Anal. Chem. **2002,** *74,* 5807−5813

Automated Postprocessing of Electrospray LC/MS Data for Profiling Protein Expression in Bacteria

Tracie L. Williams,*,† Peter Leopold,‡ and Steven Musser†

Instrumentation and Biophysics Branch, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland 20740, and BioAnalyte Inc., 264 Eastern Promenade, Portland, Maine 04101

We describe an integrated approach for automating protein analysis of bacterial cell extracts. The method uses electrospray LC/MS to generate chromatographic profiles of proteins present in an extract, along with a software program that automates the data analysis. The software program, Retana, automates the sequential summing, centroiding, and deconvolution of multiply charged proteins present in consecutive scans of the LC/MS analysis. This procedure generates a concise, single spectrum of proteins present in the extract, along with their retention time and relative abundance. A comparison of the method with "whole cell" MALDI analysis demonstrates improved mass resolution and mass accuracy, along with the appearance of a greater number of proteins. Additionally, it is possible to compare protein expression among strains of bacteria by normalizing the relative abundance of similar proteins in each analysis.

An important area of research in microbiology is the identification of specific virulence factors, which differentiate pathogenic strains of bacteria from nonpathogenic strains. Most efforts to identify factors influencing virulence have focused on the genetic differences between pathogens and closely related nonpathogenic bacteria.[1,2] Genes encode proteins that function in conserved signaling pathways, and thus, a consequence of change in the bacterial genome is altered protein expression. The difference between pathogenic and nonpathogenic bacterial strains may be reflected in expression of new proteins, differences in the amount of expressed proteins, protein sequence mutations, or differences in posttranslational protein modifications, such as phosphorylation and acetylation.

Numerous MS techniques have been used to identify macromolecular biomarkers and develop spectral fingerprints of bacteria. One successful approach has been the analysis of whole bacterial cells by matrix assisted laser desorption ionization mass spectrometry (MALDI-MS).[3−8] Identification is based upon the complex patterns observed and the characteristic nature of bacterial mass spectra. It is generally believed that most of the ions observed in this experiment are derived from proteins dissolved in the MALDI matrix.[9−10] Although the spectra reflect only a small portion of the cellular proteome, they are nevertheless sufficiently characteristic to address many taxonomic or biological questions, giving rise to the term *phyloproteomics*.[10]

The identification of differentially expressed proteins requires sample-to-sample comparisons of protein expression profiles. Reproducibility of experimental conditions and mass accuracy are imperative for successful mass spectrometric analysis of protein expression. For MALDI-TOFMS of bacteria, reproducibility of spectra has been problematic for both methodological and biological reasons.[11−14] Much emphasis has been placed on sample preparation techniques as an important factor in developing sensitive MALDI-TOFMS methods for microorganism identification.[9,13−20] In addition, the quality of the MALDI mass spectrum

* Corresponding author. Instrumentation and Biophysics Branch, U.S. Food and Drug Administration, 5100 Paint Branch Parkway, HFS-717, College Park, MD 20740.

† U.S. Food and Drug Administration.

‡ BioAnalyte Inc..

(1) Groisman, E. A., Ed. *Principles of Bacterial Pathogenesis*; Academic Press: San Diego; 2001.

(2) Salyers, A. A.; Whitt, D. D. *Bacterial Pathogenesis: A Molecular Approach*. ASM Press: Washington, DC; 1994.

(3) 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.

(4) Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1992−1996.

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

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

(7) Fenselau, C.; Demirev, P. A. *Mass Spec. Rev.* **2001**, *20*, 157−171.

(8) Lay, J. O., Jr. *Mass Spec. Rev.* **2001**, *20*, 172−194.

(9) Holland, R. D.; Duffy, C. R.; Rafii, F.; Sutherland, J. B.; Heinze, T.; Holder, C. L.; Voorhees, K. J.; Lay, J. O. *Anal. Chem.* **1999**, *71*, 3226−3230.

(10) Conway, G. C.; Smole, S. C.; Sarracino, D. A.; Arbeit, R. A.; Leopold, P. E. *J. Mol. Microbiol. Biotechnol.* **2001**, *1*, 103−112.

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

(12) Domin, M. A.; Welham, K. J.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 222−226.

(13) 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.

(14) Lay, J. O., Jr.; Holland, R. D. *Methods Mol. Biol.* **2000**, *146*, 461−488.

(15) Birmingham, J.; Demirev, P.; Ho, Y. P.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 604−607.

(16) Madonna, A. J.; Basile, F.; Ferrer, I.; Meetani, M. A.; Rees, J. C.; Voorhees, K. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2220−2229.

(17) Evason, D. J.; Claydon, M. A.; Gordon, D. B. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 669−672.

(18) Arnold, R. J.; Reilly, J. P. *Anal. Biochem.* **1999**, *269*, 105−112.

(19) Lay, J. O., Jr. *Trends Anal. Chem.* **2000**, *18*, 507−516.

(20) Smole, S. C.; King, L. A.; Leopold, P. E.; Arbeit, R. D. *J. Microbiol. Methods* **2002**, *48*, 107−15.

