Digital microbiology: detection and classification of unknown bacterial pathogens using a label-free laser light scatter-sensing system

Bartek Rajwa[a*] M. Murat Dundar[b], Ferit Akova[b], Valery Patsekin[a], Euiwon Bae[c], Yanjie Tang[d] J.Eric Dietz[e], E. Daniel Hirleman[f], J. Paul Robinson[a], and Arun K. Bhunia[d].

[a]Bindley Bioscience Center, Purdue University, 1203 W. State Street, West Lafayette, IN 47907;
[b]Computer & Information Science Dept., Indiana University Purdue University Indianapolis (IUPUI), 723 W. Michigan St., SL 280C, Indianapolis, IN 46202
[c]School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907
[d]Department of Food Science, 745 Agriculture Mall Drive, West Lafayette, IN 47907
[e]Homeland Security Institute, Purdue University, Mann Hall, 203 S. Martin Jischke Drive, West Lafayette, IN 47907
[f]School of Engineerning, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343

ABSTRACT

The majority of tools for pathogen sensing and recognition are based on physiological or genetic properties of microorganisms. However, there is enormous interest in devising label-free and reagentless biosensors that would operate utilizing the biophysical signatures of samples without the need for labeling and reporting biochemistry. Optical biosensors are closest to realizing this goal and vibrational spectroscopies are examples of well-established optical label-free biosensing techniques. A recently introduced forward-scatter phenotyping (FSP) also belongs to the broad class of optical sensors. However, in contrast to spectroscopies, the remarkable specificity of FSP derives from the morphological information that bacterial material encodes on a coherent optical wavefront passing through the colony. The system collects elastically scattered light patterns that, given a constant environment, are unique to each bacterial species and/or serovar. Both FSP technology and spectroscopies rely on statistical machine learning to perform recognition and classification. However, the commonly used methods utilize either simplistic unsupervised learning or traditional supervised techniques that assume completeness of training libraries. This restrictive assumption is known to be false for real-life conditions, resulting in unsatisfactory levels of accuracy, and consequently limited overall performance for biodetection and classification tasks. The presented work demonstrates preliminary studies on the use of FSP system to classify selected serotypes of non-O157 Shiga toxin-producing *E. coli* in a nonexhaustive framework, that is, without full knowledge about all the possible classes that can be encountered. Our study uses a Bayesian approach to learning with a nonexhaustive training dataset to allow for the automated and distributed detection of unknown bacterial classes.

Keywords: *E. coli*, STEC, non-O157, classification, light scatter, biodetection

1. INTRODUCTION

Traditional bacteria recognition methods based on antibodies or genetic matching remain labor intensive and time consuming, and involve multiple steps. Recently a number of new biosensors designed to perform classification of bacteria in a label-free manner have been reported in the literature.[1,2] These tools are label-free in

*Send correspondence to Bartek Rajwa, e-mail: brajwa@purdue.edu, telephone: 765 496 1153

Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring; and Biometric Technology for Human Identification VIII, eds. Southern, Montgomery, Taylor, Weigl, Vijaya Kumar, Prabhakar, Ross, Proc. of SPIE Vol. 8029, 80290C · © 2011 SPIE · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.884541

the sense of not employing fluorescence labels, but using alternative detection modalities such as surface plasmon resonance, amperometric and potentiometric measurements, or electrochemical impedance spectroscopy. However, these biosensors still utilize traditional biological recognition elements: enzymes, antibodies, and nucleic acids. The only well researched and broadly utilized techniques capable of true reagentless fingerprinting of bacteria are vibrational spectroscopic methods (Raman and IR),[3–9] autofluorescence-based observations,[10,11] light-scattering analysis of bacterial suspensions[12–15] or colonies,[16–20] and MALDI-TOF-based systems.[21–24]

The elastic light-scatter fingerprinting system (also called BARDOT – BActeria Rapid Detection using Optical scattering Technology) developed at Purdue University has shown great promise for distinguishing bacterial cultures at the genus, species, and strain level for *Listeria, Staphylococcus, Salmonella,* and *Vibrio*.[16–18] BARDOT uses a sophisticated light-scatter analysis to extract a forward light–scatter phenotype (FSP) from bacterial cultures on the surface of agar in a semi-solid state.[16,17,25] The FSP information is subsequently used by a pattern-recognition system to assess the possible origin of the distinctive forward-scattering pattern. It has been successfully shown that scattering properties of *Listeria, E. coli, Salmonella, Staphylococcus*, and *Vibrio* colonies can be used to differentiate the species occurring in food samples as well as those isolated from experimentally infected animals.[18]

