The Beautifully Simple but Intriguingly Complex World of Small Genomes

Sandra W. Clifton and Bruce A. Roe

Genome Res. 1998 8: 331-333
Access the most recent version at doi:10.1101/gr.8.4.331

The Beautifully Simple but Intriguingly Complex World of Small Genomes

Sandra W. Clifton and Bruce A. Roe[1]

Department of Chemistry and Biochemistry, The University of Oklahoma, Norman, Oklahoma 73017 USA

In his review of the Small Genome Meeting in 1997, Eugene Koonin remarked that we have quite a bit of ''... ignorance about the simple cell (*Escherichia coli*) that has been the primary object of molecular biology for several decades.'' This message was borne out again this year at the recent conference on Microbial Genomes II: Sequencing, Functional Characterization and Comparative Genomics (The Institute for Genomic Research Genomic Series, Hilton Head, SC, January 31–February 3 1998), where experienced sequencers and microbiologists continued to exchange information about the latest disoveries in genomics and comparative genomics, that is, polygenomics. As Claire Fraser cochair of the meeting, pointed out in her introductory remarks, the emphasis of this year's meeting was to garner new biological insights from these genomic sequences. With the sequence of seven genomes completed in 1997 and >50 small genome sequencing projects under way worldwide, it is anticipated that at least 25–30 genomes will be completed this year. The small genome sequencing projects reported at the plenary sessions are summarized as supplementary material at www.genome.org or can be obtained at either http://www.mcs.anl.gov/home/gaasterl/genomes.html or http://www.tigr.org/tdb/mdb/mdb.html.

With the availability of the relatively large number of sequences available and with many more anticipated, comparative genomics now is a fait accompli. This was exemplified by Fred Blattner's report comparing sample sequences from the bacterial pathogens *E. coli* O157:H7, *E. coli* CFT073, and *Yersinia pestis* to the completed genomic sequence of *E. coli* K-12 in an attempt to identify unique regions that might be responsible for pathogenesis. In *E. coli* O157:H7, 300 unique regions were identified that possibly could define a group of virulence genes (a pathosphere) that interact with the host; 10% are large pathogenicity islands and 90% are smaller segments. Claire Fraser reported a comparative analysis of the single circular spirochetes *Treponema pallidum* chomosome with the single linear *Borrellia burgdorferi* chromosome and its 9 circular plasmids and 12 linear plasmids. There was very little sequence similarity between each organism's individual genes, but both organisms have evolved seemingly independent regions coding for similar function. Like the mycoplasmas, both lack the biosynthetic pathways for amino acid biosynthesis, the TCA cycle, and the electron transport system. They also represent a class of minimal genomes; nearly 50% of the genes are unique, having no biological function yet defined. In his analysis of the *B. burgdorferi* plasmids, Sherwood Casjens observed a large number (65%) of plasmid-borne paralogous gene families, short repeat tracts, and 12 kb of a 63-bp repeat and rhetorically questioned whether linear plasmids should be dealt with as chromosomes or plasmids. Steven Norris presented the novel aspects of the *T. pallidum* genome, which evades host defenses by having very few surface proteins. Of note, this organism contains no genes resembling toxins, but it has genes for many surface proteins that could be possible virulence genes. Consistent with the genes of other organisms, almost one-quarter of *B. bergdorferi* genes are not similar to any known genes and thus have an unknown function. In reporting the sequence of a third mycoplasma genome, *Ureaplasma urealyticum*, John Glass noted that it is quite different than the two apparently similar *Mycoplasma genitalium* and *Mycoplasma pneumoniae* genomes. Not only does the *Ureaplasma* have a large number of genes without orthologs in either of the other two *Mycoplasma* species and vice versa, but it also has a significantly rearranged genomic organization, something that is being seen in many supposedly related microorganisms.

Just when we thought we understood a biological pathway, such as the aminoacylation of tRNA, Dieter Soll presented results that refute the dogma that aminoacyl-tRNA synthetases are highly conserved and that each amino acid has its own respective synthetase and its cognate tRNA. Comparative analysis of several microbial genomes now reveals the absence of some of the tRNA synthetase homologs. In addition, as new genomic sequence data become available, new biochemical pathways and related enzymes also will begin to emerge, such as a subsidiary pathway (GntII) in L-iodinate catabolism reported and experimentally confirmed in *E. coli* by Tyrrell Conway. The revelation of an additional pathway even in this well-studied model organism suggests that there will be many other surprises in the unique genes yet to be uncovered in additional microorganisms.

