138–144 *Nucleic Acids Research, 1998, Vol. 26, No. 1*

The Organelle Genome Database Project (GOBASE)

Maria Korab-Laskowska, Pierre Rioux, Nicolas Brossard, Timothy G. Littlejohn[1], Michael W. Gray[2], B. Franz Lang and Gertraud Burger*

Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada, [1]The Australian Genomic Information Centre, University of Sydney, NSW 2006, Australia and [2]Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Received September 16, 1997; Accepted September 26, 1997

ABSTRACT

The taxonomically broad organelle genome database (GOBASE) organizes and integrates diverse data related to organelles (mitochondria and chloroplasts). The current version of GOBASE focuses on the mitochondrial subset of data and contains molecular sequences, RNA secondary structures and genetic maps, as well as taxonomic information for all eukaryotic species represented. The database has been designed so that complex biological queries, especially ones posed in a comparative genomics context, are supported. GOBASE has been implemented as a relational database with a web-based user interface (http://megasun.bch.umontreal.ca/gobase/gobase.html). Custom software tools have been written in house to assist in the population of the database, data validation, nomenclature standardization and front-end design. The database is fully operational and publicly accessible via the World Wide Web, allowing interactive browsing, sophisticated searching and easy downloading of data.

INTRODUCTION

Research on mitochondria and chloroplasts is wide ranging, covering such diverse topics as the bacterial origin of these organelles, the functional and evolutionary relationship between organellar and nuclear genomes, the central role of organelles in energy production in eukaryotes, the involvement of mitochondria in human disease and their genetic variation within populations, and the ultrastructural characteristics of organelles in the different eukaryotic lineages. At present, 63 complete mitochondrial (mt) DNA and 13 chloroplast DNA sequences are available in public domain databanks. With the exception of viruses, organelle DNAs currently constitute the largest set of completely sequenced genomes. In addition, organellar DNAs are information-rich, typically containing 40–100 well-conserved genes, with non-coding spacer and intronic regions often comprising <10% of the overall sequence (for a recent overview, see 1). These features combine to make organelle genomes ideal for comparative genomic studies.

In the new era of genomics, concerted efforts are being made to sequence complete organelle genomes on a large scale. The Organelle Genome Megasequencing Program (OGMP), for example, has as its goal the sequencing of mitochondrial and chloroplast genomes from a phylogenetically broad range of protists (2) whereas the Fungal Mitochondrial Genome Project (FMGP) focuses on mtDNA of lower fungi (3). The body of organelle data is considerable, comprising detailed knowledge about organelle enzyme complexes and their catalytic functions, protein import and processing pathways, membrane architecture, genetic maps, translation codes, mechanisms of gene expression, protein and RNA secondary structures, and descriptions of the eukaryotic organisms that harbour organelles. However, this information is dispersed among numerous data sources (books, journals, websites, theses). For an investigator who is not directly involved in the research of a particular field, it is often difficult and time-consuming to locate organellar data of interest.

Biology in the era of genomics not only aims to investigate the coding content of DNA sequences and the potential function of newly discovered genes, but also probes interactions of gene products, genetic mechanisms that perpetuate genomes, and evolutionary forces that shape genomes at the molecular, population and ecological level. In order to fully exploit the large body of existing organelle data, we have established a database that organizes and integrates these various data types, drawing together pertinent information from dispersed data sources.

SCOPE OF GOBASE

GOBASE is intended to be more than simply a molecular sequence repository. By integrating sequence data and related biological information such as genetic maps, RNA secondary structures and organismal information, GOBASE aims to become a powerful tool in the comparative study of gene structure/function/evolutionary relationships. The database has been designed in such a way that questions can be posed in terms that are used to describe basic biological concepts. We have placed special emphasis on allowing queries that are structured at the high level of complexity inherent in biological problems.

It is a much deplored fact that sequence data in public-domain data repositories often contain errors or are incompletely annotated. The GOBASE team, which includes biology experts with longstanding experience in the field of organelle research, has made data correction and completion a main objective of this database project. Another goal of GOBASE is to assist the

*To whom correspondence should be addressed at: C.p. 6128, Succ. Centre-Ville, 2900 Boulevard Edouard-Montpetit, Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada. Tel: +1 514 343 7936; Fax: +1 514 343 2210; Email: burgerg@bch.umontreal.ca

organelle research community in data analysis. Researchers will be able to submit unpublished, confidential data and have password-protected access to view and analyze their data in the context of the integrated and enriched information contained in GOBASE.

To our knowledge, GOBASE is one of the first integrated, multiple-genome databases. The experience that we are accumulating in the construction of a database encompassing multiple organelle genomes should be applicable to bacterial and ultimately nuclear genomes in the years to come, when dozens or even hundreds of such genome sequences will have been fully determined.

