempted (probably because judged unnecessary when conducting a mere clustering study in opposition to identification).

In 2008, Samek et al.[8] analyzed 24-h old macrocolonies of *Staphylococcus epidermidis* directly on Mueller–Hinton (MH) agar in a mostly qualitative study with proposed band assignments and concluded that, based on the relative ratios of Raman peaks, differentiation of *S. epidermidis* should be achievable. In 2012, Almarashi et al.[9] analyzed colonies from at least 30 colonies of four strains directly on Columbia blood agar (CBA), using a 785-nm excitation wavelength. Despite the medium complexity due to the presence of blood, colonies of the same bacterial species and/or strains clustered together in a distinct domain, although with a higher data dispersion attributed to the heterogeneity of the colonies (in accordance with Choo-Smith results[7]).

It is worthwhile to notice that Raman spectra of solid culture media show peaks in a similar Raman-shift region of interest compared to bacteria. Marotta and Bottomley[10] clearly showed some similarities between surface-enhanced Raman spectra acquired on 14 different solid culture media and spectra of *Escherichia coli*. This is easily understandable as organic nutrients have similar chemical structures as bacterial constituents. It is very difficult to estimate a posteriori to what extent the culture medium contributed to the sample Raman signal. This extent is presumably small as the optical configuration being used is highly confocal (axial resolution smaller than the colony height).

Single cell analyses have shown the impact on the Raman signal of biochemical composition changes occurring at different bacterial growth stages. Xie et al.[11] showed, using Laser Tweezer Raman Spectroscopy (LTRS), that nucleic acid and protein Raman signals varied with the growth stage because of changes in metabolism. It was proposed that for unsynchronized cultures, growing the cells to stationary phase would help to improve the identification. Those qualitative results were confirmed and quantified by Moritz et al.[12] who conducted a single-cell analysis by LTRS. *E. coli* was sampled at 2-h intervals after inoculation, allowing for a more accurate kinetic analysis of nucleic acid and protein Raman bands (confirming the results of Talukder et al.[13] reported an increase of nucleoid proteins levels when comparing *E. coli* cultured for 5 and 24 h). When reaching the stationary stage, it is reported that both protein synthesis and cell division will eventually be reduced.

Those observations were confirmed by other studies conducted on yeasts[14] in the first 3 h of inoculation transitioning from the lag to the early exponential phase, or when comparing 6- and 18-h old single cells of *S. epidermidis*.[15] In addition, Huang et al.[16] showed that although Raman is sensitive enough to discriminate bacteria harvested from 4 and 24 h cultures (respectively, representative of the exponential and stationary phases), it did not prevent discrimination between the three species of the studied model. It has been also shown[17] that correct identification rates (CIRs) of *Bacillus* species were essentially unaffected by time of growth between 24 and 48 h (supporting the fact that most metabolic activity changes occur in the first 24 h).

Modifications of glycogens observed at the surface of *E. coli* microcolonies had also been reported to vary significantly during prolonged culture as well as formation of extra-cellular polymeric substances (EPS). Eboigbodin and Biggs[18] conducted a systematic analysis of free and bound EPS using IR vibrational spectroscopy at different growth stages after 6 and 24 h: *E. coli* was shown to produce very little EPS while *Bacillus subtilis* showed large changes in carbohydrate/protein ratio. Ciobotă et al.[19] later demonstrated that the presence of polyhydroxybutyrate in microbial cells did not prevent Raman identification as long as the microorganisms were in the exponential growth phase.

Most studies converge on the concept that direct on-agar Raman identification at the species level (and even possibly at the strain level) performs well in the limited context of each study, concluding that an effort to standardize, to extend the size of the database, and to better understand individual spectral contributions are needed to move the field forward.

In our study, we first chose not to compensate for possible interference by medium. Before going to an elaborated procedure to attempt to correct for possible media effects, we thought that the first step was to establish the level of performance achievable "as is" and we are eager to report rather high CIRs without any correction for the underlying medium. Besides, the risk of losing information by performing spectra correction cannot be ruled out. We corrected for a background, mainly consisting of fluorescence signal, with no additional attempts to quantify or identify sources of variability originating from the acquisition procedure or the sample itself. In order to reduce the biological variations, we chose instead to work at constant growth time (exactly 6 or 24 h) on the single culture medium and to process the samples immediately, without any storage period. This was taken as a precautionary measure in an effort to simplify the identification problem and to avoid confounding factors, but it is not mandatory (no decrease in the Pearson correlation coefficient between fresh cultures on one side and fresh cultures stored up to 5 days at 4°C on the other side). We chose to carry out our work on trypticase soy agar (TSA) as it is a very generic medium, but we have reasons to believe that similar results could be achieved at least on most nonchromogenic media as the prior art teaches us that discrimination is achievable, at various levels, on a variety of media: Sabouraud agar[3–5] (MH), and even CBA despite the presence of hemoglobin.[9]

Several points differentiate our study from the prior art, besides the obvious diversity of the species and strains and the large number of strains included in the database. First, no correction for possible agar contribution was attempted as done by the group of Puppels[5] by a systematic orthogonalization step, circumventing the need to perform control measurements on the culture medium and therefore saving the time and demonstrating some level of robustness. Inversely, a large panel of preprocessing and classification methods was studied for systematic comparison. Second, spectra were acquired at 532-nm excitation wavelength instead of 830 or 785 nm as done in other on-agar studies. Although it is commonly accepted that the higher fluorescence level observed at shorter wavelength should decrease the Raman signal-to-noise ratios (SNRs) and the classification performance, we demonstrated a good classification at 532 nm. Benefits of working at 532 nm include: (i) improved spatial resolution leading to a smaller confocal depth, (ii) ultimately decreased acquisition time, and (iii) staying in the convenient "visible range" (no need for NIR optics and large Raman shift range).

2 Materials and Methods

2.1 *Choice of Species and Strains*

Nine bacterial and one yeast species were selected which all belong to the ones most frequently encountered in clinical microbiology. They include six Gram-negative species comprising three *Enterobacteriaceae* species (*E. coli*, *E. aerogenes*, and *E. cloacae*) and three non*Enterobacteriaceae* species (*Acinetobacter baumannii*, *A. johnsonii*, and *Stenotrophomonas maltophilia*). Some of these species are known to be difficult to identify by the phenotypic methods. Four Gram-positive species were added, including three bacterial species (*Bacillus cereus*, *Staphylococcus aureus*, and *S. epidermidis*) and one yeast species, selected as an eukaryotic outlier (*Candida albicans*). Eight well-characterized strains were selected per species. They were provided by the American Type Culture Collection (Manassas, VA, USA), by the Centers for Disease Control (Atlanta, GA, USA) or were taken from the bioMérieux culture collection.