In this report we summarize our preliminary results demonstrating feasibility of forward-scatter phenotyping for classification of Shiga toxin–producing *Escherichia coli* (STEC) colonies, specifically non-O157 serotypes. STEC can cause diarrhea, bloody diarrhea (hemorrhagic colitis), and hemolytic uremic syndrome (HUS). The detection of non-O157 STEC is not widely practiced and few laboratories are able to detect non-O157 strains. This is primarily because many non-O157 STEC strains lack the characteristics of O157 STEC, such as delayed fermentation of sorbitol and the hemolytic activity on hemolysin agar, and therefore cannot be identified on the routinely used modified sorbitol-MacConkey agar (SMAC).[26,27]

About 70% of non-O157 STEC isolated from humans falls into six serogroups: O26, O103, O111, O121, O45, and O145. It has been estimated that in the United States about half the diarrheal illnesses caused by STEC are due to O157 and about half to subtypes of *E. coli* other than O157; and some of these illnesses due to organisms other than O157 can be manifested by symptoms just as severe as those caused by STEC O157. In this report we attempt to detect three of these serotypes (O26, O103, and O111) in addition to O157.

Previously, we demonstrated that classification of *Salmonella* phenotypes not only can be successfully performed with access to exhaustive training libraries, but can also be implemented in a nonexhaustive regimen, i.e., without assuming that access to the complete training set is always available. The rationale behind the nonexhaustive classification is based on the observation that the number of *Salmonella* serotypes is too large to allow for a practical traditional statistical machine learning approach. The *Salmonella* results showed that different serotypes indeed exhibit distinguishable forward-scatter phenotypes and can be modeled as well-defined multidimensional Gaussians in the feature space. Although this study demonstrated the feasibility of our approach, it also promoted speculation that a more complex distribution of phenotypes may lead to classification problems that ultimately would not be solvable with the help of parametric methods, i.e., techniques assuming some level of knowledge about the nature of the distributions. In this report we demonstrate

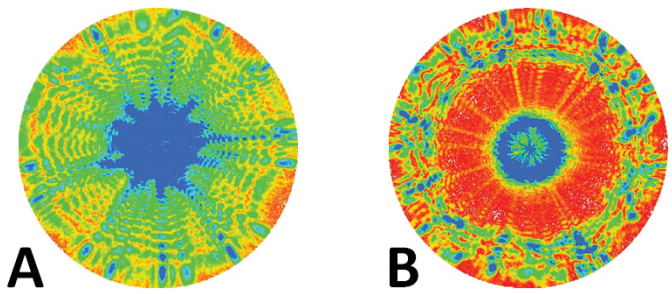

Figure 1. Examples of laser light scatter patterns formed by *E. coli* colonies. A – O157:H7, B – O26:H11. Pseudocolor look-up table added to visualize intensity differences.

our first attempt to apply a nonexhaustive learning technique to the problem of detecting non-O157 STEC. We show promising preliminary results for models assuming both exhaustive and nonexhaustive training libraries, and we conclude that the complex nature of non-O157 STEC phenotype distribution indeed may require a new generation of statistical machine learning methods to cope with the extraordinary diversity of *E. coli*.

2. METHODS

2.1 Cell cultures

Seven serotypes of *E. coli* (O157:H7, O26:H11, O103:H2, O103:H6, O103:H11, O111:H8, and O111:NM) from the USDA ARS Culture Collection and collection of the Department of Food Science, Purdue University, were plated on medium and incubated at 37°C for 12-13 h, depending on the growth speed. MacConkey agar with sorbitol (SMAC) was used as the growth medium. Plates were incubated at 37°C until colony size reached 1.3 0.2 mm in diameter before analysis using BARDOT.

