Elastic light scattering for clinical pathogens identification: application to early screening of Staphylococcus aureus on specific medium

Emmanuelle Schultz, Valentin Genuer, Pierre R. Marcoux, Olivier Gal, Chakib Belafdil, Damien Decq, Max Maurin, Sophie Morales

Elastic Light Scattering for clinical pathogens identification: application to early screening of *Staphylococcus aureus* on specific medium.

E. Schultz[a], V. Genuer[a], P. Marcoux[a], O. Gal[b], C. Belafdil[b], D. Decq[a], Max Maurin[c], S. Morales[a]
[a]Université Grenoble-alpes, CEA, LETI, Minatec-Campus, 17 avenue des Martyrs, 38054 Grenoble cedex 9, France; [b]CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, 91191 Gif-sur-Yvette, France; [c]CHU Grenoble, Bacteriology Lab, CS10217, 38043 Grenoble cedex 9, France.

ABSTRACT

Elastic Light Scattering (ELS) is an innovative technique to identify bacterial pathogens directly on culture plates. Compelling results have already been reported for agri-food applications. Here, we have developed ELS for clinical diagnosis, starting with *Staphylococcus aureus* early screening. Our goal is to bring a result (positive/negative) after only 6 h of growth to fight surgical-site infections. The method starts with the acquisition of the scattering pattern arising from the interaction between a laser beam and a single bacterial colony growing on a culture medium. Then, the resulting image, considered as the bacterial species signature, is analyzed using statistical learning techniques. We present a custom optical setup able to target bacterial colonies with various sizes (30-500 microns). This system was used to collect a reference dataset of 38 strains of *S. aureus* and other *Staphyloccocus* species (5459 images) on ChromID-SAID/MRSA bi-plates. A validation set from 20 patients has then been acquired and clinically-validated according to chromogenic enzymatic tests. The best correct-identification rate between *S. aureus* and *S.* non-*aureus* (94.7%) has been obtained using a support vector machine classifier trained on a combination of Fourier-Bessel moments and Local-Binary-Patterns extracted features. This statistical model applied to the validation set provided a sensitivity and a specificity of 90.0% and 56.9%, or alternatively, a positive predictive value of 47% and a negative predictive value of 93%. From a clinical point of view, the results head in the right direction and pave the way toward the WHO's requirements for rapid, low-cost, and automated diagnosis tools.

Keywords: Elastic Light Scattering, bacteria identification, pattern recognition, machine learning, culture-based method, *S. aureus* screening.

1. INTRODUCTION

Staphylococcus aureus is a bacterium commonly found in healthy people's skin microbiota. About 30% of the population carry it on the nose or on the open skin areas, a rate that can rise in some communities, such as retirement homes or nurseries. Most of *S. aureus* carriers do not develop any symptoms in normal conditions and settings, but they have two to nine times more risks to get an infection subsequently to a surgical operation. This infection can be minor when limited to the skin (boils, pimples, or inflamed healing), but becomes serious when developing in blood or lungs (bacteremia, pneumonia), or when localized in implanted devices. This applies particularly for cardiothoracic, orthopedic, or ear surgeries. Methicillin or mupirocin are two classes of antibiotic currently used to treat *Staphylococcus* infections. However, some subpopulations of *S. aureus* have developed resistance against these agents that, hence, can no longer treat the infection. The severity of the health problem has increased with the dissemination of these emerging resistant strains between patients and care community, between different hospitals, cities and even countries. Worldwide, *S. aureus* is at the origin of most of the surgical site infections (SSIs) and contributes to an increased morbidity, mortality and healthcare costs. Since SSIs represent the third cause of healthcare associated infections (HAIs) in developed countries – it is the first cause in low-resource countries –, *S. aureus* infections are now also classified as HAIs.

Prevention and surveillance of S. aureus infections is therefore a major public health issue. Humphreys et al.[1] review the evidence on the value of screening for nasal carriage of S. aureus and subsequent decolonization of positive patients pre-operatively. The method has been shown particularly effective to reduce S. aureus SSI rates and mortality, especially in cardiothoracic and orthopedic departments. The associated costs have also decreased, despite the additional costs of screening and decolonization. On the contrary, universal decolonization without prior screening has negative impacts,

among which the emergence of resistance mechanisms, especially against muporicin, thus widening the resistance spectrum of *S. aureus* resistant strains.

The current challenge is now to implement a large-scale screening of *S. aureus* and subsequent decolonization. Its success relies on the availability of appropriate diagnosis tools able to communicate rapidly a positive or negative result to patients, caregivers and medical departments. Presently, molecular-, also called PCR-, based approaches have demonstrated their capability to deliver a very rapid (1 h) and highly specific result directly on clinical specimens[2]. GeneOhm-MRSA (Beckton-Disckinson) and GeneXpert MRSA (Cepheid) are such examples of genomic tests. However, their cost (40$/test) is not affordable for most hospitals, even in rich countries, especially in a screening context where a large number of specimen have to be analyzed everyday. The alternative method requires a 24-h culture step on a particular broth medium, both specific and chromogenic. Such a medium enables selective isolation of *Staphylococcus* species and coloration of colonies through an enzymatic activity peculiar to *S. aureus*[3], [4]. Analysis is thus performed by visual inspection: *e.g.* in ChromID S. aureus, thanks to α-glucosidase substrates, *S. aureus* colonies turn green while most of all the other *Staphylococcus* species turn white. To rule out green non-*aureus Staphylococcus* colonies, a supplementary test shall be performed, through another identification method. In our case, confirmation tests were done on chromAgar S. aureus, a chromogenic medium targeting another enzymatic activity than the glucosidase on which ChromID relies. A negative result can then be given 24 h after the nasal swab, but a positive result requires 48 h. Culture-based methods, although affordable, are thus time-consuming and difficult to automate.

In this study we decided to apply a novel identification method based on Elastic Light Scattering (ELS) directly on the chromogenic media, or any kind of medium. The goal is to give a positive/negative result much earlier than conventional biochemical approaches, that is, after 6 h of growth only, and with no additional test. The technique has also considerable potentials in terms of cost and automation.

