Meeting report

A golden age for microbial genomics

Tom Coenye

Address: Laboratorium voor Microbiologie, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium. E-mail: Tom.Coenye@UGent.be

Published: 21 October 2003

Genome ***Biology*** 2003, **4:**347

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/347

A report on the main symposium ‘Exploiting genomes: from bases to megabases in 50 years’ at the 153rd Meeting of the Society for General Microbiology (SGM), Manchester, UK, 8-9 September 2003.

To celebrate the golden jubilee of Watson and Crick’s discovery of the structure of DNA, the main symposium of the SGM’s 153rd meeting addressed a wide range of topics related to bacterial genomics. For additional information on the meeting, see the SGM website [http://www.sgm.ac.uk/meetings/pdfabstracts/umist2003abs.pdf].

Comparing microbial genomes

George Weinstock (Human Genome Sequencing Center, Houston, USA) opened the symposium with a talk illustrating the generic approach to characterizing a microbial genome, using *Treponema pallidum*, the spirochete that causes syphilis, as an example. Following sequencing, gene prediction and functional annotation, microarrays can be developed, knockout mutants can be created, genes can be cloned and recombinant proteins produced. Weinstock pointed out several steps in the process where there is room for improvement. For example, there has been very little change in shotgun-sequencing methodology over the past few years, and future gains in time and efficiency will depend on the development of new methodologies: little sequence-assembly software specifically designed for microbial genomes has been developed, for example. Future challenges for genome annotation and analysis include the need to rely more on experimental evidence instead of automated computer-assisted predictions, and updating and correcting the databases.

The latest completed genome sequences from various groups of bacteria were reviewed by several speakers. Julian Parkhill (The Sanger Institute, Hinxton, UK) focused on the comparative analysis of the genomes of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *B. pertussis* and *B. parapertussis* cause whooping cough in humans, whereas *B. bronchiseptica* causes respiratory tract infections in a wide range of organisms. Despite their close phylogenetic relationship, there are huge differences in gene content and in the number of pseudogenes present in these three bacterial species. Detailed analyses indicate that *B. parapertussis* and *B. pertussis* are independent derivatives of *B. bronchiseptica*-like ancestors and that their evolution seems to have been driven by gene loss and gene inactivation, which could explain their limited host range. The increased virulence of *B. pertussis* for humans is thought to result from changes in the regulatory machinery that controls the expression of virulence factors. The underlying reason for the greater virulence might be that increased opportunities for transmission in a growing human population eliminated the need to keep damage to the host to a minimum. Parkhill also discussed the genome sequences of two other human pathogens, *Bacteroides fragilis*, an opportunistic pathogen, and *Tropheryma whipplei*, the bacterium associated with Whipple’s disease, which causes malabsorption of nutrients in the intestine. A remarkable feature of the *B. fragilis* genome is the presence of many invertible promoters (including promoters controlling the expression of outer-membrane proteins and exported proteins), which provides enormous possibilities for variation. *T. whipplei* has a relatively small genome (925,938 bp) but contains a large amount (approximately 5%) of non-coding repetitive DNA. This non-coding DNA is located in two clusters that are almost opposite one another in the chromosome. The genome also contains a large number of genes coding for a family of predicted surface proteins called WiSP and some of these genes are associated with non-coding repetitive DNA.

The complete genome sequences from *Yersinia pseudotuberculosis*, *Yersinia enterocolitica*, and two strains of *Yersinia pestis* have recently become available, and their detailed analysis allows us to trace the evolution of *Yersinia* species and to determine which factors are responsible for the virulence of *Y. pestis*, the causative agent of plague. Emilio Garcia (Lawrence Livermore National Laboratory, Livermore, USA) presented a comparative analysis of these genomes. It turns out that the molecular evolution of pathogenicity in *Y. pestis* has involved extensive expansion of insertion sequences, lateral gene transfer, extensive genomic rearrangements, genome reduction by deletion, or inactivation and point mutations. Annotation of the inactivated genes in *Y. pestis* revealed that most were involved in carbohydrate or amino-acid metabolism or in motility, which presents a fundamentally different situation from that in the genus *Bordetella*, where mostly regulatory genes have been inactivated.

