HOVERGEN: a database of homologous vertebrate genes

Laurent Duret*, Dominique Mouchiroud and Manolo Gouy
Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard, Lyon I, URA-CNRS 243 Bat. 741, 43 Blvd du 11 Novembre 1918, 69622 Villeurbanne cedex, France

Received February 17, 1994; Revised and Accepted May 10, 1994

ABSTRACT

Comparison of homologous genes is a major step for many studies related to genome structure, function or evolution. Similarity search programs easily find genes homologous to a given sequence. However, only very tedious manual procedures allow the retrieval of all sets of homologous genes sequenced for a given set of species. Moreover, this search often generates errors due to the complexity of data to be managed simultaneously: phylogenetic trees, alignments, taxonomy, sequences and related information. HOVERGEN helps to solve these problems by integrating all this information. HOVERGEN corresponds to GenBank sequences from all vertebrate species, with some data corrected, clarified, or completed, notably to address the problem of redundancy. Coding sequences have been classified in gene families. Protein multiple alignments and phylogenetic trees have been calculated for each family. Sequences and related information have been structured in an ACNUC database which permits complex selections. A graphical interface has been developed to visualize and edit trees. Genes are displayed in color, according to their taxonomy. Users have directly access to all information attached to sequences and to multiple alignments simply by clicking on genes. This graphical tool gives thus a rapid and simple access to all data necessary to interpret homology relationships between genes. HOVERGEN allows the user to easily select sets of homologous vertebrate genes, and thus is particularly useful for comparative sequence analysis, or molecular evolution studies.

INTRODUCTION

Comparison of homologous sequences is an essential step to reconstruct the phylogeny of species or to understand the mechanisms of molecular evolution. Such comparisons are also often the first step in understanding the function of a gene. These different research fields have greatly taken advantage of the huge accumulation of sequence data during the last decade. Notably, due to the quick development of computer tools, comparative sequence analysis has proven to be a very efficient approach to detect functionally important regions in protein or nucleic acid sequences or even to study the overall structure of mammalian genomes (for recent examples, see 1–3).

To date, about 500 pairs of man and mouse homologous genes have been sequenced (GenBank 80, December 1993). This figure will rapidly increase since it is now envisaged to jointly sequence homologous regions of human and mouse chromosomes (4). Although less studied, many other vertebrate species are represented in databases. This is particularly interesting for comparative sequence analysis or molecular evolution studies because it allows one to focus attention on recent as well as ancient events. For example, the divergence time between rat and mouse is about 12 million years (My) (5) whereas more than 400 My separate fishes from other vertebrates (6). Therefore, vertebrate sequences are appropriate for the study of genome or gene structure, function and evolution by comparison of homologous genes.

Some important concepts must be defined here. Two genes are said to be homologous if they share similarity due to common ancestry. Among homologous genes, a distinction is made between orthologous genes, i.e. genes that have diverged after a speciation event, and paralogous genes, i.e. genes that have diverged after duplication of an ancestral gene (see fig. 1a). This distinction is essential in molecular phylogeny since only orthologous genes are suitable to reconstruct the history of speciations, but it is also very important in comparative sequence analysis. Indeed, among multigene families, related genes often have different functions and regulations (for example alpha- and beta-actins). Therefore, by comparing paralogous genes one may miss some essential features specific of each gene (2).

However, the orthologous/paralogous distinction is sometimes difficult to establish, as illustrated by the following example. Suppose that you want to compare human and rat bone morphogenetic protein (BMP) genes. The GenBank sequences M22490 is defined as the human BMP2B gene and the sequence Z25868 as the rat BMP2 gene. Comparison of these two sequences show that they are highly similar. Using both criteria, they may be considered as orthologous. However, as shown in figure 1b, the rat gene is much closer to the xenope one than to the human one. Since the divergence of xenope is much older than the divergence between human and rat, the rat and human genes are very likely to be paralogous. Further examination of