Briefly, ELS reveals the complex internal structure and assembly features of a bacterial colony by exposing it to a coherent light beam. This information is collected in a few hundreds of milliseconds in a single snapshot, also called scatterogram. It constitutes a signature of the optical phenotype of the colony, which is representative of the bacterial genus or species that compose it. To provide bacterial identification, the method relies on pattern recognition algorithms trained on reference datasets, rather than on recovering algorithms as conventionally used in X-ray diffraction. The reason is that there is multiple-scale scattering processes inside the colony that are not yet well described by the models[5].

Elastic Light Scattering was first introduced in 2012 at Purdue University for the identification of common foodborne pathogenic bacteria, such as *Listeria*, *Salmonella*, *Vibrio*, *Staphylococcus* and *Escherichia coli*[6]. Genus- or species-level distinction between these bacteria was possible with an accuracy of 90-99%. These compelling performances have been obtained after 24 h of incubation, when colonies were formed of about 10^6 of cells and measured precisely (1.0±0.2) mm in diameter.

Starting from the state of the art, there were a number of challenges to overcome. First, developing the method to consider young colonies having three orders of magnitude less biomass than 24-h-old colonies. This has two major impacts on the method: first, the optical setup must allow the detection of non-visible colonies prior to the generation of comprehensive scattering patterns, and second, data analysis must be optimized to handle less complex, almost ring-shaped scatterograms. The second challenge was to collect in a clinical environment an extensive high-quality reference database of scatterograms to train and evaluate machine-learning algorithms. Optimization of feature extraction and classification were of particular importance to steeply rise the performances of the model, thus making it able to discriminate between species that all belong to the same genus *Staphylococcus*. In doing that, we would take the technique a step further than the state of the art, where studies have mainly considered bacteria with less biological proximity[7].

The study is completed by collecting a set of clinical data obtained after 6 h of cultivation from nasal swabs and double validated using 24-h cultivation chromogenic tests. We examined the performances of the model to classify positive patients vs. negative patients with phenotypes not included in the training set. This is of particular interest for the final application, where a variety of phenotypes may express, depending on the patient condition, and possibly not included in the training set.

2. MATERIAL AND METHODS

2.1 *Staphylococcus* strains and culture conditions for reference library acquisition

Sixteen *S. aureus* strains and 22 *S.* non-*aureus* strains were considered in the training set, giving a total of 38 strains (see Table 1). They were obtained from Microbiologics® (Kwik-Stik™ containing lyophilized ATCC® strain). Starting from 24-h cultures at 37°C on TSA (Trypcase Soy Agar), 0.5-McF suspensions (DensiCheck, bioMérieux) were prepared in suspension medium (bioMérieux). These suspensions were then diluted at 1/1000 in suspension medium and finally around 40 µL were plated on ChromID SAID/MRSA bi-plates (BioMérieux) to obtain approximately 100-1000 microcolonies after 6 hours of incubation at 37°C. Between 10 and 20 colonies were analyzed by ELS. These were chosen adequately isolated from each other to avoid acquisition of superimposed patterns. The analysis time per dish was always limited to 15 minutes otherwise the bacterial growth may interfere with the identification.

A total of 5459 scattering images have been recorded over the course of several months. We paid attention to acquire images from various culture preparations, namely from several Petri dishes and several weeks. The goal was to include in the training database variations due to biological settings, such as quality of the nutrient broth or of culture conditions, while avoiding the inclusion of confounding factors (*e.g.* between species and culture conditions or date).

Table 1. List of the bacterial species and strains considered in the training set. ATCC references, corresponding codes and numbers of strains, dishes and images are included.

Species	Code	Strains	Number of strains	Number of dishes	Number of images
S. aureus	"Sta-aur"	NCTC 6571, ATCC BAA-1026, ATCC BAA-2312, NCTC12493, ATCC 11632, ATCC 12600, ATCC 14775, ATCC 25923, ATCC 29213, ATCC 29737, ATCC 9144, ATCC BAA-44, ATCC 43300, ATCC 700698, NCTC 12973, ATCC BAA-976	16	182	2574
S. capitis	"Sta-cap"	ATCC 146, ATCC 35661	2	15	180
S. epidermidis	"Sta-epi"	ATCC 14990, ATCC 12228, ATCC 49461, ATCC 49134, ATCC 51625, ATCC 35984	6	59	597
S. gallinarium	"Sta-gal"	ATCC 700401	1	11	86
S. haemolyticus	"Sta-hae"	ATCC 29970	1	10	165
S. pseudintermedius	"Sta-pse"	ATCC 49444	1	7	48
S. lentus	"Sta-len"	ATCC 700403	1	9	187
S. simulans	"Sta-sim"	ATCC 27851	1	10	183
S. sciuri	"Sta-sci"	ATCC 29061	1	9	158
S. saprophyticus	"Sta-sap"	ATCC 15305, ATCC 35552, ATCC 43867, ATCC 49453, ATCC 49907, ATCC BAA-750	6	62	927
S. warneri	"Sta-war"	ATCC 49454	1	11	204
S. xylosus	"Sta-xyl"	ATCC 29971	1	11	150
		Total other than *S. aureus* ("Sta-other")	22	214	2885
		Total	**38**	**396**	**5459**

2.2 Nasal screening and double-validation protocol for clinical dataset

Patients from hospital were screened for *S. aureus* by ELS. During the 3-month clinical study period, from mid-December 2016 to mid-February 2017, a total of 20 clinical specimens were collected in the Grenoble Hospital (CHU La Tronche, France). All of them were nasal swabs (ESwab, COPAN Diagnostics) and streaked on chromogenic and specific ChromID S. aureus plates (SAID, bioMérieux). All plates were examined after 6 h of incubation at 36°C. The protocol is illustrated in Figure 1. Then, on a given plate, several microcolonies were analyzed in the MICRODIFF system. Between one and ten scattering patterns were recorded, one per colony, depending on the density of colonies found onto the surface. Location

of each investigated microcolony was also carefully recorded. The dish was then put back into the incubator for 18 h more, so as to yield the chromogenic test: a white colour meant that a negative label (*i.e. S.* non-*aureus*) could be associated to the scatterogram, whereas a green colour indicated a positive label (*i.e. S. aureus*). In case this first chromogenic reading was positive, the green microcolony was streaked onto a second chromogenic test (CHROMagar Staph. aureus), targeting another enzymatic activity, in order to get a confirmation about the *S. aureus* status.

