# Bioinformatics-Based Strategies for Rapid Microorganism Identification by Mass Spectrometry

*Plamen A. Demirev, Andrew B. Feldman, and Jeffrey S. Lin*

We review approaches for microorganism identification that exploit the wealth of information in constantly expanding proteome databases. Masses of an organism's protein biomarkers are experimentally determined and matched against sequence-derived masses of proteins, found together with their source organisms in proteome databases. The source organisms are ranked according to the matches, resulting in microorganism identification. Statistical analysis of proteome uniqueness across organisms in a database enables evaluation of the probability of false identifications based on protein mass assignments alone. Biomarkers likely to be observed can be identified based solely on microbial genome sequence information. Protein identification methodologies allow assignment of detected proteins to specific microorganisms and, by extension, allow identification of the microorganism from which those proteins originate.

## INTRODUCTION

Effective responses to bioterrorism attacks or novel emerging infectious diseases such as SARS require enhanced capabilities for rapid and accurate microorganism identification. For example, the intentional use of *Bacillus anthracis* spores in the fall of 2001 highlighted the importance of accurate bioagent surveillance and sensor technologies for the quick and reliable detection of both natural and bioengineered microorganisms.

Mass spectrometry (MS) is one emerging technology capable of meeting the challenges posed by biological threats. The current paradigm for rapid microorganism characterization by MS is based on the detection and identification of biomarkers using experimental mass spectra. This paradigm can be traced back to Anhalt and Fenselau,[1] who demonstrated that biomolecules from different pathogenic bacteria, introduced intact into a mass spectrometer, could be vaporized and ionized directly by electron impact. These chemical biomarker signatures for different organisms were structurally identified by MS. Furthermore, their signature composition and abundances allowed taxonomic distinctions among the microorganisms to be made.

After the introduction of soft ionization MS techniques[2–4] (recognized with the Nobel Prize in Chemistry in 2002), a number of new applications of MS in biology and medicine emerged. These techniques—Matrix-Assisted Laser Desorption/Ionization (MALDI) and electrospray ionization (ESI)—allowed for the first time the transfer of large (>30 kDa), intact, nonvolatile biomolecules, such as proteins, into the gas phase. In a MALDI

experiment, a low-mass photo-absorbing organic compound (matrix) is added to a sample prior to irradiation with ultraviolet (typically 337 nm) nanosecond laser pulses to desorb high-mass biomolecular ions. In ESI, large, multiply charged ions are generated by transporting the analyte solution through a capillary needle typically biased from 2 to 4 kV relative to ground. Several stages of differential pumping and suitable ion optics allow the interfacing of an ESI ion source operating at atmospheric pressure with a mass spectrometer operating in high-vacuum conditions.

Combining MALDI or ESI with MS instrumentation, the molecular masses of individual proteins larger than 100 kDa could be determined with unprecedented accuracy. Several laboratories reported applications of MALDI and ESI to studies of intact cells of microorganisms.[5–15] These ionization techniques, in particular MALDI (Fig. 1), have enabled intact protein biomarkers to be detected and exploited in microorganism characterization.[16,17]

In this article, we review MS approaches for microorganism identification, developed at APL and elsewhere, that exploit the wealth of information found in the ever-expanding genome and proteome (all proteins encoded in the genome) databases for prokaryotic organisms (bacteria and archaea) and viruses. Broadly, these approaches are based on experimentally determining the masses of the protein biomarkers detected from an unknown organism, which yields the protein's "signature" (Fig. 2). Next, *de novo* generation of the protein biomarkers is performed by computing their masses from the protein sequences for each organism in the proteome database. Microorganism identification is achieved by matching the experimental masses of the unknown against database masses and ranking the organisms according to the statistical significance of the number of matches in the mass spectrum.

The successful implementation of a bioinformatics-based approach for microorganism identification has several requirements. First, the proteome database should be complete, that is, it should contain the proteins for the unknown microorganism. Since each gene in the genome codes for a putative protein, the completeness requirement means that the genome of the organism should be sequenced. Second, statistical analysis of proteome uniqueness must be performed in terms of the mass accuracy of the MS instrument to estimate the probability of misidentifications and to assign a confidence measure to the final result. Third, improvements to proteome database fidelity should be incorporated, for example, to account for the most common "post-translational modifications" (PTMs) to the proteins coded in the genome that are not directly reflected in the databases.

We also review methods for improving the identification reliability of the bioinformatics approach by imposing constraints on the number of potential biomarkers to be matched, for example, by including only the highly expressed proteins in the protein biomarker database for each organism. *In silico* creation of protein biomarker

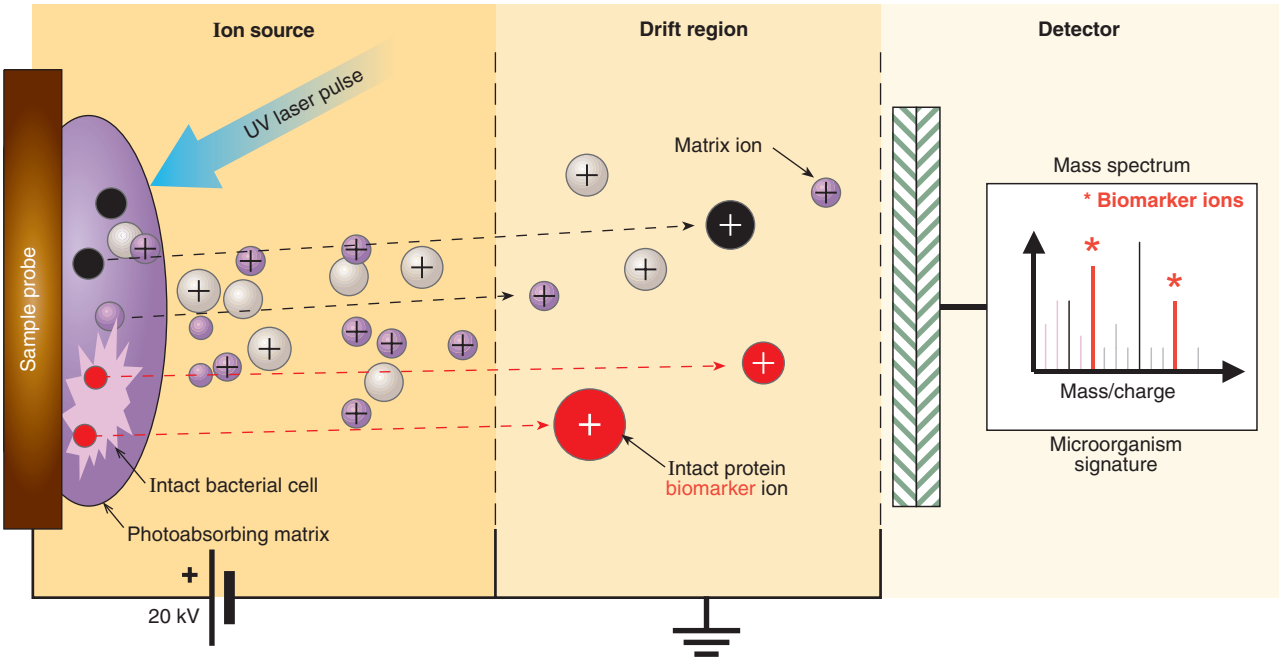



**Figure 1.** Principle of operation of a MALDI TOF mass spectrometer for microorganism identification. Intact biomarker ions are desorbed into the vacuum as a result of laser photon–matrix molecule interactions. All ions acquire kinetic energy proportional to their charge *z*. Ions with the same charge but different masses have different times of flight (TOF) through the drift region of the instrument. Calibration procedures correlate an ion's TOF, measured in an experiment, with its mass *m* (TOF $\propto \sqrt{m/z}$ ).