2.2 *Sample Preparation*

Efforts were made to minimize variation in sample handling: standardized time of growth, short elapsed time between culture and measurement, no storage, and single culture medium originating from the same lot. Strains were stored at −80°C in broth containing glycerol. Before Raman analysis, a first overnight culture was performed on TSA (bioMérieux Ref. 43011) at 37°C (except for *A. baumannii* and *A. johnsonii* that were grown at 30°C). This first culture was stored at 4°C and constituted a "stock culture" which was used as source during the measurement campaign (3 weeks of storage at most). TSA was selected as the preferred artificial solid culture medium as it expressed a weaker fluorescence or less pronounced Raman features than other tested suitable media from bioMérieux: Columbia Agar with 5% Sheep Blood (Ref. 43041); mannitol-salt agar containing 2 μg/mL of oxacillin (Ref. 43671); Drigalski agar (Ref. 43341); medium dedicated to cultures of *E. coli, Proteus, Streptococci* (CPS3, Ref. 43541); lactose agar with Bromocresol purple (Ref. 43021); and TSA + 5% sheep blood (Ref. 43001).

For the macrocolony study, the preparation consisted in picking up colonies from a stock culture, streaking on TSA and culturing for 24 h. For the microcolony study, an intermediate culture was done by picking up colonies from stock culture on TSA and culturing them overnight to revitalize the bacteria. This overnight culture was followed by a 6-h long culture to obtain microcolonies. The time elapsed between the end of growth and reading did not exceed 30 min. Culture temperature was 37°C for all species, except for *A. baumannii* and *A. johnsonii* (30°C).

2.3 *Spectroscopic Device and Measurements*

Raman spectra were acquired using a LabRam ARAMIS (Horiba Jobin Yvon, Villeneuve d'Ascq, France) micro-spectrometer equipped with a 532-nm laser (Ventus LP 532 50 mW, Laser Quantum, Stockport, UK) and a −70°C Peltier-cooled CCD detector (Synapse TE cooled, Horiba Jobin Yvon). The acquisition spectral window ranged from 395 to 3075 cm^{-1}, given the choice of a 600 line/mm grating. The 1024 channels yield a spectral resolution ranging from 3.07 to 2.65 cm^{-1} in the $[470\,\text{to}\,1700]$ cm^{-1} region selected for data processing. Optimal acquisition conditions, established experimentally, appeared to be quite different between macrocolonies and microcolonies. The parameters used, respectively, on macrocolonies and microcolonies are summarized in Table 1.

Petri dishes with bacterial cultures were directly transferred from the incubator to the spectrometer and Raman spectra were recorded directly from the grown colonies without any additional preprocessing. To account for most variations in bacterial samples, as well as to avoid significant variation of the material during Raman measurements, the following criteria were applied:

- Since bacteria continue their growth even at room temperature, total measurement time for every Petri dish did not exceed 1 h.

Table 1 Parameters and conditions of Raman spectra acquisition for microbial macrocolonies and microcolonies.

Colony type	Time of growth (h)	Microscope objective (x/NA)	Confocal hole (μm)	Axial confocal thickness (μm)	Focus offset (μm)	Laser power sample (mW)	Acquisition time (s)	Points per colony
Macro	24	50/0.5	800	60	−20	11	5×20	6 to 8
Micro	6	100/0.8	200	5	−3 to −8	36	5×15	1 to 4

- To account for possible intercolony variations, Raman spectra were recorded from several isolated colonies on the Petri dish.
- To account for possible intracolony variations, Raman spectra were taken from several points within a colony. On macrocolonies, an automated acquisition was usually possible with a distance between points of 20 μm.
- To probe as much microbial material as possible, acquisition focusing was set inside the colony by applying a systematic vertical offset relative to the surface of the colony.
- The number of spectra before averaging and the integration time of each single spectrum were optimized to provide an SNR sufficient for further data processing within the shortest possible time. Five successive spectra were acquired for every single measurement at constant focusing. This recording sequence enabled to eliminate offline saturated individual spectra, which sometimes occurred at the beginning of the sequence, as strong and rapidly decreasing fluorescence signal was observed for a limited number of species on macrocolonies due to photobleaching (chemical photodegradation of highly unsaturated organic molecules present in the sample, often observed in the conditions of acquisition). The mean of the unsaturated spectra acquired at a given position was used as the input spectrum for data processing. Hence, we were able to apply an identical acquisition protocol for all species, with constant laser power and integration time, while both coping with occasional signal saturation and avoiding too low value of SNR.

2.4 *Indicators of Spectra Quality*

Three indicators were used to assess the quality of a spectrum: the SNR, the relative standard deviation (STD), and the Pearson coefficient of correlation (R). The first quality indicator, SNR, is derived from the signal, defined as the mean of the net spectrum in the region of interest [see Sec. 2.5 (iii)], and the noise, defined as the STD of the net spectrum between 1800 and 2000 cm^{-1} (a region deemed free of Raman signal). The second index, STD, used to estimate reproducibility between spectra (preferably net spectra) of the same species, is defined as the within-species STD of each spectral channel, averaged on all channels. It directly provides a relative (mean) STD for spectra previously normalized by their own mean intensity. The third quality index, used to evaluate similarity between spectra originating from a given strain or species, is the Pearson correlation coefficient (R), calculated for each pair of spectra in a given dataset of spectra. Plotting the distribution of R enabled a quick visualization of the homogeneity of a dataset. The mean value of these coefficients is an indicator of the similarity of spectra within the dataset.

2.5 *Data Preprocessing*

The preprocessing of the initial Raman spectra is important for the subsequent data analysis and classification since it eliminates or reduces significantly the impact of the nonbacterial variability (e.g., instrumental or stochastic). Four preprocessing steps were used in this study:

- suppression of "cosmic" spikes;
- correction of possible wavenumber shift of the spectra, which has instrumental origin;
- extraction of the signal of interest (by deriving, or subtracting background, or the raw signal itself), accompanied with smoothing to reduce random spectral noise;
- normalization of spectral intensities to exclude the effect of varying laser power, focusing grade, sample density, etc.

All preprocessing was performed automatically in the R software environment[20] using the existing or developed in-house routines. Elapsed time for preprocessing of 100 spectra did not exceed 20 s, including 7 s for cosmic spikes suppression, and 12 s for the background suppression.

i. Suppression of spikes due to gamma rays from surrounding radioactivity and cosmic rays impinging the CCD detector[21] (the so-called "cosmic" spikes) was the first step of preprocessing. For each spectrum, a peak search was done using the second derivative of the spectrum. Identification of these spikes in the peak list was done from smoothed spectra as the spikes, being thinner and often of larger intensities than Raman peaks, decrease more rapidly upon smoothing than Raman peaks. Detected spikes were replaced by a linear interpolation of the surrounding signal. This method was preferred to the more usual one of detecting spikes by comparing multiple spectra acquired successively on a given spot of the colony because of possible photobleaching between successive spectra.

ii. Wavenumber shifts of the same spectral features between different spectra were observed within and between days. Their origin was mainly instrumental, as shown by their time dependence. The shift is constant for the entire spectrum if expressed as a number of spectrum channels. Selected peaks of each net spectrum [see (iii)], at approximate fixed positions, were fitted by a Gaussian function with a linear background. The applied realignment is the mean of the shift values of

the selected peaks compared to their fixed reference positions. Only peaks common to all spectra and characterized by a sufficient SNR value were selected which limited their number to two. The two peaks were at 746 and 1127 cm^{-1} for macrocolonies and at 783 and 1003 cm^{-1} for microcolonies.