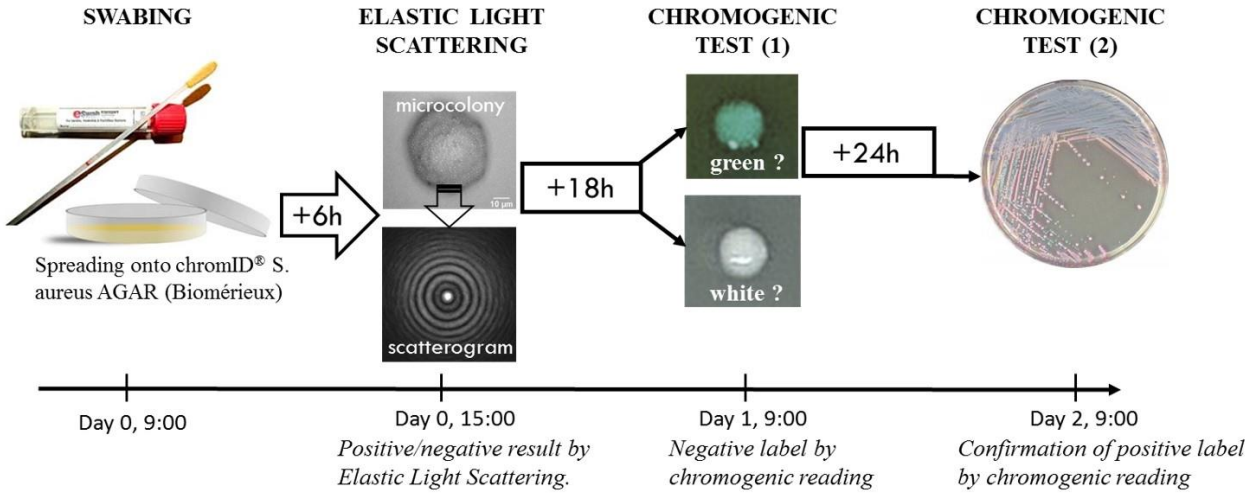

Figure 1. Illustration of the clinical protocol.

A total of 224 images were collected. The true labels are distributed into 60 “positive”, 130 “negative”, 10 “intermediate”, and 24 “unknown”. In some cases, it happened that the colony on which the scatterogram was acquired was surrounded by numerous other colonies. When these exhibited both the green and white colours in the double-validation test after 24 h of incubation, corresponding to positive and negative results, we could not label the image and decided to qualify it as “intermediate”. “Unknown” labels correspond to scatterograms that could not be associated to a colony, for instance when no colony could be found after 24-h incubation at the position were the scatterogram had been acquired. Aggregates in the nutrient medium most likely have been detected as microcolonies in such cases since they also produce scattering patterns. The improvement of the detection setup would prevent this kind of mistakes.

2.3 Elastic Light Scattering

The collection of scattering images was acquired using a custom setup that has been recently integrated in a transportable instrument called MICRODIFF. As described in Figure 2, the system allows detection of micrometric-scale colonies using a darkfield microscope arrangement, as well as the forward collection of the scattering pattern from a single colony. A 532-nm laser beam (Spectra Physics Excelsior 532-20-CDRH) with a laser power decreased down to 50 µW is spatially expanded using a double-lens system (L1-L2), then made divergent using a ×10 microscope objective.

The sample, a microbial isolation Petri dish plate, is placed upside-down in the output beam with the cover closed. It is moved so that a single bacterial colony is centered in the laser beam and a scattering pattern is produced. The latter is forwardly recorded using a CCD sensor (PixelFly QE, 1392×1024 pixels, 6.45-µm pixel size) placed 25 mm below the nutrient medium surface. It is worth noting that forward acquisition geometry is here possible since we use transparent broth media.

Two benefits are drawn from the divergent beam by comparison with a collimated beam. First, the scattering pattern is recorded with a magnifying factor on the sensor (×60). Second, the laser probe diameter can vary depending on the position of the sample along the optical axis, thus enabling to target various sizes of colonies. In this study, colonies are between 15 and 50 microns in diameter, depending on the strains (*S. aureus* strains tend to grow faster than *S.* non-*aureus* strains, leading to larger colonies). We have determined that a single beam diameter of 100 microns can address this variety of colony sizes while maintaining uniform acquisition conditions, especially in terms of wavefront shape. The position *Z* of

the sample where this beam size is available is properly set by calibration. The acquisition time is slightly adjusted by the operator to warrant non-saturated pixels. Looking at the whole collection, acquisition time is (216±87) ms. Estimated exposure levels are about 0.6 W/cm², namely low enough to keep colonies growing.

