Biomedical Optics EXPRESS

Sensitive and specific discrimination of pathogenic and nonpathogenic *Escherichia coli* using Raman spectroscopy—a comparison of two multivariate analysis techniques

Khozima Hamasha,[1,4] Qassem I. Mohaidat,[1,5] Russell A. Putnam,[2] Ryan C. Woodman,[2] Sunil Palchaudhuri,[3] and Steven J. Rehse[2,*]

[1]*Department of Physics and Astronomy, Wayne State University, Detroit, MI 48201, USA*
[2]*Department of Physics, University of Windsor, Windsor, Ontario N9B 3P4, Canada*
[3]*Department of Immunology and Microbiology, Wayne State University, Detroit, Michigan 48201, USA*
[4]*Now with Department of Basic Science, Al-Huson University College, Al-Balqa Applied University, Irbid, Jordan*
[5]*Now with Department of Physics, Yarmouk University, Irbid, Jordan*
rehse@uwindsor.ca

Abstract: The determination of bacterial identity at the strain level is still a complex and time-consuming endeavor. In this study, visible wavelength spontaneous Raman spectroscopy has been used for the discrimination of four closely related *Escherichia coli* strains: pathogenic enterohemorrhagic *E. coli* O157:H7 and non-pathogenic *E. coli* C, *E. coli* Hfr K-12, and *E. coli* HF4714. Raman spectra from 600 to 2000 cm^{-1} were analyzed with two multivariate chemometric techniques, principal component-discriminant function analysis and partial least squares-discriminant analysis, to determine optimal parameters for the discrimination of pathogenic *E. coli* from the non-pathogenic strains. Spectral preprocessing techniques such as smoothing with windows of various sizes and differentiation were investigated. The sensitivity and specificity of both techniques was in excess of 95%, determined by external testing of the chemometric models. This study suggests that spontaneous Raman spectroscopy with visible wavelength excitation is potentially useful for the rapid identification and classification of clinically-relevant bacteria at the strain level.

OCIS codes: (300.6450) Spectroscopy, Raman; (170.0170) Medical optics and biotechnology; (170.1420) Biology; (170.1580) Chemometrics; (170.5660) Raman spectroscopy; (170.1530) Cell analysis.

References and links

1. M. S. Ibelings, K. Maquelin, H. P. Endtz, H. A. Bruining, and G. J. Puppels, "Rapid identification of Candida spp. in peritonitis patients by Raman spectroscopy," Clin. Microbiol. Infect. **11**(5), 353–358 (2005).
2. R. M. Jarvis and R. Goodacre, "Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria," FEMS Microbiol. Lett. **232**(2), 127–132 (2004).
3. E. C. López-Díez and R. Goodacre, "Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics," Anal. Chem. **76**(3), 585–591 (2004).
4. K. Maquelin, L. P. Choo-Smith, T. van Vreeswijk, H. P. Endtz, B. Smith, R. Bennett, H. A. Bruining, and G. J. Puppels, "Raman spectroscopic method for identification of clinically relevant microorganisms growing on solid culture medium," Anal. Chem. **72**(1), 12–19 (2000).
5. K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. A. Ngo-Thi, T. van Vreeswijk, M. Stämmler, H. P. Endtz, H. A. Bruining, D. Naumann, and G. J. Puppels, "Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures," J. Clin. Microbiol. **41**(1), 324–329 (2003).
6. U. Neugebauer, U. Schmid, K. Baumann, U. Holzgrabe, W. Ziebuhr, S. Kozitskaya, W. Kiefer, M. Schmitt, and J. Popp, "Characterization of bacterial growth and the influence of antibiotics by means of UV resonance Raman spectroscopy," Biopolymers **82**(4), 306–311 (2006).
7. D. Pappas, B. W. Smith, and J. D. Winefordner, "Raman spectroscopy in bioanalysis," Talanta **51**(1), 131–144 (2000).

