Review TRENDS in Microbiology Vol.14 No.2 February 2006

Metaproteomics: studying functional gene expression in microbial ecosystems

Paul Wilmes[1] and Philip L. Bond[1,2]

[1]School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, UK, NR4 7TJ
[2]School of Biological Sciences, University of East Anglia, Norwich, Norfolk, UK, NR4 7TJ

The recent availability of extensive metagenomic sequences from various environmental microbial communities has extended the postgenomic era to the field of environmental microbiology. Although still restricted to a small number of studies, metaproteomic investigations have revealed interesting aspects of functional gene expression within microbial habitats that contain limited microbial diversity. These studies highlight the potential of proteomics for the study of microbial consortia. However, the application of proteomic investigations to complex microbial assemblages such as seawater and soil still presents considerable challenges. Nonetheless, metaproteomics will enhance the understanding of the microbial world and link microbial community composition to function.

Metaproteomics: the missing functional link?

During most of the geological past of the Earth, microorganisms have had primary roles in shaping the environmental conditions that exist today. They are the major drivers of the biogeochemistry, the nutrient cycles and the degradation of natural and anthropogenic waste products of the planet [1]. Many of these important reactions are catalysed by microbial enzymes that are made up of individual proteins and that can, on a grand scale, be regarded as the catalysts that ensure planetary homeostasis. Consequently, it makes sense to study the vast range of microbial proteins in the different ecosystems where they occur. This is reflected in the emerging field of metaproteomics – the proteomic analysis of mixed microbial communities.

In microbiology, the application of postgenomic techniques has been limited mainly to laboratory studies of pure cultures. Consequently, these studies do not provide information about gene expression in the complex mixtures of microorganisms as found in the biosphere, where they form the backbone of a given ecosystem. The study of environmental microbial communities has been hampered by enrichment bias, in which standard culturing techniques select mostly for easily culturable organisms and result in a distorted understanding of microbial ecology. Cultivation-independent studies of microbial community composition that have been performed during the past 25 years have provided insight into the vastness of microbial diversity [2]. These investigations highlight the problem of how best to study the complex microbial biomass. Techniques must be developed that enable the study of microbial populations *in situ*, respecting both community structure and natural habitat. The emergence of metagenomics (the genomic analysis of uncultured microorganisms [2]) has led to the initiation of community sequencing projects during the past two years, which have resulted in the publication of extensive metagenomic sequences from different microbial populations [3–6]. These studies highlight the extent of microbial genetic and functional diversity.

With the availability of metagenomic sequences and the increasing number of complete individual genome sequences, it is now possible to apply postgenomic techniques – particularly proteomics – to complex microbial communities [7]. These studies provide insight into the functional dimensions of environmental genomic datasets and will help to achieve a major goal of environmental microbiology: the ability to link individual microbial species to function. In contrast to substrate-specific approaches [e.g. fluorescence *in situ* hybridization with microautoradiography (FISH–MAR) [8] and stable isotope probing (SIP) [9]], transcriptomic and proteomic analyses can detect physiological responses to changes in various environmental conditions. So far, the application of oligonucleotide microarrays to address environmental questions has been limited to the phylogenetic analysis of microbial communities [10]. However, functional microarrays have considerable potential for environmental monitoring [11]. For example, a functional oligonucleotide array based on 2402 known genes and pathways has been designed to monitor microbial populations that are involved in biodegradation and metal resistance in soil [12]. Although microarrays provide rapid generation of large functional datasets, there are numerous challenges and potential pitfalls associated with this approach. The functional analysis described [12] was based solely on genes and pathways that had been revealed through the study of laboratory isolates. This approach can lead to an information bias. With regard to specificity, the generation of explicit functional data from oligonucleotide microarrays remains a challenge and some arbitrary similarity

Corresponding author: Wilmes, P. (p.wilmes@uea.ac.uk).
Available online 9 January 2006

 0966-842X/$ - see front matter doi:10.1016/j.tim.2005.12.006

levels must be determined empirically (the microarray in the described study [12] could differentiate sequences with <88% similarity). Furthermore, it is now widely accepted (mainly from work using yeast and mammalian cells) that mRNA expression and protein expression do not always correlate directly [13]. By contrast, proteomic investigations of microbial communities are less restrictive [7,14] because they are not based on assumptions of which genes are important. Furthermore, with the generation of *de novo* peptide sequences, proteomic studies are specific with regard to protein identity and phylogenetic origin. Because protein expression is a reflection of specific microbial activities in a given ecosystem, proteomics has great potential for the functional analysis of microbial communities. We believe that the elucidation of metaproteomic expression will be central to functional studies of microbial consortia. Metaproteomics will gain momentum with the advent of further environmental sequencing projects and it will provide a useful tool with which to focus research direction. Here, we describe recent work that demonstrates the power of proteomics in microbial ecology research.

