Anal. Chem. **1999,** *71,* 2732−2738

Microorganism Identification by Mass Spectrometry and Protein Database Searches

Plamen A. Demirev,* Yen-Peng Ho, Victor Ryzhov, and Catherine Fenselau

Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

A method for rapid identification of microorganisms is presented, which exploits the wealth of information contained in prokaryotic genome and protein sequence databases. The method is based on determining the masses of a set of ions by MALDI TOF mass spectrometry of intact or treated cells. Subsequent correlation of each ion in the set to a protein, along with the organismic source of the protein, is performed by searching an Internet-accessible protein database. Convoluting the lists for all ions and ranking the organisms corresponding to matched ions results in the identification of the microorganism. The method has been successfully demonstrated on *B. subtilis* and *E. coli*, two organisms with completely sequenced genomes. The method has been also tested for identification from mass spectra of mixtures of microorganisms, from spectra of an organism at different growth stages, and from spectra originating at other laboratories. Experimental factors such as MALDI matrix preparation, spectral reproducibility, contaminants, mass range, and measurement accuracy on the database search procedure are addressed too. The proposed method has several advantages over other MS methods for microorganism identification.

The massive effort to sequence the human genome has brought about a rapid increase in the speed with which DNA sequences from all species are being accumulated in publicly available computer databases. As a result, the complete genomes of 18 microorganisms are now known,[1] with the entire sequence of a multicellular organism—the nematode *Caenorhabditis elegans*—very recently unveiled.[2] The rate with which complete genome sequences of new microorganisms are unraveled is constantly accelerating, and it is predicted that the completed genomes of more than 50 prokaryotes will be available in less than a year.[3] One of the major tasks of the emerging discipline of bioinformatics[4] is mapping the interconnections and correlations between DNA sequences and protein sequences in an effort to understand cellular function at the molecular level. There exists complementarity between the genome of an organism and its respective proteome, i.e., the dynamic entity set of all expressed proteins. In databases, such complementarity is realized via assignment of an amino acid sequence to each "open reading frame" (ORF) in a DNA sequence. By using bioinformatics tools, the complete proteomes of microorganisms with established DNA sequences are also made available and accessible through the Internet. Unequivocal characterization of such organisms can be achieved through knowledge of their complete genomes or complementary proteomes. Concurrently, sequences of many individual proteins and genes from various organisms have also been obtained and entered into databases. Such a wealth of sequence information has prompted scientists to address previously intractable questions about the structure and functions of the proteome.[5−8]

Recent technical and methodological advances in mass spectrometry (MS) have promoted its wider use for structural elucidation of biopolymers. The expanding requirements in proteomics, e.g., for rapid identification of proteins present in a mixture in picomolar amounts, have resulted in the development of powerful MS-based procedures for identity assignment of individual proteins.[9−14] They are based on chemical/enzymatic digestion of material (obtained from a single spot in a two-dimensional gel electropherogram or by other suitable chromatographic techniques) and mass spectral determination of the molecular masses of the protein and resulting peptides (peptide mapping). Making use of already available information in protein sequence databases, a comparison is made between proteolytic peptide mass patterns generated "in silico" and experimentally observed peptide masses. A "hit-list" is compiled, ranking candidate proteins in the database, based on (among other criteria) number of matches between the proteolytic fragments. Several Web sites are accessible that provide software for protein identification on-line, based on peptide mapping and sequence

(1) National Center for Biotechnology Information (NIH), http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html.
(2) The *C. elegans* Sequencing Consortium. *Science* **1998**, *282*, 2012−2018.
(3) Arigoni, F.; Talabot, F.; Peitsch, M.; Edgerton, M.; Meldrum, E.; Allet, E.; Fish, R.; Jamotte, Th.; Curchod, M.-L.; Loferer, H. *Nat. Biotechnol.* **1998**, *16*, 851−856.
(4) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*; Baxevanis, A., Oulette, B., Eds.; Methods of Biochemical Analysis 39; Wiley-Interscience: New York, 1998.
(5) Roepstorff, P. *Curr. Opin. Biotechnol.* **1997**, *8*, 6−13.
(6) Humphery-Smith, I.; Blackstock, W. *J. Protein Chem.* **1997**, *16*, 537−544.
(7) James, P. *Biochem. Biophys. Res. Commun.* **1997**, *231*, 1−6.
(8) Kuster, B.; Mann, M. *Curr. Opin. Struct. Biol.* **1998**, *8*, 393−400.
(9) Henzel, W.; Billeci, T.; Stults, J.; Wong, S.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.
(10) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−345.
(11) Pappen, D.; Hojrup, P.; Bleasby, A. *Curr. Biol.* **1993**, *3*, 327−332.
(12) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.
(13) Yates, J. R., III; McCormack, A.; Eng, J. *Anal. Chem.* **1996**, *68*, 534A−540A.
(14) Fenyö, D.; Qin, J.; Chait, B. *Electrophoresis* **1998**, ***19***, 998−1005.

