

Novel technology for rapid species-specific detection of *Bacillus* spores

Melissa D. Krebs^{a,1}, Brian Mansfield^{b,1}, Ping Yip^b, Sarah J. Cohen^a,
Abraham L. Sonenshein^c, Ben A. Hitt^b, Cristina E. Davis^{a,*}

^aThe Charles Stark Draper Laboratory, Mechanical and Instruments Division, Bioengineering Group, United States

^bCorrelogic Systems Inc., 6701 Democracy Blvd. Suite 300, Bethesda, MD 20817, United States

^cDepartment of Molecular Biology and Microbiology, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, United States

Received in revised form 8 November 2005; accepted 14 December 2005

Abstract

There is an urgent need for a small, inexpensive sensor that can rapidly detect bio-warfare agents with high specificity. *Bacillus anthracis*, the causative agent of anthrax, would be a perilous disease-causing organism in the event of a release. Currently, most anthrax detection research is based on nucleic acid detection, immunoassays and mass spectrometry, with few detection levels reported below 10^5 spores. Here, we show the ability to distinguish *Bacillus* spores to a level approaching 10^3 spores, below the reported median infectious dose of *B. anthracis*, using pyrolysis—micromachined differential mobility spectrometry and novel pattern recognition algorithms that combine lead cluster mapping with genetic algorithms.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Differential mobility spectrometry; *Bacillus* spore detection; Genetic algorithms; Adaptive pattern recognition

1. Introduction

With increasing concern about the potential for a biological agent attack, the need for a portable, inexpensive, and durable sensor that can rapidly detect and identify biological weapons agents continues to grow. *B. anthracis*, the causative agent of anthrax, has been identified as one of the most dangerous disease-causing organisms capable of devastation in the event of a release (Inglesby et al., 1999). Anthrax spores can be inhaled and transported to lymph nodes, germinating up to 60 days later (Friedlander et al., 1993). The germinating bacteria produce a toxin that causes necrosis, edema, and hemorrhaging (Smith and Keppie, 1954; Friedlander, 1997). In the event of a release, the rapid detection of anthrax is critical for effectively treating patients that have been exposed (Brookmeyer et al., 2003). Quickly identifying the presence of environmental spores is difficult as the DNA is well-protected inside the spore and the spore structure is chemically different from that of

vegetative cells. Furthermore, *B. anthracis* is genetically similar to other *Bacillus* species, such as *B. cereus* and *B. thuringiensis* (Read et al., 2003; Radnedge et al., 2003; Helgason et al., 2000), complicating the differentiation of the potential biological weapon from non-pathogenic spores.

Since the October 2001 anthrax attacks in the United States, there has been significant research focused on finding a rapid, sensitive, and specific anthrax detection system. To date, most work has focused on nucleic acid detection (Patra et al., 1996; Lee et al., 1999; Makino et al., 2001; Higgins et al., 1999; Cooney, 2001; Uhl et al., 2002; McBride et al., 2003), which offers extremely high sensitivity. However, the spores must be at least partially germinated prior to the assay, several reagents are required, and assay times are generally greater than a half-hour. Another detection method that has been widely explored is immunodetection (Phillips et al., 1983; Phillips and Martin, 1988; De et al., 2002; Longchamp and Leighton, 1999; Zhou et al., 2002; Quinlan and Foegeding, 1997). Again, these assays can be very sensitive but require a number of qualified reagents and typically take more than 30 min to complete. An additional concern with immuno-based assays is the cross-reactivity of the antibodies and the potential for non-specific binding. Mass spectrometry has also been used to detect spores (Beverly et al., 1996; Fox et al., 1993; Goodacre et al., 2000; Smith and MacDonald, 2004; Ferguson et al., 2004), but the sensitivity is

* Corresponding author at: The University of California, Department of Mechanical and Aeronautical Engineering, One Shields Avenue, 2097 Bainer Hall, Davis, CA 95616, USA. Tel.: +1 530 754 9004; fax: +1 530 752 4158.

E-mail address: cedavis@ucdavis.edu (C.E. Davis).

¹ These authors contributed equally to this work and should both be considered first authors.

not reported to be as high as with nucleic acid or antibody detection. In addition, mass spectrometers are not readily portable, are very expensive and most do not operate at atmospheric conditions, limiting their potential for field-use.

The ID 50 (median infectious dose) for *B. anthracis* is reported to be 8000–10,000 spores (Cieslak and Eitzen, 1999), and the LD 50 (median lethal dose) is 61,800 spores in Rhesus macaques (Vasconcelos et al., 2003). To date, few rapid detection methods can detect spore levels below 100,000. Arakawa et al. (2003) report detection of 1000 spores using microcalorimetric spectroscopy, but this technique fails in the presence of water and thus requires sample lyophilization prior to analysis, which slows the analysis and is not readily deployable in the field.

Here, we demonstrate the ability to distinguish *Bacillus* spores in water to a level below the reported ID 50 using pyrolysis—differential mobility spectrometry (DMS) in combination with a novel pattern recognition and classification algorithm. The DMS is a microfabricated ion mobility spectrometer that is capable of operating at ambient temperature and at atmospheric pressure (Miller et al., 2000, 2001; Eiceman et al., 2002). We have previously shown its use for vegetative bacteria identification (Shnayderman et al., 2005), chemical weapons agent detection (Krebs et al., 2005), and preliminary spore detection experiments (Krebs et al., 2005; Davis et al., 2003). The pattern recognition algorithm was originally described in a paper describing the application of pattern recognition to serum profiles generated by mass spectrometry for the detection of ovarian cancer markers (Petricoin et al., 2002). The algorithm combines elements from genetic algorithms first described by Holland (1992) and cluster analysis methods from Kohonen (1982). It is designed to detect subtle changes in the relative abundance of multiple spectral components and is most effective in situations where data streams derived from two or more different states lack dominant classifiers. The sensitivity of this portable device combined with the algorithm offers promise for real-time detection, identification, and distinction between spores of closely related bacteria in a potentially field-deployable system.