*To whom correspondence should be addressed

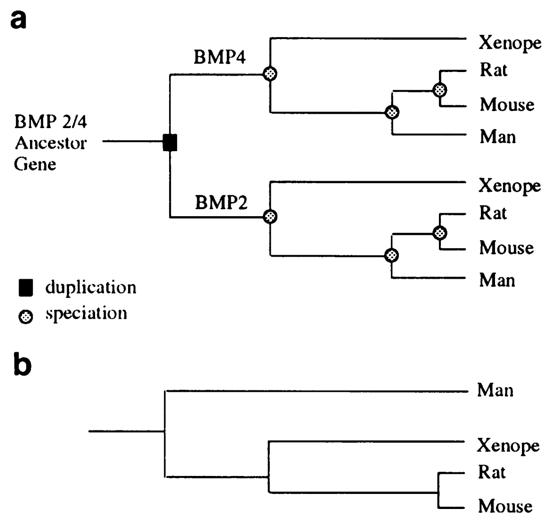


Figure 1. Partial phylogenetic tree of the bone morphogenetic protein (BMP) gene family illustrating paralogy and orthology concepts. **a**—The human, rat, mouse and xenope BMP2 genes are orthologous since they have diverged after speciation events. BMP2 and BMP4 genes are paralogous since they were derived from a duplication event. **b**— The position of the xenope gene in the tree demonstrates that human and rodent genes are paralogous.

other BMP sequences shows that the human BMP2B gene corresponds to the rat gene named BMP4. It is important to notice that when data are lacking (in the example given above, if the xenope sequence was absent), it may be impossible to detect paralogy. Thus, except for very well studied gene families, it is not possible to construct tables of orthologous genes. Rather, homology relationships must be re-interpreted each time new sequences are entered in databases.

Using similarity search programs, it is easy to find in databases all entries that are homologous to a given sequence. However, up to now, there was no tool suitable to retrieve all sets of orthologous genes among a given set of species. Therefore, this search is generally performed manually, essentially on the basis of sequence definition. Unfortunately, sequence definitions are sometimes inaccurate. Therefore, some paralogous genes may have the same definition whereas some orthologous ones may be differently named. Moreover, as this work is particularly tedious, orthologous genes are only searched among species to be studied, whereas, as shown in the above example, it may be important to examine the phylogenetic tree of all sequences of a same gene family to distinguish between paralogy and orthology. Therefore, by using this manual approach one may miss some orthologous genes and, more damaging, may consider some paralogous genes as orthologous. Finally, as sequence databases are growing exponentially, this approach will soon become impracticable.

To respond to these problems, a database of homologous vertebrate genes named HOVERGEN (HOmologous VERtebrate GENe database), which integrates phylogenetic trees, protein multiple alignments, sequences and GenBank annotations, completed with data relevant to gene structure and expression, was constructed. With its graphical interface, HOVERGEN allows the user to easily and rapidly retrieve sets of orthologous genes. The content, structure and usage of HOVERGEN is described in this paper.

DATABASE CONTENT

GenBank nuclear vertebrate sequences

HOVERGEN is primarily based on the GenBank subset of nuclear vertebrate sequences (7). A curator program is used to correct some errors, clarify some features and add information relevant to gene structure and expression. First, some complete coding sequences are mentioned as partial in GenBank (because the location of the stop codon is erroneous), or conversely (notably, because of the choice of EMBL to consider sequences coding for signal peptides as complete). The rules to correct these errors in HOVERGEN are the following: there must be an ATG initiation codon, the number of codons must be an integer, the nucleotides following the end of the coding sequence must be compatible with a stop codon. Secondly, the position of 5′- or 3′-non-coding regions is rarely declared in GenBank, and the description of introns is not totally satisfactory. In particular, when a GenBank entry includes several coding sequences (CDS) from adjacent genes or alternative splicing, it is important to know to which CDS each intron corresponds. Unfortunately, this information is not always provided, or not clearly enough to be automatically interpreted by a program. In HOVERGEN, the position of 5′- and 3′-non-coding regions, introns (either internal or 5′ or 3′ relative to the CDS) and the CDS to which they belong are allexplicitly declared. As we are interested in the study of regulation of gene expression, the information compiled by F. Larsen about the pattern of expression of human genes (8) was included in HOVERGEN. CpG islands, regions susceptible to be involved in regulation of transcription (9), are systematically searched for and annotated when found. Finally, the G+C-content in third position of codons is given for each CDS. This information can be used to predict the isochore in which genes are located (10) and is therefore interesting for the study of vertebrate genomes compositional structure.

