

# Databases, Sequence

## D Ussery

Copyright © 2001 Academic Press  
doi: 10.1006/rwgn.2001.0313

## Ussery, D

Institute of Biotechnology, Technical University of  
Denmark, Lyngby, DK-2800, Denmark

The complete genome of the bacterium *Haemophilus influenzae* was published in 1995. For the first time, it became possible to look at the complete DNA sequence of the whole circular chromosome of a bacteria. Since then, many more bacterial (and archaeal and eukaryotic) genomes have been sequenced and deposited into GenBank. At the time of writing this article (September 2000) there are currently about 86 prokaryotic genomes that have been sequenced, of which 52 (9 archaeal and 43 bacterial genomes) are publicly available. The number of sequenced genomes will continue to grow quickly, as it is now possible to sequence a bacterial genome in a single day. This is a mixed blessing for researchers in that it often feels as if there is too much information.

The purpose of this article is to provide an overview of the genome databases currently available. Due to the transient nature of the lists, all of the databases are websites, which can be updated easily and regularly, as more genomes are sequenced.

Genome databases can be divided into four broad categories:

### “Archival” Databases Which Contain Sequences of Published Genomes

There are several databases which contain sequences of published genomes, in various formats. Perhaps the most common format for many molecular biologists is GenBank, although many people also use the European Molecular Biology Laboratories (EMBL) or DDBJ (DNA Data Bank of Japan) format. GenBank, EMBL, and DDBJ all contain the same data, in slightly different formats. In all of the databases in this group, it is possible to download the complete genomic sequence, either with or without annotation of the coding regions.

The NCBI web page is updated regularly, and provides a good overview of the sequences available, with lists sorted either alphabetically or by taxonomic group. In addition, large plasmids which are part of the genome are usually included in the entries. The GenBank site is simply an ftp site, with little informa-

tion about the individual genomes, although it is good for downloading the genome sequences. The EMBL page allows one to download genome sequences in a variety of formats, including a “segment” format, where it is possible to obtain a sequence of only a small region of the chromosome. The DDBJ site uses a JAVA applet to allow the user to access a graphical view a particular region of the chromosome.

### Databases at Major Sequencing Centers, Which Contain Access to “ongoing” Genome Projects

How does one find information about which of the genomes that have been sequenced are similar to a particular organism being studied? There are a couple of good places to start. For published sequences, the NCBI page and TIGR website, mentioned at the top of the list in **Table 1**, are very good resources. However, there are many additional genomes that have been sequenced and are publicly available, even though they have not yet been published. This information can be spread amongst several different databases, and the best current method seems to be checking a number of websites on a regular basis to keep updated. The Sanger Centre regularly updates its web pages with progress on sequenced genomes, and all access to the “raw data,” before it has been fully assembled. Most of the bacterial genomes have been sequenced either by the Sanger Centre or TIGR. The TIGR website is also updated regularly, and preliminary sequencing data can be downloaded with permission from TIGR. Preliminary data, including sequenced but unpublished genomes, are also available from the University of Oklahoma and Washington University in St Louis. There are other sequencing centers; this list is meant to be an overview, and is not exhaustive.

### Databases Which Contain a Centralized Set of Links to Sequenced Genomes

There are many websites which contain lists of sequenced projects, with links about the various genomes, such as which lab the genome was sequenced in, who funded it, and taxonomic classification of the organism. The NCBI list is well maintained and current. The INFOBIOGEN website is a good place to check the status of sequencing projects; this site also has links to FASTA files of the sequences from the various genomes. The GOLD website contains listing of all sequenced genomes, including those done by industry which are likely not to be part of the public domain for several years. The Enhanced Microbial Genomes Library not only contains lists of

genomes, but provides “improved and corrected annotations.” Finally, the last two websites in this section are lists of microbial genomes funded by the National Institute of Allergy and Infectious Diseases (NIAID) and the Department of Energy (DOE), both in the USA. We also maintain a list of completed genomes that is updated on a regular basis (<http://www.cbs.dtu.dk/services/GenomeAtlas/>).

