Liquid Chromatography/Mass Spectrometry Characterization of *Escherichia coli* and *Shigella* Species

Robert A. Everley,[a,b] Tiffany M. Mott,[a] Shane A. Wyatt,[a] Denise M. Toney,[a] and Timothy R. Croley[a,b]

[a] Division of Consolidated Laboratory Services, Commonwealth of Virginia, Richmond, Virginia, USA
[b] Department of Chemistry, Virginia Commonwealth University, Richmond, Virginia, USA

Liquid chromatography/quadrupole time of flight mass spectrometry (LC/QTOF MS) utilizing electrospray ionization was employed to monitor protein expression in *Escherichia coli* and *Shigella* organisms. Comparison with MALDI/TOF-MS revealed more proteins, particularly above 15 kDa. A combination of automated charge state deconvolution, spectral mirroring, and spectral subtraction was used to reveal subtle differences in the LC/MS data. Reproducible intact protein biomarker candidates were discovered based on their unique mass, retention time, and relative intensity. These marker candidates were implemented to differentiate closely related strain types, (e.g., two distinct isolates of *E. coli* O157:H7) and to correctly identify unknown pathogens. This LC/MS approach is less labor-intensive than pulsed-field gel electrophoresis, affords greater specificity than real-time PCR, and requires no primers or antibodies. Additionally, this approach would be beneficial during outbreaks of foodborne disease or bioterrorism investigations by complementing methods typically used in diagnostic microbiology laboratories. (J Am Soc Mass Spectrom 2008, 19, 1621–1628) © 2008 American Society for Mass Spectrometry

Two of the top 10 leading causes of food and waterborne illness outbreaks reported to the Centers for Disease Control and Prevention from 1972 to 2000 were Shiga toxin producing *E. coli* (STEC) and *Shigella* organisms [1]. The threat of these pathogens being used intentionally as biowarfare agents also exists. STEC and some species of *Shigella* produce a toxin that is classified as a Category B select agent and deliberate use of these bacteria has been documented [1, 2].

Clinical manifestations (malaise, abdominal pain, diarrhea etc.) of exposure do not unambiguously identify their cause. Therefore, analytical methods are needed to gain further insight. The three most commonly used analytical methods in diagnostic microbiology laboratories are enzyme immunoassay (EIA), real-time polymerase chain reaction (PCR), and pulsed-field gel electrophoresis (PFGE). EIA methods are sensitive, often eliminating the need for cultural enrichment, but lack the specificity to be considered confirmatory. Real-time PCR is a rapid and sensitive approach requiring 0.5 to 4 h postculture to perform. However, strain-specific and often species-specific primers are unavailable or impractical and, for this reason, the specificity required in outbreak investigations is not typically afforded by this method. An important consideration is that the mere presence of a gene does not guarantee that protein is being expressed; bacterial pathogens have been shown to contain genes that are not expressed [3–5]. Finally, the gold standard for providing strain level discrimination of bacteria during outbreak investigations is PFGE. While this technique possesses the desired specificity, it is not easily automated, is labor intensive, and requires a minimum of 2 d postculture.

Several different reviews on the analysis of bacteria all had one common conclusion—the strongest approach is an integrated one combining information from several different yet complimentary techniques [6–9]. The benefit of an integrated approach and the importance of bacterial pathogens to public health and homeland security act as a driving force behind the development of new methods of detection and characterization. The approach described here is unique in that it utilizes liquid chromatography/mass spectrometry (LC/MS) of intact proteins, to monitor protein expression in bacterial cells. Key differences between closely related strains may occur within the proteome to which genetic approaches are insensitive, e.g., post-translational modifications (PTMs), making this approach potentially advantageous.

Since much of the early work regarding the mass spectrometric analysis of intact bacterial proteins has involved matrix assisted laser desorption/time of flight

Address reprint requests to Dr. T. Croley, Division of Consolidated Laboratories, Commonwealth of Virginia, 600 N. 5th St., Richmond, VA 23219, USA. E-mail: tim.croley@dgs.virginia.gov

1044-0305/08/$32.00
doi:10.1016/j.jasms.2008.07.003

Published online July 15, 2008
Received December 14, 2007
Revised July 7, 2008
Accepted July 7, 2008

Table 1. The 10 known isolates examined in this study

Family	Enterobacteriaceae					
Genus	*Escherichia*				*Shigella*	
Species	*E. coli*				*S. flexneri*	*S. sonnei*
Serotype	ND (nonpathogenic)	O111:NM	O26:H11	O157:H7	ND	NA
Accession number	06-0004 06-0006	06-1440	06-1418	06-1439 06-1464	04-0497 06-0967	06-1362 06-1364

ND = not determined; NM = nonmotile; NA = not applicable.

mass spectrometry (MALDI/TOF-MS) [10–12], a brief comparison with this technique was performed. To examine the efficacy of this approach as a tool in diagnostic microbiology, a model set of 10 *Shigella* and *E. coli* clinical isolates were studied (Table 1) as a proof of concept. From these 10 isolates, putative biomarkers or biomarker candidates were discovered based on protein mass, retention time, and relative intensity, and evaluated for their reproducibility by performing five replicate analyses. Finally, the validity of these putative markers was challenged by applying them to a blind test of clinical isolates.