Whole-genome DNA microarrays recently developed for the genus *Yersinia* are also being used to study its evolutionary genomics, as described by Stewart Hinchliffe (School of Hygiene and Tropical Medicine, London, UK). Genome-wide microarray analysis of a large number of *Y. pestis* and *Y. pseudotuberculosis* strains of diverse origins has identified a number of chromosomal differences between species, biovars, serotypes and strains of *Y. pestis* and *Y. pseudotuberculosis* that may provide insights into the evolution of these organisms at the species and sub-species level.

Siv Andersson (Uppsala University, Sweden) presented a comparative analysis of the genomes of sequenced α-proteobacteria. There is tremendous variation within the α-proteobacteria in genome size (ranging from 1 to almost 10 megabases) and structure (the genome often consists of multiple circular or linear replicons), but little is known about the underlying mechanisms that cause this diversity. Andersson presented data that showed a correlation between the lifestyle of the organism and its genome content. Plant-associated bacteria such as *Mesorhizobium loti* have undergone extreme genome expansion (up to a few thousand genes), whereas the shift to an intracellular environment and vector-mediated transmission (for example, in *Bartonella* species, which are thought to be transmitted by fleas) has resulted in extreme genome reduction.

With a 580 kilobase genome encoding 480 proteins, *Mycoplasma genitalium* is the smallest living organism that can grow by itself. Clyde Hutchison (University of North Carolina, Chapel Hill, USA) used this organism as a model to answer questions regarding the minimal genome. Comparative analysis and transposon mutagenesis in *M. genitalium* and the closely related *Mycoplasma pneumoniae* revealed that the *M. genitalium* genome contains 265 to 350 essential genes. The function of many of these genes is unknown, and studies are now underway to fully define the minimal genome and correlate the essential genes to biological functions (see the Berkeley Structural Genomics website [http://www.strgen.org]).

Data from the competitive hybridization of whole genomes of the *Streptomyces violaceoruber* clade against *Streptomyces coelicolor* were presented by Kannika Duangmal (University of Newcastle, Newcastle upon Tyne, UK) and revealed that most genes in *S. coelicolor* are conserved in members of the *S. violaceoruber* species group. It turns out that all members of the *S. violaceoruber* clade have similar patterns of genes missing. These data, from whole-genome microarray comparisons, correlate well with results obtained with DNA-DNA hybridization, or with fingerprinting by amplified fragment-length polymorphism (AFLP) or rep-PCR fingerprinting, which uses primers corresponding to repetitive elements in bacterial genomes.

Gene duplication has long been considered a mechanism for the acquisition of new functions through evolution. Dirk Gevers (Ghent University, Belgium) described the occurrence and distribution of duplicated genes in 107 bacterial whole-genome sequences. A significant degree of gene duplication was found in all genomes (up to 44%), and the amount correlates linearly with genome size. Up to 16% of all open reading frames (ORFs) in any given genome and 3.6% of the total combined proteome of all 107 genomes occurred in block duplications, rather than duplications of single genes. Approximately 94% of the blocks are the typical size of a bacterial operon (three to four genes), indicating a putative mechanism for operon duplication and evolution.