2. Materials and methods

2.1. Spore preparation

B. subtilis strain SMY, a wild-type, prototrophic, Marburg strain, was grown overnight at 30 °C on a plate of tryptose blood agar base (Difco Laboratories; Franklin Lakes, NJ) and then inoculated into 2 l of DS medium (Fouet and Sonenshein, 1990) in a 6 l Erlenmeyer flask. The flask was incubated with shaking (200 rpm) at 37 °C for 48 h. The cells were harvested by centrifugation at 13,000 × *g* for 20 min at 4 °C, washed four times with 100 ml sterile, deionized water, and resuspended in 20 ml sterile water. The suspension was estimated to contain 95% mature, refractile spores by phase contrast microscopy. The spore titer was determined by assaying colony formation on DS agar plates after heating to 80 °C for 10 min. Spores were diluted in sterile water when lower concentrations were required for testing. *B. cereus* strain CIP5832 and *B. thuringiensis* strain 407 Cry+ (both obtained from D. Lereclus, Institut Pasteur, Paris, France) were grown on DS agar plates for 48 h at 37 °C. The cultures were harvested by flooding the plates with sterile, deionized water and

scraping up the bacterial colonies. After transfer to a centrifuge tube and centrifugation at 13,000 × *g* for 10 min at 4 °C, the spores were washed, resuspended, and titered as above.

2.2. Pyrolysis-FAIMS analysis of *Bacillus* spores

The experimental setup consisted of a CDS Pyroprobe 1000 (CDS Analytical Inc., Oxford, PA) connected to a 0.53 mm i.d. × 0.5 m length deactivated fused silica column (Agilent Technologies, Palo Alto, CA) held at 200 °C. A prototype SDP-1 micromachined differential mobility spectrometer (microDMx™, Sionex Corporation, Waltham, MA) with an ionization source of radioactive nickel (⁶³Ni) was connected to the outlet of the column. Grade 5 nitrogen was used as the carrier gas to sweep the pyrolyzed sample from the pyrolysis chamber into the deactivated fused silica column and carry it into the DMS. The flow was regulated by mass flow controllers (MKS Instruments, Andover, MA), and was set to 30 ml/min for the sample to be carried through the pyrolyzer and column, where it joined a second flow of nitrogen at 300 ml/min for introduction into the DMS. The interface temperature of the pyrolyzer was set at 110 °C.

A slurry of 4 μl of *Bacillus* spores suspended in sterile water was loaded into a quartz tube. The tube was placed in the pyrolysis probe platinum coil, and the probe was then loaded into the pyrolysis unit. The spores were pyrolyzed by increasing the temperature to 650 °C at a rate of 0.01 °C/ms, and holding this temperature for 99.99 s. The DMS was programmed to sweep the compensation voltage through a voltage range from −40 to 10 V every 1.6125 s at 250 steps per scan. A single scan equals 1.6125 s. The RF field was set at 1200 V. The spectra of the pyrolyzed spores corresponding to the detected positive and negative ions were recorded on a laptop computer connected to the DMS unit. The spectra consist of two independent variables (the compensation voltage and the scan number in time) and one dependent variable (the detected ion abundance at each point).

2.3. Data processing

For each of the three species, *B. subtilis*, *B. cereus*, and *B. thuringiensis*, 100 experiments were conducted at three concentrations to yield a total of 900 experiments. The concentrations used were 2.0 × 10⁷ spores/ml (80,000 spores/experiment), 2.5 × 10⁶ spores/ml (10,000 spores/experiment), and 1.25 × 10⁶ spores/ml (5000 spores/experiment). The positive and negative spectra from each run were concatenated and then aligned across all runs so that the pyrolysis event started at exactly the same scan in each file. As the compensation voltage at which an ion elutes can be affected by the moisture content of the sample and the gas flow rate as it passes through the DMS (Miller et al., 2001; Krylova et al., 2003) the data were further aligned in the Vc-dimension by a rigid shift of a few pixels when necessary. The amount of shift was determined by comparison of the total abundances at each Vc value (across all scans) of a data file with these total abundances from a single reference file. The cross-correlation of the data and reference files was calculated to determine optimal alignment, based on the location at which this value was at a maximum. The positive and negative data are then rigidly shifted in the Vc direction based on this result. The data were then analyzed using three techniques: principal component analysis (PCA), decision tree analysis, and the ProteomeQuest[®] genetic algorithm—lead cluster analysis. First, PCA was performed using the princomp function in MATLAB Statistics Toolbox (The Mathworks Inc., Natick, MA) software version 7.0.4.365 Release 14 Service Pack 2. As the data files are quite large when including both positive and negative spectra, we first had to compress the information in order to run the princomp program successfully. We did this by summing the data in time to obtain a 1 × Vc-length (1 × 250) vector using only the positive spectra. Next, we performed decision tree analysis on the same compressed data using the treefit function, also in the MATLAB Statistics Toolbox. The entire uncompressed data were finally analyzed by ProteomeQuest[®] (Correlogic Systems Inc.) (Hitt, 2005).

ProteomeQuest[®] combines elements from genetic algorithms first described by Holland (1992) and cluster analysis methods from Kohonen (1982). These genetic algorithms function in a manner similar to natural selection. The input data for analysis are ASCII files of spectra consisting of a first column of index values, referred to as “features” and a second column of the associated amplitudes. The output of the algorithm is an *N*-dimensional

centroid map that represents the most fit subset of amplitudes at N defined features that best segregates the preliminary data. Each centroid in the map is assigned a specific state, for instance *B. thuringiensis*, *B. subtilis* or *B. cereus*, and is surrounded by a decision boundary. This bounded centroid is called a node. A model is the combination of a specific set of N features and their nodes.

Data analysis by ProteomeQuest[®] is divided into two phases: phase I in which the computer is aware of the identity of the spectra, and phase II where no information about the identity of the spectra is provided.

In phase I the available, known spectra, representing the two or more states to be classified, are divided into two sets of data – the training and testing sets – and compared. The algorithm uses an iterative search to identify a small subset of key features in the training set that completely segregate the spectra of each state. To do this, the software starts by creating hundreds of small sets of individual features, selected at random from the training set spectra. Each candidate set contains N features, where N typically varies from 3 to 20. The fitness test consists of plotting the pattern formed by the combined amplitudes of the N candidate features in N -dimensional space. The pattern formed by the relative amplitude of the spectrum data for this set of chosen values is then rated for its ability to distinguish the two preliminary populations in the testing set. The features within the highest rated sets are reshuffled by the genetic algorithm, to form new subset candidates and the resultant amplitudes are rated iteratively until the set that fully discriminates the preliminary set emerges. At the end of this process a number of different models are generated, each with a unique set of N features, nodes and decision boundaries. The training accuracy is the accuracy of a given model at the end of this process on the training set samples used in model building.