Experimental

Materials and Methods

HPLC grade solvents (acetonitrile, formic acid, and trifluoroacetic acid) were purchased from Fisher Scientific (Fairlawn, NJ) and 2-propanol was purchased from Honeywell, Burdick, and Jackson (Morristown, NJ). The water utilized for HPLC analysis was purified in-house to yield organic-free 18.3 MΩ × cm using an E-pure purification system (Barnstead International, Dubuque, IA). Sterile water that had been autoclaved and purified with a RiOs 5 Water Purification System (Millipore, Billerica, MA) was used during bacteria preparation.

Growth and Lysis

Isolates were obtained from the Virginia Division of Consolidated Laboratory Services. Biosafety level 2 (BSL2) procedures and facilities were utilized for sample handling and preparation. Trypticase soy agar plates containing 5% sheep's blood were used as the growth medium. Cells were grown for 24 h at a temperature of 37 °C in the presence of oxygen with 5% CO_2. After this growth period, cells were removed from the plate and placed in a test tube containing 1 mL of sterile water until the optical density reading reached 1.0 using a microscan turbidity meter (Dade Behring, West Sacramento, CA). A 500 μL aliquot of this suspension was washed three times with 500 μL of sterile water followed by centrifugation (6000 × *g* at room temperature for 5 min) to remove residual media. Finally, the cells were resuspended in 150 μL of the lysis solution (1:1 organic-free H_2O: acetonitrile, 0.1% vol/vol trifluoroacetic acid). After chemical lysis, the sample was again centrifuged (4100 × *g* for 4 min) at room temperature. Following centrifugation, 65 μL of supernatant was removed and placed in an autosampler vial for analysis.

MALDI/TOF-MS

Aliquots (1 μL) of (1:1) bacterial cell lysate and sinapinic acid (30 mg/mL) matrix solution were spotted onto a stainless steel MALDI target plate and allowed to air dry. Mass spectra were obtained with a Bruker Daltonics Ultraflex II (Billerica, MA) instrument operating in linear, positive ion mode. Mass spectra were acquired utilizing the following instrument parameters: pulsed ion extraction delay of 300 ns, ion source voltage one, 25 kV, ion source voltage two, 23.25 kV, and ion source lens voltage 6.20 kV. For each sample, mass spectra were acquired by accumulating 200 laser shots at 54% laser power in the *m/z* range of 4000–40,000 Da.

FlexAnalysis (Bruker Daltonics) software was used to generate peak lists after background subtraction, peak centering, and "...background subtraction, Gaussian smoothing, and peak centering." Data lists containing *m/z* values and corresponding peak intensities were exported as text files before mass spectral comparison by MS Manager (Advanced Chemistry Development Laboratories, Toronto, ON).

LC/QTOF MS Analysis

Intact proteins were separated by reversed-phase chromatography using an Acquity liquid chromatograph (Waters, Milford, MA). Gradient elution (5%–55% B in 60 min) was used at a flow rate of 0.225 mL/min, where A = H_2O (1% formic acid) and B = 2-propanol (1% formic acid). The column was a nonporous Prosphere P-HR 2.1 × 150 mm, 4 μm particle size (Alltech, Columbia, MD) operated at 50 °C. The autosampler was maintained at 15 °C before administering the injection volume of 20 μL.

A Q-TOF Premier (Waters) utilizing positive ion electrospray ionization was used for mass analysis. Ions were monitored from *m/z* 620–2450 and resolved in single reflectron (V) mode. The parameters employed in the MS method were optimized for sensitivity and resolution using bovine serum albumin. These values were +3.9 kV capillary voltage, 40 V cone voltage,

115 °C source temperature, 500 °C desolvation temperature, and 900 L/h desolvation gas flow.

Data Processing

All LC/MS data were processed using two software packages: Protrawler6 (BioAnalyte, Portland, ME) and MS Manager (Advanced Chemistry Development Laboratories). Protrawler6 software provided automated charge state deconvolution of multiply charged ions by first dividing the full-scan data from the chromatogram into time intervals (30 s) and then summing and deconvoluting the data from each interval. The process was repeated to obtain neutral masses of the proteins that eluted during each interval. A text file containing the neutral masses, intensities, and retention times was then created summarizing the results for each chromatogram [13]. Retention time information can be used for further study (e.g., fraction collection) of proteins of interest or to distinguish proteins of the same mass that differ in retention. The masses and intensities were used to create a single spectrum representing all of the proteins observed in the lysate using MS Manager.