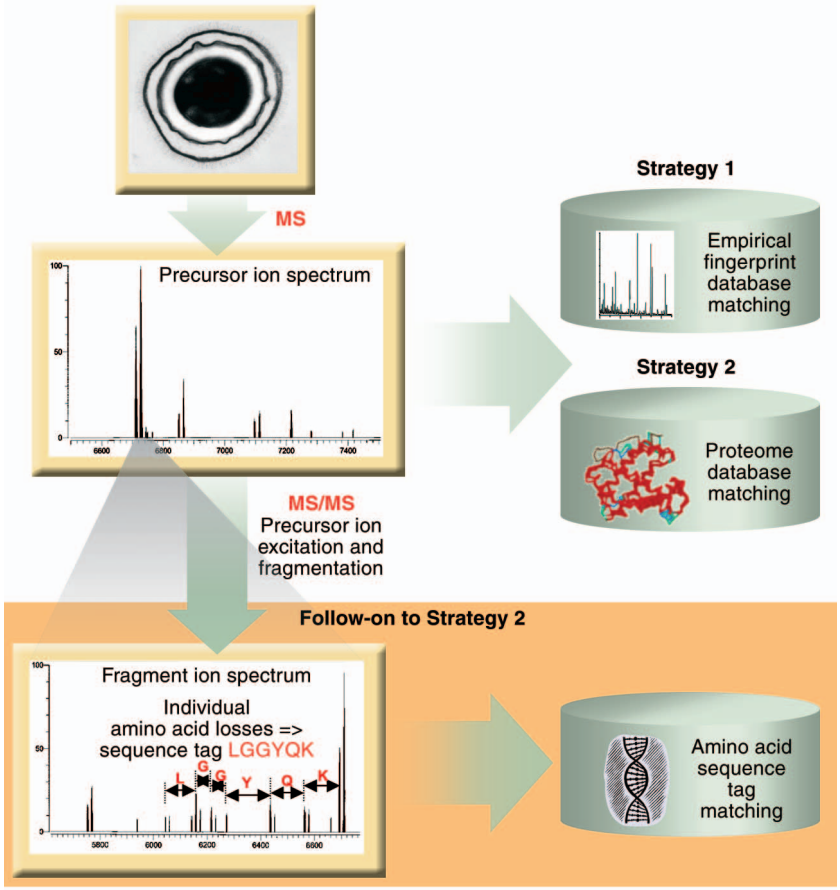



**Figure 2.** Two different strategies currently used for MS-based microorganism identification. In the first, a "fingerprint" approach, identification is achieved by matching an experimental mass spectrum to MS fingerprints from an empirically compiled mass spectral database of known organisms. In the second, a bioinformatics-based strategy, the experimental mass spectrum is matched to protein masses derived from a proteome database. MS fingerprint matching does not require the genome sequence of the organism to be known; however, the fingerprints of an organism depend on experimental conditions and various biological factors. Follow-on developments of the bioinformatics-based strategy include tandem MS (MS/MS). In an MS/MS experiment, a precursor biomarker ion in the experimental mass spectrum is isolated and excited via interactions with neutral gas molecules, electrons, or photons. The precursor ion dissociates, and the detected fragments can be correlated to the amino acid sequence of the precursor. A partial amino acid sequence (a "tag") may be sufficient to identify the precursor protein by sequence homology searches in a proteome database, and from there, enable identification of the organism from which the protein originates.

databases for microorganism identification by MS is discussed as well. In this case, each database entry is selected directly using microbial genome sequence information to deduce proteins that are highly expressed, thus flagging them as the most likely to be detected by MS. In addition, we illustrate the advantages of *in silico* generation of protein biomarker databases for particular pathogenic microorganisms, including *Bacillus* spores and the SARS virus.

## MS-BASED APPROACHES FOR MICROORGANISM IDENTIFICATION

In conventional MS-based microorganism identification approaches, experimental MALDI mass spectra from a microorganism are compared with a collection of mass spectra of known organisms—MS "fingerprints"—compiled into a reference biomarker signature library (Fig. 2). The MS-detected biomarkers can vary with sample preparation, instrumental conditions, microorganism biochemistry, and environmental conditions, such as diverse biological backgrounds.[16] For instance, depending on the organism's developmental stage, different sets of biomarker peaks are observed from the same organism.[18] Therefore, to perform effectively, this fingerprint approach requires collection of a vast number of spectra for each targeted microorganism under a variety of different conditions. Collecting fingerprints for all conditions, both for target pathogenic and nonpathogenic background organisms, is generally not feasible because of the enormous sample throughput bottleneck. In addition, this fingerprint library approach has the practical limitation that no signature data are available unless MS has been performed on the organism (this is not always feasible for novel or highly pathogenic organisms).

To provide signature robustness and to avoid the sample throughput bottlenecks associated with fingerprinting, a bioinformatics-based strategy (Fig. 3) for protein biomarker database generation was proposed.[18–20] Here, the genome, which codes for all proteins that can be expressed in an organism, provides a set of expected biomarkers that are robust against experimental variability (see the preceding paragraph). The expected biomarkers for a microorganism are determined from the masses of a subset of all of its potentially expressible proteins. This is the major difference between the "traditional" fingerprint and bioinformatics-based approaches. In both strategies, the experimental MS data are compared to expected masses (from reference spectra in the traditional approach; derived from the genome in the new approach), and the microorganism that provides the most statistically significant matches is selected.

While different sets of proteins can be expressed in a microorganism and experimentally observed by MS (depending on growth stage, growth medium, etc.), the masses of all these proteins can be independently derived from their sequences. The amino acid sequences

