Res. Microbiol. 151 (2000) 79–84

S092325080000125-X/REV

Microbial genome sequencing 2000: new insights into physiology, evolution and expression analysis

William Nierman, Jonathan A. Eisen, Claire M. Fraser*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Abstract — The complete genome sequence has been reported for 24 microbial organisms. The genome organization and gene content of these organisms has revealed an incredible diversity. Nearly half of the open reading frames identified by these sequencing projects are for potential genes with no known biological function. Efforts to make evolutionary sense and biological sense of the gene content of these organisms have been initiated. The greatest future challenge of genomics will be to determine function for the unknown genes.

sequence analysis / DNA / expression regulation / gene / phylogeny

1. Microbial genome organization and gene content

The diversity of life on earth is most spectacularly concentrated in the microbes. The niches which they occupy, the metabolic pathways they employ to derive energy from their environments, their tolerance of extremes of temperature, pressure, pH and salinity, and the oxygen and water levels to which they have adapted vastly exceeds anything found in the world of plants and animals. Yet we know of this diversity from observations of the less than 1% of Earth's 2–3 billion microbial species that we have been able to identify and study [1]. The microbes were the first living organisms to emerge on earth and after 3–4 billion years of evolution form the foundation of the modern biosphere from both an ecological and evolutionary perspective. Within the microbes are those species that have been historically and are yet today major human pathogens, inflicting sickness and death on too great a fraction of Earth's human population.

* Correspondence and reprints
Tel.: +1 301 838 3500; fax: +1 301 838 0209;
cmfraser@tigr.org

A revolution in the practice of microbiology was initiated when The Institute for Genomic Research (TIGR) published the first genome sequence for a free-living organism, *Haemophilus influenzae,* in 1995 [4]. Since that event 23 other microbial genome sequencing projects have been completed. In 1999, the genome sequence of six microbial species and two chromosomes from protozoan pathogens were reported. This progress has represented, on average, one completed genome sequence every 2 months and all indications point to this pace continuing to accelerate. Work is underway at TIGR and in other laboratories around the world on nearly 90 microbial genome sequencing projects from a diverse group of pathogens, archaea, and species of evolutionary importance (see http://www.tigr.org for a complete list). In the next 2–3 years, international efforts in microbial genome sequencing will generate more than 200 Mbp (million base pairs) of new DNA sequence containing ≈200 000 predicted genes, at least two to three times the number of genes expected from the completion of the human genome project.

Genome sequencing and analysis have revealed tremendous variability in microbial genome size and GC content, ranging from a

Table I. Summary of features from completed microbial genomes.

Organism	*Genome size (Mbp)*	*Number of ORFs*	*Unknown function*	*Unique ORFs*
A. fulgidus	2.18	2 437	1 315 (54%)	641 (26%)
M. thermotautotrophicum	1.75	1 855	1 010 (54%)	496 (27%)
M. jannaschii	1.66	1 749	1 076 (62%)	525 (30%)
P. horikoshii	1.74	2 061	859 (42%)	453 (22%)
A. aeolicus	1.50	1 521	663 (44%)	407 (27%)
B. subtilis	4.20	4 100	1 722 (42%)	1 053 (26%)
B. burgdorferi	1.44	1 751	1 132 (65%)	682 (39%)
C. trachomatis	1.04	894	290 (32%)	255 (29%)
D. radiodurans	3.28	3 192	1 715 (54%)	1 001 (31%)
E. coli	4.60	4 288	1 632 (38%)	1 114 (26%)
H. influenzae	1.83	1 692	592 (35%)	237 (14%)
H. pylori	1.66	1 657	744 (45%)	539 (33%)
M. tuberculosis	4.41	3 924	1 521 (39%)	606 (15%)
M. genitalium	0.58	470	173 (37%)	7 (2%)
M. pneumoniae	0.81	677	248 (37%)	67 (10%)
Synechocystis sp.	3.57	3 168	2 384 (75%)	1 426 (45%)
T. martima	1.86	1 877	863 (46%)	373 (26%)
T. pallidum	1.14	1 040	461 (44%)	280 (27%)
R. prowazkii	1.10	834	48 (12%)	207 (25%)
C. pneumonia	1.23	1 073	437 (40%)	186 (17%)
A. pernix	1.67	2 694	2 061 (76%)	1 538 (57%)
L. lactis	2.35	1 495	398 (27%)	83 (6%)
	45.6	44 449	20 726 (47%)	11 924 (27%)

low of 29% for *Borrelia burgdorferi* [5] to a high of 68% for *Deinococcus radiodurans* [11]. The more than two-fold difference in GC content is also reflected in differences in overall codon usage and amino acid composition among species. Genome organization is also variable with examples of single circular chromosomes, chromosomes plus one or a few plasmids or extrachromosomal elements, to the extreme seen with *B. burgdorferi*, a genome composed of a 910-kbp linear chromosome and 21 linear and circular extrachromosomal elements.

