HOBACGEN: Database System for Comparative Genomics in Bacteria

Guy Perrière,[1] Laurent Duret, and Manolo Gouy

Laboratoire de Biométrie et Biologie Évolutive, Unité Mixte de Recherche Centre National de la Recherche Scientifique (UMR CNRS) n°. 5558, Université Claude Bernard–Lyon 1, 69622 Villeurbanne Cedex, France

We present here HOBACGEN, a database system devoted to comparative genomics in bacteria. HOBACGEN contains all available protein genes from bacteria, archaea, and yeast, taken from SWISS-PROT/TrEMBL and classified into families. It also includes multiple alignments and phylogenetic trees built from these families. The database is organized under a client/server architecture with a client written in Java, which may run on any platform. This client integrates a graphical interface allowing users to select families according to various criteria and notably to select homologs common to a given set of taxa. This interface also allows users to visualize multiple alignments and trees associated to families. In tree displays, protein gene names are colored according to the taxonomy of the corresponding organisms. Users may access all information associated to sequences and multiple alignments by clicking on genes. This graphic tool thus gives a rapid and simple access to all data required to interpret homology relationships between genes and distinguish orthologs from paralogs. Instructions for installation of the client or the server are available at http://pbil.univ-lyon1.fr/databases/hobacgen.html.

Comparative genomics has become a common approach in sequence analysis, particularly with the advent of bacterial genome sequencing projects. The most common use of this technique is represented by similarity searches performed on the CDSs (coding DNA sequences) revealed by prediction tools. If homologs to these CDSs are found in the sequence databases, this is a hint that the corresponding protein exists. Another use is functional analysis; it is possible to assign a function to a protein or to detect its functional regions by homology to other sequences. The study of structural constraints can also be approached in this way. For example, secondary structure prediction methods now extensively use multiple alignments to refine their results (Cuff and Barton 1999). Several molecular phylogenetic studies also rely on comparative genomics: search for lateral gene transfers (Nelson et al. 1999), determination of metabolic pathways specific to some taxa, or ancestral genome content estimation (Mushegian and Koonin 1996).

Among sequences that are homologous over their entire length, one has to distinguish orthologous sequences (i.e., sequences that have diverged after a speciation event), from paralogous sequences (i.e., sequences that have diverged after duplication of an ancestral gene). This distinction is essential for molecular phylogeny as it is necessary to work with orthologous genes to infer species phylogeny from gene phylogeny. This distinction is also important to predict the function of a new gene or to search for functionally conserved regions by comparative analysis, because paralogous genes, even if closely related, may have different functions or regulations. The distinction between orthologous and paralogous genes requires a careful analysis of homologous sequences (Duret et al. 1994). It is not possible to rely on database definitions to identify orthologous genes, because many sequences are not annotated, and when annotations are present they are sometimes inexplicit or inaccurate. Thus, some paralogous genes may have very similar definitions, whereas orthologous genes may be given totally different names.

The process of homology determination is a complex task involving different steps. First, it is necessary to use sequence similarity criteria to search for gene families. Then, it is required to align the sequences belonging to a family and to compute the corresponding phylogenetic tree. During all this process an access to sequence database annotations and taxonomic data is needed. Some systems contain information on the similarities observed between the different proteins introduced in the general databases: the retrieval system Entrez (McEntyre 1998) allows one to select the "neighbors" of a sequence, and the MIPS database (Mewes et al. 2000) as well as the ProtoMap system (Yona et al. 2000) integrate gene families with their corresponding alignments and cladograms. The problem is that these systems do not allow one to distinguish easily between orthologous and paralogous genes. Also, they only use WWW (World-Wide Web) interfaces that lack interactivity.

To answer these problems, we decided to build the HOBACGEN (HOmologous BACterial GENes) database. This system contains all available protein sequences from bacteria, archaea, and yeast organized into families of homologous genes determined by similarity. Multiple alignments as well as phylogenetic

[1] **Corresponding author.**
E-MAIL perriere@biomserv.univ-lyon1.fr; FAX 33 478-89-27-19.

10:379–385 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

trees and taxonomic information are provided with these families that help identify orthologs. A dedicated client system with a graphical interface allows users to access the database and to exploit its data.

RESULTS

Since February 1999 the HOBACGEN interface and data are publicly available at http://pbil.univ-lyon1.fr/databases/hobacgen.html. The system is built under a client/server architecture. This kind of organization avoids the need to install the complete database on the users' computers. This also prevents the users from completing the tedious and time-consuming updating operations usually required with sequence databases. It is also possible to download the full set of files and index tables for installing a local server.

Client

The client, named HobacFetch, is a Java application. The choice of this language allows the portability of the program on any computer for which a Java virtual machine (v. 1.1.3 or higher) is available (i.e., all common platforms). HobacFetch was developed as an application to avoid known problems with applets: incompatibilities between Java versions and Internet browsers, necessity to build certified applets in a way to authorize read/write access on users' disks, and slowness of loading even with fast access network.

