
Gregory C. Conway¹,²,⁷, Sandra C. Smole³,⁵,⁷, David A. Sarracino¹,⁶, Robert D. Arbeit³,⁴,⁵, Peter E. Leopold¹∗

¹ProteiGene Inc., 44 Manning Road, Billerica, MA 01821, USA, ²Department of Cellular and Molecular Biology, Harvard University, Cambridge, MA USA, ³Department of Medicine and ⁴Microbiology, Boston University School of Medicine, MA USA, ⁵Research Service, Boston Veterans Affairs Medical Center, Boston, MA USA, ⁶Department of Chemistry, Tufts University, Medford, MA USA
⁷These authors contributed equally to this paper.

Abstract

To evaluate matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-ToF MS) as a tool for rapid identification of common clinical bacterial isolates, we analyzed 25 carefully selected isolates of pathogenic Escherichia coli (E. coli) and additional Enterobacteriaceae members. Organisms were prepared according to clinical microbiological protocols and analyzed with minimal additional processing. Spectra were reproducible from preparation to preparation and comprised 40-100 peaks primarily representing intracellular proteins with masses up to 25 kDa. Spectra of 14 genetically diverse bacteremic isolates of E. coli were compared with isolates representing other genera within the Enterobacteriaceae family. Using a new spectrum comparison algorithm, E. coli isolates were closely related to each other and were readily distinguishable from other Enterobacteriaceae, including Salmonella and Shigella. Presently, the methodology permits the analysis of 40 unknown isolates per hour per instrument. These results suggest that MALDI-ToF MS offers a rapid and reliable approach for performing phyloproteomics i.e., identification of unknown bacterial isolates based on similarities within protein biomarker databases.

Introduction

The core tasks of a clinical microbiology laboratory are to isolate, identify, and determine the antimicrobial susceptibility of the microorganisms present in clinical specimens. Currently, the most prevalent methods for species identification rely on the assessment of various bacterial metabolic activities. These systems typically require the overnight growth of organisms in a set of special indicator media. Nucleotide analyses can be used to resolve many bacterial species, and DNA probes are commercially available for determining selected species; nevertheless, genotypic methods for species identification remain technically demanding and expensive.

Recent advances in physical chemistry have facilitated the analysis of complex mixtures of macromolecules. Matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-ToF MS) is particularly effective for accurately and precisely determining the molecular weight of proteins (Allmaier et al., 1995; Kaufmann, 1995). In practice, a sample is co-crystallized with a UV-absorbing organic acid matrix. The analyte-matrix preparation is irradiated with a UV laser, which ablates the matrix and volatilizes and ionizes the analyte. The ionized analyte is accelerated through an electric field, and its time-of-flight in a 1-2 meter-long vacuum tube is recorded. The mass of the analyte is proportional to the square of the flight time. Several studies have demonstrated that whole microorganisms can provide distinct and reproducible mass spectra (Cain et al., 1994; Claydon et al., 1996; Holland et al., 1996; Krishnamurthy and Ross, 1996; Welham et al., 1998; Demirev et al., 1999). The effects of sample preparation and instrument settings are crucial factors relating to spectral reproducibility (Wang et al., 1998). In order for MALDI-ToF MS to be useful in the clinical setting, methods will have to be sufficiently sensitive, reproducible, discriminatory, and robust.

In this study, we used MALDI-ToF MS to analyze isolates of Enterobacteriaceae, with particular emphasis on a well-characterized set of genotypically diverse Escherichia coli isolates. We first investigated and optimized technical parameters, including bacterial growth conditions, matrix preparations, instrument settings, and established a rapid and practical protocol by which organisms could be taken from overnight colonies grown on routine agar media and processed within 15 minutes to obtain reproducible spectra. The observed macromolecules represented primarily intracellular proteins. Each isolate generated a distinctive set of mass spectrometric biomarkers; based on cluster analysis of these spectra, a dendrogram was generated which was essentially indistinguishable from that previously obtained using independent, extensively validated methods (Maslow et al., 1995).

