Anal. Chem. **2002,** *74,* 3174−3182

Mass Spectrometric Methods for Generation of Protein Mass Database Used for Bacterial Identification

Zhengping Wang,[†] Kevin Dunlop,[†] S. Randolph Long,[‡] and Liang Li*,[†]

Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2G2, and Edgewood Chemical Biological Center, SCBRD-RT, Aberdeen Proving Ground, Maryland 21010

The availability of a suitable database is critical in a proteomic approach for bacterial identification by mass spectrometry (MS). The major limitation of the present public proteome database is the lack of extensive low-mass bacterial protein entries with masses experimentally verified for most bacteria. Here, we present a method based on mass spectrometry to create protein mass tables specifically tailored for bacterial identification. Several issues related to the detection of bacterial proteins for the purpose of database creation are addressed. Three species of bacteria, namely, *Escherichia coli*, *Bacillus megaterium*, and *Citrobacter freundii*, which can be found in the ambient environment, were chosen for this study. Direct matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS analysis of each bacterial extract reveals 20−29 protein components in the mass range from 2000 to 20 000 Da. HPLC fractionation of bacterial extracts followed by off-line MALDI-TOF analysis of individual fractions detects 156−423 components. Analysis of the extracts by HPLC/electrospray ionization MS shows the number of detectable proteins in the range of 46−59. Although a number of components were common to the three detection schemes employed, some unique components were found using each of these techniques. In addition, for *E. coli* where a large proteome database exists in the public domain, a number of masses detected by the mass spectrometric methods do not match with the proteome database. Compared to the public proteome database, the mass tables generated in this work are demonstrated to be more useful for bacterial identification in an application where the bacteria of interest have limited protein entries in the public database. The implication of this work for future development of a comprehensive mass database is discussed.

Mass spectrometry (MS) can potentially become a very rapid and sensitive tool for bacterial identification. Current research in developing MS for bacterial identification is focusing on the detection of bacterial proteins as biomarkers. Low-mass bacterial proteins can be readily detected by matrix-assisted laser desorption/ionization (MALDI) MS from either cell lysates[1−12] or whole cells.[13−24] Liquid chromatography/electrospray ionization (LC/ESI) MS has also been applied to the analysis of low-mass bacterial proteins from cell extracts.[25−28] One approach of bacterial iden-

* To whom correspondence should be addressed. E-mail: Liang.Li@ualberta.ca. Fax: 1-780-492-8231.

[†] University of Alberta.

[‡] Edgewood Chemical Biological Center.

(1) Cain. T. C.; Lubman, D. M.; Weber, W. J. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 1026−1030.
(2) Krishnamurthy, T, Ross, P. L.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 883−888.
(3) Chong, B. E.; Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1900−1908.
(4) Van Adrichem, J. H. M.; Bornsmen, K. O.; Conzelmann, H.; Gass, M. A. S.; Eppenberger, H.; Kresbach, G. M.; Ehrat, M.; Leist, C. H. *Anal. Chem.* **1998**, *70*, 923−930.
(5) Easterling, M. L.; Colangelo, C. M.; Scott, R. A.; Ameter, I. J. *Anal. Chem.* **1998**, *70*, 2704−2709.
(6) Wang, Z.; Russon, L. M.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 456−464.
(7) Dai, Y. Q.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 73−78.
(8) Holland, R. D.; Duffy, C. R.; Rafii, F.; Sutherland, J. B.; Heinze, T. M.; Holder, C. L.; Voorhees, K. J.; Lay Jr. J. O. *Anal. Chem.* **1999**, *71*, 3226−3230.
(9) Arnold, R. J.; Reilly, J. P. *Anal. Biochem.* **1999**, *269*, 105−112.
(10) Birmingham, J.; Demirev, P.; Ho, Y.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 604−606.
(11) Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Anal. Chem.* **1999**, *71*, 3894−3900.
(12) Domin, M. A.; Welham, K. J.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 222−226.
(13) Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay. J. O., Jr. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1227−1232.
(14) Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1992−1996.
(15) Claydon, M. A.; Davey, S. N.; Edwards-Jones, V.; Gordon, D. B. *Nat. Biotechnol.* **1996**, *14*, 1584−1586.
(16) Welham, K. J.; Domin, M. A.; Scannell, D. E.; Cohen, E.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 176−180.
(17) Haag, A. M.; Taylor, S. N.; Johnston, K. H.; Cole, R. B. *J. Mass Spectrom.* **1998**, *33*, 750−756.
(18) Arnold, R. J.; Karty, J. A.; Ellington, A. D.; Reilly, J. P. *Anal. Chem.* **1999**, *71*, 1990−1996.
(19) Saenz, A. J.; Pertersen, 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.
(20) Lynn, E. C.; Chung, M.; Tsai, W.; Han, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 2022−2027.
(21) Leenders, F.; Stein, T. H.; Kablitz, B.; Franke, P.; Vater, J. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 943−949.
(22) Winkler, M. A.; Uher, J.; Cepa, S. *Anal. Chem.* **1999**, *71*, 3416−3419.
(23) Evason, D. J.; Claydon, M. A.; Gordon, D. B. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 669−672.
(24) Madonna, A. J.; Basile, F.; Ferrer, I.; Meetani, M. A.; Rees, J. C.; Voorhees, K. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2220−2229.

10.1021/ac015725f CCC: $22.00 © 2002 American Chemical Society
Published on Web 05/03/2002

tification by MS relies on producing mass spectral fingerprints of proteins and then comparing unknown spectra to the archived sets.[17,29−33] This approach, however, suffers from a lack of mass spectral reproducibility even in cases where identical cultures are used in replicate experiments.

An alternative approach, proposed by Fenselau and co-workers, involves detecting and cataloguing the masses of proteins expressed by bacteria.[34] These protein masses would then be searched against the public database.[34−36] Potential matches would be retrieved with statistically significant scores assigned to the possible candidates. This method was found to be useful for identifying *Helicobacter pylori* from a list of bacteria whose individual proteome size in the mass range from 4 to 20 kDa exceeds about 200 entries.[36] As was pointed out,[35] the usefulness of this approach is very much dependent on the extent and quality of the database available, as well as the mass data obtained from the unknown.