In phase II, species whose identity is not known to the computer are classified. To do this, the amplitudes of the relevant N features are extracted from the spectral file of interest and mapped in N -dimensions. If this point falls within a decision boundary, the species is classified according to the previously determined identity of that node. The validation accuracy of a given model is its performance on this second set of samples, which were not used in model development.

3. Results

3.1. Principal components analysis

One hundred pyrolysis-DMS experiments were conducted for each *B. subtilis*, *B. cereus*, and *B. thuringiensis* spore species at three concentrations of 80,000 (80k), 10,000 (10k), and 5000 (5k) spores, after method development to determine the optimal conditions for biomarker release (Krebs et al., 2005). Principal component analysis was performed with the 80k data from all three species, using a single vector from the positive spectra that was created by summing all abundances measured across time in each sample. These vectors served as the input for the principal component analysis, and the result is shown in Fig. 1.

3.2. Decision tree analysis

The same data used for PCA was next used for decision tree analysis. A tree of 20 nodes was built using all 100 files from each of the three species at the 80k concentration. This tree is very large and has most likely overfit the data. To fit this more appropriately we sought to prune the tree size using cross-validation and then determined the optimal tree size by comparing the error rates. This was done by withholding 10% of the files (10 of each species), building a decision tree with the other 90%, and subsequently calculating the accuracy of the remaining 10%. Cross-validation was calculated at all levels of pruning, and an optimal tree size determined. This was repeated iteratively holding out a different 10% of the files each time.

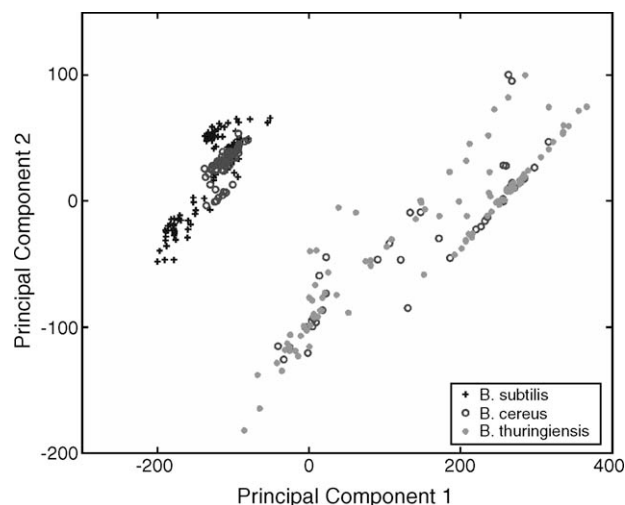


Fig. 1. Results of principal component analysis for 80,000 spores per sample, 100 samples for each species.

The pruning level that seemed to perform the best on average was selected for our original tree. In this case, we found that pruning the tree from 20 nodes to 5 nodes seemed to be optimal. This 5 node tree is shown in Fig. 2. This indicates that there are four points that aid in the separation of the three species when looking at the summed data, occurring at Vc values of -19.12 , -39.80 , -16.71 , and -16.10 . *B. subtilis* and *B. thuringiensis* were separated out completely at the first split, and *B. cereus* proved harder to separate from the two. The estimated probabilities of each class at each node are shown in the figure, and the overall accuracy of the tree for all three species is 86.9%.

3.3. Modeling the data using cluster mapping and genetic algorithms

For the next analysis technique, the data from each species was randomly divided into three categories: a training set (50 spectra of each species), a testing set (150 spectra of each species), and a validation set (100 spectra of each species). The training and testing sets consisted of files whose species identities were known by the computer and were used in the

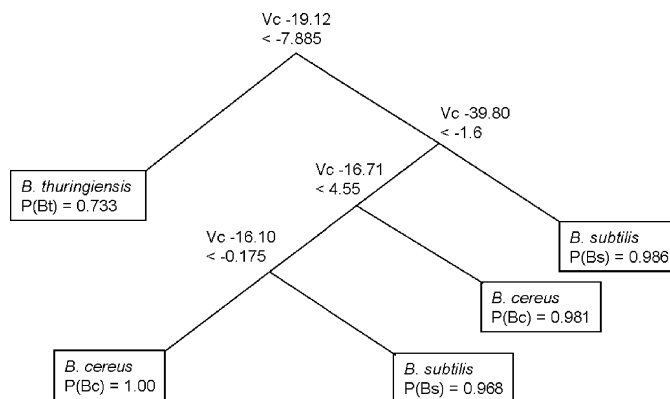


Fig. 2. Decision tree analysis using data from 80,000 spores per sample, 100 samples for each species.

model building phase (phase I) of ProteomeQuest[®]. The software was set up to generate 40 models containing from 5 to 12 features with decision boundaries varying from large (setting 0.5) to very tight (setting 0.9) around the centroid. The validation set of data, which was withheld from the iterative model building process, was then analyzed by each model to give an independent measure of the accuracy of species prediction across all three concentrations. The specificity, sensitivity, and accuracy described below were calculated from the results of the independent validation set using the following equations:

- Sensitivity = (True positives)/(True positives) + (False negatives)
- Specificity = (True negatives)/(True negatives) + (False positives)
- Accuracy = (True positives + True negatives)/(Total number of samples)

The first set of models generated were binary models comparing one bacterial species against another. For these models, data from all three spore concentrations were used together. The validation results for the six binary models with the highest accuracies are shown in Table 1. *B. subtilis* was readily distinguished from *B. cereus* and also from *B. thuringiensis* even at a level as low as 5000 spores, with accuracies higher than 90%. *B. cereus* and *B. thuringiensis* proved slightly more difficult to distinguish, with accuracies just under 70%. However, this is not surprising, as these two species are genetically very similar. The specificities and

Table 1
Comparisons of validation results based on two-way modeling across all concentrations (80k spores, 10k, and 5k)

Model	B	DB	N	Sn	Sp	A
<i>B. subtilis</i> and <i>B. thuringiensis</i>						
1	8	0.9	21	99.0	98.0	98.5
2	12	0.9	24	93.0	99.0	96.0
3	9	0.9	24	93.0	96.0	94.5
4	10	0.9	16	91.0	96.0	93.5
5	9	0.8	4	89.0	96.0	92.5
6	6	0.9	16	88.0	96.0	92.0
<i>B. cereus</i> and <i>B. subtilis</i>						
1	9	0.9	33	87.9	96.0	92.0
2	2	0.9	36	91.9	91.1	91.5
3	11	0.9	35	90.9	91.1	91.0
4	10	0.9	29	90.9	90.1	90.5
5	7	0.9	13	94.9	86.1	90.5
6	8	0.8	7	96.0	82.2	89.0
<i>B. cereus</i> and <i>B. thuringiensis</i>						
1	6	0.8	5	76.0	62.4	69.17
2	8	0.7	3	72.0	66.3	69.14
3	12	0.7	2	81.0	55.4	68.14
4	10	0.8	10	64.0	71.3	67.67
5	9	0.7	2	70.0	63.4	66.68
6	7	0.7	3	61.0	69.3	65.17

