Chapter 5
Bottom-Up Proteomics Methods for Strain-Level Typing and Identification of Bacteria

Jacek P. Dworzanski

Introduction

The microbiological methods used for the detection and identification of bacteria were by necessity based on culturing and staining techniques combined with microscopic evaluation of cells. However, over the past few decades the use of molecular methods gained importance in microbiological laboratories and led to tremendous changes in a way of detecting microorganisms, their identification at the species level, and typing of isolates to infer subspecies diversity. Although routine identification methods continue to be based on the determination of the morphology, differential staining, and physiology of a microbial isolate, currently these methods are gradually supplanted by the use of diverse genomic and proteomic-based approaches that include mass spectrometry (MS) techniques, among others.

MS-based methods represent a broad group of highly versatile approaches that use precise mass measurements to infer identity of diverse biomolecules. Although for many decades the scope of investigated molecules was limited by their molecular mass and polarity, developments in soft ionization techniques like electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) substantially broaden the range of investigated species. Nowadays not only proteins and nucleic acids but also multimolecular complexes, and even whole viruses can be mass analyzed by modern MS instruments and used to infer genomic information encoded in nucleotide and amino acid sequences. Therefore, MS-based analysis of nucleic acid amplicons and proteins is increasingly replacing the older, time-consuming, and labor-intensive approaches.

Currently, both “top-down” and “bottom-up” methods are used to analyze microbial proteins by MS. In top-down approach, proteins are analyzed to determine molecular masses of intact proteins and to characterize them by using gas-phase fragmentation techniques. The bottom-up characterization of proteins uses prote-

J. P. Dworzanski (✉)
Leidos, Inc., 1816 Queen Anne Sq, Bel Air, MD 21015, USA
e-mail: jacek.p.dworzanski@gmail.com

P. Demirev, T. R. Sandrin (eds.), *Applications of Mass Spectrometry in Microbiology,*
DOI 10.1007/978-3-319-26070-9_5

olysis and analysis of released peptides by tandem MS to reveal their amino acid sequences. Bottom-up analysis of a protein mixture is usually called shotgun proteomics, to indicate analogy to shotgun genomic sequencing (Yates 1998).

Molecular criteria used for defining bacterial species have been progressing from the determination of nucleotide content (e.g., mol% G-C), DNA–DNA hybridization (DDH) and melting point analysis, which measure the degree of genetic similarity between two genomes, 16S rRNA gene sequencing, and multilocus sequence analysis (MLSA) of housekeeping genes, up to proteomics-based analysis and sequencing the whole microbial genome.

Since the 1960s, a means for determining relatedness of strains was based on a comparison of genomic similarities measured by DDH between DNA strands. DDH has driven the construction of current bacterial taxonomy and has become a gold standard for the delineation of bacterial species, which were defined as a collection of strains with a DDH value of at least 70% (Wayne et al. 1987). However, these methods are difficult and laborious; therefore, other genomic approaches were developed to replace DDH, including DNA sequencing by hybridization with custom-designed microarrays, or comparison of 16S rRNA sequences used with the assumption that if strains share less than 97% of sequence similarity, they belong to different species (Stackebrandt and Goebel 1994). In fact, sequencing of 16S rRNA combined with searching a database (DB) with millions of entries has become the most commonly used method for identifying and classifying microbial species (Cole et al. 2009; Quast et al. 2013). However, the 16S rRNA gene has limited specificity, for example, only 80% of isolates were recently found to be unambiguously identified at the species level (Chatellier et al. 2014). Therefore, genes with less-conserved sequences from protein-coding loci, for example, DNA gyrase *(gyrB)* or RNA polymerase *(rpoB)* have to be used instead. Unfortunately, different genes may give different patterns of interspecies relationships due to horizontal gene transfer (HGT) or unequal rates of nucleotide substitution. Therefore, sequence analysis of 6−8 housekeeping genes (a multilocus approach) was designed to increase the resolution and to buffer the potential impact of the HGT on the determined relatedness. Despite being successful in phylogenetic discrimination of strains at the subspecies level, it has major drawbacks that arise from a putative bias in gene selection and amplification primer availability.

The universally adopted genomic approaches to strain subtyping still use DNA fingerprinting techniques based on: (i) analysis of restriction nuclease digested whole cell DNA fragments resolved by pulsed-field gel electrophoresis (PFGE), or (ii) polymerase chain reaction (PCR)-amplified segments targeting loci with a variable number of tandem repeats (VNTR), which reveal relatedness at a microevolutionary level by using the technique called multiple locus VNTR analysis (MLVA). These DNA fingerprinting techniques are used for high accuracy isolate characterization, for example, by the Centers for Disease Control and Prevention PulseNet program (http://www.cdc.gov/pulsenet/pathogens/pfge.html) to recognize, investigate, and control outbreaks of food infections. However these procedures are also quite lengthy. For example, a standard operating procedure of PFGE takes up to 5 days and includes the isolation and growth of the culture, cell lysis, digestion of DNA with restriction nuclease in agarose gel, followed by gel electrophoresis; staining and documentation of a gel. The other fingerprinting procedure used to discriminate between closely related strains is called "optical mapping" or whole-genome mapping, which provides maps of a chromosome based on optical analysis of DNA fragments obtained by digestion with a restriction nuclease. The experimental data obtained by this technique can be correlated directly to DNA sequence information in the public databases so that markers for resistance or virulence can be easily recognized (Miller 2013).

Reliable characterization of microorganisms at subspecies level is increasingly essential in clinical, biotechnological, environmental, and epidemiological studies. Currently, reliable characterization of microorganisms is based on genetic and genomic criteria inferred from complete genome sequences, considered as the reference standard for determining bacterial phylogenies. The most widely adopted tools for comparing and analyzing complete genome sequences are based on in silico calculation of digital DDH-type indices representing conservation of the core genome, the DNA content measured as the proportion of DNA shared by two genomes (Goris et al. 2007), and alignment-free approaches using oligonucleotide frequencies for phylogenomic inferences (Bohlin et al. 2008). The DDH-type indices include average nucleotide identity (ANI) of all orthologous genes shared by two genomes (Konstantinidis and Tiedje 2005a) or its equivalent, average amino acid identity (AAI) of protein-coding genes (Konstantinidis and Tiedje 2005b), calculated using BLAST or BLASTP algorithms; the maximal unique matches index (MUMi) (Deloger et al. 2009); and refinements of these approaches, for instance, by using the rapid alignment tool MUMer (Richter and Rosselló-Móra 2009). More recently, a similar method called the genome BLAST distance phylogeny makes use of DNA rather than genes and uses a set of local alignment tools and a special formula to calculate a genome-to-genome distance (Meier-Kolthoff et al. 2013).

Of these, ANI and AAI indices have been most widely used as possible next-generation gold standards for species delineation because they represent a robust measure of the genetic distance between two sequenced bacterial strains and are strongly correlated with DDH data. In addition, they are also strongly correlated with 16S rRNA gene sequence similarity, the percentage of conserved DNA, the mutation rate of the genome, and offer resolution at the subspecies level (Konstantinidis and Tiedje 2005a, b; Goris et al. 2007). However, the major drawback of this approach is that it is only available for a pair of strains with complete genome sequences.

Importantly, both AAI index and DDH values for bacterial strains can be predicted experimentally by using a proteomics-derived index termed the fraction of shared (tryptic) peptides (FSP, Dworzanski et al. 2010). FSP is calculated from the peptide-centric bottom-up proteomics MS data sets acquired during analysis of an unknown bacterial strain and searched, with a suitable search engine, against DB proteomes predicted from complete genome sequences of reference strains. In this approach, the high-throughput proteome identification of thousands of released peptides reveals amino acid sequence information translated from genomic sequences that may be used not only for predicting strain similarities but also for identifications of genes that are actually expressed. Consequently, bottom-up proteomics allows high-resolution typing and subspecies level identifications, reflecting both genomic similarities and supplanting traditional typing approaches based on serological (e.g., H-antigen typing) and phenotypic properties, like antibiotic resistance.

Currently, the important pieces of information about an isolated bacterial strain, that is, the species, serovar, subtype, or its antibiotic resistance are generated by separate tests. However, the bottom-up proteomics analysis potentially allows finding this kind of information in just one test comprising liquid chromatography (LC)-MS/MS analysis and data mining with a suite of bioinformatics tools. In this chapter, I will focus on a group of highly versatile bottom-up shotgun-proteomics methods allowing for the identification, classification, and characterization of microorganisms by revealing: (i) strain identity, (ii) serotype, (iii) virulence, (iv) antimicrobial resistance profile, and (v) a subtype reflecting differences in both the gene content and single amino acid variations (SAVs) of expressed proteins.

Cell Harvesting and Protein Extraction

Samples analyzed by MS-based approaches for bacteria identification are initially processed in microbiological laboratory settings; therefore, researchers should follow standard procedures used for sample collection and preconcentration, and these methods will not be discussed here.

Generally, clinical, environmental, or food samples are processed to obtain pure cultures either directly, for example, from blood samples, or by isolating them from other cells and/or food and environmental matrices using diverse enrichment techniques. Such cells are then grown to obtain pure cultures by using diverse selective or enriched liquid and agar-solidified media supporting the growth of a wide range of microorganisms. The microbial cells are then harvested, washed with buffers or distilled water, and processed to extract their proteins for further proteomic analysis. The sample processing steps usually follow standard protocols developed for shotgun-proteomics workflows that include microbial cell lysis, extraction, solubilization and preseparation of proteins, specific cleavage of proteins into peptides, and peptide purification and separation immediately prior to MS analysis (Gundry et al. 2009). However, depending on the infectability of the material, all the steps preceding peptide analysis should be carried out in a laboratory approved for working with infectious agents.

Cell Lysis and the Preparation of Whole Cell Protein Extracts

Microbial cell lysis provides access to cytosolic and the majority of membrane proteins, and therefore is a crucial step for efficient extraction of expressed proteins and their analysis by shotgun-proteomics methods. Such whole cell protein extracts are usually obtained by rupturing cells in lysis buffers containing protease inhibitors by using physical methods, such as ultrasonication, bead beating, French press, freeze-thaw, thermal lysis, pressure cycling, and (bio)chemical lysis procedures, involving murolytic enzymes like lysozyme, detergents, chaotropes, and other reagents. However, in case of biochemical and chemical methods, the compatibility of (bio) chemical reagents with the analytical technique must be considered. For example, although lysozyme is very effective in lysing Gram-positive bacteria by hydrolyzing glycosidic linkages in the bacterial wall peptidoglycan, it may interfere with the identifications of peptides obtained by global digestion of protein extracts.

In general, the obtained lysates are cleared by centrifugation to remove cellular debris, and the supernatant or "supernate" is considered a whole cell protein extract composed of a complex mixture of proteins, other cell constituents such as lipids, nucleic acids, polysaccharides, low molecular mass metabolites, and all additives. These additives include buffers, chaotropes, detergents, or cocktails of proteinase inhibitors, which are added to aid in protein extraction and preserve the integrity of a proteome.

Cell lysis can also include a combination of chemical and diverse physical methods. For example, Lee et al. (2006) demonstrated rapid lysis of bacterial cells using both thermal and mechanical lysis directly on a chip through a combination of the laser irradiation and agitation with magnetic beads. This and many other microfluidic devices for cell lysis were recently reviewed by Nan et al. (2014). In another example, Napoli et al. (2014) performed cellular lysis of bacteria through the frictional action of glass beads added to the sample solution combined with pressure waves provided by a probe sonication of a cells/glass beads mixture. However, sonication becomes problematic for lysis of pathogenic microorganisms due to safety concerns, and is not well adapted for automated, high-throughput liquid-handling platforms. An approach aimed to overcome such concerns was proposed and tested by Tanca et al. (2013) in their comparative study of sample preparation workflows. They extracted proteins from *Escherichia coli* by subjecting cells to lysis in buffered solutions of surfactants for 30 min at 95 °C by using a thermo-mixer at 500 rev/min.

For highly pathogenic microorganisms which should be handled in the biosafety level 3 (BSL-3) laboratory, Tracz et al. (2013) grew bacteria with required biocontainment precautions and after harvesting and resuspending cells in sterile water, they were then gamma-inactivated. Further, microbial suspensions were incubated at 95 °C for 5 min and vortexed with glass beads to rupture cells and release proteins. Similar safety precautions were also used by Jabbour et al. (2010b) and Wade et al. (2011) by pelleting the cells from cultures by centrifugation, washing, resuspending in a buffer, and lysing them thermally by a 1-h long heating at 95 °C. In addition, a portion of each lysed sample was plated and incubated for 5 days to ensure no growth prior to removing samples from the BSL-2 or BSL-3 laboratory. However, for lysing enterohemorrhagic and enteroaggregative *E. coli* strains, Jabbour et al. (2014) used the bead beating technique.

The choice of a lysis method may also be tailored for a specific group of microorganisms. For example, François et al. (2014) prepared a total protein extract from *Staphylococcus aureus* by resuspending harvested cells in a lysis buffer containing calcium and magnesium chlorides and protease inhibitors. By adding the murolytic enzyme lysostaphin that cleaves crosslinking pentaglycin bridges in the cell wall of *Staphylococci,* they released protoplasts that immediately underwent lysis due to hyposmotic shock. The presence of a high-molecular DNA in such samples causes high viscosity that may be reduced by adding DNase; however, this contaminates the sample and may complicate sample processing workflows.

Preparation of Subcellular Fractions

Among subcellular proteomes investigated for identification of bacterial subspecies, attention was concentrated on surface and membrane-associated proteins, especially outer membrane proteins (OMPs) of Gram-negative bacteria, surface layer (S-layer) proteins of Gram-positive bacteria, flagella, and extracellular proteins (ECPs).

Outer Membrane Proteins

After cell lysis by ultrasonication or any other method, cell debris is usually removed by centrifugation and the resulting supernatant is assumed to contain the total cellular protein fraction composed of both membrane and the soluble cytosolic proteins. Therefore, in some applications it is advantageous to separate these proteins, for example, by ultracentrifugation, to obtain the pellet corresponding to the membrane fraction. For example, Jabbour et al. (2010b) and Wade et al. (2011)—after thermal lysis and removing cell debris by centrifugation—ultracentrifuged the obtained supernatants at 100,000 g to pellet membrane proteins which they resuspended in a buffered solution of N-lauroylsarcosinate. Because OMPs of Gram-negative bacteria are insoluble in sarcosine solutions, ultracentrifugation of such mixture allows for pelleting OMPs.

A more streamlined procedure for OMP isolation was used by Damron et al. (2009). They simply suspended harvested cells in a buffered solution of sarcosyl with protease inhibitors and lysed cells by sonication on ice. The lysate was then clarified by low-speed centrifugation, and the supernatant was centrifuged at 40,000 g to obtain a pellet containing OMPs.

Among filtration methods, there is a growing popularity of using ultrafiltration centrifugal devices, for example, Microcon$^{(R)}$-type filters (EMD Millipore, Billerica, MA, USA), which allow for removal of lower molecular mass contaminants, buffer exchange, and sample concentration.

Surface Layer Proteins

Surfaces of many microbial cells are coated with a layer of proteins (known as “S-layer”) that have an important role in the cell’s growth, survival, and interaction with the host organism, and are present in a high copy number. In addition, such proteins are easily available for solvent extraction and could be used for sequence-based identification and typing of microbial cells.

For example, for the identification and typing of *Lactobacillus* spp. used as probiotic bacteria in dietary supplements and milk products, the extraction of S-layer proteins was carried out from the water washed bacterial cells by incubation with 5 M lithium chloride or guanidine hydrochloride solutions (Johnson et al. 2013; Podlesny et al. 2011). After the removal of cells by centrifugation and filtration, the extract may be concentrated by ultrafiltration. The precipitated S-layer proteins are suspended in 1 M lithium chloride to dissociate any proteins which are soluble, and the purified S-layer proteins are pelleted by centrifugation (Goh et al. 2009). Alternatively, S-layer proteins—which are characterized by a high isoelectric point (pI > 9)—may be purified by a cation-exchange chromatography (Podleśny et al. 2011).

Preparation of Flagella

Flagella are isolated from bacteria growing on plates by scraping and suspending in a suitable buffer while those cultivated on liquid media are directly harvested by centrifugation. However, centrifugation may cause cell surface damage through collisions resulting in shear forces on the bacterial cell surface; therefore, it should be performed at low speeds or even avoided. For example, Cheng et al. (2013) harvested a full loopful of enteric bacteria and gently suspended them in a lysozyme solution, followed by vigorous vortexing to shear off flagella and centrifugation to remove cells. The supernatant was filtered through a 0.2 µm pore size low protein binding membrane of a syringe filter to retain and wash flagella with deionized water. Finally, the isolated flagella were on-filter trypsinized by exposing them for a couple of hours to a trypsin solution.

However, Sun et al. (2013) did not use lysozyme in their protocol on isolation of flagella from *Shewanella* cells that produce a single polar flagellum. Therefore, after a vortexing step to shear off flagella and removing the cells by centrifugation, they passed the supernatant containing flagella through a 0.45-µm-pore filter that did not retain them. Consequently, they used ultracentrifugation to pellet purified flagella and re-suspend them in water for further analysis.

Extracellular Proteins

ECPs include proteins that are actively transported to the bacterial outer surroundings through the cytoplasmic membrane, as well as those that are simply shed from the bacterial surface. Therefore, they are prepared from spent media obtained after harvesting cells by centrifugation. The cell-free media are usually sterilized by filtration, and the ECPs are routinely isolated by precipitation with trichloroacetic acid (TCA), followed by washing with acetone to remove TCA (Sun et al. 2014; Enany et al. 2014; Halbedel et al. 2014). However, ultrafiltration may also be used for concentrating ECPs, for example, by using centrifugal ultrafiltration devices (Jabbour et al. 2014).

Processing of Bacterial Proteins for Bottom-Up Proteomics Analysis

The conventional method of proteome analysis involves gel separation of proteins as the final purification step that is followed by in-gel digestion and mass spectrometric analysis of released peptides (Tonella et al. 2001). This sample preparation method is still widely used in proteomics of bacteria (Hartmann et al. 2014) and although it has many advantages, it is a relatively lengthy and labor-intensive procedure. Therefore, the gel-free, shotgun protein digestion methods are frequently used for faster and more efficient processing of proteins for LC-MS/MS analysis of peptides. However, the shotgun protocols have to deal with highly contaminated samples because proteins extracted from bacterial cells usually contain other cell constituents and reagents, including those used for breaking interactions involved in aggregation of membrane proteins that facilitate their solubilization (see Section "Cell Lysis and the Preparation of Whole Cell Protein Extracts"). The presence of such substances may interfere with further processing and LC-MS analysis; therefore, they have to be removed from the sample before downstream processing, for example, by using solid-phase extraction or precipitation approaches. However, due to the low molecular mass of many reagents and cellular metabolites in comparison to the M_r of proteins, size-exclusion chromatography or ultrafiltration are frequently used to purify protein extracts, especially in spin-column or spin-filter formats, to minimize the time required for sample processing.

Cell Shaving

Surface proteins play a crucial role in the interaction between cells and their environment, and the outermost cell components can be digested for strain identification without previous cell rupturing. In recent years novel approaches have been developed for analysis of such proteins that include, among others, membrane washing, two-phase partitioning, and protein shaving (Zhang et al. 2013a). Protein shaving is based on the direct digestion of live, intact cells under isotonic conditions, so surface-exposed domains of membrane proteins, named the "surfome," are "shaved" by a protease and the released peptides can be analyzed by LC-MS/MS. This way, the problems with attempting to solubilize the entire membrane are avoided. Methods and approaches used in surfomics for fast identification of surface proteins have been reviewed by Olaya-Abril et al. (2014).

Recently, Karlsson et al. (2012) applied a lipid-based immobilization technique in the microfluidic format to immobilize intact cells of *Helicobacter pylori* and to obtain peptides from their surface-exposed outermost proteins by shaving them with a trypsin solution. The released peptides were successfully analyzed for strain-level discrimination of analyzed samples.

Protein Digestion Methods

Protein digestion is usually carried out through hydrolysis of the amide bonds catalyzed by chemical reagents, such as cyanogen bromide cleaving at methionine residue, acid catalyzed hydrolysis at aspartic acid (Fenselau et al. 2011), the cleavage at tryptophan and tyrosine residues induced by electrochemical oxidation (Basile and Hauser 2011), or enzymatically with endopeptidases. There are many proteolytic enzymes differing by their specificity for cleaving bonds between individual amino acid residues in a protein. However, trypsin—a serine protease which cleaves at the carboxyl side of arginine and lysine—is the most commonly used protease for protein digestion in shotgun proteomics. Such cleavage specificity gives tryptic peptides a structure that is particularly amenable to informative fragmentation, following ionization and collisional activation in a mass spectrometer. Nevertheless, a combination of highly selective proteases may improve protein and proteome coverage by creating complementary peptides (Wiśniewski and Mann 2012).

In general, the digestion process has to be optimized to achieve maximum efficiency based on a number of parameters affecting the enzymatic reaction that include: (i) solubilization and denaturation of proteins, (ii) reduction of disulfide bonds, (iii) alkylation of reduced cysteines, and (iv) digestion conditions.

Solubilization and Denaturation of Proteins

Adequate solubilization and proper unfolding of proteins in complex microbial extracts are crucial for providing a protease access to cleavage sites. It is especially important in regard to membrane proteins that comprise approximately a quarter of all open-reading frames (ORFs) in typical bacterial genome. They are usually underrepresented in LC-MS proteomics experiments due to poor solubility and lower abundance in comparison to typical cytoplasm proteins. Therefore, the use of diverse solubilization reagents like urea, detergents, and organic solvents has shown to improve digestion efficiency measured as the number of identified peptides and/or sequence coverage of proteins (Mayne et al. 2014).

Detergents are considered the best protein solubilizers, but sodium dodecyl sulfate (SDS) and other conventionally used surfactants are detrimental for LC-MS analysis and have to be completely removed before analysis. Therefore, surfactant replacement strategies have been developed and are used in many laboratories. The most popular among them are based on filter-aided sample preparation protocols (FASP, Manza et al. 2005; Jabbour et al. 2007, 2010c; Wiśniewski et al. 2009). However, many others methods could be used to remove SDS, for example, ethyl acetate extraction (Yeung et al. 2008), potassium dodecyl sulfate (KDS) precipitation (Zhou et al. 2012), or detergent removal with spin columns (Antharavally et al. 2011; Bereman et al. 2011).

Zhou et al. (2012) compared four in-solution protocols for digestion of whole cell lysates from *Shewanella oneidensis* MR-1. In the first step, they denatured proteins using (1) 8 M urea at 37 °C for 1 h, (2) 50 % trifluoroethanol at 60 °C for 2 h, (3) 1 % SDS at 95 °C for 5 min, and compared them to denaturation with 4 % SDS at 95 °C for 5 min, followed by the FASP protocol that includes SDS exchange by urea prior to sample digestion on a standard ultrafiltration device (Wiśniewski et al. 2009). Samples were then reduced using dithiothreitol (DTT) followed by cysteine alkylation by iodoacetamide (IAA) and after dilution were digested with trypsin. SDS was removed by the KDS precipitation method with KCl. They found only minor differences in sample digestion efficiency among these four methods because LC-MS/MS analyses allowed for identification of more than 4000 peptides from ca. 1000 proteins in each case. This proves that a postdigestion precipitation method could be used as an alternative to predigestion SDS removal by the ultrafiltration-based FASP.

