Anal. Chem. **2002,** *74,* 6191−6199

Analysis of Microbial Mixtures by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

Karen L. Wahl,* Sharon C. Wunschel, Kristin H. Jarman, Nancy B. Valentine, Catherine E. Petersen, Mark T. Kingsley, Kimberly A. Zartolas,† and Adam J. Saenz†

Pacific Northwest National Laboratory, Richland, Washington 99352

Many different laboratories are currently developing mass-spectrometric techniques to analyze and identify microorganisms. However, minimal work has been done with mixtures of bacteria. To demonstrate that microbial mixtures could be analyzed by matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS), mixed bacterial cultures were analyzed in a double-blind fashion. Nine different bacterial species currently in our MALDI-MS fingerprint library were used to generate 50 different simulated mixed bacterial cultures similar to that done for an initial blind study previously reported (Jarman, K. H.; Cebula, S. T.; Saenz, A. J.; Petersen, C. E.; Valentine, N. B.; Kingsley, M. T.; Wahl, K. L. *Anal. Chem.* 2000, *72,* 1217−1223). The samples were analyzed by MALDI-MS with automated data extraction and analysis algorithms developed in our laboratory. The components present in the sample were identified correctly to the species level in all but one of the samples. However, correctly eliminating closely related organisms was challenging for the current algorithms, especially in differentiating *Serratia marcescens, Escherichia coli,* and *Yersinia enterocolitica,* which have some similarities in their MALDI-MS fingerprints. Efforts to improve the specificity of the algorithms are in progress.

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI -MS) has been used to analyze intact, cultured microorganisms with minimal sample handling.[2,3,4] Two recent review articles, which include the capabilities and current limitations that need to be addressed,[5,6] provide an excellent overview of this emerging research field. The MALDI-MS technique for identifying biomolecules provides rapid analysis time (<1 min/sample analysis), low sample-volume requirements (<1 μL of fluid), and the highly selective nature of mass spectrometric analysis based on molecular weights. The m/z values for mass spectral peaks and the patterns with which they are observed can provide very specific and unbiased analysis, because they indicate molecular weights of true components of the sample. Bacterial cells can be identified by comparing MALDI-MS spectra obtained from cultured bacterial cells and simple microbial mixtures against a library of known MALDI-MS spectral fingerprints obtained from intact bacterial cells[1] or from comparison with the proteomic database.[7]

One main advantage of this MALDI-MS technique over many other bacterial analysis methods is the generic capability to classify and identify bacteria. A large number of targets can be analyzed 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, therefore, direct further, more specific testing.

We previously reported our initial blind study[1] designed to determine if MALDI-MS could be used to identify bacterial species from pure cultures or simple microbial mixtures with automated data-analysis algorithms. Data extraction and visualization algorithms developed in our laboratory were used to generate MALDI-MS fingerprints from replicate spectra. Further development of our algorithms and expansion of our bacterial fingerprint library warranted another blind study to further verify this approach for bacterial identification with more complex samples and an extended database. Nine different bacterial species currently in the MALDI-MS fingerprint library and one "unknown control" that was not in the library were chosen for this study. They were used to generate 50 simulated mixed bacterial cultures, similar to that done for the initial blind study presented previously.[1] Three of the two-component mixtures from the original study were replicated and reevaluated in the present work. Twenty-eight samples contained two bacterial species, and 18 samples contained three bacterial species. In addition, a four-component mixture was replicated four times to evaluate reproducibility in analyzing a "complex" sample. The data were evaluated at different bacterial classification levels to determine the limits of the current analysis algorithms. The original blind study was performed with a database containing only five organisms, each representing a different species, and thus, species level and strain level identifica-

† Current address: Dartmouth College, MML Building, Hanover, NH.

(1) Jarman, K. H.; Cebula, S. T.; Saenz, A. J.; Petersen, C. E.; Valentine, N. B.; Kingsley, M. T.; Wahl, K. L. *Anal. Chem.* **2000**, *72*, 1217−1223.
(2) Claydon, M. A.; Davey, S. N.; Edwards-Jones, V.; Gordon, D. B. *Nature Biotechnol.* **1996**, *14*, 1584−1586.
(3) Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J.; Persons, C.; Voorhees, K.; Lay, J. O., Jr. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1227−1232.
(4) Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1992−1996.
(5) Fenselau, C.; Demirev, P. A. *Mass Spectrom. Rev.* **2001**, *20* (4), 157−171.
(6) Lay, J. O., Jr. *Mass Spectrom. Rev.* **2001**, *20* (4), 172−194.
(7) Demirev, P. A.; Ho. Y.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999**, *71* (14), 2732−2738.

10.1021/ac0203847 CCC: $22.00
Published on Web 11/15/2002

tion was not challenged appropriately. This second blind study presented here explores the challenge of strain level identification with the current algorithms with an expanded database containing multiple strains from the same species. This study not only establishes the current capability and limitations of our current data collection and analysis protocols but also provides data for improving the algorithms for future analyses.

EXPERIMENTAL METHODS

Supplies. The cultures used in the study include *Bacillus atrophaeus* (ATCC 49337), *Bacillus cereus* (ATCC 14579), *Bacillus subtilis* (ATCC 49760), *Bacillus thuringiensis* (ATCC 13367), *Escherichia coli* W3110 (ATCC 27325), *Pantoea agglomerans* (ATCC 33243), *Pseudomonas putida* (PP0301), *Pseudomonas stutzeri* (ATCC 39524), *Serratia marcescens* (ATCC 13880), and the "unknown control", *Bacillus sphaericus* (ATCC 14577). They were obtained from the American Type Culture Collection, Manassas, VA. Bacto Luria Bertani (LB) Broth Miller (Difco), Bacto tryptic soy broth (TSB) without dextrose (Difco), and Bacto nutrient broth (Difco) were purchased from Becton Dickinson (Sparks, MD). Horse heart cytochrome *c* and angiotensin I were obtained from Sigma (St. Louis, MO). Ferulic acid and trifluoroacetic acid were purchased from Aldrich (Milwaukee, WI). Acetonitrile and ammonium chloride were obtained from J. T. Baker (Phillipsburg, NJ). The water was from a Milli-Q Plus purification system (Millipore Corp., Bedford, MA).