One hundred one *B. cereus*, 99 *B. subtilis*, and 100 *B. thuringiensis* files were used. Data shown for each binary comparison include number of biomarker features (B), decision boundary (DB), number of nodes (N), sensitivity (Sn), specificity (Sp), and percent accuracy (A). Sensitivity and specificity are calculated with respect to the first species named in each comparison.

sensitivities for each model are also reported in this table. For example, for the model with the highest accuracy (92.0%) in the comparison of *B. cereus* and *B. subtilis*, the sensitivity and specificity for the files of each species used in validation were 87.9% and 96%, respectively, as calculated with respect to *B. cereus*. This means that for the 101 *B. cereus* files in the blind testing, 89 were classified as *B. cereus* and the remaining 12 were classified as *B. subtilis*, whereas of the 99 *B. subtilis* files, 95 were classified correctly while 4 were classified as *B. cereus*.

The biomarker features found across 40 independent models are displayed in Fig. 3. Panel (a) shows the biomarker features found in 40 models that allowed discrimination of *B. subtilis* and *B. thuringiensis*. Note that there is one biomarker feature that was selected in many of the models, which suggests that it has a classification value that is important for the discrimination of these two species. Panel (b) shows a similar plot for *B. subtilis* and *B. cereus*, and again we see that the same biomarker feature appears in many of these models as well. When comparing the models of *B. cereus* and *B. thuringiensis* (Fig. 3c), no biomarkers appear as frequently across all models, which is consistent with these two species being difficult to separate. To further examine the one biomarker feature that appears to be important in distinguishing *B. subtilis* from the other two species, we graphed the abundance value at that marker location in the raw data for *B. subtilis* and *B. thuringiensis* (Fig. 4). Even at the level of 5000 spores as shown, there is a clear trend of separation in the raw data. When the data are normalized to give the same total ion count for each spectrum, an identical plot is obtained.

3.4. Modeling of *B. cereus* and *B. thuringiensis*

To verify that *B. cereus* and *B. thuringiensis* tend to be harder to separate from each other than from *B. subtilis* due to their relatedness, we created 40 binary models to distinguish *B. subtilis* from a pool of *B. cereus* and *B. thuringiensis* files. Again the 5k, 10k and 80k files for each species were combined and randomized prior to modeling. For model building we used 200 files for each species split into training and testing data sets of 50:100 and 150:300 (*B. subtilis*: *B. cereus* and *B. thuringiensis*), respectively. For the independent validation set we kept back an additional 100 spectra for each species. The results for the six models yielding the highest accuracies are shown in Table 2. The classification obtained with these models shows that *B. cereus* and *B. thuringiensis* have biomarker features common to each other but that differ from *B. subtilis*.

As *B. cereus* and *B. thuringiensis* are the most difficult to classify, we modeled these two species at each concentration individually to determine if there is a concentration limit below which the species become indistinguishable. To generate these models, spectra were randomized and assigned into sets of 75 spectra for each species for model building (25 *B. cereus*: 25 *B. thuringiensis* training; 50 *B. cereus*: 50 *B. thuringiensis* testing), and 25 of each spectra for the independent validation set. The models offering the highest classification had 60.8% accuracy at 5k concentration, 64% accuracy at 10k concentration, and 88% accuracy at 80k concentration. Therefore,