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

8. R. Petry, M. Schmitt, and J. Popp, "Raman spectroscopy—a prospective tool in the life sciences," ChemPhysChem **4**(1), 14–30 (2003).
9. Q. Wu, T. Hamilton, W. H. Nelson, S. Elliott, J. F. Sperry, and M. Wu, "UV Raman spectral intensities of *E. coli* and other bacteria excited at 228.9, 244.0, and 248.2 nm," Anal. Chem. **73**(14), 3432–3440 (2001).
10. H. Yang and J. Irudayaraj, "Rapid detection of foodborne microorganisms on food surface using Fourier transform Raman spectroscopy," J. Mol. Struct. **646**(1-3), 35–43 (2003).
11. R. Goodacre, E. M. Timmins, R. Burton, N. Kaderbhai, A. M. Woodward, D. B. Kell, and P. J. Rooney, "Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks," Microbiology **144**(5), 1157–1170 (1998).
12. G. E. Fox, J. D. Wisotzkey, and P. Jurtshuk, Jr., "How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity," Int. J. Syst. Bacteriol. **42**(1), 166–170 (1992).
13. D. Hutsebaut, J. Vandroemme, J. Heyrman, P. Dawyndt, P. Vandenabeele, L. Moens, and P. de Vos, "Raman microspectroscopy as an identification tool within the phylogenetically homogeneous '*Bacillus subtilis*' group," Syst. Appl. Microbiol. **29**(8), 650–660 (2006).
14. P. Rösch, M. Harz, M. Schmitt, K.-D. Peschke, O. Ronneberger, H. Burkhardt, H. W. Motzkus, M. Lankers, S. Hofer, H. Thiele, and J. Popp, "Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations," Appl. Environ. Microbiol. **71**(3), 1626–1637 (2005).
15. M. Harz, P. Rösch, K.-D. Peschke, O. Ronneberger, H. Burkhardt, and J. Popp, "Micro-Raman spectroscopic identification of bacterial cells of the genus *Staphylococcus* and dependence on their cultivation conditions," Analyst (Lond.) **130**(11), 1543–1550 (2005).
16. M. S. Donnenberg and T. S. Whittam, "Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*," J. Clin. Invest. **107**(5), 539–548 (2001).
17. H. M. Al-Qadiri, M. Lin, A. G. Cavinato, and B. A. Rasco, "Fourier transform infrared spectroscopy, detection and identification of *Escherichia coli* O157:H7 and *Alicyclobacillus* strains in apple juice," Int. J. Food Microbiol. **111**(1), 73–80 (2006).
18. P. S. Mead and P. M. Griffin, "*Escherichia coli* O157:H7," Lancet **352**(9135), 1207–1212 (1998).
19. R. Claender, *The Bacteriophages* (Plenum, 1998).
20. T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.-G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa, "Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12," DNA Res. **8**(1), 11–22 (2001).
21. S. Palchaudhuri, S. J. Rehse, K. Hamasha, T. Syed, E. Kurtovic, E. Kurtovic, and J. Stenger, "Raman spectroscopy of xylitol uptake and metabolism in Gram-positive and Gram-negative bacteria," Appl. Environ. Microbiol. **77**(1), 131–137 (2011).
22. K. Hamasha, M. B. Sahana, C. Jani, S. Nyayapathy, C.-M. Kang, and S. J. Rehse, "The effect of Wag31 phosphorylation on the cells and the cell envelope fraction of wild-type and conditional mutants of *Mycobacterium smegmatis* studied by visible-wavelength Raman spectroscopy," Biochem. Biophys. Res. Commun. **391**(1), 664–668 (2010).
23. I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, 1986).
24. J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis* (Prentice Hall, 2009).
25. M. Barker and W. Rayens, "Partial least squares for discrimination," J. Chemometr. **17**(3), 166–173 (2003).
26. C. Mello, D. Ribeiro, F. Novaes, and R. J. Poppi, "Rapid differentiation among bacteria that cause gastroenteritis by use of low-resolution Raman spectroscopy and PLS discriminant analysis," Anal. Bioanal. Chem. **383**(4), 701–706 (2005).
27. C. Mello, E. Sevéri, E. Ricci, A. Marangoni, L. Coelho, D. Ribeiro, and R. J. Poppi, "Fast differentiation of bacteria causing pharyngitis by low resolution Raman spectroscopy and PLS-discriminant analysis," J. Braz. Chem. Soc. **19**(1), 29–34 (2008).
28. A. Cao, A. K. Pandya, G. K. Serhatkulu, R. E. Weber, H. Dai, J. S. Thakur, V. M. Naik, R. Naik, G. W. Auner, R. Rabah, and D. C. Freeman, "A robust method for automated background subtraction of tissue fluorescence," J. Raman Spectrosc. **38**(9), 1199–1205 (2007).
29. W. E. Huang, R. I. Griffiths, I. P. Thompson, M. J. Bailey, and A. S. Whiteley, "Raman microscopic analysis of single microbial cells," Anal. Chem. **76**(15), 4452–4458 (2004).
30. M. L. Laucks, A. Sengupta, K. Junge, E. J. Davis, and B. D. Swanson, "Comparison of psychro-active arctic marine bacteria and common mesophillic bacteria using surface-enhanced Raman spectroscopy," Appl. Spectrosc. **59**(10), 1222–1228 (2005).
31. Y. Liu, L. He, A. Mustapha, H. Li, Z. Q. Hu, and M. Lin, "Antibacterial activities of zinc oxide nanoparticles against *Escherichia coli* O157:H7," J. Appl. Microbiol. **107**(4), 1193–1201 (2009).
32. J. W. Chan, H. Winhold, M. H. Corzett, J. M. Ulloa, M. Cosman, R. Balhorn, and T. Huser, "Monitoring dynamic protein expression in living *E. coli*. Bacterial cells by laser tweezers Raman spectroscopy," Cytometry A **71A**(7), 468–474 (2007).
33. I. Notingher, "Raman spectroscopy cell-based biosensors," Sensors (Basel Switzerland) **7**(8), 1343–1358 (2007).
34. L. J. Goeller and M. R. Riley, "Discrimination of bacteria and bacteriophages by Raman spectroscopy and surface-enhanced Raman spectroscopy," Appl. Spectrosc. **61**(7), 679–685 (2007).

