Talanta 84 (2011) 730–737

Contents lists available at ScienceDirect

Talanta

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

Identification and discrimination of bacterial strains by laser induced breakdown spectroscopy and neural networks

D. Marcos-Martinez[a], J.A. Ayala[b], R.C. Izquierdo-Hornillos[a], F.J. Manuel de Villena[a], J.O. Caceres[a,*]

[a] *Departamento de Química Analítica, Facultad de Ciencias Químicas Universidad Complutense, 28040 Madrid, Spain*
[b] *Centro de Biología Molecular "Severo Ochoa", CSIC, C/Nicolás Cabrera, 1, Cantoblanco, 28049 Madrid, Spain*

ARTICLE INFO

Article history:
Received 13 October 2010
Received in revised form 17 January 2011
Accepted 30 January 2011
Available online 4 March 2011

Keywords:
Laser induced breakdown spectroscopy
Neural networks
Bacteria

ABSTRACT

A method based on laser induced breakdown spectroscopy (LIBS) and neural networks (NNs) has been developed and applied to the identification and discrimination of specific bacteria strains (*Pseudomonas aeroginosa*, *Escherichia coli* and *Salmonella typhimurium*). Instant identification of the samples is achieved using a spectral library, which was obtained by analysis using a single laser pulse of representative samples and treatment by neural networks. The samples used in this study were divided into three groups, which were prepared on three different days. The results obtained allow the identification of the bacteria tested with a certainty of over 95%, and show that only a difference between the bacteria can cause identification. Single-shot measurements were sufficient for clear identification of the bacterial strains studied. The method can be developed for automatic real time, fast, reliable and robust measurements and can be packaged in portable systems for non-specialist users.

1. Introduction

Laser induced breakdown spectroscopy (LIBS) analysis by direct measurement of the optical emission from laser-induced plasma has been the subject of research for many years [1,2]. This technique provides a quick and versatile method for analyzing different types of samples that can be inaccessible or tedious using conventional analytical techniques, and is particularly useful for analyzing samples with a complex matrix [3,4].

In many situations, a considerable number of alternative techniques with a higher sensitivity are available. However, LIBS provides several advantages over conventional methods for elemental analysis: (a) LIBS eliminates the sample preparation step for analysis; (b) The analysis can be performed in any state of matter (solid, liquid, gas); (c) The analysis is performed in a few seconds; (d) A very small amount of sample, on the order of micrograms, is vaporized from the surface of the sample; and (e) The analysis detects all elements without bias, including those present in molecules (which are atomized during the process) [5,6]. Two specific advantages of the application of LIBS in microbiological systems are: (a) Measurements can be performed remotely, allowing the analysis of hazardous, highly contagious, or pathogenic targets; and (b) The analysis can be computerized, no longer requiring the expertise of a trained microbiologist for the identification of bacteria or bio-agents.

Detection and identification of biological samples and, in particular, bacteria using the LIBS technique has been studied by several research groups [6–10]. The first of these studies focused on the use of LIBS as a system for early real-time detection of biological weapons. The motivation for most of these studies was LIBS' ability to provide extremely fast identification compared to traditional methods of identifying bacteria. For example, in preliminary experiments performed in 2003, Morel et al. investigated the detection of six strains of bacteria and two pollens [7]. They placed particular emphasis on *Bacillus globigii*, which acts as a non-pathogenic surrogate for *Bacilus anthracis* (anthrax), demonstrating the ability of LIBS to detect bacteria. That same year, Hybl et al. investigated the technique's potential for detecting and discriminating aerosolized bacterial spores from more common background fungal/mold spores and pollens [11]. More recently, the possibility of using LIBS to detect and identify bacteria in clinical diagnosis and public health has prompted investigations into the use of more realistic samples from a clinical analysis perspective. For example, the study by Rehse et al. [6] focusing on the identification of *Escherichia coli* and *Pseudomona aeruginosa* cultured in different media has been analyzed. This study succeeds in identifying or separating the two bacteria, producing positive results despite the modification of the bacterial wall of *P. aeruginosa* observed in some culture media. However, the correct identification rate of some cases studied fell below 90%. Thus, there is a clear need for more thorough and systematic studies that include new approaches.

The aim of this paper is to use a simple and direct method, based on LIBS and NNs, to identify and discriminate biological samples. In this case, no detailed chemical analysis was sought, but rather an

* Corresponding author. Tel.: +34 913944322.
E-mail address: jcaceres@quim.ucm.es (J.O. Caceres).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.01.069

instant identification of the sample using a unique characteristic of LIBS, which is its ability to generate a spectral “fingerprint” of the sample. This is due to the nature of the emission spectra, representative of the main compounds and the matrix that constituted the sample. The matrix structure and composition strongly affect the intensity of the emission lines, and have often inhibited a possible direct relationship between the elemental concentration of a sample and the intensities of its spectral lines. Thus, LIBS provides a unique spectrum, corresponding only to the sample under analysis. Using a correlation procedure, the LIBS–NN system developed can be trained to recognize spectra from different samples, which means evaluating the similarity of unknown spectra against a spectral library of classified samples. It is necessary to point out that the optical transmission properties of the optical fiber, the wavelength dependence of the spectrometer, and the wavelength efficiency of the detector elements all contribute to an overall wavelength dependence in the sample signal.

