

Complete genomes in WWW Entrez: data representation and analysis

Tatiana A. Tatusova, Ilene Karsch-Mizrachi and James A. Ostell

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received on November 16, 1998; accepted on December 10, 1998

Abstract

Motivation: The large amount of genome sequence data now publicly available can be accessed through the National Center for Biotechnology Information (NCBI) Entrez search and retrieval system, making it possible to explore data of a breadth and scope exceeding traditional flatfile views.

Results: Here we report recent improvements for completely sequenced genomes from viruses, bacteria, and yeast. Flexible web based views, precomputed relationships, and immediate access to analytical tools provide scientists with a portal into the new insights to be gained from completed genome sequences.

Availability: Entrez Genomes can be accessed on the World Wide Web at <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>

Contact: tatiana@ncbi.nlm.nih.gov

Introduction

In September 1995, the National Center for Biotechnology Information (NCBI) created the Genomes division of Entrez for handling data obtained from large-scale sequencing projects for genomes and chromosomes. This coincided with the release of the complete *Haemophilus influenzae* genome sequence (Fleischmann *et al.*, 1995), marking the start of a new era of megabase sequence generation and analysis. The first complete eukaryotic genome, *Saccharomyces cerevisiae*, was published in 1996 (Goffeau *et al.*, 1996). By the end of 1998, 16 complete microbial genomes will have been published (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995, 1997, 1998; Bult *et al.*, 1996; Himmelreich *et al.*, 1996; Kaneko *et al.*, 1996; Klenk *et al.*, 1996; Kunst *et al.*, 1997; Blattner *et al.*, 1997; Tomb *et al.*, 1997; Smith *et al.*, 1997; Deckert *et al.*, 1998; Cole *et al.*, 1998; Kawarabayasi *et al.*, 1998; Stephens *et al.*, 1998) and the complete genome of nematode may be published soon after.

The Entrez search and retrieval engine at NCBI provides access to nucleotide, protein and bibliographic information for the genome scale sequence and presents text and graphical display for a variety of completely sequenced genomes and chromosomes, contiged sequence maps, and integrated genetic and physical maps. The sequence information is

presented at several levels of detail including: (1) all completed genomes (Figure 1), (2) all chromosomes in a single organism (Figure 2), (3) single chromosomes (Figure 3), (4) detailed views of parts of chromosomes (Figures 4 and 5), and (5) single genes (Figure 6). At each level are (1) one or more presentation views, (2) a precomputed synthesis or summary, and (3) links and analyses appropriate to that level.

All completed genomes

Entrez Genomes displays data from small viral and organelle genomes, complete and near-complete genomes from bacteria and lower eukaryotes. It also presents maps and other views for higher eukaryotes (Kuzio, 1996), but that is not the focus of this article. Currently the Genome division of Entrez contains 781 entries divided into six large taxonomy groups: Archaea, Bacteria, Eukaryotae, Viroids, Viruses and Plasmids. Within each group there are subdivisions for genomes and plasmids. There is also an organelle subdivision for Eukaryotae, which contains complete organelle (mitochondrial, chloroplast) genome sequences. The currently available genomes are presented by Entrez Genomes in their taxonomic context (Figure 1).

A genome per organism

The complete genome view for a single organism includes multiple chromosomes and extrachromosomal elements if available. It also contains links to other organism specific sites. Figure 2 shows the overview of the complete genome of *Saccharomyces cerevisiae* with its 16 chromosomes. In the overview the user may search the entire genome for a gene of interest or select a specific chromosome on which to focus.

Chromosome view

The small genomes group contains chromosomes from viruses, viroids, organelles and plasmids. These sequences are deposited in a single GenBank (Benson *et al.*, 1998) record due to their small size. Because of the large degree of redundancy (multiple versions, population variants) in the GenBank database, we present a single complete chromosome as