Safety Precautions. Trifluoroacetic acid is corrosive and causes severe burns. Suitable protective clothing, including gloves, laboratory coat, and eye/face protection, should be worn when working with the concentrated solution.

Laboratory Methods. *B. subtilis*, *B. thuringiensis*, *B. atrophaeus*, *B. cereus*, *S. marcescens*, and the "unknown control", *B. sphaericus*, were cultured in tryptic soy broth (TSB) and incubated overnight in a shaker incubator at 30 °C, 120 rpm. *Ps. stutzeri*, *P. agglomerans*, and *Ps. putida* were cultured in TSB and incubated at 37 °C, 130 rpm. *E. coli* was cultured in LB broth and incubated at 37 °C, 130 rpm. The cells were centrifuged at 14 000 rpm for 2 min, decanted, and washed twice with 1 mL of 2% ammonium chloride. The concentration of cells was $\sim 10^6$ cells/μL when deposited onto the MALDI sample plate.[8] Bacterial samples were cultured on 10 different days with five samples run each day.

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 experiment. 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.[9] 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 solution was applied to the bacterial sample spot and allowed to air-dry. Each spectrum was obtained averaging 128 laser shots. Five replicate spectra from each sample were collected for the "unknown" sample analysis. Each spectrum was internally calibrated with the monomer ion of cytochrome *c* (m/z 12 361) and the monomer ion of angiotensin I (m/z 1297). The data files were then transferred to the data analyst for automated peak extraction and analysis.

FINGERPRINT LIBRARY DEVELOPMENT AND UNKNOWN IDENTIFICATION

Library reference fingerprints were created using the method previously described[1] in which 10 replicate spectra were collected from each divided culture on each of 3 days, yielding a total of 60 spectra/bacterium. Peak extraction was performed independently for each of these 60 replicate spectra using the PNNL peak-extraction algorithm,[10] and the average m/z value and standard deviation for each reproducible peak from these replicates was calculated. The reproducible fingerprint for each bacterium contains peaks found in at least 70% of the replicate spectra. This threshold is designed to allow only the most reproducible (and probably by default the most intense) biomarkers to appear in a MALDI-MS fingerprint. The value of 70% for the threshold is chosen empirically, and it provides the maximum same-species degree of association and the minimum across-species degree of association for the fingerprints in our reference library. The fingerprint library constructed in this manner contains over 30 bacteria, as shown in Table 1. We note that the spectra used to generate the MALDI fingerprints in our library were collected over a period of 2 yrs, during which time the laser was replaced and the instrument was serviced. Although these events were not carefully controlled and monitored, this blind study does provide a cursory test of the reproducibility of bacterial fingerprints over time under normal laboratory operating conditions.

Once the library of organisms is generated, a potentially "unknown" sample is analyzed by collecting only five replicate spectra per sample and then combining them to form a composite peak table, which is then compared to each library fingerprint. For each library entry, the set of fingerprint peaks is extracted from this sample peak table and used to determine the likelihood of that library species/strain being present in the sample. This approach allows for the capability of identifying individual components present in a mixed microbial sample.

Two related identification algorithms were applied in the current study. The first identification algorithm applied in a blind fashion is the one outlined in Jarman et al.[1] in which a degree of association with each fingerprint is computed and used to make an identification. The same set of parameters is used, with the exception that the degree of association is automatically set to zero if less than 20% of fingerprint peaks are observed in a blind sample. This identification algorithm is referred to as the *original method*.

The second identification algorithm will be referred to as the *modified method*. This method is applied in the same manner as the original method, but it uses a two-sided hypothesis test rather than the one-sided test outlined in Jarman et al.[1] The modified identification algorithm is described below.

MODIFIED METHOD FOR BACTERIAL IDENTIFICATION

The method of Jarman et al.[1] for bacterial identification is analogous to a significance test against a single null hypothesis

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

(9) 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.

(10) 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.

Table 1. Bacteria Represented in Fingerprint Library

***Bacillus atrophaeus* ATCC 49337**	***Escherichia coli* W3110**
***Bacillus cereus* ATCC 14579**	*Escherichia coli* O157:H7 ATCC 43890
***Bacillus subtilis* ATCC 49760**	*Escherichia coli* O157:H7 ATCC 43894
Bacillus subtilis ATCC 49822	
Bacillus simplex ZAN 044	***Pantoea agglomerans* ATCC 33243**
Bacillus thuringiensis ATCC 10792	*Pseudomonas aeruginosa* B0267
***Bacillus thuringiensis* ATCC 13367**	***Pseudomonas putida* PP0301**
Bacillus thuringiensis ATCC 19265	*Pseudomonas putida* ATCC 39169
Bacillus thuringiensis ATCC 29730	***Pseudomonas stutzeri* ATCC 39524**
Bacillus thuringiensis subsp. *kurstaki* ATCC 33679	
Bacillus thuringiensis subsp. *kurstaki* ATCC 33680	***Serratia marcescens* ATCC 13880**
Bacillus thuringiensis subsp. *kurstaki* ATCC 35866	
Bacillus thuringinsis subsp. *israelensis* ATCC 35646	*Shewanella alga*
Bacillus thuringiensis subsp. *israelensis* ATCC 39152	
	Staphylococcus aureus ATCC 14458
Escherichia coli ATCC 33694	*Staphylococcus aureus* ATCC 19095
Escherichia coli HB101	*Staphylococcus aureus* ATCC 14990
Escherichia coli JM109	
Escherichia coli RZ1032	*Yersinia enterocolitica* ATCC 51871

that the reference microorganism is present in the sample. The original method has been modified to a simple hypothesis test in which

H_0: reference microorganism is not present in the sample

H_A: reference microorganism is present in the sample

The hypothesis test is a likelihood ratio test for H_0 versus H_A and proceeds in three steps. In the first step, a peak table is constructed from the five original spectra that contains a list of the peak locations and intensities of any peaks in the unknown sample. In the second step, any reference fingerprint peaks appearing in the peak table of the unknown are extracted using a prediction interval based on the Student's *t*-distribution.

