REVIEW

Environmental proteomics: Analysis of structure and function of microbial communities

Thomas Schneider and Kathrin Riedel

Department of Microbiology, Institute of Plant Biology, University of Zurich, Zurich, Switzerland

Received: June 28, 2009
Revised: November 12, 2009
Accepted: November 12, 2009

Prokaryotic and eukaryotic microorganisms make a vital contribution to biogeochemical cycles by decomposing virtually all natural compounds and thereby exert a lasting effect on biosphere and climate. The rapidly growing number of metagenomic sequences together with revolutionary advances in bioinformatics and protein analyses have opened completely new horizons to investigate the molecular basis of such complex processes. Proteomics has contributed substantially to our understanding of individual organisms at the cellular level as it offers excellent possibilities to probe many protein functions and responses simultaneously. However, it has not yet been widely applied in microbial ecology, although most proteins have an intrinsic metabolic function which can be used to relate microbial activities to the identity of defined organisms in multispecies communities. Albeit still in its infancy, environmental proteomics enables simple protein cataloging, comparative and semi-quantitative proteomics, analyses of protein localization, discovery of post-translational modifications, and even determination of amino-acid sequences and genotypes by strain-resolved proteogenomics. This review traces the historical development of environmental proteomics and summarizes milestone publications in the field. In conclusion, we briefly discuss current limitations of microbial community proteomics but also the potential of emerging technologies to shape the future of metaproteome analyses.

Keywords:
Community function / Community structure / Environmental proteomics / Microbial ecology / Microbiology

1 Introduction

1.1 Microbial ecology – studying structure and function of microbial communities in the environment

Microbes such as bacteria, fungi, and viruses are omnipresent. They play an essential role in biogeochemical cycles and can decompose virtually all natural compounds, thereby exerting a lasting effect on biosphere and climate. About 20 years ago, ecologists started to realize that microbial activity and physiology in a certain environment is strongly dependent on the composition of the present community and the interactions of the community members during nutrient competition, predation, and cellular signaling [1]. However, the fact that more than 90% of the microorganisms in a given environment are not readily cultured using standard methods [2] hampered investigations aiming toward a deeper insight into the structure and function of biological systems for a long time and individual contributions of different species to a certain environment remained largely unknown. The recent development of numerous molecular tools that bypass the need to isolate and culture individual microbial species has afforded promising new insights into microbial ecology that might revolutionize our concepts of microbial diversity and physiology within complex consortia and up to entire ecosystems: (i) 16S rRNA sequencing approaches provide important information about species

Correspondence: Dr. Thomas Schneider, Department of Microbiology, Institute of Plant Biology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
E-mail: Thomas.Schneider@botinst.uzh.ch
Fax: +44-63-50384

Abbreviations: AMD, acid mine drainage; **EBPR,** enhanced biological phosphorus removal; **SAF,** spectral abundance factor

composition and evolution (reviewed in [3]); (ii) novel shotgun sequencing and pyrosequencing techniques enable the mapping of whole metagenomes (reviewed in [4–6]) as well as the study of transcriptional profiles of microbial consortia; (iii) proteomics methods allow qualitative and quantitative assessment of the protein complement in a given environment (reviewed in [7–12]). In particular, metagenomics has emerged as a powerful tool to investigate structural, evolutionary, and metabolic properties of complex microbial communities. An important milestone in the history of metagenome analyses was the Sargasso Sea sequencing project of Venter *et al.* in 2004 [13]. In the meantime, metagenomes from numerous habitats, *e.g.* soil, global ocean, human gut, and feces, have been sequenced (for an overview see, *e.g.* IMG/M webpage [14]). Thanks to the recent progress in sequencing technologies, microbial genomic, and metagenomic sequence information will continue to grow exponentially and will thus offer a solid basis for any post-genomic research [15–17].

Proteome coverage and ''resolving power'' of environmental proteomics analyses strongly depend on size and quality of reference protein databases against which MS and/or MS/MS data have to be searched. As shown in numerous recently published environmental proteomics studies, *e.g.* strain resolved community proteomics of acid mine drainage (AMD) biofilms [18–20], activated sludge [21], and community proteomics of the leaf phyllosphere [22], the combination of metagenomics and metaproteomics can provide valuable insights into structure and physiology of different phylogenetic groups present in a specific environment.

In the present review, we outline the historical development of environmental proteomics before introducing the reader to current state-of-the-art proteomics methodologies with a strong emphasis on method-critical techniques. We highlight the most important publications in the field as well as recent developments in quantitative environmental proteomics and conclude with a view toward potential future applications.

1.2 Historical retrospective of ''omics'' technologies

Prior to the last decade global analyses of microbial genomes, transcriptomes, proteomes, or even metabolomes were restricted to species amenable to isolated cultivation. More recently, rapid advances in ''omics'' technologies have made it possible to study not only hitherto uncultivable species, but complex microbial communities and even entire ecosystems. Figure 1 depicts the historical evolution and technical milestones of these global molecular approaches. At the turn of the millennium, novel shotgun DNA sequencing technologies such as 454 pyrosequencing [23] coupled with significant cost reductions gave a tremendous boost to culture-independent metagenomics research and began to reveal the diversity and distribution of indigenous microbial populations in natural environments (reviewed in [17]). Metagenomics strategies alone cannot elucidate the functionality of microorganisms present in the respective ecosystem. Moreover, an enormous number of newly identified ORFs with no homology to well-characterized genes still await functional assignment. These limitations have stimulated the development of environmental transcriptome analyses, although the short half-life of mRNA molecules, challenging extraction protocols due to interfering organic and inorganic compounds, and the often low correlation between transcription levels and actual protein expression still appear to be major drawbacks of metatranscriptome studies [24]. Increasingly, proteomics has emerged as a promising technique to characterize microbial activities at the molecular level. Proteomics, originally defined as ''the large-scale study of proteins expressed by an organism'' [25], started to develop in the 1970s when protein profiles of single organisms were analyzed by 2-DE [26]. At that time protein identification was, if at all possible, time consuming and cost-intensive due to a lack of genomic sequence information and advanced protein sequence analyses. Since the 1990s proteomics has become much more widespread, feasible, and reliable thanks to three technical revolutions: (i) the enormous increase of genomic and metagenomic data provides a solid basis for protein identification; (ii) tremendous progress in sensitivity and accuracy of mass spectrometers enables a correct, high-throughput protein identification, relative and absolute quantification of proteins, and the determination of post-translational modifications; and (iii) formidable improvements in computing power and bioinformatics allow processing and evaluation of substantial datasets. Global analyses of proteins involved in biotransformation, *i.e.* enzymes, finally allow a holistic characterization of microbial metabolic dynamics and shed light on the regulation of the metabolome, the complete set of metabolic intermediates, signaling molecules, and secondary metabolites found within a biological sample [27].

