Analysis of Hazardous Biological Materials by MALDI Mass Spectrometry
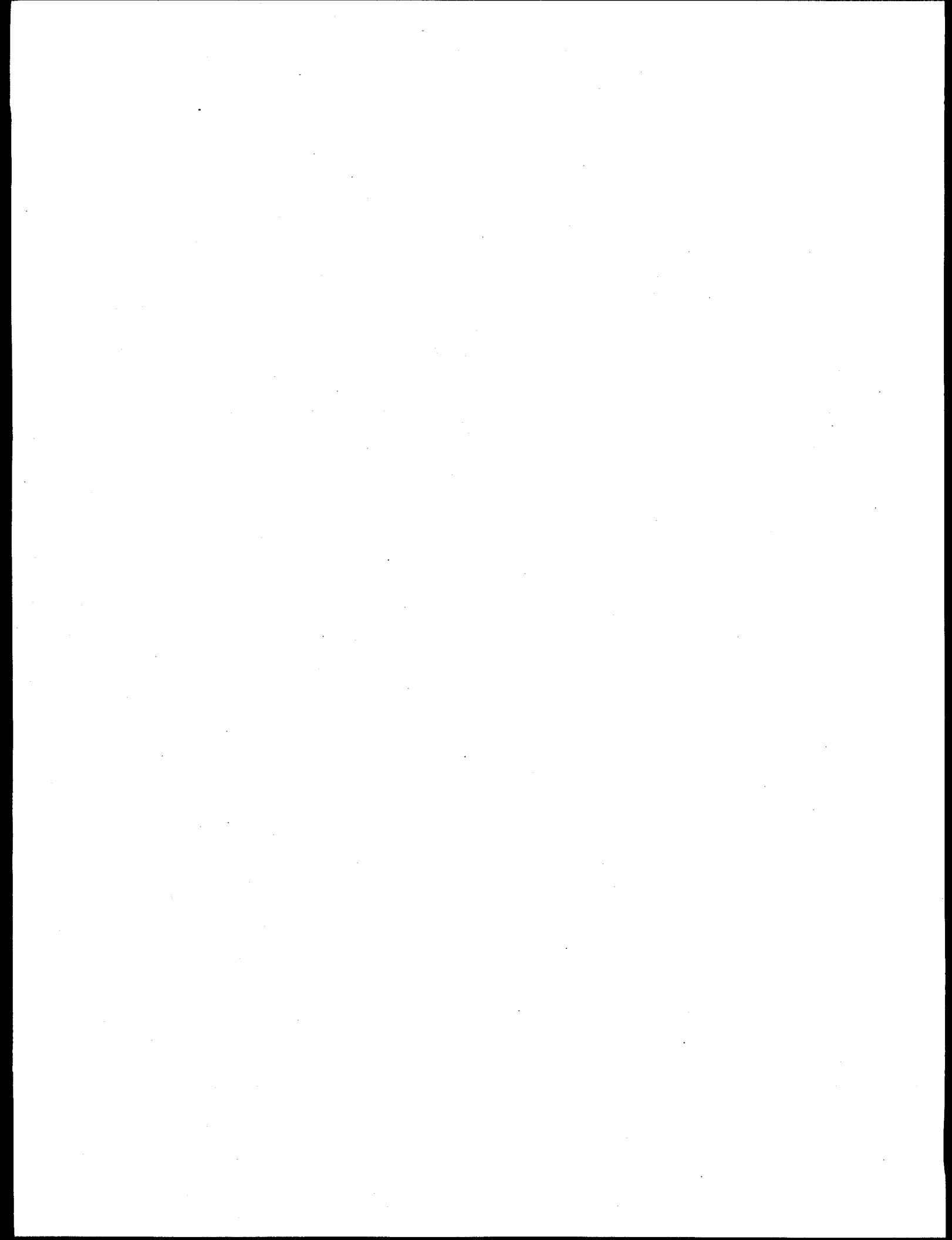
K. L. Wahl
K. H. Jarman
N. B. Valentine
M. T. Kingsley

C. E. Petersen
S. T. Cebula
A. J. Saenz

December 1999

Pacific Northwest National Laboratory
Richland, Washington 99352

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Summary

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) is being developed as a rapid screening tool for pathogen identification. Advantages of MALDI-MS for biomolecule identification include rapid analysis time (<1 min per sample analysis), low sample volume requirements (< 1 μL fluid), and the highly selective and specific nature of mass spectrometric analysis. Bacterial cells are identified by comparing MALDI-MS spectra obtained from cultured bacterial cells and simple microbial mixtures against a library of known MALDI-MS spectral fingerprints of intact bacterial cells.

Objectives of this research include 1) evaluation and standardization of a MALDI-MS analysis protocol, 2) reproducibility assessment of the MALDI-MS spectra, and 3) feasibility demonstration of this approach for bacterial identification in the forensic sciences. Note that initial efforts have not attempted to reach the sensitivity limits of the MALDI-MS technique. In the research presented here, approximately 10^6 cells are deposited onto the MALDI target slide. However, only about 10% of the sample is used to obtain the MALDI-MS spectra. Options to improve sensitivity and optimize this technique for complex samples are provided as recommendations for future work.

A MALDI-MS analysis protocol was evaluated and standardized through a combination of empirical experience and designed experiments. Empirical experience was used to standardize portions of the protocol such as bacterial culture preparation and the technique for spotting the MALDI sample plate. A series of designed experiments was used to evaluate effects of the MALDI matrix and layering technique, and select the matrix producing the most reproducible results. Ferulic acid (FA) was selected as the standard matrix for this research. A single layering technique was initially selected as a sample spotting technique and was later modified slightly to provide stronger MS signal intensities.

Reproducibility of MALDI-MS spectra obtained from replicate bacterial cultures was demonstrated on 10 different days over the course of 3 months. Three sources of variability were isolated in this study: culture-to-culture, operator-to-operator, and replicate-to-replicate. Evaluation was based primarily on the peak locations, peak intensities, and presence or absence of peaks in the MALDI-MS spectra. Based on peak location and intensity of the MALDI-MS spectra, replicate-to-replicate variability was the greatest, and operator-to-operator variability was the smallest. As for the difference in the presence or absence of peaks, the culture and operator had the largest effect.

A method for statistical identification of unknown bacteria was developed. This comparison method for mass spectral fingerprints is fully automated and statistically based, providing objective analysis of unknown samples. Based on extraction of reference fingerprint ions from spectra, this approach lends itself well to real-world applications where samples to be analyzed are likely to be impure. This approach was illustrated using a blind study. In the study, MALDI-MS fingerprints for *Bacillus atrophaeus, Bacillus cereus, Escherichia coli, Pantoea agglomerans*, and *Pseudomonas putida* were collected and form a reference library. The identification of test samples containing one or more reference bacteria, potentially mixed with one species not in the library (*Shewanella alga*), was performed by comparison to the reference
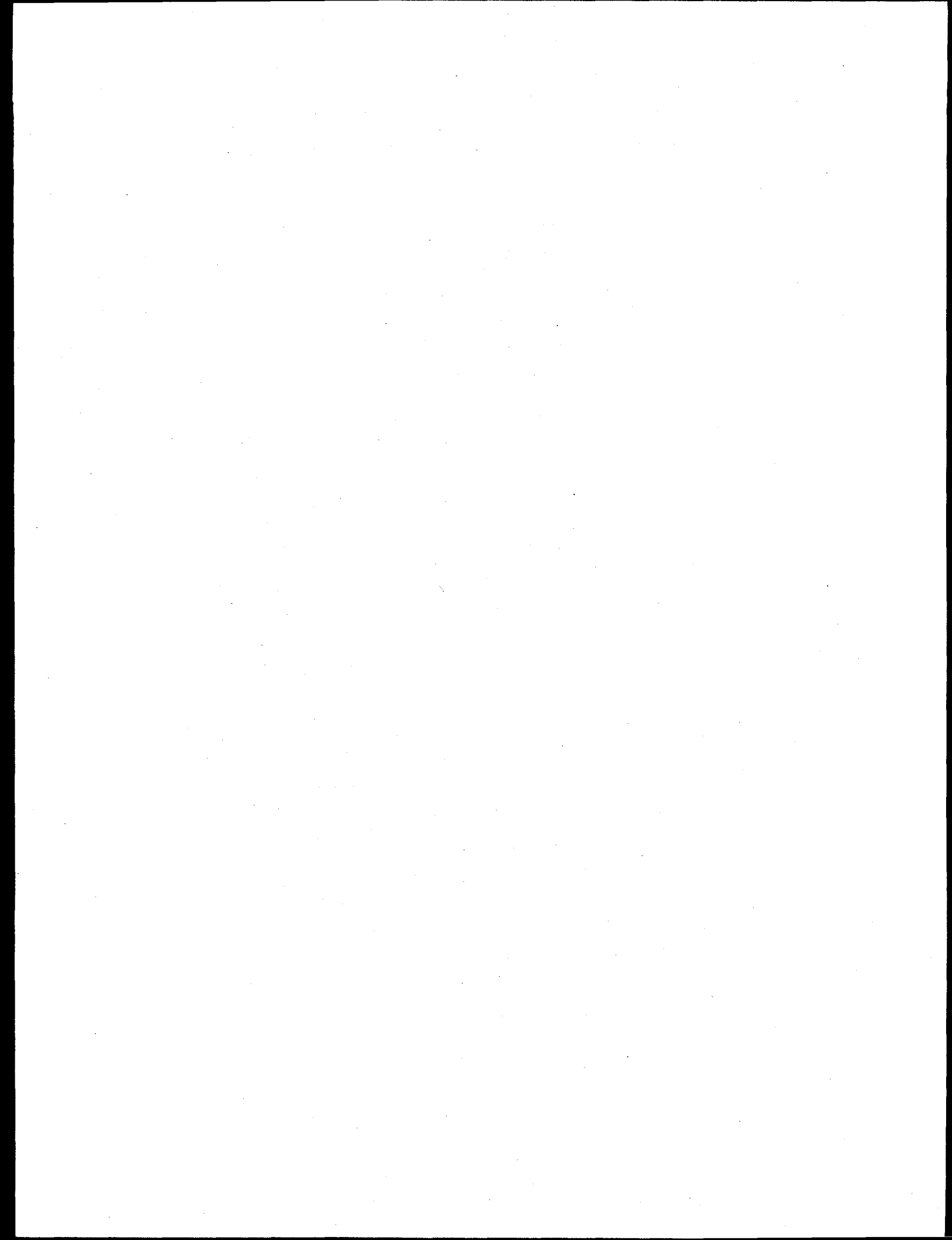
library with a calculated degree of association. Out of 60 samples, no false positives were present, and the correct identification rate was 75%. Missed identifications were largely due to a weak *B. cereus* presence in bacterial mixtures. Modifications for improved performance in the presence of weak signals are discussed, and results suggest these improvements might yield a higher than 90% correct identification rate, with a minimal false-positive rate.

Additional details of this research can be found in the four journal articles submitted during this research effort. References and copies of articles not in press can be found in Appendix A. A web site was prepared to distribute additional information requested due to extreme interest from the mass spectrometry research community in two poster presentations presented at the 47th ASMS Conference on Mass Spectrometry and Allied Topics in Dallas, Texas, June, 1999. The web site address is *http://www.pnl.gov/maldi-tofms*. Links to copies of the extended abstracts and poster presentations are provided as well as other information.

This 2-year research effort culminated in a successful blind-study demonstration of the feasibility of MALDI-MS for bacterial identification. The automated peak detection and statistically-based data analysis and comparison algorithms developed at Pacific Northwest National Laboratory provide a non-biased approach to data interpretation. With cultured samples under laboratory conditions, this combined MALDI-MS and data analysis approach provides the potential for rapid, accurate bacterial identification. In addition, the approach developed under this research was designed to extend beyond the well-controlled laboratory environment. For example, the bacterial identification approach is based on extracting and comparing key biomarkers from an unknown spectrum. Therefore, it should be relatively insensitive to dirty or contaminated samples that might be collected in the field. Additional blind studies should be performed to confirm or refute this hypothesis.

Acknowledgments

Contents

Figures

Tables

1.0 Introduction

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-MS) has become a valuable tool for analyzing microorganisms. The speed with which data can be obtained from MALDI-MS makes this a potentially important tool for biological health hazard monitoring and forensic applications.

The excitement in the mass spectrometry community in this potential field of application is evident by the expanding list of research laboratories pursuing development of MALDI-MS for bacterial identification. Numerous research groups have demonstrated the ability to obtain unique MALDI-MS spectra from intact bacterial cells[1-7] and bacterial cell extracts[8-14]. The ability to differentiate strains of the same species has been investigated[7, 11, 15, 16]. Reproducibility of MALDI-MS spectra from bacterial species under carefully controlled experimental conditions has also been demonstrated[13, 17]. Wang et. al. have reported on inter-laboratory reproducibility of the MALDI-MS analysis of several bacterial species[13]. However, there are still issues that need to be addressed, including the careful control of experimental parameters for reproducible spectra and selection of optimal experimental parameters such as solvent and matrix.

One of the remaining challenges of this approach is the efficient and effective analysis of the data. Most previous work has used qualitative comparisons or tabulations of ions rather than automated statistical techniques. Recently, a method for numerical comparison of MALDI-MS spectra has been developed[7]. This technique, based on the cross correlation between two spectra over the mass range of interest, is effective in comparing spectra under laboratory conditions when the samples to be compared are pure and controlled. Another recent approach is to compare the molecular masses obtained in the mass spectrum from bacterial analysis by MALDI-MS with the information contained in the prokaryotic genome and protein sequence databases available on the world wide web[18]. This approach depends less on reproducibility issues and experimental parameters that may affect the spectral appearance. However, it does require that the organism of interest have information catalogued in these databases.

An alternative approach for automated identification using MALDI-MS has also been developed. Rather than comparing entire spectra, the method focuses on extracting key biomarkers from the spectrum and using those biomarkers to construct MALDI fingerprints and make identifications. The method for constructing MALDI fingerprints is developed in Jarman et. al.[19], where a MALDI fingerprint consists of a collection of estimated peak heights and locations along with their corresponding uncertainties. In addition, the frequency with which each biomarker appears is included in the MALDI-MS fingerprint. In this way, it is acknowledged that biomarkers do not always appear in 100% of the replicates, due to a number of causes such as very small protein concentrations or peaks missed by the peak detection algorithm.

Because it is based on extracting key biomarkers from spectra, this approach can be used to identify bacteria in dirty samples or samples containing more than one analyte. The power of this approach is illustrated using a blind study. In the blind study, test samples containing one or more bacteria from a reference library are identified with MALDI-MS fingerprints for *Bacillus*

atrophaeus, *Bacillus cereus*, *Escherichia coli*, *Pantoea agglomerans*, and *Pseudomonas putida*. In addition, some test samples also contain a species (*Shewanella alga*) not in the reference library to simulate an unknown environmental organism. Results suggest that reliable and accurate bacterial identification is possible with this approach.

Two major goals of this work are to 1) demonstrate reproducibility and standardization of the MALDI-MS technique for bacterial analysis and 2) develop and demonstrate statistical methods to objectively analyze the MALDI-MS spectra. To perform this task, the spectral analysis tools must be able to effectively characterize and account for the variability that exists between replicate bacterial cultures as well as MALDI-MS analyses. Methods for doing this were developed as part of this program and have been published[17, 19]. The outcome is to demonstrate this capability in a blind study where unknowns are correctly identified based on comparison with comprised fingerprints of known bacterial species.

2.0 Background

2.1 Background of MALDI Mass Spectrometry

Mass spectrometry is a powerful analytical tool for providing rapid molecular-weight information for a wide variety of analytes in various sample forms. Since mass spectrometry is a vacuum-based technique, one challenge is to transfer the analyte intact into the gas phase while promoting ionization. Matrix-assisted laser desorption/ionization is an ionization technique ideally suited for rapid molecular-weight screening of large biomolecules, such as proteins present in bacteria.