A number of LC- and MS-compatible surfactants, for example, ProteaseMAX, Invitrosol, Rapigest, and PPS Silent Surfactant have also been developed and evaluated to improve protein digestion efficiency. Structures of these commercially available surfactants have an acid labile moiety and, therefore, can be easily degraded prior to LC-MS into components that do not interfere with peptides analysis. For example, Wu et al. (2011) investigated three surfactant-assisted shotgun methods for their applicability to membrane proteome analysis of *E. coli* using acid labile surfactants, sodium 3-[(2-methyl-2-undecyl-1,3-dioxolan-4-yl)-methoxyl]-1-propanesulfonate (RapiGest), PPS, and SDS. They found RapiGest as a preferred reagent for LC-MS/MS analysis of tryptic digests based on the higher number of identified peptides (5799 unique peptides) in comparison to SDS and PPS methods. However, in the study of whole cell protein extracts obtained from *E. coli* cells, Tanca et al. (2013) found that SDS-based buffer outperformed RapiGest in terms of protein extraction yield, and the number of MS identifications and proteome coverage. Therefore, they further tested SDS extracts in five different MS sample preparation workflows, among them, the spin-column detergent removal, followed by in-solution digestion and the FASP method. Although the number of proteins identified among the five tested protocols was comparable (between 1007 and 1104), the FASP dramatically outperformed the competing workflows in the number of identified peptides. For example, with FASP they identified, on average, 7.7 peptides per protein, while the SDS spin-column workflow gave only 4.6 peptides per protein. This indicates the superiority of the FASP procedure for strain identification due to much better proteome coverage.

Waas et al. (2014) investigated the effect of eight commercially available MS-compatible surfactants, two organic solvents, and two chaotropes on the enzymatic digestion efficiency of membrane protein-enriched extract. They found that Progenta^{TM} anionic surfactants—easily cleaved with trifluoroacetic acid (TFA) into small organic molecules that do not exhibit surfactant activity or interfere with analysis by mass spectrometry—outperform other surfactants when tested alone. However, in combinations with guanidine and acetonitrile, all surfactants improved their performance to near similar levels. Nevertheless, the highest number of unique peptides (exceeding 5000) was observed with Invitrosol^{TM}, a proprietary surfactant blend manufactured by ThermoFisher Scientific (Waltham, MA, USA), which does not interfere with protease activity and is compatible with reversed-phase (RP) LC-ESI-MS analysis.

The other group of surfactants proven useful for solubilization and digestion of membrane-bound proteins are volatile surfactants like perfluorooctanoic acid that can be easily evaporated prior to LC-MS analysis. As an alternative to surfactants, trifluoroethanol has proven useful for concurrent protein extraction and denaturation for mass-limited samples where sample cleanup is usually detrimental to sensitivity (Wang et al. 2005; Fleurbaaij et al. 2014).

Reduction of Disulfide Bonds and Alkylation of Reduced Cysteines

Thorough protein digestion requires protease access to as many proteolytic sites as possible and is aided by the inclusion of good protein denaturing agents combined with reduction and blocking of free sulfhydryl groups by the alkylation step. Proteins are usually reduced with DTT and cysteines are alkylated with IAA at room temperature to form carbamidomethylated derivatives. Because IAA is unstable in light, it must be prepared immediately before alkylation of reduced proteins and protease digestion for MS analysis. However, in some cases the overalkylation with IAA may modify lysine, histidine, and N-terminal residues (Boja and Fales 2001). Therefore, to avoid these side effects caused by IAA, some researchers suggest alternate approaches, such as the use of 4-vinylpyridine to alkylate cysteine sulfhydryl groups of proteins after previous reduction with tris(2-carboxyethyl)phosphine (Erde et al. 2014). Generally, the concentrations of reagents are selected in consideration of the enzyme optimal activity and overalkylation side effects, and they are removed before MS analysis by ultrafiltration, solid-phase extraction, or in-line RP chromatography.

Protein Digestion Conditions

Trypsin, a work horse in bottom-up proteomics, is the protease of choice as it has a high specificity and is stable under a wide range of conditions, including 40 % acetonitrile and 2 M urea. However, its cleavage sites are not always predictable due to frequent miscleavages caused by skipping a cleavable residue (Lys or Arg) when the successive Lys/Arg are present, or due to low trypsin digestion efficiency when these residues are followed by Pro. Miscleavages may also occur due to incomplete protein denaturation or post-translational modifications (PTMs) on amino acid residues near protease cleavage sites. In addition, auto-proteolysis can generate pseudotrypsin exhibiting chymotrypsin-like specificity. Hence, the modified trypsin, for example, through dimethylation of lysine residues, is commonly used which has better cleavage specificity and maintains optimal activity at higher temperatures. However, trypsin preparations usually contain some contaminating chymotrypsin; therefore, commercial products known as "sequencing grade" are treated with N-tosyl-phenylalanyl chloromethyl ketone (TPCK) to inhibit chymotrypsin activity. Nevertheless, to avoid auto-digestion, trypsin is used at low concentrations and the reaction is typically carried out at 37 °C for a few hours or even overnight before termination. Therefore, many approaches have been developed focusing on increasing the speed, yield, and robustness of the digestion process through optimization of reaction conditions, immobilization of trypsin on solid supports, or by addition of other proteolytic enzymes. For example, Glatter et al. (2012) found superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin alone to yield fully cleaved peptides while reducing the abundance of miscleaved peptides. The overview of the available techniques and digestion methods for shotgun-proteomics applications can be found in recent literature (e.g., Switzar et al. 2013a; Vuckovic et al. 2013).

Various reagents have been reported as enhancers used to accelerate protease digestion, as well as to improve the digestion efficiency for membrane proteins. For example, Masuda et al. (2008) compared 27 enhancers, including surfactants, organic solvents, and chaotropic agents, and examined their influence on the protease activity of trypsin and protease Lys-C as well as on the solubility of membrane proteins. They found that bile salts, like sodium deoxycholate even at 0.01 % concentration, increased trypsin activity more than fivefold; hence they developed a new protocol based on the use of this surfactant for protein extraction, solubilization, and trypsin activation. Their protocol, which included extraction of cholic acid from the acidified sample with ethyl acetate (phase transfer) before LC-MS analysis, improved substantially the efficiency of protein identification for membrane-enriched fractions of *E. coli* (Masuda et al. 2008).

To shorten the digestion time of *E. coli* protein extracts down to 15 min, Masuda et al. (2009) used immobilized trypsin in a spin-column format. Moreover, they increased the digestion efficiency even further by the presence of sample solubilizers, that is, lauroylsarcosine and deoxycholate that act as a natural trypsin activity-enhancing agent present in bile acids secreted into a small intestine. Overall, by using this approach they identified 1453 proteins, including 545 membranes proteins.

Recently, Erde et al. (2014) used 0.2 % deoxycholic acid to enhance trypsin performance during the FASP digestion of a whole cell protein extract from *E. coli* cells and showed that this modified protocol, referred to as enhanced or "eFASP," increased tryptic digestion efficiency for both cytosolic and membrane proteins.

Modification of protein digestions using physical methods has also contributed to improved digestion efficiency and proteomic coverage. Covalent and dynamic immobilization of trypsin on micro- and nanoparticles, the use of pressure cycling technology, high-intensity ultrasound, and the microwave heating have improved the kinetics of tryptic digestion by reducing digestion time and enhancing the cleavage specificity, especially for hydrophobic and membrane proteins (Vaezzadeh et al. 2010).

Sample Digestion Strategies

The currently available digestion strategies and recent developments in the acceleration of the digestion process allowing for reduction of the digestion time from hours to minutes or even seconds have been reviewed by Switzar et al. (2013a).

In recent years, the FASP method (Manza et al. 2005; Jabbour et al. 2007, 2010c; Wiśniewski et al. 2009) has emerged as a key tool for processing microbial protein extracts for strain identification (Jabbour et al. 2010a, b, 2014; Wade et al. 2010, 2011). It enables the integration of all sample processing steps required for efficient on-filter enzymatic cleavage of proteins and removal of contaminants by using a filtration unit as a "one-pot" proteomics reactor, thereby reducing the risk of sample loss. Nevertheless, these commercially available units should be passivated before the use to avoid peptide losses from low copy number proteins. For example, Erde et al. (2014) found that overnight incubation of both filter units and collection tubes in the passivation solution of a nonionic surfactant Tween-20 increased dramatically the peptide recovery from small samples (up to 300 %).

Although diverse types of ultrafiltration devices are used, generally, the 30 kDa units are best suited for FASP because they retain small proteins ($M_r < 10$ kDa)—due to the large Stokes radii of proteins unfolded in urea that prevents them from passing the filter—and pass more larger peptides (with $M_r > 1500$ Da) than the 10 kDa filters. In addition, the centrifugation time needed to concentrate samples is 3–4 times shorter than that with the 10 kDa units (Wiśniewski et al. 2011).

To increase the proteome coverage, Wiśniewski and Mann (2012) suggested a consecutive sample digestion procedure carried out in a filtration unit proteomic reactor and developed a protocol, enabling consecutive digestion of the sample with two or more enzymes, referred to as multienzyme digestion (MED)-FASP. In this "extended" FASP method, peptides are liberated by centrifugation after each digestion step and the remaining material is subsequently cleaved with the next proteinase. Therefore, orthogonal populations of peptides are created from the same sample that can be jointly or separately analyzed using LC-MS/MS to increase substantially the number of identified peptides in comparison to the single enzyme digestion protocol applied to the same amount of sample. For example, they found that consecutive use of endoproteinase Lys-C and trypsin enabled in some cases to double the number of identified unique peptides (Wiśniewski and Mann 2012). The application of MED-FASP to analysis of *E. coli* ATCC 25922 strain whole cell lysates—by using digestion with endoproteinase LysC, followed by filter washes and trypsin digestion—allowed the identification of 8206 ± 270 unique peptides in the LysC fraction, and 10,728 ± 319 tryptic peptides per sample (Wiśniewski and Rakus 2014).

The FASP protocol for shotgun proteomics of whole cell lysates was also extended to a high-throughput sample preparation procedure based on simultaneous processing of samples in 96-well filter plates (Switzar et al. 2013b). Their protocol enabled all sample preparation steps, including cell lysis, buffer exchange, protein denaturation, reduction, alkylation, and proteolytic digestion to be carried out for a large number of samples. The protocol would be suitable for diagnostic analysis, for example, in a clinical laboratory or for processing large numbers of fractions resulting from prefractionation of microbial proteomes in a research lab. They pointed out that the usage of a single plate for all sample preparation steps following cell lysis reduces potential samples losses, increases sensitivity, and allows for automation.

Yu et al. (2012) combined FASP, used for an efficient depletion of detergents, with the ultrafast and efficient microwave-assisted on-filter enzymatic digestion by transferring proteins mixed with trypsin on filter units to a microwave oven where they were digested for less than 1 min. Also, Chang et al. (2013) used the FASP method for processing the *Acinetobacter baumannii* whole cell protein extract, followed by a 15-min-long microwave-assisted protein digestion with trypsin.

However, according to Reddy et al. (2013) the faster reaction rate is not caused by the microwave quantum effect but the thermal one. Therefore, both microwave and conventional heating at high temperatures (50 °C) can be used to accelerate digestion reactions. For example, Tracz et al. (2013) trypsin-digested whole cell protein extracts from pathogenic strains of *Yersinia, Francisella,* and *Bacillus* at 53 °C for a couple of hours and used thousands of released and confidently identified peptides for successful bacterial identifications.

Liquid Phase Separation and Ionization of Peptides Followed by Acquisition of Tandem Mass Spectra

In classical bottom-up methods, separated proteins are in-gel trypsinized, and the released peptides are identified by mass mapping or by analyzing product ion mass spectra obtained through the collision-induced dissociation or postsource decay (Chalmers and Gaskel 2000). In the shotgun approach, peptides are released during proteome-wide digestion of microbial proteins with proteolytic enzymes and in some applications they are directly analyzed using MALDI time-of-flight (TOF) MS for peptides mass fingerprinting (PMF) of microbes, or peptides from dominating proteins are sequenced using MALDI-MS/MS technologies. For example, Warscheid and Fenselau (2003) investigated the PMF concept for analysis of small acid-soluble proteins in *Bacillus* species by on-probe shotgun trypsin digestion of spores from this genus. The released peptides were also identified by tandem-MS techniques for distinguishing *B. cereus, B. thuringiensis, B. subtilis, B. globigii,* and *B. anthracis* Sterne strains. More recently, Balážová et al. (2014) demonstrated that microwave-accelerated shotgun tryptic digestion of cellular material combined with MALDI-TOF MS profiling of released peptides allowed for subspecies differentiation of *Staphylococcus* and *Bacillus* strains. However, substantial improvements in the scope of sequence coverage and reliability can be achieved through separation of peptides by LC or capillary electrophoresis (CE) prior to ESI-MS/MS analysis (Wolters et al. 2001).

Liquid Chromatography-ESI-MS

Several strategies have been developed to fractionate peptides prior to MS analysis that include separation based on one-dimensional (1D) nano-LC and multidimensional separation systems. In the former approach, the resolution of peptide separation can be increased through the use of RP columns with smaller particle sizes, for example, below 2 µm in diameter, and submicroliter flow rates (Fröhlich and Arnold 2009). This technique gives higher efficiency but requires higher pressure separation, and is therefore referred to as ultrahigh-pressure liquid chromatography (UPLC). However, nanospray is more sensitive than approaches using higher flow rates because electrospray is a concentration-sensitive process. Consequently, the use of a narrower column and lower flow rates will cause the elution of peptides as narrower peaks with higher maximal concentrations. In addition, the use of longer columns operated at higher temperatures may increase both high-resolution and high-peak capacity separations even further.

For example, Hebert et al. (2014) identified more than 34,000 peptides with unique sequences over a 70-min run by using a 35-cm long RP column with 75 µm internal diameter. This column was packed with 1.7 µm C18 particles and was operated at 60 °C by using the mobile phase containing 5 % of dimethyl sulfoxide, in addition to the standard components, that is, formic acid/water and formic acid/acetonitrile. Eluting peptide cations were electrospray ionized and analyzed on a hybrid mass spectrometer (quadrupole-orbitrap-quadrupole-ion trap, Q-OT-qIT; Orbitrap Fusion, Thermo Scientific, San Jose, CA, USA).

The most popular multidimensional separation systems use: (i) a combination of peptide separation according to their isoelectric point by isoelectric focusing on immobilized pH gradient, followed by RP LC separation according to their hydrophobic properties and MS/MS analysis (Vaezzadeh et al. 2010; Geiser et al. 2011b), (ii) off-gel electrophoresis and RP LC-MS/MS (Geiser et al. 2011a), or (iii) multidimensional liquid chromatography (MDLC). In the latter group of methods, the most commonly used is the multidimensional protein identification technology, termed Mud-PIT, which was introduced by Washburn et al. (2001). Mud-PIT consists of two orthogonal separation systems—strong cation exchange (SCX) and RP—coupled online in an automated fashion and offering the possibility to analyze highly complex peptide mixtures in a single experiment. Most commonly, and as originally published, an RP-precolumn is followed by an SCX-precolumn, and finally the main RP-separation column; thus forming a triphasic column packed into an ESI-emitter tip directly coupled to a mass spectrometer; however, there are many variations of this basic format (Lohrig and Wolters 2009). Recently, a detailed protocol has been described for the construction of a simple and flexible online

RP-SCX-RP LC system and its implementation for deep proteome profiling on a common shotgun-proteomics platform (Lam et al. 2014).

In addition to the online MDLC format, offline approaches are quite popular and each of them has its advantages and disadvantages, that is, reduced labor time in case of online separation and the flexibility of offline fraction collection. However, online methods are not optimal for peptide separation due to the elution of peptides with a solvent step gradient during ion-exchange chromatography. Therefore, offline techniques based on a continuous gradient ion-exchange separation of peptides, which are subsequently analyzed by RP LC coupled with ESI-MS/MS, represent a better choice for the comprehensive analysis of the bacterial proteome. By using this approach, Jaffe et al. (2004) found almost 10,000 unique tryptic peptides corresponding to 81 % of the predicted ORFs for a small, wall-less bacterium *Mycoplasma pneumoniae*.

Although very high proteome coverage can be achieved, in the past it usually required a long data acquisition time. For example, Hendrickson et al. (2010) reported detection of 1671 proteins representing 64 % of all genome predicted proteins of *Methylobacillus flagellatus*. However, they achieved it by analyzing five prefractions, resolved by using 2D capillary high-performance LC (HPLC) analysis that consisted of a seven-part step gradient from the cation-exchange portion of the biphasic column, followed by the reverse phase elution and MS analysis. This gave a total of 35 separate HPLC runs per technical replicate with 60 min effective acquisition time per run.

Capillary Electrophoresis-ESI-MS

ESI-MS/MS allows for online detection and identification of peptides separated by CE (Janini et al. 2003). This approach is rarely used for microbial identification purposes; however, Hu et al. (2005 and 2006) described a successful application of this technique for identification of microbial mixtures using a quadrupole ion trap operated in a selective tandem-MS mode. They trypsin-digested bacterial proteins and analyzed released peptides with CE-MS/MS by targeting species-unique tryptic peptide ions. For that purpose they first created a small DB of proteotypic tryptic peptides derived from abundant proteins that are species-specific biomarkers for targeted strains. Isolated ions of such peptides were analyzed by using a selective reaction monitoring approach. The overall identification success for this method was 97 % on the basis of analysis of 34 clinical samples with a total analysis time of 8 h that included a 6-h long cultivation step. Moreover, they shortened the time-consuming digestion process to 15 min by the application of microwave-assisted proteolysis (Lin et al. 2005).

Recently, Fleurbaaij et al. (2014) developed a CE-ESI-MS/MS bottom-up proteomics workflow for sensitive and specific peptide analysis with the emphasis on the identification of β-lactamases in various Gram-negative bacterial species even from single colonies. They demonstrated the ability of the system to successfully assess multidrug-resistant bacterial clinical isolates.

Liquid Chromatography MALDI-MS/MS

The separation of complex peptide mixtures using LC columns is usually coupled to mass spectrometric analysis by electrospraying column effluent directly into the mass spectrometer. However, peptides separated by nanoscale LC may be coupled to a collector that deposits microfractions onto a MALDI plate, thus allowing for the MALDI-MS/MS analysis of the fractions by instruments with TOF/TOF ion optics or/and LTQ-Orbitraps (Yang et al. 2007; Baeumlisberger et al. 2011). For example, Lasaosa et al. (2009) found that the MALDI-based platform led to a significantly increased number of peptides identified from a tryptic digest of the cytosolic proteome of the bacterium *Corynebacterium glutamicum;* probably due to the fact that the size of the unique peptides identified by MALDI was, on average, 25 % larger and more hydrophilic than the unique peptides identified by ESI (Yang et al. 2007).

Generally, there are several benefits associated with the LC-MALDI-MS/MS approach. First, the collection of MS/MS data is decoupled from the chromatographic separation, so the sample can be reanalyzed using optimized MS/MS parameters. Second, the relative insensitivity to interfering compounds in the sample matrix and/or mobile phases allows carrying the chromatography under optimized conditions. Third, this approach provides the ability to archive the sample plate (Fernández-Puente et al. 2014).

In conclusion, nano-LC combined with further improvements in MS sensitivity and speed will continue to reduce whole proteome analysis time for microbial strains by producing tens of thousands of peptide sequence-to-spectrum matches (PSMs) in less than 1 h (Hebert et al. 2014). However, LC-MALDI-MS/MS analyses may be better suited for specific applications requiring sample archiving.

Database (DB) Construction and Searching

The prevailing approach for peptide, protein, and microbial strain identification in shotgun proteomics is based on decoding amino acid sequences by using combined information of the tryptic peptide mass and its fragmentation spectrum matched against DB sequences. Therefore, the success of identifying any ionizable peptide depends on the availability of suitable DB reference sequences, and by no means can it be assumed that sets of reference genomes/proteomes available in the public DBs are complete or fully representative for any isolated strain. Therefore, the use of an appropriate DB is crucial for subspecies typing and identification of strains.

Bacterial DBs

The construction of protein sequence DB plays a crucial role in proteomic workflows; however, the DB should contain all possible sequences while on the other hand, if the DB is too large, the search engine may introduce false positive identifications (Vaudel et al. 2014).

There are almost 15,000 bacterial strains with sequenced genomes, including 4000 with complete genome sequences, available in public DBs, as of fall of 2014, and chromosome and plasmid-encoded protein sequences predicted from these genomes can be downloaded from the National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov/genomes/Bacteria) or from the Universal Protein Resource (UniProt) Knowledgebase (UniProtKB; www.uniprot.org). However, microbial proteomes in these DBs vary greatly in terms of their curation, completeness, and comprehensiveness; hence, the use of most recent versions translated from complete genome sequencing projects is strongly recommended. Amino acid sequences in these DBs represent a translation of nucleotide sequences in computationally determined ORFs that potentially encode proteins. ORF begins with an initiation codon and ends with a stop codon and has the potential to encode a single polypeptide expressed as a protein; however, many may not actually do so. In addition, different bioinformatics approaches for automatic annotation of genes are currently used and this affects the quality of protein lists used in proteomics. For example, different annotation tools may predict different translational start sites (TSS) for ORFs that will affect the N-terminal peptides generated during in silico digestion (de Souza et al. 2010 and 2011; Armengaud et al. 2013). Furthermore, a protein should be understood as one of many isoforms representing the expressed gene and may differ from a polypeptide specified by a nucleotide sequence due to co-translational modifications or PTMs of a nascent polypeptide. Co-translational modification refers to the removal of N-terminal methionine by N-methionyl aminopeptidase and affects the majority of bacterial proteins. PTMs comprise both the proteolytic processing of a polypeptide, for example, to generate appropriate targeting signals, and covalent modifications of its amino acids (Hesketh et al. 2002; Bonissone et al. 2013; Zhang et al. 2013b). Therefore, the available DB searching algorithms, in fact, identify ORFs, not proteins. Moreover, during analysis of an unknown microbial strain the confirmation of the full amino acid sequence or “100 % coverage” of a potential protein would be required for the identification of an ORF, because sequences of orthologous proteins from a closely related strain may only differ due to an SAV. Consequently, the true identification of proteins is rarely achievable during high-throughput analyses of microbial proteomes.

In the early studies on identification of bacteria using shotgun proteomics, Dworzanski et al. (2004) constructed a prototype proteome DB from genome sequences downloaded from the NCBI site. They used a computational Gene Locator and Interpolated Markov Modeler (Glimmer) developed by Salzberg et al. (1998) to identify protein-coding ORFs and translated them into amino acid sequences of all putative proteins. All these sequences were used for assembling a microbial proteome DB in a FASTA format.

A sequence in FASTA format begins with a single-line description distinguished from the sequence data by a greater-than (">") symbol and ends with a carriage return. Although the description is generally considered as a free form, software applications such as search engines assume that the first word or string after the ">" symbol is a real sequence identifier and use it for processing while the remainder of the line is a supplementary description. Therefore, Dworzanski et al. (2004) modified header lines of each protein in a DB, by using a header replacer script written in Perl, and added abbreviated strain names in header lines, so the search engine was recognizing and assigning PSMs directly to reference DB strains instead to particular proteins in each proteome. Consequently, the search engine SEQUEST was recognizing each proteome as a single "pseudo-polyprotein" and could be used for ranking all peptide-to-strain matches while retaining complete information about protein sources with each peptide. Although the above DB could be searched directly, the search efficiency may be substantially improved by in silico digestion of all sequences to create an indexed peptide sequence DB derived from all DB proteomes.