### Bioinformatic Databases, Which Analyze Various Forms of Data from Genome Projects

We maintain the DNA Structural Atlas of Genomes web page, which is updated on a regular basis (<http://www.cbs.dtu.dk/services/GenomeAtlas/>); we use a graphical representation of the whole genome on a single page to summarize structural properties. An example of this is shown in **Figure 1**, which is a DNA Genome Atlas for chromosome 3 of *Plasmodium falciparum*. Note that the telomere regions contain a curved band (deep blue in band A in the figure), and are generally more thermostable, i.e., will melt at a higher temperature (green in band B) and are more rigid (dark green in band C). Also they contain a direct repeat (blue in band E) and a different, larger region of inverted repeats (red in band F). Although both telomeres are GC rich (lighter red in band H), one end contains primarily Gs (turquoise at the right hand end of band G) whilst the other is enriched in Cs (purple in the left hand side of band G). The atlases are a method of obtaining an overview of an entire genome.

There are many other web sites devoted to bioinformatics of whole genomes. One of the most comprehensive projects for analysis of complete genomes is the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which has entries on metabolic

pathways, regulatory pathways, and gene expression in whole genomes. The BioMolecular Engineering Research Center (BMERC) contains tools for comparison of different genomes, as well as the next two web sites in the table. The final link (“What Is There?”) attempts to produce metabolic reconstructions for sequenced (or partially sequenced) genomes.

### Summary

There are hundreds of genome databases available; key web pages are shown in **Table 1**. Many of these will allow blast searches to be done, both against the published genomes as well as the “current ongoing” genome projects. DNA Structural Atlases are a way of viewing whole genomes, in terms of DNA structures, and are useful for finding regions of unusual DNA structures. The number of sequenced genomes will soon reach more than a hundred. Genome databases are necessary to track and better utilize this information.

### Further Reading

- Baxevanis AD (2000) The Molecular Biology Database Collection: an online compilation of relevant database resources. *Nucleic Acids Research* 28: 1–7. (Note. *Nucleic Acids Research* traditionally devotes the first issue in January to sequence databases.)
- Nelson KE, Paulsen IT, Heidelberg JF and Fraser CM (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nature Biotechnology* 18: 1049–1054.
- Pedersen AG, Jensen LJ, Stærfeldt HH, Brunak S and Ussery DW (2000) A DNA structural atlas for *Escherichia coli*. *Journal of Molecular Biology* 299: 907–930.

