REVIEW ARTICLE

Metagenomics approaches in systems microbiology
José M. Vieites1, María-Eugenia Guazzaroni2, Ana Beloqui1, Peter N. Golyshin3,4,5 & Manuel Ferrer1

1CSIC, Instituto de Catalysis, Madrid, Spain; 2CSIC, Estación Experimental del Zaidín, Granada, Spain; 3Environmental Microbiology Laboratory, HZI-Helmholtz Centre for Infection Research, Braunschweig, Germany; 4School of Biological Sciences, Bangor University, Gwynedd, UK; and 5Centre for Integrated Research in the Rural Environment, Aberystwyth University-Bangor University Partnership (CIRRE), Aberystwyth, UK

Correspondence: Manuel Ferrer, CSIC, Institute of Catalysis, Marie Curie 2, 28049 Madrid, Spain. Tel.: +34 91 5854928; fax: +34 91 5854760; e-mail: mferrer@icp.csic.es

Received 30 June 2008; revised 23 September 2008; accepted 26 October 2008. First published online 1 December 2008.
DOI:10.1111/j.1574-6976.2008.00152.x

Editor: Michael Galperin

Keywords systems microbiology; systems biology; metagenomics; functional networks; ecosystem functioning; microbial communities.

Abstract

The world of microorganisms comprises a vast diversity of live organisms, each with its individual set of genes, cellular components and metabolic reactions that interact within the cell and communicate with the environment in many different ways. There is a strong imperative to gain a broader view of the wired and interconnected cellular and environmental processes as a whole. Therefore, the systems microbiology approach is necessary to understand and predict ecosystem functioning. On the other hand, currently we experience a rise of metagenomics as an emerging tool to study communities of uncultured microorganisms. In this review, we conducted a survey of important methodologies in metagenomics and describe systems microbiology-like approaches for gaining a mechanistic understanding of complex microbial systems to interrogate compositional, evolutionary and metabolic properties. The review also discusses how metagenomics can be used as a holistic indicator for ecosystem response in terms of matter, nutrient and energy sources and functional networking.

Introduction

The broad aim of systems microbiology, a subset of systems biology, is to acquire an understanding of the wiring diagrams of life, to gain knowledge about the relationships between the individual components that build a cellular organism, a community and an ecological niche. As a complement to the long-standing trend towards reductionism, systems microbiology seeks to treat the community as a whole, integrating fundamental biological knowledge to ultimately create an integrated picture of how a microbial cell or community operates. This is not an easy task as we know that microbial communities vary dynamically in response to both spatial and temporal gradients as well as to nutrient and energy fluxes and, more importantly, that most microorganisms are members of complex communities, with diversity including 100s or 1000s of distinct taxa (Lozupone & Knight, 2008). Whereas, traditionally, systems microbiology has focused on pure cultures or enrichments (McHardy & Rigoutsos, 2007), it is becoming increasingly apparent that to understand ‘whole’ microbial communities one must study the ‘whole’. Such understanding requires proper techniques that reveal (1) the size and genome information of whole microbial communities to access the phylogenetic and genomic complexity (the so-called ‘part list’), (2) the role of predominant members or individual phylotypes within particular environmental niches to identify major players and microbial signatures related to the bio- and geochemical characteristics and (3) the functional composition and metabolic networks (connectivity between ‘parts’) to understand and predict ecosystem functioning and microbial impact on ecosystems. Only by considering an environmental sample as a whole, that is, as a ‘macro-organism’, will we understand the network complexity and the whole set of metabolic routes that profoundly affect the ecosystem structure and hence influence major environmental processes such as organic matter mineralization.

For technical reasons, the diversity and population density of microorganisms inhabiting the biosphere (6.3 × 10^30 kg total biomass) is too high to be studied using traditional cultivation strategies (Whitman et al., 1998; McHardy & Rigoutsos, 2007; Raes et al., 2007; Beloqui et al., 2008). Therefore, one should consider other alternatives for studying microbial communities. Members of a community are often metabolically interdependent or utilize specific resources, which limit their cultivability and the success of isolation. Following on from this, where the phylogenetic diversity has been compared or analysed, the
number of microbial interactions and processes seems to be underestimated (see e.g. by Warnecke et al., 2007) because the size, nature, structure and organization of metabolic circuits is up to 1000-fold higher than the microbial context: 1 g of soil may contain up to $10^9$ bacteria (Schloss & Handelsman, 2006), and assuming 3000 genes per single genome, there will be up to $3 \times 10^{12}$ genes mediating 1.2 $\times$ 10$^{12}$ putative reactions (assuming that 40% genes have catalytic activity) (Dinsdale et al., 2008). In this review, we describe the application of the emerging field of metagenomics to systems microbiology to answer the following questions: How diverse are metabolic pathways and networks? How do microorganisms and protein-coding genes interact with each other to lead to the overall function of the system? How many specific matter, nutrient and energy sources are metabolized by different microorganisms? How many specific microorganisms and functions metabolize different substrates? How do environmental stimuli impact ecosystem functioning and long-term stability as a whole? How can we obtain this global information?

**Interplay of microbial complexity and metagenomics**

The global outcome of microbial metabolic processes is the integration of interactions with a global significance on very small scales. Two forms of interactions may be established to accomplish biogeochemical cycles that are necessary to sustain life on Earth. The first is microbiological. Microbial communities behave as groups of microorganisms that interact and, together, accomplish more than those same organisms would separately (Lozupone & Knight, 2008). The second form of interaction is chemical interdependence. A continuum of interactions, ranging from obligate to minimal, is thought to exist among members of microbial communities. Whatever the case, microbial communities, in which members interact, are distinct from microbial assemblages, in which members merely co-exist. Apart from understanding the vast diversity of global microbial species, the issue of whether microbial species are ubiquitous or are more endemic and limited to certain geographical areas, was raised about 100 years ago. Current evidence indicates that a large fraction of microorganisms are not cosmopolitan, but are instead restricted to a specific habitat type and geographic location. However, there are a few examples of truly cosmopolitan organisms, for example the Deep-Sea Marine Group I Archaea (DeLong, 2006) and marine obligate hydrocarbon-degrading Gammaproteobacteria, i.e. Alcanivorax (Yakimov et al., 2007). As new, higher resolution technologies become available, further research may reveal other microorganisms to be more widespread than previously thought.

It is essential to note that the relative abundance of representatives of a certain group of microorganisms is not necessarily linked to the importance of that group in the functioning of the community. Common organisms may not necessarily play a critical role in a community despite their numbers, and organisms that only muster 0.1% of the community (e.g. nitrogen fixers) can be of pivotal importance (Dinsdale et al., 2008). Systematic characterization of this pivotal diversity will provide new understanding of metabolic activities and interdependencies underlying microbial life, and the role of individual organisms in the ecosystems. In this context, it should be important to find the microbial and enzymatic complements in different niches and how they determine community functioning. Further, current and future systems microbiology approaches can provide an approach to understand the complex properties of microbial communities, their dynamics and their impacts in natural systems. Systems approaches to microbial communities could also help answer the fundamental questions of environmental microbiology: which organisms are out there and what are they doing. As a first step in these investigations it is necessary to identify the members of the community under study, as well as the interactions in which they are engaged. However, as we mentioned above, according to the well-established dilemma of environmental microbiology, only a minority of microorganisms are readily culturable (see e.g. in Ingham et al., 2007). For this reason, a wide range of approaches collectively described as environmental genomics (‘metagenomics’) have been developed to study such communities without culturing individual organisms, in spite of their complexity and variability (Ferrer et al., 2008). The term ‘metagenomics’ has been used broadly to encompass research ranging from examining environmental DNA in functional screenings and drug discovery (for recent example see Schmeisser et al., 2007) to randomly sampling the genomes from a small subset of organisms present in an environment (Tringe et al., 2005). The main task of the sequencing-based metagenomics is to reconstruct the metabolism of the organisms making up the community, and to predict their functional roles in the ecosystem. A large-scale genomic data analysis may be combined with gene expression analysis (transcriptome) to further identify the genes associated with causal interactions between genes and traits and generate co-expression networks using a collection of DNA sequences (Cavalieri & Grosu, 2004; Ferrara et al., 2008) or for taxon-specific probing (Urisman et al., 2005). However, the sequence variability of genes in organisms even belonging to the same species and the incomplete genomic data due to the complexity of communities, limit application of this method. Recent advances in mass spectrometry have contributed to the solution of this problem, allowing an enormous progress to be made in proteomics.
and metabolomics/interactomics (see an example in Lo et al., 2007). All of these complementary techniques have been applied traditionally to pure cultures of microorganisms, but the recent advances in metagenomics, whole-genome sequencing, high-throughput proteomics and new computational tools now provide a new toolbox for systems microbiology to get an integrated insight into the functioning of the microbial community as a whole and to study cellular and physiological functions of individual organisms in the community in depth (Farber & Lusis, 2008).