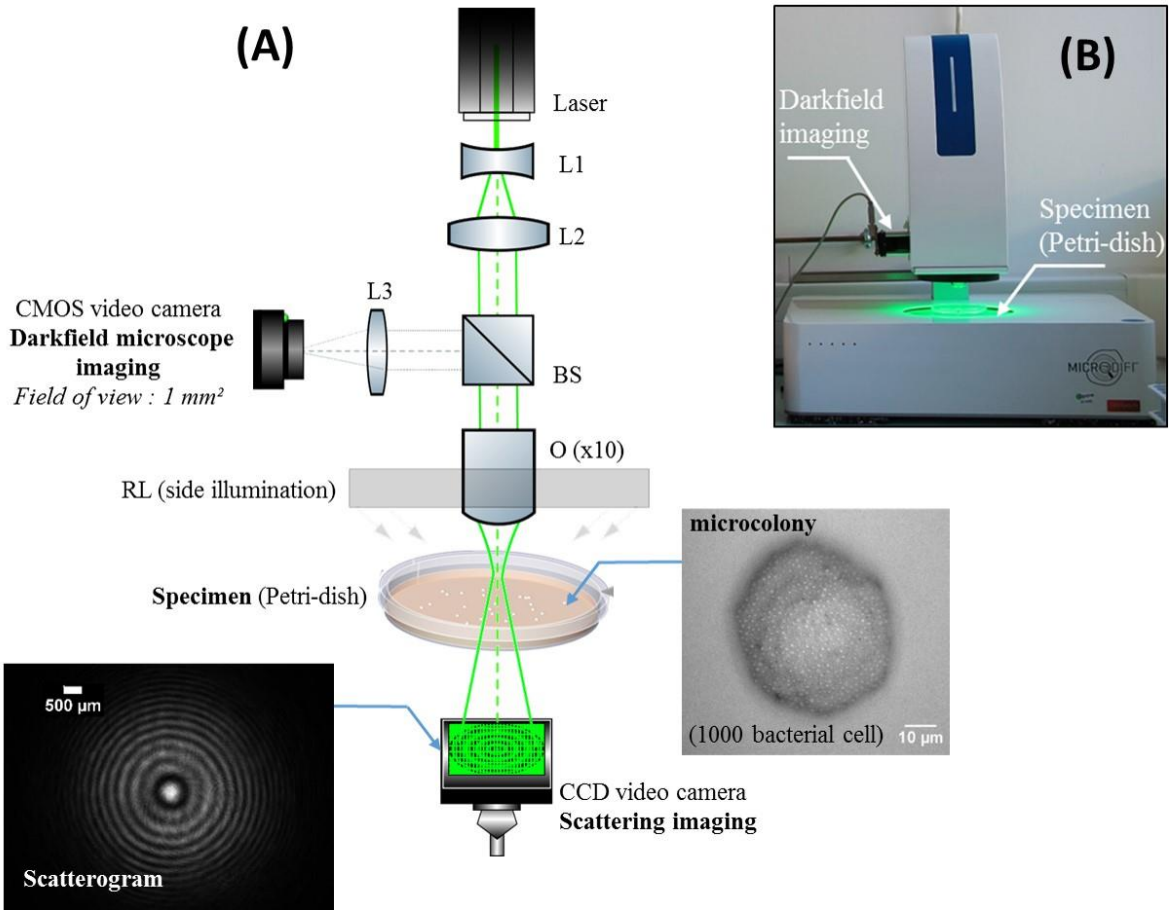

Figure 2. Elastic Light Scattering: (A) optical schematic and (B) photograph of the integrated piece of equipment called MICRODIFF, 25 kg, 40×50×60 cm³.

2.4 Data analysis

In order to test the ability of the MICRODIFF instrument to identify *Staphylococcus* species, we applied machine-learning techniques on the reference database of 5459 scatterograms from 12 *Staphylococcus* species recorded on Petri dishes after 6 h of culture. As the classification cannot be performed directly on the raw images, mainly because of their very high dimensionality (ca. 1.4×10^6 pixels per image), the first step consists in calculating a vector of descriptors for each image that catches most of pertinent information for the identification. Two types of descriptors have been tested in this study.

The first type is based on Fourier-Bessel moments[8], [9]: an orthonormal Fourier-Bessel basis is defined on a unitary disk with 1000-pixel diameter and, for each image, we calculate the complex coefficients a_{nm} of the development for radial ranks n = 1, 2,… 100 and azimuthal ranks m = 0, 1,… 10:

$$a_{nm} = \frac{1}{c_{nm}} \int_0^{2\pi} \int_0^1 F(r,\theta) J_m(j_{mn} r) e^{-jm\theta} r dr d\theta$$

where c_{nm} is a normalization constant, j_{mn} is the n-th zero of the Bessel function J_m of the first kind and rank m, and $F(r,\theta)$ is the map of the image grey levels inside the unitary disk. This development is very well fitted to efficiently

describe a 2D map with circular and concentric patterns, provided that the center of the Fourier-Bessel basis is placed at the center of the patterns. The 1100 complex coefficients a_{nm} are transformed into 1100 rotation-invariant moments $\tilde{a}_{nm}$ by considering their modulus. The computation time is about 10 s per image on a laptop PC, once the Fourier-Bessel basis is loaded in the memory. Note that the development was limited at $n = 100$ and $m = 10$ mainly for memory considerations and not for computational or performance limitations: both the development precision and the classification performance still increase with rank.

The other type of descriptor tested in this study is a pattern descriptor: it only characterizes the shape of the local intensity variations encountered in the image regardless of orientation or position. We used the "local binary patterns" operator (LBP) defined by Ojala[10] and known to be simple yet very efficient for pattern-recognition applications, and its implementation in Matlab by N. Skarbnik[11]. The idea consists, for each pixel in the image, in comparing its grey-level value with those of the N neighbor pixels (or the interpolated values) symmetrically distributed around it on a circle of radius R, and coding the sequence of results in an N-bit binary number (+ is coded by 1 and – by 0). The list of the 2^N possible sequences is reduced to P sequences by grouping the sequences differing by just a rotation, which makes the result rotation-invariant. From this new image, where each pixel has been replaced by the corresponding sequence in [1, P], a histogram of length P is calculated. The P values of this histogram constitutes the LBP descriptor vector used for the classification. For our study, we considered $N = 8$ neighbors, which corresponds to $P = 36$ rotation-invariant sequences, and 6 radius values (R = 1, 2, 4, 8, 16, 32 in pixels). The optimal radius for classification was found between 8 and 16 but classification performance is still better when the concatenated vector of the 216 values is taken as the descriptor. The computation time for the 6 radius values is about 30 s per image.

The second step of the classification consists in choosing a classifier and training it on the descriptors of the reference database images. For our study, we chose the "support-vector machine" (SVM) because of its simplicity of use and well-known ability to work with high-dimension data, based on theoretical guarantees[12] and producing good results in practice. We used the libsvm implementation and its Matlab interface[13]. The classification performance was estimated through a 5-fold cross-validation procedure: the image database was split into 5 balanced groups and each group was successively used as test data for the classifier trained on the rest of the data. Classification results can be summarized in the form of a confusion matrix giving the classification rates in each species for images from each species, averaged on all test groups. The correct-identification rates (CIRs) in each species are the figures on the diagonal. The overall CIR is the mean of the CIRs for each species.

The classifier is built following the "one-versus-one" procedure: for each of the 5 folds, an SVM classifier is trained for each pair of species. Then each test image is classified according to a vote between the results of the $K(K-1)/2 = 66$ classifiers (for $K = 12$ species). As the species are structured in strains, with between 1 and 16 stains per species, the species identification can also be obtained by training the classifiers at the level of the 38 strains and simply deducing the species from the found strain. This latter scheme proved to give better results on our *Staphylococcus* database than directly identifying the species. As our final goal is to build an *S. aureus* screening test, we also implemented a classification with only two classes: *S. aureus* vs. *S.* non-*aureus*. Once again, the classification was tested both by directly learning the two classes and by learning the strains and deducing the class from the found strain.

