A study on identification of bacteria in environmental samples using single-cell Raman spectroscopy: feasibility and reference libraries

Jean-Charles Baritaux, Anne-Catherine Simon, Emmanuelle Schultz, C. Emain, P. Laurent, Jean-Marc Dinten

A study on identification of bacteria in environmental samples using single-cell Raman spectroscopy: feasibility and reference libraries

Jean-Charles Baritaux[1], Anne-Catherine Simon[2], Emmanuelle Schultz[1], C. Emain[1], P. Laurent[1], Jean-Marc Dinten[1]

1. Université Grenoble-alpes, CEA, LETI, Minatec-Campus, F-38000, Grenoble, France
2. CEA, LIST, Gif-sur-Yvette, F-91191, France

Abstract

We report on our recent efforts towards identifying bacteria in environmental samples by means of Raman spectroscopy. We established a database of Raman spectra from bacteria submitted to various environmental conditions. This dataset was used to verify that Raman typing is possible from measurements performed in non-ideal conditions. Starting from the same dataset, we then varied the phenotype and matrix diversity content included in the reference library used to train the statistical model. The results show that it is possible to obtain models with an extended coverage of spectral variabilities, compared to environment-specific models trained on spectra from a restricted set of conditions. Broad coverage models are desirable for environmental samples since the exact conditions of the bacteria cannot be controlled.

Keywords : Raman spectroscopy, Single bacterial cell identification, Classification, Reference libraries, Outliers removal, Environmental samples

Introduction

Raman spectroscopy is a widespread chemical profiling technique that is gaining popularity in bacteria monitoring applications. In particular, it is very promising for biological threat detection activities (Stöckel et al. 2012a). Spectra from individual bacterial cells are collected in a few seconds and can be used for identification down to the strain level (Huang et al. 2004; Rosch et al. 2005; Willemse-Erix et al. 2009). Because the Raman fingerprint is phenotype dependent, it is challenging to assemble a collection of Raman spectra suitable for training the chemometric model used in identification. Ideally, all phenotypes considered in a given application are represented in the training set. Besides, one asset of Raman spectroscopy is the possibility to make measurements with minimal sample preparation or even directly in the environmental matrix of the bacteria. Correspondingly, the training set includes the matrix contribution. This approach has been successfully applied to cerebrospinal fluid analysis (Harz et al. 2009), urinary tract infection (Kloss et al. 2013), or pathogen detection in water (Tripathi et al. 2008; Van de Vossenberg et al. 2013; Kusić et al. 2014), for instance. Samples investigated in biological threat detection are heterogeneous, both in terms of phenotype content and environmental matrix. For such complex samples, purification protocols are usually required to isolate the cells before informative spectra can be collected (Meisel et al. 2011; Stöckel et al. 2012b; Stöckel et al. 2014). Even so, a matrix component is still present in the data and included in the model. The Raman analysis of such heterogeneous samples requires a compromise between database scope, keeping the creation of such a database tractable and minimizing sample preparation.

The purpose of this study is twofold. First, verify the feasibility and robustness of bacteria identification from Raman measurements performed in non-ideal conditions. To this end, a comprehensive database of single-cell Raman spectra was collected on a panel of five bacteria species (*Bacillus cereus*, *Bacillus subtilis*, *Staphylococcus epidermidis*, *Escherichia coli*, *Serratia marcescens*). In these experiments, the cells were submitted to various growth conditions and embedded in two environmental matrices relevant to field applications: a solution of atmospheric air dissolved in water and condensed water from a cooling-tower filtration system. A statistical model was trained on this dataset, and classification was performed at the species level. The dependence of Raman identification on cultivation conditions, namely medium, temperature, and age, has been the subject of earlier studies (Hutsebaut et al. 2004; Harz et al. 2005). This work is targeted at environmental samples and we include the matrix contribution as an additional variable. Our second goal is to investigate the content required in the training set of the chemometric model used for identification. We discuss the benefits of outlier detection on our dataset which was acquired over several months. We then examine identification robustness to deviations between the conditions (both in terms of phenotype and matrix) included in the statistical model and the conditions of the sample. Two limit cases are considered as follows: a model trained on a single phenotype and single matrix for each species, or conversely a comprehensive model including all available diversity for each species. The latter type is of particular interest in applications where the model has to approximate a phenotype possibly missing from the training set. This is the case of bio-threat detection or for identification of non-cultivable strains (Kumar et al. 2015).

Materials and methods

Strains and culture conditions

Three Gram-positive strains, *B. cereus* ATCC10702 (BC), *Bacillus subtilis* ATCC23857 (BS), and *S. epidermidis* ATCC14990 (SE), and three Gram-negative strains *E. coli* ATCC9637 (EC), *E. coli* ATCC11775 (EC), and *S. marcescens* ATCC27137 (SM), were considered in this study. Overnight liquid cultures (16 h incubation time) were prepared in a volume of 25 mL. In order to ensure that all species were in exponential phase at the time of measurement, the fastest growing species (*E. coli* and *S. marcescens*) were re-cultured for an additional 4 h prior to Raman analysis by transferring 100 uL of overnight culture in 10 mL of fresh medium. Bacterial growth was monitored by optical density.

