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

ARTICLE *in* ANALYTICAL CHEMISTRY · APRIL 2000

Impact Factor: 5.64 · DOI: 10.1021/ac990832j · Source: PubMed

7 AUTHORS, INCLUDING:

Kristin H. Jarman
Pacific Northwest National Laboratory
37 PUBLICATIONS **941** CITATIONS
SEE PROFILE

Karen L Wahl
Pacific Northwest National Laboratory
39 PUBLICATIONS **966** CITATIONS
SEE PROFILE

Anal. Chem. **2000,** *72,* 1217−1223

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, and Karen L. Wahl

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

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.

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 interlaboratory 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 worldwide web.[19] This approach is less dependent on reproducibility issues and experimental parameters

(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 Biotechnol.* **1996**, *14*, 1584−1586.
(4) Erhard, M.; von Dohren, H.; Jungblut, P. *Nature Biotechnol.* **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; p 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.

10.1021/ac990832j CCC: $19.00
Published on Web 02/18/2000

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, 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 *B. atrophaeus* ATCC 49337, *B. cereus* ATCC 14579^T, *E. coli* ATCC 33694, *P. agglomerans* ATCC 33243 (American Type Culture Collection, Manassas, VA), *P. putida* F1,[21] and *S. 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). Horse heart cytochrome *c* and angiotensin I were obtained from Sigma (St. Louis, MO). Ferulic acid and trifluoroacetic acid (TFA) were purchased from Aldrich (Milwaukee, WI). Acetonitrile and ammonium chloride were obtained from J. T. Baker (Phillipsburg, NJ). The water was obtained from a Milli-Q Plus purification system (Millipore Corp., Bedford, MA).

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 h in a shaker incubator at the appropriate temperatures. Each culture was divided in half and centrifuged at 14 000 rpm for 2 min, decanted, and washed twice with 2% ammonium chloride. Cells were reconstituted in 2% ammonium chloride, and the optical density was measured at 600 nm.

B. atrophaeus, *B. cereus*, *P. agglomerans*, and *S. alga* were cultured in TSB and *P. putida* in nutrient broth for 15 h at 30 °C in a shaker incubator. *E. coli* was cultured in LB 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 2 min to form a cell pellet. The supernatant was discarded, 1.0 mL of 2% ammonium chloride was then added to the pellet, and the mixture was 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 10^6-10^7 cells were delivered to the MALDI target for analysis.[23] This value is estimated by 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 at 600 nm. Samples were then coded and delivered to the MALDI-MS laboratory for analysis.

MALDI-MS 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 300, the delay time was 60 ns, the accelerating voltage was 23 kV, 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

(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. Am. Soc. Mass Spectrom.* **1999**, *10*, 1131−1137.

Table 1. Blind Study Bacterial Combinations Tested

B. atrophaeus (days 2, 4)
B. cereus (days 1, 5)
E. coli (days 2, 6)
P. agglomerans (days 3, 6)
P. putida (days 1, 5)
S. alga (days 1, 6)
B. cereus and *S. alga* (days 2, 5)
E. coli and *S. alga* (days 2, 4)
P. putida and *S. alga* (days 3, 4)
B. atrophaeus and *B. cereus* (days 3, 4)
B. cereus and *P. agglomerans* (days 3, 5)
E. coli and *P. agglomerans* (days 1, 6)
P. agglomerans and *P. putida* (days 2, 4)
B. atrophaeus, E. coli, and *P. putida* (days 3, 6)
B. cereus, E. coli, and *S. alga* (days 1, 5)

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 10 replicate spectra were collected from each divided culture on each of 3 days, yielding a total of 60 spectra/bacterium. For the blind study, 5 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 *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 3 days and once in the last 3 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.

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 frequency of occurrence for each peak.[20] More specifically, a MALDI fingerprint is defined by $\boldsymbol{F} = \{l_i, s_{li}, h_i, s_{hi}, p_i\}$, 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 on the basis of 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 0 and 1. If the likelihood is close to 1, then the blinded sample contains the significant fingerprint biomarkers, and the reference bacterium is determined to be present. If the likelihood is close to 0, 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 u of length N_{fp} is constructed. The elements of u contain 0's and 1's. The ith element of u is 0 if the ith fingerprint peak is not observed in the blinded sample spectrum and 1 if the ith fingerprint peak is observed in the blinded sample spectrum. The number of 1's in u (or sum of all elements of 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

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

third stage of this algorithm uses p_i for all peaks i and the vector u to estimate the degree of match between the fingerprint and the blinded sample.

