Meeting report
Bacterial comparative genomics
Michael B Prentice

Address: Department of Microbiology, University College Cork, Cork, Ireland. E-mail: m.prentice@ucc.ie

A report on 'Genomes 2004: International Conference on the Analysis of Microbial and Other Genomes', Hinxton, UK, 14-17 April 2004.

A measure of the maturity of the technology for sequencing small genomes was the notable absence of detail about the sequencing process in presentations at the Genomes 2004 meeting. Instead, the focus was very much on the use of genome-sequence data to answer biological questions about bacterial evolution, physiology and pathogenicity. The value of obtaining multiple genome sequences from closely related organisms and the use of genome arrays to compile genome inventories (gene lists) of unsequenced species of different origin and phenotype were themes that were repeated in many presentations and posters.

The keynote presentation by Philip Hugenholtz (University of California, Berkeley, USA) reminded us of the limitations of organisms cultured on agar plates for representing microbial communities, and of the inbuilt bias of the current excess of proteobacteria among sequenced organisms. 'Metagenomic sequencing' of specific environments and ecosystems could correct this bias, but massive data-handling capacity would be required. In a step away from *in vitro* pure cultures and towards metagenomic sequencing, Hugenholtz showed that shotgun-sequencing reads of total DNA from a simple microbial community - a biofilm from an acid-mine drainage system - could be compiled into scaffolds (genome reconstructions with permitted gaps of a known size range) using a conventional single-genome assembly program. These scaffolds, when grouped (binned) by overall G+C content, could be assigned to the *Leptospirillum* and *Ferroplasma* bacterial species that were known, from DNA-probing experiments, to be present in the biofilm. The gene complement of the *Ferroplasma*-like scaffolds was comparable to that of a previously sequenced strain, but 22% divergence was seen at the nucleotide level. This sequence and assembly process allows community-based metabolic reconstruction and can provide insights into bacterial evolution, but at present it works only for a simple underlying bacterial community.

Insights into pathogenicity
Understanding what distinguishes pathogenic from non-pathogenic species is a major focus of bacterial comparative genomics. Jacques Ravel (The Institute for Genomic Research, Rockville, USA) presented work by Claire Fraser's group to sequence three strains of *Bacillus anthracis*, the cause of anthrax, and to search for sites of nucleotide polymorphism that could then be assessed for their discriminating power across a wide collection of *B. anthracis* strains. He also showed rather worrying sequence data from a naturally occurring *Bacillus cereus* strain. The *B. anthracis* genome is closely related to that of *B. cereus* - a ubiquitous soil organism and opportunistic human pathogen - but it contains some chromosomal differences and two additional plasmids: pX01 encoding a PA toxin complex, and pX02 encoding a capsule biosynthesis operon. *B. cereus* strain G9241 has been linked to fatal pneumonia in the USA; its genome sequence shows that it contains plasmid pBCX01 encoding the PA toxin of *B. anthracis*, and another plasmid, pBC218, apparently specifying a novel capsule. In a mouse model, this *B. cereus* strain was at least as lethal as *B. anthracis*. The implication of these findings is that the organism we know as *B. anthracis* is merely one heavily sampled subtype of *B. cereus* and that there may be other lethal surprises waiting in the soil.

The genome sequence of *Neisseria meningitidis* Group A Z2491 was used in the pathogenicity studies described by Xavier Nassif (Faculté de Médecine Necker-Enfants Malades, Paris, France). Using a microarray to compare strains of *N. meningitidis* isolated from asymptomatic carriers with strains associated with disease, Nassif has identified a 9 kilobase (kb) meningococcal disease island. This island includes an outer membrane protein gene and was strongly associated with strains causing disease in patients over ten years of age. Preliminary experimental data suggest that this

region is capable of self excision from the chromosome and could be a filamentous phage.

Continuing the theme of comparing closely related species, George Weinstock (Baylor College of Medicine, Houston, USA) set out various approaches using the *Treponema pallidum* genome sequence to evaluate the pathogenesis of syphilis and to develop potential new diagnostic tests and vaccine candidates. Systematic cloning of the vast majority of *T. pallidum* open reading frames has allowed the expression of their products as fusion proteins with glutathione-S-transferase (GST). Screening these fusion proteins for immunogenicity with antisera from rabbits and humans infected with *T. pallidum* detected 34 novel proteins and possible vaccine targets, the majority comprising exported proteins. Using a *T. pallidum* genomic array to compare *T. pallidum* with the closely related Yaws agent, *Treponema pertenue*, showed that all *T. pallidum* genes are present in *T. pertenue*, and that they have a very similar expression profile in animal models. Any differences in phenotype between the two species are presumably due to very minor sequence differences (less than 6 base-pairs in 10 kb) and the absence of four genes in *T. pallidum* that are present in *T. pertenue*.