Culture medium and temperature were varied for the purpose of this study. As a starting point, all microorganisms were cultured in the standard conditions prescribed by the supplier (see Table 1). Note that standard conditions differ from bacteria to bacteria. In addition to standard conditions, the combinations of media (LB, TSB) with temperatures (30 and 37 °C) were applied to the species BS, EC, and SM. Finally, three custom media of increasing nutrient content were prepared and used to grow BS and EC (while keeping their standard temperature). Medium 1, the least nutritive, was composed of 0.5 g/L NaCl (SIGMA S5886), 0.186 g/L KCl (PROLABO 26764.298), 4.8 g/L MgSO4 (SIGMA M2643), and 3.603 g/L alpha-d-glucose (ALDRICH 15,896-8). Medium 2 was the same composition with the addition of 20 g/L soy peptone (enzymatic digest FLUKA 87972). Medium 3 was the same as medium 2, plus 5 g/L yeast extract (FLUKA 70161).

Strain	Abbreviation used in text	Standard culture conditions
B. cereus ATCC10702	BC	TSB, 30 °C
B.sSubtilis ATCC23857	BS	TSB, 30 °C
E. coli ATCC9637	EC	LB, 37 °C
E. coli ATCC11775	EC	LB, 37 °C
S. epidermidis ATCC14990	SE	LB, 30 °C
S. marscesens ATCC27137	SM	LB, 30 °C

TSB trypticase soy broth, *LB* Luria-Bertani broth

Table 1. Bacterial strains and standard culture conditions

Sample preparation

After culture the cells were washed in sterile water (AGUETTANT, OTEC Sterile water) using 3500 rpm (822*g*) centrifugation for 2 min and re-suspended at a concentration of about 10^5–10^6 cells/μL. We considered three re-suspension solutions: water, AIR, and TARH. AIR and TARH are characterized real-world environmental matrices corresponding to atmospheric air dissolved in water, and condensed water from a cooling-tower filtration system, respectively. Cells grown in non-standard conditions were always suspended in water, while cells from standard conditions were suspended in AIR, TARH, or water. In this way, growth conditions effects were decoupled from environmental matrix effects. Spectra from cells cultured in standard conditions and re-suspended in water are referred to as standard, in contrast with spectra where culture, or matrix, was altered. One microliter of suspension was sampled and deposited on a Quartz slide (TedPella Inc. 19 × 19 × 0.5 mm). The smear was evaporated for 1 min at room temperature and immediately taken to our instrument for examination and Raman spectrum collection. Ten spectra were acquired in each smear.

A collection of spectra—or equivalently, a database—was assembled over the course of three measurement campaigns. The first campaign consisted in varying environmental matrices (namely AIR and TARH). This campaign involved all five species and was divided in two sessions separated by 3 months. The second campaign varied the culture medium (media 1, 2, and 3, see "Strains and culture conditions" section), and was realized on BS and EC over two 1-week sessions separated by 1 month. Experiments involving BS, EC, and SM in the media (LB, TSB) at temperatures (30 and 37 °C) were done in the third campaign. Meanwhile, a database of standard spectra from the five species was continuously enriched resulting in a collection acquired over the course of 10 months. An overview of the whole dataset is available in Supp. 1.

Confocal Raman microspectroscopy

The collection of spectra was acquired using a custom Raman instrument recently developed in our lab. A detailed description can be found in Strola et al. (2014). Briefly, the system allows fast detection and targeting of bacterial cells, as well as the measurement of single-cell Raman spectra using a confocal arrangement. The beam of a 532-nm, 50-mW laser (Spectra Physics Excelsior 532-50-CDRH) is attenuated and focused by a microscope objective (×100, 0.8 NA, Olympus LMPLFLN) in order to provide a spot size of 1 μm in diameter at the sample. Raman back-scattered light from an individual bacterium is collected by the same objective, filtered from Rayleigh light, and focused into the entrance fiber of a dispersive spectrometer (Hyperflux U1-532, Tornado Spectral systems, Toronto Canada). The spectrometer featured a −15 °C TE-cooled CCD, and spectral resolution of 7 cm^{-1} over the

band 500–3400 cm^{-1}. Shot-noise limited spectra were acquired using 10-s integration time. The system has recently been integrated in a transportable instrument called Bacram (30 kg, $70 \times 44 \times 71$ cm^3), currently in use to build a relevant bio-defense database.

Samples from real-world matrices contained non-specific particles and other impurities alongside the bacteria. We also observed the formation of a film covering the cells in the TARH matrix, in some cases. In this work, bacterial cells were discriminated from other particles based on morphology and reflectivity. It is known that a drying droplet exhibits convective micro-flow towards the contact line. This caused most non-specific particles to accumulate at the border of the smear (the so-called coffee-stain effect). The center regions of the smears therefore displayed a reduced number of particles. Unambiguous localization of bacteria was straightforward in the center because of the high cell concentration of our samples (10^5–10^6 cells/μL). The confocal arrangement provided spatial filtering which allowed maximizing the bacterial signal with respect to signal from other Raman-active substance that may be present in the sample.