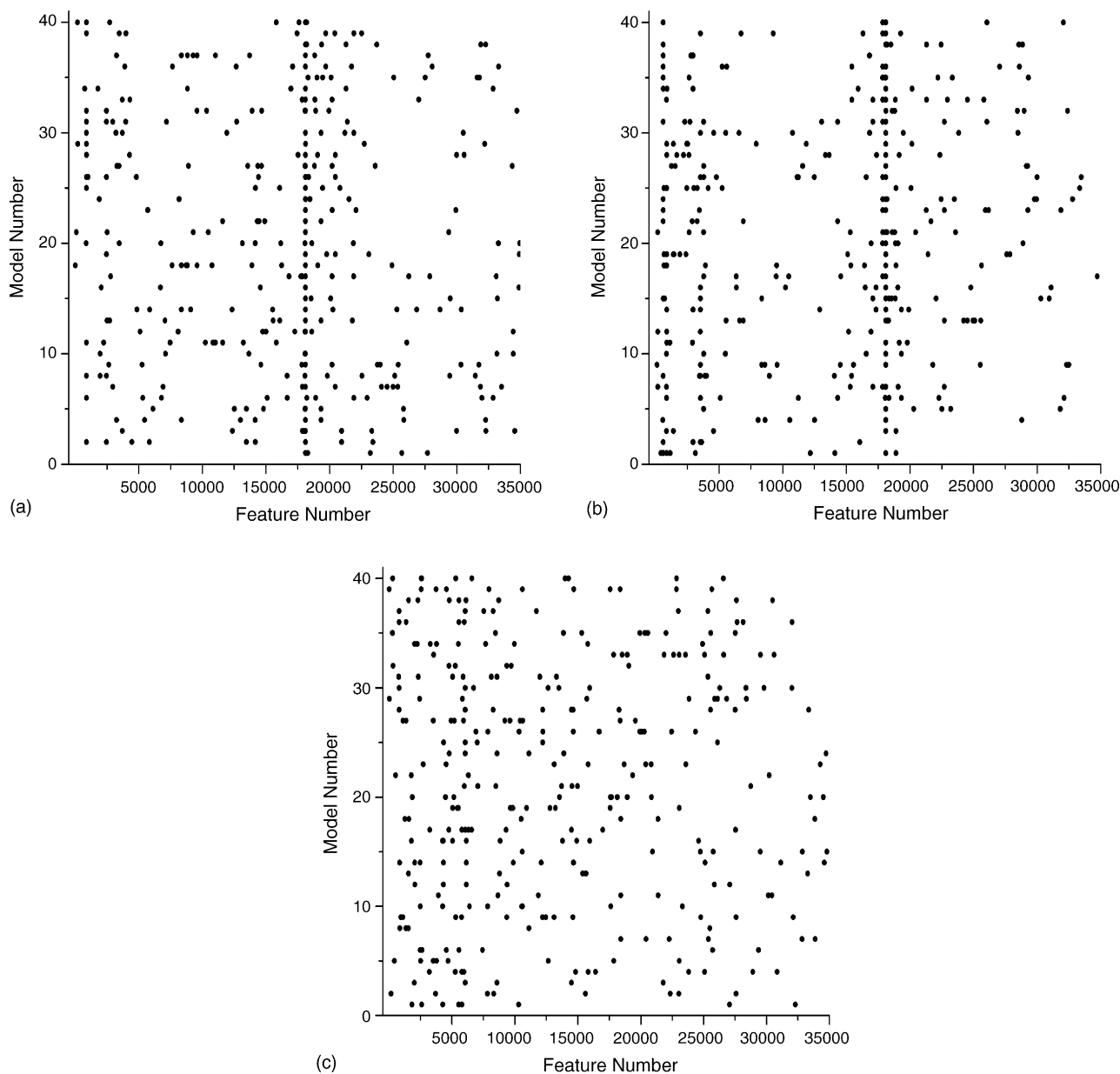


Fig. 3. Distribution of features across 40 models. (a) *B. subtilis* vs. *B. thuringiensis*, (b) *B. subtilis* vs. *B. cereus*, and (c) *B. cereus* vs. *B. thuringiensis*.

Table 2

Comparison of validation results from modeling across the three spore concentrations *B. subtilis* vs. *B. cereus* and *B. thuringiensis* together

Model	B	DB	N	Sn	Sp	A
<i>B. subtilis</i> and (<i>B. cereus</i> and <i>B. thuringiensis</i>)						
1	10	0.9	39	95	86.9	92.3
2	11	0.9	32	95	85.9	92
3	12	0.9	35	93.5	83.8	90.3
4	8	0.9	23	92.5	84.8	90
5	7	0.9	27	93	81.8	89.3
6	11	0.8	8	90.5	84.8	88.7

One hundred files of *B. subtilis* were modeled against 200 files of the other species (100 files each of *B. cereus* and *B. thuringiensis*). Abbreviations as in Table 1. Sensitivity and specificity are calculated with respect to *B. subtilis*.

classification accuracies are higher for these two closely related species when more spores are present.

3.5. Three-way modeling

Next, a set of three-way comparisons were performed to classify all three groups from one another in a single model. For these models only the 80k data were used, since we determined that below that concentration *B. cereus* and *B. thuringiensis* are more difficult to distinguish. For each species the spectra were randomly assigned to a training set of 25, a testing set of 50, and an independent validation set of 25 spectra. The results are shown in Table 3. In Table 3a of 25 *B. thuringiensis* in the validation set, 2 were classified as *B. cereus*, 0 were classified as *B. subtilis* and

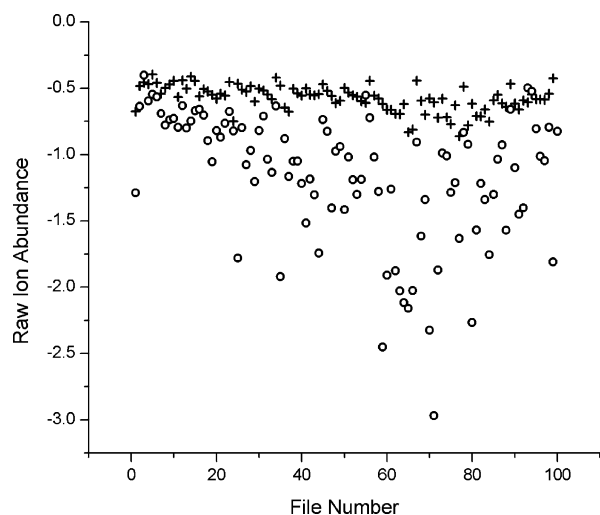


Fig. 4. Plot of the dominant classifier feature 18097. The intensity of this biomarker feature was extracted from each of the 100 raw data files for the 5k concentration of *B. subtilis* (+) and *B. thuringiensis* (O). The data for each species shows a different distribution at this point. The classification algorithm finds data points such as this to aid in decision-making. While one feature alone cannot completely discriminate the two species, the unique combination of features within a model does.

23 were correctly classified, an overall accuracy of 92%. Similarly the accuracy for *B. subtilis* was 88% and for *B. cereus* 52%. An overall accuracy of 77.3% was obtained. The overall accuracy of the second model is 73.3%, and the species accuracies are: *B. subtilis* 68%, *B. thuringiensis* 92%, and *B. cereus* 60%.

Representative spectra from the three species at 5000 spore concentration are shown in Fig. 5. The raw data for the biomarker features resulting from the three-way model (a) are indicated with circles. The data from these experiments look

Table 3
Results of modeling *B. cereus* vs. *B. subtilis* vs. *B. thuringiensis* in a single three-way model

Predicted	Actual			Total classified
	<i>B. cereus</i>	<i>B. subtilis</i>	<i>B. thuringiensis</i>	
Model (a)				
<i>B. cereus</i>	13	3	2	18
<i>B. subtilis</i>	5	22	0	27
<i>B. thuringiensis</i>	7	0	23	30
Total actual files	25	25	25	75
Accuracy (%)	52.0	88.0	92.0	77.3
Model (b)				
<i>B. cereus</i>	15	8	2	25
<i>B. subtilis</i>	5	17	0	22
<i>B. thuringiensis</i>	5	0	23	28
Total actual files	25	25	25	75
Accuracy (%)	60.0	68.0	92.0	73.3

Twenty five files of each species at the 80k concentration were used for validation. Reading down the columns, one can determine how those 25 files were classified. Two models are shown. Model (a) using a decision boundary setting of 0.9, contains 9 features and 22 clusters, with an overall accuracy of 77.3%. Model (b) using a decision boundary setting of 0.8, contains 12 features and 4 clusters, with an overall accuracy of 73.3%.

very similar by eye, but when the data are normalized and expressed as a ratio relative to each other, by ProteomeQuest[®] they are significantly different enough to create a pattern that can reliably distinguish the species from one another.