10.1021/ac0258958 CCC: $22.00
Published on Web 10/10/2002

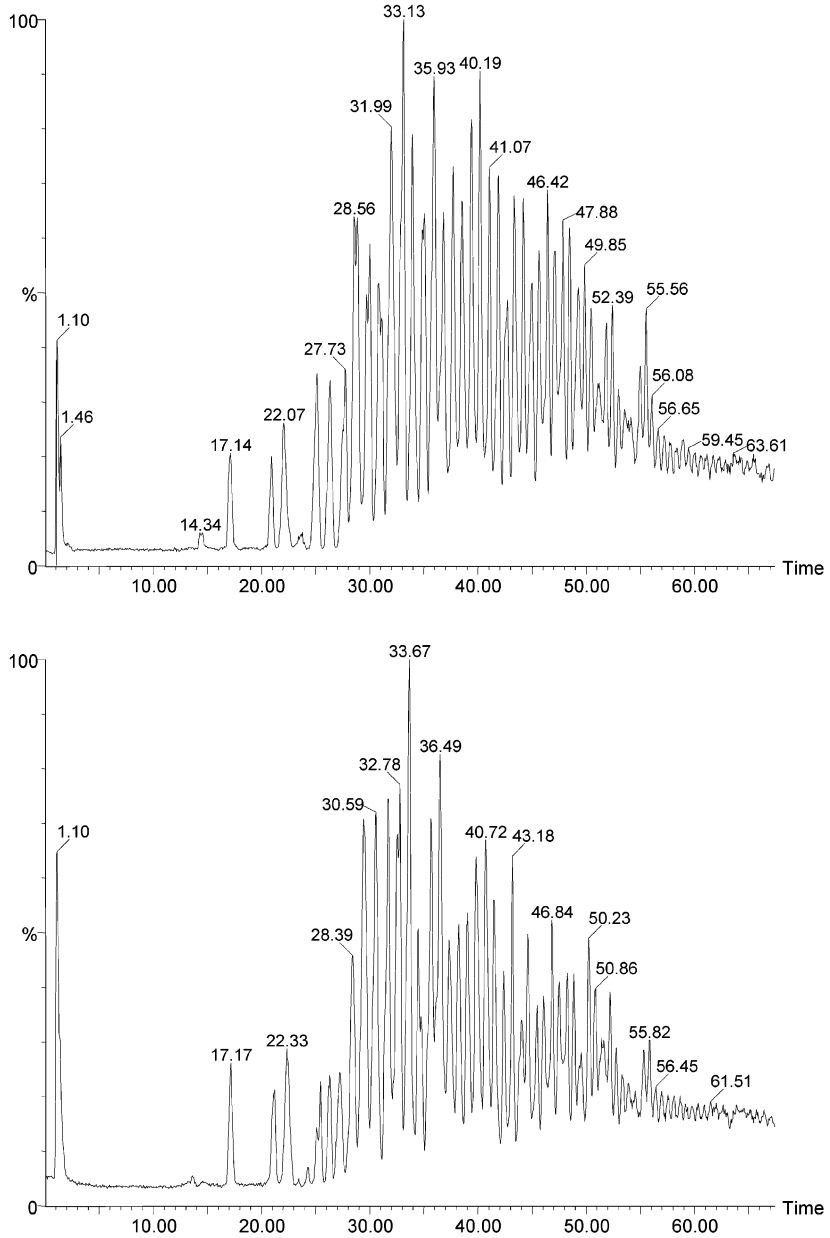

Figure 1. Chromatograms of the bacterial cell lysates extracted from two different strains of *V. parahaemolyticus*. The top chromatogram is that of BAC-98-3547, and the bottom is that of VP47 (a pathogenic strain).

is dependent upon the concentration of the analyte.[21] Other issues that can cause problems for MALDI-MS analysis of bacteria include salt or detergent interference, incorrect matrix-to-analyte ratio, or poor choice of matrix material. In addition, as the size of the analyte increases, the resolution decreases and a mass accuracy of only 1−5 ppt is expected.[22] Combined, these experimental factors influence both the number and type of proteins observed by MALDI-MS, thereby limiting the amount of information obtained and decreasing the probability that unique proteins will be observed.

Electrospray ionization mass spectrometry (ESI-MS) has been developed concurrently with MALDI as an alternative means of biological analysis. ESI-MS is easily interfaced with liquid chromatography (LC), offering better run-to-run reproducibility, the capability for on-line sample processing,[23] and increased sensitivity. Additionally, by combining chromatography with ESI-MS, salt adducts are reduced or eliminated, allowing for improved spectral quality, resulting in more accurate deconvolution of the multiply charged protein peaks. The improvement in sensitivity and overall response for proteins in LC/ESI-MS is the product of sequential elution of proteins from the column, which allows the detection of more proteins by greatly reducing ion suppression, since fewer components are present during the ionization process. Therefore, although chromatography adds to the time required for data acquisition, the spectral output is of higher quality and produces a more accurate representation of the overall abundance of specific proteins present in a mixture.

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

(22) Jensen, P. K.; Pasa-Tolic, L.; Peden, K. K.; Martinovic, S.; Lipton, M. S.; Anderson, G. A.; Tolic, N.; Wong, K. K.; Smith, R. D. *Electrophoresis* **2000**, *21*, 1372−1380.

