D332–D334 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj145

The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide

Konstantinos Liolios[1,2], Nektarios Tavernarakis[3], Philip Hugenholtz[4] and Nikos C. Kyrpides[5,*]

[1]Department of Pathology and [2]Department of Microbiology-Immunology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA, [3]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Heraklion, Crete, Greece, [4]Microbial Ecology Program and [5]Microbial Genome Analysis Program, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA

Received September 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

The Genomes On Line Database (GOLD) is a web resource for comprehensive access to information regarding complete and ongoing genome sequencing projects worldwide. The database currently incorporates information on over 1500 sequencing projects, of which 294 have been completed and the data deposited in the public databases. GOLD v.2 has been expanded to provide information related to organism properties such as phenotype, ecotype and disease. Furthermore, project relevance and availability information is now included. GOLD is available at http://www.genomesonline.org. It is also mirrored at the Institute of Molecular Biology and Biotechnology, Crete, Greece at http://gold.imbb.forth.gr/

HISTORY AND GROWTH

GOLD (1,2) was created in 1997 with the aim to (i) monitor all genome sequencing projects from instigation to completion and (ii) provide the community with a centralized database integrating diverse information related to those projects in the form of hyper-text links to disparate web-based resources.

Few would have predicted that less than a decade later the rate of genome sequencing would surpass even Moore's law for the increase of microprocessor computational power. We anticipate that, as new sequencing technologies are introduced, such as pyrosequencing (3), and the cost of existing technologies continues to decline, the number of genome sequences will continue to grow exponentially and the number of research groups able to contribute genome sequences also will dramatically increase. Therefore, the need for a searchable database that comprehensively tracks genome projects to help guide selection of new projects and provide up-to-date overview statistics will only increase.

CURRENT STATUS OF THE DATABASE

Published complete genomes

From 350 projects at the time of its previous report (2), GOLD has grown in providing information for 1575 genome projects worldwide, today. Almost 300 of those projects are currently being reported as completed with their sequences submitted to public databases. These are reported in GOLD as Published Complete Genomes. A genome publication is not always available in the literature for these projects as quite often submitters choose to release their sequence data to the community prior to publication. From the 297 complete and published genome projects, 235 are bacterial, 23 are archaeal and 39 are eukaryotic.

Ongoing genome projects

In addition to the completed projects, there are currently 1263 ongoing sequencing projects. Of those, 697 are bacterial, 38 archaeal and 526 are eukaryotic projects. The latter includes 208 EST and 10 RST projects, in addition to the 308 genome projects. These can be retrieved by using GOLD's search engine, selecting 'EST' or 'RST' or 'Genome' at the *Type* field.

From the 1263 ongoing projects, 114 are also considered complete at this point, i.e. the sequencing phase has been completed but the data are not yet submitted to the public sequencing repositories. These can be retrieved using the search engine by selecting 'Complete Unpublished' at the *Status* field.

GOLD is not limited to providing information on sequencing projects for which results will become publicly available at some point in the future. Rather, it seeks out and displays all publicly reported projects, whether the actual data will become public at some point or remain proprietary. It is our hope that this will better serve researchers, agencies and sequencing centres in the process of selecting new projects, or identifying sources of currently existing ones. These projects can be

*To whom correspondence should be addressed. Tel: +1 925 296 5718; Fax: +1 925 296 5720; Email: NCKyrpides@lbl.gov retrieved by selecting ‘Proprietary’ at the *Availability* field of the Search page. GOLD currently has information for the sequencing of 56 proprietary genome projects running at various private companies. Usually only the information for the sequencing project itself has been made available in these cases. A total of 25 such projects are also considered completed.

Sequencing is currently being performed in a large variety of sequencing centres, through a variety of funding sources and analysis is presented in many different databases. As displayed in the indexing link of the database, GOLD reports sequencing projects from 566 sequencing centres, funded from 186 agencies, and links to 427 distinct databases that provide sequence data analysis and information for the above genome projects.