4. Discussion

There is an increased interest in the development of portable, sensitive, real-time devices for the detection of potential biological weapons agents. One particularly attractive development is the micromachined DMS, a small device that detects ions which are separated by their mobility through an electric field. In previous work we showed that distinct differential mobility spectra can be derived for three chemicals present in high concentrations in spores: dipicolinic acid, picolinic acid, and pyridine (Davis et al., 2003), and we showed that the signal from spores differed from that of the solvent background (Krebs et al., 2005). We now demonstrate the ability to fractionate complex biological samples in a reproducible pattern that contains sufficient information to discriminate between closely related species of *Bacillus* spores. In particular, we have shown the ability to detect and distinguish *B. subtilis*, a spore-forming bacterium commonly found in environmental samples, from *B. cereus* and *B. thuringiensis*, which are closely related to *B. anthracis*, the causative agent of anthrax.

We have shown this using three analysis techniques. We decreased the data resolution in the PCA and decision tree analysis to make the data easier to handle computationally in a readily available commercial software package. At this resolution, PCA showed the ability to separate *B. thuringiensis* and *B. subtilis*, but *B. cereus* proved much more difficult to separate from these two. The decision tree analysis showed similar results in that *B. subtilis* and *B. thuringiensis* were completely separated at the first split, while *B. cereus* was present on both sides of the tree. The decision tree indicated four points in the summed data that were helpful in the separation of the species, and showed an overall accuracy of 86.9% in separating the three species from each other.

The final method of analysis was using a pattern recognition classification tool that couples cluster mapping with genetic algorithms. For this analysis we had the advantage of using the entire data, including both positive and negative spectra at full resolution. The results show an ability to distinguish *B. subtilis* from *B. thuringiensis* at an accuracy of 98.5%, *B. subtilis* from *B. cereus* at an accuracy of 92%, and *B. thuringiensis* and *B. cereus* at an accuracy of 69%. We can also distinguish *B. subtilis* from *B. cereus* and *B. thuringiensis* when the latter two are grouped together, indicating that there are biomarker features present in both *B. cereus* and *B. thuringiensis* that are the same, but different from the more distantly related *B. subtilis*. As *B. subtilis* and *B. thuringiensis* can be more commonly found in the environment, the ability to distinguish these species from *B. anthracis* is very important when designing a detection system to minimize the number of false positives. For this reason, we continue to work towards finding methods to improve the distinction of the species. The biomarker pattern recognition models were created across three concentrations so that marker features present across

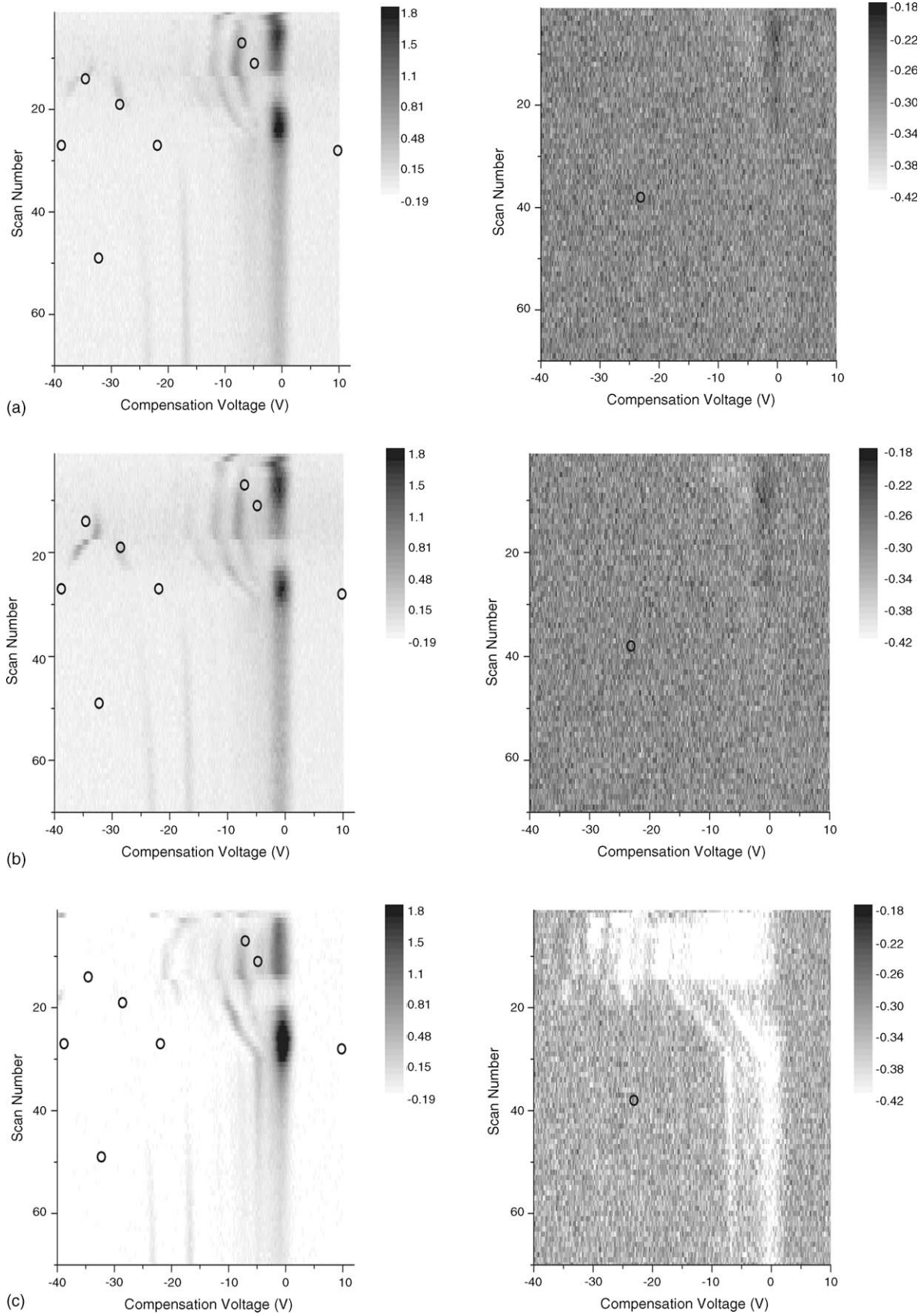


Fig. 5. Representative differential mobility spectra of 80,000 spores undergoing pyrolysis at 650 °C for 99.99 s. Positive ion spectrum left, negative ion spectrum right. X-axis represents V_c (V), Y-axis represents scan number. The nine features from three-way model (a) are circled in black. (a) *B. subtilis*, (b) *B. cereus*, and (c) *B. thuringiensis*.

this entire range could be found. This approach aims to find biomarker features that will not dilute out at the lower concentrations and will not saturate the detector at higher concentrations.