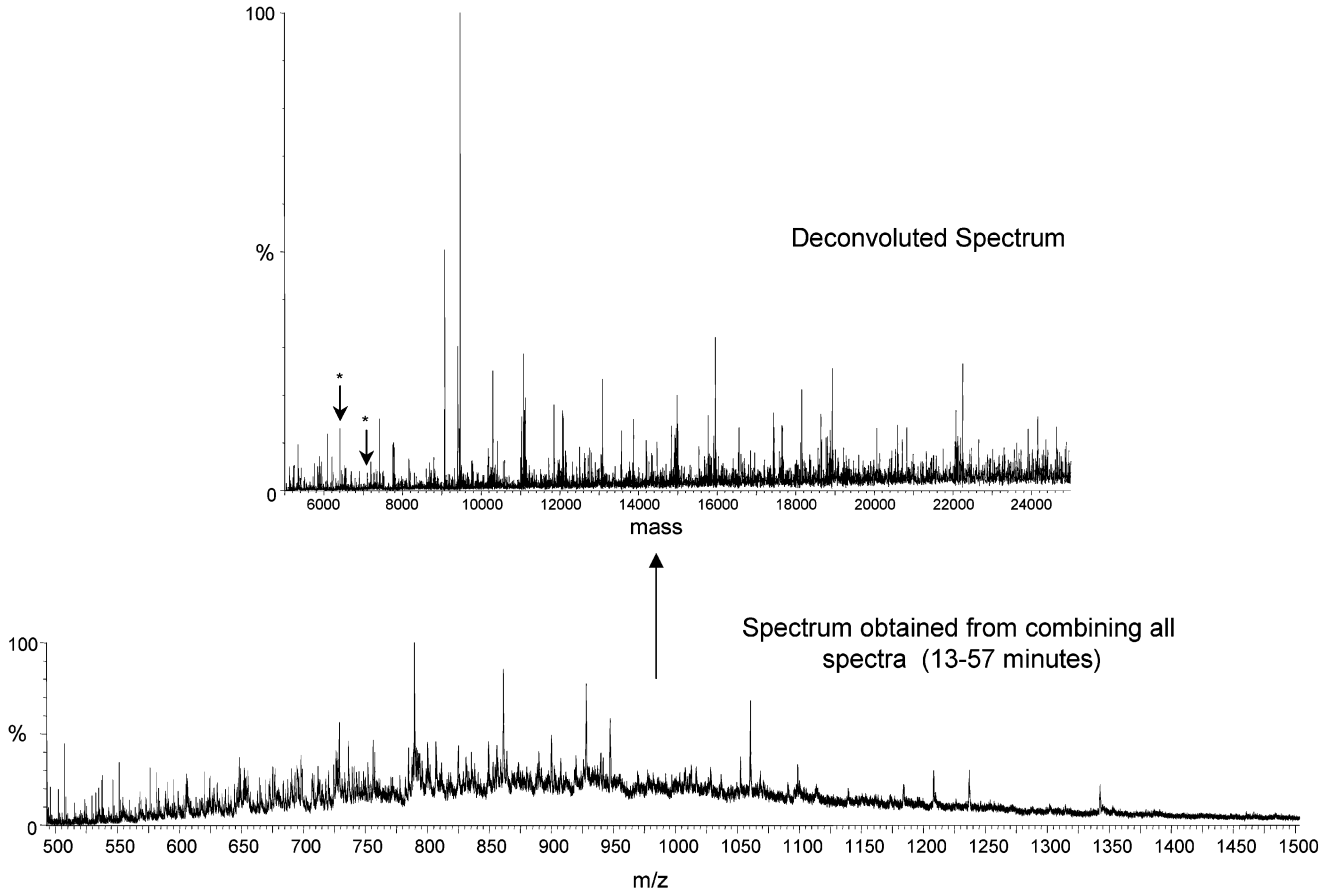

Figure 2. All spectra obtained between 13 and 57 min are combined into a single spectrum and then deconvoluted using MaxEnt 1. The resulting molecular weight spectrum is noisy and hampers the identification of low abundance proteins.

Although capillary LC/MS overcomes many of the limitations found in MALDI experiments with whole bacteria, the chromatography does not easily lend itself to provide a useful "profile" of the bacterial strain. Nevertheless, several studies have used LC/ES-MS for the rapid separation and mass analysis of whole-cell lysates.[24−30] Lubman and co-workers have also reported several 2-D separation methods in order to improve the chromatographic resolution for this type of analysis.[31−33] The goal of these studies was to generate a two-dimensional image of proteins expressed in cells, which resembles information provided by two-dimensional polyacrylamide gel electrophoresis (2-D PAGE). These virtual images show a distinctive reproducible protein pattern that is associated with a particular cell line. The limiting factor in these experiments is the generation of very large data sets, which require extensive postacquisition processing, a procedure difficult to reproduce manually and not yet automated.

This article describes an automated procedure for generating bacterial protein profiles from the LC/ES-MS chromatogram of bacterial cell lysates. The method translates the chromatographic and multiply charged protein information into one comprehensive mass versus intensity spectrum, a process that is time- and labor-intensive to complete manually. Once the data is converted, a number of software tools can be used to compare bacterial profiles and monitor changes in protein expression that are linked to pathogenicity. Once significant changes in protein expression have been identified, the proteins of interest can be singled out, identified, and examined for alterations in sequence or posttranslational modifications.