The study is completed by testing the learned classifiers on the clinical dataset of 224 scatterograms: each classifier is trained on the whole reference database and directly applied to each of the clinical scatterograms. The results are considered in the form of a test *S. aureus* vs. *S.* non-*aureus*. Three classification schemes were tested: by directly learning the two classes, or by learning the 12 species and deducing the class from the found species, or by learning the 38 strains and deducing the class from the found strain. As we deal here with a dichotomous test, the CIRs are called the sensitivity and specificity of the test, respectively for the positive (*S. aureus*) and negative (*S.* non-*aureus*) classes. If the prevalence of *S. aureus* is known among the targeted population, the classification performance can be alternatively assessed through the positive and negative predictive values, *PPV* and *NPV*, which can indeed be written in the form

$$PPV = \frac{Se \times p}{Se \times p + (1 - Sp) \times (1 - p)}$$

$$NPV = \frac{Sp \times (1 - p)}{(1 - Se) \times p + Sp \times (1 - p)}$$

where p is the estimated prevalence of the positive condition and Se and Sp are respectively the sensitivity and the specificity of the test.

3. RESULTS AND DISCUSSION

3.1 Reference dataset

Scatterograms

Examples of scatterograms for four strains, two of them being *S. aureus*, are displayed in Figure 3. For each of the strains, we show 5 images recorded during independent experiments. In this way, we both illustrate the high inter-species (inter-class) similarity and high intra-class variability, both of them representing a considerable challenge for the classification task.

The scatterograms are ring-shaped, and compatible with a light-propagation model based on numerical integration of Rayleigh-Sommerfeld equations in the case of a monolayer assembly of bacterial cells[5]. Scrutinizing the images, we first notice variations in the number of rings and in the area covered by the scattered signal. These are closely related to colony size and height. Modeling and experimental studies indeed reported that the number of rings rise when the amount of cells and number of layers composing the colony increase[14]. The angle of diffraction also increases, resulting in larger interfringes distances and larger scattering image areas. The relative intensity of the rings also varies from one scatterogram to another. This is a manifestation of variations in the colony texture, or, in other terms, of differences in the arrangement of the individual cells inside the colony. The impact of variations in size, height and texture of the colonies is particularly strong when considering 6-h-old colonies being in the beginning of their exponential growth phase. Elastic light scattering is found less sensitive to these parameters when colonies are incubated as long as 24 h and are in a stationary growth phase. We also observed that some broth media batches exhibit poor optical quality in the form of cell aggregates, leading to small superimposed patterns (see for example image 4) or even blurred images (image 11). We chose to nevertheless keep them in the dataset since such interferences may occur in real conditions.

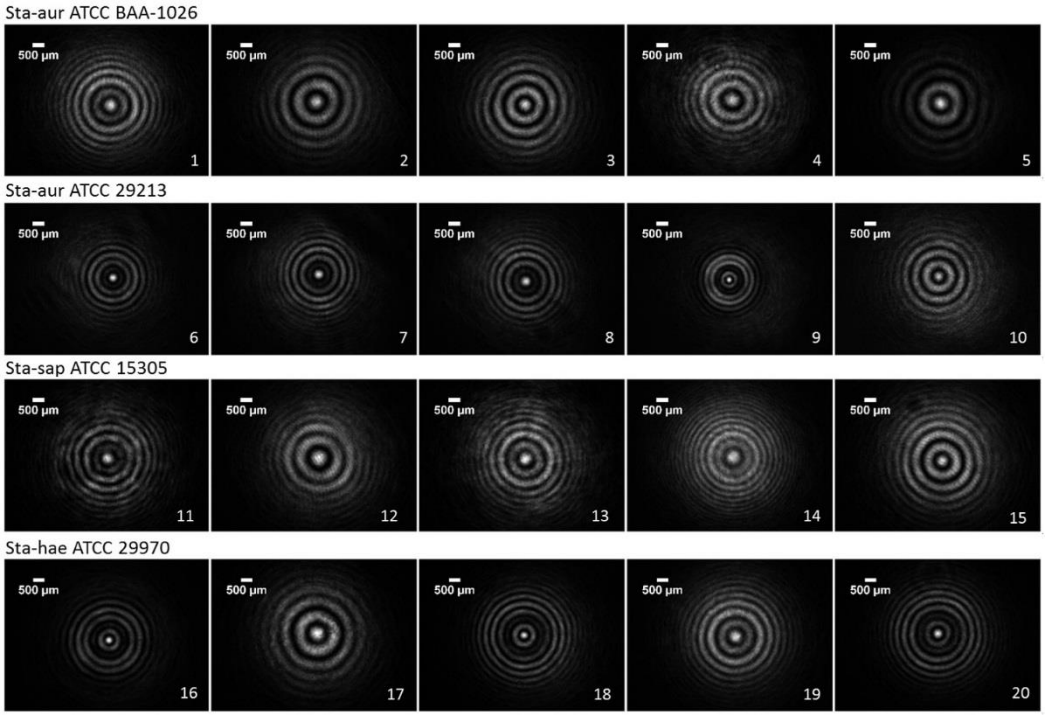

Figure 3. Examples of raw scattering images measured on microcolonies from 4 strains of *S. aureus* (1-10) and of other *Staphylococcus* species (11-20). Five images are presented for each strain to illustrate the intra-class variations.

Classification

As described in Section 2.4, an SVM classifier was trained on different types of descriptors extracted from images of the whole database. Different classification schemes have been assessed: classification on the 12 species, by directly learning the 12 species or by learning the 38 strains and deducing the species from the found strain, and classification on two classes (*S. aureus* vs. *S.* non-*aureus*), by directly learning the two classes or by learning the 38 strains and deducing the class from the found strain. The CIRs, estimated using a 5-fold cross-validation procedure, are listed in Table 2. They are given for the four classification schemes and the two types of descriptors (Fourier-Bessel moments and LBP descriptors) as well as for a third descriptor obtained by concatenating both types. Better CIRs are obtained with the LBP descriptors compared to the FB moments. The CIR is still better when both vectors are merged, at the price of a much greater number of variables. Besides, it is always preferable to learn the identification on subclasses (i.e. the strains) and then consolidate at the species (or *S. aureus* vs. *S.* non-*aureus*) level rather than directly learn on species. Indeed, learning at the strain level gives access to a finer model of classes. But this is also partly due to the database being far from balanced at the species level.