1.3 Terminology of environmental proteomics

Less than 5 years ago, Wilmes and Bond [28] defined *metaproteomics* as ''the large-scale characterization of the entire protein complement of environmental microbiota at a given point in time''; meanwhile, rapid advances and multifold applications of high-throughput ''omics'' technologies have led to many novel denominations including *environmental proteomics, metaproteomics, community proteomics*, or *community proteogenomics*. These terms are often used as synonyms; however, as rightly stated by Verberkmoes *et al.* [9], they stand in fact for slightly different experimental setups and outcomes. While *environmental proteomics* should be regarded as a generic term simply describing proteome analyses of environmental samples, *metaproteomics* comprises studies of highly complex biological systems which do not allow assigning large numbers of proteins to specific species within phylotypes. In contrast, the term *community proteomics* implies that most of the identified proteins can be related to specific members of the community; thus far such

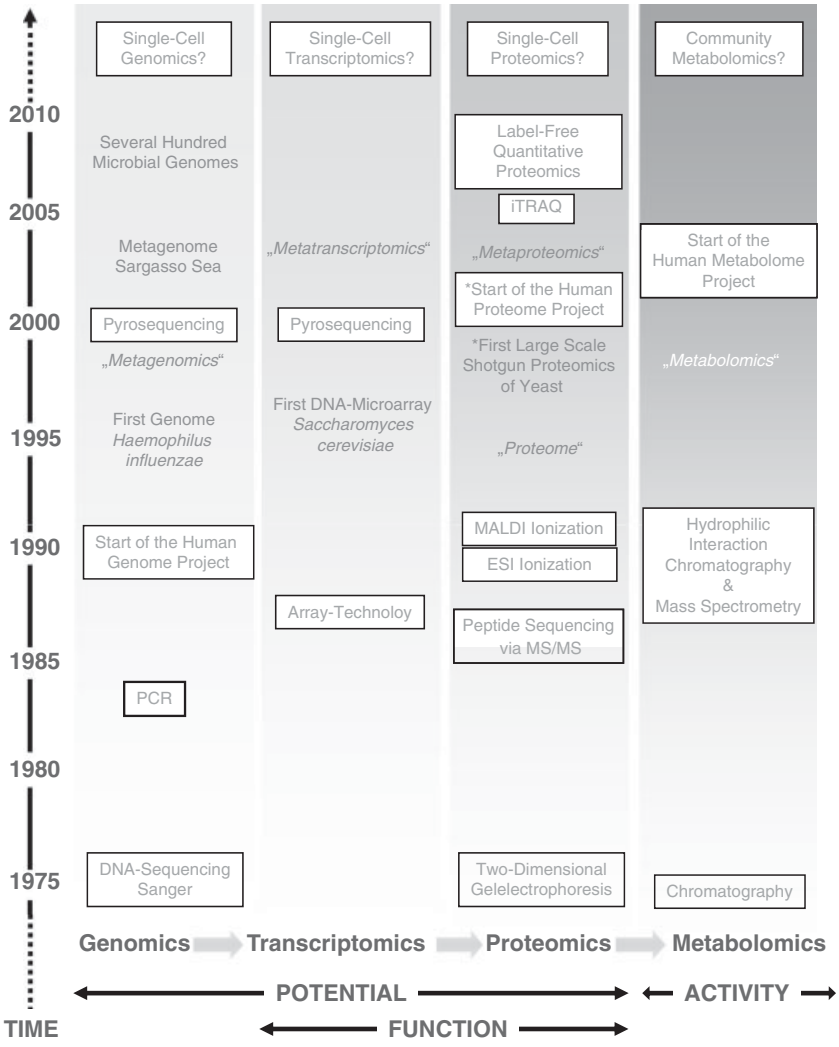

Figure 1. Overview of the historical development and interrelationship of the different ''omics'' technologies including methodological innovations (white boxes), introduction of novel terms (quotation marks), and milestone publications or initiatives. *Initiated/published in 2001.

studies have been limited to low- or medium-complexity environments. The term *proteogenomics*, which was initially used to describe the application of proteomics for the enhancement of gene annotations, does nowadays also define the assessment of strain or species variations and the evolutionary development of the genomic makeup of certain environments [9]. Furthermore, proteogenomics contributes to the understanding of the actual gene function by linking identified protein information to the DNA level.

1.4 Potential applications of environmental proteomics

In their natural habitat, microorganisms are often facing expeditious and harsh changes of environmental parameters such as temperature, humidity, nutrient availability, and predators. A common strategy of microbes to overcome these challenges is an alteration of their protein expression profiles. Consequently, the mere study of individual genes and their regulation is not sufficient to fully understand microbial adaptation strategies and post-genomic analyses including transcriptomics and proteomics are urgently needed to investigate the physiology of complex microbial consortia at a molecular level.

Even though still in its infancy, environmental proteomics already comprises a real ''treasure chest'' of technologies ranging from simple protein cataloging (*e.g.* by mapping the protein complement of an ecosystem at a certain time point) to comparative and quantitative proteomics (*e.g.* by evaluating how different environmental conditions affect protein expression), analyses of protein localizations, discovery of post-translational modifications which might affect protein functionality, investigation of protein-protein interactions, and even determination of amino-acid sequences and genotypes, (*e.g.* by strain-resolved proteogenomics).

Hence, potential applications of the above listed technologies in microbial ecology are numerous and include the description of novel functional genes, the identification of