10.1021/ac990165u CCC: $18.00
Published on Web 05/27/1999

database search strategies.[15] Additional mass spectral information—increased mass accuracy or peptide sequence "tagging" (i.e., knowledge of the partial sequence of a peptide in the mixture, obtained by tandem mass spectrometry or peptide ladder techniques)—increases the specificity of the approach.[16−18]

In a separate development, the suitability of several MS approaches for rapid identification and characterization of microorganisms has been probed.[19−23] A common denominator of all such approaches is the detection of organism-specific ions, indicative of the presence of unique "biomarkers" and combinations thereof. Cell wall lipids have been successfully employed as bacterial biomarkers.[19,20] Recently, many reports have suggested that proteins, expressed in microorganisms, can be also used as biomarkers. Mass spectra of protein mixtures, obtained on matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) instruments, have been employed for organism identification and classification.[21−30] Thus far, this approach has been based on differences in the "fingerprint" protein profile for different organisms, i.e., in the patterns of observed mass spectral peaks, typically in the mass range of 4−15 kDa. A crucial requirement for successful identification by that approach is mass spectral reproducibility. However, it is known that spectra of such complex mixtures depend in intricate ways on a number of instrumental and microbiological factors—among these are bacterial culture growth times,[31] sample pretreatment,[29] and MALDI matrixes. Although a report on interlaboratory reproducibility of MALDI spectra of several samples has appeared,[29] it is clear that a fingerprint mass spectral library of microorganisms is shaped by both the ionization technique and mass analyzer used to compile it.

Here, we address the possibility of exploiting the wealth of information contained in prokaryotic genome and protein sequence databases to rapidly identify microorganisms. Specifically, we have demonstrated that microorganisms can be identified when the constituents of a set of proteins represented by their molecular masses in a mass spectrum are each identified in a sequence database along with their organismic sources. Two organisms with completely sequenced genomes, *Bacillus subtilis* and *Escherichia coli* (a Gram-positive and a Gram-negative bacterium, respectively), are studied here. A mixture of the two organisms and the capability of the database search approach for identifying individual microorganisms have been studied as well. Experimental factors, such as culture time, choice of MALDI matrix, mass measurement accuracy, and mass range on the database search identification, are addressed. We argue that the proposed method has several advantages over the fingerprint library methods for microorganism identification. It is independent of the specific ionization technique and mass analyzer, and it alleviates the requirement for rigorous reproducibility, crucial in currently used fingerprint-based approaches.

EXPERIMENTAL SECTION

For the purpose of illustrating the feasibility of the method, MALDI TOF mass spectrometry was employed. The described database search method is not restricted to that specific instrument combination and sample preparation. Sinapinic acid (SA) or α-cyano-4-hydroxycinnamic acid (CHCA) 50 mM in 70:30 CH_3CN/H_2O and an equimolar mixture of SA and 4-methoxycinnamic acid (MCA) in 70:30 CH_3CN/H_2O were used as matrixes. The microorganisms studied were *B. subtilis* (strain 168, ATCC No. 23857) and *E. coli* (ATCC No. 11775). They were grown in-house according to standard procedures: 8 g/L nutrient broth (Difco Labs, Detroit, MI) was used as a growth medium; after harvesting the material was centrifuged for 10 min at 10^4g and washed with water three times prior to lyophilization for prolonged storage at −10 °C. Lyophilized vegetative cells were suspended in a 70:30 solution of CH_3CN/0.1% trifluoroacetic acid at a concentration of 5 mg/mL. *B. subtilis* suspension (0.2 μL) was deposited on the sample slide of a Kompact MALDI 4 TOF instrument (Kratos Analytical Instruments, Manchester, U.K.). The sample was allowed to dry and was subjected to corona plasma discharge (CPD) treatment for 15 s according to a published procedure.[32] After that treatment (aimed at release of higher mass biomarkers from the microorganism), a SA matrix solution (0.2 μL) was deposited on the sample slide before MALDI mass spectrometry. For *E. coli*, sample and matrix solutions were mixed directly on the sample slide. A sample of *E. coli* and *B. subtilis* was prepared by mixing suspensions of the two microorganisms on the slide prior to CPD treatment and MALDI mass spectrometry. In some experiments, an internal mass calibration standard (a solution of bovine insulin and bovine ubiquitin) was added to the *E. coli* sample/matrix mixture on the sample slide in order to increase the accuracy of mass determination. For *B. subtilis*, external calibration of the instrument using a mixture of proteins (bovine insulin, bovine ubiquitin, and horse heart cytochrome *c*) was performed prior to running the samples. All proteins were obtained from Sigma Chemical Co. (St. Louis, MO).

