Big Time for Small Genomes

Eugene V. Koonin

Genome Res. 1997 7: 418-421
Access the most recent version at doi:10.1101/gr.7.5.418

REVIEW

Big Time for Small Genomes

Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

A new field of research has emerged that would have been unthinkable to most people as recently as 2 years ago. There is no generally accepted name for this discipline, but ''genome-based microbiology'' seems to be appropriate. As indicated in the title of the recent meeting ''Small Genomes: Sequencing, Functional Characterization and Comparative Genetics'' (The Institute for Genomic Research Genomic Series, Hilton Head, SC, January 25–28, 1997), this field capitalizes on complete sequences of ''small'' genomes, which effectively means genomes of unicellular organisms. The information derived from complete genome sequences and particularly from their comparative analysis is used explicitly to study the biology of the microbial cells and to infer phylogenetic conclusions. At the meeting, approximately one-half of the talks still belonged to the genomic era, that is, they were progress reports on projects aimed at complete sequencing of a particular genome or, in some cases, several genomes. The remaining talks ventured into the postgenomic era and presented either genome comparison studies or functional analysis based on the knowledge of a complete genome sequence. As a substantive indication of the importance attached to the new field, the meeting featured brief presentations by representatives of three major funding agencies, National Institutes of Health (USA), Department of Energy (USA), and the Wellcome Trust (UK), all of which provide significant support for further development and worldwide coordination of the microbial genomics effort.

In his introductory overview of the small genome sequencing efforts, J. Craig Venter (The Institute for Genomic Research, Rockville, MD, hereafter TIGR) indicated that even though the number of completely sequenced genomes is currently very small, the trend toward exponential (if not faster) growth is apparent—two complete genomes appeared in 1995, four in 1996, and it is realistic to expect eight or more in 1997. Therefore, there is no doubt that within a short period of time, we will have ample material for comparative genomics. It is anticipated that in the next 1–2 years, ~50 complete genome sequences of bacteria and archaea will become available. The advanced genome projects presented at the meeting are listed in Table 1. According to Dr. Venter, TIGR is committed to maintaining an updated list of sequenced genomes at its World Wide Web site (http://www.tigr.org). Conceptually and methodologically most of the genome projects appear similar. The shotgun cloning and sequencing approach originally applied by TIGR to sequencing of the *Haemophilus influenzae* genome clearly has conquered the small genome community. The individual reports differed primarily in the details of the methods used for contig assembly and gap closing. It is worth noting that the computer system called MAGPIE, which was developed by Terry Gaasterland (Argonne National Laboratory, IL) and was described in her talk at the meeting, is becoming an increasingly important tool for organizing the data obtained in genome projects and performing their initial analysis.

The singular landmark for which this meeting will be remembered, however, was achieved using the more traditional approach of directed sequencing on the basis of a detailed genome map. This was the completion of the *Escherichia coli* genome sequence announced by Fred Blattner (University of Wisconsin, Madison) and backed up by a poster presentation from Hirotada Mori (Nara Institute of Science and Technology, Japan). The significance of the fact that we now have the complete genome structure of the ultimate model organism, which probably has been studied experimentally in more detail than all the rest taken together, is hard to overestimate. Even if, in the general context of this meeting, the determination of a bacterial genome sequence may seem a relatively mundane project, it is to be remembered that the *E. coli* genome project started in a different epoch (only 6 years ago!), and its completion certainly is a major scientific feat. Besides, *E. coli* is still by a good margin the longest prokaryotic genome sequence available. The genome consists of 4,638,858 bp and according to the preliminary annotation reported by Fred Blattner, contains 4286 protein-coding genes. The next num-

[1]E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301) 480-9241.