Typically, MALDI is performed by imparting a pulsed laser beam onto a laser-absorbing matrix material co-dispersed with analyte molecules. A brief overview of the technique is outlined in Figure 2.1. The solution containing the analyte is co-deposited onto a stationary support with an organic "matrix" molecule of a low molecular weight, such as ferulic acid, that absorbs energy at the wavelength of the laser. The matrix acts as a medium for energy transfer
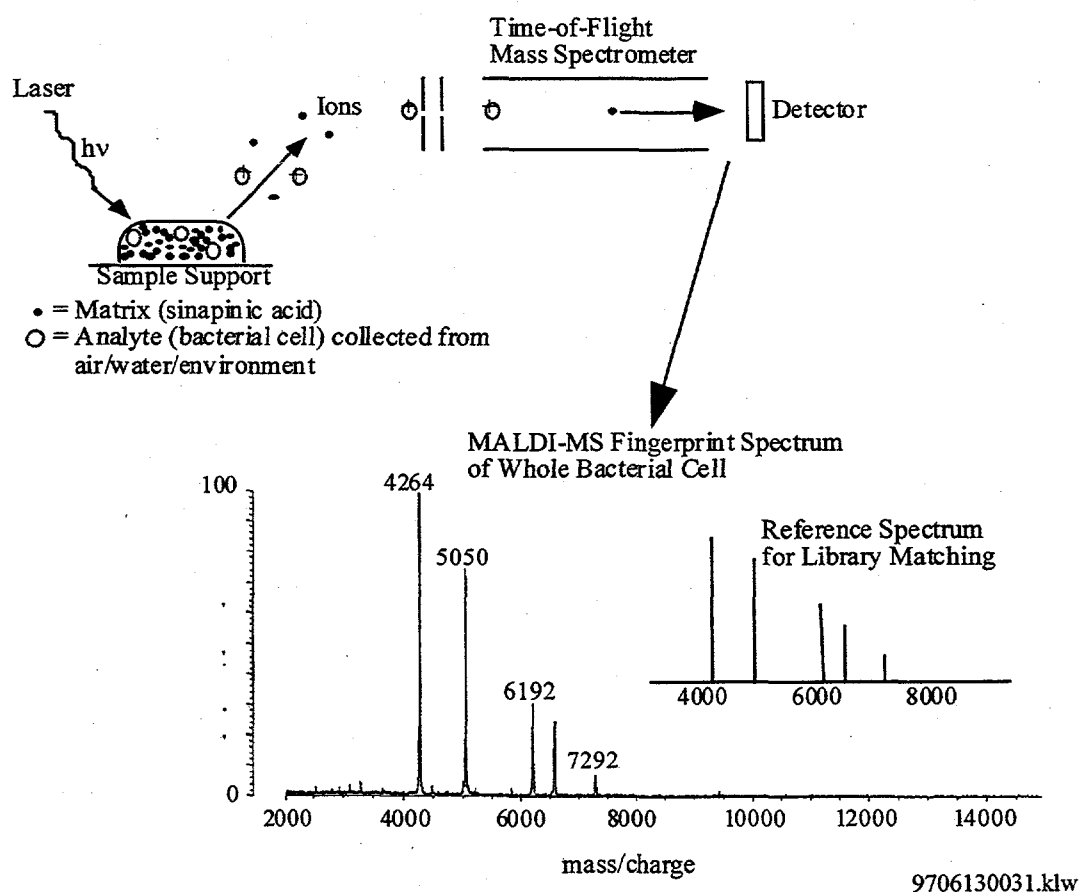
Figure 2.1. Schematic of the MALDI-MS Process

from the laser to the biomolecules of interest, aiding in desorption of the intact biomolecules. The analyte and matrix co-crystallize on the sample support surface before insertion into the vacuum system of the mass spectrometer for analysis. Pulsing the laser onto this crystalline mixture causes desorption and ionization of intact biomolecules in excess of 300,000 daltons. The desorbed ions are then separated in the mass analyzer (e.g., time-of-flight) based on their mass-to-charge ratio (m/z) and sequentially detected by the mass spectrometer. The MALDI-MS analysis of protein standards typically produces an $(M+H)^+$ ion corresponding to the intact protein with no degradation.

In addition to biomolecule analysis, MALDI-MS can provide information on small molecules that may also be of concern in terms of chemical hazards, such as explosives or toxins. Recent work has shown the utility of MALDI-MS for analyzing small molecules such as chelators and small organic acids[20].

3.0 Experimental Methods

One critical component of the MALDI-MS method is to transfer the sample to the MALDI sample plate and prepare the sample spot. This requires appropriate selection of the MALDI matrix and solvent mixture compatible with the analyte of interest and selection of a deposition method for co-crystallization of the matrix and analyte on the MALDI sample plate. It has been shown that the choice of both matrix material and solvent conditions can affect MALDI mass spectra[13, 21-24]. We also observed differences in the MALDI spectra obtained from the same bacterial sample with what was thought to be only subtle changes in the sample spotting procedure [17]. Currently, there is no single method for this deposition process within the mass spectrometry community. Several different methods exist with different laboratories having their method of choice. At least half of the laboratories publishing results of MALDI-MS for microorganism analysis are pre-mixing the bacterial cells with the matrix before spotting onto the MALDI plate. In addition, most of the laboratories are using either sinapinic acid or α-cyano-4-hydrocycinnamic acid (ACHC) as matrices that are traditional matrices for proteins and peptides. In our laboratory, ACHC resulted in increased doubly charged ions increasing the complexity of the spectra and decreasing the sensitivity of the singly charged ions as compared to either sinapinic or ferulic acid. Selection of the matrix compound used in these studies is described in more detail in Section 3.1.

A method to aid in analyzing bacterial samples of unknown concentration by MALDI mass spectrometry was also developed during this project [25]. It is shown that in MALDI analysis of bacteria, the intensities of resulting peaks in spectra are sensitive to the microbial concentration. At the high and low ends of the concentration range, no signal can be obtained, leaving very concentrated or very dilute samples indistinguishable. This is true for MALDI-MS analysis in general where the relative ratio of analyte to matrix is important for successful analysis[25-27]. The addition of cytochrome c as an internal control allows the differentiation of these concentrated and dilute samples. This can be seen in Figure 3.1, giving the results of MALDI-MS analysis of three different concentrations of bacterial cells. It was found that *E. coli* produced good MALDI mass spectra at concentrations between approximately 7×10^5 cells/μL and 2×10^7 cells/μL [25]. Again this is the range of the current technique. Improving the sensitivity is the task of future research.
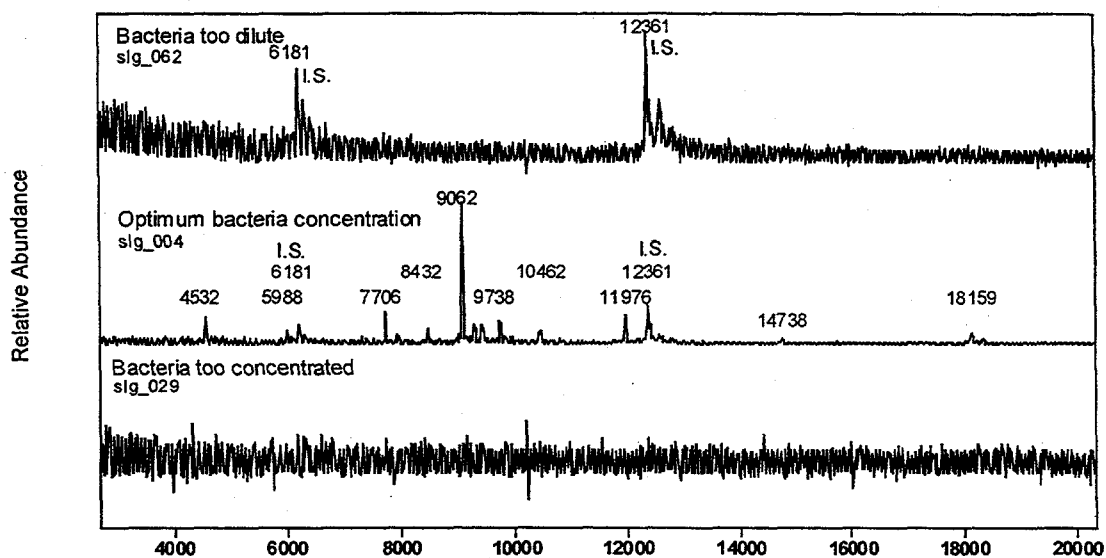
Figure 3.1. Comparison of MALDI-MS Spectra at Three Different *E. coli* Concentrations with Internal Control

3.1 Matrix Spotting Study and Evolving Method

Experiments were conducted to identify a reproducible matrix-spotting combination for MALDI analysis (as outlined in Experiment I.B. of research plan). The first experiment was conducted using a single *E. coli* culture. A parallel experiment was conducted using a protein mix of aprotinin (M_r 6512 Daltons), ubiquitin (M_r 8565 Daltons), and myoglobin (M_r 16951 Daltons) that produces peaks at known locations.

Data were collected on nine matrix-spotting combinations in this experiment. These combinations are listed in Table 3.1.

The data from the MCA-Layer and Pre-coat-Layer combinations were not used in the analysis because there were too few replicate spectra collected (due to difficulties in obtaining reasonable spectra). Peaks throughout the mass/charge range of interest were selected for comparison. Reproducibility was measured by the following criteria: variability in peak location, overall peak height, and frequency of occurrence of key protein peaks. The results for the *E. coli* experiment are summarized. Seven peak locations were selected for the comparison based on the large number of peak occurrences: 664, 787, 843, 1297, 9064, 9740, and 12361. Six peaks appeared with the ferulic acid matrix, but failed to appear with one or more of the other spotting-matrix combinations.

3.2

Table 3.1. Spotting-Matrix Combinations for *E. coli*

	Spotting	
Matrix	FA-Layer	FA-Premix
	SA-Layer	SA-Premix
	MCA-Layer	MCA-Premix
	MCA/SA-Layer	MCA/SA-Premix
	Pre-coat-Layer	

Matrices: FA = ferulic acid; SA = sinapinic acid;
    MCA = 4-methoxycinnamic acid
Pre-coat = commercially prepared sinapinic acid
    coated MALDI plate
Spotting Techniques: Premix matrix and sample
    before spotting
Layer sample droplet; then matrix on MALDI
    plate.

In summary, the ferulic acid layer and premix method both yield fairly small variability in peak location at all locations. The ferulic acid layering method also tends to yield the most intense peaks at all *m/z* values and has the highest presence rate of peaks among the different spotting techniques evaluated. Therefore, based on the results of this matrix and spotting study, the ferulic acid matrix with the simple layering technique was the recommended MALDI analysis protocol. However, in practice, adding a second layer of matrix actually improved the signal strength as well as increased the number of different peaks related to the bacterial cells observed. This method was hence adopted for the blind study and is detailed in the experimental section below. A comparison of the fingerprints obtained with the layering and modified-layering technique will be presented in the fingerprinting section (5.4).

3.2 Supplies

The cultures used in this research include *Bacillus atrophaeus* ATCC 49337, *Bacillus cereus* ATCC 14579[T], *Escherichia coli* ATCC 33694, *Pantoea agglomerans* ATCC 33243 (American Type Culture Collection, Manassas, VA), *Pseudomonas putida* F1[28], and *Shewanella alga* BrY[29]. Bacto Luria-Bertani (LB) Broth, Miller (Difco), Bacto Tryptic Soy Broth w/o Dextrose (Difco), and Bacto Nutrient broth (Difco) were purchased from Fisher Scientific (Pittsburgh, PA, USA). Horse heart cytochrome *c* and angiotensin I were obtained from Sigma (St. Louis, MO, USA). Ferulic acid and trifluoroacetic acid (TFA) were purchased from Aldrich (Gillingham-Dorset, United Kingdom and Milwaukee, WI, USA). Acetonitrile and

ammonium chloride were obtained from J.T. Baker (Phillipsburg, NJ, USA). The water was obtained from a Milli-Q Plus purification system (Millipore Corporation, Bedford, MA, USA).

3.3 Bacterial Growth and Standardization

Bacterial culturing is the general procedure for obtaining bacterial samples for analysis. For standardization and reproducibility, bacteria were cultured in separate tubes, two tubes of 3.5 mL each per organism, and incubated ~15 h in a shaker incubator at the appropriate temperatures. Each culture was divided in half and centrifuged at 14,000 rpm for 2 minutes, then decanted and washed twice with 2% ammonium chloride. Cells were reconstituted in 2% ammonium chloride, and the optical density was measured at 600 nm.

Bacillus atrophaeus, Bacillus cereus, Pantoea agglomerans, and *Shewanella* alga were cultured in TSB and *Pseudomonas putida* in nutrient broth for 15 h at 30°C in a shaker incubator. *Escherichia coli* was cultured in Luria-Bertani broth with streptomycin for 15 h in a 37°C shaker incubator. Each culture was obtained from the same respective stock solutions during the blind study.

Before mass spectrometric analysis, the broth was washed from the cells with 2% ammonium chloride. For example, a 1-mL aliquot of the cells was centrifuged (14,000 rpm) for 2 minutes to form a cell pellet. The supernatant was discarded; 1.0 mL of 2% ammonium chloride was then added to the pellet and resuspended by vortexing. The suspension was pelleted and washed once more. The final pellet was resuspended with 0.2 mL of 2% ammonium chloride, and this suspension was used for MALDI analysis.

Blind study samples containing two (three) different microorganisms were generated by mixing cell suspensions in approximately 1:1 (1:1:1) concentration ratios measured using optical density @ 600 nm. Samples were then coded and delivered to the MALDI-MS laboratory for analysis.

The optical-density measurements at 600 nm can be related back to relative cell concentrations with generated growth curves tracking the bacterial cell counts versus optical density measurements over the entire growth phase. These curves were generated only for *E. coli* and *B. atrophaeus* as shown in Figure 3.2 and Figure 3.3, respectively.
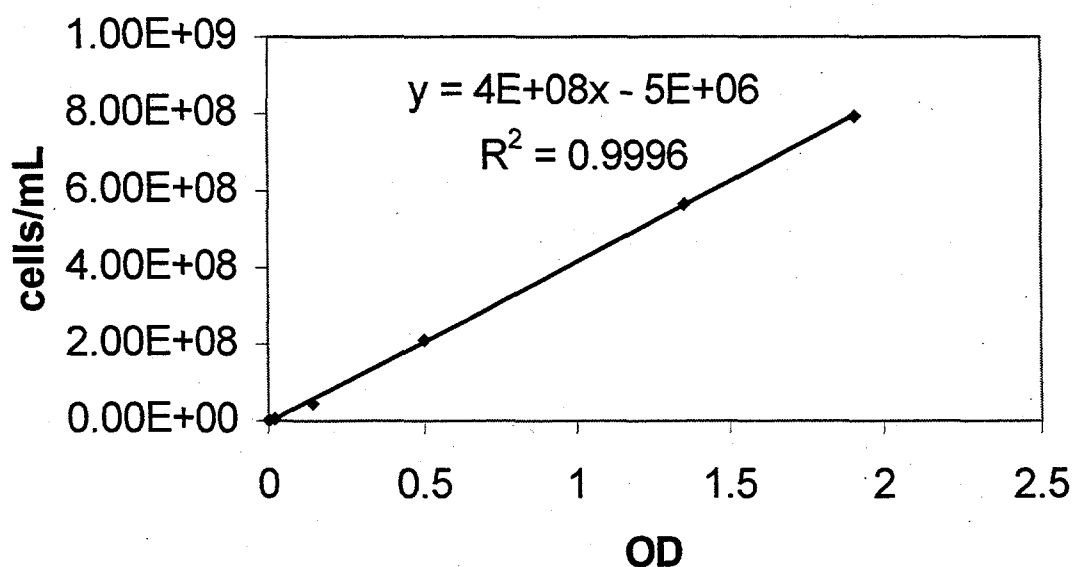
Figure 3.2. Growth Curve for *E. coli*
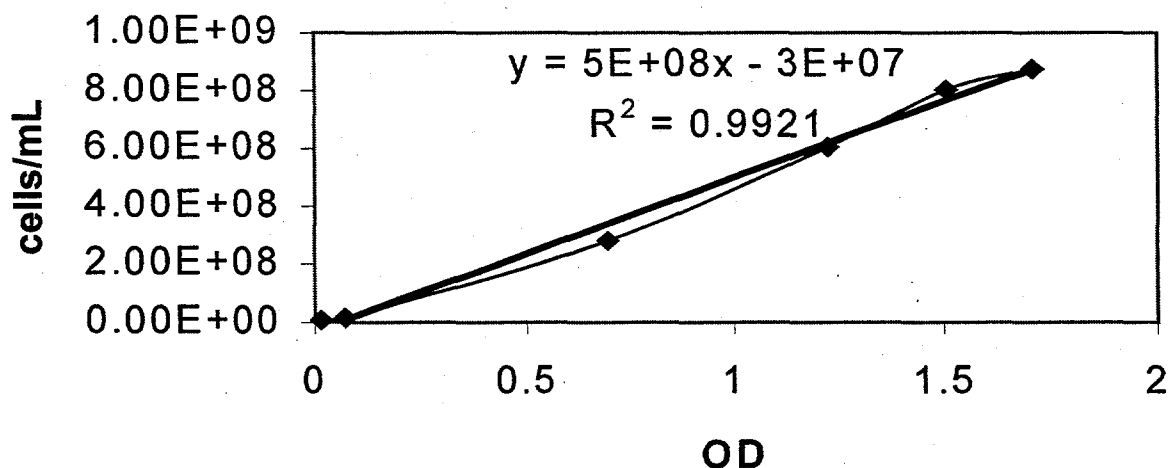
Figure 3.3. Growth Curve for *Bacillus atrophaeus*