iii. To select the signal of interest before classification, several preprocessing methods were tested and compared, including very simple ones: simple smoothing without further extraction (the signal thus preprocessed is called "raw spectra"), background estimation by peak-clipping and subtraction ("net spectra"), and first and second (smoothed) derivative spectra. Smoothed spectra as well as first and second derivatives were calculated using Savitzky–Golay filters[22] (degree 2, on 13 points). The peak-clipping algorithm for background suppression was based on the SNIP algorithm,[23] already successfully applied to Raman spectra processing,[24,25] here with an initial smoothing done by the same Savitzky–Golay filter and the neighborhood window reduced to a radius of 1 (this implies more iterations but less parameters). By "background" is meant a broad, slowly varying signal combined with the Raman signal of interest and presumably very variable and poorly informative. It is mainly due to the fluorescence of the microorganisms themselves and possibly the underlying medium, to the CCD background signal, and to various diffusive, parasite light sources. The subtracted signal unavoidably also includes some part of the Raman signal itself, especially in the presence of broad bands, with presumably little interesting information.

iv. The last step of preprocessing was the spectrum normalization which was essential to compare spectra as the laser power or the thickness and density of the observed colony may vary, therefore preventing a robust control of the signal intensity. The region of interest providing the best classification results was the $[470\,\text{to}\,1600]$ cm^{-1} region for macrocolonies and the $[650\,\text{to}\,1700]$ cm^{-1} region for microcolonies. Normalization was done by dividing each signal by its mean (or by the mean of its absolute value in the case of first or second derivatives) in this region.

2.6 *Classification*

Two types of cross-validation were carried out in the so-called "stringent" and "nonstringent" modes.

When performing a classification at the species level in the stringent mode, all spectra belonging to the strain being classified were previously removed from the reference database, thus preventing an artificially almost perfect match. Our biological model contained eight strains per species and 10 species, so 10 strains (one per species) were randomly chosen and simultaneously removed from the reference spectra to constitute the test group. An eightfold cross-validation was therefore sufficient to test all spectra.

In the nonstringent mode, all strains were represented in the reference database. In this case, one-eighth of the spectra of each strain was randomly chosen, removed from the reference database, and tested. In those conditions, an eightfold cross-validation was also sufficient to test all spectra.

The stringent mode is thought to be more representative of a clinical situation, where the exact microbial strain present in the sample of interest is most often absent from the reference database used for the identification. In each mode, 10 cross-validations were done, with randomly chosen eightfold partitions. It allowed for the calculation of mean and STD for the CIR. It has been verified that these values are not significantly modified when the number of cross-validations is increased (e.g., to 100).

Several classification algorithms were tested:

i. The Euclidean distance (ED), where an averaged reference spectrum (or signal of interest), is calculated for each species, and the nearest reference spectrum gives the selected species.

ii. The k-nearest neighbors (KNN), with $k = 3$, where all reference spectra are kept (without averaging), and a vote between the k-nearest (in the sense of ED) reference spectra decides the selected species (function "knn" of the *R* package "class"[26]).

iii. The LDA provided by function "lda" of the *R* package "MASS."[26]

iv. The regularized quadratic discriminant analysis (rQDA), where the selected species is given by the smallest Mahalanobis distance, based on the variance-covariance matrix calculated (hence regularized), for each species, using the $n - 1$ discriminant variables provided by LDA (where $n = 10$ is the number of species).

v. The support vector machine (SVM) with a vote between the $n(n-1)/2$ one-versus-one SVMs (function "svm" of the *R* package "e1071," interfacing the "LIBSVM" library[27]).

Classification results of each method are summarized by the mean of the CIRs of all species. They are also given with more details in the form of a confusion matrix which consists in a cross-table of actual and found species membership, with classification rates expressed in percentages of the number of tested spectra in the actual species. Sensitivity and specificity for a given species can be obtained from this confusion matrix, since sensitivity simply is the corresponding CIR, and specificity is given, after removing the row and column of that species, by summing in each row and then averaging. Nevertheless, since the classification scheme consists of choosing one class among 10, the specificity is naturally high (90% for a random classifier) whereas getting a high sensitivity (or CIR) is much more challenging (a random classifier would give 10%).

3 Results

3.1 *Databases Description*

The databases of macrocolonies and microcolonies contain 2533 spectra and 1813 spectra, respectively (Table 2), acquired for the same 80 strains from 10 species. Each raw spectrum is subdivided into 1024 channels corresponding to a Raman shift ranging from 395 to 3075 cm^{-1}.

3.2 *Indicators of Spectral Quality*

The mean SNR values per species, for both macrocolony and microcolony bases, are listed in Table 3. They show that single spectra acquired on microcolonies are of better quality than those from macrocolonies since their SNRs are approximately

Table 2 List of species, code, number of strains, and number of acquired spectra per species, in macrocolonies and microcolonies databases.

Species	Code	No. of strains	No. of spectra	
			Macrocolonies	Microcolonies
Acinetobacter baumannii	ACN-BAU	8	257	182
Acinetobacter johnsonii	ACN-JOH	8	236	192
Bacillus cereus	BAC-CEU	8	190	180
Candida albicans	CAN-ALB	8	261	143
Enterobacter aerogenes	ENT-AER	8	260	213
Enterobacter cloacae	ENT-CLC	8	264	191
Escherichia coli	ESH-COL	8	288	193
Staphylococcus aureus	STA-AUA	8	235	202
Staphylococcus epidermidis	STA-EPI	8	288	180
Stenotrophomonas maltophilia	STE-MLT	8	254	137
Total		80	2533	1813

2.7-fold higher than those of the latter (average SNR of 10.3 for microcolonies versus 3.8 for macrocolonies). Moreover, the relative STD of spectra in each species (Table 4) clearly demonstrates that the within-species reproducibility is significantly higher with microcolonies (STD of 0.10) than with macrocolonies (STD of 0.17). This is further illustrated by the Pearson correlation analysis of the *E. coli* spectra (Fig. 1) quantifying the observed similarities among macrocolonies, among microcolonies, and between macrocolonies and microcolonies. It was observed that spectra correlated with an average R value of 0.97 and 0.98 for macrocolonies and microcolonies, respectively, while the correlation dropped dramatically to 0.48 when comparing the macrocolonies to microcolonies, which clearly shows that their spectra are very different. The Pearson coefficient distribution was narrower for microcolonies, a further indication of the better reproducibility of the microcolony dataset.

3.3 *Macrocolonies*

Figure 2 presents raw spectra of *E. coli* strain API 9203096, before and after normalization. Because of the high variability in spectral background in both cases, it was deemed essential to remove the background before classifying, as is commonly performed in spectroscopic analysis.

The mean, normalized, preprocessed spectra for each species are shown in Fig. 3, for the (smoothed) raw spectra, the net spectra after background suppression, and the first derivative (second derivative is not shown). It is important to notice that nearly the same peaks are present in all species. The main differences between species lie in the relative abundances, an observation easily explained by the fact that all species have similar biochemical compositions and are therefore characterized by identical chemical bounds, differing mostly by the relative abundance of particular biogroups. One exception is presented by *S. aureus*, where six of the eight strains had the characteristic golden pigmentation, due to carotenoids and associated with the intense and very specific Raman peaks[28] at 1160 and 1520 cm^{-1}, clearly visible in the averaged signal (Fig. 3). The corresponding principal component analysis (PCA) plots (Fig. 3, bottom) suggest that the background subtraction step, by the peak-clipping algorithm or by deriving the spectra, substantially improves the discrimination between species.