In principle, this statistical approach is not limited to the use of a public proteome database for mass comparison. It should also be useful in combination with other proteome mass databases for bacterial identification. We are interested in developing a mass spectrometric approach for applications where a target bacterium needs to be differentiated from a handful of background bacteria. In these applications, the bacteria of interest may not have a large number of low-mass protein entries in the public database. As an example, a model system involved in the differentiation of three bacteria, namely, *Escherichia coli*, *Bacillus megaterium*, and *Citrobacter freundii,* which can be found in the ambient environment, has been investigated. We illustrate that the current public database, with its very limited number of low-mass protein entries for two of the bacteria, is not adequate for bacterial differentiation based on mass data comparison. We present our studies of using mass spectrometric techniques to create protein mass tables that are tailored for bacterial identification and demonstrate their utility for differentiating bacteria in a small data set consisting of these three environmental bacteria.

EXPERIMENTAL SECTION

Materials. Bacterial cells, *E. coli* (ATCC9637), *B. megaterium* (ATCC14581), and *C. freundii* (ATCC8090), used in this work were either from the Edgewood Chemical Biological Center (ECBC), Aberdeen Proving Ground, MD, or cultured in-house. The ECBC bacterial cells were grown in Luria Broth (LB) (BBL, Becton Dickinson, Cockeysville, MD) at 35 °C and harvested after sufficient growth had occurred (usually 18−48 h). The cells were washed with sterile water and freeze-dried. The lyophilized cells were shipped to the University of Alberta under dry ice. Bacterial samples grown at the University of Alberta laboratory were incubated in LB (BBL, Becton Dickinson) at 30 °C with shaking. The *E. coli* cells were harvested at 36 h, and *B. megaterium* cells were harvested at 16 h; the cells were then washed with sterile water, lyophilized, and stored below 0 °C before extraction. HPLC grade acetonitrile, 2-propanol, and glacial acetic acid were from Fisher Scientific Canada (Edmonton, AB, Canada). Water was obtained from a Milli-Q Plus purification system (Millipore Corp., Bedford, MA). Bradykinin, bovine insulin chain B, horse cytochrome *c*, trifluoroacetic acid (TFA), and α-cyano-4-hydroxycinnamic acid (HCCA) were from Sigma-Aldrich-Fluka (Oakville, ON, Canada). HCCA was purified by recrystallization from ethanol prior to use.

Sample Preparation. Bacterial extracts were prepared by a solvent suspension method as described previously.[6] Briefly, for LC/ESI and LC/off-line MALDI analysis, ~25 mg of lyophilized bacterial cells (*E. coli*, *B. megaterium* or *C. freundii*) was suspended in 1 mL of 0.1% TFA, vortexed for ~3 min, and centrifuged at 11750*g*. The supernatant (i.e., the clear solution above the cell debris) was then transferred into a fresh vial. This extraction process was repeated 3 times per sample, and the extracts were pooled to maximize the extraction efficiency. The pooled extracts were filtered using Mirocon-3 filters with 3000-Da molecular mass cutoff (Millipore) and then concentrated to ~0.5 mL by Speed-Vac. For direct MALDI analysis, ~1 mg of lyophilized bacterial cells was extracted in 500 μL of 0.1% TFA by vortexing, and the supernatant was directly taken for analysis without further cleaning.

HPLC. Solvent delivery and separations were performed on an Agilent (Palo Alto, CA) HP1100 HPLC system. The solvents used for reversed-phase HPLC separation were acetonitrile (A) and water (B) with 0.1% (v/v) TFA in both phases. The gradient elution profile was optimized for each bacterial extract. For *E. coli* and *B. megaterium*, the gradient was 2−20% B in 10 min, 20−40% B in 30 min, 40−55% B in 5 min, and 55−90% B in 15 min. For *C. freundii*, the gradient was 2−20% in 10 min, 20−40% B in 20 min, 40−55% B in 30 min, and 55−90% B in 5 min. For on-line LC/ESI, 40 μL of bacterial extract, as prepared using the procedure described above, was separated on a 150 × 2.1 mm i.d. C_8 column (5-μm particles with 300-Å pore size; Vydac, Hesperia, CA) at a flow rate of 200 μL/min. For LC/off-line MALDI, 100 μL of bacterial extract was separated on a 250 × 4.6 mm i.d. C_8 column (5-μm particles with 300-Å pore size, Vydac) at a flow rate of 500 μL/min. Fractions were collected at 1-min intervals using a Gilson FC 203B fractionation collector (Gilson, Middleton, WI).

ESI. The HPLC effluent was analyzed with a HP 1100 MSD quadrupole mass spectrometer (Agilent). The m/z range from 500 to 3000 was scanned in 1.90 s, and the ions were detected with a high-energy dynode detector. Control of both the HPLC and MS systems was accomplished with HP ChemStation software. To compensate for the signal suppressing effect from TFA in the mobile phase during ESI analysis, glacial acetic acid was added

(25) MacNair, J. E.; Opiteck, G. J.; Jorgenson, J. W.; Moseley, A. M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1279−1285.

(26) Krishnamurthy, T.; Davis, M. T.; Stahl, D. C.; Lee, T. D. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 39−49.

(27) Dalluge, J. J.; Reddy, P. *Biotechniques* **2000**, *28*, 156−160.

(28) Dunlop, K. Y.; Li, L. *J. Chromatogr., A* **2001**, *925*, 123−132.

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

(30) 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.

(31) Nisson, C. L. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 1067−1071.

(32) 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.

(33) 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.

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

(35) Pineda, F. J.; Lin, J. S.; Fenselau, C.; Demirev, P. A. *Anal. Chem.* **2000**, *72*, 3739−3744.