Results

Optimization of Sample Preparation and Instrument Parameters

Our goal was to develop a simple method for preparing samples that would yield reproducible spectra of distinctive
Further Reading

Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at caister.com

• **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof

• **Aspergillus and Penicillium in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2

• **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorfit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins

• **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr

• **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles

• **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate

• **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem

• **Microalgae: Current Research and Applications**
  Edited by: MN Tsaioglou (2016)
  www.caister.com/microalgae

• **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma

• **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol

• **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo

• **Shigella: Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella

• **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romani, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms

• **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alphah

• **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile

• **Flow Cytometry in Microbiology**: Technology and Applications
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow

• **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics

• **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015

• **Corynebacterium glutamicum: From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory

• **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines

• **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals

• **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  www.caister.com/bacteria-plant

• **Aeromonas**
  Edited by: J Graf (2015)
  www.caister.com/aeromonas

• **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics

• **Leishmania: Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2

• **Acanthamoeba: Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2

• **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2

• **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order
biomarkers. Initially, bacteria were grown overnight in broth, washed with deionized water by centrifugation, and lyophilized to provide a stable bacterial preparation. Subsequently, aliquots were reconstituted in water as needed to a final concentration of 1 mg/ml, mixed with matrix, air-dried, and analyzed. This approach was sufficient for MALDI analysis and obviated the need for more complex fractionation schemes. We investigated two widely used MALDI matrices, α-cyano-4-hydroxycinnamic acid (α-cyano) and sinapinic acid, each dissolved in either of two solvents: 30% acetonitrile/0.1% trifluoroacetic acid in water (acetonitrile/TFA) or isopropyl alcohol. For a variety of bacteria tested, the greatest number and size range of mass spectral peaks were obtained using sinapinic acid (10 mg/ml) dissolved in acetonitrile/TFA (Figure 1); consequently, all subsequent studies were performed using this combination.

Instrument parameters (e.g., extraction voltage, laser attenuation, delay time) and digitizer parameters (e.g., delay, time base, spectral size) were also varied to

![Figure 1. MALDI analysis of E. coli 947A-17 using two different matrices. Aliquots of organisms were mixed with α-cyano-4-hydroxycinnamic acid or sinapinic acid and analyzed. Sinapinic acid consistently produced more spectral peaks and signals of higher mass than α-cyano-4-hydroxycinnamic acid.](image)

![Figure 2. Detection of pap operon expression by MALDI analysis. E. coli C1a transfected with either the nonrecombinant (pACYC184) or the recombinant pap operon vector (pRHU-845) were analyzed. A spectral peak appears at 16556.9 Da for E. coli (pRHU-845); PapA, the major structural protein of the pap operon has a calculated mass of 16554 Da based on the nucleotide sequence.](image)
Figure 3. Sensitivity of MALDI for detection of bacterial biomarkers. Serial dilutions of *E. coli* Bos36 were analyzed to determine the sensitivity of MALDI mass spectrometry. To insure equivalent data acquisition for each sample, ubiquitin was added to each dilution as an internal standard and data was acquired until a standard ubiquitin signal was obtained for each sample. Distinct mass spectral peaks were obtained with as little as 250 µg dry weight of organisms per ml, which represents ~1.8 x 10^5 bacteria per 0.5 µl aliquot analyzed.