The CIRs of the five classification methods tested on these four sets of raw or preprocessed spectra are shown for stringent and (partially) for nonstringent modes in Table 5 (upper part). Figure 4 shows the confusion matrix obtained for the stringent mode in the best configuration, which is the rQDA method applied to the first derivative spectra. In this case, the CIR is $94.1 \pm 0.6\%$. It decreases to $93.3 \pm 0.9\%$, when the realignment step is skipped. Same results are presented in terms of sensitivity (i.e., CIR) and specificity in Table 6. As anticipated in Sec. 3, specificities are very high, with an average value of 99.3%.

We observe that the rQDA method still provides the best CIRs for the net and second derivative signals. It is also noticeable that the CIRs obtained with the raw spectra and

Table 3 Average signal-to-noise ratio (SNR) per species for macrocolonies and microcolonies studies.

	ACN-BAU	ACN-JOH	BAC-CEU	CAN-ALB	ENT-AER	ENT-CLC	ESH-COL	STA-AUA	STA-EPI	STE-MLT	Av.
Macro	3.7	3.5	2.2	3.9	4.1	4.2	4.0	4.6	2.9	5.0	3.8
Micro	12.9	9.7	12.5	6.8	13.5	10.2	12.4	10.1	8.6	5.8	10.3

Table 4 Mean standard deviation (STD) of normalized net spectra per species for macrocolonies and microcolonies studies.

	ACN-BAU	ACN-JOH	BAC-CEU	CAN-ALB	ENT-AER	ENT-CLC	ESH-COL	STA-AUA	STA-EPI	STE-MLT	Av.
Macro	0.14	0.21	0.14	0.12	0.11	0.11	0.14	0.47	0.13	0.08	0.17
Micro	0.09	0.09	0.08	0.15	0.07	0.11	0.08	0.08	0.11	0.11	0.10

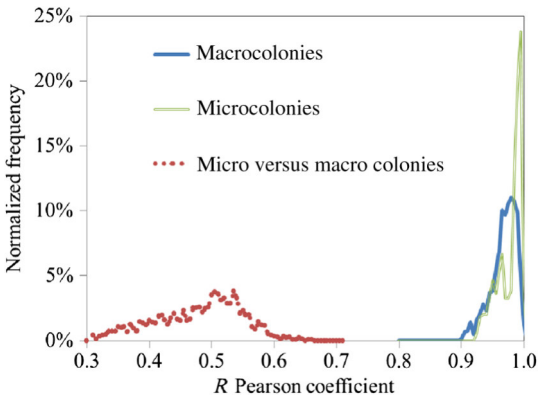

Fig. 1 Frequency distribution of *R* Pearson coefficients when correlating normalized net spectra of *Escherichia coli* (i) from microcolonies 2×2, (ii) from macrocolonies 2×2, and (iii) from microcolonies with spectra from macrocolonies.

the more advanced classification methods (92.0% with LDA, 92.7% with rQDA, and 93.3% with SVM) are barely lower than the best CIR. This is very different from the results obtained with the ED and KNN, where subtracting the background (by the peak-clipping algorithm or by deriving) clearly improves the CIR, as suggested by the PCA plots. This is a direct consequence of the close relationship between PCA and ED.

In the nonstringent mode, the best CIR is $99.8 \pm 0.1\%$ and was obtained with SVM. It is worthwhile to notice that it was obtained for the raw spectra and is very close to the one obtained with SVM on first derivative ($99.7 \pm 0.1\%$).

3.4 *Microcolonies*

Figure 5 shows the same set of raw and preprocessed spectra as in Fig. 3, also with the corresponding PCA plots, here for the microcolonies. We also observed in the PCA plots that the background subtraction by peak-clipping or the derivative improves the discrimination between species, although it is seemingly poorer than with macrocolonies. Interestingly, the yellow strains of *S. aureus* did not have detectable Raman peaks specific to carotenoids. This is in agreement with the absence of observable pigmentation of those strains at the microcolony stage.

The CIRs for the microcolonies are shown in the lower part of Table 5 for the stringent and nonstringent modes. Figure 6 shows the confusion matrix obtained in the best stringent configuration, which is applying the SVM method to the first derivative spectra. The corresponding CIR is $91.5 \pm 0.4\%$ (decreasing to $88.7 \pm 0.4\%$ when the realignment is skipped). Once again, the ED and KNN are the only classification methods that showed a clear improvement when background signal was removed compared to raw spectra. In the nonstringent mode, the best CIR is $98.0 \pm 0.1\%$ and was obtained with SVM on raw spectra, as for macrocolonies.

4 Discussion

4.1 *Misclassifications and Taxonomy*

Species that are the most difficult to differentiate by Raman spectroscopy are also the ones being very close in their taxonomic position, as defined by using conventional phenotypic and molecular methods. With macrocolonies as well as with microcolonies, the lowest CIRs were observed inside the *Enterobacteriaceae* family (for *E. aerogenes*, *E. cloacae*, and *E. coli*) and inside the *Acinetobacter* genus (*A. johnsonii* and *A. baumannii*). Other significant errors occurred with *E. cloacae* instead of *A. johnsonii* for macrocolonies and *S. maltophilia* instead of *E. cloacae* for microcolonies. Confusions confined under the genus level accounted for 89% of all errors for macrocolonies and 44% for microcolonies. When errors inside the *Enterobacteriaceae* family are included, the proportion increased to 93% of all errors with macrocolonies and 85% with microcolonies. These results are similar to earlier findings[28] showing confusions between *Enterococcus faecalis* and *E. faecium*.

4.2 *Comparison Between Macrocolonies and Microcolonies and Influence of the Agar Signal*

For any given species, spectra acquired on microcolonies were very different from spectra acquired on macrocolonies. This was illustrated by the Pearson correlation analysis of the *E. coli* spectra (Fig. 1), which clearly implies that identification is only possible if the spectra of the tested sample are acquired at the same culture age as the reference spectra forming the database, at least for times of culture where the growth stage is expected to be very different as is the case in this study.

These differences are more directly shown in Fig. 7 for species *E. coli* by comparing the mean net spectra of macrocolonies and microcolonies. The Raman spectrum of TSA is also shown to demonstrate that the difference cannot be due to the contribution of the underlying agar as the TSA medium has a few characteristics unique peaks (although it also clearly

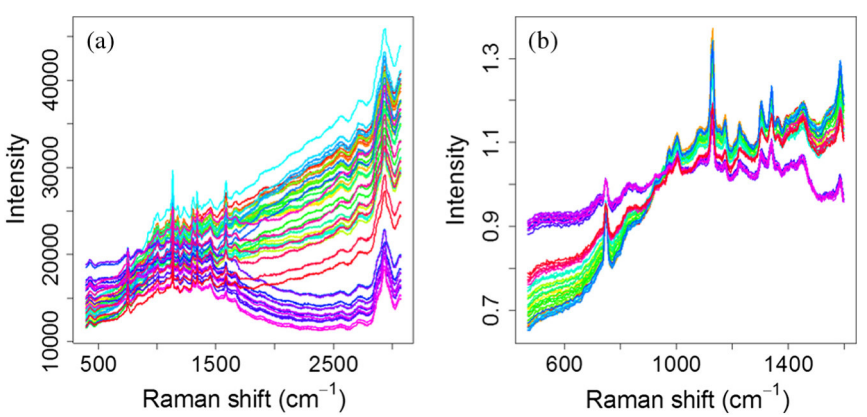

Fig. 2 Raw spectra of strain "ESH-COL API 9203096," (a) on the full Raman-shift range without normalization and (b) restricted and normalized on the region of interest (macrocolonies).