(36) Demirev, P. A.; Lin, J. S.; Pineda, F. J.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 4566−4573

Table 1. $(M + H)^+$ from *E. coli* Extract by On-Line LC/ESI-MS

2008.0	**2123.0**	**2140.0**	**2375.3**	**2416.2**	**2431.5**	**2504.9**	2735.9	**3509.6**	**3624.7**
3793.1	4480.8	**5037**	**5097**	5171	**5550**	**6255**	**6316**	6330	**7140**
7273	**7707**	**7782**	**7855**	**9064**	**9192**	**9209**	**9226**	**9231**	**9265**
9518	**9536**	**9610**	9684	**9740**	**10386**	**10459**	11224	11297	**11782**
13094	15693	15767							

Table 2. $(M + H)^+$ from *B. megaterium* Extract by On-Line LC/ESI-MS

3047.5	**3186.9**	**4379.2**	**4741.1**	**4816.3**	**6262**	**6276**	6316	**6336**	6352
6389	**6393**	**6449**	**6578**	**6897**	**7111**	7157	**7280**	**7425**	**7451**
7467	**7519**	7649	**7711**	**9334**	**9351**	**9620**	**9694**	**9747**	9754
9768	9829	9884	10026	**10045**	**10453**	10696	**11063**	**11538**	**11612**
11726	12046	12120	12408						

Table 3. $(M + H)^+$ from *C. freundii* Extract by On-Line LC/ESI-MS

2063.1	**2391.3**	**2431.0**	2473.2	**2515.2**	**2676.0**	**2793.0**	**2852.8**	3038.5	**3236.6**
3449.6	**3565.0**	**3580.8**	**4007.3**	**4420.5**	**4436.7**	**4504.2**	**5240**	**5254**	**5935**
6530	**6721**	**7335**	**7736**	**7752**	**8278**	8548	9057	**9196**	9224
9523	**9527**	**10107**	**10301**	**10682**	11162	11252	11675	11691	**11750**
11870	11956	**12157**	12239	16745	**17178**				

to the column effluent at 100 μL/min by a polyetheretherketone (PEEK) "Y" connector and a syringe pump (Cole Parmer, Vernon Hills, IL). The "Y" was connected to the electrospray interface by a 30-cm piece of PEEK tubing (0.005-in. i.d.). It is found that ~10 times signal enhancement can be achieved by the postcolumn addition of acetic acid. Early results showed the total ion chromatogram (TIC) intensities varied according to the voltage at the capillary exit. A large voltage drop in the region between the capillary exit and the first skimmer produced more intense signals in the TIC. However, it also had the effect of stripping charges from the analytes as well as producing severe fragmentation due to collision-induced dissociation (CID). To overcome the detrimental CID effects caused by a large and constant fragmentation voltage, the voltage was varied as the quadrupole scanned across the selected mass range. The optimized voltage ramp was found to be at m/z 500, 60 V; at m/z 1000, 120 V; at m/z 3000, 220 V. The resulting mass spectra showed a greater number of peaks and a higher signal-to-noise ratio compared to those obtained using a constant fragmentation voltage.

MALDI. The MALDI results were obtained in a time-lag focusing MALDI-TOF mass spectrometer which has been described in detail elsewhere.[37] A 337-nm and 3-ns pulse width of laser beam from a nitrogen laser (model VSL 337ND, Laser Sciences Inc., Newton, MA) was used for desorption. In general, 50−100 laser shots (3−5 μJ pulse energy) were averaged to produce a mass spectrum. Spectra were acquired and processed with Hewlett-Packard supporting software and reprocessed with the Igor Pro software package (WaveMetrics, Inc., Lake Oswego, OR).

For LC/off-line MALDI analysis, the HPLC fractions were concentrated by 50 times to ~10 μL before mixing with matrix for analysis. A two-layer method was used for MALDI sample preparation.[38] HCCA was used as the matrix. About 1 μL of 0.1 M HCCA in acetone/water (99:1, by volume) was applied to the MALDI probe tip to quickly form the first layer. For the second layer, the sample solution was mixed in 1:1 ratio with saturated HCCA in formic acid/2-propanol/water (1:2:3, by volume). About 0.6−1 μL of the second layer solution was then applied onto the first layer and allowed to dry. On-probe washing of the MALDI sample with water was performed to remove the salts. External mass calibration was done in the mass range of 2−20 kDa using insulin chain B, horse cytochrome *c*, and its multiply charged species and dimer.

RESULTS

LC/ESI. Tables 1−3 list the protein masses detected from the three bacterial samples. In this work, protein mass refers to the mass of a singly protonated protein ion. The boldface mass values correspond to those found using LC/off-line MALDI and the underlined values are those whose molecular masses match with proteins in the SWISS-PROT or TrEMBL databases. The mass measurement accuracy in LC/ESI MS is typically better than 0.02%.

Direct MALDI. Figure 1 shows the MALDI spectra of the extracts from the three bacteria. The mass values of singly protonated molecular ions are listed in Tables 4−6. The boldface masses match the LC/off-line MALDI data, and the underlined masses matched the proteins in the corresponding proteome database. With external calibration, the mass measurement accuracy is normally better than 0.05%.

LC/Off-Line MALDI. The mass values of the protonated protein ions from the *E. coli* sample detected by LC/off-line MALDI are listed in Table 7, and the mass values from *B. megaterium* and *C. freundii* are listed in Tables SA and SB (Supporting Information). The boldface masses match those found by LC/ESI (Tables 1−3). In Table 7, the underlined masses are the ones matching with those found in the proteome database of *E. coli*.

For the MALDI analysis, clusters of peaks with mass difference of 16 Da in the molecular ion region of the protein were sometimes observed. These peaks correspond to the methionine-oxidized

[37] Whittal R. M.; Li, L. *Anal. Chem.* **1995**, *67*, 1950−1954.

[38] Dai, Y.; Whittal, R. M.; Li, L. *Anal. Chem.* **1996**, *68*, 2494−2500.