Functional profiling of complex environments at opposite ends of Crick's central dogma

Genomic fingerprinting

Tringe *et al.* [6] used a gene-based bioinformatic approach to interpret environmental genomic data obtained from different microbial ecosystems. By clustering the various functional groups, the authors found that the predicted protein complement of a community is influenced by its environment. However, little can be concluded regarding the importance of genes in microbial ecosystems based on their presence alone. Consequently, the expression of individual genes in different microbial niches must be demonstrated before conclusions can be reached regarding the interpretation and diagnosis of microbial activities in these environments.

Proteomic fingerprinting

Schulze *et al.* [15] used mass spectrometry (MS)-based proteomics [high mass-accuracy tandem MS (MS–MS)] to analyse the protein complement of water that contained high levels of dissolved organic matter from four different environments (a peat bog lake, soil from an unmanaged deciduous forest, soil from a managed evergreen spruce forest and acidic soil from beneath a spruce). Their aim was to determine the phylogenetic groups from which individual proteins originated and to elucidate the potential catalytic function of these proteins in the sampled ecosystems. The overall soil protein composition was investigated and the majority of proteins was of bacterial origin; for example, 78% of proteins in surface water obtained from a peat bog lake were bacterial. In forest leachate, the number of proteins that originated from plants, fungi and vertebrates was approximately twice the number in the sampled lake water. Interestingly, the number of bacterial proteins in forest soil was greater in winter than in summer. When all of the detected enzymes with known functions were compared, differences between the proteins from the lake and those from the deciduous forest were observed; for example, peroxidases involved in the degradation of complex molecules were found only in water from forest soil, whereas transferases that might be involved in methane production were found only in the water from the peat bog. Although this study detected only a small fraction of the proteins that are present in the sampled environments, it reflects the presence and activity of different taxonomic groups and demonstrates the potential of proteomic fingerprints by indicating overall changes in ecosystem biology that were, in this case, not limited to archaeal and bacterial constituents. Metaproteomic investigations of environmental samples will be most potent when coupled with information about species diversity and richness within the ecosystem. Thus, ecological niches with simpler microbial communities [e.g. acid mine drainage (AMD) biofilm, whale-fall sites and activated sludge] should be easier to understand than more-complex ecosystems (e.g. forest, ocean and soil) from their proteomic signatures.

Although Tringe *et al.* [6] and Schulze *et al.* [15] investigated the fingerprinting of environmental samples at opposite ends of the 'molecular biology spectrum' and although the scale of investigation differed, both studies provide interesting insights into the biology of ecosystems and demonstrate the large potential of environmental fingerprinting approaches based on either DNA [6] or protein [15].

Protein expression in microbial communities with limited diversity

Metaproteomics has been used to decipher the metabolic details of a treatment process for activated sludge wastewater that operates for phosphorus removal [7,16]. The process, known as enhanced biological phosphorus removal (EBPR), is desirable because it enables the microbiological removal of phosphate from refuse water and, thus, limits eutrophication in surface waters that receive effluent. The treatment process is characterized by alternating anaerobic and aerobic phases, with most of the sludge being recycled. This arrangement selects for specific polyphosphate-accumulating organisms (PAOs) such as uncultured organisms that are related to *Rhodocyclus* species. EBPR is well established in civil engineering; however, numerous details of the microbiology and biochemistry of the process are unknown. A comparative study of metaproteomic expression between the anaerobic and aerobic phases in two distinct activated sludges (one with 'good' EBPR performance and one with 'poor' EBPR performance) has led to interesting insights into differential protein expression patterns [16]. In this study, it was demonstrated that protein expression within the well-performing sludge (dominated by the typical *Rhodocyclus*-type PAOs) did not alternate as much during the anaerobic–aerobic sludge cycling compared with the poor-performing sludge [16]. An interesting possibility is that the more equilibrated protein expression in the well-performing sludge provides a bioenergetic advantage to PAOs in these alternating systems.

