176–178 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg090

GOBASE—a database of mitochondrial and chloroplast information

Emmet A. O'Brien*, Elarbi Badidi, Ania Barbasiewicz, Cristina deSousa, B. Franz Lang and Gertraud Burger

Program in Evolutionary Biology, Canadian Institute for Advanced Research, Departement de Biochimie, Universite de Montreal, 2900 Boulevard Edouard-Montpetit, Montreal, Quebec H3T 1J4, Canada

Received August 30, 2002; Accepted September 20, 2002

ABSTRACT

GOBASE is a relational database containing integrated sequence, RNA secondary structure and biochemical and taxonomic information about organelles. GOBASE release 6 (summer 2002) contains over 130 000 mitochondrial sequences, an increase of 37% over the previous release, and more than 30 000 chloroplast sequences in a new auxiliary database. To handle this flood of new data, we have designed and implemented GOpop, a Java system for population and verification of the database. We have also implemented a more powerful and flexible user interface using the PHP programming language. http://megasun.bch.umontreal.ca/gobase/gobase.html.

INTRODUCTION

During the last decade, the increasing capacity of computers to retrieve, organise and analyse information has led to a qualitative change in the ways in which it is possible to do biological research. The volume of information available to the community of biological scientists is constantly increasing and diversifying. While general databases such as GenBank (1) are essential to the storage of this information, they cannot be expected to address the specialised requirements of particular fields. There is a necessity for specialist databases to meet the individual needs of the various research communities within biological science, in order to collect and annotate different kinds of data related to particular subjects, to add the knowledge of experts to the raw data, and to provide a depth of information complementary to the breadth available in general databases. The organelle genome database GOBASE is one such specialised database. Since 1995, the GOBASE team has been gathering biological information from a number of sources (2,3) and augmenting this with data generated locally. Nucleotide and protein sequences, genetic and physical maps, RNA secondary structures and biochemical and taxonomic data have been verified and integrated in GOBASE for mitochondria and more recently for chloroplast information and sophisticated searches can be run on this collection of data.

Mitochondria and chloroplasts are of biological interest for many reasons. They play a key role in energy utilisation within eukaryotic cells, which is interesting both at a biochemical level and because mitochondrial dysfunction is linked, for example, to various human diseases and male sterility in plants. These organelles also contain many unique molecular mechanisms such as transsplicing and post-transcriptional RNA editing. GOBASE was created to address questions specific to organelle biochemistry and evolutionary origins, as well as broader issues in comparative biology such as the relationship between genome structure and function and questions relating to genome and organismal evolution.

Organelles are particularly suited to evolutionary studies because of the large number of complete genomes available. There are 408 complete mitochondrial genomes and 21 complete chloroplast genomes in GOBASE release 6, and many more are currently being sequenced [e.g. by the Organelle Genome Megasequencing Program (4) (http://megasun.bch.umontreal.ca/ogmpproj.html) and Fungal Mitochondrial Genome Project (5) (http://megasun.bch.umontreal.ca/People/lang/FMGP/FMGP.html)]. Comparative genome analyses can be carried out on this dataset in ways that are only beginning to become possible for nuclear genomes. For example, investigation of genome evolution and identification of complex gene rearrangements and their roles in evolutionary divergence and speciation can both be carried out more effectively on this collection of data.

GOBASE provides a central, comprehensive, well-validated resource where researchers can access integrated organelle data of many types, including nucleic acid and protein sequences, RNA secondary structures, standardised gene and product names assigned by GOBASE's biological experts, genetic maps and taxonomic information.

GOBASE release 6 (summer 2002) contains analysed and annotated sequences derived from GenBank release 129 (March 2002). This represent a major increase in the amount of data contained in GOBASE; where release 5 contained 83 270 sequences including 31 620 proteins, GOBASE release 6 contains a total of 130 780 mitochondrial sequences, of which 50 948 are proteins, an increase of ~37%.