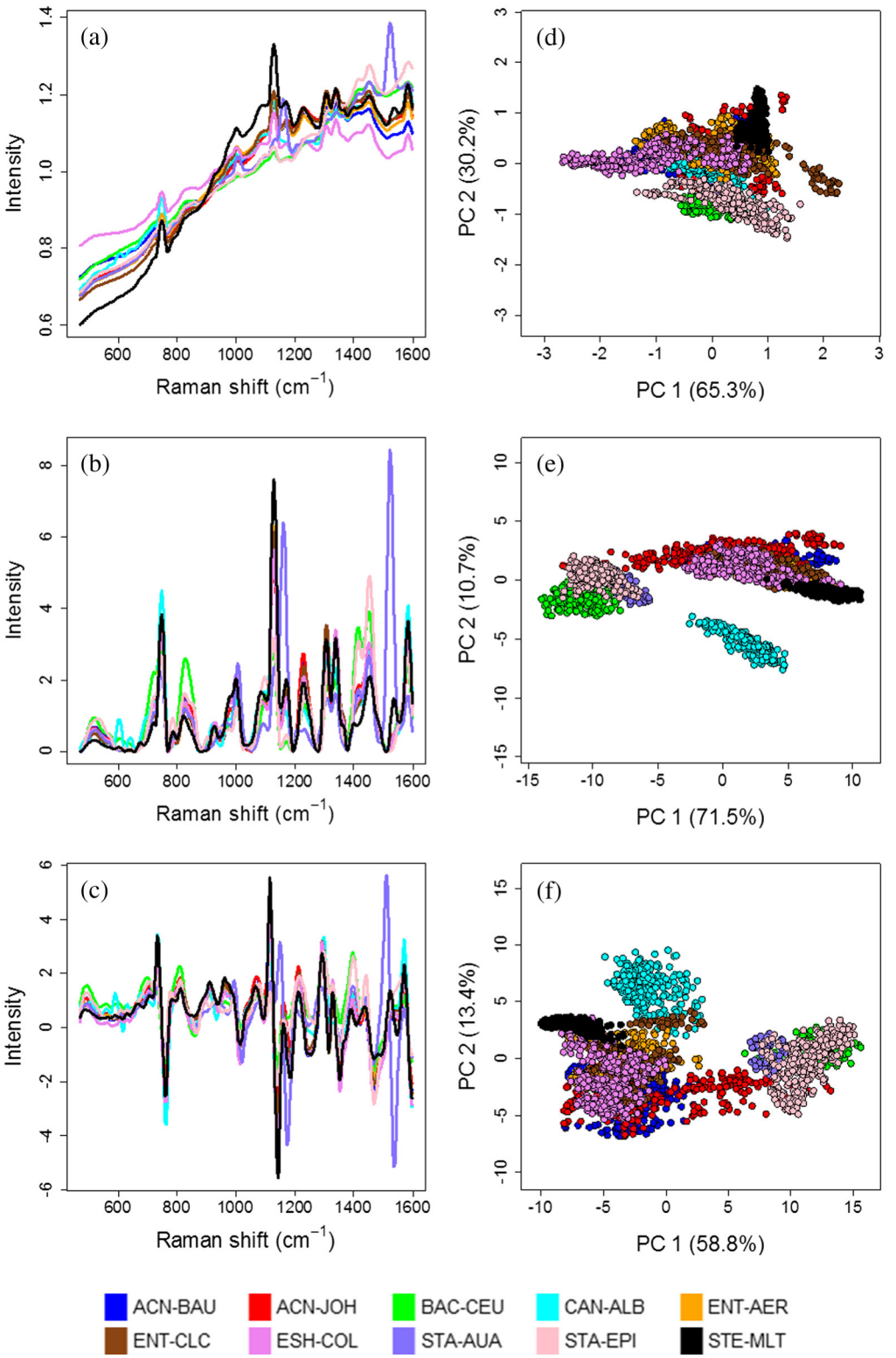

Fig. 3 Mean spectra per species for macrocolonies, for (a) raw spectra, (b) net spectra, (c) first derivative, and (d, e, f) corresponding PCA plots (first two components, with one color per species). NB: Orange strains of *Staphylococcus aureus* have been excluded from PCA to provide more details.

shows multiple peaks in the same region of interest as the colonies). We suggest that the large differences observed between microcolonies and macrocolonies are indeed due to the biological differences and not to the underlying growth medium (a result to be expected for microorganisms at different growth stages).

The following three arguments support our proposition:

- First, we noticed that the TSA medium [Fig. 7(a)] is showing a peak at 1414 cm^{-1} which is absent from spectra of both macrocolonies and microcolonies, an indication that

Table 5 Correct identification rates (CIR) obtained for the macrocolonies and microcolonies in stringent and nonstringent modes with the five classification methods and the four sets of preprocessed spectra.

	Stringent CV					Nonstringent CV	
	ED	KNN	LDA	rQDA	SVM	ED	SVM
				Macrocolonies			
Raw spectra	58.5 ± 0.3	77.7 ± 0.7	92.0 ± 0.6	92.7 ± 0.4	93.3 ± 0.4	62.1 ± 0.1	99.8 ± 0.1
Net spectra	74.1 ± 1.0	81.4 ± 0.5	89.1 ± 0.7	91.6 ± 0.6	88.3 ± 0.7	82.8 ± 0.1	99.0 ± 0.1
First derivative	74.4 ± 0.8	84.5 ± 0.5	91.2 ± 0.8	94.1 ± 0.6	91.6 ± 0.7	80.2 ± 0.2	99.7 ± 0.1
Second derivative	81.3 ± 0.2	80.0 ± 0.5	90.3 ± 1.0	92.4 ± 0.4	89.8 ± 0.5	84.8 ± 0.1	98.6 ± 0.1
				Microcolonies			
Raw spectra	66.9 ± 0.4	77.1 ± 0.2	88.3 ± 0.6	88.4 ± 0.7	90.8 ± 0.3	69.7 ± 0.2	98.0 ± 0.1
Net spectra	77.3 ± 0.4	83.9 ± 0.6	88.1 ± 0.5	87.2 ± 0.4	90.7 ± 0.4	81.7 ± 0.2	97.7 ± 0.3
First derivative	78.4 ± 0.4	85.2 ± 0.3	88.5 ± 0.6	88.2 ± 0.6	91.5 ± 0.4	82.7 ± 0.2	97.8 ± 0.2
Second derivative	79.3 ± 0.4	81.8 ± 0.5	88.7 ± 0.6	88.2 ± 0.7	88.9 ± 0.4	84.7 ± 0.1	95.8 ± 0.2

the culture medium might indeed not be significantly contributing to the measured Raman signal.