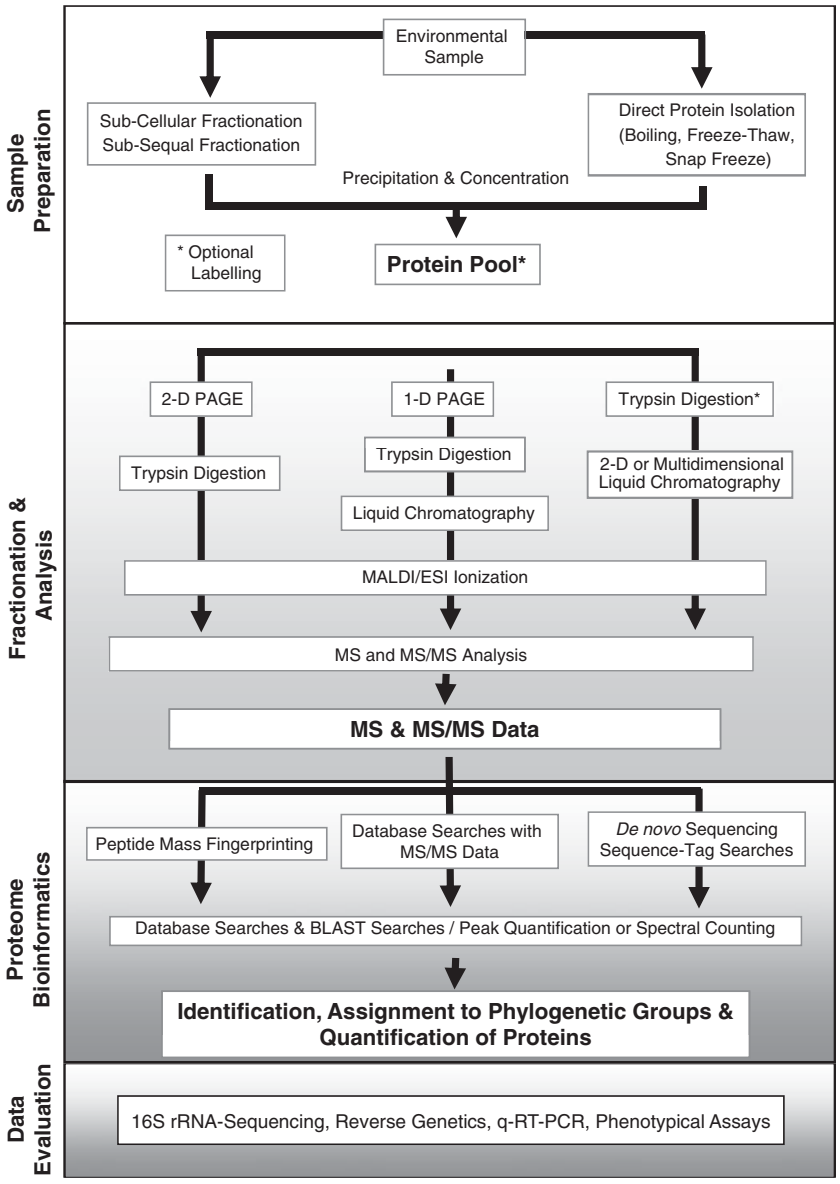

Figure 2. Schematic flow-chart summarizing different environmental proteomics methodologies. Discrete steps, *i.e.* sample preparation, protein fractionation and analysis, protein identification and optional quantification (indicated by *), proteome bioinformatics and data evaluation are depicted in boxes. The various protein analysis strategies exhibit significant differences in throughput, sensitivity, and reliability of identification, which are discussed in detail in the text.

completely new catalytic enzymes or entire metabolic pathways, and the description of functional bioindicators to monitor dynamics and sustainability of environmental quality (reviewed in [29]). Further improvement and concerted usage of the complete set of "omics" technologies will allow us to revisit microbial ecology concepts by linking genetic and functional diversity in microbial communities and relating taxonomic and functional diversity to ecosystem stability.

2 State-of-the-art proteomics technologies

A standard proteomics experiment typically comprises four basic steps (Fig. 2): (i) sample preparation including protein extraction, purification, and concentration; (ii) protein denaturation and reduction; (iii) protein (or peptide) separation, enzymatic digestion, and MS analysis; and (iv) protein identification based on the obtained MS and/or MS/MS data. The trustworthiness of an environmental proteome analysis can be further increased if the obtained data is validated by complementary methods, *e.g.* transcriptome analyses or (if applicable) phenotypical assays. Metaproteomics encompasses similar experimental setups although it needs to overcome additional challenges inherent in samples from natural environments, *e.g.* high organism/protein complexity, over- or under-representation of certain organisms/proteins, heterogeneity of organic and inorganic contaminants, *etc.* The following paragraphs will give an overview of state-of-the-art proteomics techniques focusing on the requirements of environmental proteomics and will discuss weaknesses and strengths of different experimental strategies.

2.1 Sample preparation

The first critical step in a metaproteome study is the comprehensive extraction of the entire protein complement of a given sample; the protocol for this should be as efficient, non-biased, and reproducible as possible. Moreover, it is crucial to avoid the addition of organic or inorganic compounds/solvents that might interfere with sequential protein separation and MS. Depending on sample type and complexity, different extraction strategies have to be employed. Generally, a pre-fractionation of the protein complement of an environment prior to analysis, *e.g.* based on protein solubility, phylogenetic origin, or cellular localization of proteins is recommended to reduce sample complexity. Examples for the application of such pre-fractionation strategies in environmental proteomics have been recently published: (i) the extraction and analysis of soluble protein fractions of sheep rumen and termite hindgut [30, 31] or the partial solubilization of proteins from bacterial cells collected from ocean water samples [32]; (ii) the enrichment of bacterial cells *via* sequential centrifugation steps from *Riftia pachyptila* or the human distal gut [33, 34]; and (iii) the separation of extracellular and cytoplasmic protein fractions from an AMD biofilm [35]. Where intracellular or cell-bound proteins will be investigated, microbial cells have to be lysed either by detergent-containing buffers, sonication, or French press treatment.

Some environments with a low spatial distribution of microbes or where the organisms or proteins of interest are bound to a matrix such as soil or sediment thwart any pre-fractionation strategies and the entire protein complement has to be extracted at once. For activated sludge, a multi-step protocol has been developed which includes various washing buffers, French press lyses and precipitation [21, 28, 36]. Protein extraction from soil is mainly hampered by the presence of perturbing matrix compounds (*e.g.* humic acids), and requires even harsher extraction procedures, *e.g.* snap-freeze protein extraction [37], hydrofluoric acid to dissolve soil minerals [38], or NaOH extraction followed by phenol treatment [39]. The NaOH extraction method was also used for sediments [40]. The relatively low number of proteins identified so far from soil- and sediment-derived samples (see following paragraphs) demonstrates the need for improved protein extraction methods in order to obtain sufficiently concentrated and purified protein samples from complex environments for downstream analyses.

Before the extracted proteins can be further analyzed, compounds which might hamper the separation, enzymatic digestion, and/or MS measurements, *i.e.* nucleic acids, lipids, or polysaccharides, have to be removed. This can be achieved either by precipitation with trichloroacetic acid, acetone or ethanol followed by resolubilizing the proteins in an appropriate buffer, or *via* 1-D SDS-PAGE. Even though precipitation often leads to high protein losses, the resolubilized proteins or, more precisely, the subsequently generated peptides can be directly subjected to gel-free multi-dimensional LC analysis.

2.2 Protein denaturation and reduction

After protein extraction and purification, individual polypeptides must be (i) denatured to disrupt intra- and intermolecular interactions, (ii) reduced to prevent the re-oxidation of disulfide bonds, and (iii) protected against unspecific proteolysis. This can be achieved by employing sample buffers that contain chaotropes (*e.g.* urea and/or thiourea), nonionic and/or zwitterionic detergents (*e.g.* Triton X-100 or CHAPS) or ionic detergents (*e.g.* SDS), reducing agents (*e.g.* DTT, dithioerythritol, or tributylphosphine), and protease inhibitors (reviewed in [41]).

2.3 Protein/peptide separation and MS analyses

2.3.1 Protein or peptide separation by gel-based or chromatographic techniques

Protein samples derived from natural environments do not lend themselves to direct MS analysis; rather, sample complexity has to be reduced first by gel-based or chromatographic techniques. This can be accomplished either on the protein level or on the peptide level after proteolytic degradation of sample proteins. For many years 2-D PAGE was regarded as the ''gold standard'' of proteomics research [26, 42]. With this method proteins are first separated along a pH gradient by IEF, followed by a second separation according to mass on SDS-PAGE gels. In this way, over a thousand proteins can be resolved on a single gel as discrete spots. Staining the gels (*e.g.* with silver, Coomassie blue, or fluorescent dyes) allows the relative determination of protein abundances based on protein spot size and intensity. Protein spots can be subsequently excised and digested in-gel (most commonly with trypsin, see below) prior to mass spectrometric analysis. A significant improvement of this technology was introduced in the late 1990s, when it became possible to label different samples with fluorescent dyes and pool these samples before PAGE (DIGE [43]), thereby reducing gel-to-gel variations. This method is commonly used in combination with 2-D PAGE (2-D DIGE). Despite having been frequently employed in various environmental studies (see following paragraphs), 2-D PAGE suffers from several weaknesses. Most notably, proteins with extreme molecular masses and isoelectric points as well as membrane proteins are difficult to analyze, and co-migration of proteins and protein isoforms hampers accurate identification and quantification. The method is also labor-intensive and consequently hardly automatable and not suited for high-throughput analyses. In the last decade, one- or multi-dimensional LC coupled to MS has emerged as a promising alternative to 2-D PAGE (reviewed in [44–46]). An experimental strategy that has proven extremely useful for the analysis of membrane proteins or highly polluted samples (where contaminants might interfere with trypsin digestion) is the separation of proteins by 1-D PAGE,