(15) (a) prospector.ucsf.edu www.proteometrics.com. (b) www.mann.embl-heidelberg.de/Services/PeptideSearch. (c) cbrg.inf.ethz.ch/MassSearch.html expasy.hcuge.ch. (d) www.seqnet.dl.ac.uk/mowse.html.

(16) Jensen, O.; Podtelejnikov, A.; Mann, M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1371−1378.

(17) Mortz, E.; O'Connor, P.; Roepstorff, P.; Kelleher, N.; Wood, T.; McLafferty, F.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8264−8267.

(18) Yates, J. R., III; Eng, J. Use of Mass Spectrometry Fragmentation Patterns of Peptides to Identify Amino Acid Sequences in Databases. U.S. Patent No. 553897, July 23, 1996.

(19) Anhalt, J. P.; Fenselau, C. *Anal. Chem.* **1975**, *47*, 219−225.

(20) Heller, D.; Fenselau, C.; Cotter, R.; Demirev, P.; Olthoff, J.; Honovich, J.; Uy, M.; Tanaka, T.; Kishimoto, Y. *Biochem. Biophys. Res. Commun.* **1987**, *142*, 194−199.

(21) *Mass Spectrometry for the Characterization of Microorganisms*; Fenselau, C., Ed.; ACS Symposium Series 541; American Chemical Society: Washington, DC, 1994.

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

(23) Claydon, M.; Davey, S.; Edwards-Jones, V.; Gordon, D. *Nat. Biotechnol.* **1996**, *14*, 1584−1586.

(24) Holland, R.; Wilkes, J.; Rafii, F.; Sutherland, J.; Person, C.; Voorhees, K.; Lay, J. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1227−1232.

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

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

(27) Welham, K.; Domin, M.; Scannell, D.; Cohen, E.; Ashton, D. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 176−180.

(28) Haag, A.; Taylor, S.; Johnston, K.; Cole, R. *J. Mass Spectrom.* **1998**, *33*, 750−756.

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

(30) Dai, Y.; Li, L.; Roser, D.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 73−78.

(31) Arnold, R.; Reilly, J. *A Study of Bacterial Culture Growth by MALDI-MS of Whole Cells*; Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL, May 31−June 4, 1998; p 180.

(32) Birmingham, J.; Demirev, P.; Ho, Y.-P.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 604−606.