- Second, the microcolony average spectrum [Fig. 7(c)] shows some characteristic peaks (664, 781, 808, 1095, and 1569 cm^{-1}) that match the position and tendency (higher nucleic acid content observed in the exponential phase compared to the stationary phase) of the metabolic activity markers identified by Moritz et al.[12] and assigned to DNA and RNA nucleic acid bands (at 668, 783, 811, 1099, and 1578 cm^{-1}). Of the two very intense peaks present in macrocolonies at 745 and 1126 cm^{-1}, we have assigned the peak at 1126 cm^{-1} to C—N and C—C stretching mainly associated with proteins[12,29] but were not able to assign the peak at 745 cm^{-1}.
- Third, we have tested and rejected the assumption that differences between microcolonies and macrocolonies could be due solely to the underlying growth medium (to be expected if the confocal height happens to be too large or in case of improper focalization); as we failed to reconstruct the spectra of microcolonies using a simple linear combination of the TSA signal and of the pure macrocolony bacteria signal. The highest coefficient of the correlation R between the closest modeled spectrum and the microcolony was only of 0.53 for 60% bacteria/40% agar).

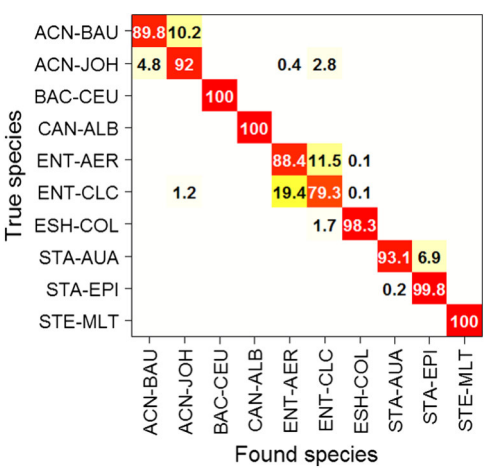

Fig. 4 Confusion matrix for macrocolonies with rQDA method in stringent mode on first derivative spectra (CIR = 94.1 ± 0.6%).

4.3 *Comparison of CIRs Between Macrocolonies and Microcolonies*

Another clear difference between macrocolonies and microcolonies lies in the fact that the CIRs from microcolonies were

Table 6 Sensitivity and specificity calculated from confusion matrix of Fig. 4 (rQDA on first derivative spectra for macrocolonies and stringent cross-validation).

	ACN-BAU (%)	ACN-JOH (%)	BAC-CEU (%)	CAN-ALB (%)	ENT-AER (%)	ENT-CLC (%)	ESH-COL (%)	STA-AUA (%)	STA-EPI (%)	STE-MLT (%)	Av. (%)
Sensitivity	89.8	92.0	100.0	100.0	88.4	79.3	98.3	93.1	99.8	100.0	94.1
Specificity	99.5	98.7	100.0	100.0	97.8	98.2	100.0	100.0	99.2	100.0	99.3

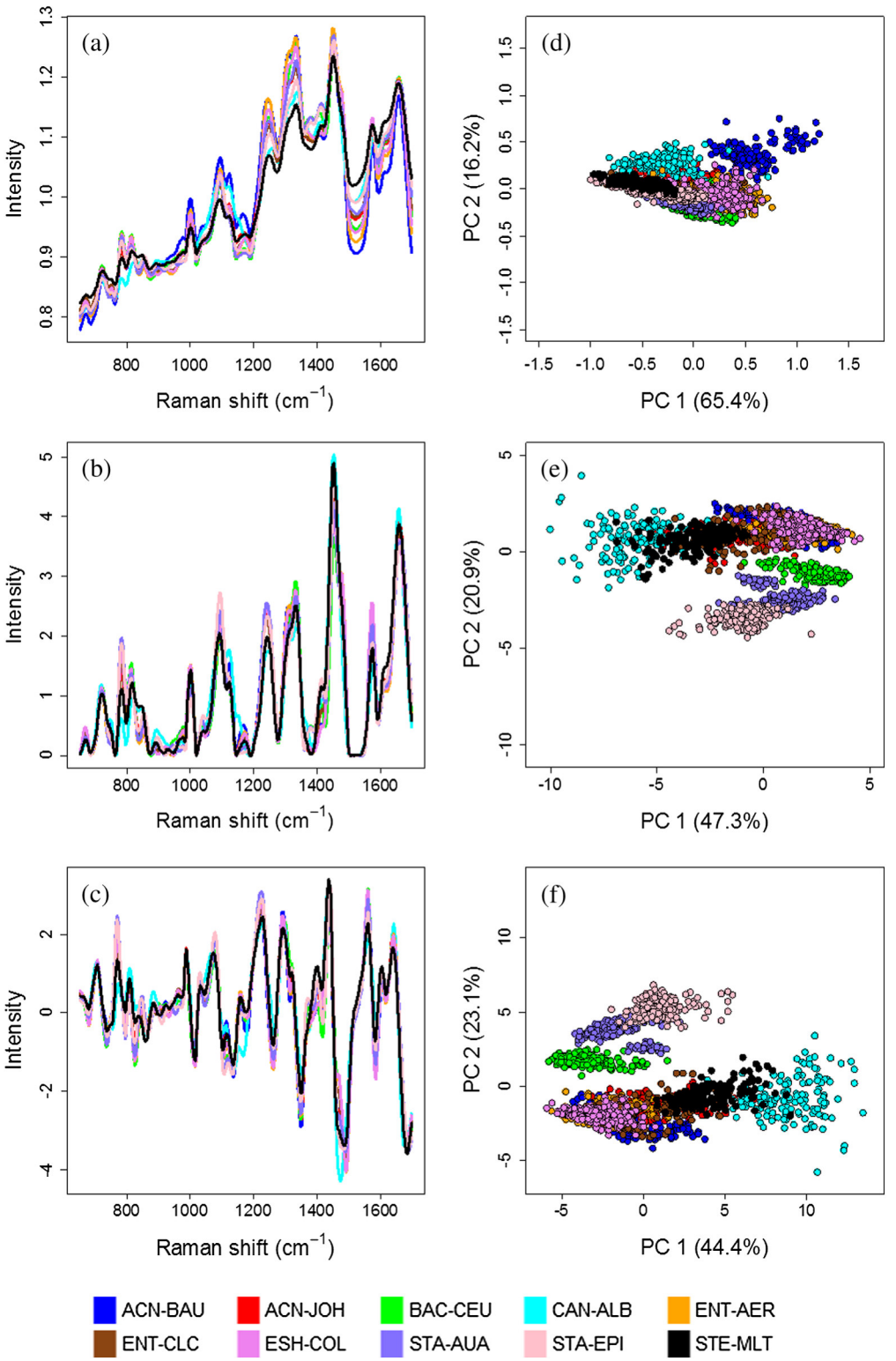

Fig. 5 Mean spectra per species for microcolonies, for (a) raw spectra, (b) net spectra, (c) first derivative, and (d, e, f) corresponding PCA plots (first two components, with one color per species).