Recently, extensive metagenomic sequences from similar EBPR sludges that are enriched with *Rhodocyclus*-type PAOs cultured in Australia and the USA have become

available (P. Hugenholtz *et al.*, unpublished). Previously, protein identification relied on time-consuming and expensive *de novo* peptide sequences being generated using MS [7] but this is now less constrained because of the availability of relevant metagenomic sequence data. For example, using such data, ~30% of proteins excised from 2D-PAGE gels can be identified by comparing their respective peptide mass fingerprints with DNA sequences using the Mascot algorithm (P. Wilmes and P.L. Bond, unpublished; http://www.matrixscience.com). Using this approach, numerous proteins have been identified that can be directly linked to previously suggested metabolic models of EBPR (P. Wilmes and P.L. Bond, unpublished). For example, strong expression of a potential phosphate-selective membrane porin, proteins involved in polyhydroxyalkanoate synthesis and proteins linked to glycogen synthesis and degradation has been revealed.

The most extensive metaproteomic investigation successfully combined 'shotgun' MS proteomics with community genomic analyses [14]. This comprehensive study analysed the protein complement of a low-complexity natural biofilm growing inside the Richmond Mine at Iron Mountain, CA (USA), an environment that is characterized by low pH (~0.8), medium temperature (~42°C) and high levels of heavy metals. Community analysis revealed that the sampled AMD biofilm was dominated by *Leptospirillum* group II but also contained *Leptospirillum* group III, *Sulfobacillus* and Archaea related to *Ferroplasma acidarmanus* and 'G-plasma'. (G-plasma' is a member of the 'alphabet-plasmas': the putatively proposed phylogenetic group names of clones obtained from Iron Mountain biofilms [17]) Using the extensive genomic dataset of a similar biofilm [4], a database of 12 148 predicted protein sequences was constructed. This enabled the proteins to be identified from their respective mass spectra and assigned to the different constituent organisms. Using stringent filter parameters that required at least two peptides per gene from the generated mass spectra dataset, 2033 individual proteins were positively identified, 48% of which originated from the dominant *Leptospirillum* group II organism. Proteins directly linked to survival challenges in this extreme environment (e.g. chaperones, thioredoxins and peroxiredoxins) were abundant. A large proportion of proteins could not be assigned to particular functional categories and were termed novel gene products. For example, the extracellular fraction contained 52% unique proteins (that were not significantly similar to any known proteins) and ~14% conserved novel proteins (that were annotated as 'hypotheticals' in the original genomic dataset [4]). Among these extracellular proteins, the three most highly expressed were encoded by genes of unknown function. One of the proteins originated from *Leptospirillum* group II and was identified as a novel cytochrome (Cyt_{579}) central to iron oxidation and, consequently, AMD formation.

The metaproteomic investigations of EBPR and the AMD biofilm both studied low-complexity microbial communities. The application of metaproteomics to more-complex microbial communities presents a considerable challenge. Nonetheless, the EBPR and AMD studies demonstrate the power of metaproteomics for the elucidation of functional details that are crucial to investigations of microbial assemblages.