The hypothesis test is performed in the third step of the process. In particular, under the null hypothesis, H_0, the frequency of appearance of a peak at fingerprint peak location, l_j, is given by some probability, q_j, estimated to be the probability of spurious peaks appearing due to chemical or background noise. Under the alternate hypothesis, H_A, the frequency of appearance of a peak at fingerprint peak location, l_j, is given by the probability, p_j, estimated from the reference fingerprint.[10] We note that in this study, we have set $q_j = 0.1$ for all peak locations, j, and all fingerprints, *F*.

Let $x_j = 0$ if fingerprint peak j is not observed in the unknown sample, and $x_j = 1$ if fingerprint peak j is observed in the unknown sample. Assuming that the appearance of peaks at different locations is independent, then the likelihood ratio for H_0 versus H_A is given by the probability of observing the outcome under H_A divided by the probability of observing the outcome under H_0. Specifically,

$$L = \frac{P\{\text{outcome under } H_A\}}{P\{\text{outcome under } H_0\}} \qquad (1)$$

$$= \frac{\Pi_{1\le j\le N} p_j^{x_j} \Pi_{1\le j\le N}(1-p_j)^{1-x_j}}{\Pi_{1\le j\le N} q_j^{x_j} \Pi_{1\le j\le N}(1-q_j)^{1-x_j}}$$

where *N* is the number of fingerprint peaks in the fingerprint *F*. The P {outcome under H_A} reduces to the product of the p_j's estimated from the fingerprint for those peaks in the fingerprint appearing in the sample multiplied by the product of the $(1 - p_j)$'s for those peaks in the fingerprint not appearing in the sample. Likewise, P\{outcome under H_A\} is reduced in the same manner by substituting q_j for p_j.

In practice, the log-likelihood ratio $\lambda = \log(L)$ is used where

$$\lambda = \sum_{1\le j\le N} \log\left(\frac{1-p_j}{1-q_j}\right) + \sum_{1\le j\le N} x_j \log\left[\frac{p_j(1-q_j)}{q_j(1-p_j)}\right] \qquad (2)$$

In performing the test, the following decision rule is applied:

If $\lambda \le K_c$, then accept H_0

If $\lambda > K_c$, then reject H_0

If H_0 is rejected, the reference species is determined to be present in the unknown sample.

The critical threshold K_c is determined by the desired significance level of the test as follows. The probability of falsely rejecting the null hypothesis is given by

$$\alpha = P\{\lambda > K_c | H_0\}$$

$$= P\left\{\sum_{1\le j\le N} \log\left(\frac{1-p_j}{1-q_j}\right) + \sum_{1\le j\le N} x_j \log\left[\frac{p_j(1-q_j)}{q_j(1-p_j)}\right] > K_c \,\middle|\, H_0\right\} \qquad (3)$$

The threshold K_c is set by fixing a desired false alarm (type I error) probability, α, and finding the smallest value of K_c that yields

$$P\{\lambda > K_C | H_0\} \le \alpha \qquad (4)$$

When the number of fingerprint peaks *N* is small, K_c can easily be obtained by enumerating and computing the probability of all