To further facilitate biomarker candidate discovery, MS Manager was employed for spectral mirroring and spectral subtraction. Spectral mirroring allowed spectra to be mirrored along the abscissa, placing the baseline at the center of view. Spectral subtraction removed all common peaks between two spectra within a given mass accuracy so that only unique ones remained. For group and strain level comparisons involving multiple spectra, the text files of all isolates not in that group or strain were combined to create a cumulative spectrum, which was then used for subtraction. A subtraction window of ±2 Da was utilized as a baseline for initial discovery and then optimized for specific datasets with larger mass ions (>30 kDa).

Results and Discussion

LC/ESI-MS Versus MALDI-TOF/MS Comparison

MALDI-TOF exhibits certain advantages over LC/MS. One, by producing primarily singly charged ions, data interpretation is greatly simplified relative to ESI-MS. MALDI-TOF is also better suited for the analysis of complex mixtures; therefore prior separation (e.g., chromatography) is not required. Consequently, MALDI-TOF has a considerable throughput advantage ~2 min/sample (after deposition and drying) compared with ~2 h/sample (data acquisition and deconvolution) over LC/MS.

However, as can be seen in Figure 1, after automated charge state deconvolution with Protrawler6, spectra from LC/MS are as simple to interpret as MALDI data, and contain more high mass information. In addition to providing more proteins (particularly >15 kDa), other advantages to using LC/MS exist. These include improved mass resolution and mass accuracy, reproducibility, and more reliable quantitative data. The last two advantages stem from the uneven distribution of the sample across the spot on the MALDI target and variance in the placement and number of laser shots acquired on the sample.

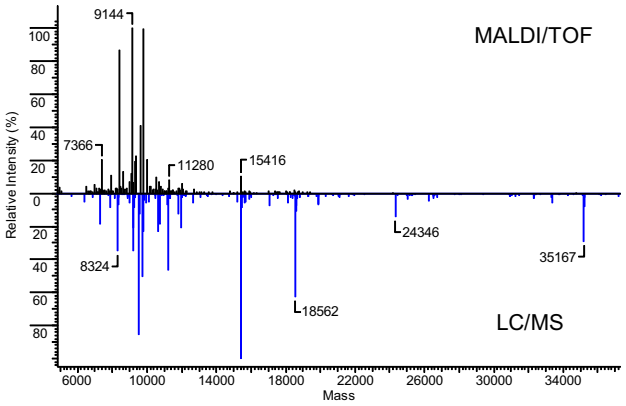

Figure 1. Comparison of MALDI/TOF-MS and LC/MS of the same *S. sonnei* lysate.

Having retention time information allows more to be known about the biomarker candidates. MALDI-TOF data is analogous to that obtained from a 1D gel, while LC/MS data is comparable to that acquired from a 2D gel (with obvious improvements in mass resolution and mass accuracy over gel-based approaches). The LC/MS approach also allows for distinctions of proteins of the same mass that differ in retention time. If the effluent from the LC is split, simultaneous fraction collection and MS analysis can be performed, and the collected fractions can be used for further study (e.g., sequencing). Protein isolation for further study can not be performed by MALDI-TOF; either LC or tandem mass spectrometry would be required.

Figure 1 depicts a comparison of MALDI/TOF and LC/QTOF data using the same sample preparation and protein extraction procedures for a *S. sonnei* isolate. Reasons for the differences in observed proteins may include difficulty in optimizing MS conditions over such a wide *m/z* range (4000–40,000 Da, Δ = 36,000 Da) with MALDI-TOF compared with (620–2450 Da, Δ = 1830 Da) during ESI. The complexity of the lysate is another issue that may lead to ion suppression and/or detector saturation. Other factors such as the matrix and the acid content of the matrix solution can also play a role in the observation of higher mass ions [14], and finding optimal matrix conditions over a wide *m/z* range can be challenging. Although the LC step causes decreased throughput, this step is likely part of the reason more proteins are observed. Often, distinctions between closely related strains may involve only one or a few proteins and, for this reason, the increased information content and protein yield observed by the LC/MS approach is likely advantageous and deemed worthy of further investigation.