In its initial phase, GOBASE is focusing on the mitochondrial subset of organelle data. In a second stage, it will also include data on chloroplasts and on representative bacteria that are thought to be specifically related to the bacterial ancestors of mitochondria and chloroplasts (α-Proteobacteria and Cyanobacteria, respectively).

DISTINCTIVE FEATURES OF GOBASE

GOBASE is not the only database that focuses on mitochondria. There are, for example, MmtDB (4), which compiles metazoan mtDNA variants, as well as MITOMAP (5) and MitoDat (6), which gather data on human mtDNA structure, function and pathogenic mutations and variations. However, because the spectrum of data contained in GOBASE is taxonomically broad, this database fills a niche that is not addressed by the foregoing mitochondrial databases. Only one other mitochondrial database, Mitbase (7), has adopted a comprehensive approach similar to that of GOBASE; however, Mitbase is still at an early stage in its development.

Databases other than GOBASE also support multi-genome comparisons; some of these databases include sophisticated graphical interfaces, e.g., GSDB (8), TDB (9) and the NCBI genome databases (10). In contrast to these other databases, GOBASE data have been corrected and standardized according to naming conventions, which permits the identification of homologous counterparts in a given set of genomes. This task is otherwise not easily carried out because different abbreviations are often used as synonyms for the same gene name in different organisms (e.g., the GenBank database employs five different abbreviations for the same mitochondrial gene, *cox1*).

In addition, particular emphasis has been placed on allowing complex biological questions to be asked in a context that is currently difficult or simply not supported in other genome databases. The Sequence Retrieval System (SRS; 11), for example, offers one of the most complex queries in the sense that multiple parameters can be set to specify and narrow down a particular question. In contrast to GOBASE, the searches apply to public-domain databanks only and the search terms are restricted to the feature keys used by the corresponding databanks. Questions supported in GOBASE include presence/absence of introns in genes; genes contained in other genes; genes that occur in one species of a given group only; standard names for homologous genes from different taxa; homologous genes encoded in different genomes (nucleus, mitochondrion and chloroplast); and searches using taxonomically broad terms such as 'fungi', 'protists' or 'arthropods', in addition to the particular species names that are supported in other databases. As molecular biology research continues to discover and unravel novel mechanisms of gene expression, especially in the organelle field, databases must be able to store and permit retrieval of information about such phenomena as *trans*-splicing, RNA editing, genes within genes, genes in pieces, and horizontal transfer of genetic elements.

CONTENTS OF THE DATABASE

GOBASE collects and integrates different types of data such as images, text and molecular sequences (Fig. 1). The current release contains all published sequences (DNA, RNA and protein) encoded by mtDNA; nucleus-encoded components that are imported into mitochondria are not included at present. Together with the sequences is stored information about them, such as molecule type, cellular location, completeness, submission date and sequence features (annotations) such as genes encoded, expression signals, etc. Also, general taxonomic information is available for all species for which sequence is contained in the database. This includes information concerning the entire taxonomic hierarchy, from species (e.g., *Paramecium aurelia*) up through phylum (Ciliates) to division (Protists) level. The particular rank for each taxonomic group is stored as well.

Whereas the above-mentioned information is stored locally, further data are integrated into GOBASE via links to external databases. Information from these sources includes secondary structures of mitochondrial introns and structural RNAs. Currently, 41 ribosomal RNA structures from plant, fungal, animal and protist mitochondria are available. The number of intron secondary structures is as yet limited; all belong to group I and are exclusively from fungi. In addition, GOBASE contains genetic maps of completely sequenced mitochondrial genomes. At the time of writing, the collection consists of 21 maps of protist and fungal mtDNAs, available in GIF and Postscript format, with new maps being added regularly. Organismal descriptions and ultrastructural images are also accessible in GOBASE, as is information about metabolic pathways, general enzyme function and E.C. number.

Tables compiling sequences, sequence features and taxonomic data have been populated from the NCBI Entrez system and Taxon database (12). Genetic maps, standard gene name assignments and general gene product information have been contributed by the GOBASE team. Secondary structure diagrams are accessible through links to the ribosomal RNA secondary structure database, maintained at the University of Colorado [R.R.Gutell (13); M.N.Schnare, M.W.Gray and R.R.Gutell (14)]. Organismal information is retrieved from the Protist Image Database [C.J.O'Kelly and T.G.Littlejohn (15)], and enzyme information comes from Mendel (16), ENZYME (17), EcoCyc (18), WIT(Puma) (19), PIR (20), SWISS-PROT (21) and Flybase (22).