- *Archaea*
 - *Euryarchaeota*
 - *Archaeoglobales*
 - *Archaeoglobus fulgidus*
 - *Methanobacteriales*
 - *Methanobacterium thermoautotrophicum*
 - *Methanococcales*
 - *Methanococcus jannaschii*
 - *Thermococcales*
 - *Pyrococcus horikoshii*
- *Bacteria*
 - *Aquificales*
 - *Aquifex aeolicus*
 - *Firmicutes*
 - *Bacillus/Clostridium group*
 - *Bacillaceae*
 - *Bacillus subtilis*
 - *Mycoplasmas and walled relatives*
 - *Mycoplasma*
 - *Mycoplasma genitalium*
 - *Mycoplasma pneumoniae*
 - *Actinobacteria*
 - *Mycobacterium tuberculosis*
 - *Spirochaetales*
 - *Spirochaetaceae*
 - *Borrelia*
 - *Borrelia burgdorferi*
 - *Treponema*
 - *Treponema pallidum subsp. pallidum*
 - *Planctomyces/Chlamydia/Verrucocomicrobium group*
 - *Chlamydia trachomatis*
 - *Proteobacteria*
 - *gamma* subdivision
 - *Enterobacteriaceae*
 - *Escherichia coli*
 - *Pasteurellaceae*
 - *Haemophilus influenzae Rd*
 - *delta/epsilon* subdivisions
 - *Helicobacter pylori*
 - *alpha* subdivision
 - *Rickettsia prowazekii*
 - *Cyanobacteria*
 - *Synechocystis PCC6803*
 - *Eukaryota*
 - *Fungi/Metazoa group*
 - *Saccharomyces cerevisiae*

Fig. 1. Taxonomic presentation of complete microbial genomes. 17 complete genome sequences deposited to EMBL/GenBank/DDBJ databases represent the species from the main taxonomic domains: Archaea and Bacteria. The phylogenetic relationships are derived from NCBI Taxonomy database (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>).

a reference sequence in these cases and then align other versions of the sequence to the reference sequence.

Complete microbial chromosomes and other large scale sequences such as *Leishmania major* (Myler, unpublished) and *Plasmodium falciparum* chromosome 2 (Gardner *et al.*, 1998) are presented as virtual, or segmented, records in the Genome division of Entrez. Rather than creating single large

entries, chromosome-size submissions are divided into several GenBank annotated entries, each no more than 350 kb long (a limit set by the International Nucleotide Database Collaboration). The individual segments can be assembled by retrieval software so that the users can view on demand the complete genome, chromosome, or other unit of interest. Entrez can present the chromosome and its components