From a summary of results from the completed microbial genome sequences, representing more that 45 Mbp of DNA sequence and 44 000 predicted open reading frames (ORFs), it is immediately apparent that almost one-half of all ORFs identified to date are of unknown biological function (*table I*). Perhaps even more surprising is the fact that approximately one-quarter of the ORFs in each species studied to date, with the exception of *Rickettsia prowazekii*, *Mycoplasma genitalium*, and *Mycoplasma pneumonia*, are unique in having no significant sequence similarity to any other available protein sequence. *R. prowazekii* is interesting in that many of its genes are most similar to mitochondrial genes [2]. It is likely one of the closest living relatives to the ancestral predecessors of the organisms that through symbiotic colonization of ancestral eukaryotes evolved to the modern mitochondria. The mycoplasmas are obligate intercellular parasites with greatly reduced genomes. Taken together, these data indicate that there is a substantial amount of microbial biology yet to be understood. The idea of a 'model organism' in the microbial world may not be a valid concept given the vast differences that we have observed, even between related species.

Genome analysis is revealing other patterns with regard to proteins for which one can make putative assignments based on sequence similarity searching. Within certain categories of genes, such as those involved in transcription and translation, for example, the total number of genes present in each genome is quite similar, even when genome size differs by five-fold or

Table II. Summary of paralogous genes.

Organism	*Genome size (Mbp)*	*Number of ORFs*	*Paralogous ORFs*[a]
T. pallidum	1.14	1 040	129 (12%)
B. burgdorferi	1.44	1 751	707 (40%)
H. pylori	1.66	1 657	266 (16%)
A. fulgidus	2.18	2 437	719 (30%)
B. subtilis	4.20	4 100	1 947 (47%)
M. tuberculosis	4.41	3 924	2 000 (51%)
E. coli	4.60	4 288	2 272 (53%)

[a] ORFs that share at least 30% sequence identity over more than 60% of their lengths.

more. This observation suggests that a basic complement of proteins is absolutely required for these cellular processes. In contrast, the number of proteins in other functional categories, such as biosynthesis of amino acids, energy metabolism, transporters, and regulatory functions, for example, is more variable and tends to increase as genome size increases. Thus, as genome size increases so too does biochemical complexity for a given organism.

As microbial genomes become larger a significant proportion of their genes are observed to be members of families of paralogous genes. These are genes related by duplication rather than by vertical inheritance. As shown in *table II*, the number of genes that are contained in these families increases as genome size increases. The one exception to this rule is seen with *B. burgdorferi* [5], but this organism is unusual in that it contains a large number of plasmid-encoded lipoprotein paralogs. The largest classes of paralogs in essentially all genomes studied to date are the ATP-binding proteins associated with ABC transporters.

2. Microbial evolution

The availability of 24 completed microbial genome sequences has provided new insights on microbial evolution and diversity. The molecular picture of evolution for the past 20 years has been dominated by the small subunit ribosomal RNA phylogentic tree that proposes three monophyletic (or non-overlapping) groups of living organisms, the bacteria, the archaea, and the eukaryotes [13]. Although the archaea resemble bacteria in many aspects of their appearance (e.g. they do not have nuclei), the three-domain proposal suggests that they are a unique evolutionary lineage, on equal par to bacteria and eukaryotes. In fact, many lines of evidence suggest that the archaea and the eukaryotes shared a common ancestor exclusive of bacteria, or in other words, that the archaea and eukaryotes are more closely related to each other than either is to bacteria.

As a result of the completion of genome sequences from representatives of all three domains of life, it is now possible to examine evolutionary relationships among living organisms in a more comprehensive way. However, this task has turned out to be anything but straightforward. Incongruities can be seen everywhere in the phylogenetic tree from its root to the major branchings when single protein phylogenies are examined. It has become clear that gene evolution does not equal species evolution. This, in large part, is a result of extensive lateral gene transfer, not only between bacteria but also between bacteria and archaea [7]. Additional reasons cited to account for these observations include gene displacement, gene duplication followed by specialization or extinction, and convergence at the molecular level.