DATA CORRECTION AND ANNOTATION

With the exponentially growing volume of sequence data submitted to public-domain databanks, it is hardly surprising that errors and inconsistencies remain undetected. Within the 20 488 sequence records retrieved during the process of populating GOBASE, we have found misidentified genes (e.g., *nad1* labelled as *cox1*), typing errors in species and product names (e.g., 'Threonone-tRNA'), free-style product names ('putative ORF with unnamed protein product'), missing information about the

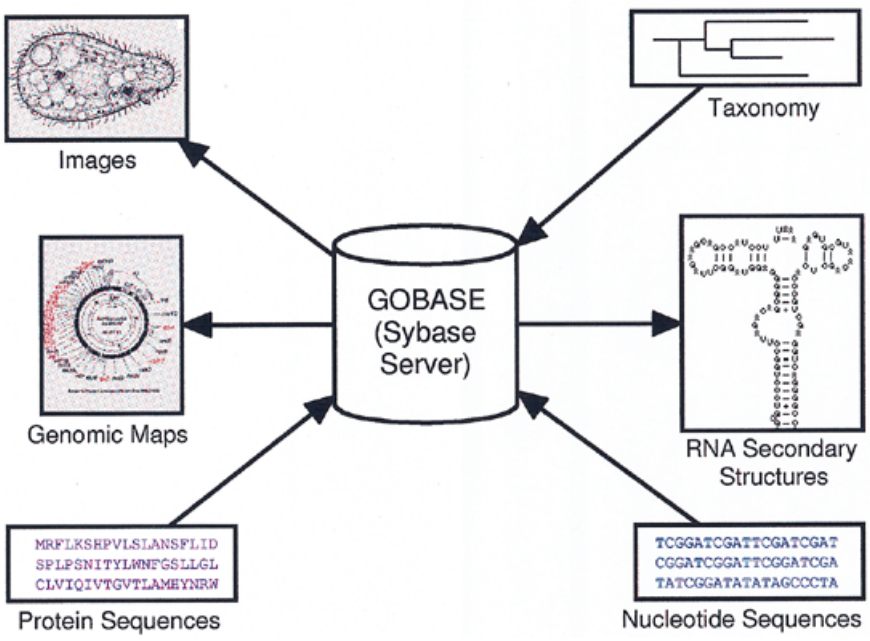

Figure 1. Data types integrated in GOBASE. The organismal images, RNA secondary structures and genetic maps are stored in other databases and accessed via hypertext links (arrows pointing to these resources).

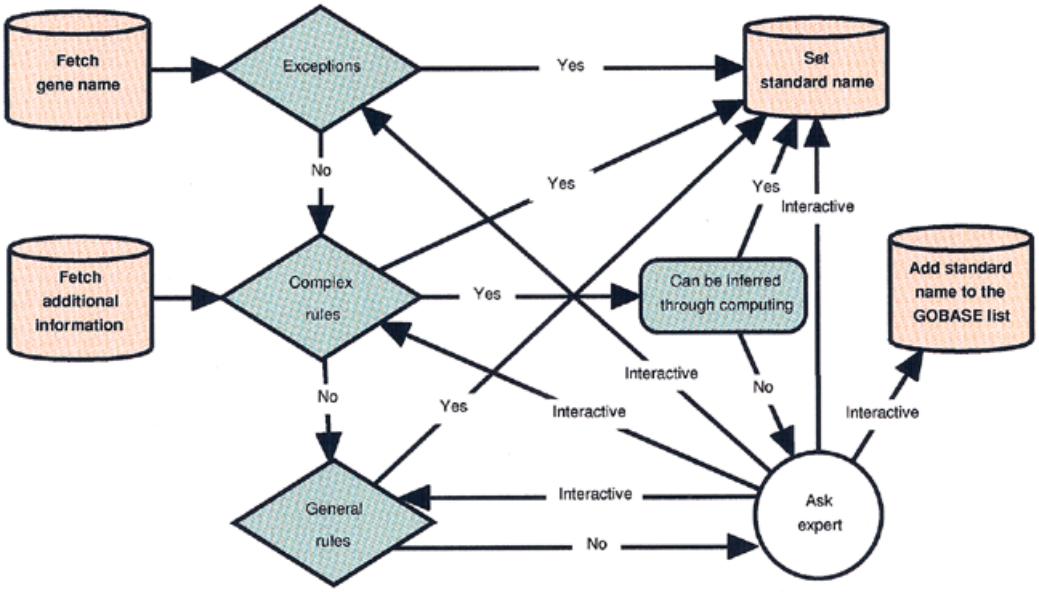

Figure 2. Assignment of standard gene names by the POP2LINKS program. The arrows indicate the next step to be taken, depending on whether or not a particular rule applies. ***Exceptions***, special cases, where an expert has assigned a name to an individual feature. ***General rules***, a list of synonyms or text pre-processing rules. ***Complex rules***, rules using additional information such as taxon names, anticodon information, length of the open reading frame (orf) and cellular location of the gene.