Starting from the main window of the interface it is possible to access the entire list or a personal subset of families and make queries to retrieve those matching specific criteria (Fig. 1). For instance users may select families containing a given number of genes or taxa. Also the user may select sequences from SWISS-PROT/TrEMBL that fit particular criteria using any of the retrieval system available on the net, and then get the families containing these genes in HOBACGEN. In this case, genes belonging to the list used to make the selection are highlighted in the tree window. Searches using taxa crossing permits one to retrieve all the families containing at least one representative of a list of taxa entered by the user. Any taxonomic level and taxa number can be used to compose the query. For example, it is possible to retrieve all families containing at least one gene from *Bacillus subtilis, Escherichia coli,* and an archaea.

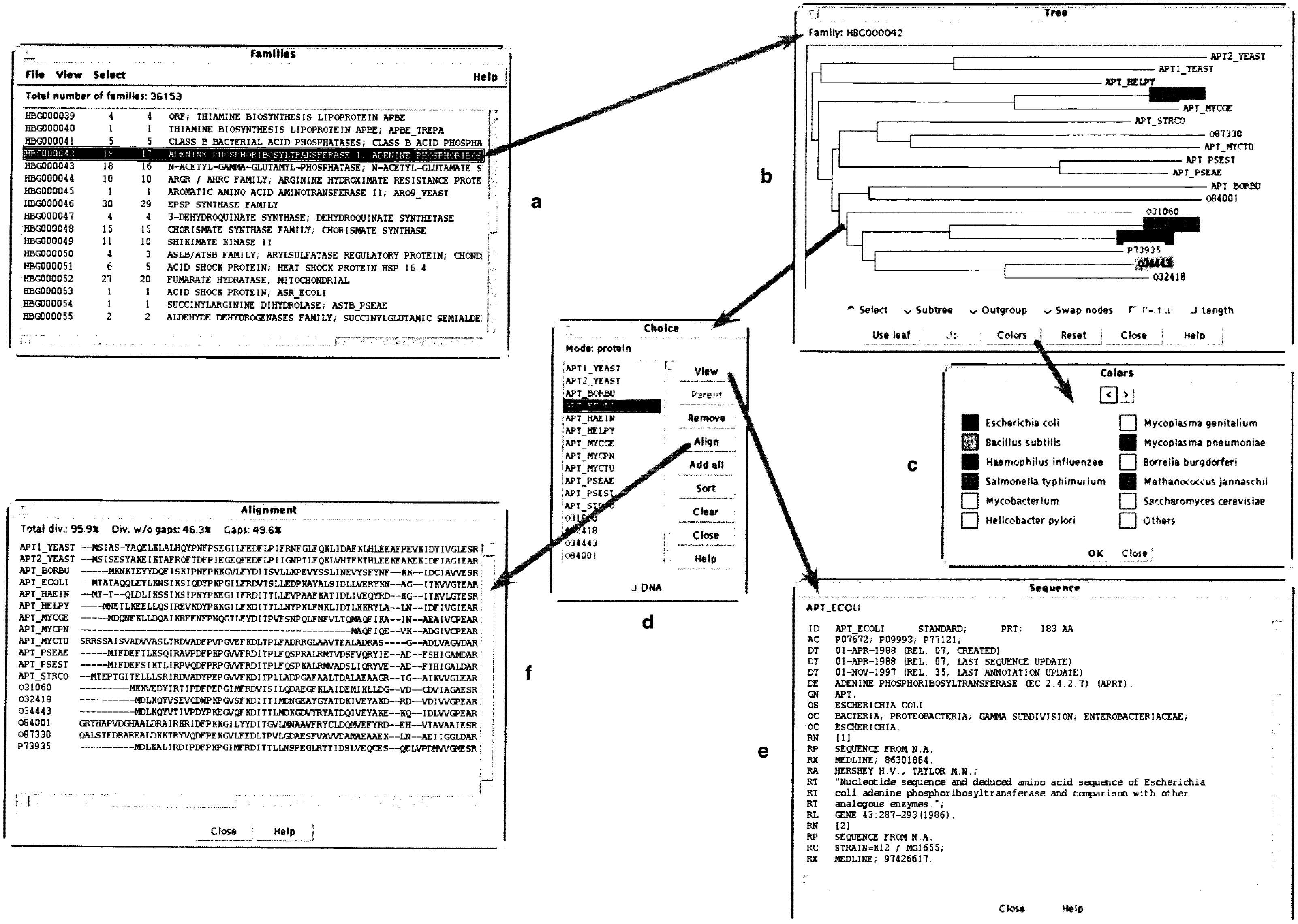
After the selection of a family, the corresponding phylogenetic tree is displayed in the tree window (Fig. 1). In this tree, sequences are colored using a code related to the taxonomic position of the species from which they have been isolated. This code helps to identify paralogs in a family. Four color sets are provided, each of them may be edited by the user who may change the taxon name or its associated color. The tree display is active, with options of rerooting, node swapping, or subtree selection. For subtree selection, the user may give a sequence name and a depth in a way to display only the part of the tree that starts from the leaf corresponding to the sequence and that goes up to a number of nodes equal to the depth entered. This possibility is particularly useful when browsing through large families. Clicking on leaves allows the user to visualize the entries from SWISS-PROT/TrEMBL and EMBL or the multiple alignments of the selected sequences. For nucleotide sequences, as a result of the important redundancy found in EMBL, frequently there is more than one CDS associated to a SWISS-PROT/TrEMBL entry. In that case the user has the possibility of selecting one of the CDSs linked to a protein to visualize it. Alignments between selected sequences are not computed but reconstructed from the preexisting whole family multiple alignment. The divergence with or without gaps and the gap frequency are calculated, thus giving a measure of similarity between aligned sequences.