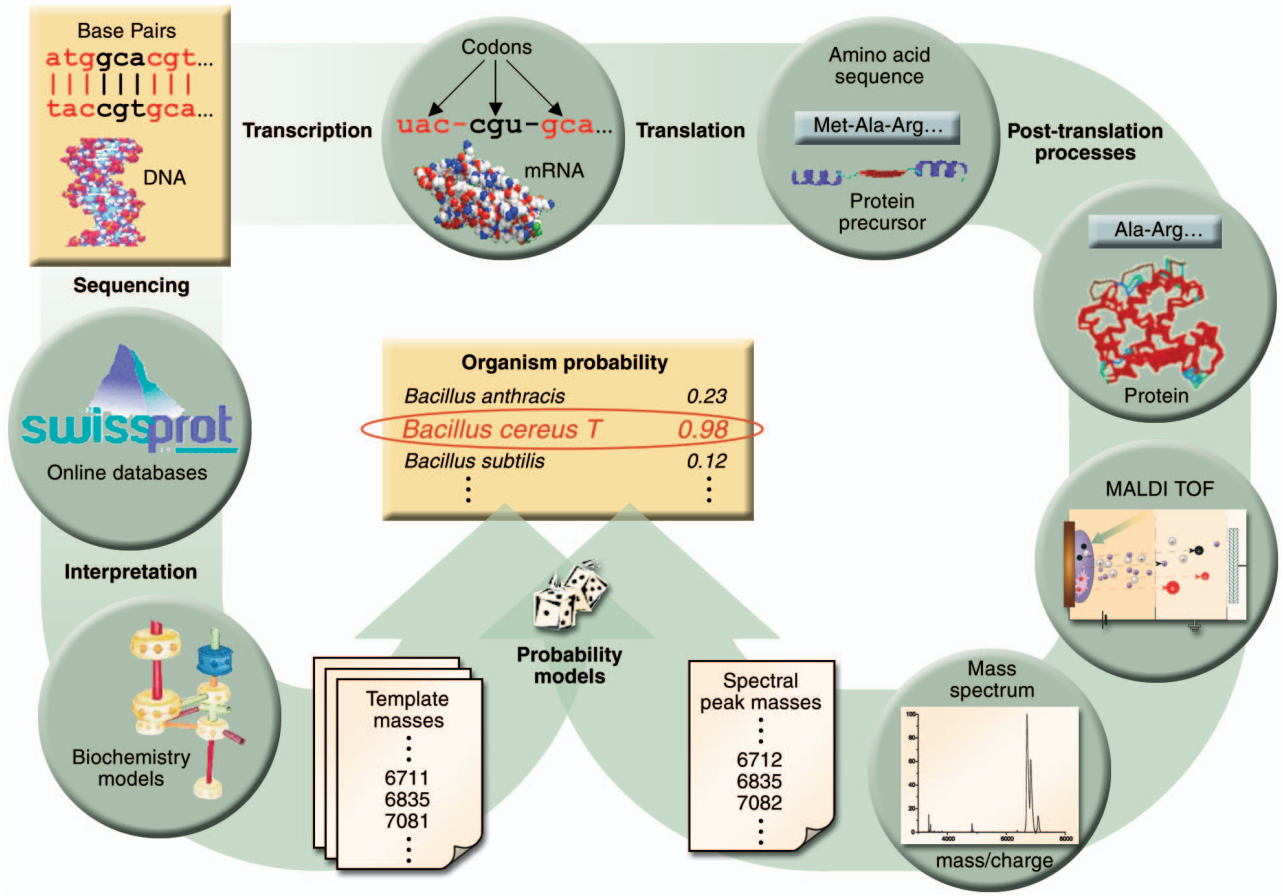



**Figure 3.** DNA, a linear biopolymer, contains four different nucleotides (bases) in a specified sequence. This sequence provides the blueprint for the "expression" of individual genes into proteins. Proteins, which are also linear biopolymers, are composed of 20 different amino acids. The genetic code maps nucleotide triplets ("codons") to each individual amino acid. Since there are 64 ($4^3$) different codons coding for 20 amino acids, a redundancy in codon usage exists (i.e., several "alternative synonymous codons" coding for the same amino acid). The sequences of all proteins for an organism can be predicted ("translated") by bioinformatics tools from that organism's genome sequence. Depending on their function, proteins are expressed (synthesized) in different "copy numbers" (from a single copy to $10^5$ copies for highly expressed proteins per cell). Furthermore, proteins may undergo *in vivo* "post-translational modification" (PTM), resulting in changes to the original amino acid sequence (translated from the genome). We determine by MS the masses of the "final" protein biomarker product (after the PTM).

of all proteins can be found (together with their source organisms) in Internet-accessible proteome databases (e.g., Refs. 21 and 22). The experimental biomarker signature (the masses of a set of ions) is determined by MS. Identification is achieved by comparing the masses from the experimental spectrum of the unknown organism with the masses of proteins in proteome databases and then ranking the candidate microorganisms according to the number of matches. The scoring method, initially proposed by Demirev et al.,[18] identified the highest ranked organism in the matching procedure as the source of the experimental mass spectrum. The method was successfully demonstrated with experimental mass spectra from Gram-positive as well as Gram-negative microorganisms with completely sequenced genomes.

Since constant intensities of the set of experimentally observed protein biomarkers are not required when using the bioinformatics-based approach, various types of MS instruments can be used for obtaining the biomarker spectra. Early on, it was also noted that several factors, such as mass range, mass measurement accuracy, and database size, influence successful microorganism identification.[18,19]

## REQUIREMENTS FOR BIOINFORMATICS-BASED APPROACHES

### Proteome Database Completeness

To identify a particular organism with this bioinformatics-based approach, it is obvious that the set of protein biomarkers for that organism must be available in the proteome database. Initially, the protein sequences found in protein sequence databases were obtained through isolation and sequencing of the proteins in each organism. In the last few years, as a result of advances in gene sequencing technologies, the proteome databases have rapidly expanded via translation of the

open reading frames of the respective organism's genome. By the end of 2003, the genomes of 134 microorganisms (bacteria and archaea) were publicly available.[23] The rate of microorganism sequencing as a function of time (Fig. 4) has increased since the first bacterial species was sequenced in 1995. The genomes of all bacteria on the CDC listing of potential bioterrorism agents have been or are currently being sequenced and are also publicly available. In some instances, these include multiple pathogenic and/or nonpathogenic strains of the same organism. Most notably, the sequences of *B. anthracis* and its closest "relative," *B. cereus*, have been published.[24,25] The speed with which the sequences of newly emerging threats to public health are made available is well illustrated by the example of the SARS virus, whose sequence was determined through an international collaboration weeks after the initial outbreak.[26]

### Proteome Statistics

Statistical analysis of the uniqueness of proteome-derived protein biomarkers as a function of experimental mass accuracy and proteome density (number of proteins per mass interval) has been performed.[19] The analysis provides a means to evaluate the rate of false identifications (organisms other than the target organism) attributable to randomly matching experimentally derived masses to database-derived masses of proteins from an organism. It also suggests that simple ranking by the number of matches is not functional for microorganism identification when the experimental mass accuracy is below 10 ppm. The probability of false identification goes up as a function of the proteome density for a given organism, since random matches will be more frequent for organisms that have more dense proteomes.

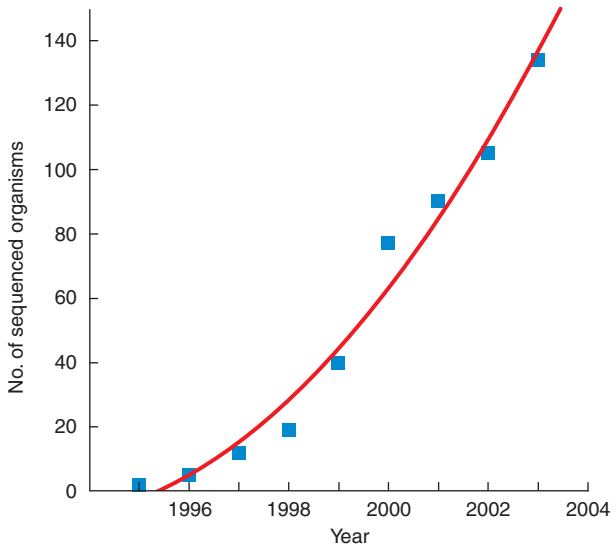



**Figure 4.** Increase in the number of completely sequenced genomes of prokaryotic organisms in publicly available genome databases as a function of time.

An analytical expression, the *p*-value, has been derived to calculate a numerical estimate of the probability for false identification due to matches between the experimentally observed mass values in an organism's biomarker signature and protein database masses of a different organism. The *p*-values vary from 0 to 1, with the lower values reflecting lower probability of matching by chance. Both theoretical analysis and *in silico* simulations confirm that *p*-values can be significantly improved by reducing the number of expected microorganism biomarkers from the entire set of potentially expressible proteins, based on rationally derived constraints (e.g., using biological domain knowledge about the expression levels of different proteins). Improving the mass accuracy of the experimentally observed biomarker signatures using, for example, Fourier transform ion cyclotron resonance (FTICR) MS[27,28] and/or isotope depletion in conjunction with TOF MS[29] also improves microorganism identification.