Figure 4. Evaluation of nature and origin of MALDI biomarkers. Aliquots of *E. coli* Bos36 were analyzed after being treated with (A) buffer alone; (B) proteinase K; (C) lysozyme; and (D) lysozyme followed by proteinase K. (E) represents proteinase K without bacteria, and (F) represents lysozyme without bacteria. The number of mass peaks detected was not increased by lysozyme treatment (C vs. A), indicating that organisms were effectively lysed by the MALDI preparation procedure alone; a few low mass peaks were absent following lysozyme treatment, suggesting that they represented cell wall constituents that were degraded by the enzyme. Aliquots were also analyzed with or without prior digestion with proteinase K (B vs. A). In this experiment with Bos36, protease digestion had little effect; however, other *E. coli* strains showed a substantial decrease in the intensity of the mass peaks detected. Aliquots of strain Bos36 were also analyzed after digestion first with lysozyme (to effectively lyse all organisms and expose intracellular constituents) and then with protease (to digest both intracellular and extracellular proteins). This dual treatment eliminated detectable mass peaks (D). These results suggest that most of the bacterial biomarkers detected by MALDI are intracellular proteins.
The possibility for the absence of these larger species is that in the complex bacterial milieu, ionization of larger molecules may not occur efficiently due to either insolubility or suppression (Krishnamurthy et al., 1996). Nevertheless, despite this limitation, a sufficient number of unique spectral peaks were detected to allow discrimination of organisms as detailed below.

Detection of a Single Protein Difference Between Two *E. coli*

To examine the expression of a unique protein against a particular genetic background, *E. coli* C1a containing either, pRHU-845, a recombinant plasmid carrying the *pap* operon, or pACYC184, the parent vector alone (Normark et al., 1983), were compared. The spectra for the *E. coli* (pRHU-845/pap+) and *E. coli* (pACYC184/pap-) differ by a single peak with a mass of 16,556.9 Da (Figure 2). This mass corresponds closely to the mass of *papA* (16554 Da) as calculated from the primary sequence; *papA* is the major structural subunit pilin protein and is present in multiple copies for each assembled pilus (Figure 2). Thus, whole cell MALDI analysis was sufficiently sensitive to detect the expression of a single, unique protein when the mass was within the sensitivity range of the method (2-25 kDa).

Sensitivity of MALDI for Detection of Bacterial Biomarkers

To quantitate the sensitivity of MALDI-ToF MS analysis, serial dilutions of bacteria ranging from 100 to 1,000 µg/ml were prepared. A known quantity of ubiquitin, a protein mass standard, was mixed with each sample prior to analysis and spectra were collected until a standard ubiquitin signal intensity was achieved for each sample. This permitted quantitative comparison of the spectra. MALDI-ToF MS yielded detectable spectra at concentrations of bacteria as low as 250 µg/ml (Figure 3). The 0.5 µl sample aliquot transferred to the plate (sample target) contained 125 ng of bacteria, or approximately 1.8 x 10^5 organisms (Bremer and Dennis, 1987) over the entire surface of the each spot (~1.76 mm^2). Based on the laser spot area (~0.015 mm^2), we estimate that 1000-5000 organisms are ablated during the acquisition of each spectrum.

Characterization of MALDI Biomarkers

Welham *et al.* (Welham *et al.*, 1998) used electron microscopy to show that bacteria remain intact in matrix microcrystals. Consequently, the observed mass peaks might represent only outer membrane or cell wall-associated biomarkers. To determine if the MALDI process ionized intracellular proteins, samples were analyzed with and without prior treatment with lysozyme (Figure 4C and 4A) which digests the cell wall of gram-negative bacteria. We hypothesized that if only extracellular proteins were ionized directly, then lysozyme treatment would liberate cytoplasmic constituents and increase the number of mass spectral peaks. However, we observed that prior lysozyme treatment had little effect on the spectra, and actually resulted in the disappearance of some lower molecular weight species (Figure 4C versus 4A), possibly due to ion suppression effects of the added lysozyme. These results suggest that the sample preparation procedure and laser
irradiation alone effectively lyses gram-negative organisms and allows the detection of intracellular constituents.

To determine if the biomarkers detected represented proteins outside the cell wall, bacteria were analyzed with and without prior proteinase K treatment. Proteinase K is a broadly active serine protease which would be expected to degrade exposed proteins, thereby eliminating their mass peaks. We observed that pretreatment with protease alone had little effect on the observed spectra (Figure 4A versus 4B).