In phylogenetic and evolutionary studies, Robert Feldman provided evolutionary placement of genes from *Aquifex aeolicus*, a marine hyperthermophile, with other organisms by grouping them based on their cellular role. Evidence presented suggested that gene trees and species trees do not necessarily match, as species trees can have their own phylogenetic characteristics while gene divergence usually precedes species divergence. John Reeve discussed the phylogenetics of methanogenesis, comparing the methane genes from the two now complete genomes of *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*. Because both organisms have a common ancestry their metha-

[1]Corresponding author.
E-MAIL broe@ou.edu; FAX (405) 325-4762.

nogenesis pathway was used to construct a phylogenetic tree. However, the resulting tree was bidirectional raising the possiblity that such comparisons might not result in legitimate trees as there might be a different selective pressure on each organism. Alternatively, the phylogeny of methanogenesis may not parallel that of the respective organisms that carry out the process. Siv Andersson's analysis of gene order and phylogeny in the recently completed *Rickettsia prowasekii* genome indicated that both the *R. prowasekii* nuclear and mitochondrial genomes result from deletions and rearrangements of a larger common ancestral aproteobacterial genome. Milton Saier provided a phylogenetic classification of transport proteins in the genomes of five bacteria, one archaeon, and one eukaryote, using a novel system based on mode of transport and energy metabolism, phylogenetic family, phylogenetic cluster, and substrate specificity. Selective pressures presumably determined the evolutionary fate of a transporter family as the evolutionary trees reflected an evolutionary history in which all the transporters studied had evolved from a common ancestor; this supports the notion, as seen in antibiotic resistance genes, that there is little horizontal transfer between, but significant gene transfer within, kingdoms.

Another major topic was genomic functional analysis. Here, Andre Goffeau provided an update on the status of the systematic analysis of 1000 orphan genes of *Saccharomyces cerevisiae* by the 130 European laboratories of the EUROFAN consortium. They have produced 600–700 PCR-driven mutants using a kanamycin-resistant cassette to disrupt each gene, which is subsequently tagged with green fluorescent protein for localization. They plan to continue these studies in collaboration with several U.S. and Canadian investigators with the goal of deleting and tagging all 5585 yeast genes before the turn of the century, with microarrays of the entire yeast proteome being prepared at Stanford University. Goffeau also mentioned that the *Schizosaccharomyces pombe* genome is almost complete, and the start of sequencing of a sporozoan and of two fungi, *Aspergillus nidulans* and *Neurospora crassa,* will begin this year. An interesting discussion of the calcium-related proteins in *S. cerevisae,* even though many ABC proteins, P-type ATPases, and permeases have already been previously cataloged, showed new and still unexpected related genes now have been found. For example, Karen Ketchum's presentation of the evolutionarily conserved transport proteins showed that in the genomes of Archae and Eubacteria at least 30% of the ORFs in each species were predicted membrane proteins and that the transporter profiles of autotrophs and heterotrophs differ. Heterotrophs have an abundance of amino acid and carbohydrate transporters, whereas autotrophs have more ionic transporters, indicating that the transporter profile determines the environmental niche occupied by and the biosynthetic capability of microorganisms. Hydropathy profiles for hypothetical transporter genes in *M. genitalium, M. jannaschii,* and *Haemophilus influenzae* helped classify a total of 256 new putative transporters. Richard Roberts discussed restriction modification (R/M) systems. Many previously unknown methylases were discovered, often with the corresponding restriction enzyme coding gene not always apparent, although a nearby ORF of unknown function is seen. These putative genes likely confer a new specificity, such as involvement in recombination or gene expression, or they could be genes of unknown function. Partial genomic sequence data also are useful as the available data from the *Neisseria gonorrhoeae* genome reveals putative restriction/modification genes that subsequently were cloned and tested. Two of the putative genes were active in vivo, indicating that new restriction/modification genes will still be discovered as new sequences become available. James Musser used target genes for molecular population genetic analysis in the pathogenic mycobacteria and streptococci studies where 165 strains of the highly virulent serotype M-1 *Streptococcus pyogenes* showed that 65 alleles existed with a total of 26 insertions or deletions in-frame for the SIC (complement inhibitor) gene. Because virulence genes generally are not polymorphic, these results are rather startling. Analysis of 35 target genes from ~800 *Mycobacterium* strains indicated that *Mycobactarium tuberculosis* is a ''new'' pathogen in evolutionary terms and that >95% of the structural gene polymorphism is linked directly to drug selection. The mechanisms for mycobacterial drug resistance were discussed by Lynn Meisel, who focused on the mycobactericidal drug isoniazid (INH). In an attempt to locate new targets, in addition to KatG, which activates the INH prodrug, and InhA, which is involved in the production of mycolic acid, a mycobacterium cell wall component, temperature-sensitive *ndh* mutants, revealed that some resistant organisms had a reduced capacity to bind NADH (Ndh). Defects in the Ndh enzyme could lead to defects in NADH oxidation causing NADH accumulation or NAD^+ depletion, which may inhibit KatG enzyme-mediated activation of INH or, alternatively, they may impede target inhibition by the activated INH. The temperature sensitivity of some *ndh* mutants suggested that this enzyme might be yet another target for antituberculosis drugs.