7:418–421 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

Table 1. Advanced Microbial Genome Projects and Their Status[a]

Species and taxonomic position	Genome size (Mb)	Status	Presenter (institution)
		Bacteria	
Escherichia coli (Proteobacteria)	4.64	complete; released, submitted to GenBank with preliminary annotation; detailed annotation in progress	Fred Blattner (University of Wisconsin, Madison)
Helicobacter pylori (Proteobacteria)	1.67	complete; annotation in progress	Jean-Francois Tomb (TIGR)
Treponema pallidum (Spirochaetales)	1.05	nearly complete (several gaps remaining)	Claire Fraser (TIGR)
Borrelia burgdorferi (Spirochaetales)	1.3	in progress, completion expected in 1997	Claire Fraser (TIGR)
Ureaplasma urealyticum (Gram-positive bacteria, Mycoplasmas)	0.76	nearly complete	Elliot J. Lefkowitz (University of Alabama, Birmingham)
Streptococcus pneumoniae (low G+C Gram-positive bacteria)	2.05	nearly complete (several gaps remaining)	Brian A. Dougherty, TIGR
Neisseria gonorrhoeae (Proteobacteria)	2.05	>95% complete; contig sequences available through WWW	Bruce A. Roe (University of Oklahoma, Oklahoma City)
Streptococcus pyogenes (low G+C Gram-positive bacteria)	1.92	>95% complete; contig sequences available through WWW	Bruce A. Roe
Mycobacterium tuberculosis (Gram-positive bacteria, Actinomycetes)	4.4	>75% complete	Robert D. Fleischmann (TIGR); Bart Barrell (The Sanger Center, Hinxton, Cambridge, UK)
Rickettsia prowazekii (Proteobacteria)	1.1	>80% complete; detailed analysis in progress	Siv G.E. Andersson (Uppsala University, Sweden)
Aquifex aeolicus (oxygen-reducing bacteria)	1.55	nearly complete (2 short gaps remaining)	Ronald V. Swanson (Recombinant Biocatalysis Inc., Sharon Hill, PA)
		Archaea	
Methanobacterium thermoautotrophicum (Euryarchaeaota, Methanobacteriales)	1.7	complete; annotation in progress	Douglas R. Smith (Genome Therapeutics Corporation, Waltham, MA)
Archaeoglobus fulvidus (Euryarchaeaota, Archaeaoglobales)	2.2	nearly complete (one short gap remaining); annotation in progress	Hans-Peter Klenk (TIGR)
Pyrobaculum aerophilum (Crenarchaeota, Thermoproteales)	2.2	nearly complete; annotation in progress	Sorel Fitz-Gibbon (University of California, Los Angeles)
Pyrococcus furiosus (Euryarchaeaota, Thermococcales)	2.1	in progress; complete genomes covered by clones	Robert B. Weiss (University of Utah, Salt Lake City)
Sulfolobus solfataricus (Crenarchaeota, Sulfolobales)	3	>50% sequenced; completion expected by the end of 1997	Christoph W. Sensen (Institute for Marine Biosciences, National Research Council of Canada, Halifax, Nova Scotia)

[a]Complete genomes released and published prior to the meeting are not included. Genome projects that are only in their early stages are not mentioned either.

ber mentioned by Fred Blattner is striking: Only 1817 *E. coli* genes (42%) have a known function. Although strong functional predictions can be derived by computer analysis, for many of the remaining genes this is a good measure of our ignorance about the simple cell that has been the primary object of molecular biology for several decades.

In contrast to the remarkable pace of bacterial and archaeal genome sequencing, relatively little progress has been reported for genomes of unicellular eukaryotes. The genome of the baker's yeast *Saccharomyces cerevisiae*, a synopsis of which was presented at the Hilton Head meeting by Bernard Dujon (Institut Pasteur, Paris, France; see below), is going to be the only one available in this category for at least several years to come. The project aimed at complete sequencing of the ~30-Mb genome of *Plasmodium falciparum* genome, described by S.L. Hoffman (Naval Medical Research Institute, Bethesda, MD), is only beginning to gain momentum, and there are significant difficulties with the isolation of individual chromosomes for subsequent shotgun sequencing and with cloning of very AT-rich sequences that abound in the *Plasmodium* genome. Somewhat paradoxically, it is practically certain that the next eukaryotic genome to be sequenced in its entirety will not be a ''small'' one from a unicellular eukaryote but the relatively large genome of the nematode *Caenorhabditis elegans*.

