The Enhanced Microbial Genomes Library

Guy Perrière*, Philippe Bessières¹ and Bernard Labedan²

Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard – Lyon 1, 43, boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France, ¹Laboratoire de Génétique Microbiennne, Département de Microbiologie, Institut National de la Recherche en Agronomie, 78350 Jouy-en-Josas, France and ²Institut de Génétique et Microbiologie, Université Paris-Sud, Bâtiment 409, 91405 Orsay Cedex, France

Received August 24, 1998; Revised October 13, 1998; Accepted October 21, 1998

ABSTRACT

Since the obtention of the complete sequence of Haemophilus influenzae Rd in 1995, the number of bacterial genomes entirely sequenced has regularly increased. A problem is that the quality of the annotations of these very large sequences is usually lower than those of the shorter entries encountered in the repository collections. Moreover, classical sequence database management systems have difficulties in handling entries of that size. In this context, we have decided to build the Enhanced Microbial Genomes Library (EMGLib) in which these two problems are alleviated. This library contains all the complete genomes from bacteria already sequenced and the yeast genome in GenBank format. The annotations are improved by the introduction of data on codon usage, gene orientation on the chromosome and gene families. It is possible to access EMGLib through two database systems set up on World Wide Web servers: the PBIL server at http://pbil.univ-lyon1.fr/emglib/emglib.html and the MICADO server at http://locus.jouy.inra.fr/micado

INTRODUCTION

The obtention of the complete genome of Haemophilus influenzae Rd in 1995 (1) opened a new era in molecular biology and particularly in the bioinformatic field. From this date, sequence database management systems and analysis programs would have to deal with complete genomes and not only with genomic fragments. The main problem of these genomes is their relative lack of annotations, compared with the "classical" sequences. This is due, of course, to the difference in scale and to the automation of sequence obtention, but also to an important deficit in biological information. Even if some efforts have been made in parallel to develop computer systems able to automate some parts of the annotation procedure (2), the level of information available for a single gene in a complete genome is usually lower than in shorter entries. Moreover, some database management systems, which were developed in the 1980s, are not able to handle sequences with a length >300 kb. Due to this fact, complete genomes are always split into smaller, overlapping entries of ~100 kb in the repository databases (3–5).

In this context, on the basis of the work our groups have performed during the Bacillus subtilis genome sequencing project (6–8), we decided to develop an Enhanced Microbial Genome Library (EMGLib). This library contains all bacterial genomes completely sequenced and the yeast genome with some improvements in their annotations. It is possible to access EMGLib through two databases set up on World Wide Web servers that allow efficient querying of genome sequences. These two systems provide easy access to the sequences and their annotations but also to complementary data such as genetic maps.

DATABASE CONTENT

The sequences, except the one corresponding to the B.subtilis genome, are taken from the genome division of GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes 1). In the case of B.subtilis, we used the sequence from the NRSub database (8). Release 1.1 (October 1998) of EMGLib contains 31 entries, these sequences total 44 040 174 bases and they allow access to 36 226 CDS, 943 rRNA, 132 tRNA, 25 snRNA and 20 other miscellaneous RNA.

We have performed many additions and corrections on the original GenBank genome entries. First, new identifiers are given for each genome (LOCUS field). Their names are based on the format CGxxxx (for bacteria) or SCCHRxxxx (for yeast). In the case of bacteria, xxxx stands for an abbreviation of the systematic name of the organism (e.g., BACSUCG is the name of the B.subtilis genome entry). In the case of yeast, xxxx stands for the chromosome number in roman numerals (e.g., SCCHRIX for chromosome IX). We also change the GenBank accession number(s) to use our own number(s), which are based on the format CGxxxx.

Features for CDS are completed with various information (Fig. 1). If the location of the replication origin and terminus are known or could be predicted (9), we add the orientation of the CDS on the chromosome (leading or lagging) under a '/strand' qualifier. Then we add a Codon Adaptation Index (CAI) (10) value under a '/CAI' qualifier and a cross-reference pointing to the corresponding entry in SWISS-PROT/TrEMBL (11) under a '/db_xref' qualifier. We rewrite or complete the 'product' qualifier using data from SWISS-PROT. In the case of ORFs, we use 'hypothetical protein' for the product associated with these putative genes. When the encoded protein is an enzyme, we add its EC number.

*To whom correspondence should be addressed. Tel: +33 472 44 62 96; Fax: +33 478 89 27 19; Email: perriere@biomserv.univ-lyon1.fr
taken from the ENZYME database (12), in an ‘/EC_number’ qualifier. Finally, when the gene was known to belong to a family defined in the HOBACGEN database (13), we added the accession number of this family in a ‘/gene_family’ qualifier.