cellular location of a sequence or anticodon position in tRNA, and numerous other errors or omissions. It is obvious that such deficiencies compromise the identification and retrieval of sequence entries in data repositories, especially when using computer-based methods. Therefore, in addition to organizing and structuring organelle data, a major effort has been directed toward data verification. Before the data in GOBASE are made accessible to the public, they are scrutinized systematically. To assist the biology expert in the verification, correction and annotation of data, specialized informatics tools have been designed (Fig. 2). With the aid of the POP2LINKS program (see below), the names of 13 883 genes have been examined; based on the verified gene names, the product names have been assigned, missing information about cellular location has been added in 84 entries and corrected in 27, and the anticodon position has been determined for 176 tRNAs.

In addition to correcting errors or completing information that is typically contained in GenBank entries, new attributes are added to the GOBASE objects. Some of these (e.g., the group to which an intron belongs or whether a gene is trans-spliced) are

assigned by an in-house expert. Other features can be calculated or determined from further attributes or objects, for example, whether or not a gene is located within another gene and number of species whose genome contains a particular gene.

DATABASE FRONT END

The user interface has been designed with the goal of allowing remote public access through pre-existing and widely accessible client-server software, thereby permitting easy interactive browsing, sophisticated and at the same time intuitive query possibilities, and facile downloading of single or multiple query results. Details about the GOBASE front end are provided below.

GOBASE is accessible through the WWW using any web browser, such as Netscape or Microsoft Internet Explorer. Currently, nine query forms are available, i.e., SEQUENCE, GENE, PROTEIN, RNA, EXON, INTRON, GENE & PRODUCT, MAP and TAXON; the implementation of SIGNAL and MODIFICATION is under way. In these query forms, various parameters can be set to specify and narrow down a particular biological question. In the GENE form, for example, one can search with the gene name (e.g., *atp8*) or, if the gene name is unknown, choose within the list of product names (= ATP synthase subunit 8) or select the product type (= protein) (Fig. 3A). Either the species name (e.g., *Boletus satanas*) to which the gene belongs can be specified, or any higher taxonomic grouping [e.g., Hymenomycetes (class) or Fungi (division)] when all homologous genes from that assemblage are of interest. The user may also specify whether or not the requested genes must contain introns, whether they should conform to a particular genetic code convention, whether partial sequences should be excluded, etc. All query forms allow retrieval by NCBI/Entrez record number, GenBank accession number or internal GOBASE number. For instance, upon entering the NCBI/Entrez gi number 786182 in the context of the GENE form, one obtains all 97 genes contained in the sequence entry for liverwort mtDNA. By specifying the GOBASE number (e.g., 104391) in the GENE query, a particular gene sequence can be retrieved, which is especially useful when more than one record exists for a particular gene.

When a query returns multiple items, e.g., 13 in the query for all available *atp8* gene sequences in fungi, a summary page is presented (Fig. 3B), within which the user can select between two major output options. One is to obtain information about the gene, such as nature and function of its product, the position in the published sequence, whether partial or complete, intron-containing or not, and also taxonomic classification from species to kingdom level (Fig. 3D). The second option enables the user to view and download the gene or protein sequences in FASTA format (Fig. 3C).

In contrast to many other databases, the display names (labels) of all items contained in GOBASE are not just uninformative numbers. Labels are composed of the gene name (e.g., *rnl*), cellular location (mt) and species name (e.g., rnl:mt:Bole.sata.176), thereby allowing the user to deduce the identity of the items readily.

STRUCTURE OF THE DATABASE

Although implemented in a relational database management system where the information is organized in multiple tables, GOBASE possesses the advantages inherent in an object-oriented database architecture, which is more intuitive for biological queries. To permit an object-oriented view of a relational database, the Genera package (23, see also below) has been chosen as the Sybase database/Web interface software.

The body of organellar data has been subdivided into a total of 11 categories that reflect the major components of molecular biology concepts (Fig. 4) and that correspond to the query forms mentioned above: *Sequence* (RNA and DNA), *Gene* (coding region), *Protein*, *RNA*, *Exon*, *Gene & Product* (general gene and product information), *Map* (genetic, physical), *Taxon*, *Signal* (regulatory elements such as promoters), and *Modification* (modified residues of DNA, RNA or protein). With the exception of *Signal* and *Modification*, which are currently being implemented, these categories constitute the entities of the database and have been populated and validated. Each entity possesses a set of attributes that describe the contained objects: e.g., in the case of the entity *Sequence*, attributes are *length*, *type of molecule*, *cellular location*, *encoded genes*, etc. Attributes may be single values, as in the case of *type of molecule* (e.g., DNA), or multiple values, as in the case of *encoded genes* (e.g., *cox1*, *nad2*, ...).