Data analysis

Data analysis (spectra pre-processing, calculation of indicators and classification) was performed using the R software environment. Custom software was written in complement of the existing routines.

Pre-processing of spectra consisted in cosmic spikes removal, smoothing, restriction to a region-of-interest (ROI), and finally, normalization by the mean signal in the ROI. Smoothing was performed using Savitzky-Golay polynomial filters (degree 4, on 9 points). A 9-point filter corresponds to 18 cm^{-1}, while the typical full-width-at-half-maximum of the Raman peaks ranges from 20 to 60 cm^{-1}. This smoothing approach enables to increase the signal-to-noise ratio (SNR) with minimum peak distortion and loss of intensity. We chose a ROI composed of the two regions 650–1800 and 2600–3200 cm^{-1}.

Classification was performed using the support vector machine (SVM) implementation "svm" of the R package "e1071," interfacing the "LIBSVM" library. We used SVM with a linear kernel, and a "C" parameter value of 10 ("C" being the regularization parameter in the Lagrange formulation of SVM). Classification performance was assessed by cross-validation. Note that the same cross-validation procedure was consistently employed for all the results presented in this work. It consists of an external *leave-one-date-out* cross-validation, with training set balancing. In leave-one-date-out cross-validation the test set is composed of spectra from a single date. This date is omitted in the training set. This way it is ensured that training set and test set are independent. Besides, correlations due to nonspecific day-to-day variations are avoided. Balancing was implemented at the level of species, growth conditions, and environmental matrices, when applicable. Larger classes were randomly sub-sampled in order to ensure a same number of spectra in each class. In order to improve SNR, the test spectra were averaged by groups of five in each cross-validation round. Classification stability was evaluated by repeating every cross-validation round ten times with 90 % of the training set sub-sampled randomly. We report the average sensitivity (true-positive rate), standard deviation of sensitivity, and average specificity (true-negative rate). Classification was performed at the species level.

An outlier detection procedure was implemented to ensure the consistency of the dataset. Outliers are defined with respect to the spectra of a single strain in a single condition. This

means that a given spectrum will be compared to the spectra of the same strain in the same conditions. The procedure consists in two steps. The first step considers the group of spectra acquired on a single date. Spectra with a large Euclidean distance to the average of the group are tagged as outliers. In the second step, all spectra of a given condition are treated together, irrespective of the date. We compute the Mahalanobis distance (Mahalanobis 1936) from each spectrum to the distribution of spectra in the same conditions. Spectra with a large Mahalanobis distance are considered unlikely to belong to the distribution and are marked outliers. The first step enforces homogeneity among acquisitions performed the same day, while the second step ensures statistical consistency across all dates.

Results

Standard spectra

Average standard spectra of each species are displayed in Fig. 1. The most prominent spectral bands in the signature region are attributed to the ring breathing modes of phenylalanine at 1004 cm^{-1}, amid III at 1245 cm^{-1}, the deformation vibration of CH_2 in proteins at 1337 cm^{-1}, the deformation vibration of CH_2/CH_3 in lipids at 1451 cm^{-1}, and amid I at 1666 cm^{-1}. We also note the strong contribution of the CH stretching vibration in the band 2800–3150 cm^{-1}. A more detailed assignment can be found elsewhere (Movasaghi et al. 2007). The two large bands around 800 and 1100 cm^{-1} are background signal from the quartz substrate. It is apparent that Raman spectra of each of the species BC, BS, and SE are well distinct from the spectra of the four other species. By contrast, there is a close resemblance of the spectra measured on EC and SM. A SVM model was trained on the standard spectra for a classification at the species level. Cross-validation results are presented in the form of a confusion matrix in Table 2. Note that the number of spectra reported in the confusion matrix differs from species to species. These spectra correspond to test spectra, while the SVM model was fit on a balanced training set in each cross-validation round (see “Data analysis” section). The mean classification rate at the species level is 96.8 %, which is in accordance with performance generally reported for Raman spectroscopy on lab cultures (Harz et al. 2005). We observe that misclassified spectra originate from confusions between EC and SM, which reflects the high similarity between Raman spectra recorded on these two species.