### Proteome Database Fidelity

A problem in matching the experimental masses ($M_{ex}$) with database-derived biomarker protein masses ($M_{th}$) is the occurrence of PTMs not reflected in the database. For instance, the most common PTM in prokaryota, cleavage of the *N*-terminal amino acid methionine (Met), if not explicitly reflected in the database, would lead to a discrepancy between observed ($M_{ex}$) and predicted ($M_{th}$) masses. Here, the discrepancy is 131 Da, the mass of a Met residue ($M_{th} = M_{ex} + 131$ Da).

A procedure to account for this specific PTM in putative protein biomarkers has been developed.[20] It is based on experimentally determined cleavage rules for the enzyme executing the PTM,[30] *N*-terminal aminopeptidase. This enzyme's activity is regulated by the type of the penultimate amino acid in the protein (i.e., the next-to-last amino acid in the translated sequence). Thus, the probability for this PTM to occur in a particular protein can be deduced by examining its amino acid sequence.

This rule has been implemented in an APL-developed Internet-accessible microorganism identification algorithm based on MS and proteome database queries (Fig. 5, see Ref. 31). The algorithm was used to characterize intact *Helicobacter pylori* Gram-negative bacteria, the most ubiquitous human pathogen. Including this PTM improves the identification reliability (the *p*-values) by at least an order of magnitude, from $10^{-2}$ to $10^{-3}$ in the case of *H. pylori*.[20]

## METHODS FOR IMPROVING IDENTIFICATION RELIABILITY

### Proteome Density: Rational Database Truncation

Statistical modeling predicts that reducing the proteome density (number of biomarkers) in the database

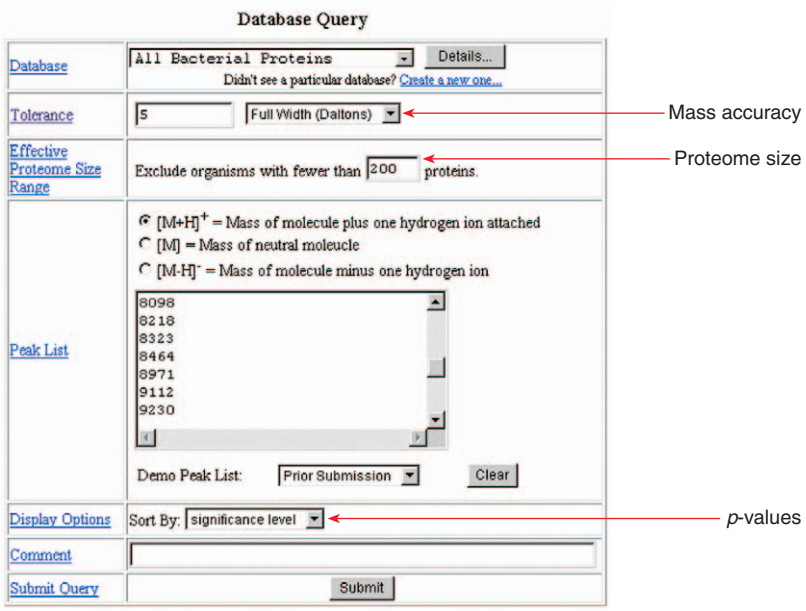



**Figure 5.** APL-developed Internet-accessible Web site with software for microorganism identification by a proteome database query.[31] The user enters experimentally observed biomarker masses as well as experimental mass accuracy, proteome database to be queried (e.g., complete or truncated), scoring method (ranking by, e.g., number of matches or *p*-values), etc.

by, for example, taking into consideration the differences in expression levels ("copy numbers") for various proteins, is a viable method for decreasing the probability of false identification. Recently, Pineda et al.[32] demonstrated experimentally that rational constraints on the number of potential microorganism biomarkers in a database can successfully scale up the permissible database size for a 95% detection confidence to more than 1000 microorganisms (for mass accuracies typical for a linear MALDI MS instrument).

Using biological domain knowledge about the expression levels of different proteins, a biomarker database has been generated by including only the highly expressed ribosomal proteins.[33] Typically this reduces the number of biomarkers in the range of 4 to 20 kDa by around 2 orders of magnitude. In a blind study, microorganisms represented in the database with 20 or more ribosomal biomarkers were correctly identified from their experimental MALDI mass spectra 100% of the time at the 95% confidence level, with no incorrect identifications.

Robustness with respect to variations in sample preparation protocol and mass analysis protocol was also demonstrated. Statistical analysis suggests that database truncation (i.e., rationally constraining the entire proteome of a sequenced organism to less than 50 highly expressed protein biomarkers) would allow successful identification of a microorganism from its experimental mass spectrum even without further improvement in the mass accuracy typical for linear MALDI TOF instruments. We also point out the existence of alternative methods, such as those involving the use of electrospray MS experiments to identify highly expressed protein biomarkers and to compile truncated databases for use in rapid microorganism identification.[34,35]

Statistical gene sequence analysis[36] has been used to create a truncated database of protein biomarkers for microorganism identification.[37] Each database entry is selected directly using only bioinformatics tools and microbial genome sequence information. To constrain the number of potential biomarkers, we exploit the fact that the alternative synonymous codons for a particular amino acid in a gene are not used randomly.[36,38] In addition, we use the positive correlation between the degree of bias in the codon usage frequencies in a gene and the expression level of the protein coded by that gene.[39–41]

We estimate codon biases by a simple statistical measure—the codon adaptation index (CAI).[38] This index uses a reference set of genes from a species to assess the usage frequencies of each codon, and a score for a particular gene is calculated from the usage frequency of the specific codons in that gene. The European Molecular Biology Open Software Suite[42] is used to calculate CAI values for the proteins of a microorganism from its genome sequence. The protein biomarkers for a microorganism are derived from the genes with the top 10 CAI values. Experimental data show that these CAI-derived protein biomarkers do indeed match peaks from experimental MALDI spectra from each microorganism. Therefore, the codon adaptation indices are a useful measure to predict highly expressed proteins for constructing a microorganism protein biomarker database, without the requirement for *a priori* protein annotation.

## Top-Down Proteomics

In a tandem MS (MS/MS) experiment, a precursor ion is selected, isolated, and excited by interaction with neutral gas molecules, electrons, or photons. The increased internal energy of the precursor ion causes its dissociation into sequence-specific fragments, which can be correlated to the precursor ion sequence.

The possibility of identifying an intact protein by deducing its partial amino acid sequence (tag) in an MS/MS experiment and subsequent homology search in a proteome database was first demonstrated by Mortz et al.[43] This top-down approach in proteomics ("top-down" and "bottom-up" indicate whether or not intact proteins are initially identified) was developed

further in a number of studies, all involving either an FTICR[44–48] or a quadrupole[49] ion trap. As already noted (Fig. 2), unambiguous identification of one or more intact protein biomarkers by this method will allow successful microorganism identification as well (provided the proteome database contains the microorganism). For instance, protein biomarkers from *B. cereus T* spores were analyzed by high-resolution tandem FTICR MS.[50] Fragmentation-derived sequence tags and BLAST sequence similarity searches[51] in a proteome database allowed unequivocal identification of the major protein biomarker as a small acid-soluble spore protein (SASP). From there, the organism itself could be unambiguously identified.[50]

Tandem MS of intact protein biomarkers still requires rather complex instrumentation (most often electrospray ionization/FTICR). However, recent developments in, e.g., TOF/TOF instruments, combined with bioinformatics, can drastically improve the specificity of individual microorganism identification, particularly in complex outdoor environments with high biological background.

