Spectrochimica Acta Part B 87 (2013) 161–167

Contents lists available at ScienceDirect

Spectrochimica Acta Part B

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

A comparison of multivariate analysis techniques and variable selection strategies in a laser-induced breakdown spectroscopy bacterial classification

Russell A. Putnam [a,1], Qassem I. Mohaidat [b], Andrew Daabous [a,1], Steven J. Rehse [a,*]

[a] *Department of Physics, University of Windsor, Windsor, Ontario N9B 3P4, Canada*
[b] *Department of Physics, Yarmouk University, Irbid 21163, Jordan*

ARTICLE INFO

Article history:
Received 15 November 2012
Accepted 12 May 2013
Available online 20 May 2013

Keywords:
Laser-induced breakdown spectroscopy
Bacterium
Discriminant function analysis
Partial least squares discriminant analysis
Multivariate analysis

ABSTRACT

Laser-induced breakdown spectroscopy has been used to obtain spectral fingerprints from live bacterial specimens from thirteen distinct taxonomic bacterial classes representative of five bacterial genera. By taking sums, ratios, and complex ratios of measured atomic emission line intensities three unique sets of independent variables (models) were constructed to determine which choice of independent variables provided optimal genus-level classification of unknown specimens utilizing a discriminant function analysis. A model composed of 80 independent variables constructed from simple and complex ratios of the measured emission line intensities was found to provide the greatest sensitivity and specificity. This model was then used in a partial least squares discriminant analysis to compare the performance of this multivariate technique with a discriminant function analysis. The partial least squares discriminant analysis possessed a higher true positive rate, possessed a higher false positive rate, and was more effective at distinguishing between highly similar spectra from closely related bacterial genera. This suggests it may be the preferred multivariate technique in future species-level or strain-level classifications.

1. Introduction

Since the initial demonstrations of bacterial identification with laser-induced breakdown spectroscopy (LIBS) in 2003, significant progress has been made in the use of multivariate chemometric analyses to classify unknown bacterial LIBS spectra [1–4]. Over the last five years we and others have demonstrated a sensitive and specific identification of live bacterial biospecimens utilizing a discriminant function analysis (DFA) to classify LIBS spectra [5–8]. The intensities of strong specific elemental atomic emission lines normalized by the total observed spectral power have been utilized as independent variables in this multivariate analysis [9]. The selection of specific spectral lines to serve as independent variables in the multivariate analysis is known as variable down-selection [10]. However it is not yet known whether the use of down-selected variables or the entire LIBS spectrum provides optimal discrimination and classification of unknown LIBS spectra, and this is an ongoing area of investigation [11,12]. It is also not known which multivariate analysis technique, if any, provides superior classification given a choice of independent variables, and multiple chemometric algorithms are still widely utilized for bacterial identification including principal component analysis (PCA), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), neural network (NN) analysis, partial least squares (PLS) regression, and support vector machine classification (SVM) [13–18].

To investigate these various strategies, we have compared the use of three different down-selected variable models consisting of emission intensities, the sum of observed intensities from the elements P, Ca, Mg, Na, and C, and complex ratios of those intensities in identical external validation tests. Variables were down-selected from bacterial LIBS spectra obtained from five different genera and 13 distinct taxonomic classes of species and strains [8]. Model performance was quantified by calculating truth tables (and the resulting sensitivity and specificity) from the external validation tests. Lastly, the down selected variable model which provided the most accurate classification was tested in a PLS-DA multivariate analysis to provide a direct comparison with the performance of the DFA.

2. Experimental

2.1. Experimental setup

The LIBS apparatus used to obtain the bacterial spectra, as well as our bacterial sample preparation and mounting protocols, have been described at length elsewhere [5,19]. Briefly, 1064 nm infrared laser pulses 10 ns in duration were used to ablate the bacterial specimens mounted on a 0.7% nutrient-free agar substrate in an argon environment. LIBS emission was collected 2 μs after the ablation pulse and dispersed in an Échelle spectrograph, and the spectra were recorded

[*] Corresponding author. Tel.: +1 519 253 3000; fax: +1 519 973 7075.
E-mail addresses: putnamr@uwindsor.ca (R.A. Putnam), q.muhaidat@yu.edu.jo (Q.I. Mohaidat), daabousa@uwindsor.ca (A. Daabous), rehse@uwindsor.ca (S.J. Rehse).
[1] Fax: +1 519 973 7075.

0584-8547/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sab.2013.05.014

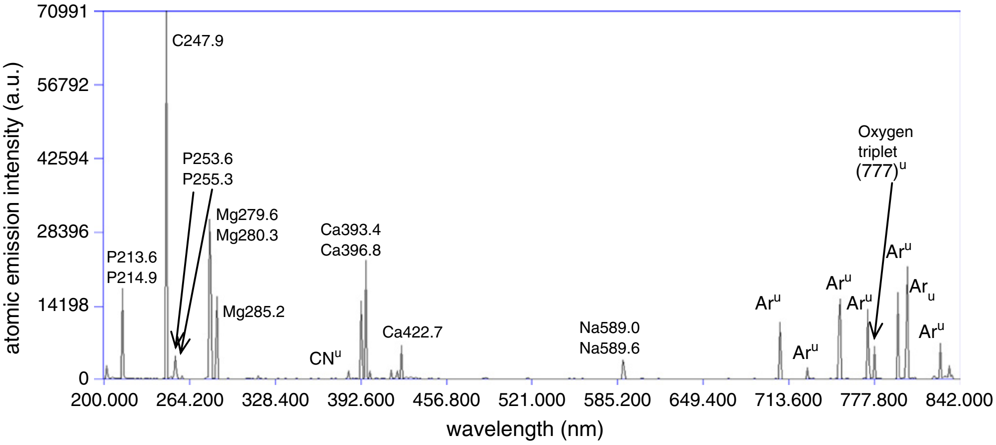

Fig. 1. A representative LIBS spectrum of a bacterial target ablated in an argon environment at atmospheric pressure. The atomic emission lines used in the bacterial discrimination indicated by an "*" in Table 3 are indicated in this spectrum. Emission features that were seen but were unused in the discrimination are indicated with a superscript "u".