Challenges and pitfalls

As mentioned, using the proteomic approach to elucidate functional components of microbial ecosystems has huge potential for the study of environmental microbiology. However, unlike other cellular macromolecules, proteins exist in many different biological and physical conformations. Consequently, there is no universal extraction protocol that can be followed. The exhaustive extraction of proteins, especially from complex environments such as seawater and soil, poses considerable challenges [18]. Another important consideration is the *in situ* activity of isolated proteins. Several techniques, including native gels and activity stains, will be used to assess the activity of particular proteins in future metaproteomic investigations. The power of using multidimensional protein-separation systems followed by MS for environmental proteomics is highlighted by Ram *et al.* [14]. The percentage of proteins assigned to the dominant biofilm organism *Leptospirillum* group II exceeded that of previous proteomic studies using pure cultures. The combination of fractionation (using liquid isoelectric focusing) and large-format (24 cm) 2D-PAGE gels with single pH unit immobilized pH gradient strips can resolve ~5000 individual protein spots from activated sludges enriched with *Rhodocyclus*-type PAOs (P. Wilmes and P.L. Bond, unpublished). The two proteomic investigation strategies are highlighted in Figure 1. Although both methods have strengths and weaknesses, the 'old-fashioned' proteomic approach based on 2D-PAGE retains a clear advantage with regard to quantification. Protein spot intensities and sizes on 2D gels reflect expression levels more accurately than do abundances inferred from MS data. Whereas some investigations have focused mainly on cytosolic proteins [7,16], Ram *et al.* [14] included cellular fractionations of their biofilm metaproteome (extracellular and periplasmic, whole-cellular, soluble and membrane fractions). Therefore, future metaproteomic investigations that include cellular fractionation could determine interesting functional aspects of protein localization.

Although the number of proteins that have been identified and separated using available proteomic analysis methods is encouraging, the diversity of other microbial ecosystems still poses enormous challenges. Based on published estimates of prokaryotic diversity in a range of different environments and by using crude arithmetic, we have estimated the number of proteins expressed in samples from microbial habitats across the whole diversity spectrum (Table 1). Both the AMD biofilm and the activated sludge enriched with *Rhodocyclus*-type PAOs are at the lower end of the diversity scale. For complex environments (e.g. seawater and soil) that are typically of more interest to microbial ecologists, one can expect to resolve only a minute fraction ($\ll$1%) of the metaproteome with the methods that are currently available. Furthermore, in addition to the vast dynamic range of microbial species distribution within one sample,

