Review

Correspondence

Philippe N. Bertin

philippe.bertin@gem.u-strasbg.fr

Advances in environmental genomics: towards an integrated view of micro-organisms and ecosystems

Philippe N. Bertin,¹ Claudine Médigue² and Philippe Normand³

¹Génétique Moléculaire, Génomique et Microbiologie, Université Louis Pasteur, UMR7156 CNRS, Strasbourg, France

²Génomique Métabolique, Génoscope, UMR8030 CNRS, Evry, France

³Ecologie Microbienne, Université Claude Bernard – Lyon 1, UMR5557 CNRS, Villeurbanne, France

Microbial genome sequencing has, for the first time, made accessible all the components needed for both the elaboration and the functioning of a cell. Associated with other global methods such as protein and mRNA profiling, genomics has considerably extended our knowledge of physiological processes and their diversity not only in human, animal and plant pathogens but also in environmental isolates. At a higher level of complexity, the so-called meta approaches have recently shown great promise in investigating microbial communities, including uncultured micro-organisms. Combined with classical methods of physico-chemistry and microbiology, these endeavours should provide us with an integrated view of how micro-organisms adapt to particular ecological niches and participate in the dynamics of ecosystems.

Introduction

Genomics is a recent conceptual approach for the biological study of micro-organisms, which relies on the analysis of the complete genetic information they contain. This scientific discipline really emerged with the characterization of the first complete genome of the autonomous organisms *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and *Methanococcus jannaschii* (Bult *et al.*, 1996). These papers were almost immediately followed by others describing laboratory model eukaryotes, i.e. *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996), and bacteria, i.e. *Escherichia coli* (Blattner *et al.*, 1997) and *Bacillus subtilis* (Kunst *et al.*, 1997). Such projects, which initially required many years and a huge amount of work, were markedly facilitated by numerous technological developments that came of age at the end of the 1990s. A concomitant reduction of sequencing costs led to an explosion of genomic programmes (Fig. 1A); these now concern organisms in all domains of life, from viruses to higher eukaryotes including man. Indeed, more than 2000 genomes are currently being sequenced and more than 500 have been already published (<http://www.genomesonline.org>).

In the microbial field, most of the studied genomes belong to the Proteobacteria, which represent the majority of known bacteria in taxonomy, pathology or biotechnology, followed by Firmicutes, Actinobacteria and others (Fig. 1B). Nowadays, the genomics of micro-organisms not only concerns strains cultivated under laboratory conditions but

also microbial communities that may contain uncultured microbes. Several research centres throughout the world have played a major role in these sequencing programmes, for example the Joint Genome Institute (JGI) and The Institute for Genomic Research (TIGR) in the USA, whose projects account for more than 30 % of the total, and the Sanger Institute and Genoscope in Europe, which have undertaken about 10 % of the projects under way. We can therefore foresee that most of the taxa described in Bergey's Manual, the reference in bacterial taxonomy, will have had their genomes sequenced within the next 5 years. They number about 5000 (Holt *et al.*, 1994), which corresponds approximately to the number of bacterial species associated with high-quality 16S rDNA sequences (<http://www.msu.edu/~garrity/taxoweb/datasets.html>). This number is evidently an underestimate due to the numerous uncultured taxa and the constant exploration of new biotopes, including extreme environments, as well as the reanalysis of known taxa with modern tools. The growing sequencing capacities will greatly increase the exploration of the microbial world in the future, in particular in the field of ecology, extending our knowledge of the diversity of microbial metabolic processes.

Origin and concepts

The origin of genomics for the study of organisms can be situated at the end of World War II (Watson & Cook-Deegan, 1991). At that time, the US Department of

Fig. 1. Exponential increase in complete sequenced genomes published from 1995 to 2007 (A) and the phylogenetic distribution of the corresponding organisms (B), from the GOLD website statistics (<http://www.genomesonline.org>).

Energy, then called the Atomic Energy Commission, considered the possibility of identifying genetic alterations resulting from the high levels of radiation generated by the explosion of atomic bombs in Hiroshima and Nagasaki. This ambitious project was marked by several crucial steps, not only conceptual but also technical (Fig. 2). First, the structure of the DNA molecule was described in 1953 by Watson & Crick (1953), who were able to interpret the biophysical data of Rosalind Franklin. They proposed a double-helix structural model including semi-conservative mechanisms based on DNA polymerases that came to be studied in detail. From then on, it became theoretically possible to study mutations generated by mutagenic radiations and their impact on genome functioning due to the cracking of the genetic code in the 1960s (Nirenberg & Matthaei, 1961). Second, the method enabling access to the DNA sequence was published about 20 years later and

to this day remains the basis of modern sequencing procedures, except for the novel technologies (see below). In the so-called Sanger method (Sanger *et al.*, 1977), multiple DNA fragments differing by only one base are produced by incorporation of modified nucleotides, which stop the polymerization reaction. The analysis of the resulting radiolabelled fragments on a polyacrylamide gel allows the sequence of the various nucleotides that constitute the DNA molecule to be determined.

The Human Genome Project (HGP) took shape in the 1980s and an international consortium of public institutions was created. The technology available at that time seemed insufficient to decipher a large genome but several major technological innovations appeared within a 10-year period. The first one was the replacement of radioactivity by fluorescent dyes, which allowed the co-migration of the reactions involving the A, T, C and G bases on the same gel track and their automated laser detection. The second one was the development of capillary sequencers progressively allowing the analysis of up to 96 samples in a single step with the concomitant development of adequate computing capacities. For those reasons, sequencing performances were multiplied 10-fold between 1995 and 1997 and 10-fold again between 1997 and 1999. The early sequencing projects of reference micro-organisms such as *B. subtilis* and *S. cerevisiae* involved many laboratories for up to almost 10 years. Nowadays, similar projects do not extend beyond 1 or 2 years, a large amount of the data usually being acquired in only a few days. The increased sequencing capacity of specialized centres, combined with a major reduction in costs, has led to the publication of the genome sequences of many organisms in recent years. The development of new technologies such as pyrosequencing should result in a further increase in genome characterization in the future. This approach, which is based on light detection resulting from an enzymic reaction cascade generated by DNA polymerization, does however have some limitations, in particular small-size reads and uncertainties with poly(N) stretches (Ronaghi, 2001).