Five groups of tables are contained in GOBASE (Fig. 5).

(i) Primary tables reflect the NCBI/DDBJ/EBI data model, including features such as CDS, RNA, SITE, REGION; these have been populated with GenBank records in ASN.1 format. The elements in the NCBI data model coincide only partially with the GOBASE entities.

(ii) Secondary tables are required to correctly represent the GOBASE objects and to integrate the various types of information about these objects. Secondary tables contain two different types of attributes: the information contained in the primary tables, but annotated/corrected by biology experts and, in addition, new attributes that have been calculated or inferred via SQL scripts/procedures from other attributes (e.g., whether or not an open reading frame is contained within an intron). Secondary tables also contain new features that had not been explicitly annotated in ASN.1 records and that have been determined from related features (e.g., the existence of introns in genes is inferred from NCBI's CDS feature).

(iii) Tables of the third group contain standard gene and product names and general information about genes and products supplied by biology experts.

(iv) URL addresses of postscript and GIF files (genetic maps, RNA secondary structures) and links to objects in external databases are compiled in a fourth group of tables.

(v) Finally, a fifth group of tables stores the taxonomy data extracted from the NCBI Taxon database. A flat file (dump file) of this relational database has been used to import the taxon tables into GOBASE.

IMPLEMENTATION

As noted earlier, GOBASE has been implemented in a relational database management system, using Sybase. The query interface applies WWW forms generated by the Web/Genera gateway. In addition to employing third-party software, we have developed several custom software tools that perform the population of GOBASE as well as data annotation and correction, and that establish the page design of query forms. These programs have been written in Perl, using the Sybperl module and the Sybase Open-Client library of programming tools. The most important of the tools developed in house are briefly described below.