Recently, Tracz et al. (2013) described a similar approach by tricking Mascot to assign PSMs directly to reference DB strains instead of proteins. They achieved it by creating a custom database, named "Genome AA," containing protein sequences deduced from 2,026 completed bacterial genomes available from the NCBI Reference Sequence (RefSeq) DB. However, each entry in the GenomeAA DB consisted of the strain name followed by a "pseudo-polyprotein" created by concatenation of all individual protein sequences separated only with the letter code J. Therefore, to preserve the integrity of peptide termini, trypsin digestion rules used by the search engine were always supplemented with information to cleave on the C- and N-terminal sides of the letter code "J." Consequently, Mascot searches against this DB report PSMs to reference strains represented by DB proteomes, instead of particular proteins.

In proteogenomic studies, six-frame translated nucleotide sequences from investigated genomes are used (Armengaud et al. 2013). However, DBs used for strain identification are usually downloaded from NCBI or UniProKB as FASTA formatted protein sequences. Nevertheless, they may be additionally cured. For example, Dworzanski et al. (2006, 2010), Jabbour et al. (2010a, b, c,) and Deshpande et al. (2011) continued to create prototype microbial DBs by adding abbreviated strain names to header lines for each downloaded protein, as described above. These abbreviated strain names were also used as specific codes that linked strains to taxonomic information derived from the NCBI taxonomy DB (http://www.ncbi.nlm.nih.gov/Taxonomy/). Finally, they indexed the DB by performing in silico digestion of proteins using a TurboSEQUEST utility program (Thermo Scientific) by assuming (trypsin) endoprotease digestion rules and allowing up to two missed cleavages per peptide; however, only peptides with M_r in the 700–3500 Da range were accepted.

It is also important to append any protein DB with sequences of common laboratory contaminants. For instance, the following FASTA formatted DBs of contaminants are available via the Internet: (i) the common Repository of Adventitious Proteins, cRAP, can be downloaded from the Global Proteome Machine FTP site (ftp://ftp.thegpm.org/fasta/cRAP) or (ii) a contaminants.fasta file containing common contaminants is available at http://maxquant.org/downloads.htm and could be appended to any target DB and used as a control for environmental and common laboratory contaminants.

DBs of Virulence Factors, Toxins, and Antibiotic Resistance Determinants

In addition to the identification of bacteria, it is also helpful to subtype isolated strains in regard to their functional capabilities such as virulence, antibiotic resistance, or production of toxins which are of high epidemiological, clinical, and agricultural or biosecurity importance. However, the search engines usually disregard this type of information contained in well-annotated DBs or it is difficult to retrieve it in an easy-to-interpret format. Therefore, based on inputs from publicly available sequences, it is advantageous to create customized DBs that are configured to facilitate subtyping of strains based on the presence of sequences associated with specific factors, for example, responsible for virulence or antibiotic resistance. Although some researchers prefer to create their own DBs customized for specific needs, there are also a few well-annotated sequence DBs targeting virulence and antibiotic resistance proteins which are available for downloading via the Internet (Chen et al. 2012; Winnenburg et al. 2008; Gupta et al. 2014).

Virulence factors (VFs) help pathogens to evade host-specific defensive mechanisms to establish infection. They include bacterial toxins, secreted effectors, for example, hydrolytic enzymes that may contribute to the pathogenicity of the bacterium, cell surface proteins that mediate bacterial attachment, and cell surface carbohydrates and proteins that protect a bacterium, among others. There are a few DBs available with protein sequences of such VFs. For example, a DB of protein VFs (VFDB) in the FASTA format was compiled based on information from more than 2000 related publications and can be downloaded from the http://www.mgc.ac.cn/VFs/main.htm Website (Chen et al. 2012). It contains sequences of 460 VFs, 24 pathogenicity islands and ca. 2500 VF-related proteins (as of November, 2014) gathered from 429 chromosomes and 93 plasmids of pathogenic bacterial strains belonging to 26 bacterial genera. The other DB of VFs, named “Victors Virulence Factors” DB currently includes 5173 VFs from strains of 125 microbial species known as pathogenic to humans and animals (50 bacterial species, 54 viruses, 13 parasites, and 8 fungi). A FASTA file with protein sequences of all these VFs is available for download from the http://www.phidias.us/victors/download.php website. The data within Victors are manually curated and comes from peer-reviewed literature and existing DBs (e.g., NCBI RefSeq). The other DB with protein sequences available for download is known as a pathogen–host interaction DB (PHI-base) (Winnenburg et al. 2008) that contains curated information on genes proven to affect the outcome of PHIs. It catalogs experimentally verified pathogenicity, virulence, and effector genes from fungal, fungus-like eukaryotic microorganisms (*Oomycete*), and bacterial pathogens infecting animal, plant, and insect hosts. PHI-base is therefore an invaluable resource in the discovery of these genes in medically and agronomically important pathogens. PHI-base contains 3012 entries with protein sequences translated from so-called pathogenicity genes (if the effect on the phenotype is qualitative) or virulence/aggressiveness genes (if the effect is quantitative) or effector genes (either activate or suppress plant defense responses) and can be downloaded in the FASTA format at http://www.phi-base.org/.

Antibiotic resistance (AR) Gene-ANNOTation (ARG-ANNOT) DB was developed by Gupta et al. (2014) and consists of a single file with amino acid sequences of existing and putative antibiotic resistance-associated proteins in a FASTA format that can be downloaded from http://www.mediterranee-infection.com/article.php?laref=282titre=arg-annot. They collected information about 1689 AR-associated genes from published works and online resources, and sequences of these gene products were retrieved from the NCBI GenBank DB. AR-associated proteins in ARG-ANNOT DB are linked to diverse antibiotics classes, including aminoglycosides, beta-lactamases, fosfomycin, fluoroquinolones, glycopeptides, macrolide-lincosamide-streptogramin, phenicols, rifampicin, sulfonamides, tetracyclines, and trimethoprim. There are also other available DBs like the Antibiotic Resistance Genes DB (ARDB, Liu and Pop 2009) or MvirDB—a microbial DB of protein toxins, virulence factors, and AR genes for bio-defense applications—that integrates DNA and protein sequence information from other sources (Zhou et al. 2007), however, they were not recently updated.

A number of web-services are available for identification of known or predicted bacterial toxins, for example, BTXpred, which makes available a FASTA formatted file of 185 bacterial toxins (http://www.imtech.res.in/raghava/btxpred/supplementary.html); and a DB of Bacterial ExoToxins for Human (DBETH, http://www.hpppi.iicb.res.in/btox/) with FASTA files of "Human Pathogenic Bacterial Exotoxin Fasta Sequences" and "Human Pathogenic Bacterial Exotoxin Homologs."

Chang et al. (2013) created the β-lactam-resistance protein DB of *A. baumannii* (abbreviated as "BRPDAB") and used it to develop an accurate and rapid shotgun-proteomics method for the identification of β-lactam-resistant *A. baumannii* pathogens. They used a serious of gene ontology (go) terms (Ashburner et al. 2000) such as beta-lactamase activity (go:0008800), penicillin binding (go:008658) or response to antibiotics (go:0046677 used as a synonym to antibiotic susceptibility/resistance), names of all β-lactam antibiotics and the name of a bacterium "*A. baumannii*" to identify in the Uniprot DBs proteins associated with the resistance of this pathogens to antibiotics. They downloaded these sequences and incorporated them into the FASTA formatted BRPDAB.

Creation/Correction of Microbial Protein DBs Through Re-sequencing and Analysis of Genomes

Despite the availability of thousands of completely sequenced genomes, many species are still represented in sequence DBs by only a single or a few strains. Therefore, for analysis of strains from DB underrepresented species, improved DBs are needed to compensate for the missing sequence variations reflecting intraspecies strain diversity that may affect the identification of organisms at the subspecies level. Such DBs could be constructed by the "maturation" of sequences, for example, by N-methionine excision, removal of N-terminal signal peptides based on annotations in the DB, or cleavage site predictions determined with the help of suitable algorithms, such as SignalP (Petersen et al. 2011) or Phobius (Käll et al. 2007b). Additional DB improvements can be achieved by correcting some sequencing errors such as incorrect predictions of TSS during an in silico-driven annotation process to make the N termini of homologs as consistent as possible within the DBs (Sato and Tajima 2012).

For example, the identification of protein variants could be improved by using the multistrain MS prokaryotic DB builder (MSMSpdbb) (de Souza et al. 2010). In this approach, a combined protein DB of closely related microorganisms is created that provides two important advantages. First, it allows for streamlining the initial DB searching by combining groups of phylogenetically close organisms, and second, it provides protein annotation improvements by correcting sequence TSS which are frequently incorrectly annotated, especially for older submissions. Recently, Bland et al. (2014) characterized 534 N termini of the marine bacterium *Roseobacter denitrificans* and found that 10 % of them were incorrectly annotated in regard to TSS. They also found five previously un-annotated proteins and eight proteins with multiple translational starts, thus showing the value of empirical evaluation of every sequenced organism for maximum annotation accuracy (Bland et al. 2014).

However, the most reliable solution to overcome the problem of relatively large sequence deviation of an unknown isolate from reference strains should be based on de novo sequencing on protein or nucleic acid levels or ultimately by performing whole-genome sequencing of additional strains from the underrepresented species. Unfortunately, de novo peptide sequencing is still impractical; therefore, both mRNA (RNA-seq, Wang et al. 2012) and genomic DNA sequencing have been used to generate customized DBs for MS identifications in proteomics studies. Because the RNA-seq approach is more appropriate for metaproteomic approaches, DNA sequencing of strains from species underrepresented in public DBs is better suited for expanding the potential sequence variation repertoire for high-resolution discrimination of unknown strains. Furthermore, the Food and Drug administration (FDA) authorization for the first next-generation sequencer, Illumina's MiSeqDx (Collins and Hamburg 2013), will allow not only the development of new human genome-based tests but will also open the way to high-throughput sequencing (HTS) being used in clinical microbiology.

There are two major approaches that have been used: de novo assembly from raw sequence reads and the reference-guided assembly if the closest reference genome is available. However, HTS technologies are error prone; for instance, the Illumina reversible dye-terminator sequencing technology (HiSeq) caused substitutions (Meacham et al. 2011) while ion semiconductor sequencing technology (Ion Torrent, Life Technologies) produced indel errors associated with homopolymer regions (Loman et al. 2012). In addition, despite many computational advances, the complete and accurate genome assembly from second-generation short-read data remains a major challenge. Therefore, instead of de novo genome assembly, the better strategy for proteomics would be re-sequencing based on mapping reads to the whole genome sequence of a strain from the same species followed by searches for single nucleotide variations (SNVs) (Caboche et al. 2014).

Recently, Wu et al. (2014) described a very efficient strategy to overcome sequence variations between the reference genome and the closely related species based on mapping sequencing reads utilizing the error-tolerant FANSe mapping algorithm (Zhang et al. 2012). FANSe corrects the SNVs for the genome deviating from the reference genome ~5 %, and exports them as corrected proteome sequences that can be used in searching peptide fragmentation spectra, and thus efficiently improves peptide and protein identification in nonmodel bacteria without complete genomic sequence. FANSe is a seed-based algorithm which uses the entire information from a sequencing read divided into small seeds of 6–8 nucleotides and aligns all of them to the reference genome sequence. The adjacent seeds mapped to the same segment are combined if they fulfill certain criteria and are used to define so-called hotspots. The alignment for each hotspot is scored and refined based on the least number of mismatches. Consequently, by reducing the number of hotspots FANSe achieves the increased sensitivity, that is, the proportion of actual positives which are correctly identified as such while maintaining a reasonable speed.

For example, sequencing of 1350 bp of 16S rDNA of an environmental isolate, Wu et al. (2014) found 100 % identity to the reference sequence of *Bacillus pumilus* SAFR-032; the only *B. pumilus* strain with complete genome sequence in public DBs. However, 16S rDNA sequence may not distinguish separate strains; therefore, they decided to re-sequence the whole genome and identified 158,407 SNVs. Among these SNVs, 143,263 were identified as substitutions, 221 insertions, and 349 as deletions in protein-coding sequences (CDS). In total, 4.93 % of the mappable region was different in comparison to the reference genome of *B. pumilus* SAFR-032, that is, in the expected range of differences between the same species strains (Goris et al. 2007). This correction allowed them to identify 14.2 % more tryptic peptides from the isolate and they will use this corrected proteome as a reference for the identification and discrimination of other strains from *B. pumilus*. In conclusion, this approach is suitable for the preparation of a set of reference proteomes for DB searching of MS/MS fragment ions derived from the unknown strain proteome for subspecies identification and strain typing.

Finally, it should be remembered that during proteomic analyses, only a fraction of genome predicted proteins and proteotypic peptides are identified and there are a number of reasons why this happens. First, peptides from undetected proteins may fall into a category of false negatives due to bioanalytical factors inherent to bottom-up proteomics: post-translationally modified peptides, peptides too short, too long or from small and low-abundance proteins are difficult to observe. Second, some predicted proteins are not real, due to incorrect genome interpretation including annotations marked as putative or hypothetical. For example, Hendrickson et al. (2010) noted that many of the nondetectable proteins of *M. flagellatus* may represent artifacts of genome annotation while a portion of the nonexpressed proteins appear to correspond to silent genomic islands. Third, some proteins must be true negatives, that is, they are not expressed under the growth conditions used because the expression of many genes is tightly regulated and/or inducible only under specific conditions. For example, Ansong et al. (2009) showed that as much as a third of the *Salmonella enterica* serovar Typhimurium strain's proteome has to be regulated at the translational level by the single virulence regulator Hfq. Nevertheless, nearly 40 % of predicted proteome was covered by peptide identifications in this work.

Custom DBs of E. coli and Salmonella Flagellins

E. coli bacteria are short rods with flagella that rotate to allow movement in liquid environments. The flagellar filament is the largest portion of the flagellum and consists of repeating subunits of the protein flagellin that induces immune responses. These immune responses have been widely utilized for serological typing of *E. coli* strains, which produce 53 distinct sequence types of flagellar H antigens. Recently, Cheng et al. (2013) developed an MS-based typing method of flagellar H antigens (MS-H). For this purpose, they constructed a FASTA-formatted DB of *E. coli* H types using the sequences and serotype information found in the NCBI nr protein DB. In this *E. coli* flagellin DB redundant sequences were collapsed into a single entry with the H-type listed in each sequence description headerline. If the H-type was not specified in the NCBI nr DB, they compared it against sequences with known H serotypes and assigned the top-scoring one. In some cases, the H-type was manually assigned (based on literature search) to sequences with missing H-type in NCBI annotation, or with incorrect H-type listed in the NCBI entry. Incorrect H-types were also discovered by finding outliers in a phylogenetic analysis of all *E. coli* flagellin sequences in the DB.

The final curated *E. coli* flagellin DB can be downloaded at http://www.biomedcentral.com/1756-0500/7/444 as a FASTA file (KC_Flagellin_20130425.fasta). This DB contains 195 unique sequence entries representing all 53 known *E. coli* H serotypes, that is, averaging close to 4 sequences per serotype. However, some serotypes were represented by only one entry (H4; H15; H23; H24; H30; H32; H39; HH43; H51; and H56) while the most common types, such as H6, H11, and H7 were represented by 10, 12, and 16 flagellin sequences, respectively (Cheng et al. 2014a).

For typing flagellin H-antigens of multiphasic *Salmonella* reference strains, Cheng et al. (2014b) created also a curated *Salmonella* flagellum DB containing 385 entries of flagellin sequences available in the literature and the NCBI nr DB. However, this DB is not available for downloading.

Search Engines

Mass spectrometric analysis of peptides released by shotgun digestion of microbial proteins generates high-resolution and high accuracy data sets of product ion spectra that can be used for decoding their amino acid sequences by three classes of approaches. First, by spectral library searches which compare the acquired spectra with a library of previously identified spectra; second, by de novo sequencing to infer the sequence directly from the mass differences of fragment ions in the spectra; and third, by DB searches which compare how well an acquired spectrum matches to a theoretical spectrum of a peptide deduced from protein sequence in the DB (Cottrell 2011; Ma and Johnson 2012). In the latter case, the search engine constructs a theoretical spectrum for each candidate peptide sequence and compares them to experimentally observed fragment ion spectra.

Although de novo approaches would be the best for decoding sequence information from peptide fragmentation spectra, they show sufficient reliability to infer only short sequence tags and thus currently cannot provide a full solution to the identification problem. Nevertheless, they are used in so-called error-tolerant searches that relax the specificity, for instance, by removing molecular mass constraint and thus allowing for matches to DB sequences when there are sequence variations due to mutations or PTMs.

Therefore, the most popular approach to interpret such MS/MS spectra in a high-throughput manner uses DB searches with software tools known as “search engines” to find the best PSMs. As input, a search engine takes MS/MS spectra and searches them against reference proteomes of strains that are expected to be related to the sample with a twofold purpose: first, to find PSMs which confidently decode tandem mass spectra; and second, to quantify the contribution of DB reference microbes to the decoded spectral data set.

There are many well-established software applications for searches with uninterpreted fragmentation spectra against DB proteomes that include SEQUEST (Eng et al. 1994), Mascot (Perkins et al. 1999), X!Tandem (Craig and Beavis 2004), MyriMatch (Tabb et al. 2007), OMSSA (Geer et al. 2004), and Andromeda (Cox et al. 2011) among many others listed in the review article by Nesvizhskii (2010). In addition, due to different approaches used in search engine algorithms, one can maximize the number of peptide identifications by using multiple search engines and combining the results. For example, the PSM gains (at 1 % error level) observed by starting with Mascot and adding SEQUEST search results may exceed 38 %, and by adding MyriMatch and X!Tandem to the combination, the gain can reach 53 % (Shteynberg et al. 2013). These outcomes were obtained by modeling each search result with PeptideProphet (Keller et al. 2002) and combining them with the iProphet tool (Shteynberg et al. 2011) that uses linear discriminant analysis to obtain more accurate PSM scores. Nevertheless, to maximize the number of correct PSMs it is important to run DB search engines using appropriate search parameters.

Setting Search Parameters

The search parameters include, among others, ions' mass tolerances appropriate for the type of instrument used and expected peptide modifications. For example, 5-ppm precursor mass tolerances for a high-resolution mass spectrometer and 0.5-Da fragment tolerance for the ion trap fragmentation. The expected peptide modifications include: (i) static (fixed) that apply to all amino acid residues in a sample, for example, cysteine modification due to the alkylation step, and (ii) dynamic (variable) which may or may not be present at each amino acid site.

Most variable modifications of amino acids are dependent on the sample processing and may include the oxidation of methionine and tryptophan; deamidation of asparagine and glutamine to their acidic counterparts, aspartate and glutamate (Yang and Zubarev 2010); carbamylation of free amino groups; and diverse modifications of N-terminal amino group. For example, the common artifact of using gel electrophoresis during the sample preparation is formation of cysteine propionamide (C[+71]). Furthermore, cysteine residues are usually carbamidomethylated (C[+57)]) by treatment with IAA to block free sulfhydryl groups. In addition, overalkylation with IAA frequently also gives modified lysine [K+(57)], histidine [H+(57)], and N-terminal residues (+57) and (+114) and may affect even substantial fraction of peptides (Boja and Fales 2001).

Commonly used protocols include urea for the solubilization and *denaturation* of proteins. However, urea in solution is in equilibrium with ammonium cyanate which decomposes to isocyanic acid reacting with protein primary amino groups and resulting in their carbamylation (Lippincott and Apostol 1999). This modification (~NH-CO-NH$_2$) gives a mass increment of 43 Da per modified amino group. Therefore, long-term exposure of proteins to high urea concentrations can lead to unfavorable heterogeneity in downstream MS analyses due to carbamylation of lysine, arginine, and N-terminal residues. Consequently, a variable modification for carbamylation of arginine and lysine residues should be taken into account whenever urea is used for sample processing. To minimize the extent of carbamylation, urea solutions should always be used fresh and all operations performed at a temperature below 30 °C. In addition, it is recommended to add methylamine to the urea solution prior to use. However, one should also investigate whether replacing urea by other chaotropic agents, such as sodium deoxycholate or surfactants is appropriate (Proc et al. 2010).

There are also other common biological modifications that should be taken into account. For example, although N-terminal acetylation is rare in bacteria, acetylated N termini are common in archaea and may affect even 15 % of their proteins (Falb et al. 2006). Therefore, the use of appropriate data-mining procedures may increase the number of identified peptides. In addition, it should be remembered that not all peptides in a sample are represented in the DB, while even spectra derived from non-peptide background constituents can be matched to peptides by a search engine.

In conclusion, the setting of proper search parameters is not trivial because taking into account the above-mentioned modifications will enlarge the search space and thus may prolong the search time substantially. Therefore, the best solution would be to estimate the prevalence of known modifications before setting the parameters for a conventional search engine. For example, the software tool "Preview" (Kil et al. 2011) performs a fast full protein DB searches with a set of product ion spectra in a fraction of time needed by a conventional search engine. It reports: (i) the amount and type of nonspecific digestion, (ii) assays the prevalence of known modifications, and (iii) recognized modifications. Such information not only allows choosing the most appropriate search parameters to maximize the number of correct matches, but also provides timely feedback for the laboratory on sample preparation artifacts, thus improving the overall efficiency and reproducibility of the shotgun-proteomic approach.

DB Searches

It is crucial that matches to all reference proteomes are reported instead of a subset of best hits as commonly done by many search engines. Therefore, the acquired fragmentation spectra could be searched separately against each reference proteome or subsets of combined proteomes, depending on the reporting capabilities of the search engine or experimental needs. For example, SEQUEST, the first software developed for searching MS/MS spectra against sequence DBs and commercially available from Thermo Scientific (San Jose, CA, USA) and Sage-N Research (Milpitas, CA, USA), reports up to 100 matches. Hence, depending on the preliminary information about the sample, the searches should be arranged appropriately. However, searches against a large DB increase the error rate and contribute to the increased rate for false-positive identifications (Cargile et al. 2004). Therefore, there are advantages of using a two-step matching process by performing the initial search against a large DB, followed by a focused DB search against DB strain proteomes with a statistically significant number of matches assigned during the initial search (Jagtap et al. 2013). Moreover, during an initial search to produce a focused DB—it is best to enable only the most common modifications (e.g., oxidized methionine and deamidated asparagine).

In the second step, a smaller and better manageable DB may be used, which may be generated by selecting as the target a set of microbial proteomes representing only one phylum, family, or even genus, and appending these forward sequences with decoy DB sequences. In addition, the smaller DB should include sequences of commonly found contaminants (see Section "Bacterial DBs") and could be searched with an extended list of expected peptide modifications.

Processing of DB Search Results

It is important to remember that an identified peptide may be a false positive regardless of its uniqueness. Moreover, a peptide that is unique throughout the protein sequence DB may be the result of a sequencing error. Therefore, quality assessments of PSMs have to be based on a solid statistical ground by using postprocessors such as PeptideProhet (Keller et al. 2002), Percolator (Käll et al. 2007a; http://www.matrixscience.com/help/percolator_help.html), or q-ranker (Spivak et al. 2009) to apply an optimal scoring function for a particular data set (Granholm and Käll 2011).