All the new genomes are important for comparative genomics, and some are of particular interest judging from their apparent phylogenetic position; for example, representatives of new archaeal taxa such as *Archaeoglobus fulgidus* (Hans-Peter Klenk, TIGR), *Pyrobaculum aerophilum* (Sorel Fitz-Gibbon, University of California, Los Angeles), and *Sulfolobus solftaricus* (Christoph Sensen, National Research Council of Canada, Halifax, Nova Scotia), or the early branching bacterium *Aquifex aeolicus* (Ronald Swanson, Recombinant Biocatalysis, Inc., Sharon Hill, PA). Most of the speakers, however, did not present results of detailed comparisons or phylogenetic analyses, apparently deferring such in-depth studies to the time when the sequences are completed and the basic annotation is finished. Two notable exceptions were the presentations by Richard Herrmann (University of Heidelberg, Germany) on the complete genome sequence of *Mycoplasma pneumoniae* and its comparison with *Mycoplasma genitalium*, and by Siv Andersson (Uppsala University, Sweden) on sequencing the *Rickettsia prowazekii* genome and comparing it to other rickettsial species. The two mycoplasmas provide the first opportunity for researchers to compare complete, closely related bacterial genomes. The gene set of the smaller *M. genitalium* is essentially a subset of the larger *M. pneumoniae* gene set, and there is significant synteny between the gene orders in these two species. Much of the difference in the gene number between *M. pneumoniae* and *M. genitalium* (677 and 470 genes, respectively) can be attributed to gene duplications in the former. The results of the *Rickettsia* genome project were presented in the general context of the evolution of intracellular parasites. Andersson and colleagues observed several features that appear to connect *Rickettsia* with mitochondria and distinguish them from mycoplasmas, for example, the presence, in the rickettsial genome, of genes for TCA cycle enzymes, cytochromes, and ATP/ADP translocases, which are all typical mitochondrial functions.

Computational biology as applied to genome analysis was represented by five plenary lectures, as well as a whole session of short, more technical talks on computational methods. Hamilton Smith (Johns Hopkins University, Baltimore, MD) analyzed regulatory signals in intergenic regions of *H. influenzae*. Monica Riley (Marine Biology Laboratory, Woods Hole, MA) discussed issues of functional classifications of gene products relative to orthologous and paralogous relationships identified by protein sequence comparison. Peter Karp (Center for Artificial Intelligence, Supercomputer Research Institute, Menlo Park, CA) presented his reconstruction of *H. influenzae* metabolism and discussed general construction principles of metabolic pathway databases. Antoine Danchin (Institut Pasteur) extended his earlier observations, originally made on a limited set of *E. coli* sequences, that bacterial genes can be partitioned into three classes using multivariate analysis of codon composition. I described the results of comparative analysis of complete sets of proteins encoded in bacterial and archaeal genomes, emphasizing the importance of combined application of several sensitive computer methods. Although it is obvious that theoretical comparative genomics is still struggling to develop a coherent language as well as an appropriate set of methods and significance criteria, a consensus appears to be emerging on some important issues. Thus, it has already become clear that gene order conservation in distantly related bacteria and archaea is almost nonexistent, with only a few essential operons (primarily those encoding ribosomal proteins) conserved. This conclusion, based on multiple genome comparisons, was made by J.L. Siefert (University of Houston, TX) and was substantiated in several other talks, particularly by Jean-Francois Tomb (TIGR),

who demonstrated the lack of synteny between *H. influenzae* and *Helicobacter pylori*, two bacterial species that are relatively close phylogenetically and have about the same number of genes. The comparison of the *M. genitalium* and *M. pneumoniae* genomes described by Richard Herrmann can be considered a positive control for these analyses, as in these closely related species, pronounced conservation of the gene order was detected in spite of the difference in the number of genes. Another emerging generalization is that in each of the small genomes a large fraction of proteins—up to 50%—belong to families of paralogs.