3.4 MALDI-TOFMS Analysis

A PerSeptive Biosystems Voyager-DE RP MALDI time-of-flight mass spectrometer with a nitrogen laser (337 nm) operated in the linear, delayed extraction and positive ion mode was used during the experiments. This instrument was purchased by the U.S. Department of Energy as a joint collaboration with the Federal Bureau of Investigation for this project. External calibration with the monomer ion of cytochrome *c* (*m/z* 12 361) and the monomer ion of angiotensin I (*m/z* 1297) was used along with an internal calibration consisting only of the

3.5

monomer ion of cytochrome c. The use of cytochrome c as an internal calibrant and also as an internal control for relative bacterial concentration is detailed in Gantt et al.1999[25]. Each spectrum was obtained by averaging 128 laser shots. The ferulic acid matrix solution was a 10 mg/mL solution in acetonitrile (30%) and 0.1% TFA (70%) along with 5 µg/mL cytochrome c and 2.5 µg/mL angiotensin I. In addition, another ferulic acid matrix solution was made similarly but without the two protein internal standards. A modified layering method was used for the bacterial analysis in which 1 µL of the bacterial sample was applied to the sample plate and allowed to air dry. Then 1 µL of the ferulic acid matrix with internal standards was applied to the bacterial sample spot and allowed to air dry. An additional 1 µL of ferulic acid without the internal standards was applied to the dried sample spot and allowed to dry before analysis. During the analysis, the bacterial samples were stored at room temperature. The operator applied the bacterial samples to the sample plate at approximately the same time and collected replicate spectra. The data files were then transferred to the data analyst for fingerprint construction or blind-study comparison.

To construct a reference fingerprint, the above procedure for MALDI-MS analysis was conducted, where ten replicate spectra were collected for each of 3 days, yielding a total of 60 spectra per bacterium. For the blind study, five replicate spectra were obtained for each of the 60 test samples.

4.0 MALDI-MS Reproducibility

An experiment was conducted to quantify the variability of MALDI-MS replicates. The objective was to make an initial assessment of the impact of random factors on the variability of MALDI-MS measurements. The overall variability due to random factors inherent in MALDI spectra needed to be characterized to help assess reproducibility and design future experiments. There was no attempt to quantify the components of variance or examine interactions between random or random and fixed effects. The probability of peak presence and precision of peak location was quantified.

Reproducibility of MALDI-MS spectra obtained from replicate bacterial cultures was demonstrated on 10 different days over the course of 3 months. Three major sources of variability were included: culture-to-culture, operator-to-operator, and replicate-to-replicate. Evaluation was based primarily on the peak locations, peak intensities, and the presence or absence of peaks in the MALDI-MS spectra. Based on peak location and intensity of the MALDI-MS spectra, replicate-to-replicate variability was the greatest and operator-to-operator variability was the smallest. As for the difference in the presence or absence of peaks, the culture and operator had the largest effect.

For each of 10 days, two replicate cultures of *E. coli* and two replicate cultures of *B. atrophaeus* were grown and analyzed by each of two operators. Each operator followed the same procedure of using ferulic acid, performed his/her own spotting following the layering technique, and performed his/her own mass axis calibration. A single matrix solution was used for each day. Each operator collected 5 replicate spectra from each sample. A chart detailing this information is provided in Table 4.1.

Table 4.1. Design of Reproducibility Study: Sample Preparation and MALDI Analysis for Each Species

Day	Analyte	Operator 1		Operator 2		Total Data Files
		Culture 1 Data Files	Culture 2 Data Files	Culture 1 Data Files	Culture 2 Data Files	
1	*E. coli*	5	5	5	5	20
	B. atrophaeus	5	5	5	5	20
2	*E. coli*	5	5	5	5	20
	B. atrophaeus	5	5	5	5	20
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
10	*E. coli*	5	5	5	5	20
	B. atrophaeus	5	5	5	5	20
	Total	100	100	100	100	400

4.1 Summary of Results

Results of this study indicate that direct MALDI-MS analysis of intact microbial cells is reproducible over an extended period of time using the same sample preparation and analysis technique. The method described led to the detection of bacterial "fingerprints" in the m/z 2000-20,000 range. *Escherichia coli* and *Bacillus atrophaeus* show only minor variations in their spectra over time, which could be due to variations in growth or to handling procedures and the efficiency of solvent extraction of bacterial proteins before analysis. Based on peak location and intensity of the MALDI-MS spectra, replicate-to-replicate variability was the greatest and operator-to-operator variability was the smallest. As for the difference in the presence or absence of peaks, the culture and operator had the largest effect. A plot of the standard deviation of the peak location versus the number of replicate spectra collected, shown in Figure 4.1, was used to help determine the number of replicate spectra needed to account for the majority of the variability expected. From this plot in Figure 4.1, it can be seen that most of the variability in the peak location is accounted for within 60 replicates. Therefore 60 replicates were chosen to generate bacterial fingerprints for the remainder of the research. More details can be found elsewhere[17].
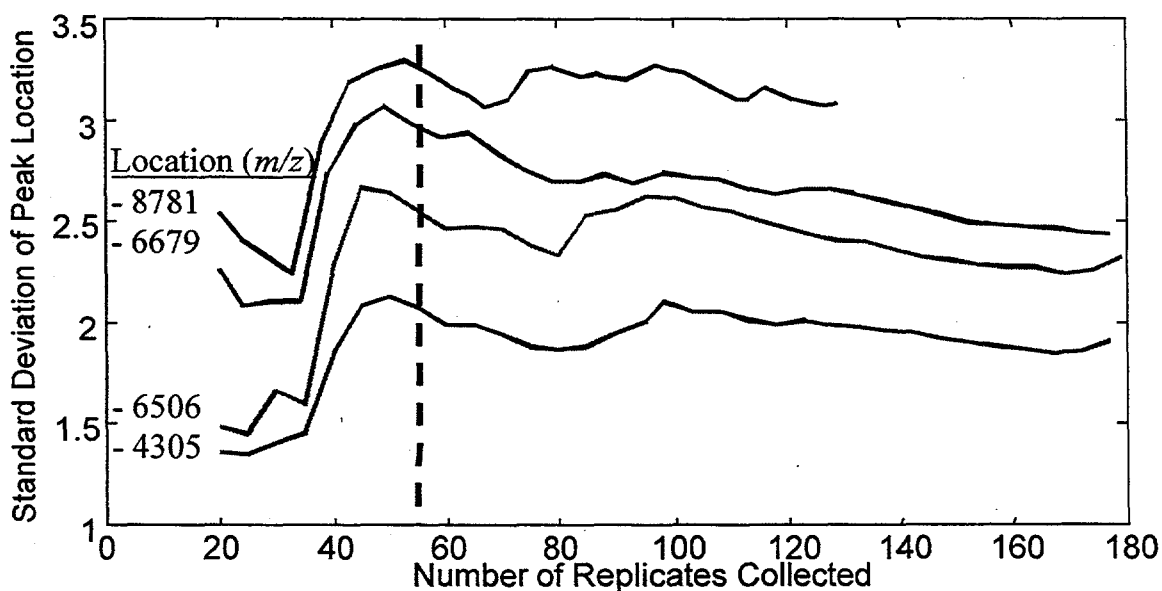

Figure 4.1. Standard Deviation of Peak Location Versus Number of Replicate Spectra

5.0 Development of Statistical Algorithms to Construct, Extract, and Compare MALDI Fingerprints

5.1 Peak Detection/Characterization

Existing mass-spectral peak detection and characterization algorithms are difficult to fully automate because these algorithms most often rely on sets of arbitrary, ad-hoc rules and require significant human interaction. For example, most methods depend upon the absolute signal intensity (peak height) exceeding some threshold set either arbitrarily above the noise level or manually by the user. For rapid bacterial identification using MALDI mass spectrometry, the peak intensities, baseline, and noise level tend to vary widely over the mass range of interest. Consequently, the threshold-based methods usually require significant user interaction where the analyst must set different thresholds for different regions of the spectrum and must often select some of the peaks manually. Due to the level of user involvement, current peak detection is more art than science and cannot easily be automated for widespread use.

Under this research, a novel method for detecting and characterizing spectral peaks has been developed. This method uses the statistical concept of a histogram to locate and characterize peaks. In time-of-flight mass spectrometry, particles are counted over time and converted to counts of particles versus mass/charge ratio. Consequently, we can describe a spectrum as a noisy histogram of mass/charge (m/z) ratios where the signal intensity at a given m/z value represents the relative number of particles counted at that value perturbed by measurement error. When no peak is present in some window of interest, the intensity is relatively constant and resembles the histogram for a discrete uniform distribution (i.e., the number of particles counted is roughly the same for all m/z values in the window). On the other hand, if a peak is present in a window of interest, the intensity profile is concentrated at the center of the window and resembles the histogram for a highly concentrated distribution (such as a normal distribution).

By viewing the spectrum as a histogram of m/z values, an effective method for peak detection and characterization can be developed. In particular, within some window of interest, the variance of m/z values under a uniform distribution (when no peak is present) will always be larger than for a centralized distribution (when a peak is present). Consequently, the variance of m/z values inside some window of interest is compared to the corresponding variance of a discrete uniform distribution. If the variance falls below some critical value, then a peak is detected. This variance estimator, called a *method of moments estimator*, is applied sequentially through a series of overlapping windows that span the spectrum.

Once a peak is identified, the peak's features, such as location, height, width, asymmetry, goodness-of-fit or probability of being a peak, can also be estimated using method-of-moments estimation on the data within the identified window. Of particular interest are 1) the peak location computed using a trimmed average or centroid of m/z values, and (2) the peak height computed using the mode, or maximum signal intensity. The peak detection and characterization procedure is illustrated in Figure 5.1. Results of a comparison between existing peak detection methods and this new approach are provided in Table 5.1.
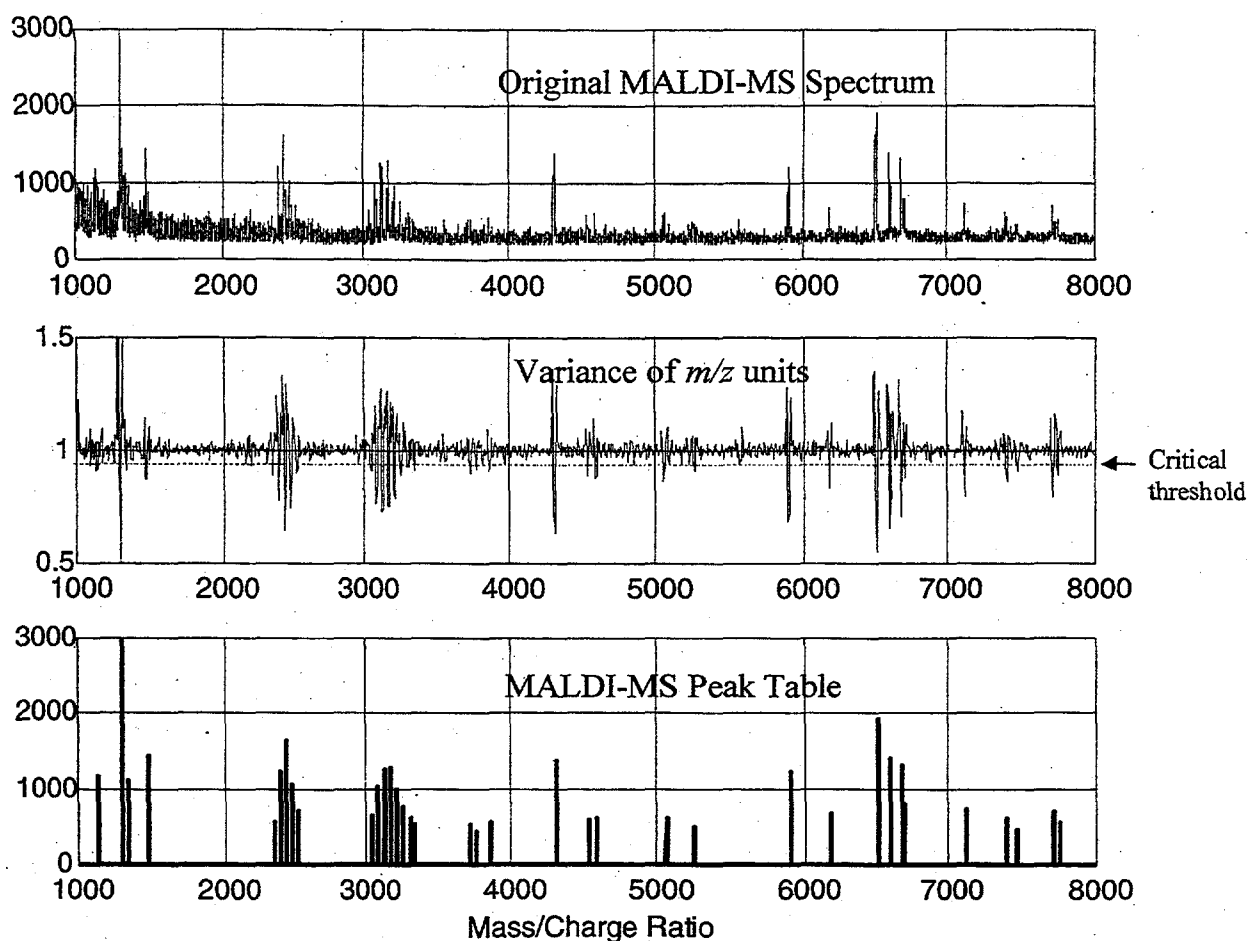
Figure 5.1. Peak Detection for MALDI-MS of *E. coli* and *B. atrophaeus* Mixture

Table 5.1. Number of False and Missed Peaks for Five Replicate MALDI-MS Spectra of *E. coli/B. atrophaeus* Mixture

Method	Percentage of Peaks Missed	Avg. Number of False Peaks per Spectrum
Variance-based	16% (n=31)[a]	0.6 (n=3)
Threshold-based (high threshold)	74% (n=135)	2 (n=10)
Threshold-based (low threshold)	31% (n=56)	5 (n=50)

(a) Number in parenthesis indicates the total number of false or missed peaks over all 5 spectra.

This peak detection and characterization method is objective, consistent, and sensitive to small peaks. It is immune to variations in baseline and fully automated, requiring no human

interaction. As a result, it lends itself well to field instrumentation wherein a user would like to have a system that analyzes a sample and performs the appropriate data analysis in a fully automated fashion.

5.2 Fingerprint Construction

In this research, a method has been developed for quantitative construction and visualization of MALDI fingerprints. Rather than viewing entire spectra, this method focuses on extracting key biomarkers from the spectrum and using those biomarkers to construct MALDI fingerprints. Constructed from replicate spectra, the MALDI fingerprints consist of estimated peak heights and locations along with their corresponding uncertainties. A key benefit of this approach is that it can be used to extract and compare bacterial fingerprints from dirty samples or samples containing more than one analyte.

A reference fingerprint is constructed for a set of replicate spectra for the bacteria of interest. A MALDI fingerprint is currently defined to be the peak location, peak height, uncertainty in location and height, and the frequency of occurrence for each peak i. Only peaks that appear in more than 70% of fingerprint replicates are included in the fingerprint. More specifically, a MALDI fingerprint is defined by $F = \{l_i, s_{li}, h_i, s_{hi}, p_i\}$ where for each peak i, l_i is the average peak location, s_{li} is the standard deviation in peak location, h_i is the average peak height (normalized to the maximum peak height), s_{hi} is the standard deviation of peak height, and p_i is the fraction of replicates in which peak i appears.