Functions allow the user to save lists of families, sequence entries, alignments, or trees in text files. The formats used are those of SWISS-PROT and EMBL for the sequences, PHYLIP for the trees, and CLUSTAL for the alignments. All of these formats are recognized standards and they can be read by most sequence analysis packages. Saved lists of families can be loaded instead of using the complete list downloaded from the server. This may save time when the user is only interested in a subset of families and when the remote server is distant from the client.

Server

The server side is made of three components: a WWW service, a dedicated C program to access the data, and the database itself. A WWW service must run on the computer serving HOBACGEN because all client/server transactions use http (HyperText Transfer Protocol). All requests sent by the client are handled by a WWW daemon able to decode them and to start cgi (common gateway interface) scripts. The second component of the server is a program, which is started by the daemon as a cgi script. It is written in standard ANSI C and may be installed on any Unix computer. This program reads data (sequences, alignments, and trees) and transmits them to the WWW daemon, which sends them back to the client. The data consist of two ACNUC databases containing, respectively, the SWISS-PROT/TrEMBL and the EMBL sequences. Annotations of the entries themselves are slightly modified to incorporate complementary data related to families and protein domains. For each protein or CDS, we add a reference to which the HOBACGEN family number it belongs. For protein sequences, we also add data on the localization of ProDom domains (Corpet et al. 2000) that are found in a given entry.

From the 109,077 proteins extracted from SWISS-



Total number of families: 36153

```
HBG000039    4    4   ORF; THIAMINE BIOSYNTHESIS LIPOPROTEIN APBE
HBG000040    1    1   THIAMINE BIOSYNTHESIS LIPOPROTEIN APBE; APBE_TREPA
HBG000041    5    5   CLASS B BACTERIAL ACID PHOSPHATASES; CLASS B ACID PHOSPHA
HBG000042   18   17   ADENINE PHOSPHORIBOSYLTRANSFERASE 1; ADENINE PHOSPHORIBOS
HBG000043   18   16   N-ACETYL-GAMMA-GLUTAMYL-PHOSPHATASE; N-ACETYL-GLUTAMATE S
HBG000044   10   10   ARGR / AHRC FAMILY; ARGININE HYDROXIMATE RESISTANCE PROTE
HBG000045    1    1   AROMATIC AMINO ACID AMINOTRANSFERASE II; ARO9_YEAST
HBG000046   30   29   EPSP SYNTHASE FAMILY
HBG000047    4    4   3-DEHYDROQUINATE SYNTHASE; DEHYDROQUINATE SYNTHETASE
HBG000048   15   15   CHORISMATE SYNTHASE FAMILY; CHORISMATE SYNTHASE
HBG000049   11   10   SHIKIMATE KINASE II
HBG000050    4    3   ASLB/ATSB FAMILY; ARYLSULFATASE REGULATORY PROTEIN; CHOND
HBG000051    6    5   ACID SHOCK PROTEIN; HEAT SHOCK PROTEIN HSP.16.4
HBG000052   27   20   FUMARATE HYDRATASE, MITOCHONDRIAL
HBG000053    1    1   ACID SHOCK PROTEIN; ASR_ECOLI
HBG000054    1    1   SUCCINYLARGININE DIHYDROLASE; ASTB_PSEAE
HBG000055    2    2   ALDEHYDE DEHYDROGENASES FAMILY; SUCCINYLGLUTAMIC SEMIALDE
```