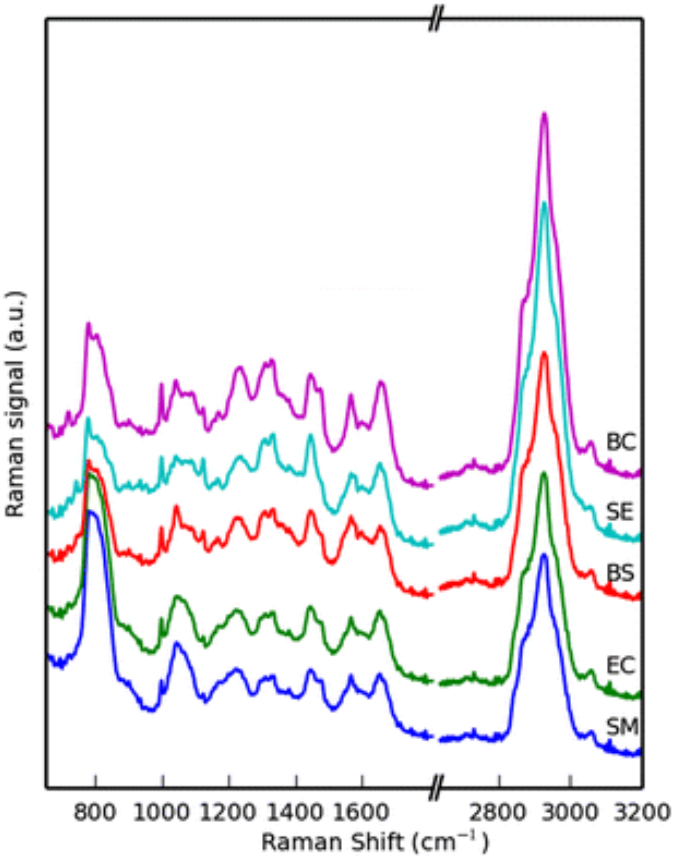

Fig. 1. Average Raman spectra in standard conditions. The spectra were smoothed and normalized to unit mean

Predicted label								
		BC	**BS**	**SE**	**EC**	**SM**	**Sens.**	**Spec.**
Test label	BC	265					100	100
	BS		380				100	100
	SE			151			100	100
	EC				446	29	93.8	97.6
	SM				25	235	90.3	98.5

Average sensitivity 96.8 %. Note classifier was trained on balance dataset
Sens. sensitivity (true-positive), *Spec*. specificity (true-negative)

Table 2. Confusion matrix for a SVM model applied to spectra in standard conditions

The bar diagram in Fig. 2 (see also Supp. 2) assesses the strategy of outlier rejection and spectral averaging outlined in “Data analysis” section. Outlier rejection applies to the training set, while averaging by groups of five applies to the spectra in test. Sensitivity (true-positive rate) and the corresponding standard deviation were evaluated by cross-validation, with and without applying the proposed strategy. Substantial gain in sensitivity and reduction of dispersion (standard deviation) are provided by averaging and rejecting outliers. Averaging spectra helps reducing instrumental variations, while outlier rejection reduces non-specific variations in the dataset. This is particularly beneficial in the case of EC and SM which display very close spectral signatures.

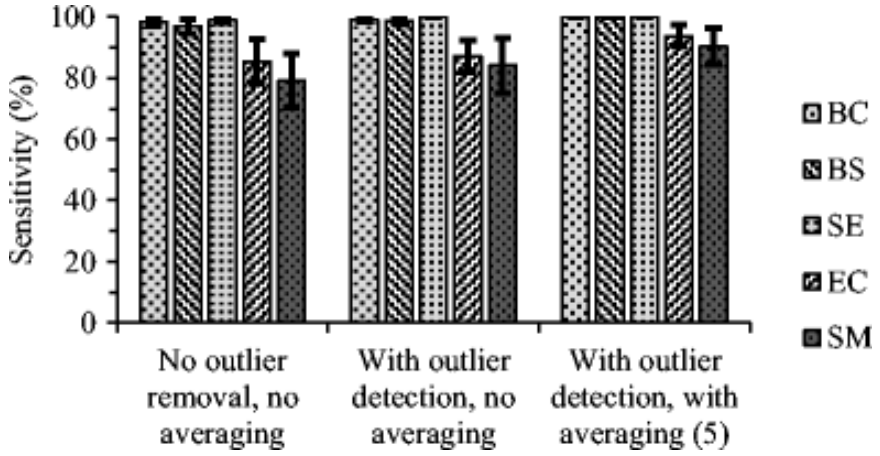

Fig. 2. Effect of outlier detection and spectra averaging. Results obtained on standard spectra

Figure 3 shows the sensitivity of the classifier for an increasing number of dates in the training set. Each new date contributes ten spectra to the training set. For each species, we keep increasing the number of dates whenever new dates are available and use the maximum number of dates otherwise. This way, the training set includes a new date for at least one species in between two points of the curves in Fig. 3. We chose to grow the training set by dates rather than by spectra because spectra were acquired in daily sessions, with day to day biological variations. Sensitivities are reaching a plateau as the number of dates increases. Sensitivities for the BC, BS, and SE converge to 100 %, while EC and SM converge to 93.8 and 90.3 %, respectively. Convergence is effective after 6 dates for the Gram positive, while EC and SM require about 20 dates. The larger number of dates required by EC and SM is not surprising since a finer model has to be established to discriminate these species. We emphasize that these results were obtained on a dataset recorded over the course of several months, which shows the stability of the Raman approach, provided that all relevant biological diversity is incorporated in the classification model.

