C3 GENOME AND ORGANISM-SPECIFIC DATABASES

Key Notes

Organism-specific resources

As well as general databases that serve the entire biology community, there are many organism-specific databases that provide information and resources for those researches working on particular species. The number of organism-specific databases is growing as more genome projects are initiated, and many can be accessed from general genomics gateway sites such as GOLD.

Database formats

There is no universally agreed format for genome databases and several viewers and browsers have been developed with graphical displays for genomic sequence analysis and annotation. One of the most versatile formats is ACeDB (originally designed for the nematode *Caenorhabditis elegans*), which has an object-orientated database architecture and is now used in many applications outside the field of genomic bioinformatics.

Finding organism-specific databases

Organism-specific data are widely distributed on the Internet. In order to find and interrogate databases on specific organisms, it is necessary to use a gateway site to access relevant databases and information resources. Worked examples are provided, using GOLD as the gateway and illustrated with Ebola virus, the bacterium *Escherichia coli*, the fruit fly *Drosophila melanogaster* and the human genome.

Related topics

File formats (C1)
Annotated sequence databases (C2)
Miscellaneous databases (C4)
Database management (O4)

Organism-specific resources

The annotated sequence databases discussed in Topic C2 are general to all organisms, and contain data relevant to viruses, bacteria, microbial eukaryotes, animals and plants, as well as recombinant molecules produced in the laboratory. However, there are also many databases devoted to particular organisms, and their numbers are increasing as further genome projects are initiated. Typically, such databases contain not only sequence data but also information on gene expression, mutant phenotypes, genome maps, genome sequencing projects and relevant scientific literature, and provide links to resources for obtaining clones, mutants and for contacting researchers. A selected list of organism-specific databases is provided in *Table 1*, but this represents only a small fraction of the resources available. The user interested in an organism that is not listed in *Table 1* could try using a search engine (Topic A3) to find a useful resource, but there are also a number of excellent gateways available on the WWW, which provide information on multiple organism-specific resources and links to the relevant sites. A number of these gateways are listed in *Table 2*.

Table 1. A small selection of organism-specific genomic databases available on the WWW.

Organism	Database/resource	URL
Escherichia coli	EcoGene	http://bmb.med.miami.edu/EcoGene/EcoWeb/
	EcoCyc (Encyclopedia of *E. coli* genes and metabolism)	http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html
	Colibri	http://genolist.pasteur.fr/Colibri/
Bacillus subtilis	SubtiList	http://genolist.pasteur.fr/SubtiList/
Saccharomyces cerevisiae	*Saccharomyces* Genome Database (SGD)	http://genome-www.stanford.edu/Saccharomyces/
Plasmodium falciparum	PlasmoDB	http://PlasmoDB.org
Arabidopsis thaliana	MIPS *Arabidopsis* thaliana Database (MAtDB)	http://mips.gsf.de/proj/thal/db
	The *Arabidopsis* information resource (TAIR)	http://www.arabidopsis.org/
Drosophila melanogaster	FlyBase	http://flybase.bio.indiana.edu/
Caenorhabditis elegans	A *C. elegans* DataBase (ACeDB)	http://www.acedb.org/
Mouse	Mouse Genome Database (MGD)	http://www.informatics.jax.org/
Human	OnLine Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/omim

These databases are actively curated by members of the research community working on the particular organism of interest and generally include links to organism-specific resources such as clone sets and mutant strains.

Table 2. Useful gateway sites providing information and links to multiple, organism-specific and genomic resources.

Gateway site	URL
NCBI Genomic Biology	http://www.ncbi.nlm.nih.gov/Genomes/index.html
GOLD (Genomes OnLine Database)	http://wit.integratedgenomics.com/GOLD/
Organism-specific genome databases	http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Bacterial genomes	http://genolist.pasteur.fr/
Yeast databases	http://genome-www.stanford.edu/Saccharomyces/yeast_info.html
EnsEMBL genome database project	http://www.ensembl.org/
MIPS (Munich Information Center for Protein Sequences)	http://mips.gsf.de

Database formats

Genomic databases need to facilitate the storage and analysis of large amounts of data, but must also have a user-friendly front-end graphical display to allow relevant data to be displayed and analyzed. A number of viewers and browsers have been developed for genomic sequence analysis and annotation (*Table 3*). These include Artemis, Apollo, EnsEMBL and GoldenPath. One of the most versatile database formats is **ACeDB**. This was originally designed for research on the nematode worm *Caenorhabditis elegans* (**A *C. elegans* DataBase**). ACeDB has an object orientated database architecture (Topic O4) rather than a simple collection of data and it has been used to handle data on other organisms, including the yeast *Scizosaccharomyces pombe* and humans. ACeDB has a graph-

Table 3. Database tools for displaying and annotating genomic sequence data.