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

1. Introduction

Spontaneous Raman spectroscopy (RS) utilizing ultraviolet, visible, or near-infrared laser excitation has recently gained popularity as an attractive technique for the biochemical characterization, rapid identification, and accurate classification of a wide range of bacterial species, including *Escherichia coli* [1–10]. Specifically, RS provides a reproducible spectrum which contains rich information about the molecular content of bacterial cells. This is useful for generating a molecular "whole-organism fingerprint" for rapid identification [11]. Due to the diversity, pathogenesis, and evolution of strains that belong to the same species, the detection and identification of bacteria at the strain level is crucial and of great importance for clinical diagnoses, food safety, and water contamination measurements. Sensitive and specific discrimination between very closely related strains of a single species – performed rapidly or at the point of contact – is challenging for classical methods. Even the genetic technique of 16S rDNA sequence analysis does not always exhibit significant variations for closely related bacterial strains [12,13]. However, previous works have demonstrated that RS can be used for a rapid discrimination between different bacterial strains. For example, Rosch et al. found that the recognition rate of bacteria at the strain level by RS was 89.2% for one study [14] and 94.9% for another [15].

The aim of this work was to evaluate the effectiveness of RS combined with suitable chemometric data analysis methods as a sensitive and specific discrimination tool of *Escherichia coli* at the strain level. Specifically, it was desired to discriminate a single pathogenic strain of *E. coli* from multiple non-pathogenic strains. *E. coli* represents a good model of bacteria that has a wide range of pathogenic and non-pathogenic strains where some strains are genetically very closely related and have the capability of generating new pathogens that may cause new disease syndromes [16]. Four closely-related well-characterized strains of *E. coli* were selected for this study. These strains included an avirulent laboratory derivative of the pathogenic strain *E. coli* O157:H7 (also called enterohemorrhagic *E. coli*, or EHEC) that can cause major public health concerns including the hemolytic-uremic syndrome which can be fatal in young children [17,18]. Three non-pathogenic laboratory strains (*E coli* C, *E. coli* Hfr K-12, and *E. coli* HF4714, a hybrid of strains K-12 and C), were also studied. A previous study of the chromosome sequence of Hfr K-12 and O157:H7 revealed that O157:H7 has evolved from Hfr K-12 after the attack of a lysogenic bacteriophage [19,20]. We have previously utilized RS to study the uptake of the 5-carbon polyol xylitol by several of these *E. coli* strains [21] as well as to measure the molecular differences of conditional mutants of *Mycobacterium smegmatis* expressing three different alleles relating to the phosphorylation of Wag31, a key cell-division protein [22].

Many Raman spectra were obtained from multiple colonies of each strain to investigate the reproducibility of the spectra and to provide a large database for this study. Due to the high dimensionality of the data (each Raman spectrum consisted of 2071 channels) classification was performed by using the multivariate analysis methods of principal component analysis (PCA) [4,6,23] which reduced the dimensionality of the data, followed by a discriminant function analysis (DFA) [24] which classified all the PCA-reduced spectra into independent categories depending on similarities and differences in the molecular composition of the bacterial strains. When done sequentially, this is referred to as a PC-DFA [2,3]. This PC-DFA technique was compared to a partial least squares discriminant analysis (PLS-DA) which reduced the dimensionality of the data to 20 latent variables which maximized the variance between the data, followed by a classification of the bacterial spectra as pathogenic or non-pathogenic [25–27]. Such chemometric techniques are now routinely used for rapid and autonomous classification or grouping of bacteria on the basis of their spectral fingerprints [11]. This study was able to demonstrate sensitive and specific *E. coli* strain differentiation using only spontaneous Raman scattering without the need for surface-enhanced (SERS) or coherent (CARS) techniques.