**Integrating environmental sequences into systems biology**

There are two distinct strategies in metagenomics, according to the primary goal. Firstly, the activity-based approach involves construction of small- to large-insert expression libraries, especially those made in lambda phage, cosmids or copy-control fosmid vectors, which are further implemented for a direct activity screening (Lorenz & Eck, 2005). The shortcoming of this approach to systems microbiology is that metagenomic libraries do have a limit in size; if one assumes that the mean size of a bacterial genome is 4 Mbp, and that 1 g of soil can contain about 10^7 different genomes and that a metagenomic library is composed of c. 40 000 fosmids each containing 40-kbp DNA fragments, then the library may cover a maximum of 400 different complete bacterial genomes. Once a library is constructed, a critical step is to screen for clones that contain target genes among a large number of clones. In the case of activity-based screening, dozens of 1000s of clones may be analysed in a single screen. Certainly, owing to the limitation of efficient expression of the metagenome-derived genes in the selected host, the numbers of positive clones will not be high. Furthermore, in activity-based screening, it is necessary to develop specialized screening systems to detect the activity of the products of the gene of interest (Ferrer et al., 2008). On the other hand, the application of the alternative, sequence-based approach involves the design of PCR primers or hybridization probes for the target genes that are derived from conserved regions of already known protein families, which a priori limits the chances for obtaining fundamentally new proteins (Ferrer et al., 2008). In contrast to the activity-based approach, the large-scale sequencing of bulk DNA through pyrosequencing techniques, or sequencing the DNA libraries constructed for archiving and sequence homology screening purposes, is aimed at capturing the largest amount of available genetic resources present in the sample or archive for its further data mining, mostly homology-based (see first examples in Schmeisser et al., 2003; Tyson et al., 2004). The clear disadvantage of this approach is its complete reliance on the existing genome annotations (Hallin et al., 2008). However, it treats the environment as a whole microbial community rather than as a single dynamic entity. There is no doubt that genome sequence information is not only the final goal of a particular metagenome project, it also serves as a starting point of so-called functional ‘-omics’ research, and it is now becoming the cornerstone of systems microbiology research even beyond the individual genome sequencing (Church, 2005; Stranger et al., 2007).

Two fundamental differences between a cultured microorganism and an environmental microbial sample have to be considered when analysing sequence data. Firstly, in the case of cultured organisms, the cells used for DNA isolation represent a clonal population and will have the same genomic sequence. In environmental samples, even for organisms that represent the same ‘species’, there are many independent lineages that result in varying degrees of sequence variation (polymorphisms) within that ‘species’ population. That heterogeneity has a significant impact on the assessment of sequence quality, especially on sequence assemblies. Current assemblers are designed to handle single, homogeneous genomes and generate various errors when presented with environmental sequences. Depending on the complexity of the population structure, individual whole genome assemblies of the different lineages may not be feasible even with sufficient sequence coverage. Instead, assembled scaffolds are binned into a population ‘pan genome’ that encodes the metabolic core of the species. This provides valuable genetic data for understanding the evolutionary processes that affect the structure and dynamics of the population (Holmes et al., 2003; Whitaker & Banfield, 2006). The second fundamental difference between cultured organisms and environmental samples is that natural microbial communities are usually highly diverse at multiple taxonomic levels. Although several of the recently analysed communities were simple and allowed genomic and metabolic reconstruction of most of their species members, others inferred diversities in the 1000s of species, without containing dominant ones (Tringe et al., 2005). A whole continuum of species diversities and abundance can exist in-between. What this means in terms of assembly is that the more complex a community is, the less chance there is of getting larger contigs of any single represented genomes. For example, in the case of 150 Mbp of soil metagenomic data, which represented an estimated 3000 species with no dominant ones, the largest scaffold is < 10 kbp and over 99% of the sequence reads do not assemble into contigs (Tringe et al., 2005). Because of the many variables involved in large-scale metagenomic data processing, the overall goals of a sequencing project have to be balanced against the diversity and structure of the community, allowing an estimation of the necessary amount of sequence to achieve a desired assembly depth (Chen & Pachter, 2005; Johnson & Slatkin, 2006; Huson et al., 2007). Therefore, genomic
sequencing of single species has made a great contribution to our understanding of microbial systems, and metagenomics approaches (in which the genomes of a group of organisms are lumped and studied together) will extend that knowledge and add to our understanding of microbial processes. Metagenomics confers the potential to map the metabolisms of microorganisms in space and time (e.g. see Woyke et al., 2006). It may even be possible to identify differences in metabolic potential between different regions using metagenomics approaches. Archiving of genomic and metagenomic samples will be necessary to provide baseline data against which microbial community structures can be assessed over time. In this context, the amount of available genomic information from complete microbial genomes, as well as environmental genomic projects, is increasing exponentially. Comparative analyses of the types, abundance and distribution of two-component system genes across genomes and metagenomes can reveal important contributions of various signal transduction systems to specific organisms and communities and allow correlations with environmental parameters (see examples below).

Even though, historically, shotgun sequencing has been implemented as a standard technique for individual genomes since the mid-1990s, its implication for metagenomes is rather new, with one of the first complete analyses of a microbial community being published in 2003–2004 (Schmeisser et al., 2003; Tyson et al., 2004). In the shotgun sequencing approach, random sampling of DNA fragments are sequenced and then assembled into genomes. Theoretically, a metagenomic library will contain DNA sequences for the majority of the genes in the microbial community. With the aid of powerful assembler computer programs, the snippets of DNA sequences are aligned and reassembled into their original order. Of special interest are two recent papers (Rusch et al., 2007; Yooshef et al., 2007) reporting the analysis of 7.7 million sequencing reads (6.3 billion bp) from microorganisms inhabiting seawater from the Northwest Atlantic to the Eastern Tropical Pacific. Other recent examples of the shotgun sequencing approach are shown in Table 1. In one of the prominent metagenomic studies conducted by Venter et al. (2004) about 1.6 Gbp of unique metagenomic DNA sequences were sequenced from Sargasso Sea samples using the shotgun approach (Venter et al., 2004). That study clearly indicated the existence of far more diverse microbial communities than previously thought (see Fig. 1). However, most of the environmental genomes sequenced to date contain only a few high-abundance species but many low-abundance species in the communities, accounting for a large portion of the total genome size of an environmental sample. The presence of a large amount of DNA fragments from the low-abundance species poses a problem for assembling genomes. To infer the biological functions of a microbial community from sequences, a process named ‘binning’ is used to group these unassembled DNA sequence fragments and small contigs into biologically meaningful ‘bins’, such as phylogenetic groups (Eisen, 2007; Podar, 2007; Chan et al., 2008). As the conventional Sanger method provides large datasets of unassembled DNA sequences and small contigs which are not amenable to genome assembly (for review see Shendure et al., 2004), new methodologies relying on innovative principles have been developed to sequence genomes far faster. Most of them are optimized for obtaining short nucleotide sequences usually several 10s of bases long, often accompanied by mate-pair information, which makes them the best choices for genome resequencing or comparative sequencing when used with the available reference genomes (Bentely, 2006). Among them, 454 pyrosequencing, now commercialized by Roche and being sold under the product names of GS 20 and GS FLX systems, is also suitable for de novo sequencing of microbial genomes due to its higher throughput and coverage, and despite somewhat a shorter read length (Margulies et al., 2005). The development of this fast and inexpensive method for sequencing bacterial genomes has led to an increase in the number of new microbial sequencing projects (see examples in Table 2). In fact, 134 whole metagenomic projects using the 454 technology platform have been reported (Liolios et al., 2008). The first clear application of these tools was to reveal groups of peptide fragments with a relatively high abundance and no known function (Angly et al., 2006; Woyke et al., 2006; Williamson et al., 2008) (Fig. 2). The culture-independent recovery of near-complete microbial genomes and partial recovery of other minor genomes from an environmental sample therefore represent an advance in the study of natural microbial communities.