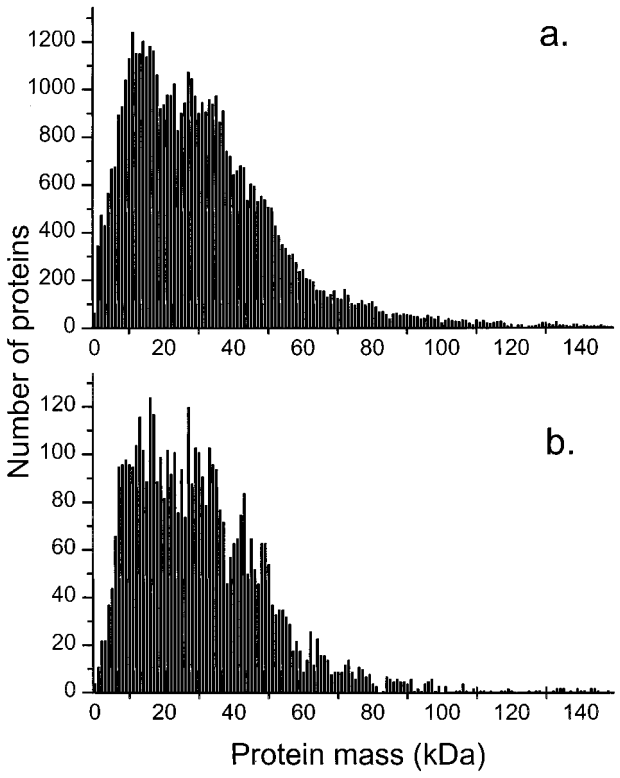

Figure 1. Molecular mass distribution (in bins of 1 kDa) of proteins deposited in the SwissPROT/TrEMBL sequence database: (a) all prokaryotic proteins; (b) all proteins from *B. subtilis*.

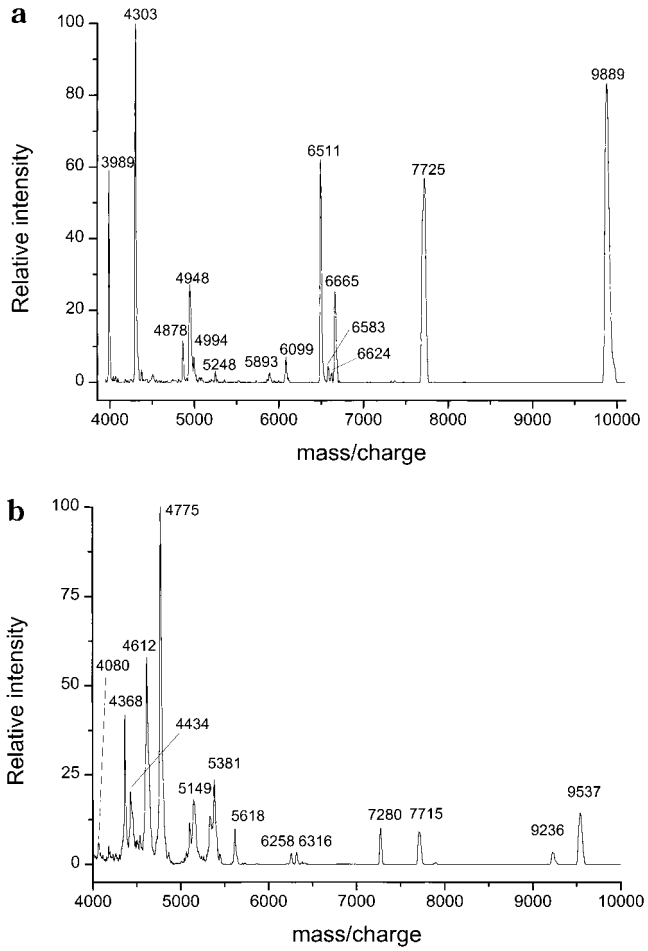

Figure 2. Positive ion MALDI spectra from (a) *B. subtilis* (8 h growth time), matrix SA; (b) *E. coli* (32 h growth time), matrix CHCA (see Experimental Section for more details).

Positive ion mass spectra (typically from 50 single laser shots rastered uniformly across the sample spot) were recorded in linear mode at 20 kV accelerating voltage and a delay of 0.3 μs. The estimated N_2 laser fluence was around 10 mJ·cm^{-2}.

A search by protein molecular mass (M_r) and based on the set of protein molecular weights in the spectra was carried out in the SwissProt/TrEMBL database (Expasy, Swiss Bioinformatics Institute) using the Sequence Retrieval System (SRSWWW) module at http://expasy.hcuge.ch/srs5/. An interactive window (Alternative Query Form) allows search by a number of classifiers. In this case, we chose averaged protein MW as the primary classifier. We selected a ±3 Da MW window, and the only restriction applied in the query was the choice of the "bacteria" protein subset of the database (in an earlier release of SwissPROT the identifier "prokaryota" was also available). Thus, protein identities and organismic sources were tentatively assigned for all peaks from the experimental spectra within the range from 4 to 15 kDa.