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

2. Materials and method

2.1 Bacterial culture conditions and sample preparation

All *E. coli* samples were prepared in a similar manner. Bacterial cells were cultured overnight in a nutrient broth medium at 37 °C, then 1 μL of the suspension was streaked on a trypticase soy agar TSA plate using a sterilized inoculating loop. The plates were incubated at 37 °C for 24 hours. Single colonies of cells were harvested from the plates using an inoculating loop and suspended in 1.5 ml of deionized water. These aliquots were centrifuged for 3 minutes at 5000 rev/min at room temperature to create a watery pellet. The supernatant and traces of the media were discarded. In all cases, a final bacterial titer of approximately 10^8 cells was utilized, as determined by a measurement of optical density.

Prior to Raman measurements, 10 μL of each of the centrifuged suspensions was transferred to a low-fluorescence quartz microscope slide which was allowed to air-dry at room temperature. After each use the slides were cleaned with deionized water, acetone, and methanol to remove any organic contamination or residue from the tested bacteria.

2.2 Raman spectroscopy measurements

Raman measurements were performed with a Jobin-Yvon Horiba Triax 550 spectrometer, a Modu-Laser (Stellar-Pro-L) argon-ion laser, and a modified Olympus model BX41 microscope. A 100X objective was used to focus about 8 mW of the 514.5 nm laser light onto the sample. The Raman scattered light was collected through the same microscope objective, dispersed with a 1200 lines/mm grating, then focused onto a liquid-nitrogen cooled charge-coupled device (CCD) detector.

A computer running LabSpec software (Jobin-Yvon Horiba) was used to record the spectra and control the experimental parameters. Each spectrum was constructed from the average of three ten-second exposures on the same spot for a total exposure time of 30 s. About 25 spectra could be obtained from each dried 10 μL bacterial pad by translation of the microscope sample stage. Raman spectra were collected between 600 and 2000 cm^{-1} and measured in 2071 channels. All Raman spectra were calibrated using the well-known Raman peak of a crystalline Si wafer.

2.3 Data processing and multivariate data analysis

Data processing of the Raman spectra included baseline correction and normalization. Background fluorescence was subtracted from each Raman spectrum with a custom Matlab program utilizing an "adaptive minmax" method which used two different polynomials of different order (one constrained and one unconstrained) to fit the fluorescence background spectrum followed by a "minmax" algorithm to prevent overfitting and underfitting of the data [28]. This program also normalized each spectrum to its maximum intensity channel. The processed Raman spectra were then analyzed via a PC-DFA (IBM SPSS Statistics v19, SPSS, Inc.). In subsequent trials, spectra were smoothed utilizing a Matlab Savitzky-Golay filter using a variety of window sizes and smoothing functions (quadratic and cubic) and also by differentiating the data prior to smoothing.

The PCA reduced the dimensionality of each spectrum from 2071 variables to 22 principal component scores (PCs) that accounted for greater than 98% of the variance in the data. DFA was performed on the 22 PCs which served as input independent variables for this analysis. In the PC-DFA data were classified as one of the four *E. coli* strains.

A PLS-DA was performed with the PLS_Toolbox v6.7.1 running under Matlab v7.6 (Eigenvector Research, Inc.). All Raman spectra were mean-centered prior to analysis. The PLS-DA reduced the dimensionality of each spectrum from 2071 independent variables to 20 latent variables. In the alternate data preprocessing, spectra were smoothed with a Savitzky-Golay filter and derivatives were taken with the PLS_Toolbox. In the PLS-DA spectra were classified as belonging to one of two groups: pathogenic or non-pathogenic *E. coli*.

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

2.4 External validation

Classification tests were performed on 478 Raman spectra from the four strains of *E. coli* acquired in 12 separate experiments. Each experiment generated a data set of spectra and each data set was acquired from bacteria cultured on separate media and harvested and tested on different days spanning several months. The data sets are shown in Table 1.