Total div.: 95.9% Div. w/o gaps: 46.3% Caps: 49.6%

```
APT1_YEAST  --MSIAS-YAQELKLALHQYPNFPSEGILFEDFLPIFRNFGLFQKLIDAFKLHLEEAFPEVKIDYIVGLESR
APT2_YEAST  --MSISESYAKEIKTAFRQFTDFPIEGEQFEDFLPIIGNPTLFQKLVHTFKTHLEEKFAKEKIDFIAGIEAR
APT_BORBU   -----MKNKTEYYDQFISKIPNFPKKGVLFYDITSVLLKPEVYSSLINEVYSFYNF--KK--IDCIAVVESR
APT_ECOLI   --MTATAQQLEYLKNSIKSIQDYPKPGILFRDVTSLLEDPKAYALSIDLLVERYKN--AG--IIKVVGIEAR
APT_HAEIN   --MT-T--QLDLIKSSIKSIPNYPKEGIIFRDITTLLEVPAAFKATIDLIVEQYRD--KG--IIKVLGTESR
APT_HELPY   ----MNETLKEELLQSIREVKDYPKKGILFKDITTLLNYPKLFNKLIDTLKKRYLA--LN--IDFIVGIEAR
APT_MYCGE   -----MDQNFKLLDQAIKRFENFPNQGTLFYDITPVFSNPQLFNFVLTQMAQFIKA--IN--AEAIVCPEAR
APT_MYCPN   ---------------------------------------------------MAQFIQE--VK--ADGIVCPEAR
APT_MYCTU   SRRSSAISVADVVASLTRDVADFPVPGVEFKDLTPLFADRRGLAAVTEALADRAS----G--ADLVAGVDAR
APT_PSEAE   -----MIFDEFTLKSQIRAVPDFPKPGVVFRDITPLFQSPRALRMTVDSFVQRYIE--AD--FSHIGAMDAR
APT_PSEST   -----MIFDEFSIKTLIRPVQDFPRPGVVFRDITPLFQSPKALRMVADSLIQRYVE--AD--FTHIGALDAR
APT_STRCO   --MTEPTGITELLLSRIRDVADYPEPGVVFKDITPLLADPGAFAALTDALAEAAGR--TG--ATKVVGLEAR
031060      ---------MKKVEDYIRTIPDFPEPGIMFRDVISILQDAEGFKLAIDEMIKLLDG--VD--CDVIAGAESR
032418      ----------MDLKQYVSEVQDWPKPGVSFKDITTIMDNGEAYGYATDKIVEYAKD--RD--VDIVVGPEAR
034443      ----------MDLKQYVTIVPDYPKEGVQFKDITTLMDKGDVYRYATDQIVEYAKE--KQ--IDLVVGPEAR
084001      GRYHAPVDGHAALDRAIRKRIDFPKKGILYYDITGVLMNAAVFRYCLDQMVEFYRD--EH--VTAVAAIESR
087330      QALSTFDRAREALDKKTRYVQDFPEKGVLFEDLTPVLGDAESFVAVVDAMAEAAEK--LN--AEIIGGLDAR
P73935      ----------MDLKALIRDIPDFPKPGIMFRDITTLLNSPEGLRYTIDSLVEQCES--QELVPDHVVGMESR
```

APT_ECOLI

```
ID   APT_ECOLI      STANDARD;      PRT;   183 AA.
AC   P07672; P09993; P77121;
DT   01-APR-1988 (REL. 07, CREATED)
DT   01-APR-1988 (REL. 07, LAST SEQUENCE UPDATE)
DT   01-NOV-1997 (REL. 35, LAST ANNOTATION UPDATE)
DE   ADENINE PHOSPHORIBOSYLTRANSFERASE (EC 2.4.2.7) (APRT).
GN   APT.
OS   ESCHERICHIA COLI.
OC   BACTERIA; PROTEOBACTERIA; GAMMA SUBDIVISION; ENTEROBACTERIACEAE;
OC   ESCHERICHIA.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 86301884.
RA   HERSHEY H.V., TAYLOR M.W.;
RT   "Nucleotide sequence and deduced amino acid sequence of Escherichia
RT   coli adenine phosphoribosyltransferase and comparison with other
RT   analogous enzymes.";
RL   GENE 43:287-293(1986).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   STRAIN=K12 / MG1655;
RX   MEDLINE; 97426617.
```

Figure 1 Organization of the HobacFetch interface. The "Families" window allows one to perform queries and to select a given family (*a*). In the Tree window, the phylogenetic tree associated with the selected family is displayed (*b*). The Colors dialog box shows the correspondence between colors and taxa (*c*). Clicking on a leaf on the Tree window starts the display of the Choice dialog box (*d*), from which it is possible to view one of the selected entry from SWISS-PROT/TrEMBL or EMBL (*e*) or the alignment of all selected sequences (*f*).

PROT/TrEMBL, a total of 71,530 (65%) were classified into 12,780 families containing at least two sequences; 32,118 proteins are unique in their family (29%), and there are 5,429 partial sequences not attached to a family (5%). The distribution of families according to the number of sequences is presented in Table 1 and the 10 largest families are reported in Table 2. When looking at the families common to the different kingdoms, we found that 190 families were common to bacteria, archaea, and yeast. Not surprisingly these families contain mainly genes coding for proteins involved in protein translation (e.g., ribosomal proteins and amino-acyl-tRNA synthetases), nucleotide biosynthesis, and glycolysis.

Example of Use

We will give here an example of HOBACGEN use, linked to the detection of annotation and sequence errors. This example was taken from the family of the 6PGD (6-phosphogluconate dehydrogenase) enzyme. Among the proteins included in this family, two were paralogs in *B. subtilis* and their entry names were 6PGD_BACSU (accession no. P12013) and YQJI_BACSU (accession no. P80859). The protein annotated in SWISS-PROT (up to release 37) as corresponding to the 6PGD in *B. subtilis* was 6PGD_BACSU, whereas YQJI_BACSU was annotated as hypothetical protein. The part of the phylogenetic tree containing these two *B. subtilis* sequences is shown in Figure 2. Judging from this tree, YQJI_BACSU is closer to the 6PGD_ECOLI (accession no. P00350) obtained from *E. coli* than 6PGD_BACSU. What is interesting is that the function of 6PGD_BACSU was not determined experimentally but assigned by similarity to other 6PGD, particularly to the 6PGD_ECOLI sequence (Reizer et al. 1991). When this assignation was made, the gene corresponding to the YQJI_BACSU entry was not sequenced yet. It means that we may have here a mistaken function assignation, but the determination of which of these two proteins really is 6PGD requires laboratory experiments because sequence similarities alone are not sufficient to close the case.