followed by in-gel digestion of excised protein bands and separation of the resulting peptides by RP chromatography. Finally, the so-called multidimensional protein identification technology (MudPIT) first described by Washburn *et al.* [47] and reviewed in [45], employs two- or multidimensional chromatography to separate peptides generated by trypsin digestion, and is gaining more and more momentum for environmental proteomics studies. The combination of various chromatographic modes that separate peptides according to different properties, *e.g.* strong-cation exchange/RP, strong-anion exchange/RP, hydrophilic liquid interaction chromatography/RP, or high-pH-RP/low-pH-RP, yields a very significant increase in resolving power suitable to address the enormous complexity of environmental samples. Moreover, these approaches allow a high level of automation, *i.e.* an online connection of chromatography and MS enabling the generation of thousands of mass spectra *per* hour and thus greatly facilitating high-throughput analyses.

2.3.2 Enzymatic digestion of proteins

Depending on the proteomics approach, peptides will be generated by enzymatic digestion (most often trypsin, but also chymotrypsin, Glu-C, Lys-C, and Asp-N are used) before or after proteins have been separated by either gel- or LC-based methods. In the former case, proteins can be digested in-gel, which is largely impervious to contaminants that might interfere with digestion; however, peptide recovery from the gel may be incomplete and the method cannot be easily automated. In the latter case, proteolysis can be performed in solution, which is more readily automatable and minimizes sample handling, but is also more susceptible to interfering substances.

2.3.3 Ionization and MS

After sample complexity has been reduced sufficiently as described above, ionization of analytes and MS can be applied to either proteins (''top-down approach'') or peptides (''bottom-up'' approach) [48]. Bottom-up approaches determine the mass of intact peptides as well as peptide fragment ions, which adds information on peptide amino acid composition and sequence and thus provide much more detailed results. Regardless of the approach, the most frequently used ionization techniques are (i) MALDI ([49]), where the ionization of matrix-embedded peptides is triggered by a laser beam and (ii) ESI ([50]), where the ionization is achieved by dispersing a peptide-containing liquid by electrospray; these ion sources can then be coupled to various mass analyzers, most commonly TOF or ion traps. Successful environmental proteomics studies require tandem mass spectrometers with high resolution, sensitivity, and mass accuracy; moreover, automation, *e.g.* coupling LC separation directly to ESI-MS (''shotgun proteomics''), greatly facilitates the analyses of numerous, complex samples. State-of-the-art mass spectrometers such as hybrid quadrupole TOF analysers, FT-ICR mass spectrometers [51], or LTQ-Orbitrap mass spectrometers [52] allow rapid and sensitive targeted MS/MS on LC time scales and highly accurate mass determination in the low-ppm to sub-ppm range. Prospective improvements of mass spectrometer accuracy, resolution, and sensitivity will further boost the application of MS for metaproteome analyses.

2.4 Data analysis and protein identification

There are generally two main routes for protein identification based on database searches: (i) PMF matches peptide masses measured by MS with those calculated *in silico* for each protein entry in the database and (ii) MS/MS determines peptide masses and generates additional peptide sequence information. Both methods rely on the presence of the respective protein sequence information in the reference database; such data are mainly generated by genome or metagenome sequencing projects.

2.4.1 PMF

As mentioned above, PMF depends on database entries of almost complete and correct coding sequences. Moreover, considering the complexity of environmental samples, many peptides might share similar mass-to-charge ratios. Thus, PMF is not well suited for large-scale environmental proteomics studies, especially where only scarce sequence information is available. In spite of these limitations, PMF was used for the identification of (i) proteins from the human infant gastrointestinal tract, which were separated by 2-D PAGE [53] and (ii) of proteins from activated sludge [36]. While the first study only identified few proteins due to the poor genomic data available for the gut microbiome at that time, the latter study could benefit from a metagenomic library generated from similar samples.

2.4.2 Identification of proteins based on MS and MS/MS data

MS/MS has emerged as a highly reliable tool to identify proteins and was employed in most of the environmental proteome analyses (Table 1). In contrast to PMF, MS/MS considers the masses of fragment ions that have been generated by the fragmentation of specific parent ions. This information together with the mass of intact peptides can then be used to generate peptide sequence information if a protein/peptide with similar fragmentation characteristics is present in the database [59]. Software packages such as

Table 1. Overview of milestone proteomics approaches to study structure and function of microbial communities (modified from [9])

Environment	Estimated no. of species or phylotypes/expressed proteins[a)]	Methodology	No. of identified proteins	Reference
Trophosome of *R. pachyptila*	1/ 3.0×10^3	2-D PAGE and MALDI-ToF MS 1-D PAGE, 2-D LC and Q-ToF MS/MS	220	[33]
Acid mine drainage biofilm	6/ 1.8×10^4	2-D LC and LTQ MS/MS	2033	[35]
Acid mine drainage biofilm		2-D LC and LTQ MS/MS	n.d.	[20]
Acid mine drainage biofilm		2-D LC and LTQ MS/MS	2752	[18]
Acid mine drainage biofilm		2-D LC and LTQ MS/MS	2382	[19]
Waste water treatment reactor	17–268/ 5.1×10^4–8.0×10^5	2-D PAGE and MALDI-ToF MS/MS	109	[54]
Sludge EPS		1-D PAGE, LC and Qtrap MS/MS	10	[55]
Sludge		2-D PAGE and MALDI-ToF MS, Q-ToF MS/MS	46	[28]
Sludge		2-D LC and LTQ MS/MS, Orbitrap, MS/MS	2378	[21, 36]
Leaf phyllosphere	~100/ 3.0×10^5	1-D PAGE, LC and Orbitrap MS/MS	2883	[22]
Higher termite hindgut	~200/ 6.0×10^5	3-D LC and LCQ-MS/MS	n.d.	[31]
Sheep rumen	~20 dominant bacterial species/ 6.0×10^4	1-D PAGE and MS/MS	4	[30]
Infant gastrointestinal tract	100–1000/ 3.0×10^5–3.0×10^6	2-D PAGE and MALDI-Tof MS	1	[53]
Human distal gut		2-D LC and Orbitrap MS/MS	3234	[34]
Estuary	100–100 000/ 3.0×10^5–3.0×10^7	2-D PAGE + LC and Q-T of MS/MS	3	[56]
Ocean		2-D LC and LTQ MS/MS	1042	[32]
Lake and soil	1×10^6/ 3.0×10^9	2-D LC and Q-Tof MS/MS	513	[38]
Contaminated soil, groundwater		1-D or 2-D PAGE + LC and MS/MS	59	[39]
Contaminated aquifer sediment		2-D PAGE and Nano-LC MS/MS + LTQ MS/MS	23	[40]
Groundwater		2-D LC and Orbitrap MS/MS	>2500	[57]

Studies are listed according to increasing habitat complexity. n.d., not described.

a) Estimated numbers of species and proteins present in the respective environment are based on average environmental microbial genome size of 3 Mbp and 1 kbp of sequence coding for one gene (modified from [58]).