Table 1. Bacterial data sets tested in this study

Data Set Number	Bacterial ID (*E. coli* strain)	# of Spectra in Set	Data Set Number	Bacterial ID (*E. coli* strain)	# of Spectra in Set
1	C	50	7	HF4714	50
2	C	50	8	HF4714	25
3	C	13	9	HfrK-12	48
4	C	30	10	HfrK-12	52
5	O157:H7	35	11	HfrK-12	25
6	O157:H7	65	12	HfrK-12	35

Classification efficacy was determined via external validation of the chemometric models. In an external validation each spectrum was tested against a model containing no other spectra from its data set. This was done by sequentially eliminating each data set listed in Table 1 from the model and testing each spectrum in it individually against a model built with the other 11 data sets. In comparison, a "cross-validated" test (commonly called a "leave-one-out" or LOO analysis) only removes one spectrum at a time from the model. In this way, each spectrum is tested against a model containing the other 477 spectra – including spectra taken at the same time and from the same specimen as the unknown spectrum. Because of this testing methodology, a LOO analysis will most likely always return overly-optimistic results and a sensitivity measured using a cross-validation test should be understood to be an upper limit on the ultimate sensitivity of the test. Results reported here were all obtained with the more realistic "external validation" methodology.

3. Results

Figure 1(a) shows the average of all the processed Raman spectra acquired for each of the four strains under investigation.

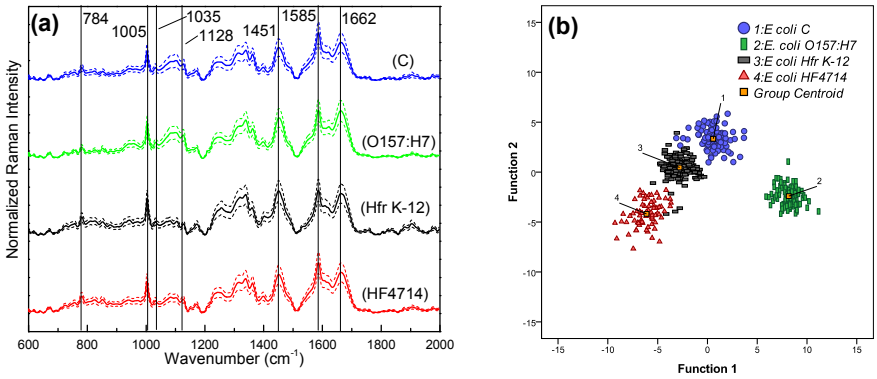

Fig. 1. (a) Normalized averaged Raman spectra of four strains of *E. coli* (top to bottom): C, O157:H7, Hfr K-12, and HF4714. Spectra have been offset vertically for clarity and the wavenumbers of important spectral features are indicated by a vertical line. The standard deviation of all the averaged spectra is indicated by dashed lines above and below the averaged spectrum. (b) A PC-DFA plot showing the first two discriminant function scores of the 478 Raman spectra. No smoothing was performed in this analysis.

Bacterial Raman spectra as shown in Fig. 1(a) consisted of bands representing the cell contents, primarily proteins, lipids, carbohydrates, and nucleic acids. For example, the strong

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

Raman peaks located at 1005 and 1662 cm^{-1} were assigned previously to proteins [15,29], the peak at 1585 cm^{-1} assigned to lipids [30], and the peak at 1451 cm^{-1} assigned to carbohydrates or lipids [15,29]. The smaller Raman peaks located at 1035 and 1128 cm^{-1} have been previously attributed to carbohydrates [31], while the peak at 784 cm^{-1} was assigned to nucleic acids [29]. The Raman spectroscopic bands observed in the *E. coli* strains were found to be consistent with those published previously [31,32]. Spectral variance was low amongst spectra acquired from a single aliquot mounted on a single microscope slide. The variance increased between spectra acquired on different days from different aliquots and cultures, as anticipated. The standard deviation of the averaged spectra for any one strain is indicated in Fig. 1(a) by the dashed lines around the average.

