EDITORIAL

Sea change for metagenomics?

Ed DeLong comments on the impact that the Global Ocean sampling expedition will have on marine and Earth science.

Large-scale genomic surveys of microbial communities are currently expanding massively in number, scope and pace. Recent genomic forays into complex microbial communities include acid-mine drainage sites[1], symbiotic associations[2], pollutant removing bioreactors[3] and the human microbiome[4]. As microbial community genomic surveys accumulate, deciphering the genetic and functional "differences that make a difference" within and between different microbial habitats is becoming ever more feasible.

In their most recent metagenomics tour de force, Craig Venter and colleagues reported a (mostly) ocean surface water microbial sequencing survey that has nearly doubled the number of known protein sequences[5,6]. The 41 randomly collected microbial samples in the 'Global Ocean Sampling' (GOS) cumulatively encompass ~6.6 billion base pairs of DNA, translating into ~6 million predicted protein sequences[6]. One impressive point that derives from comparative analyses is that even at this vast scale of sampling, the rate of new protein-family discovery is still linear, as new sequences are sampled. This is the same trend that was pointed out early on in whole-genome sequencing efforts[7]. So, even after a super-sized survey like this, we are nowhere near saturation with respect to sampling extant sequence space[6]. As a consequence, for example, the GOS study showed that the apparently limited taxonomic distribution of some protein families is probably just an artefact of gross under-sampling.

However, many of the observations and conclusions from the GOS study were largely confirmatory. For example, the nature and extent of genome sequence variation among dominant, closely related planktonic bacteria (such as *Prochlorococcus* species) had already been reported[8]. Similarly, proteorhodopsin amino-acid sequence variation, previously identified and experimentally shown to have potential adaptive significance in variable light fields[9], also showed similar trends in the GOS dataset.

What is new in the GOS study is the sheer size of the dataset, and the novel methods and tools that the authors needed to develop to deal with its magnitude. Size truly matters. These datasets raise new issues regarding data management, computational resources, sampling and analytical strategies, and the downstream analyses that will be necessary to begin to decipher the biological significance of Nature's nucleic acid parts list. Simply on the basis of size alone, the GOS dataset is a milestone in the endeavour to understand the magnitude and scope of efforts that will be required to make sense of microbial genomic and functional diversity in the sea.

'Impedance mismatch' refers to an inadequate (or excessive) ability of one system to accommodate the input from another. The phenomenal increase in metagenomics data, although extraordinarily useful, is also accelerating a type of impedance mismatch as the amount of genomic data outstrips our abilities to interpret it and to test and confirm functional hypotheses. Similarly rich, quantitative datasets at other levels of biological organization, that together represent the expression of genomic information — from the transcriptome, to the proteome and the 'metabolome', to the cell, populations, communities and ecosystems — need to be developed, and soon. Only by gathering quantitative datasets that traverse this biological hierarchy, along with associated environmental information, will biological systems on our planet be more comprehensively understood.

Interpreting these new trans-hierarchical datasets in the context of system behaviour will present significant new challenges for microbiologists, theoretical ecologists and Earth systems scientists alike. Physicists, genome biologists, biochemists, physiologists, computational biologists, mathematicians, environmental scientists — and yes, microbiologists — will all contribute to a more quantitative and integrated interpretation of the microbial Earth system.

As new methods and technologies begin to alleviate some of microbiology's current impedance mismatch, a much deeper understanding of the inner workings of our microbial planet is certain to emerge. Metagenomics is an important part of this journey, but is surely not the final destination.

Edward F. DeLong is a Professor in the Division of Biological Engineering & Department of Civil and Environmental Engineering, Room 48-427, Massachusetts Institute of Technology, 77 Massachusetts 02139, USA.
e-mail: DeLong@mit.edu

1. Tyson, G. W. *et al. Nature* **428**, 37–43 (2004).
2. Woyke, T. *et al. Nature* **443**, 950–955 (2006).
3. Garcia Martin, H. *et al. Nature Biotechnol.* **24**, 1263–1269 (2006).
4. Turnbaugh, P. J. *et al. Nature* **444**, 1027–1031 (2006).
5. Rusch, D. B. *et al. Plos Biol.* **5**, e77 (2007).
6. Yooseph, S. *et al. PLoS Biol.* **5**, e16 (2007).
7. Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V. & Ouzounis, C. A. *Genome Biol.* **4**, 401 (2003).
8. Coleman, M. L. *et al. Science* **311**, 1768–1770 (2006).
9. Beja, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. *Nature* **411**, 786–789 (2001).