This method for extracting and visualizing MALDI-MS fingerprints was submitted, accepted, and published in the journal *Rapid Communications in Mass Spectrometry* in August, 1999[19]. For a complete description of this method, the reader is referred to this article titled "Extracting and Visualizing Matrix-assisted Laser Desorption Ionization Time of Flight Mass Spectral Fingerprints" by Jarman et. al [19].

5.3 Identification of an Unknown

The technique for comparing MALDI fingerprints is based on extracting MALDI-MS peaks from an unknown spectrum and is tailored to the problem of associating an unknown source spectrum with the appropriate reference (Figure 5.2).

Figure 5.2. Illustration of Fingerprint Comparison Method

Comparison of an unknown spectrum to a reference fingerprint takes place in three stages. First, peaks in the unknown spectrum are detected and characterized, and a peak table is generated consisting of peak locations and their corresponding heights. In the second stage, peak locations of the unknown are compared to the peak locations for the fingerprint peaks. Any peak locations for the unknown that fall inside uncertainty regions for a fingerprint peak are extracted. The result is a vector u containing 0s and 1s of length equal to the number of fingerprint peaks. The ith element of u is zero if the ith fingerprint peak was not found in the unknown spectrum, and one if the ith fingerprint peak was found in the unknown spectrum. The vector u indicates the degree of match between the unknown and reference fingerprint. If all the elements of u are 1, then all fingerprint peaks are in the unknown. If all the elements of u are zero, then none of the fingerprint peaks are in the unknown.

All fingerprint peaks do not always appear when the reference bacterium is present. Therefore, the third stage of this algorithm uses the frequency of appearance of each fingerprint

peak and the vector $\underline{u}$ to estimate the significance of the observed fingerprint peaks in the unknown. In particular, consider the following hypotheses:

H_0:   unknown contains species k

H_A: unknown does not contain species k.

Experimental results indicate that all fingerprint peaks are not equally important. Some peaks are very strong and appear in virtually all replicates, while others are much weaker and tend to drop out of some replicates. Under H_0 (the sample contains species k), the probability that the unknown contains reference peak i is p_i where the value p_i is the frequency of appearance of peak i, estimated from the fingerprint. A significance of the fingerprint peaks observed in the unknown is estimated from these probabilities p_i. In particular, we use

P{fewer fingerprint peaks than observed in the unknown | H_0}

to measure whether the difference between the observed fingerprint peaks in the unknown and the reference fingerprint is significant. This represents a likelihood of the MALDI spectrum of an unknown, given species k is present. In this way, our method of comparing an unknown to a reference fingerprint is similar to a statistical test of significance[30], where a likelihood of the observed outcome of the experiment is computed, and the null hypothesis is accepted or rejected based on this likelihood. If this likelihood is large, then the unobserved fingerprint peaks are not important, and we conclude that the sample contains reference species k. If this likelihood is small, then the unobserved fingerprint peaks are important, and we conclude that the sample does not contain reference species k. For a more complete description of this comparison method, the reader is referred to Jarman et al. in Appendix A.

5.4 Variability of MALDI Spectra

As previously mentioned, the MALDI-MS experimental method was evolving during this research. As a result, three different fingerprints for *E. coli* and *B. atrophaeus* were developed on three separate occasions and with three different sample-spotting procedures. The Venn diagram in Figure 5.3 shows the three different fingerprints for *E. coli*. The pre-mixed fingerprint represents 200 spectra and was obtained during the 10-day reproducibility study to initially determine the sources of variability in the overall experiment from bacterial culturing through MALDI-MS analysis. The other two fingerprints, layer and layer+layer, represent 60 replicate spectra obtained over 3 days. The different spotting methods are as follows:

- Pre-mix: matrix with internal standards pre-mixed with bacteria in a 2:1 volume ratio
- Layer: bacteria slurry deposited onto the MALDI plate and air dried; then matrix with internal standards deposited on top of the bacteria spot
- Layer + layer: same as Layer, then additional matrix without internal standards deposited on top of the layered spot.
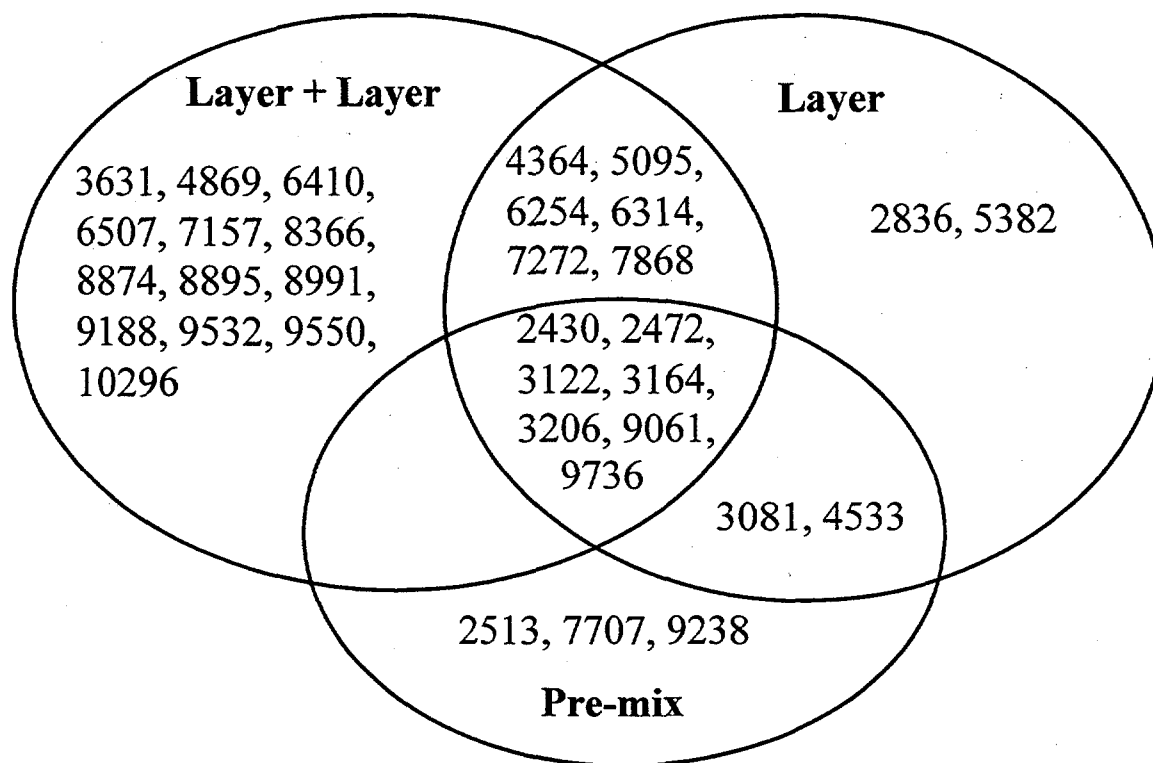
Figure 5.3. Venn Diagram of *E. coli* Fingerprints from Different Spotting Methods

As can be seen in Figure 5.3, there is significant overlap in the majority of the ions observed for the simple pre-mixed and layering method of sample spotting. However, the more recent modified layer-plus-layer method where an additional layer of matrix is added on top of the original layer method generates a significant number of additional ions. Similar results were also observed for the three fingerprints of *B. atrophaeus* as shown in Figure 5.4. Based on the statistical analysis of the replicate data, each single *E. coli* fingerprint seems to have very good reproducibility. However, fingerprints constructed from different data sets look somewhat different. It is unclear at this stage how significant these differences will be and what effect they might have on the overall scheme. Preliminary tests between the two layering methods still produced strong degrees of association with *E. coli* data as desired.
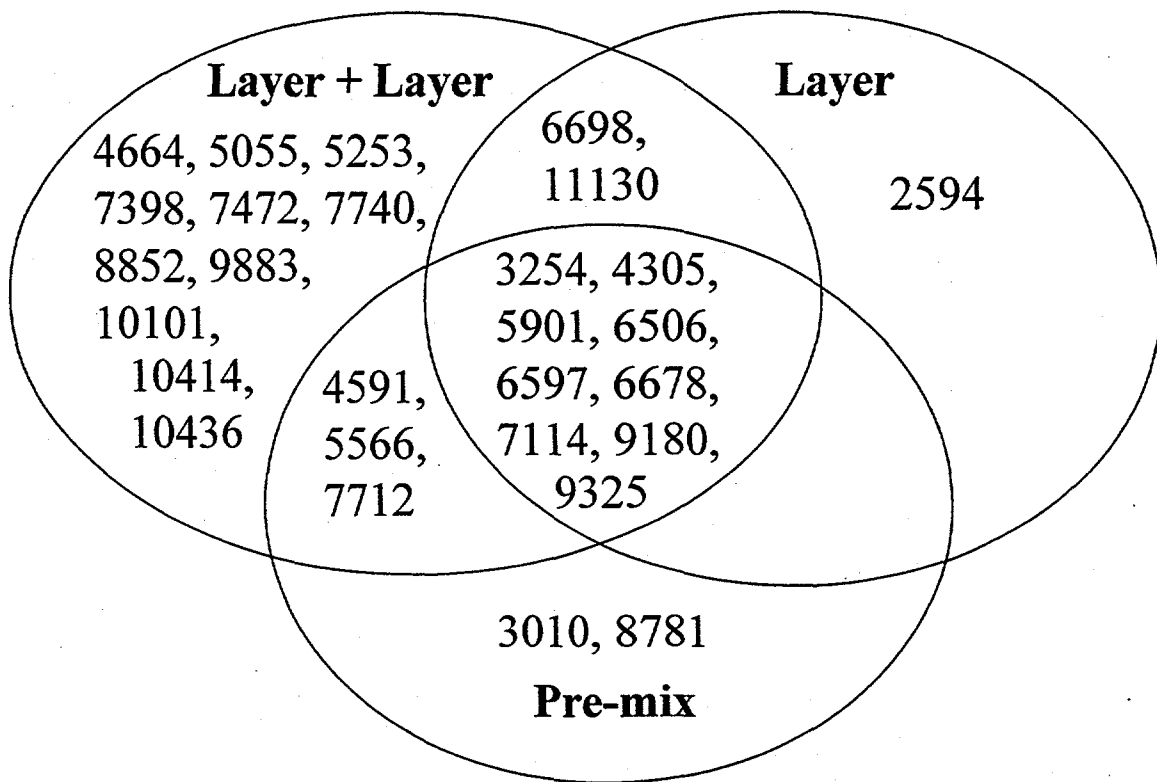
Figure 5.4. Venn Diagram of *B. atrophaeus* Fingerprints from Different Spotting Methods

6.0 Blind Study Proof-of-Principal

6.1 Objective

This study was done to determine if MALDI-MS can be used to differentiate between bacterial species in a blind test situation. Replicate data were generated to produce error rates of identification for single culture samples, mixtures, and unknowns.

6.2 Fingerprint Library Development

The bacteria chosen for an initial proof-of-concept experiment were as follows: *E. coli*, *Pantoea agglomerans* (formerly known as *Erwinia herbicola*), *Pseudomonas putida*, *Bacillus atrophaeus*, and *Bacillus cereus*. This provided candidates from Gram negative and Gram positive, different genera, and two different *Bacillus* species. These five bacterial species were fingerprinted following the protocols established and outlined in the experimental section. Ten MALDI-MS spectra from duplicate samples of each bacteria were obtained on three different culture days. Therefore, 60 spectra of each bacterial species were used to determine the fingerprint. All data collected were directly transferred to the statisticians for automated data processing, including peak detection. The standard deviation of location is usually <2.0 *m/z* units, and the frequency of appearance is usually >95%. A Venn diagram showing the overlap of the fingerprints obtained is shown in Figure 6.1. A table of the fingerprints generated for this study can be found in Jarman et al. in Appendix A.

6.3 Blind Study Experimental Design

6.3.1 Methods

Bacteria were cultured in separate tubes as previously described, washed, and then mixed in ratios that approximate equal concentration based on optical density readings. The samples were labeled with the coded numbers 1-60 as described below and then delivered to the MALDI-MS laboratory for analysis of unknowns. Five spectra were obtained for each of the 60 unknown samples. The data files were transferred to the statisticians for automated data processing.
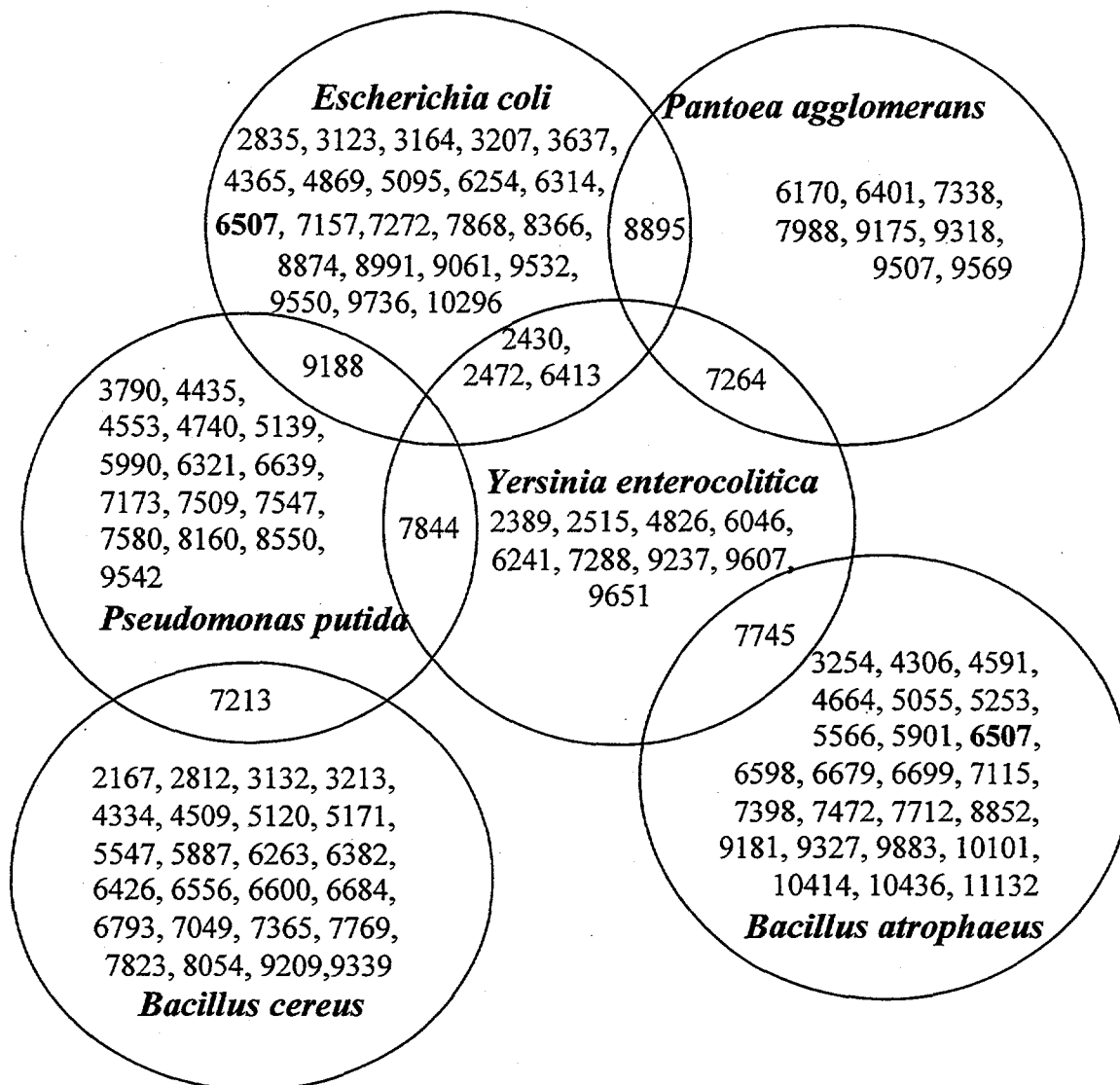
Figure 6.1. Venn Diagram of the MALDI-MS Fingerprints of Bacterial Species Used in the Blind Study

6.3.2 Outline of Design

Combinations of the five fingerprint library species along with *Shewanella alga* (*S. alga*) were analyzed with MALDI-MS. In this study, *S. alga* is not contained in the reference library and serves the role of an unknown environmental microbe. Fifteen bacterial combinations were selected from the set of all possible combinations of one, two, or three of the bacteria to meet the following objectives. Each fingerprint library species appeared alone to evaluate the ability of the algorithm to correctly identify a species in the absence of other bacteria types. *S. alga* appeared alone to test the ability of the algorithm to correctly eliminate every member of the fingerprint

library. Pairs of two bacterial species were selected to identify the strength of the algorithm with samples containing more than one bacterial species. (Namely, three fingerprint library species appeared with *S. alga* to determine if the presence of an unknown environmental microbe would affect the ability to correctly identify a species in the library. Each fingerprint library species also appeared in a pair with at least one other fingerprint library species to assess whether the algorithm could correctly identify both species). Finally, two bacterial combinations were selected containing three bacterial species to investigate further the ability of the algorithm to detect multiple bacterial species in a sample. The specific combinations were selected such that each bacterial species appeared in at least three combinations, or 12 total samples. Bacterial combinations used in this study are listed in Table 6.1.