Biomarker Candidate Discovery

The word "discovery" here denotes the process of going from a sample containing hundreds of proteins to a small list of unique proteins that have potential identification utility. These proteins are still in the candidacy phase as they require further validation by examining a large list of unknowns. Further structural characterization of the qualitative candidates to determine whether they are unique by sequence or by modification may also prove beneficial. In the first step of discovery, chromatographic data is collected in full-scan mode. Next, automated charge state deconvolution is performed to yield a single mass spectrum representing all of the proteins observed in the chromatogram. The spectra are then mirrored (Figure 2a) and subtracted, revealing unique masses (Figure 2b). As seen in Figure 2b, numerous peaks appear to be unique to each isolate after subtraction. However, many of these peaks were not reproducible and may have been artifacts from the deconvolution process. For this reason, a protein was deemed a biomarker candidate only if its unique mass, retention time, and or relative intensity was observed in each of the five repeated experiments. Spectra from the 10 isolates listed in Table 1 were examined to find reproducible biomarker candidates whose presence or absence could be used to identify unknown samples. To determine the specificity of the technique, a search for biomarker candidates was made at each taxonomic level (e.g., species, serotype, and strain).

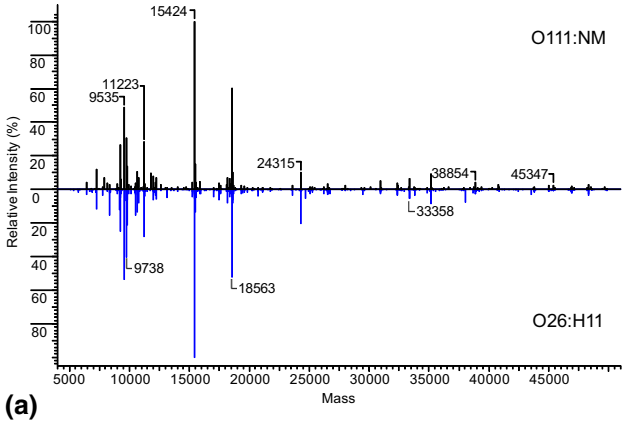

Figure 2. (**a**) Integrated mass spectra for all proteins observed after automated charge state deconvolution for the indicated serotypes. (**b**) Results after mass spectral subtraction using a ±2 Da window leaving only unique masses for each serotype.

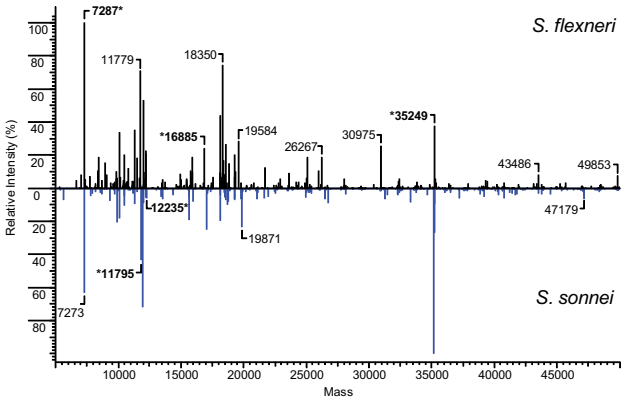

Figure 3. Mass spectral comparison of two species of *Shigella* treated as in Figure 2. Masses unique to each species are denoted in bold-type and with an asterisk.

Qualitative Markers: Mass and Retention Time

Real-time PCR primers for *Shigella* speciation are commercially unavailable. However, with LC/MS, distinctions between *Shigella* species were observed. Figure 3 depicts mirrored and subtracted spectra for *S. flexneri* and *S. sonnei*. The masses in bold marked with asterisks were found in both isolates of that *Shigella* species during each of the five replicate analyses, yet were not observed in any of the other eight isolates studied. These species-specific marker candidates also enabled the genera *Escherichia* and *Shigella* to be distinguished, even though no genus-specific biomarker candidates were observed. For example, the protein at mass 12,235 was unique to *S. sonnei* and was therefore not associated with the species *S. flexneri* or the genus *Escherichia*.

The protein at mass 7287 unique to *S. flexneri* has the same retention time (26.4 min) and nearly the same mass as a 7273 Da protein present in all of the *E. coli* and *S. sonnei* isolates studied. This mass difference of 14 Da could be due to a PTM (e.g., methylation), an amino acid substitution (e.g., I for V), or some combination of the two. Either way, such a small difference would likely go unnoticed in a gel-based approach, or when using a detector with less specificity such as ultra-violet or fluorescence spectroscopy.

As an example of serotype differentiation by this approach, the two *E. coli* O157:H7 isolates were compared against the other eight isolates. During this comparison, a protein at mass 18,996 was discovered unique to this serotype, thereby demonstrating the ability of this method to distinguish enterohemorrhagic *E. coli* (EHEC) serotypes (e.g., O157:H7, O26:H11, and O111:NM), which are otherwise indistinguishable by

Table 2. Group-specific qualitative biomarker candidates. Mass (±2 Da) is listed first followed by retention time (±0.5 min) in parentheses

Group	*E. coli* O157:H7	non-O157:H7 EHEC	*S. flexneri*	*S. sonnei*
Unique proteins	18,996 (43.3)	15,478 (27.1)	35,250 (31.4)[a,b]	11,795 (27.3)
		24,315 (38.5)	16,886 (26.8)	12,235 (45.4)
			7287 (27.9)[b]	

[a]Mass tolerance of ±3 Da.
[b]Proteins that may be quantitative markers. These marker candidates were present in all five replicates.

clinical symptoms [15]. Table 2 contains the masses and retention times for proteins that were found unique to a group such as to both O157:H7 or to both *S. sonnei* isolates etc. Using the experimental conditions described here, the proteins listed in Table 2 could be used to identify unknowns based on their presence or absence.