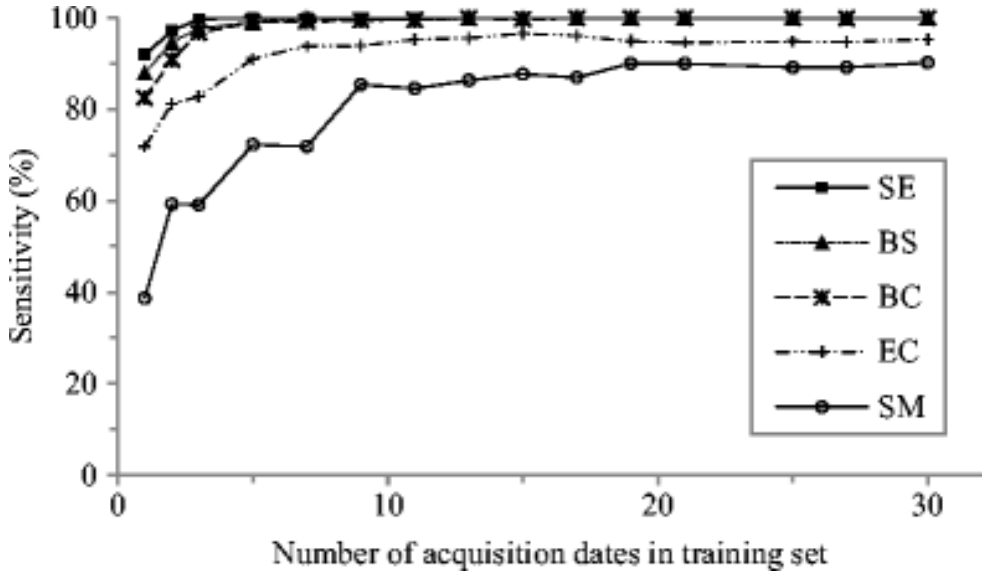

Fig. 3. Classification performance as a function of number of acquisition dates in the training set of the SVM model. Each acquisition date corresponds to ten spectra. In between two points the number of dates increases for at least one species

Varying culture conditions

In this section, we investigate the classification at the species level of bacteria submitted to various growth conditions using Raman spectroscopy and a SVM model. BS, EC, and SM were cultured in LB and TSB at two temperatures (30 and 37 °C). In addition, BS and EC were grown in three media of increasing nutrient content. Cells were washed and re-suspended in FreeWater prior to Raman measurements. We refer to "Materials and methods" section for details. This part of the study was limited to BS, EC, and SM since most classification errors were due to confusions between EC and SM. BS was kept as the unique element of the groups BC, BS, and SE. Although the test set was restricted to BS, EC, and SM, the training set still contained spectra from the five species. We examine the performance of several training sets for the SVM model. A specific training set (denoted SPEC) is composed exclusively of spectra from the same conditions as the test set. At the opposite, a standard training set (denoted STD) contains only standard spectra, independently of the test set. We additionally define an extended training set (EXT) by making the union of SPEC and STD, and an exhaustive training set (EXH) which includes all conditions at our disposal (including environmental matrices studied in the sequel).

The results for BS, EC, and SM in LB and TSB at temperatures 30 and 37 °C are presented in the bar diagram in Fig. 4 (see also Supp. 3). No data was available for BS in LB at 37 °C. Sensitivities and corresponding standard deviations are plotted. BS is always identified correctly, independently of growth conditions and training set. By contrast, the performance of SVM on EC and SM is strongly dependent on the training set. We note that performance depends on growth conditions to a lesser extent. Similarly to what was observed on standard spectra, misclassifications correspond to confusions between EC and SM. The most favorable condition for these two species is LB 37 °C (standard condition of EC), with 96 % sensitivity for EC and 93.5 % sensitivity for SM on the standard training set. For all other conditions, a standard training set leads to as classifier that is biased towards SM and unable to identify EC (30 % success in average, with 19.4 % standard deviation). This problem is mitigated by replacing the training set STD with SPEC. This results in 94.3 % average sensitivity on EC and SM. Training sets EXT and EXH are essentially equivalent to SPEC in this case. Their advantage is a slight reduction in performance gap between EC and SM.