Typically, the analysis and annotation of metagenomic datasets involve grouping (binning) of sequences using sequence characteristics (e.g. GC content, codon, or oligonucleotide frequencies) or inferred taxonomic affiliation (Teeling et al., 2004). Binning is useful when the diversity is relatively low and can lead to a separation of the various types of cellular metabolisms present in the community. If this is the case, whole-metagenome sequencing will recover a few near-complete microbial genomes and other minor genomes. A good example of the successful use of binning to study such communities comes from work on the acid mine biofilms (Tyson et al., 2004). Also, community genome sequencing and assembly of acid mine drainage (AMD) revealed the dominance of the biofilm by a handful of distinct genome types. This may be attributed to a small number of niches within the AMD system at any time, possibly because the ecosystem is relatively geochemically simple: for example, the dominant electron donors and acceptors are iron, sulphur and oxygen, and the temperature and fluid composition cycle were within a relatively narrow range on an annual time-scale. Once metagenomic
sequences have been assembled and binned, and the coding regions (genes and gene fragments) identified, the next important step in environmental sequence analysis is also laden with challenges when compared with gene prediction in completed microbial genomes. In this context, a number of novel gene prediction approaches have been published, tailored to the specific features of the metagenomic data, including one that can handle short sequences generated using the 454 technology (Krause et al., 2006; Noguchi et al., 2006). However, because environmental genomic data

Table 1. Analysis of microbial communities through shotgun metagenomic sequencing

<table>
<thead>
<tr>
<th>Sample</th>
<th>Library size*</th>
<th>Host or vector system used</th>
<th>Average insert size (kbp)</th>
<th>Biodiversity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sargasso Sea</td>
<td>1 985 561</td>
<td>Bst XI linearized pBR322 derivative</td>
<td>2–6</td>
<td>Samples were dominated by genes from Proteobacteria (primarily subgroups Alpha, Beta, and Gamma) with moderate contributions from Firmicutes, Cyanobacteria, and species in the CFB phyla (Cyanophaga, Flavobacterium, and Bacteroides). Poor sequencing coverage enabled the assembly of only two near-complete genomes. Here, 1.6 Gbp of unique metagenomic DNA sequences were obtained</td>
<td>Venter et al. (2004)</td>
</tr>
<tr>
<td>Human feces</td>
<td>36 769</td>
<td>Zero Blunt TOPO PCR cloning pHOS2 (S3)</td>
<td>0.5–1.0</td>
<td>Study of uncultured viruses in human feces. The most abundant fecal virus was pepper mild mottle virus 72 bacterial phylotypes and one archaeal phylotype were identified. The bacterial phylotypes were assigned to only two divisions, the Firmicutes and the Actinobacteria</td>
<td>Zhang et al. (2006)</td>
</tr>
<tr>
<td>Human distal gut</td>
<td>139 521</td>
<td>pHOS2 (S3)</td>
<td>2–3</td>
<td>Study of uncultured viruses in human feces. The most abundant fecal virus was pepper mild mottle virus 72 bacterial phylotypes and one archaeal phylotype were identified. The bacterial phylotypes were assigned to only two divisions, the Firmicutes and the Actinobacteria</td>
<td>Gill et al. (2006)</td>
</tr>
<tr>
<td>Soil</td>
<td>1129 (Bacteria) 527 (Archaea) 919 (Fungi) 4577 (Viruses)</td>
<td>pCR® 2.1-TOPO (Bacterial, Archaean and Fungal) pSMART (viral)</td>
<td>0.5</td>
<td>This is the first study to use sequencing to characterize soil viral communities. Within each of the four microbial groups, data showed minimal taxonomic overlap between sites, suggesting that soil archaea, bacteria, fungi, and viruses are globally as well as locally diverse</td>
<td>Fierer et al. (2007)</td>
</tr>
<tr>
<td>Acid mine drainage biofilm</td>
<td>103 462</td>
<td>pUC18</td>
<td>3.2</td>
<td>Authors report the reconstruction of near-complete genomes of Leptospirillum group II and Ferroplasma type II, and partial recovery of three other genomes</td>
<td>Tyson et al. (2004)</td>
</tr>
<tr>
<td>Chesapeake Bay viroplankton</td>
<td>564</td>
<td>pSMART-HCK</td>
<td>1.3</td>
<td>This report describes the first detailed examination of an estuarine double-stranded DNA viral metagenome. This analysis suggests that dsDNA viruses are likely one of the largest reservoirs of unknown genetic diversity in the biosphere</td>
<td>Bench et al. (2007a, b)</td>
</tr>
<tr>
<td>Global Ocean</td>
<td>7 697 926</td>
<td>Bst XI linearized pBR322 derivative</td>
<td>2</td>
<td>Authors report a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analysed as part of the Sorcerer II Global Ocean Sampling expedition. The resulting 7.7 million sequencing reads from 41 samples provide an unprecedented look at the great diversity and heterogeneity in naturally occurring microbial populations</td>
<td>Rusch et al. (2007)</td>
</tr>
<tr>
<td>Soil</td>
<td>1 186 200</td>
<td>pJN105,pCF430 &amp; pBeloBAC11</td>
<td>2.7–45</td>
<td>Authors designed a metagenomic analysis to isolate antibiotic resistance genes from six libraries of soil. They identified nine clones expressing resistance to aminoglycoside antibiotics and one expressing tetracycline resistance</td>
<td>Riesenfeld et al. (2004)</td>
</tr>
<tr>
<td>Worm lacking a mouth, gut &amp; nephridia</td>
<td>279 157 (3 kb library) 36 095 (35 kb library)</td>
<td>pMCL200 (3 kb library) pCC1FOS™ (35 kb library)</td>
<td>3–35</td>
<td>Metagenomic approach to describe four co-occurring symbionts from the marine oligochaete worm Olavius algarvensis. Data revealed that the symbionts are sulphur-oxidizing and sulphate-reducing bacteria, all of which are capable of carbon fixation, thus providing the host with multiple sources of nutrition</td>
<td>Woyke et al. (2006)</td>
</tr>
</tbody>
</table>

*Number of reads produced.
usually consist of complex mixtures of sequences (also containing only short fragments of genes, with no start and/or stop codons) from various genomes, care needs to be exercised when predicting genes based on such sequences, as not every potential ORF may code for a protein. To solve this problem, tools to describe and compare the richness, membership and structure of microbial communities using peptide fragment sequences extracted from metagenomic sequence data have been developed recently (Schloss & Handelsman, 2008). Another important issue of metagenome data handling is the capacity of the public databases for data storage. Currently, there are a few public databases that store annotated microbial genomes and allow various types of searches and sequence analyses (Galperin, 2006). However, the number of data repositories for environmental sequences is limited. Although GenBank serves as the main repository for all public sequences, the annotation quality for environmental data and the options for comparative analyses are limited.