The two non-O157:H7 EHEC have peaks that identify them as a group as well (Table 2), but when each individual isolate (06-1440 or 06-1418) was compared against the other nine, no unique peaks were found. It was suspected, however, that one of the isolates might share a genetic similarity with some of the other eight isolates that was not shared with the other non-O157:H7 EHEC. For this reason, the O111:NM (nonmotile) and O26:H11 spectra were subtracted only against each other. During this comparison, a protein at 11,779 Da having a retention time of 27.0 min was found unique to O111:NM (Figure 2b). A protein of this same mass and retention time has also been observed in *S. flexneri* and *E. coli* O157:H7 isolates. Accordingly, during a blind test these two *E. coli* serotypes could be distinguished first by looking for the group-specific peaks listed in Table 2, which would classify them as a non-O157:H7 EHEC, then observing a protein at 11,779 Da with a retention time of 27.0 min, which would indicate that the sample was *E. coli* O111:NM.

In epidemiological and forensic investigations, techniques that can characterize bacteria at the strain level are desirable for establishing cluster or outbreak relationships via strain relatedness. Highly specific characterization is needed to detect and pinpoint the source of an outbreak, such as a particular produce manufacturer or suspected bioweapons facility. To this end, strain level comparisons between *E. coli* O157:H7 isolates were made. One O157:H7 isolate studied, accession no. 06-1464, has shown a reproducible protein at 14,880 Da eluting at 26.9 min not observed in the other O157:H7 isolate, accession no. 06-1439, or any of the other *E. coli* or *Shigella* samples. The differences observed between these two O157:H7 isolates indicate that the method described here is not only capable of identifying bacteria, but also of discerning small phenotypic differences, which could be indicative of the pathogen's origin and growth environment. With the exception of PFGE, which indicated ~98% similarity, other established techniques (e.g., serology) found these two isolates to be identical. In addition to the value of establishing strain relatedness during outbreak investigations, the ability to distinguish two strains (such as the ones described above) that while genetically similar are epidemiologically unrelated is also significant. Figure 4 depicts the comparison of the two *E. coli* O157:H7 spectra with PFGE results in the inset.

Analogous to PFGE, in which sequencing of the chromosomal fragments is not performed [16], this approach does not involve sequencing of the biomarker candidates. The justification is that reproducible biomarker candidates have been observed allowing for characterization at the strain level without knowing the actual identity of the proteins involved. Therefore, this approach could potentially be applied to bacteria whose genomes have not been sequenced. In contrast, proteomic approaches that rely upon database results for identification purposes would have little utility for such bacteria. Finally, circumventing protein sequencing allows the avoidance of a timely digestion step, resulting in a reduced analysis time.

Distinction of Isobars Differing in Retention

Since proteins with larger quantities of, or more easily accessible, hydrophobic regions will stay adsorbed to the column longer [17], retention time can therefore be used to distinguish two different isobaric proteins. This is critical when a sample has two or more different proteins of approximately the same mass. Such was

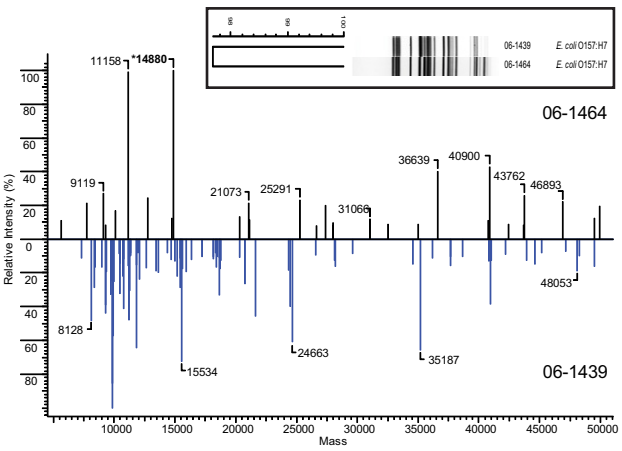

Figure 4. Strain-level comparison of two *E. coli* O157:H7 after treatment as in Figure 2. The gold standard method, PFGE analysis, determined ~98% similarity between the two isolates after 48 h (inset).

found to be the case for *S. flexneri* 04-0497. This strain of *S. flexneri* has two proteins within 2 Da of mass 18,121 that differ in retention time by nearly 16 min (approximately 13% B at a gradient slope of 0.83% B/min). One protein that eluted at 37.1 min had been observed in both *E. coli* O157:H7 and both *S. flexneri* isolates studied. The other protein, however, eluted at 21.3 min and was present only in *S. flexneri* 04-0497. This is an example where techniques providing mass alone (e.g., MALDI/TOF-MS) can be insufficient to recognize some biomarker candidates.