Table 6.1. Bacteria/Bacteria Combinations

Single Bacteria Type	Test Set Bacteria Type with Bacteria Type not in Test Set	Pair of Bacteria Types in Test Set	Three Bacteria Types (Test Set and Non-test Set)
Bacillus atrophaeus *Bacillus cereus* *E. coli* *Pantoea agglomerans* *Pseudomonas putida* *Shewanella alga (not in test set, "unknown control")*	*Bacillus cereus & Shewanella alga* *E. coli & Shewanella alga* *Pseudomonas putida & Shewanella alga*	*Bacillus atrophaeus & Bacillus cereus* *Bacillus cereus & Pantoea agglomerans* *E. coli & Pantoea agglomerans* *Pantoea agglomerans & Pseudomonas putida*	*Bacillus atrophaeus & E. coli & Pseudomonas putida* *Bacillus cereus & E. coli & Shewanella alga*

6.3.3 Experimental Design

Test samples were collected over a 6-day period. For each bacterial combination in Table 6.1, replicate test samples were prepared on two different days. For replicate samples prepared on the same day, one or more cultures were divided and labeled. For replicate samples prepared on different days, independent bacterial cultures and independent MALDI-MS sample preparations were used. Note that the samples were numbered, and neither the MALDI-MS operators nor the data analyst knew the contents of the samples.

Each bacterial combination was used once in the first 3 days and once in the last 3 days to capture day-to-day variability. Each sample was used to generate five replicate MALDI spectra. These five replicate MALDI spectra were then used to form an unknown source composite spectrum. The unknown source composite spectrum was then compared to each of the library fingerprints using a calculated degree of association.

6.3.4 Results and Discussion

Details of the blind study are presented in a submitted journal article "Fully Automated Bacterial Identification using Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry" provided in Appendix A.

Using the fingerprint comparison technique outlined in Jarman et al. 1999[19], all 60 unknown composite spectra were compared to the fingerprint library consisting of *Eshcerichia coli, Pantoea agglomerans, Pseudomonas putida, Bacillus atrophaeus, and Bacillus cereus*. For each unknown, the five replicate spectra were combined to a single composite spectrum. By comparing a composite of five spectra to each fingerprint, we attempted to minimize the effects of poor-quality spectra due to high noise or low bacterial concentrations. Each comparison was made by computing a *degree of association* (d. a.) between the unknown and every library fingerprint.

The results of the blind study are given in Table 6.2. In the table, the percent of true positives or true negatives is given. The total number of samples used to compute these percentages is given in parentheses. The detailed results containing the degree of association between every unknown and every fingerprint are given in Jarman et al. in Appendix A.

Overall, the contents in 45 out of all 60 samples (75%) are correctly identified. Of the 15 errors made, 13 were caused by a failure to detect *Bacillus cereus* in a mixture of two or more bacterial species. Of the 40 samples not containing *Bacillus cereus*, all bacteria in 38 samples (95%) are correctly identified. One of these errors is a failure to identify *E. coli* in a sample with no other species, and the other error is a failure to identify *B. atrophaeus* in a mixture with *E. coli* and *P. putida*. We note that there were no false positives in this study.

6.3.5 Diagnosing the Errors

The results of the blind study are very encouraging, with the exception of *B. cereus*. Overall, *B. cereus* is correctly detected only 35% of the time. All of these errors occur in samples containing mixtures of 2 or more species. This failure to identify *B. cereus* in mixtures appears to be due to a combination of the fingerprint comparison method and the fact that the *B. cereus* ions are relatively weaker than other species.

6.4

Table 6.2. Summary of Blind Study Results

Species	Percent of Samples Correctly Identified (True Positives)	Percent of Samples Correctly Eliminated (True Negatives)
Bacillus atrophaeus	94% (n=18)	100% (n=42)
Bacillus cereus	35% (n=20)	100% (n=40)
E. coli	95% (n=20)	100% (n=40)
Pantoea agglomerans	100% (n=16)	100% (n=44)
Pseudomonas putida	100% (n=16)	100% (n=44)

Table 6.3 plots the results generated by comparing the blind study samples to the reference fingerprint library using a threshold of 20%, rather than the 50% used in the blind study. This means that only 20% of the ions in the fingerprint for a given bacteria must be present in the unknown sample to consider that these bacteria are present in the unknown. In this case, only three errors are made (5% of all blind study samples), all three of which correspond to a failure to identify *B. cereus*. We note that the algorithm has been modified after analyzing the blind study data, so these results do not reflect a blind comparison. However, they do suggest future improvement for this numerical approach.

Table 6.3. Comparison Using 20% Threshold and Blind Study Data

Species	Percent of Samples Correctly Identified (True Positives)	Percent of Samples Correctly Eliminated (True Negatives)
Bacillus atrophaeus	100% (n=18)	100% (n=42)
Bacillus cereus	85% (n=20)	100% (n=40)
E. coli	100% (n=20)	100% (n=40)
Pantoea agglomerans	100% (n=16)	100% (n=44)
Pseudomonas putida	100% (n=16)	100% (n=44)

7.0 Related Work

An effort was made under internal DOE Laboratory Directed Research and Development (LDRD) funding to determine the uniqueness of this MALDI-MS fingerprinting technique. Series of strains for both *E. coli* and *B. thuringiensis* were analyzed by MALDI-MS to evaluate the ability to distinguish bacterial strains. The *E. coli* analyses were performed before the systematically determined fingerprinting method, and thus sufficient replicates of those strains were not obtained for rigorous comparisons. However, casual observation suggests that strain differentiation will be challenging with this technique. Similar results were obtained with the fingerprinted *B. thuringiensis* strains. Nine different strains were fingerprinted, and close overlap was observed. The current spectral and fingerprint comparison methods fail to characterize significant differences in the strains, even though there is a hint of uniqueness. Table 7.1 contains a compiled listing of the bacterial species fingerprinted to date within our laboratory at Pacific Northwest National Laboratory. The fingerprints were generated following the method outlined previously of obtaining a total of 60 replicate MALDI-MS spectra, collecting 20 per day over 3 replicate culturing days.

Table 7.1. Sixteen Different Bacterial Species and Strains Fingerprinted by the MALDI-MS Method

Bacteria	ATCC Number
Escherichia coli, HB101	33694
Pantoea agglomerans (Erwinia herbicola)	33243
Pseudomona putida	F1[a]
Yersinia enterocolitica	51871
Bacillus atrophaeus	49337
Bacillus cereus	14579
Shewanella alga BrY	(a)
Bacillus thuringiensis	10792
Bacillus thuringiensis	13367
Bacillus thuringiensis	19265
Bacillus thuringiensis	29730
Bacillus thuringiensis	33679
Bacillus thuringiensis	35646
Bacillus thuringiensis subsp *kurstaki*	33680
Bacillus thuringiensis subsp *kurstaki*	35866
Bacillus thuringiensis subsp *israelensis*	39152

(a) Source other than ATCC as indicated in Experimental Supplies section.

8.0 Summary

The feasibility of using MALDI-MS to rapidly identify bacteria was demonstrated based on laboratory-cultured samples and simple bacterial mixtures. Good reproducibility of the MALDI-MS spectra for a single bacterial species was obtained under controlled conditions. Completely automated data extraction as well as comparison and analysis methods were developed for statistically-based defensible data analysis.

The fingerprint extraction technique developed to analyze the MALDI spectra offer many benefits for bacterial analysis as listed below:

- Speed of analysis
- Generic nature of identification (i.e. broad range of targets simultaneously)
- Different detection mechanism from other methods including PCR and antibody-based detection for confirmation
- Information rich data
- Ability to provide classification of unknown or new organism
- Fingerprint-extraction technique works on mixtures/multiple analytes
- Obtaining the degree of association enables a quantitative, statistically-based match

One main advantage of this MALDI-MS technique over many other bacterial analysis methods is the generic capability for bacterial identification or species classification. In other words, a large number of targets can be analyzed for simultaneously and do not require an a priori selection of specific antibody or primer for identification. It may be possible that genetically altered microorganisms can at least be classified with their nearest neighbors in the database and direct further more specific and lengthy testing.

Another main advantage of this technique is the information that is obtained. The resulting m/z values and the patterns with which they are observed can provide very specific and unbiased analysis, as they indicate molecular weights of true components of the sample. These help to provide more defensible identification.

The successful blind study demonstration of this MALDI-MS technique offers great expectations for use in forensic analysis. However, further development and evaluation is needed before acceptance as a routine forensic analysis tool. Improvements in sensitivity and determination of the concentration limits of this application of MALDI-MS are needed. The current application of approximately 10^6 cells to the MALDI spot is more than should be ultimately needed once the method has been optimized for sensitivity. There has been some indication that as few as 5,000 cells may be sufficient for MALDI-MS analysis[31].

Initial application of this technology will most likely be in a laboratory setting for sample screening for biological composition. The ability to prepare samples on-site and then ship the MALDI plate with dried samples to the laboratory or a regional facility for fast turnaround and avoidance of some hazardous shipping requirements will be advantageous. However, the advantages of this technology and the on-going development of a truly portable MALDI-TOFMS

instrument will eventually allow for off-site or field use. It is important to continue to demonstrate the capabilities of this technology and test its limitations.

9.0 References

(1) R. Holland, J. Wilkes, F. Rafii, J. Sutherland, C. Persons, K. Voorhees, J.O. Lay. 1996. *Rapid Commun. Mass Spectrom. 10*, 1227-1232.

(2) T. Krishnamurthy and P. Ross. 1996. *Rapid Commun. Mass Spectrom., 10*, 1992-1996.

(3) M. Claydon, S. Davey, V. Edwards-Jones, and D. Gordon. 1996. *Nature Biotechnology. 14*, 1584-1586.

(4) M. Erhard, H.V. Dohren, and P. Jungblut. 1997. *Nature Biotechnology, 15*, 906-909.

(5) K.J. Welham, M.A. Domin, D. Eoin Scannell, E. Cohen, and David S. Ashton. 1998. *Rapid Commun. Mass Spectrom. 12*, 176-180.

(6) J. Karty, S. Lato, and J. Reilly. 1998. *Rapid Commun. Mass Spectrom. 1998, 12*, 625-629.

(7) R. Arnold and J. Reilly. 1998. *J. Rapid Commun. Mass Spectrom. 12*, 630-636.

(8) T. Cain, D. Lubman, and W. Weber. 1994. *Rapid Commun. Mass Spectrom. 8*, 1026-1030.

(9) T. Krishnamurthy, P. Ross, and U. Rajamani. 1996. *Rapid Commun. Mass Spectrom. 10*, 883-888.

(10) B. Chong, D. Wall, D. Lubman, and S. Flynn. 1997. *Rapid Commun. Mass Spectrom. 11*, 1900-1908.

(11) A.M. Haag, S.N. Taylor, K.H. Johnston, and R.B. Cole. 1998. *J. Mass Spectrom. 33*, 750-756.

(12) J.v. Adrichem, K. Bornsen, H. Conzelmann, M. Gass, H. Eppenberger, G. Kreshbach, M. Ehrat, and C. Leist. 1998. *Anal. Chem. 70*, 923-930.

(13) Z. Wang, L. Russon, L. Li, D. Roser, and S. Long. 1998. *Rapid Commun. Mass Spectrom. 1998, 12*, 456-464.

(14) M.L. Easterling, C.M. Colangelo, R.A. Scott, and I. J. Amster. 1998. *Anal. Chem. 1998, 70*, 2704-2709.

(15) W.F. Haddon, G. Full, R.E. Mandrell, M.R. Wachtel, A.H. Bates, and L.A. Harden. 1998. Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL, p. 177.

(16) C.L. Nilsson. 1999. *Rapid Commun. Mass Spectrom. 13*, 1067-1071.

(17) A.J. Saenz, C.E. Petersen, N.B. Valentine, S.L. Gantt, K.H. Jarman, M.T. Kingsley, and K.L. Wahl. 1999. *Rapid Commun. Mass Spectrom. 13*, 1580-1585.

(18) P.A. Demirev, Y.-P. Ho, V. Ryzhov, and C. Fenselau. 1999. *Anal. Chem. 71*, 2732-2738.

(19) K.H. Jarman, D.S. Daly, C.E. Petersen, A.J. Saenz, N.B. Valentine, K.L. Wahl. 1999. *Rapid Commun. Mass Spectrom. 13*, 1586-1594.

(20) S.C. Goheen, K.L. Wahl, J.A. Campbell, and W.P. Hess. 1997. *J. Mass Spectrom. 32*, 820-828.

(21) R. Beavis and B. Chait. 1990. *Proc. Natl. Acad. Sci. 87*, 6873-6877.

(22) K. Bornsen, M. Gass, G. Bruin, J.v. Adrichem, M. Biro, G. Kresbach, and M. Ehrat. 1997. *Rapid Commun. Mass Spectrom. 11*, 603-609.

(23) S. Cohen and B. Chait. 1996. *Anal. Chem. 68*, 31-37.

(24) D.W. Ashton, M.a. Domin, and K.J. Welham. 1999. *Rapid Commun. Mass Spectrom. 13*, 222-226.

(25) S.L. Gantt, N.B. Valentine, A.J. Saenz, M.T. Kingsley, and K.L. Wahl. 1999. *J. Amer. Soc. Mass Spectrom. 10*, 1131-1137.

(26) K.K. Murray, L.M. Preston, and D.H. Russell. 1993. *Biol. Mass Spectrom. 22*, 544-550.

(27) W. Wilkinson, A. Gusev, A. Proctor, M. Houalla, and D. Hercules. 1997. *Fresenius J. Anal. Chem. 357*, 241-248.

(28) G.J. Zylstra, W.R. McCombie, D.T. Gibson, and B.A. Finette. 1988. *Appl. Environ. Microbiol. 54*, 1498-1503.

(29) F.J. Caccavo, R.P. Blakemore, and D.R. Lovely. 1992. *Appl. Environ. Microbiol. 58*, 3211-3216.

(30) P. Bickel and K. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics.*; Prentice-Hall.

(31) J. Bundy and C. Fenselau. 1999. *Anal. Chem. 71*, 1460-1463.

Appendix A
Journal Articles Resulting from this Research Project

Appendix A
Journal Articles Resulting from this Research Project

Gantt, S.L., N. B. Valentine, A.J. Saenz, M.T. Kingsley, and K.L. Wahl. 1999. Use of an internal control for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis of bacteria. *J. Amer. Soc. Mass Spectrom.* 10, 1131-1137.

Jarman, K.H., D.S. Daly, C.E. Petersen, A.J. Saenz, N.B. Valentine, and K.L. Wahl. 1999. "Extracting and visualizing matrix-assisted laser desorption/ionization time-of-flight mass spectral fingerprints". *Rapid Commun. Mass Spectrom.* 13, 1586-1594.

Saenz, A.J., C.E. Petersen, N.B. Valentine, S.L. Gantt, K.H. Jarman, M.T. Kingsley, and K.L. Wahl. 1999. "Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis", *Rapid Commun. Mass Spectrom.* 13, 1580-1585.

NOTE: The following article is not yet publicly available, but is being included in this appendix.

Jarman, K.H., S.T. Cebula, A.J. Saenz, C.E. Petersen, N.B. Valentine, M.T. Kingsley, K.L. Wahl. 1999. "Fully automated bacterial identification using matrix-assisted laser desorption/ionization mass spectrometry". *Anal. Chem.* In Press.