Mascot [60], SEQUEST [61], or X!tandem [62] allow high-throughput analyses of thousands of uninterpreted experimental MS and MS/MS spectra that can be generated by a single MS run. The newest generation of mass spectrometers provides MS/MS data with sufficient mass accuracy to deduce the exact amino acid sequence of peptides [63].

2.4.3 Sequence tagging and *de novo* sequencing

Less than 10 years ago, an error-tolerant methodology termed *peptide sequence tagging* was developed that collates partial peptide sequence information to peptide mass for database searching and thus allows for differences caused by post-translational modifications, amino acid substitutions, or other variations between the theoretical and measured peptide mass [64].

Peptide *de novo* sequencing seeks to predict the entire sequence of a peptide based on MS/MS spectra of peptide fragment ions. The methodology is especially useful for the identification/characterization of proteins for which no homologue exists in the database and highly valuable for metaproteome analyses of unexplored microbial communities. Examples of prominent *de novo* software packages are PEAKS [65] and Sequit [66], which are able to reconstruct an entire peptide sequence from MS/MS spectra without a reference database. However, reliable full-length *de novo* peptide sequencing remains an elusive goal, and even the most accurate algorithms can only reconstruct

30–45% of peptides [67], as often a complete set of peptide fragment-ions needed to create a full sequence is missing.

2.4.4 Comparative proteome analysis and protein quantitation

High-throughput acquisition of qualitative and especially quantitative environmental proteome data critically depends on bioinformatic tools capable of handling large and heterogeneous data sets. Several recently developed software packages, *e.g.* DTAselect [68] or Scaffold [69–71], are able to sort and filter enormous amounts of data and to compare different samples by counting the peptide spectra that were assigned to a particular protein; this allows a (semi-)quantitative evaluation of protein abundances in environmental samples [72]. A method often used in combination with spectral counting is the calculation of the normalized spectral abundance factor, which takes into account that larger proteins tend to contribute more peptides to a MS analysis than smaller ones [73, 74]. Spectral counts are divided by protein length giving the so-called spectral abundance factor (SAF). The SAF is then normalized by dividing it by the sum of all SAFs that have been obtained in one analysis, which allows a comparison of protein levels across different MS measurements.

2.5 Data evaluation

Analogous to conventional proteome analyses, metaproteomics data should be evaluated by complementary approaches, *e.g.* 16S rRNA-pyrosequencing or fluorescent *in situ* hybridization to assess community composition, and also transcriptome analyses and RT-PCR approaches to confirm the expression of protein coding genes at the mRNA level. If applicable, phenotypical assays (*e.g.* measurement of enzyme activities) should be employed to confirm the expression and functionality of certain proteins. Naturally, the isolation of RNA and active proteins from environmental samples is hampered by sample complexity and interfering contaminants, thus a comprehensive validation of data obtained by metaproteome analyses remains challenging.

3 Current environmental proteomics studies – where are we so far?

Even though microorganisms are of major importance for every biological system as they contribute to global nutrient cycling, organic matter decomposition, eutrophization, and many other processes, the application of community or metaproteome analyses to study structure and function of uncultivable microorganisms or microbial communities in their natural environment is still limited (Table 1). In the following paragraphs we will discuss milestone publications in detail, starting with low-complexity communities, *e.g.* biofilms in AMDs, followed by medium-complexity habitats, *e.g.* animal and human intestinal tracts, up to highly complex habitats, *e.g.* aqueous or soil environments.

3.1 Community proteomics of marine symbionts of *R. pachyptila*

The deep-sea tube worm *R. pachyptila* harbors a specialized organ, the trophosome, filled with sulphide-oxidizing endosymbiotic bacteria that provide the worm with carbon, nitrogen, and other nutrients. An intracellular and membrane protein reference map based on metagenomic data of the endosymbionts [75] was created by 2-D PAGE coupled to MALDI-TOF-MS and 1-D PAGE combined with 2-D LC-MS/MS [33]. It showed that the bacteria simultaneously express enzymes of the Calvin cycle and the reductive tricarboxylic acid (TCA) cycle to fix CO_2. Moreover, the comparison of protein profiles derived from sulphide-rich and sulphide-depleted environments indicated that the *Riftia* endosymbionts repress the expression of energetically costly sulphide-oxidation-related enzymes and the key Calvin cycle enzyme RubisCo in favor of less ATP-consuming TCA cycle enzymes when H_2S is limited [33].

3.2 Whole-community proteomics of natural AMD mixed biofilms

An outstanding example for a comprehensive shotgun proteomics approach (2-D LC-MS/MS) is represented by the study of Ram *et al.* [35] who compared the protein complement of two natural biofilms present in an acid mine drainage. These biofilms conveniently exhibit a comparatively low complexity due to the extreme conditions of their habitat. The authors identified more than 2000 proteins from the five most abundant species and obtained a remarkable 48% protein coverage for the dominant biofilm organism *Leptospirillum* group II. In further analyses of the same biofilms, Lo *et al.* [20] were able to differentiate between peptides of discrete AMD populations and found strong evidence for interpopulation recombination – an approach strongly dependent on a database containing strain-specific genome information. This method is referred to as ''strain-resolving proteogenomics.'' The study was expanded by Denef *et al.* [18], whose extensive semi-quantitative analysis of 27 distinct AMD biofilm protein profiles revealed that specific environmental conditions select for particular recombinant types thus leading to a fine-scale tuning of microbial populations. More recently, Goltsman *et al.* [19] employed both metagenomics and semi-quantitative community proteomics to analyze a Richmond mine biofilm and identified 64.6 and 44.9% of the predicted

proteins of *Leptospirillum* Groups II and III; the study nicely demonstrates the potential of a simultaneous genome and proteome approach.

3.3 Proteome analyses of waste water treatment plants and activated sludge

Lacerda *et al.* [54] investigated the response of a natural community in a continuous-flow wastewater treatment bioreactor to an inhibitory level of cadmium by 2-D PAGE combined with MALDI-TOF/TOF-MS and *de novo* sequencing. The authors observed a significant shift in the community proteome after cadmium shock, as indicated by the differential expression of more than 100 proteins including ATPases, oxidoreductases, and transport proteins. Park *et al.* [55] analyzed the protein complement of extracellular polymeric substances of activated sludge flocs by 1-D PAGE combined with LC-MS/MS and identified a limited number of bacterial but also human polypeptides, among them proteins associated with bacterial defense, cell appendages, outer membrane proteins and a human elastase. In 2004, Wilmes and Bond [28] studied the molecular mechanisms of enhanced biological phosphorus removal (EBPR) by a comparative metaproteome analysis of two laboratory wastewater sludge microbial communities with and without EBPR performance by 2-D PAGE combined to MALDI-TOF-MS. Major differences in protein expression profiles between the two reactors were detected. A short time later, more than 2300 proteins were identified by 2-D LC-MS/MS analyses of activated sludge [21, 36], aided by reference metagenomic data from studies of EBPR sludge [76]. The obtained data indicated that the uncultured polyphosphate-accumulating bacterium ''Candidatus *Accumulibacter phosphatis*'' is dominating the microbial community of the EBPR reactor and further enabled an extensive analysis of metabolic pathways, *e.g.* denitrification, fatty acid cycling, and glyoxylate bypass, all central to EBPR.