**Application of holistic metagenomic indicators**

As mentioned before, one of the main obstacles to understanding complex microbial communities is that the number of interactions, processes and activities in the living cells is enormous (1 g of soil may contain up to $10^{12}$ interconnected putative reactions). Therefore, there is a need to obtain metagenomic fingerprints that could contribute to the holistic overview of fine-grained eco-genomic data into systems biology. In this section, we recapitulate and comment on different strategies for this.

The first requirement is to have a comprehensive and large volume of DNA sequences from a number of environments which differ in regard to the species richness and main environmental constraints (both geographic and orographic). Further, the community genome (bulk DNA) or genome of singular organisms (enriched DNA) is needed to perform, for example, the 454 pyrosequencing, Sanger shotgun sequencing or combinations thereof, sequence assembling and *in silico* annotation of these community genomes. Obviously, genome reconstruction is much easier if the community is of a low complexity and is dominated by populations of a few genomically very distinct taxa (Tyson *et al*., 2004). However, even in more complex environments, it should be possible to extend the random shotgun or pyrosequencing approach to recover the genomes of uncultivated strains and species. These data can then be used to explore the nature of the community metabolic network, to find conditions for cultivating previously uncultivated organisms, to monitor community structure over time, and to construct DNA microarrays or monitor global community gene expression patterns. As an example, whole genome shotgun sequencing and metabolic pathway reconstruction revealed that the symbionts from the marine oligochaete *Olavius algarvensis*, a worm lacking a mouth, gut and nephridia, are sulphur-oxidizing and sulphate-reducing bacteria, all of which are capable of carbon fixation, thus providing the host with multiple sources of nutrition (Woyke *et al*., 2006). Molecular evidence for the uptake and recycling of worm-waste products by the symbionts suggests how the worm could eliminate its excretory system, an adaptation unique among annelid worms. This study suggests how the versatile metabolism within this symbiotic
Table 2. Metagenomic populations characterized through the 454 pyrosequencing technology

<table>
<thead>
<tr>
<th>Sample</th>
<th>454-library size*</th>
<th>Average length of reads</th>
<th>Biodiversity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coral reef</td>
<td>188 445</td>
<td>c.100 bp</td>
<td>Four fragments of 16S rRNA gene were dominated by Proteobacteria</td>
<td>Krause et al. (2008)</td>
</tr>
<tr>
<td>Solar saltern</td>
<td>582 681</td>
<td>c.100 bp</td>
<td>151 genomic fragments were dominated by different halophilic archaea and by Salinibacter</td>
<td>Krause et al. (2008)</td>
</tr>
<tr>
<td>Stromatolite</td>
<td>124 694</td>
<td>c.100 bp</td>
<td>Nine fragments of 16S rRNA gene were dominated by Cyanobacteria</td>
<td>Krause et al. (2008)</td>
</tr>
<tr>
<td>Soudan Mine</td>
<td>334 386 (RS) 388 627 (BS)</td>
<td>106 bp (RS) 99.1 bp (BS)</td>
<td>76 16S rRNA gene fragments dominated by Alpha- and Gammaproteobacteria (Red Sample) and 24 16S rRNA gene fragments dominated by Actinobacteria (Black Sample)</td>
<td>Edwards et al. (2006)</td>
</tr>
<tr>
<td>Mouse and human distal gut</td>
<td>345 000</td>
<td>93.1 bp</td>
<td>Obesity is associated with changes in the relative abundance of the two dominant bacterial divisions, the Bacteroidetes and the Firmicutes</td>
<td>Turnbaugh et al. (2006)</td>
</tr>
<tr>
<td>Normal and CCD (Colonies collapse disorder) hives</td>
<td>–</td>
<td>150 bp</td>
<td>The clusters included an abundant member of the Gammaproteobacteria and several less frequent but widespread organisms from the Betaproteobacteria, Alphaproteobacteria, Firmicutes, and Actinobacteria groups</td>
<td>Cox-Foster et al. (2007)</td>
</tr>
<tr>
<td>Coral Porites astroides</td>
<td>316 279</td>
<td>102 bp</td>
<td>The most prominent bacterial groups were Proteobacteria, Firmicutes, Cyanobacteria and Actinobacteria</td>
<td>Wegley et al. (2007)</td>
</tr>
<tr>
<td>North Atlantic Deep Water and Axial Seamount Ocean surface waters</td>
<td>414 323 (DNA) 128 324 (cDNA)</td>
<td>110 bp (DNA) 114 bp (cDNA)</td>
<td>The genus Prochlorococcus and Alphaproteobacteria (genus Pelagibacter) were the two most highly represented taxonomic groups in both DNA and cDNA libraries</td>
<td>Frias-Lopez et al. (2008)</td>
</tr>
<tr>
<td>Soil</td>
<td>314 041</td>
<td>96.4 bp</td>
<td>Results indicate that crenarchaeota may be the most abundant ammonia-oxidizing organisms in soil ecosystems on Earth</td>
<td>Leininger et al. (2006)</td>
</tr>
<tr>
<td>Human mouth</td>
<td>100 000–390 000</td>
<td>c.100 bp</td>
<td>The 28 sequences from this study are associated with five different bacterial phyla, with most sequences located in the phylum Fusobacteria and specifically related to the genus Leptotrichia</td>
<td>Marcy et al. (2007)</td>
</tr>
<tr>
<td>Marine virome of four oceanic regions</td>
<td>1 768 297</td>
<td>102 bp</td>
<td>Metagenomic analyses of 184 viral assemblages collected over a decade and representing 68 sites in four major oceanic regions. This work provides evidence that the composition of viral assemblages varies in different geographic regions</td>
<td>Angly et al. (2006)</td>
</tr>
<tr>
<td>Northwest Atlantic &amp; Eastern Tropical Pacific Seawater Surface and hypersaline marine, freshwater samples</td>
<td>–</td>
<td>c.100 bp</td>
<td>Analysis of 7.7 million sequencing reads (6.3 billion bp) from the microbes collected across a several-1000-km marine transects</td>
<td>Rusch et al. (2007)</td>
</tr>
<tr>
<td>Termite hindgut</td>
<td>–</td>
<td>250 bp</td>
<td>Metagenomic analysis of 37 samples. Results showed that most of the 154 662 viral peptide sequences identified were not similar to those in the current database and that only few 1000 genes encoding metabolic and cellular functions could be unambiguously identified</td>
<td>Williamson et al. (2008)</td>
</tr>
</tbody>
</table>

*Number of reads produced.
consortium provides the host with an optimal energy supply. In another prominent example, new insights into other important symbiotic functions, including H₂ metabolism, CO₂-reductive acetogenesis and N₂ fixation, are also provided by this first system-wide gene analysis of a wood-feeding higher termite microbial community specialized towards plant lignocellulose degradation (Warnecke et al., 2007). A particularly vexing aspect of microbial metagenomics is the common observation of high genome variability among strains of a species within an environmental sample (Thompson et al., 2005). Such observations have raised significant questions about the validity of the microbial species concept, and the value of single genome sequences for comparisons between taxa (Konstandinidis & Tiedje, 2005). To reconcile this dilemma, it has been suggested that bacterial species have a ‘core-genome’, consisting of genes that are always present, and a ‘pan-genome’ of genes that are variably present (Tettelin et al., 2005).