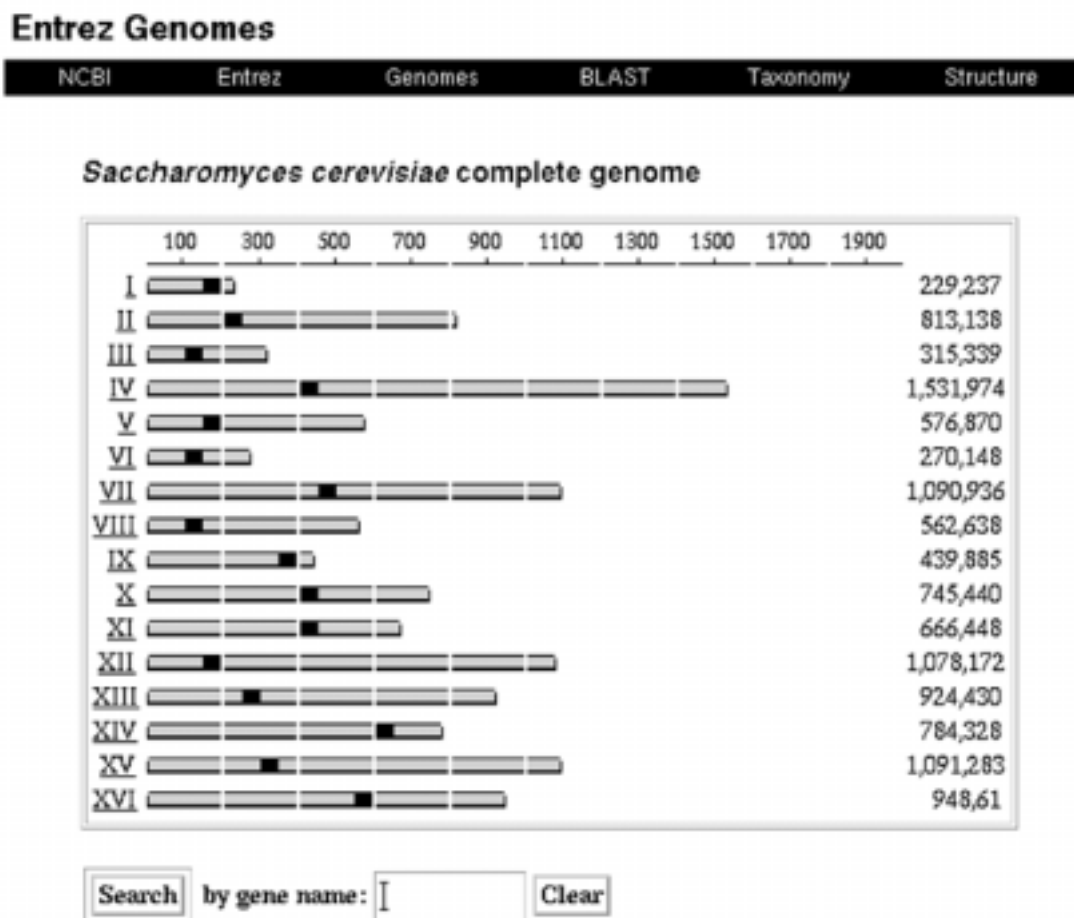


Fig. 2. *Saccharomyces cerevisiae* complete genome. Grey rectangles represent sixteen chromosomes with the length proportional to the size of the chromosome; Black dots depict the centromeres. Gene search over the whole genome is provided.

graphically or in a GenBank format which lists the information necessary to assemble the genome from the underlying records (Figure 3).

Specific regions or objects can be viewed by using the mouse to click on the site of interest. Regions may be selected by clicking the mouse on the picture to zoom in, or by entering one or two markers of interest in the Search query boxes and letting the display automatically locate the appropriate region of the chromosome.

Chromosome regions

A number of alternate views are provided for regions of chromosomes. Common elements are provided in each type of view, such as the ability to search, to select individual objects, to move left or right, zoom in or out, and to see the currently displayed region on an overview of the chromosome. The selected region stays the same when one switches between views. The views differ in emphasizing different aspects of the data or an underlying analysis of the data.

Figure 4 shows a region of the *Aquifex aeolicus* genome, which emphasizes protein coding genes. The genome is a segmented, virtual record, so the underlying GenBank records are retrieved and their coding region and gene annotations projected onto the chromosome region on-the-fly. Only coding regions are shown in this compact representation. The region is shown as a red wedge in the small overview in the lower right.

Figure 5 shows the same region in the TaxTable view, which emphasizes a precomputed analysis of the evolutionary context of the protein coding genes. The figure at the top provides a genome overview with the currently selected region displayed in a box. The colored dots represent protein coding genes and the color of the dots indicates their closest homolog by evolutionary group. The table at the bottom displays the details of this analysis for the genes in the selected region.

Chromosome regions can also be viewed as tables of proteins, tables of any marker, GenBank format, FASTA format, and alternate graphical views (not shown).

Entrez Genomes

BLAST Genomes Nucleotides Proteins Structure Taxonomy PubMed Help

Search by gene name:

Bacillus subtilis complete genome

CONTRIBUTING GENOME CENTERS:

BSNR

DATA and ANALYSIS:

PROTABLE

TAXTABLE

FTP

BLAST WITH MICROBIAL GENOMES

Accession: [AL009126](#)
Total Bases Sequenced: 4214814 bp
Completed: Nov 20, 1997.



Taxonomy Id: [1423](#)
Genetic code: [11](#)
Lineage: Eubacteria; Firmicutes; Low G+C gram-positive bacteria;
Bacillaceae; Bacillus.