**See also: 0556, 0557, 1475**

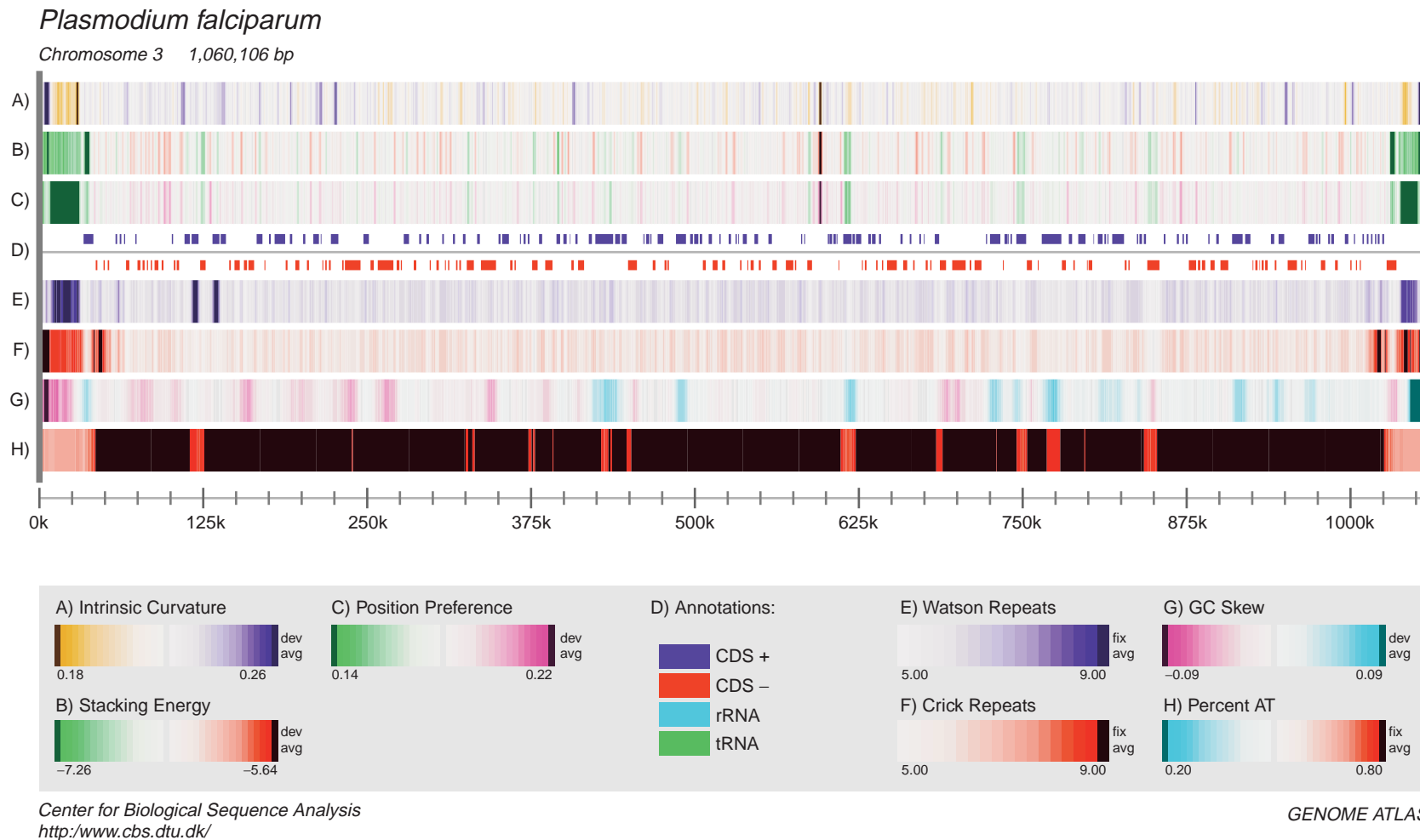
**Table I** A list of genome databases<sup>a</sup>