Table 1. Distribution of Families (of Size >1) According to Their Number of Protein Sequences

Number		
proteins	families	Frequency (%)
2	5,964	46.7
3–4	3,181	24.9
5–9	1,987	15.5
10–19	906	7.1
20–39	491	3.8
40–99	204	1.6
≥100	47	0.4
Total	12,780	100

Table 2. HOBACGEN 10 Largest Families

Family name	No. of proteins
ABC transporters	874
Bacterial flagellin	559
NifH/FrxC family	439
Outer membrane protein (OmpC/OmpK)	390
LuxR/UhpA transcriptional regulators	338
Porin	338
RNA polymerase σ^{70} factor	314
Type II topoisomerases	305
Short-chain dehydrogenases/reductases	293
Major outer membrane lipoprotein	291
Total	4,141

The alignment of amino-terminal region of YQJI_BACSU with 6PGD_BACSU, 6PGD_ECOLI, and another protein belonging to the same family, 6PGD_SALTY (accession no. P14062) is shown in Figure 3. The alignment shows a large deletion of 63 amino acids in the amino terminus of YQJI_BACSU. This deletion is observed in none of the other proteins of this family. But translation in one of the three possible frames of the DNA region located up to 190 residues before the initiation codon documented in the complete genome sequence presents a protein sequence that matches the

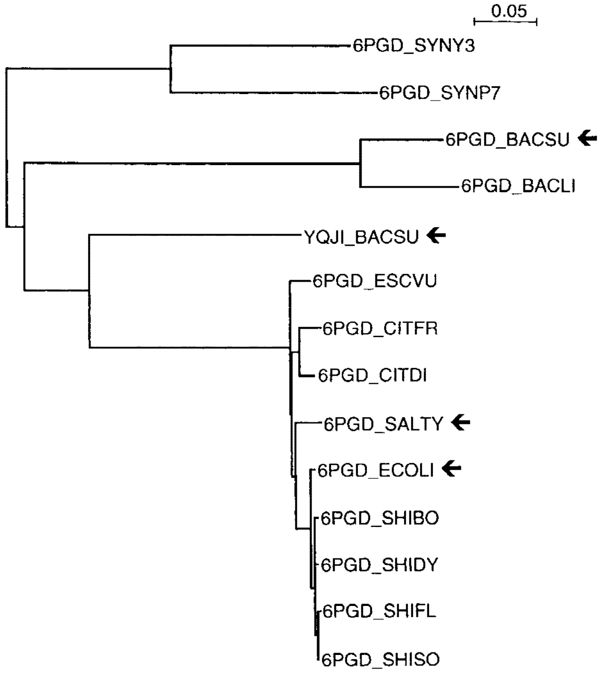
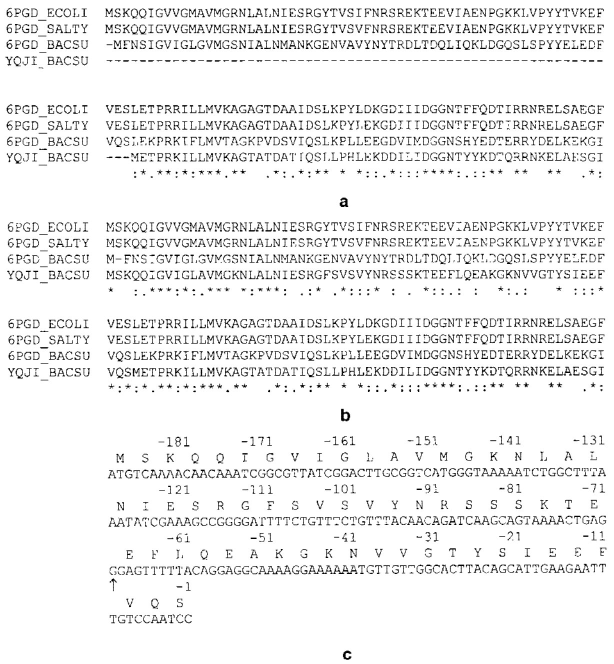


Figure 2 Part of the HOBACGEN tree for the 6PGD (6-phosphogluconate dehydrogenase) family. The tree has been rooted with two cyanobacterial sequences (6PGD_SYNY3 and 6PGD_SYNP7). The four sequences for which the multiple alignment is shown in Fig. 3 are marked with arrows.



```
6PGD_ECOLI   MSKQQIGVVGMAVMGRNLALNIESRGYTVSIFNRSREKTEEVIAENPGKKLVPYYTVKEF
6PGD_SALTY   MSKQQIGVVGMAVMGRNLALNIESRGYTVSVFNRSREKTEEVIAENPGKKLVPYYTVKEF
6PGD_BACSU   -MFNSIGVIGLGVMGSNIALNMANKGENVAVYNYTRDLTDQLIQKLDGQSLSPYYELEDF
YQJI_BACSU   ------------------------------------------------------------

6PGD_ECOLI   VESLETPRRILLMVKAGAGTDAAIDSLKPYLDKGDIIIDGGNTFFQDTIRRNRELSAEGF
6PGD_SALTY   VESLETPRRILLMVKAGAGTDAAIDSLKPYLEKGDIIIDGGNTFFQDTIRRNRELSAEGF
6PGD_BACSU   VQSLEKPRKIFLMVTAGKPVDSVIQSLKPLLEEGDVIMDGGNSHYEDTERRYDELKEKGI
YQJI_BACSU   ---METPRKILLMVKAGTATDATIQSLLPHLEKDDILIDGGNTYYKDTQRRNKELAESGI
                :*.**:*:***.**  .*:.*:** * *::.*:::****:.::** **  ** .*:
```