**DATABASE ACCESS**

The easiest way to access EMGLib is through two databases installed on World Wide Web servers that provide graphical interfaces and cover complementary aspects. The first one is hosted at the Bioinformatic Pole of Lyon (PBIL) and its URL is http://pbil.univ-lyon1.fr/emglib/emglib.html. Its main purpose is to provide a powerful retrieval system for gene sequences in a way to compose complex queries. Indeed, on this server EMGLib is indexed with the ACNUC sequence database management system (14). This system allows us to index all collections in EMBL/GenBank/DDBJ, SWISS-PROT/TrEMBL or NBRF/PIR (15) formats. Under ACNUC, each CDS or structural RNA gene can be seen as an independent sequence, and keywords corresponding to the contents of the different qualifiers are attached to these subsequences.

The home page of the server gives access to entry points allowing one to make simple or complex queries. Simple queries are made by keyword, sequence name and accession number. More sophisticated access is possible through WWW-Query, a general interface for the ACNUC databases installed on this server (16). WWW-Query permits the selection of sequences using different criteria like entry names, accession numbers, keywords, bibliographic references, dates of insertion in the bank, or the nature of the molecule sequenced (e.g., DNA, mRNA, tRNA, etc.). It is possible to combine many criteria using logical operators, and to use the results of previous queries to build new ones. Each kind of feature can be accessed and extracted as well as its flanking regions, and CDS can be translated into proteins. Different formats are available for extracting sequences including FASTA, MASE and GenBank. The server also gives access to various documents such as release notes, a history of the database, an on-line documentation, a list of keywords and a list of all the protein genes accessible in EMGLib. This list includes the cross-references to the corresponding proteins in SWISS-PROT.

It is also possible to install a local copy of the ACNUC version of EMGLib. This version is available at ftp://pbil.univ-lyon1.fr/pub/emglib and it requires the graphical retrieval system Query_win to run. This program is written in C and uses the Vibrant library, which is a part of the toolbox distributed by the National Center for Biotechnology Information (NCBI). Binaries of Query_win are available through our server for UNIX workstations (Sun, DEC Alpha, IBM RISC, HP/UX, Silicon Graphics), and for all microcomputers under MacOS 7.x/8.x and Windows 95/98 operating systems. Query_win integrates the same functionalities as the WWW-Query server plus more sophisticated options (like keywords and species browsing and keywords projections on sequences).

The second database is hosted at the INRA Microbial Genetics Laboratory, and its URL is http://locus.jouy.inra.fr/micado. MICADO is a relational database devoted to microbial genomes (6), on the WWW since 1994, it is now accessed 10 000 times a week. This horizontally integrates all DNA sequences information for archaea and eubacteria, including complete genomes, with genetic maps of *B.subtilis* (7) and *Escherichia coli* (17,18).

As MICADO is the data repository of the functional analysis of *B.subtilis* unknown genes, a vertical integration of information is achieved for the bacterium. Physiological data concerning the disruption of 1200 genes is actually collected in the database by a European consortium of 17 laboratories (19), and linked to metabolic pathways classifications, DNA sequence and the genetic map. Information is searched through WWW by text annotations (from DNA features to authors), sequence comparisons (BLAST and FASTA), browsing classification trees (metabolism and taxa), and finally by navigating genome maps.

A graphical navigation on genome maps has been programmed to offer selective information retrieval on large-scale data sets, thus providing global overview and easy access to genome information. The maps have hierarchical relationships; they display chromosomal information at different scales, from the chromosome and the genetic map, to featured physical maps, and finally text of the sequences and their annotations (Fig. 2). Two versions of the graphical interface to MICADO are now accessible on WWW, the most complete and reliable in Perl 5 language, running on the WWW server, and derived from this Perl version, a new one in Java (20). The latter, running on the client side, brings more interactivity to users, and allows multi-windowing, with synchronized browsing of map representations at different scales. The Java client communicates with the database through an object server in C++, and uses the CORBA application interface protocol. This new standard is relevant to our growing needs, especially for future extensions. It is providing an interesting alternative to physical integration, by allowing us to establish virtual federations of shared databases.

**PERSPECTIVES**

In the short term we want to continue the process of annotations improvement by systematically checking gene names and the functions associated to their products, as there are often historical assignations that have been proven to be wrong in light of the complete genome obtention. We want also to assign functions to hypothetical proteins as the proportion of these proteins in completely sequenced bacterial genomes is high (around 30%). This can be done using data from comparative genomics provided by the HOBACGEN and COLIPAGE (http://www-colipage.igmors.u-psud.fr) databases which we want to introduce in EMGLib.

In the long run, as the MICADO system is able to manage different sources of data due to its intrinsic extension capabilities, we want to integrate in EMGLib many other aspects of bacterial molecular biology. We have already planned to include the metabolic classification of genes and functional analysis data for...
some species (like *B. subtilis*). At last, data on comparative genomics more complex than simple cross-references to specialized databases will be introduced. This will be the case of different data compiled in COLIPAGE where a lot of information has been assembled about the paralogous proteins and especially their constitutive segments of homology (modules) (21,22). These evolutionary data will complete the information on gene families obtained from HOBACGEN and will also be useful for people interested in understanding protein evolution.

REFERENCES