Based on the few metagenome surveys of viral communities completed to date, i.e. from Chesapeake Bay (Bench et al., 2007a,b), near-shore waters, sediments, and four oceanic regions over 4600 km in distance and 3000 m in depth (Breitbart et al., 2002, 2004; Angly et al., 2005, 2006), and stromatolites (Desnues et al., 2008), a consensus is emerging that viral communities are also extraordinarily diverse (5 × 10⁹ cells mL⁻¹ surface seawater, Sano et al., 2004), and contain a high proportion of novel sequences with unknown functions (Fig. 3), some of which have significant impacts on microbial diversity and on global biogeochemical cycling. Moreover, the assemblages contained up to 130 000 genotypes that were found to overlap in composition between disparate geographic regions, and local environmental conditions enriched for certain viral types through selective pressure. Against the typical backdrop of over 60% unknown sequences, viral metagenome libraries tend to contain a collection of gene homologs that are relatively distant from better known representatives in bacterial genomes (Breitbart et al., 2002, 2003). In two extensive studies, metagenomic analyses of 184 viral assemblages collected over a decade and representing 68 sites in four major oceanic regions (Angly et al., 2006) and 37 new surface marine, freshwater and hypersaline samples collected during the Sorcerer II Global Ocean Sampling Expeditions (Williamson et al., 2008) showed that most of the viral sequences were not similar to those in the current database and that only a few thousands of genes encoding metabolic and cellular functions could be unambiguously identified. Distributional and phylogenetic analyses of these host-derived viral sequences also suggested that viral acquisition of environmentally relevant genes of host origin is a more abundant and widespread phenomenon than previously appreciated. Recruitment of GOS viral sequence fragments...
against 27 complete aquatic viral genomes revealed that only a few reference bacteriophage genomes were highly abundant and closely related, but not identical, to cyanophages, whereas prophage-like sequences were most common in the Arctic. Through these studies it is clear that the abundance and broad geographical distribution of viral sequences within microbial fractions, the prevalence of genes among viral sequences that encode microbial physiological function and their distinct phylogenetic distribution lend strong support to the notion that viral-mediated gene acquisition is a common and ongoing mechanism for generating microbial diversity in the marine environment.

The large-scale sequence data has a somewhat lower resolution, but can provide access to much greater genomic information of untapped microbial biodiversity. However, to some extent the data are unhelpful and inaccurate as far as linking specific genomes to specific roles, as we know that > 60% of genes are widespread and have similar functions in different organisms. For this reason, new developments involve the direct separation (Venter et al., 2004) or enrichment of cells per gene of interest (see example in Hoster et al., 2005) or preferably enrichment using the 13C-labelled compound directly related to primary ecological functions. The enrichment of environmental samples, for example those with scarce biomass, by amendment with specific substrates is a proven method to obtain a general reduction of microbial, and thus genetic, diversity due to the selective increase of population densities of fast-growing organisms utilizing the amended substrates (the chance to obtain a specific organism per gene is very high). As a recent example of the first approach is the sequencing of up to 62 Mbp metagenomic signatures of the Peru Margin sub-seafloor (Biddle et al., 2008). Here the authors found Crenarchaeota as a major microbial group, and therefore nitrification, sulphate reduction and methane generation the major processes in this particular environment. Although that study provided information on microbial diversity which is less biased than in the datasets implementing a high-copy number plasmid cloning step, there is a problem in interpreting the data in terms of the functions of deduced protein

<table>
<thead>
<tr>
<th>No. of specific activities</th>
<th>0</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-Polyproenyl-4-hydroxybenzoate carboxylase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formate dehydrogenase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenine-specific methyltransferase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTP cyclohydrolase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribonucleotide reductase I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribonucleotide reductase II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylcrotonyl-CoA carboxylase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dihydroxyacetone synthetase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytochrome C oxidase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoleucyl-tRNA synthetase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylcrotonyl-CoA carboxylase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA polymerase III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Composition of the viral protein diversity in different environmental samples. (a) Protein composition based on protein homology. As shown, viral communities contain a high diversity of unknown sequences (modified from Angly et al., 2006). (b) Total number of major ‘known’ proteins found in viral communities (modified from Angly et al., 2006). As shown, the majority of protein-coding genes are related to the acquisition of nucleotides and enzymes for transfection. (c) Protein distribution of ‘known’ functions by functional categories (modified from Williamson et al., 2008).
metabolic reconstruction of the system. Sorting the proteins into large groups, and taking a census according to their COGs of Pfam affiliations, does not provide the necessary resolution for deciphering the metabolic potential of the microbial system in question, and therefore the application of genomic signatures from this kind of project to systems biology will likely remain problematic. A particularly elegant strategy avoids this problem through a combination of the extraction of complete uncultured genomes in complex communities (with up to 5000 species) using high-resolution SIP to reconstruct their metabolisms and to link specific microorganisms, whose DNA is separated by ultracentrifugation, to specific ecological functions (Kalyuzhnaya et al., 2008). Here, the authors have provided a number of shotgun sequences of microbial communities harbouring a few dominant organisms under dynamic utilization of different nutrients. Accordingly, environment-specific organisms and processes catalyzed by the corresponding microorganisms have been pinpointed. An important lesson from this work is that environmental samples can dynamically vary from phylogetic and process equilibrium: the population of a closed niche may be modified until the systems reach a new equilibrium where the key microorganisms have the working capacity to make the overall community work, which one can consider a big step towards in situ systems biology. Even though some technical issues could be improved in the future (the enrichment itself, i.e. addition of extra substrate quantities in amounts high enough to extract the ‘heavy’ DNA, cloning bias and relatively low sequencing coverage), all in all, this work can be considered a significant contribution towards the understanding of the system as a whole.

The second requirement is to classify the different contributions of the different classes of microorganisms for global element cycling in specific niches and to determine the ratio of the number of microorganisms and processes in each category to detect predominant members and thereby predominant biochemical transformation; that is, to identify the connected poles of microorganisms and activities that shape the internal structure of an ecological niche and that progress independently of the others. Here, the problem is also exacerbated because of a relatively low share of metagenomic DNA-encoded pathways for unusual and valuable ecological functions, in the midst of a vast abundance of DNA for housekeeping biological functions. Novel techniques that make possible a numerical description of the specific biological functions unique to specific niches and acting against particular elements are required. Experimental platforms are therefore strongly needed for testing, dynamically analysing, surveying and visualizing the type of metabolic activity in metagenomic DNA of any origin, not only that based on large-scale sequence analysis (for recent review see Raes & Bork, 2008).

**Meta-transcriptomics: global analysis of gene expression in the microbial community**

Given current research trends, it seems likely that metagenomic datasets will continue to grow rapidly and soon will dwarf whole-genome sequence datasets derived from cultivated microorganisms. The nature, size, and complexity of this information present formidable challenges to analysis and interpretation. In addition, although these data provide information about genome content, there is no clear indication of gene expression or expression dynamics. Although techniques such as quantitative PCR can be used to quantify gene expression in natural samples, these are limited usually to measurement of a small number of known genes. Many questions remain therefore to be answered. What fraction of the new genes discovered in metagenomic datasets are actually expressed? What hypothetical proteins are expressed and what is their function? What are the dynamics and timescale for gene expression in individual microbial species, and communities? With the completion of the sequencing of over 134 metagenomes, the monitoring of global changes in gene expression, so-called transcriptomics, is an increasingly attractive method for dissecting the molecular basis of metabolic and ecological aspects (Liu & Zhu, 2005). Special interest is the use of metagenomic information for microarray-based technologies to reveal complex microbial functions. Techniques for metagenome-wide or high-throughput analysis of gene expression (transcriptomics techniques) can be divided into two broad categories: (1) techniques such as differential display (DD), cDNA-amplified fragment-length polymorphism (cDNA-AFLP), suppression subtractive hybridization (SSH), and serial analysis of gene expression (SAGE) that require no a priori knowledge of gene sequence (‘open architecture systems’) and (2) techniques such as catalysed reporter deposition (CARD)-FISH and microarrays that require pre-existing knowledge of the metagenome sequence (‘closed architecture systems’). These techniques are also used to identify the genes whose expression is differentially regulated in response to various natural or artificial stresses. In the following, we describe briefly the metagenomic interest of all these techniques, with the exception of SAGE, which has not as yet been applied to the microbial communities.