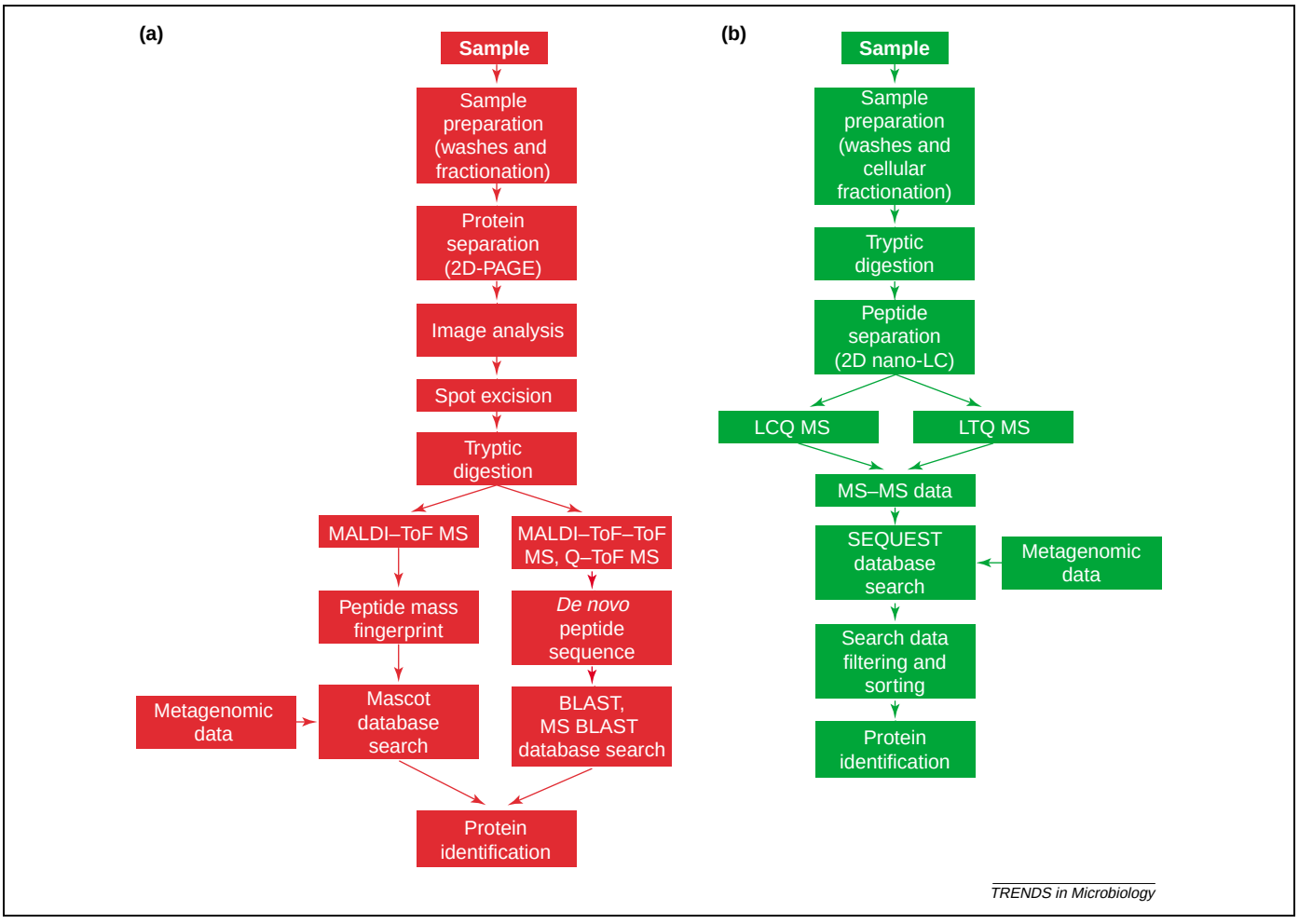

Figure 1. Metaproteomic analysis strategies. **(a)** 2D-PAGE-based analysis [7] (P. Wilmes and P.L. Bond, unpublished). **(b)** Liquid-chromatography-based analysis [14]. Abbreviations: LCQ: liquid chromatography electrospray 3D quadrupole ion trap; LTQ: liquid chromatography electrospray 2D linear ion trap; MALDI–ToF: matrix-assisted laser desorption–ionization time-of-flight; nano-LC: nano-liquid chromatography; Q–ToF: quadrupole time-of-flight. Mascot (http://www.matrixscience.com) and SEQUEST (http://fields.scripps.edu/sequest/) are algorithms that enable MS data to be searched against amino acid and/or nucleotide databases in order to obtain positive protein identifications.

protein expression levels within one cell might differ by six orders of magnitude [19]. Consequently, substantial improvements in the technologies for protein extraction, separation and identification are necessary to encompass these two vast dynamic ranges and are an important consideration for environmental microbiologists and analytical chemists.

The assembly of metagenomic sequences from diverse environments into distinct genomes is a considerable challenge [6]. Protein identification is easier with the availability of extensive metagenomic sequences. However, without environmental sequence data, proteins obtained from environmental samples can be identified reliably from their respective *de novo* peptide sequences by searching against the current databases using the MS basic local alignment search tool (BLAST) algorithm [20]. This offers an alternative metaproteomic identification route that does not require complete metagenomic sequences.

Future perspectives

Despite the limited number of environmental proteomic investigations that has been carried out, it is clear that metaproteomics has huge potential in the field of environmental microbiology. Although it remains a daunting task to elucidate all of the functional proteins that are contained within an environmental sample, metaproteomics will find immediate use in studies

Table 1. Microbial diversity and metaproteomic expression

Ecosystem	Estimated number of individual taxa	Estimated number of expressed proteins[a]	Refs
AMD biofilm	6	1.8×10^4	[4]
Activated sludge	17–268	5.1×10^4–8.0×10^5	[24]
Ocean water (1 ml)	160	4.8×10^5	[25]
Sargasso sea (combined sample, i.e. 1730 l)	1824–47 733	5.5×10^6–1.4×10^8	[5]
Soil (1 g)	1×10^6	3.0×10^9	[26]

[a]Estimated numbers of proteins based on average environmental microbial genome size of 3 Mbp and 1 kbp of sequence coding for one gene.