*To whom correspondence should be addressed. Email: eobrien@bch.umontreal.ca

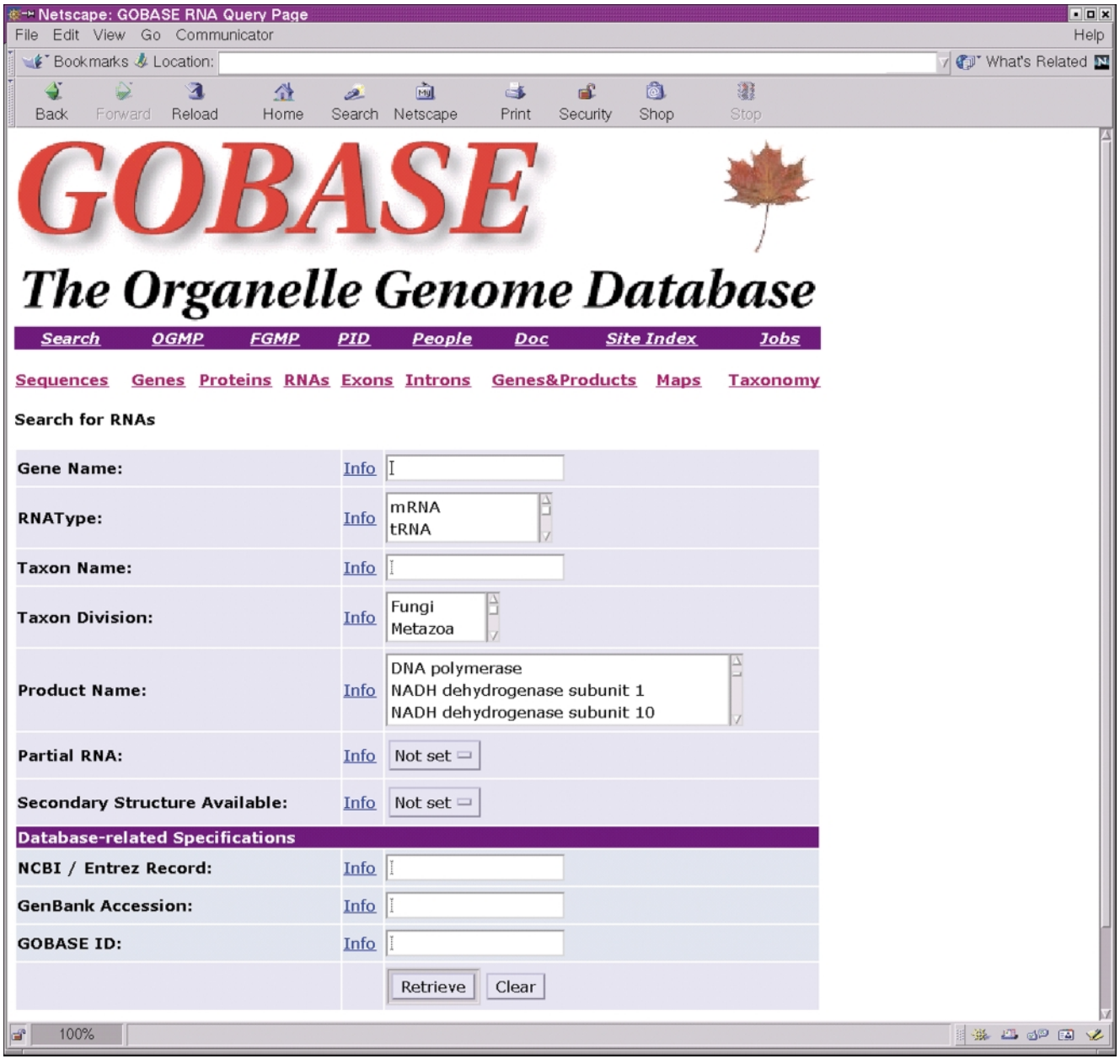

Figure 1. Example GOdigger interface page. This is the RNA search page. Parameters for the search are divided into biological features and database-related properties.

GOpop: MAINTENANCE AND UPDATING OF GOBASE CONTENTS

The database is currently maintained by a suite of programs written in Perl, Java and SQL. The programs retrieve data from GenBank, cross-reference sequence and taxonomy information, standardise gene and product names and validate the data in numerous other ways. Initial data gathering and parsing is carried out by Perl programs, while subsequent verification steps are carried out using a Java application, GOpop (6), developed in-house. GOpop allows new records to be verified based on a series of logical test, representing biological criteria in a simple Boolean fashion. For example, a new protein gene record is tested for the presence of an appropriate gene name and the existence of stop and start codons. GOpop also tests whether any record is complete and whether it already exists in the database. Reference lists are maintained to allow the programs to identify gene synonyms or recurrence of previously handled errors. GOBASE records are populated automatically as far as possible, but the complexity of these biological data is such that a complete and unambiguous system of automatic annotation is not feasible. Records are flagged for intervention when the appropriate annotation cannot be unambiguously carried out by automated means. In such cases the records are modified as necessary by a human expert before being included in a GOBASE release.