RESULTS AND DISCUSSION

Compilation of electron ionization (EI) mass spectra of volatile organics in a computer database has formed the basis of mass spectral search strategies for identification of unknown low-molecular-weight compounds (often in complex mixtures). Various search algorithms have been developed for the positive identification of *unknowns* by matching their characteristic EI mass spectra against the spectra of *known* compounds in a database and vice versa.[33,34] The mass spectral approach to proteomics, which emerged in the early 1990s,[9−12] is based on a similar paradigm. In this case, the overlap of proteolytic peptide masses is used as a matching criterion. The relative intensity of each peptide in a mass spectrum depends on both intrinsic (e.g., basicity, hydrophobicity) and extrinsic (instrumental) factors, and relative signal intensities are not part of these searches.

Here we apply a new approach for MS identification of microorganisms. We compare the set of protein molecular masses in a mass spectrum of an *unknown* organism against a database containing the sequence-derived molecular masses of all proteins present in *known* organisms. Currently, the number of microorganisms whose proteomes are completely known is limited.[1] On the other hand, several publicly accessible databases contain each more than 50 000 protein sequences of prokaryota proteins derived from genomic ORF as well as nongenomic entries. *B. subtilis* (strain 168) has a total of 4590 protein entries, and *E. coli* has 6521, including redundancies, in the SwissPROT/TrEMBL database. The molecular mass distribution of known prokaryotic proteins (Figure 1a) has a peak centered around 12 kDa. Similarly, the M_r distribution of proteins from the *B. subtils* proteome has its maximum around 12 kDa (Figure 1b). From these data it is clear that many microbial proteins have masses in that range

(33) Chapman, J. *Computers in Mass Spectrometry*; Academic Press: London, 1978.

(34) McLafferty, F.; Turecek, F. *Interpretation of Mass Spectra*, 4th ed.; University Science Books: Mill Valley, CA, 1993.

Table 1. Ranking of Organisms According to Matched Peaks in *B. subtilis* Spectrum (Figure 2a)

	observed mass (Da)														
organism[a]	3988	4302	4506	4877	4947	4993	5247	5892	6098	6510	6582	6623	6664	7724	9888
B. subtilis	x	x	x	x		x	x	x	x	x		x		x	x
E. coli		x		x			x	x		x	x				
B. burgdorferi	x	x		x	x			x							
M. tuberculosis							x		x		x			x	
P. aeruginosa		x				x									
M. leprae										x					x

[a] Only organisms with more than one matching peak (within ±3 Da) are listed.

Table 2. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 2b)

	observed mass (Da)																
organism[a]	4079	4367	4433	4538	4611	4774	5101	5148	5335	5380	5617	6257	6315	7279	7714	9235	9536
E. coli	x	x	x	x	x		x			x	x		x	x	x	x	x
H. influenza							x					x	x		x	x	
B. subtilis												x	x		x		x
M. leprae										x	x			x		x	
B. burgdorferi	x			x		x									x		
S. typhimurium	x			x				x								x	
H. pyroli									x					x			
Synechococcus sp.		x													x		
M. tuberculosis															x	x	

*Only organisms with more than one matching peak (within ±3 Da) are listed.

Table 3. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 1b of ref 29)

	observed mass (Da)											
organism[a]	4362	4711	5076	5752	6255	7272	7708	8447	9067	9424	10464	10760
E. coli	x	x	x			x	x	x	x	x	x	x
H. influenza	x		x		x		x	x		x	x	
B. subtilis					x	x	x	x	x			x
Synechococcus sp.	x					x	x				x	x
H. pyroli		x		x		x						
M. leprae										x	x	x
Rhizobium sp.						x				x	x	
B. burgdorferi		x		x								
M. tuberculosis							x				x	
S. typhimurium				x								x

[a] Only organisms with more than one matching peak (within ±5 Da) are listed.

(although the exact nature of these low-mass-range proteins is still debated[35]). Therefore, it may be also expected that unique combinations of protein masses in the mass range from 4 to 15 kDa can serve to identify procaryotic microorganisms.