Type	Name of database
1. Lists of published genomes	<p>NCBI list of sequenced genomes  <a href="http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.html">http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.html</a>            GenBank  <a href="ftp://ncbi.nlm.nih.gov/genbank/genomes/">ftp://ncbi.nlm.nih.gov/genbank/genomes/</a>            EMBL  <a href="http://www.ebi.ac.uk/genomes/">http://www.ebi.ac.uk/genomes/</a>            DDBJ (DNA Data Bank of Japan)  <a href="http://gjb.genes.nig.ac.jp/">http://gjb.genes.nig.ac.jp/</a></p>
2. Links to genome sequencing centres	<p>Sanger Center  <a href="http://www.sanger.ac.uk/Projects/">http://www.sanger.ac.uk/Projects/</a>            TIGR  <a href="http://www.tigr.org/tdb/mdb/mdbcomplete.html">http://www.tigr.org/tdb/mdb/mdbcomplete.html</a>            TIGR Latest Update for Unfinished Microbial Genome Data  <a href="http://www.tigr.org/cgi-bin/BlastSearch/ReleaseDate.cgi">http://www.tigr.org/cgi-bin/BlastSearch/ReleaseDate.cgi</a>            TIGR's "ongoing projects"  <a href="http://www.tigr.org/tdb/mdb/mdbinprogress.html">http://www.tigr.org/tdb/mdb/mdbinprogress.html</a>            University of Oklahoma's Advanced Center for Genome Technology  <a href="http://www.genome.ou.edu/">http://www.genome.ou.edu/</a>            Washington University in St Louis Genome Sequencing Center  <a href="http://genome.wustl.edu/gsc/C_elegans/navcelegans.pl">http://genome.wustl.edu/gsc/C_elegans/navcelegans.pl</a></p>
3. Lists of links about sequenced genomes	<p>NCBI list of bacterial genomes that are complete but not published  <a href="http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html">http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html</a>            NCBI list of completed and ongoing projects  <a href="http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html">http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html</a>            Blast NCBI genomes  <a href="http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html#GENOMES">http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html#GENOMES</a>            Complete genomes in KEGG  <a href="http://www.genome.ad.jp/kegg/catalog/org_list.html">http://www.genome.ad.jp/kegg/catalog/org_list.html</a>            GOLD – Genomes OnLine Database  <a href="http://216.190.101.28/GOLD/completegenomes.html">http://216.190.101.28/GOLD/completegenomes.html</a>            GOLD – "Ongoing" Genomes OnLine Database  <a href="http://216.190.101.28/GOLD/prokaryagenomes.html">http://216.190.101.28/GOLD/prokaryagenomes.html</a>            Infobiogen list of complete genomes  <a href="http://www.infobiogen.fr/doc/data/complete_genome.html">http://www.infobiogen.fr/doc/data/complete_genome.html</a>            Infobiogen list of incomplete genomes  <a href="http://www.infobiogen.fr/doc/data/uncomplete_genome.html">http://www.infobiogen.fr/doc/data/uncomplete_genome.html</a>            The Enhanced Microbial Genomes Library  <a href="http://pbil.univ-lyon1.fr/emglib/emglib.html">http://pbil.univ-lyon1.fr/emglib/emglib.html</a>            NIH (National Institute of Allergy and Infectious Diseases) supported projects  <a href="http://www.niaid.nih.gov/dmid/genomes/genome.htm">http://www.niaid.nih.gov/dmid/genomes/genome.htm</a>            Department of Energy (DOE) funded Microbial Genomes, completed and ongoing projects  <a href="http://www.er.doe.gov/production/ober/EPR/mig_cont.html">http://www.er.doe.gov/production/ober/EPR/mig_cont.html</a></p>
4. Lists of genome analysis web pages	<p>KEGG: Kyoto Encyclopedia of Genes and Genomes  <a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>            BMERC – Completed genomes search and analysis  <a href="http://bmerc-www.bu.edu/bioinformatics/bioinformatics.html">http://bmerc-www.bu.edu/bioinformatics/bioinformatics.html</a>            Comparative sequence analysis of whole genomes  <a href="http://www.bork.embl-heidelberg.de/Genome/">http://www.bork.embl-heidelberg.de/Genome/</a></p>

“What Is There” – Interactive Metabolic Reconstruction on the Web  
<http://129.15.12.51:8080/WIT2/CGI/index.cgi?user=>  
NCBI’s Complete Genomes Page  
[http://www.ncbi.nlm.nih.gov/Complete\\_Genomes/](http://www.ncbi.nlm.nih.gov/Complete_Genomes/)  
CBS DNA Structural Atlases for Complete Genomes  
<http://www.cbs.dtu.dk/services/GenomeAtlas/>

---

note: An updated version of this list can be found at the following URL:  
<http://www.cbs.dtu.dk/services/GenomeAtlas/TableI.html>



**Figure 1** DNA “Genome Atlas” for *Plasmodium falciparum*. The different colored lines are as described in the text and at our website (<http://www.cbs.dtu.dk/services/GenomeAtlas/>). (Note: this figure can also be seen at the following URL: [http://www.cbs.dtu.dk/services/GenomeAtlas/Pfalciparum/pfal\\_3.genomeatlas.lin.html](http://www.cbs.dtu.dk/services/GenomeAtlas/Pfalciparum/pfal_3.genomeatlas.lin.html).)