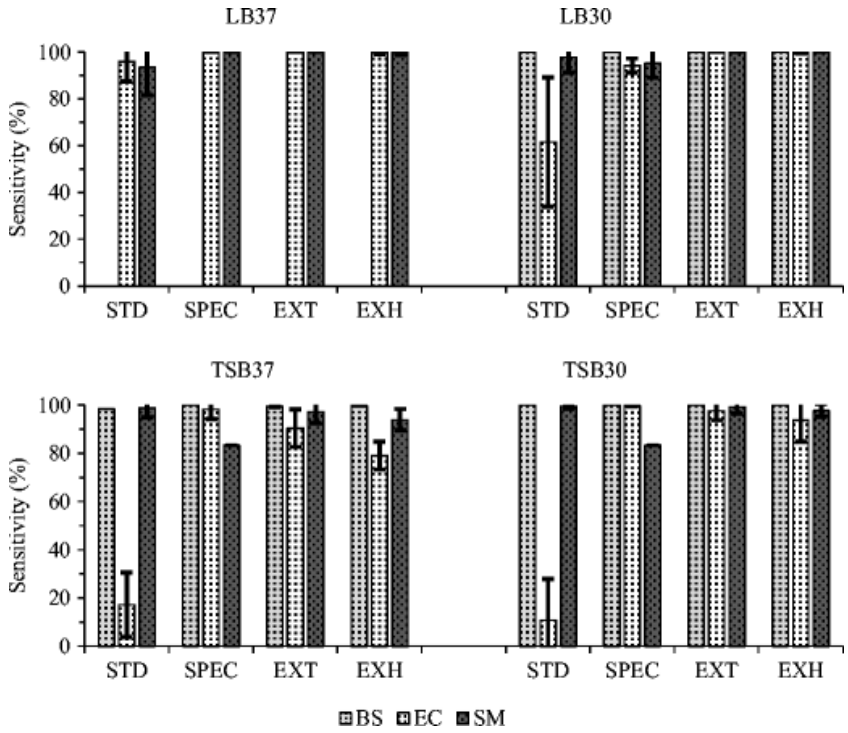

Fig. 4. Classification of BS, EC, and SM cultured in (LB, TSB) at temperatures (30 and 37 °C). Sensitivity and standard deviation are represented. Standard (*STD*), specific (*SPEC*), extended (*EXT*), and exhaustive (*EXH*) training sets are evaluated

Figure 5 shows the classification results for BS and EC grown in the three custom media (see also Supp. 4). The training set SPEC was not considered for this data. We note that medium 1 was not nutritive enough to permit bacterial growth. The spectra from both species in medium 1 are poorly identified when the model is trained on STD. For BS, this problem is addressed by switching the training set to EXT or EXH. The same strategy applied to EC improves the results, yet the sensitivity does not exceed 63 % (EXT training set). The success rate for BS in medium 2 and medium 3 is 100 % regardless of the training set (including STD). On the other hand, EC spectra from medium 2 and medium 3 require the training sets EXT or EXH in order to raise the sensitivity from about 50 % in STD to nearly 95 %.

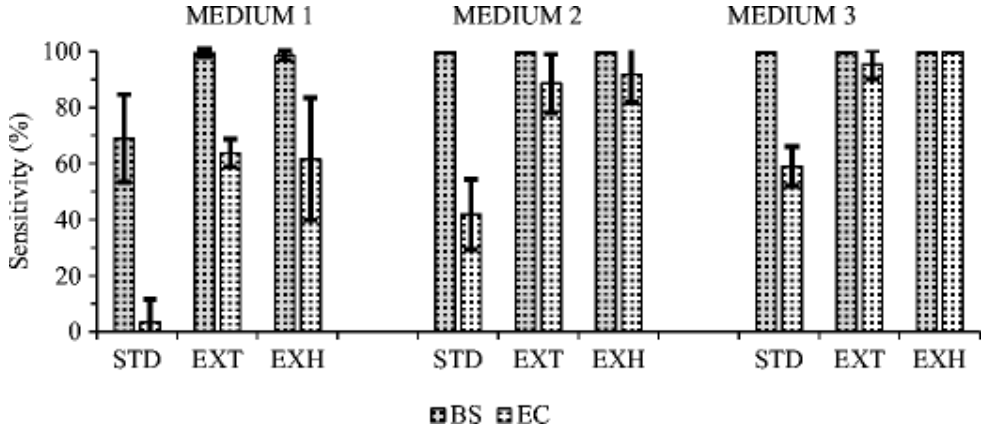

Fig. 5. Classification results on BS and EC cultured in three media of increasing nutrient content. Sensitivity and standard deviation are represented. Standard (*STD*), extended (*EXT*), and exhaustive (*EXH*) training sets are evaluated

In all these results, EXT or EXH training sets improve average sensitivity, and sometimes, robustness (by reducing the standard deviation).

Varying environmental matrix

This section discusses the classification of bacteria measured directly in environmental matrices. The five species BC, BS, SE, EC, and SM were embedded in AIR and TARH prior to Raman acquisitions (see "Sample sreparation" section). SVM was trained on the four datasets STD, SPEC, EXT, and EXH described in "Varying culture conditions." The sensitivities and corresponding standard deviations of the four models are displayed in Fig. 6 (data available in Supp. 5, Supp. 6, Supp. 7). We begin with noting that the four models perform equally well on the classification of species BC, BS, and SE, with 99.7 % average sensitivity. This is not the case for the discrimination of the two spectrally close species EC and SM. While the STD training set is insufficient to identify EC in AIR (4 % success), the SPEC, EXT, and EXH training sets lead to satisfactory results on the AIR matrix. The corresponding scores exceed 75 % for EC, and 92 % for SM in AIR. We note, however, that the TARH matrix remains problematic for SM, even when specific spectra are added to the training set. In this case, the model is strongly biased towards EC, since the majority of SM spectra end up identified as EC, while EC scores 98.1 %, in average.