An Algorithm for Automated Bacterial Identification using Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry

Kristin H. Jarman*, Sharon T. Cebula, Adam J. Saenz, Catherine E. Petersen,
Nancy B. Valentine, Mark T. Kingsley, Karen L. Wahl

Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352

ABSTRACT

An algorithm for bacterial identification using matrix-assisted laser desorption/ionization (MALDI) mass spectrometry is being developed. This mass spectral fingerprint comparison algorithm is fully automated and statistically based, providing objective analysis of samples to be identified. Based on extraction of reference fingerprint ions from test spectra, this approach should lend itself well to real-world applications where samples are likely to be impure. This algorithm is illustrated using a blind study. In the study, MALDI-MS fingerprints for *Bacillus atrophaeus* ATCC 49337, *Bacillus cereus* ATCC 14579[T], *Escherichia coli* ATCC 33694, *Pantoea agglomerans* ATCC 33243, and *Pseudomonas putida* F1 are collected and form a reference library. The identification of test samples containing one or more reference bacteria, potentially mixed with one species not in the library (*Shewanella alga* BrY) is performed by comparison to the reference library with a calculated degree of association. Out of 60 samples, no false positives are present, and the correct identification rate is 75%. Missed identifications are largely due to a weak *B. cereus* signal in the bacterial mixtures. Potential modifications to the algorithm are presented and result in a higher than 90% correct identification rate for the blind study data, suggesting that this approach has the potential for reliable and accurate automated data analysis of MALDI-MS.

INTRODUCTION

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-MS) has become a valuable tool for the analysis of microorganisms. The speed with which data can be obtained from MALDI-MS makes this a potentially important tool for biological health hazard monitoring, food processing, blood screening, and disease diagnoses. Numerous research groups have demonstrated the ability to obtain unique MALDI-MS spectra from intact bacterial cells[1-7], and bacterial cell extracts[8-14]. The ability to differentiate strains of the same species has been investigated[7, 11, 15, 16]. Reproducibility of MALDI-MS spectra from bacterial species under carefully controlled experimental conditions has also been demonstrated[13, 17]. Wang et. al. have reported on inter-laboratory reproducibility of the MALDI-MS analysis of several bacterial species[13]. While these results are encouraging for the applicability of MALDI-MS to bacterial identification, many issues still need to be addressed including spectral variability due to culture growth time [18].

Another challenge of this MALDI-MS method as a tool for bacterial identification is the efficient and effective analysis of the data. Most previous work has used qualitative comparisons

or tabulations of ions rather than statistical techniques. Recently, a method for numerical comparison of MALDI-MS spectra has been developed[7]. This technique, based on the cross correlation between two spectra over the mass range of interest, is effective in comparing spectra under laboratory conditions when the samples to be compared are pure and controlled. Another recent approach is to compare the molecular masses obtained in the mass spectrum from bacterial analysis by MALDI-MS with the information contained in the prokaryotic genome and protein sequence databases available on the world wide web[19]. This approach is less dependent on reproducibility issues and experimental parameters that may affect the spectral appearance. However, it is difficult to fully automate, requires that the organism of interest has information catalogued in these databases, and does not provide an estimate of the uncertainty associated with identifications.

Two major goals of this ongoing work include the development and demonstration of statistical algorithms to objectively analyze the MALDI-MS spectra and to correctly identify samples based on comparison with fingerprints of known bacterial species. The spectral analysis tools must be able to effectively characterize and account for variability between replicate bacterial cultures as well as MALDI-MS analyses. Approaches for doing this have been addressed previously[17, 20].

In this work, a new algorithm for bacterial identification using MALDI-MS is presented. Rather than comparing entire spectra, the algorithm extracts key biomarkers from the spectrum, and uses those biomarkers to construct MALDI fingerprints and make identifications. The algorithm for constructing MALDI fingerprints is presented in Jarman et. al.[20], where a MALDI fingerprint consists of a collection of estimated peak heights and locations along with their corresponding uncertainties. In addition, the frequency with which each biomarker appears is included in the MALDI-MS fingerprint. In this way, it is acknowledged that biomarkers do not always appear in 100% of the replicates, due to a number of causes such as very small protein concentrations, or peaks missed by the peak detection algorithm.

The identification algorithm compares biomarkers from spectra of test samples to MALDI-MS fingerprints in a reference library and calculates a degree of match. By isolating and comparing specific biomarkers, this approach lends itself well to real world applications, where test samples are likely to be impure. The algorithm presented here is illustrated through a blind study. The library used in this study contains MALDI-MS fingerprints for single strains each of *Bacillus atrophaeus*, *Bacillus cereus*, *Escherichia coli*, *Pantoea agglomerans*, and *Pseudomonas putida*. In addition, some test samples also contain a bacterium (*Shewanella alga*) not in the reference library, in order to simulate an uncharacterized environmental organism. Although this study is limited, results provide evidence of feasibility of this algorithm and MALDI-MS for reliable bacterial identification.

EXPERIMENTAL METHODS

Supplies. The cultures used in this study include *Bacillus atrophaeus* ATCC 49337, *Bacillus cereus* ATCC 14579[T], *Escherichia coli* ATCC 33694, *Pantoea agglomerans* ATCC 33243 (American Type Culture Collection, Manassas, VA), *Pseudomonas putida* F1[21], and *Shewanella alga* BrY[22]. Bacto Luria-Bertani (LB) Broth Miller (Difco), Bacto Tryptic Soy

Broth(TSB) w/o Dextrose (Difco) and Bacto Nutrient broth (Difco) were purchased from Fisher Scientific (Pittsburgh, PA, USA). Horse heart cytochrome *c* and angiotensin I were obtained from Sigma (St. Louis, MO, USA). Ferulic acid and trifluoroacetic acid (TFA) were purchased from Aldrich (Milwaukee, WI, USA). Acetonitrile and ammonium chloride were obtained from J.T. Baker (Phillipsburg, NJ, USA). The water was obtained from a Milli-Q Plus purification system (Millipore Corporation, Bedford, MA, USA).

Safety Precautions. TFA is corrosive and causes severe burns. It is toxic by inhalation, in contact with skin and if swallowed. Suitable protective clothing including lab coat, gloves and eye/face protection should be worn when working with the stock solution.

Laboratory Methods. Bacteria were cultured in separate tubes, two tubes of 3.5 mL each per organism and incubated ~15 hours in a shaker incubator at the appropriate temperatures. Each culture was divided in half and centrifuged at 14,000 rpm for 2 minutes, decanted and washed twice with 2% ammonium chloride. Cells were reconstituted in 2% ammonium chloride and the optical density was measured at 600 nm.

Bacillus atrophaeus, Bacillus cereus, Pantoea agglomerans and *Shewanella alga* were cultured in TSB and *Pseudomonas putida* in Nutrient broth for 15h at 30°C in a shaker incubator. *Escherichia coli* was cultured in Luria-Bertani broth with streptomycin for 15 h in a 37°C shaker incubator. Each culture was obtained from the same respective stock solutions during the blind study.

Prior to mass spectrometric analysis, the broth was washed from the cells with 2% ammonium chloride. For example, a 1.5 mL aliquot of the cells was centrifuged (14,000 rpm) for two minutes to form a cell pellet. The supernatant was discarded, 1.0 mL of 2% ammonium chloride was then added to the pellet, and resuspended by vortexing. The suspension was pelleted and washed once more. The final pellet was resuspended with 0.2 mL of 2% ammonium chloride and this suspension was used for MALDI analysis. Approximately 10^6 to 10^7 cells were delivered to the MALDI target for analysis[23]. This value is estimated based on comparing the optical density of the *E. coli* bacterial culture at 600 nm to the *E. coli* growth curve.

Blinded samples containing two (three) different microorganisms were generated by mixing cell suspensions in approximately 1:1 (1:1:1) concentration ratios measured using optical density @ 600nm. Samples were then coded and delivered to the MALDI-MS laboratory for analysis.

MALDI-TOFMS Analysis. A PerSeptive Biosystems Voyager-DE RP MALDI time-of-flight mass spectrometer with a nitrogen laser (337 nm) operated in the linear, delayed extraction, and positive ion mode was used during the experiments. The low mass gate was set to m/z 150, the delay time was 60 ns, the accelerating voltage was 23kV, and the grid voltage and guide wire voltage were set to 90% and 0.2% of the accelerating voltage respectively. External calibration with the monomer ion of cytochrome *c* (*m/z* 12 361) and the monomer ion of angiotensin I (*m/z* 1297) was used along with an internal calibration consisting only of the monomer ion of cytochrome *c*. Each spectrum was obtained by averaging 128 laser shots.

The ferulic acid matrix solution was a 10 mg/mL solution in acetonitrile (30%) and 0.1% TFA (70%) along with 5 µg/mL cytochrome c and 2.5 µg/mL angtiotensin I. In addition, another ferulic acid matrix solution was made similarly but without the two protein internal standards. A layering method was used for the bacterial analysis in which 1 µL of the bacterial sample was applied to the sample plate and allowed to air dry. Then 1 µL of the ferulic acid matrix with internal standards was applied to the bacterial sample spot and allowed to air dry. An additional 1 µL of ferulic acid without the internal standards was applied to the dried sample spot and allowed to dry before analysis. During the analysis, the bacterial samples were stored at room temperature. The operator applied the bacterial samples to the sample plate at approximately the same time and collected replicate spectra. The data files were then transferred to the data analyst for fingerprint construction or blind study comparison.

To construct a reference fingerprint, the above procedure for MALDI-MS analysis was conducted, where ten replicate spectra were collected from each divided culture on each of three days, yielding a total of 60 spectra per bacterium. For the blind study, five replicate spectra were obtained for each of the 60 test samples.

Blind Study Experimental Design. MALDI-MS analysis was performed on combinations of the five fingerprint library species along with *Shewanella alga* (*S. alga*). In this study, *S. alga* is not contained in the reference library and serves the role of an uncharacterized environmental microbe. Fifteen bacterial combinations were selected from the set of all possible combinations of one, two or three bacteria to meet the following objectives. Each fingerprint library species appeared alone to evaluate the ability of the algorithm to correctly identify a species in the absence of other bacteria types. *S. alga* appeared alone to test the ability of the algorithm to correctly eliminate every member of the fingerprint library. Samples containing two or three bacterial species were selected to assess performance of the algorithm with samples containing more than one bacterial species. The specific combinations were selected such that each bacterial species appeared in at least three combinations, or 12 total samples. Bacterial combinations used in this study are listed in Table 1.

Blinded samples were collected over a 6-day period. Each bacterial combination in Table 1 was used to generate four blinded replicate samples for MALDI-MS analysis. For a given combination, replicate samples were prepared from independent cultures on the two different days as indicated in parentheses in the table. On a given day, cultures were divided and labeled, so that two separate MALDI-MS analyses were run. We note that the samples were numbered and neither the MALDI-MS operators nor the data analyst knew the contents of the samples.

Each bacterial combination was used once in the first three days and once in the last three days to incorporate day to day variability. Each blinded sample was used to generate five replicate MALDI spectra. These five replicate MALDI spectra were then used to form a composite spectrum. The composite spectrum was then compared to each of the library fingerprints using a calculated degree of association.

Table 1. Blind Study Bacterial Combinations Tested

B. atrophaeus (Day 2 & 4)	P. putida & S. alga (Day 3 & 4)
B. cereus (Day 1 & 5)	B. atrophaeus & B. cereus (Day 3 & 4)
E. coli (Day 2 & 6)	B. cereus & P. agglomerans (Day 3 & 5)
P. agglomerans (Day 3 & 6)	E. coli & P. agglomerans (Day 1 & 6)
P. putida (Day 1 & 5)	P. agglomerans & P. putida (Day 2 & 4)
S. alga (Day 1 & 6)	B. atrophaeus & E. coli & P. putida (Day 3 & 6)
B. cereus & S. alga (Day 2 & 5)	B. cereus & E. coli & S. alga (Day 1 & 5)
E. coli & S. alga (Day 2 & 4)	

NUMERICAL APPROACH

MALDI-MS Fingerprint Construction. In this work, a MALDI-MS fingerprint is defined to be the peak location, peak height, uncertainties in location and height, and the frequency of occurrence for each peak[20]. More specifically, a MALDI fingerprint is defined by $F = \{l_i, s_{li}, h_i, s_{hi}, p_i\}$ where for each peak i, l_i is the average peak location, s_{li} is the standard deviation in peak location, h_i is the average peak height (normalized to the maximum peak height), s_{hi} is the standard deviation of peak height, and p_i is the fraction of replicates in which peak i appears. For this study, only peaks that appear in more than 70% of fingerprint replicates are included in the fingerprint. Selected based on past experience, this 70% threshold is designed to allow only the most reproducible biomarkers to appear in a MALDI-MS fingerprint. However, further investigation is needed to better determine the most reliable criteria for allowing peaks to appear in a MALDI-MS fingerprint.

Bacterial Identification. For each blinded sample and each reference fingerprint, a likelihood is computed based on the number of fingerprint ions observed in the blinded sample. This likelihood is a value between zero and one. If the likelihood is close to one, then the blinded sample contains the significant fingerprint biomarkers, and the reference bacterium is determined to be present. If the likelihood is close to zero, then the blinded sample does not contain the significant fingerprint biomarkers, and the reference is determined to be absent.

Identification takes place in three stages. First, peaks in the blinded sample spectra are detected, characterized, and averaged across replicates obtained, and a table consisting of peak locations and their corresponding heights is generated. Second, peak locations of the blinded sample are compared to peak locations for a given reference fingerprint. Blinded sample spectral peak locations falling inside the uncertainty region for a fingerprint peak are labeled "observed", where the uncertainty region for each fingerprint peak is given by the $1-\alpha$ prediction interval for that peak constructed from the average and standard deviation in peak location, and the Student's t-distribution[24].

Let N_{fp} denote the total number of ions in a given fingerprint. A vector $\underline{u}$ of length N_{fp} is constructed. The elements of $\underline{u}$ contain 0's and 1's. The ith element of $\underline{u}$ is zero if the ith fingerprint peak is not observed in the blinded sample spectrum, and one if the ith fingerprint

peak is observed in the blinded sample spectrum. The number of ones in $\underline{u}$ (or sum of all elements of $\underline{u}$) indicates the number of fingerprint biomarkers observed in the blinded sample.

Experimental results indicate that all fingerprint biomarkers are not equally important. Some are very strong and appear in virtually all replicates, while others are much weaker and tend to drop out of some replicates. The importance of each fingerprint biomarker i is indicated by its frequency of appearance p_i. The third stage of this algorithm uses p_i for all peaks i and the vector $\underline{u}$ to estimate the degree of match between the fingerprint and the blinded sample.

Consider the following hypotheses

H_0:   blinded sample contains species k

H_A: blinded sample does not contain species k.

Under H_0 (the sample contains species k), the probability the blinded sample spectrum contains fingerprint peak i is p_i. Our algorithm for comparing a blinded sample to a reference fingerprint is similar to a statistical test of significance[24], where a significance of the observed outcome of the experiment is computed, and the null hypothesis is accepted or rejected based on this significance. In particular, the outcome is the vector $\underline{u}$ indicating the fingerprint biomarkers observed in the blinded sample. The significance is the probability of having fewer fingerprint biomarkers (elements of $\underline{u}$ equal to 1) than observed, given H_0 is true.

For fingerprint k, let M represent the set of fingerprint peaks not observed (missing) in the blinded sample (elements of $\underline{u}$ equal to zero). The set of peaks observed in the blinded sample (elements of $\underline{u}$ equal to one) is represented by the complement of the set M, denoted M^c. The significance of M is measured using the *degree of association* with fingerprint k (denoted *d.a.(k)*), and can be expressed by

$$
\begin{aligned}
d.a.(k) &= 1 - P\{\text{ all peaks in } M^{\,C} \text{ observed and } \geq 1 \text{ peak in } M \text{ observed } | H_0\} \\
&= 1 - P\{\text{all peaks in } M^{\,C} \text{ observed } | H_0\} P\{\geq 1 \text{ peak in } M \text{ observed } | H_0\} \\
&= 1 - P\{\text{all peaks in } M^{\,C} \text{ observed } | H_0\}(1 - P\{\text{no peaks in } M \text{ observed } | H_0\}) \quad (2) \\
&= 1 - \prod_{i \in M^{\,C}} p_i [1 - \prod_{i \in M}(1 - p_i)]
\end{aligned}
$$

When $M = \varnothing$, all the fingerprint biomarkers are observed in the blinded sample and we define

$$
P\{\text{no peaks in } M \text{ present} | H_0\} = \prod_{i \in M}(1 - p_i) = P\{\varnothing = \varnothing\} = 1
$$

When $M^c = \varnothing$, none of the fingerprint biomarkers are observed in the blinded sample and we define

$$P\{\text{all peaks in } M^C \text{ present} \mid H_0\} = \prod_{i \in M^C} p_i = P\{\varnothing = \varnothing\} = 1$$

In practice, Eqn. (2) is modified slightly. In particular, we automatically set $d.a.(k) = 0$ if less than some percentage of the fingerprint peaks are observed in the blinded sample. Doing this prevents a very small number of fingerprint ions to result in the false conclusion that a given reference bacterium is present. This gives the following expression for the degree of association

$$d.a.\ (k) \ = \ \begin{cases} \text{Eqn (2)} & \geq x\% \text{ f.p. peaks observed in blinded sample} \\ 0 & < x\% \text{ f.p. peaks observed in blinded sample} \end{cases} \qquad (3)$$

Clearly, $d.a.(k)$ can range from zero to one. When all of the fingerprint peaks are present in a blinded sample, $M = \varnothing$ and $d.a.(k) = 1$. When none of the fingerprint peaks are present in a blinded sample, $d.a.(k) = 0$ from Eqn. 3. When some of the fingerprint peaks are present in the blinded sample, $0 < d.a.(k) < 1$.

