Chemometric Data Analysis Strategies for Optimizing Pathogen Discrimination and Classification Using Laser-Induced Breakdown Spectroscopy (LIBS) Emission Spectra

Khadija Sheikh, Andrew Daabous, Russell Putnam, and Steven J. Rehse

Department of Physics, University of Windsor, Windsor, Ontario, CA

University of Windsor *thinking forward*

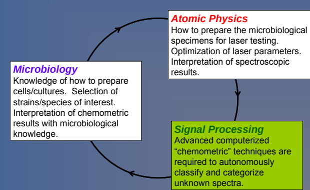
What we do

We use a laser-based optical emission spectroscopy technique know as "***laser-induced breakdown spectroscopy***" (**LIBS**) to rapidly identify pathogenic bacteria.

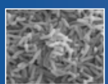
Our aim is to develop a real-time point-of-care medical diagnostic technology for the improvement of human health and safety.

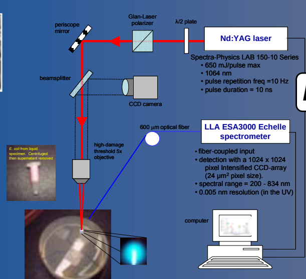
By allowing the in vitro determination of bacterial cell elemental composition and uptake, we are also developing a new microbiological tool for the 21st Century.

An interdisciplinary student experience



How we do it





Goal of this project

Because of the similarity in bacterial LIBS spectra, computerized chemometric signal-processing algorithms must be used to distinguish LIBS spectra from different bacteria.

(1) We investigated the use of two different chemometric techniques: *discriminant function analysis* (DFA) and *partial least squares-discriminant analysis* (PLS-DA) to quantify their sensitivity and specificity.

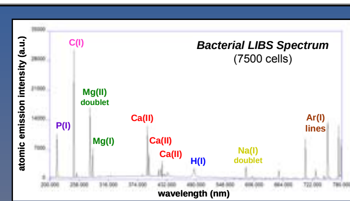
(2) We investigated three different models for converting LIBS spectral data into the independent variables used by the chemometric algorithms.

SPECTRAL LIBRARY

A library of spectral data sets for five genera has been constructed (see table at right) from over 32 separate LIBS experiments spanning 3 years.

Within each genus, data sets are further classified into 13 bacterial identifications by species or strain.

Genus	Species/Strain ID	Data set
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
Enterobacter	6: *E. cloacae* ATCC 13047	16: *E. cloacae* ATCC 13047
Staphylococcus	7: *S. saprophyticus*	17: *S. saprophyticus*
	8: *S. aureus*	18: *S. aureus*
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
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% conc.
	12: *M. smegmatis* TE	31: *M. smegmatis* TE
	13: *M. smegmatis* TA	32: *M. smegmatis* TA



In the LIBS spectrum at left there are 13 strong emission lines which are used as data: four phosphorus lines, one carbon line, two magnesium II (Mg^{2+}), one magnesium line, three calcium II (Ca^{2+}) lines, and two sodium (Na) lines. Argon, nitrogen, hydrogen, and oxygen are NOT used.

These 13 emission intensities are combined in different ways in the three "models" developed below.

Line ID	P I	P I	P I	P I	C	Mg II	Mg II	Mg I	Ca II	Ca II	Ca II	Na I	Na I
Wavelength (nm)	213.61	214.91	253.56	255.32	247.85	279.55	280.26	285.21	393.36	396.84	422.67	588.99	589.59

Discriminant Function Analysis vs. Partial Least Squares-Discriminant Analysis

DFA is a statistical analysis used to predict a categorical dependent variable by one or more continuous or binary independent variables. It uses the spectral data with their group assignments and discriminates them by maximizing between-class variance. It does so by constructing a unique set of discriminant functions. The program used for the DFA analysis was SPSS v19 (IBM, Inc.).

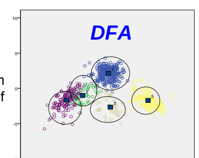
PLSDA is a multivariate inverse least squares discrimination method used to classify samples. PLSDA is used to find the maximum variance or separation between classes and not the variance of the data set. This leads to the conclusion that PLSDA is designed to work best with a "yes or no" test. The PLSDA software used was a MATLAB toolbox called PLS_toolbox (Eigenvector Research, Inc.)

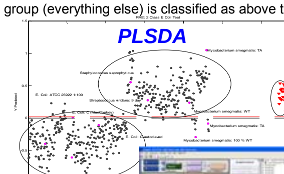
What is external validation?

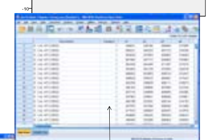
External validation is performed when the test data is kept separate from the model so that it is completely unknown to the model. This is the only true test of a model, to see how it will classify samples that it has never seen before. All numbers below are from external validation tests.

A 5 class genus-level DFA was performed on Ratio Model 2 data and compared with a 2 class PLSDA "yes or no" test. The external tests left one group out of the model and tested that group against the model for classification. Using PLSDA, five tests were run where each genus was classified as group 1 and all other genera classified as group 2. DFA was used to classify all groups with one model where PLSDA uses one model per genus.