By comparing the protein sequences encoded in the four archaeal species whose genomes have been completely sequenced, Makarova et al. [6] have defined an evolutionary core of genes in archaeal genomes based on their presence in all four. These core genes (31–35% of the genome content) code primarily for proteins

involved in genome replication and expression. Specific metabolic functions are more sporadically present in the four genomes. An additional observation is that the core genes are more similar to eukaryotic counterparts while the genes present in only two or three of the species are most similar to bacterial homologs. The authors suggest that this may be due to lateral gene transfer of metabolic genes from bacteria in the evolution of the archaea. The corollary to this observation is that the core genome replication and expression genes have not been laterally transferred. Perhaps the complexity of the multi-protein structures required for replication and expression, replication complexes, transcription complexes, and ribosomes, make efficient lateral transfer and fixation of the transferred gene or genes statistically extremely improbable.

An alternative to single gene phylogenies is to build 'average' phylogenetic trees for the whole genome based on gene content. The first such 'gene content' phylogenetic analysis was reported by Snel et al. [9] who showed that a distance tree based on number of genes shared between genomes is remarkably similar to the rRNA tree for those same species. Subsequently, Fitz-Gibbon and House [3] showed similar results using a parsimony analysis of presence and absence of genes. In addition, Tekaia et al. [10] showed that a hierarchical classification method (which is a clustering-based method and thus not a true phylogenetic method) also gives similar results. These studies show that there is a basic 'average' phylogenetic history for each species that is recoverable even though individual genes may not follow this pattern.

3. Microbial genome expression analysis

Beyond trying to decipher molecular evolution, another formidable challenge in microbial genomics will be how to make use of the new sequence information on a large scale to better understand biology. By using approaches that include oligonucleotide chips, microarrays, and proteome analysis it should be possible to move from a static picture of a genome, as captured in a set of DNA and protein sequences, to an identification of gene networks and a better understanding of the dynamic nature of the regulation of gene expression in the microbial cell.

At TIGR and in many other laboratories, whole genome expression analysis based on the ORFs identified in genome sequencing projects is underway. PCR products prepared from each ORF are spotted in a high density array on a glass microscope slide. These microarrays are probed with fluorescently labeled whole organism cDNA prepared from mRNA. High resolution image scanners and analysis software quantify the signal intensity from each spot on the slide to determine the mRNA level from the cell of the ORF represented by the spot. This microarray methodology is particularly powerful in quantifying differential levels of expression of each ORF for cells grown under different conditions.

Whole genome expression analysis for *Escherichia coli* and *Mycobacterium tuberculosis* using whole genome microarrays were reported by Richmond et al. [8] and by Wilson et al. [12], respectively. The *E. coli* project measured changes in RNA levels before and after exposure to heat shock and following treatment with isopropyl-β-D-thiogalactopyranoside (IPTG). Treatment with IPTG resulted in induction of the *lacZYA* and *melAB* operons. Heat shock significantly altered the expression levels of 119 genes including 35 OFRs that were previously uncharacterized. Analysis of signal intensities suggested that at least 25% of the *E. coli* genes were expressed at detectable levels during growth in rich medium. This analysis was intended to be an initial validation of the whole genome *E. coli* microarray.

The *M. tuberculosis* analysis determined alterations in RNA levels after treatment with the drug isoniazid (INH). This drug was selected for this first study because it is given to more TB patients than any other, and because it is the drug to which resistance emerges most frequently. INH was found to induce several genes physiologically relevant to the drug's mode of action. INH selectively inhibits the synthesis of

mycolic acids, the major component of the waxy, outer lipid envelope of mycobacteria. The genes are induced within a fraction of a generation time of addition of the INH and are some of those directly involved in the processes inhibited by INH. Because the affected enzymatic pathway contains proven drug targets, perhaps other proteins operating in the same pathway, as revealed by the microarray data, might also be appropriate targets for new drug development. Patterns of induction and repression of gene expression may also prove valuable in designing screens for novel compounds that exert similar effects. These kinds of applications are two of the valuable potential uses of microarray technology in drug discovery and validation.

4. Conclusion and future considerations

The early events of the first phase of the development of the science of genomics can perhaps be listed as initiation of the Human Genome Project, high throughput human EST sequencing, and the completion of the total sequence of *Haemophilus influenzae*. This first phase can now be declared at an end with the routine accomplishment of whole genome sequencing. An avalanche of genome sequence data is being generated and a second phase development of the science of genomics is necessary for producing the tools for dealing with these data. It is not merely a matter of software tools for managing sequence data or software tools for finding genes in eukaryotic DNA sequence, but more importantly, tools for determining what all this sequence data means.