Uncertainty in peak extraction is also incorporated into the comparison technique. For each reference fingerprint biomarker i, a corresponding blinded sample peak is extracted if it falls inside the $1-\alpha$ prediction interval determined by the average and standard deviation of the reference biomarker (assuming a normal distribution). This implies that if a given peak is present in 100 spectra, on average it will fail to be extracted 100α times. This reduces the probability that $u_i = 1$ under H_0 by a factor of $1-\alpha$, so that

$$p_i \leftarrow (1-\alpha)p_i$$

for all fingerprint peaks i.

The relative intensities of the ions are not taken into account in the identification algorithm presented here. We realize this is an important parameter, however there is currently considerable variability between the relative intensity of replicate MALDI spectra[13] and determining a reliable, objective way to deal with this variability is still in progress.

Finally, for each blinded sample, five replicate MALDI spectra are combined to form a single composite peak table for which a degree of association is computed. By comparing a composite of five spectra to each fingerprint, the effects of poor quality spectra due to high noise or low bacterial concentrations are reduced. We acknowledge that combining five spectra in this manner increases the probability of fingerprint biomarkers being present under the null hypothesis. If the five replicates are independent, this increased probability can easily be computed and incorporated into the comparison procedure. However, empirical evidence suggests that the replicates are not independent. Several reasons for this are possible, including: a weak signal due to low overall bacterial concentration across all five replicates, and calibration errors causing a systematic m/z shift for all replicates. To enable a more concise presentation, computing an accurate estimate of the probability of a fingerprint peak being present in one of five replicates will be addressed in future work.

Based on previous empirical experience and statistical convention for interpretation of results, a five point determination scale is used to interpret the results based on the degree of association with a given species k. This scale is given below:

d.a.(k)	Conclusion
0.7-1.0	k highly likely to be present
0.15-0.7	k likely to be present
0.05-0.15	inconclusive
0.01-0.05	k unlikely to be present
0.0-0.01	k highly unlikely to be present

In this study, a species k is identified in the sample for likely or highly likely conclusions (*d.a(k)* between 0.15-1.0). All other conclusions result in a determination that species k is absent from the blinded sample.

RESULTS AND DISCUSSION

For this study, the MALDI-MS reference fingerprints for *B. atrophaeus, B. cereus, E. coli, P. agglomerans*, and *P. putida* and the computed degree of association between each blinded sample and each reference fingerprint are provided as Supplementary Materials. The blind study comparison results are summarized in Table 2, where the lower threshold in Eqn (3) is set to $x=50\%$, selected arbitrarily. In the Table, the percent of true positives and true negatives is given. The total number of samples used to compute these percentages is given in parentheses.

With the initial, completely blinded application of this comparison method, the entire contents in 45 out of all 60 samples (75%) are correctly identified. Of the 15 errors made, 13 are caused by a failure to detect *Bacillus cereus* (ATCC 14579[T]) in a mixture of two or more bacterial species. Of the 40 samples not containing *B. cereus*, all bacteria in 38 samples (95%) are correctly identified. One of these errors is a failure to identify *E. coli* in a sample with no other species, and the other error is a failure to identify *B. atrophaeus* in a mixture with *E. coli* and *P. putida*. No false positives occurred in this study.

Single representative MALDI-MS spectra of three different blinded samples are shown in Figure 1. Spectrum (a) is from blinded sample #49 containing *P. putida*. Spectrum (b) is from blinded sample #56 containing a mixture of *B. atrophaeus*, *E. coli* and *P. putida*. Spectrum (c) is from blinded sample #58 containing *E. coli*. Ions from *E. coli* and *P. putida* can be visually observed in the spectrum (b) of Figure 1. However, visual comparison can be influenced by complexity of the mixture spectrum and differences in relative intensity between ions from different species, making it more difficult for confident comparisons to be made.

Table 2. Summary of Blind Study Results

Species	% Samples Correctly Identified (True Positives)		% Samples Correctly Eliminated (True Negatives)
	Threshold = 50%	Threshold = 20%[†]	Threshold=50%, 20%
Bacillus atrophaeus	94% (n=18)	100%	100% (n=42)
Bacillus cereus	35% (n=20)	85%	100% (n=40)
E. coli	95% (n=20)	100%	100% (n=40)
Pantoea agglomerans	100% (n=16)	100%	100% (n=44)
Pseudomonas putida	100% (n=16)	100%	100% (n=44)

† Results for post-blind study comparison.
n = number of samples

While visual inspection of the MALDI-MS spectra would reveal some of these identifications, the success rates and confidence in conclusions would surely be much lower than with this automated approach. In addition, the approach we are using never relies on a single mass spectrum but rather the compilation of at least five replicates to minimize the normal variability observed with MALDI-MS spectra, an approach that is difficult to implement visually.

Diagnosing the Errors. The results of the blind study are promising, with the exception of *B. cereus*. Overall, *B. cereus* is correctly detected only 35% of the time. All of these errors occur in samples containing mixtures of 2 or more species. This failure to identify *B. cereus* appears to be due to the fact that in the presence of the other blind study species, fewer than 50% of the *B. cereus* fingerprint peaks typically appear in a spectrum. Figure 2 plots the percentage of fingerprint peaks present for each species in the blind study mixture samples containing *B. cereus*. For example, the first sample plotted contains a mixture of *B. cereus*, *E. coli*, and *S. alga*. In this particular sample, approximately 16% of the *B. cereus* fingerprint peaks appear, approximately 65% of the *E. coli* peaks appear, and 100% of the *S. alga* peaks appear.

By examining Figure 2, it is clear that the percentage of *B. cereus* (ATCC 14579[T]) peaks appearing in mixtures is consistently lower than for other species. It is unclear at this point whether this is due to a relative concentration difference in the mixtures or because this particular strain of *B. cereus* does not compete as well as other species for ionization. Similar competition for ionization is known to occur in prepared protein mixtures[25], and more effort to understand this potential competition for ionization versus relative cell concentration is needed.
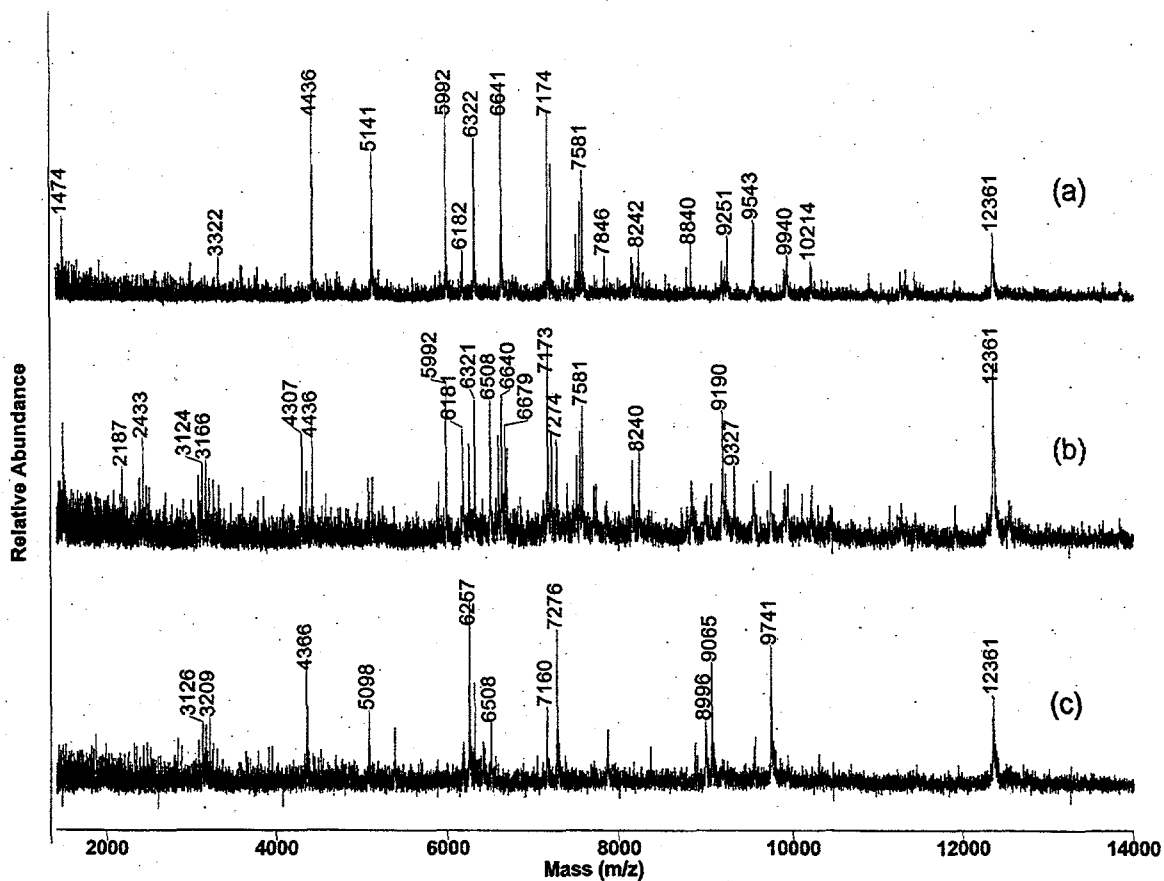
Figure 1. Representative MALDI-MS spectra for three samples from the blind study. Contents of samples are as follows: (a) *P. putida*, (b) *B. atrophaeus*, *E. coli* and *P. putida*, and (c) *E. coli*.

All of the *B. cereus* errors made in the blind study are due to the fact that fewer than 50% of the fingerprint peaks appear in the spectrum, so that the comparison algorithm automatically sets the degree of association to zero. We note that this 50% value was selected somewhat arbitrarily as a first guess until enough data was available to test such threshold settings. This suggests that by setting the threshold in Eqn 3 lower, fewer *B. cereus* errors will be made. Table 2 gives the results generated by comparing the blind study samples to the reference fingerprint library using a threshold of 20%, rather than the 50% used in the blinded comparison. In this case, only three errors are made; all three errors are false negatives corresponding to a failure to identify *B. cereus* in a mixture. We note that the algorithm has been modified after analyzing the blind study data, so these results do not reflect a blind comparison. However, they do suggest potential improvement for this numerical approach.
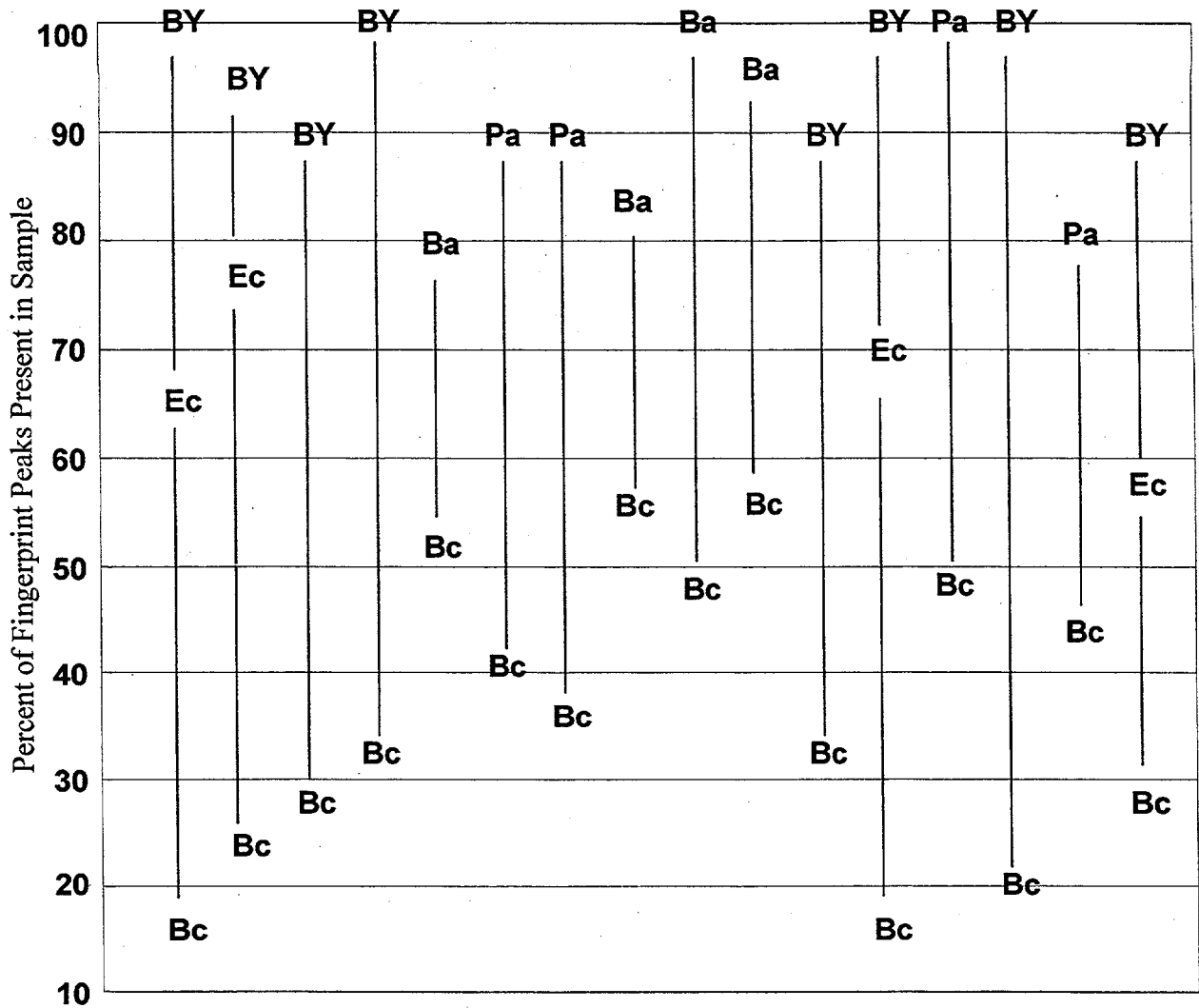
Figure 2. Percent of fingerprint peaks present in blind study mixtures containing *Bacillus cereus*. Bc = *B. cereus*, Ec = *E. coli*, Ba = *B. atrophaeus*, and Pa = *P. agglomerans*, BY = *S. alga*

CONCLUSIONS

A statistically-based algorithm for bacterial identification using MALDI-MS with automated data extraction and analysis has been introduced. The blind study results indicate (1) reproducible MALDI-MS fingerprints can be constructed, (2) MALDI-MS fingerprints are unique for the limited number of organisms in this study, and (3) the potential exists for fully automated bacterial identification using MALDI-MS. A benefit of this approach is that it is not susceptible to human bias present in qualitative, visual comparison. Therefore, it can be used in future studies to help assess the utility of MALDI-MS for bacterial identification. In addition, by isolating and extracting biomarkers of interest, this algorithm has the potential for identification in situations where the samples are likely to be impure. As a result, this approach lends itself to

implementation into field-deployable instrumentation, where a user would like to perform rapid, on site bacterial identification in a fully automated fashion.

The blind study presented here is quite limited in scope, a single strain of each species, and many research questions remain to be answered. First, the number of organisms included in this work is very small. A key question is how well this approach will work when a larger library of organisms is included. The authors are currently building a more extensive fingerprint library including several *Bacillus* strains, several *E. coli* strains, and a number of other genera. Unpublished results on a library containing 22 organisms indicate that this approach works well in identifying organisms at least to the genera level, and to the species level in most cases. This ongoing research will be addressed in future publications.