Quantitative Markers

In addition to qualitative aspects (e.g., mass and retention time) that signify biomarker candidates, proteins that differ in intensity are also informative and may be caused by up or down regulation or possibly genetic engineering (to produce more toxin etc.). The utility of quantitative biomarker candidates was evident during the analysis of the two non-O157:H7 EHECs. In the other eight samples, the intensity of a protein at 15,406 Da was much greater than one at 15,423 Da, but the trend was reversed for the two non-O157:H7 EHECs. Interestingly, this difference involved two of the most abundant proteins in the lysate and, for this reason, the quantitative difference was immediately obvious meaning no spectral subtraction of common peaks was required.

Strain level quantitative differences were also observed. In *S. flexneri* 04-0497, a protein at mass 9737 eluting at 26.4 min and highly abundant in all nine other isolates was low in intensity. Additionally, in *E. coli* 06-0006, a protein that elutes at 30.6 min weighing 35,171 Da, common to other *E. coli* and *S. sonnei* isolates, is completely absent, most likely underexpressed below the limit of detection. Alternatively, the gene for this protein could be damaged or turned off or possibly absent so that no protein is being expressed at all, making this a qualitative distinction.

Analysis of Unknowns

To challenge the validity of the biomarker candidates discussed above, a blind study of 13 isolates distinct from the original 10 was performed. In an attempt to identify each unknown, the mass spectra obtained from each of the 13 isolates were individually screened for the biomarker candidates listed in Table 2. Upon inspection of the blind study data, one initial observation was shifting retention times for the markers. During the early investigation of known isolates, a retention time window of ± 0.5 min was observed. However, during the blind study, analyte retention times seemed more variable, indicating an average window of ± 1.0 min was more suitable. Possible explanations for this variation include degradation of the column or minor differences in the mobile phase composition. However, this variation was consistent within each run, thus not affecting the relative retention times of the analytes. When used in conjunction, the retention time, mass, and relative intensity (RI) information (for quantitative markers) allowed the biomarker candidates to be detected with confidence.

Another observation was made concerning two of the three biomarker candidates for *S. flexneri*, one at 7287 and one at 35,250 Da. These proteins have counterparts in *E. coli* and *S. sonnei* exhibiting the same retention times but at decreased masses of 7273 and 35,170 Da. During the blind study, these two *S. flexneri* proteins were observed in small amounts (2%–5% RI, relative to the base peak) in some of the *E. coli* O157:H7 and *S. sonnei* isolates. There were a few possible reasons for this. (1) If the genes encoding these two proteins reside on extrachromosomal elements, they may have been horizontally transferred. (2) There may have been a small amount of *S. flexneri* present in these samples and they were therefore technically a mixture. (3) Since the mass spectrometer displayed higher total ion counts during the blind study than in any of the five previous replicates of the known isolates, these low abundance ions were now within the limit of detection. However, even with the greater ion counts, the *E. coli* and *S. sonnei* counterparts were not observed in any *S. flexneri* isolates, and the 16,886 Da *S. flexneri* marker, which was the least intense of the three *S. flexneri* markers (Figure 3), was not observed in any of the non-*S. flexneri* isolates. Since the exact reason(s) was not determined, these two *S. flexneri* proteins were, at least for the unknown isolates examined here, best used as quantitative biomarker candidates rather than qualitative.

Using the two *S. flexneri* proteins as quantitative markers, all biomarker candidates were present *and* absent as expected allowing all 13 unknown isolates to be correctly identified. In total, there were three *S. sonnei*, three *S. flexneri*, four *E. coli* O157:H7, one *E. coli* O26:H11, and two *E. coli* O111:NM isolates identified. Since the reproducibility of the biomarker candidates had already been established, only a single analysis was required to correctly identify the unknowns. The time required to collect, process, and examine the data to determine the identity of the unknown isolates was approximately 2 h per sample postculture.

Currently, identification of unknowns by this method is achieved through association with previously examined (known) pathogens, and is therefore limited to the types of pathogens listed in Table 1. For each type of bacterium that has been studied, an unknown of that same type can be identified in 2 h postculture. If a previously unstudied bacterium were analyzed, e.g., *S. boydii*, it would likely be recognized as the *Escherichia/Shigella* genus. Based on absence of marker masses, previously studied species, serotypes, and strains of the *Escherichia/Shigella* genus would be eliminated from consideration in 2 h postculture, providing a general classification of the bacterium.