To classify these samples we analyzed the spectra generated by pyrolysis of spores detected with differential mobility spectrometry using ProteomeQuest[®], an algorithm that combines the generation of sets of biomarker features with a genetic algorithm to search for combinations of markers in the spectra which, taken together, can discriminate between the different species with high accuracy. Each resulting set of biomarker features represents a classification model. The six models with the highest accuracies for the binary comparisons are presented in Tables 1 and 2. Each model consists of centroids surrounded by a decision boundary (nodes) into which the data will fall and which are unique to one species or another. Classification of unknown samples is made by examining the spectrum for the unknown at the specific biomarker feature locations, using these to determine into which node it will fall, and thus determining the identity of the species. Different models differ in the number of biomarker features in the spectra, the number of centroids, and the size of the decision boundary around the centroid. While many models of similar accuracy can be generated from the data, depending on the particular combinations of number of features and size of the decision boundary parameter (Tables 1 and 2) selected during model building, models with a tight decision boundary (setting = 0.9) and fewer centroids will be built from biomarker features with the least variance within a species and may represent more robust models. However, the number of nodes can also reflect the number of discrete differences within the spectra of a species and we have seen models with a high number of nodes that prove to be robust across many samples (data not shown). The decision as to which model is best to use becomes clearer as the models are challenged with more and more independent sets of spectra. Within the spectral datasets any features which are strong classifiers will be selected more frequently. With the *Bacillus* species examined there was one dominant classifier, feature 18,097, corresponding to a compensation voltage at -20.92 V and scan 3 (4.8375 s after start of pyrolysis) within the negative ion region, that was selected by most of the 40 models created (Fig. 3), and there were also a number of less dominant ones. The dominant feature appears many times in models distinguishing *B. subtilis* from one of the other two species. Examining this feature across many files of *B. subtilis* and *B. thuringiensis* shows that indeed there is a trend of separation when plotting the raw abundance value at this point.

In addition to the binary comparisons, we have also shown the ability to create a single model that can discriminate between three species (Table 3 and Fig. 5). In this case, three-way modeling is generally less accurate than the binary modeling due to the high similarity of *B. cereus* and *B. thuringiensis* which makes these spores very difficult to discriminate, especially when present in low quantities.

For classification problems in which there are clear dominant markers or a combination of very clear cut expression levels typical of a state, most traditional methods of classification, such as *k*-nearest neighbors, principal component

analysis and decision tree analysis are very powerful. The strength of an algorithm such as ProteomeQuest[®] is when the state differences arise from multiple subtle changes in the relative amplitudes of features. Through the internal normalization schema used within ProteomeQuest[®], subtle relative amplitude differences are amplified. When these differences are then searched by a powerful iterative procedure that combines lead cluster mapping with a genetic algorithm, those subtle, reproducible changes in features that characterize a state, rise above the differences seen within other features caused by natural, random variation. The resulting models reflect the combination of features and the extent of their variation, defined by the decision boundary that separate the states.

The methodology we have demonstrated here should be widely applicable to scientists in many fields that are interested in the sensitive and specific detection of a particular pathogen or disease. In addition to *Bacillus* spores, we envision its application to other spore formers that would be important to monitor including *B. cereus* (a causative agent of food poisoning), *Clostridium botulinum* (botulism), *C. perfringens* (gas gangrene and food poisoning), *C. tetani* (tetanus), *C. sordellii* (diarrheal disease), and *C. difficile* (antibiotic-associated diarrhea and pseudomembranous colitis). The experimental setup offers the potential for even further miniaturization. We are currently developing a small pyrolysis oven that is mounted directly in-line with the DMS to make the entire setup handheld. Compressed air can replace the nitrogen, which would allow for a more fieldable setup. System control from an external computer can also be implemented readily, which would allow many of these units to be monitored from a single location. Finally, using other species it will be possible to build a database of species-specific biomarker features. This database could be remotely updated as emerging threats appear. From the spectrum derived from a single environmental sampling a variety of biological agents might be identified against the database in seconds. It should be noted that so far we have only tested spores in sterile water, without any interferents present. Adding interferents that would be present in any sampling environment will increase the complexity of the analysis and may affect the sensitivity and specificity rates.

The DMS is a portable and reagent-free device that offers rapid, real-time spectral analysis of ionized compounds. We have shown that, in combination with powerful pattern recognition algorithms, it can be applied to environmental sampling for biological agents. Using three different species of *Bacillus* spores, one a common environmental contaminant and two others related to *B. anthracis*, we have demonstrated the ability to distinguish between these species of *Bacillus* spores at levels as low as 5000 spores. This holds promise for the development of sensitive detectors for biological agents.

Acknowledgements

The authors gratefully acknowledge the input of Dr. Michael Callahan and Dr. Jeffrey Gelfand from the Massachusetts General Hospital on this work, and also the support and collaborative efforts of the entire Draper Bioengineering group.

This project was partially sponsored by the Department of the Army, Cooperative Agreement DAMD-17-02-2-0006. The content of this paper does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