by an intensified charge-coupled device (ESA3000, LLA Instruments, GmbH). Pulse energies were approximately 10 mJ/pulse and each spectrum was averaged from spectra acquired at five sampling locations, each approximately 100 μm in diameter. Approximately 7500 bacterial cells in total were ablated for each spectrum [5]. A representative LIBS spectrum of a bacterial target ablated on an agar substrate in an argon atmosphere is shown in Fig. 1. This spectrum is the averaged accumulation of five separate sampling locations. Five spectra were acquired at each sampling location, thus twenty-five laser pulses were used to obtain this spectrum.

The bacteria were chosen to represent a fairly wide taxonomic range. Spectra were acquired from representative Gram-negative phenotypes (*Escherichia coli* and *Enterobacter cloacae*), Gram-positive phenotypes (two species of *Staphylococci* and two species of *Streptococci*), and the atypical acid-fast *Mycobacterium* phenotype (three strains of *Mycobacterium smegmatis*). In total, LIBS spectra from 13 unique bacterial strains were obtained in 32 completely distinct experiments (e.g. cultured in different media, grown on different days over the course of 18 months, and exposed to different environmental stresses) [8]. This is shown in Table 1.

The five representative bacterial genera that were tested are listed in the first column of Table 1 and the thirteen bacterial taxonomic groups tested (e.g. *E. coli* strain C, *E. coli* strain HF4714, *Staphylococcus aureus*, *Staphylococcus saprophyticus*) are listed in column two. The 32 distinct experiments that were performed yielded the 32 data sets shown in column three of Table 1. Each distinct experiment was

Table 1
Identities of the 32 data sets used to construct a spectral library composed of 669 bacterial LIBS spectra.

Genus	Bacterial ID	Data set
1: *Escherichia*	1: *E. coli* ATCC 25922	1: *E. coli* ATCC 25922
	1: *E. coli* ATCC 25922	2: *E. coli* ATCC 25922/*E. cloacae* (10:1)
	1: *E. coli* ATCC 25922	3: *E. coli* ATCC 25922/*E. cloacae* (100:1)
	1: *E. coli* ATCC 25922	4: *E. coli* ATCC 25922/*E. cloacae* (1000:1)
	2: *E. coli* O157:H7 (EHEC)	5: *E. coli* O157:H7
	3: *E. coli* C	6: *E. coli* C
	3: *E. coli* C	7: *E. coli* C — cultured on MacConkey agar
	3: *E. coli* C	8: *E. coli* C — starved for 1 day
	3: *E. coli* C	9: *E. coli* C — starved for 4 days
	3: *E. coli* C	10: *E. coli* C — starved for 6 days
	3: *E. coli* C	11: *E. coli* C — starved for 8 days
	3: *E. coli* C	12: *E. coli* C — autoclaved
	3: *E. coli* C	13: *E. coli* C — UV exposed/killed
	4: *E. coli* HF4714	14: *E. coli* HF4714
	5: *E. coli* Hfr-K12	15: *E. coli* Hfr-K12
2: *Enterobacter*	6: *E. cloacae* ATCC 13047	16: *E. cloacae* ATCC 13047
3: *Staphylococcus*	7: *S. saprophyticus*	17: *S. saprophyticus*
	8: *S. aureus*	18: *S. aureus*
4: *Streptococcus*	9: *S. mutans*	19: *S. mutans*
	10: *S. viridans*	20: *S. viridans*
	10: *S. viridans*	21: *S. viridans* — starved for 1 day
	10: *S. viridans*	22: *S. viridans* — starved for 6 days
	10: *S. viridans*	23: *S. viridans* — starved for 9 days
	10: *S. viridans*	24: *S. viridans* — UV exposed/killed
	10: *S. viridans*	25: *S. viridans* — autoclaved
5: *Mycobacterium*	11: *M. smegmatis* WT	26: *M. smegmatis* WT — 90% dilution
	11: *M. smegmatis* WT	27: *M. smegmatis* WT — 60% dilution
	11: *M. smegmatis* WT	28: *M. smegmatis* WT — 50% dilution
	11: *M. smegmatis* WT	29: *M. smegmatis* WT
	11: *M. smegmatis* WT	30: *M. smegmatis* WT — 100% concentration
	12: *M. smegmatis* TE	31: *M. smegmatis* TE
	13: *M. smegmatis* TA	32: *M. smegmatis* TA

performed with one aliquot of bacteria prepared separately from the others and thus each data set represents completely unique experimental data. For example, data set 6, "*E. coli* C" which would have yielded approximately 20 spectra and data set 12, "*E. coli* C — autoclaved" which would have yielded another 20 spectra, were all obtained from aliquots ultimately derived from the same mother strain of bacteria, but tested many months apart from each other, grown from completely different cultures each using freshly prepared nutrient media, and handled differently. In this case one of the aliquots was placed in a microbiological autoclave prior to testing to render the sample inactive. Also, the LIBS apparatus would have been cycled dozens of times in between the acquisition of these data sets (including the cleaning of optics, realignment of beams, and adjusting of laser pulse energy for use in other experiments) This point should be emphasized, as the high degree of reproducibility through time evidenced by the chemometric classification of these spectra suggests that these results were not very sensitive to uncontrollable experimental fluctuations that would be expected in measurements taken over such a long period of time and with bacterial specimens handled in such disparate ways. We believe this is an indicator of the highly robust nature of the LIBS-based identification method.