EXPERIMENTAL SECTION

Acetonitrile and HPLC grade water were purchased from J. T. Baker (Phillipsburg, NJ). Acetic acid was purchased from Sigma-Aldrich Chemical Co. (St. Louis, MO). All were used without further purification.

Cells from two strains of *Vibrio parahaemolyticus* (VP47 and Bac-98-3547) were used for these experiments. Strain VP45 is a

(23) Xiang, F.; Anderson, G. A.; Veenstra, T. D.; Lipton, M. S.; Smith, R. D. *Anal. Chem* **2000**, *72*, 2475−2481.
(24) Liang, X.; Zheng, K.; Qian, M. G.; Lubman, D. M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1219−1226.
(25) Opitek, G. J.; Lewis, K. C.; Jorgenson, J. W.; Anderegg, R. J. *Anal. Chem.* **1997**, *69*, 1518−1524.
(26) Chen, Y.; Wall, D.; Lubman, D. M. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 1994−2003.
(27) Chong, B. E.; Lubman, D. M.; Miller, F. R.; Rosenspire, A. J. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 1808−1812.
(28) Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Anal. Chem.* **1999**, *71*, 3894−3900.
(29) Chong, B. E.; Kim, J.; Lubman, D. M.; Tiedje, J. M.; Kathariou, S. *J. Chromatogr., B* **2000**, 167−177.
(30) Chong, B. E.; Hamler, R. L.; Lubman, D. M.; Ethier, S. P.; Rosenspire, A. J.; Miller, F. R. *Anal. Chem.* **2001**, 1219−1227.
(31) Wall, D. B.; Kachman, M. T.; Gong, S.; Hinderer, R.; Parus, S.; Misek, D. E.; Hanash, S. M.; Lubman, D. M. *Anal. Chem.* **2000**, *72*, 1099−1111.
(32) Chong, B. E.; Yan, F.; Lubman, D. M.; Miller, F. R. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 291−296.
(33) Kachman, M. T.; Wang, H.; Schwartz, D. R.; Cho, K. R.; Lubman, D. M. *Anal. Chem.* **2002**, *74*, 1779−1791.