Table 2. Correct-identification rates (mean per species) obtained on the *Staphylococcus* database with different descriptors and classification schemes (12 species, 38 strains, 396 dishes, 5459 images, L2-normalization of descriptor vectors, linear SVM, 5-fold cross-validation).

Descriptor	Fourier-Bessel (FB) on centered images [n=100, m=10]	Local binary patterns (LBP) on raw images [N=8, R=1, 2, 4, 8, 16, 32]	FB + LBP merged
Number of variables	1100	216	1316
On 12 species	44.6%	72.3%	73.3%
On 12 sp., learned on 38 str.	55.4%	**79.9%**	**80.4%**
On 2 classes	78.3%	87.4%	88.5%
On 2 cl., learned on 38 str.	84.1%	**93.7%**	**94.7%**

The best configurations are also presented in the form of a confusion matrix in Figure 4. The best CIR is 94.7%, which is in accordance with performance generally reported for chromogenic tests on 24-h cultures[8], [15]. Most of the confusions are due to *S. pseudintermedius* images being misclassified as *S. aureus* in 20.8% of the cases. This result may reflect the high biological similarity between these two coagulase-positive species that have been considered as a single *S. aureus* species up to 2005[16], [17]. Considering coagulase-negative staphylococci species, we can note that identification at the species level remains problematic.

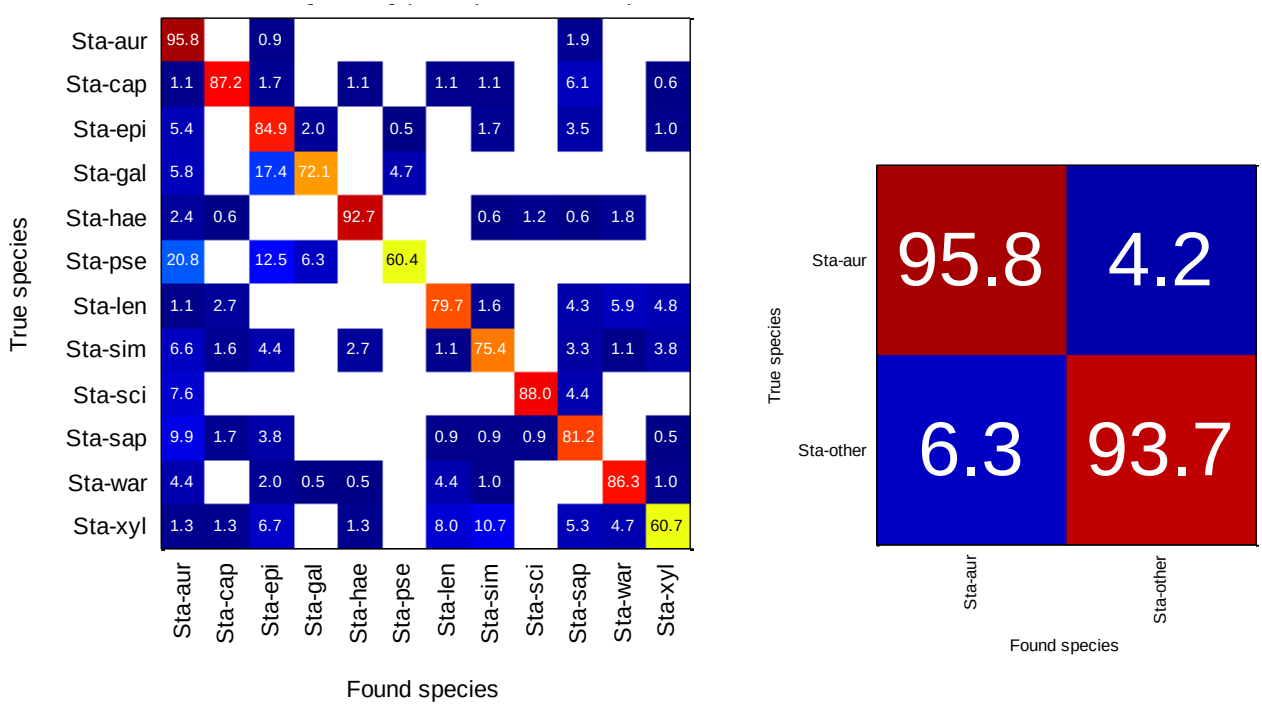

Figure 4. Confusion matrices obtained in the best configuration, for a classification on 12 *Staphylococcus* species (left; CIR = 80.4%) and on 2 classes *S. aureus* vs. *S.* non-*aureus* (right; CIR = 94.7%), both learned on the 38 strains. The figures are the classification rates in %.

3.2 Clinical dataset

Once the model trained on the reference database, it was applied to classify images collected from clinical samples as *S. aureus* or other *S. spp*. This step is crucial in assessing the performances of the method, since clinical samples necessarily differ from standard lab cultures due the strain identity and/or stress. The classification performance can be assessed through the sensitivity/specificity of the test or, alternatively, if the prevalence of S. *aureus* is known among the targeted population, through the positive and negative predictive values (PPV and NPV; see Section 2.4). The PPV and NPV of a test are important for clinicians since they give the probability that the result of the test, either positive or negative, is true, i.e. the probability that a patient has really the disease if the result is positive, and the probability that a patient has not the disease when the result is negative, respectively.

The results on the 224 images collected on patients are listed in Table 3 and Figure 5. One can observe that, on the patient images too, the LBP descriptors are better than the FB moments, but they are not outperformed by the merged FB+LBP descriptors. And, unlike the results obtained on the *Staphylococcus* database, learning on the 12 species or on the 38 strains degrades the test performance: a direct learning on the 2 classes gives much better performance.

Table 3. Sensitivity and specificity values, and corresponding positive and negative predictive values in brackets for an estimated prevalence of 30%, obtained in a *S. aureus* vs. *S.* non-*aureus* test performed on the clinical images (19 patients, 45 dishes, 224 images), based on the SVM learned on the *Staphylococcus* database, with same descriptors and classification schemes, except that the identification result is systematically examined as a dichotomous test *S. aureus* vs. *S.* non-*aureus* .