Finally, a combination of lysozyme treatment followed by proteinase K digestion was used to determine if the MALDI biomarkers detected represented intracellular proteins (Figure 4D). This procedure would result first in lysis of the cell wall, followed by protease digestion of the released intracellular constituents. For those microorganisms for which proteinase K treatment alone had little effect on the observed spectra, prior lysozyme treatment followed by proteolysis eliminated all biomarkers (Figure 4D). Taken together, these results indicate that most of the biomarkers detected are proteins, probably of intracellular or cytoplasmic origin.

Figure 6. MALDI analysis of 14 genotypically diverse E. coli isolates. (A) At low resolution, visual analysis suggests that some peaks are common to all E. coli examined and some peaks are unique to individual isolates. (B) Enlarged spectra for E. coli strains 949A-70 and 949A-102. Mass peaks common to the spectra of both strains are marked (+); peaks unique to the individual strains are marked (*). The peaks marked are typical of those used to construct the reference database for quantitative comparison of spectra from different strains.
Effect of Growth Conditions on Bacterial Spectra
It is well known that protein expression can be influenced by growth media and growth phase (Arnold et al., 1999). To determine how growth media and incubation times affect mass spectra, we analyzed organisms grown in two different media and organisms sampled over time in culture. *E. coli* strain Bos36 grown overnight in either trypticase soy (TS) or Luria-Bertani (LB) broth yielded spectra with a similarity of ~0.80 (Figure 5A). Spectra for organisms grown for 1.5, 3, 6, and 16 hrs were >0.65 similar (Figure 5B). The relatively high similarities indicate that (a) MALDI-ToF MS is sensitive enough to detect variations in biomarkers associated with growth phase and choice of media; (b) spectral variations associated with different media are comparable to the variability associated with different strains of species; and (c) spectral variations related to growth phase can exceed the variations typically seen within a species.

Finally, we compared spectra obtained using organisms grown in LB broth with those obtained using organisms taken from a small colony grown on LB agar media. The mass peaks obtained from bacteria grown on solid media differed from those of the same strain grown in liquid culture (Figure 5C); nevertheless, organisms from both sources yielded complex but still reproducible spectra.

Figure 7. MALDI analysis of clinical isolates representing nine additional genera of *Enterobacteriaceae*. (A) Although some peaks were common to most organisms, a large number of genus-specific markers were detected. (B) Enlarged spectra of isolates of *Proteus mirabilis* and *Salmonella typhi*. The symbols (+ and *) indicate peaks that are common and unique, respectively, for these two isolates.
Different Isolates and Species Yield Unique Biomarkers

To determine the variation of mass spectra among strains within an individual bacterial species we analyzed 14 *E. coli* isolates (Figure 6A, B); the strains were carefully selected on the basis of previous studies to represent a wide range of genetic diversity (Arthur et al., 1990; Maslow et al., 1995). All strains were grown overnight on LB agar, processed using 10 mg/ml sinapinic acid in acetonitrile/TFA, and analyzed using optimized instrument parameters. Each strain yielded complex spectra comprising ~50 mass peaks. While there were differences among the spectra, many mass peaks detected were common to all 14 *E. coli* strains, suggesting that species specific biomarkers are present for development of species identification.

To assess the ability of the MALDI-ToF MS spectra to differentiate *E. coli* from other bacteria, we analyzed isolates representing nine additional genera within the *Enterobacteriaceae* family and, as a true outlier, an isolate of *S. capitis* representing a gram-positive species from a different bacterial family (*Micrococcaceae*). As with the *E. coli* isolates, approximately 50 mass peaks were detected for each organism (Figure 7A and B) with the exception of *S. capitis* which generated ~20 peaks.