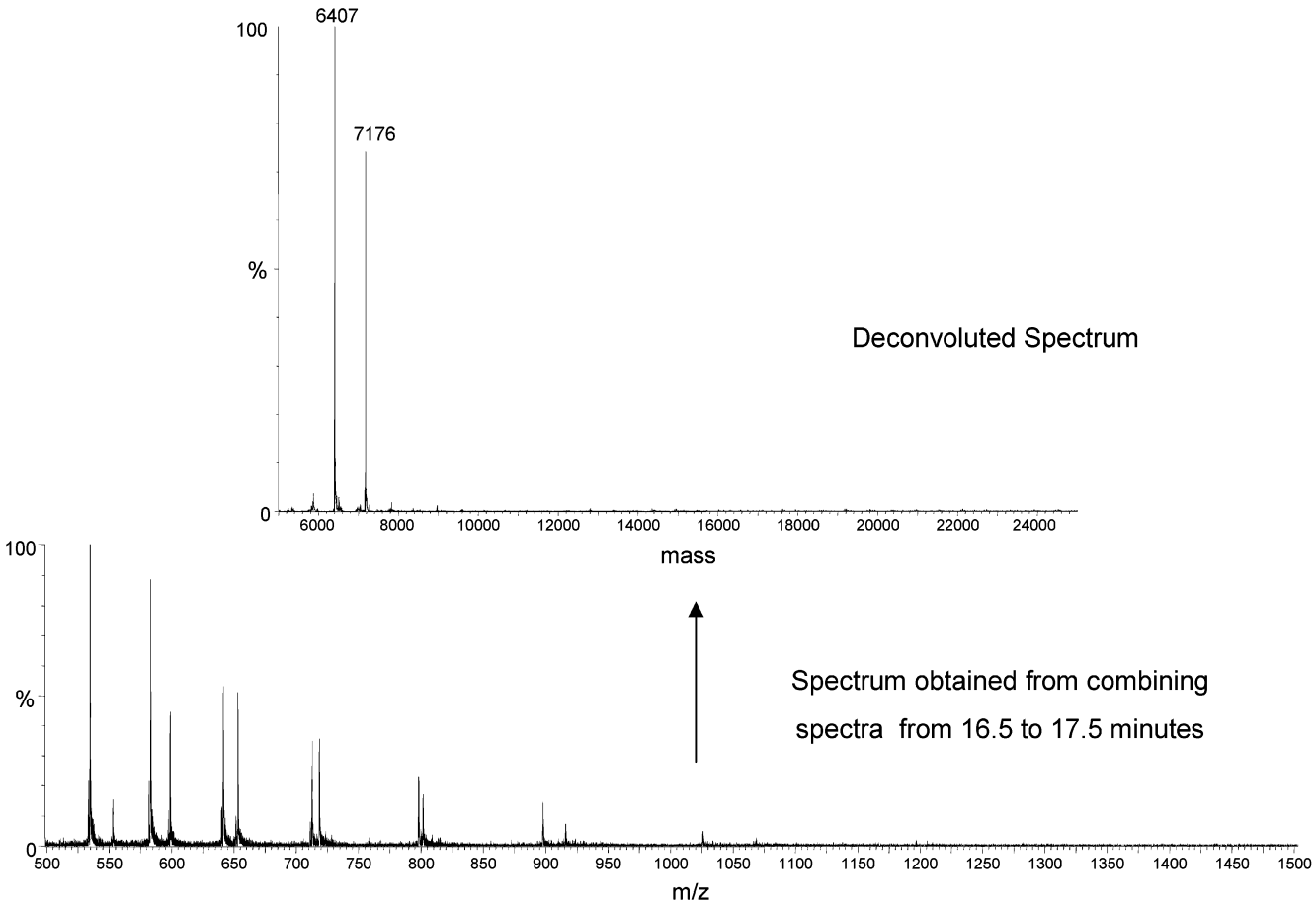

Figure 3. Spectra obtained between 16.5 and 17.5 min are combined into a single spectrum and then deconvoluted using MaxEnt 1. Proteins with a molecular weight of 6407 and 7176 are easily observed.

known human pathogen (03:K6), whereas strain Bac-98-3547 is an environmental strain (O4:K55) not associated with human pathogenicity. Cultures were grown for 24 h on tryptic soy agar purchased from Difco Laboratories (Detroit, MI) and supplemented with 2% NaCl. The cell isolates were suspended in ethanol and stored at 4 °C until needed.

Bacterial cells were lysed using a solution of acetonitrile, water, and formic acid (50:45:5). The whole cells were first centrifuged to form a loosely packed cell pellet. A 100-μL portion of this pellet was removed and mixed with 1 mL of the extraction solution. The cells were sonicated in this solution for 30 min, centrifuged to a pellet, and removed. The remaining solution was vacuum-concentrated to a total volume of 500 μL.

A Hewlett-Packard 1100 HPLC system (Palo Alto, CA) fitted with a 20-cm × 320-μm-i.d. LC column packed in-house with POROS 10 R2 packing (PerSeptive Biosystems, Framingham, MA) was used to separate the proteins of the whole-cell bacterial extract. A 2-μL sample was injected onto the column, and the separation was carried out at a flow of 20 μL/min with a shallow gradient (10−50% B in 50 min). Solvent A contained 0.5% acetic acid in water, and solvent B was 0.5% acetic acid in acetonitrile.

A Micromass (Manchester, U.K.) QTOF II was used to acquire the data in full-scan continuum mode with an m/z range from 100 to 2000. Data were processed using Masslynx v.3.5 software, and the multiply charged protein spectrum was deconvoluted into a molecular weight spectrum using Micromass's MaxEnt 1 program.

Automated analysis of the data files was performed with Retana, custom software written for this purpose by BioAnalyte Inc. The function of this program is to automate data processing subroutines within the data processing program and to produce a combined time and intensity text output file. Briefly, the program sums all data within a specified time interval, uses MaxEnt 1 to deconvolute the multiply charged ions, centers the result, performs a threshold selection, and reports the mass, intensity, and retention time of the protein in a text file. It continues this process across sequential portions of the chromatogram. All aspects of the subroutines, including retention times, mass windows, number of MaxEnt 1 iterations, and spectra to combine, can be controlled through Retana.

Upon completion of the Retana program, the text file contains a cumulative list of all the masses that were observed upon deconvolution of the individual summed spectra. This text file records mass, intensity, and retention time. The retention time information is held in the text file for the user to reference if a protein is singled out or deemed significant for further study, and thereby facilitates the isolation and purification process. It can also be used to verify that proteins of the same mass are actually two unrelated proteins, as indicated by their different retention times. The mass and intensity list is converted into a file that Masslynx is able to read via the Databridge program. Alternatively, the text file can be read by a graphing program, such as Grapher. In this paper, Grapher v.3 (Golden Software, Inc., Colorado) was used to manipulate and display the data.