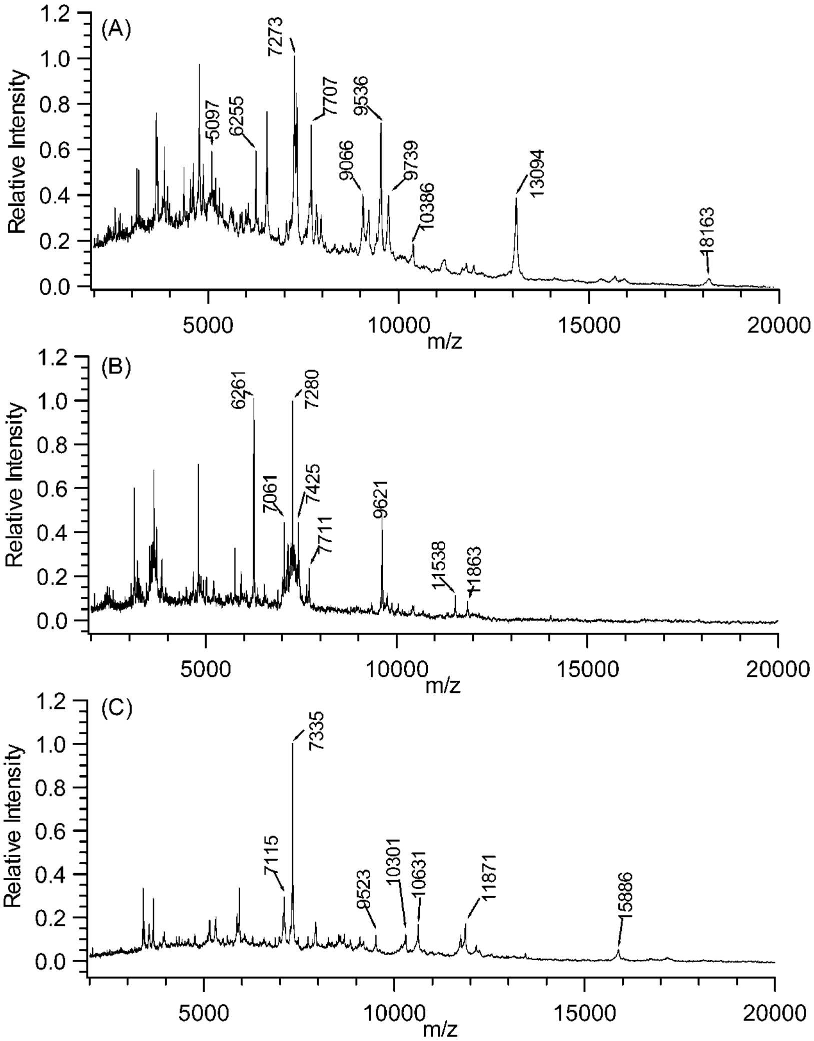

Figure 1. MALDI spectra of proteins from cells (grown in ECBC) extracted by 0.1% TFA aqueous solution. (A) *E. coli*, (B) *B. megaterium*, and (C) *C. freundii*.

Table 4. $(M + H)^+$ from *E. coli* Extract by Direct MALDI

2126	**2374**	**2383**	**5097**	**5293**	**5381**	**5993**	6057	**6255**	**6316**
6413	**6853**	**7061**	**7273**	**7333**	**7707**	**7849**	7970	**9066**	**9226**
9536	**9739**	**10386**	**11206**	**11783**	**11977**	**13094**	15682	**18163**	

Table 5. $(M + H)^+$ from *B. megaterium* Extract by Direct MALDI

6261	6540	**6897**	**7061**	7140	7158	**7280**	**7425**	**7450**	7648
7711	**9336**	**9351**	**9621**	9756	9882	**10044**	**10414**	**10453**	**11538**
11863									

Table 6. $(M + H)^+$ from *C. freundii* Extract by Direct MALDI

7090	7115	7141	**7335**	**7736**	**7865**	**8278**	8547	**8695**	8851
9107	**9523**	10287	**10301**	**10631**	**11751**	**11871**	**12156**	13448	**15886**

(Met-Ox) proteins. The oxidation most likely occurred during protein extraction or MALDI sample preparation. For the mass tables listed in this work, only the nonoxidized protein masses were included.

Data Comparison. As seen from Tables 4–6, for *E. coli*, *B. megaterium*, and *C. freundii*, direct MALDI analysis shows 29, 21, and 20 components, respectively. LC/ESI shows 43, 44, and 46 components (see Tables 1–3) and LC/off-line MALDI shows 423, 276, and 156 components (see Table 7 and Tables SA and SB, Supporting Information). Based on the different ionization methods and sample introduction procedures, it appears very likely that different degrees of ion suppression as well as variation in sample concentration have lead to the observed disparity in the number of detected species. Although direct MALDI analysis is