**Differential display**

DD-PCR is a PCR-based technique that allows side-by-side comparison of multiple RNA samples and can facilitate the identification of both suppressed and induced genes. There are two main steps in DD-PCR: (1) reverse transcription (RT) of RNAs isolated from different populations with a set of degenerate, anchored oligo(dT) primers to generate...
SSH method is based on a suppression PCR effect and combines normalization and subtraction in a single procedure. The normalization step equalizes the abundance of DNA fragments within the target population, and the subtraction step excludes sequences that are common to the populations being compared. This dramatically increases the probability of obtaining low-abundance differentially expressed cDNA or metagenomic DNA fragments, and simplifies analysis of the subtracted library (Rebrikov et al., 2004). In an elegant study, Galbraith et al. (2004) used this tool combined with a metagenomic approach to find an unexpectedly large difference in the archaeal community structure between the rumen microbial populations of two steers fed identical diets and housed together, which would be difficult to detect using other common techniques.

SSH may be applied to produce subtracted cDNA libraries to identify genes that are differentially expressed among environmental samples. This approach will lead to the isolation of niche-specific novel and active metabolic pathways (Rebrikov et al., 2004). To construct subtractive libraries, mRNA is isolated from the different samples to be compared, cDNA is then generated, and these cDNA populations are subtracted. Preliminary experiments revealed that 1–2 Gbp of the metagenomic information of polluted vs. pristine sites are converted into 30–200 SSH clones of c. 20 000 bp each, i.e. 0.001% subtractive clones. The resulting subtracted DNA fragments may be cloned to set up small SSH libraries, providing a plethora of gene targets active against pollutants in a fashion entirely independent of the overlapping functions in pristine and pollutant sites. Here, cDNA prepared, for example, from contaminated samples is used as the ‘tester’ and that from the pristine samples as the ‘driver’ for the forward subtraction to isolate fragments corresponding to genes whose expression level was increased in these conditions.

**CARD-FISH**

CARD-FISH is also a valuable tool for measuring gene activity in vivo (see example in Gittet et al., 2008). However, it is only relevant for quantitation of the transcripts of known genes (to make the probe for CARD-FISH, the exact gene sequence must be known). This limits the use of this technique in activity-based analysis of metagenomes as, in many cases, one should work with unknown genes and normally one tries to reveal new functions rather than doing a census of already known ones.

**DNA microarrays**

DNA microarray technology has vast potential for our understanding of microbial systems. Microarray-based genomic technology is a powerful tool for viewing the expression of thousands of genes simultaneously in a single
experiment (Hoheisel, 2006). Although this technology was initially designed for transcriptional profiling of a single species, its applications have been dramatically extended to environmental applications in recent years (Zhou & Thompson, 2002, 2004; Adamczyk et al., 2003; El Fantroussi et al., 2003; Taroncher-Oldenburg et al., 2003; Zhou, 2003; Loy et al., 2004; Tiquia et al., 2004; Bodrossy et al., 2006; An & Parsek, 2007). One of the greatest challenges in using microarrays for analysing environmental samples is the low detection sensitivity of microarray-based hybridization, in combination with the low biomass often present in samples from environmental settings. Microarrays for expression profiling can be divided into two broad categories – microarrays based on the deposition of preassembled DNA probes (cDNA microarrays) and those based on in situ synthesis of oligonucleotide probes (e.g. Affymetrix arrays, oligonucleotide microarrays). Applications employing DNA microarrays include, for example, characterization of microbial communities from environmental samples such as soil and water (Zhou, 2003; Eyers et al., 2004), pathogen detection in clinical specimens and field isolates (Bodrossy & Sessitsch, 2004), and monitoring of bacterial contamination of food and water (Lemarchand et al., 2004). Various types of DNA microarray have been applied to study the microbial diversity of various environments. Those include, for example, oligonucleotide comprising 20–70 bp (Ward et al., 2007), cDNA (PCR-amplified DNA fragments) (Wu et al., 2004), and whole genome DNA (Bae et al., 2005). To date, metadNA microarray studies have examined global gene expression in over 20 different environments covering a wide variety of research areas, some of which will be discussed below.

The use of microarrays to profile metagenomic libraries may also offer an effective approach for rapid characterization of many clones. As an example, a fosmid library was obtained and further arrayed on a glass slide (Sebat et al., 2003). This format is referred to as a metagenome microarray (MGA). In the MGA format, the ‘probe’ and ‘target’ concept is the reverse of those of general cDNA and oligonucleotide microarrays. Targets (fosmid clones) are spotted on a slide and a specific gene probe is labelled and used for hybridization. This microarray format may offer an effective metagenome-screening approach for identifying clones from metagenome libraries rapidly without the need of laborious procedures for screening various target genes. Sebat et al. (2003) and Park et al. (2008) used a microarray platform to screen a metagenomic library with whole microbial genomes and community genomes. To evaluate the functional diversity of communities of soil eukaryotic microorganisms, Bailly et al. (2007) evaluated an experimental approach based on the construction and screening of a cDNA library from a metatranscriptome using polyadenylated mRNA extracted from a forest soil. The diversity of organisms was evaluated by sequencing a portion of the 18S rRNA genes and cDNA. The metatranscriptome analysis revealed that the taxonomic distribution did not coincide; however, 180 species were found not to be present in the soil and 70% of the sequences were related to fungi and unicellular eukaryotes (protists). DNA-based microarray detection approaches coupled with whole-community genome amplification have been used to analyse microbial community structure in low-biomass groundwater microbial communities (Wu et al., 2006). However, this approach could not be directly adapted and used for mRNA-based activity analyses. A practical problem in detecting mRNAs from environmental samples by microarray hybridization is obtaining a sufficient amount of mRNAs for analysis. Some type of signal amplification before hybridization is needed. However, random PCR-based amplification is not an appropriate choice due to amplification bias and thus the loss of quantitative information (Nygaard & Hovig, 2006). Additionally, the gene-by-gene nature of conventional PCR (while rather useful for rRNA) severely restricts the throughput advantages of microarray analyses for functional genes. To solve this problem, a new method, termed whole-community RNA amplification, was developed for randomly amplifying whole-community RNAs (Gao et al., 2006) to provide sufficient amounts of mRNAs from environmental samples for microarray analysis.