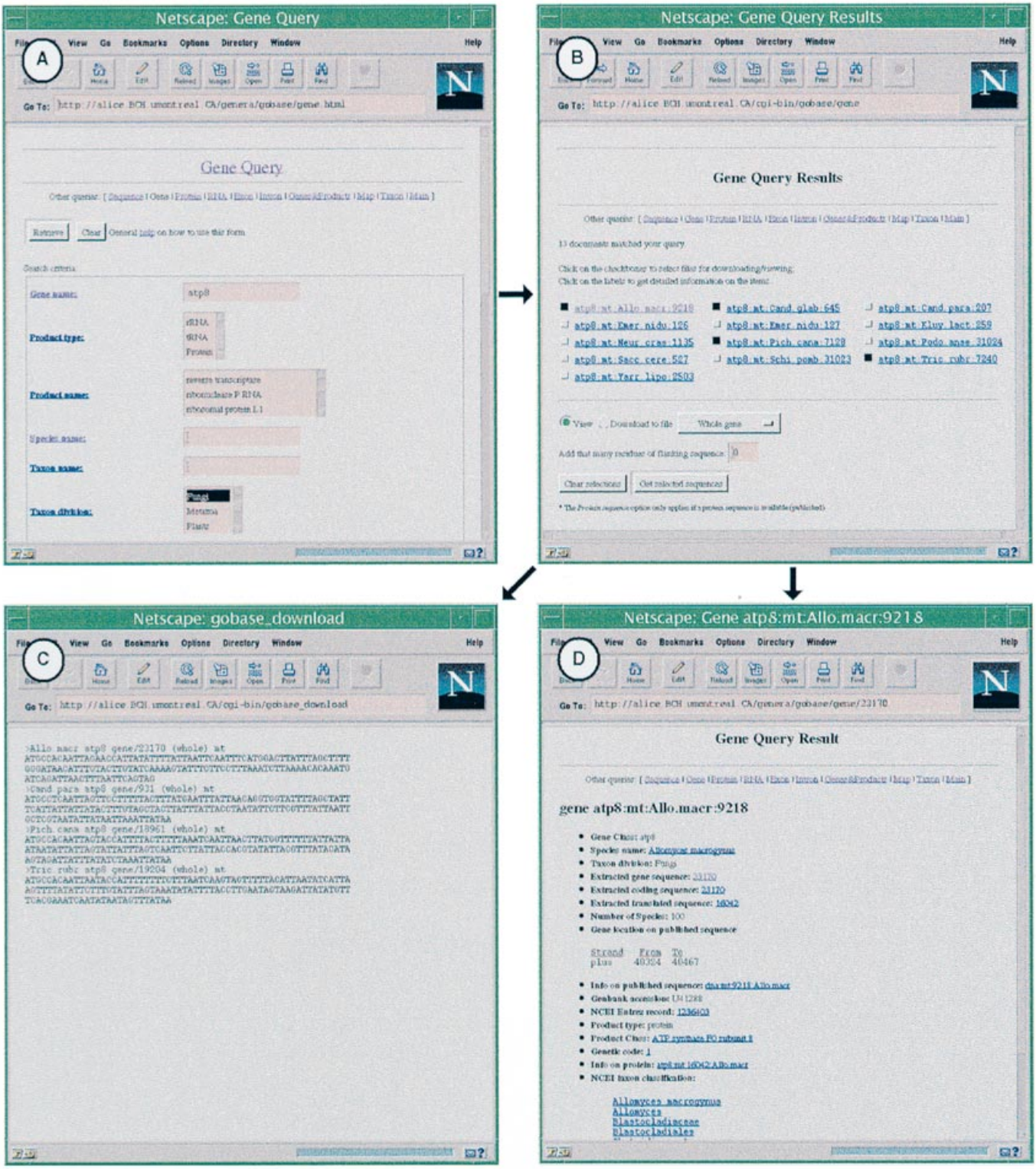

Figure 3. Sample GOBASE query, using the GENE query form. (**A)** The gene name *atp8* and the taxon division *Fungi* have been selected in the GENE query form. Upon activating the *Retrieve* button, page (B**)** is shown. (**B)** The 'GENE query results' page lists 13 genes that satisfy the query specified in (A). On this page, the user can opt to download/view one or multiple gene sequences by clicking on the checkboxes. Four items have been selected (black squares). Upon clicking on the *Get selected sequences* button, page **(C)** is displayed. When clicking on a gene label, here *atp8:mt:Allo.macr.9218*, page **(D)** is shown.

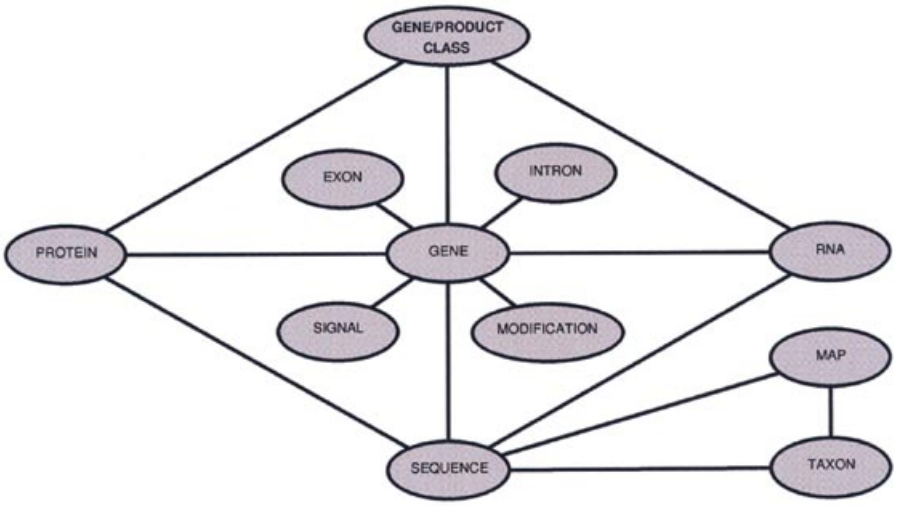

Figure 4. Entities in GOBASE and their interrelation. The lines between entities represent how the user can 'jump' from one form to another. Currently, SIGNAL and MODIFICATION are not fully implemented.

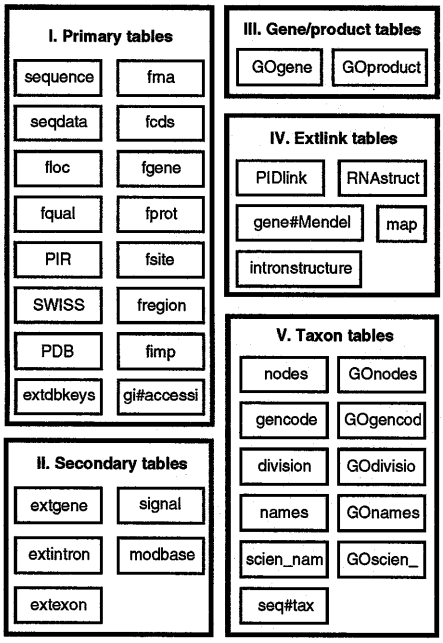

Figure 5. The five groups of GOBASE tables. Each small box represents a table. For more details, see text. The specifications of the individual tables can be examined at http://alice.bch.umontreal.ca/genera/gobase/table.html