As for molecular biology several years earlier, genomics constitutes a deep conceptual revolution. For many years, biology focused on the description of distinct elements and their classification. The cell was for a long time considered to be a collection of objects. Genomics now allows the physiology of organisms to be addressed in a global way, considering them as a large set of components working together within a complex network of interactions. Such a perception is fundamental for an understanding of biological mechanisms, as no living organism may be restricted to a single gene or to a family of genes expressed at different times during the cell cycle. Nevertheless, the first sequencing projects have demonstrated that, even in reference organisms such as *E. coli* and *B. subtilis* (Blattner *et al.*, 1997; Kunst *et al.*, 1997), 25–40 % of the coding sequences (CDS) have no known function and are related to no gene of known function. Therefore it became clear early on that entire aspects of organisms' physiology

Fig. 2. Some important milestones in the history of genomics and DNA sequencing. 1953, DNA double helix structure discovery; 1961, first step in the cracking of the genetic code; 1975, chain-termination nucleic acid sequencing method developed by Sanger; 1985, international consortium set up to sequence the human genome (HGP); 1986, increased sequencing capacities by the use of fluorescent dyes; 1995, first microbial genomes published by J. C. Venter and colleagues; 1997, development and commercialization of capillary sequencers; 2005, development and commercialization of GS20 pyrosequencer.

remained to be discovered. The knowledge of the complete genome sequence of numerous organisms has led to profound changes in investigation methods, aiming to characterize genes whose function was not identified by conventional strategies or to study stress responses because of their multifactorial character. The use of global analysis methods, usually called large-scale or high-throughput technologies, has spread to many laboratories. Instead of studying individual genes, proteins or metabolic products, the integral profile of organisms can now be established. In particular, the differential analysis in various growth conditions of the whole protein content ('proteome'), transcripts ('transcriptome') or metabolites ('metabolome') allows the simultaneous quantification of gene expression or the accumulation of the corresponding products in an organism (Delneri *et al.*, 2001; Kahn, 1995; Velculescu *et al.*, 1997). Integrative biology, which aims at including all physiological processes in a bioinformatics predictive approach, will also benefit from these expanded genomic possibilities (Bonneau *et al.*, 2006).

Most of the sequencing efforts were initially focused on human pathogens, such as bacteria and parasites, and later on higher eukaryotes, to establish a basis for a better understanding of physiological processes in human beings. More recently, there has been a growing interest in micro-organisms isolated from various environments. The objective is to exploit available methodologies to identify specific properties that allow these organisms to grow in

diverse environments, including extreme ecological niches. These studies should also lead to a better understanding of ecosystems themselves. Moreover, they should lead to the identification of novel functions that may be exploitable for biotechnological applications, in particular the bioremediation of contaminated environments. The wide availability of these data to the scientific community and their comparison offer unprecedented opportunities for studying how the components of a living organism, also the organisms themselves, function together and respond to environmental challenges. A deep knowledge of the elements involved, their spatial and temporal distribution, and the metabolic pathways they belong to, would allow an integrated picture to be drawn of the biological processes under study and could lead to an optimal use of micro-organisms' properties, favouring the desired effects.

Strategies and results

DNA sequencing and comparative genomics

The size of bacterial genomes extends from 0.16 Mb in insect-symbiotic bacteria such as *Carsonella ruddii* (Nakabachi *et al.*, 2006) to more than 12 Mb in soil saprophytic myxobacteria such as *Sorangium cellulosum* (Pradella *et al.*, 2002). The genomic sequence of an organism results from the repeated determination of the nucleotide sequence from several thousand DNA fragments

(‘run’). The result of a run performed from a sequencing reaction carried out by the classical dideoxy chain-termination method usually corresponds to approximately 750 nt. To date, the cost of such a run can be estimated, all included, as approximately 1 €, which is notably lower than at the beginning of the genomic era, due to technical improvements described above and also to the many thoroughly robotized steps. The budget available will determine, at least in part, the number of sequencing reactions that will be carried out and read. It is usually assumed that the sequencing of five equivalent-genomes ($5 \times$ coverage) is sufficient to get a good idea of the genomic content. However, the coverage is usually higher, i.e. $12\text{--}15 \times$, in order to reduce the work-intensive gap-closing and finishing steps. In theory, this corresponds to approximately 100 000 runs for a genome of about 5 Mb.

Whole-genome random sequencing, the so-called shotgun strategy, was used by TIGR to decipher the chromosome of *H. influenzae* published in 1995 (Fleischmann *et al.*, 1995) and helped them overtake the other genome projects started earlier with more stepwise and time-consuming strategies. The shotgun approach presents several advantages such as simplicity and rapidity. However, it requires a major bioinformatics effort to analyse the data resulting from the multiple runs and assemble them with precision into a complete molecule. Consequently, such a procedure was initially not chosen because it was feared repeats (ribosomal genes, insertion sequences, etc.) would make assembly impossible. In this approach, genomic DNA is usually fragmented by nebulization, i.e. physical breakage, rather than molecular digestion by endonucleases, which may show some location bias. Several genomic libraries are usually constructed in various cloning vectors to make plasmid DNA preparation easier, to avoid the toxicity that may result from the overexpression of some genes, e.g. those encoding membrane proteins, and to facilitate the construction of a scaffold, an important step in the reconstruction of the molecule (Frangeul *et al.*, 1999). Small DNA fragments (around 3 kb), are cloned into high-copy-number vectors, while medium-sized (around 10 kb) and large (>100 kb) fragments are cloned into low-copy-number vectors and bacterial artificial chromosomes (BACs), respectively, helping to reduce the toxicity associated to some sequences.

Sequences obtained by such a procedure are read, aligned and progressively assembled into larger fragments (contigs) using appropriate algorithms such as the Phrep/Phrap/Consed suite, which results in a progressive reduction in the number of contigs (Green, 2001). The regions not yet sequenced at this stage constitute gaps in the chromosome(s) that will be filled in during the last step of the sequencing project. The so-called finishing step can be carried out by different methods, such as genome walking or sequencing of fragments obtained by PCR with oligonucleotides designed to target the extremities of adjacent contigs. The final objective is the sequence of a complete DNA molecule that respects the international

norms of the so-called Bermuda Accord, i.e. 99.99 % exactness, which corresponds to <1 error per 10 000 bases (<http://www.genome.gov/10000923>).

The topology of the genome is a parameter that needs to be determined experimentally for at least some representatives of each taxonomic group. It had been assumed from work on *E. coli* and *B. subtilis* that bacterial chromosomes were circular, and the mechanism of replication based on a leading strand and a lagging strand is consistent with this topology. Conversely, it is well known that eukaryotic chromosomes are linear, with telomeres and telomere-associated proteins to ensure full-length replication. However, it progressively emerged that some prokaryotic plasmids are linear, as in *Streptomyces* species (Hirochika *et al.*, 1984). More recently, it has been shown that the *Streptomyces lividans* chromosome itself is linear (Lin *et al.*, 1993), that it is possible to circularize it at the cost of instability (Wan *et al.*, 2004), and that the arms contain rearrangements and secondary metabolism genes (Dary *et al.*, 2000) as well as long inverted repeats. In addition, it came to be hypothesized that in some lineages such as Actinobacteria, linearity was the norm, which explains the need to determine experimentally the circularity of the recently sequenced *Frankia* genomes (Normand *et al.*, 2007). Finally, the genome of the actinobacterium *Saccharopolyspora erythraea* has also been shown to be circular, which further supports the current vision that most genomes are circular but that experimental determination is still an essential step in their reconstruction (Oliynyk *et al.*, 2007).