One of the major problems associated with microarrays is derived from the short half-life of mRNA (Selinger et al., 2003; Andersson et al., 2006), and that mRNA in bacteria and Archaea usually constitute only a small fraction of total RNA. Several approaches for overcoming these challenges have been developed recently. In one approach, rRNA subtraction was used in combination with randomly primed RT-PCR to generate microbial community cDNA for cloning and downstream sequence analysis (Poresky et al., 2005). Frias-Lopez et al. (2008) developed a method for the polyadenylation of bacterial mRNA using Escherichia coli poly(A) polymerase, which facilitates preferential isolation of bacterial mRNA from rRNA in crude extracts, to synthesize and sequence microbial community cDNA from the North Pacific Subtropical Gyre by pyrosequencing. This avoids the need to prepare clone libraries and their associated biases. Genes associated with key metabolic pathways in open ocean microbial species, including those involved in photosynthesis, carbon fixation and nitrogen acquisition, and a number of genes encoding hypothetical proteins were highly represented in the cDNA pool. Moreover, genes present in the variable regions of Prochlorococcus genomes were among the most highly expressed, suggesting that these encode proteins central to cellular processes in specific genotypes. Although many transcripts detected were highly similar to genes previously detected in ocean metagenomic surveys, a significant fraction (50%) were unique. Universal
DNA microarrays have also been used to comprehensively determine the binding capacities over a full range of affinities for five transcriptional factors (TF) of different structural classes from yeast, worm, mouse and humans (Berger et al., 2006). The authors designed a universal microarray containing all possible k-mers (c. one million probes) covering all 10-bp binding sites, by converting high-density single-stranded oligonucleotide arrays in double-stranded DNA arrays, identifying more than 30 TF binding sequences (and thereby regulatory elements) from whole DNA samples from yeast, worm, mouse and humans.

With the completed genome sequences in hand it is possible to analyse the expression of all genes in each sample under various environmental conditions using whole-genome DNA microarrays (Shut et al., 2003; Gao et al., 2006). Such genome-wide expression analysis provides important data for identifying regulatory circuits in uncultured organisms (Lovley, 2003). In an elegant study based on most of the known genes and pathways involved in biodegradation and metal resistance, a comprehensive 50-mer-based oligonucleotide microarray was developed for effective monitoring of biodegrading populations (Rhee et al., 2004). This type of DNA microarray was used effectively to analyse naphthalene-amended enrichment and soil microcosms and demonstrated that microbial communities changed differentially depending on the incubation conditions (Cho & Tiedje, 2002). Also, a global gene expression analysis revealed the co-regulation of several so far-unknown genes during the degradation of alkylbenzenes (Kuhner et al., 2005). Besides this, DNA microarrays have been used to determine bacterial species, in quantitative applications of stress gene analysis of microbial genomes and in genome-wide transcriptional profiles (Mufler et al., 2002; Greene & Vooroudou, 2003).

The study of the gene expression from an environmental sample using DNA microarrays is a challenging task. Firstly, the sensitivity may often be a part of the problem in PCR-based cDNA microarrays, as only genes from populations contributing to > 5% of the community DNA can be detected. Second, samples often contain a variety of environmental contaminants that affect the quality of RNA and DNA hybridization (Zhou & Thompson, 2002) and make it difficult to extract undegraded mRNA (Burgmann et al., 2003). The specificity of the extraction method plays a central role and should vary depending on the site of sampling, as there must be sufficient discrimination between probes. However, there is a promising perspective for microarrays in determining the relative abundance of a microorganism bearing a specific functional gene in a complex environment. Over a range of 1–100 ng of target genomic DNA concentration, Wu et al. (2004) observed a linear relationship between signal intensity and target DNA from pure and mixed culture communities. However, specificity is a key issue, as one needs to distinguish the differences in hybridization signals due to population abundance from those due to sequence divergence. Furthermore, annotation and the comprehensive functional characterization of proteins or RNA molecules remain difficult, error-prone processes, as systems microbiology relies heavily on a thorough understanding of the functions of gene products (Morrison et al., 2006). Better annotation quality and curated functional information would enable improved gene and gene function predictions in newly sequenced organisms and environmental samples, and allow evaluation of context-dependent expression and function. In this context, user-friendly data mining and visualization tools such as MADNET (Segota et al., 2008) for rapid analysis of diverse high-throughput biological data such as MGA experiments have been developed. These tools contain, for example, information on metabolic and signalling pathways and transcription factors.

**Microbial community proteomics**

Metagenome techniques have been employed to evaluate the modifications of the microbial community structure and diversity in contrasted situations, and to identify populations associated with these modifications (Ranjard et al., 2000). However, DNA-based approaches only provide indications of the genetic microbial potential, they do not allow evaluation of the expression of this potential in ecosystems (DeLong, 2004; Rodriguez-Valera, 2004). Following this observation, and in spite of the increasing awareness of the need of a functional analysis of indigenous microbial communities, there is actually a lack of methods for characterizing the in situ expression of the microbial metagenome. In this context, the terms ‘proteomics’ and ‘proteome’, which were introduced in 1995 (Wasinger et al., 1995), emerge now as a key complement for functional genomics (Hegde et al., 2003; Maron et al., 2007). In this context, the protein molecules, but not the mRNAs, are the key players in real-time metabolic processes: the mRNA is a highly unstable transmitter on the path from the genes to the ribosome, but each protein molecule represents the end product of gene expression (Kuhner et al., 2005). Community proteomics will therefore gain momentum with the advent of further environmental sequencing projects and will become a useful tool for systems microbiology. Successful application of protein-based techniques relies on effective recovery of proteins from environmental samples. Ideally, a procedure for recovering microbial proteins from the environmental matrix should allow highly efficient protein recovery to obtain a protein pool that is (1) of sufficient purity for analysis using the biochemical methods available, and (2) representative of the total proteins within the natural microbial community.
Recent examples include the characterization of the community proteome of indigenous bacterial communities in freshwater samples (Benndorf et al., 2007; Pierre-Alain et al., 2007). Authors analysed the variations of protein fingerprints according to perturbations (cadmium or mercury contamination) and compared these with the DNA fingerprint. Both analyses were statistically similar, showing that both the functional and genetic structures of the freshwater bacterial community were complex and varied with perturbation. A few other studies demonstrated the possibility to extract and resolve the diversity of proteins directly from diverse natural environments (Ogunseitan, 1993, 1996, 1997). These works evidenced that the microbial community proteome is rather complex considering environmental conditions, and varies significantly with change in these conditions (Ogunseitan, 1993, 1996). Considerable lack of gene sequence data at that time made it difficult phylogenetically to attribute the detected proteins. Newly seeded techniques such as transcriptomics, proteomics and interactomics also offer remarkable promise as tools to address long-standing questions, for example regarding the molecular mechanisms involved in the control of microbiobly driven xenobiotic degradation pathways. During the process of biodegradation, gene transcripts have been studied using high-throughput transcriptomic techniques with microarrays. Generally, however, transcripts have no ability to produce any physiological response; rather, they must be translated into proteins with significant functional impact. Exploring the differential expression of a wide variety of proteins and screening of the entire genome for proteins that interact with particular mineralization regulatory factors would help us to gain insights into bioremediation (Singh, 2006). In an elegant example, proteomic analysis revealed significant shifts in the microbial community physiology within 15 min of cadmium exposure, a rapid change not detectable using the phylogenetic profiling tools common to molecular microbial ecology (Lacerda et al., 2007). This study demonstrated that two-dimensional electrophoresis can be used to separate and resolve hundreds of proteins from a microbial community of more than 50 species, and that it is possible to identify proteins from a mixture of bacteria with un-sequenced genomes. Metaproteomics has also been used to decipher the metabolic details of a treatment process for activated sludge wastewater that operates for phosphorus removal (Wilmes & Bond, 2006; Wilmes et al., 2008a, b). A comparative study of protein expression between the anaerobic and aerobic phases in two distinct activated sludges demonstrated that protein expression within the sludge performing well did not alternate as much during the anaerobic–aerobic sludge cycling compared with the sludge performing poorly. An interesting possibility is that the more equilibrated protein expression in the sludge performing well provides a bioenergetic advantage to phosphate-accumulating organisms in these alternating systems. Further, Krüger’s group have also used metaproteomic analysis to link microbial processes or environmental relevance, i.e. anaerobic oxidation of methane, with the presence of certain proteins and compounds (Krüger et al., 2003), thus indicating that metaproteomics can also be used as holistic indicator.