Many chemometric methods have been evaluated by other research groups, such as principal components analysis (PCA), soft independent modeling of class analogy (SIMCA), and partial least-squares discriminant analysis (PLS-DA). Those methods are not able to give satisfactory solutions to many practical problems that can be attributed mainly to uncertainty in identification, which can be even higher than 30% [5,12]. Research exists demonstrating that the use of NNs can provide better results. An interesting comparison of some of these methods with the use of NNs has been performed [13,14]. The NN was selected because it can significantly improve the identification capability without considerably increasing the difficulty of implementation. Specifically, in this work, the aim was to improve the recognition capacity by developing a method capable of identifying extremely similar samples that have few physical and spectral differences between them.

Thus, strong requirements were imposed on the identification model, where two fundamental aspects were introduced as follows. On the one hand, the broadest spectral range was selected in order to cover the greatest number of spectral characteristics of the sample. Tests performed using shorter spectral ranges with few peaks, selected by PCA, show that the model's recognition ability decreases. On the other hand, because the size of the data for the NN (denoted hereafter as input-data) can be quite large, a mathematical algorithm optimized to efficiently and effectively handle a large amount of data was used.

2. Materials and methods

2.1. Experimental set-up

Fig. 1 shows a schematic view of the LIBS technique. Experiments were performed by using a Q-switched Nd:YAG laser (Quantel, Brio model) operating at 1064 nm, with a pulse duration of 4 ns full width at half maximum (FWHM), 4 mm beam diameter and 0.6 mrad divergence. The samples were placed directly over an *X*–*Y*–*Z* manual micrometric positionator with a 0.5 μm stage of travel at every coordinate to ensure that each laser pulse impinged on a fresh sample. The laser beam was focused onto the sample surface with a 100 mm focal-distance lens, producing a spot of about 100 μm in diameter. This large working distance allowed easy sample manipulation and plasma light collection while the focusing provided by the lens enabled extremely precise placement of the beam within the bacterial target, but not in the adjacent substrate material. The pulse energy was 20 mJ, and the repetition rate was 1 Hz. The emission from the plasma created was collected with a 4-mm aperture, with a 7 mm focus fused silica collimator placed at a distance of 3 cm from the sample, and then focused into an optical fiber (with a 1000 μm core diameter and 0.22 numerical aperture), which was coupled to the entrance of the spectrometer. The spectrometer system was a user-configured miniature single-fiber system (EPP2000, StellarNet, Tampa, FL, U.S.A.) with a gated CCD detector. A grating of 300 l/mm was selected; a spectral resolution of 0.5 nm was achieved with a 7 μm entrance slit. The spectral range from 200 to 1000 nm was used. The detector integration time was set to 1 ms. To prevent the detection of bremsstrahlung, the detector was triggered with a 5 μs delay time between the laser pulse and the acquired plasma radiation using a digital delay generator (Stanford model DG535). The spectrometer was computer-controlled using an interface developed with Matlab, which allowed for data processing and real-time NN analysis.

Table 1
Nomenclature used for the samples.

Bacterial strains	Culture media 1 LB agar	Culture media 2 MacConkey agar	Culture media 3 Brucella anaerobic agar
Pseudomonas aeruginosa (B1)	B1M1 (11)	B1M2 (12)	B1M3 (13)
Escherichia coli (B2)	B2M1 (21)	B2M2 (22)	B2M3 (23)
Salmonella typhymurium (B3)	B3M1 (31)	B3M2 (32)	B3M3 (33)

2.2. Bacterial samples

All samples were used with no further preparation than that described herein. The samples were taken directly from a frozen culture, placed into the common Petri dish (8.9 cm in diameter), and incubated at 37 °C for 18 h. They were prepared on three different days, with a 10-day gap in between. The bacterial samples were of wild-type strains, *E. coli* OV2, *Salmonella typhymurium* LB5010, and *P. aeruginosa* M841. The media were LB agar (from Difco Microbiology, Lawrence, KS, U.S.A.), MacConkey agar (from Difco Microbiology, Lawrence, KS, U.S.A.), and Brucella anaerobic agar (from bioMerieux SA, Marcy l'Etoile, France), designated as culture media M1, M2 and M3, respectively. Because the NNs handle numbers, numerical and alphanumerical nomenclatures were used to label each sample. Table 1 shows the name structure used for the samples. The alphanumerical names indicate the bacterial strain and culture medium. The numbers in brackets correspond to the identification number of the bacterial sample. For example, sample 23 corresponds to B2M3, which means, B2 = *E. coli* and M3 = Brucella anaerobic agar. In addition, the first number in this numerical nomenclature correlates with the bacteria, and the second with the culture medium. A minimum of 3 replicate tests for each sample were performed, with a total of 81 samples analyzed.

2.3. LIBS measurements and spectral libraries

It has been shown that the selection of some wavelengths for elements such as P, C, Mg, Ca, and Na provides sufficient information to achieve the identification of bacteria [10]. However, in this work, a broad spectral range was selected in order to cover all spectral characteristics of the samples. This has the disadvantage of decreasing the spectral resolution, which makes it even more difficult to unequivocally identify the elements responsible for those lines, but has the advantage of reducing the analysis time. Taking into account that the plasma is generated from evaporated and ionized sample materials mixed with ambient gas [15], some spectral lines from air may be present in the spectra, making the elemental identification less effective. On the other hand, it has been demonstrated [16], that spectra from air can be correlated with the use