The complete DNA molecule will then be annotated not only to identify the various CDS but also to determine as precisely as possible their role, the localization of transcription (promoters, terminators) and translation signals (ribosome-binding sites), the existence of operons, the conservation of gene order between micro-organisms (synteny) or the presence of duplications. The procedure carried out automatically should be, in principle, followed by an expert annotation, i.e. a careful examination by researchers of the data automatically obtained and their validation according to experimental results and/or data in the literature. Unfortunately, such an approach is most often not performed, which has contributed to a pyramidal increase in the misinterpretation of gene functions in databases: an initial misinterpretation of the function of a protein or a motif can be transmitted to their closest neighbours identified by BLAST searching, resulting in a cascade of errors. Some powerful interfaces, such as MaGe (Vallenet *et al.*, 2006), exist that hopefully will help produce a high-quality annotation, and their use should become more widespread in the future. In addition, the identification of domains as subsets of proteins has been a very promising approach, implemented by databases such as InterPro (<http://www.ebi.ac.uk/interpro/>). Evolutionarily, domains are short fragments of DNA that are swapped between genes (Servant *et al.*, 2002) and probably constitute the random inventiveness of bacterial genomes to preadapt

to unforeseen conditions. Operationally, the identification of such domains may help address the function of a given protein. Finally, besides generating predicted protein functions, some systems such as PUMA2 (Maltsev *et al.*, 2006) and BioCyc (Krummenacker *et al.*, 2005) provide extensive metabolic pathway information. The set of predicted enzymes in a newly annotated genome allows the reconstruction of metabolic networks that potentially exist in this organism. These reconstructions are based on reference metabolic databases such as KEGG (Kanehisa *et al.*, 2006) and MetaCyc (Caspi *et al.*, 2006) and comparisons of pathways allow the identification of conserved and missing reactions. This information is helpful for inferring functional coupling of genes participating in the same cellular process. In addition, such metabolic models represent valuable scientific hypotheses regarding an organism's physiology and life style, and facilitate experimental planning (Opperdoes & Coombs, 2007; Sivachenko & Yuryev, 2007).

Whatever the organisms under study, i.e. pathogenic microbes and environmental isolates explored in the first genomic programmes or more recently, the results of genome sequencing are always of importance, as they allow organisms to be examined in a global way. In many cases, these data have greatly increased the previously accumulated knowledge of the organisms, e.g. secondary metabolite biosynthesis in *Pseudomonas fluorescens* (Paulsen *et al.*, 2005) or the acquisition of virulence factors in *Yersinia enterocolitica* (Thomson *et al.*, 2006). When genomic approaches are applied to organisms that have been the subject of a limited number of studies, the results obtained are often remarkable. For example, the sequencing of *Cyanidioschyzon merolae*, a unicellular red alga living in sulphur-rich hot springs, revealed the existence of a compact genome. Its constituent elements gave important information on genes essential to the development of photosynthetic activity, including in higher plants (Matsuzaki *et al.*, 2004). Similarly, the deep-sea bacterium *Idiomarina loihiensis* presents many adaptive traits as compared to other Gammaproteobacteria. The existence in its genome of numerous peptidase-encoding genes as well as amino acid and peptide uptake systems suggests that these nutrients are present in the environment as proteinaceous particles and are the main source of carbon and energy metabolism rather than sugar fermentation as initially thought (Hou *et al.*, 2004). The genome exploration of photosynthetic *Bradyrhizobium* strains also showed that, at least in some rhizobia, host-plant nodulation occurred via a mechanism not involving *nod* genes (Giraud *et al.*, 2007). Finally, the recent genome deciphering of metallophilic micro-organisms such as *Herminiimonas arsenicoxydans* has demonstrated the existence of novel strategies in colonizing arsenic-rich environments (Muller *et al.*, 2007). By detoxification of the milieu, these mechanisms may have played a key role in the early stages of life development on Earth, contributing to the emergence and establishment of other micro-organisms.

A major consequence of the increase in the number of known genomes is that a comparative analysis, including the comparison of several members of a given species (Binnewies *et al.*, 2006), is becoming the norm. As a first approach, comparative analyses such as codon usage or mol% G + C content are possible. In addition, the in-depth comparison of two or more closely related genomes should permit identification of genes correlating with their differential physiological properties and enable a function to be attributed to several uncharacterized genes. By comparing a large number of genomes, it has been possible to identify genes common to all lineages and thus to propose the existence of a minimal set of about 250 genes for a prokaryotic cell (Mushegian & Koonin, 1996). Such a group of genes is however not sufficient to characterize a living organism, in particular in the context of microbial ecosystems, where specific and extreme conditions are often the rule. It is also possible for a set of genes to identify those subject to positive selection (diversifying) or negative selection (homogenizing), genes for which selection brings about a different number of mutations in the protein chain relative to what is expected according to the neutral evolution model (Kimura, 1979). Such a situation has been demonstrated for human papilloma virus (Chen *et al.*, 2005), the late blight agent *Phytophthora infestans* (Liu *et al.*, 2005), and *Sinorhizobium meliloti* (Bailly *et al.*, 2006).

Synteny (conservation of gene order) also yields information on selection pressures that have played a role over the course of evolution since the time when two given genomes had a common ancestor. Genes with related functions are often colocalized on the genome so as to have comparable expression levels (as in the case of polycistronic messengers), to be synthesized in the same part of the cell, and be co-transferred during conjugation (Ettema *et al.*, 2005). There are numerous examples of related genes that are far apart, nevertheless colocalization of genes on a genome, and even more so on a set of genomes, is an indication that their functions are related. Synteny can be quantified, for example by the number of genes in pairs of genomes that have less than a given number of intervening non-homologues. It is known that the rate of synteny is in general a function of the genetic distance between two strains, except when strains have undergone marked lifestyle modification such as when a lineage becomes symbiotic and consequently suffers massive gene loss (minimizing selection) or conversely for taxa inhabiting soil or other comparably permissive biotopes where genomes are more free to accumulate genes (permissive selection). For example the exploration of the genomes of three strains of the facultatively symbiotic *Frankia* sp., which vary in size from 5.43 Mbp for a narrow-host-range strain (Cci3) to 9.05 Mbp for a broad-host-range strain (EAN1pec), showed that gene conservation is quite high among the three genomes, especially at the origin of replication (Fig. 3). This picture is also supported by the pattern of synteny, with conservation decreasing at the replication terminus, which corresponds to a high degree of gene rearrangement, duplication or deletion in this region. It

Fig. 3. Comparative genomics of three *Frankia* sp. genomes. (A) Conventional circular representation of the *Frankia alni* genome, strain ACN14a (Normand *et al.*, 2007). Dark blue, backbone (conserved) genes found in all three strains; Red, strain-specific genes present in ACN14a, but absent from EaN1pec and Cci3 genomes; Green, genes common between *Frankia* and *Streptomyces*. (B) Synteny plot between the three *Frankia* chromosomes. This line plot was obtained by using synteny results between *Frankia* sp. Cci3 and ACN14a, as well as results obtained from comparison of *Frankia* sp. ACN14a and EaN1pec.

has been hypothesized that much of the size differences can be accounted for by expansion in this area of the genomes of strains EAN1pec and ACN14a (a medium-host-range strain), in concert with the biogeographic history of the symbioses and host plant speciation (Normand *et al.*, 2007).

