Nucleic Acids Research, 2006, Vol. 34, Database issue ***D697–D699***
doi:10.1093/nar/gkj098

GOBASE—a database of organelle and bacterial genome information

Emmet A. O'Brien*, Yue Zhang, LiuSong Yang, Eric Wang, Veronique Marie, B. Franz Lang and Gertraud Burger

Center Robert-Cedergren for Bioinformatics and Genomics, Département de Biochimie, Pavillon Roger-Gaudry, Université de Montréal, 2900 Edouard-Montpetit, Montréal, QC, Canada H3T 1J4

Received September 13, 2005; Revised and Accepted October 16, 2005

ABSTRACT

The organelle genome database GOBASE is now in its twelfth release, and includes 350 000 mitochondrial sequences and 118 000 chloroplast sequences, roughly a 3-fold expansion since previously documented. GOBASE also includes a fully reannotated genome sequence of *Rickettsia prowazekii*, one of the closest bacterial relatives of mitochondria, and will shortly expand to contain more data from bacteria from which organelles originated. All these sequences are now accessible through a single unified interface. Enhancements to the functionality of GOBASE include addition of pages for RNA structures and a page compiling data about the taxonomic distribution of organelle-encoded genes; incorporation of Gene Ontology terms; addition of features deduced from incomplete annotations to sequences in GenBank; marking of type examples in cases where single genes in single species are oversampled within GenBank; and addition of graphics illustrating gene structure and the position of neighbouring genes on a sequence. The database has been reimplemented in PostgreSQL to facilitate development and maintenance, and structural modifications have been made to speed up queries, particularly those related to taxonomy. The GOBASE database can be queried at http://gobase.bcm.umontreal.ca/ and inquiries should be directed to gobase@bch.umontreal.ca.