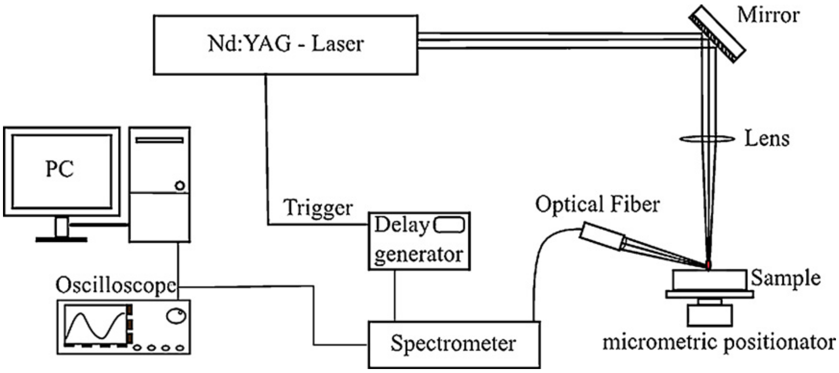

Fig. 1. The experimental setup, including a Nd:YAG laser, a delay generator, a micrometric positionator to move the sample, a 1 m optical fiber, and an optical charge coupled device (CCD) spectrometer.

of a simple multiplicative scaling factor, thus demonstrating that changes in relative spectral contributions from oxygen and nitrogen are not occurring. Furthermore, the contribution of different atomic ionization levels in the spectra does not change significantly. It is also necessary to consider the kinetic effect of the native CN molecular band usually observed in LIBS spectra for organic and biological samples [17,18], which corresponds to either native CN molecules vaporized from the sample, or CN bond formations due to recombination with the air.

Spectral emission lines and a continuum background emission are typical components of the spectra created in LIBS. In addition, the intensity of the spectra can change from pulse to pulse and from day to day, but does not affect the system's ability to identify the sample. There are two reasons for this. First, the system is trained with a set of spectra that were recorded using all possible variations of the ablation parameters, such as changes in the lens-sample distance and the laser pulse energy. Second, both the intensity ratio and the bandwidth were analyzed by the NNs, which made discrimination and identification possible.

Each sample was irradiated with 100 laser pulses. For each pulse, the generated plasma spectrum was acquired and stored as a column on a dataset. Thus, the dataset contains the intensity at different wavelengths in rows and the spectra in columns. Thus, our dataset has 2048 rows (one for each wavelength) and 100 columns or spectra for each sample. In order to avoid data variations due to changes in the laser pulse energy, each spectrum was normalized by the intensity of one specific spectral line (i.e., with the most intense assigned to hydrogen Hα, Fig. 2). Each of these individual worksheets containing the spectra for a specific sample constitutes a spectral library. 80 spectra were used to create the fingerprint, and the other 20 spectra were used to test the identification model. The greater the number of spectra used in the fingerprint of a sample, the better the recognition capacity of the method. A more thorough study of how recognition affected model identification is shown in Section 4. Because the acquisition of these 100 spectra is very fast (<2 min, taking into account the integration time and 1 Hz laser pulse repetition), 80 spectra were selected for the creation of the fingerprint. Although the data matrix can be considerably large, the computation time for training the NNs was always below 10 s.

2.4. NN model

The NNs used were based on a multilayer perceptron, feed-forward, supervised network. They consist of several neurons (information processing units) arranged in two or more layers. Each one receives information from all of the neurons in the previous layer. The connections are controlled by a weight that modulates the output from the neuron before inputting its numerical content into a neuron in the next layer. The process that optimizes the weights is called the learning or training process [19,20], and is based on a back-propagation (BP) algorithm [20]. The inputs from each neuron are added by an activation function, and the result is transformed by a transfer function that limits the amplitude of the neuron output. In this work, the hyperbolic tangent sigmoid function was used as the NN transfer function. When the NN parameters are adjusted by slightly refreshing the weights, the NN is able to learn from its environment. Every NN model was designed using Matlab software (Matworks, 2010a).

2.4.1. Description of the learning and verification set

Because the NNs were based on a supervised algorithm, in order to optimize the weight matrix, it was necessary to use input and output data that adequately characterized the process to be modeled. The input data was a linear combination of the dataset libraries. Data was randomly distributed into the learning (80%) and verification (20%). Once the learning and verification process was carried out, the network parameters were transferred to a program developed in Matlab that provided real-time identification during data acquisition.

2.4.2. NN model optimization and verification process

The NN model consisted of three layers (input, hidden and output), a topology widely used to model systems with a similar level of complexity [21]. In particular, the input layer consisted of 2048 nodes (intensity values in the 200–1000 nm wavelength range).

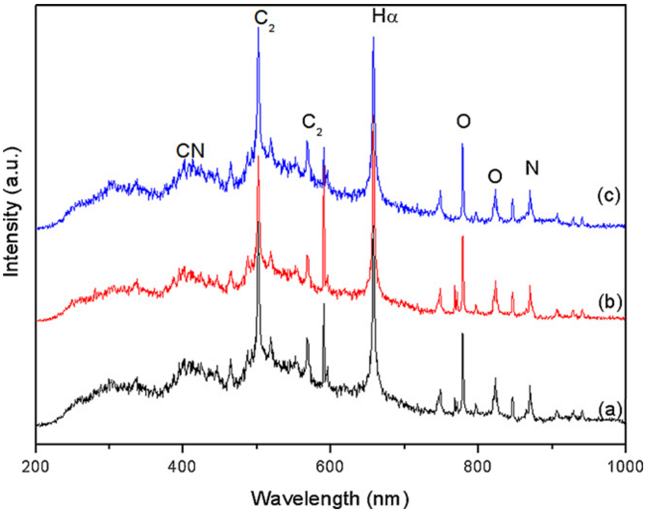

Fig. 2. Normalized LIBS single-shot spectra for 3 bacterial strains from day 1: (a) sample 11(B1M1); (b) sample 21 (B2M1); (c) sample 31 (B3M1). Plots (b) and (c) have been shifted for better observation.