Descriptors	Fourier-Bessel (FB) on centered images [n=100, m=10]	Local binary patterns (LBP) on raw images [N=8, R=1, 2, 4, 8, 16, 32]	FB + LBP merged
Classification on 2 classes, learned on 2 classes	76.7% / 44.6% [37.2% / 81.7%]	**65.0% / 83.8%** **[63.2% / 84.8%]**	65.0% / 80.8% [59.2% / 84.3%]
Classification on 2 classes, learned of 12 species	85.0% / 31.5% [34.7% / 83.1%]	71.7% / 63.1% [45.4% / 83.9%]	71.7% / 63.8% [45.9% / 84.0%]
Classification on 2 classes, learned on 38 strains	76.7% / 29.2% [31.7% / 74.5%]	65.0% / 56.9% [39.3% / 79.1%]	71.7% / 49.2% [37.7% / 80.2%]

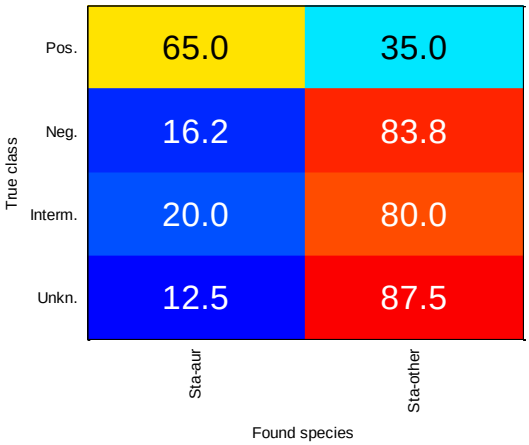

Figure 5. Confusion matrix obtained in the best configuration, for a classification learned on the 2 classes (*S. aureus* vs. *S.* non-*aureus*) with the LBP descriptor.

For that classification scheme with only two classes, one can optimize the test performance by scanning the discrimination threshold of the classifier, namely, the oriented distance from the SVM hyperplane. One thus plots the corresponding ROC curve (Receiver Operator Characteristic curve), which is the sensitivity value, *Se*, as a function of the complement of the specificity, 1 – *Sp* (see Figure 5, in the case of the LBP descriptor). For instance, if the sensitivity should be increased up to 90%, one could set the discrimination threshold at –1.6, which leads to a specificity of 56.9% (see Figure 5) and to positive and negative predictive values of 47,2% and 93.0%, respectively.

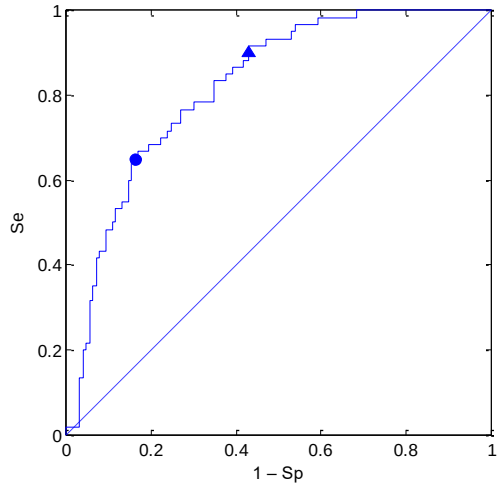

Figure 6. ROC curve plotted for a classification on 2 classes (*S. aureus* vs. *S.* non-*aureus*) with the LBP descriptor, by scanning the oriented distance from the SVM hyperplane. The circle (resp. triangle) shows the classification performance obtained with a discrimination threshold set at 0 (resp. –1.6).

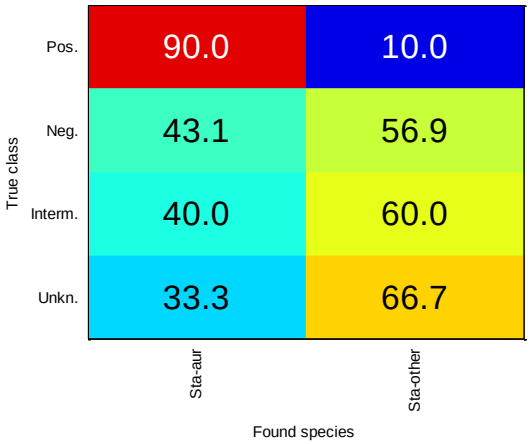

Figure 7. Confusion matrix obtained in the same configuration as in Figure 5, for a classification on 2 classes (*S. aureus* vs. *S.* non-*aureus*) with the LBP descriptor, when the discrimination threshold is lowered down to –1.6.

These results therefore indicate that, although we could not achieve both sensitivity and specificity at the same level as obtained with the reference database, we can adjust the classifier to optimize either the sensitivity or the specificity while keeping the other one at an acceptable level. From a clinician point of view, optimizing the NPV is a priority. Indeed it is preferable to have fewer false negatives but more false positives. Also, a high NPV value, such as the one reported here

(93%), means that a high degree of confidence can be assigned to the test. In other terms, a patient detected negative is unlikely to develop the infection in 93% of the cases.

Nevertheless, the performance is much lower than with the *Staphylococcus* database. This can be partly due to the strains present if the clinical samples, necessarily different from those included in the learning database. A likely heavier reason lies in the fact that the clinical samples generally present some latency before the exponential growth when cultivated in a Petri dish. This should account for microcolonies with sizes and development stages rather various and different from those obtained from standardized laboratory cultures, although the culture duration (6 h) is the same. We can assume that this has a strong impact on the diffraction patterns, thus on the calculated descriptors. A way to circumvent or reduce this inconvenience could be adding a pre-culture step for the clinical samples compared to the laboratory reference samples. The goal is to take account of the varying duration of the lag phase, which is induced by the adaptation required for bacterial cells to begin to exploit new cultivation conditions. The duration of such a pre-culture would remain to be optimized. Another, more direct way would consist in mimicking the stress conditions of the clinical samples for the reference database, or even including real clinical samples in the database.