Cluster Analysis of MALDI Biomarkers Yields a Phylogenetic Dendrogram Comparable to that Determined by Independent Validated Methods

To assess quantitatively whether MALDI-ToF MS could distinguish isolates of different species, we performed a single linkage cluster analysis of all the spectra and derived a dendrogram to represent the results graphically. The numerical similarity between two isolates or two species is the similarity value of the point at which the two branches diverge. The dendrogram demonstrates two critical results (Figure 8). First, essentially all of the *E. coli* isolates are tightly clustered, with similarity typically ~0.8. Second, all the isolates representing other *Enterobacteriaceae* genera branched deeply from the *E. coli* cluster with similarities
~0.65. Of note, the spectra of the *Shigella* and *Salmonella* isolates were most closely similar to the *E. coli* cluster. The isolate of *S. capitis* was a true outlier, with similarity ~0.3 compared to all of the gram-negative isolates.

Repeated MALDI-ToF MS analyses of a single culture (technical reproducibility) (Arbeit, 1999) and of multiple, independent cultures of the same isolate (biologic reproducibility) demonstrated spectral similarity of ~0.90 (data not shown). The product of these two effects (technical and biologic reproducibility) is ~0.80; this is comparable to the degree of similarity among the diverse *E. coli* strains examined (Maslow *et al*., 1995). Consequently, we conclude that at this time strain-specific identification is not yet possible using the current protocols and data analysis. However, species identification is possible.

MALDI-ToF MS indicated that the isolate designated “*E. coli* Afr102” was highly divergent from the *E. coli* cluster (Figure 8). In an earlier study, multilocus enzyme electrophoresis (MLEE) assigned this isolate as the sole representative of a deeply branching lineage; by genotypic analyses, it had a unique ribotype and lacked all of the common *E. coli* virulence genes (Maslow *et al*., 1995). On reevaluation, Afr102 could not be confirmed as an *E. coli*. Biochemical testing performed with the API 20E kit (bioMerieux Vitek, Hazelwood, MO) provided low selective species identification; when additional biochemical tests were included, the isolate was designated as an *E. coli* with a 90% probability. The automated VITEK (bioMerieux Vitek) identified the isolate as an *Enterobacter intermedius* (probability 59%). Analysis of *E. coli* specific housekeeping genes using PCR also suggests that Afr102 may not be an *E. coli* (Smole, unpublished observations). In agreement with other bacterial identification methods, MALDI-ToF MS indicates that Afr102 is probably not an *E. coli*.

**Discussion**

Species identification of bacterial isolates represents a major activity of a clinical microbiology laboratory. Biotyping, the most commonly used method, requires an overnight subculture to assess the metabolic activities of the isolates. Although DNA-based approaches are more rapid, commercial systems have high unit costs and are currently available only for selected species. Previous reports have indicated that whole-cell analysis of bacteria by MALDI-ToF MS is feasible and can provide reproducible spectra (Claydon *et al*., 1996; Holland *et al*., 1996; Krishnamurthy *et al*., 1996; Welham *et al*., 1998; Demirev *et al*., 1999). These studies typically analyzed laboratory isolates and focused on technique. Although, quantitative cluster analysis of several lab strains of *E. coli* was attempted by other investigators (Easterling *et al*., 1998), diagnostic utility was not evaluated.

In this study we applied MALDI-ToF MS to an extensive and clinically-relevant group of pathogenic *E. coli* as well as an additional group of clinically significant genera. Although these other genera are significant in their own right, *E. coli* are commonly isolated bacterial pathogens. We and others have used multiple techniques, including multilocus enzyme electrophoresis (MLEE), analysis of ribosomal operon polymorphisms (ribotyping) and antigenic variation (serotyping), to define the phylogenetic and genotypic characteristics of large collections of isolates causing urinary tract infection and sepsis (Arthur *et al*., 1990; Maslow *et al*., 1995). The 14 *E. coli* strains used in this study were carefully selected to represent widely divergent genotypes associated with clinical infection (Maslow *et al*., 1995).