Twenty to thirty spectra were obtained in approximately 30 min in each experiment yielding the data sets shown for a total of 669 LIBS spectra. The number of spectra obtained in any one experiment was limited only by the ability to translate the laser spot around the approximately 1 cm^2 bacterial deposition. Although efforts were taken to try to obtain highly similar spectra from each bacterial deposition, no data "outliers" were omitted from our data sets and efforts were made to maximize the number of spectra from every bacterial deposition rather than to standardize the number of spectra taken.

2.2. Models for chemometric analysis (lines, RM1, and RM2)

The three independent variable models that were tested are referred to here as the "lines" model, ratio model one (RM1), and ratio model two (RM2). The lines model was the simplest of the three, having been used in all our previous work. It consisted of the intensities of thirteen strong emission lines normalized by the total spectral power of the LIBS spectrum. The intensity of a line was taken to be the total integrated area under the curve of the background-subtracted emission line profile and the total spectral power was the sum of the thirteen intensities. The identities of the thirteen lines are provided in the detailed discussion of RM2 below and are shown in the spectrum in Fig. 1.

RM1 consisted of 24 independent variables, shown in Table 2. The first five variables were the sums of the measured intensities for each element including the sum of four phosphorus lines, one carbon line, three magnesium lines, three calcium lines, and two sodium lines. No distinction was made between lines from neutral and singly-ionized species in these sums. This strategy was briefly investigated, but was found to add little to the analysis. Aside from the fact that these lines were highly robust and exhibited excellent signal-to-noise in the bacterial LIBS spectrum, these five specific elements (P, C, Ca, Mg, and Na) are very important to bacterial function and physiology, and thus to the LIBS-based identification. This has been discussed by us in depth previously [9].

Table 2
The twenty-four independent variables used in ratio model one (RM1).

P (sum)	Mg/Ca
C (sum)	Mg/Na
Mg (sum)	Ca/Na
Ca (sum)	Ca/(P + Mg)
Na (sum)	Mg/(Ca + P)
P/C	P/(Ca + Mg)
P/Mg	Ca/(C + Na)
P/Ca	Mg/(C + Na)
P/Na	P/(C + Na)
C/Mg	(Ca + P + Mg)/C
C/Ca	(Ca + P + Mg)/Na
C/Na	(Ca + P + Mg)/(C + Na)

Table 3
The 80 independent variables used in ratio model two (RM2).

P213.618 (p1)[a]	p1/na1	p4/c	mg2/na2
P214.914 (p2)[a]	p1/na2	p4/mg1	mg3/c
P255.326 (p3)[a]	p2/c	p4/mg2	mg3/ca1
P253.560 (p4)[a]	p2/mg1	p4/mg3	mg3/ca2
C247.856 (c)[a]	p2/mg2	p4/ca1	mg3/ca3
Mg279.553 (mg1)[a]	p2/mg3	p4/ca2	mg3/na1
Mg280.271 (mg2)[a]	p2/ca1	p4/ca3	mg3/na2
Mg285.213 (mg3)[a]	p2/ca2	p4/na1	ca1/c
Ca393.361 (ca1)[a]	p2/ca3	p4/na2	ca1/na1
Ca396.837 (ca2)[a]	p2/na1	mg1/c	ca1/na2
Ca422.666 (ca3)[a]	p2/na2	mg1/ca1	ca2/c
Na588.995 (na1)[a]	p3/c	mg1/ca2	ca2/na1
Na589.593 (na2)[a]	p3/mg1	mg1/ca3	ca2/na2
p1/c	p3/mg2	mg1/na1	ca3/c
p1/mg1	p3/mg3	mg1/na2	ca3/na1
p1/mg2	p3/ca1	mg2/c	ca3/na2
p1/mg3	p3/ca2	mg2/ca1	c/na1
p1/ca1	p3/ca3	mg2/ca2	c/na2
p1/ca2	p3/na1	mg2/ca3	mg3/mg1
p1/ca3	p3/na2	mg2/na1	mg3/mg2

[a] Indicates a line used in the "lines" model.

The remaining nineteen variables were composed of ratios of these sums (ten independent variables) and also unique combinations of the summed intensities forming complex ratios (nine independent variables). This approach has been utilized with success by Gottfried et al. to discriminate LIBS spectra obtained from explosive residues [14,20].

RM2 consisted of 80 independent variables, shown in Table 3. The first thirteen variables were merely the intensities of the thirteen strong emission lines used in the lines model (indicated by an asterisk). These variables are identified by their element symbol and their wavelength in nanometers, as well as a shorthand identifier in parentheses. The remaining 67 variables were simple ratios of these thirteen intensities. Although complex ratios of these variables can be constructed as was done in RM1, this quickly raised the total number of independent variables in the model to such a large number that it was deemed not practical both for computational reasons and to avoid over-determining the data. It was decided that when the dimensionality of the original data was not reduced significantly then the benefits of performing a down-selection were reduced and the more appropriate model would be to use the entire spectrum. This was not done by us due to the size of the spectrum (>54,000 channels) and the presence of spectral "gaps" in the spectrum due to optical design constraints within the Échelle spectrometer. Only down-selected models were investigated.

2.3. Chemometric analysis techniques

Two multivariate chemometric analysis techniques were compared for discrimination between different bacterial genera based on the LIBS emission spectra. The two techniques compared in this study were a discriminant function analysis (DFA) performed with SPSS v.19 (IBM, Inc.) and a partial least squares discriminant analysis (PLS-DA) performed with the PLS_toolbox v6.7.1 running under Matlab v7.6 (Eigenvector Research, Inc.). These two analysis techniques were compared using the down selected variables in RM2.