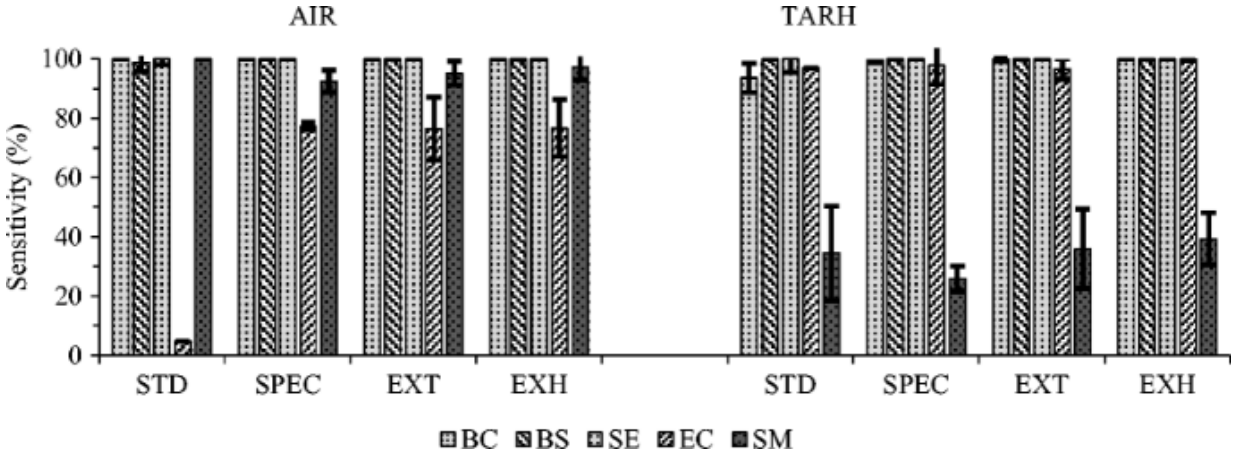

Fig. 6. Classification of BC, BS, SE, EC, and SM measured in AIR and TARH environmental matrices. Sensitivity and standard deviation are represented. Standard (*STD*), specific (*SPEC*), extended (*EXT*), and exhaustive (*EXH*) training sets are evaluated

Discussion

Datasets acquired in heterogeneous conditions, and over extended periods of time, are likely to contain non-specific variations. It is essential to ensure consistency of the data entering the statistical model, both globally and condition-wise. In our case, this was implemented by applying an outlier detection procedure to the data. With this approach, stable classification results were obtained on data assembled over a 10-month period. Outlier detection is not only useful to prepare training data but also to treat test data. This is extremely valuable for samples from non-ideal conditions. We noted that outlier detection was mostly beneficial to the spectrally closest species of our study. We also showed that averaging several test spectra helps reducing the dispersion of the results. Although not valid in the general case, this averaging is nonetheless relevant in many practical cases. In biological threat detection, for instance, a suspicious sample can be assumed to be highly concentrated in a single pathogen.

As expected, the requirements for establishing a reliable classifier are largely determined by the spectral distances of the bacteria under scrutiny. For very distinct phenotypes, intra-species variations induced by changing conditions remain lower than inter-species distances. Consequently, a standard training set suffices to define the classes of an SVM model. This was the case of BC, BS, and SE in our work. For all conditions considered in this study, these species were correctly classified using a model trained on standard spectra. This observation may prove very useful in applications involving a small number of spectrally well-distinct microorganisms because the construction of a specific database is avoided. Our panel of bacteria contained also the spectrally close species EC and SM. For these two, intra-species variations dominated inter-species variations, and specific spectra had to be included in the training set. We demonstrated high sensitivity on EC and SM submitted to various growth conditions or measured directly in the AIR matrix when the training set included diversity. The TARH matrix, however, remained problematic. One reason for this was the higher impurity content of TARH, compared to AIR. This caused not only spectral distortions but also a sampling bias since the cells which were well distinct from non-specific particles—often based on size and contrast—ended up measured more frequently. As a result, a tighter cluster of spectra was collected, making classification more challenging. The TARH case thus shows the limits of measurements made without sample preparation.

Specific databases may not be well adapted for analyzing environmental samples, for which the precise history is most likely unknown. One may consider instead a comprehensive dataset consolidating all species and conditions available. This training set was referred to as exhaustive in this work, and was found to outperform the specific training set in most cases. Still, it should be kept in mind that adding diversity to the training set possibly leads to a degradation of the results when inter-species distances are small compared to the additional variability. Although this situation was not observed in the present work, it was reported in the case of a study involving several *Bacillus* strains (Hutsebaut et al. 2004). For this reason, we also considered an extended training set. It is a midpoint between specific and exhaustive training sets, with the goal of balancing diversity and performance. In our evaluations, this extended training set performed comparably to the exhaustive training set. This strategy is well adapted to samples where the exact matrix or exact bacterial strain is possibly missing from the training set, and has to be approximated (Stöckel et al. 2014; Kumar et al. 2015).