Table 7. (M + H)$^+$ from *E. coli* Extract by LC/Off-Line MALDI

2004	2006	**2009**	2014	2026	2041	2044	2050	2054	2067
2075	2085	2106	**2123**	2127	2132	**2141**	2166	2197	2208
2232	2237	2240	2260	2264	2279	2282	2293	2297	2301
2305	2312	2331	2339	2353	2360	2374	**2376**	2382	2386
2395	2400	2403	**2417**	**2426**	**2431**	2453	2463	2474	2487
2507	2513	2526	2530	2545	2560	2563	2574	2577	2588
2599	2602	2608	2612	2621	2630	2635	2640	2649	2656
2661	2665	2668	2675	2682	2693	**2737**	2745	2748	2754
2780	2786	2795	2809	2814	2819	2843	2846	2858	2860
2874	2878	2884	2910	2970	2975	2984	2999	3003	3014
3022	3037	3054	3073	3090	3101	3119	3127	3135	3159
3192	3212	3226	3251	3259	3286	3289	3295	3302	3309
3324	3331	3389	3401	3406	3417	3424	3476	3486	3492
3500	**3510**	3534	3538	3547	3557	3587	3592	3599	3615
3624	3648	3667	3750	3764	3769	3782	**3793**	3796	3908
3930	3954	3984	3996	4003	4008	4012	4024	4034	4053
4069	4086	4095	4104	4112	4128	4145	4251	4260	4366
4372	4384	4394	4399	4407	4446	4456	4465	4473	4484
4500	4562	4716	4722	4792	4895	4903	4940	4985	5006
5017	**5035**	5040	5052	5073	5087	**5097**	5144	5154	5180
5202	5253	5295	5362	5378	5382	5396	5414	5430	5449
5469	5480	5493	**5551**	5566	5585	5644	5650	5659	5667
5695	5709	5720	5725	5736	5747	5754	5772	5800	5808
5818	5850	5857	5868	5873	5901	5994	6012	6140	6180
6196	6224	6243	**6255**	6261	6265	6283	6298	**6316**	6324
6338	6344	6369	6411	6453	6486	6493	6555	6600	6617
6669	6684	6702	6723	6772	6787	6827	6854	6866	6890
6942	6951	6958	7060	7070	7095	7109	**7139**	7158	7169
7185	7255	7265	7269	**7274**	7281	7292	7298	7307	7314
7321	7326	7334	7412	7570	7600	7618	**7708**	7735	7740
7752	**7781**	7799	**7852**	7868	8189	8206	8216	8220	8228
8242	8258	8267	8280	8291	8305	8325	8342	8369	8376
8398	8450	8525	8591	8635	8670	8782	8796	8800	8814
8820	8859	8868	8877	8883	8892	8897	8965	8978	8994
8997	9048	9055	**9066**	9080	9160	9175	**9192**	**9209**	**9228**
9252	**9264**	9281	9296	9369	9424	9431	9439	9458	9478
9520	9527	**9537**	9545	9555	9573	9583	**9611**	**9740**	9752
9766	9785	9834	9852	9885	9952	9982	9996	10045	10106
10123	10176	10257	10301	10315	10332	10372	10378	10387	10437
10453	**10463**	10477	10618	10653	10662	10945	11035	11170	11186
11208	11216	11240	11472	11653	11776	**11783**	11794	11870	11977
12233	12446	12769	13077	**13095**	13109	13127	13240	13650	14749
14839	**15693**	18162							

the most straightforward method, it clearly suffers from the inability to detect a large number of components present in the cell extracts. With on-line HPLC separation, ESI MS partially overcomes the problem. However, ion suppression still exists since the fractions still contain multiple proteins with varying concentrations. Most chromatographic peaks in LC/ESI exhibited mass spectral peaks from at most two or three components. In contrast, LC/off-line MALDI detected 5−10 components in many fractions. Note that, unlike on-line LC/ESI, each fraction used for the off-line MALDI analysis was concentrated by a factor of 50 before mixing with the second-layer matrix in a two-layer sample preparation. This concentration step certainly helped to detect many more components in the cell extract by LC/off-line MALDI.

It should be noted that not all of the masses detected from direct MALDI are present in the mass tables produced by LC/off-line MALDI. Only the boldface numbers in Tables 4−6 from direct MALDI are also detected in LC/off-line MALDI. Using one-dimensional HPLC separation, each fraction likely consists of many protein components. Thus, ion suppression still remains an obstacle, although not to the same extent as with direct MALDI. Another possible reason is that some proteins could have eluted into several adjacent HPLC fractions and become too dilute to be detected after HPLC fractionation. The potential for protein loss during sample workup, such as protein concentration by solvent evaporation in individual fractions, can also contribute to the differences observed. Although LC/off-line MALDI detects more peaks than LC/ESI for a given bacterial extract, the masses detected by LC/ESI are not necessarily a subset of those produced by LC/off-line MALDI. For *E. coli*, LC/ESI-MS produces 34 masses (see the boldface numbers in Table 1) matching with the LC/off-line MALDI table. For *B. megaterium*, 29 of the 44 components (Table 2) match with the results found by LC/off-line MALDI. Comparing Table 3 to Table SB (Supporting Information), it can be seen that *C. freundii* has 33 out of 46 components matched with the off-line MALDI data. When the results from direct MALDI and LC/ESI-MS analysis are compared, it is found that, for *E. coli* (Table 4 vs Table 1), 13 out of 29 masses observed in direct MALDI are also detected in LC/ESI-MS. A total of 16 out of 21 masses observed in direct MALDI are found in LC/ESI-MS table for *B. megaterium* (Table 5 vs Table 2), and 9 out of 20 masses from direct MALDI match the LC/ESI-MS data for *C. freundii* (Table 6 vs Table 1).

The above results show that different sets of proteins are detected by different MS ionization techniques (MALDI and ESI) due to the differences in ion suppression and sensitivity. To generate a mass table better representing the protein components in the cell extracts, the results from different MS techniques may be combined.

DISCUSSION

The selection and the availability of suitable databases are critical issues in a proteomic approach for bacterial identification. This strategy is based on searching a set of protein masses detected from an unknown bacterium against a protein mass database of known bacteria. One option is to use the public proteome database available on the Internet.[34] Unfortunately, for a vast majority of bacteria, the number of entries in the current proteome database is still very limited, particularly in the low-mass range (2−20 kDa), which is the main focus of sensitive bacterial identification by MS methods. This is true even for bacteria with their genome sequencing completed. The lack of low-mass protein information in the current proteome database is attributed to the fact that it is difficult to accurately predict and detect small genes.[39] The proteomes of some bacteria, such as *E. coli* and *Bacillus subtilis*, are relatively well studied, since they are often used as models to study Gram-negative and Gram-positive bacteria, respectively. According to the National Centre for Biotechnology Information (NCBI),[40] the genome of 55 bacteria has been completely sequenced (as of November 2001). The number of protein entries in the proteome database of these bacteria is dramatically different. There are 2997 entries for *E. coli* between 2 and 20 kDa, 1515 entries for *B. subtilis*, whereas for some other bacteria only very few entries exist. For instance, *Ureaplasma urealyticum* has only 30 entries. This dramatic difference in the number of protein entries is partially related to the different genome sizes but, to a greater extent, is due to the poor characterization of small gene products for most bacterial species. Consequently, any attempt to identify bacteria based on searching the current proteome database will be biased toward those bacterial species with relatively complete proteome databases. To overcome this problem, Fenselau and co-workers have developed a statistical method that accounts for this proteome size bias.[35] The applicability of this method is clearly dependent on the availability of a suitable proteome database—public or otherwise. Providing that a reasonable size of low-mass proteome exists in the database for the bacterium to be identified, the statistical method has been shown to be much more reliable than simply counting the number of matches.[35]