Viewer format	URL for further information and tutorials
Artemis	http://www.sanger.ac.uk/Software/Artemis
ACeDB	http://www.acedb.org/Tutorial/brief-tutorial.shtml
Apollo	http://www.ensembl.org/apollo/
EnsEMBL	http://www.ensembl.org
NCBI map viewer	http://www.ncbi.nlm.nih.gov/
GoldenPath	http://genome.ucsc.edu/

ical user interface with displays and tools designed for genomic data. Other features of ACeDB include **AQL (ACeDB query language)**, interfaces with Perl and Java, WWW interfaces (of which AceBrowser is the current and supported version), WinAce (the Windows95/NT version of ACeDB), CITA (a CORBA interface to the database) and Acembly (a sequence assembly system). The URL is http://www.acedb.org/.

Finding organism-specific databases

This section provides a number of worked examples of how to find organism-specific databases, resources and information. A good starting point for this type of search is the Genomes OnLine Database (GOLD). Once the GOLD top page has been accessed (http://wit.integratedgenomics.com/GOLD/), the site can be searched for information on any organism, for example Ebola virus, the bacterium *Escherichia coli*, the fruit fly *Drosophila melanogaster*, and humans.

Ebola virus

Information on Ebola virus can be found by clicking on the *Viruses* hyperlink under 'Other Links' on the GOLD top page. This accesses the European Bioinformatics Institute page of completed viral genomes, which currently lists nearly 700 viruses whose genomes have been sequenced, together with the corresponding European Molecular Biology Laboratory Nucleotide Sequence Database file (Topic C2), the sequence in FASTA format (Topic C1) and the sequence retrieval system entry (Topic D2). The URL is http://www.ebi.ac.uk/cgi-bin/genomes/genomes.cgi?genomes=viruses.

Escherichia coli

Entering *Escherichia coli* as a search term after accessing the SEARCH GOLD query form pulls records on nine different bacterial strains. Resources are listed in a table with the following headings: *Organism, Tree, Information, Size/ORF-number, Data Search, Institution, Funding, Genome Database, Status* and *Publication*. Of these, the links provided under the headings *Information* and *Genome Database* are the most useful. There are more than 20 resources listed including some general ones [e.g. National Center for Biotechnology Information (NCBI), SWISS-PROT, TIGR (The Institute for Genomic Research) and some specific ones (e.g. EcoCyc, EcoGene, Colibri, which are also shown in *Table 1*). The *E. coli* resources can be accessed directly via hyperlinks, and many of the sites have lists of further resources. For example, the EcoCyc page, the Encyclopedia of *E. coli* genes and metabolism, contains a link to 'other information on *E. coli*' which is a page containing 13 further links to *E. coli* resources on the WWW. Similarly, the Colibri site has a link 'other sites related to *E. coli*' which lists over 20 additional resources.

Drosophila melanogaster

When *Drosophila melanogaster* is used as a search term, the *Genome Database* links provided by GOLD are as follows: NCBI, BDGP (Berkeley *Drosophila* Genome Project), FlyBase-UK, FlyBase-USA, EBI-Proteome, KEGG and IBM-Annotation. Under *Information*, a taxonomy resource and another comprehensive *Drosophila* site, the Interactive Fly, are also listed. FlyBase-UK links to http://www.edgp.ebi.ac.uk/ and provides a large number of *Drosophila* resources, including clone orders, annotated sequences, raw sequence data in FASTA format, sequence sets and access to sequence analysis tools.

Human

When *Homo sapiens* is used as a search term, links provided by GOLD are as follows: NCBI, ORNL, RIKEN, EnsEMBL, Proteome-EBI, Sanger Centre, HOWDY and IBM-Annotation. Clicking on NCBI links to the NCBI human genome resources page, which contains images of human chromosomes. The user can select from a drop-down menu of clones, genes, physical maps, genetic maps and variation and then click on any of the chromosomes to see the information available for the chromosome chosen under that heading. The NCBI server also provides access to UniGene, Online Mendelian Inheritance in Man (OMIM) and other miscellaneous databases (Topic C4).