a

```
6PGD_ECOLI   MSKQQIGVVGMAVMGRNLALNIESRGYTVSIFNRSREKTEEVIAENPGKKLVPYYTVKEF
6PGD_SALTY   MSKQQIGVVGMAVMGRNLALNIESRGYTVSVFNRSREKTEEVIAENPGKKLVPYYTVKEF
6PGD_BACSU   M-FNSIGVIGLGVMGSNIALNMANKGENVAVYNYTRDLTDQLIQKLDGQSLSPYYELEDF
YQJI_BACSU   MSKQQIGVIGLAVMGKNLALNIESRGFSVSVYNRSSSKTEEFLQEAKGKNVVGTYSIEEF
             *  :.***:*:.*** *:***: .:* .*:::* : . *::.: :  *:.:   * :::*

6PGD_ECOLI   VESLETPRRILLMVKAGAGTDAAIDSLKPYLDKGDIIIDGGNTFFQDTIRRNRELSAEGF
6PGD_SALTY   VESLETPRRILLMVKAGAGTDAAIDSLKPYLEKGDIIIDGGNTFFQDTIRRNRELSAEGF
6PGD_BACSU   VQSLEKPRKIFLMVTAGKPVDSVIQSLKPLLEEGDVIMDGGNSHYEDTERRYDELKEKGI
YQJI_BACSU   VQSMETPRKILLMVKAGTATDATIQSLLPHLEKDDILIDGGNTYYKDTQRRNKELAESGI
             *:*:*.**:*:***.**  .*:.*:** * *::.*:::****:.::** **  ** .*:
```

b

```
    -181      -171      -161      -151      -141      -131
  M  S  K  Q  Q  I  G  V  I  G  L  A  V  M  G  K  N  L  A  L
ATGTCAAAACAACAAATCGGCGTTATCGGACTTGCGGTCATGGGTAAAAATCTGGCTTTA
    -121      -111      -101       -91       -81       -71
  N  I  E  S  R  G  F  S  V  S  V  Y  N  R  S  S  S  K  T  E
AATATCGAAAGCCGGGGATTTTCTGTTTCTGTTTACAACAGATCAAGCAGTAAAACTGAG
     -61       -51       -41       -31       -21       -11
  E  F  L  Q  E  A  K  G  K  N  V  V  G  T  Y  S  I  E  E  F
GGAGTTTTTACAGGAGGCAAAAGGAAAAAATGTTGTTGGCACTTACAGCATTGAAGAATT
↑        -1
  V  Q  S
TGTCCAATCC
```

c

Figure 3 Alignment of the amino-terminal region of the four *B. subtilis, Salmonella typhimurium,* and *E. coli* sequences selected from the tree shown in Fig. 2. In *a,* the original YQJI_BACSU sequence is used; in *b* the corrected sequence is used; *c* shows the translation we have used to correct YQJI_BACSU. This translation starts at 190 nucleotides before the position of the initiation codon given in the sequence annotations of *B. subtilis* genome. The insertion of a G at position −70 (arrow), is responsible of the frameshift that prevented the annotators to find the real initiation codon.

alignment with the four other sequences. It means that a frameshift has been introduced in the CDS used to build the YQJI_BACSU entry. This frameshift is due to insertion of a G at position −70 before the translation start of the gene corresponding to the YQJI_BACSU protein. After that analysis, the corresponding SWISS-PROT entry has been modified according to our information and the entry itself has been renamed 6PG2_BACSU. It is now annotated as a possible 6PGD enzyme.

DISCUSSION

HOBACGEN allows users to rapidly select gene families according to various criteria and notably to select homologs common to a given set of taxa. The color graphical interface provides an easy access to all the data (multiple alignments, phylogenetic trees, taxonomic data, sequence annotations) required to interpret homology relationships between genes and thus to distinguish orthologs from paralogs. Thus, HOBACGEN should be a useful tool for comparative genomics, phylogeny, or molecular evolution studies.

Several other systems include information on homology relationships between sequences. Entrez (McEntyre 1998) shows all significant similarities detected between protein or DNA sequences. However, Entrez does not provide multiple alignments or phylogenetic trees. In the MIPS database (Mewes et al. 2000) and the ProtoMap system (Yona et al. 2000), protein sequences are also classified into families, but with different definitions and objectives. With MIPS, proteins are first grouped in superfamilies on the basis of sequence similarity and functional relationships. Superfamilies are then subdivided in families, subfamilies, or entries on the basis of their similarity level (equal to 50%, 80%, and 95%, respectively). MIPS and ProtoMap are both limited to proteins, and do not give access to the gene sequences like HOBACGEN. Another major problem with MIPS is a lack of integrated and user friendly interface to access the data. As with the majority of sequence databases, access is realized only through a WWW interface. This kind of interface presents several limitations, the most important one being a lack of interactivity. Indeed, if the WWW server for the MIPS database allows users to visualize multiple alignments and dendrograms, display is completely static. For instance, it is not possible to exclude sequences from the multiple alignment or to manipulate the trees (for node swapping, rerooting, or subtree selection) unlike what is possible with HobacFetch. The ProtoMap system presents a slightly better interactivity thanks to a set of Java applets, but this interface still does not permit tree or alignment manipulation. At last, it is not possible to query MIPS or ProtoMap according to taxonomic relationships (e.g., select all homologs common to a given set of taxa).