## Bottom-Up Proteomics

Similar to the top-down proteomics methodology for microorganism identification, the bottom-up approaches are based on initial identification of individual proteins. In bottom-up proteomics, proteolysis (enzymatic digestion) of the proteins is first performed, resulting in several peptide fragments ("proteolytic" peptides) from each protein (see Fig. 6). The specificity of the proteolytic enzymes, complementary to or concurrently with peptide sequence tag information obtained by tandem mass spectrometry, improves the capability for unequivocal protein identification in classical bottom-up proteomics.[52–54]

The recent, rapid identification of *Bacillus* spores was achieved by selective solubilization of the SASP biomarkers and their subsequent proteolytic digestion *in situ* by using trypsin (immobilized on agarose beads) as a proteolytic enzyme. The proteolytic peptides were then analyzed by two different types of tandem mass spectrometers—MALDI TOF MS with a curved-field reflectron[55] or a hybrid ion trap/TOF mass spectrometer.[56] In the former, protein identification was obtained by partial sequencing of the proteolytic peptides in postsource decay experiments combined with proteome-based database searches. Similarly, precursor ions of interest were isolated and excited by collisions in the quadrupole ion trap. High-mass-accuracy fragment ions were detected in the TOF analyzer, allowing sequence-specific information to be obtained. The protein, and from there the microorganism source, were again identified by proteome database searching. The applicability of this bottom-up proteomics approach for the rapid identification of *Bacillus* spore mixtures was also illustrated.[55]

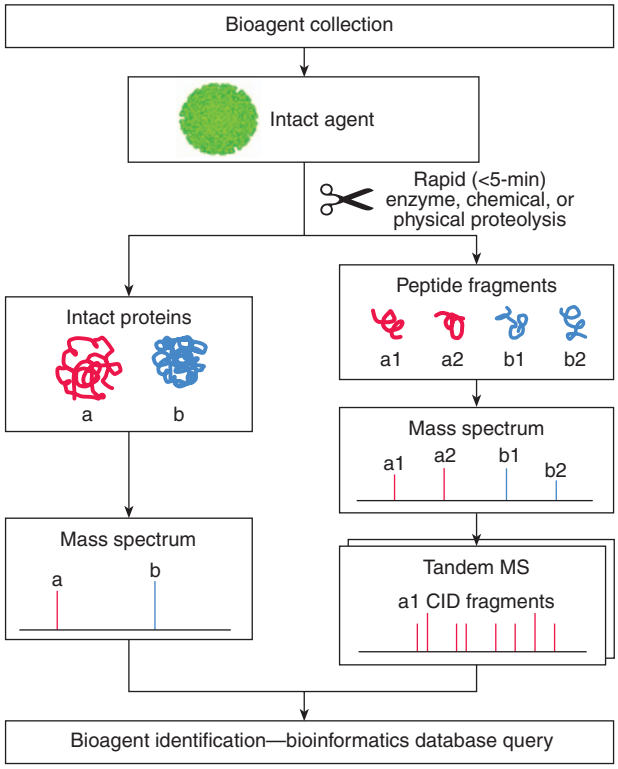



**Figure 6.** Two pathways for bioinformatics-based rapid microorganism identification by MS: either the intact protein biomarkers or the proteolytic peptides obtained after protein digestion are detected. Along the second pathway, tryptic enzymes are used to cleave the intact protein biomarkers at specific amino acid sites along the sequence, generating a peptide mass map of the intact protein. Subsequently, the proteolytic peptides can be further fragmented in an MS/MS experiment for protein identification by proteome database searches. Both pathways can be incorporated in parallel, with the second one serving to reduce the number of false positive microorganism identifications and/or confirming a positive identification (CID = collision-induced dissociation).

## *In Silico*–Generated Biomarker Databases

An approach based on experimental data from digested protein biomarkers, but employing a specially constructed database of organism-specific proteolytic peptide masses, has already been demonstrated.[57,58] Experimentally, the unknown microorganism is digested with a selective protease for a short time. The products of such incomplete digestion are then analyzed by MALDI TOF MS and compared to an *in silico*–generated database of proteolytic peptide masses. This approach can be applied for the rapid identification of viruses and other organisms (e.g., bacterial spores) that show a low number of biomarkers. Some examples are as follows.

The Sindbis virus AR 339 was successfully identified by using the masses of observed proteolytic peptides to query an *in silico*–generated database composed of proteolytic peptide masses for six viruses whose genomes have been sequenced.[57] Similarly, *Bacillus* spores were identified by this approach, again by creating *in silico* a database with the proteolytic peptide masses from all *Bacillus* and

*Clostridium* SASPs, with sequences available in public databases.[58] Further illustration of the capability for *in silico* protein biomarker database generation is provided by the comparison between experimentally observed proteolytic peptide fragments of the nucleocapsid protein in the SARS virus[59] and *in silico*–predicted products (Fig. 7).

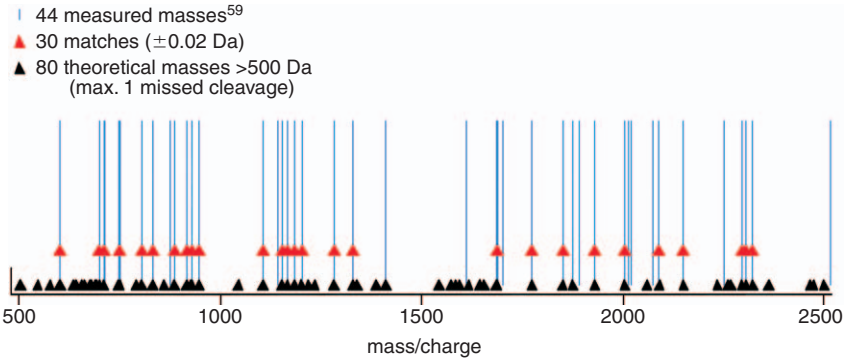



**Figure 7.** Comparison between experimentally determined masses of the proteolytic peptides of the nucleocapsid protein of the human SARS virus and an *in silico*–generated peptide mass signature map for the same protein. The nucleocapsid (*N*-structural) protein is expressed in the highest copy number in the virus and is the virion's major structural component, forming the helical nucleocapsid by associating with the viral RNA. The experimental mass values are from Krokhin et al.,[59] while the *in silico* digestion of the sequence (SwissProt entry P59595) was performed with the "PeptideMass" software tool, accessible from the SwissProt Web site.[22] Trypsin was used as a cleaving enzyme, and one missed cleavage was allowed.