Consider the following hypotheses: (1) H_0, blinded sample contains species k; (2) 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 u indicating the fingerprint biomarkers observed in the blinded sample. The significance is the probability of having fewer fingerprint biomarkers (elements of 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 u equal to 0). The set of peaks observed in the blinded sample (elements of u equal to 1) 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 da(k)) and can be expressed by

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

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

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

In practice, eq 1 is modified slightly. In particular, we automatically set da(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

$$\mathrm{da}(k) = \begin{bmatrix} \text{eq 1} & \geq x\% \text{ fp peaks observed in blinded sample} \\ 0 & < x\% \text{ fp. peaks observed in blinded sample} \end{bmatrix} \quad (2)$$

Clearly, da(k) can range from 0 to 1. When all of the fingerprint peaks are present in a blinded sample, $M = \varnothing$ and da(k) = 1. When none of the fingerprint peaks are present in a blinded sample, da(k) = 0 from eq 2. When some of the fingerprint peaks are present in the blinded sample, $0 < \mathrm{da}(k) < 1$.

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 the following: 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:

da(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

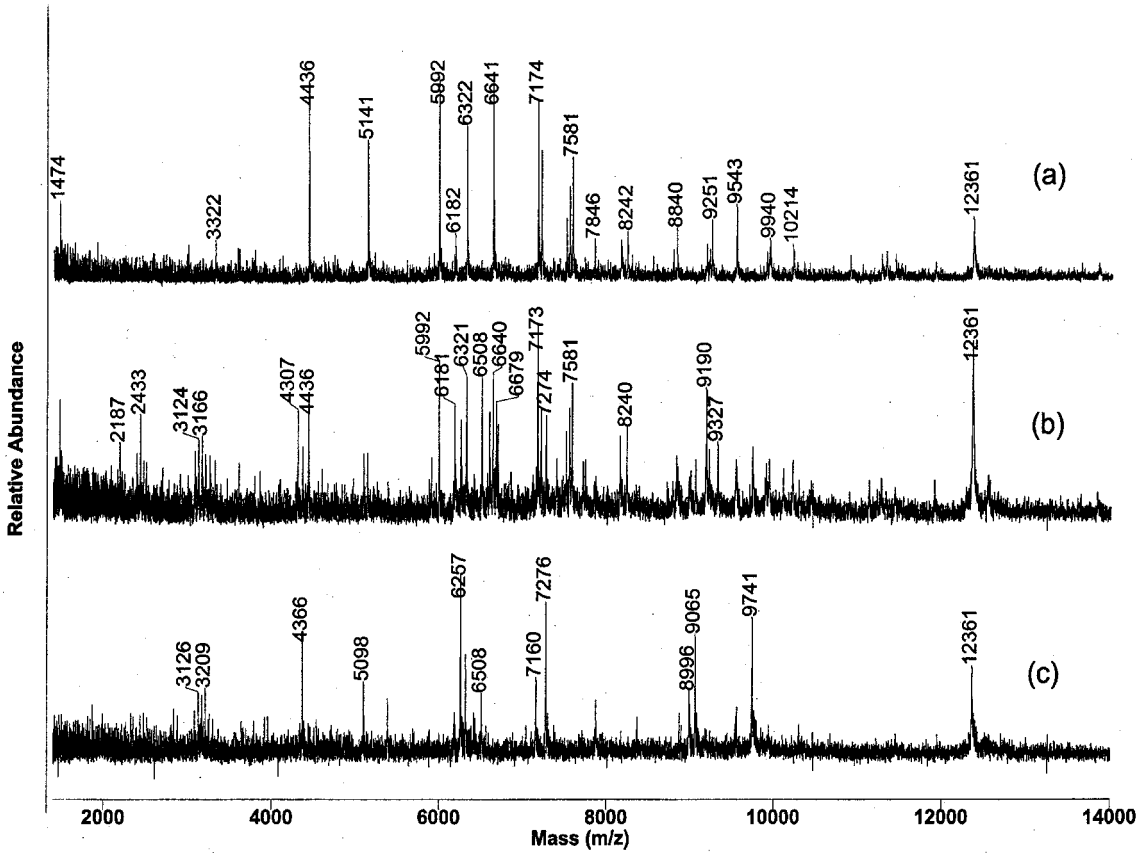

Figure 1. Representative MALDI-MS spectra for three samples from the blind study. Contents of samples are as follows: (a) *P. putida*, (b) *B. atrophaeus*, *E. coli*, and *P. putida*, and (c) *E. coli*.

