Environmental Microbiology (2003) doi:10.1046/j.1462-2920.2003.00505.x

Editorial

The future of microbial genomics

Since the first complete microbial genome was published in 1995, more than 100 microbial genomes have been completely sequenced and published, and another 300 microbial genome sequencing projects are estimated to be in progress worldwide. The significance of the information that has been derived from these complete individual genomes cannot be underestimated. Sequencing technologies have improved considerably, and the overall costs for sequencing have reduced to the point that sequencing a complete microbial genome, although still sometimes accompanied by various difficulties, is now almost routine. The field of microbial genomics has moved away from the primary initial focus on pathogen genomes to include the sequencing of diverse prokaryotes that occupy a range of environmental niches, and which are responsible for an array of environmental processes. Every genome that has been sequenced to date has provided new insight into biological processes, activities, and potential of these species that was not evident before the availability of the genome sequence. We have gained unprecedented insight into gene transfer (Nelson *et al*., 1999; Perna *et al*., 2001), environmental applications (Nelson *et al*., 2002), and the virulence mechanisms in many of these species (Tettelin *et al*., 2000; Tettelin *et al*., 2001; Tettelin *et al*., 2002). Sequence databases and comparative tools are now more easily accessible and allow for successful comparisons of different genomes, the identification of metabolic pathways and the analysis of transporter profiles across various species. Most significantly, the tremendous success of genome sequencing has allowed us to pursue other avenues where we can now derive genomic information from the multitudes of uncultivable prokaryotic species and complex microbial populations that exist in nature.

Accompanying the improved technologies that have resulted in lowered costs and improved bioinformatics tools, are new technologies that have bolstered our abilities to work with non-traditional species. Organisms for which limited quantities of DNA are available can now be sequenced due to the availability of reagents that allow for exponential DNA amplifications. For example, *Epulopiscium*, the largest known heterotrophic bacterium (600 μm by 60 μm), is currently being sequenced by The Institute for Genomic Research (TIGR) in Rockville, Maryland. Even though we are still unable to culture this bacterium in the laboratory, the genome sequence stands to provide important information on the unusual ability of *Epulopiscium* to produce live multiple offspring internally, as well as the changes that occurred early in the transition from the prokaryotic cell to eukaryotic cell. In addition, we anticipate that metabolic reconstruction and successful growth of this organism based on the functionally annotated genome will be possible. Generation of sufficient DNA for sequencing of *Epulopiscium* has been made possible, in part, by the development of bacteriophage ϕ29 DNA polymerase as a tool for multiply primed rolling circle amplification of DNA directly from cells or plaques (Dean *et al*., 2001; Dean *et al*., 2002; Repli-G™, Molecular Staging; TempliPhi™, Amersham Biosciences). The anticipated success of this project suggests that many uncultivable species for which we can sort a limited number of cells and generate small quantities of DNA can now be sequenced to completion.

Although we are making progress in working with uncultivable species, it is evident that we will have to develop large-scale technologies to handle the cultivation of the large number of microbial species in nature. Zengler and workers have described a high-throughput cultivation method that employs the encapsulation of cells in gel microdroplets under low nutrient flux conditions followed by flow cytometry, to detect microdroplets that contain microcolonies (Zengler *et al*., 2002). By trying to mimic the concentrations of these nutrients in their natural environments, some species can be successfully cultivated where they could not previously. They have shown that this technique can successfully be applied to multiple environments. The authors correctly point out that although the use of bacterial artificial chromosomes (BACs) allow for the identification of novel metabolic processes, cultivation will ultimately be necessary if a comprehensive understanding of these species is desired.

The improvement of tools for sequencing and assembly, accompanied by a reduction in costs, has also been a significant boost to the field of environmental genomics. We know that the associations of microorganisms in the environment are significant, and many play major roles in elemental recycling, the conversion of biomass and the onset of disease. Molecular techniques such as 16S rDNA

sequencing and phylogenetic analysis, restriction fragment length polymorphism (RFLP) and fluorescent *in situ* hybridization (FISH) analysis have allowed us to make tremendous advances in terms of being able to identify the extent of microbial diversity in various environments. These techniques, however, have revealed limited to no information on the genetic diversity contained within these environments. For example, the physiological role that is being played by individual species identified by 16S rDNA sequencing cannot be defined. Some level of analogy may be drawn by extrapolation to available genetic information from related species, but it is apparent from whole genome sequencing that species that appear to be closely related from 16S rDNA sequences may have tremendous differences in genome composition (Perna *et al*., 2001). Without being able to cultivate these organisms, the option of sequencing and analysing large DNA fragments retrieved directly from the environment becomes more attractive in an attempt to increase our understanding on how communities control various processes. Deciphering the genetic information of uncultured species can currently be achieved by sequencing genomic libraries that are created directly from environmental DNA (Beja *et al*., 2000a; Beja *et al*., 2002a; Quaiser *et al*., 2002; Liles *et al*., 2003). This development has taken advantage primarily of the successful construction of large insert BAC and fosmid libraries (up to 120 kb in size). Initial surveys have demonstrated an unanticipated level of microbial diversity that remains to be explored (Beja *et al*., 2002a; 2002b). Currently, this technique of sequencing and analysing genetic information obtained directly from the environment has been extended to the analysis of soils, the human oral cavity and gastrointestinal tract (GIT), and the Sargasso Sea. Although different styles of libraries may be used, whether a combination of small and medium insert libraries of BACs and fosmids, one major challenge in working with some of these environments is anticipated to be the successful assembly of all the sequence data so that large contiguous pieces of DNA which contain operons or links to phylogenetic markers can be generated. Micro-heterogeneity in these environments may also be an issue. The ideal would be to regenerate extremely large contiguous pieces of DNA (overlapping fosmids for example) or, if possible, complete microbial genomes of previously uncultivated species from these samples. It remains to be seen however, if current assembly software can handle this challenge.