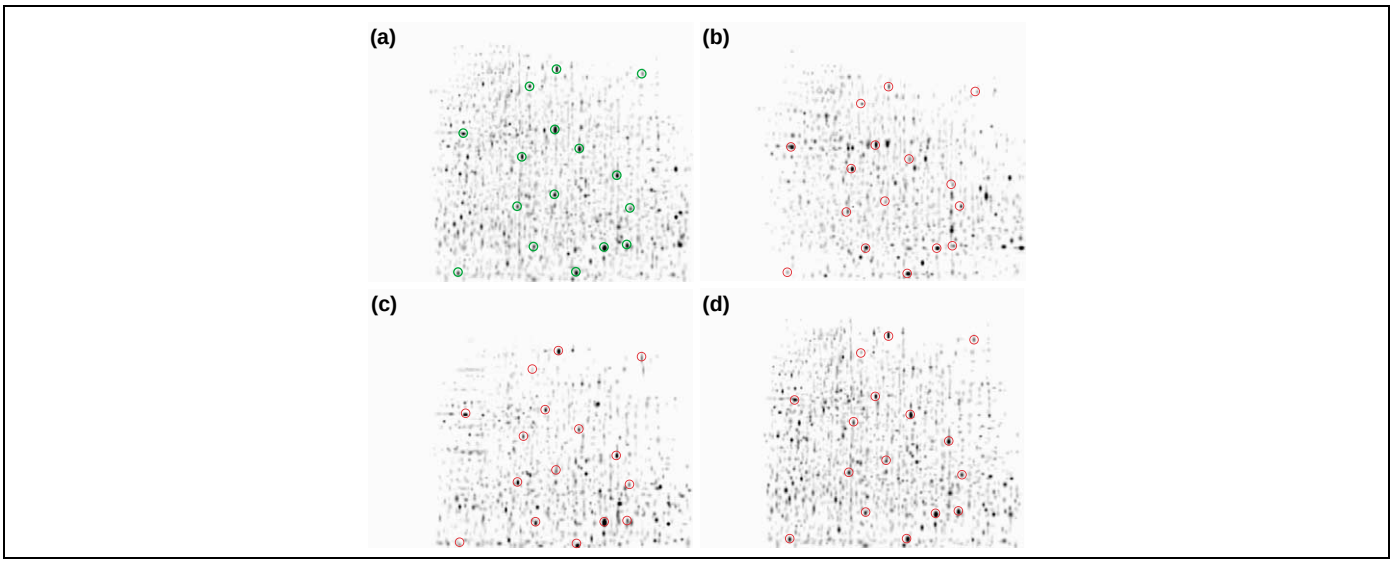

Figure 2. Metaproteomic fingerprint of activated sludge after enhanced biological phosphorus removal. 2D-PAGE gels have a pH range of 4–7. **(a)** Composite gel with conserved protein constellations circled in green; **(b–d)** representative gels at 25, 40 and 55 mg l^{-1} phosphorus removal with conserved protein constellations circled in red.

focusing only on parts of expressed environmental proteomes. Investigations that focus on limited numbers of highly expressed proteins can have immediate impacts on developments in the field. For example, activated sludges with sequentially increased phosphorus removal performance were analysed by 2D-PAGE. The resulting gels contained several conserved protein constellations that provided a proteomic fingerprint of EBPR (Figure 2). The identification of such proteomic markers could lead to the development of functional protein arrays for monitoring the activity of microbial communities in environmental systems or, in the case of activated sludge, accurate process control. Alternatively, these markers form good targets for identification by western blot analysis.

Metagenomic sequence data comprise many genes of unknown function, implying a high abundance of previously unknown functional systems [6]. This is where metaproteomics might prove invaluable in future investigations. Functional elucidation attempts must be prioritized and metaproteomics could provide the relevant decision criteria (i.e. the functions of highly expressed unknown genes should be investigated first).

Environments that undergo periodic fluctuations such as diurnal changes in redox and energetic conditions are obvious targets for comparative metaproteomic investigations. Examples include tidal sediments and phytoplankton blooms, which are important in nutrient cycling. The detection of differential expression and protein modifications such as phosphorylations and thiol–disulfide conversions that are important for enzymatic activity [21] could be performed to compare metaproteomic changes in response to such fluctuating environmental conditions.

Along with metaproteomics, several other functional methods that are aimed at directly linking microbial diversity to function *in situ* have recently become available to microbial ecologists. These methods include simultaneous mRNA and rRNA FISH [22], functional microarrays [12] and stable isotope probing (for review, see Ref. [23]). The integrative use of these techniques will provide more-comprehensive views of microbial function.