NEW DEVELOPMENTS

Since the last report (2), a number of additional data fields have been added to the database. These include new data fields available in the project tables, as well as in the search engine.

The project tables now have the following additional fields: (i) *GOLDSTAMP*, a unique identifier for each project in the database. This ensures that multiple projects for the same species can be more easily tracked and distinguished. All completed published projects receive a *Gc-ID* (‘*c*’ standing for complete) project number that follows the order of completion of each project. All draft projects receive a *Gi-ID* (‘*i*’ standing for incomplete) project number which follows the entry to the database order. In a similar manner, all metagenomics projects receive a *Gm-ID*. In the future all EST projects will receive a *Ge-ID*. Accordingly, there are three major types of projects presented in the database, corresponding to the four types of unique identifiers (i.e. genome, metagenome and EST or RST). The search engine allows queries for each of these three major data types separately, through the *Type* field. A search or browsing can also be performed on the Goldstamp IDs through the corresponding search field. (ii) *GC content*, which displays the GC percentage of the organism, when known. (iii) *Contact* information, which provides the name and contact information of the PI responsible for a given project. All these fields are available for either search or browsing through the search engine.

The most important new development in the database is the addition of new data types pertinent to the properties of the organism or the sequencing project.

Under the first category three new fields have been added in the Search page: Phenotype, Ecotype and Disease. Each of these three fields has been populated with a number of attributes or keywords extracted from the literature or other public sources such as NCBI’s Entrez Genome Project Database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) or TIGR’s Genome Properties (4). *Phenotype* includes attributes that describe phenotypic properties of the organisms such as Motility, Temperature growth, Oxygen requirements, etc. *Ecotype* includes attributes related to the environment or habitat the organism is usually found in (i.e. Aquatic, Soil, Host, etc.). Finally *Disease*, includes attributes related to disease names or patterns Although a standardized controlled vocabulary to describe the above fields has not yet been developed, where ever a similar attribute was identified to other public resources, every effort was made to maintain the already existing vocabulary. As shown on the GOLD indexing information page, there are currently 203 phenotypic related attributes, 117 related to Ecotype and 188 are disease related keywords.

Under the second category (i.e. properties of the sequencing projects) two new fields have been added in the search page: Relevance and Availability. *Relevance* refers to attributes that provide information for the motivation behind the selection of different projects (i.e. biotechnological, medical, environmental or phylogenetic). *Availability* of the project, refers to the anticipated data release plan that may become either public or remain proprietary.

OVERVIEW STATISTICS

Although several different types of statistics, related to each of the data fields, can be derived from the user at any point using the search engine, the database also provides readily available graphical overviews for specific data types. These are provided through the link ‘*Gold Statisitcs*’ available on the home page of the database and include the following data types.

Sequencing centers

More than half of the 1500 currently available sequencing projects on GOLD are distributed among only six major sequencing centres. On top of the list is the Joint Genome Institute which is the Department of Energy sequencing facility with 18% of world production at this point. These represent number of unique individual projects and do not correspond in any way with the actual size of the project in number of sequenced bases.

Phylogenetic distribution

The sampling bias (5) towards only three major bacterial lineages (Proteobacteria 52%, Firmicutes 23%, Actinobacteria 8%) continues to persist despite the large increase in sequencing projects (Figure 1). However, the total number of bacterial phyla sampled for genome sequences has improved dramatically since the last release of GOLD. The coverage of archaeal and eukaryotic diversity by genome projects is also improving, albeit largely from recently announced projects for which no sequence data yet exist. With the advent of metagenomics (6,7), it becomes even more imperative to have a large and diverse set of reference genomes for separation and comparative analysis of these complex, multi-genomic datasets.

Project relevance

Biomedical relevance continues to be the dominant motivation for genome sequencing projects (44%) followed by biotechnological relevance (41%). The number of projects with Environmental relevance (12%) has dramatically increased over the last two years, mostly due to the Moore’s Foundation Initiative (http://www.moore.org/microgenome/default.asp) on sequencing of marine microbes. Finally projects with Phylogenetic relevance are restricted to just 3% and are mostly supported from the NSFs Tree of Life program, and more recently from the DOE community sequencing program.