Finally, having several genomes at one's disposal also allows identification and quantification of laterally transferred genes, i.e. those that have a phylogeny different from that of the 16S gene (Daubin *et al.*, 2003) and that may take part in a microbial evolution process in relation to specific ecological niches (Collyn *et al.*, 2006). In this respect, an integrated analysis based on a theoretical and experimental comparison, i.e. array-based genomic hybridization and expression and proteomic profiling, of two *Burkholderia pseudomallei* strains saprophytic and pathogenic to human beings and animals has revealed the presence or absence in these organisms of numerous regions acquired by horizontal gene transfer. Such clusters of genes greatly contribute to the phenotypic diversity of this pathogen (Ou *et al.*, 2005). Remarkably, comparative genomics has given rise to a new concept highlighting the great diversity between closely related strains. A species can be described by its pangenome, i.e. the sum of a core genome containing genes present in all strains, and a dispensable genome, with genes absent from one or more strains and genes unique to each strain (Medini *et al.*, 2005).

mRNA expression profiling

As soon as the first genomes were published, it became clear that a high proportion of the genes identified by annotation strategies had no known function and could

not be related to any gene of known function. This observation justifies by itself the development of functional genomic programmes, designed to discover new functions and their underlying regulatory mechanisms. Knowing the complete genome of many micro-organisms allows their expression profile to be determined in a global way. Rather than studying individually the expression of one gene or even of a group of genes, it is now possible to perform a differential analysis of mRNA content (transcriptome) and evaluate the effect of multiple stresses or growth conditions on microbial physiology. Macroarrays or microarrays (or DNA chips) can be used for these purposes. The former consist of nylon membranes similar to those commonly used in hybridization experiments; the latter consist of glass slides chemically treated to increase the binding of DNA fragments. The probes corresponding to the genes and/or the intergenic regions of the organism under study can be spotted in duplicate: they can consist of DNA fragments obtained by PCR amplification of some or all of the genes, short oligonucleotides directly synthesized on the support or long oligonucleotides (~70-mers); DNA fragments recovered from genomic libraries may also be used when the genome of the organism has not yet been fully sequenced (Dharmadi & Gonzalez, 2004).

In each case, total RNAs are extracted from cells and then used to synthesize cDNAs with a reverse transcriptase. The labelling that will serve to detect the signal is performed during this step with radiolabelled or fluorescent group-bearing nucleotides. The radioactive labelling is similar in the two RNA pools analysed on macroarrays and all steps of the experiment must then be performed in parallel. In contrast, the use of distinct fluorochromes, e.g. Cy3 and

Cy5, in the DNA chip technology allows competitive hybridization to be performed. After hybridization of the DNA targets with the probes, the signal intensity can be measured for each spot with appropriate detection instruments (e.g. Phosphorimager for radioactivity and fluorescence scanner for fluorochromes) and dedicated software (e.g. ArrayVision for macroarrays and Genepix for microarrays). Statistical methods then allow the comparison of the signal obtained from the two RNA pools and the identification of genes overexpressed or repressed between the two conditions tested (Nadon & Shoemaker, 2002). It should be emphasized, however, that even though these methods can be applied without too much difficulty for studying natural isolates cultivated under laboratory conditions, the major limiting factor in microbial ecology results from the difficulty of extracting mRNAs from complex samples (Dennis *et al.*, 2003).

Functional genomics methods such as transcriptome differential analysis are of considerable interest in the study of interactions between organisms because of their capacity to reveal the set of mechanisms induced in both partners. In many instances, the major biological determinants for a given phenotype are known, but not the minor ones, and transcriptomic analysis can help obtain a more global view of the involved processes. For example, such experiments have been performed in *Sinorhizobium meliloti* to investigate the symbiotic physiological modifications resulting from growth inside plant nodules (Becker *et al.*, 2004). However, most of the studies to date devoted to either human pathogens or symbiotic micro-organisms have focused on the host response rather than on the physiology of the microbe itself. For example, DNA microarrays have been used to study the inflammatory reaction of various cell lines resulting from the interaction with organisms such as bacteria, viruses or protozoa (Jenner & Young, 2005).

On the other hand, many studies have investigated the adaptive mechanisms in micro-organisms either cultured independently or subjected to variations in their environmental conditions. Heat-shock resistance, anaerobic growth, oxidative stress or nutrient starvation have been studied under laboratory conditions, in particular on reference organisms such as *E. coli* (Dharmadi & Gonzalez, 2004). More recently, similar studies relying on transcriptome analysis were performed on bacteria of interest in microbial ecology, e.g. UV irradiation resistance and heavy metal reduction in *Shewanella oneidensis* (Bencheikh-Latmani *et al.*, 2005; Qiu *et al.*, 2005). The genus *Shewanella* contains numerous species characterized by great respiratory versatility and believed to play an important role in the carbon biogeochemical cycle. Several data published recently on other environmental isolates further support the innovative character of functional genomics approaches. For example, the genome of the piezophilic bacterium *Photobacterium profundum* has been sequenced and expression profiles obtained under low and high pressure have been compared. The results demonstrate

that this micro-organism has developed specific mechanisms to adapt to these perturbations. These processes result not only in the variation of stress factors as described for many other organisms, but also in deep changes in carbohydrate metabolism (Vezzi *et al.*, 2005).

Finally, a regulatory role for small RNAs has been recently demonstrated in various micro-organisms, including *E. coli* (Massé & Gottesman, 2002) and *Pseudomonas fluorescens* (Kay *et al.*, 2005). This emerging paradigm could deeply modify our vision of bacterial control and adaptation mechanisms, and transcriptomic studies could help us understand the full extent of these regulatory RNAs. In this respect, the use of tiling arrays of overlapping probes spanning the entire genome could lead to the identification of intergenic regions potentially coding for small RNAs and playing a role in the microbial response to particular stresses.

Protein and metabolite profiling

mRNAs are not usually considered as the main effectors of biological responses of organisms under specific environmental conditions, but proteins are. Moreover, a modulation in their activity does not always depend on a modified abundance level of the corresponding genes but rather on post-translational modification or proteolysis. In some cases, transcriptome analysis may therefore not constitute an appropriate approach as far as the mechanisms allowing adaptation to environmental stresses or growth under extreme conditions are concerned.