Table 2. Summary of Blind Study Results

species	% samples correctly identified (true positives) threshold 50%[a]	% samples correctly identified (true positives) threshold 20%[b]	% samples correctly eliminated (true negatives) threshold 50%, 20%[a]
B. atrophaeus	94 ($n = 18$)	100	100 ($n = 42$)
B. cereus	35 ($n = 20$)	85	100 ($n = 40$)
E. coli	95 ($n = 20$)	100	100 ($n = 40$)
P. agglomerans	100 ($n = 16$)	100	100 ($n = 44$)
P. putida	100 ($n = 16$)	100	100 ($n = 44$)

[a] n, number of samples. [b] Results for post-blind study comparison.

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

RESULTS AND DISCUSSION

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 Supporting Information. The blind study comparison results are summarized in Table 2, where the lower threshold in eq 2 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 *B. 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 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.

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.

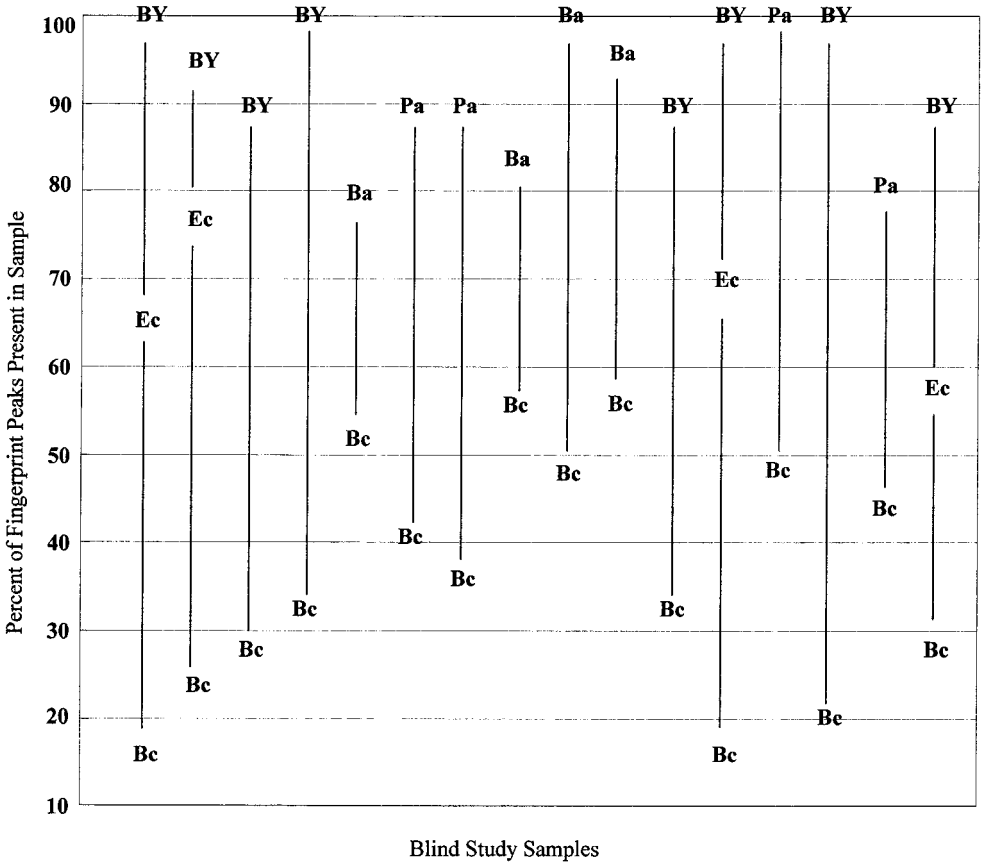

Figure 2. Percent of fingerprint peaks present in blind study mixtures containing *B. cereus*. Key: Bc, *B. cereus*; Ec, *E. coli*; Ba, *B. atrophaeus*; and Pa, *P. agglomerans*; BY, *S. alga*.

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 two 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, ~16% of the *B. cereus* fingerprint peaks appear, ~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.

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 were available to test such threshold settings. This suggests that by setting the threshold in eq 2 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.

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 instrumenta-

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

tion, 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 with only 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 some 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.

ACKNOWLEDGMENT

The U.S. Department of Energy (D.O.E.) and the Federal Bureau of Investigation supported this work. Ideas and opinions expressed here do not necessarily represent those of the FBI. Initial funding was supplied by the D.O.E. through the Laboratory Directed Research and Development program. Battelle Memorial Institute under Contract DE-AC06-76RLO 1830 operates Pacific Northwest National Laboratory for the U.S. Department of Energy.

SUPPORTING INFORMATION 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. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review July 27, 1999. Accepted December 16, 1999.

AC990832J