One can envision that, with detailed bioinformatic analysis, an entire microbial community could be handled in a manner similar to the pipeline developed for annotating a microbial genome. As such, a single database could contain all of the predicted ORFs, their putative annotation, and biological role categories from a particular environment sample at a particular point in time. By examining the role assignments, it then becomes possible to identify all the biochemical pathways, and characterize the metabolic capacity of the entire microbial community that may be present. Just as a photorhodopsin was identified in the study by DeLong and co-workers (Beja *et al*., 2000b) based on the analysis of a single fosmid clone, an unprecendented amount of biochemical and physiochemical data undoubtedly will be obtained from the sequencing of other complex environments. The soil metagenome definitely will yield new antibiotics, as well as catabolic pathways for ring-based compounds and pathways for the synthesis of other secondary metabolites. The human GIT metagenome will yield significant information on the metabolic potential of the species that inhabit the GIT, their fermentation potential and the end products that they are capable of producing. This in turn will have implications for human GIT function, human health and the potential for these species to cause diseases. By having the metagenome of a healthy individual, we will be in a position to compare with the metagenome of a diseased individual allowing for the identification of possible microbial species and factors that are responsible for the onset of various diseases.

In June of this year, the first metagenomics meeting was held in Darmstadt Germany. This timely meeting was the innovation of groups at Darmstadt University of Technology and BRAIN Biotechnology Research and Information Network in Darmstadt, Germany, and the VAAM Functional Genomics Group in Feldafin, Germany. The meeting brought together some of the most respected leaders in the field including Ed DeLong from the Monterey Bay Aquarium Research Institute (California, USA), Oded Beja (Technicon-Israel Institute for Technology), Michael Wagner (University of Vienna, Austria) and Bill Martin (University of Dusseldorf, Germany). Representatives from a number of American and European funding agencies were also present. The discussions and presentations highlighted progress in the field of metagenomics, studies which have spread to include a range of soil ecosystems, biofilms, and sites from a number of oceans and seas. Soils in particular are being heavily studied for important enzyme activities, industrial biocatalysts and novel antibiotics. It is anticipated that this meeting is only the forerunner of many more of its kind that will be held on both sides of the Atlantic.

Regardless of all the progress that we are making as environmental microbiologists, we are still presented with the situation where at the individual microbe level, close to 40% of each genome remains as hypothetical or conserved hypothetical proteins. In addition, it is humbling to realize that no single prokaryote has been studied to the point that all the gene functions within that organism are known. This demands that high-throughput methodologies be developed for the efficient analysis of these large data sets, including high-throughput proteomics, gene

expression and protein–protein interaction studies. It is anticipated that these types of methodologies will further extend into an analysis of microbial communities where new techniques for studying the complex communities need to be developed, such that the associations between the previously cultivated and the estimated greater than 99% of uncultivated species can be evaluated. Microarrays are rapidly becoming standard laboratory tools for investigating gene expression under different conditions, as well as for looking at the presence and absence of genes in different strains or species that are related to a reference genome. It is encouraging to note that some recent studies have demonstrated the success of using microarrays in detecting gene expression in very complex microbial communities. Suppressive subtractive hybridation (SSH) has also been used successfully to identify differences in community composition from environmental samples such as the rumen (B. White, pers. comm.), and this methodology also holds significant promise as a tool for environmental genomics studies.

The tremendous successes of genomics and our foray into metagenomics means that scientists will need to continue their discussions so that newly developing techniques can be rapidly exchanged, and efforts are not duplicated. Funding and fruitful collaborations, both domestic and international, will need to be continuously expanded, and it is anticipated that the high quality of work which has come out of the initial studies in community genomics will be kept to as a high a standard as we have seen for the complete whole genome sequencing of various microbial species.

Karen E. Nelson
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

References

Beja, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P., *et al.* (2000a) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2:** 516–529.

Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al.* (2000b) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902–1906.

Beja, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L., *et al.* (2002a) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microbiol* **68:** 335–345.

Beja, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., *et al.* (2002b) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415:** 630–633.

Dean, F.B., Nelson, J.R., Giesler, T.L., and Lasken, R.S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11:** 1095–1099.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99:** 5261–5266.

Liles, M.R., Manske, B.F., Bintrim, S.B., Handelsman, J., and Goodman, R.M. (2003) A Census of rRNA Genes and Linked Genomic Sequences within a Soil Metagenomic Library. *Appl Environ Microbiol* **69:** 2684–2691.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399:** 323–329.

Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Fouts, D., *et al.* (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4:** 799–808.

Perna, N.T., Plunkett, G., III, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., *et al.* (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157: H7. *Nature* **409:** 529–533.

Quaiser, A., Ochsenreiter, T., Klenk, H.P., Kletzin, A., Treusch, A.H., Meurer, G., *et al.* (2002) First insight into the genome of an uncultivated crenarchaeote from soil. *Environ Microbiol* **4:** 603–611.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287:** 1809–1815.

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., *et al.* (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293:** 498–506.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Eisen, J.A., Peterson, S., Wessels, M.R., *et al.* (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci USA* **99:** 12391–12396.

Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E.J., Short, J.M., and Keller, M. (2002) Cultivating the uncultured. *Proc Natl Acad Sci USA* **99:** 15681–15686.