For each isolate, MALDI-ToF MS analysis yielded a spectrum representing a large set of bacterial products. We used selective enzymatic digestion to demonstrate that the biomarkers detected are primarily proteins from intracellular origin but may include some cell wall or membrane associated proteins. Growth medium and times had no qualitative effect on the spectra obtained, indicating that the method was robust and reproducible for our observed conditions. As anticipated, spectral differences were greater for the growth phase of a specific organism than our choice of growth medium. Although we have compared only LB and TS broth in this study, we have subsequently compared LB agar versus trypticase soy with 5% sheep blood agar and found an additional three peaks in the spectra of organisms grown on the blood agar (Smole, unpublished observations). Markers larger than greater than 25 kDa were observed infrequently, probably due to low ionization yields or poor solubility for higher mass proteins. Nevertheless, the number, quality, and variation of the markers detected was sufficient to reproducibly distinguish bacterial species and demonstrate the potential diagnostic value of this approach.

The complex spectra obtained by whole-cell MALDI-ToF were compared by single-linkage cluster analysis of spectral peak lists. Single-linkage cluster analysis is conceptually and operationally simpler than principal component analysis. We are currently evaluating this alternative approach to determine whether it provides additional discriminatory power. We used peak lists rather than whole spectra in order to reduce data handling complexity from 50,000 amplitudes to a discrete set of numbers corresponding to detectable biomarkers.

The resulting data set, although simplified, contained sufficient information to generate a phylogenetic dendrogram that was topologically congruent with that previously obtained by MLEE and also supported by ribotype and serotype data (Maslow *et al*., 1995). Ribotyping, MLEE, and serotyping detect variations at the level of nucleotide sequence, amino acid sequence, and polysaccharide structure, respectively. As noted above, we have shown that the majority of biomarkers detected by MALDI-ToF MS are proteins.

A striking aspect of our dendrogram is the deep-rooting of *E. coli* compared with other *Enterobacteriaceae*. *Shigella*, considered to be a close evolutionary relative of *E. coli*, are often difficult to distinguish from *E. coli* even by DNA-based methods. MALDI-ToF MS resolved Shigella as highly similar to *E. coli*, but nevertheless distinctly different. The correct and unanticipated determination that isolate Afr102 was not a bona fide *E. coli* also demonstrated the discriminatory power of the MALDI-ToF MS spectra and the cluster analysis. We respectfully suggest that the term “phyloproteomics” be applied to the development and application of this technology.

The organisms we analyzed had been passaged in
vitro and spectra suitable for analysis were obtained with as few as 1.8 x 10^5 organisms. Potentially, MALDI-ToF MS might be suitable for detection and identification of microorganisms directly from clinical samples, especially those that are normally sterile and are not biochemically complex, such as urine and cerebrospinal fluid. The addition of simple concentration and fractionation methods might permit analysis of organisms from blood and other complex materials.

Implementation of MALDI-TOF MS analysis in a clinical microbiology laboratory should be straightforward. Sample preparation is simple and requires little operator time. A mass spectrometer specifically tailored for bacterial identification could be more compact and economical than current general purpose instruments. Similarly, task-specific software would enable automatic data acquisition, peak picking, and comparison to a database of bacterial mass spectra. With these improvements mass spectrometry could become a rapid, high-throughput, cost-effective method for bacterial identification in clinical microbiology laboratories with suitable volume or microbiology reference laboratories. Efforts are currently in progress to tailor this method specifically for problematic organisms for which current methods take days to weeks to provide definitive species identification.

Experimental Procedures

Bacterial Isolates and Growth Conditions

The previously characterized E. coli isolates (n=14) used in this study were originally obtained from bacteremic patients (Arthur et al., 1990; Maslow et al., 1995). Isolates (n=9) of other genera in the Enterobacteriaceae family (Citrobacter freundii, Enterobacter cloacae, Klebsiella oxytoca, Morganella morgani, Proteus mirabilis, Providencia stuartii, Salmonella typhi, Serratia marcescens, and Shigella sonnei) were obtained from the reference stocks of the Pathology and Laboratory Medicine Service at the Boston Veterans Affairs (VA) Medical Center (courtesy of Dr. Stephen M. Brecher). The sole gram positive organism, Staphylococcus capitis (ATCC# 56961), was obtained from the ATCC (Manassas, VA). E. coli strain C1a (Sasaki and Bertani, 1965) derivatives carrying either pRHU-845, which expresses the pap (pyelonephritis-associated pilus) operon (Normark et al., 1983), or pACYC184, the parental cloning vector, have been described elsewhere (Arthur et al., 1990).