Figure 1(b) shows the PC-DFA plot for the four *E. coli* strains. Each colored point represents a spectrum which is plotted against its DF1 and DF2 scores. The four main groups were recovered with high reproducibility and the avirulent pathogenic strain (*E. coli* O157:H7) was recovered in a cluster much separated from the other strains, demonstrating a greater variation than the other three strains. The excellent discrimination of the DFA data in Fig. 1(b) is the result of the "internal validation" that the DFA routine performs during the construction of the discriminant functions. Since the identity of every spectrum is known, the classification is excellent, as expected. This classification should not be interpreted as the accuracy of the test when the identities of the spectra are unknown. In that case accuracy drops. The classification accuracy of this PC-DFA as determined by the "leave one out" analysis was still excellent: 100% of *E. coli* O157:H7, 99.3% of *E. coli* C, 99.4% of *E. coli* Hfr K-12, and 98.7% of *E. coli* HF4714 spectra were correctly identified. As shown below, this dropped to more realistic values when entire data sets were withheld in the external validation test.

To investigate the reason for the large difference between the *E. coli* O157:H7 spectra and the spectra from the non-pathogenic strains, the spectral regions most important for discrimination were identified by comparing the plot of the first PC loading with the spectral differences between the average spectra. This difference was obtained by simply subtracting the spectra from each other. Figure 2 reveals the similarity between the PC1 loading plot and the difference between the average spectra of *E. coli* O157:H7 and *E coli* C.

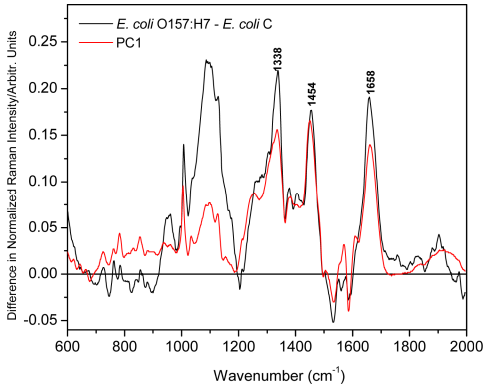

Fig. 2. The first principal component loading of the PCA (red) plotted with the difference of the average Raman spectrum of *E. coli* O157:H7 and *E coli* C (black). A strong correlation between these two shows that the difference between pathogenic *E. coli* O157:H7 and *E coli* C particularly in the important Raman bands at 1658, 1454, and 1338 cm^{-1} accounts for a significant amount of the overall variance in the data. No smoothing was performed in this analysis.

The differences in the intensities of the strong Raman bands located at 1338, 1454, and 1658 cm^{-1} represent the spectral features that possessed the most between-group variance in the data and were thus utilized as the primary basis for discrimination. Those peaks were observed in all spectra and have been previously assigned for protein or DNA, carbohydrates,

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

and protein respectively [33,34]. Although the difference of the EHEC and *E. coli* C Raman band around 1100 cm^{-1} was fairly large, its fairly small principal component loading showed that it was not responsible for reliable discrimination amongst all four strains which were used in the principal component analysis. This is further evidence that simple "subtraction" or "overlaying" of spectra is not an adequate indicator of reliable discrimination or classification in a multiple-group analysis.

Lastly, an EHEC-identification assay was simulated by performing an external validation test on both the PLS-DA and PC-DFA models. Each of the twelve data sets listed in Table 1 was sequentially withheld from the model to serve as a test group and then one by one spectra from that withheld test group were tested to see if they were classified as EHEC or not. The results of the external validation PLS-DA and PC-DFA tests are shown in Table 2, where the percentage of spectra in each data set identified as EHEC (O157:H7) is given. The overall sensitivity (percentage of true positives) and the specificity (one-hundred minus the percentage of false positives) for each analysis are given.

Table 2. Classification accuracy of PLS-DA and PC-DFA *E. coli* O157:H7 tests with no spectral preprocessing

Bacterial ID (# in group)	PLS-DA % ID'd as EHEC	PC-DFA % ID'd as EHEC	Bacterial ID (# in group)	PLS-DA % ID'd as EHEC	PC-DFA % ID'd as EHEC
C(50)	0.00%	0.00%	HF4714(50)	0.00%	0.00%
C(50)	0.00%	0.00%	HF4714(25)	36.00%	76.00%
C(13)	15.38%	30.77%	HfrK-12(48)	0.00%	0.00%
C(30)	36.67%	40.00%	HfrK-12(52)	0.00%	0.00%
O157:H7(35)	100.00%	100.00%	HfrK-12(25)	0.00%	0.00%
O157:H7(65)	100.00%	100.00%	HfrK-12(35)	0.00%	0.00%

PLS-DA		PC-DFA	
Sensitivity	100.00% ± 0.00%	Sensitivity	100.00% ± 0.00%
Specificity	94.18% ± 13.61%	Specificity	90.74% ± 22.50%