2.2 Forward-scatter phenotyping system

The current implementation of BARDOT consists of three major components: colony counter/locator, forward-scatter measurement device, and 2-D motorized stage. The forward-scatter measurement device is composed of a laser diode module (VHK Circular Beam Visible Laser Diode 0.95 mW, 635 nm), and a monochromatic CMOS image sensor (PL-B741U-BL, 1.3 MP, PixeLINK, Ottawa, ON, Canada) with 1280 x 1024 resolution and equipped with a USB interface. The PL-B741 is a high-performance camera based on the Cypress IBIS5 CMOS global shutter progressive scan sensor with a 2/3 optical format and 27 fps output. The BARDOT system also uses the camera to operate in colony-counting mode with the assistance of a Tamron Mega Pixel M118FM08 lens with manual iris (Tamron USA, Inc, Commack, NY). The motion control is provided by stepping motors (Velmex Vexta Type 17, Velmex, Bloomfield, NY) which are driven by the 2-axis programmable stepping-motor controller (VXM1-1, Velmex, Bloomfield, NY).

2.3 Forward-scatter pattern analysis

Although the resultant scatter patterns could in theory be analyzed on the basis of a rigorous mathematical model of scatter, the complexity of the problem precludes easy application of this straightforward solution. Instead we developed a processing pipeline based on the idea of employing a statistical machine learning model to solve the inverse light-scatter problem for bacterial colonies. This approach has previously been successfully used with smaller-scale problems, such as evaluation of size and refractive indices of single particles.

The features extracted from scatter patterns are, in principle, readily quantified. The polar nature of the scattering signatures as exemplified in Figure 1 suggests decomposition using a set of orthogonal polynomials with radial characteristics. Our current implementation of the classification system employs pseudo-Zernike moments (PZMs).[28–30]

To compute the pseudo-Zernike moments of a given image, the center of the image is taken as the origin and pixel coordinates are mapped to the range of the unit circle. Rotational invariance is obtained by using the magnitudes of the pseudo-Zernike moments as features. The details of pseudo-Zernike moment computation can be found in our published reports.[16,18]

Another set of tools used in the reported work for quantifying patterns are gray-level co-occurrence matrices (GLCMs). The GLCMs are routinely employed to quantify the number of occurrences at various distances and angles of pixel intensity values with respect to each other.[31,32] The GLCMs are used to extract 14 low- and high-frequency (depending on the pixel-to-pixel distance used in the co-occurrence matrix) texture properties (so-called Haralick texture features). We used a total of 84 GLCMs-based features computed for distances between 1 to 6 pixels.

The best features were selected using a wrapper approach provided by `varSelRF` package for R.[33,34] This package implements a feature selection technique using random forests, and it is built upon the `randomForest` package which implemented the algorithm proposed by L. Breiman.[35]

2.4 Automated classification

Two different classification systems were used. The exhaustive system assumed that all the classes which may be encountered are accounted for in the training library. Therefore an "exhaustive" library is expected. This detection/classification system has been implemented using both Bayesian methods (data not shown) and a support vector machine algorithm with an RBF kernel using `e1071` package for R.[36,37]

The nonexhaustive classification regimen does not make an assumption regarding the completeness of the classes list in the training set. Instead it formulates the problem as a combination of classification and detection. A new instance is assumed to be either a member of one of the existing classes or a member of a new "unknown" class. The new class is progressively recovered from the emerging distribution with the arrival of new instances. The nonexhaustive statistical machine learning approach solves the problem of encountering emerging pathogens for which there are no available examples in the database. The details of the implementation and description of the classification methodology are provided in our recent publications.[19,38]

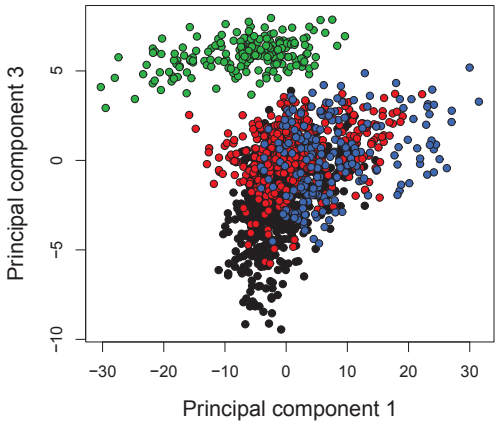

Figure 2. Principle component analysis plot demonstrating mixture of *E. coli* STEC scatter patterns. Note that O157:H7 (shown as green) is the only class which can be separated using this unsupervised approach in linear space.