On the other hand, the MIPS already integrates an access to protein homology domains alignments. This is important as it is now well known that the majority of proteins, even in bacteria, have a modular structure (Patthy 1991, 1994; Riley and Labedan 1997). For each domain a multiple alignment, containing only the homologous region, is furnished. Up to now, even if we have already introduced ProDom data in the sequence annotations of HOBACGEN, it is not possible to access the corresponding alignments and trees. This is why a next step in the development of the interface will be the introduction of the possibility to handle multiple alignments and trees based on ProDom data.

To use HOBACGEN for structural studies based on comparative genomics, data on protein structure are needed. Among the data we plan to introduce are the location of specific regions like transmembrane segments, antigenic sites, or PROSITE signatures (Hofman et al. 1999) and the prediction of secondary structures produced by the SOPMA program (Geourjon and Deléage 1995).

METHODS

HOBACGEN release 6 (December 1999) was built with all protein sequences from bacteria, archaea, and yeast taken from SWISS-PROT 38 and TrEMBL 12 (Bairoch and Apweiler 2000). The choice of SWISS-PROT/TrEMBL to develop our system was guided by the fact that this database is nearly nonredundant. We thus avoided the problem of redundancy declaration and management encountered before with the HOVERGEN database (Duret et al. 1994, 1999). Among the 109,077 proteins used to build this release of HOBACGEN, 85,059 were from bacteria, 14,998 from archaea, and 6873 from yeast. This version of the database integrates protein and DNA sequences of all the genes from 21 complete genomes (Table 3). Thanks to the cross-references put in the DR field of SWISS-PROT/TrEMBL annotations, the corresponding nucleotide sequences from EMBL (Baker et al. 2000) were also integrated in the system structure. Protein and nucleotide sequences were organized into two separated ACNUC databases (Gouy et al. 1985).

To build the families, a similarity search of all proteins against themselves was performed with the BLASTP2 program (Altschul et al. 1997). Low complexity regions were filtered with SEG (Wootton and Fedheren 1996), a threshold of 10^{-4} for E-values and the BLOSUM62 matrix (Henikoff and Henikoff 1992) for amino acid similarity measures were used. BLAST output was filtered in a way to remove homologous segment pairs (HSPs) not compatible with a global alignment (Fig. 4). For complete protein sequences, two sequences in a pair were included in the same family if the remaining HSPs covered ≥80% of the protein length and if their similarity was ≥50% (two amino acids are considered similar if their BLOSUM62 similarity score is positive). We used simple transitive links to build families. Once families of complete protein sequences were built, partial sequences were included in the classification. A partial sequence having similarity with a complete protein was included in a family if it fulfilled the two conditions required for a complete sequence and if its length was ≥100 amino acids or greater or equal to half the length of the complete protein. Also, partial sequences may be included in more than one family.

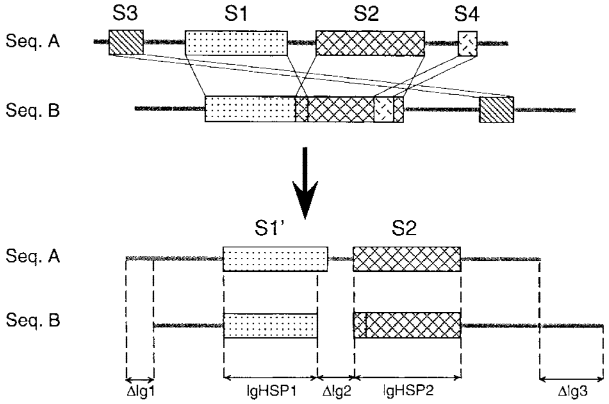


Figure 4 Removal of incompatible HSPs. For each couple of homologous sequences found by BLASTP, HSPs that are not compatible with a global alignment are removed. In this example, segments S1 and S2 are compatible, but not segments S3 and S4. Therefore, they are ignored by further computations on similarity measures that allow one to classify (or not) these two sequences in the same family.

Family names were created using a semiautomated procedure that parsed the DE and SIMILARITY fields of the SWISS-PROT/TrEMBL annotations. The program first created a nonredundant table obtained by merging all DE fields taken from the sequences belonging to a given family. If the length of this table was shorter than a given threshold, it was used to build the family name. If not, the program then created a second table obtained by merging all the SIMILARITY fields. Again, if the length of this table was shorter than a threshold, it was used to build the family name. If this second table was also too long, a manual expertise was performed to set up the name.