Declaration of redundancy

It is frequent to find in GenBank several entries that correspond to a same gene. Typically, one finds a report that corresponds to the cDNA and another derived from genomic sequencing, and sometimes many other complete or partial sequences independently obtained by different laboratories. This redundancy has many drawbacks. For example, it gives a confuse view of the current state of knowledge about a given gene: it makes difficult to distinguish between paralogous genes, different alleles or alternative splicing products of a same gene. Furthermore, similarity search programs will pick up the same gene many times during searches. More annoying, redundancy may distort the interpretation of statistical analyses. Therefore, redundancy among coding sequences was systematically searched for. The aim of this procedure is to group all sequence entries that are real or artifactual variants of a single gene.

An important problem is that two sequences of a same gene are not necessarily strictly identical. Two otherwise identical coding sequences may appear with different lengths in GenBank because the stop codon is included in one but excluded in the other. Other differences arise due to polymorphism in populations or to experimental errors. In average, the frequency of polymorphic sites in coding sequences of *Homo sapiens* is about 0.04% (calculated from 11). However, for a given gene, differences between alleles may reach up to 0.4% (Apolipoprotein E) (11). Little is known about other vertebrate species; however, a higher polymorphism is expected for species with larger effective population size. Sequencing errors are also a problem since sequences submitted to GenBank are not necessarily always carefully checked. Estimates of errors in databases, based on different assumptions, range from 0.29% to 3.55%, or 0.19% to 1.42% if only substitutions are considered (12, 13). These latter figures are probably more realistic for coding sequences, since insertions or deletions cause frame-shifts and hence, are easily detected. Noticeably, paralogous genes that result from a recent gene duplication or conversion may be less divergent. For example, coding sequences of human α-1 and α-2 globin genes are strictly identical. Therefore, when fixing criteria to define redundancy, one must choose a compromise between two sources of error: 1— considering very similar paralogous genes as redundant, and 2— considering slightly divergent redundant sequences as different genes. In order to minimize this latter risk, two coding sequences with less than 1% divergence and 3 nt of difference in length were defined as redundant. However, the user must be aware that two coding sequences declared as redundant in HOVERGEN may correspond to homologous genes resulting from a recent speciation, duplication or conversion event (since the average rate of evolution in coding sequences (silent or non silent) in mammals is about 0.41 ± 0.14% per My after divergence (calculated from 14), 1% difference may correspond to 2.7 ± 1 My).

Using these criteria, 19% of redundancy among the 30687 GenBank vertebrate coding sequences (release 80, December 1993) was detected. This redundancy is not eliminated from HOVERGEN since each sequence may be associated with useful information. Rather, redundancy is explicitly declared which allows the user to discard it when he judges necessary.

Classification in gene families

Our first aim is to facilitate the search of orthologous vertebrate genes. Therefore coding sequences were classified in families, according to similarity criteria. A family is defined here as the set of genes descending from a unique gene in the ancestor of vertebrates. Thus, paralogous genes that result from a duplication anterior to the divergence of vertebrates will be classified in different families.

First, to simplify classification, we have excluded redundancy and partial coding sequences shorter than 300 nt. The remaining 18704 coding sequences were translated in proteins, and compared to each other with the similarity search program blastp (15), using the substitution matrix PAM120. The threshold score to report similarity (S in blastp program) was set according to proteins length (L): $S=150$ for $L \geq 170$ aa, $S=L-20$ for $L<170$ aa and $S=35$ for $L<55$ aa. These thresholds are low enough to detect all homologous sequences that diverged since vertebrate radiation and are high enough to avoid excessive noise due to the presence of segments of low compositional complexity shared by many gene families. Then, sequences were automatically sorted according to similarity criteria, with a program based on the classification algorithm UPGMA (16). The distance metric used for this process was the Poisson probability given by blastp. Afterwards, a program was used to group genes that have at least one homologous sequence in common. Finally, sequences within each group were partitioned in gene families. The difficulty of this classification process is that gene families do not evolve at the same rate. Thus, highly divergent genes may belong to a same family whereas other very similar genes may belong to different families. Therefore, a manual expertise was necessary to perform this last step.

Table 1. Distribution of gene families according to their number of sequences

Number of non redundant sequences	Number of gene families gene families	Frequency
1	1182	39,90%
2	482	16,27%
3–4	500	16,88%
5–9	442	14,92%
10–19	237	8,00%
20–39	83	2,80%
40–99	26	0,88%
≥ 100	10	0,34%
Total	2962	

Table 2. The 10 biggest gene families

Gene family	Number of non redundant sequences
Immunoglobulin	3267
T-cell receptor alpha-chain	320
T-cell receptor beta-chain	310
MHC class I	281
MHC class II	269
cytochrome	137
collagene	120
homeobox	118
beta globin	116
T-cell receptor delta-chain	102

The 18704 protein coding genes have been classified in 2962 families, of which 40% are represented by a single sequence. The distribution of families according to the number of sequences is presented in table 1, and the 10 biggest families are reported in table 2.