Briefly, we assume that each sample, characterized by a feature vector, x_i, is distributed according to a normal distribution with mean vector, μ, and a covariance matrix, Σ, i.e., $x_i \sim N(\mu_k, \Sigma_k)$. This approach uses maximum likelihood detection evaluated for all the classes in the training dataset to determine presence of emerging class. Once a sample is determined to be a novelty, a new class is generated and the current set of known classes is augmented with this class. If a sample is not a novelty, then the algorithm check if the class maximizing the likelihood is a previously discovered class that is not one of the initially known classes. If yes, the class parameters are updated; otherwise they are not updated. The mean vectors, μ_k, are estimated by the sample mean. An inverted Wishart prior with m degrees of freedom and a scale matrix, Ψ is defined over the covariance matrices, i.e., $\Sigma_k \sim W^{-1}(\Psi, m)$ and the covariance matrices for each class are estimated by the posterior mean.

In the described approach, the labels of the newly created classes are not known until they are analyzed by an independent method. To differentiate these classes from those initially available in the training library we use the concept of labeled vs. unlabeled classes, where the terms *labeled* and *unlabeled* refer to existing and newly generated classes, respectively. The parameters of the labeled classes are estimated once in the beginning, whereas those of unlabeled classes are recursively updated as more colonies are assigned to these classes via sequential classification. When a new class is generated it will initially contain one sample. Therefore the sample estimate of Σ_k in this case will be ill conditioned. Estimating Σ_k by the posterior mean alleviates this situation as long as the number of labeled classes in the training dataset allows for a robust estimate of Ψ. Thus, the core assumption of our algorithm is that the parameter sets defining the class-conditional distributions originate from a common prior distribution and that the number of labeled classes is large enough to obtain a robust estimate of this distribution.

3. RESULTS

The tested dataset contained FSPs of several thousand colonies collected over a number of days. A complete set of features was extracted using a custom-developed BugBuster software package. A total of 149 features, including 65 pseudo-Zernike moments and 84 GLMC-based features, was computed for every colony. The FSPs of all the colonies growing on different plates and on different days were pooled together to account for natural biological variability. Three hundred and twenty random phenotypes (instances) per class were drawn from the data pool for training and crossvalidation.

Two different classification models were used for training. In the first one 4 classes of STEC were defined by pooling all the organism expressing the same O-antigen. The second model defined classes on the basis of O and H serotypes. Additionally, every model could be used directly in a multi-class fashion or employed in a two-step process in which O157:H7 is classified (and eliminated from the dataset) first, followed by classification of the remaining non-O157 samples. For every model an overall multi-class classification score was calculated, as well as individual scores that were obtained assuming a one-vs-all case. Therefore, the individual classification scores show the ability of a classifier to correctly select only the given class from the mixture of all the other classes, effectively reducing the task to a two-class problem.

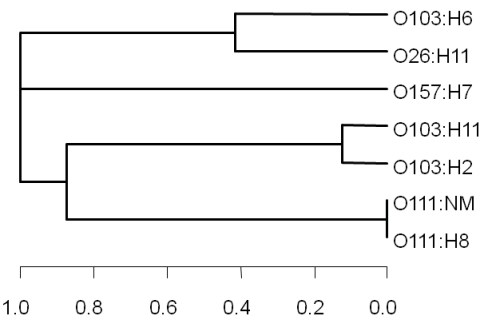

Figure 3. A dendrogram illustrating phenotypic similarities between tested serotypes of *E. coli*

The unsupervised feature reduction by PCA demonstrated that O157 was easily separable in the linear space. Therefore, it was expected that detection of this group can be performed with a very high level of sensitivity and specificity regardless of the assumed model. Indeed, as Tables 1 and 2 demonstrate O157:H7 can be classified with accuracy, sensitivity, and specificity of 100% for both the 4-class model and the 7-class model. The overall accuracy for the 4-class model is indeed slightly higher higher than 7-class model, but the difference is not large. Owing to the characteristic change in colony color when grown on SMAC the O157 serotype is the easiest to detect and separate using light-scatter properties. The individual classification success for non-O157 classes depends on a model and expressed as F-scores varies from 0.8 to 0.95. The lowest classification success was noted for O26:H11 serotype.