References

- Arakawa, E.T., Lavrik, N.V., Datskos, P.G., 2003. *Appl. Optics* 42 (10) 1757–1762.
- Beverly, M.B., Basile, F., Voorhees, K.J., 1996. *Rapid Commun. Mass. Spectrom.* 10, 455–458.
- Brookmeyer, R., Johnson, E., Bollinger, R., 2003. *Proc. Natl. Acad. Sci. U.S.A.* 100 (17) 10129–10132.
- Cieslak, T.J., Eitzen, E.M., 1999. *Emerg. Infect. Dis.* 5 (4) 552–555.
- Cooney, S., 2001. *Nat. Med.* 7 (12) 1265.
- Davis, C.E., Kang, J.M., Dube, C.E., Borenstein, J.T., Nazarov, E.G., Miller, R.A., Zapata, A.M. S., 2003. In: *Proceedings of the 12th International Conference on Transducers Solid-State Sensors Actuators and Microsystems*, vol. 2, June 9–12, 2003, pp. 1233–1237.
- De, B.K., Bragg, S.L., Sanden, G.N., Wilson, K.E., Diem, L.A., Marston, C.K., Hoffmaster, A.R., Barnett, G.A., Weyant, R.S., Abshire, T.G., Ezzell, J.W., Popovic, T., 2002. *Emerg. Infect. Dis.* 8 (10) 1060–1065.
- Eiceman, G.A., Nazarov, E.G., Miller, R.A., Krylov, E., Zapata, A., 2002. *Analyst* 127 (4) 466–471.
- Fergenson, D.P., Pitesky, M.E., Tobias, H.J., Steele, P.T., Czerwieniec, G.A., Russell, S.C., Lebrilla, C.B., Horn, J.M., Coffee, K.R., Srivastava, A., Pillai, S.P., Shih, M.T.P., Hall, H.L., Ramponi, A.J., Chang, J.T., Langlois, R.G., Estacio, P.L., Hadley, R.T., Frank, M., Gard, E.E., 2004. *Anal. Chem.* 76 (2) 373–378.
- Fouet, A., Sonenshein, A.L., 1990. *J. Bacteriol.* 172, 835–844.
- Fox, A., Black, G.E., Fox, K., Rostovtseva, S., 1993. *J. Clin. Microbiol.* 31 (4) 887–894.
- Friedlander, A.M., 1997. In: Zajtcuk, R., Bellamy, R.F. (Eds.), *Textbook of Military Medicine: Medical Aspects of Chemical and Biological Warfare*. Office of the Surgeon General, U.S. Department of the Army, Washington, DC, pp. 467–478.
- Friedlander, A.M., Welkos, S.L., Pitt, M.L., Ezzell, J.W., Worsham, P.L., Rose, K.J., Ivins, B.E., Lowe, J.R., Howe, G.B., Mikesell, P., et al., 1993. *J. Infect. Dis.* 167 (5) 1239–1243.
- Goodacre, R., Shann, B., Gilbert, R.J., Timmins, E.M., McGovern, A.C., Alsberg, B.K., Kell, D.B., Logan, N.A., 2000. *Anal. Chem.* 72 (1) 119–127.
- Helgason, E., Okstad, O.A., Caugant, D.A., Johansen, H.A., Fouet, A., Mock, M., Hegna, I., Kolsto, A.B., 2000. *Appl. Environ. Microbiol.* 66 (6) 2627–2630.
- Higgins, J.A., Ibrahim, M.S., Knauert, F.K., 1999. *Ann. NY Acad. Sci.* 894, 130–148.
- Hitt B.A., 2005. Knowledge Discovery Engine, U.S. Patent and Trademark Office Notice of Allowance, Application number 09/883,196.
- Holland, J., 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA.
- Inglesby, T.V., Henderson, D.A., Bartlett, J.G., Ascher, M.S., Eitzen, E., Friedlander, A.M., Hauer, J., McDade, J., Osterholm, M.T., O'Toole, T., Parker, G., Perl, T.M., Russell, P.K., Tonat, K., 1999. *JAMA* 271 (18) 1735–1963.
- Kohonen, T., 1982. *Biol. Cybern.* 43, 59–69.
- Krebs, M.D., Zapata, A.M., Nazarov, E.G., Miller, R.A., Costa, I.S., Sonenshein, A.L., Davis, C.E., 2005. *IEEE Sens. J.* 5 (4) 696–703.
- Krylova, N.S., Krylov, E., Eiceman, G.A., Stone, J.A., 2003. *J. Phys. Chem. A* 107 (19) 3648–3654.
- Lee, M.A., Brightwell, G., Leslie, D., Bird, H., Hamilton, A., 1999. *J. Appl. Microbiol.* 87 (2) 218–223.
- Longchamp, P., Leighton, T., 1999. *J. Appl. Microbiol.* 87 (2) 246–249.
- Makino, S.I., Cheun, H.I., Watarai, M., Uchida, I., Takeshi, K., 2001. *Lett. Appl. Microbiol.* 33 (3) 237–240.
- McBride, M.T., Masquelier, D., Hindson, B.J., Makarewicz, A.J., Brown, S., Burris, K., Metz, T., Langlois, R.G., Tsang, K.W., Bryan, R., Anderson, D.A., Venkateswaran, K.S., Milanovich, F.P., Colston, B.W., 2003. *Anal. Chem.* 75 (20) 5293–5299.
- Miller, R.A., Eiceman, G.A., Nazarov, E.G., King, A.T., 2000. *Sens. Actuators* 67, 300–306.
- Miller, R.A., Nazarov, E.G., Eiceman, G.A., King, A.T., 2001. *Sens. Actuators* 91, 307–318.
- Patra, G., Sylvestre, P., Ramisse, V., Therasse, J., Guesdon, J.L., 1996. *FEMS Immunol. Med. Microbiol.* 15 (4) 223–231.
- Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A., 2002. *Lancet* 359, 572–577.
- Phillips, A.P., Martin, K.L., 1988. *J. Appl. Bacteriol.* 64 (1) 47–55.
- Phillips, A.P., Martin, K.L., Broster, M.G., 1983. *J. Clin. Microbiol.* 17 (1) 41–47.
- Quinlan, J.J., Foegeding, P.M., 1997. *Appl. Environ. Microbiol.* 63 (2) 482–487.
- Radnedge, L., Agron, P.G., Hill, K.K., Jackson, P.J., Ticknor, L.O., Keim, P., Andersen, G.L., 2003. *Appl. Environ. Microbiol.* 69 (5) 2755–2764.
- Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B., Fraser, C.M., 2003. *Nature* 423 (6935) 81–86.
- Shnayderman, M., Mansfield, B., Yip, P., Clark, H.A., Krebs, M.D., Cohen, S.J., Zeskind, J.E., Ryan, E.T., Dorkin, H.L., Callahan, M.V., Stair, T.O., Gelfand, J.A., Gill, C.J., Hitt, B., Davis, C.E., 2005. *Anal. Chem.* 77 (18) 5930–5937.
- Smith, H., Keppie, J., 1954. *Nature* 173 (4410) 869–870.
- Smith, P.A., MacDonald, S., 2004. *J. Chromatogr. A* 1036, 249–253.
- Uhl, J.R., Bell, C.A., Sloan, L.M., Espy, M.J., Smith, T.F., Rosenblatt, J.E., Cockerill, F.R., 2002. *Mayo Clin. Proc.* 77 (7) 673–680.
- Vasconcelos, D., Barnewall, R., Babin, M., Hunt, R., Estep, J., Nielsen, C., Carnes, R., Carney, J., 2003. *Lab. Invest.* 83 (8) 1201–1209.
- Zhou, B., Wirsching, P., Janda, K.D., 2002. *Proc. Natl. Acad. Sci. U.S.A.* 99 (8) 5241–5246.