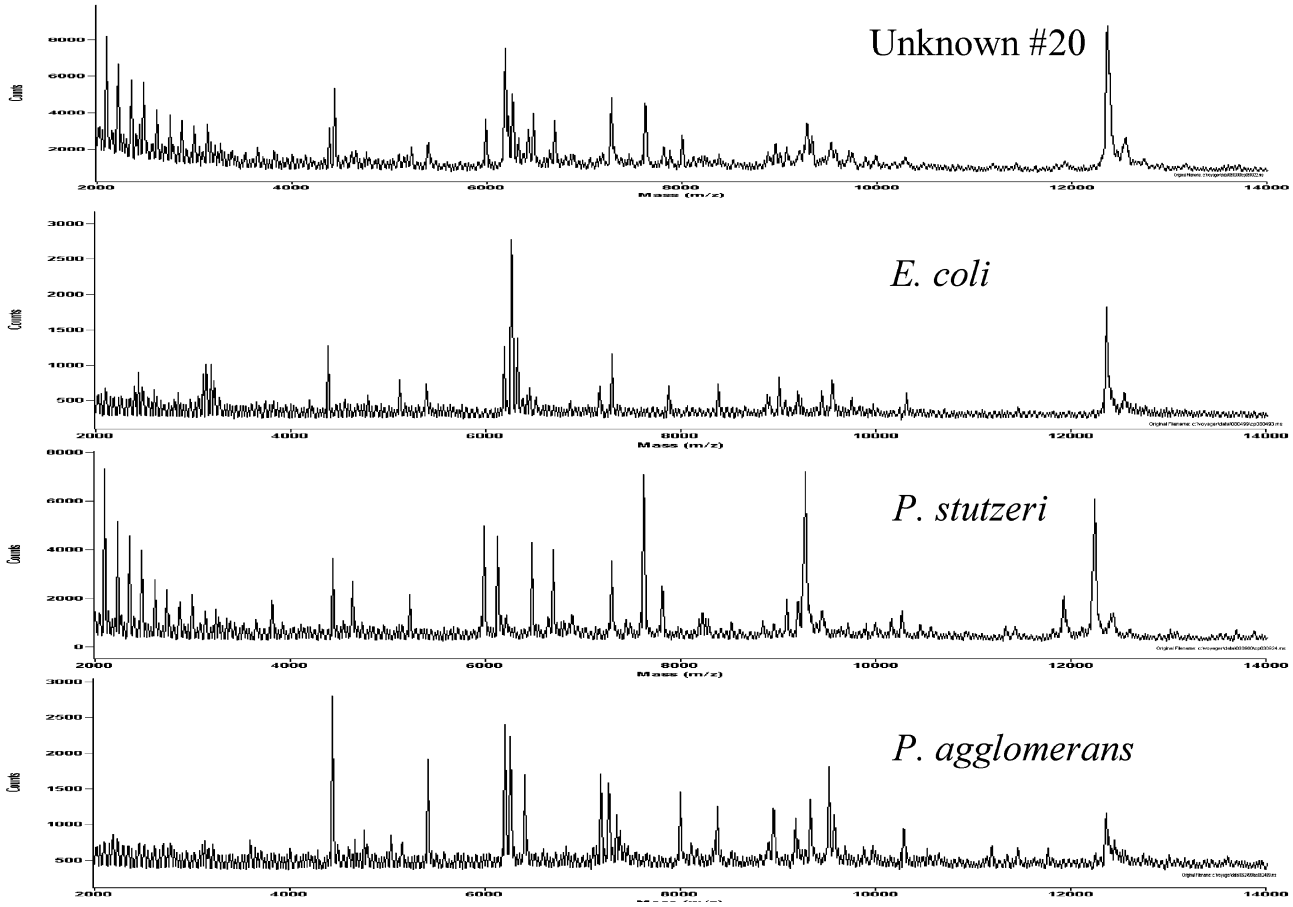

Figure 1. Representative MALDI spectrum *(m/z* 2000−14000) for each fingerprint organism and for unknown sample no. 20.

possible combinations of outcomes $\{x_j; j = 1, 2, ..., N\}$ under the null hypothesis and finding the K_c that meets eq 5. When the number of fingerprint peaks, N, is large, however, enumerating all 2^N possible outcomes becomes computationally infeasible. Therefore, when $N \geq 10$, we approximate the statistical distribution of the log-likelihood ratio λ with a normal approximation, where the mean and variance are expressed as

$$E[\lambda] = \sum_{1\leq j\leq N} \log\left(\frac{1-p_j}{1-q_j}\right) + \sum_{1\leq j\leq N} q_j \log\left[\frac{p_j(1-q_j)}{q_j(1-p_j)}\right]$$

$$\mathrm{Var}[\lambda] = \sum_{1\leq j\leq N} q_j(1-q_j) \log\left[\frac{p_j(1-q_j)}{q_j(1-p_j)}\right]^2 \qquad (5)$$

This statistical distribution of the log-likelihood ratio can then be used to select the K_c that meets eq 4. (We note that under the alternative hypothesis, the q_j and $(1 - q_j)$ immediately following the second summation in $E[\lambda]$ and $\mathrm{Var}[\lambda]$ are replaced by p_j and $1 - p_j$, respectively. By modifying eq 5 in this way, expressions for the type II error and power of the test can be estimated when N is large).

Blind Study Experimental Design. Bacterial combinations were analyzed using MALDI-MS on 10 different days. To prepare each sample for analysis, the bacteria in each combination were cultured individually and later mixed on the basis of approximately equal absorbance at 600 nm. The samples were coded, and the mass spectrometer operator and statistician performing data analysis did not know the content of the samples. Bacterial combinations used in this study are shown in Table 2.

Each blind sample was used to generate five replicate MALDI spectra. Peaks from these five replicate MALDI spectra were extracted with our PNNL peak extraction algorithm[1] and then used to form a composite peak table containing the average m/z value for each reproducible peak observed in the replicate spectra. The composite peak table was then compared to each of the library fingerprints using the original and unmodified methods.

RESULTS AND DISCUSSION

The fingerprint library used in this blind study contains fingerprints for over 30 bacteria from eight different genera, two different subspecies within a species, and seven different strains within a given species (Table 1). The organisms used in these experiments are noted in bold.

Representative single MALDI spectra from three of our MALDI-MS fingerprints for *E. coli*, *P. agglomerans*, and *Ps. stutzeri* are shown in Figure 1 along with a spectrum for one of the unknown samples (no. 20) for example. The large ion just beyond m/z 12 000 is for the singly charged cytochrome *c* that is used as an internal mass axis calibrant. The data for *Ps. stutzeri* was collected with bovine cytochrome *c* (MW = 12 231) instead of the usual equine cytochrome *c* (MW = 12 360), which explains the mass shift observed in Figure 1. As discussed previously,[1] our fingerprints are created from compiling 60 replicate spectra