The most comprehensive community proteome investigation thus far, successfully combined mass spectrometry proteomics with community genomic analyses (Ram et al., 2005) of a low species richness (mainly Leptospirillum species) and abundant biomass natural biofilm growing inside the Richmond Mine at Iron Mountain. Using the extensive genomic dataset from microbial biofilm, a database of 12 148 predicted protein sequences was constructed. More than 2000 individual proteins (52% with no significant similarity to any known proteins and 14% annotated as ‘hypotheticals’ in the original genomic dataset) were positively identified and linked to survival challenges in this extreme environment (e.g. chaperones, thioredoxins and peroxiredoxins). Genomic datasets of the same environment have been used to identify, with strain specificity, expressed proteins from the dominant members of a genomically uncharacterized, natural, acidophilic biofilm (Tyson et al., 2004; Lo et al., 2007). Proteomics results reveal large blocks of gene variants crucial for the adaptation to specific ecological niches within the very acidic, metal-rich environment. Mass spectrometry-based discrimination of expressed protein products that differ by as little as a single amino acid enables us to distinguish the behaviour of closely related coexisting organisms. Therefore, because proteomic data simultaneously convey information about genome type and activity, strain-resolved community proteomics is an important complement to cultivation-independent genomic analysis of microorganisms in the natural environment. Obviously, 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. 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 (Tringe et al., 2005; Schweder et al., 2008).

Community proteomics approaches have also been used to observe expressed protein profiles of natural Chesapeake Bay microbial communities (Kan et al., 2005). The authors found that the proteins were of marine microbial origin and were correlated with abundant Chesapeake Bay microbial lineages, Bacteroides and Alphaproteobacteria. A combination of genomic information and metaproteomes has also been used to explain the horizontal genetic exchanges in the microbial evolutionary history (Hatfull et al., 2006). Here, the genomic analysis of 30 complete mycobacteriophages
(viruses that infect mycobacterial hosts) revealed them to be genetically diverse and to contain many previously unidentified genes. The clustering of protein-encoding genes into ‘phamilies’ of related sequences provides a clue as to which genes are most prevalent in these phages, and how phamily size corresponds to phage and bacterial homologs outside of the mycobacteriophage group. This analysis revealed 101 tRNAs, 3 tmRNAs, and 3357 proteins belonging to 1536 phamilies of related sequences, and moreover that of the 1536 phamilies, only 230 (15%) have amino acid sequence similarity to previously reported proteins, reflecting the enormous genetic diversity of the entire phage population. Further, in an elegant example, Klaassens et al. (2007) obtained peptide mass fingerprints from sequence data to produce 6000 peptide fragments similar to DNA sequences of an accompanying metagenomic library from a low-complexity natural microbial biofilm in the gastrointestinal track. Markert et al. (2007) also employed the proteome analysis of the symbionts of Riftia pachyptila, the tube-worms living in hydrothermal vents. The studies suggested that the reductive tricarboxylic acid cycle, enzymes of which were found in abundance in cell extracts, is an important pathway in carbon fixation, along with the Calvin cycle, within the worm–symbiont system, which explains the high carbon isotope values which have been puzzling researchers within the worm–symbiont system, which explains the high carbon isotope variations on known sequence themes, but rather the existence of entirely new sequence classes and novel functionalities (Ferrer et al., 2007). Whole-sequencing of bulk environmental DNA has recovered near-complete microbial genomes and partial genomes of other minor community members, which opens up new avenues for systems microbiology studies, but also underscores the genetic diversity potentially available in environmental communities (Arber, 2000, 2004). This is of special interest because the inability to measure biochemical parameters in single cells poses a serious restriction. Whole sequencing of tens, hundreds, or even millions of cultured and uncultured cells via metagenomic analysis may be useful to better understand the range of variation and the true dynamics of organisms. However, quantitative measurements of gene expression, protein level, metabolites and other cellular constituents are needed to complete the picture of biological systems. Furthermore, measurements of activity in single cells will be necessary and for this reason it is highly desirable to be able to carry out these measurements non-destructively and in real time. Tracking cell measurements over time would be particularly helpful in following components that are located at specific points in the cells. For that, high-throughput methods for cell sorting such as in vitro compartmentalization (Griffiths & Tawfik, 2006) and fluorescence activated flow cytometry approaches followed by DNA amplification (for review see Lasken, 2007) promise to greatly accelerate the pace of microbiology research using uncultured microorganisms. Single-cell sequencing of those selected environmentally major microorganisms will make an important contribution to revealing the extent and nature of the diversity, the core or consensus genome that is essential to ‘species’, the role of horizontal gene transfer in evolution, and the relationship of diversity to environmental factors (Ward, 2006; Stepanauskas & Sieracki, 2007; Ishoe y et al., 2008). This is of special interest as we know that a range of microbial communities are represented by a small number of highly abundant organism types. It is furthermore clear that to access the microorganisms in their natural milieu there is a strong need to elaborate a ‘Community Systems Microbiology’.

Concluding remarks

Recent progress has revealed that the capture of genetic resources of complex microbial communities in metagenome libraries allows the discovery of a richness of new genomic and metabolic diversity that had not previously been imagined. Activity-based screening of such libraries has demonstrated that this new diversity is not simply variations on known sequence themes, but rather the existence of entirely new sequence classes and novel functionalities (Ferrer et al., 2007). Whole-sequencing of bulk environmental DNA has recovered near-complete microbial genomes and partial genomes of other minor community members, which opens up new avenues for systems microbiology studies, but also underscores the genetic diversity potentially available in environmental communities (Arber, 2000, 2004). This is of special interest because the inability to measure biochemical parameters in single cells poses a serious restriction. Whole sequencing of tens, hundreds, or even millions of cultured and uncultured cells via metagenomic analysis may be useful to better understand the range of variation and the true dynamics of organisms. However, quantitative measurements of gene expression, protein level, metabolites and other cellular constituents are needed to complete the picture of biological systems. Furthermore, measurements of activity in single cells will be necessary and for this reason it is highly desirable to be able to carry out these measurements non-destructively and in real time. Tracking cell measurements over time would be particularly helpful in following components that are located at specific points in the cells. For that, high-throughput methods for cell sorting such as in vitro compartmentalization (Griffiths & Tawfik, 2006) and fluorescence activated flow cytometry approaches followed by DNA amplification (for review see Lasken, 2007) promise to greatly accelerate the pace of microbiology research using uncultured microorganisms. Single-cell sequencing of those selected environmentally major microorganisms will make an important contribution to revealing the extent and nature of the diversity, the core or consensus genome that is essential to ‘species’, the role of horizontal gene transfer in evolution, and the relationship of diversity to environmental factors (Ward, 2006; Stepanauskas & Sieracki, 2007; Ishoe y et al., 2008). This is of special interest as we know that a range of microbial communities are represented by a small number of highly abundant organism types. It is furthermore clear that to access the microorganisms in their natural milieu there is a strong need to elaborate a ‘Community Systems Microbiology’.
concept that builds on the -omics approaches and aims at the understanding of functioning of microbial communities as a whole. Together this will help us to provide a description of ecosystem development which can be covered by three key parameters: (1) microbial community structure; (2) ecological networking; and (3) ecological genomic signatures that can be used as holistic indicators of a particular ecosystem.

**Acknowledgements**

This research was supported by the Spanish MEC BIO2006-11738, CSD2007-00005 and GEN2006-27750-C-4-E and CDTI I+DEA (CENIT-2007-1031) projects. A.B. thanks the Spanish MEC for a FPU fellowship. M.-E.G. thanks the Junta de Andalucía grant. P.N.G. was supported by Grant 0313751K from the Federal Ministry for Science and Education (BMBF) within the GenoMikPlus initiative.

**References**