Protein sequence multiple alignments and phylogenetic trees

Each set of non-redundant protein sequences belonging to a same family has been aligned with the multiple alignment program CLUSTALV (17), except the 5 most numerous families (table 2) that correspond to genes of the immune system and which evolutionary histories are particularly complex. Then, phylogenetic trees have been obtained from each multiple alignment using the 'neighbor joining' method developed by Saitou and Nei (18), and stored in parenthesized form (Phylip format (19)).

Note that alignments provided by HOVERGEN are crude results of CLUSTALV, without manual correction. Although little errors in alignments do not disturb the overall structure of trees, they may locally modify their topology. Hence, trees in HOVERGEN are an efficient working tool to detect paralogy but may not correspond to exact phylogenetic trees, particularly when short branches are considered.

DATABASE STRUCTURE

Sequences and associated information are stored using ACNUC, a management system especially devoted to biological sequence databases (20). Each set of redundant CDS, and each gene family is assigned a specific reference number. Two tables make the link between CDS and their corresponding redundancy reference number, and between redundancy reference numbers and gene families (fig. 2). Each alignment and phylogenetic tree is stored

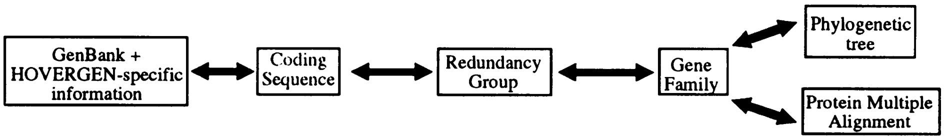
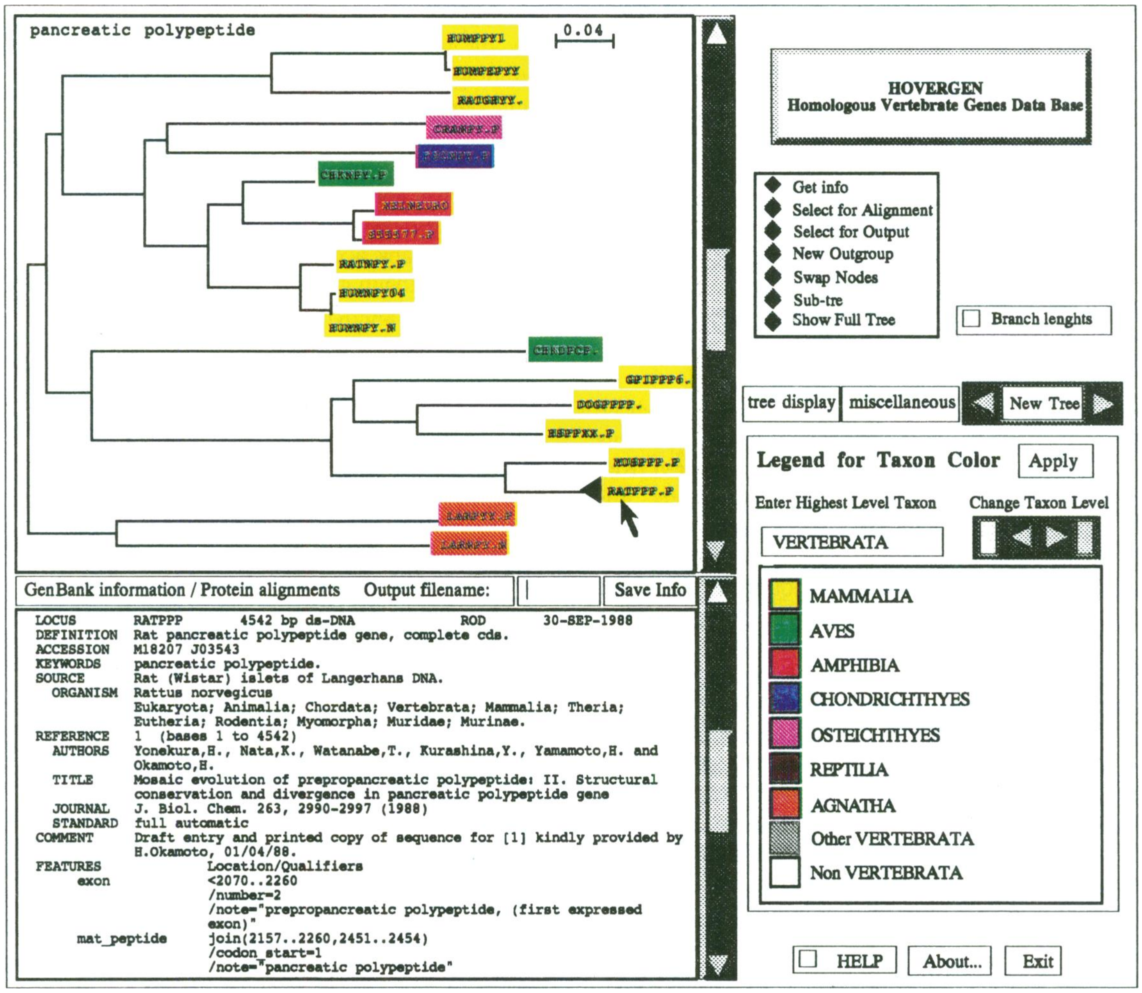