The global analysis of proteins synthesized by any organism (the proteome) usually relies on protein profiles obtained by two-dimensional (2D) electrophoresis. In a first step, this procedure allows the separation of proteins according to their charge. Next, they are separated according to their molecular mass on polyacrylamide gel. The proteins are then visualized by an organic dye (Coomassie blue), by metallic salt reduction (silver nitrate) or more recently by fluorescent labelling (DIGE). Protein separation by 2D electrophoresis has been used for many years. The development of functional genomics approaches associated with genome sequencing has however led to considerable progress in identification methods (Rabilloud, 2002). Polypeptide characterization, which previously relied on chemical procedures such as Edman degradation, can now be performed by the cheap, rapid and precise mass spectrometric analysis of peptide fingerprints. The development of mass spectrometers able to ionize and precisely determine the mass of peptides is what made it possible to link proteins with genome data. With this aim, proteins of interest are recovered from gels and enzymically digested, e.g. by trypsin, which specifically cuts the polypeptide chain at lysine or arginine residues, or with cyanogen bromide, which cuts at methionine residues. For each protein of interest, the whole set of generated peptides is analysed by mass spectrometry (MALDI-TOF) to define precisely the molecular mass of most of them. The peptide

fingerprint is then compared to sequences present in databases. Appropriate software, such as Profound, MS-fit and Mascot, generates theoretical protein profiles from all sequences in databases and compares these data to experimental fingerprints obtained by mass spectrometry. The probability that the protein of interest corresponds to one of these sequences is then evaluated. If the genome of the organism under study has not yet been sequenced, or in the case of unconvincing results, other approaches are used (Carapito *et al.*, 2006). Uncertain data can be due to unlikely sequencing errors, DNA rearrangements, translational readthrough of stop codons, translational frame-shifting and other 'recoding' events, or to post-translational modification (e.g. glycosylation or phosphorylation). In such instances, a tandem mass spectrometry method can be used, e.g. ESI MS/MS. In this procedure, several peptides obtained by trypsin digestion are randomly fragmented by collision. The mass of the various fragments is determined and their comparative analysis allows short amino acid sequences to be deduced (Aebersold & Mann, 2003). These polypeptide sequences are then used to identify the protein of interest by comparison to homologous proteins in databases with algorithms such as BLAST.

Proteomic analysis coupled with mass spectrometry does however have several limitations. Indeed, all proteins of an organism cannot be visualized whatever the detection method used because some proteins are present and effective at very low levels. Abundance levels of proteins in a yeast cell for instance can vary from one to a million per cell (Ghaemmaghami *et al.*, 2003) and it is of course difficult to monitor over such a large dynamic range. Furthermore, some proteins are quite labile while others are not. Moreover, membrane proteins are usually difficult to detect on 2D gels because of a low solubility level. Nevertheless, a global or partial proteomic map (cytoplasmic, membrane or extracellular fraction) of several organisms has been drawn. These bacteria are often model organisms, e.g. *B. subtilis* (Hecker & Volker, 2004) or human pathogens, e.g. *Mycobacterium tuberculosis*, the aetiological agent of tuberculosis, or *Helicobacter pylori*, responsible for stomach ulcer (Jungblut, 2001). Recently, the physiology of an increasing number of environmental isolates has been studied by a proteomic approach. For example, arsenic metabolism and the adaptation to cold have been investigated in *Herminiimonas arsenicoxydans* (Muller *et al.*, 2007) and in the archaeon *Methanococcoides burtonii* (Saunders *et al.*, 2005), respectively. Similarly, several proteomic approaches addressed the adaptive response of plants interacting with mycorrhizae or rhizobacteria (Kim *et al.*, 2004). The development of interactive databases, such as SWISS-2DPAGE (<http://expasy.org/ch2d/>) or InPact (Muller *et al.*, 2007) (<http://inact.u-strasbg.fr/~db/index.php>), should greatly facilitate the exploitation of these data (Fig. 4). Finally, using specific staining procedures, a 2D map of iron-metalloproteins has been recently drawn for the acidophilic archaeon *Ferroplasma acidiphilum* (Ferrer *et al.*, 2007).

The results suggest that the high content of metalloproteins in this organism may not only reflect its current iron-rich habitat but may also indicate that it represents a relic of early life on Earth, where metals were more prevalent than today due to widespread volcanic and hydrothermal activities.

Microbial cells usually contain several thousand genes. Many of them code for proteins with an enzymic activity, which therefore are suspected to generate hundreds of metabolic reactions. The constant progress in techniques of analytical chemistry, such as nuclear magnetic resonance (NMR), Fourier transform mass spectrometry (FT-MS) or high-resolution liquid chromatography (HPLC), should lead to the global identification of the various categories of metabolites present in a cell, such as saccharides (glycome), lipids (lipidome) and amino acids (Hollywood *et al.*, 2006). In model organisms, such procedures might for example enable the relationship between a mutation and a specific biosynthetic pathway to be established. In this respect, they would play an important role in functional genomics, which aims to attribute functions to the numerous hypothetical proteins identified in sequenced organisms (Kell *et al.*, 2005). What is applicable to model organisms such as *E. coli* and *S. cerevisiae* will hopefully be used in the future to characterize environmental isolates or microbial communities in their environment. Moreover, the relationship between metabolites and proteins, between proteins and nucleic acids and between proteins and proteins would constitute another development in the understanding of molecular interactions in a cell (Pu *et al.*, 2007).

Uncultured organisms and microbial communities

Microbial communities are complex and variable biological assemblies whose study has been difficult for a long time because of the unculturability of many of their components. Indeed, numerous micro-organisms are known to be recalcitrant to culture for different and often unknown reasons. This is true of many pathogens and symbionts of plants or animals. The genome of several of them has however been recently deciphered, e.g. *Mycobacterium leprae*, the aetiological agent of leprosy (Cole *et al.*, 2001), *Tropheryma whipplei*, the causative agent of Whipple's chronic disease (Bentley *et al.*, 2003), *Buchnera aphidicola*, a symbiotic organism of aphids (Shigenobu *et al.*, 2000), and *Rickettsia conorii*, an obligate intracellular parasite (Ogata *et al.*, 2001). One of the most interesting results obtained from these studies is the strong reduction in size as well as the marked global enrichment in A and T as compared to those of their phylogenetic relatives, which further supports the major impact that descriptive and functional genomics can have in the understanding of biological processes and genome evolution.