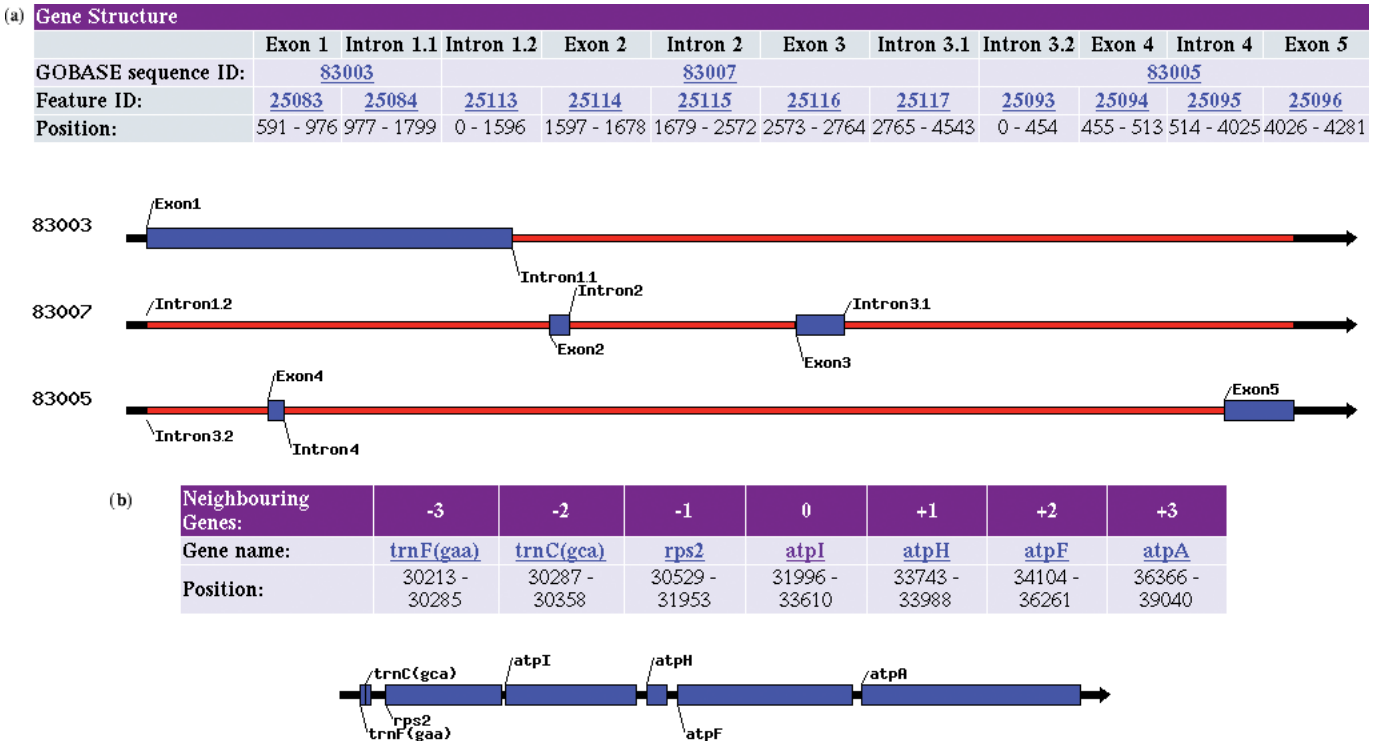
INTRODUCTION

Mitochondria and chloroplasts are of interest to biologists for studies as diverse as population genetics, molecular taxonomy and understanding metabolism-related disease in humans. The volume of information available concerning these organelles is constantly increasing and diversifying. There is therefore a growing need for specialist databases to collect, cross-reference and annotate this information for the requirements of different research communities, to set raw data in the context of expert knowledge and to complementing the role of general databases such as GenBank (1). GOBASE was designed primarily to address broad issues of comparative biology, such as the evolutionary origins of organelle endosymbiosis, gene migration to the nucleus and diversity of genome architecture, gene structure and gene expression mechanisms in organelles (2,3). Organelles are well suited to evolutionary studies because of the large number of complete genomes available. GOBASE release 12 (May 2005) contains 1517 complete mitochondrial genomes and 43 complete chloroplast genomes.

GOBASE has been gathering biological information related to mitochondria and chloroplasts, curating this information and making it publicly available through a web-based interface since 1995 (4–7). GOBASE contains a number of different categories of data (nucleotide and protein sequences, taxonomical data, RNA secondary structures, genetic maps) all of which have been collected and verified by expert curators. Gene and product names are assigned from a standardized list maintained internally, to allow for ease of searching and sorting. This assembly of data is made available to researchers through an intuitive interface allowing for a wide range of precisely specified searches. GOBASE release 12 (May 2005) contains ∼350 000 mitochondrial sequences, including 150 000 proteins, and 120 000 chloroplast sequences including 43 000 proteins, derived mostly from GenBank release 145. This represents a roughly 3-fold increase in the contents of GOBASE since the last documented release (7).

To further enhance the database's utility as a tool for evolutionary comparison, we have recently started to integrate data from bacteria closely related to the ancestors of mitochondria and chloroplasts into GOBASE. A genome sequence for *Rickettsia prowazekii* was obtained from GenBank (1) and has been comprehensively re-annotated by a combination of the AutoFACT automated annotation

*To whom correspondence should be addressed. Tel: +1 514 343 6111; Fax: +1 514 343 7936; Email: eobrien@bch.umontreal.ca



(a)

Gene Structure											
	Exon 1	Intron 1.1	Intron 1.2	Exon 2	Intron 2	Exon 3	Intron 3.1	Intron 3.2	Exon 4	Intron 4	Exon 5
GOBASE sequence ID:	83003		83007					83005			
Feature ID:	25083	25084	25113	25114	25115	25116	25117	25093	25094	25095	25096
Position:	591 - 976	977 - 1799	0 - 1596	1597 - 1678	1679 - 2572	2573 - 2764	2765 - 4543	0 - 454	455 - 513	514 - 4025	4026 - 4281

(b)