THE GOdigger INTERFACE

GOdigger is a PHP/JavaScript web-based interface to the GOBASE database. The Genera web interface (http://gdbdoc.gdb.org/letovsky/wgen.html), which GOBASE has used in the past, is of limited use with large datasets, as it can only display 500 results for any given query, and does not allow the graphical interface to be customised. The objective of the GOdigger project was to replace this interface and to provide users with a more intuitive and user-friendly method of accessing GOBASE.

The layout of the search pages in GOdigger is based on those in the Genera interface and has been generated using the PHP3 programming language. GOdigger is composed of scripts corresponding to each biological entity defined in GOBASE; these are Sequence, Gene, RNA, Protein, Exon, Intron, Product, GeneProductClass, Genome and Taxon. Each script permits the user to specify query parameters and retrieve data for the corresponding feature from the underlying database. A help script is also provided which explains the biological meaning of terms and the structure of specific data types and provides numerous examples.

The new interface has been designed to maximise the amount of information shown for multiple entries and limit the repetition of data. The appearance of the pages has been tailored to be more user friendly, distinguishing biological information from database-specific queries (Fig. 1). The complex GOBASE-specific labels previously used to identify individual result files have been removed as they are no longer necessary. The limits on the number of entries that could be retrieved from the database have also been eliminated from the new interface. GOdigger displays results on multiple pages, with a maximum of 100 entries per page, and page number hyperlinks allow the user to navigate easily between pages. The new interface was completely designed and programmed in-house and will therefore, be easier and more efficient to maintain.

CHLOROPLAST DATA IN GOBASE

We have gathered chloroplast data and made it available in GOBASE as a separate dataset. Our chloroplast dataset currently (summer 2002) contains 30 734 chloroplast sequences, including 13 839 proteins. This information can be accessed via a link from the GOBASE main page. The PHP interface for the chloroplast data has been derived from the Godigger interface for mitochondrial data.

FUTURE PLANS

We are in the process of transferring the GOBASE database from Sybase to PostgreSQL, an open-source relational database which is equally powerful and robust and more easy to manage. We are streamlining and redesigning our upgrade procedures accordingly. When this process is completed, we expect to be able to issue more frequent GOBASE updates.

In addition, we intend to make it possible to carry out joint queries of the mitochondrial and chloroplast datasets, in order to address questions relating to both organelles; these datasets are currently entirely separate. We also envision adding data, such as genome sequences from bacteria closely related to the ancestors of mitochondria and chloroplasts, in order to enhance GOBASE's capacity to handle broad evolutionary questions related to the origin of organelles and of the eukaryotic cell as a whole.

Finally, we are in the process of designing and implementing a Web-based interface to allow integrated access to the analytical tools developed by our bioinformatics group and others for the investigation of datasets retrieved from GOBASE. This 'workbench' employs CORBA architecture for accessing distributed tools and data, and will include such functions as translation from nucleotide to protein sequence, multiple sequence alignment, phylogenetic analysis, protein structure analysis and detection of introns and transfer RNAs. Users of the workbench will have to register with GOBASE in order to have access to project management operations and analysis tools. The workbench prototype is currently under development. It is written primarily in Java and carries out internal management through a MySQL relational database.

ACKNOWLEDGEMENTS

We wish to thank Alex Nip for his contribution to the technological framework of the database, Mathieu Coudert and Bruno Duclouet for development of the GOdigger PHP interface, Liusong Yang and Lin Liu for programming support in PHP and Java, and Allan Sun for system administration. This project was funded by grant GOP-15331 from the Canadian Institutes for Health Research. Salary and interaction support from the Canadian Institute for Advanced Research to G.B. and B.F.L. is gratefully acknowledged.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
2. 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.
3. Shimko,N., Liu,L., Lang,B.F. and Burger,G. (2001) GOBASE: the organelle genome database. *Nucleic Acids Res.*, **29**, 128–132.
4. Lang,B.F., Seif,E., Gray,M.W., O'Kelly,C.J. and Burger,G. (1999) A comparative genomics approach to the evolution of Eukaryotes and their mitochondria. *J. Eukaryot. Microbiol.*, **46**, 320–326.
5. Paquin,B., Laforest,M.J., Forget,L., Roewer,I., Wang,Z., Longcore,J. and Lang,B.F. (1997) The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. *Curr. Genet.*, **31**, 380–395.
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**, 0020.