Table 2. Blind-Study Bacterial Combinations

sample	combination
	Day 1
1	*B. subtilis, B. thuringiensis, Ps. stutzeri, S. marcescens*
2	*P. agglomerans, Ps. putida*
3	*B. atrophaeus, B. subtilis, E. coli*
4	*B. thuringiensis, P. agglomerans*
5	*B. thuringiensis, Ps. putida, S. marcescens*
	Day 2
1	*B. sphaericus, E. coli, P. agglomerans*
2	*B. subtilis, Ps. putida*
3	*B. atrophaeus, B. cereus*
4	*B. cereus, B. subtilis, P. agglomerans*
5	*B. subtilis, Ps. stutzeri*
	Day 3
1	*P. agglomerans, Ps. putida*
2	*B. cereus, B. sphaericus, Ps. putida*
3	*B. subtilis, B. thuringiensis, Ps. stutzeri, S. marcescens*
4	*B. sphaericus, B. subtilis, S. marcescens*
5	*B. cereus, P. agglomerans*
	Day 4
1	*B. sphaericus, Ps. putida*
2	*B. cereus, Ps. putida, S. marcescens*
3	*B. cereus, B. thuringiensis, E. coli*
4	*B. sphaericus, E. coli*
5	*E. coli, P. agglomerans, Ps. stutzeri*
	Day 5
1	*B. atrophaeus, E. coli*
2	*B. sphaericus, B. thuringiensis, Ps. putida*
3	*B. sphaericus, B. thuringiensis, E. coli*
4	*B. atrophaeus, B. cereus*
5	*B. cereus, P. agglomerans*
	Day 6
1	*B. atrophaeus, B. sphaericus, B. subtilis*
2	*P. agglomerans, Ps. putida*
3	*Ps. putida, S. marcescens*
4	*B. atrophaeus, B. cereus*
5	*B. atrophaeus, B. thuringiensis*
	Day 7
1	*B. atrophaeus, B. cereus*
2	*B. atrophaeus, Ps. stutzeri*
3	*E. coli, S. marcescens*
4	*B. subtilis, B. thuringiensis, Ps. stutzeri, S. marcescens*
5	*B. thuringiensis, Ps. stutzeri*
	Day 8
1	*B. subtilis, S. marcescens*
2	*B. subtilis, B. thuringiensis, Ps. putida*
3	*P. agglomerans, Ps. stutzeri*
4	*P. agglomerans, Ps. putida*
5	*B. atrophaeus, E. coli, Ps. putida*
	Day 9
1	*E. coli, Ps. putida, Ps. stutzeri*
2	*B. cereus, P. agglomerans*
3	*B. atrophaeus, P. agglomerans, S. marcescens*
4	*B. sphaericus, S. marcescens*
5	*B. atrophaeus, Ps. stutzeri, S. marcescens*
	Day 10
1	*B. subtilis, E. coli*
2	*B. cereus, B. sphaericus, Ps. stutzeri*
3	*B. sphaericus, B. thuringiensis*
4	*B. cereus, P. agglomerans*
5	*B. subtilis, B. thuringiensis, Ps. stutzeri, S. marcescens*

collected over 3 different days on separate cultures for each organism. Ions appearing in at least 70% of these replicate spectra are included in the fingerprint for that organism. The fingerprint ions for the three organisms shown in Figure 1 are given in Table 3. The standard deviations in the m/z value for each ion are

Table 3. MALDI-MS Fingerprints of Three Organisms and Composite Spectrum for Unknown Sample No. 20[a]

E. coli W3110		*P. agglomerans*		*Ps. stutzeri* 39524		unknown sample no. 20		
m/z	freq	m/z	freq	m/z	freq	m/z		
2431	0.74	**6401**	0.70	**2084**	1.00	**2085**	6255	9882
3081	0.86	7264	0.70	**2214**	1.00	2186	**6258**	9971
3125	0.98	**7989**	0.70	**2343**	1.00	**2214**	**6317**	9985
3165	0.97	9175	0.70	**2472**	1.00	**2343**	6400	9988
3207	0.91	**9318**	0.70	**2601**	1.00	2387	6404	10163
3638	0.74	**9507**	0.98	**2730**	1.00	**2431**	**6409**	10280
4366	1.00			**2859**	1.00	**2472**	**6464**	11189
5097	0.97			**2988**	1.00	2512	**6627**	11193
5150	0.74			**3118**	0.87	**2602**	**6683**	11441
5382	0.83			**3247**	0.77	**2731**	7182	11929
6256	1.00			**3637**	0.87	2772	**7273**	12356
6316	0.98			**3809**	1.00	2775	7443	12536
6412	0.93			**4421**	0.97	**2860**	**7616**	12540
6857	0.79			**4634**	0.98	**2989**	**7809**	12544
7159	0.98			**5212**	0.82	3039	7813	12735
7274	1.00			**5975**	0.98	**3080**	**7872**	13153
7871	1.00			**6463**	0.98	**3120**	**7990**	
8370	0.98			**6627**	0.92	**3165**	**8220**	
8877	0.91			**6683**	1.00	**3206**	**8368**	
8994	0.98			6795	0.75	**3248**	8523	
9190	0.90			6882	0.72	3377	**8825**	
9226	0.84			**7274**	0.98	3506	**8877**	
9428	0.95			7438	0.92	3510	8952	
9535	0.79			**7616**	0.97	**3637**	**9071**	
9554	0.79			**7808**	0.98	**3811**	9075	
10299	1.00			**8219**	0.95	4123	9195	
11450	0.79			8518	0.88	**4366**	9199	
				8826	0.93	**4422**	9201	
				9073	0.97	4605	**9266**	
				9201	0.97	**4636**	9270	
				9267	0.97	4755	9276	
				9701	0.78	**5098**	**9318**	
				10159	0.97	**5212**	9323	
				10266	0.93	5397	9504	
				10451	0.80	5401	**9508**	
				10558	0.77	5592	9564	
				11412	0.80	**5975**	**9703**	
				11925	0.98	6181	9733	
				13871	0.77	6201	9738	

[a] Bold ions appeared in mixture shown in Figure 1.

computed from the data and are all <2, with the exception of five of the higher-molecular-weight ions between m/z 10 000 and 14 000.