Exploiting the genome with bioinformatics, microarrays, proteomics and glycomics

Whole-genome microarrays from fully sequenced genomes are a powerful tool for identifying differences in gene content between organisms and for studying gene expression. Strains of *Staphylococcus aureus* resistant to the antibiotic vancomycin present a potentially serious public-health problem. In this context, Terry Gaasterland (Rockefeller University, New York, USA) described the development of a multistrain *S. aureus* microarray. Pairwise comparisons of the available genomes of strains of *S. aureus* have revealed considerable variation in gene content across the epidemiological landscape. The ultimate goal of this work is to identify changes in protein-coding potential that correlate with antibiotic resistance; this will be achieved by measuring differences in gene expression in vancomycin-sensitive and vancomycin-resistant pairs of *S. aureus* isolates. Philip Butcher (St George's Hospital Medical School, London, UK) uses microarrays to help understand the complex pathophysiology of *Mycobacterium tuberculosis* infection. Besides discussing some methodological aspects of microarray work (for example, the many problems associated with isolating mRNA from *M. tuberculosis* cells infecting macrophages), Butcher focused on the use of *M. tuberculosis* microarrays to investigate the intracellular lifestyle of this organism and its interaction with host macrophages. In the future, results from work like this can be combined with results from microarray work on the mammalian host and will provide a transcriptome analysis of host-pathogen interactions.

Properties of DNA sequences other than their simple coding potential can be used to help characterize genomes, as David Ussery (Center for Biological Sequence Analysis, Lyngby, Denmark) illustrated in his 'DNA-centric' perspective on whole-genome sequences. Several physical properties of DNA sequences, including curvature, stacking energy and position preference, can be used to detect specific features in whole-genome sequences. One example is the detection of promoter sequences on the basis of their structural properties. Structural parameters also correlate well with gene-expression levels as revealed by microarray data, and thus can be used to predict which genes are highly expressed. Several tools are available on the Center for Biological Sequence Analysis website [http://www.cbs.dtu.dk].

David O'Connor (Center for Proteomic Research, Southampton, UK) highlighted the many applications of proteomics, which include validation of genome sequences by confirming gene function, identification of protein function, monitoring expression levels and studying post-translational modifications. He highlighted several new tools for proteomics including techniques to absolutely quantify proteins from mass spectrometry (MS) data and desorption/ionization on silicon time-of flight MS (DIOS-TOF/TOF-MS). This latter technique provides high accuracy and significant fragmentation information and will be particularly useful for the characterization of biomolecules. O'Connor illustrated the power of proteomics to characterize regulatory systems in bacteria, using the ribosome-binding GTPase protein BipA, a regulatory protein that coordinates the expression of multiple regions in the genome, as an example.

Genome sequences provide not only an opportunity for deciphering bacterial proteomes but also indirect insights into the generation of the cell-surface carbohydrate structures - the glycome - that contribute to the virulence of many bacterial pathogens. A detailed description of cell-surface polysaccharide structures of the food-borne pathogen *Campylobacter jejuni* was presented by Brendan Wren (London School of Hygiene and Tropical Medicine). Despite the limited genetic repertoire of this organism, *C. jejuni* produces two types of glycolipid structure and possesses two glycosylation systems. Besides the general *N*-linked glycosylation process, *C. jejuni* also has an *O*-linked flagellin glycosylation pathway. Together, these pathways are responsible for glycosylating more than 40 proteins, which may play important roles in adherence and invasion.

Two posters from the Bacterial Pathogenesis and Genomics Unit at the University of Birmingham (Birmingham, UK) described new online facilities for genome research. Arshad Khan presented ViruloGenome [http://www.vge.ac.uk], a facility providing free tools for analyzing and exploiting data from completed and partial (in progress) bacterial genomes, including the ability to perform PSI-BLAST searches against proteins from unfinished genomes. Roy Chaudhuri described *coli*BASE, an online database for *Escherichia coli*, *Salmonella* and *Shigella* comparative genomics [http://colibase.bham.ac.uk]. Similar databases have been developed for *Clostridium* (*Clostri*DB) [http://clostri.bham.ac.uk] and *Campylobacter* (*Campy*DB) [http://campy.bham.ac.uk].

Overall, the talks in this symposium showed that we are truly in a golden age for microbial genomics, as the availability of many genome sequences and numerous tools to analyze and compare them enables great steps forward to be made.