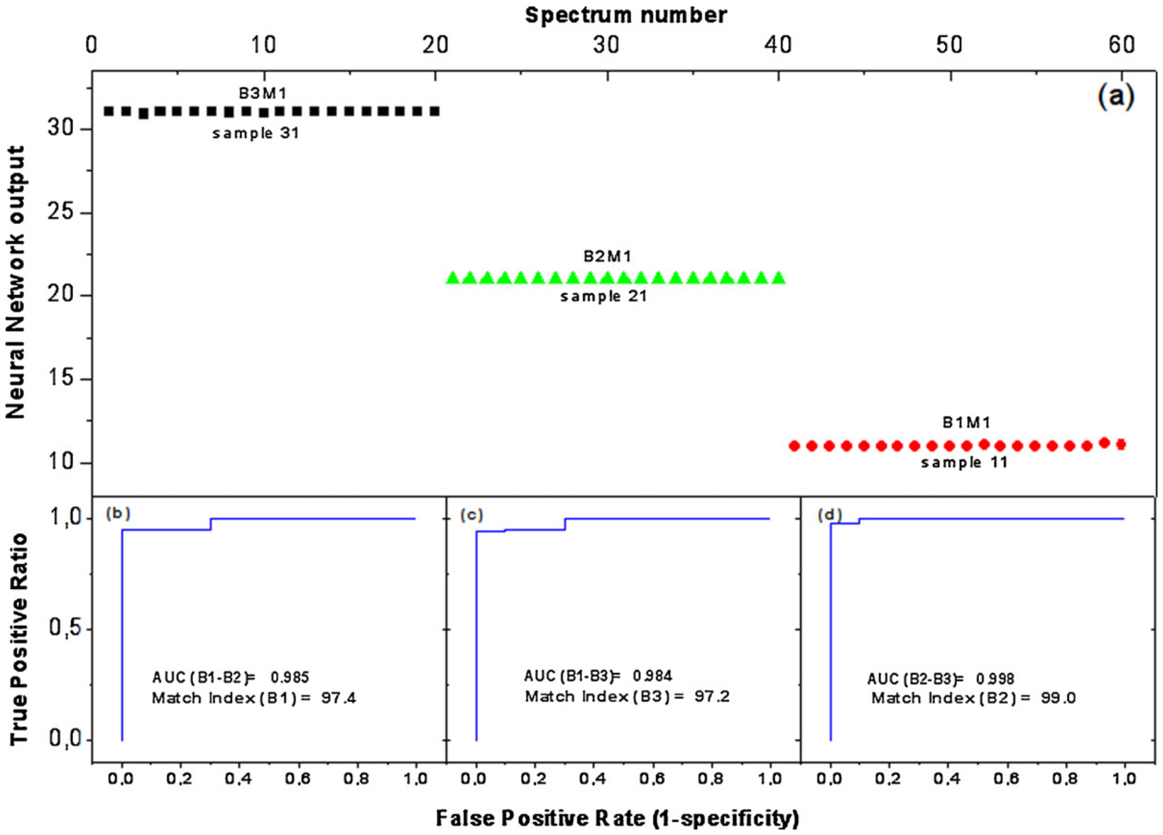

Fig. 3. (a) LIBS–NN correlation method applied to the identification of bacterial strains from day 1. Squares correspond to B3M1 (sample 31), triangles correspond to B2M1 (sample 21), Circles correspond to B1M1 (sample 11), (b), (c) and (d) correspond to ROC plot from Section 3, application to NN result for those bacterial strains; B1–B2 (AUC = 0.985 match index = 97.4%), B1–B3 (AUC = 0.984 match index = 97.2%) and B2–B3 (AUC = 0.998 match index = 99.0%) respectively.

The output layer was comprised of *J* neurons (where *J* = number of reference samples used) for estimating the similarity between the reference sample spectrum and the testing sample spectrum. The identification process was based on the ability of the NNs to detect the degree of similarity between the new spectrum and each of the reference spectra used in the learning process.

During the training process, each sample used as a reference was associated with an identification number (usually the same number assigned to the sample) in the output layer. Thus, a perfect identification was obtained if the output from the NN model for the test sample matched the identification number assigned to the reference. It is possible to use more than two identification numbers simultaneously, e.g., when analyzing a large number of samples. Zero was always used to indicate no match at all.

NN training was achieved by applying the BP algorithm, based on the conjugate gradient method [22], one of the general-purpose, second-order techniques that help minimize the goal functions of several variables. Second order indicates that such methods use the second derivatives of the error function, whereas a first-order