Under the conditions used, the MALDI spectra (Figure 2) of *B. subtilis* or *E. coli* contain multiple peaks between 4 and 10 kDa with a signal-to-noise ratio better than 3. They are listed in Tables 1 and 2, respectively. A database search was performed, based on the observed masses. It was assumed that singly protonated molecules were detected; i.e., a proton mass was subtracted from the observed mass in order to obtain the average M_r. In assigning the respective peaks (i.e., proteins with M_r within the ΔM_r window chosen, ±3 Da), the organisms from which each potential protein originates are also determined. These are presented in Tables 1 and 2. From Table 1, one microorganism, *B. subtilis*, is identified as the source of 12 of the 15 peaks. There are two "runnerups" in that example, which provide matches for 6 and 5 of the 15 major peaks. It is evident from Table 2 that 13 *E. coli* proteins match observed peaks (out of 17 total), while one microorganism matches 5 of the 17 peaks. The possibility that unmatched peaks can correspond to alkali cation adducts and/or posttranslationally modified products (including proteolytic fragments) of proteins already present in the database will be explored in a software implementation of the described apporach.

As already pointed out, there exist inherent problems with the reproducibility of MALDI mass spectra from microorganisms, which depend on a multitude of factors. For instance, a direct comparison of published MALDI mass spectra of the same organism, *E. coli*,[26,29,30] shows that they do not match each other or the spectrum in Figure 2b. However, searching the proteome database for masses observed in each spectrum leads to the positive identification of the bacteria in each case (Tables 3−5).

(35) Das, S.; Yu, L.; Gaitatzes, C.; Roger, R.; Freeman, J.; Blenkowska, J.; Adams, R. M.; Smith, T. F. *Nature* **1997**, *385*, 29−30.

Table 4. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 1a of ref 30)

	observed mass (Da)										
organism[a]	3636	4365	4532	4769	6547	7271	7333	9061	9535	9737	13093
E. coli		x	x			x	x	x	x	x	x
B. subtilis				x		x	x	x	x	x	x
Synechococcus sp.		x	x			x			x	x	x
B. burgdorferi	x		x				x		x		
H. influenza		x			x					x	
Rhizobium sp.					x	x				x	
H. pyroli				x		x					
M. tuberculosis										x	x
S. typhimurium	x		x								

[a] Only organisms with more than one matching peak (within ±5 Da) are listed.

Table 5. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 4 of ref 26)

	observed mass (Da)						
organism[a]	5100	5380	7280	8320	9070	9530	9740
E. coli	x	x	x	x	x	x	x
B. subtilis					x	x	x
M. tuberculosis					x	x	x
H. pyroli			x	x			
B. burgdorferi					x	x	
H. influenza	x				x		
E. cloacae	x		x				
Synechocystis sp.					x		x

[a] Only organisms with more than one matching peak (within ±5 Da) are listed.

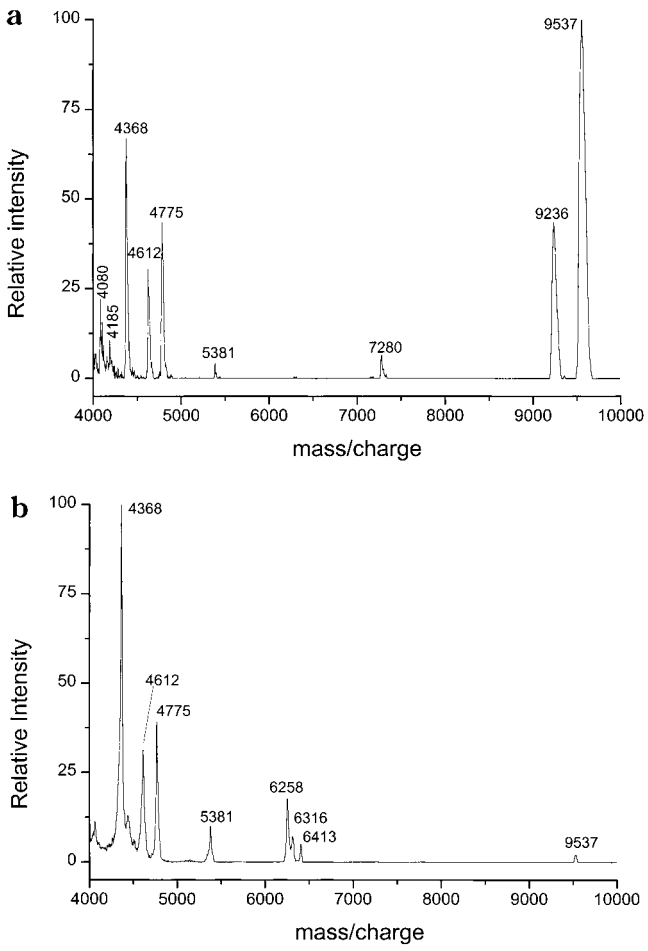

Figure 3. Positive ion MALDI spectra from (a) *E. coli* (32 h growth time), matrix MCA/SA mixture; (b) *E. coli* (8 h growth time), matrix CHCA.