DB search algorithms attempt to match every experimental spectrum to DB peptides and report parameters to determine correctness of each PSM. For example, the information contained in each output file generated by a search engine SEQUEST includes: (i) PSMs, (ii) peptide assignments to reference microbial proteins or (if headers were appropriately modified) proteomes in the DB, referred here as "peptide-to-bacterial" (PTB) strain assignments, and (iii) parameters estimating the correctness of PSMs (Xcorr, ΔCn, Sp, RSp, ΔM). However, a better way to express the accuracy of such assignments would be to calculate probabilities that each PSM is correct. For example, Dworzanski et al. (2004) interpreted the above SEQUEST matching parameters using discriminant function (DF) analysis. They arrived at probability scores for PSMs by modeling distributions of correctly and incorrectly identified peptides from a training data set obtained from analysis of a known bacterial strain.

Among many other computational ways to determine such probabilities, the PeptideProphet algorithm (Keller et al. 2002) gained a wide acceptance in the field of proteomics. It may be used as a standalone application or as part of a suite of software tools for the analysis of tandem MS data sets known as the Trans-Proteomic Pipeline (Deutsch et al. 2010). PeptideProphet was also incorporated into BacID/ABOid software and applied for the selection of correct PSMs used for discrimination of diverse microbial strains (Dworzanski et al. 2006, 2010; Jabbour 2010a, b, c, 2014; Wade et al. 2010, 2011). In this approach, BacID/ABOid retrieves and organizes both SEQUEST and PeptideProphet output files by creating a binary matrix of PTB assignments which can be generated using all raw data, or any subset of PSMs selected to ensure high confidence results. The final PTB matrix is created by "filtering out" not only the low-quality PSMs but also identifications matching common contaminants and sequences from the decoy DB, and retaining only a sequence unique set of peptides which are then combined and archived for further processing using a comma separated value (CSV) file format.

Recently, Koskinen et al. (2011) described the approach that "seeks to present DB search results in a more logical format", that is, by creating a minimal set of proteins, grouped into families on the basis of shared peptide matches and by using hierarchical clustering with scores of non-shared peptide matches as a distance metric. This approach is very similar to that used by BacID/ABOid software (Dworzanski et al. 2006, 2010; Deshpande et al. 2011) for presenting DB search results of unknown microbial strains represented as "pseudo-polyproteins." Unfortunately, the BacID/ABOid software is not available to the scientific community; therefore incorporation of this approach into a family of DB processing tools by Mascot will allow researchers to illustrate how families of strains are related and thus making it easier to make taxonomic or diagnostic decisions.

Subspecies Differentiation and Strain-Level Typing of Bacteria Based on Searching Protein DBs with Peptide MS/MS Spectra

The availability of commercial LC-MS software tools for full characterization of microorganisms and their physiological capabilities have lagged behind the technological advances in MS instrumentation. Although significant effort has been put into development of bioinformatics tools to identify mixtures of proteins, software applications that focus specifically on the identification of microbial strains and characterization of its proteome have been lacking. While commercial search engines combined with available data-mining methods can be used to identify microorganisms, the majority of these tools do not have the ability to take into account intricate phylogenetic relationships among strains which are an important part of characterizing both isolates and microbial mixtures. Therefore, customized DBs and data-mining approaches have been developed by a few research groups to overcome these shortcomings.

Searches of fragmentation mass spectra from trypsinized microbial proteins against DB of reference proteomes return PSMs identifying peptide sequences and can be used for revealing the distribution of PSMs among DB species. Further processing of such assignments allows to: (i) deduce the identity of an isolated organism based on analysis of taxon-specific and taxon-shared sequences, and (ii) uncover intraspecies relatedness based on genomic similarities revealed by analysis of the multidimensional structure of peptide conservation profiles across DB strains. However, there are no standardized approaches on how to perform such analyses; therefore, I will outline only the methods most frequently used by diverse research groups.

Classification and Identification of Bacteria Based on the Number of Shared Peptides

The need for rapid detection, identification, and classification of pathogenic microorganisms is vital for clinical, epidemiological, agricultural, and public health emergencies that include a potential biological terrorist attack. Therefore, the efforts to achieve such objectives were substantially intensified after the October 2001 anthrax attack in the USA. Many methods were proposed for this purpose and some of them were based on mass spectrometry, for example, analysis of microbial cell pyrolysis products, lipid extracts, nucleic acids, proteins, or amino acid sequences of protein digestion products, that is, peptides.

The use of protein sequences for the identification of species is not new (Sanger 1959) and was underscored by Frederic Sanger during his Nobel Prize Lecture in 1952. His idea was next revitalized with the advent of high-throughput proteomics era by C. Fensealu, P. Demirev, J. Yates, and others (Yates 1998; Demirev et al. 1999; Fenselau and Demirev 2001).

One of the bottom-up proteomic methods aimed for rapid identification and classification of microbes based on the concept of the number shared peptides was invented by J. P. Dworzanski and L. Li (Dworzanski et al. 2004). They coupled LC/MS/MS analysis of peptides obtained by trypsin digestion of whole cell bacterial extracts with searching an in-house created DB obtained by translating available genomic sequences with the ORF finding software Glimmer. The analysis of peptide sequences and their matches to proteomes of reference bacteria in the DB allowed them to identify selected bacterial samples down to the species and strain levels. Furthermore, they could identify the isolates regardless of the culture growth phase and with no prior knowledge of the test sample (Dworzanski et al. 2006). This procedure was next automated by using algorithms BacID/ABOid developed by J.P. Dworzanski and implemented by S. Deshpande in Visual Basic and Perl (Dworzanski et al. 2006; Deshpande et al. 2011) and applied for analysis of diverse agents of biological origin (ABO) (Dworzanski et al. 2010; Jabbour et al. 2010a, b, c, 2014; Wade et al. 2010, 2011).

Peptide-to-Taxa Assignments: Determination of the Closest Neighbor

The shotgun-proteomic analysis of an unknown strain followed by DB searches and validation of determined PSMs gives a peptide profile of an unknown strain (u). That type of peptide profile can be represented as a column vector with each component indicating that the specified sequence is encoded in its genome. On the other hand, each of these peptides may match only one DB reference proteome (unique peptides) or many (shared peptides). Thus, each peptide is characterized by a “phylogenetic” profile across DB reference strains and may be represented as a row vector with each component taking a value of either one or zero, where one/zero indicates the presence/absence of the exact matching peptide sequence in the corresponding DB proteome. These row vectors form a matrix of assignments that may be visualized as a virtual array of peptides assigned to theoretical proteomes of DB strains (Fig. 5.1, where 1/0 are represented as closed/open circles, respectively). This way, the results of MS/MS analysis may be represented as a binary map of PTB strain assignments where similar peptide profiles per reference strains indicate a correlated pattern of relatedness among such DB strains while similar “phylogenetic” profiles of peptides across strains suggest that they originate from homologous proteins (Dworzanski et al. 2006).

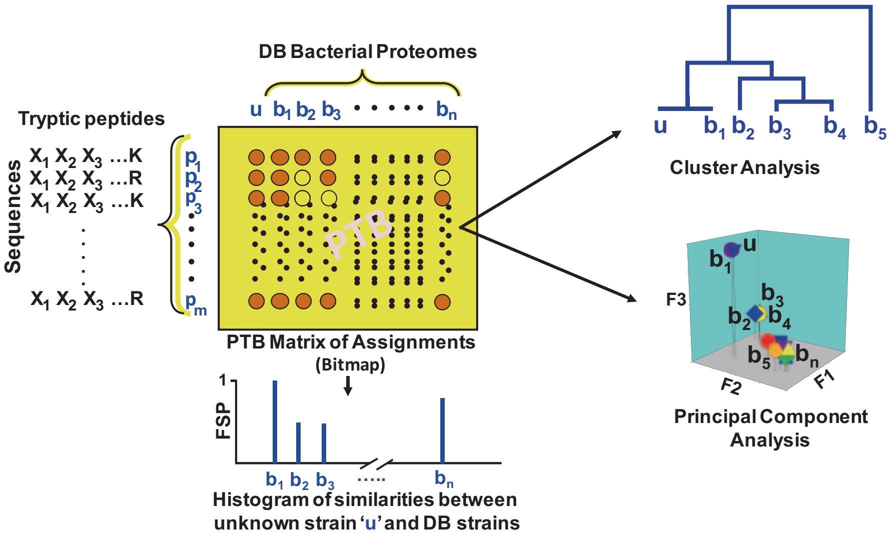


Fig. 5.1 Schematic representation of mapping tryptic peptides sequences (p_1, p_2, p_3, …, p_m) identified by shotgun-proteomics analysis of an unknown *(u)* strain to database *(DB)* proteomes of reference strains (b_1, b_2, b_3, …, b_n) and analysis of the created matrix of peptide-to-bacterial *(PTB)* strain assignments using multivariate statistical methods to reveal the closest DB neighbor. "FSP" in the "Histogram of similarities" stands for fractions of shared peptide sequences between *u* and each DB bacterial proteome. (Reprinted with permission from Dworzanski et al. (2006, pp. 76–87). Copyright 2006 American Chemical Society)

Dworzanski et al. (2004) carried out 1D HPLC-MS/MS analysis of tryptic digests derived from protein extracts of selected bacterial strains with fully sequenced genomes and used a statistical scoring algorithm to rank MS/MS spectral matching results for bacterial identification. Peptides with scores exceeding a threshold probability value were accepted and assigned to the bacterial proteomes represented in the DB. Because they used modified header lines of each protein in a DB (see Section "Bacterial DBs" for details), SEQUEST was recognizing each proteome as a single "pseudo-polyprotein" and assigning PSMs directly to reference DB strains instead to particular proteins in each proteome.

All the PTB strain assignments reported by SEQUEST were then organized as PTB matrices allowing for easy transformations and presentation of results that included: (i) ranking of assignments in the form of histograms showing the number of matching peptides per reference proteome (similarity scores) or (ii) displaying the distribution of unique peptides to further improve identification by the removal of "degenerate peptides," that is, peptides shared by reference proteomes (Dworzanski et al. 2004).

The selection of unique peptides was carried out by assuming that a DB strain proteome with the highest number of matching peptides is deemed to be the most likely candidate of a true match. With this assumption, deconvolution can be performed iteratively by selecting the highest scoring bacterium and filtering out shared peptides from histogram bins associated with all the remaining bacteria, which generates a new histogram of peptide matches per strain. A subsequent step involves the removal of peptides from the second highest scoring organism in the newly assembled histogram, and so on. Such a deconvolution filter acts as the "Occham razor" that removes shared peptide sequences from the PTB matrix, usually associated with orthologous proteins, and reveals the minimum set of strains capable of explaining all accepted PSMs.

Initially, the approach developed by Dworzanski et al. (2004) was focused on the identification of bacteria with fully sequenced genomes and therefore represented in the DB. However, they also reported that although unique peptides from the correctly identified strain can explain all high scoring PSMs, it is not the case for a strain not represented in the DB. In such cases, the Occham razor-type filter reveals the nearest DB neighbors of an unknown strain, reflecting taxonomical position of an unknown microorganism.

Recently, Tracz et al. (2013) reported a novel variation of the above method for bacteria identification implemented by using the Mascot search engine. In this approach, they compared the number of peptides shared between the unknown and DB strains by tricking Mascot to report such assignments. They achieved it by creating "pseudo-polyproteins" of concatenated protein sequences of each DB strain proteome (see Section "Bacterial DBs" for details), so searches against such a DB report PSMs to DB strains instead to particular proteins.

In proof-of-concept experiments, they analyzed whole cell protein extracts from selected *Bacillus, Francisella, Yersinia,* and *Clostridium* strains with complete genome sequences, or their close neighbors with the same status, which were chosen as surrogates for highly pathogenic species (Fig. 5.2). To speed up the sample preparation process, the reduced and C-alkylated proteins were digested with trypsin at elevated temperature and analyzed with a nano-LC-LTQ Orbitrap mass spectrom-

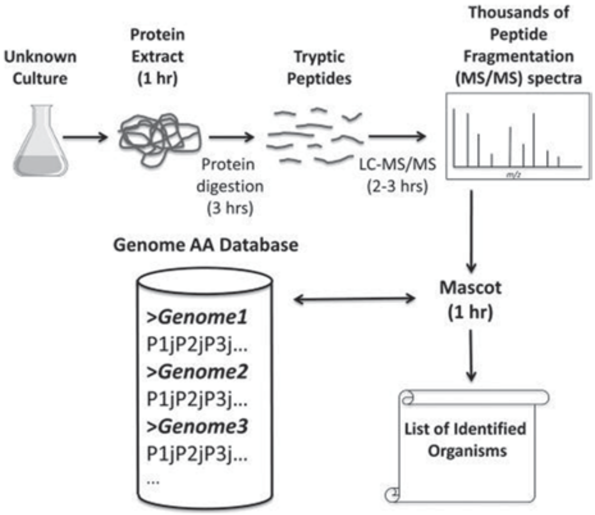


Fig. 5.2 Schematic of a shotgun-proteomics "genome identification" method that in less than 8 h (postculture) allows for strain identification. This method involves: (i) protein extraction and in-solution trypsin digestion, (ii) analysis of tryptic peptides by LC-MS/MS, and (iii) using MS data to search against a novel DB of genomes represented by concatenated proteins of genome-predicted proteomes. (Reprinted with permission from Tracz et al. (2013, pp. 54–57). Copyright 2013 Elsevier B.V.)

eter. The acquired tandem mass spectra were searched against “Genome AA” DB, and the search results were exported from the Mascot Website interface in a CSV file format.

The results file contains all accepted PSMs and DB strains ranked according to Mascot scores reflecting, among others, the total number of peptide matches per strain. These numbers are graphed as black bars in Fig. 5.3 depicting results from identification of strains by LC-MS/MS. However, peptides assigned to the highest scoring strain (so-called “red bold” matches in the Mascot jargon) could be divided into strain specific or “unique” peptides and those shared with other strains and called “degenerate.” Mascot flags the latter peptides when they are assigned to any other strain in the report as “not bold red”; thus allowing to filter out matches to degenerate peptides from all remaining strains. This process is repeated in regard to the second highest ranking strain and so on, allowing counting only unique matches and preparing a minimal list of strains contributing to the pool of identified peptides. The numbers of such peptides per strain are presented as gray bars in Fig. 5.3 and allow for clear identification of analyzed *Francisella* strains as *Francisella tularensis* LVS and *Francisella philomiragia* subsp. philomiragia. Note that the analyzed strain of *F. tularensis* LVS is represented in the DB while the strain of *F. philomiragia* subsp. philomiragia (ATCC 251015) is represented in the DB only by a different strain of this subspecies, that is, strain ATCC 25017.

Under these circumstances the highest scoring strain, LVS, was correctly identified because among all unique peptides, the matches to other closely related strains were lower than 0.2 %. However, in case of ATCC 25015^T strain matches to other strains were substantially higher because 37 (2.6 %) of unique peptides matched *Francisella noatunensis* subsp. orientalis str. Toba 04 and 30 (2 %) matched the *Francisella sp.* TX077308 strain. This indicates minor sequence differences between strains ATCC 25015 and ATCC 25017^T, and proves that LC-MS/MS can potentially discriminate isolates from the subspecies philomiragia.

Tracz et al. (2013) pointed out the advantages of their approach such as the lack of any prior knowledge of the analyzed microorganism, and the capability of generating organism-specific sequence data. In addition, their method can provide relative protein expression levels, including the confirmation of virulence factor expression, which has relatively low cost of consumables per sample; and a relatively fast turnaround time (< 8 h postculture). Moreover, it can be easily implemented in a typical proteomics laboratory.

Analysis of Subproteomes

Although subproteome analyses, by definition, are limited in scope, they usually provide comprehensive representation and coverage of specific protein types in comparison to whole cell proteome approaches. Among subproteomes investigated for subspecies identification of bacteria, the most attention attracted surface and membrane proteins, especially OMPs of Gram-negative bacteria, surface layer (S-layer) proteins of Gram-positive bacteria, and ECPs.

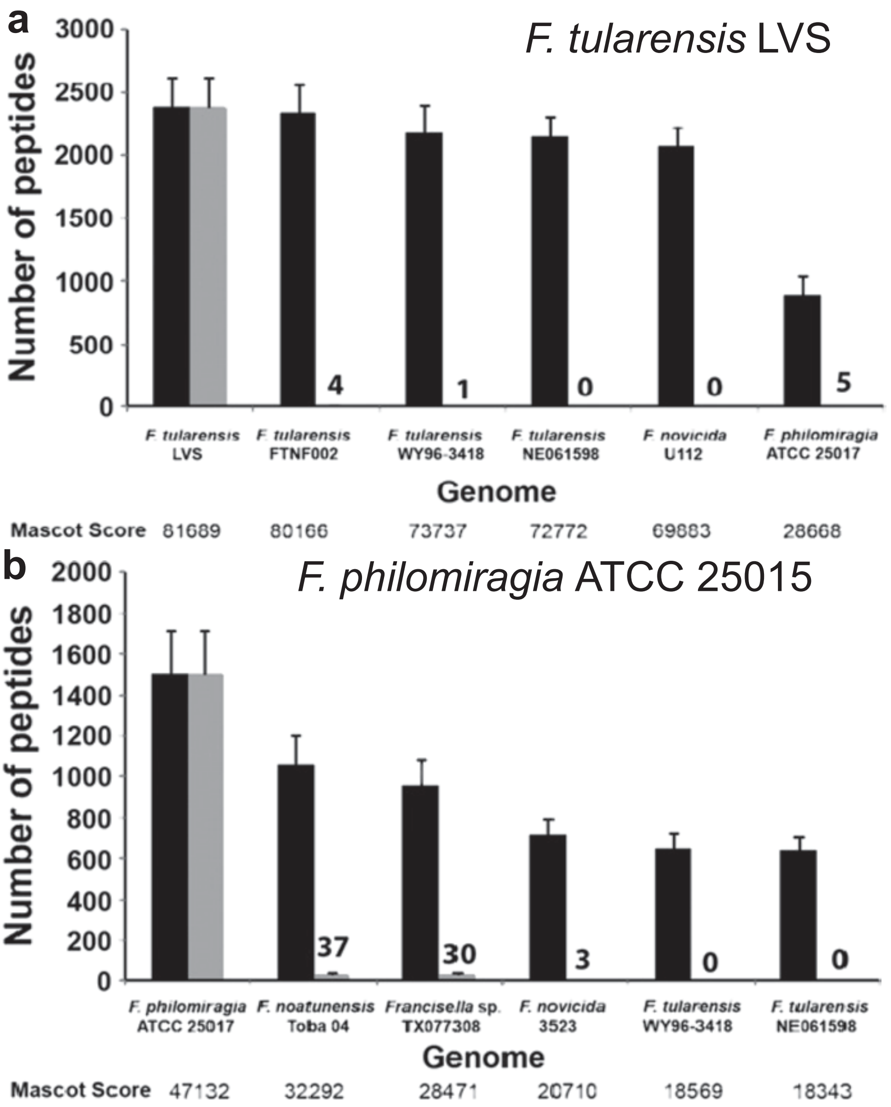


Fig. 5.3 Representative results from identification of bacterial species by LC-MS/MS. DB search results are plotted for *Francisella tularensis* LVS (**a**) and *Francisella philomiragia* ATCC 25015 (**b**). The total number of shared peptides *(black bars)* and the number of strain unique peptides *(gray bars)* for identified bacterial genome-predicted proteomes were sorted by Mascot scores. (Reprinted with permission from Tracz et al. (2013, pp. 54–57). Copyright 2013 Elsevier B.V.)

Surface proteins, including OMPs play a critical role in processes leading to pathogenicity by mediating interaction with a host, evasion of the immune system, efflux of antibiotics, and import of nutrients. Due to their location, they interface the cell and the environment and are candidate targets for developing protective strategies (vaccines and therapeutics) as well as detection and identification strategies for microbial strains. For example, Jabbour et al. 2010b showed that shotgun-proteomics analysis of OMPs from the *Yersinia pestis* CO92 strain provided unambiguous strain-level identification with all identified tryptic peptides matching the correct DB reference strain, while the remaining DB reference strains of *Y. pestis,* that is, 91001, Antiqua, Nepal 516, Kim, and *Yersinia pseudotuberculosis* IP 32953 were ranked as distant matches based on the number of shared peptides. In addition to strain identification, the results of the same analysis also provided a list of proteins known as being associated with established *Yersinia* virulence factors, like plasmid-encoded plasminogen activator protease precursor and the toxin protein.

Karlsson et al. (2012) analyzed surface-exposed proteins of fully sequenced *Helicobacter pylori* strains J99, ATCC 26695, and the type strain of this species, CCUG 17874^{T}, by using shotgun-proteomics method applied to intact cells immobilized in the flow channel of a microfluidic device called lipid/protein interaction (LPI)-FlowCell. The released and identified peptides were matched to 38 reference strains with complete genome sequences, including 26 *H. pylori* and 12 strains from other species of the *Helicobacter* genus. They showed that this method worked well for discriminating different strains of *H. pylori,* including the strain not represented in the DB.

Wade et al. (2011)investigated the discrimination of pathogenic and nonpathogenic strains of *Francisella tularensis* and *Burkholderia pseudomallei* by using shotgun-proteomics analyses. They found that LC-MS/MS analysis of trypsinized, OMP-enriched subproteomes of these microorganisms combined with data processing that included the BACid software application that allowed for confident subspecies identification and discrimination between pathogenic and nonpathogenic strains of the same species. For example, they analyzed the OMP extract of a highly virulent strain *F. tularensis* subsp. *tularensis* Schu S4, considered a potential bioterrorism agent because it causes the severe disease called type A tularemia, and the analysis of the attenuated strain of *F. tularensis* subsp. *holarctica,* known as the only live vaccine strain (LVS). These strains represent two of multiple recognized *F. tularensis* subspecies that differ in virulence and lethality following infection, that is, *tularensis,* causing the most severe disease, moderately virulent subspecies *holarctica,* followed by *mediasiatica* and *novicida* causing infections only in immunocompromised individuals (Steiner et al. 2014). Genomic analysis suggests that the subspecies of *Francisella tularensis* have evolved by vertical descent, through unidirectional gene losses from the highly virulent strain of *F. tularensis* subsp. *tularensis* which gave the less virulent *F. tularensis* subsp. *holarctica* strains. Furthermore, the attenuated LVS strain also evolved from the *holarctica* strain through gene losses, because complementation of LVS with genes *pilA* and FTT0918 restored its virulence to the level of virulent *holarctica* strains (Forslund et al. 2006; Salomonsson et al. 2009).

Shotgun proteomics analyses took advantage of such differences by allowing not only distinguishing different *Francisella tularensis* subspecies but also for confident discrimination between similar strains. For instance, analysis of the strain Schu S4 allowed for correct identification of this strain at the subspecies level *(tularensis).*

The analysis also discriminated this strain from other *F. tularensis* subsp. *tularensis* strains, that is, FSC 198, WY96198, and identified strain unique peptides from proteins associated with known virulence factors, like type-IV pili fiber building block protein (Lindgren et al. 2009).

Sequence variability is a common feature in surface and secreted proteins of microorganisms because such variability may confer increased fitness allowing the pathogen to use alternative receptors and infect different tissues or even different species. In most cases the variability probably reflects antigenic variation, which allows the pathogen to evade protective immunity in an infected host. It is commonly assumed that conservation of a limited number of residues is sufficient to promote correct protein folding and/or to confer a specific function, while other residues may vary and cause changes in antigenic properties of the protein. For example, Jabbour et al. (2014) reported that ECPs that include both actively secreted and those originating from leaking through or shedding cellular membranes could be used for the characterization of pathogenic *E. coli* strains. These included enterohemorrhagic *E. coli* (EHEC) that cause hemorrhagic colitis and enteroaggregative (EAEC) strains, like the serotype O104H4 that caused the fatal outbreak which occurred in Germany in 2011. They found shotgun-proteomics analysis of ECPs very useful and practical for differentiation among EHEC and EAEC strains due to the increased number of strain-unique peptides identified in comparison to their results obtained with whole cell protein extracts.