significantly lower than those obtained from macrocolonies (see CIRs in Table 5 and confusion matrices in Figs. 4 and 6). The reasons for this observed drop of performance were not elucidated. Logical explanations could be a lower specific Raman signal from microcolonies because of the smaller quantity of biological material, hence a lower SNR, or a higher contribution of the underlying culture medium, due to the confocal volume extending beyond microcolony depth, hence a less specific signal. These assumptions are denied by the fact that despite the reduced overall signal, the spectra quality actually appeared to be better for microcolonies than for macrocolonies, since they showed a higher SNR (10 versus 4; see Table 3). Moreover, Fig. 1 illustrates (for *E. coli*) that the normalized net spectra were slightly better correlated within species for the

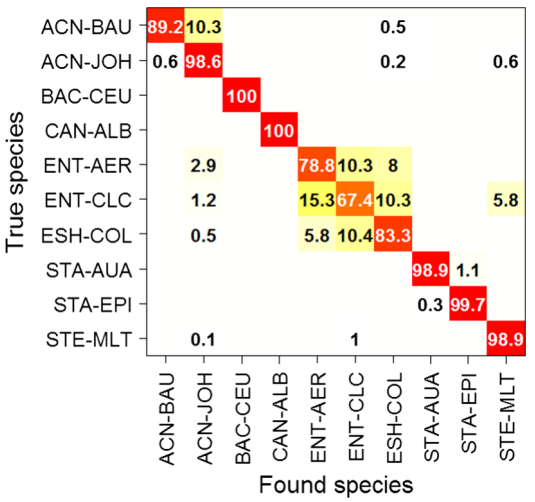

Fig. 6 Confusion matrix for microcolonies with SVM method in stringent mode on first derivative spectra (CIR = 91.5 ± 0.4%).

microcolonies than for the macrocolonies and Table 4 shows (for all species) that they had a lower STD on average (0.10 versus 0.17). This is in agreement with the already cited study by Choo-Smith et al.[7] who observed that microcolonies are more homogeneous in their composition than macrocolonies.

A possible way to reconcile the fact that lower CIRs are observed for microcolonies despite a lower dispersion inside each class and a higher SNR could be to assume that the microcolonies show less chemical composition differences between species than macrocolonies, therefore making the discrimination between species more difficult. Possibly the Raman peaks associated with a high metabolic activity are not that species-specific, otherwise discrimination would be improved as the relative importance of those marker peaks decreases in the stationary phase.

4.4 *Influence of Background Subtraction on Classification*

It was already mentioned that the best classification results (in stringent mode) were obtained for the derivative spectra, which are deemed to be background subtracted, with the rQDA and SVM methods (for macrocolonies and microcolonies, respectively), but that the same methods give barely lower CIRs on raw spectra (see Table 5). This seems contradictory with the PCA plots of Figs. 3 and 5 which suggest that the background subtraction substantially improves discrimination between species, as actually observed in the CIRs obtained with the ED and KNN methods. This logically questions the astonishingly small improvement in the classification performance of LDA, rQDA, and SVM methods between raw and background-subtracted spectra. Or, one could equivalently ask why they proceed so well

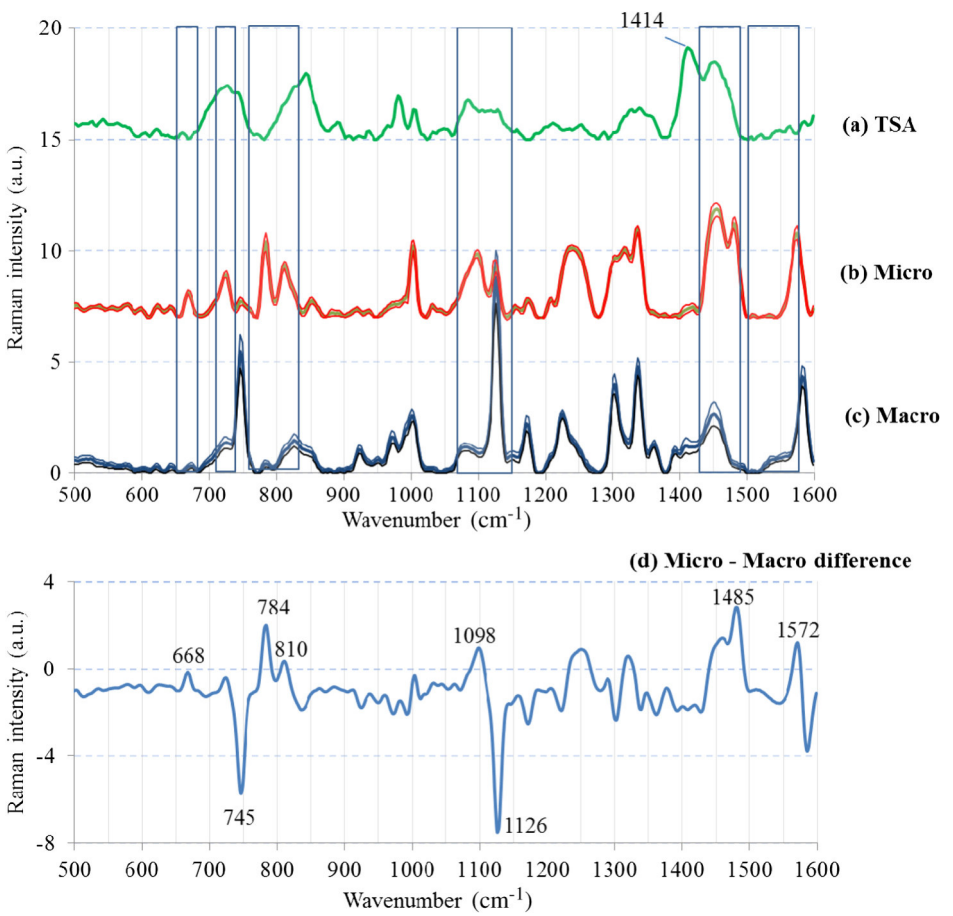

Fig. 7 From top to bottom: average normalized net Raman spectra (offset for clarity) of (a) TSA culture medium, (b) *E. coli* microcolonies (±1 SD in thin traces), (c) *E. coli* macrocolonies, and (d) difference between microcolonies and macrocolonies.

with the raw spectra. A part of the answer lies in the fact that, contrary to ED and KNN (closely related with PCA through the ED), those methods are not based on constant weights for all channels throughout the spectrum. They instead determine optimized weights (by taking account of the common or specific variance-covariance matrix for LDA and rQDA, or by searching for the separating hyperplane with maximum margin and minimum classification errors for SVM), which shows equivalent performance in terms of classification to explicitly subtracting the background.

Now we come to the question: is it really necessary to subtract the background? In fact, even if raw spectra give good classification results, it seems dangerous to keep background in the signal of interest. Indeed, background depends a lot on experimental conditions. In our study, sources of variability were minimized by for instance choosing a constant acquisition time for spectra for all 80 strains of our experimental collection and selecting a unique culture medium. This is a valuable approach here, as we wanted to evaluate the ultimate discriminatory power of Raman spectroscopy under nearly ideal conditions while minimizing possible confounding factors. For obvious reasons, it might be not very practical, as for instance a common time of acquisition is unlikely to be found as the number of species and strains will be significantly expanded. Variable acquisition time will very likely cause an increase in fluorescence background variability as photobleaching is concurrent to acquisition. Also, multiple media and low temperature storage will certainly be in frequent use in real life. For those reasons, removing the background seems mandatory because it is likely to render the Raman procedure more robust.