We have also applied *in silico* biomarker analysis to *B. anthracis* spores. The rapid and reliable identification of these spores is a major task for successful countermeasures against their use as an instrument of bioterror. Recently, several studies illustrated the usefulness of the MS approach (e.g., MALDI TOF MS) for rapid identification of *Bacillus* spores and discrimination between different species such as *B. anthracis* and its close relative *B. cereus*.[60,61] Differences in the masses of detected SASP biomarkers for spores of *B. anthracis* and *B. cereus* were experimentally measured. However, the nature of these differences did not become clear until the genomes of the two organisms became publicly available.[24,25]

Table 1 lists spore-associated protein biomarkers in the mass range from 6 to 10 kDa from *B. cereus* and *B. anthracis*. It is clear that two of the three major SASP biomarkers (also observed experimentally in MALDI TOF spectra from spores[60,61]) have the same masses for both species. Only one of them—the major SASP B—differs in mass, by about 32 Da, between the two species. Comparison between the sequences corresponding to these biomarkers points to differences in only two amino acid positions (see the boxed insert). These differences can be traced back to single nucleotide polymorphisms in the corresponding genes for the two proteins. The availability of the complete genomes allows us to

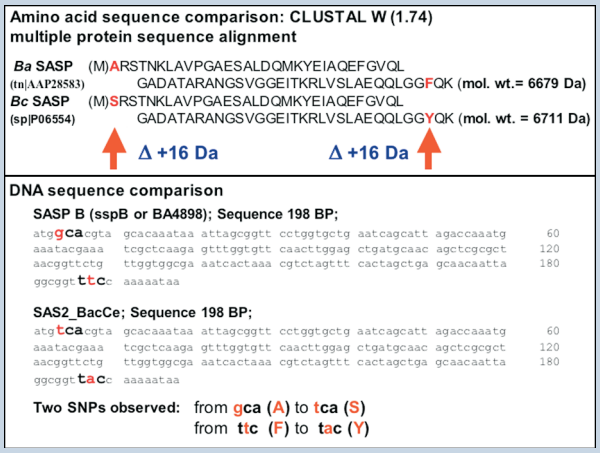
**Table 1.** *Bacillus* spore–associated protein biomarkers.[22]

| TrEMBL identification | Description | MS-detected molecular masses | |
|---|---|---|---|
| | | *B. anthracis* | *B. cereus* |
| AAP24546 | Small acid-soluble spore protein (SASP), gamma-type | 9738 | 9507 |
| AAP27427 | Spore coat protein K | | |
| AAP28906 | Spore coat protein F–related protein | | |
| AAP24097 | SASP | 6669 | |
| AAP24857 | SASP | **6835** → | **6835** |
| AAP25114 | Spore germination protein GerPF | | |
| AAP25116 | Spore germination protein GerPD | | |
| AAP25118 | Spore germination protein GerPB | | |
| AAP25119 | Spore germination protein GerPA | | |
| AAP25276 | SASP, alpha/beta family | | |
| AAP25879 | SASP, alpha/beta family | **7081** → | **7081** |
| AAP26210 | Spore germination protein GerPA | | |
| AAP26936 | SASP, alpha/beta family | 7163 | |
| AAP26939 | SASP | 7349 | |
| AAP28583 | SASP B | **6679** → | **6711** |
| AAP28729 | Spore germination protein, GerPF-like protein | | |

Note: Experimentally observed masses are matched to the sequence-derived masses by considering a post-translational modification: *N*-terminal Met cleavage. The three major biomarker peaks observed are shown in bold.

### COMPARISON OF AMINO ACID AND DNA SEQUENCES

The following shows the alignment of the respective amino acid and DNA sequences for the major biomarker SASPs detected in the MALDI spectra of *B. anthracis* or *B. cereus* spores. A single nucleotide polymorphism (SNP) is a codon that differs by only one base (one letter) in each of the genes being compared. Here, the two SNPs code for two different amino acids, reflected in the observed experimental mass difference of 32 Da between the two proteins. (Clustal W is a general-purpose multiple-sequence alignment program for DNA or proteins.)




reconstruct *in silico* and confidently predict the expected protein biomarkers for these two *Bacillus* species, including mass differences, which are important for successful discrimination between the two organisms.

## OUTLOOK

Miniaturized field-portable MS systems currently under development for biodefense at APL rely on the capability to predict biomarker signatures for biological pathogens and toxins under varying conditions. The set of MS-detectable biomarkers typically varies with the growth conditions and growth state of the organisms, sample collection and preparation protocols, and the presence of other organisms (whether in the normal background or deliberately inserted). Bioinformatics tools can be used to provide robustness with respect to such variability and are key to the successful deployment of MS-based instruments for counterproliferation, homeland security,[62,63] and biomedical applications[64] (see also articles by Ecelberger et al. and Antoine et al., this issue).

Future portable tandem MS systems that permit analysis of amino acid sequences within peptides from rapid enzymatic and/or chemical digests can enable rapid detection of agents followed by high-specificity validation (low false alarms). This could be important in outdoor environments with high biological background. Here, proteome database searches play a critical role in identifying the agent-specific proteins from experimentally derived partial amino acid sequence information. Such systems will also have the potential to detect certain classes of engineered organisms or novel, naturally emerging strains through identification of the modified protein biomarkers. The emergent organisms can potentially be classified (bacterial species or even viral class) according to biomarker similarities and/or specific protein sequence homologies to known organisms. The development of highly integrated and field-portable MS systems with bioinformatics capabilities will clearly impact an even wider range of applications beyond biodefense, including clinical microbiology, point-of-care medical diagnostics, and food and water safety.

## REFERENCES


[1]Anhalt, J. P., and Fenselau, C., "Identification of Bacteria Using Mass Spectrometry," *Anal. Chem.* **47**, 219–225 (1975).

[2]Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M., "Electrospray Ionization for Mass Spectrometry of Large Biomolecules," *Science* **246**, 64–71 (1989).

[3]Tanaka, K., Waki, H., Ido, Y., Akita, S., and Yoshida, Y., "Protein and Polymer Analysis up to m/z 100,000 by Laser Ionization Time-of-Flight Mass Spectrometry," *Rapid Commun. Mass Spectrom.* **2**, 151–153 (1988).

[4]Karas, M., and Hillenkamp, F., "Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons," *Anal. Chem.* **60**(20), 2299–2301 (1988).

[5]Bryden, W. A., Benson, R. C., Ecelberger, S. A., Phillips, T. E., Cotter, R. J., and Fenselau, C., "The Tiny-TOF Mass-Spectrometer for Chemical and Biological Sensing," *Johns Hopkins APL Tech. Dig.* **16**, 296–310 (1995).

[6]Claydon, M. A., Davey, S. N., Edwards Jones, V., and Gordon, D. B., "The Rapid Identification of Intact Microorganisms Using Mass Spectrometry," *Nat. Biotechnol.* **14**, 1584–1586 (1996).

[7]Holland, R. D., Wilkes, J. G., Raffi, F., Sutherland, J. B., Persons, C. C., et al., "Rapid Identification of Intact Whole Bacteria Based on Spectral Patterns Using Matrix-Assisted Laser Desorption/Ionization with Time-of-Flight Mass Spectrometry," *Rapid Commun. Mass Spectrom.* **10**, 1227–1232 (1996).

[8]Krishnamurthy, T., Ross, P. L., and Rajamani, U., "Detection of Pathogenic and Non-Pathogenic Bacteria by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry," *Rapid Commun. Mass Spectrom.* **10**, 883–888 (1996).

[9]Welham, K. J., Domin, M. A., Scannell, D. E., Cohen, E., and Ashton, D. S., "The Characterization of Micro-Organisms by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry," *Rapid Commun. Mass Spectrom.* **12**, 176–180 (1998).

[10]Wang, Z. P., Russon, L., Li, L., Roser, D. C., and Long, S. R., "Investigation of Spectral Reproducibility in Direct Analysis of Bacteria Proteins by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry," *Rapid Commun. Mass Spectrom.* **12**, 456–464 (1998).