Approaches for making phylogenic sense of the gene content of organisms are starting to be explored as are the use of microarrays for expression analysis on a whole genome scale. A substantial fraction of the gene content of even the most extensively studied organisms is for genes with no known function. The greatest challenge remains in finding efficient ways to identify the function for these genes and in determining how all the genes work together to make an organism what it is. This is the great challenge that our success in genomics has brought to us. Our future accomplishments in this endeavor will likely reveal things about biology and biological systems that has to date been out of reach of even the human imagination.

References

[1] American Academy of Microbiology Colloquium, The Microbial World, American Society for Microbiology, 1997.

[2] Andersson S.G., Zomorodipour A., Andersson J.O., Sicheritz-Ponten T., Alsmark U.C., Podowski R.M., Naslund A.K., Eriksson A.S., Winkler H.H., Kurland C.G., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, Nature 396 (1998) 133–140.

[3] Fitz-Gibbon S.T., House C.H., Whole genome-based phylogenetic analysis of free-living microorganisms, Nucleic Acids Res. 27 (1999) 4218–4222.

[4] Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.-F., Dougherty B.A., Merrick J.M., McKenney K., Sutton G., Fitzhugh W., Fields C., Gocayne J.D., Scott J., Shirley R., Liu L.-I., Glodek A., Kelley J.M., Weidman J.F., Phillips C.A., Spriggs T., Hedblom E., Cotton M.D., Utterback T.R., Hanna M.C., Nguyen D.T., Saudek D.M., Brandon R.C., Fine L.D., Fritchman J.L., Fuhrmann J.L., Geoghagen N.S.M., Gnehm C.L., McDonald L.A., Small K.V., Fraser C.M., Smith H.O., Venter J.C., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd., Science 269 (1995) 496–512.

[5] Fraser C.M., Casjens S., Huang W.M., Sutton G.G., Clayton R., Lathigra R., White O., Ketchum K.A., Dodson R., Hickey E.K., Gwinn M., Dougherty B., Tomb J.F., Fleischmann R.D., Richardson D., Peterson J., Kerlavage A.R., Quackenbush J., Salzberg S., Hanson M., Van Vugt R., Palmer N., Adams M.D., Gocayne J., Weidman J., Utterback T., Watthey L., McDonald L., Artiach P., Bowman C., Garland S., Fujii C., Cotton M.D., Horst K., Roberts K., Hatch B., Smith H.O., Venter J.C., Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*, Nature 390 (1997) 580–586.

[6] Makarova K.S., Aravind L., Galperin M.Y., Grishin N.V., Tatusov R.L., Wolf Y.I., Koonin E.V., Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell, Genome Res. 9 (1999) 608–628.

[7] Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C., Fraser C.M., Genome sequencing of *Thermotoga maritima*: Evidence for lateral gene transfer between archaea and bacteria, Nature 399 (1999) 323–329.

[8] Richmond C.S., Glasner J.D., Mau R., Jin H., Blattner F.R., Genome-wide expression profiling in *Escherichia coli* K-12, Nucleic Acids Res. 27 (1999) 3821–3835.

[9] Snel B., Bork P., Huynen M.A., Genome phylogeny based on gene content, Nat. Genet. 21 (1999) 108–110.

[10] Tekaia F., Lazcano A., Dujon B., The genomic tree as revealed from whole proteome comparisons, Genome Res. 9 (1999) 550–557.

[11] White O., Eisen J.A., Heidelberg J.F., Hickey E.K., Peterson J.D., Dodson R.J., Haft D.H., Gwinn M.L., Nelson W.C., Richardson D.L., Moffat K.S., Qin H., Jiang L., Pamphile W., Crosby M., Shen M., Vamathevan J.J., Lam P., McDonald L., Utterback T., Zalewski C., Makarova K.S., Aravind L., Daly M.J., Minton K.W., Fleischmann R.D., Ketchum K.A., Nelson K.E., Salzberg S., Smith H.O., Venter J.C., Fraser C.M., Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1, Science 286 (1999) 1571–1577.

[12] Wilson M., Derisi J., Kristensen H.H., Imboden P., Rane S., Brown P.O., Schoolnik G.K., Exploring drug-induced alterations in gene expression in mycobacterium tuberculosis by microarray hybridization, Proc. Natl. Acad. Sci. USA 96 (1999) 12833–12838.

[13] Woese C.R., Fox G.E., Phylogenetic structure of the prokaryotic domain: the primary kingdoms, Proc. Natl. Acad. Sci. USA 74 (1977) 5088–5090.