DFA is a multivariate analysis technique that uses independent variables (atomic emission intensities) to calculate a dependant variable (bacterial identity) to classify or discriminate between two or more groups [21]. The independent variables (contained in the model) are used to construct a set of discriminant functions which maximize the variance between known data sets in a library. These discriminant functions are then used to calculate discriminant function scores which determine the identity of an unknown spectrum. In our DFA comparison, the library was composed of five genera of bacteria, as shown in column one of Table 1.

In each test of the DFA all the spectra in each of the 32 data sets (typically 20–30 spectra per data set) were withheld and classified one-by-one by a DFA library composed of the other 31 data sets. Therefore 32 separate tests were performed. This is known as external validation, because each spectrum was tested against a library where no other spectra acquired at the same time or under the same conditions were present. In comparison, a cross-validated test only removes one spectrum at a time from the library and will most likely return overly-optimistic results. Because only one data set existed for *E. cloacae* ATCC 13047, this data set could not be withheld for external testing, but the genus remained in the analysis to provide a possible "false positive" result for similar bacteria. Thus each spectrum, with no similar spectra in the training library, was classified as belonging to either genus *Escherichia*, *Enterobacter*, *Staphylococcus*, *Streptococcus*, or *Mycobacterium* in a series of 31 separate tests of the library. There is no "null test" in this analysis, as every unknown spectrum must be assigned to one of those five groups.

PLS-DA is a multivariate technique that finds the maximum variance between two groups. PLS-DA takes a set of independent variables as determined by our models and constructs latent variables to maximize the variance between the two groups. The latent variables are predictor variables which are used to classify each spectrum. The PLS-DA then calculates a discrimination line (or this can be user-determined) to predict the class of each spectrum based on Bayesian statistics by minimizing the number of false positives and negatives [22]. In all of our results, the Bayesian-determined discrimination line was utilized for spectral classification. The identity of unknown spectra was then predicted based on this discrimination line in the pre-compiled library. It is essentially a yes or no test where one genus was grouped as the "yes group" and the remaining four genera were grouped together as a "no group." For example, we could utilize this PLS-DA to determine if an unknown spectrum belonged to genus *Staphylococcus* or not. If it was classified as "no," the PLS-DA did not tell us which of the other four genera it most closely resembled. This analysis therefore allowed for a null test. All unknown samples were classified in a PLS-DA test specific for each genus, and if the test group was classified as belonging to the "no group" for each model, it remained unknown and was not classified as belonging to any genus. In this test of the PLS-DA, every spectrum in the 31 data sets (again excluding *E. cloacae*) was tested in five different PLS-DA models, one for each genus. Because each of the 31 data sets was withheld from the library in turn, this resulted in 155 separate tests being performed. No preprocessing was used on the lines or ratio models in the PLS-DA since the variables had already been down-selected from the whole spectrum model.

3. Results and discussion

3.1. Model comparison: lines, RM1, and RM2

The DFA technique was used to compare the three independent variable models described in Section 2.2. The accuracy of classification was reported in the form of truth tables which provide true positive and negative results, as well as false positives and negatives. As mentioned earlier, since there was only one set of *Enterobacter* data no external validation could be performed so there are no truth tables for this genus. Results were tabulated for every spectrum, then totaled for each genus. The truth tables for the three models are shown in Table 4.

In each of the DFA results, four discriminant functions (DF1 through DF4) were constructed to determine the classification of each spectrum. When using the lines model DF1 accounted for approximately 74% of the variance among the data as determined by averaging over the 31 tests. DF2 accounted for 20% of the variance in the data on average, while DF3 and DF4 played a less-important role (accounting for less than 6% of the combined variance). In these analyses the independent variables C, Mg279, and Mg280 played important roles in the construction of both DF1 and DF2 as revealed by their structure matrix scores, while all four P lines accounted for much less of the variance.

When using RM1, DF1 captured less of the variance of the data than in the lines model accounting for 71% of the variance. DF2 accounted for 19% of the variance in the data while DF3 and DF4 played a more important role in discriminating between genera (approximately 10% of the total variance in the data). When using RM1, the independent variables containing ratios with phosphorus played a much larger role in the construction of DF1. P/(C + Na) and P/C were the variables contributing most significantly to the construction of DF1 as determined by the structure matrix. Since Na plays little to no role in bacterial discrimination (often being a residue from the nutrition medium) these two variables are highly similar and in the future it may be possible to eliminate complex ratios containing Na such as P/(C + Na). Calcium ratios such as Ca/(C + Na) were significant in the construction of DF1 and DF2. Truth table results for the RM1 model are shown in Table 4.

When using RM2, DF1 on average accounted for approximately 68% of the variance of the data, DF2 accounted for 18%, DF3 for 9%, and DF4 for 5% of the variance of the data. As expected, when a greater number of independent variables were used, the DFA was able to construct more effective discriminant functions (less of the variance accounted for by just one function). DF3 and DF4 played a larger role in discriminating between the classes (14% of the variance), when using RM2 than the other models, but still constituted a

Table 4
Truth table results for three independent variable models utilized in a genus-level discriminant function analysis of bacterial LIBS spectra.

Lines model			Ratio model 1			Ratio model 2		
Escherichia	True	False	*Escherichia*	True	False	*Escherichia*	True	False
Positive	89.97%	4.28%	Positive	96.32%	7.95%	Positive	95.65%	9.17%
Negative	95.72%	10.03%	Negative	92.05%	3.68%	Negative	90.83%	4.35%
Staphylococcus	True	False	*Staphylococcus*	True	False	*Staphylococcus*	True	False
Positive	62.16%	2.55%	Positive	51.35%	1.70%	Positive	54.05%	0.51%
Negative	97.45%	37.84%	Negative	98.30%	48.65%	Negative	99.49%	45.95%
Streptococcus	True	False	*Streptococcus*	True	False	*Streptococcus*	True	False
Positive	83.82%	2.24%	Positive	88.24%	0.41%	Positive	95.59%	1.02%
Negative	97.76%	16.18%	Negative	99.59%	11.76%	Negative	98.98%	4.41%
Mycobacterium	True	False	*Mycobacterium*	True	False	*Mycobacterium*	True	False
Positive	89.61%	1.27%	Positive	89.61%	1.06%	Positive	88.31%	1.06%
Negative	98.73%	10.39%	Negative	98.94%	10.39%	Negative	98.94%	11.69%