The analysis of our model system consisting of *E. coli*, *B. megaterium*, and *C. freundii* can be used as an example to demonstrate the importance of having a sufficient number of proteome entries in the database for reliable bacterial identification.

Direct MALDI analysis of the extract from *B. megaterium* detected 21 proteins (see Table 5). Out of these 21 proteins, only 4 match with the *B. megaterium* proteome database within the experimental mass error of ±0.05% (the matched masses are underlined in Table 5). Note that there are only 55 protein entries in the mass range between 2 and 20 kDa for this bacterium in the public proteome database. When these protein masses from Table 5 are searched against the proteome database of other bacteria, which have a relatively larger number of entries, it is found that there is an increase in the number of matches. For example, 16 masses in Table 5 match *E. coli* proteins, 15 match *X. fasridiosa*, and 11 match *B. subtilis* as well as *B. halodurans*. With the use of the statistical scoring method (http://infobacters-vr.jhuapl.edu/),[35] the top candidate is *B. megaterium* with a significance level of 0.003 987. The second candidate is *Chlamydia pneumoniae* with a significance level of 0.072. This search was limited to the bacteria with individual proteome size of >20 entries. In this case, the statistical method correctly identifies the bacterium.

The simple number-matching method from the LC/ESI data from *B. megaterium* (see Table 2) is also biased toward those bacteria with a relatively complete proteome database. Of the 44 proteins detected by LC/ESI, only 4 match with the *B. megaterium* database, while 19 masses match with the *E. coli* proteins and 12 match with *B. subtilis*. Using the statistical scoring method, among the searched bacteria with individual proteome size of >20 entries, the top candidates are found to be *Streptomyces coelicolor* (with a significance level of 0.006 749) and *B. megaterium* (with a significance level of 0.017 43).

For *C. freundii*, none of the masses detected by direct MALDI or LC/ESI matches with the proteins in *C. freundii* proteome database. There are 47 entries in the mass range of 2−20 kDa in the public database. However, 17 out of the 20 detected masses in direct MALDI match with *E. coli* proteome database, and 11 match *B. subtilis* proteome database. In this case, both simple number counting and statistical scoring fail to identify the bacterium.

For the *E. coli* data listed in Table 4 that was obtained by direct MALDI analysis, if we use the same searching parameters as for *B. megaterium* and *C. freundii*, the statistical method gives the top candidates as *Klebsiella oxytoca* (with a significance level of 0.014 07) and *Pseudomonas alcaligenes* (with a significance level of 0.040 43). But, if we limit the search to those bacteria with proteome size of greater than 100, instead of 20, the statistical method correctly identifies *E. coli* as the top candidate.

It is clear that the reliability of both the simple number matching and the statistical method is very much dependent on the availability of a suitable proteome database for mass comparison. The current public database may not be adequate for many applications, particularly in cases where the bacterium of interest has a limited number of protein entries in the database. We note that another limitation of the public bacterial proteome database is that many of the protein masses listed in the database were simply derived from their genome sequences and were not experimentally confirmed. In reality, it is difficult to translate the genome sequence or gene information into the correct protein mass information, particularly for genes with small open reading frames (ORFs).[41,42] While more sophisticated bioinformatics tools

[39] Rudd, K. E.; Humphery-Smith, I.; Wasinger, V. C.; Bairoch, A. *Electrophoresis* **1998**, *19*, 536−544.

[40] http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html.

[41] Blattner, F. R.; Plunkett, G.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. *Science* **1997**, *277*, 1453−1474.

[42] Yamamoto, Y.; Aiba, H.; Baba, T.; Hayashi, K.; Inada, T.; Isono, K.; Itoh, T.; Kimura, S.; Kitagawa, M.; Makino, K.; Miki, T.; Mitsuhashi, N.; Mizobuchi, K.; Mori, H.; Nakade, S.; Nakamura, Y.; Nashimoto, H.; Oshima, T.; Oyama, S.; Saito, N.; Sampei, G.; Satoh, Y.; Sivasundaram, S.; Tagami, H.; Horiushi, T. *DNA Res.* **1997**, *4*, 169−178.

are being developed in a rapid pace,[43] the *exact* starting and ending sequences of a gene for protein expression is still difficult to predict at present. Posttranslational modifications can change the molecular mass of a protein. Even though the genome database for many organisms is expanding rapidly, which will greatly facilitate the establishment of the genome-derived *proteome sequence* database, the establishment of the *proteome mass* database for the organism will always lag behind. Moreover, in vitro processes, such as protein modification or fragmentation during protein extraction or MS sample preparation, can also alter the protein apparent masses.

Instead of relying on the use of the public proteome database for bacterial identification, a complementary approach is to establish a protein mass database using MS techniques. Such a database should provide more reliable protein mass information and account for the possible posttranslational modifications or other modifications related to sample workup. In addition, the MS-derived mass database would only include the proteins that are expressed in moderate quantity in cells. A database not including the low-expression proteins would decrease the chance of falsely matching a detected mass to a protein from an unrelated bacterial species.