O-antigen	Sensitivity	Precision	F-measure	Specificity	Accuracy
O103	0.86	0.90	0.88	0.98	0.96
O111	0.95	0.94	0.95	0.99	0.98
O157	1.00	1.00	1.00	1.00	1.00
O26	0.88	0.87	0.87	0.98	0.96

Table 1. Classification success for 4-class model.

The models used in the second part of the study did not assume access to complete information regarding the classes. It was expected that any new instance submitted for classification might in fact belong to a new class for which the current system had no examples. This type of classification reflected a real-life situation in which a laboratory would

Serotype	Sensitivity	Precision	F-measure	Specificity	Accuracy
O157:H7	1.00	1.00	1.00	1.00	1.00
O26:H11	0.81	0.81	0.81	0.98	0.96
O103:H2	0.84	0.88	0.86	0.99	0.97
O103:H6	0.92	0.93	0.92	0.99	0.99
O103:H11	0.87	0.88	0.87	0.99	0.97
O111:H8	0.92	0.88	0.90	0.99	0.98
O111:NM	0.92	0.94	0.93	0.99	0.99

Table 2. Classification success for 7-class model.

not have information available about an emerging pathogen that might be encountered in tested samples at the outset of an outbreak. The problem of library nonexhaustiveness is not limited to biophysical methods such as BARDOT, Raman spectroscopy, or IC-MALDI-TOF. In fact, this issue is even more limiting in the case of molecular methods that rely on particular genetic information to produce a meaningful outcome. The biophysical techniques will produce a measurement regardless of the quality of a training library. However, the presence of an unknown class will render all the automated classification results untrustworthy unless a mechanism exists to detect the presence of a new population.

In order to simulate such a detection process we removed one known class at a time from the training library and performed an *in silico* experiment during which unknown instances mixed with data from known classes were analyzed by our classifier. The results summarized in Table 3 show that the BARDOT system was indeed able to recognize a large portion of the unknown instances and raise an alarm about the presence of an emerging pathogen not represented in the library. The detection success is reported as area under the receiver operating characteristic (ROC) curve, which is a plot of the detector sensitivity, or true positive rate, vs. the false positive rate ($1 -$ specificity or $1 -$ true negative rate), as the detector discrimination threshold is varied. The area under the curve (AUC) value is equal to the probability that our emerging-pathogen detector will rank a randomly chosen unknown pathogenic colony higher than a randomly chosen known one. The performance of our detector varies from 0.79 for O103:H11 and O103:H2 to over 0.94 for O111.

Detected class	AUC	SD
4-class model		
O103	0.871	0.009
O111	0.943	0.006
O26	0.829	0.009
7-class model		
O103:H11	0.797	0.020
O103:H2	0.797	0.012
O103:H6	0.873	0.008
O111:H8	0.918	0.011
O111:NM	0.949	0.006
O26:H11	0.852	0.015

Table 3. AUC values illustrating performance of the emerging pathogen-detection system for 4-class and 7-class models. The O157 serotype was assumed to be always known.

In order to compare the relationship between the 7-class and 4-class models we computed an FSP-based 7-class dissimilarity score. The dissimilarity between classes was defined as a normalized classification error for a two-class linear SVM classification problem, with no prior feature selection applied. Therefore the dissimilarity score reflected the overlap between classes in feature space. Classes with higher similarity (lower dissimilarity) overlap more than do classes with lower similarity. The resultant dendrogram is presented in Figure 3. The most interesting feature of the constructed hierarchy is the fact that O103 class carrying the same O-antigen but different H-antigens did not cluster together. This may indicate an influence of H-antigens on the resultant phenotype or problems with reproducibility.

4. DISCUSSION AND CONCLUSIONS

Traditionally the presence of *E. coli* O157 is detected employing the SMAC technique. However, in contrast to O157:H7, which grows easy-to-distinguish pale colonies when plated on SMAC, most non-O157 STEC colonies appear pink and are visually indistinguishable. Owing to the color change, the light-scattering pattern of pale O157 colonies is dramatically different from other collected patterns and can be readily distinguished, as illustrated in Fig 3.