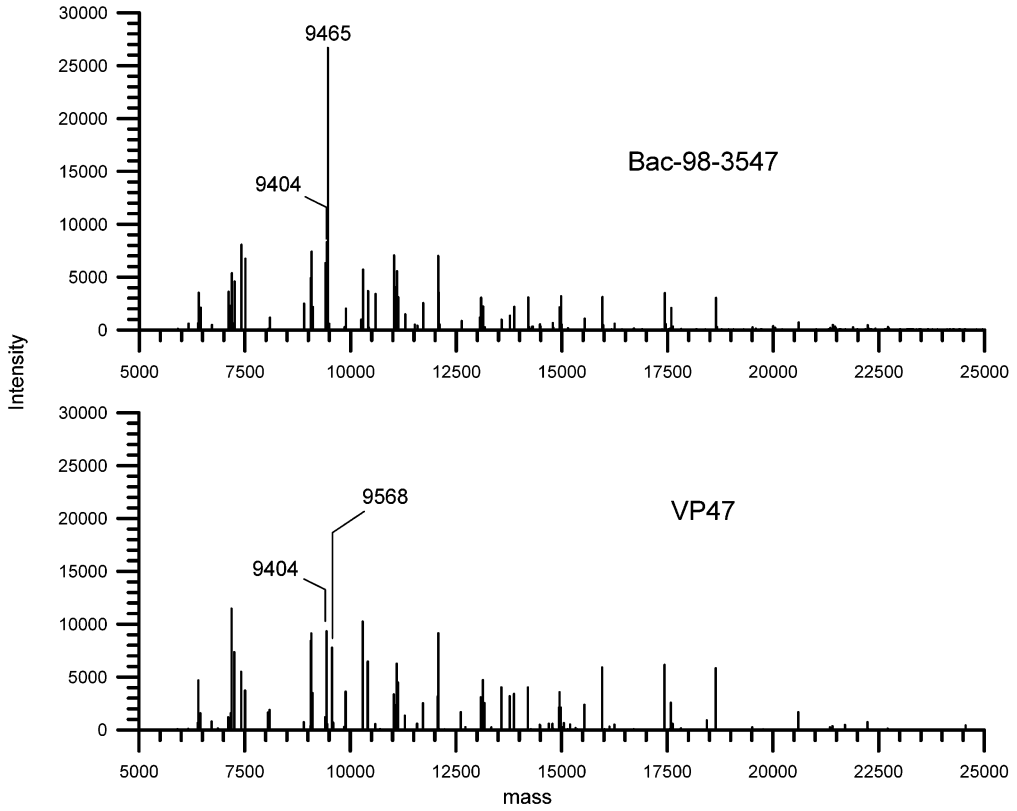

Figure 4. Single representative cellular protein spectrum of the nonpathogenic strain (top) and pathogenic strain (bottom) of *V. parahaemolyticus*.

RESULTS AND DISCUSSION

Proteins were extracted from cells of two different strains of *V. parahaemolyticus* and injected onto the HPLC column. The chromatograms are shown in Figure 1. Although chromatographic peak intensity is somewhat lower in the VP47 sample, the retention times and peak patterns are similar. However, some of the retention times vary by as much as 0.5 min, which is most likely due to poor reproducibility of the LC gradient. In addition, proteins may coelute, and if peak resolution is not sufficient, they will not be detected by examination of the chromatogram alone. Therefore, simply comparing chromatograms without associated mass information is not a sufficient means of determining the presence of proteins that are unique to a bacterial strain.

To obtain a high-quality final spectrum, it is essential that spectra be combined in short intervals and subsequently deconvoluted, rather than summing the scans for the entire chromatographic run. To illustrate this point, all mass spectra obtained from 13 to 57 min were summed and combined into a single spectrum and deconvoluted using MaxEnt 1. The resulting deconvoluted spectrum (Figure 2) appears noisy, contains a small number of well-defined protein masses, and makes it impossible to distinguish between less abundant proteins and noise. The poor quality of the final spectrum can be attributed to the great number of multiply charged ions obtained in the summed spectrum, creating a peak at every mass. As the number of successive scans added to the spectrum increases, so does the overall noise, the result of which is the successful deconvolution of only the most intense proteins present in the sample. Analyzing smaller time segments of the LC analysis (0.5−1 min) provides for an unequivocal identification of the protein, because there are fewer multiply charged ions present in the summed spectrum and improved signal-to-noise. Figure 3 shows the spectrum obtained from combining all spectra between 16.5 and 17.5 min and the resulting deconvolution spectrum showing the presence of proteins with masses of 6407 and 7176 Da. These proteins are indicated by an asterisk in Figure 2. When the entire experiment is combined into one spectrum and then deconvoluted, proteins such as these are lost in the noise. For this reason, the spectra must be combined in smaller intervals and separate deconvolution processes must be performed on the individual intervals.

Manually processing each chromatographic peak is not only time- and labor-intensive, but also difficult to reproduce. To overcome these problems and to provide a consistent data format that was independent of retention time, a number of data processing subroutines were automated to produce a single representative cellular protein spectrum. There are three processes (summing, deconvoluting, and centering) associated with each iteration of the Retana program.

In the chromatograms shown in Figure 1, proteins began eluting after 13 min and concluded by 60 min. On the basis of this information, Retana began processing the data at 13.0 min in 60-s intervals. Although smaller time intervals could have been used, most of the chromatographic peak widths are greater than 30 s, and therefore, no data is lost as a result of using 1-min intervals for processing. To prove this point, the postprocessing analysis was performed on the data set using 30-s time intervals. No change in the final spectrum was observed (data not shown). Instead, the processing time simply lengthened with no observable benefit. In addition, no change in the final spectrum is observed