Our automated bacterial identification approach is illustrated for sample no. 20 with the results of the blinded analysis of the unknown mixture given in Figure 2. The top spectrum in Figure 2 is the composite ion spectrum of the unknown sample obtained by our peak detection algorithms from five replicate MALDI spectra of that sample[1]. Our automated bacterial identification algorithm correctly identified this sample as a mixture of *E. coli*, *P. agglomerans*, and *Ps. stutzeri*. The ion table generated automatically for this unknown sample no. 20 with our peak detection algorithm is given in Table 3 to compare to the fingerprint ions for each organism. For comparison, the MALDI-MS fingerprints for each of these three organisms between m/z 5000 and 10 000 are also shown in Figure 2. The top composite spectrum exhibits ions from each of these components as well as some additional ions not assigned as a reproducible fingerprint ion for any of the organisms. Because our algorithms rely on extracting fingerprint ions from the test spectrum, the automated analysis of this mixture

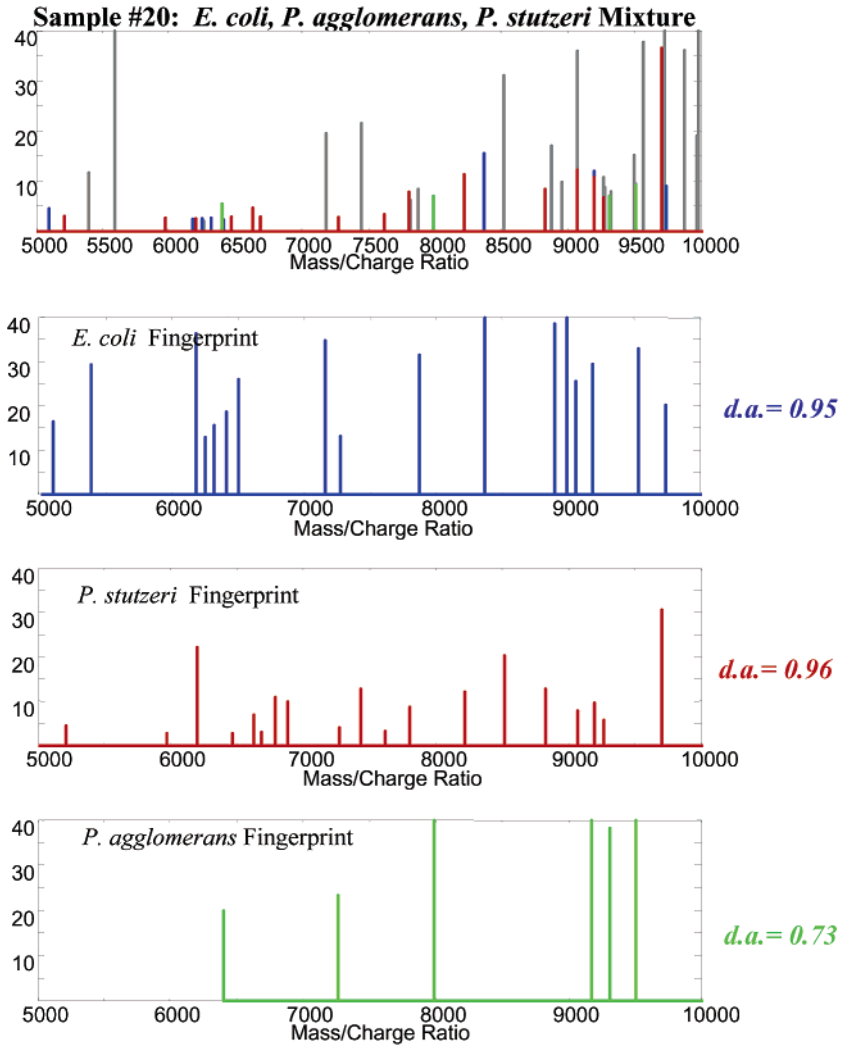

Figure 2. Microorganism identification of a mixed culture using MALDI-MS.

Table 4. Genus-Level Blind Study Results

	original method		modified method	
genus	% correctly identified	% correctly eliminated	% correctly identified	% correctly eliminated
Bacillus	100	56	100	89
Escherichia	100	51	100	56
Pantoea	100	94	100	97
Pseudomona	100	100	100	100
Serratia	100	84	100	89
Shewanella	*a*	100	*a*	100
Staphylococcus	*a*	100	*a*	100
Yersinia	*a*	52	*a*	74

[a] Not analyzed in blind study.

Table 5. Species-Level Blind Study Results

	original method		modified method	
species	% correctly identified	% correctly eliminated	% correctly identified	% correctly eliminated
B. atrophaeus	100	74	92	76
B. cereus	100	59	100	70
B. subtilis	100	24	100	73
B. simplex	*a*	96	*a*	100
B. thuringiensis	100	16	100	46
E. coli	100	51	100	56
P. agglomerans	100	94	100	97
Ps. aeruginosa	*a*	92	*a*	96
Ps. putida	100	100	100	100
Ps. stutzeri	100	100	100	100
S. marcescens	100	84	100	89
Sh. alga	*a*	100	*a*	100
Staph. aureus	*a*	100	*a*	100
Staph. epidermidis	*a*	100	*a*	100
Y. enterocolitica	*a*	52	*a*	74

[a] Not analyzed in blind study.

with the original algorithm identifies all three microbial species present in this sample with a degree of association for each match greater than 0.7 (with d.a. = 1 being an ideal match). A degree of association greater than 0.15 is considered a positive match, or presence in the sample.

Detailed analysis of all of the mixtures analyzed is presented in a stepwise approach. Previous experience has shown that the existing identification algorithms can effectively identify organisms to the genus level, but they can have some difficulty differentiating organisms at the species and subspecies level.