The distribution of forward-scatter features of the non-O157 STEC serotypes of *E. coli* reflects a very higher level of biological phenotypic diversity. Our 4-class model assumed that only the O antigens could effectively affect the FSP; however, the classification results obtained from studies of the 7-class model demonstrated that this assumption might be wrong. We were able to detect minor phenotype changes in classes carrying various H antigens. Although our previous work indicated that differences in lipopolysaccharide (LPS) content could be implicated in formation of distinguishable FSPs, the new data suggest that the phenotypic variability may be

also affected by various H antigens. This problem requires future research, as it is too early to speculate on the role of H antigen on colony micro- and nanostructure.

Although there is no standardized detection method for non-O157 STEC serotypes, there are techniques available that can be used to serotype non-O157 organisms. Among them are a new set of differential media for O26, 103, O111, and O145, immunomagnetic separation with specific antibodies, and molecular methods.[27,39] Regrettably, none of these methods is fast and simple, and all require an enrichment step. Moreover, the molecular methods do not really identify the organism causing illness – they can merely detect the presence of particular genes. Therefore unless the organism is isolated the molecular methods cannot determine which organism carries these virulence genes. This constraint places additional stress on the importance of cultural approaches and illustrates why alternative rapid methods working in conjunction with selective media for classification of *E.coli* are highly desirable.

Recently it has been demonstrated that some Shiga-toxin producing non-O157 serotypes (O165:H25, O26:H11 and H32, O156:H25) can be recognized using IC-MALDI-TOF technique.[24] These results provide evidence that direct label-free phenotypic determination of O-serogroups is indeed possible. Since the O-antigen is part of the LPS of the Gram-negative outer membrane, it is expected that optical methods sensitive to subtle differences in LPS content can also be utilized to detect organisms carrying various O-antigens. It has been previously demonstrated that difference in LPS content can indeed be implicated as one of the sources of diversity in forward-scatter patterns formed by interaction of a laser beam with bacterial colonies. Compared to IC-MALDI-TOF the FSP-base classification required no laborious pre-treatment of samples. The bacterial colonies could be interrogated while intact on agar plates, and no additional manipulation or handling of samples was required. The sample-analysis cost as well as the cost of required instrumentation also favors the optical approach.

The limited scope of the demonstrated classification system, and the classification accuracy of the FSP methodology at the current stage of technique development would not allow it to be competitive with established PCR-based or immunological methods as the final determination protocol of *E. coli* serotype. However, since FSP method shows a lot of promise we work on expanding or FSP-based classifier to other common non-O157 STEC including O121, O145, and O45. The access to an inexpensive, label-free, and nondestructive detection system would provide the chance to communicate preliminary positive results to submitting physicians without delay, although with the clarification that results should be correlated with patient symptoms, and a follow-up test would be performed using molecular methods.[26,27] The fact that FSP can be performed simultaneously with standard automated colony-counting procedures, does not require any special media nor any biochemical treatment, and is nondestructive, allowing the plates to be further processed by any confirmatory techniques makes it a good candidate for inexpensive and easy to implement STEC detection method.

ACKNOWLEDGMENTS

The project was supported by grants number 5R21AI085531-02 and 1R56AI089511-01 from the National Institute of Allergy and Infectious Diseases (NIAID), and through a cooperative agreement with the Agricultural Research Service of the US Department of Agriculture project number 1935-42000-035 and the Center for Food Safety and Engineering at Purdue University.

REFERENCES

1. O. Lazcka, F. J. D. Campo, and F. X. Muoz, "Pathogen detection: a perspective of traditional methods and biosensors," *Biosensors & Bioelectronics* **22**, pp. 1205–1217, Feb 2007. PMID: 16934970.
2. A. K. Bhunia, "Biosensors and bio-based methods for the separation and detection of foodborne pathogens," *Advances in Food and Nutrition Research* **54**, pp. 1–44, 2008. PMID: 18291303.