Even more than comparative genomics, analysis of gene functions, based on the knowledge of complete genome sequence, is still in its infancy. An impressive illustration of the potential of this approach was presented by Richard Moxon (Oxford University, UK), who described the use of the complete genome sequence of *H. influenzae* for the identification of a number of new genes involved in the biosynthesis of the lipopolysaccharide, an important determinant of pathogenicity. Clyde A. Hutchison III (University of North Carolina, Chapel Hill) presented the initial results of an ambitious project aimed at the construction of a microorganism with a minimal gene set, that is, containing no dispensable genes. The first phase of the project includes transposon-mediated mutagenesis of *M. genitalium* genes and identification of those genes that can be mutated without rendering the cell nonviable. Theoretical estimates suggest that the minimal set may consist of ~250 genes, and so >200 mutants must be identified. So far, eight have been detected. And even when all of the nonessential gene are defined—a significant achievement in itself—the formidable task of actually constructing the minimal cell will remain. But the goal appears to be worth the effort, both as a demonstration of a principle and as an approach to the creation of an extremely interesting model system. Bernard Dujon outlined the research program of EUROFAN, the European consortium whose ultimate goal is the complete functional characterization of the yeast genome. In its first phase the program concentrates on ORPHANS, genes that have no obvious homologs in other organisms. Daniel Shoemaker (Stanford University, Palo Alto, CA) described a methodology for rapidly constructing a large number of yeast mutants with deleted open reading frames based on a modified PCR targeting method and including introduction of unique tags for strain tracking in mixed populations. So far, deletion mutants have been produced for 500 genes, and the method can be scaled up to the complete genome. Lisa Wodicka (Affymetrix, Santa Clara, CA) presented the initial results of complete gene expression monitoring in yeast using high-density oligonucleotide arrays. It has been shown that 85% of the genes are expressed on a rich medium and 60% on a minimal medium. Preliminary as they are, these communications showed that novel, genome-oriented technologies for studying gene functions are already here, and their repertoire should be expected to grow in parallel with the accumulation of complete genome sequences.

Last but not least, several presentations discussed genome-oriented approaches to the analysis of microbial diversity in natural habitats. Eric Matur (Recombinant Biocatalysis, Inc., La Jolla, CA) described a strategy for enzyme discovery based on expression cloning with nucleic acids extracted directly from the environment, without prior cultivation of microorganisms. The approach has already yielded several hundred enzymes with potential industrial applications. Edward DeLong (University of California, Santa Barbara) described genome analysis of uncultivated nonthermophilic crenarchaeota using various techniques for cloning DNA directly from the environment. The preliminary results clearly show that nonthermophilic crenarchaeota are closely related to thermophilic species. Finally, Norman Pace (University of California, Berkeley) outlined the large-scale microbial diversity studies performed by his laboratory in a variety of natural habitats using universal primers to PCR rRNA directly from environmental samples. These studies reveal a dramatic diversity of bacteria and archaea and indicate that only 0.001%–0.1% of the prokaryotic species can be cultivated with standard techniques. Furthermore, this estimate is based on the assumption, far from proven, that the ''universal'' rRNA primers are actually universally conserved in all life forms. It is impossible to rule out that a fourth and a fifth domain of life exist, but with current techniques, they may be discovered only by accident. This is an apt reminder that with all of the impressive advances of microbial genomics, we are only touching the tip of the iceberg, and new technological breakthroughs such as methods allowing the determination of long sequences using DNA extracted from a single cell may still change our view of the microbial world.

In her concluding remarks, Claire Fraser (TIGR) announced that from now on the Small Genome meeting will be held every year, thus solidifying this new field of genome-based microbiology.

The abstracts of the meeting are published in *Microbial & Comparative Genomics* (1996) Volume 1, Number 4.