In any given environment, the biological diversity data obtained for example by the analysis of 16S RNA suggests that only a small fraction of organisms present can be

Fig. 4. The InPact proteomic database (<http://inpact.u-strasbg.fr/~db/index.php>). (A) The various functionalities of the interface (right) allow the exploration of specific areas of the 2D gel by moving the scanning box (inner square in the top image) or by using a zoom-in/out function (bottom). Spots present in the selected area can be outlined, and the corresponding MS results can be seen for each by using the Show Details function (middle). In addition, more information can be seen (left) by hovering the mouse over any spot and/or clicking on it [name, molecular mass (Mw), pI, MS peptide sequence]. (B) One of the numerous arsenate reductases identified within the genome of *Herminiimonas arsenicoxydans* (Muller *et al.*, 2007) illustrates the data that can be obtained for any protein identified by MS, e.g. the label of the corresponding CDS and the spot numbers where the protein has been identified (top), the size and the location within the genome sequence of the MS peptide fragments obtained as well as their percentage coverage with respect to the full-length CDS (bottom).

cultivated. This complexity varies from only a few taxonomic groups in non-organic environments characterized by a limited amount of energy sources to several thousands of taxa in soils and oceans where a large amount of carbon molecules are available. These communities can now be explored as a whole by the sequencing of their genomic DNA content (metagenome). Despite the difficulties in analysing, interpreting and comparing metagenomic data (Foerstner *et al.*, 2006), and depending on the diversity of microbes present in the environment under study, a reconstruction of the genome of some of them can be envisaged. Questions such as which organism is present and what metabolic functions are involved in biogeochemical transformations can now be addressed at a molecular level. For example, the metagenome of an acid mine drainage biofilm has been characterized. This work has led to the nearly complete genomic description of two chemolithotrophic micro-organisms, i.e. the bacterium *Leptospirillum* sp. and the archaeon *Ferroplasma acidarmanus* (Tyson *et al.*, 2004). Similarly, environmental genomics

has been used to reconstruct the genome of the uncultured bacterium *Kuenenia stuttgartiensis*, which plays a role in anaerobic ammonium oxidation, and to identify genes involved in various aspects of nitrogen metabolism (Strous *et al.*, 2006). The genome of several sulphur-metabolizing symbionts of the marine oligochaete *Olavius algarvensis* has also been deciphered recently (Woyke *et al.*, 2006). On the other hand, the Sargasso Sea metagenome has shown the existence of a large diversity in proteorhodopsin and phosphate transport encoding genes and led to the identification of a huge number of genes of unknown function (Rusch *et al.*, 2007). A similar metagenomic approach has also been used to address the stratification of microbial communities in the ocean and revealed important variations in gene distribution, in particular with regard to carbon and energy metabolism (DeLong *et al.*, 2006).

The interest in environmental isolates in the search for novel biocatalysts predates the development of microbial

genomics. Indeed, independently of genomic programmes, numerous enzymes have been isolated, including ones from extremophiles, for their higher thermal stability and/or their high specific activity, allowing them to function in industrial processes of degradation or synthesis (Demirjian *et al.*, 2001). Moreover, the properties of enzymes such as proteases, lipases, glycosidases and polymerases, key players in the development of molecular biology, have often been optimized by methods such as mutagenesis or molecular directed evolution. These techniques, associated with the exploration of the data resulting from genome sequencing, have led to the elaboration of new molecules able to degrade recalcitrant products, including many pollutants such as organohalogenated compounds. These strategies are important not only in the development of bioremediation processes but also in the design of new biosensors able to detect the presence of toxic chemical elements in the environment (Parales & Ditty, 2005; Zylstra & Kukor, 2005). In this perspective, metagenomic analyses have opened the way to the exploration of uncultured organisms aiming to discover novel genes (Streit & Schmitz, 2004). Despite their limitations, due to the biases inherent in screening of metagenomic DNA libraries on the basis of enzymic activities or sequence similarities, several enzymes with a high potential for industrial and/or pharmaceutical applications have been identified, including esterases and oxidoreductases as well as proteins involved in the synthesis of vitamins and molecules with antibacterial properties (Lorenz & Eck, 2005).

Finally, a novel proteomic approach is now emerging to characterize the protein content in a global way. It focuses not on specific micro-organisms but rather on the whole microbial community. From a theoretical point of view, such an approach appeared to be doomed by the major differences in molecular mass and pI for a given protein in various micro-organisms, which would result in not having a single spot for all nitrogenase proteins in a biotope, for instance. Recently, however, such a so-called metaproteomic analysis of microbes growing in water-plant sludge has been performed (Wilmes & Bond, 2004). Although preliminary, this work led to the identification of several proteins belonging to an uncultured organism of the *Rhodocyclus* lineage, which is known to accumulate polyphosphates. Such a strategy can also be used not only to describe an ecosystem, but also to study its response to perturbations. For example, a metaproteomic approach has been successfully used to study the temporal dynamics of microbial communities subjected to cadmium exposure and to characterize the resulting response in terms of toxicity and resistance (Lacerda *et al.*, 2007). Similarly, the spatial dynamics of bacterioplankton was evaluated along the Chesapeake Bay, the largest estuary in the USA, and the proteins identified were shown to correlate with major microbial lineages, i.e. Bacteroides and Alphaproteobacteria, present in this ecosystem (Kan *et al.*, 2005). Such a functional approach therefore offers considerable potential for the comprehensive analysis of ecosystem

function, even if in all likelihood only the dominant organisms' proteins will be visible. The recent development of microfluidic HPLC-Chip/MS technology will significantly improve its sensitivity (Hardouin *et al.*, 2006). Consequently, and despite the technical difficulties inherent in this method, the metaproteomic characterization of microbial communities is expected to develop widely in the next few years.

Perspectives

In the last few years, a huge amount of genomic sequences have been published in databases. There is no reason to consider that the growth in such data will slacken in the near future, in particular due to the recent development of new sequencing technologies. Quite to the contrary, these new methods will probably increase the genomic exploration of organisms thanks to a major reduction in both cost and time. Consequently, most of the known microbial taxonomic groups should soon have been investigated, at least in part. Moreover, the increasing capacity of sequencing centres associated with the development of technological platforms will greatly improve our knowledge of the structure, the functioning, the diversity and the evolution of microbial genomes.

A similar tendency should also be observed in the study of microbial communities. In this respect, metagenomic approaches based on new sequencing technologies will be of great interest for investigating complex microbe consortia as a whole and addressing the important aspect of uncultured micro-organisms growing in environments as diverse as polar ice, the human intestine or industrial sites. They will provide a marked acceleration in the generation of raw genome sequence data, rendering more acute the need for novel automated computational approaches. Associated with the development of large-scale functional methods, including metaproteomics and metabolomics, such studies should give an important insight into the extreme biodiversity of microbes inhabiting terrestrial and marine environments. All these genomic approaches, *a fortiori* if they are combined with more classical methods of geochemistry, microscopy, genetics, molecular biology and/or biochemistry, will give an integrated view of the organisms present in any environment, their role and their relationships. They will lead to a better understanding of how micro-organisms adapt to and colonize as yet largely unexplored ecological niches.

References

- Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* **422**, 198–207.
- Bailly, X., Olivieri, I., De Mita, S., Cleyet-Marel, J. C. & Bena, G. (2006). Recombination and selection shape the molecular diversity pattern of nitrogen-fixing *Sinorhizobium* sp. associated to *Medicago*. *Mol Ecol* **15**, 2719–2734.