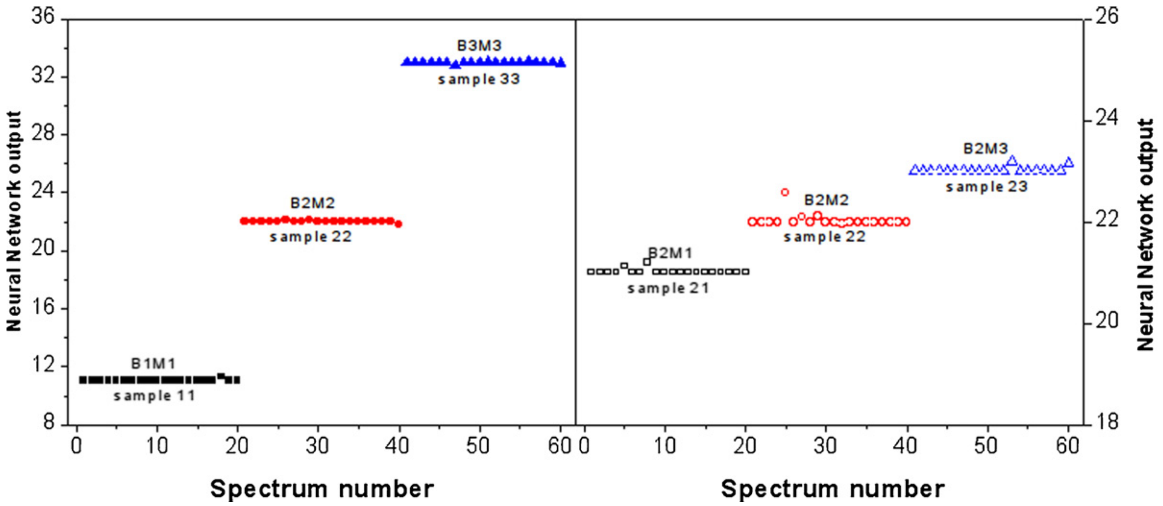

Fig. 4. (a) NN output for 20 spectra for sample B1M1, B2M2 and B3M3 from day 2 (not included in the training), and (b) the NN output for 20 spectra for samples B2M1, B2M2 and B2M3 from day 3. This confirms that the spectral library used in the training process was useful for all samples analyzed.

technique, such as standard back-propagation, uses only the first derivatives.

To determine when the training should be stopped, an early stopping criteria based on the validation set was used [23]. The number of epochs was not relevant in this case. To avoid an overfitting of the NN model, the learning process was repeated while the verification mean square error (MSE), defined in Eq. (1), was decreased:

$$\mathrm{MSE} = \frac{1}{N}\sum_{k}^{N}(r_k - y_k)^2 \tag{1}$$

where N, y_k, and r_k are the number of input data, the response from each output neuron, and the real output response, respectively. A detailed description of the calculation process is provided in the literature [19,23].

3. Process for testing the optimized LIBS/NN method

An identification process has only two possibilities: positive (P) or negative (N). There are therefore four possible results from this binary classifier. If the outcome from a prediction is P and the actual value is also P, this is a true positive (TP), but if the actual value is N, it is a false positive (FP). Conversely, a true negative (TN) occurs when the predicted outcome and the actual value are both N, and a false negative (FN) is when the predicted outcome is N, but the actual value is P.

Sensitivity (S) and *specificity* (SP) are two performance metrics for a screening process which are based on the number of TP, FP, TN and FN in the validation set. The S of a detection system can be calculated according to Eq. (2) [24,25]:

$$S = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$S = 1$ indicates that all samples are correctly identified. S is also called the true positive rate (TPR).

SP is defined as the proportion of correctly identified, fault-free recognitions [23]. The SP of a detection system can be calculated according to Eq. (3):

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

The probability of a false alarm (false positive rate, or FPR) is the proportion of fault-free recognitions that are classified erroneously. Obviously, FPR = 1 − SP. A more complete description of these metrics is provided in Ref. [24].

To evaluate a screening process, the values of both metrics (S and SP) are required, as neither one can properly evaluate the process as a single metric. This is because it is possible to force $S = 1$ for our detection system if all cases are reported as positive (SP = 0), or SP = 1 if the system reports all cases as negative ($S = 0$).

Accuracy (A) is the main parameter of a recognition procedure for decision making, and the reason the metrics for assessing detection processes are so important. These involve the relative frequency of correct and incorrect identification obtained from the results, and can be calculated according to Eq. (4) [24,26]:

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

The receiver operating characteristic (ROC) curve is the standard tool for plotting all possible combinations of sensitivity and specificity for a screening process. ROC plot analysis was developed in the context of electronic signal detection in the early 1950s [25] to evaluate classification performance [26]. The TPR is used as the y axis in an ROC plot, while the FPR(1 − SP) is used as the x axis. A ROC curve is at the ideal operating point when TPR = 1 and FPR = 0. It is widely accepted [24] that the area under curve (AUC) provides a better measure than accuracy for evaluating the predictive ability of the classification. This can be calculated using Eq. (5) [27,28]:

$$\mathrm{AUC} = \frac{S_0 - n_0(n_0 + 1)/2}{n_0 n_1} \tag{5}$$

where n_0 and n_1 are the numbers of positives and negatives, respectively, and $S_0 = \sum r_i$, where r_i is the i_{th} positive.

The confidence of the prediction can be expressed by a conditional probability, i.e. the rate of correct classification within the classified spectra (accuracy). The confidence was estimated by match index (MI). The higher the match index, the better the efficiency of the network for identifying a bacterial strain.

The robustness of the model was tested by assessing the ability of LIBS–NN to estimate the correct result when an unknown sample was input into the network model. In other words, the higher the robustness, the better the efficiency of the network model for identifying a sample not included in the training step as an “unknown,” and not identifying it as another bacteria [29]. In order to test this, one dataset (input data spectra) from the training set was removed. The results obtained with the remaining ones were then checked, in terms of the probability of correct identification. This was alternately repeated for each bacterial strain.

3.1. LIBS–NN model validation process

Three external validation processes were carried out, each with its respective set of input data. The first validation set was carried out to evaluate the model’s capacity to recognize and identify different bacterial strains in the same culture media. The test samples were comprised of the day 1 bacterial samples. Each bacterial sample was individually analyzed and contrasted with another from the same day in a binary manner.