To better determine the capabilities of our approach, results are reported at the genus, species, and strain level. In particular, for a given genus (species, strain), true positives are reported as the fraction of blind samples in which the identification algorithm

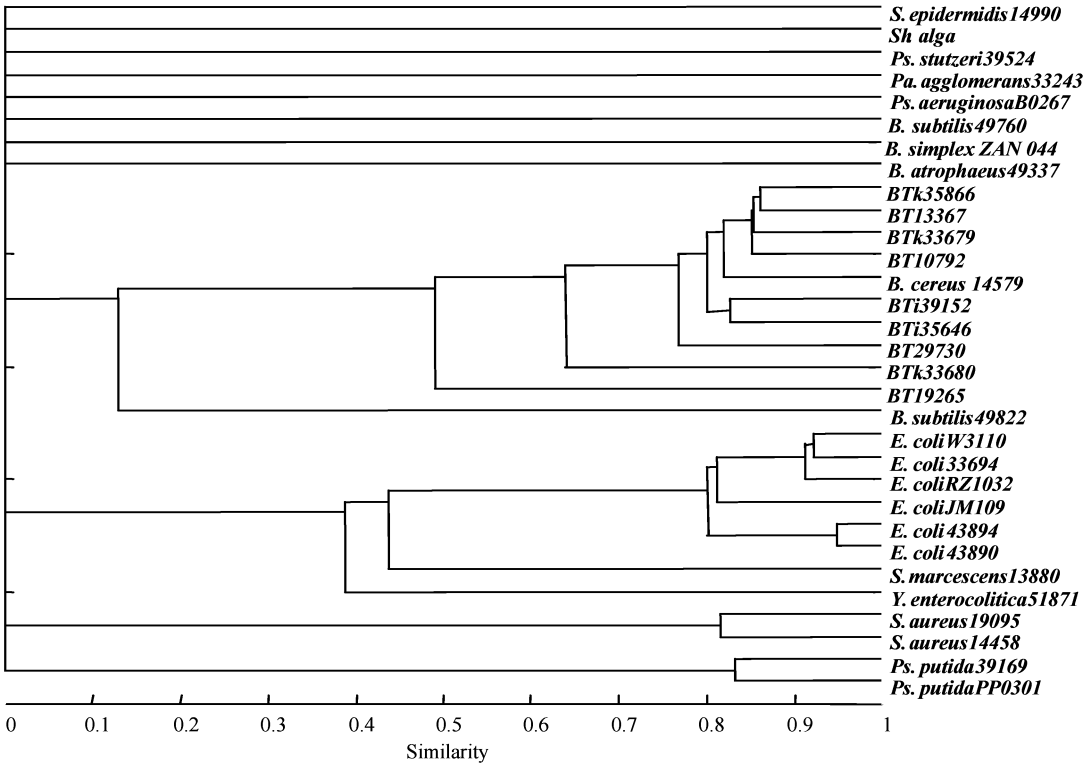

Figure 3. Dendrogram of MALDI-MS bacterial fingerprints.

determined any member of that genus (species, strain) to be present when a member is actually present in the sample. Analogously, true negatives are reported as the fraction of blind samples in which the identification algorithm eliminated that genus (species, subspecies, strain) when no member is present in the sample. The modified method performs uniformly the same or better than the original method (Tables 4 and 5). Therefore, we focus our discussion on the modified method; however, the general observations and conclusions are the same for both methods.

Table 4 contains the results of the blind sample analysis at the genus level. In Table 4, no false negatives are reported for either identification method (i.e., percent correctly identified is 100% for all genera tested). This suggests that MALDI-MS and the identification methods developed here can sensitively detect the presence of bacteria at the genus level. On the other hand, a number of false positives are reported (i.e., percent correctly eliminated is less than 100%) for several of the genera. These false positives tend to be caused by an apparent similarity between MALDI-MS fingerprints of *E. coli*, *S. marcescens*, and *Y. enterocolitica*. For example, the vast majority of *E. coli* false positives and all of the *Y. enterocolitica* false positives occur for samples containing either *E. coli*, *S. marcescens*, or both. Of the 12 samples containing *S. marcescens* and not *E. coli*, the genus *Escherichia* is identified as being present in all 12, and *Y. enterocolitica* is identified as being present in 9 of them. Of the 10 samples containing *E. coli* and not *S. marcescens*, *S. marcescens* is identified as present in all 10, and *Y. enterocolitica* is identified as present in 4. This apparent cross-genera similarity between MALDI-MS fingerprints is supported by Figure 3, which plots a dendrogram of the fingerprint library used in the study. The dendrogram is constructed using the degree of association[1] and classical hierarchical clustering techniques that help visualize similarities between different fingerprints. In general, different MALDI-MS fingerprints are combined into groups one at a time until only a single group remains. Conceptually, the two fingerprints with the highest degree of association are joined together first, followed by those with the next highest degree of association, and so on. When a "new" fingerprint is compared to a group of fingerprints (or two groups of fingerprints are compared), the similarity is estimated from the individual degrees of association between the group members and the new fingerprint to be added (or the members of the group to be added). The lines connecting different fingerprints represent the order in which two different fingerprints or groups of fingerprints are joined. Most of the different genera in the fingerprint library are not grouped together (i.e., degree of association is 0). The various *E. coli* strains have a high degree of association among each other and are therefore grouped at a high similarity ($\sim$0.8). Additionally, the *E. coli* group have a higher-than-expected similarity with *S. marcescens* and *Y. enterocolitica*. This suggests that the MALDI-MS fingerprints for *E. coli*, *S. marsescens*, and *Y. enterocolitica* share some common peaks and explains, in part, the inability of the identification algorithms in the current development stage to differentiate between *Serratia* and *Escherichia* in the blind study.