- Becker, A., Bergès, H., Krol, E., Bruand, C., Rüberg, S., Capela, D., Lauber, E., Meilhoc, E., Ampe, F. & other authors (2004). Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. *Mol Plant Microbe Interact* 17, 292–303.
- Bencheikh-Latmani, R., Williams, S. M., Haucke, L., Criddle, C. S., Wu, L., Zhou, J. & Tebo, B. M. (2005). Global transcriptional profiling of *Shewanella oneidensis* MR-1 during Cr(VI) and U(VI) reduction. *Appl Environ Microbiol* 71, 7453–7460.
- Bentley, S. D., Maiwald, M., Murphy, L. D., Pallen, M. J., Yeats, C. A., Dover, L. G., Norbertczak, H. T., Besra, G. S., Quail, M. A. & other authors (2003). Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whippelii*. *Lancet* 361, 637–644.
- Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J., Bellgard, M., Wassenaar, T. M. & Ussery, D. W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* 6, 165–185.
- Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.
- Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S. & Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7, R36.
- Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A. & other authors (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073.
- Carapito, C., Muller, D., Turlin, E., Koechler, S., Danchin, A., Van Dorsselaer, A., Leize-Wagner, E., Bertin, P. N. & Lett, M. C. (2006). Identification of genes and proteins involved in the pleiotropic response to arsenic stress in *Caenibacter arsenoxydans*, a metalloresistant beta-proteobacterium with an unsequenced genome. *Biochimie* 88, 595–606.
- Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J. & other authors (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 34, D511–D516.
- Chen, Z., Terai, M., Fu, L., Herrero, R., DeSalle, R. & Burk, R. D. (2005). Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. *J Virol* 79, 7014–7023.
- Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honoré, N., Garnier, T., Churcher, C. & other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011.
- Collyn, F., Guy, L., Marceau, M., Simonet, M. & Roten, C. A. (2006). Describing ancient horizontal gene transfers at the nucleotide and gene levels by comparative pathogenicity island genomics. *Bioinformatics* 22, 1072–1079.
- Dary, A., Martin, P., Wenner, T., Decaris, B. & Leblond, P. (2000). DNA rearrangements at the extremities of the *Streptomyces ambifaciens* linear chromosome: evidence for developmental control. *Biochimie* 82, 29–34.
- Daubin, V., Lerat, E. & Perriere, G. (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol* 4, R57.
- Delneri, D., Brancia, F. L. & Oliver, S. G. (2001). Towards a truly integrative biology through the functional genomics of yeast. *Curr Opin Biotechnol* 12, 87–91.
- DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U., Martinez, A., Sullivan, M. B., Edwards, R. & other authors (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496–503.
- Demirjian, D. C., Moris-Varas, F. & Cassidy, C. S. (2001). Enzymes from extremophiles. *Curr Opin Chem Biol* 5, 144–151.
- Dennis, P., Edwards, E. A., Liss, S. N. & Fulthorpe, R. (2003). Monitoring gene expression in mixed microbial communities by using DNA microarrays. *Appl Environ Microbiol* 69, 769–778.
- Dharmadi, Y. & Gonzalez, R. (2004). DNA microarrays: experimental issues, data analysis, and application to bacterial systems. *Biotechnol Prog* 20, 1309–1324.
- Ettema, T. J. G., de Vos, W. M. & van der Oost, J. (2005). Discovering novel biology by *in silico* archaeology. *Nat Rev Microbiol* 3, 859–869.
- Ferrer, M., Golyshina, O. V., Beloqui, A., Golyshin, P. N. & Timmis, K. N. (2007). The cellular machinery of *Ferroplasma acidiphilum* is iron-protein-dominated. *Nature* 445, 91–94.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.
- Foerster, K. U., Von Mering, C. & Bork, P. (2006). Comparative analysis of environmental sequences: potential and challenges. *Philos Trans R Soc Lond B Biol Sci* 361, 519–523.
- Frangeul, L., Nelson, K. E., Buchreiser, C., Danchin, A., Glaser, P. & Kunst, F. (1999). Cloning and assembling strategies in microbial genome projects. *Microbiology* 145, 2625–2634.
- Ghaemmhami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. & Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature* 425, 737–741.
- Giraud, E., Moulin, L., Vallenet, D., Barbe, V., Cytryn, E., Avarre, J. C., Jaubert, M., Simon, D., Cartieaux, F. & other authors (2007). Legumes bioses: absence of *nod* genes in photosynthetic bradyrhizobia. *Science* 316, 1307–1312.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C. & other authors (1996). Life with 6000 genes. *Science* 274, 563–567.
- Green, E. D. (2001). Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet* 2, 573–583.
- Hardouin, J., Duchateau, M., Joubert-Caron, R. & Caron, M. (2006). Usefulness of an integrated microfluidic device (HPLC-Chip-MS) to enhance confidence in protein identification by proteomics. *Rapid Commun Mass Spectrom* 20, 3236–3244.
- Hecker, M. & Volker, U. (2004). Towards a comprehensive understanding of *Bacillus subtilis* cell physiology by physiological proteomics. *Proteomics* 4, 3727–3750.
- Hirochika, H., Nakamura, K. & Sakaguchi, K. (1984). A linear DNA plasmid from *Streptomyces rochei* with an inverted terminal repetition of 614 base pairs. *EMBO J* 3, 761–766.
- Hollywood, K., Brison, D. R. & Goodacre, R. (2006). Metabolomics: current technologies and future trends. *Proteomics* 6, 4716–4723.
- Holt, J. G., Krieg, N. R., Sneath, P. H. A., Staley, J. T. & Williams, S. T. (1994). *Bergey's Manual of Determinative Bacteriology*, 9th edn. Baltimore: Williams & Wilkins.
- Hou, S., Saw, J. H., Lee, K. S., Freitas, T. A., Belisle, C., Kawarabayasi, Y., Donachie, S. P., Pikina, A., Galperin, M. Y. & other authors (2004). Genome sequence of the deep-sea gamma-proteobacterium *Idiomarina loihiensis* reveals amino acid fermentation as a source of carbon and energy. *Proc Natl Acad Sci U S A* 101, 18036–18041.
- Jenner, R. G. & Young, R. A. (2005). Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol* 3, 281–294.