Confirmation of the Taxonomic Position of an Unknown Strain

Generally, analysis of a strain not represented in the DB indicates that a single DB strain cannot explain all accepted PSMs. However, a similar output could also be obtained by analysis of a mixed-culture sample or by contamination with other microbial proteins/peptides in the analytical laboratory, for example, through sample carryover. Therefore, to exclude the risk of cross-contamination or sample carryover, the profiles of all identified peptides represented as a binary matrix of PTB assignments can be further analyzed to infer taxonomic positions of contributing strains. The approach devised by Dworzanski et al. (2006) is based on the lowest common ancestor (LCA) strategy of inferring taxonomic position from peptide sequences by mapping them to "pseudo-super-proteomes" of DB strains grouped into hierarchical taxonomic units. A very similar strategy was later incorporated into the MEGAN algorithm (Huson et al. 2007) and other software tools for analysis of metagenomic data and metaproteomic data, like the Unipept web application that—based on submitted tryptic peptides—returns an interactive tree map by providing an insight into the sample biodiversity (Mesuere et al. 2012).

In this approach all DB strains are classified in accordance with the established taxonomy of prokaryotic microorganisms where similar bacterial strains are grouped into species while groupings of very similar species form genera. These species/genus levels in the taxonomic position within the classification scheme is reflected in the binomial name of bacteria. However, groupings do not stop at this level, but also include higher taxonomic arrangements of organisms into hierarchical classifications based on similarities. Namely, similar genera are placed in the same family; similar families in the same order; similar orders in the same class; similar classes in the same phylum; and finally all bacterial and archaeal phyla form the domains (or "kingdoms") of *Bacteria* and *Archaea,* respectively. Consequently, the classification of an unknown strain involves mapping of its peptides to taxa represented by "pseudo-super-proteomes" composed of DB strains grouped into the descending taxonomic ranks: phyla, classes, orders, families, genera, and species in accordance with the NCBI taxonomic classification hierarchy (Federhen 2012).

According to this peptide-centric LCA algorithm, a peptide is assigned to a lower level taxon only if its sequence is unique to this taxon; otherwise it remain assigned only to the higher level taxon and the process proceeds from domains to phyla, classes, orders, families, genera, species, and subspecies levels. For example, a peptide is assigned to a given species only if it does not match with any other species contained in the sequence DB; conversely, if the sequence is shared among several species contained in the DB, all belonging to the same genus, the sequence is unambiguously assigned only at the genus level. This way, widely conserved peptide sequences are always assigned to high-order taxa and highly variable provide the most accurate results for discrimination at the subspecies level. This type of analysis takes into account the error rate determined for the accepted set of PSMs and can be executed in a few seconds by a software application (ABOid, Deshpande et al. 2011). Moreover, this approach could also be applied to metaproteomic analyses of microbial mixtures.

The above described procedure is quite useful because it focuses the final classification process on a group of reference strains that are closest relatives of the isolated one. For example, shotgun-proteomic analysis of the whole cell protein extract of a poisonous strain isolated from the Indonesian rice dish followed by the above classification method indicated that it can be classified as *Firmicutes* → *Bacilli* → *Bacillales* → *Bacillaceae* → *Bacillus* → *B. cereus* group strain with the highest number of unique assignments matching the *B. cereus* ATCC 14579 strain. The correctness of this identification was confirmed by DDH analysis, and sequencing of the 16S rRNA and *gyrB* phylogenetic markers. In addition, this strain was serotyped based on the polymorphism of flagellar H-antigen as H-10 (Dworzanski et al. 2010). Bitmap representation of the PTB matrix of 599 peptide sequences from this strain assigned to the nearest DB strains is shown in Fig. 5.4. The displayed DB strains and peptide sequences were rearranged and analyzed by two-way hierarchical cluster analysis (HCA) with PermutMatrix (http://www.atgc-montpellier.fr/permutmatrix/) using Euclidean distances and unweighted pair group averages as the aggregation method (Caraux and Pinloche 2005). Dworzanski et al. (2010) found that the rice isolate shared 526 peptides (FSP = 0.88) with the closest DB neighbor strain (*B. cereus* ATCC 14579) while other members of the *B. cereus* group, and especially a clade of *B. anthracis* strains, were more distant (FSP = 0.82). Furthermore, the remaining *Bacillaceae* strains shared only 3–6 % of peptides with the serotype H-10.

The bitmap representation of PTB matches simplifies comparative analysis of strains by focusing on peptides with high discriminative power. For example,

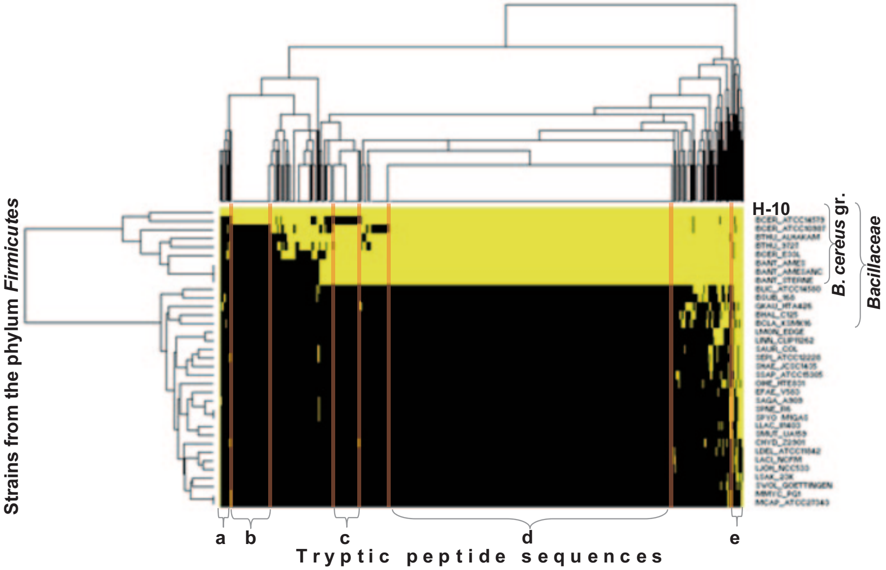


Fig. 5.4 Bitmap representation of the clustered data matrix of 599 peptide sequences from the *B. cereus* serotype H-10 isolate assigned to the nearest neighbors in the DB. Each *yellow (white)* cell represents the presence and each *black* cell the absence of a peptide-to-bacterium match. Two-way HCA was performed with PermutMatrix (Caraux and Pinloche 2005) using Euclidean distances and unweighted pair group averages as the aggregation method. The dendrogram of bacterial strains shows that the H-10 strain clusters with the *B. cereus* group of bacteria and forms a subcluster with a type strain *B. cereus* ATCC 14579. The dendrogram of peptides allows visual selection of sequences. For instance, clusters marked "a" through "e" indicate groupings of peptides with different discriminative/diagnostic power. Abbreviations of DB bacterial strains: XYYY_Z…Z, where X represents the first letter of a genus name, YYY represent the first three letters of a species name, and Z…Z represent the strain name. (Reprinted with permission from Dworzanski et al. (2010, pp. 145–155). Copyright 2010 American Chemical Society)

peptides marked as cluster "d" in Fig. 5.4 represent the majority of identified peptides while they only discriminate between the *B. cereus* group and remaining DB strains. On the other side, clusters "a" and "c" reveal sequences that discriminate serotype H-10 and its closest DB neighbor while cluster "e" indicates peptides with low discriminatory power. Indeed, peptides grouped in the latter cluster were derived from proteins with highly conserved sequences, that is, ribosomal proteins, elongation factors, and chaperones.

Relationship Between the Fraction of Shared Peptides (FSP) and Conservation of the Genome/Proteome

Currently, public DBs list multiple genome sequences for many microbial species and this increasing number of complete genome sequences together with next-generation sequencing capabilities available in many laboratories provides a wealth of new data for analysis of genomic similarities. Among many attempts to use such data to find similarities between strains, currently the best approach seems to be by quantifying the DNA conservation of bacterial genomes. Accordingly, the relatedness between two bacterial strains can be determined by comparing sequences of all homologous genes or their protein products through the computation of sequence-derived parameters that estimate ANI or AAI indices that correspond to the traditional DDH standard of the current species definition (Konstantinidis and Tiedje 2005a, b; Goris et al. 2007). Several programs are available for calculating the ANI; for example, JSpecies can be found at the Website: http://www.imedea.uib.es/jspecies/ (Richter and Rosselló-Móra 2009).

Despite their taxonomic value as a robust and universal measure of strain similarities, these indices are not applicable to nonsequenced species, such as clinical, food, or environmental isolates. Therefore, shotgun-proteomics methods that can indirectly measure or can estimate an AAI and DDH indices from experimentally determined FSP values could be applied for strain-level discrimination and typing of bacteria.

Inter-relationships between FSPs determined from shotgun-proteomic experiments and widely used genome conservation measures, that is, DDH and the ANI/AAI indices were estimated by Dworzanski et al. (2010). Their approach was based on the Kimura (1969) model for the estimation of amino acid substitution rates for homologous proteins. However, they extended this model to short DNA segments (used for the determination of DDH values) and (tryptic) peptides viewed as expression products of DNA segments, by making the following assumptions. First, they assumed that "homologous proteins" in the Kimura model could be substituted by "pseudo-polyproteins" of closely related strains. Second, "amino acid substitutions" arising from genomic mutations in strains (e.g., SNPs) could be replaced on the (tryptic) peptidome level by "no longer shared peptides" between "pseudo-polyproteins" representing bacterial strains. Consequently, differences between strains manifested as amino acid substitutions and quantified as AAI indices, are reflected at the DNA level by DDH values, and on the peptidome level by FSPs determined from shotgun-proteomics experiments.

With the above assumptions the time *(t)* since the divergence of any two strains from a common ancestor can be expressed as $t=-2.3\log(\mathrm{AAI})/2k_{\mathrm{aa}}$, where k_{aa} is the rate of substitution per amino acid per time. However, by substituting "amino acid sites" in each proteome with "peptide sites" T_{p} of length L (where L represents the number of amino acid residues), the "fraction of identical amino acid sites" (AAI) could be substituted by the "fraction of identical peptide sites" (FSP index), and the time since the strain divergence could be calculated from the equation $t=-2.3\log(\mathrm{FSP})/2k_{\mathrm{p}}$, where k_{p} refers to the rate constant for peptide substitutions. Obviously, for any given pair of microorganisms, the time since divergence is independent from the similarity measures used to express it. Hence, by equating time, expressed using the above shown equations, the relationship between FSP and genome/proteome conservation index can be estimated in the exponential form as $\mathrm{FSP}=(\mathrm{AAI})^{\mathrm{L}}$, where the peptide length L is equivalent to the ratio of substitution rates ($\mathrm{L}=k_{\mathrm{p}/}k_{\mathrm{aa}}$) (Dworzanski et al. 2010).

In accordance with this model, the fraction of peptides shared between two microbial proteomes is always lower than the AAI index value and depends on the peptide length. These inter-relationships are depicted in Fig. 5.5 for peptides with 8, 15, and 30 amino acids that represent a typical range of peptide lengths identified in shotgun experiments.

Due to logarithmic relationships between FSP and AAI indexes, this model predicts that relatively small differences in the amino acid identities are associated with substantially decreased values of the FSP index. Indeed, as shown above (see Section "Confirmation of the Taxonomic Position of an Unknown Strain") for a *B. cereus* strain isolated from food (H-10), the FSP with its nearest neighbor was 0.82, or 82 %; however, for more distant strains from the same genus, the FSP values dropped to only a few percent. Furthermore, this model predicts that for proteomes characterized by 94 % sequence identity on the amino acid level (AAI), that is assumed as a cutoff value for strains belonging to the same species (Konstantinidis and Tiedje 2005b), the FSP for 15 amino acid residues long peptides is only 40 % and much lower for longer peptides (Fig. 5.5). Therefore, the FSP index is char-

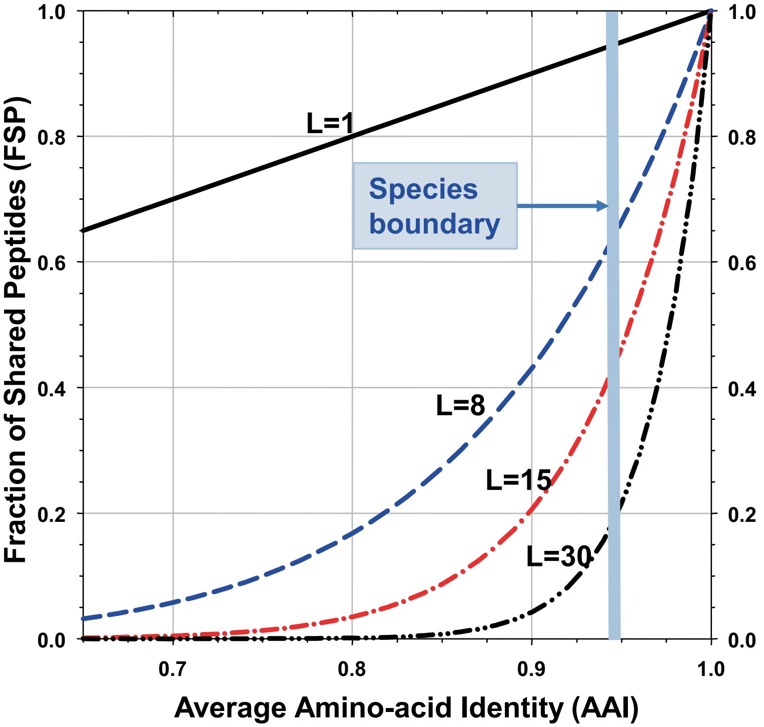


Fig. 5.5 Relationships between proteome conservation expressed as the averaged amino acid identity (AAI) index and the fraction of shared peptides (FSP) calculated for peptides of different length (L) by using the equation FSP=(AAI)^{L}. (Reprinted with permission from Dworzanski et al. (2010, pp. 145–155). Copyright 2010 American Chemical Society)

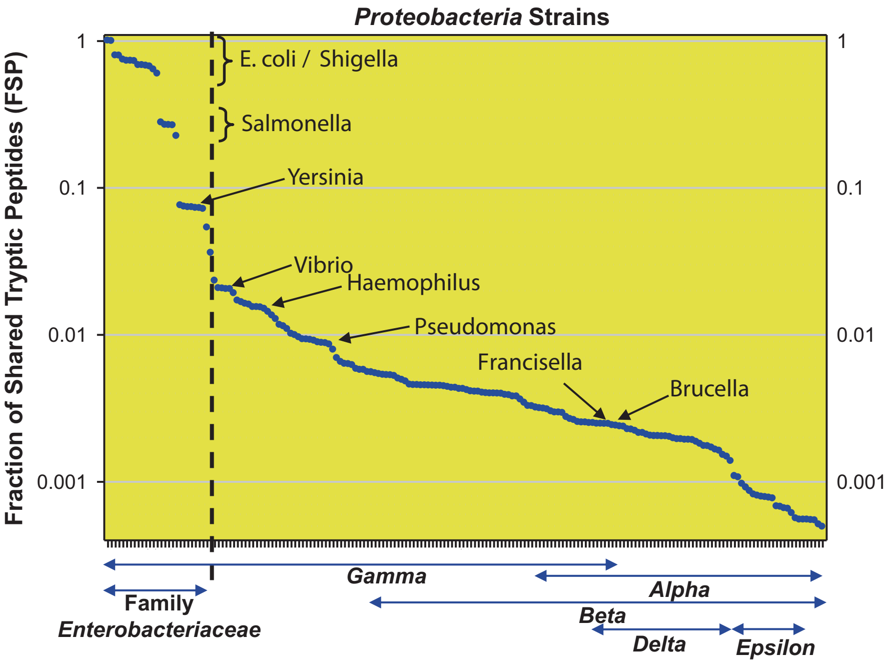


Fig. 5.6 Theoretical proteome similarities between *E. coli* K-12 and other DB alpha-, beta-, gamma-, delta-, and epsilon-proteobacterial strains, expressed by tryptic peptide FSPs calculated as Dice indices that take into account the proportion of shared (common) peptides to the averaged number of unique peptides found in both proteomes. Note that only bacterial strains from the same family (*Enterobacteriaceae*) share more than 2 % of tryptic peptides

acterized by a good resolving power required for discrimination of closely related strains, as demonstrated by whole proteome similarities between *E. coli* K-12 strain and other DB strains from the phylum *Proteobacteria* (Fig. 5.6). These theoretical FSPs were calculated as Dice similarity indices based on in silico digestion of reference DB proteomes following trypsin specificity rules, allowing up to two missed cleavages per peptide, and counting only tryptic peptides with M_r in the 700–3500 Da range (Dworzanski et al. 2010).

Conceptually, the overall genomic similarity between two strains expressed as a DDH value is equivalent to the FSP index because the matched peptides between two strains reflect DNA segments which would potentially form perfect hybrid pairings. Therefore, for consistency with the FSP term, the DDH value could be considered as a fraction of shared DNA segments between strains. On the basis of available data, relationships between these similarities were approximated by a linear function (DDH = 1.597 × FSP − 0.707, $R^{2=}$0.78) and used to calibrate proteomic similarities against DDH values (Dworzanski et al. 2010). According to this equation, the DDH cutoff of 70 %, which is used for species discrimination (Wayne et al. 1987), is equivalent to experimentally determine proteomic similarities of 88 % (FSP, 0.86–0.90). Accordingly, strains with the FSP values higher than

88 % should be treated as one species. However, the FSP values used in the above work were obtained from trypsinized whole cell protein extracts which may overrepresent peptides derived from highly conserved, high copy number proteins like those involved in the information processing. Therefore, these FSP values may be biased toward a higher FSP values in comparison to complete proteomes or some subproteomes.

Indeed, based on in silico digestion of all predicted tryptic peptides between reference proteomes of *B. anthracis* Sterne or *B. cereus* ATCC 14579 and other *Bacillaceae* strains with sequenced genomes, Dworzanski et al. (2010) found that theoretical FSPs were substantially lower than their experimental values. For example, for a pair of theoretical proteomes with a calculated FSP of 0.7, the experimentally determined values were found in the range 0.83−0.89. However, in the case of tryptic peptides released from surface-exposed proteins of *H. pylori* strains J99 and 26695, Karlsson et al. (2012) found intraspecies FSPs between these and more than 20 other *H. pylori* strains to be in the range of 0.65–0.82, that is, closer to the expected theoretical value.

It is well known that HGT, gene duplications, indels, and nucleotide substitutions are major evolutionary processes shaping microbial genomes, and closely related organisms engage in genetic exchange more frequently than distantly related ones. Recently, Caro-Quintero and Konstantinidis (2014) quantified HGT between bacterial genomes representing different phyla and found that inter-phylum HGT may affect up to ~ 16 % of the total genes. However, ribosomal and other conserved protein-coding genes were subjected to HGT at least 150 times less frequently than genes encoding metabolic enzymes or ATP-binding cassette transporters (ABC transporters). Therefore, sequences of the latter genes and their products have more discriminatory power for strain differentiation that is reflected in lower FSP values.

Turse et al. (2010) carried out investigations aimed to find FSPs between bacterial strains as a function of separating them evolutionary distances determined from 16S rDNA sequences with CLUSTAL W. In the first stage ("proof of concept") they performed LC-MS/MS analyzes of trypsin digested whole cell protein extracts from *Shewanella* strains and phylogenetically distant strains of *S. enterica* subsp. *enterica* and *Deinococcus radiodurans*. Although *Shewanella* and *Salmonella* strains are both classified as *Gamma-Proteobacteria*, Deinococcus is much more distant from both of these genera because it belongs to the separate phylum, *Deinococcus-Thermus*. They found that with increasing evolutionary distances between bacteria, the determined FSPs decrease exponentially, that is, in a fashion expected from relationships between FSPs and evolutionary similarities expressed as AAI indexes. For example, FSPs between most genetically distant *Shewanella* strains was only 6 %, while strains from this genus shared less than 1 % peptides with the *Salmonella* strain.

In the second stage Turse et al. (2010) analyzed four Columbia River environmental isolates designated as HRCR-1, 2, 4, and 5, which based on 16S rDNA sequences, showed phylogenetic affiliation with *Shewanella oneidensis* MR-1 or *Shewanella putrefaciens* CN32 strains. These findings were confirmed by the determined FSPs calculated from LC-MS/MS spectra acquired during analyses of these

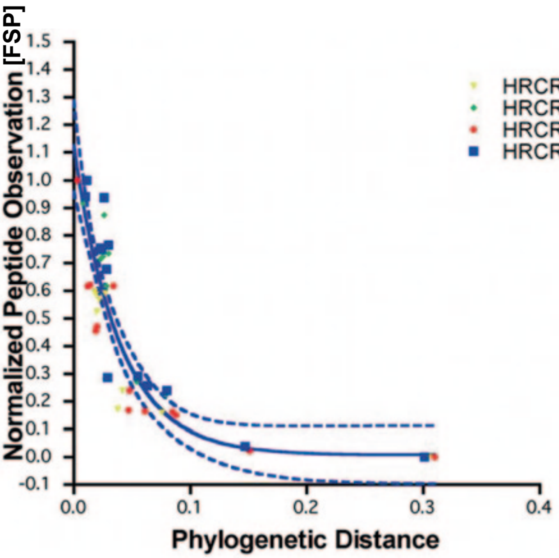


Fig. 5.7 The fraction of shared peptides (FSP is denoted here as *Normalized Peptide Observation*) between the environmental *Shewanella* isolates *(HRCR-1 through 5)* and DB strains plotted against phylogenetic (evolutionary) distances determined from 16S rDNA sequences. Reproduced from Turse et al. (2010). Open access journal

isolates. They also found that in all cases FSPs plotted against evolutionary distances were decreasing exponentially (Fig. 5.7). Note that according to the terminology used by Turse et al. FSPs are called "normalized peptide observations" indicating that the number of observed shared peptides was normalized, that is, divided by the total number of identified peptides determined by analyzing the actual reference strain under identical conditions.

Karlsson et al. (2012) analyzed surface proteins of *H. pylori* strains J99 ATCC 26695, and CCUG 17874^{T} by "shaving" surface-exposed domains of these proteins directly from intact cells immobilized in the flow channel of a microfluidic device. The released and identified peptides were matched to 38 reference strains with complete genome sequences, including 26 of *H. pylori* and 12 strains from other species of the *Helicobacter* genus. In the above-mentioned study the authors compared genomic similarities between *Helicobacter* strains based on the number of shared peptides with the well-established methods based on analysis of DNA sequences: (i) the ANI index (Konstantinidis and Tiedje 2005b) calculated using both BLAST (ANIb) and MUMmer algorithms; and (ii) tetra-nucleotide frequency correlation coefficient (TETRA, Bohlin et al 2008) that bypasses the complexity of performing multiple sequence alignments and avoids the ambiguity of choosing individual genes by inferring evolutionary relationships between species directly from their complete genomic sequences.