This is not surprising since all spectra should reflect the presence of expressed proteins from the same genome. The same type of robustness can be illustrated by comparing the MALDI spectra from the same sample of *E. coli*, obtained in different matrixes. The spectra have different fingerprints—peaks above 5 kDa are more prominent in the spectrum obtained with MCA/SA matrix (Figure 3a), in comparison to spectra with CHCA matrix (Figure 2b). However, the database search method results in positive identification of the species in each spectrum (Tables 2 and 7). Effects of incubation time on experimentally obtained mass spectra from *E. coli* have been discussed in the literature.[31] Spectra from *E. coli* harvested after 8 and 32 h of growth are compared in Figures 3b and 2b. Again the overall spectral appearance is different for the two samples. Nevertheless, the identification is straightforward in both cases (Tables 2 and 7). It appears that experimental factors such as choice of an "appropriate" MALDI matrix, variability in the levels of protein expression, etc., will have limited influence when microorganisms are identified by searching the proteome database.

It is clear that an increased number of detected peaks across a broader mass range and higher mass accuracy in determining their masses would boost the specificity of microorganism identification. There should exist an optimum in the number of proteins (i.e., detected masses in a mass spectrum) needed for an unequivocal and nonredundant identification of an organism, and this number will vary with mass accuracy. That number will also be a function of the constantly increasing number of entries from various organisms in protein databases, the selected mass range, etc. An illustration of such dependence is provided by comparing the distributions of molecular masses of individual proteins in the combined proteomes of *E. coli* and *B. subtilis* (Figure 4). There are 1261 and 535 different proteins listed in SwissPROT (update 36) in the range between 4 and 20 kDa for each organism, respectively. Mass accuracy of 10 ppm is sufficient to differentiate between most of the individual proteins from each of these two subsets (Figure 4). However, such high accuracy is not a necessary precondition for successful discrimination between the two organisms and their unique identification from a larger set of proteins, as described above. A spectrum (Figure 5) obtained from a mixture of the two organisms, *E. coli* and *B.*

Table 6. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 3a)

	observed mass (Da)								
organism[a]	4079	4184	4367	4612	4774	5380	7279	9235	9536
E. coli	x		x	x		x	x	x	x
B. burgdorferi	x	x			x				
M. leprae						x	x	x	
Synechococcus sp.		x	x						
S. typhimurium	x							x	

[a] Only organisms with more than one matching peak (within ±3 Da) are listed.

Table 7. Ranking of Organisms According to Matched Peaks in *E. coli* Spectrum (Figure 3b)

	observed mass (Da)									
organism[a]	4079	4367	4433	4611	4774	5380	6257	6315	6412	9536
E. coli	x	x	x	x		x		x	x	x
B. subtilis							x	x		x
H. influenza							x	x		
B. burgdorferi	x				x					
M. leprae						x			x	
Synechocystis sp.		x							x	

[a] Only organisms with more than one matching peak (within ±3 Da) are listed.

Table 8. Ranking of Organisms According to Matched Peaks in Spectrum of *B. subtilis* and *E. coli* Mixture (Figure 5)

	observed mass (Da)															
organism[a]	4611	4774	4877	4963	5013	6098	6412	6510	7203	7279	7334	7724	9235	9536	9888	10002
E. coli	x		x	x			x	x	x	x	x		x	x		x
B. subtilis			x		x	x		x	x			x		x	x	x
M. leprae							x	x	x				x		x	x
Synechocystis sp.							x		x	x						x
B. burgdorferi		x	x								x					
H. influenza				x									x			
M. tuberculosis						x						x				
S. typhimurium			x										x			

[a] Only organisms with more than one matching peak (within ±3 Da) are listed.