4. CONCLUSION

The goal of the study was to assess Elastic Light Scattering as an alternative method to enzymatic chromogenic tests for the screening of *Staphylococcus aureus*. Indeed, the technique has major assets, such as the potential to accelerate the time-to-result to patients and caregivers while reducing analysis costs. We started to collect with our custom setup a reference database of 5459 images from 6-h-old microcolonies belonging to 12 species from *Staphylococcus* genus, corresponding to 22 strains of *S. aureus* and 16 strains of other *Staphylococcus* species. We investigated different feature extraction methods and different classification schemes to optimize the model. The beneficial of a texture-based descriptor in addition to morphological features has been shown, despite the fact the images look at the first glance mainly sharply ring-shaped and with few textural information. The best configuration has been obtained when an SVM classifier was applied on the 38 strains level, then consolidated up to a two-class problem, which led to a correct-identification rate of 94.7%. This result is encouraging for the deployment of Elastic Light Scattering in a clinical setting since it is in accordance with the culture-based chromogenic tests (93%). In order to assess the robustness of the model vis-à-vis the deviations between standardized lab cultures, such as the ones constituting the reference database, and real samples, we collected a clinical set of 224 images extracted from nasal swabs from 20 patients. The results in terms of sensitivity/specificity or negative and positive predictive values head in the right direction. Yet, difficulties appeared to provide a test with concomitant high sensitivity and specificity, and, more broadly, to reach the same level of performances as for the reference database. This suggests that light scattering technique applied as early as 6 h during the growth phase is particularly sensitive to the stress state of the bacterial cells. This manifests itself by various sizes and heights of the microcolonies, and thus differences in the scattering patterns. A way to overcome this diversity would be to include real samples in the reference database. For example, this has been successively implemented in commercially-available MALDI-TOF mass spectroscopy identification devices. One has yet to verify that a higher diversity in the data used to train the model will not deteriorate the performances.

The result of this study positions Elastic Light Scattering favourably on the major public heath application represented by *S. aureus* screening. By optimizing the dataset on which the model will be trained, and pursuing the investigation on feature extractors, we expect to increase the performances on real samples. The instrument will also be upgraded to improve the detection of microcolonies and accelerate the throughtput. The technique will indeed gain from producing and analyzing larger datasets, so opening up to the investigation of new classifiers such as neural networks and deep-learning algorithms.

REFERENCES

[1] Humphreys, H., Becker, K., Dohmen, P.M., Petrosillo, N., Spencer, M., van Rijen, M., Wechsler-Fördös, A., Pujol, M. Dubouix, A., and Garau, J., "*Staphylococcus aureus* and surgical site infections: benefits of screening and decolonization before surgery," Journal of Hospital Infection 94, 295-304 (2017).

[2] Liu, Y., Zhang,J., and J, Y., "PCR-based Approaches for the Detection of Clinical Methicillin resistant *Staphylococcus aureus*," The Open Microbiology Journal 10, 45-56 (2016).

[3] Perry, J.D, Rennison, C., Butterworth, L.A., Hopley, A.L.J., and Gould, F.K., "Evaluation of S. aureus ID, a New Chromogenic Agar Medium for Detection of *Staphylococcus aureus*," Journal of Clinical Microbiology 41, 5695-5698 (2003).

[4] Roberts, P., and Scopes, E., "Evaluation of Brilliance Staph 24 Agar For Detection of Staphylococci In A Clinical Setting," 17 June 2012, https://tools.thermofisher.com/content/sfs/brochures/Evaluation-of-Brilliance-Staph-24-Agar-For-Detection-of-Staphylococci-In-A-Clinical-Setting-LT2087-GLOBAL-EN-low-res.pdf (June 2012).

[5] Genuer, V., "Elastic light scattering for fast identification of pathogens," PhD thesis, Grenoble Alpes University, France (2017).

[6] Bae, E., Kim, H., McNally, H.A., Bhunia, A. K., Hirleman, E. D., "Investigation of the presence of rod-shaped bacteria on food surface via elastic light scattering," Advances in Bioscience and Biotechnology 3, 344-352 (2012).

[7] Sing, A.K., Bhunia, A.K., "Optical scatter patterns facilitate rapid differentiation of *Enterobacteriaceae* on CHROMagarTM Orientation medium," Microbial Biotechnology 9 (1), 127–135 (2016).

[8] Genuer, V., Gal, O., Méteau, J., Marcoux, P., Schultz, E., Lacot, É., Maurin, M., and Dinten, J.-M., "Optical elastic scattering for early label-free identification of clinical pathogens," Proc. SPIE 9698, 96980A (2016).

[9] Zhao, Z., and Singer, A., "Fourier-Bessel rotational invariant eigenimages," J. Opt. Soc. A Opt. Image Sci. Vis. 30, 871-877 (2013).

[10] Ojala, T., Pietikäinen, M., and Mäenpää, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Analysis and Machine Intelligence 24, 971-987 (2002).

[11] Skarbnik, N., "Local binary patterns," Matlab Central File Exchange (2012-2017), https://fr.mathworks.com/matlabcentral/fileexchange/36484-local-binary-patterns

[12] Vapnik, V., "Statistical Learning Theory," John Wiley and Sons, New York (1998).

[13] Chang, C.C., and Lin, C.J., "LIBSVM – A Library for Support Vector Machines" (2003-2016), http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] Bai, N., "Computational Modeling and Experimental Characterization of Bacterial Microcolonies for Rapid Detection using Light Scattering," PhD thesis, Purdue University, USA (2012).

[15] BioMérieux, "ChromID S. aureus Agar (SAID). Chromogenic medium for the selective isolation of staphylococci and the direct identification of S. aureus. Instructions for use, REF 43371", (2013), www.biomerieux.co.kr/upload/Package_Insert_-_12047_-_H_-_en_-_43371%5B1%5D-2.pdf (March 2013).

[16] Dmitrenko, O.A., Balbutskaya, A.A., Skvortsov, V.N., "Ecological features, pathogenic properties, and role of *Staphylococcus intermedius* group representatives in animal and human infectious pathology," Mol. Genet. Microbiol. Virol. 31, 117 (2016).

[17] Bannoehr, J., Ben Zakour, N. L., Waller, A. S., Guardabassi, L., Thoday, K.L.,Van den Broek, A. H. M., Fitzgerald, J.R., "Population Genetic Structure of the *Staphylococcus intermedius* Group: Insights into agr Diversification and the Emergence of Methicillin-Resistant Strains," J. Bacteriol. 189, 8685-8692 (2007).