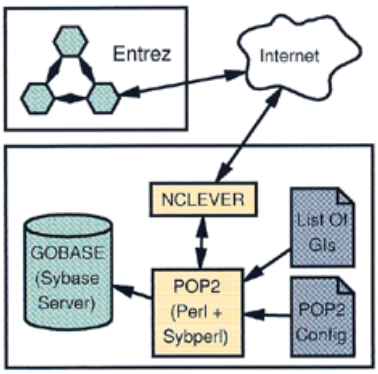

Figure 6. Population process of GOBASE with sequence data from GenBank. POP2 controls the population process, NCLEVER retrieves the sequences through the Entrez system. UIDs, unique sequence identifiers. POP2config, configuration file for POP2. For more details, see text.

POP2 (24) performs data acquisition from the Entrez system (Fig. 6). This tool parses ASN.1 records and writes the contents of the various fields into the corresponding GOBASE tables. A configuration file contains the Sybase table definitions, ASN.1 specifications and access methods to ensure that the population function is performed accurately. For locating all mitochondrial records in GenBank in the course of populating GOBASE and to establish a list of unique identifiers of Entrez records (UIDs), we interrogated the Entrez query system. NCLEVER (25), a command line version of the Network Entrez query system that permits batch processing, is used to fetch data from the Entrez databases via the Internet.

MF2ASN (Rioux,P. *et al.*, unpublished) converts a DNA sequence file in 'masterfile' (MF) format into an ASN.1 record (submission format) for direct submission to NCBI. The MF format, a sequence file format developed by the OGMP, integrates DNA sequence, feature annotations and experimental or analytical notes into a single FASTA-compatible ascii file. A masterfile is both human- and computer-readable (26). MF2ASN verifies the syntax and logic of annotations, intron/exon boundaries and translation of protein-coding genes, and it adds submission information such as references, taxonomy, and submitter's name and address.

SUB2PUB (Rioux,P., unpublished) converts an ASN.1 record in GenBank submission format into one in ASN.1 publication format. Together with MF2ASN, this tool allows us to populate GOBASE with sequence data that are confidential and have not yet been submitted to GenBank for publication.

MKGOFORMS (27) customizes the page layout of the GOBASE query forms. Fields can be hidden, grouped together and reordered, and the field descriptors are automatically hyperlinked to a corresponding entry in a help file. MKGOFORMS post-processes the basic HTML pages produced by the Genera software.

POP2LINKS (Rioux,P. *et al.*, in preparation), as noted earlier, is a tool that assists in the standardization of gene and product names. Based on a set of rules, this interactive learning program recognizes the various synonymous or erroneous names of genes and products and supplies the corresponding standard name. The rules make use of (i) a list of synonyms or text pre-processing procedures; (ii) additional information such as taxon name, anticodon, protein length (for naming ORFs) and cellular location; and (iii) a list of unique cases, in which the expert has assigned a particular name to an individual feature (e.g., that the incorrectly annotated *nad1* of record #1234 be renamed *cox1*). Certain missing or erroneous information detected by the program is completed automatically; in other cases, the biology expert is consulted. A more detailed description of POP2LINKS will be published elsewhere.

WORK IN PROGRESS

An improved version of GOBASE is in preparation with a projected release date of January 1998. At that time, queries with the current gene name synonyms will be possible, gene order information will be extractable, and links to a tRNA database will be added. We also plan to include secondary structures of RNAse P RNAs and 5S RNAs. Furthermore, security access will be implemented so that researchers can submit their confidential data for querying through GOBASE and be able to view the results in the context of published data. Finally, a data currency manager will be in place to ensure that the information available through GOBASE is updated regularly and automatically.

DATABASE ACCESS

The GOBASE database is publicly accessible through the megasun World Wide Web server at the URL:
http://megasun.bch.umontreal.ca/gobase/gobase.html

SOFTWARE DISTRIBUTION

Software developed by us is available free of charge. The programs together with installation instructions can be downloaded as binary files from the FTP server via anonymous file transfer (ftp://megasun.bch.umontreal.ca/pub).

ACKNOWLEDGEMENTS

The Organelle Genome Database Project (GOBASE) has been supported by a grant (GO-12984) from the Canadian Genome Analysis and Technology Program (CGAT).