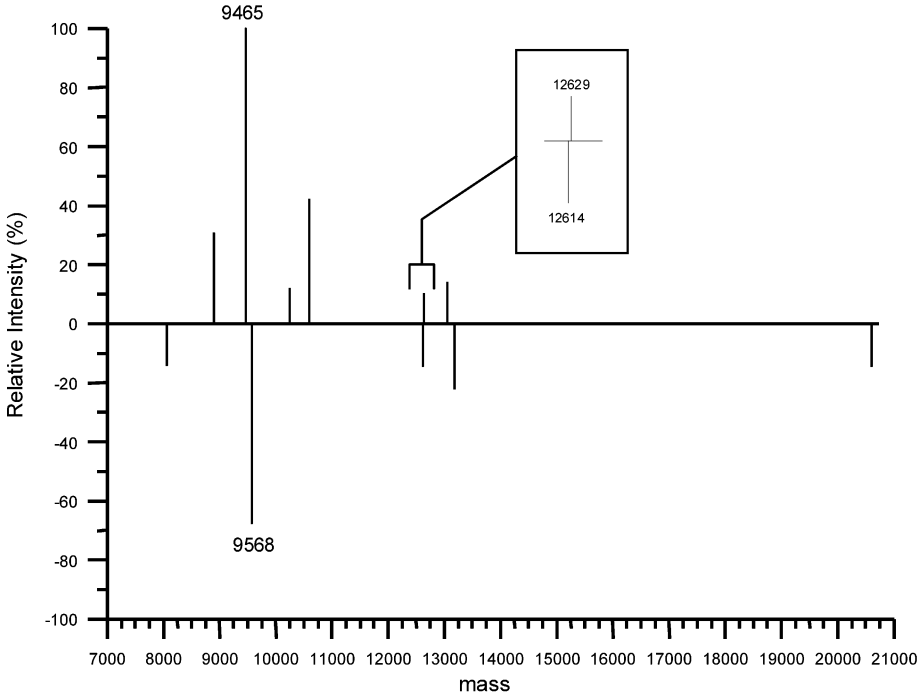

Figure 5. Representative protein profile of the two bacterial strains is observed as differing only in 11 peaks.

when the start time of the Retana program is varied. For example, the spectrum showed no differences, regardless of whether Retana began analysis at 13.0 min or at 13.5 min. Provided that the time intervals are small, even proteins that are low in abundance are detected and identified by Retana and are unaffected by changes in time interval or processing start time.

When deconvoluting multiply charged spectra with MaxEnt 1, it is important to recognize sources of artifacts and noise created by the process so that they can be removed from the final list of masses used to compile the reconstructed bacterial profile. One important source of artifacts created by MaxEnt 1 is incomplete deconvolution of the multiply charged spectrum, the result of which is the introduction of masses corresponding to half and twice the molecular weight of the correct mass. This problem becomes more significant as the signal-to-noise ratio decreases. To prevent these artifacts of the deconvolution program from appearing in the final spectrum, Retana regards the most intense peak in the MaxEnt 1 display as the protein mass of interest. Any peak with either one-half or twice the molecular weight of this peak is ignored and not written to the text file.

Noise is another source of erroneous peaks in a deconvoluted spectrum. MaxEnt 1 will attempt to deconvolute a multiply charged spectrum, regardless of whether there is sufficient signal present, resulting in a deconvoluted spectrum with a large number of apparent protein masses that are due only to noise. For example, if a retention time window is specified that falls between two chromatographic peaks, the combined spectra will produce a deconvoluted spectrum that can consist of hundreds of masses (noise). Retana not only allows the user to specify a peak intensity threshold (in this case, 10%) used to differentiate signal from noise, but also allows the user to specify a maximum number of peaks that should be observed in the deconvoluted spectrum. If the user-specified number of peaks (in this case, 100) reported by MaxEnt 1 is surpassed, Retana disregards the data as noise and does not write the peak masses to the text file. In this manner, noise is differentiated and removed.

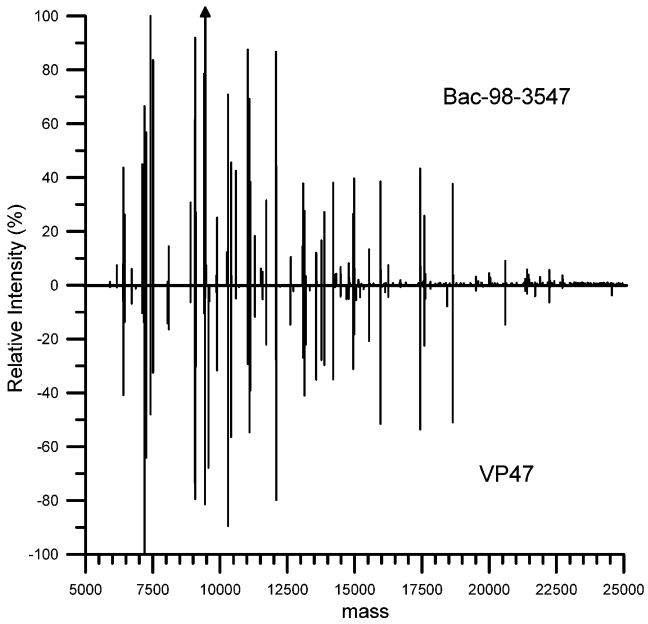

Figure 6. When the spectra of Bac-98-3547 is normalized to the second largest peak in the display, it is evident that there is little difference in the relative intensities of the majority of the proteins in the sample.