One potential utility for this method would involve uploading LC/MS results from various public health

laboratories into a public database such as PulseNet, which utilizes PFGE data to detect outbreaks around the country [18]. Before LC/MS data could be used in this fashion, the issue of inter- and intralaboratory variability must be addressed. The first step to address both types of variability would be the formation of a standard operating procedure (SOP) document. The SOP would provide detailed methods and instructions beginning with cell growth and ending with data analysis, and would be followed as closely as possible by partner laboratories. The same LC/MS instrumentation would not be required but the same LC conditions (column, gradient, etc.) and a TOF mass analyzer should be used.

Besides implementing an SOP across laboratories, another way to address inter- and intralaboratory variability would be to add a standard protein to the lysate before analysis. Being a standard, this protein's mass and retention time would be well characterized before use. To address retention time shifts, the biomarker retention times would not only be reported as ± 0.5 min, but also as from a system where on average myoglobin eluted at 32.4 min. For example, if another laboratory uses a system with more dead volume than the one described here and myoglobin eluted at 32.9 min, 0.5 min could be subtracted from all of the results in that laboratory before searching for the biomarkers observed in this laboratory. This standard could also be utilized as an internal calibrant to correct for mass shifts. Finally, since the same concentration would be added to all lysates, the intensities of the proteins in the lysates could be held relative to this standard, and these relative intensities would be used to better ascertain quantitative differences between lysates. This is in contrast to using absolute intensities, which may vary based on chromatographic peak shape and source cleanliness etc.

Conclusions

LC/MS characterization of *Shigella* and *Escherichia* demonstrated greater specificity than obtainable using current real-time PCR protocols, allowed for distinctions at the strain level, and was less labor-intensive compared with PFGE, the gold standard for subtyping bacteria. Reproducible intact protein biomarker candidates were observed and successfully implemented for the identification of unknown pathogens without the use of primers, antibodies, or proteomic database searches. This was of particular interest for *Shigella* speciation for which PCR primers are commercially unavailable. These protein biomarker candidates could be sequenced and used to reverse-engineer novel PCR primers [19–21]. Likewise, these biomarker candidates could be purified for the production of antibodies to enhance serological investigations (e.g., protein microarrays) [22, 23].

Future work will include the expansion of the database of bacteria that have been studied. Based on the proof of concept work described above this should be quite straightforward. Statistical analysis such as principle components or hierarchical cluster analysis of LC/MS data may be investigated to asses their utility for establishing similarity/strain relatedness between known and unknown bacteria. Since submission, it was determined that three repeat experiments provided sufficient reproducibility for determining biomarker candidates rather than five as reported here. Additionally, a threefold increase in throughput was achieved using ultra-performance liquid chromatography [24]. This work would add another dimension to an integrated approach for more comprehensive bacterial identification.

Acknowledgments

The authors acknowledge funding of a portion of this research by the Centers for Disease Control and Prevention (CDC) grant no. U90/CCU317014. T.M.M. was supported by an appointment to the Emerging Infectious Disease (EID) Fellowship Program administered by the Association of Public Health Laboratories (APHL) and funded by the CDC. The authors thank Michael Martin and Michelle Sheldon for their review of the manuscript.