With the protein mass tables produced, we first try to address the question of to what extent the protein masses detected by MS differ from the protein masses in the public proteome database. *E. coli* is used as an example for comparison since the current *E. coli* proteome database is quite extensive. In Tables 1, 4, and 7, the underlined numbers are the masses matched with those listed in the proteome database. It is clear that a large fraction of the masses observed by MS are not shown in the proteome database. This is not totally surprising. It has been reported that a significant percentage of small *E. coli* genes or ORFs remains either unassigned or poorly characterized.[41,42] A variety of posttranslational modifications, including methylation, acetylation, and carboxylation, may have occurred on many low-mass proteins.[44,45] Moreover, more than 60% of *E. coli* proteins are found to be proteolytically processed.[44,46] Therefore, many of the observed proteins are likely the products of in vivo fragmentation products from larger proteins with the cleavage of N-teminal methionine (Met) and signal peptides. In vitro fragmentation occurring during sample preparation might also attribute to the discrepancy of the observed and predicted protein masses. The formation of disulfide bonds in some proteins will also change the apparent masses of the proteins. For example, a protein with an apparent mass of 7867 Da (obtained by internal mass calibration with ubiquitin and its doubly charged species) has been identified as 50S ribosomal protein S20.[47] The theoretical molecular weight is 7871; the 4-Da mass difference is due to the formation of two disulfide bonds.

A close look at the mass data shown in Table 7 reveals an interesting fact that most of the proteins with masses below 5000 Da detected by MS are not listed in the proteome database. Identification of the origins of these small proteins is currently underway and will be reported elsewhere. Our preliminary work suggests that some of these low-mass components are peptides generated by degradation of the intact putative proteins.

The utility of mass tables created by LC/off-line MALDI and LC/ESI-MS (i.e., Tables 1−3 and 7, and SA and SB, Supporting Information) for differentiating bacteria within a small data set was evaluated. Several experiments were carried out by direct MALDI analysis of different bacterial samples. The obtained protein masses were then compared to the mass tables.

Figure 2A shows the mass spectrum of a batch of *E. coli* sample cultured in-house. The protein mass values are listed in Table 8. Compared to the tables generated by LC/off-line MALDI (Table 7 and Tables SA and SB, Supporting Information), 24 masses (boldface type) match with those listed in the *E. coli* table. Four match with *B. megaterium*, and seven match with *C. freundii*. When the MALDI data listed in Table 8 are compared to the mass tables from LC/ESI (Tables 1−3), 11 masses (italicized) match *E. coli*, 2 masses match *B. megaterium*, and 3 masses match *C. freunidii*.

A similar comparison can be made for the *B. megaterium* sample analyzed by direct MALDI (see Table 5). A total of 16 out of the 21 peaks match with those listed in the LC/MALDI table of *B. megaterium*, 11 masses match *E. coli*, and 2 match *C. freundii*. When the LC/ESI mass tables are used for mass comparison, 15 of the masses detected by direct MALDI match *B. megaterium*, 2 match *E. coli*, and none of these masses matches with any *C. freundii* proteins.

Another batch of *B. megaterium* cells grown in-house was also analyzed by direct MALDI. Figure 2B shows the MALDI spectrum, and the protein mass values are listed in Table 9. When the LC/off-line MALDI mass tables (Table 7 and Tables SA and SB, Supporting Information) are used, 24 of the 46 components detected by direct MALDI match *B. megaterium* (bolded), 12 match *E. coli*, and 3 match *C. freundii*. Compared to the LC/ESI table, 10 of the 46 masses match *B. megaterium* (italicized), 2 match *E. coli*, and none match *C. freundii*. In contrast, when the public database was used for mass comparison, 35 out of the 46 components matched *E. coli* database. Only three matched *Bacillus megaterium*.

The above examples illustrate that, even based on simple number matching, MS-derived mass tables can be used to differentiate a bacterium from the three bacteria examined. The application of a statistical method for data comparison[36,37] is expected to improve the reliability in bacterial identification. These results are very encouraging, and we envision that this method will be useful for applications where the identity of a target bacterium needs to be rapidly established against a handful of background bacteria. For future work, careful annotation of MS-derived mass tables should result in a powerful protein mass database suitable for bacterial identification in various different applications. In this regard, several other issues will be considered in our future work. Since one MS technique only detects a subset of the entire low-mass proteome, different detection methods and extraction protocols need to be used to examine a bacterium in order to detect most of the relatively highly expressed proteins. Another important consideration is that protein expression is very sensitive to the growth condition. For example, our previous work[13] showed a mass table generated by LC/off-line MALDI on

[43] Yada T.; Totoki, T.; Takagi, T.; Nakai, K. *DNA Res.* **2001**, *8*, 97−106.

[44] Link, A. J.; Robison, K.; Church, G. M. *Electrophoreis* **1997**, *18*, 1259−1313.

[45] Wang, Z.; Doucette, A.; Li, L. Manuscript in preparation.

[46] Wasinger, V. C.; Humphery-Smith, I. *FEMS Microbiol. lett.* **1998**, *169*, 375−382.

[47] Keller, B. O.; Wang, Z.; Li, L. *J. Chromatogr., B,* in press.