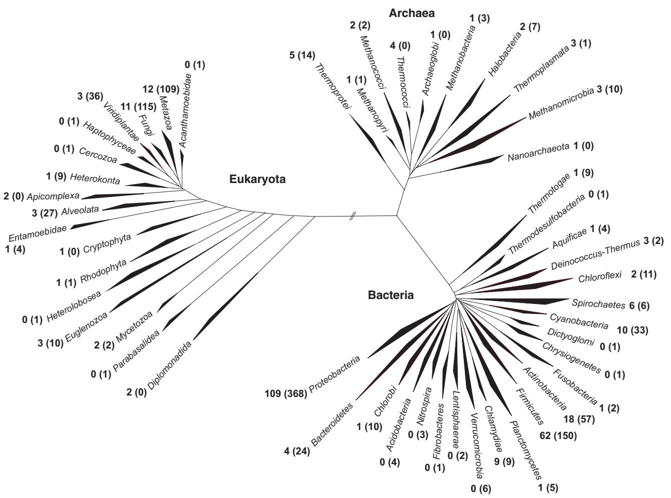

Figure 1. Schematic representation of the genome sequencing coverage of the major lineages in the tree of life as of September 15, 2005. The numbers next to each lineage denotes number of published complete genomes (and ongoing projects).

DATABASE WEB AND SEARCH ENGINE

The exponential increase of the size of the data and the amount of viewer hits brought the need for a transformation of GOLD and adaptation of a three-tier architecture. This architecture includes a Relational Database Management System, which holds the data, web clients that view the data and a middle interface that makes the connection possible. The new architecture was implemented by the Postgres RDBMS because it is very robust and also free, Perl which is the language of preference when it comes to regular expressions and displaying database content on a web interface. Postgres provides 'binary tree (B-tree)' indexing that make retrieval of data fast and modules like the Perl DBI make getting data from a database trivial. The looks of the web interface were retained so as not to confuse the usual viewer but were enhanced by more vivid readable fonts and images. The viewer also has the ability to sort the displayed data according to almost any possible data type. Finally the backend flat text files, out of which the database tables get generated, also retained the old format for the ease of use for the curators but now are holding a lot more new data fields. The actual engine of GOLD is completely different. Instead of pre-creating static html pages out of flat text files, GOLD now is using PERL CGI scripts that access the data from the Postgres database and dynamically create the corresponding page according to the viewers request. A very powerful search engine page gives the ability to the user, of not only searching by any data field possible, but also selecting which data types should be displayed for the returning results. This is possible by the construction of the SQL query created from the selections/deselections of the client. The SQL query takes advantage of the relations between the different tables, retrieves the results and passes them to the web CGI scripts, which then display the results.

AVAILABILITY

GOLD can be accessed at http://www.genomesonline.org/.

Further comments and feedback are welcome at mail@genomesonline.org.

ACKNOWLEDGEMENTS

We are grateful to all the colleagues who kindly provide information for the more accurate monitoring of the genome projects. The list of contributors is available at http://www.genomesonline.org/acknowledgments.html. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48, Lawrence Berkeley National Laboraory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. Funding for the Open Access publications was provided by the Lawrence Berkeley National Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Kyrpides,N.C. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.
3. Diggle,M.A. and Clarke,S.C. (2004) Pyrosequencing: sequence typing at the speed of light. *Mol. Biotechnol.*, **28**, 129–137.
4. Haft,D.H., Selengut,J.D., Brinkac,L.M., Zafar,N. and White,O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, **21**, 293–306.
5. Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, REVIEWS0003.
6. DeLong,E.F. (2004) Microbial population genomics and ecology: the road ahead. *Environ. Microbiol.*, **6**, 875–878.
7. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.