relatively small fraction of the total variance. The independent variables Ca2/C, Ca1/C, and Ca3/C played the largest role in constructing DF1 to discriminate between genera, with a large structure matrix value for all 31 tests. P played a much smaller role in the construction of the functions and many of the P lines and ratios had low correlations with DF1–DF3. A graphical representation of the first two discriminant function scores of all the spectra in an external-validation DFA performed on data set 32 (*M. smegmatis* strain TA) is shown in Fig. 2. The "unknown" bacterial spectra are represented by the "*x*" symbols and 34 of 34 unknown spectra were correctly classified as *Mycobacterium*, even though the model contained no other spectra from strain TA. Truth table results for RM2 are shown in Table 4.

3.2. Chemometric technique comparison: DFA vs. PLS-DA

Based on its performance in the DFA model comparison tests, RM2 was used in a comparison of the two analysis techniques of PLS-DA and DFA. Utilizing RM2, the PLS-DA was performed as described in Section 2.3 and a truth table of the results is shown in Table 5 (with the DFA truth tables for RM2 repeated for ease of comparison). A graphical representation of the external-validation PLS-DA performed on data set 32 (*M. smegmatis* strain TA) is shown in Fig. 3. Again, the "unknown" bacterial spectra are represented by the "*x*" symbols. In Fig. 3(a) 34 of 34 unknown spectra were correctly classified as *Mycobacterium* in a "*Mycobacterium*" test where all other data sets were grouped as "non-*Mycobacterium*." In Fig. 3(b) the same 34 spectra were tested in a "*Streptococcus*" test and 34 of 34 were correctly identified as not belonging to genus *Streptococcus* (a true negative). The 34 spectra were tested against the other genera as well (not shown). In all cases the discrimination line was chosen by the PLS_toolbox to minimize the number of false positives and negatives in the library (model), as mentioned earlier. The sensitivity and specificity of each method were calculated and are given on the bottom of Table 5. Sensitivity equals the number of true positives divided by the total number of true positives and false negatives times 100% and specificity equals the number of true negatives divided by the total number of true negatives and false positives times 100%.

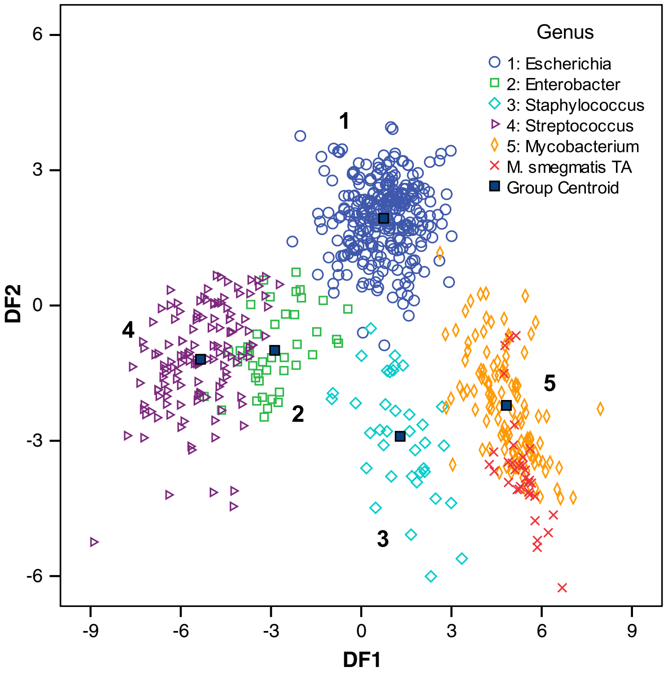

Fig. 2. The first two discriminant function scores of all the spectra in an external-validation DFA utilizing ratio model two (RM2) performed on data set 32 (*M. smegmatis* strain TA). The "unknown" bacterial spectra are represented by the "*x*" symbols and 34 of 34 unknown spectra were correctly classified as belonging to genus *Mycobacterium*, even though the model contained no other spectra from strain TA.

Table 5
Truth table results for two multivariate techniques (DFA and PLS-DA) utilized in a genus-level classification of bacterial LIBS spectra.

DFA: RM2			PLS-DA: RM2		
Escherichia	True	False	*Escherichia*	True	False
Positive	95.65%	9.17%	Positive	89.63%	15.95%
Negative	90.83%	4.35%	Negative	84.05%	10.37%
Staphylococcus	True	False	*Staphylococcus*	True	False
Positive	54.05%	0.51%	Positive	86.49%	5.85%
Negative	99.49%	45.95%	Negative	94.15%	13.51%
Streptococcus	True	False	*Streptococcus*	True	False
Positive	95.59%	1.02%	Positive	99.26%	13.32%
Negative	98.98%	4.41%	Negative	88.68%	0.74%
Mycobacterium	True	False	*Mycobacterium*	True	False
Positive	88.31%	1.06%	Positive	96.10%	4.08%
Negative	98.94%	11.69%	Negative	95.92%	3.90%
Sensitivity	91.4 ± 16.4%		Sensitivity	93.1 ± 10.3%	
Specificity	97.5 ± 9.4%		Specificity	90.6 ± 21.3%	