The anaerobic opportunistic pathogen *Bacteroides fragilis* was the subject of the presentation by Julian Parkhill (Wellcome Trust Sanger Institute, Hinxton, UK). Comparison of the genome sequences of two strains, NCTC9343 and 638R, shows over 12% of nucleotides lie in unique sequences. Although promoter inversion is a common mechanism for generating phase variation in bacteria, a striking feature of both of these genomes is the number and variety of loci under the control of invertible promoters, including genes encoding polysaccharide synthetases, two-component regulators, restriction endonucleases and outer membrane proteins. In comparison, the published sequence of the related gut commensal organism *Bacteroides thetaiotaomicron* did not show these particular phase-variable systems, and 50% of its genes were not shared with *B. fragilis*.

In a brief focus on a eukaryotic pathogen, Nina Agabian (University of California, San Francisco, USA) presented genomic and experimental data suggesting that the opportunist fungal pathogen *Candida albicans*, in contrast to the non-pathogenic yeast *Saccharomyces cerevisiae*, may preferentially grow by oxidative metabolism using lipid substrates, producing lipoxin-like substances which affect the host immune system.

Bacterial evolution

A second major focus for comparative studies is attempting to understand the evolution of present day bacteria species. Stephen Gordon (Veterinary Laboratories Agency, Weybridge, UK) described a comparison of the genomes of *Mycobacterium tuberculosis* and *Mycobacterium bovis* (the

cause of tuberculosis in cattle). He found deletions in the *M. bovis* genome compared with *M. tuberculosis*, suggesting that popular anthropological theories of the origin of tuberculosis coinciding with the domestication of cattle are unlikely to be correct. Comparison of multiple genome sequences of *M. tuberculosis* and the *M. bovis* sequence shows an unexpectedly high ratio of non-synonymous to synonymous mutations across all coding sequences. A possible evolutionary hypothesis to explain this would be that the divergence of *M. bovis* and *M. tuberculosis* was so recent that there has been insufficient time for purifying selection to operate against non-synonymous mutations.

Bdellovibrio bacteriovorus is a bacterial predator of Gram-negative bacteria. Its genome has been analyzed by Stephan Schuster (Max-Planck Institute for Developmental Biology, Tübingen, Germany) and was found to encode many hydrolases, motility genes and transporters, but key amino-acid synthesis pathways are absent; these features fit the organism's predatory, parasitic lifestyle. Schuster and colleagues found no evidence of recent gene transfer to *B. bacteriovorus* from its prey.

Databases for genome-scale analyses

Minoru Kanehisa (Kyoto University, Japan) gave an update on the Kyoto Encyclopedia of Genes and Genome (KEGG) suite of databases [http://www.genome.ad.jp/kegg/], including iKEG, a custom-made KEGG database feature that is being developed to enable automatic annotation and pathway mapping for any species. Additionally, further integration of the chemical and protein databases using a common graph-based analytical approach is in progress. Ross Overbeek (The Fellowship for the Interpretation of Genomes, Burr Ridge, USA) suggested that single-genome annotation was inevitably less efficient than pan-genomic annotation of subsystems; the latter process is shortly to be facilitated by his SEED annotation tool, which will be publicly available. A similar comparative approach allowing the systematic filling of pathway holes was described by Peter Karp (SRI International, Menlo Park, USA) as a new feature of his Pathway Tools software [http://bioinformatics.ai.sri.com/ptools/].

Alex Bateman (Wellcome Trust Sanger Institute) presented the Rfam database (available at [http://www.sanger.ac.uk/Software/Rfam/] and [http://rfam.wustl.edu/]), which is a collection of multiple sequence alignments and covariance models to aid recognition of non-coding RNA families, many of which were previously omitted from bacterial genome-sequence annotations because of their small size. The HAMAP project [http://www.expasy.org/sprot/hamap/] was presented by Amos Bairoch (Swiss Institute of Bioinformatics, Geneva, Switzerland). This project aims automatically to annotate a significant percentage of proteins originating from microbial genome-sequencing projects, using manually curated protein families.

The meeting confirmed that the shotgun-sequencing approach pioneered by Fred Sanger is moving on to a third phase of its revolutionizing effect on microbiology. First came gene sequences, then genome sequences of key human pathogens and laboratory organisms, and now multiple genome sequences of closely related organisms and metagenomic sequences of microbial communities are attainable. These data will provide us with a wealth of new insights into the biology of prokaryotes.