[11]Arnold, R. J., and Reilly, J. P., "Fingerprint Matching of *E. coli* Strains with Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry of Whole Cells Using a Modified Correlation Approach," *Rapid Commun. Mass Spectrom.* **12**, 630–636 (1998).

[12]Thomas, J. J., Falk, B., Fenselau, C., Jackman, J., and Ezzell, J., "Viral Characterization by Direct Analysis of Capsid Proteins," *Anal. Chem.* **70**, 3863–3867 (1998).

[13]Winkler, M. A., Uher, J., and Cepa, S., "Direct Analysis and Identification of *Helicobacter* and *Campylobacter* Species by MALDI-TOF Mass Spectrometry," *Anal. Chem.* **71**, 3416–3419 (1999).

[14]Jarman, K. H., Cebula, S. T., Saenz, A. J., Petersen, C. E., Valentine, N. B., et al., "An Algorithm for Automated Bacterial Identification Using Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry," *Anal. Chem.* **72**, 1217–1223 (2002).

[15]Scholl, P. F., Leonardo, M. A., Rule, A. M., Carlson, M. A., Antoine, M. D., and Buckley, T. J., "The Development of Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry for the Detection of Biological Warfare Agent Aerosols," *Johns Hopkins APL Tech. Dig.* **20**, 343–351 (1999).

[16]Fenselau, C., and Demirev, P., "Characterization of Intact Microorganisms by MALDI Mass Spectrometry," *Mass Spectrom. Rev.* **20**, 157–171 (2001).

[17]Lay, J., "MALDI-TOF Mass Spectrometry of Bacteria," *Mass Spectrom. Rev.* **20**, 172–194 (2001).

[18]Demirev, P., Ho, Y. P., Ryzhov, V., and Fenselau, C., "Microorganism Identification by Mass Spectrometry and Protein Database Searches," *Anal. Chem.* **71,** 2732–2738 (1999).

[19]Pineda, F., Lin, J. S., Fenselau, C., and Demirev, P., "Testing the Significance of Microorganism Identification by Mass Spectrometry and Proteome Database Search," *Anal. Chem.* **72**, 3739–3744 (2000).

[20]Demirev, P., Lin, J. S., Pineda, F. J., and Fenselau, C., "Bioinformatics and Mass Spectrometry for Microorganism Identification: Proteome-Wide Post-Translational Modifications and Database Search Algorithms for Characterization of Intact *H. pylori*," *Anal. Chem.* **73**, 4566–4573 (2001).

[21]Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., et al., "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL," *Nucl. Acids Res.* **31**(1), 365–370 (1 Jan 2003).

[22]ExPASy Molecular Biology Server Web site, http://expasy.ch/.

[23]Institute for Genomic Research Web site, http:// www.tigr.org.

[24]Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., et al., "The Genome Sequence of *Bacillus anthracis Ames* and Comparison to Closely Related Bacteria," *Nature* **423**, 81–86 (2003).

[25]Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., et al., "Genome Sequence of *Bacillus cereus* and Comparative Analysis with *Bacillus anthracis*," *Nature* **423**, 87–91 (2003).

[26]Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A. R., et al., "The Genome Sequence of the SARS-Associated Coronavirus," *Science* **300**, 1399–1404 (2003).

[27]Lee, S. W., Berger, S. J., Martinovic, S., Pasa-Tolic, L., Anderson, G. A., et al., "Direct Mass Spectrometric Analysis of Intact Proteins of the Yeast Large Ribosomal Subunit Using Capillary LC/FTICR," *Proc. Nat. Acad. Sci.* **99**, 5942–5947 (2002).

[28]Jones, J. J., Stump, M. J., Fleming, R. C., Lay, J. O., and Wilkins, C. L., "Investigation of MALDI-TOF and FT-MS Techniques for Analysis of *Escherichia coli* Whole Cells," *Anal. Chem.* **75**, 1340–1347 (2003).

[29]Stump, M. J., Jones, J. J., Fleming, R. C., Lay, J. O., and Wilkins, C. L., "Use of Double-Depleted C-13 and N-15 Culture Media for Analysis of Whole Cell Bacteria by MALDI Time-of-Flight and Fourier Transform Mass Spectrometry," *J. Am. Soc. Mass Spectrom.* **14**, 1306–1314 (2003).

[30]Gonzales, T., and Baudouy, J. R., "Bacterial Aminopeptidases: Properties and Functions," *FEMS Microbiol. Rev.* **18**, 319–344 (1996).

[31]http://infobacter.jhuapl.edu.

[32]Pineda, F., Antoine, M., Demirev, P., Feldman, A., Jackman, J., et al., "Rapid Microorganism Identification by MALDI Mass Spectrometry and Model-Derived Ribosomal Protein Biomarkers," *Anal. Chem.* **75**, 3817–3822 (2003).

[33]Arnold, R. J., and Reilly, J. P., "Observation of *E. coli* Ribosomal Proteins and Their Posttranslational Modifications by Mass Spectrometry," *Anal. Biochem.* **269**, 105–112 (1999).

[34]Wang, Z. P., Dunlop, K., Long, S. R., and Li, L., "Mass Spectrometric Methods for Generation of Protein Mass Database Used for Bacterial Identification," *Anal. Chem.* **74**, 3174–3182 (1 Jul 2002).

[35]Williams, T. L., Leopold, P., and Musser, S., "Automated Postprocessing of Electrospray LC/MS Data for Profiling Protein Expression in Bacteria," *Anal. Chem.* **74**(22), 5807–5813 (15 Nov 2002).

[36]Karlin, S., Campbell, A. M., and Mrazek, J., "Comparative DNA Analysis Across Diverse Genomes," *Ann. Rev. Genetics* **32**, 185–225 (1998).

[37]Demirev, P. A., Feldman, A. B., Lin, J. S., Pineda, F. J., and Resch, C. L., "Microorganism Identification by Mass Spectrometry and Bioinformatics-Generated Biomarker Databases," in *Proc. 51st Ann. Conf. of the ASMS*, CD-ROM, Montreal, Canada (Jun 2003).

[38]Sharp, P. M., and Li, W. H., "The Codon Adaptation Index: A Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications," *Nucl. Acids Res.* **15**, 1281–1295 (1987).

[39]Perrot, M., Sagliocco, F., Mini, T., Monribot, C., Schneider, U., et al., "Two-Dimensional Gel Protein Database of *S. cerevisiae*," *Electrophoresis* **20**, 2280–2298 (Aug 1999).

[40]Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., et al., "An Accurate Mass Tag Strategy for Quantitative and High-Throughput Proteome Measurements," *Proteomics* **2**, 513–523 (2002).

[41]Hunter, T. C., Andon, N. L., Koller, A., Yates, J. R., and Haynes, P. A., "The Functional Proteomics Toolbox: Methods and Applications," *J. Chromatogr.* **B782**, 165–181 (2002).

[42]European Molecular Biology Open Software Suite Web site, http://www.hgmp.mrc.ac.uk/Software/EMBOSS/

[43]Mortz, E., O'Connor, P., Roepstorff, P., Kelleher, N., Wood, T., et al., "Sequence Tag Identification of Intact Proteins by Matching Tandem Mass Spectral Data Against Sequence Data Bases," *Proc. Nat. Acad. Sci.* **93**, 8264–8267 (1999).

[44]Zubarev, R., Kelleher, N., and McLafferty, F. W., "Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process," *J. Am. Chem. Soc.* **120**, 3265–3266 (1998).