Conclusion

The promise of culture-free identification of bacteria in environmental samples by Raman spectroscopy is largely conditioned upon our ability to build robust statistical models with sufficient coverage. We demonstrated the feasibility of Raman typing from measurements performed non-ideal conditions, and investigated the identification robustness to deviations between conditions included in the statistical model and conditions of the sample. We started with a condition-specific model trained on a dataset matching at best the sample conditions. This approach lead to very satisfying identification results on a panel of five species measured directly in two environmental matrices relevant for field applications: atmospheric air dissolved in water, and condensed water from a cooling tower. This result is encouraging for the deployment of a Raman instrument in the field. Yet, difficulties appeared for the bacteria with the closest spectral signatures. This suggests a tradeoff between sample preparation and direct in situ measurements. We then varied the diversity content of the training set, with the goal of building a model with broad coverage. Two limit cases were considered: the least amount of diversity and all the diversity available. Interestingly, the model with least amount of diversity (trained solely on standard spectra) was sufficient to classify the most spectrally

distinct bacteria, independently of their conditions. The exhaustive model, on the other hand, resulted in very satisfying performance on the entire dataset. This last approach is promising for environmental samples since the investigated phenotype, or environmental matrix may be missing from the training set and need to be approximated.

The results of this study will help refining the content of a bio-pathogens reference library currently under construction. By carefully selecting a library that accounts for the spectral distances between the pathogens of interest, as well as spectral variabilities resulting from non-ideal measurements, we anticipate that our Raman instrument will efficiently integrate the chain of analytical tools deployed in the field in response to biological threat.

References

Harz M, Rosch P, Peschke K-D et al (2005) Micro-Raman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions. Analyst 130:1543–1550. doi:10.1039/b507715j

Harz M, Kiehntopf M, Stockel S et al (2009) Direct analysis of clinical relevant single bacterial cells from cerebrospinal fluid during bacterial meningitis by means of micro-Raman spectroscopy. J Biophotonics. doi:10.1002/jbio.200810068

Huang WE, Griffiths RI, Thompson IP et al (2004) Raman microscopic analysis of single microbial cells. Anal Chem 76:4452–4458. doi:10.1021/ac049753k

Hutsebaut D, Maquelin K, De Vos P et al (2004) Effect of culture conditions on the achievable taxonomic resolution of Raman spectroscopy disclosed by three Bacillus species. Anal Chem 76:6274–6281. doi:10.1021/ac049228l

Kloss S, Kampe B, Sachse S et al (2013) Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study. Anal Chem 85:9610–9616. doi:10.1021/ac401806f

Kumar V, Kampe B, Rösch P, Popp J (2015) Classification and identification of pigmented cocci bacteria relevant to the soil environment via Raman spectroscopy. Environ Sci Pollut Res. doi:10.1007/s11356-015-4593-5

Kusić D, Kampe B, Rösch P, Popp J (2014) Identification of water pathogens by Raman microspectroscopy. Water Res 48:179–189. doi:10.1016/j.watres.2013.09.030

Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci Calcutta 2:49–55

Meisel S, Stockel S, Elschner M et al (2011) Assessment of two isolation techniques for bacteria in milk towards their compatibility with Raman spectroscopy. Analyst 136:4997–5005. doi:10.1039/C1AN15761B

Movasaghi Z, Rehman S, Rehman IU (2007) Raman spectroscopy of biological tissues. Appl Spectrosc Rev 42:493–541. doi:10.1080/05704920701551530

Rosch P, Harz M, Schmitt M et al (2005) Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. Appl Environ Microbiol 71:1626–1637. doi:10.1128/AEM.71.3.1626-1637.2005

Stöckel S, Meisel S, Elschner M et al (2012a) Identification of bacillus anthracis via Raman spectroscopy and chemometric approaches. Anal Chem 84:9873–9880. doi:10.1021/ac302250t

Stöckel S, Meisel S, Elschner M et al (2012b) Raman spectroscopic detection of anthrax endospores in powder samples. Angew Chem Int Ed 51:5339–5342. doi:10.1002/anie.201201266

Stöckel S, Meisel S, Elschner M et al (2014) Raman spectroscopic detection and identification of Burkholderia mallei and Burkholderia pseudomallei in feedstuff. Anal Bioanal Chem. doi:10.1007/s00216-014-7906-5

Strola SA, Baritaux J-C, Schultz E et al (2014) Single bacteria identification by Raman spectroscopy. J Biomed Opt. doi:10.1117/1.JBO.19.11.111610

Tripathi A, Jabbour RE, Treado PJ et al (2008) Waterborne pathogen detection using Raman spectroscopy. Appl Spectrosc 62:1–9. doi:10.1366/000370208783412546

Van de Vossenberg J, Tervahauta H, Maquelin K et al (2013) Identification of bacteria in drinking water with Raman spectroscopy. Anal Methods 5:2679–2687. doi:10.1039/C3AY40289D

Willemse-Erix DFM, Scholtes-Timmerman MJ, Jachtenberg J-W et al (2009) Optical fingerprinting in bacterial epidemiology: Raman spectroscopy as a real-time typing method. J Clin Microbiol 47:652–659. doi:10.1128/JCM.01900-08

Acknowledgments

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