To evaluate the validity of the model and reference datasets for identifying a bacterial strain cultivated on different days, a second validation set was performed using only the library spectra of the bacterial samples from day 1 as reference samples. Bacterial samples from the other two days were then individually contrasted with the library reference spectra, both at the same and in a different culture media.

A third independent validation set was carried out to test the model’s ability to identify unknown samples. To test the robustness of the model, the spectra of samples from day 1 were used as the library reference, and one bacterial dataset was removed and alternately repeated for each bacterial strain. The model was considered robust if the sample removed was identified as unknown and the NN output for these bacteria was zero.

4. Results and discussion

The NN was trained for spectral characteristics using the input data of known samples. The subsequent network model parameters were then handled by an identification program that performs real-time identification during data acquisition. Fig. 2 shows an example of a normalized single-shot spectrum for three bacterial strains in different culture media. The compositions of these samples were very similar, and their spectra look very similar, making visual identification difficult. Due to very fast testing conditions, and bearing in mind the high spectral similarity between the samples (for which a single shot might be the only sampling event), the spectra were not averaged.

4.1. LIBS–NN model validation

4.1.1. First validation process

The LIBS/NN correlation method was first applied to the identification of bacteria in the same day. A collection of single-shot

Table 2
Classification results from neural networks analysis.

Bacteria/medium	Classification results					
	Identification test				Robustness test	
	Match index	Unidentified	Misidentified	Correct identification	Misidentified	Correct as unknown
Pseudomonas aeruginosa/LB	97.4	0	0	100	0	100
Pseudomonas aeruginosa/MacConkey	98.2	0	0	100	0	100
Pseudomonas aeruginosa/Brucella	98.0	0	0	100	0	100
E. coli/LB	99.0	0	0	100	0	100
E. coli/MacConkey	100	0	0	100	0	100
E. coli/Brucella	100	0	0	100	0	100
Salmonella typhymurium/LB	97.2	0	0	100	0	100
Salmonella typhymurium/MacConkey	98.0	0	0	100	0	100
Salmonella typhymurium/Brucella	100	0	0	100	0	100

spectra for each sample were analyzed and contrasted with the spectra for reference samples in the same day, in a binary manner. In each binary analysis, part of the library for the sample was used as reference. As an example, Fig. 3a shows the analysis of sample B1M1 and B2M1 (*Pseudomona* and *E. coli* on the same cultured media LB agar). 80 spectra for each sample were used as a reference sample set, where 11 and 21 were assigned as identification numbers, respectively. 20 spectra for samples not included in the reference were tested. The references were assigned the same bacterial identification numbers for the NN output, taking into account the bacteria and matrix. Thus, if the NN has an output equal to 11, this corresponds to the identification of bacterial strains B1M1 shown in Table 1. If the NN output contains an output equal to 21, this corresponds to the identification of bacteria B2M1, and so on.

With the first laser shot, the network recognizes the spectrum as belonging to a reference dataset, and the NN output is 21. Therefore, the NN model cannot “see” the difference between the analyzed sample and the reference sample with an identification number equal to 21. The same result was obtained up to spectrum 20. Then, when test sample B2M1 is replaced by sample B1M1, the network output is 11. At this point, the NNs correctly assign the spectra to the identification number equal to 11 used for this sample. Most of the spectra for this sample were assigned correctly, and only two deviated from the expected behavior, returning values in between the two values assigned. This behavior strongly affects both the match index and the AUC, which were 97.4% and 0.985, respectively. Given that the spectra analyzed came from a single laser shot, the disturbance spectra observed in only 2 out of 20 is not only more than acceptable, but is essential for taking into account the match index. Exactly the same results were obtained when using the dataset for all other bacterial strains from day 1 as a reference, and updating the NN parameters, as shown in Table 2.

Fig. 3b shows the ROC plot (Eqs. (2) and (3)) and match index (Eq. (4)) obtained for this case, and the capacity of the LIBS/NN to identify those bacterial strains. It means that the differences between the two colonies cannot be attributed to matrix variations. Only differences between bacteria can cause identification within the same matrix. Another important result is obtained from this test. The high match index shows that, despite not taking into account the contributions of both air and the culture medium, a correct identification can be achieved, and helps to decrease the analysis time without significantly affecting the model discrimination capacity.

4.1.2. Second validation process

As part of a tough test for evaluating the validity of the reference matrix with regard to time and the method's ability to identify bacteria cultivated on different days, the spectra for the bacterial strains in day 1 were introduced into the NN model as references. The mathematical procedure followed was similar to the training and validation process described above (see Table 2). Fig. 4a shows the results obtained for 3 samples from day 2 (B1M1, B2M2 and B3M3). Twenty single-shot spectra for each sample were measured. It can be seen that the samples have been correctly identified from the first laser shot. Results also show the network's capacity to work simultaneously with more than one fingerprint, without significantly increasing the computing time. The training process for all samples from day 1 used as references required 7.2 s of execution time on a standard computer. Fig. 4b shows the results obtained for *E. coli* (B2) from day 3 (last date) in three different media. The bacteria were correctly identified, even in this case where the identification was made using the first library (the oldest) as a reference, and comparing to libraries obtained subsequently. This confirmed the temporal validity of the libraries, at least during this study. In this case, shown in Fig. 4b, the reader might assume that identification of the bacteria should not be dependent on the culture medium. However, as described previously by Rehse [6], significant differences were observed in samples obtained from bacteria grown on MacConkey agar plates. This difference was interpreted as a real and not unexpected elemental alteration of the membrane of the bacteria cultured in that medium, and did not represent an inherent limitation of the LIBS technology. The alteration of fundamental bacteria chemistry was attributed to the presence of bile salts in the MacConkey medium, which is known from biochemistry to disrupt membrane integrity. Because standard serological (antibody-based) diagnostics are also membrane-based, rather than genetically based, these competing microbiological techniques might misidentify such bacteria if their outer membrane or surface has been significantly altered.