Kunst, F., Ogasawara, N. et al.
The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*
Nature 390, 249–256 (1997)

[98044033](#)

CONTIG join([Z99104](#):1..213080,[Z99105](#):18431..221160,
[Z99106](#):13061..209100,[Z99107](#):11151..213190,[Z99108](#):11071..208430,
[Z99109](#):11751..210440,[Z99110](#):15551..216750,[Z99111](#):16351..208230,
[Z99112](#):4601..208780,[Z99113](#):26001..233780,[Z99114](#):14811..207730,
[Z99115](#):12361..213680,[Z99116](#):13961..218470,[Z99117](#):14281..213420,
[Z99118](#):17741..218410,[Z99119](#):15771..215640,[Z99120](#):16411..217420,
[Z99121](#):14871..209510,[Z99122](#):11971..212610,[Z99123](#):11301..212150,
[Z99124](#):11271..215534)

Fig. 3. *Bacillus subtilis* complete genome: single circular chromosome. The GenBank records are collectively combined into a virtual chromosome. Entrez provides a graphical presentation of the chromosome, with bands of alternating colors indicating the regions of consolidation. CONTIG report shows how the complete sequence can be assembled from individual segments.

Single gene views

Figure 6 shows a detailed view of a single protein coding region as analyzed for the TaxTable view (Figure 5). The single protein sequence is shown at the top. Links from it to other parts of Entrez appear as buttons on the top line. At the bottom is a graphical display showing the alignments of this protein with similar proteins in other genomes, revealing the experimental

details behind the summary in the TaxTable for this gene. Each of the proteins shown graphically in the alignment is accessible by clicking the hot links. In addition, any pairwise alignment can be examined at a residue by residue level in another window by clicking on this display.

Other single gene views include GenBank/GenPept format or FASTA format.