Table 3. Number of Proteins, Genomic Sequences and CDSs from Completely Sequenced Genomes That Can Be Accessed in HOBACGEN Release 6

	Number		
Species	**prots.**	**seqs.**	**CDSs**
Aeropyrum pernix	2698	15	2699
Aquifex aeolicus	1550	109	1522
Archaeoglobus fulgidus	2411	185	2419
Bacillus subtilis	4642	1093	9434
Borrelia burgdorferi	2218	821	1686
Chlamydia pneumoniae	1115	148	1104
Chlamydia trachomatis	1436	661	1432
Escherichia coli	8295	5021	16179
Haemophilus influenzae	1989	521	2140
Helicobacter pylori J99	1464	132	1491
Methanobacterium thermoautotrophicum	2071	239	2098
Methanococcus jannaschii	1771	155	1772
Mycobacterium tuberculosis	4089	742	4338
Mycoplasma genitalium	576	392	917
Mycoplasma pneumoniae	701	157	790
Pyrococcus horikoshii	2061	10	2065
Rickettsia prowazekii	847	529	920
Saccharomyces cerevisiae	6873	14702	12770
Synechocystis sp.	3248	151	3378
Thermotoga maritima	1892	201	1979
Treponema pallidum	1077	179	1166

The number of CDSs often exceeds the number of proteins because of the redundancy in the EMBL database.

For each family, multiple sequence alignments and phylogenetic trees were built. Protein sequences were aligned with CLUSTALW 1.7 (Higgins et al. 1996). All default parameters were used except that the "Fast/Approximate" option was preferred for pairwise alignments. Alignments were stored into text files in CLUSTAL format, with one file per family. Phylogenetic trees were built with these alignments using the observed divergence as a measure of distance. When the distance matrix was complete, trees were computed with BIONJ, an improved version of the NJ algorithm (Gascuel 1997). When the matrix was incomplete (i.e., when there were partial sequences in the family that did not overlap with each others), we used a method derived from Lapointe and Kirsch (1995) that estimates missing distances (A. Guénoche, unpubl.). Trees were rooted with the midpoint method that minimizes the differences between branch lengths averages between the two sides of root. They were stored into text files in the standard parenthesized PHYLIP format (Felsenstein 1989), with one file per family.

The alignments provided by HOBACGEN were not manually corrected and trees were built from observed divergence levels between protein sequences. Therefore, although trees in HOBACGEN are efficient to detect paralogies, they may not correspond to exact phylogenetic trees, especially when short branches are considered or when evolutionary rates greatly differ between lineages.

ACKNOWLEDGMENTS

This work was supported by grants from the PRFMMIP (Projet de Recherche Fondamentale pour les Maladies Microbiennes Infectieuses et Pathogènes) and the Genome Program from CNRS.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bairoch, A. and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Baker, W., A. van den Broek, E. Camon, P. Hingamp, P. Sterk, G. Stoesser, and M.A. Tuli. 2000. The EMBL nucleotide sequence database. *Nucleic Acids Res.* **28:** 19–23.

Corpet, F., F. Servant, J. Gouzy, and D. Kahn. 2000. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28:** 267–269.

Cuff, J.A. and G.J. Barton. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34:** 508–519.

Duret, L, D. Mouchiroud, and M. Gouy. 1994. Hovergen: A database of homologuous vertebrate genes. *Nucleic Acids Res.* **22:** 2360–2365.

Duret, L., G. Perrière, and M. Gouy. 1999. Hovergen: Database and software for comparative analysis of homologous vertebrate genes. In *Bioinformatics Databases and Systems* (ed. S.I. Letovsky), pp. 13–29. Kluwer Academic Publishers, Boston, MA.

Felsenstein, J. 1989. PHYLIP: Phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14:** 685–695.

Geourjon C. and G. Deléage. 1995. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* **11:** 681–684.

Gouy, M., C. Gautier, M. Attimonelli, C. Lanave, and G. di Paola. 1985. ACNUC—A portable retrieval system for nucleic acid sequence databases: Logical and physical designs and usage. *Comput. Applic. Biosci.* **1:** 167–172.

Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Higgins, D.G., J.D. Thompson, and T.J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Lapointe, F.J. and J.A.W. Kirsch. 1995. Estimating phylogenies from lacunose distances matrices, with special reference to DNA hybridization data. *Mol. Biol. Evol.* **12:** 266–284.

McEntyre, J. 1998. Linking up with Entrez. *Trends Genet.* **14:** 39–40.

Mewes, H.W., D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schüller et al. 2000. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **28:** 37–40.

Mushegian, A.R. and E.V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci.* **93:** 10268–10273.

Nelson, K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum et al. 1999. Evidence of lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima. Nature* **399:** 323–329.

Patthy, L. 1991. Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1:** 351–361.

———. 1994. Introns and exons. *Curr. Opin. Struct. Biol.* **4:** 383–392.

Reizer, A., J. Deutscher, M.H. Saier, Jr., and J. Reizer. 1991 Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*. *Mol. Microbiol.* **5:** 1081–1089.

Riley, M. and B. Labedan. 1997. Protein evolution viewed through *Escherichia coli* protein sequences: Introducing the notion of structural segment of homology, the module. *J. Mol. Biol.* **269:** 1–12.

Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.

Yona, G., N. Linial, and M. Linial. 2000. ProtoMap: Automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* **28:** 49–55.

Received August 11, 1999; accepted in revised form January 20, 2000.