3.4 Community proteogenomics of phyllosphere bacteria

Very recently, Delmotte *et al.* [22] combined a culture-independent metagenome and metaproteome approach to study the microbiota associated with leaves of soybean, clover and *Arabidopsis thaliana* plants. Phyllosphere bacteria were washed from the leaves and DNA and proteins were extracted and analyzed by pyrosequencing and 1-D PAGE LC-MS/MS resulting in the identification of 2883 proteins. The majority of the proteins were related to *Methylobacterium*, *Sphingomonas*, and *Pseudomonas*, indicating the predominance of these genera within the phyllosphere community. Functional assignments of the proteins suggested that phyllosphere Methylobacteria are able to exploit methanol as carbon and energy source and that Sphingomonads possess a particularly large substrate spectrum on plant leaves.

3.5 Community proteomics of animal intestinal tracts

Warnecke *et al.* [31] employed a combined genomics and multidimensional-LC-MS/MS proteomics approach to investigate the microbial community present in the hindgut of higher wood-feeding termites. For peptide separation, they used a 2-D approach consisting of three steps: RP LC followed by an SCX-chromatography and an additional RP LC step. This three step system has been used successfully to improve the resolving power of LC leading to an increased number of identified proteins in a complex sample [77].] The authors reported the presence of a large set of bacterial enzymes involved in the degradation of cellulose and xylan and other important symbiotic functions such as H_2 metabolism, CO_2-reductive acetogenesis, and N_2 fixation. In a more recent study, Toyoda *et al.* [30] used 1-D PAGE coupled to MS/MS to identify cellulose-binding proteins derived from sheep rumen microorganisms; among these proteins were endoglucanase F of the cellulolytic bacterium *Fibrobacter succinogenes* and exoglucanase Cel6A of the fungus *Piromyces equi*.

3.6 Community proteomics of human intestinal tracts

Klaassens *et al.* [53] studied the functionality of the uncultured microbiota of human infant stool samples by 2-D PAGE combined with MALDI-TOF-MS. The authors observed time-dependent changes in the gut metaproteome, but were not able to identify more than one protein exhibiting high similarity to a bifidobacterial transaldolase due to at that time limited microbiome sequence information. Finally, Verberkmoes *et al.* [34] identified several thousand proteins present in two female twin fecal samples by an extensive semi-quantitative shotgun proteome analysis, among them bacterial proteins involved in well-known but also undescribed microbial pathways and human antimicrobial peptides.

3.7 Metaproteome analyses of ocean water

One of the very first metaproteome analyses was presented by Kan *et al.* [56], who compared protein profiles from various sample origins of the Chesapeake Bay by 2-D PAGE and tried to identify protein spots excised from the gels by an LC-MS/MS approach; however, the obtained information was rather limited, as a substantial DNA sequence background was still lacking. Recently, Sowell *et al.* [32] published a comprehensive study of the Sargasso Sea surface metaproteome. The authors employed 2-D LC

coupled to MS/MS and identified over 1000 proteins, among them an overwhelming number of SAR11 periplasmic substrate-binding proteins as well as *Prochlorococcus* and *Synechococcus* proteins involved in photosynthesis and carbon fixation. High abundance of SAR11 transporters as determined by spectral counting (see Section 2.4.4) suggests that cells endeavor to maximize nutrient uptake activity and thus gain a competitive advantage in nutrient-depleted environments.

3.8 Metaproteome studies of highly complex groundwater and soil environments

Schulze *et al.* [38] presented an interesting functional insight into the complex microbial communities present in dissolved organic matter from lake water and seepage water adhering to soil micro-particles. Although the number of proteins identified by 2-D LC-MS/MS was comparatively low, the authors were able to assign functional proteins to broad taxonomic groups and observed rather unexpected seasonal variations of the protein complement. Notably, decomposing enzymes were only found among proteins extracted from soil particles, thereby indicating that the degradation of soil organic matter mainly takes place in biofilm-associated communities. More recently, Benndorf *et al.* [39] published a metaproteome analysis of protein extracts from contaminated soil and groundwater employing either 1-D or 2-D PAGE combined with LC and MS/MS. Proteome analyses of soils mainly suffer from numerous inorganic and organic contaminants, which hamper protein separation and identification; thus, only 59 proteins could be identified although the authors presented a multi-step purification protocol combining NaOH treatment and phenol extraction. A similar approach was employed to investigate the metaproteome of an anaerobic benzene degrading community inhabiting aquifier sediments [40]. Even though only a handful of proteins were identified – among them an enoyl-CoA hydratase involved in the anoxic degradation of xenobiotics – the authors demonstrated that their metaproteome extraction method is potentially valuable to investigate sediment microbial communities. Recently, Wilkins *et al.* [57] studied a microbial community present in uranium-contaminated groundwater and identified more then 2500 proteins by a shotgun analysis. MS/MS data were searched against a database containing all predicted ORFs of the genomes of community dominating Geobacter strains; moreover, the relative abundance of proteins from samples collected during acetate amendment was analyzed by spectral-counting. Relative protein quantitation reflected major changes in community metabolism in response to biostimulation and indicated the importance of energy generation during enhanced growth on acetate. Thus, the authors propose that proteogenomics can be used to diagnose the metabolic state of microbial communities involved in bioremediation.

4 Future perspectives

When viewed in relation to its enormous potential, the actual output of environmental proteomics appears so far to be disappointingly limited. Present studies have mainly focused on microbial communities with a relatively low diversity or dominated by a particular phylogenetic group. The main obstacles toward a comprehensive metaproteome coverage seem to be (i) the irregular species distribution within environmental samples, (ii) the wide range of protein expression levels within microbial cells, and (iii) the enormous genetic heterogeneity within microbial populations [8]. It is encouraging to note, however, that constantly improving extraction methods alongside advances in downstream MS technology and a steadily growing pool of bioinformatics data might soon help to overcome the current challenges and limitations of metaproteomics research.

4.1 Improvements of mass spectrometer sensitivity and accuracy

A successful environmental proteomics experiment entails the reliable identification not only of the predominant but also of low-abundance proteins, which intimately depends on ultrasensitive and highly accurate mass spectrometers. A significant and foreseeable improvement in MS performance will enable us to (i) identify low-abundance but important gene products, *e.g.* proteins involved in transcriptional and translational regulation, (ii) evaluate the spatial distribution of proteins within complex habitats, and (iii) investigate proteins on such a fine scale that we might even envision single cell proteome analyses.

