Genomes OnLine Database (GOLD): a monitor of genome projects world-wide

Axel Bernal, Uy Ear and Nikos Kyrpides*

Integrated Genomics, Chicago Technology Park, 2201 West Campbell Park Drive, Chicago, IL 60612, USA

Received September 5, 2000; Revised and Accepted November 8, 2000

ABSTRACT

GOLD is a comprehensive resource for accessing information related to completed and ongoing genome projects world-wide. The database currently provides information on 350 genome projects, of which 48 have been completely sequenced and their analysis published. GOLD was created in 1997 and since April 2000 it has been licensed to Integrated Genomics. The database is freely available through the URL: http://igweb.integratedgenomics.com/GOLD/.

INTRODUCTION

The year 1995 marked the birth of a new era in biology. During that year, the complete DNA sequence of an autonomous living organism, *Haemophilus influenzae,* became available (1), although the entire genetic material of a phage (bacteriophage ΦX174) was completely sequenced eighteen years earlier (2). Since 1995, >30 genomes have been completely sequenced and published (3) including three eukaryotic, six archaeal and 26 bacterial. Each of these sequencing projects has been accompanied by a variety of analyses and studies performed by institutes and universities around the world. In many cases, the results are made freely available to the public through Internet-based databases, such as WIT (http://wit.integratedgenomics.com/IGwit/) (4), KEGG (http://www.genome.ad.jp/kegg/kegg2.html) (5), GeneQuiz (http://www.sander.ebi.ac.uk/gqsrv/submit/) (6) and MIPS (http://www.mips.biochem.mpg.de/) (7).

SCOPE OF THE DATABASE

Researchers often face the problem of not having adequate information easily accessible or of not becoming easily aware of the new developments and available data in their field. Genomics is arguably one of the most rapidly developing areas in biology today, with actual data from sequencing projects increasing exponentially over the past five years. Coupled to this development is the growing number of databases that explore and analyze these data. The drop in sequencing cost and the improvement in sequencing technology has made possible a very large number of sequencing projects around the world (8). This proliferation of sequencing projects has created a need for constructing and maintaining a resource that would monitor and display information pertinent to a genome project.

HISTORY AND GROWTH

GOLD was established in 1997 as a web resource that would collect and provide information for all publicly available genome projects. At its inception, GOLD held information on six complete genomes and a handful of ongoing genome projects. Today, GOLD provides information on 197 genomic sequencing centers and 74 funding agencies covering 350 genome projects. Of those genomes, 48 have been completely sequenced and their analysis published. GOLD also reports as ‘in progress’ 176 prokaryotic and 126 eukaryotic genome projects. Furthermore, the database provides over 3200 hypertext links.

DATABASE ACCESS

GOLD is available through Integrated Genomics’s web server at the entry point: http://igweb.integratedgenomics.com/GOLD/. Users have direct access to the database’s pre-computed tables that include the complete published genomes and the ongoing prokaryotic and eukaryotic genome projects. They can also use the search form to query specific features or information about a genome project. Corrections, suggestions and feedback are most welcome at gold@integratedgenomics.com.

DATABASE FORMAT

GOLD is currently built as a set of ASCII (text) files. The information for every reported genome is organized into specific fields to record the various data types: the organism name (Organism), the phylogenetic position of an organism (Tree), general information on an organism (Information) and the size of the genome and the number of the predicted ORFs (Size). In addition, there are hyperlinks to the actual data such as the DNA sequence, the lists of ORFs and functions (DATA), the organizations that completed (or are completing) the sequencing of an organism (Institution) and the agencies that provided the funds for the sequencing project (Funding). There is also a collection of links to different databases that provide online analysis for a particular genome (Genome Database). Finally, the current status of ongoing projects (Status) and the references for completed and published genomes (Publication) are also reported.

DATABASE SEARCH ENGINE

As of March 2000, GOLD has taken on a relational model that allows for indexing of the fields described above. An engine

*To whom correspondence should be addressed. Tel: +1 312 226 9435; Fax: +1 312 226 9446; Email: nikos@integratedgenomics.com that generates Text-Based Relational Database (TBRD engine) was created using Perl. The TBRD engine is capable of resolving any Sentence Query Language (SQL) statement and displaying the index in the GOLD relational database. Before the TBRD generation, a program was written to extract as much data as possible from the former GOLD database. Then, manual curation followed to complete omissions or fix inconsistencies. This process eliminates duplications of data and standardizes the storage and display formats of the tables and their contents. The present GOLD interface, built by CGI scripts, allows a user to input queries and receive results. Once the input is received, the CGI scripts build an SQL statement and send it to the TBRD engine. The CGI scripts then wrap the results from the TBRD engine in HTML format and give them to the users.

ACKNOWLEDGEMENTS

We are grateful to our colleagues who have contributed information for a more accurate monitoring of genome projects. GOLD has been supported by Integrated Genomics.

REFERENCES

1. 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. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,C.A., Hutchison,C.A., Slocombe,P.M. and Smith,M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, **265**, 687–695.
3. Kyrpides,N.C. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.
4. Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.,Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
5. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
6. Hoersch,S., Leroy,C., Brown,N.P., Andrade,M.A. and Sander,C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem. Sci.*, **25**, 33–35.
7. Mewes,H.W., Frishman,D., Gruber,C., Geier,B., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F., Schuller,C., Stocker,S. and Well,B. (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 37–40.
8. Overbeek,R. (2000) Genomics: what is realistically achievable. *Genome Biol.*, **1**, 2002.1–2002.3.