Blind-study results for the species and strain level are provided in Tables 5 and 6, respectively. In these tables, the true positive rate is consistently high with two exceptions. First, the percent of *B. subtilis* 49760 correctly identified is 0% at the strain level. Blind samples containing *B. subtilis* 49760 are identified as *B. subtilis* 49822. Investigation into this problem discovered a contamination problem with our *B. subtilis* 49760 stock culture and fingerprint data. In retrospect, the location of *B. subtilis* 49760 in the dendrogram shown in Figure 3 should have been suspicious

Table 6. Strain-Level Blind-Study Results

strain	original method		modified method	
	% correctly identified	% correctly eliminated	% correctly identified	% correctly eliminated
B. atrophaeus (ATTC 49337)	100	74	92	72
B. cereus (ATCC 14579)	100	59	100	70
B. subtilis (ATCC 49760)	0	100	0	100
B. subtilis (ATCC 49822)	*a*	18	*a*	54
B. simplex (ZAN044)	*a*	96	*a*	100
B. thuringiensis (ATCC 10792)	*a*	24	*a*	44
B. thuringiensis (ATCC 13367)	100	38	100	57
B. thuringiensis (ATCC 19265)	*a*	10	*a*	40
B. thuringiensis (ATCC 29730)	*a*	34	*a*	46
B. thuringiensis kurstaki (ATCC 33679)	*a*	26	*a*	44
B. thuringiensis kurstaki (ATCC33680)	*a*	28	*a*	40
B. thuringiensis kurstaki (ATCC 35866)	*a*	38	*a*	46
B. thuringiensis israel. (ATCC 35646)	*a*	18	*a*	42
B. thuringiensis israel. (ATCC 39152)	*a*	34	*a*	40
E. coli (ATCC 33694)	*a*	44	*a*	48
E. coli (ATCC 43890)	*a*	70	*a*	76
E. coli (ATCC 43894)	*a*	68	*a*	70
E. coli (JM109)	*a*	48	*a*	58
E. coli (RZ1032)	*a*	50	*a*	62
E. coli (W3110)	100	62	100	82
P. agglomerans (ATTC 33243)	100	94	100	96
Ps. aeruginosa (B0267)	*a*	92	*a*	97
Ps. putida (PP0301)	86	100	86	100
Ps. putida (ATCC 39169)	*a*	72	*a*	72
Ps. stutzeri (ATCC 39524)	100	100	100	100
S. marcescens (ATCC 13880)	100	84	100	89
Sh. alga	*a*	100	*a*	100
Staph. aureus (ATCC 14458)	*a*	100	*a*	100
Staph. aureus (ATCC 19095)	*a*	100	*a*	100
Staph. epidermidis (ATCC 14990)	*a*	100	*a*	100
Y. entercolitica (ATCC 51871)	*a*	52	*a*	74

[a] Not analyzed in blind study.

since it did not show the expected similarity to the other *B. subtilis* strains. Once the blind study data was analyzed and the problem discovered, we could not adjust the fingerprint for this analysis and still maintain the blind analysis. This problem that we discovered during the blind study clearly stresses the need for careful evaluation of cultures for purity and correct identification with other microbiology methods during the fingerprinting stage.

The second problem with true positive identification was for *Ps. putida* PP0301. In this case, the strain was falsely identified as *Ps. putida* ATCC 39169 in several cases. This is not surprising, since the fingerprints of these two strains are so similar, as represented by their proximity in the dendrogram in Figure 3.

The percent of samples correctly eliminated decreases from the genus to the strain level. We note that 84% of all the false positives in the study occurred when a different organism of the same genus was present in the sample. Consider the species-level results of Table 5. Nine of the blind samples resulted in false positives for *B. atrophaeus*. *B. subtilis* appeared in all nine of these blind samples. *B. cereus* was erroneously detected in 10 of the blind samples, and 9 out of 10 of these samples contained *B. thuringiensis*. Ten of the 20 *B. thuringiensis* false positives occurred when *B. cereus* appeared in the sample, 5 occurred when *B. sphaericus* (not in the library) appeared in the sample, and an additional 3 occurred when both *B. thuringiensis* and *B. cereus* occurred in the sample. The cross-genera confusion between *Escherichia*, *Yersinia*, and *Serratia* accounts for an additional 13% of all false positives in this blind study. The remaining 3% of false positives are spurious and likely due to chance alone. This tendency to confuse organisms within a genus is expected when considering the dendrogram in Figure 3. In particular, the species and strains within a genus tend to group tightly together in the dendrogram, indicating a strong similarity between MALDI-MS fingerprints at this level. Therefore, we expect the identification algorithms to have more difficulty differentiating organisms within a genus than between genera, leading to an increase in false positives (or a corresponding decrease in the percent correctly eliminated) for strains and species within a genus. We are currently developing an alternative data analysis approach to be more sensitive to species and strain differentiation.

SUMMARY

Results demonstrating the potential of MALDI mass spectrometry and automated fingerprint extraction algorithms to analyze microbial mixtures in a laboratory setting are presented. The components present in 50 different microbial mixtures were correctly identified in a double-blind study at the genus level in all samples and to the species level in the majority of the samples. As expected with the current algorithms, strain-level identification is challenging at best. Improvements to the algorithms are in progress, and initial testing with the blind-study data shows significant improvement in identification at the strain level and better elimination of closely related strains. This new approach will be presented in detail when completely developed. As the MALDI-MS fingerprint library expands, continued fine-tuning and

improvements will be made with the data analysis algorithms to improve the performance in the specificity of the identification. One key advantage to our current approach is that a family association can be made between unknown strains not currently in the fingerprint library and closely related organisms that are present in the fingerprint library.

ACKNOWLEDGMENT

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 U.S. Department of Energy through the Laboratory Directed Research and Development program. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC06-76RL01830.

Received for review June 10, 2002. Accepted September 26, 2002.

AC0203847