The ANI values between the same species strains are typically 94 % or greater while between strains of distinct species exhibit values below 94 %. The ANI values observed between *H. pylori* strain J99 and the other *H. pylori* strains were at a similar level of ca. 94%, that is, typical for intraspecies diversity, with the exception of *H. pylori* Shi 470, which had a lower ANI value of 93% (Fig. 5.8). However, the FSP values with strain J99 showed slightly better resolution for other *H. pylori* strains (FSPs at the level 0.75–0.69) with the lowest FSP value (0.686) for the Shi 470 strain. Also, the comparison of ANI and FSP values between *H. pylori* 26695 and other *H. pylori* strains showed a similar trend; however, ANIs were slightly higher (95%), with the exception of strain J99. When comparing to other species of *Helicobacter,* such as *Helicobacter acinonychis* and *Helicobacter hepaticus,* the ANI values for J99 dropped to approximately 89% and 66%, respectively, which was reflected in a lower peptide matches per strain that is equivalent to FSPs dropping down to 0.52 and 0.07, respectively (Fig. 5.8). These values correlated well with TETRA results between genomes which also have been shown to be high (>0.99) when ANI and FSP values are high, although stronger correlation was observed for interspecies genome comparisons, for example, in the case of *H. hepaticus* and other 11 strains outside the *H. pylori* species (Karlsson et al 2012).

The numbers of peptides shared between the *H. pylori* strains J99 and 26695 and strains of *H. pylori* for which genome sequences exist were also compared to the

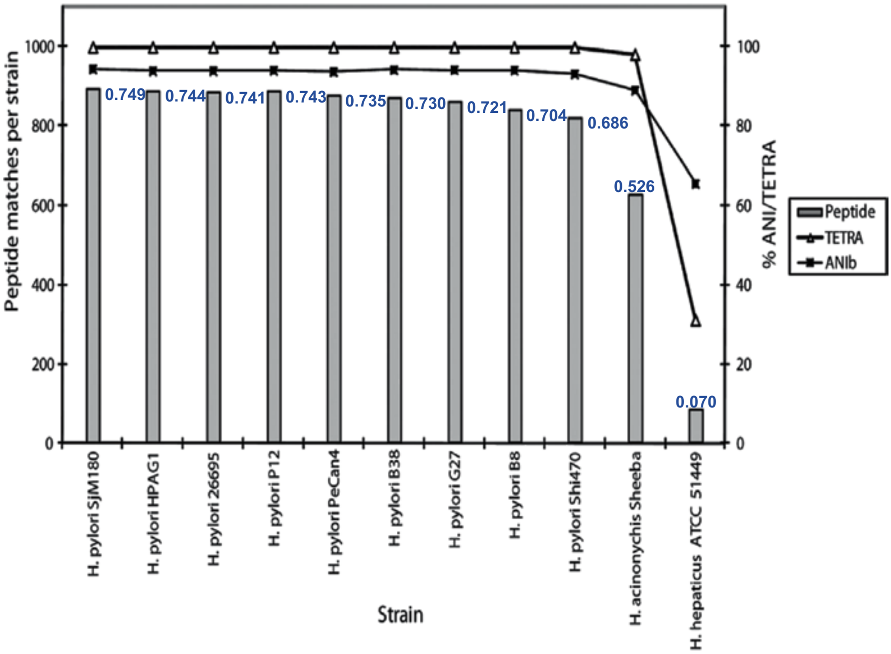


Fig. 5.8 Peptide matches per *Helicobacter* strains for the *H. pylori* J99 sample, compared to whole-genome analyses using TETRA and ANI. The peptide matches per strain are shown as bars accompanied by the FSPs values, and the TETRA (multiplied by 100) and ANI indices are depicted by lines connected by symbols as indicated by the legend box. (Reprinted with permission from Karlsson et al. (2012, pp. 2710–2720). Copyright 2012 American Chemical Society)

results from multilocus sequence typing (MLST) analyses. Karlsson et al. (2012) carried such sequence analyses of internal fragments for the seven housekeeping genes by using *H. pylori* MLST Website (http://pubmlst.org/helicobacter/). They found that similarities of the concatenated MLST sequences of *H. pylori* strains in relation to the reference strains ranged from 94.6 to 97.0%, with no direct correlation between the number of strain-specific peptides and MLST sequence similarities for *H. pylori* strains. However, interspecies comparisons showed that the decrease in the number of strain-specific peptides was accompanied by a marked decrease also in MLST sequence similarities.

In conclusion, the FSP index provides a sensitive metric for measuring genomic relatedness between microorganisms that outperforms commonly used methods for quantifying genome conservation between microbial strains.

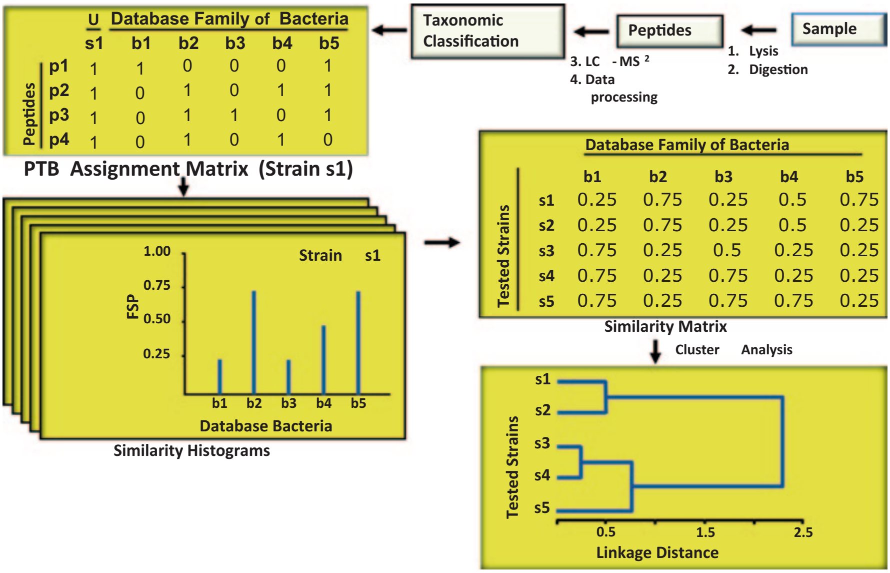
Genomic Interrelationships Among Unknown Strains Revealed by Shotgun Proteomics

Shotgun-proteomics analysis of strains isolated from clinical, food, or environmental matrices usually indicates that many DB strain proteomes could explain the determined PSMs. In general, this situation is analogous to a protein inference problem frequently encountered in bottom-up proteomics, although in this case we are focusing on "pseudo-polyproteins" representing strains instead of regular proteins. Therefore, there are two basic ways of finding the solution. The first approach is based on the parsimony principle and seeks to find the minimal list of DB strains that could explain all identified peptides (Tracz et al. 2013). The second approach is based on the creation of a maximal exploratory list of strains containing all DB strains matching at least one peptide; equivalent to selecting the whole matrix of PTB strain assignments (Dworzanski et al. 2010). However, the optimal solution could rely on using the "trimmed" matrix of PTB assignments, obtained by keeping only reference strains from the closest taxonomic units, for example, on the species, genus or family level.

In proteomics the most popular is the first approach, that is, the construction of minimal explanatory list of proteins and several tools, including ProteinProphet (Nesvizhskii et al. 2003) and IDPicker (Ma et al. 2009) are able to extract such lists automatically from the identified peptides. However, for a strain typing purposes all reference proteomes matching an isolate could be used as coordinates representing their similarities to an unknown strain (Dworzanski et al. 2010).

For example, let us assume for the sake of clarity that LC-MS/MS analysis of an unknown (U) strain s1 returned four confidently identified peptides p1 through p4 which were assigned to the closest DB neighbors represented by reference strains b_1–b_5, as shown schematically in Fig. 5.9.

As discussed in Section "Peptide-to-Taxa Assignments: Determination of the Closest Neighbor", the results of such PTB matches are arranged into the presence/absence assignment matrix and, in general, similarities between the analyzed



PTB Assignment Matrix (Strain s1)

	U	Database Family of Bacteria				
	s1	b1	b2	b3	b4	b5
p1	1	1	0	0	0	1
p2	1	0	1	0	1	1
p3	1	0	1	1	0	1
p4	1	0	1	0	1	0

Similarity Matrix

Tested Strains	Database Family of Bacteria				
	b1	b2	b3	b4	b5
s1	0.25	0.75	0.25	0.5	0.75
s2	0.25	0.75	0.25	0.5	0.25
s3	0.75	0.25	0.5	0.25	0.25
s4	0.75	0.25	0.75	0.25	0.25
s5	0.75	0.25	0.75	0.75	0.25

Fig. 5.9 Schematic representation of the sample and data processing workflow for shotgun-proteomics-based analysis of unknown strains (s1, s2, s3, s4, s5) revealing genomic inter-relationships among them. *U* unknown strain, *PTB* peptide sequence-to-bacterial strains assignments. (Reprinted with permission from Dworzanski et al. (2010, pp. 145–155). Copyright 2010 American Chemical Society)

microbial isolate and *n* reference strains $(b_1, b_2, b_3, \ldots, b_n)$ in the DB are measured as FSPs that may be presented as a similarity histogram. Moreover, these FSP indices are also considered as elements of a row vector representing that isolate in an *n*-dimensional vector space of reference strains. In Fig. 5.9 the similarity histogram for s1 indicates that this strain is not identical with any reference strain; however strains b_2 and b_5 are its closest relatives in this micro-DB, sharing 75 % of peptides (FSP = 3/4) while b_1 and b_3 are the least similar (FSP = 1/4). In the case of analyzing numerous isolates (e.g., strains s1–s5), each isolate is characterized by a set of FSP values which are elements of a row vector; and all such row vectors form a similarity matrix that can be analyzed using multivariable analysis methods, such as HCA to reveal genomic relatedness among unknown strains, for example, s1–s5.

Dworzanski et al. (2010) used this approach for phylogenomic analysis of isolates from poisonous food samples. The results of their analysis are shown as the upper diagram in Fig. 5.10 and are contrasted with a dendrogram obtained by cluster analysis of the DDH data, lower diagram, for the same strains.

The topologies of both dendrograms are very similar. Moreover, both trees closely resemble clusters and subclusters of strains revealed by HCA of concatenated nucleotide sequences of *gyrB* genes superimposed on both trees and marked as *gyrB* “Groups 1−3” to facilitate a three-way comparison of analyzed strain groupings. For instance, these topologies indicate that strains belonging to *gyrB* “Group 1” include *B. anthracis* Sterne and ten food isolates, and the same pattern was inferred from both DNA hybridization results and the proteomics data. As can be noted, two

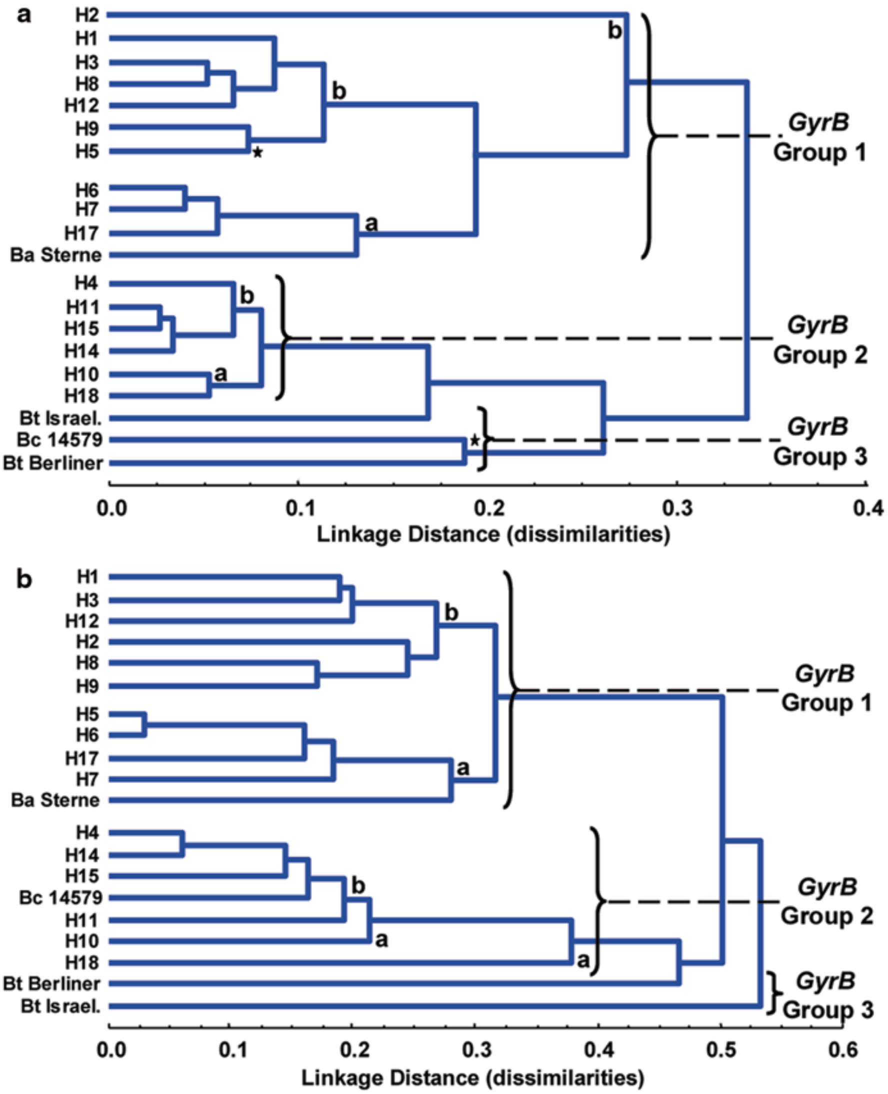


Fig. 5.10 Relatedness among *B. cereus* strains isolated from poisonous food samples (serotypes H1 through H18) and selected *Bacillus* type strains determined by hierarchical cluster analysis of distance matrices obtained from **(a)** proteomic and (**b**) DNA−DNA reassociation data. *gyrB* groups 1−3 stand for clusters of H-serotypes revealed by the analysis of concatenated nucleotide sequences of *gyrB* genes. (Reprinted with permission from Dworzanski et al. (2010, pp. 145–155). Copyright 2010 American Chemical Society

distinct subgroupings emerge from “Group 1.” The subcluster marked as “a” indicates strains highly similar to *B. anthracis,* while the subcluster “b” agglomerates strains only moderately similar to this reference strain. It is interesting to note that serotypes H1, H3, and H12 of the “b” subcluster are known as cereulide-producing strains. However, the comparison of strains grouped as members of *gyrB* subclusters "a" and "b" shows a biologically interesting disagreement between proteomics and DDH-based data. On the basis of proteomic similarities, strain H5 (marked with an asterisk in Fig. 5.10) was assigned to subcluster "b" while it was placed, together with serotype H9, into subcluster "a" on the basis of both *gyrB* sequences and hybridization values. Nevertheless, phylogenetic trees built using sequences of many housekeeping proteins and the *B. cereus* virulence factor sphingomyelinase indicate a substantial similarity of H5 with serotypes H3 and H12 and thus support findings revealed by proteomic similarities.

Overall, the obtained data indicate that proteomic similarities, DDH and *gyrB* sequencing provide very similar strain classification results, thus validating the proteomics-based approach developed by Dworzanski et al. (2010). Therefore, proteomic similarities expressed as FSP values could potentially replace DDH, as well as the gyrB or 16S rRNA sequencing in revealing phylogenomic affiliations and interrelationships among the *B. cereus* group.

Discrimination of Microbial Strains Based on Typing of Flagellin and Surface Layer Proteins

Flagellar filaments are composed of as many as 20,000 structural subunits of a 40–60 kDa protein flagellin—expressed by many bacteria, including pathogenic strains of *E. coli* and *Salmonella* spp.—and is characterized by highly variable sequences associated with the surface-exposed domains (also known as H antigens), and the conserved sequences that are crucial for filament assembly. These filaments are acting as propellers allowing cells to be motile and thus to respond to environmental stimuli; however, flagella may also contribute to bacterial pathogenicity and host immune responses (Ramos et al. 2004).

Typing of *E. coli* and *Salmonella* Strains Based on Flagellin H Antigen Sequences

Typing of *E. coli* Strains Antigenicity of flagellar H antigens and lipopolysaccharides (O antigens) were used for serotyping of *E. coli* strains for decades and this approach is widely adopted in classification of strains for taxonomic and epidemiological purposes. Moreover, serotyping based on the examination of 53 distinct H antigens is regarded as the gold standard for classification of isolates, especially during the investigation of outbreaks caused by *E. coli* pathogenic strains. However, serotyping of surface antigens is associated with some difficulties because on the one side, flagellum expression may depend on several environmental factors and on the other, diagnostic H-sera are not commercially available and therefore differ in quality. In addition, preparatory steps and serological protocols involved are laborious and lengthy because, in addition to multistep agglutination reactions, may involve extra procedures, like motility induction. The procedures involved usually take a few days to complete, therefore, molecular methods capable of replacing or to support the serotyping have been developed. They take advantage of sequence polymorphism of flagellin encoding gene *fli* C (Prager et al. 2003) or its product, that is, flagellin H-antigen (Cheng et al. 2013) to provide clear cut classification with very good correlation to serotyping.

The shotgun-proteomics-based approach based on flagellin sequencing was recently reported by Cheng et al. (2013). In this approach, referred to as "MS-H," Cheng and co-workers isolated flagella and typed the *E. coli* H antigens by searching fragmentation spectra of flagellin tryptic peptides against a custom flagellin DB of 195 unique sequence entries representing all 53 known *E. coli* H serotypes (Section "Custom DBs of *E. coli* and S*almonella* Flagellins"). More importantly, they also developed a new procedure for flagella isolation and sample processing prior to LC-MS/MS analysis. This procedure includes a simplified workflow of vortexing bacterial cells to shear off flagella, combined with their isolation by filtration that is followed by on-filter trypsin digestion (see Section "Preparation of Flagella") and LC-MS/MS analysis. The H-serotype assignments to 41 clinical isolates of *E. coli* carried out by proteomics and serological methods showed that they were concordant in 92.7 % of cases. Interestingly, the discrepancies included two strains which were untypeable by serological methods while the MS-H approach assigned their types as H7 and H21. One of these strains was previously typed as H7 and later became untypeable by agglutination, while the correctness of the second assignment (H21) to the sero-untypeable strain was confirmed by DNA sequencing of *fli* C.

The sequence coverage of flagellin depends on many factors, and one of them relates to the amount of flagellin digest used for the MS-H procedure. For example, LC-MS/MS analyses of 0.15 µg of flagellin from serotype O157:H7 with a quadrupole-TOF instrument were associated with 60 % of sequence coverage which was increased up to 88 % for a 7.5 µg sample. Therefore by replacing a quadrupole TOF instrument in the LC-MS/MS system with a higher resolution Orbitrap, they found that both diagnostic specificity and sensitivity parameters for MS-H method reached 100 %. The example of complete concordance between serotyping and proteomics results obtained by searching MS data against a curated *E. coli* flagellin DB (Custom DB) is shown in Table 5.1. In addition, the comparison of top hits returned by searches against the custom and public DBs shows the superiority of using the curated DB for strain typing based on flagellin sequences. Consequently, Cheng et al. (2013) concluded that MS-H generates results much faster and with greater simplicity in comparison to antibody-based agglutination or primer-based PCR methods and pointed out that the MS-H method should be particularly useful during *E. coli* outbreak by providing rapid presumptive H-type classification of strains.

Typing of Salmonella Strains Cheng et al. (2014b) explored the MS-H platform also for typing *Salmonella* flagella by using the same sample preparation method as for *E. coli* samples (Section "Preparation of Flagella"), followed by LC-MS/MS analysis of peptides, and searching their fragmentation spectra against a curated *Salmonella* flagellum DB containing 385 entries. However, *Salmonella* flagellins

Table 5.1 Top hits produced by searching *E. coli* flagellin MS data against a curated *E. coli* flagellin custom DB and the public DBs: Swiss-prot and NCBI nr[a]. (Cheng et al. 2014a)

Strain number	Confirmed serotype	Custom DB (195 sequences) top hit	Swiss-prot (331,337 sequences) top hit	NCBInr (25,303,445 sequences) top hit
E169	H1	H1	*Shigella* flagellin	flagellin [*E. coli*]
E170	H2	H2	*E. coli* Elongation factor	flagellin [*E. coli*]
E171	H3	H3	*Salmonella* flagellin	flagellin [*E. coli*]
E172	H4	H4	*E. coli* K12 flagellin	flagellin [*E. coli*]
E173	H5	H5	*E. coli* K12 flagellin	*E. coli* flagellar protein FliC
E174	H6	H6	*Shigella* flagellin	FliC [*E. Coli*]
EDL933	H7	H7	*Shigella* flagellin	flagellin [*E. coli*]
E176	H8	H8	*Shigella* flagellin	flagellin [*E. coli*]
E177	H9	H9	*Shigella* flagellin	flagellin [*E. coli*]
E659	H10	H10	*E. coli* K12 flagellin	flagellin [*E. coli*]

[a] An Orbitrap system was used with 30 ppm peptide mass tolerance, 0.5 Da MS/MS tolerance, one missed tryptic cleavage for all DB searches. Oxidation on methionine and deamidation on glutamine and asparagine were chosen as a possible modification.

are more diversified in comparison to *E. coli,* therefore the Kauffmann-White-Le Minor serotyping scheme for designation of *Salmonella* serotypes recognizes 119 *Salmonella* flagellum H antigens composed of combinations of distinct antigenic factors. The antigenic portion of the *Salmonella* flagellar structure is encoded by two genes—*fliC* with homologs in other enteric bacteria and *Salmonella* specific *fljB—which encode two types of flagellins, known as phase 1 and phase flagellins, respectively.* Although diphasic cells express only one type of flagellar protein at a time, some serovars always express only one flagellar antigen and are considered monophasic (e.g., *S. enterica* subspecies IIIa, IV, VII and *Salmonella bongori*). Nevertheless, in rare instances *Salmonella* may be also triphasic by expressing one-third, plasmid-encoded flagellar H antigen, thus providing a mechanism for the generation of new serovars through the horizontal transfer and recombination of flagellin genes (Li et al. 1994; McQuiston et al. 2004). Flagellar antigens that are immunologically related are also known as “antigen complexes” and exhibit very similar sequences (Ranieri et al. 2013).

To validate the MS-H approach for typing *Salmonella* strains, Cheng et al. (2014b) analyzed 24 serovars from 43 strains that included 25 diphasic, one triphasic, and 17 monophasic isolates; and obtained identification results for the first strain in only a few hours after sample preparation from the culture based on sequence coverage and the associated identification confidence scores. They found that all 17 monophasic flagella were correctly and reproducibly identified, however, complications were noticed during the characterization of phase 2 factor 1 complexes (1,2; 1,5; 1,6; and 1,2,7) and phase 1 antigen groups (“r,” “i,” and “r, i”) due to their extremely close sequence similarities (McQuiston et al. 2004). In addition, a phase 3 antigen z49 of serovar Infantis (6,7:r:1,5:z49) was not identified because the z49 sequence was not available for comparison. Overall, for 25 diphasic strains, there was 75 % accuracy for phase 1 antigens and 69 % accuracy for unstable phase 2 antigens; however, the results were 100 % accurate at the antigen cluster/complex level (Cheng et al. 2014b). In conclusion, with the increasing number of sequenced flagellar genes, the resolution of the MS-H method for some diphasic strains should also be improved in the near future.

Typing of *Lactobacillus* Strains Based on Surface Layer (S-Layer) Protein Sequences

Cell envelopes in numerous bacteria and archaea are covered by a porous layer of proteins. Moreover, for the majority of bacteria this proteinaceous surface layer is de facto composed from numerous identical protein subunits with M_r in the range of 25–200 kDa, and with a copy numbers exceeding 5×10^5 subunits (Sleytr and Messner 1983), thus making them an attractive target for extraction (see Section "Surface Layer Proteins") and sequence-based discrimination of microbial strains.