Regarding environmental genomics, Christa Schleper reported preparing and walking on fosmid libraries from the uncultivated archaeon *Cenarchaem sybiosum,* which lives in association with a marine sponge. Sequence analysis of 16S rRNA and the DNA polymerase indicated a close relationship of the organism to other members of the extremely thermophilic crenarcheota. Because the DNA polymerase, which can be expressed in *E. coli,* has a structure similar to other thermophilic DNA polymerases, this organism has been classified as a thermophilic archaeon. David Reiman, in his discussion of genomics and the identification of uncultivated microbial pathogens, pointed out that sequence-based environmental genomics can and should have a role in the identification of unculturable pathogens directly from the host as long as thought and care are given to the choice of strains, sequence targets, and clinical specimen.

A series of presentations described studies of individual enzymes. David Gelfand characterized several new thermostable polymerases to obtain new enzymes with more desirable DNA sequencing capabilities, specifically, the ability to accept new dye-labeled terminators as substrates. In another discussion, indicating the importance of the ability of some microorganisms to recycle metals such as iron and manganese, Kenneth Nealson pointed out that oxidation of the reduced iron and manganese plentiful in ancient anoxic

oceans might be responsible for the formation of our present-day oxygen-containing atmosphere. Rita Colwell felt that pathogenesis should be considered an environmental event, pointing out that *Vibrio cholerae*, a pathogen that has been associated with plankton copepods, can become dormant and nonculturable when environmental conditions are negative, subsequently surviving in ocean currents for up to 3 years. Thus, by correlating global climatic predictions and the presence of the copepod host, the likelihood of an outbreak or epidemic of cholera could be anticipated.