4.2 Quantitative environmental proteomics

Another important future line of research will take advantage of the increasing power of metaproteomics tools to quantitatively analyze/compare protein expression rates in environmental samples. Despite its well-known drawbacks, 2-D PAGE has dominated quantitative protein expression studies until recently; moreover, quantitative proteomics has been restricted to biological systems of low or limited complexity [33, 53, 54, 56].

Nowadays, 2-D gel-free LC-MS-based technologies have emerged as powerful tools for comparative/quantitative proteome studies and might be applied to in-depth, quantitative proteome profiling of complex environments. Recently, label-free techniques [72], which are based on counting fragment spectra of peptides used to identify a certain protein have been employed for quantitative environmental proteome analyses (see Section 2.4.4, [18–21, 32, 34, 35]). An advantage of label-free approaches is their comparably large dynamic range, which is of particular

importance when multifaceted and large protein changes within different samples have to be anticipated [78]. Because samples have to be analyzed separately and are therefore liable to experimental variations, it is crucial that sample preparation and analysis become highly standardized and reproducible [78].

A revolutionary development in the field of quantitative proteomics was the introduction of isotope- or isobar-tag based technologies, *e.g.* ICAT [79], iTRAQ [80], and ANIBAL [81], which enable the analysis of different samples in a single MS measurement. However, so far none of these methods was used to assess protein expression in complex environmental samples, which might be due to the fact that these techniques are relatively costly, can only be applied to a limited number of samples and need considerable post-processing of the original samples. Moreover, the development of commensurate hardware and, even more importantly, software tools for label-based quantitative proteomics lags behind the advances in MS [82, 83].

All these quantitative approaches rely on highly accurate MS data to reduce interfering signals and on reproducible peptide chromatography to be able to correlate similar (identical) peptides across different samples. Promising tools that meet these requirements are the newest generation of electrospray ultra-high resolution TOF mass spectrometers, which might even have the potential to generate more reliable quantitative data from complex environments.

5 Concluding remarks

Proteomics is one of today's fastest developing research areas and has contributed substantially to our understanding of individual organisms at the cellular level. Its attractiveness stems from being able to probe many protein functions and responses simultaneously, and seems also ideally suited to improve our knowledge of the complex interplay between the constitution of a habitat, diversity, and architecture of microbial communities and ecosystem functioning. Recently, a limited number of studies describing large-scale proteome analyses of environmental samples have demonstrated the huge potential of metaproteomics to unveil the molecular mechanisms involved in function, interaction, physiology, and evolution of microbial communities. Moreover, the rapidly growing number of genomic and metagenomic sequences together with revolutionary advances in protein analysis and bioinformatics have opened up a completely new range of applications, *e.g.* studying the impact of environmental changes upon protein expression profiles of entire microbial communities (''quantitative metaproteomics'') or measuring low-level protein expression differences in order to resolve the functional significance of spatial protein distribution within a given environment. In conclusion, the comprehensive knowledge gained by the concerted application of system-level approaches such as genomics, transcriptomics, proteomics and metabolomics will greatly advance our understanding of biogeochemical cycles and will facilitate the biotechnological harnessing of microbial communities or uncultivable organisms.

The authors would like to thank Alexander Grunau for critically reading the manuscript. Parts of the work presented here were generated in the frame of the Austrian research network MICDIF and were supported by the Austrian Science Foundation (FWF).

The authors have declared no conflict of interest.

6 References

[1] Brock, T. D., The study of microorganisms *in situ*: progress and problems. *Symp. Soc. Gen. Microbiol.* 1987, *41*, 1–17.

[2] Amann, R. I., Ludwig, W., Schleifer, K. H., Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* 1995, *59*, 143–169.

[3] Schloss, P. D., Handelsman, J., Toward a census of bacteria in soil. *PLoS Comput. Biol.* 2006, *2*, e92.

[4] Vieites, J. M., Guazzaroni, M. E., Beloqui, A., Golyshin, P. N., Ferrer, M., Metagenomics approaches in systems microbiology. *FEMS Microbiol. Rev.* 2009, *33*, 236–255.

[5] Cardenas, E., Tiedje, J. M., New tools for discovering and characterizing microbial diversity. *Curr. Opin. Biotechnol.* 2008, *19*, 544–549.

[6] Singh, J., Behal, A., Singla, N., Joshi, A. *et al.*, Metagenomics: concept, methodology, ecological inference and recent advances. *Biotechnol. J.* 2009, *4*, 480–494.

[7] Keller, M., Hettich, R., Environmental proteomics: a paradigm shift in characterizing microbial activities at the molecular level. *Microbiol. Mol. Biol. Rev.* 2009, *73*, 62–70.

[8] Wilmes, P., Bond, P. L., Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. *Curr. Opin. Microbiol.* 2009, *12*, 310–317.

[9] VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., Banfield, J. F., Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* 2009, *7*, 196–205.

[10] Lacerda, C. M., Reardon, K. F., Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology. *Brief. Funct. Genomic. Proteomic.* 2009, *8*, 75–87.

[11] Lopez-Barea, J., Gomez-Ariza, J. L., Environmental proteomics and metallomics. *Proteomics* 2006, *6*, 51–62.

[12] Schulze, W. X., Environmental proteomics – what proteins from soil and surface water can tell us: a perspective. *Biogeosci. Discuss.* 2004, *1*, 195–218.

[13] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L. *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004, *304*, 66–74.

[14] Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K. *et al.*, IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008, *36*, D534–538.

[15] Pignatelli, M., Aparicio, G., Blanquer, I., Hernandez, V. *et al.*, Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics* 2008, *24*, 2124–2125.

[16] Tringe, S. G., Rubin, E. M., Metagenomics: DNA sequencing of environmental samples. *Nat. Rev.* 2005, *6*, 805–814.

[17] Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A. *et al.*, Comparative metagenomics of microbial communities. *Science* 2005, *308*, 554–557.

[18] Denef, V. J., VerBerkmoes, N. C., Shah, M. B., Abraham, P. *et al.*, Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environ. Microbiol.* 2009, *11*, 313–325.

[19] Goltsman, D. S. A., Denef, V. J., Singer, S. W., VerBerkmoes, N. C. *et al.*, Community genomic and proteomic analysis of chemoautotrophic, iron-oxidizing "*Leptospirillum rubarum*" (Group II) and *Leptospirillum ferrodiazotrophum* (Group III) in acid mine drainage biofilms. *Appl. Environ. Microbiol.* 2009, *75*, 4599–4615.

[20] Lo, I., Denef, V. J., Verberkmoes, N. C., Shah, M. B. *et al.*, Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 2007, *446*, 537–541.

[21] Wilmes, P., Andersson, A. F., Lefsrud, M. G., Wexler, M. *et al.*, Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J.* 2008, *2*, 853–864.

[22] Delmotte, N., Knief, C., Chaffron, S., Innerebner, G. *et al.*, Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl. Acad. Sci.* USA 2009, *106*, 16428–16433.

[23] Ahmadian, A., Ehn, M., Hober, S., Pyrosequencing: history, biochemistry and future. *Clin. Chim. Acta* 2006, *363*, 83–94.

[24] Zhou, J., Thompson, D. K., Challenges in applying microarrays to environmental studies. *Curr. Opin. Biotechnol.* 2002, *13*, 204–207.