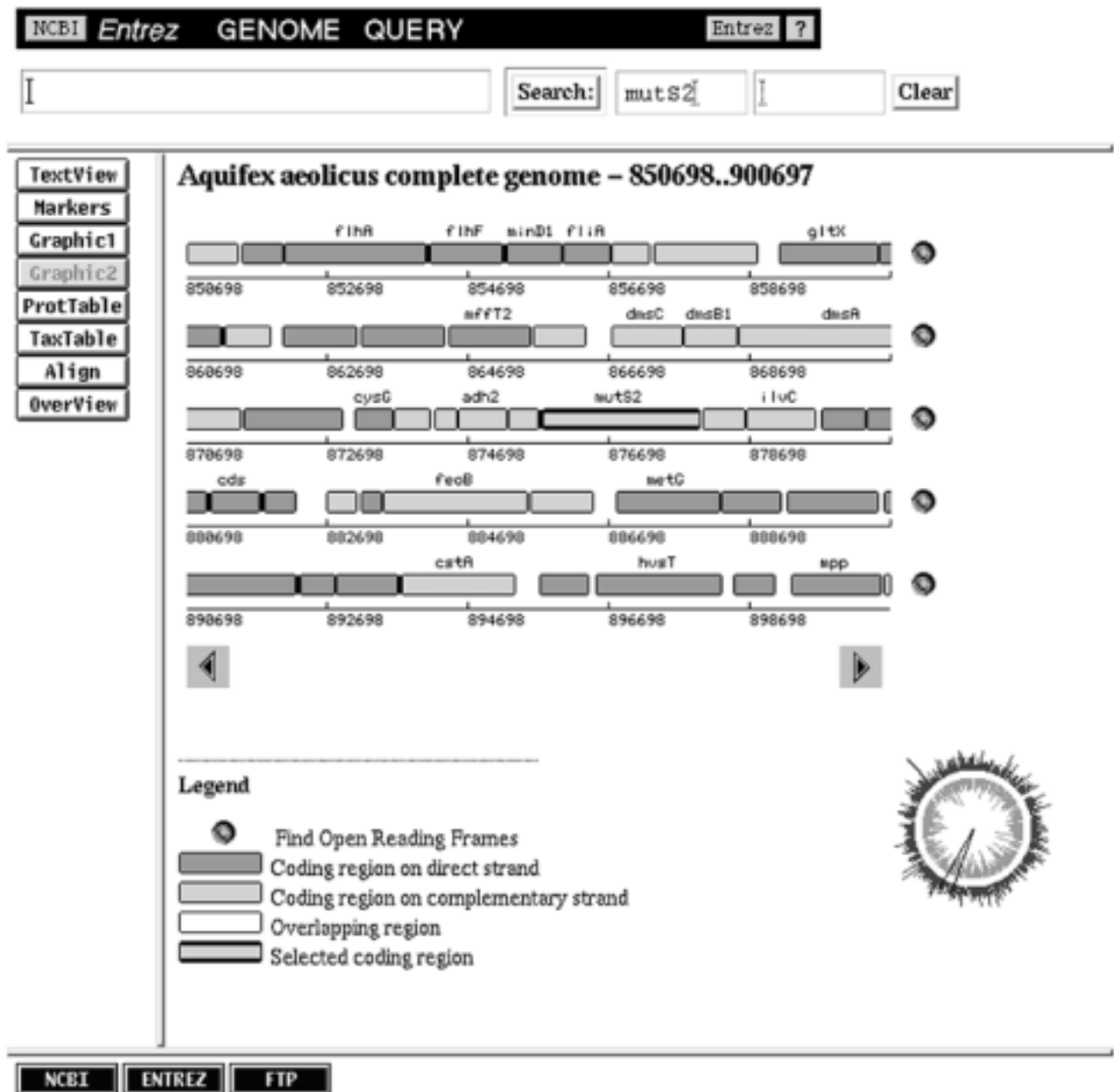


Fig. 4. Aquifex aeolicus protein view. The region is selected by gene query: mutS2 with query gene highlighted. Dark and light grey color represent coding regions on direct and complementary strands accordingly. The overview at the right bottom of the picture shows all the coding regions in the chromosome. The length of the line is proportional to protein length. Clicking on the compact overview allows you to select another distant region without returning to the main overview page.

Analyses

A number of tools and analyses have been developed which underlie and are connected to the genome views above. Some are accessible directly for use in other projects as well.

BLAST with microbial genomes

The sequences of partially completed genomes are available on the web for searching by similarity to a query sequence. Sequences from the 16 complete microbial genomes and 20 unfinished genomes can be searched using the Blast 2.0 algo-



Fig. 5. *Aquifex aeolicus* taxtable view. The genome map shows a graphical representation of all of the protein coding genes. The 50 kb region centered by gene *mutS2* (see Figure 4) is highlighted. The dots depicting gene products are color-coded based on the taxonomic classification of the protein to which they show greatest degree of similarity (see legend to Figure 6). The Cut-off bit score for reporting high-scoring segment pairs can be set to be more or less stringent: Cut-off+ bit score specifies the margin in choosing the best 'hit'. The genes in the selected region are listed in the table. The second column contains the taxonomic classification of the organism from which the protein with the strongest similarity was derived. A black dot indicates that there were no similarity scores greater than the cutoff value. The GenBank IDs (with a link to the protein record in Entrez) and the bit scores are given for the best match in each of the three taxonomic groups.

rithm (Altschul *et al.*, 1997). The user can compare known protein sequences to the nucleotide sequences from completed and unfinished microbial genomes using *tblastn* or nucleotide sequences to the nucleotide sequences using *blastn*. The user also has the ability to choose to search the sequences from individual organisms or all Archaea or all Bacteria or any combination of sequence databases that the user chooses. The Blast with Microbial Genomes page can be found at <http://www.ncbi.nlm.nih.gov/unfinishedgenome.html>.

Neighbors and variants

Small virus complete genomes in the Genomes division are presented with a reference sequence. Variants of the reference sequence are shown in the alignment view. For the retroviruses, the alignment of mutation and population variants are precomputed. In other cases, the alignments are constructed 'on the fly' using the information from Highest-scoring Segment Pair (HSP) database. The HSP database is

developed and maintained at NCBI. It contains all similar segments, also called neighbors, for every sequence in GenBank database produced by BLAST similarity search. Neighbors of the viral genome sequences are filtered by the organism and the length of the reference sequence (to eliminate partial sequence entries). The reference sequence is aligned with the filtered neighbors and the resulting alignment is presented graphically and as text.

Taxonomic classification of protein neighbors

The analysis at the single gene level includes the comparison of amino-acid sequences of proteins encoded by complete genomes to proteins in current databases. All proteins of the genome are compared with the non-redundant (nr) database using BLAST 2.0 (Altschul *et al.*, 1997) and the hits were classified according to taxonomic origin.