3. D. Helm, H. Labischinski, G. Schallehn, and D. Naumann, "Classification and identification of bacteria by Fourier-transform infrared spectroscopy," *Journal of General Microbiology* **137**(1), pp. 69–79, 1991.
4. D. Naumann, D. Helm, and H. Labischinski, "Microbiological characterizations by FT-IR spectroscopy," *Nature* **351**, pp. 81–82, May 1991. PMID: 1902911.
5. P. Rsch, M. Schmitt, W. Kiefer, and J. Popp, "The identification of microorganisms by micro-Raman spectroscopy," *Journal of Molecular Structure* **661-662**, pp. 363–369, Dec 2003.
6. C. Yu, J. Irudayaraj, C. Debroy, Z. Schmilovtich, and A. Mizrach, "Spectroscopic differentiation and quantification of microorganisms in apple juice," *Journal of Food Science* **69**(7), pp. 268–272, 2004.
7. M. Gupta, J. Irudayaraj, C. Debroy, Z. Schmilovitch, and A. Mizrach, "Differentiation of food pathogens using FTIR and artificial neural networks," *Transactions of the ASAE* **48**(5), pp. 1889–1892, 2006.
8. M. Harz, P. Rösch, and J. Popp, "Vibrational spectroscopy-a powerful tool for the rapid identification of microbial cells at the single-cell level," *Cytometry Part A* **75A**(2), pp. 104–113, 2009.
9. D. F. M. Willemse-Erix, M. J. Scholtes-Timmerman, J. Jachtenberg, W. B. van Leeuwen, D. Horst-Kreft, T. C. B. Schut, R. H. Deurenberg, G. J. Puppels, A. van Belkum, M. C. Vos, and K. Maquelin, "Optical fingerprinting in bacterial epidemiology: Raman spectroscopy as a real-time typing method," *Journal of Clinical Microbiology* **47**, pp. 652–659, mar 2009.
10. C. Estes, A. Duncan, B. Wade, C. Lloyd, W. Ellis, and L. Powers, "Reagentless detection of microorganisms by intrinsic fluorescence," *Biosensors and Bioelectronics* **18**, pp. 511–519, May 2003.
11. M. S. Ammor, S. Delgado, P. Alvarez-Martin, A. Margolles, and B. Mayo, "Reagentless identification of human bifidobacteria by intrinsic fluorescence," *Journal of Microbiological Methods* **69**, pp. 100–106, apr 2007.
12. P. J. Wyatt, "Identification of bacteria by differential light scattering," *Nature* **221**, pp. 1257–8, Mar 1969. PMID: 4886590.
13. P. J. Wyatt and D. T. Phillips, "Structure of single bacteria from light scattering," *Journal of Theoretical Biology* **37**, pp. 493–501, Dec 1972. PMID: 4645360.
14. H. B. Steen, "Light scattering measurement in an arc lamp-based flow cytometer," *Cytometry* **11**(2), pp. 223–30, 1990. PMID: 2180652.
15. B. Rajwa, M. Venkatapathi, K. Ragheb, P. P. Banada, E. D. Hirleman, T. Lary, and J. P. Robinson, "Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier," *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **73**, pp. 369–79, apr 2008. PMID: 18163466.
16. B. Bayraktar, P. P. Banada, E. D. Hirleman, A. K. Bhunia, J. P. Robinson, and B. Rajwa, "Feature extraction from light-scatter patterns of Listeria colonies for identification and classification," *Journal of Biomedical Optics* **11**(3), p. 34006, 2006. PMID: 16822056.
17. P. P. Banada, S. Guo, B. Bayraktar, E. Bae, B. Rajwa, J. P. Robinson, E. D. Hirleman, and A. K. Bhunia, "Optical forward-scattering for detection of Listeria monocytogenes and other Listeria species," *Biosensors & Bioelectronics* **22**, pp. 1664–71, Mar 2007. PMID: 16949268.
18. P. P. Banada, K. Huff, E. Bae, B. Rajwa, A. Aroonnual, B. Bayraktar, A. Adil, J. P. Robinson, E. D. Hirleman, and A. K. Bhunia, "Label-free detection of multiple bacterial pathogens using light-scattering sensor," *Biosensors & Bioelectronics* **24**, pp. 1685–92, Feb 2009. PMID: 18945607.
19. B. Rajwa, M. M. Dundar, F. Akova, A. Bettasso, V. Patsekin, E. D. Hirleman, A. K. Bhunia, and J. P. Robinson, "Discovering the unknown: Detection of emerging pathogens using a label-free light-scattering system," *Cytometry Part A* **77A**(12), pp. 1103–1112, 2010.
20. E. Bae, N. Bai, A. Aroonnual, J. P. Robinson, A. K. Bhunia, and E. D. Hirleman, "Modeling light propagation through bacterial colonies and its correlation with forward scattering patterns," *Journal of Biomedical Optics* **15**(4), p. 045001, 2010.
21. J. O. Lay, "MALDI-TOF mass spectrometry of bacteria," *Mass Spectrometry Reviews* **20**(4), pp. 172–194, 2001.
22. D. Dare, "Rapid bacterial characterization and identification by MALDI-TOF mass spectrometry," in *Advanced Techniques in Diagnostic Microbiology*, Y. Tang and C. W. Stratton, eds., pp. 117–133, Springer US, New York, 2006.