[25] Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D. *et al.*, Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* 1995, *13*, 19–50.

[26] O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 1975, *250*, 4007–4021.

[27] Bochner, B. R., Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* 2009, *33*, 191–205.

[28] Wilmes, P., Bond, P. L., The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* 2004, *6*, 911–920.

[29] Maron, P. A., Ranjard, L., Mougel, C., Lemanceau, P., Metaproteomics: a new approach for studying functional microbial ecology. *Microb. Ecol.* 2007, *53*, 486–493.

[30] Toyoda, A., Iio, W., Mitsumori, M., Minato, H., Isolation and identification of cellulose-binding proteins from sheep rumen contents. *Appl. Environ. Microbiol.* 2009, *75*, 1667–1673.

[31] Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M. *et al.*, Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 2007, *450*, 560–565.

[32] Sowell, S. M., Wilhelm, L. J., Norbeck, A. D., Lipton, M. S. *et al.*, Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* 2009, *3*, 93–105.

[33] Markert, S., Arndt, C., Felbeck, H., Becher, D. *et al.*, Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* 2007, *315*, 247–250.

[34] Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A. *et al.*, Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 2009, *3*, 179–189.

[35] Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W. *et al.*, Community proteomics of a natural microbial biofilm. *Science* 2005, *308*, 1915–1920.

[36] Wilmes, P., Wexler, M., Bond, P. L., Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS ONE* 2008, *3*, e1778.

[37] Singleton, I., Merrington, G., Colvan, S., Delahunty, J. S., The potential of soil protein-based methods to indicate metal contamination *Appl. Soil Ecol.* 2003, *23*, 25–32.

[38] Schulze, W. X., Gleixner, G., Kaiser, K., Guggenberger, G. *et al.*, A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia* 2005, *142*, 335–343.

[39] Benndorf, D., Balcke, G. U., Harms, H., von Bergen, M., Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J.* 2007, *1*, 224–234.

[40] Benndorf, D., Vogt, C., Jehmlich, N., Schmidt, Y. *et al.*, Improving protein extraction and separation methods for investigating the metaproteome of anaerobic benzene communities within sediments. *Biodegradation* 2009, *20*, 737–750.

[41] Görg, A., Weiss, W., Dunn, M. J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics* 2004, *4*, 3665–3685.

[42] Klose, J., Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 1975, *26*, 231–243.

[43] Ünlü, M., Morgan, M. E., Minden, J. S., Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 1997, *18*, 2071–2077.

[44] Lane, C. S., Mass spectrometry-based proteomics in the life sciences. *Cell. Mol. Life Sci.* 2005, *62*, 848–869.

[45] Motoyama, A., Yates, J. R., III, Multidimensional LC separations in shotgun proteomics. *Anal. Chem.* 2008, *80*, 7187–7193.

[46] Peng, J., Gygi, S. P., Proteomics: the move to mixtures. *J. Mass Spectrom.* 2001, *36*, 1083–1091.

[47] Washburn, M. P., Wolters, D., Yates, J. R., III, arge-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotech.* 2001, *19*, 242–247.

[48] Chait, B. T., CHEMISTRY: Mass Spectrometry: Bottom-Up or Top-Down? *Science* 2006, *314*, 65–66.

[49] Hillenkamp, F., Karas, M., Beavis, R. C., Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* 1991, *63*, 1193A–1203A.

[50] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989, *246*, 64–71.

[51] Marshall, A. G., Hendrickson, C. L., Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* 1998, *17*, 1–35.

[52] Hu, Q., Noll, R. J., Li, H., Makarov, A. *et al.*, The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* 2005, *40*, 430–443.

[53] Klaassens, E. S., de Vos, W. M., Vaughan, E. E., Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl. Environ. Microbiol.* 2007, *73*, 1388–1392.

[54] Lacerda, C. M., Choe, L. H., Reardon, K. F., Metaproteomic analysis of a bacterial community response to cadmium exposure. *J. Proteome Res.* 2007, *6*, 1145–1152.

[55] Park, C., Helm, R. F., Novak, J. T., Investigating the fate of activated sludge extracellular proteins in sludge digestion using sodium dodecyl sulfate polyacrylamide gel electrophoresis. *Water Environ. Res.* 2008, *80*, 2219–2227.

[56] Kan, J., Hanson, T. E., Ginter, J. M., Wang, K., Chen, F., Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Syst.* 2005, *1*, 7.

[57] Wilkins, M. J., Verberkmoes, N. C., Williams, K. H., Callister, S. J. *et al.*, Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation. *Appl. Environ. Microbiol.* 2009, *75*, 6591–6599.

[58] Wilmes, P., Bond, P. L., Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* 2006, *14*, 92–97.

[59] Hunt, D. F., Yates, J. R., III, Shabanowitz, J., Winston, S., Hauer, C. R., Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* 1986, *83*, 6233–6237.

[60] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[61] Eng, J. K., McCormack, A. L., Yates, J. R., III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[62] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[63] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007, *4*, 787–797.

[64] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 2002, *66*, 4390–4399.

[65] Ma, B., Zhang, K., Hendrie, C., Liang, C. *et al.*, PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2337–2342.

[66] Demine, R., Walden, P., Sequit: software for *de novo* peptide sequencing by matrix-assisted laser desorption/ ionization post-source decay mass spectrometry. *Rapid Commun. Mass Spectrom.* 2004, *18*, 907–913.

[67] Frank, A., Pevzner, P., PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[68] Tabb, D. L., McDonald, W. H., Yates, J. R., DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002, *1*, 21–26.

[69] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[70] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.

[71] Craig, R., Beavis, R. C., A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2306–2310.

[72] Gilchrist, A., Au, C. E., Hiding, J., Bell, A. W. *et al.*, Quantitative proteomics analysis of the secretory pathway. *Cell* 2006, *127*, 1265–1281.

[73] Florens, L., Carozza, M. J., Swanson, S. K., Fournier, M. *et al.*, Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 2006, *40*, 303–311.

[74] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. *et al.*, Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[75] Robidart, J. C., Bench, S. R., Feldman, R. A., Novoradovsky, A. *et al.*, Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environ. Microbiol.* 2008, *10*, 727–737.

[76] Garcia Martin, H., Ivanova, N., Kunin, V., Warnecke, F. *et al.*, Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.* 2006, *24*, 1263–1269.

[77] Wei, J., Sun, J., Yu, W., Jones, A. *et al.*, Global proteome discovery using an online three-dimensional LC-MS/MS. *J. Proteome Res.* 2005, *4*, 801–808.

[78] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 2007, *389*, 1017–1031.

[79] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 1999, *17*, 994–999.

[80] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. *et al.*, Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[81] Panchaud, A., Hansson, J., Affolter, M., Bel Rhlid, R. *et al.*, ANIBAL, stable isotope-based quantitative proteomics by aniline and benzoic acid labeling of amino and carboxylic groups. *Mol. Cell. Proteomics* 2008, *7*, 800–812.

[82] Choi, H., Fermin, D., Nesvizhskii, A. I., Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* 2008, *7*, 2373–2385.

[83] Ono, M., Shitashige, M., Honda, K., Isobe, T. *et al.*, Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics* 2006, *5*, 1338–1347.