Neighbouring Genes:	-3	-2	-1	0	+1	+2	+3
Gene name:	trnF(gaa)	trnC(gca)	rps2	atpI	atpH	atpF	atpA
Position:	30213 - 30285	30287 - 30358	30529 - 31953	31996 - 33610	33743 - 33988	34104 - 36261	36366 - 39040

Figure 1. (**a**) Gene structure diagram for the mitochondrial *nad1* gene from *Arabidopsis thaliana*, GOBASE feature ID 25091. The table shows links to the sequence entries containing this information, the intron and exon feature entries assigned to each sequence, and the positions of each feature on the respective sequence. Exons are shown in blue and introns in red. The images are each scaled to a standard width; in cases where exons are widely separated on a sequence, a breakpoint is indicated in the image. (**b**) Diagram of genes in the vicinity of chloroplast *atpI* gene from *Euglena gracilis*, GOBASE feature ID 784380. The table contains the position of each gene on the sequence and links to the entries for each of the neighbouring genes. The diagram indicates genes in blue and intergenic regions in black, scaled to a standard width. Strand direction is indicated by an arrow; in cases where there are neighbouring genes on both strands, one image is shown for each strand.

system (8) and expert manual curation. Additional bacterial data will follow in GOBASE release 13, due later in 2005.

STRUCTURAL REDESIGN/NEW FEATURES

The GOBASE interface contains PHP query interface pages corresponding to the classes of biological entity represented within the GOBASE database: these are Gene, Gene&ProductClass, Sequence, Protein, Exon, Intron, RNA, RNAStructure, Taxonomy, Map and GeneDistribution. Since release 8.1 (October 2003) access to the mitochondrial and chloroplast datasets in GOBASE has been combined, and the previously independent interface to chloroplast data has been retired. In each page of the GOBASE interface, users can now select the origin of the data which they will query, with the options of searching mitochondrial, chloroplast or bacterial data, or all of these datasets.

GOBASE now includes several new pages, allowing the contents of the database to be interrogated in different ways. The GeneDistribution page shows an overall summary of the distribution of mitochondrial genes, sorted into columns by functional category and into rows by species ordered by taxonomic division, in order to facilitate assessments of the distribution of specific genes or functional gene classes, in specific organisms or across clades. An RNAStructure page has also been added to the interface, providing direct links to .pdf files containing diagrams illustrating the structures of many of the ribosomal and RNAse P RNA sequences contained in the database, with links to the appropriate sequence and RNA feature entries. Finally, an updated Taxonomy interface page has been added, making use of a novel database architecture (manuscript under preparation) to provide rapid and efficient navigation of a structure representing the NCBI taxonomic tree and access to all GOBASE data relating to any clade of interest at any level (taxonomic rank) in the tree.

The GOBASE interface has been redesigned to enhance querying and representation of results. The Gene query result page now contains graphics illustrating the internal structure of complex genes (Figure 1a) and neighbouring genes on the chromosome (Figure 1b). This also allows for a more sophisticated representation of *trans*-spliced genes than has previously been possible.

Information from the Gene Ontology project (9) has also been integrated into the GOBASE database. Every gene and gene product defined in GOBASE is associated with a suitable set of Gene Ontology terms as determined by our curators. This Gene Ontology information is accessible to GOBASE users through the Gene&ProductClass interface.

GOBASE has recently started to include deduced features, taken from data that are only implicit in a GenBank entry. For example, while exons are usually explicitly identified in an entry in GenBank, this may not be the case for their cognate introns. In such instances, the presence of an intron is inferred from the positions of the exons bounding it, and an entry for that intron is generated internally in GOBASE. Also, new sequence entries are checked for transfer RNA sequences using the program tRNAscanSE (10) and any putative RNAs identified by this method which have not already been annotated are marked in GOBASE as deduced features. Deduced features are distinguished by colour on the appropriate query results page. Examples can be seen from the RNA query page for gene name 'trnK' and taxon name 'Panax', or from the intron page for gene name 'trnG' and taxon name 'Prunus'.