REFERENCES

1 Gray,M.W, Lang,B.F., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M., Delage,E., Brossard,N., Littlejohn,T., *et al.* (1998) *Nucleic Acids Res*., in press.
2 The Organelle Genome Megasequencing Program. OGMP URL: http://megasun.bch.umontreal.ca/ogmp/ogmp.html
3 Lang,B.F. (1997). The Fungal Mitochondrial Genome Project. FMGP URL: http://megasun.bch.umontreal.ca/People/lang/FMGP/FMGP.html
4 Calò,D., De Pascali,A., Sasanelli,D, Tanzariello,F., Ponzetta,M.T., Saccone,C. and Attimonelli,M. (1997) *Nucleic Acids Res*. **25**, 200–205 [see also this issue (1998) *Nucleic Acids Res*. **26**, 120–125]. MmtDB URL: http://area.ba.cnr.it/~areamt08/MmtDBWWW.htm
5 Kogelnik,A.M., Lott,M.T., Brown,M.D., Navathe,S.B. and Wallace,D.C. (1997) *Nucleic Acids Res*. **25**, 196–199 [see also this issue (1998) *Nucleic Acids Res*. **26**, 112–115]. Mitomap URL: http://www.gen.emory.edu/mitomap.html
6 Zullo,S. (1997) MitoDat: URL: http://www-lmmb.ncifcrf.gov:80/mitoDat/
7 Mitbase URL: http://www.ebi.ac.uk/htbin/Mitbase/mitbase.pl
8 Harger,C., Skupski,M., Allen,E., Clark,C., Crowley,D., Dickenson,E., Easley,D., Epinosa-Lujan,A., Farmer,A. and Fields,C. (1997) *Nucleic Acids Res*. **25**, 18–23 [see also this issue (1998) *Nucleic Acids Res*. **26**, 21–26]. GSDB URL: http://www.ncgr.org/gsdb/
9 TIGR database. TDB URL: http://www.tigr.org/tigr_home/tdb/tdb.html
10 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res*. **25**, 1–6 [see also this issue (1998) *Nucleic Acids Res*. **26**, 1–7]. Entrez URL: http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
11 Sequence Retrieval System. SRS URL: http://mcbi-34.med.nyu.edu/srs/srsc
12 NCBI Taxon database. URL: http:///www.ncbi.nlm.nih.gov/Taxonomy/tax.html
13 Gutell,R.R. (1994) *Nucleic Acids Res.* **22**, 3502–3507. URL: http://pundit.colorado.edu:8080/RNA/16S/mitochondria.html
14 Gutell,R.R., Gray, M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.* **21**, 3055–3074. URL: http://pundit.colorado.edu:8080/RNA/23S/mitochondria.html
15 Protist Image Database. PID URL: http://megasun.bch.umontreal.ca/protists/protists.html
16 CPGN (1994) *Plant Mol. Biol. Reporter* **12**, S81–S88. Mendel URL: http://probe.nal.usda.gov:8000/plant/aboutmendel.html
17 Enzym URL: http://teosinte.agron.missouri.edu/enzyme.html
18 Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1997) *Nucleic Acids Res*. **25**, 43–50 [see also this issue (1998) *Nucleic Acids Res*. **26**, 50–53]. EcoCyc URL: http://www.ai.sri.com/ecocyc/ecocyc.html
19 WIT URL: http://www.cme.msu.edu/WIT
20 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,D.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L.-S.L. *et al.* (1997) *Nucleic Acids Res*. **25**, 24–28 [see also this issue (1998) *Nucleic Acids Res*. **26**, 27–32]. PIR URL: http://www.psc.edu/general/software/packages/nbrf-pir/nbrf.html
21 Bairoch,A. and Apweiler.R. (1997) *Nucleic Acids Res*. **25**, 31–36 [see also this issue (1998) *Nucleic Acids Res*. **26**, 38–42]. SWISS-PROT URL: http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
22 The FlyBase Consortium (1997) *Nucleic Acids Res*. **25**, 63–66 [see also this issue (1998) *Nucleic Acids Res*. **26**, 85–88]. FlyBase URL: http://flybase.bio.indiana.edu/
23 Letovsky,S. (1994) Web/Genera. URL: http://gdbdoc.gdb.org/letovsky/wgen.html
24 Brossard,N. And Rioux,P. (1997) POP2 - bionet.announce, bionet.software, Message-ID: 34830D6C.167eB0E7@bch.montreal.edu
25 Rioux,P., Gilbert,W.A. and Littlejohn,T.G. (1994) *J. Comp. Biol*. **1**, 293–295.
26 OGMP (1995) The mastefile format. URL: http://megasun.bch.umontreal.ca/People/lang/ogmp-mf/intro.html
27 Rioux,P. and Brossard,N. (1997) MKGOFORMS - bionet.announce, bionet.software, Message-ID: 347B32AA.167EB0E7@bch.umontreal.ca