5 Conclusion

Our original intent was to evaluate spectral classification performance while minimizing data preprocessing to establish a benchmark for the performance of identification. We have shown that it was possible to discriminate, at the species level, 80 strains belonging to 10 different bacterial and yeast species, with a CIR ranging from 91.5% for microcolonies to 94.1% for macrocolonies, via direct measurements on the culture medium. Importantly, these numbers were obtained in a stringent cross-validation analysis. This opens the door to an innovative clinical diagnostic workflow, allowing the possible interrogation of cultures as early as 6 h from culture start with the possibility of resuming the culture after Raman spectroscopy in order to facilitate other forms of downstream microbiological analysis. Interference from the underlying medium is most likely absent and we are strongly suggesting that most of the difference observed between microcolonies and macrocolonies are of biological origin. Without any attempt to correct for the medium contribution, results are judged excellent as significantly above the 90% cut-off limit routinely accepted in IVD identification.

In real clinical settings, the nature of a sample is likely to be important, whether or not a patient was treated with antibiotics for instance. If bacteria are fastidious and long culture periods are required, the medium composition may change drastically. As TSA is not the most frequently used medium in a clinical laboratory, the study should be extended to other media but there is little risk that the performances will be affected (at least for nonchromogenic media) as shown by the diversity of media used in the published prior art. The power of discrimination of Raman might decrease with a larger number of species included in the database, but more efforts are needed to confirm or refute this proposition.

The simplicity of the preprocessing method used in this study as well as the absence of any sample preparation after culture, coupled with low biomass requirements, low invasiveness, and real-time measurement make Raman spectroscopy an outstanding technology candidate for rapid and automated IVD.

Acknowledgments

This work has been supported by the French National Agency (ANR) in the framework of its program "Recherche technologique Nano-INNOV/RT" (project DIAGRAM ANR-09-NIRT-002). The authors wish to thank their collaborators from Horiba Jobin-Yvon for discussions and providing access to their facilities for the initial measurements conducted at the beginning of the project, and in particular Philippe de Bettignies for fruitful discussions on Raman instrument design and performance.

References

1. K. L. Davis, J. M. Tedesco, and J. M. Shaver, "Advances in fiber optic Raman instrumentation," *Proc. SPIE* **3608**, 148–156 (1999).
2. M. Harz, P. Rosch, and J. Popp, "Powerful tool for the rapid identification of microbial cells at the single-cell level," *Cytometry Part A* **75A**(2), 104–113 (2009).
3. K. Maquelin et al., "Raman spectroscopic method for identification of clinically relevant microorganisms growing on solid culture medium," *Anal. Chem.* **72**(1), 12–19 (2000).
4. K. Maquelin et al., "Rapid identification of Candida species by confocal Raman microspectroscopy," *J. Clin. Microbiol.* **40**(2), 594–600 (2002).
5. K. Maquelin et al., "Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures," *J. Clin. Microbiol.* **41**(1), 324–329 (2003).
6. B. D. Beier and A. J. Berger, "Method for automated background subtraction from Raman spectra containing known contaminants," *Analyst* **134**(6), 1198–1202 (2009).
7. L. P. Choo-Smith et al., "Investigating microbial (micro)colony heterogeneity by vibrational spectroscopy," *Appl. Environ. Microbiol.* **67**(4), 1461–1469 (2001).
8. O. Samek et al., "Raman spectroscopy for rapid discrimination of Staphylococcus epidermidis clones related to medical device-associated infections," *Laser Phys. Lett.* **5**(6), 465–470 (2008).
9. J. F. M. Almarashi et al., "Raman spectroscopy of bacterial species and strains cultivated under reproducible conditions," *Spectrosc. Int. J.* **27**(5–6), 361–365 (2012).
10. N. E. Marotta and L. A. Bottomley, "Surface-enhanced Raman scattering of bacterial cell culture growth media," *Appl. Spectrosc.* **64**(6), 601–606 (2010).
11. C. Xie et al., "Identification of single bacterial cells in aqueous solution using confocal laser tweezers Raman spectroscopy," *Anal. Chem.* **77**(14), 4390–4397 (2005).
12. T. J. Moritz et al., "Effect of cefazolin treatment on the nonresonant Raman signatures of the metabolic state of individual Escherichia coli cells," *Anal. Chem.* **82**(7), 2703–2710 (2010).
13. A. A. Talukder et al., "Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid," *J. Bacteriol.* **181**(20), 6361–6370 (1999).
14. G. P. Singh et al., "The lag phase and G1 phase of a single yeast cell monitored by Raman microspectroscopy," *J. Raman Spectrosc.* **37**(8), 858–864 (2006).
15. M. Harz et al., "MicroRaman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions," *Analyst* **130**(11), 1543–1550 (2005).
16. W. E. Huang et al., "Raman microscopic analysis of single microbial cells," *Anal. Chem.* **76**(15), 4452–4458 (2004).
17. D. Hutsebaut et al., "Effect of culture conditions on the achievable taxonomic resolution of Raman spectroscopy disclosed by three Bacillus species," *Anal. Chem.* **76**(21), 6274–6281 (2004).

18. K. E. Eboigbodin and C. A. Biggs, "Characterization of the extracellular polymeric substances produced by Escherichia coli using infrared spectroscopic, proteomic, and aggregation studies," *Biomacromolecules* **9**(2), 686–695 (2008).
19. V. Ciobotă et al., "The influence of intracellular storage material on bacterial identification by means of Raman spectroscopy," *Anal. Bioanal. Chem.* **397**(7), 2929–2937 (2010).
20. W. N. Venables, D. M. Smith, and R Core Team, "R: a language and environment for statistical computing," http://www.r-project.org (2013).
21. D. Groom, "Cosmic rays and other nonsense in astronomical CCD imagers," *Exp. Astron.* **14**(1), 45–55 (2002).
22. A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.* **36**(8), 1627–1639 (1964).
23. C. G. Ryan et al., "SNIP, a statistics-sensitive background treatment for the quantitative of PIXE spectra in geoscience applications," *Nucl. Instrum. Methods* **B34**(3), 396–402 (1988).
24. T. Bocklitz et al., "How to pre-process Raman spectra for reliable and stable models?," *Anal. Chim. Acta* **704**(1–2), 47–56 (2011).
25. S. Stöckel et al., "Identification of Bacillus anthracis via Raman spectroscopy and chemometric approaches," *Anal. Chem.* **84**(22), 9873–9880 (2012).
26. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer, New York (2002).
27. C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," (2013), http://www.csie.ntu.edu.tw/~cjlin/libsvm.
28. M. L. Paret et al., "Biochemical characterization of Gram-positive and Gram-negative plant-associated bacteria with micro-Raman spectroscopy," *Appl. Spectrosc.* **64**(4), 433–441 (2010).
29. L. J. Goeller and M. R. Riley, "Discrimination of bacteria and bacteriophages by Raman spectroscopy and surface-enhanced Raman spectroscopy," *Appl. Spectrosc.* **61**(7), 679–685 (2007).

Biographies of the authors are not available.