[45]Li, W., Hendrickson, C., Emmett, M., and Marshall, A. G., "Identification of Intact Proteins in Mixtures by Alternated Capillary Liquid Chromatography Electrospray Ionization and LC ESI Infrared Multiphoton Dissociation Fourier Transform Ion Cyclotron Resonance Mass Spectrometry," *Anal. Chem.* **71**, 4397–4402 (1999).

[46]Xiang, F., Anderson, G. A., Veenstra, T., Lipton, M., and Smith, R. D., "Characterization of Microorganisms and Biomarker Development from Global ESI-MS/MS Analyses of Cell Lysates," *Anal. Chem.* **72**, 2475–2481 (2000).

[47]Meng, F. Y., Cargile, B. J., Miller, L. M., Forbes, A. J., Johnson, J. R., and Kelleher, N., "Informatics and Multiplexing of Intact Protein Identification in Bacteria and the Archaea," *Nat. Biotechnol.* **19**(10), 952–957 (Oct 2001).

[48]Taylor, G. K., Kim, Y. B., Forbes, A. J., Meng, F. Y., McCarthy, R., and Kelleher, N. L., "Web and Database Software for Identification of Intact Proteins Using 'Top Down' Mass Spectrometry," *Anal. Chem.* **75**(16), 4081–4086 (15 Aug 2003).

[49]Cargile, B., McLuckey, S. A., and Stephenson, J. L., "Identification of Bacteriophage MS2 Coat Protein from *E. coli* Lysates via Ion Trap Collisional Activation of Intact Protein Ions," *Anal. Chem.* **73**, 1277–1285 (2001).

[50]Demirev, P., Ramirez, J., and Fenselau, C., "Tandem Mass Spectrometry of Intact Proteins for Characterization of Biomarkers from *Bacillus cereus T* Spores," *Anal. Chem.* **73,** 5725–5731 (2001).

[51]Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., et al., "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Res.* **25**, 3389–3402 (1997).

[52]Wilkins, M. R., Appel, R. D., Hochstrasse, D. F., and Williams, K. L. (eds.), *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag, New York (1997).

[53]Liebler, D. C., *Introduction to Proteomics: Tools for the New Biology*, Humana Press (2001).

[54]Aebersold, R., and Mann, M., "Mass Spectrometry-Based Proteomics," *Nature* **422**(6928), 198–207 (13 Mar 2003).

[55]Warscheid, B., and Fenselau, C., "Characterization of *Bacillus* Spore Species and Their Mixtures Using Postsource Decay with a Curved-Field Reflectron," *Anal. Chem.* **75**, 5608–5617 (2003).

[56]Warscheid, B., Jackson, K., Sutton, C., and Fenselau, C., "MALDI Analysis of *Bacilli* in Spore Mixtures by Applying a Quadrupole Ion Trap Time-of-Flight Tandem Mass Spectrometer," *Anal. Chem.* **75**, 5608–5617 (2003).

[57]Yao, Z., Demirev, P., and Fenselau, C., "Mass Spectrometry-Based Proteolytic Mapping for Rapid Virus Identification," *Anal. Chem.* **74**, 2529–2534 (2002).

[58]English, R. D., Warscheid, B., Fenselau, C., and Cotter, R. J., "*Bacillus* Spore Identification via Proteolytic Peptide Mapping with a Miniaturized MALDI TOF Mass Spectrometer," *Anal. Chem.* **75**, 6886–6893 (2003).

[59]Krokhin, O., Li, Y., Andonov, A., Feldmann, H., Flick, R., et al., "Mass Spectrometric Characterization of Proteins from the SARS Virus," *Mol. Cell. Proteomics* **2**, 346–356 (2003).

[60]Elhanany, E., Barak, R., Fisher, M., Kobiler, D., and Altboum, Z., "Detection of Specific *Bacillus anthracis* Spore Biomarkers by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry," *Rapid Comm. Mass Spectrom.* **15**, 2110–2116 (2001).

[61]Hathout, Y., Setlow, B., Cabrera-Martinez, R. M., Fenselau, C., and Setlow, P., "Small, Acid-Soluble Proteins as Biomarkers in Mass Spectrometry Analysis of *Bacillus* Spores," *Appl. Environ. Microbiol.* **69**, 1100–107 (2003).

[62]Cornish, T. J., and Bryden, W. A., "Miniature Time-of-Flight Mass Spectrometer for a Field-Portable Biodetection System," *Johns Hopkins APL Tech. Dig.* **20**(3), 335–342 (1999).

[63]McLoughlin, M. P., Allmon, W. R., Anderson, C. W., Carlson, M. A., DeCicco, D. J., and Evancich, N. H., "Development of a Field-Portable Time-of-Flight Mass Spectrometer System," *Johns Hopkins APL Tech. Dig.* **20**(3), 326–332 (1999).

[64]Ko, H. W., "Biomedical and Biochemical Technology at APL," *Johns Hopkins APL Tech. Dig.* **24**(1), 41–51 (2003).

ACKNOWLEDGMENTS: We are grateful to Catherine Fenselau (Department of Chemistry and Biochemistry, University of Maryland, College Park) and Fernando Pineda (Department of Molecular Microbiology, The Johns Hopkins University Bloomberg School of Public Health) for fruitful collaborations. We also acknowledge Miquel Antoine, Joany Jackman, and Cheryl Resch, all at APL, for their contributions to the work reviewed here. Funding at APL has been provided by internal R&D grants.

## THE AUTHORS

PLAMEN A. DEMIREV is a Senior Professional Staff member in the Sensor Sciences Group of APL's Research and Technology Development Center. He has an M.S. (physics, 1979) from the University of Sofia and a Ph.D. (chemistry, 1988) from the Bulgarian Academy of Sciences. In 1990 he joined the faculty of Uppsala University, Sweden, where he became a docent in ion physics (1995). Before joining APL in 2001, Dr. Demirev was a research scientist at the University of Maryland. His current interests include physical methods for rapid detection of human pathogens in complex environments. He has co-authored more than 90 scientific papers in fields ranging from ion–solid interactions to atomic and molecular clusters and mass spectral quantification of organics. His e-mail is plamen.demirev@jhuapl.edu.

ANDREW B. FELDMAN is the supervisor of the Bioinformatics Section in APL's Research and Technology Development Center. He received A.B. (1986), A.M. (1988), and Ph.D. (1997) degrees in physics from Harvard University. He was a postdoctoral fellow at the Harvard University–Massachusetts Institute of Technology Division of Health Sciences and Technology from 1996 to 2000, working in the field of computational biophysics. Dr. Feldman joined APL in 2000. He has contributed to numerous programs in biodetection and defense bioinformatics and has developed a multi-institutional program for rapid detection of malaria using mass spectrometry. He is a member of the American Physical Society and currently serves as APL's technical lead for the National Chemical, Biological, and Radiological Technology Alliance. His e-mail address is andrew.feldman@jhuapl.edu.

JEFFREY S. LIN received a B.S.E. in mechanical/aerospace engineering from Princeton University in 1986 and joined APL's Aeronautics Department in that same year. He received an M.S. in computer science from The Johns Hopkins University in 1989 and subsequently developed automated diagnostic and nondestructive evaluation systems. Mr. Lin joined the System and Information Sciences Group of the Research and Technology Development Center in 1996, where he recently has been developing and applying bioinformatics algorithms for the detection of biological warfare agents. He is a member of the American Society for Mass Spectrometry. His e-mail address is jeff.lin@jhuapl.edu.