Figure 3 graphically shows PLS-DA results for two of the twelve tests. In these tests, PLS-DA built a spectral model that maximized the variance between EHEC data and non-EHEC data. It then assigned each spectrum a single "Y predictor variable" that was used to classify the spectrum as EHEC or not. In these tests spectra from non-EHEC *E. coli* possessed a Y predictor variable of approximately 0.2 (arbitrary units) and spectra from EHEC possessed a Y predictor variable of approximately −0.8 (arbitrary units). The difference in the means was 1.0. Unknown spectra were tested by calculating their Y predictor values and determining if the value fell above or below a PLS-DA calculated discrimination line. This line (which can be user-determined if desired) is calculated to minimize the number of false positives and negatives in the model.

In Fig. 3(a) an *E. coli* C data set was withheld from the model and was used to test the two-class classification. 100% of the 50 spectra were correctly classified (above the PLS_Toolbox-selected predictor line) as belonging in the non-pathogenic group. All eleven other data sets were used to create the model. In Fig. 3(b) an EHEC data set was withheld from the model and was used to test the two-class classification. 100% of the 35 spectra were correctly classified (below the PLS_Toolbox selected predictor line) as belonging in the pathogenic group. Because the models were created with different spectral data sets, the predictor values and discrimination lines were not the same in the two tests.

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

4. Discussion

All of the results presented up to this point, including in Table 2, were obtained with no spectral preprocessing other than the background subtraction and normalization already described. To determine if the specificity could be improved reproducibly above 95%, spectra were smoothed to eliminate spurious noise spikes using a Savitzky-Golay filter with various window sizes from one channel (no smoothing) to 45 channels, using a cubic and quadratic analytic function in the window, and also taking a second derivative prior to smoothing.

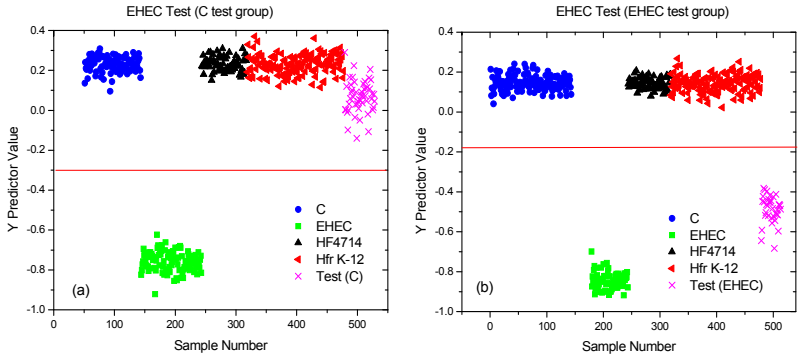

Fig. 3. PLS-DA results of an EHEC test on (a) non-pathogenic *E. coli* strain C and (b) pathogenic O157:H7. In (a) 100% of the non-pathogenic strain C spectra in the test group were correctly identified as being nonpathogenic *E. coli*, possessing Y predictor values above the determined discrimination line. In (b) 100% of the pathogenic strain O157:H7 spectra in the test group were correctly identified as being pathogenic, possessing Y predictor values below the determined discrimination line. No smoothing was performed in this analysis.

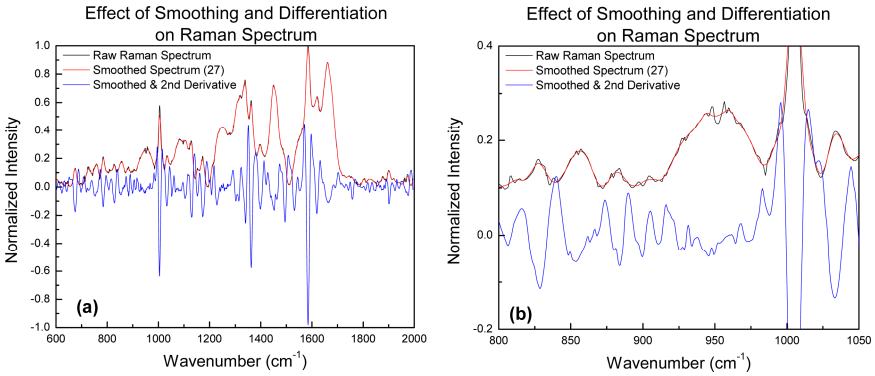

Fig. 4. (a) The unprocessed (raw), smoothed (with a window size of 27), and smoothed 2nd derivative spectrum of *E. coli* O157:H7. (b) The same spectra zoomed in on the region from 800 – 1050 cm^{-1}.