Bacterial broth cultures were inoculated from plates containing single colonies grown on Luria-Bertani (LB) agar (Difco Laboratories, Chicago, IL). The bacteria were cultured uniformly for 14 hours and were harvested by centrifugation, washed thrice in water, lyophilized, and stored at -20°C. Stock cultures of a bacteria culture by MALDI-MS of whole cells. Anal. Chem. 71: 1990-1995. The authors acknowledge Geoffrey J. Barton (Laboratory of Molecular Microbiology, Oxford University, Oxford, UK) for the software package, “OC-a cluster analysis program” in relation to single linkage cluster analysis and its graphical representation.

References

This work was supported by a SBIR/NIH phase I (grant R43-08857-01). The authors acknowledge Geoffrey J. Barton (Laboratory of Molecular Microbiology, Oxford University, Oxford, UK) for the software package, “OC-a cluster analysis program” in relation to single linkage cluster analysis and its graphical representation.

Acknowledgements

Phyloproteomics: Species Identification Using MALDI ToF MS 111

Spectra, Peak List and Cluster Analysis

Spectra were reduced to peak lists based on a peak picking algorithm in the Bruker Xmass software using apex peak detection with sensitivities of 2.5-4. The Bruker algorithm annotates local maxima according to a criterion of signal-to-noise cutoff of four times the value of sensitivity, where the noise value is calculated globally based on amplitudes at the low mass end of the spectrum, and the signal is determined locally from an implicit baseline subtraction routine. Peak lists were inspected by eye to ensure completeness and accuracy, and then analyzed using a similarity metric that provides a scalar distance between peak lists based on allowed error in exact measurement of individual peak masses. If the mass difference between peaks m_i in peak list i and m_j in peak list j is less than the value p, then the two mass peaks are assumed to be the same. The similarity measure s_{i,j}(p) between peak list i and j is defined as

\[ s_{i,j}(p) = \frac{c_{i,j}(p)}{\sqrt{n_i n_j}} \]

where \( n_i \) and \( n_j \) are the total number of peaks identified in the individual peak lists, and \( c_{i,j}(p) = \text{the number of peaks in common between peak list } i \text{ and } j \). The value of \( s_{i,j}(p) \) is limited to the domain \( 0 < s_{i,j}(p) \leq 1 \). When the peak list i and j are indistinguishable, \( s_{i,j}(p)=1 \); then the peak list are completely dissimilar, \( s_{i,j}(p)=0 \). The values of \( n_i \) and \( n_j \) are typically between 40 and 50; \( c_{i,j}(p) \), the number of peaks in common, ranges from only a few to nearly all.

The allowable mass difference \( p \) is a free parameter. When \( p \) is arbitrarily large, \( s_{i,j}(p) = 1 \) for all pairs of peak lists, and the similarity measure is unable to discriminate between peak lists. Likewise, when \( p \) is arbitrarily small, \( s_{i,j}(p) = 0 \) for all pairs of peak lists, and all peak lists are equally dissimilar. Empirically, clustering is consistent for \( p \) in the range of 10<p<50; the cluster diagram branches optimally, i.e., deeply, when \( p = 25 \). This corresponds to errors in reproducibility of ~25 Da, or the presence or absence of a Na"^+ (mass = 23) adduct.

The matrix of peak list-peak list similarities are represented as a dendrogram according to the single linkage cluster algorithm. Specifically, the two most similar peak lists are designated a cluster and a new (n-1) x (n-1) matrix of similarities is constructed by comparing each of the remaining n-2 peak lists with the first cluster. The similarity between a peak list and a cluster is defined as the largest similarity between the peak list and any peak list within that cluster. The process is repeated until the nth peak list has been compared with a cluster comprising n-1 peak lists.