The results of the protein sequence comparison for each complete genome is summarized in a Taxtable (Figure 5). For each protein in the selected genome, the table shows the



Fig. 6. Neighbors of putative protein (PID:2983686) coded by aq_1244 gene in *Aquifex aeolicus*. The protein homologs are classified into three major domains, Eukaryota, Eubacteria and Archaea (see Figure 5). Best match (for cut-off bit score 47.0) in each domain is shown in a table. Each bar in the graphical alignment is color coded to indicate the taxonomic classification of the homolog: red color depicts Eukarya; blue — Eubacteria; yellow — Archaea. Grey color indicates that the homolog doesn't belong to any of three main classes. First column in the table shows alignment bit score, second column contains Genbank ID, and the third one indicates the genome if the homolog belongs to one of the Complete genomes. The detailed pairwise alignment can be seen by clicking on the bit score, in that case a small red circle appears next to the selected bit score.

most similar protein in Archae, Bacteria, Eukaryotae. For each organism, the 'hits' from closely related species are excluded from consideration. For example, for *E. coli*, only bacteria outside the Proteobacteria domain were included. The genome map shows a graphical representation of all of the protein coding genes in the genome. The dots depicting gene products are color-coded based on the taxonomic classification of the protein to which they show greatest (even by a very small margin) degree of similarity. The large-scale integration of the protein alignments provides an insight into genome evolution and organization. This approach makes it possible to perform comprehensive comparisons between major phylogenetic groups and produce an informative outline of these relationships.

Looking for open reading frames

The ORF Finder (Open Reading Frame Finder) (Tatusova, unpublished) is a graphical analysis tool, which finds all

open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool has been integrated into some of the views (Figure 4). For each 10 kb region of bacterial genomes the user can perform an open reading frame search with OrfFinder. With the average size of a gene in bacterial genomes at about 1000 nucleotides, 10 kb seems to be the most reasonable genome region size for analysis. The deduced amino acid sequence can then also be searched against the sequence database using the WWW BLAST server. The OrfFinder analysis sometimes helps to reveal the errors in GenBank annotation and allows the researcher to find and to analyze small unannotated ORFs in intergenic regions of bacterial genomes.

Summary

This integrated approach to displaying, exploring, and analyzing genome scale sequence data gives scientists an essential platform for understanding the information and making

discoveries. It provides a means of moving from genome to gene, across organisms from viruses to eukaryotes, in a seamless way. Yet it also provides the flexibility necessary to take advantage of perspectives and analyses appropriate to only certain organisms or certain levels of organization. With the huge amount of data coming from the large-scale sequencing projects, presenting these enormous amounts of data in a coherent, concise way becomes more and more important.

Acknowledgments

The World Wide Web Entrez Genomes project represents the efforts of many NCBI staff members along with the collective contributions of many dedicated scientists world wide.

References

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Benson,D.A. *et al.* (1998) Genbank. *Nucleic Acids Res.*, **26**, 1–7.
- Blattner,F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
- Bult,C.J. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.
- Cole,S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
- Deckert,G., *et al.* (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, **392**, 353–358.
- Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Fraser,C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
- Fraser,C.M. *et al.* (1997) Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.
- Fraser,C.M. *et al.* (1998) Complete genome sequence of *Treponema pallidum*, the Syphilis Spirochete. *Science*, **281**, 375–388.
- Gardner,M.J. *et al.* (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, (in press).
- Goffeau,A. *et al.* (1996) Life of 6000 genes. *Science*, **274**, 546–567.
- Himmelreich,R., Hilbert,H., Plagens,H., Pirkel,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–4449.
- Kaneko,T. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**, 109–136.
- Kawarabayasi,Y. *et al.* (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3. *DNA Res.*, **5**, 55–76.
- Klenk,H.P. *et al.* (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.
- Kunst,F. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
- Kuzio,J. (1996) Genomic sequence presentation. *Trends Genet.*, **12**, 321–322.
- Smith,D.R. *et al.* (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *J. Bacteriol.*, **179**, 7135–7155.
- Stephens,R.S. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans. *Science*, **282**, 754–759.
- Tomb,J.-F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.