Shown is an example of output from each program as well as the input user interface. The DFA graph shows how each class is grouped by multiple functions (x and y axes are functions 1 and 2) and an unknown spectrum is classified by where it lies with respect to the center of each group. The PLSDA graph displays how the "yes" group is classified by being below the line and the "no" group (everything else) is classified as above the line.







SPSS

PLS Toolbox

Conclusions

- Ratio model 1 and the lines model yielded similar results. Ratio model 2 resulted in improvement in classification.
- DFA techniques work well when all samples are known and exceeds PLSDA if groups to be discriminated are very similar.
- PLSDA techniques work well when new samples are introduced into the test set (not in the library) and when a yes or no result is desired.
- PLSDA showed increased sensitivity but decreased specificity compared to DFA.

Conveying Results: Truth Tables

In *external validation tests*, truth tables convey the results of the classification of an unknown "blind" sample against the model once the "true" identity is revealed. Say you are trying to identify the MRSA bacteria. Four results of the test are possible:

Positive/Negative: indicates whether the sample was classified as the sample in question (positive = "yes, it is MRSA") or as something else (negative = "no, it is not MRSA")
True/False: indicates whether the identification was correct (true = "you said it was MRSA, and it actually was") or incorrect (false = "you said it was MRSA and it was not")

Values in the truth tables indicate the percentage of test results which returned the given result. Desired is 100% true positives (sensitivity) and 100% true negatives (specificity)

Results: DFA on 3 Models

LINES MODEL

This model (the basis of all our previous work) used the intensity of each of the 13 observed emission lines in the LIBS spectrum as independent variables.

E. COLI	True	False
Positive	89.97%	4.28%
Negative	95.72%	10.03%
STAPHYLOCOCCUS	**True**	**False**
Positive	62.16%	2.55%
Negative	97.45%	37.84%
STREPTOCOCCUS	**True**	**False**
Positive	83.82%	2.74%
Negative	97.76%	16.18%
MYCOBACTERIUM	**True**	**False**
Positive	83.82%	2.74%
Negative	97.76%	16.18%

RATIO MODEL 1

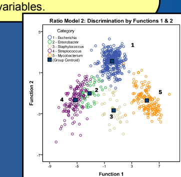
This model first summed the intensities of all lines from a given element (creating values for C, P, Mg, Ca, Na). Each of these five groups was used as an independent variable as were ratios of the elemental sums and ratios of combinations of those sums. Total of 24 independent variables.

E. COLI	True	False
Positive	96.32%	7.95%
Negative	92.05%	3.68%
STAPHYLOCOCCUS	**True**	**False**
Positive	51.35%	1.70%
Negative	98.30%	48.65%
STREPTOCOCCUS	**True**	**False**
Positive	88.24%	0.41%
Negative	99.59%	11.76%
MYCOBACTERIUM	**True**	**False**
Positive	89.61%	1.06%
Negative	98.94%	10.39%

RATIO MODEL 2

Intensities of the emission lines comprised the first 13 variables. The remaining independent variables were obtained by taking ratios of the emission line intensities and ratios of various sums of intensities. No reciprocals included. Total of 80 independent variables.

E. COLI	True	False
Positive	95.65%	9.17%
Negative	90.83%	4.35%
STAPHYLOCOCCUS	**True**	**False**
Positive	54.05%	0.51%
Negative	99.49%	45.95%
STREPTOCOCCUS	**True**	**False**
Positive	95.59%	1.02%
Negative	98.98%	4.41%
MYCOBACTERIUM	**True**	**False**
Positive	88.31%	1.06%
Negative	98.94%	11.69%



Results: DFA vs. PLSDA

Ratio model 2 was used to compare the two analysis techniques (DFA and PLSDA). Listed below are the truth tables from the DFA 5 Class and PLSDA 2 Class "yes or no" test.

DFA 5 Class:

E. COLI	True	False
Positive	95.65%	9.17%
Negative	90.83%	4.35%
STAPHYLOCOCCUS	**True**	**False**
Positive	54.05%	0.51%
Negative	99.49%	45.95%
STREPTOCOCCUS	**True**	**False**
Positive	95.59%	1.02%
Negative	98.98%	4.41%
MYCOBACTERIUM	**True**	**False**
Positive	88.31%	1.06%
Negative	98.94%	11.69%

PLSDA 2 Class:

E. COLI	True	False
Positive	89.63%	15.95%
Negative	84.05%	10.37%
STAPHYLOCOCCUS	**True**	**False**
Positive	86.49%	5.85%
Negative	94.15%	13.51%
STREPTOCOCCUS	**True**	**False**
Positive	99.26%	13.32%
Negative	88.68%	0.74%
MYCOBACTERIUM	**True**	**False**
Positive	96.10%	4.08%
Negative	95.92%	3.90%

DFA: *Sensitivity*: 91.37 ± 16.39 % *Specificity*: 97.46 ± 9.35 %
PLSDA: *Sensitivity*: 93.13 ± 10.25 % *Specificity*: 90.60 ± 21.33 %