- Jungblut, P. R. (2001). Proteome analysis of bacterial pathogens. *Microbes Infect* 3, 831–840.
- Kahn, P. (1995). From genome to proteome: looking at a cell's proteins. *Science* 270, 369–370.
- Kan, J., Hanson, T. E., Ginter, J. M., Wang, K. & Chen, F. (2005). Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems* 1, 7.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34, D354–D357.
- Kay, E., Dubuis, C. & Haas, D. (2005). Three small RNAs jointly ensure secondary metabolism and biocontrol in *Pseudomonas fluorescens* CHA0. *Proc Natl Acad Sci U S A* 102, 17136–17141.
- Kell, D. B., Brown, M., Davey, H. M., Dunn, W. B., Spasic, I. & Oliver, S. G. (2005). Metabolic footprinting and systems biology: the medium is the message. *Nat Rev Microbiol* 3, 557–565.
- Kim, S. T., Kim, S. G., Hwang, D. H., Kang, S. Y., Kim, H. J., Lee, B. H., Lee, J. J. & Kang, K. Y. (2004). Proteomic analysis of pathogen-responsive proteins from rice leaves induced by rice blast fungus, *Magnaporthe grisea*. *Proteomics* 4, 3569–3578.
- Kimura, M. (1979). The neutral theory of molecular evolution. *Sci Am* 241, 98–100.
- Krummenacker, M., Paley, S., Mueller, L., Yan, T. & Karp, P. D. (2005). Querying and computing with BioCyc databases. *Bioinformatics* 21, 3454–3455.
- Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A. & other authors (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256.
- Lacerda, C. M. R., Choe, L. H. & Reardon, K. F. (2007). Metaproteomic analysis of a bacterial community response to cadmium exposure. *J Proteome Res* 6, 1145–1152.
- Lin, Y. S., Kieser, H. M., Hopwood, D. A. & Chen, C. W. (1993). The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol Microbiol* 10, 923–933.
- Liu, Z., Bos, J. I., Armstrong, M., Whisson, S. C., da Cunha, L., Torto-Alalibo, T., Win, J., Avrova, A. O., Wright, F. & other authors (2005). Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. *Mol Biol Evol* 22, 659–672.
- Lorenz, P. & Eck, J. (2005). Metagenomics and industrial applications. *Nat Rev Microbiol* 3, 510–516.
- Maltsev, N., Glass, E., Sulakhe, D., Rodriguez, A., Syed, M. H., Bompada, T., Zhang, Y. & D'Souza, M. (2006). PUMA2 – grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res* 34, D369–D372.
- Masse, E. & Gottesman, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A* 99, 4620–4625.
- Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S. Y., Mori, T., Nishida, K., Yagisawa, F. & other authors (2004). Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428, 653–657.
- Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. (2005). The microbial pan-genome. *Curr Opin Genet Dev* 15, 589–594.
- Muller, D., Medigue, C., Koechler, S., Barbe, V., Barakat, M., Talla, E., Bonnefoy, V., Krin, E., Arsene-Ploetze, F. & other authors (2007). A tale of two oxidation states: bacterial colonization of arsenic-rich environment. *PLoS Genet* 3, e53.
- Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 93, 10268–10273.
- Nadon, R. & Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends Genet* 18, 265–271.
- Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A. & Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314, 267.
- Nirenberg, M. W. & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* 47, 1588–1602.
- Normand, P., Lapierre, P., Tisa, L. S., Gogarten, J. P., Alloisio, N., Bagnarol, E., Bassi, C. A., Berry, A. M., Bickhart, D. M. & other authors (2007). Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res* 17, 7–15.
- Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P. E., Barbe, V., Samson, D., Roux, V., Cossart, P., Weissenbach, J. & other authors (2001). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293, 2093–2098.
- Oliynyk, M., Samborsky, M., Lester, J. B., Mironenko, T., Scott, N., Dickens, S., Haydock, S. F. & Leadlay, P. F. (2007). Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat Biotechnol* 25, 447–453.
- Oppendoes, F. R. & Coombs, G. H. (2007). Metabolism of *Leishmania*: proven and predicted. *Trends Parasitol* 23, 149–158.
- Ou, K., Ong, C., Koh, S. Y., Rodrigues, F., Sim, S. H., Wong, D., Ooi, C. H., Ng, K. C., Jikuya, H. & other authors (2005). Integrative genomic, transcriptional, and proteomic diversity in natural isolates of the human pathogen *Burkholderia pseudomallei*. *J Bacteriol* 187, 4276–4285.
- Parales, R. E. & Ditty, J. L. (2005). Laboratory evolution of catabolic enzymes and pathways. *Curr Opin Biotechnol* 16, 315–325.
- Paulsen, I. T., Press, C. M., Ravel, J., Kobayashi, D. Y., Myers, G. S., Mavrodi, D. V., DeBoy, R. T., Seshadri, R., Ren, Q. & other authors (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat Biotechnol* 23, 873–878.
- Pradella, S., Hans, A., Sproer, C., Reichenbach, H., Gerth, K. & Beyer, S. (2002). Characterisation, genome size and genetic manipulation of the myxobacterium *Sorangium cellulosum* So ce56. *Arch Microbiol* 178, 484–492.
- Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S. J. (2007). Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. *Proteomics* 7, 944–960.
- Qiu, X., Tiedje, J. M. & Sundin, G. W. (2005). Genome-wide examination of the natural solar radiation response in *Shewanella oneidensis* MR-1. *Photochem Photobiol* 81, 1559–1568.
- Rabilloud, T. (2002). Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* 2, 3–10.
- Ronaghi, M. (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Res* 11, 3–11.
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M. & other authors (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5, e77.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74, 5463–5467.
- Saunders, N. F. W., Goodchild, A., Raftery, M., Guilhaus, M., Curmi, P. M. G. & Cavicchioli, R. (2005). Predicted roles for hypothetical

proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanococcoides burtonii*. *J Proteome Res* 4, 464–472.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. & Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief Bioinform* 3, 246–251.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407, 81–86.

Sivachenko, A. Y. & Yuryev, A. (2007). Pathway analysis software as a tool for drug target selection, prioritization and validation of drug mechanism. *Expert Opin Ther Targets* 11, 411–421.

Streit, W. R. & Schmitz, R. A. (2004). Metagenomics – the key to the uncultured microbes. *Curr Opin Microbiol* 7, 492–498.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M. W., Horn, M., Daims, H., Bartol-Mavel, D. & other authors (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440, 790–794.

Thomson, N. R., Howard, S., Wren, B. W., Holden, M. T., Crossman, L., Challis, G. L., Churcher, C., Mungall, K., Brooks, K. & other authors (2006). The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. *PLoS Genet* 2, e206.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C. (2006). MaGe: a

microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 34, 53–65.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr, Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell* 88, 243–251.

Vezzi, A., Campanaro, S., D'Angelo, M., Simonato, F., Vitulo, N., Lauro, F. M., Cestaro, A., Malacrida, G., Simionati, B. & other authors (2005). Life at depth: *Photobacterium profundum* genome sequence and expression analysis. *Science* 307, 1459–1461.

Wan, X.-F., Verberkmoes, N. C., McCue, L. A., Stanek, D., Connelly, H., Hauser, L. J., Wu, L., Liu, X., Yan, T. & other authors (2004). Transcriptomic and proteomic characterization of the Fur regulon in the metal-reducing bacterium *Shewanella oneidensis*. *J Bacteriol* 186, 8385–8400.

Watson, J. D. & Cook-Deegan, R. M. (1991). Origins of the human genome project. *FASEB J* 5, 8–11.

Watson, J. D. & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature* 171, 737–738.

Wilmes, P. & 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.

Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., Boffelli, D., Anderson, I. J., Barry, K. W. & other authors (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443, 950–955.

Zylstra, G. J. & Kukor, J. J. (2005). What is environmental biotechnology? *Curr Opin Biotechnol* 16, 243–245.