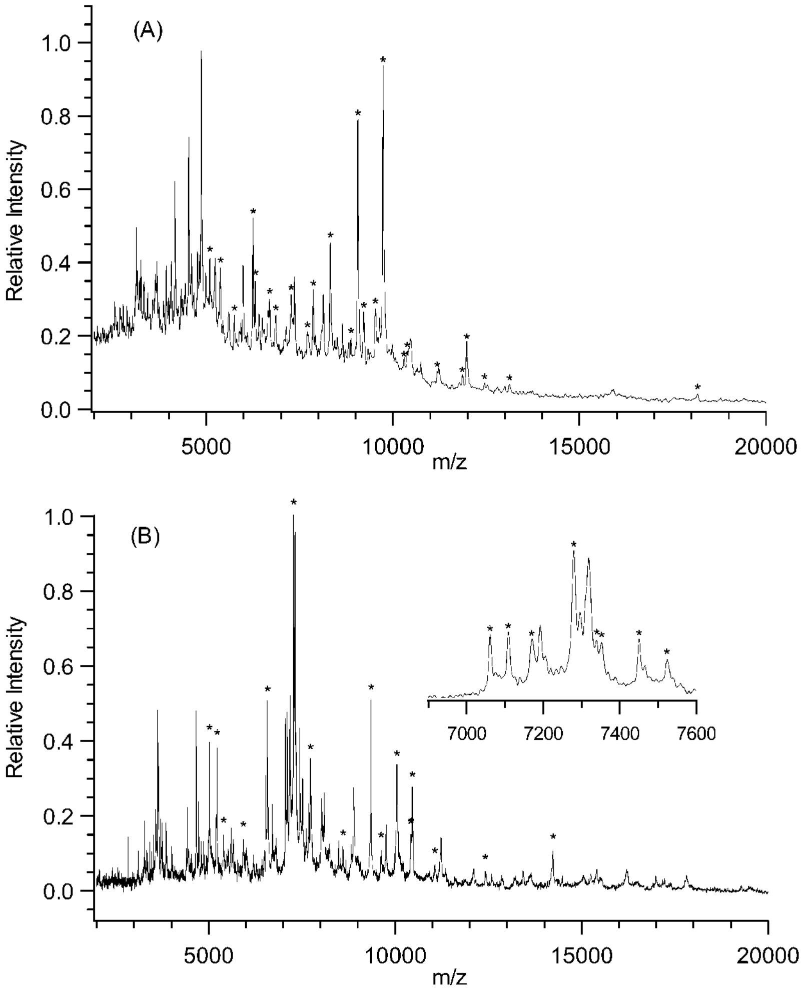

Figure 2. (A) MALDI spectrum of *E. coli* (grown in-house) proteins extracted by 0.1% TFA aqueous solution. The peaks with asterisks matched the masses in LC/off-line MALDI table (Table 7). (B) MALDI spectrum of *B. megaterium* (grown in-house) proteins extracted by 0.1% TFA aqueous solution. The peaks with asterisks matched the masses in LC/off-line MALDI table (Table SA, Supporting Information).

Table 8. $(M + H)^+$ of *E. coli* Cells Harvested at 36 h by Direct MALDI[a]

5097	5242	**5381**	**5755**	***6255***	***6317***	6420	6511	**6697**	**6864**
7145	***7272***	7369	***7708***	**7868**	8096	8139	**8327**	8649	**8884**
9065	***9225***	***9536***	9642	***9741***	9988	**10300**	***10386***	10496	10750
11186	*11229*	**11865**	**11977**	**12451**	13010	**13127**	15906	**18163**	

[a] Boldface masses match the LC/off-line MALDI table for *E. coli* (Table 7); italicized masses match the LC/ESI data (Table 1).

Table 9. $(M + H)^+$ of *B. megaterium* Grown In-House by Direct MALDI[a]

5018	**5223**	**5397**	5605	5653	**5929**	6537	***6577***	6709	6816
7061	***7108***	**7171**	7191	***7280***	7318	**7340**	**7353**	***7450***	***7521***
7618	7693	**7730**	8025	8100	8487	**8604**	8887	***9355***	***9621***
9760	***10047***	**10414**	**10434**	***10455***	***11061***	11229	12083	**12417**	12434
12584	**14223**	15407	16212	16989	17812				

[a] Boldface masses match the LC/off-line MALDI table for *B. megaterium* (Table SA, Supporting Information); italicized masses match the LC/ESI data (Table 2).

a batch of *E. coli* sample that was different from the one used to generate the mass table shown in Table 7. A total of 169 out of 307 protein masses in that table match with those shown in Table 7. The large numbers of different protein masses from the different batch sample suggest that bacteria grown under various conditions and harvested at different growth times need to be examined. In a mass database composed of all masses detected by the mass spectrometric method, the protein masses *consistently* observed under each growth condition or each sample preparation condition should be given more weight in an algorithm that searches these database to make an identification, compared to others only detected under a specific condition.

CONCLUSIONS

We have shown that, based on protein masses alone, the present public proteome database has limited use for differentiating bacteria that have small numbers of low-mass entries. A method based on mass spectrometric techniques to generate protein mass database is proposed. The MS-derived mass database should complement the public proteome database for bacterial identification. There are several merits of creating protein mass database by using mass spectrometric methods. First, the mass database is used for a single purpose and thus the initial database can be composed of a number of mass tables each containing the consistently MS-detectable proteins from an individual bacterium. These mass tables can be rapidly created by MS in conjunction with different sample preparation and handling methods. The resulting database should be bias free from the sample preparation method used for preparing an unknown sample. Second, the MS protocol used to create mass tables can be readily adapted and used to rapidly generate mass data for a new strain of bacteria. This is important in order to keep the database current and meet special needs. Third, the mass database generated by MS can include many posttranslationally modified proteins or proteins being altered during sample workup. Many of these proteins are not included in the public proteome database, but they should be useful, providing they are consistently detected. Fourth, the MS-derived mass database would only include the proteins that are detectable by MS, which should enable us to use the limited mass range more effectively. Finally, the mass database can be expanded, if needed, to include other comparative parameters such as MS/MS spectra of intact proteins or peptides to increase the confidence level of identification.

ACKNOWLEDGMENT

This work was funded by the U.S. Army ECBC. We thank Drs. Peter Snyder and Dennis Roser, ECBC, for the helpful discussion during the course of this work. We also thank Ms. Lidan Tao for her assistance in bacterial cell growth and sample preparation.

SUPPORTING INFORMATION AVAILABLE

Mass values of the protonated protein ions from *B. megaterium* and *C. freundii* detected by LC/off-line MALDI, Tables SA and SB. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review December 7, 2001. Accepted February 15, 2002.

AC015725F