References

1. Khardori, N. Potential Agents of Bioterrorism: Historical Prospective and an Overview. In *Bioterrorism Preparedness* Chap. I, Khardori, N., Ed.; Wiley-VCH Verlag GmbH and Co. KGaA: Weinheim, 2006.
2. Carus, S. W. *Bioterrorism and Biocrimes: The Illicit Use of Biological Agents in the 20th Century*, Center for Counterproliferation Research, National Defense University: Washington DC, 2002.
3. Monday, S. R.; Minnich, S. A.; Feng, P. C. A 12-Base-Pair Deletion in the Flagellar Master Control Gene flhC Causes Nonmotility of the Pathogenic German Sorbitol-fermenting *Escherichia coli* O157:H Strains. *J. Bacteriol.* **2004,** *186,* 2319–2327.
4. Tominaga, A.; Mahmoud, M. A.; Mukaihara, T.; Enomoto, M. Molecular Characterization of Intact, but Cryptic, Flagellin Genes in the Genus *Shigella*. *Mol. Microbiol.* **1994,** *12,* 277–285.
5. Monday, S. R.; Whittam T. S.; Feng, P. C. Genetic and Evolutionary Analysis of Mutations in the gusA Gene that Cause the Absence of β-Glucuronidase Activity in *Escherichia coli* O157:H7. *J. Infect. Dis.* **2001,** *184,* 918–921.
6. Sutherland, J. B.; Rafii, F. Cultural, Serological and Genetic Methods for Identification of Bacteria. In *Identification of Microorganisms by Mass Spectrometry*, Chap. I, Wilkins, C. L.; Lay, J. O., Eds.; John Wiley and Sons, Inc.: Hoboken, 2006.
7. Houpikian, P.; Raoult, D. Traditional and Molecular Techniques for the Study of Emerging Bacterial Diseases: One Laboratory's Perspective. *Emerg. Infect. Dis.* **2002,** *8,* 122–131.
8. Fenselau, C. Mass Spectrometry for Characterization of Microorganisms: An Overview. In *Mass Spectrometry for the Characterization of Microorganisms*, Chap. I, ACS Symposium Series 541, Fenselau, C., Ed.; American Chemical Society: Washington DC, 1994.
9. Busse, H. J.; Denner, E. B. M.; Lubitz, W. Classification and Identification of Bacteria: Current Approaches to an Old Problem. Overview of Methods Used in Bacterial Systematics. *J. Biotechnol.* **1996,** *47,* 3–38.
10. Fox, A. Mass Spectrometry: Identification and Biodetection, Lessons Learned, and Future Developments. In *Identification of Microorganisms by Mass Spectrometry*, Chap. II, Wilkins, C. L.; Lay, J. O., Eds.; John Wiley and Sons, Inc.: Hoboken, 2006.
11. Fenselau, C.; Demirev, P. A. Characterization of Intact Microorganisms by MALDI Mass Spectrometry. *Mass Spectrom. Rev.* **2001,** *20,* 157–171.
12. Lay, J. O. MALDI-TOF Mass Spectrometry of Bacteria. *Mass Spectrom. Rev.* **2001,** *20,* 172–194.
13. Williams, T. L.; Leopold, P.; Musser, S. Automated Post-Processing of Electrospray LC/MS Data for Profiling Protein Expression in Bacteria. *Anal. Chem.* **2002,** *74,* 5807–5813.
14. Williams, T. L.; Andrzejewski, D.; Lay, J. O.; Musser, S. M. Experimental Factors Affecting the Quality and Reproducibility of MALDI TOF Mass Spectra Obtained from Whole Bacteria. *J. Am. Soc. Mass Spectrom.* **2003,** *14,* 342–351.

15. Nataro J. P.; Kaper, J. B. Diarrheagenic *Escherichia coli. Clin. Microbiol. Rev.* **1998,** *11*(1), 142–201.
16. Schwartz, D. C.; Saffran, W.; Welsh, J.; Haas, R.; Goldenberg, M.; Cantor, C. R. New Techniques for Purifying Large DNAs and Studying Their Properties and Packaging. Cold Spring Harbor Symposium. *Quant. Biol.* **1983,** *47*, 189–195.
17. Simpson, C. F. Introduction to Chromatography in Biotechnology. In *High Performance Liquid Chromatography: Principles and Methods in Biotechnology*, Chap. I, Katz, E. D., Ed.; John Wiley and Sons: West Sussex, 1996.
18. Swaminathan, B.; Barett, T. J.; Hunter, S. B.; Tauxe R. V. PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States. *Emerg. Infect. Dis.* **2001,** *7*(3), 382–389.
19. Williams, T. L.; Musser, S. M.; Nordstrom, J. L.; DePaola, A.; Monday, S. R. Identification of a Protein Biomarker Unique to the Pandemic O3:K6 Clone of *Vibrio parahaemolyticus*. *J. Clin. Microbiol.* **2004,** *42,* 1657–1665.
20. Williams, T. L.; Monday, S. R.; Feng, P. C. H.; Musser, S. M. Identifying New PCR Targets for Pathogenic Bacteria Using Top-Down LC/MS Protein Discovery. *J. Biomol. Tech.* **2005,** *16,* 134–142.
21. Williams, T. L.; Monday, S. R.; Edelson-Mammel, S.; Buchanan, R.; Musser, S. M. A Top-Down Proteomics Approach for Differentiating Thermal Resistant Strains of *Enterobacter sakazakii*. *Proteomics* **2005,** *5,* 4161–4169.
22. Mezzasoma, L.; Bacarese-Hamilton, T.; Di Cristina, M; Rossi, R.; Bistoni, F.; Crisanti, A. Antigen Microarrays for Serodiagnosis of Infectious Disease. *Clin. Chem.* **2002,** *48*(1), 121–130.
23. Steller, S.; Angenendt, P.; Cahill, D. J.; Heuberger, S.; Lehrach, H.; Kreutzberger, J. Bacterial Protein Microarrays for Identification of New Potential Diagnostic Markers for *Neisseria meningitides* infections. *Proteomics* **2005,** *5,* 2048–2055.
24. Everley, R. A.; Croley, T. R. Ultra-Performance Liquid Chromatography/Mass Spectrometry of Intact Proteins. *J. Chromatogr. A* **2008,** *1192,* 239–247.