4.1.3. Third validation process

In order to complete the validation of the prediction capability of the optimized NN, a third independent validation set was carried out in order to test the model's ability to identify unknown samples as unknown. To test the robustness of the model, the spectra of samples from day 1 were used as the library reference, and one bacterial dataset was removed and alternately repeated for each bacterial strain.

The NN compares the analyzed sample spectra with those stored as a reference (like two fingerprints). Therefore, if they match, the output from the network is satisfactory, and the value assigned to the reference and NN output match. If the fingerprints (spectra) differ slightly, the NN output must be zero. This means that the bacteria are not present in the training set, and are therefore unknown. As an example, Fig. 5 shows a possible stronger test. In this case, all samples from day 1 were used as a reference, and the spectra for *Salmonella* (B3) were removed. The NN output for 20 spectra from each sample in day 2 (B3M1, B2M1 and B1M1) was tested. The first 20 spectra correspond to sample B2M1, the

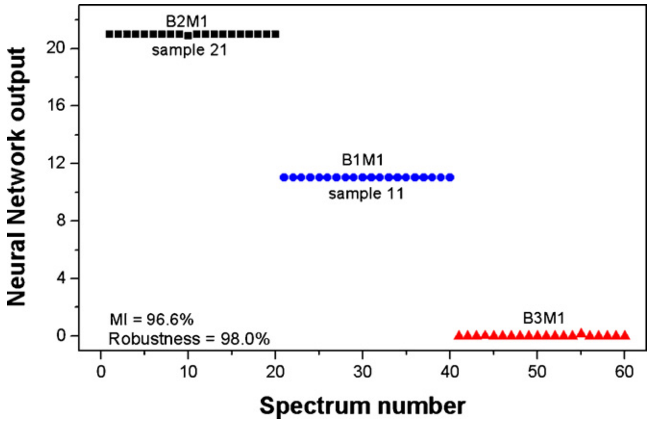

Fig. 5. NN outputs for B1M1, B2M1 and B3M1, where the latter corresponds to a sample set not included in the training set. Thus, this B3M1 sample was unknown for the NN model. Sample B3M3 was correctly assigned a zero, the identification number assigned to unknown samples. The match index was 96.6%, and the robustness was 98.0%.

next 20 to sample B1M1, and the last 20 to sample B3M1. All samples were correctly identified. Even sample B3M3 was correctly identified as unknown, and not as another sample present during training, which demonstrates the robustness of the method used.

Some interesting modes of identification can be observed. To improve the correct identification rate, the number of spectra for each sample must increase to improve the match index. It is therefore necessary to select a high number of spectra, and because each spectrum comes from a single laser pulse, the time taken to collect the data (1 s per spectrum) is not an issue. The broad spectral range used also plays a key role in correct identification.

4.2. *Number of spectra used in the training set*

Finally, to evaluate the optimum number of spectra used in the training process, the variations in robustness as a function of the number of spectra used in the training matrix was studied. Fig. 6 shows a plot of these results. As we can observe, robustness increases rapidly with the number of spectra. Even for very low numbers (8 spectra), the robustness is acceptable.

The time required to obtain the spectra is very fast. Once stored, they are selected for further analysis in real time. Analyses carried out with different types of bacteria (generated at different times) showed that the libraries were adequate for correct identification of the bacterial strains, even with small variations in the experimental conditions, such as changes in laser energy, room temperature, or sample distance.

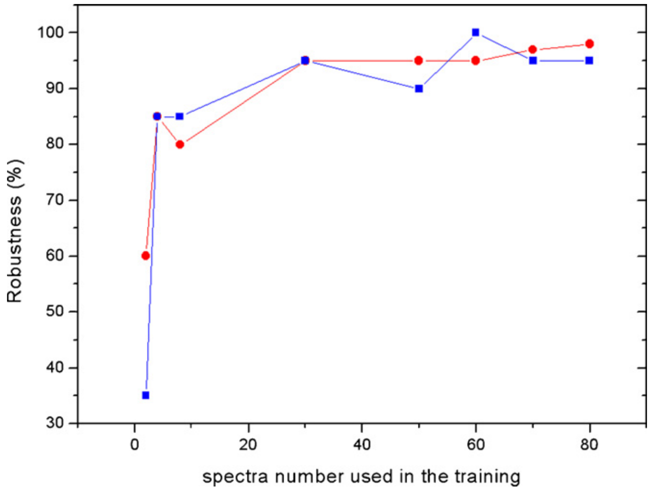

Fig. 6. Robustness obtained for the samples B1M1 (squares) and B2M2 (circles), as a function of the number of spectra used in the process of training the NN.

Single-shot measurements were sufficient for clear identification of the bacterial strains studied. In light of these results, the optimized NN model provides reliable results (sample identification) for all samples analyzed. This result is the best indicator of the capacity of the methodology presented.

5. Conclusions

It has been shown that accurate sample analyses can be obtained using LIBS/NN. Tests performed on bacteria samples demonstrated 100% reliable identification of known and unknown samples with very similar spectral characteristics. In addition, in studies where the only variation was the type of bacteria, the identification was correct, and therefore did not depend on the culture medium. Only differences between bacteria resulted in identification. Despite not taking into account the contributions of the air and culture medium, a correct identification can be achieved, which helps decrease the analysis time without significantly affecting the model's discrimination capacity.