GenBank + HOVERGEN-specific information
Coding Sequence
Redundancy Group
Gene Family
Phylogenetic tree
Protein Multiple Alignment

Figure 2. Schematic representation of data relationships in HOVERGEN.



```
LOCUS       RATPPP       4542 bp ds-DNA            ROD       30-SEP-1988
DEFINITION  Rat pancreatic polypeptide gene, complete cds.
ACCESSION   M18207 J03543
KEYWORDS    pancreatic polypeptide.
SOURCE      Rat (Wistar) islets of Langerhans DNA.
  ORGANISM  Rattus norvegicus
            Eukaryota; Animalia; Chordata; Vertebrata; Mammalia; Theria;
            Eutheria; Rodentia; Myomorpha; Muridae; Murinae.
REFERENCE   1  (bases 1 to 4542)
  AUTHORS   Yonekura,H., Nata,K., Watanabe,T., Kurashina,Y., Yamamoto,H. and
            Okamoto,H.
  TITLE     Mosaic evolution of prepropancreatic polypeptide: II. Structural
            conservation and divergence in pancreatic polypeptide gene
  JOURNAL   J. Biol. Chem. 263, 2990-2997 (1988)
  STANDARD  full automatic
COMMENT     Draft entry and printed copy of sequence for [1] kindly provided by
            H.Okamoto, 01/04/88.
FEATURES             Location/Qualifiers
     exon            <2070..2260
                     /number=2
                     /note="prepropancreatic polypeptide, (first expressed
                     exon)"
     mat_peptide     join(2157..2260,2451..2454)
                     /codon_start=1
                     /note="pancreatic polypeptide"
```

Figure 3. Graphical interface of HOVERGEN.

in a separate file, named according to the gene family to which it corresponds.

USER INTERFACE

Sequence retrieval software

A program named 'query' (20) allows the user to make complex selections in the ACNUC database according to any information declared in a structured form (keywords, bibliography, taxonomy, etc.). Selection criteria can be combined with boolean operators. For example, since G+C-content is indicated in HOVERGEN, it is possible to select all coding sequences with more than 30% but less than 60% G+C-content. ACNUC permits to handle easily both sequences and their biologically meaningful subfragments when explicitly declared (e.g. in HOVERGEN: CDS, exons, introns, non-coding regions, CpG islands ...). Therefore, it is possible with 'query' to select, for example, all complete introns shorter than 100 nt. Finally, 'query' allows the user to extract selected sequences (with the possibility to translate protein coding sequences) or parts of them. Interestingly, one may extract, for example, sequences situated 10 nt upstream and downstream of intron junctions.

Noticeably, and because information about gene families has been included in HOVERGEN, it is possible to select with 'query' all homologous genes that have been simultaneously sequenced in a list of species.

Graphical interface for multiple alignments and phylogenetic trees

The search of orthologous genes necessitates to visualize the phylogenetic tree of the family to which they belong. Therefore, we constructed a graphical interface with which the user can interact with a computer mouse (fig. 3). The central part of this graphical interface is a tree editor that allows one to pivot the two daughter branches around each node, and to position the root anywhere. This latter feature is particularly important since the 'neighbor-joining' method gives unrooted trees. Furthermore, the user can reduce or magnify trees, or even split trees in subtrees, which can be useful when studying complex families.