S-layers have been found in numerous *Lactobacillus* species, such as *L. helveticus*, *L. brevis*, and the former *L. acidophilus* group, that is, *L. acidophilus, L. amylovorus, L. crispatus, and L. gallinarum.* Moreover, phylogenetic trees based on *Lactobacillus* S-layer protein sequences provide much better strain resolution than those constructed on the basis of 16S rRNA or the elongation factor Tu sequences (Hynönen and Palva 2013). Therefore, Podleśny et al. (2011) took advantage of these sequence differences between the S-layer proteins by using a proteomics-based approach to identify and type strains isolated from a Canadian dairy product. They also compared proteomics results with genomic data obtained by sequencing genes encoding 16S rRNA, the RNA polymerase alpha subunit *(rpoA),* phenylalanyl-tRNA synthase alpha subunit *(pheS),* translational elongation factor Tu *(tuf),* and Hsp60 chaperonins *(groEL)* and found them in full agreement. For instance, the sequence analysis of 16S rRNA gene from the isolated strain confirmed the affiliation of an isolate with the *Lactobacillus acidophilus* group bacteria, while the MLSA data revealed the close relationships with *L. helveticus* and *L. gallinarum.* However, the determination of the partial sequences for *pheS* and *groEL* showed higher similarity with *L. helveticus* (98 %) than with *L. gallinarum* (*phes*, 96 %, *groEL* 94 %). On the contrary to these lengthy genomic procedures, the nano-LC-linear quadrupole ion trap-Fourier transform ion cyclotron resonance (LTQ-FT-ICR) MS analysis of tryptic peptides from S-layer proteins combined with searching the NCBI nonredundant DB allowed not only for high confidence identification of the source organism as *L. helveticus,* but also for typing and strain rankings based on the number of matched peptides. These data placed "surface layer protein precursor" protein—encoded by the gene *slp*—from *L. helveticus* R0052 as the best match which suggests that this strain is the nearest neighbor among six *L. helveticus* strains available in the DB. Moreover, 53 unique peptides (71 % sequence coverage) matched this surface protein from the strain R00052 while the number of matches to the remaining five *L. helveticus* strains (JCM1003, GCL1001, CP790, M4, and DPC4571) was in the range of only 14–22 peptides. This proteomics-based strain-level classification was finally validated by sequencing the *slp* gene encoding surface layer protein of the isolate and showing its 99.8 % sequence identity with the corresponding slp gene of *L. helveticus* R0052 (Podleśny et al. 2011).

In conclusion, LC-MS/MS analysis of surface layer proteins proved that the proteomics method is the appropriate molecular tool for the identification of S-layer-possessing lactobacilli at the subspecies level.

Discrimination of Strains Based on Antibiotic Resistance

The term “antibiotic resistance” implies that isolates are not inhibited by the usually achievable concentrations of a drug and may fall in the range where specific microbial resistance mechanisms are likely. In general, the resistance to a given antibiotic may be intrinsic or acquired. Therefore, the correct identification of a pathogen could be used to predict its intrinsic resistance as a naturally occurring trait characteristic for a given subspecies, species or genus. However, the conventional identification process provides no information about the acquired resistance derived either from genetic mutations or acquisition of foreign DNA from other bacteria and therefore it has to be determined experimentally by measuring the ability of an isolate to grow in the presence of commonly used antibiotics.

The automated systems for simultaneous microbial identification and antimicrobial susceptibility testing are commercially available. However, although the microbial identification may be performed in less than 1 h, for example, by MALDI-TOF-MS-based systems, the time of full panel antimicrobial susceptibility testing usually requires up to 24 h (Machen et al. 2014).

Although the antibiotic resistance could be detected by analysis of specific genes, the question remains: are these genes functional and will they be expressed? The bottom-up proteomics approach can easily address these issues by searching for specific proteins associated with antibiotic resistance (see Section “DBs of Virulence Factors, Toxins, and Antibiotic Resistance Determinants”). More importantly, the mass spectra acquired during proteomic analysis may be used to provide information both on strain identity and the expression of genes associated with antibiotic resistance.

For example, Chang et al. (2013) developed a rapid shotgun-proteomics method for the identification of β-lactam-resistant *A. baumannii* pathogenic strains based on searching a custom DB of resistance-associated proteins, referred to as “BRPDAB” (see Section “Creation/Correction of Microbial Protein DBs Through Re-sequencing and Analysis of Genomes”). They disrupted bacterial cells with a bead-beater homogenizer and processed the protein extract using a FASP method (see Section “Sample Digestion Strategies”) combined with a 15-min long microwave-assisted protein digestion with trypsin. The released peptides were analyzed using a nano-LC-ESI-MS/MS platform and the acquired fragmentation spectra were searched against the BRPDAB DB with SEQUEST.

They used data from shotgun-proteomics analyses of both multidrug resistant strain MDRAB1, and sensitive to antibiotics strains ATCC17978 and ATCC19606, to identify strain-specific peptides for *A. baumannii* which were added to the BRP-DB DB. By combining all the β-lactam resistance-related proteins and *A. baumannii* specific proteins in the same DB, they used the same search results both for the identification of *A. baumannii* and the evaluation of its antibiotic resistance potential. To validate this approach they analyzed 20 clinical isolates and found: (i) all of them correctly identified as *A. baumannii* strains; and (ii) all the 20 *A. baumannii* strains as potentially antibiotic resistant due to detection of at least two β-lactam-resistance associated proteins in each isolate. For example, all the clinical isolates expressed AmpC cephalosporinase, known as a strong antibiotic resistance enzyme that hydrolyzes most β-lactams, including penicillin, monobactam, and cephalosporins. Nineteen strains expressed carbapenem-associated resistance protein, while the *Acinetobacter*-derived cephalosporinase-53 and beta-lactamase OXA-69-like protein (named for its greater activity against oxacillin) were identified in extracts from 7 and 6 clinical isolates, respectively. Moreover, the entire procedure, including LC-MS/MS analysis and DB searching only requires 5–6 h to simultaneously identify *A. baumannii* strains and their antibiotic resistance mechanisms.

Overall, the shotgun-proteomics findings were consistent with the minimal inhibitory concentration (MIC) determination results because all 20 *A. baumannii* clinical isolates were found resistant to carbapenem, monobactam, cephalosporin, and to a combination treatment of penicillin and β-lactamase inhibitors. The results obtained demonstrate that by augmenting the custom DB with strain-specific unique peptide sequences, it is possible to obtain simultaneously both strain-level identification of *A. baumannii* clinical isolates and their antibiotic resistance mechanism information within 5–6 h. Therefore, the approach developed by Chang et al (2013) could be used for a rapid, sensitive, and specific detection of β-lactam-resistant strains of *A. baumannii*.

The bottom-up proteomic method based on CE-ESI-MS/MS of tryptic peptides was also used for the detection of a class of β-lactamases called carbapenemases in multidrug-resistant Gram-negative bacteria (*Klebsiella pneumoniae, E. coli*, and *Enterococcus cloacae*) from 27 clinical isolates (Fleurbaaij et al. 2014). For this purpose, bacteria harvested from liquid growth media, or even picked from single colonies were resuspended in 50% solution of trifluoroethanol in deionized water and lyzed by sonication, followed by protein reduction, alkylation, and the overnight digestion with trypsin. Data from MS analysis were searched against a custom DB composed of bacterial sequences downloaded from the Microbial Proteomic Resource at the University of Bergen Gade Institute Website (http://org.uib.no/prokaryotedb; de Souza et al. 2010) supplemented in-house with various β-lactamase sequences.

Overall, using a CE-ESI-MS/MS platform, Fleurbaaij et al. (2014) identified OXA-48 carbapenemase in 17 samples and demonstrated the *Klebsiella pneumoniae* carbapenemase (KPC) in 10 samples. Moreover, they found that some of these isolates also expressed a number of extended spectrum β-lactamases such as CTX (named for their greater activity against cefotaxime) which were co-expressed in 11 out of 17 OXA-48 positive strains. All these findings were confirmed by a battery of phenotypic and genomic tests (PCR-based test targeting carbapenemase; MIC analysis with meropenem, the phenotypic Hodge test).

However, they pointed out that in the case of PCR methods specific primers are needed, requiring a priori knowledge that may become problematic in case specific mutations occur in the corresponding target sequences. They also performed the MALDI-TOF MS-based ertapenem breakdown assay (Sparbier et al. 2012) with all clinical samples, and while KPC was easily detected with this method (10/10), they only correctly identified three out of 17 (3/17) OXA-48 producers.

Finally, Fleurbaaij et al. (2014) noticed that analysis of as little as 10 ng of a tryptic digest results in the identification of 300–500 unique peptides from 100 to 200 proteins. Therefore, it is obvious that the same analysis can reveal not only β-lactamase resistance but also the identity of bacterial species harboring the resistance phenotype.

It should be noted that although the antibiotic resistance in pathogenic bacteria can even cause death, the antibiotic resistance might be a useful property in case of probiotic strains used as prophylactic agents in the treatment of antibiotic-associated diarrhea. However, even probiotic strains should be free of transmissible genes that can cause the dissemination of antibiotic resistance to pathogenic bacteria and this way may reduce the therapeutic possibilities in infectious diseases. For example, Jacobsen et al. (2007) reported on in vivo transfer of wild-type AR plasmids from food strains of *Lactobacillus plantarum* to *Enterococcus faecalis* strain in the gastrointestinal tract of rats. This and other findings of acquired AR genes in isolates intended for probiotic or nutritional use highlight the importance of antimicrobial susceptibility testing in industrial laboratories for documenting the safety of commercial lactic acid bacteria in our food and the potential role of shotgun proteomics in this process (Klare et al. 2007; Gueimonde et al. 2010).

Concluding Remarks

Currently, the total number of prokaryotic genomes available in public DBs approaches 15,000 and exceeds the number of known species with validly published names (12,391); although, numerous taxa are still underrepresented in public DBs. However, species most important from the pathological, biotechnological, and epidemiological standpoint are represented by many strains, thus assuring a solid foundation for a growing use of bottom-up proteomics methods for the subspecies-level identification and typing of strains. For example, 964 and 150 genome sequences are available for *E. coli* and *B. cereus* strains, respectively. Therefore, the very large and still-growing number of sequenced microbial genomes makes it likely that identical or very similar sequences from a given species have been investigated.

On the heels of this genomic revolution, bottom-up proteomics methods allow for comparison of microbial genomes through the lens of tens of thousands of peptide sequences, providing high coverage of predicted proteomes on a routine basis. Such comprehensive readout of sequence information from genes that are actually expressed can be used for subspecies identification and sequence-based typing of microbial strains not included in whole-genome DBs. In addition, in a fraction of time needed for the whole-genome sequencing, shotgun-proteomics methods may provide comparable depth of information about genomic-level relatedness among investigated strains, thus bridging the gap between the whole-genome sequencing and other genomic methods.

The principal factor motivating the implementation of shotgun-proteomics methods is a high-information-content output provided by this approach, in comparison to MALDI-TOF-based platforms, allowing not only for high-resolution strain-level identification through finding the nearest-neighbor strains in the DB and assessment of their relatedness, but also for a comprehensive analysis of proteomes.

Such analysis of microbial strain proteomes may be performed simultaneously with strain identification and used for the characterization of strain serological and biological properties affecting pathological potential or disease outcomes, which may be revealed by the identification of virulence and AR-associated proteins as biomarkers of high diagnostic and prognostic value. Therefore, in the era of high-throughput proteomics and online bioinformatics, rapid genome-based proteomic typing of infecting agents, and especially highly virulent and potentially antibiotic resistant resistance strains, holds promise for guiding proper clinical care and to prevent potential local or global outbreaks.

Bottom-up proteomics methods still need refinement of protocols, and improvements in the standardization and availability of bioinformatics tools for comprehensive data analysis on a routine basis. Although recent innovations in mass spectrometric instrumentation have accelerated the speed and sensitivity of proteome analysis (Hebert et al. 2014), further improvements can be obtained by emphasizing the optimization, simplification, and automation of sample preparation, for example, through single-tube proteomics approaches integrating all steps from cell lysis to peptide fractionation (Hughes et al. 2014; Fan et al. 2014), peptide separation techniques, and bioinformatics tools for fast, automated data interpretation for strain-level identification of cultivable bacteria and comprehensive characterization of each isolated microbial strain in the near future.

Acknowledgment The author thanks Dr. A Peter Snyder for fruitful collaborations during various phases of the research partially described here, for critical reading of the manuscript, constructive comments on this article, and checking infelicities of expression.

References

Ansong C, Yoon H, Porwollik S, et al. Global systems-level analysis of Hfq and SmpB deletion mutants in *Salmonella*: implications for virulence and global protein translation. PLoS ONE 2009;4(3):e4809. doi:10.1371/journal.pone.0004809.

Antharavally BS, Mallia KA, Rosenblatt MM, et al. Efficient removal of detergents from proteins and peptides in a spin column format. Anal Biochem. 2011;416:39–44. doi:10.1016/j.ab.2011.05.013.

Armengaud J, Hartmann EM, Bland C. Proteogenomics for environmental microbiology. Proteomics. 2013;13:2731–42. doi:10.1002/pmic.201200576.

Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000:25:25–9. doi:10.1038/75556.

Baeumlisberger D, Rohmer M, Arrey TN, et al. Simple dual-spotting procedure enhances nLC–MALDI MS/MS analysis of digests with less specific enzymes. J Proteome Res. 2011:10:2889–94. doi:10.1021/pr2001644.

Balážová T, Šedo O, Štefanić P, et al. Improvement in *Staphylococcus* and *Bacillus* strain differentiation by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry profiling by using microwave-assisted enzymatic digestion. Rapid Commun Mass Spectrom. 2014;28:1855–61. doi:10.1002/rcm.6966.

Basile F, Hauser N. Rapid online nonenzymatic protein digestion combining microwave heating acid hydrolysis and electrochemical oxidation. Anal Chem. 2011;83:359–67. doi:10.1021/ac1024705.

Bereman MS, Egertson JD, MacCoss MJ. Comparison between procedures using SDS for shotgun proteomic analyses of complex samples. Proteomics. 2011:11:2931–35. doi:10.1002/pmic.201100045.

Bland C, Hartmann EM, Christie-Oleza JA, et al. N-terminal-oriented proteogenomics of the marine bacterium *Roseobacter denitrificans* OCh114 using N-Succinimidyloxycarbonylmethyl) tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography. Mol Cell Proteomics. 2014;13:1369–81. doi:10.1074/mcp.O113.032854.

Bohlin J, Skjerve E, Ussery DW. Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. BMC Genomics. 2008;9:104. doi:10.1186/1471-2164-9-104.

Boja ES, Fales HM. Overalkylation of a protein digest with iodoacetamide. Anal Chem. 2001;73:3576–82. doi:10.1021/ac0103423.

Bonissone S, Gupta N, Romine M, et al. N-terminal protein processing: a comparative proteogenomic analysis. Mol Cell Proteomics. 2013;12:14–28. doi:10.1074/mcp.M112.019075.

Caboche S, Audebert C, Lemoine Y, et al. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics. 2014;15:264. doi:10.1186/1471-2164-15-264.

Caraux G, Pinloche S. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. Bioinformatics. 2005;21:1280–1. doi:10.1093/bioinformatics/bti141.

Cargile BJ, Bundy JL, Stephenson JL Jr Potential for false positive identifications from large databases through tandem mass spectrometry. J Proteome Res. 2004;3:1082–5. doi:10.1021/pr049946o.

Caro-Quintero A, Konstantinidis KT. Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. ISME J. 2015;9(4):958–67. doi:10.1038/ismej.2014.193.

Chalmers MJ, Gaskel SJ. Advances in mass spectrometry for proteome analysis. Curr Opin Biotechnol. 2000;11:384–90.

Chang CJ, Lin JH, Chang KC, et al. Diagnosis of β-lactam resistance in *Acinetobacter baumannii* using shotgun proteomics and LC-nano-electrospray ionization ion trap mass spectrometry. Anal Chem. 2013;85:2802–8. doi:10.1021/ac303326a.

Chatellier S, Mugnier N, Allard F, et al. Comparison of two approaches for the classification of 16S rRNA gene sequences. J Med Microbiol. 2014;63:1311–5. doi:10.1099/jmm.0.074377-0.

Chen LH, Xiong ZH, Sun LL, et al. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucl Acids Res. 2012;40:D641–5. doi:10.1093/nar/gkr989.

Cheng K, Drebot M, McCrea J, et al. MS-H: a novel proteomic approach to isolate and type the *E. coli* H antigen using membrane filtration and liquid chromatography-tandem mass spectrometry (LC-MS/MS). PLoS ONE. 2013;8:e57339. doi:10.1371/journal.pone.0057339.

Cheng K, Sloan A, McCorrister S, et al. Fit-for-purpose curated database application in mass spectrometry-based targeted protein identification and validation. BMC Res Notes. 2014a;7:444. doi:10.1186/1756-0500-7-444.

Cheng K, Sloan A, Meakin J, et al. Sequence-level and dual-phase identification of *Salmonella* flagellum antigens by liquid chromatography-tandem mass spectrometry (LC-MS/MS). J Clin Microbiol. 2014b;52:2189–92. doi:10.1128/JCM.00242-14.

Cole JR, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucl Acids Res. 2009;37:D141–5. doi:10.1093/nar/gkp353.

Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. N Engl J Med. 2013;369:2369–71.

Cottrell JS. Protein identification using MS/MS data. J Proteomics. 2011;74:1842–51. doi:10.1016/j.jprot.2011.05.014.

Cox J, Neuhauser N, Michalski A, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10:1794–1805. doi:10.1021/pr101065j.

Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7. doi:10.1093/bioinformatics/bth092.

Damron FH, Napper J, Teter MA, et al. Lipotoxin F of *Pseudomonas aeruginosa* is an AlgU-dependent and alginate-independent outer membrane protein involved in resistance to oxidative stress and adhesion to A549 human lung epithelia. Microbiology. 2009;155:1028–38. doi:10.1099/mic.0.025833-0.

Deloger M, El Karoui M, Petit MA. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. 2009;191:91–9. doi:10.1128/JB.01202-08.

Demirev PA, Ho YP, Ryzhov V, et al. Microorganism identification by mass spectrometry and protein database searches. Anal Chem. 1999;71:2732–8. doi:10.1021/ac990165u.

Deshpande SV, Jabbour RE, Snyder PA, et al. ABOid: a software for automated identification and phyloproteomics classification of tandem mass spectrometric data. J Chromatograph Separat Techniq. 2011;S5:001. doi:10.4172/2157-7064.S5-001.

de Souza GA, Arntzen MØ, Wiker HG. MSMSpdbb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. Bioinformatics. 2010;26:698–9. doi:10.1093/bioinformatics/btq004.

de Souza GA, Arntzen MØ, Fortuin S, et al. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. Mol Cell Proteomics. 2011;10:M110-002527. doi:10.1074/mcp.M110.002527.

Deutsch EW, Mendoza L, Shteynberg D, et al. A guided tour of the trans-proteomic pipeline. Proteomics. 2010;10:1150–9. doi:10.1002/pmic.200900375.

Dworzanski JP, Snyder AP, Chen R, et al. Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. Anal Chem. 2004;7:2355–66. doi:10.1021/ac0349781.

Dworzanski JP, Deshpande SV, Chen R, et al. Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. J Proteome Res. 2006;5:76–87. doi:10.1021/pr050294t.

Dworzanski JP, Dickinson DN, Deshpande SV, et al. Discrimination and phylogenomic classification of *Bacillus anthracis-cereus-thuringiensis* strains based on LC-MS/MS analysis of whole cell protein digests. Anal Chem. 2010;82:145–55. doi:10.1021/ac9015648.

Falb M, Aivaliotis M, Garcia-Rizo C, et al. Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey. J Mol Biol. 2006;362:915–24. doi:10.1016/j.jmb.2006.07.086.

Enany S, Yoshida Y, Yamamoto T. Exploring extra-cellular proteins in methicillin susceptible and methicillin resistant *Staphylococcus aureus* by liquid chromatography–tandem mass spectrometry. World J Microbiol Biotechnol. 2014;30:1269–83. doi:10.1007/s11274-013-1550-7.

Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Amer Soc Mass Spectrom. 1994;5:976–89. doi:10.1016/1044-0305(94)80016-2.

Erde J, Ogorzalek Loo RR, Loo JA. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. J Proteome Res. 2014;13:1885–95. doi:10.1021/pr4010019.

Fan C, Shi Z, Pan Y, et al. Dual matrix-based immobilized trypsin for complementary proteolytic digestion and fast proteomics analysis with higher protein sequence coverage. Anal Chem. 2014;86:1452–8. doi:10.1021/ac402696b.

Federhen S. The NCBI Taxonomy database. Nucl Acids Res. 2012;40:D136–43. doi:10.1093/nar/gkr1178.

Fenselau C, Demirev PA. Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrom Rev. 2001;20:157–71. doi:10.1002/mas.10004.

Fenselau C, Laine O, Swatkoski S. Microwave assisted acid cleavage for denaturation and proteolysis of intact human adenovirus. Internat J Mass Spectrom. 2011;301:7–11.doi:10.1016/j.ijms.2010.05.026.

Fernández-Puente P, Mateos J, Blanco FJ, et al. LC-MALDI-TOF/TOF for shotgun proteomics. In: Martins-de-Souza D editor. Shotgun proteomics: methods and protocols, methods in molecular biology. Vol. 1156. New York: Springer; 2014. p 27–38. doi:10.1007/978-1-4939-0685-7_2.

Fleurbaaij F, Heemskerk AA, Russcher A, et al. Capillary-electrophoresis mass spectrometry for the detection of carbapenemases in (multi-) drug-resistant Gram-negative bacteria. Anal Chem. 2014;86:9154–61. doi:10.1021/ac502049p.

Forslund AL, Kuoppa K, Svensson K, et al. Direct repeat-mediated deletion of a type IV pilin gene results in major virulence attenuation of *Francisella tularensis*. Mol Microbiol. 2006;59:1818–30. doi:10.1111/j.1365-2958.2006.05061.x.

François P, Scherl A, Hochstrasser D, et al. Proteomic Approach to Investigate Pathogenicity and Metabolism of Methicillin-Resistant Staphylococcus aureus. In:Yindo Ji (ed.), Methicillin-resistant *Staphylococcus aureus* (MRSA) protocols, Methods in Molecular Biology. Vol. 1085. New York: Springer; 2014. p. 231–50. doi:10.1007/978-1-62703-664-1_14.

Fröhlich T, Arnold GJ. A newcomer's guide to nano-liquid-chromatography of peptides. In: Reinders J, Sickmann A, editors. Proteomics, methods in molecular biology. Vol. 564. Springer, Heidelberg; 2009. p. 123–41. doi:10.1007/978-1-60761-157-8_7.

Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. J Proteome Res. 2004;3:958–64. doi:10.1021/pr0499491.

Geiser L, Dayon L, Vaezzadeh AR, et al. Shotgun proteomics: a relative quantitative approach using Off-Gel electrophoresis and LC-MS/MS. In: Walls D, Loughran ST, editors. Protein chromatography: methods and protocols, methods in molecular biology. Vol. 681. Springer, Heidelberg; 2011a. pp 459–72. doi:10.1007/978-1-60761-913-0_27.