The 80 independent variables used in RM2 were used in the PLS-DA. These 80 down-selected independent variables were further reduced to 20 latent variables (LVs). An investigation of the PLS-DA was conducted to compare the number of LVs and the corresponding rates of true positives and true negatives. Using a leave-one-out analysis performed by the PLS_toolbox, the PLS-DA chose the number of latent variables to be consistently 4 or 5 for all the tests. Using various data sets of *Mycobacterium* and *Escherichia* the latent variables were then manually set from 0 to 20 and the number of true positives and true negatives respectively were observed and plotted as a function of the number of LVs. Fig. 4 shows the rates of true positives as a function of the number of LVs for data sets 26, 28, and 32 (*M. smegmatis* strain WT — 90% dilution, *M. smegmatis* strain WT — 50% dilution, and *M. smegmatis* strain TA). Data set 26 showed that true positives increased up to 14 LVs, data set 28 showed increased true positives up to16 LVs, and data set 32 showed increased true positives to only 3 LVs. Similar results were seen for other data sets and the true positives and true negatives were maximized for all data sets when at least 20 LVs were used. For each test run thereafter the number of LVs was forced to 20 in the PLS-DA. Ongoing research is being conducted to further maximize the number of latent variables while considering the root mean squared error of calibration.

4. Discussion

A comparison of the DFA performed with the three different models consisting of lines, RM1, and RM2 showed that RM2 yielded the overall highest true positive and true negative rates with true positive rates of 95%, 54%, 95%, and 88% for the four genera and true negative rates of 91%, 99%, 99%, and 99%. Overall the sensitivity was 91.4 ± 16.4% and the specificity was 97.5 ± 9.4%. The sensitivity and specificity were obtained by averaging the results from the 31 tests and the standard deviation is reported as the uncertainty. RM1 performed similarly, but slightly worse than RM2, with RM2 offering a noted improvement in the performance of the *Staphylococcus* and *Streptococcus* tests. In comparison, the lines model performed worst with true positive rates of 90%, 62%, 83%, and 83% for the four genera and true negative rates of 96%, 97%, 98%, and 98%. Although many of these true positive rates are similar, it can be seen that the rates of false positives and false negatives were reduced substantially by the use of RM2. Having 80 independent variables allowed for more variance of the data to be expressed resulting in a better statistical classification of the unknown bacterial spectra. It should be mentioned that prior knowledge of which elemental lines contributed most significantly to accurate classification when using the lines model allowed

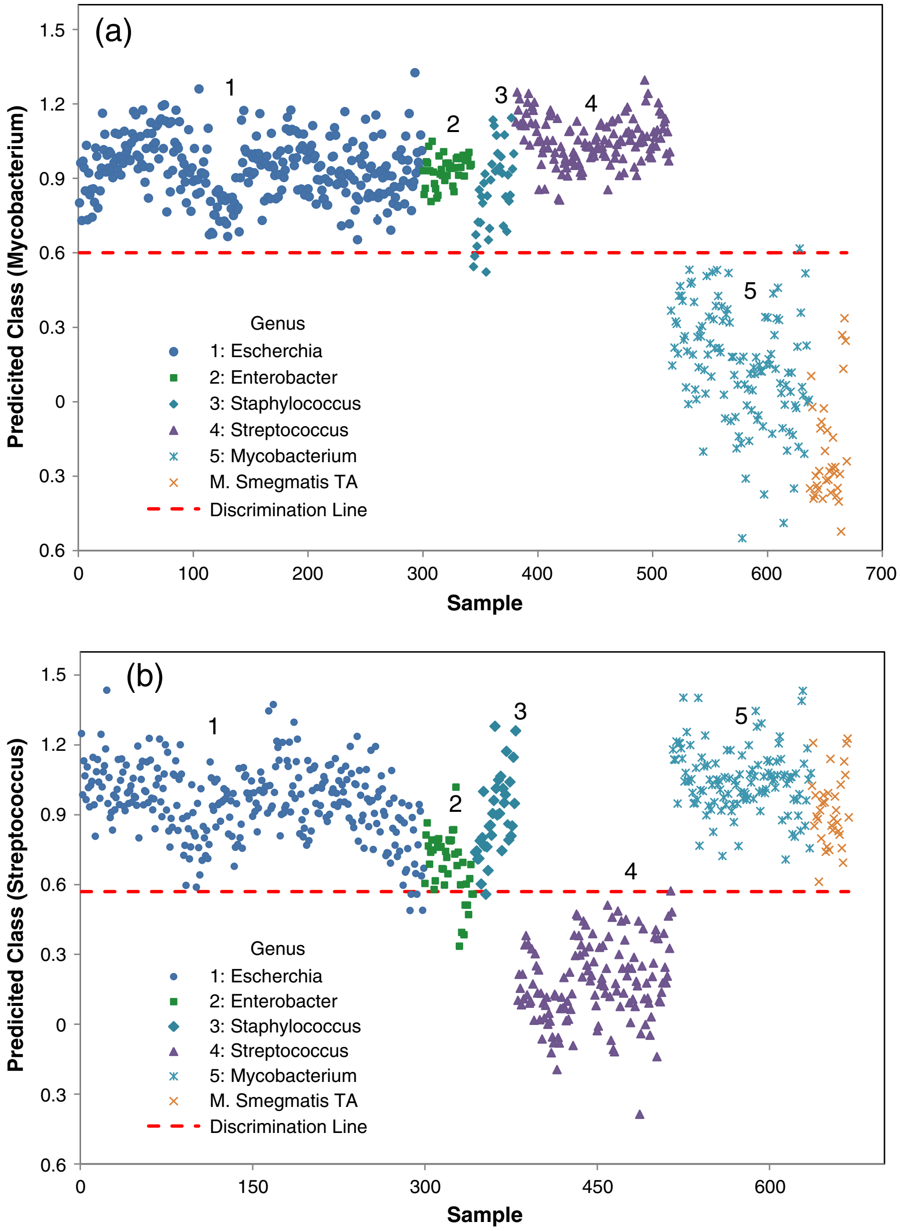

Fig. 3. A graphical representation of the external-validation PLS-DA performed on data set 32 (*M. smegmatis* strain TA). The "unknown" bacterial spectra are represented by the "*x*" symbols. (a) 34 of 34 unknown spectra were correctly classified as *Mycobacterium* (true positives) in a "*Mycobacterium*" test where all other data sets were grouped as "non-*Mycobacterium*". (b) 34 of 34 unknown spectra were correctly classified as not belonging to genus *Streptococcus* (true negatives) in a "*Streptococcus*" test where all other data sets were grouped as "non-*Streptococcus*".

the construction of appropriate ratios in RM2 which then resulted in the improved classification demonstrated by RM2.