The identification analysis was stable over a long period of time, and minor changes in experimental conditions, such as the intensity of the LIBS single-shot regimen and continuum background, were not relevant for sample identification. The system was able to perform a correct identification even with a single laser shot. The most important conclusion is that in the 200–1000 nm range, each spectrum is a true fingerprint of the sample, allowing correct differentiation of each bacterial strain using the NN model.

Multivariate techniques are known to be efficient methods for sorting and classifying data. However, the results of this study show that better reproducibility data and the introduction of advanced statistical models are needed to produce robust classification models. Clearly, the sample size was small, representing the lower limit of practical application. However, the verification test emphasized that the methodology used in this work can provide a measure of confidence classification that may have practical significance. The study will be extended to characterize different bacteria and, more importantly, to differentiate pathogenic bacterial strains, thus demonstrating medical diagnosis potential. This work is currently underway in our laboratories. The equipment and methods used in this work can be developed for quick, automatic, reliable and robust measurements in real time.

Acknowledgements

This work was supported by the Complutense University of Madrid (grant CCG08-UCM/PPQ-3713) granted to JO Caceres, and grant BFU2006-04574 from the Ministry of Science and Innovation of Spain, granted to JA Ayala. The authors acknowledge the work performed by the anonymous referees who helped improve this paper.

References

[1] D.A. Cremers, L.J. Radziemski, Handbook of Laser-induced Breakdown Spectroscopy, John Wiley & Sons, Chichester, UK, 2006.
[2] A.W. Miziolek, V. Palleschi, I. Schechter, Laser-induced Breakdown Spectroscopy (LIBS), Cambridge University Press, Cambridge, UK, 2006.
[3] J.O. Cáceres, J. Tornero López, H.H. Telle, A. González Ureña, Spectrochim. Acta B 56 (2001) 831–838.
[4] W. Lee, J. Wu, Y. Lee, J. Sneddon, Appl. Spectrosc. Rev. 39 (2004) 27–97.
[5] S.J. Rehse, N. Jeyasingham, J. Diedrich, S. Palchaudhuri, J. Appl. Phys. 105 (2009) 102034–102113.
[6] S.J. Rehse, J. Diedrich, S. Palchaudhuri, Spectrochim. Acta B 62 (2007) 1169–1176.

[7] S. Morel, N. Leone, P. Adam, J. Amouroux, Appl. Opt. 42 (2003) 6184–6191.
[8] M. Mordmueller, C. Bohling, A. John, W. Schade, in: J.C. Carrano, C.J. Collins (Eds.), Optically Based Biological and Chemical Detection for Defence, vol. V, SPIE, Berlin, Germany, 2009, p. 74840F-10.
[9] R.A. Multari, D.A. Cremers, J.M. Dupre, J.E. Gustafson, Appl. Spectrosc. 64 (2010) 750–759.
[10] S.J. Rehse, Q.I. Mohaidat, S. Palchaudhuri, Appl. Opt. 49 (2010) C27–C35.
[11] J.D. Hybl, G.A. Lithgow, S.G. Buckley, Appl. Spectrosc. 57 (2003) 1207–1215.
[12] F. Yueh, H. Zheng, J.P. Singh, S. Burgess, Spectrochim. Acta B 64 (2009) 1059–1067.
[13] P. Inakollu, T. Philip, A.K. Rai, F. Yueh, J.P. Singh, Spectrochim. Acta B 64 (2009) 99–104.
[14] R.E. Shaffer, S.L. Rose-Pehrsson, R.A. McGill, Anal. Chim. Acta 384 (1999) 305–317.
[15] A. Sarzyski, W. Skrzeczanowski, J. Marczak, Proceedings of SPIE, Munich, Germany, 2007, pp. 66180V–66180V-10.
[16] R.J. Nordstrom, Appl. Spectrosc. 49 (1995) 1490–1499.
[17] A. Portnov, S. Rosenwaks, I. Bar, Appl. Opt. 42 (2003) 2835–2842.
[18] M. Baudelet, M. Boueri, J. Yu, S.S. Mao, V. Piscitelli, X. Mao, R.E. Russo, Spectrochim. Acta B 62 (2007) 1329–1334.
[19] H.B. Demuth, M.H. Beale, M.T. Hagan, Neural Network Toolbox for Use with MATLAB: User's Guide 9th for Version 6.0 (Release 2008a), Math Works, 2007.
[20] A.J. Maren, C.T. Harston, Handbook of Neural Computing Applications, Academic Press, San Diego, USA, 1990.
[21] J. Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Tellier, M. Potin-Gautier, I.L. Hecho, Anal. Bioanal. Chem. 385 (2006) 256–262.
[22] M.F. Møller, Neural Networks 6 (1993) 525–533.
[23] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, USA, 1996.
[24] Y. Jiang, B. Cukic, Y. Ma, Empir. Softw. Eng. 13 (2008) 561–595.
[25] T. Fawcett, Pattern Recogn. Lett. 27 (2006) 861–874.
[26] G.I. Webb, K.M. Ting, Mach. Learn. 58 (2005) 25–32.
[27] C. Ling, J. Huang, H. Zhang, Advances in Artificial Intelligence, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 329–341.
[28] D.J. Hand, R.J. Till, Mach. Learn. 45 (2001) 171–186.
[29] J. Sirven, B. Salle, P. Mauchien, J. Lacour, S. Maurice, G. Manhes, J. Anal. At. Spectrom. 22 (2007) 1471–1480.