The PLS-DA showed a small increase in specificity for a smoothing window of 27 channels, with a slight improvement when the PLS_Toolbox was allowed to take the derivatives prior to smoothing. These were chosen to be the optimal conditions for PLS-DA by virtue of yielding the best sensitivity and specificity. The PC-DFA demonstrated little dependence upon smoothing window size or the use of quadratic or cubic functions, except for when a 15 channel quadratic smoothing was performed. In this case, the specificity increased to greater than 98%, reducing the numbers of false positives from 40% to 6.67% in data set four and from 76% to 4% in data set eight. This was chosen to be the optimal condition for PC-DFA, although it is likely that with a larger number of data sets the anomalous behavior of the 15 channel window smoothing would be removed. Therefore although the specificity of the PC-DFA is reported to be in excess of 98%, it is likely from the other tests conducted that 91% is a more realistic specificity. Differentiation performed

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013

outside of the PLS_Toolbox prior to smoothing did not yield efficient classification with PC-DFA and this preprocessing was not investigated further. This will be an ongoing area of investigation. In Fig. 4 the effect of smoothing and second differentiation on a representative Raman spectrum of *E. coli* C is presented. The final performance of the two multivariate techniques using the optimized preprocessing routines is shown in Table 3.

Table 3. Classification accuracy of PLS-DA and PC-DFA *E. coli* O157:H7 tests using optimal preprocessing

Bacterial ID (# in group)	PLS-DA % ID'd as EHEC	PC-DFA % ID'd as EHEC	Bacterial ID (# in group)	PLS-DA % ID'd as EHEC	PC-DFA % ID'd as EHEC
C(50)	0.00%	2.00%	HF4714(50)	0.00%	0.00%
C(50)	0.00%	0.00%	HF4714(25)	4.17%	4.00%
C(13)	23.08%	23.08%	HfrK-12(48)	0.00%	0.00%
C(30)	0.00%	6.67%	HfrK-12(52)	3.85%	0.00%
O157:H7(35)	100.00%	100.00%	HfrK-12(25)	0.00%	0.00%
O157:H7(65)	100.00%	100.00%	HfrK-12(35)	0.00%	0.00%

PLS-DA		PC-DFA	
Sensitivity	100.00% ± 0.00%	Sensitivity	100.00% ± 0.00%
Specificity	98.41% ± 4.59%	Specificity	98.15% ± 4.71%

5. Conclusions and future work

Spontaneous Raman scattering can effectively discriminate pathogenic *E. coli* O157:H7 from nonpathogenic strains of *E. coli* given the use of an appropriate multivariate chemometric technique to classify unknown spectra. *E. coli* O157:H7 is a pathogen of significant public-health interest but the nonpathogenic strains are quite common, suggesting the need for a rapid non-genetic test to differentiate the pathogenic from nonpathogenic strains.

Partial least squares discriminant analysis and principal component discriminant function analysis both showed sensitive (high rates of true positives) and specific (low rates of false positives) classification of a strain of pathogenic *E. coli* from three strains of nonpathogenic *E. coli*. Although these results are compelling, additional strains of *E. coli* need to be added to this model including additional pathogens such as enterotoxigenic (ETEC) *E. coli*, enteropathogenic (EPEC) *E. coli*, and enteroaggregative (EAggEC) *E. coli*, and additional nonpathogenic strains, including the very common coliform types commonly used as indicators of water quality. In addition, strains cultured on a wide variety of nutrient media should be included in the model to provide even greater possibilities for molecular diversity in the recorded spectra.

Lastly, it remains to be seen what number of colony-forming units (CFU's) is necessary to provide such sensitive and specific discrimination. An unrealistically high bacterial titer for in situ identification was used in this study, but the possibility of using Raman spectroscopy in a microbiology laboratory for strain-identification after culturing remains. In such an application RS could be performed as well as other more traditional methods, to perhaps improve on the post-culture speed of species and strain identification.

Acknowledgments

The authors would like to acknowledge Talha Syed, Eldar Kurtovic, and Emir Kurtovic for assistance with the microbiological sample preparation. This work was supported in part by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN/418254-2012). RAP was supported by the University of Windsor's Outstanding Scholars program. KH was supported in part by the financial contributions of Mr. Richard Barber.

#181975 - $15.00 USD Received 21 Dec 2012; revised 15 Feb 2013; accepted 19 Feb 2013; published 1 Mar 2013