In the DFA tests it was shown that a DFA was able to effectively classify a sample between five different genera. Lower sensitivity was seen with *Staphylococci* data sets, but this is not indicative of any issues related specifically to *Staphylococci* or to the multivariate techniques. This was merely a result of there being only two representative *Staphylococci* data sets to include in the analysis, as can be seen in Table 1, with one of these data sets being among the earliest experiments performed in the construction of the spectral library. It is believed that the addition of newer and more varied *Staphylococci* spectra will increase the sensitivity and specificity of this genus to values seen in other genera. When the DFA was given an unknown bacterial spectrum using any of the 31 libraries tested it was able to classify the bacteria as one of the five classes with high sensitivity, whereas our PLS-DA was effective in determining if the unknown spectrum belonged to a specific class or not. If information is needed about whether an unknown bacterium is or is not a certain class, PLS-DA is the preferred method (i.e. in an online test of beef products searching for spectra consistent with the presence of enterohemorrhagic *E. coli*). If the bacterial type needs to be known from among multiple competing possibilities (i.e. in a clinical diagnostic) DFA is probably the preferred technique, although it must be said that it is possible to efficiently run a number of PLS-DA tests in sequence to arrive at a statistical classification of the unknown spectrum. Therefore both analyses can perform both functions, if necessary. In our classification tests PLS-DA yielded higher sensitivity (93.1%) than the DFA (91.4%) with a smaller uncertainty on this value, but possessed lower specificity (90.6%) than the DFA (97.5%) with a larger uncertainty. Importantly, marked improvement was demonstrated by the PLS-DA with the problematic *Staphylococci* data sets. PLS-DA was able to identify more bacteria correctly, possessing a higher true positive rate but identified more bacteria incorrectly, possessing a higher false positive rate than the DFA.

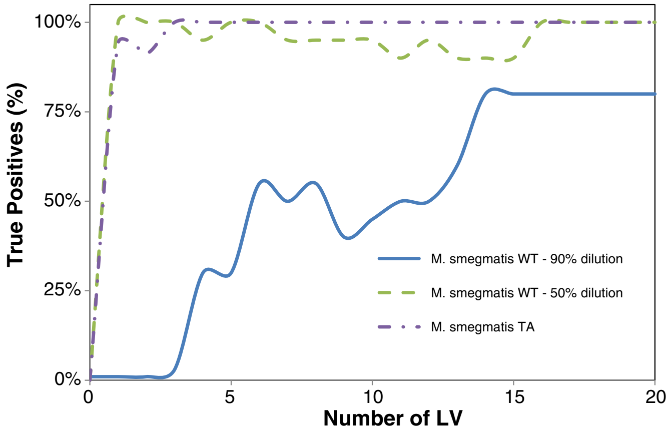

Fig. 4. Percentage of true positives plotted as a function of the number of LVs used by PLS-DA to predict class. The PLS-DA model was constructed using *Mycobacterium* as the "yes group" and the remaining genera as the "no group." Three representative data sets of *Mycobacterium* were tested for true positives (*M. smegmatis* strain TA, *M. smegmatis* strain WT — 50% dilution, and *M. smegmatis* strain WT — 90% dilution). Rates of true positives increased as the number of LVs increased until approximately 20.

PLS-DA seems to be more effective at distinguishing bacteria from similar genera. For example, *M. smegmatis* and *E. coli* are similar in composition and were identified incorrectly as each other more commonly in the DFA than in the PLS-DA. PLS-DA was able to statistically find the variance between LIBS spectra from similar bacteria and reliably discriminate them. It may therefore be true that a DFA is more effective in genus-level discrimination on bacterial specimens with a wide range of potential identities, but discrimination at the species- or strain-level once the genus is accurately identified may require the use of PLS-DA. Work is ongoing to investigate this possibility.

5. Conclusion

We have shown that a sensitive and specific genus level classification of LIBS spectra from live bacterial specimens can be performed with a DFA or a PLS-DA using several different independent variable models. The three models constructed from down-selected independent variables possessed similar sensitivities and specificities when utilized in a genus-level five-class DFA, but the model consisting of 80 independent variables constructed from the normalized emission intensities of thirteen lines of P, Ca, Mg, Na, C, and complex ratios of those intensities performed best. It possessed a sensitivity of 91.4% and a specificity of 97.5%. All results were obtained using external-validation tests. When this model was utilized in a PLS-DA, it possessed a sensitivity of 93.1% and a specificity of 90.6%. The number of latent variables required for efficient classification using this model was investigated, and chosen to be 20 in all subsequent tests.

It is apparent that both multivariate techniques provide effective classification of unknown bacterial LIBS spectra. From the performance in this five genus classification, it is possible that DFA may be an appropriate technique to use when the identity of a specimen is completely unknown and genus-level discrimination is required. More precise identification at the species-level or strain-level may be subsequently performed with a PLS-DA, which demonstrated improved performance at discriminating highly similar spectra. Ultimately, the sensitivity and specificity of the two techniques were similar in this investigation, although they classify based on fundamentally different mathematical principles. Because the same spectral library was efficacious in both techniques, it is possible that both analyses could be performed simultaneously on an unknown sample to provide an independent verification of specimen identity. It is likely that computational processing power would easily allow such a verification, as the classification of one unknown spectrum against a pre-compiled library model is performed rapidly by both techniques. Such a confirmation will need to be investigated in future work.