Geiser L, Vaezzadeh AR, Deshusses JM, et al. Shotgun proteomics: a qualitative approach applying isoelectric focusing on immobilized pH gradient and LC-MS/MS. In: Walls D, Loughran ST, editors. Protein chromatography: methods and protocols, methods in molecular biology. Vol. 681. Springer, Heidelberg; 2011b. pp 449–58. doi:10.1007/978-1-60761-913-0_26.

Glatter T, Ludwig C, Ahrné E, et al. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. J Proteome Res. 2012;11:5145–56. doi:10.1021/pr300273g.

Goh YJ, Azcarate-Peril MA, O'Flaherty S, et al. Development and application of a *upp*-based counterselective gene replacement system for the study of the S-layer protein SlpX of *Lactobacillus acidophilus* NCFM. Appl Environ Microbiol. 2009;75:3093–105. doi:10.1128/AEM.02502-08.

Goris J, Konstantinidis KT, Klappenbach JA, et al. DNA–DNA hybridization values and their relation to whole genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:81–91. doi:10.1099/ijs.0.64483-0.

Granholm V, Käll L. Quality assessments of peptide–spectrum matches in shotgun proteomics. Proteomics. 2011;11:1086–93. doi:10.1002/pmic.201000432.

Gueimonde M, Flórez AB, van Hoek AHAM, et al. Genetic basis of tetracycline resistance in *Bifidobacterium animalis* subsp. *Lactis*. Appl Environ Microbiol. 2010;76:3364–9. doi:10.1128/AEM.03096-09.

Gundry RL, White MY, Murray CI, et al. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. Curr Protoc Mol Biol. 2009;88:10.25.1–10.25.23. doi:10.1002/0471142727.mb1025s88.

Gupta SK, Padmanabhan BR, Diene SM, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58:212–20. doi:10.1128/AAC.01310-13.

Halbedel S, Reiss S, Hahn B, et al. A systematic proteomic analysis of *Listeria monocytogenes* house-keeping protein secretion systems. Mol Cell Proteomics. 2014;13:3063–81. doi:10.1074/mcp.M114.041327.

Hartmann EM, Allain F, Gaillard JC, et al. Taking the shortcut for high-throughput shotgun proteomic analysis of bacteria. In: Vergunst AC, O'Callaghan D, editors. Host-bacteria interactions, methods in molecular biology. Vol. 1197. New York: Springer, 2014. p. 275–85. doi:10.1007/978-1-4939-1261-2_16.

Hebert AS, Richards AL, Bailey DJ, et al. The one hour yeast proteome. Mol Cell Proteomics. 2014;13:339–47. doi:10.1074/mcp.M113.034769.

Hendrickson EL, Beck DA, Wang T, et al. Expressed genome of *Methylobacillus flagellatus* as defined through comprehensive proteomics and new insights into methylotrophy. J Bacteriol. 2010;192:4859–67.doi:10.1128/JB.00512-10.

Hesketh AR, Chandra G, Shaw AD, et al. Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of *Streptomyces coelicolor*. Mol Microbiol. 2002;46:917–32. doi:10.1046/j.1365-2958.2002.03219.x.

Hu A, Tsai PJ, Ho YP. Identification of microbial mixtures by capillary electrophoresis/selective tandem mass spectrometry. Anal Chem. 2005;77:1488–95.

Hu A, Chen CT, Tsai PJ, et al. Using capillary electrophoresis-selective tandem mass spectrometry to identify pathogens in clinical samples. Anal Chem. 2006;78:5124–33.

Hughes CS, Foehr S, Garfield DA, et al. Ultrasensitive proteome analysis using paramagnetic bead technology. Mol Syst Biol. 2014;10:757. doi:10.15252/msb.20145625.

Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86. doi:10.1101/gr.5969107.

Hynönen U, Palva A. *Lactobacillus* surface layer proteins: structure, function and applications. Appl Microbiol Biotechnol. 2013;97:5225–43. doi:10.1007/s00253-013-4962-2.

Jabbour RE, Dworzanski JP, Deshpande SV, Wick CH, Zulich AW. Effect of microbial sample processing conditions on bacterial identification using mass spectrometry-based proteomics approach. Proceedings of the 55th Conference of the American Society for Mass Spectrometry, Indianapolis, IN, June 3–7, 2007.

Jabbour RE, Deshpande SV, Wade MM, et al. Double-blind characterization of non-genome-sequenced bacteria by mass spectrometry-based proteomics. Appl Environ Microbiol. 2010a;76:3637–44. doi:10.1128/AEM.00055-10.

Jabbour RE, Wade MM, Deshpande SV, et al. Identification of *Yersinia pestis* and *Escherichia coli* strains by whole cell and outer membrane protein extracts with mass spectrometry-based proteomics. J Proteome Res. 2010b;9:3647–55. doi:10.1021/pr100402y.

Jabbour RE, Deshpande SV, Stanford MF, et al. A protein processing filter method for bacterial identification by mass spectrometry-based proteomics. J Proteome Res. 2010c;10:907–12. doi:10.1021/pr101086a.

Jabbour RE, Deshpande SV, McCubbin PE, et al. Extracellular protein biomarkers for the characterization of enterohemorrhagic and enteroaggregative *Escherichia coli* strains. J Microbiol Methods. 2014;98:76–83. doi:10.1016/j.mimet.2013.12.017.

Jacobsen L, Wilcks A, Hammer K, et al. Horizontal transfer of *tet*(M) and *erm*(B) resistance plasmids from food strains of *Lactobacillus plantarum* to *Enterococcus faecalis* JH2-2 in the gastrointestinal tract of gnotobiotic rats. FEMS Microbiology Ecology. 2007;59:158–166. doi:10.1111/j.1574-6941.2006.00212.x.

Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004;4:59–77. doi:10.1002/pmic.200300511.

Jagtap P, Goslinga J, Kooren JA, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics. 2013;13:1352–7. doi:10.1002/pmic.201200352.

Janini GM, Zhou M, Yu LR, et al. On-column sample enrichment for capillary electrophoresis sheathless electrospray ionization mass spectrometry: evaluation for peptide analysis and protein identification. Anal Chem. 2003;75:5984–93. doi:10.1021/ac0301548.

Johnson B, Selle K, O'Flaherty S, et al. Identification of extracellular surface-layer associated proteins *in Lactobacillus acidophilus* NCFM. Microbiology. 2013;159:2269–82.doi:10.1099/mic.0.070755-0.

Käll L, Canterbury JD, Weston J, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007a;4:923–5. doi:10.1038/nmeth1113.

Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucl Acids Res. 2007b;35 Suppl 2:W429–32. doi:10.1093/nar/gkm256.

Karlsson R, Davidson M, Svensson-Stadler L, et al. Strain-level typing and identification of bacteria using mass spectrometry-based proteomics. J Proteome Res. 2012;11:2710–20. doi:10.1021/pr2010633.

Keller A, Nesvizhskii AI, Kolker E, et al. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74:5383–92. doi:10.1021/ac025747h.

Kil YJ, Becker C, Sandoval W, et al. Preview: a program for surveying shotgun proteomics tandem mass spectrometry data. Anal Chem. 2011;83:5259–67. doi:10.1021/ac200609a.

Kimura M. The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci U S A. 1969;63:1181–8.

Klare I, Konstabel C, Werner G, et al. Antimicrobial susceptibilities of *Lactobacillus, Pediococcus* and *Lactococcus* human isolates and cultures intended for probiotic or nutritional use. J Antimicrob Chemother. 2007;59:900–12. doi:10.1093/jac/dkm035.

Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005a;102:2567–72. doi:10.1073/pnas.0409727102.

Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005b;187:6258–64. doi:10.1128/JB.187.18.6258-6264.2005.

Koskinen VR, Emery PA, Creasy DM, et al. Hierarchical clustering of shotgun proteomics data. Mol Cell Proteomics. 2011;10(6):M110.003822. doi:10.1074/mcp.M110.003822.

Lam MPY, Law CH, Quan Q, et al. Fully automatable multidimensional reversed-phase liquid chromatography with online tandem mass spectrometry. In: Martins-de-Souza D editor. Shotgun proteomics: methods and protocols, methods in molecular biology. Vol. 1156. New York; Springer; 2014. p. 39–51. doi:10.1007/978-1-4939-0685-7_3.

Lasaosa M, Delmotte N, Huber CG, et al. A 2D reversed-phase x ion-pair reversed-phase HPLC-MALDI TOF/TOF-MS approach for shotgun proteome analysis. Anal Bioanal Chem. 2009;393:1245–56. doi:10.1007/s00216-008-2539-1.

Lee J-G, Cheong KH, Huh N, et al. Microchip-based one step DNA extraction and real-time PCR in one chamber for rapid pathogen identification. Lab Chip. 2006:886–95. doi:10.1039/B515876A.

Li J, Nelson K, McWhorter AC, et al. Recombinational basis of serovar diversity in *Salmonella enterica*. Proc Natl Acad Sci U S A. 1994;91:2552–6.

Lin SS, Wu CH, Sun MC, et al. Microwave-assisted enzyme-catalyzed reactions in various solvent system. J Am Soc Mass Spectrom. 2005;16:581–8. doi:10.1016/j.jasms.2005.01.012.

Lindgren H, Honn M, Golovlev I, et al. The 58-kilodalton major virulence factor of *Francisella tularensis* is required for efficient utilization of iron. Infect Immun. 2009;77:4429–36. doi:10.1128/IAI.00702-09.

Lippincott J, Apostol I. Carbamylation of cysteine: a potential artifact in peptide mapping of hemoglobins in the presence of urea. Anal Biochem. 1999;267:57–64. doi:10.1006/abio.1998.2970.

Liu B, Pop M. ARDB-antibiotic resistance genes database. Nucl Acids Res. 2009;37:D443–7. doi:10.1093/nar/gkn656.

Lohrig K, Wolters D. Multidimensional protein identification technology. In: Reinders J, Sickmann A, editors. Proteomics, methods in molecular biology. Vol. 564. Heidelberg: Springer; 2009. p. 143–53. doi:10.1007/978-1-60761-157-8_8.

Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30:434–9. doi:10.1038/nbt.2198.

Ma B, Johnson R. *De novo* sequencing and homology searching. Mol Cell Proteomics. 2012;11(2):O111.014902. doi:10.1074/mcp.O111.014902.

Ma Z, Dasari S, Chambers MC, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009;8:3872–81. doi:10.1021/pr900360j.

Machen A, Drake T, Wang YF. Same day identification and full panel antimicrobial susceptibility testing of bacteria from positive blood culture bottles made possible by a combined lysis-filtration method with MALDI-TOF VITEK mass spectrometry and the VITEK2 system. PLoS ONE. 2014;9(2):e87870. doi:10.1371/journal.pone.0087870.

Manza LL, Stamer SL, Ham AJL, et al. Sample preparation and digestion for proteomic analyses using spin filters. Proteomics. 2005;5:1742–5.

Masuda T, Tomita M, Ishihama Y. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. J Proteome Res. 2008;7:731–40. doi:10.1021/pr700658q.

Masuda T, Saito N, Tomita M, et al. Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. Mol Cell Proteomics. 2009;8:2770–7. doi:10.1074/mcp.M900240-MCP200.

Mayne J, Starr AE, Ning Z, et al. Fine tuning of proteomic technologies to improve biological findings: advancements in 2011–2013. Anal Chem 2014;86:176–95. doi:10.1021/ac403551f.

McQuiston JR, Parrenas R, Ortiz-Rivera M, et al. Sequencing and comparative analysis of flagellin genes *fliC, fljB,* and *flap* from *Salmonella*. J Clin Microbiol. 2004;42:1923–32. doi:10.1128/JCM.42.5.1923-1932.2004.

Meacham F, Boffelli D, Dhahbi J, et al. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics. 2011;12:451. doi:10.1186/1471-2105-12-451.

Meier-Kolthoff JP, Auch AF, Klenk HP, et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013;14:60. doi:10.1186/1471-2105-14-60.

Mesuere B, Devreese B, Debyser G, et al. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res. 2012;11:5773–80. doi:10.1021/pr300576s.

Miller JM. Whole-genome mapping: a new paradigm in strain-typing technology. J Clin Microbiol. 2013;51:1066–70. doi:10.1128/JCM.00093-13.

Nan L, Jiang Z, Wei X. Emerging microfluidic devices for cell lysis: a review. Lab Chip. 2014;14:1060–73. doi:10.1039/c3lc51133b.

Napoli A, Aiello D, Aiello G, et al. Mass Spectrometry-based proteomic approach in *Oenococcus oeni* (*O. oeni*) enological starter. J Proteome Res. 2014;13:2856–66. doi:10.1021/pr4012798.

Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010;73:2092–123. doi:10.1016/j.jprot.2010.08.009.

Nesvizhskii AI, Keller A, Kolker E, et al. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75:4646–58. doi:10.1021/ac0341261.

Olaya-Abril A, Jimenez-Munguia I, Gomez-Gascon L, et al. Surfomics: shaving live organisms for a fast proteomic identification of surface proteins. J Proteomics. 2013;97:164–76. doi:10.1016/j.jprot.2013.03.035.

Perkins DN, Pappin DJC, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.

Petersen TN, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6. doi:10.1038/nmeth.1701.

Podlesny M, Jarocki P, Komon E, et al. LC-MS/MS analysis of surface layer proteins as a useful method for the identification of lactobacilli from the *Lactobacillus acidophilus* group. J Microbiol Biotechnol. 2011;21:421–9. doi:10.4014/jmb.1009.09036.

Prager R, Strutz U, Fruth A, et al. Subtyping of pathogenic *Escherichia coli* strains using flagellar (H)-antigens: serotyping versus fliC polymorphisms. Int J Med Microbiol. 2003;292:477–86. doi:10.1078/1438-4221-00226.

Proc JL, Kuzyk MA, Hardie DB, et al. A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. J Proteome Res. 2010;9:5422–37. doi:10.1021/pr100656u.

Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl Acids Res. 2013;41:D590–6. doi:10.1093/nar/gks1219.

Ramos HC, Rumbo M, Sirard JC. Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. Trends Microbiol. 2004;12:509–17. doi:10.1016/j.tim.2004.09.002.

Ranieri ML, Shi C, Switt AIM, et al. Comparison of typing methods with a new procedure based on sequence characterization for *Salmonella* serovar prediction. J Clin Microbiol. 2013;51:1786–97. doi:10.1128/JCM.03201-12.

Reddy PM, Huang YS, Chen CT, et al. Evaluating the potential nonthermal microwave effects of microwave-assisted proteolytic reactions. J Proteomics. 2013;80:160–70. doi:10.1016/j.jprot.2013.01.005.

Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106:19126–31. doi:10.1073/pnas.0906412106.

Salomonsson E, Kuoppa K, Forslund AL, et al. Reintroduction of two deleted virulence loci restores full virulence to the live vaccine strain of *Francisella tularensis*. Infect Immun. 2009;77:3424–31. doi:10.1128/IAI.00196-09.

Salzberg SL, Delcher AL, Kasif S, et al. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 1998;26:544–8. doi:10.1093/nar/26.2.544.

Sanger F. Chemistry of insulin. Science. 1959;129:1340–4. doi:10.1126/science.129.3359.1340.

Sato N, Tajima N. Statistics of N-terminal alignment as a guide for refining prokaryotic gene annotation. Genomics. 2012;99:138–43. doi:10.1016/j.ygeno.2011.12.004.

Shteynberg D, Deutsch EW, Lam H, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics. 2011;10(12):M111.007690. doi:10.1074/mcp.M111.007690.

Shteynberg D, Nesvizhskii AI, Moritz RL, et al. Combining results of multiple search engines in proteomics. Mol Cell Proteomics. 2013;12:2383–93. doi:10.1074/mcp.R113.027797.

Sleytr UB, Messner P. Crystalline surface layers on bacteria. Annu Rev Microbiol. 1983;37:311–39.

Sparbier K, Schubert S, Weller U, et al. Matrix-assisted laser desorption ionization–time of flight mass spectrometry-based functional assay for rapid detection of resistance against β-lactam antibiotics. J Clin Microbiol. 2012;50:927–37. doi:10.1128/JCM.05737-11.

Spivak M, Weston J, Bottou L, et al. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. J Proteome Res. 2009;8:3737–45. doi:10.1021/pr801109k.

Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol. 1994;44:846–9. doi:10.1099/00207713-44-4-846.

Steiner DJ, Furuya Y, Metzger DW. Host–pathogen interactions and immune evasion strategies in *Francisella tularensis* pathogenicity. Infect Drug Resist. 2014;7:239–51. doi:10.2147/idr.s53700.

Sun L, Jin M, Ding W, et al. Posttranslational modification of flagellin FlaB in *Shewanella oneidensis*. J Bacteriol. 2013;195:2550–61. doi:10.1128/JB.00015-13.

Sun L, Dong Y, Shi M, et al. Two residues predominantly dictate functional difference in motility between *Shewanella oneidensis* flagellins flaA and flaB. J Biol Chem. 2014;289:14547–59. doi:10.1074/jbc.M114.552000.

Switzar L, Giera M, Niessen WM. Protein digestion: an overview of the available techniques and recent developments. J Proteome Res. 2013a;12:1067–77. doi:10.1021/pr301201x.

Switzar L, van Angeren J, Pinkse M, et al. A high-throughput sample preparation method for cellular proteomics using 96-well filter plates. Proteomics. 2013b;13:2980–3. doi:10.1002/pmic.201300080.

Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res. 2007;6:654–61.

Tanca A, Biosa G, Pagnozzi D, et al. Comparison of detergent-based sample preparation workflows for LTQ-Orbitrap analysis of the *Escherichia coli* proteome. Proteomics. 2013;13:2597–607. doi:10.1002/pmic.201200478.

Tonella L, Hoogland C, Binz PA, et al. New perspectives in the *Escherichia coli* proteome investigation. Proteomics. 2001;1:409–23. doi:10.1002/1615-9861(200103).

Tracz DM, McCorrister SJ, Chong PM, et al. A simple shotgun proteomics method for rapid bacterial identification. J Microbiol Methods. 2013;94:54–7. doi:10.1016/j.mimet.2013.04.008.

Turse JE, Marshall MJ, Fredrickson JK, et al. An empirical strategy for characterizing bacterial proteomes across species in the absence of genomic sequences. PLoS ONE. 2010;5(11):e13968. doi:10.1371/journal.pone.0013968.

Vaezzadeh AR, Deshusses JM, Waridel P, et al. Accelerated digestion for high-throughput proteomics analysis of whole bacterial proteomes. J Microbiol Methods. 2010;80:56–62. doi:10.1016/j.mimet.2009.10.019.

Vaudel M, Venne AS, Berven FS, et al. Shedding light on black boxes in protein identification. Proteomics. 2014;14:1001–5. doi:10.1002/pmic.201300488.

Vuckovic D, Dagley LF, Purcell A, et al. Membrane proteomics by high performance liquid chromatography–tandem mass spectrometry: analytical approaches and challenges. Proteomics. 2013;13:404–23. doi:10.1002/pmic.201200340.

Waas M, Bhattacharya S, Chuppa S, et al. Combine and conquer: surfactants, solvents, and chaotropes for robust mass spectrometry based analyses of membrane proteins. Anal Chem. 2014;86:1551–9. doi:10.1021/ac403185a.

Wade MM, Zulich AW, Wick CH, et al. Discrimination of pathogenic versus non-pathogenic *Yersinia pestis* and *Escherichia coli* using proteomics mass spectrometry (No. ECBC-TR-771). Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD. 2010. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA522639. Accessed 7 Nov 2014.

Wade MM, Wick CH, Zulich AW, et al. Discrimination of pathogenic vs. nonpathogenic *Francisella tularensis* and *Burkholderia pseudomallei* using proteomics mass spectrometry (No. ECBC-TR-857). Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD. 2011. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA540823. Accessed 7 Nov 2014.

Wang H, Qian WJ, Mottaz HM, et al. Development and evaluation of a micro- and nanoscale proteomic sample preparation method. Proteome Res. 2005;4:2397–403. doi:10.1021/pr050160f.

Wang X, Slebos RJ, Wang D, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res. 2012;11:1009–17. doi:10.1021/pr200766z.

Warscheid B, Fenselau C. Characterization of *Bacillus* spore species and their mixtures using postsource decay with a curved-field reflectron. Anal Chem. 2003;75:5618–27.

Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001;19:242–7.

Wayne LG, Brenner DJ, Colwell RR, et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol. 1987;37:463–4. doi:10.1099/00207713-37-4-463.

Winnenburg R, Urban M, Beacham A, et al. PHI-base update: additions to the pathogen-host interaction database. Nucl Acids Res. 2008;36:D572–6. doi:10.1093/nar/gkm858.

Wiśniewski JR, Mann M. Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. Anal Chem. 2012;84:2631–7. doi:10.1021/ac300006b.

Wiśniewski JR, Rakus D. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the *Escherichia coli* proteome. J Proteomics. 2014;10:322–31. doi:10.1016/j.dib.2014.08.004.

Wiśniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods. 2009;6:359–62.

Wiśniewski JR, Zielinska DF, Mann M. Comparison of ultrafiltration units for proteomic and N-glycoproteomic analysis by the filter-aided sample preparation method. Anal Biochem. 2011;410:307–9. doi:10.1016/j.ab.2010.12.004.

Wolters DA, Washburn MP, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. 2001;73:5683–90.

Wu F, Sun D, Wang N, et al. Comparison of surfactant-assisted shotgun methods using acid-labile surfactants and sodium dodecyl sulfate for membrane proteome analysis. Anal Chim Acta. 2011;698:36–43. doi:10.1016/j.aca.2011.04.039.

Wu X, Xu L, Gu W, et al. Iterative genome correction largely improves proteomic analysis of non-model organisms. J Proteome Res. 2014;13:2724–34. doi:dx.doi.org/10.1021/pr500369b.

Yang H, Zubarev RA. Mass spectrometric analysis of asparagine deamidation and aspartate isomerization in polypeptides. Electrophoresis. 2010;31:1764–72. doi:10.1002/elps.201000027.

Yang Y, Zhang S, Howe K, et al. A comparison of nLC-ESI-MS/MS and nLC-MALDI-MS/MS for GeLC-based protein identification and iTRAQ-based shotgun quantitative proteomics. J Biomol Tech. 2007;18:226–37.

Yates JR. Mass spectrometry and the age of the proteome. J Mass Spectrom. 1998;33:1–19.

Yeung YG, Nieves E, Angeletti RH, Stanley ER. Removal of detergents from protein digests for mass spectrometry analysis. Anal Biochem. 2008;382:135–7. doi:10.1016/j.ab.2008.07.034.

Yu Y, Xie L, Gunawardena HP, et al. GOFAST: an integrated approach for efficient and comprehensive membrane proteome analysis. Anal Chem. 2012;84:9008–14. doi:10.1021/ac300134e.

Zhang G, Fedyunin I, Kirchner S, et al. FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. Nucl Acids Res. 2012;40(11):e83. doi:10.1093/nar/gks196.

Zhang CX, Creskey MC, Cyr TD, et al. Proteomic identification of *Listeria monocytogenes* surface-associated proteins. Proteomics. 2013a;13:3040–5. doi:10.1002/pmic.201200449.

Zhang K, Zheng S, Yang JS, et al. Comprehensive profiling of protein lysine acetylation in *Escherichia coli*. J Proteome Res. 2013b;12:844–51. doi:10.1021/pr300912q.

Zhou CE, Smith J, Lam M, et al. MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucl Acids Res. 2007;35:D391–4. doi:10.1093/nar/gkl791.

Zhou JY, Dann GP, Shi T, et al. Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. Anal Chem. 2012;84:2862–7. doi:10.1021/ac203394r.