In addition to verifying hypothetical metabolic processes gained from empirical observations and metagenomic data, metaproteomics provides a tool for identifying functional genes of primary importance. This is illustrated by the identification of enzymes that are central to EBPR biochemistry and by the discovery of a novel cytochrome (Cyt_{579}) that is central to iron oxidation in the AMD biofilm investigation. Metaproteomics can also be used as an important decision tool to prioritize the investigation of elusive functional genes from large metagenomic datasets. Future metaproteomic investigations of complex microbial communities still face numerous challenges with regard to techniques of proteomic extraction, separation and identification. By identifying functional components of the biogeochemical cycles of the Earth, metaproteomics will provide a new dimension to environmental catalysis.

Acknowledgements

We thank the reviewers and Margaret Wexler for helpful suggestions regarding the manuscript. P.W. was supported by a School of Environmental Sciences studentship, a British Council Chevening Scholarship, a short-term research scholarship from the German Academic Exchange Service and a research scholarship from the Luxembourg Ministry of Culture, Higher Education and Research.

References

1 Rodriguez-Valera, F. (2004) Environmental genomics, the big picture? *FEMS Microbiol. Lett.* 231, 153–158
2 Schloss, P.D. and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14, 303–310
3 Schmeisser, C. *et al*. (2003) Metagenome survey of biofilms in drinking water. *Appl. Environ. Microbiol.* 69, 7298–7309
4 Tyson, G.W. *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43
5 Venter, J.C. *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74
6 Tringe, S.G. *et al*. (2005) Comparative metagenomics of microbial communities. *Science* 308, 554–557

7 Wilmes, P. and Bond, P.L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* 6, 911–920
8 Lee, N. *et al*. (1999) Combination of fluorescent in situ hybridization and microautoradiography – a new tool for structure–function analyses in microbial ecology. *Appl. Environ. Microbiol.* 65, 1289–1297
9 Radajewski, S. *et al*. (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403, 646–649
10 Loy, A. *et al*. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognised lineages of sulfate-reducing prokaryotes in the environment. *Appl. Environ. Microbiol.* 68, 5064–5081
11 Sharkey, F.H. *et al*. (2004) Detection and quantification of gene expression in environmental bacteriology. *Appl. Environ. Microbiol.* 70, 3795–3806
12 Rhee, S-K. *et al*. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl. Environ. Microbiol.* 70, 4303–4317
13 Pradet-Balade, B. *et al*. (2001) Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* 26, 225–229
14 Ram, R.J. *et al*. (2005) Community proteomics of a natural microbial biofilm. *Science* 308, 1915–1920
15 Schulze, W.X. *et al*. (2005) A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia* 142, 335–343
16 Wilmes, P. and Bond, P.L. Towards exposure of elusive metabolic mixed-culture processes: the application of metaproteomic analyses to activated sludge. *Water Sci. Technol.* (in press)
17 Baker, B.J. and Banfield, J.F. (2003) Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* 44, 139–152
18 Ogunseitan, O.A. and LeBlanc, J. (2004) Environmental proteomics: methods and applications for aquatic ecosystems. In *Molecular Microbial Ecology Manual* (2nd edn) (Kowalchuk, G.A. *et al*., eds), pp. 1027–1046, Springer-Verlag
19 Tyers, M. and Mann, M. (2003) From genomics to proteomics. *Nature* 422, 193–197
20 Shevchenko, A. *et al*. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI–quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 73, 1917–1926
21 Leichert, L.I. and Jakob, U. (2004) Protein thiol modifications visualized *in vivo*. *PLoS Biol.* 2, e333
22 Pernthaler, A. and Amann, R. (2004) Simultaneous fluorescence *in situ* hybridisation of mRNA and rRNA in environmental bacteria. *Appl. Environ. Microbiol.* 70, 5426–5433
23 Dumont, M.G. and Murrell, J.C. (2005) Stable isotope probing – linking microbial identity to function. *Nat. Rev. Microbiol.* 3, 499–504
24 Wagner, M. and Loy, A. (2002) Bacterial community composition and function in sewage treatment systems. *Curr. Opin. Biotechnol.* 13, 218–227
25 Curtis, T.P. *et al*. (2002) Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. U. S. A.* 99, 10494–10499
26 Gans, J. *et al*. (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* 309, 1387–1390