23. P. Seng, M. Drancourt, F. Gouriet, B. L. Scola, P. Fournier, J. M. Rolain, and D. Raoult, "Ongoing revolution in bacteriology: Routine identification of bacteria by Matrix-Assisted laser desorption ionization Time-of-Flight mass spectrometry," *Clinical Infectious Diseases* **49**(4), pp. 543–551, 2009.
24. A. Karger, M. Ziller, B. Bettin, B. Mintel, S. Schares, and L. Geue, "Determination of serotypes of Shiga toxin-producing escherichia coli isolates by intact cell matrix-assisted laser desorption ionization-time of flight mass spectrometry," *Applied and Environmental Microbiology* **77**, pp. 896–905, Feb. 2011. PMID: 21115707.
25. E. Bae, P. P. Banada, K. Huff, A. K. Bhunia, J. P. Robinson, and E. D. Hirleman, "Biophysical modeling of forward scattering from bacterial colonies using scalar diffraction theory," *Applied Optics* **46**, pp. 3639–48, jun 2007. PMID: 17514326.
26. J. T. Brooks, E. G. Sowers, J. G. Wells, K. D. Greene, P. M. Griffin, R. M. Hoekstra, and N. A. Strockbine, "Non-O157 Shiga toxinproducing Escherichia coli infections in the united states, 19832002," *Journal of Infectious Diseases* **192**, pp. 1422 –1429, Oct. 2005.
27. E. C. Mathusa, Y. Chen, E. Enache, and L. Hontz, "Non-O157 Shiga toxin-producing Escherichia coli in foods," *Journal of Food Protection* **73**, pp. 1721–1736, Sept. 2010.
28. A. Khotanzad and Y. Hong, "Invariant image recognition by Zernike moments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **12**(5), pp. 489–497, 1990.
29. A. Khotanzad and Y. H. Hong, "Rotation invariant image recognition using features selected via a systematic method," *Pattern Recognition* **23**(10), pp. 1089–1101, 1990.
30. R. Mukundan and K. R. Ramakrishnan, "Fast computation of Legendre and Zernike moments," *Pattern Recognition* **28**, pp. 1433–1442, Sep 1995.
31. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on* **3**(6), pp. 610–621, 1973.
32. R. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE* **67**(5), pp. 786–804, 1979.
33. R. Díaz-Uriarte and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics* **7**, p. 3, 2006. PMID: 16398926.
34. R. Diaz-Uriarte, "GeneSrF and varSelRF: a web-based tool and r package for gene selection and classification using random forest," *BMC Bioinformatics* **8**(1), p. 328, 2007.
35. L. Breiman, "Random forests," *Machine Learning* **45**, pp. 5–32, Oct. 2001.
36. C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
37. D. Meyer, *Support Vector Machines: The interface to libsvm in Package e1071*, 2011. Paper available at `http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf`.
38. F. Akova, M. Dundar, V. J. Davisson, E. D. Hirleman, A. K. Bhunia, J. P. Robinson, and B. Rajwa, "A machine-learning approach to detecting unmatched bacterial serovars," *Statistical Analysis and Data Mining Journal* **3(5)**, pp. 289–301, 2010.
39. B. Poss, L. D. Zutter, M. Heyndrickx, and L. Herman, "Novel differential and confirmation plating media for Shiga toxin-producing Escherichia coli serotypes O26, O103, O111, O145 and sorbitol-positive and -negative o157," *FEMS Microbiology Letters* **282**(1), pp. 124–131, 2008.