Acknowledgements

Funding for this project was provided in part by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN/418254-2012). RAP and AD were supported by the University of Windsor's Outstanding Scholars program.

References

[1] A. Assion, M. Wollenhaupt, L. Haag, F. Mayorov, C. Sarpe-Tudoran, M. Winter, U. Kutschera, T. Baumert, Femtosecond laser-induced breakdown spectrometry for Ca^{2+} analysis of samples with high spatial resolution, Appl. Phys. B. 77 (2003) 391–397.

[2] J.D. Hybl, G.A. Lithgow, S.G. Buckley, Laser-induced breakdown spectroscopy detection and classification of biological aerosols, Appl. Spectrosc. 57 (2003) 1207–1215.

[3] S. Morel, M. Leone, P. Adam, J. Amouroux, Detection of bacteria by time-resolved laser-induced breakdown spectroscopy, Appl. Opt. 42 (2003) 6184–6191.

[4] A.C. Samuels, F.C. DeLucia Jr., K.L. McNesby, A.W. Miziolek, Laser-induced breakdown spectroscopy of bacterial spores, molds, pollens, and protein: initial studies of discrimination potential, Appl. Opt. 42 (2003) 6205–6209.

[5] S.J. Rehse, Q.I. Mohaidat, S. Palchaudhuri, Towards the clinical application of laser-induced breakdown spectroscopy for rapid pathogen diagnosis: the effect of mixed cultures and sample dilution on bacterial identification, Appl. Opt. 49 (2010) C27–C35.

[6] Q. Mohaidat, S. Palchaudhuri, S.J. Rehse, The effect of bacterial environmental and metabolic stresses on a LIBS-based identification of *Escherichia coli* and *Streptococcus viridans*, Appl. Spectrosc. 65 (2011) 386–392.

[7] C. Barnett, C. Bell, K. Vig, A.C. Akpovo, L. Johnson, S. Pillai, S. Singh, Development of a LIBS assay for the detection of *Salmonella enterica* serovar Typhimurium from food, Anal. Bioanal. Chem. 400 (2011) 3323–3330.

[8] Q.I. Mohaidat, K. Sheikh, S. Palchaudhuri, S.J. Rehse, Pathogen identification with laser-induced breakdown spectroscopy: the effect of bacterial and biofluid specimen contamination, Appl. Opt. 51 (2012) B99–B107.

[9] S.J. Rehse, N. Jeyasingham, J. Diedrich, S. Palchaudhuri, A membrane basis for bacterial identification and discrimination using laser-induced breakdown spectroscopy, J. Appl. Phys. 105 (2009) 102034.

[10] F.C. De Lucia Jr., J.L. Gottfried, C.A. Munson, A.W. Miziolek, Multivariate analysis of standoff laser-induced breakdown spectroscopy spectra for classification of explosive-containing residues, Appl. Opt. 47 (2008) G112–G121.

[11] J.L. Gottfried, F.C. De Lucia Jr., A.W. Miziolek, Discrimination of explosive residues on organic and inorganic substrates using laser-induced breakdown spectroscopy, J. Anal. Atom. Spectrom. 24 (2009) 249–356.

[12] F.C. De Lucia Jr., J.L. Gottfried, Influence of variable selection on partial least squares discriminant analysis models for explosive residue classification, Spectrochim. Acta Part B 66 (2011) 122–128.

[13] D.W. Merdes, J.M. Suhan, J.M. Keay, D.M. Hadka, W.R. Bradley, The investigation of laser-induced breakdown spectroscopy for detection of biological contaminants on surfaces, Spectroscopy 22 (2007) 28–38.

[14] J.L. Gottfried, F.C. De Lucia Jr., C.A. Munson, A.W. Miziolek, Standoff detection of chemical and biological threats using laser-induced breakdown spectroscopy, Appl. Spectrosc. 62 (2008) 353–363.

[15] D. Marcos-Martinez, J.A. Ayala, R.C. Izquierdo-Hornillos, F.J. Manuel de Villena, J.O. Caceres, Identification and discrimination of bacterial strains by laser induced breakdown spectroscopy and neural networks, Talanta 84 (2011) 730–737.

[16] D.E. Lewis, J. Martinez, C.A. Akpovo, L. Johnson, A. Chauhan, M.D. Edington, Discrimination of bacteria from Jamaican bauxite soils using laser-induced breakdown spectroscopy, Anal. Bioanal. Chem. 401 (2011) 2225–2236.

[17] R. Multari, D.A. Cremers, M.L. Bostian, Use of laser-induced breakdown spectroscopy for the differentiation of pathogens and viruses on substrates, Appl. Opt. 51 (2012) B57–B64.

[18] J. Cisewski, E. Snyder, J. Hannig, L. Oudejans, Support vector machine classification of suspect powders using laser-induced breakdown spectroscopy (LIBS) spectral data, J. Chemometrics 26 (2012) 143–149.

[19] S.J. Rehse, J. Diedrich, S. Palchaudhuri, Identification and discrimination of *Pseudomonas aeruginosa* bacteria grown in blood and bile by laser-induced breakdown spectroscopy, Spectrochim. Acta Part B 62 (2007) 1169–1176.

[20] J.L. Gottfried, F.C. DeLucia Jr., C.A. Munson, A.W. Miziolek, Laser-induced breakdown spectroscopy for detection of explosives residues: a review of recent advances, challenges, and future prospects, Anal. Bioanal. Chem. 395 (2009) 283–300.

[21] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate Data Analysis, 7th edition Prentice Hall, 2009.

[22] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemometrics 17 (2003) 166–173.