Reconstructed spectra of all the proteins observed in bacterial cell lysates of two different strains of *V. parahaemolyticus* are shown in Figure 4. When the spectra are superimposed, most of the protein masses are the same, as would be expected. However, some significant differences in protein masses are observed (Figure 5) including the absence of the peak at 9465 Da in the pathogenic strain of *V. parahaemolyticus*, a peak that dominates the spectrum of the nonpathogenic strain of the bacteria. Since

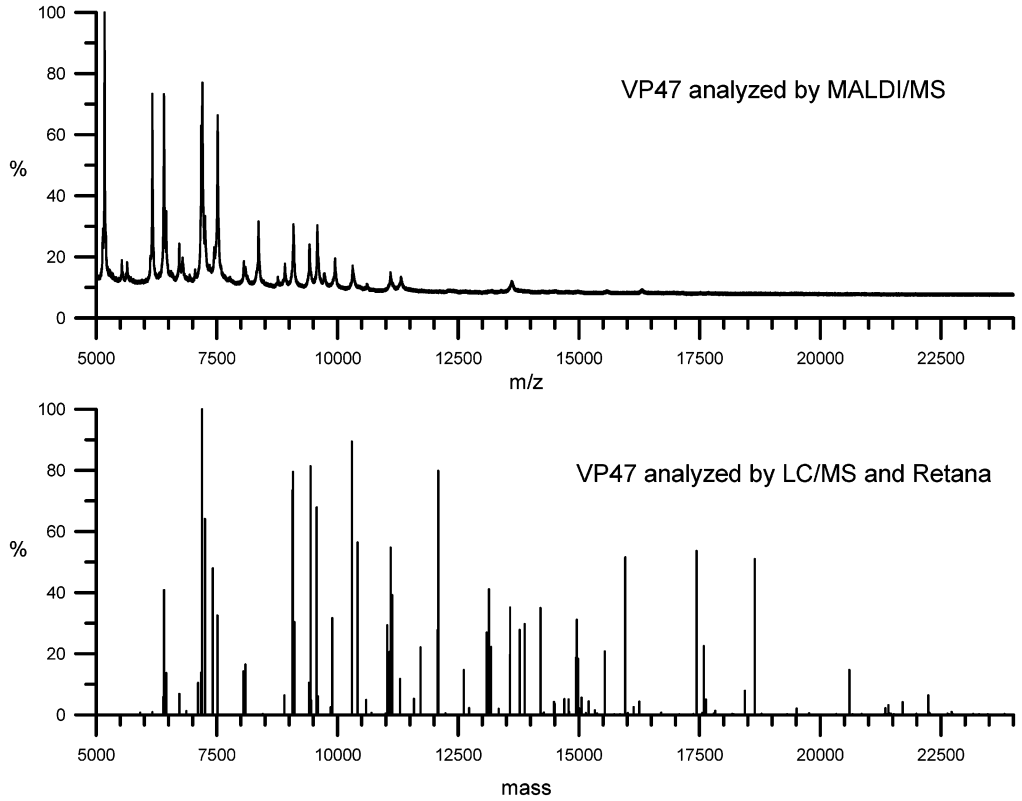

Figure 7. Comparison of the data obtained from MALDI-MS and that obtained using LC/MS with Retana.

Retana also records retention time in its text file, these proteins can be easily isolated and further studied to determine whether they are an indicator of bacterial pathogenicity or other biological significance.

The presence or absence of a protein is not the only possible indicator of pathogenicity. Instead, the amount of a particular protein in a sample may serve as a trigger for a signaling process to turn on or off that which makes the strain pathogenic. Therefore, comparing the protein peak intensities of the two bacterial strains may identify a characteristic marker of bacterial pathogenicity. If the spectra of the two *V. parahaemolyticus* are normalized, the peak at 9465 Da dominates the spectrum of Bac-98-3547, and it is difficult to compare the less intense peaks. However, if the spectrum of Bac-98-3547 is normalized to the second largest peak in the display (Figure 6), significant changes in the intensities of only a few peaks are observed. There are some differences in the intensities of the protein peaks in the range of 7000−8000 Da. Again, since Retana records retention time, these proteins can be easily isolated and further studied. Reanalysis of samples prepared on different days demonstrates that the protein ratios are highly reproducible. Therefore, by normalizing the final profile spectrum to proteins with the same mass in several different strains, it is not only possible to identify proteins of different mass but also identify significant changes in level of protein expression.

Retana is used to automate the data processing of LC/MS data to generate a comprehensive spectrum that can be used as a bacterial protein profile. Although this method is neither as rapid nor as simple as that conducted by MALDI-MS, the information obtained is richer, because the proteins do not experience mass discrimination nor selective ionization, as is common in MALDI experiments. Figure 7 compares the spectra of VP47 observed via whole cell analysis by MALDI with that obtained by LC/MS and processed with Retana. Notable differences between the two spectra include a high mass discrimination effect and fewer observed masses in the MALDI spectrum. Although the protein profile in the MALDI experiment is sufficient to identify the bacteria as a *V. parahaemolyticus*, there is insufficient information to differentiate the strain. As a research tool for profiling bacterial protein profiles, LC/MS with postacquisition processing offers a number of advantages, including display of a greater number of proteins, highly reproducible protein mass, and abundance information, in addition to improved mass accuracy and resolution. It is also important to note that should a unique protein be identified in the MALDI experiment, it would then be necessary to develop a methodology to purify enough protein for sequence analysis, an unnecessary step in the LC/MS experiment, since the retention time of the unique protein is known and fraction collection can be easily performed on a subsequent LC run.

ACKNOWLEDGMENT

We thank Dr. Angelo DePaola of the Food and Drug Administration's Gulf Coast Seafood Laboratory (Dauphin Island, AL) for providing the *Vibrio parahaemolyticus* cells.

Received for review June 27, 2002. Accepted September 11, 2002.

AC0258958