There are cases where the available data in GenBank contain numerous identical or near-identical sequences derived from population studies, and this sample bias can be inconvenient in certain queries. For example, GOBASE contains more than a thousand entries for the *cox1* gene from *Homo sapiens* mitochondria, and to retrieve all of these by default may be inappropriate for researchers interested in the evolution of *cox1* in a taxonomically broad context. We have therefore implemented a procedure for marking type examples (selected subsets of sequences to accurately represent the range of larger datasets) within GOBASE, such that for any situation where more than five copies of the same gene exist from a given species, the sequences are aligned using CLUSTAL W (11), and the most distantly related five are selected as type examples and marked as such in the interface. By default, GOBASE query results show only the type examples for these highly sampled genes, but the user may select the option of retrieving all sequences. Type examples are recalculated with every new population of the database.

IMPLEMENTATION

The GOBASE database is implemented in version 7.4.1 of the PostgreSQL relational database management system with a web interface written in v4.3.8 of the PHP scripting language. The graphics on the gene pages are generated using the GD module for Perl/PHP, version 2.0.25. Perl (5.8.0) scripts are used to download data from GenBank and process it into GOBASE. All procedures are executed on PCs with two 2.4 GHz or 2.8 GHz Intel Xeon CPUs.

FUTURE PLANS

The addition of bacterial data to GOBASE will continue, with the inclusion and reannotation of genomes from cyanobacteria, closely related to the ancestors of plastids, more α-proteobacterial sequences, which are closely related to the ancestors of mitochondria, and *E.coli* strain K12, the biochemically best-studied eubacterium. The presence of these sequences will permit more comprehensive comparative analysis of gene structure and function in organelles and the evolutionary relationships between organelles and their bacterial predecessors.

We also intend to include information related to RNA editing in GOBASE in the near future. RNA editing is the programmed alteration of a transcript relative to the gene from which it is transcribed, and occurs in a broad range of biological contexts but is best documented in mitochondria (12). We have developed techniques for parsing information related to RNA editing from GenBank entries, with the intent of storing this information in GOBASE and making it available to users in a clear and consistent fashion.

ACKNOWLEDGEMENTS

The authors would like to thank Ning Luo for work on modifying the GOBASE population procedures for the PostgreSQL implementation of the database, Carole Ricault for verification on the automatic assignment of gene names in GOBASE, Maria Yu and Arnaud Joubet for contribution to the development of the GOBASE interface and Allan Sun for system administration. This project was funded by grant MOP-15331 from the Canadian Institutes for Health Research (CIHR, Genetics Institute). Salary and interaction support from the Canadian Institute for Advanced Research (CIAR, Program in Evolutionary Biology) to G.B. and B.F.L. is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the Canadian Institutes for Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
2. Gray,M.W., Lang,B.F. and Burger,G. (2004) Mitochondria of protists. *Annu. Rev. Genet*., **38**, 477–524.
3. Burger,G., Gray,M.W. and Lang,B.F. (2003) Mitochondrial genomes—anything goes. *Trends Genet*., **19**, 709–716.
4. Korab-Laskowska,M., Rioux,P., Brossard,N., Littlejohn,T.G., Gray,M.W., Lang,B.F. and Burger,G. (1998) The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res*., **26**, 138–144.
5. Shimko,N., Liu,L., Lang,B.F. and Burger,G. (2001) GOBASE: the organelle genome database. *Nucleic Acids Res*., **29**, 128–132.
6. Barbasiewicz,A., Liu,L., Lang,B.F. and Burger,G. (2002) Building a genome database using an object-oriented approach. *In Silico Biol*., **2**, 213–217.
7. O' Brien,E.A., Badidi,E., Barbasiewicz,A., deSousa,C., Lang,B.F. and Burger,G. (2003) GOBASE—a database of mitochondrial and chloroplast information. *Nucleic Acids Res*., **31**, 176–178.
8. Koski,L.B., Gray,M.W., Lang,B.F. and Burger,G. (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics*, **6**, 151.
9. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
10. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., **25**, 955–964.
11. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
12. Gray,M.W. (2003) Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life*, **55**, 227–233.