Although no separate session was held for informatics, it is an underlying theme of any sequencing and/or analysis project, and several talks dealt exclusively with this subject. In an inspiring lecture, Piotr Slonimski proposed several laws of genomics. Using new mathematical and computer tools, all of the ORFs of the completed genomes were aligned, analyzed, and grouped as paralog and ortholog clusters. A number of regularities were observed and used as the basis for these laws of genomics. The ''first law of genomics'' was that the frequency of paralogous duplications follow a simple distribution, that is, one-eighth of the genes is present in clusters of two homologous sequences, one-sixteenth in clusters of three, and so on. The proposed ''second law of genomics'' is based on deriving a Global Genomic Proximity Index, that is, orthologs, using the protein sequece and protein families common to different species. This second law states that the number of orthologous protein sequences is directly proportional to the genome size and that Eucarya are evolutionarily closer to the Archea than to the Eubacteria. Initial results from Masaru Timita and colleagues using E-CELL, a new software environment for whole cell modeling that is being used to simulate the cellular processes of *M. genitalium*, indicated that the minimal cell requires 127 genes. They anticipate future work involving modeling of cell division, amino acid and nucleotide biosynthesis, the TCA cycle, electron transport, and pH and gene regulation pathways. In contrast, Scott Peterson experimentally determined a minimal microbial gene complement, using transposon mutagenesis of the *M. genitalium* and *M. pneumoniae* genomes, where 285 protein-coding genes were identified as minimal gene set. In his discussion of the evolutionary implications of the composition biases of bacterial genomes, Samuel Kaplan pointed out that because the dinucleotide genome signatures of *Chlamydomonas* and *Sulfolobus* were very similar to that of mitochondria, it is likely that a fusion of a *Chlamydomonas* and *Sulfolobus*-like cell resulted in the formation of a eukaryotic cell with a large chromosome and a smaller mitochondrial genome. Nigel Saunders dicussed a potential adaptive measure that could allow successful movement of microorganisms to new hosts and/or microenvironments and enable genome evolution. He analyzed repetitive DNA to identify important genes in host–parasite interactions and identify unstable repeats that mediate phase variations. David Schwartz and colleagues presented advances in optical mapping, a useful tool to automatically construct restriction maps of the *E. coli* and *Deinoccus radiodurans* genomes. Their imaging system is now fully automated and their map construction algorithms now result in constructing maps from the visual data that in the future might be miniaturized to increase throughput and decrease cost.

Single cell and small multicellular eukaryotes also were represented at the meeting. In addition to the familiar *Saccharomyces* species, Daniel Lawson reported that sequencing by the Malaria Genome Sequencing Consortium of 8 of the 14 chromosomes of the malarial parasite *Plasmodium falciparum* is in progress, with Chromosome 3 nearing completion; Malcolm Gardner reported that Chromosome 2 is in the final stages of closure as well. Doris Kupfer presented a poster regarding the preparation of an EST database for the ascomycete *Aspergillus nidulans*. She also indicated that the fungal community plans to sequence the eight chromosomes of the 30-Mbp genome of *A. nidulans* once funding is obtained. In conjunction with the *Arabidopsis thaliana* genomic sequencing project, Waclaw Syzbalski described constructing a new modification to the bacterial artificial chromosome pBeloBAC11 cloning vector, allowing the conditional amplification of inserts immediately prior to cell harvesting, thereby reducing the levels of host genomic DNA contamination. The issue of modifying microbial genome to produce novel strains for large-scale production of small molecules was addressed by Dolf van Loon to exploit *Bacillus subtilis* as a producer of commercial quantities of riboflavin.

Finally, discussion of the role of microbial genome sequencing in relation to drug discovery hinged on the point that all known antibiotics are rapidly becoming ineffectual because most pathogenic microorganisms have devised biochemical means to evade the antimicrobial mechanisms, Richard Goold addressed the importance of comparative genomics for drug discovery as it aids in detailing changes in gene expression, in target selection for knockout mutations, in identification of drug targets, and in the drug profiling of known and new products. Martin Rosenberg and Gerald Vovis described the process of identifying new antimicrobials after the genomic sequence is analyzed. Here, rapid testing for gene essentiality and in vivo gene expression can identify novel broad spectrum targets, which then can be used in high-throughput screens. The value of the genomic sequence is that it confers the ability to look in the right place to unveil all potential drug targets, but as Molly Schmid related, the path from ''gene to screen'' in *Staphyloccuus aureus* is long and tedious. After identifying ~18% of the *S. aureus* genes as pharmacologically important, conditionally lethal (temperature-sensitive) mutants in 100 essential genes—26 with unknown function—were isolated and screened for susceptibility to inhibitors to discern the specific mechanism of action of these compounds. Roberta Hare also discussed the choice of antimicrobial targets in *S. aureus*, pointing out that genome sequencing and analysis can indicate the genes common to both humans and pathogens so they can be eliminated as antimicrobial targets.

With the excitement of this year's Small Genome Meeting pointing to the importance of microbial genomic sequencing as a tool for discovering new biochemical pathways, new industrial products, and new drugs, as well as for comparative genomics and evolutionary tree building, it is clear that we have only begun to uncover the information that resides in the beautifully simple but intriguingly complex world of microbial and small multicellular genomes.