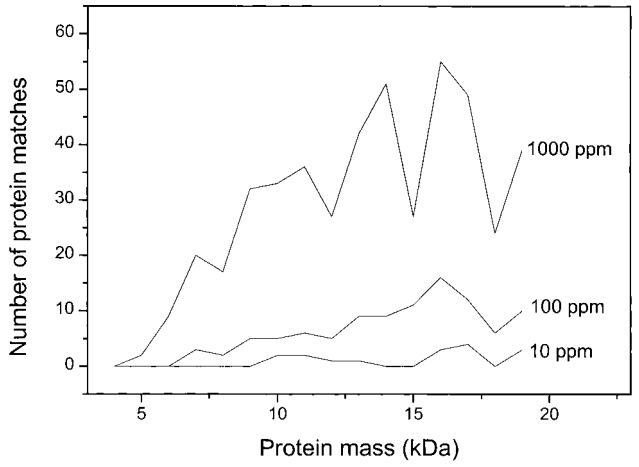

Figure 4. Number of proteins combined from *B. subtilis* and *E. coli* with masses within a predetermined mass window (in ppm) as a function of molecular mass.

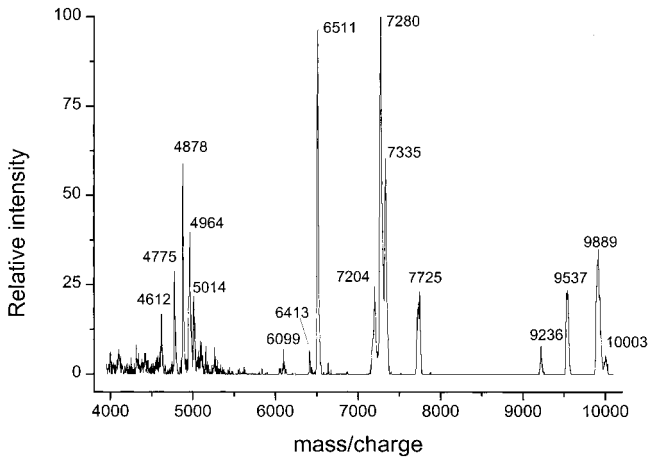

Figure 5. Positive ion MALDI spectrum from a mixture of *B. subtilis* and *E. coli*, matrix SA.

subtilis, exhibits peaks that are characteristic of each species. From 16 major peaks in that spectrum, 11 are assigned to *E. coli* proteins and 9 to *B. subtilis* (Table 8). A more stringent mass determination will reduce the interorganism peak overlaps. Also, more advanced ranking strategies, based for example on the number of overlapping proteins from different species within a mass window, may improve identification of individual microorganisms in a mixture

as well as the success rate of the database search approach in general. Theoretical bioinformatics work in progress will quantitatively address these issues, including in-depth statistical and probabilistic analysis.

CONCLUSIONS AND FUTURE PROSPECTS

We have demonstrated with the examples of *B. subtilis* and *E. coli* the feasibility of an approach for microorganism identification at the species level by mass spectrometry-based protein database searching. Future studies with other microorganisms will continue to test the general validity of the method. The approach introduced here is independent of relative signal intensities in the mass spectrum. It does not even require that the same set of proteins be expressed and/or detected in each analysis of the same organism, only that a set is characteristic so that it can be associated with a microorganism source. Consequently, this kind of database search approach will be more robust than comparison to fingerprint libraries. Other attractive features of the proposed approach are its potential capabilities to identify individual organisms present in mixtures of organisms and, from there, the possibility to identify genetically engineered organisms, containing biomarkers of more than one species. Again, the particular choices of sample preparation, ionization, and mass analysis for obtaining mass spectra are not restrictive for the described approach, which also has a potential to be used for identification of cells from individual tissues.

In the examples described here, the tentative identities of the individual protein biomarkers have not been explored, although this information is contained in the database. As already demonstrated, positive identification of an individual protein would require tryptic and/or mass spectral fragmentation. We note that tandem mass spectrometry (coupled to efficient precursor ion excitation methods) can provide sequence-specific fragments of the individual proteins. Such information, which can improve sensitivity and signal/noise ratio, is "orthogonal" to the MW information and especially useful for identifying posttranslationally modified proteins. Obviously, it can serve as an additional constraint for microorganism identification based on database search strategies. A user-friendly software for implementation of the above general approach is currently being developed and will incorporate options for introduction of such orthogonal information among several other features.

ACKNOWLEDGMENT

This work has been supported by the Applied Physics Lab/Johns Hopkins University and DARPA. Danying Zhu is gratefully acknowledged for growing the microorganisms used in this study.

Received for review February 15, 1999. Accepted April 20, 1999.

AC990165U