Several other important questions remain open, including the sensitivity of the method, effects of bacterial growth phase and culture media on the MALDI-MS fingerprint, and the extension of this capability to the identification of bacteria in more realistic environmental or forensic samples. The method we have established here for data extraction and comparison will allow for controlled evaluation of these variables in a statistically-based fashion.

Supplementary Material Available. MALDI-MS Fingerprints for *B. atrophaeus, B. cereus, E. coli, P. agglomerans*, and *P. putida* are contained in Tables 3-7, respectively. The degree of association between test samples and each reference fingerprint in the blind study is provided in Table 8.

ACKNOWLEDGEMENTS

REFERENCES

(1)   Holland, R.; Wilkes, J.; Rafii, F.; Sutherland, J.; Persons, C.; Voorhees, K.; Lay, J., Jr. *Rapid Commun. Mass Spectrom.* **1996,** *10*, 1227-1232.

(2)   Krishnamurthy, T.; Ross, P. *Rapid Commun. Mass Spectrom.* **1996,** *10*, 1992-1996.

(3)   Claydon, M.; Davey, S.; Edwards-Jones, V.; Gordon, D. *Nature Biotechnology* **1996,** *14*, 1584-1586.

(4)   Erhard, M.; von Dohren, H.; Jungblut, P. *Nature Biotechnology* **1997,** *15*, 906-909.

(5)   Welham, K. J.; Domin, M. A.; Scannell, D. E.; Cohen, E.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1998,** *12*, 176-180.

(6)   Karty, J.; Lato, S.; Reilly, J. *Rapid Commun. Mass Spectrom.* **1998,** *12*, 625-629.

(7)   Arnold, R.; Reilly, J. *Rapid Commun. Mass Spectrom.* **1998,** *12*, 630-636.

(8)   Cain, T.; Lubman, D.; Weber, W. *Rapid Commun. Mass Spectrom.* **1994,** *8*, 1026-1030.

(9)   Krishnamurthy, T.; Ross, P.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996,** *10*, 883-888.

(10)   Chong, B.; Wall, D.; Lubman, D.; Flynn, S. *Rapid Commun. Mass Spectrom.* **1997,** *11*, 1900-1908.

(11)   Haag, A. M.; Taylor, S. N.; Johnston, K. H.; Cole, R. B. *J. Mass Spectrom.* **1998,** *33*, 750-756.

(12)   van Adrichem, J.; Bornsen, K.; Conzelmann, H.; Gass, M.; Eppenberger, H.; Kreshbach, G.; Ehrat, M.; Leist, C. *Anal. Chem.* **1998,** *70*, 923-930.

(13)   Wang, Z.; Russon, L.; Li, L.; Roser, D.; Long, S. *Rapid Commun. Mass Spectrom.* **1998,** *12*, 456-464.

(14)   Easterling, M. L.; Colangelo, C. M.; Scott, R. A.; Amster, I. J. *Anal. Chem.* **1998,** *70*, 2704-2709.

(15)   Haddon, W. F.; Full, G.; Mandrell, R. E.; Wachtel, M. R.; Bates, A. H.; Harden, L. A. , in Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL, **1998**, 177.

(16)   Nilsson, C. L. *Rapid Commun. Mass Spectrom.* **1999,** *13*, 1067-1071.

(17)   Saenz, A. J.; Petersen, C. E.; Valentine, N. B.; Gantt, S. L.; Jarman, K. H.; Kingsley, M. T.; Wahl, K. L. *Rapid Commun. Mass Spectrom.* **1999,** *13*, 1580-1585.

(18)   Arnold, R.; Karty, J.; Ellington, A.; Reilly, J. *Anal. Chem.* **1999,** *71*, 1990-1996.

(19)   Demirev, P. A.; Ho, Y.-P.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999,** *71*, 2732-2738.

(20)   Jarman, K. H.; Daly, D. S.; Petersen, C. E.; Saenz, A. J.; Valentine, N. B.; Wahl, K. L. *Rapid Commun. Mass Spectrom.* **1999,** *13*, 1586-1594.

(21)   Zylstra, G. J.; McCombie, W. R.; Gibson, D. T.; Finette, B. A. *Appl. Environ. Microbiol.* **1988,** *54*, 1498-1503.

(22)   Caccavo, F. J.; Blakemore, R. P.; Lovely, D. R. *Appl. Environ. Microbiol.* **1992,** *58*, 3211-3216.

(23)   Gantt, S. L.; Valentine, N. B.; Saenz, A. J.; Kingsley, M. T.; Wahl, K. L. *J. Amer. Soc. Mass Spectrom.* **1999,** *10*, 1131-1137.

(24)   Bickel, P.; Doksum, K. *Mathematical Statistics: Basic Ideas and Selected Topics.*; Prentice-Hall., **1977.**

(25)   Cohen, S.; Chait, B. *Anal. Chem.* **1996,** *68*, 31-37.

Supplementary Material

An Algorithm for Automated Bacterial Identification using Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry

Kristin H. Jarman*, Sharon T. Cebula, Adam J. Saenz, Catherine E. Petersen,
Nancy B. Valentine, Mark T. Kingsley, Karen L. Wahl

Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352

SUPPLEMENTARY MATERIAL

Tables 3-7 list the fingerprints for *B. atrophaeus*, *B. cereus*, *E.coli*, *P. agglomerans*, and *P. putida*, respectively. In each table, "Average." refers to the average, "Std Dev" refers to the standard deviation, and "Frequency" refers to the relative frequency of appearance of each biomarker as a fraction of the total number of replicate spectra.

Table 3. *Bacillus atrophaeus* MALDI-MS Fingerprint

Peak Location		Peak Height		
Average	St. Dev	Average	St. Dev	Frequency
3254.01	1.25	12.00	4.04	0.93
4305.40	1.04	35.86	12.31	1.00
4590.52	1.44	9.47	3.34	0.81
4663.79	1.38	10.24	4.49	0.86
5055.14	3.60	9.94	3.67	0.86
5253.15	1.68	9.18	2.84	0.78
5566.27	1.94	10.00	4.17	0.90
5900.65	1.26	21.05	7.31	1.00
6506.22	1.50	41.83	14.11	1.00
6597.24	1.48	32.68	11.04	1.00
6677.65	1.58	28.34	9.81	1.00
6698.33	1.72	21.82	7.97	1.00
7114.12	2.28	14.19	6.15	1.00
7398.02	2.10	18.55	6.84	1.00
7472.09	2.41	11.27	4.18	0.92
7711.88	1.88	23.22	12.20	1.00
7739.87	2.22	14.69	5.54	0.98
8851.83	2.23	12.04	6.14	0.98
9179.85	2.33	19.04	7.47	1.00
9325.10	2.16	17.74	6.53	1.00
9882.62	2.92	10.87	4.60	0.78
10100.84	3.01	12.73	5.94	0.93
10413.87	2.66	11.06	5.21	0.73
10435.73	2.92	13.35	6.66	0.98
11129.69	2.94	14.32	6.38	1.00

Table 4. *Bacillus cereus* MALDI-MS Fingerprint

Peak Location		Peak Height		
Average	St. Dev	Average	St. Dev	Frequency
2167.33	0.76	9.46	3.14	0.72
2812.16	0.96	11.50	4.42	0.93
3132.05	1.05	9.79	4.02	0.90
3213.07	0.91	12.46	5.02	1.00
4334.07	0.85	48.32	18.01	1.00
4509.91	1.54	7.69	2.66	0.88
5120.11	1.33	8.75	2.64	1.00
5170.81	1.69	8.18	3.58	0.87
5546.88	1.30	9.77	4.15	1.00
5886.01	1.43	15.17	5.65	1.00
6262.47	1.18	32.79	14.39	1.00
6382.43	2.15	8.05	2.68	0.83
6424.59	1.21	49.09	21.54	1.00
6555.75	1.47	9.44	3.43	0.78
6600.33	1.65	8.89	3.47	0.88
6684.34	1.85	12.36	7.40	0.97
6792.54	1.70	12.23	4.85	0.98
7049.22	2.38	7.66	2.78	0.95
7211.46	2.29	7.37	2.82	0.82
7365.03	2.22	9.51	3.77	1.00
7767.86	1.93	10.03	4.80	1.00
7822.61	1.99	9.45	3.57	0.98
8053.51	2.31	7.19	3.13	0.75
9209.32	2.26	6.79	2.67	0.80
9339.43	2.49	7.63	3.06	0.92

Table 5. *Escherichia coli* MALDI-MS Fingerprint

Peak Location		Peak Height		
Average	Stand. Dev.	Average	Stand. Dev.	Frequency
2430.32	1.39	18.24	10.52	0.95
2471.69	2.04	15.57	9.84	0.81
3122.32	1.63	22.30	12.24	0.98
3164.21	1.02	24.37	13.08	0.98
3206.07	1.22	21.04	11.34	0.95
3636.90	1.29	16.51	9.03	0.85
4364.19	1.71	25.74	14.71	0.97
4869.28	2.07	16.21	9.75	0.83
5095.31	1.52	20.49	10.97	0.95
6253.51	1.89	38.98	21.64	0.98
6314.33	2.39	24.64	13.68	0.97
6409.72	2.12	21.76	12.05	0.98
6506.62	2.17	18.13	10.30	0.95
7156.58	2.47	20.18	13.71	0.83
7272.00	2.12	52.65	29.35	0.98
7868.20	2.59	21.21	11.54	0.93
8366.31	3.03	15.66	9.61	0.92
8873.65	3.48	17.02	9.91	0.95
8895.39	3.06	19.43	12.01	0.92
8991.21	3.13	18.45	10.31	0.95
9060.52	2.53	25.22	14.88	1.00
9188.23	3.51	24.70	15.23	0.98
9532.22	3.40	21.06	11.58	0.97
9550.39	3.36	22.06	11.25	0.73
9736.32	2.81	41.16	25.08	0.97
10296.26	3.85	19.05	11.83	0.98

Table 6. *Pantoea agglomerans* MALDI-MS Fingerprint

Peak Location		Peak Height		
Average	Std. Dev.	Average	Std. Dev.	Frequency
6199.66	0.97	42.13	14.49	0.70
6400.94	1.26	32.46	13.92	0.70
7263.80	0.97	50.12	25.63	0.70
7338.40	3.55	35.42	23.66	0.90
7988.29	2.23	27.62	11.88	0.70
8898.23	4.11	33.69	24.26	0.70
9175.34	1.64	26.89	13.89	0.70
9318.39	1.55	32.83	14.20	0.70
9506.82	1.59	46.24	23.92	0.98
9568.51	2.27	25.91	12.78	0.70

Table 7. *Pseudomonas putida* MALDI-MS Fingerprint

Peak Location		Peak Height		
Average	Std. Dev.	Average	Std. Dev.	Frequency
4434.98	1.04	28.95	15.96	1.00
5138.93	1.50	17.34	8.47	1.00
5990.28	1.08	33.10	18.17	1.00
6320.79	1.16	27.54	14.62	1.00
6638.57	1.32	29.58	15.97	1.00
7172.37	1.46	33.49	18.47	1.00
7213.51	1.62	24.87	16.02	1.00
7508.24	1.56	15.65	8.09	0.93
7546.43	1.71	26.97	14.47	1.00
7579.38	1.34	25.81	13.60	1.00
7843.84	1.93	14.64	7.28	0.90
8160.19	2.01	17.03	8.90	0.98
8238.12	2.01	18.65	11.03	0.97
8548.91	1.95	13.62	7.21	0.87
8791.34	2.22	13.58	6.53	0.80
8838.41	2.11	17.43	9.46	0.98
9188.41	1.80	15.21	8.22	0.98
9222.76	2.37	15.18	6.83	0.75
9248.93	2.11	17.52	10.15	0.98
9541.39	1.89	18.00	9.76	1.00
9898.93	2.12	14.50	7.83	0.87
9936.42	2.25	16.86	8.84	0.80
10217.00	2.65	16.49	8.99	0.98
11271.14	2.72	13.73	6.58	0.75
11909.92	2.79	12.29	7.02	0.72

Table 8 reports the degree of association between each test sample in the blind study, and each reference fingerprint given in Tables 3-7. Guidelines used for making identifications in the study are outlined in the Results and Discussion section.

Table 8. Blind study results using a 50% comparison threshold

Sample ‡	Contents	Degree of Association with Library Fingerprint				
		B. atrophaeus	B. cereus	E. coli	P. agglomerans	P. putida
1	S. alga	0	0	0	0	0
2	B. cereus & E. coli & S. alga	0	0	0.9	0	0
3	B. cereus	0	0.93	0	0	0
4	E. coli & P. agglomerans	0	0	0.91	0.97	0
5	P. putida	0	0	0	0	1
6	E. coli & P. agglomerans	0	0	0.91	1	0
7	B. cereus & E. coli & S. alga	0	0	0.93	0	0
8	P. putida	0	0	0	0	0.94
9	S. alga	0	0	0	0	0
10	B. cereus	0	1	0	0	0
11	E. coli & S. alga	0	0	0.84	0	0
12	P. agglomerans & P. putida	0	0	0	0.97	0.95
13	B. atrophaeus	0.94	0	0	0	0
14	E. coli	0	0	0.96	0	0
15	B. cereus & S. alga	0	0	0	0	0
16	E. coli & S. alga	0	0	0	0	0
17	B. atrophaeus	0.93	0	0	0	0
18	P. agglomerans & P. putida	0	0	0	0.94	0.93
19	E. coli	0	0	1	0	0
20	B. cereus & S. alga	0	0	0	0	0
21	P. agglomerans	0	0	0	1	0
22	B. atrophaeus & E. coli & P.putida	0.66	0	0.79	0	0.95
23	B. atrophaeus & B. cereus	0.9	0.64	0	0	0
24	P. putida & S. alga	0	0	0	0	0.73
25	B. cereus & P. agglomerans	0	0	0	0.97	0
26	B. cereus & P. agglomerans	0	0	0	0.97	0
27	P. putida & S. alga	0	0	0	0	0.91
28	B. atrophaeus & B. cereus	0.89	0.82	0	0	0
29	B. atrophaeus & E. coli & P.putida	0	0	0.83	0	1
30	P. agglomerans	0	0	0	0.97	0

Table 8 (cont.). Blind study results using a 50% comparison threshold

Sample	Contents	Degree of Association with Library Fingerprint				
		B. atrophaeus	B. cereus	E. coli	P. agglomerans	P. putida
31	P. agglomerans & P. putida	0	0	0	1	1
32	B. atrophaeus & B. cereus	1	0	0	0	0
33	E. coli & S. alga	0	0	0.86	0	0
34	B. atrophaeus	0.95	0	0	0	0
35	P. putida & S. alga	0	0	0	0	0.71
36	E. coli & S. alga	0	0	0.84	0	0
37	P. agglomerans & P. putida	0	0	0	1	0.94
38	B. atrophaeus & B. cereus	0.95	0.77	0	0	0
39	P. putida & S. alga	0	0	0	0	0.92
40	B. atrophaeus	0.95	0	0	0	0
41	P. putida	0	0	0	0	1
42	B. cereus	0	0.96	0	0	0
43	B. cereus & S. alga	0	0	0	0	0
44	B. cereus & E. coli & S. alga	0	0	0.88	0	0
45	B. cereus & P. agglomerans	0	0	0	1	0
46	B. cereus	0	0.97	0	0	0
47	B. cereus & S. alga	0	0	0	0	0
48	B. cereus & P. agglomerans	0	0	0	0.94	0
49	P. putida	0	0	0	0	0.95
50	B. cereus & E. coli & S. alga	0	0	0.79	0	0
51	B. atrophaeus & E. coli & P.putida	0.75	0	0.77	0	0.81
52	S. alga	0	0	0	0	0
53	P. agglomerans	0	0	0	0.94	0
54	E. coli	0	0	0.95	0	0
55	E. coli & P. agglomerans	0	0	0.92	1	0
56	B. atrophaeus & E. coli & P.putida	0.66	0	0.81	0	0.94
57	P. agglomerans	0	0	0	1	0
58	E. coli	0	0	0.95	0	0
59	S. alga	0	0	0	0	0
60	E. coli & P. agglomerans	0	0	0.87	1	0

‡ Samples are listed in the order collected. Ten samples per day were collected for 6 days.