Genes are displayed in color, according to the taxonomic position of the species from which they have been isolated. The user may also choose to affect color according to genera, orders, classes, or any other taxonomic classification level.

By clicking on a gene in the tree, the user can visualize in another window the GenBank definition of the corresponding sequence. If there exist multiple redundant coding sequences for this gene, all definitions are displayed. Then, by clicking on a definition, the user has directly access to the full GenBank record of each sequence, completed with HOVERGEN-specific information.

In a similar fashion, the user can select genes for which he wants to visualize the protein sequence alignment. Alignments between selected genes are not calculated but reconstructed from the pre-existing whole family multiple alignment. Thus, this operation is almost instantaneous. The divergence (number of differences / aligned sequence length - gaps excluded) and the gap frequency (number of gaps / aligned sequence length) are calculated, thus giving a measure of similarity between aligned sequences.

It is possible to highlight genes belonging to a list saved in a file. Thus the user may use 'query' to select sequences that fit particular criteria (see above) and then visualize the corresponding genes in the tree. This option may be useful when trees are complex and one would like to ask a question such as: 'in which genes of this tree are the introns known?'.

Functions allow the user to save sets of selected genes, multiple alignments or GenBank information in output files. Phylogenetic trees can be saved in postscript format for printing.

The user can select a family tree by different means. He can visualize trees from a list of families or sequence names selected and saved with the 'query' program. Otherwise he can manually enter a sequence name or a keyword matching the definition of a given family. It is also possible to select gene families according to the number of different vertebrate species present in these families.

DISCUSSION

Search for orthologous genes

The comparison of orthologous genes is a very efficient approach to reconstruct species phylogenies, to study evolutionary processes at the molecular level or to find functionally important regions in biological sequences. Interesting new findings are expected to emerge from the increasing amount of sequences available through databases (1–3). However, the search of orthologous genes in a given set of species is a complex task that, up to now, could be managed only manually (see methods in 21,22).

The first step of this manual approach consists in searching for similarity between genes of each species. It is easy and rapid, using similarity search program such as blast (15), to find genes homologous to a given sequence. However, the time needed for comparing all sequences of several species between each other increases with the square of the number of sequences. The number of significant similarities found also increases dramatically and thus the analysis of similarity search outputs is particularly tedious. Therefore, this approach is generally limited to a small number of species. The second step consists in verifying that related genes are not paralogous but orthologous. As discussed in the introduction, the distinction between orthologous and paralogous genes requires to consider the entire phylogenetic tree of each gene family (fig. 1b). However, since this data is not available, the distinction can only be made on the basis of sequence definitions. Unfortunately, sequence annotation is not always provided, may be inexplicit or inaccurate. For example the sequence of human endothelin 1, published in 1989 (GenBank accession number M25380) is defined as 'the' human endothelin gene (23), whereas it was recognized later that two other endothelin genes (endothelin 2 and 3) exist in mammalian genomes (24). Other example, the murine endothelin 2 (X59556) is defined as a 'vasoactive intestinal contractor'. Therefore some orthologous sequences are missed whereas some paralogous genes may be considered as orthologous. This procedure is thus error-prone, particularly tedious, and, furthermore, must be repeated at each new database release, or each time one wants to change genes under study. Finally, this approach will soon become impossible because of the exponential growth of published sequences.

Two main features of HOVERGEN help solving these problems:

1— HOVERGEN provides users with a partition of sequences in gene families resulting from a systematic search of similarity between vertebrate genes. Moreover, multiple protein alignments and phylogenetic trees of each gene families are provided.

2— A graphical interface allows the user to easily handle simultaneously phylogenetic trees, protein multiple alignments, sequences and associated information. Tree editing functions (modification of the root, extraction of sub-tree) have proven to be particularly efficient for analysis of complex families.

Thus, HOVERGEN gives a rapid and simple access to all data necessary to interpret homology relationships between genes. We insist on the fact that HOVERGEN does not give lists of orthologous but homologous genes and that the user has to interpret phylogenetic trees to exclude paralogy.

In the PIR-NBRF database (25), protein sequences are also classified in families, but with different definitions and objectives. In PIR-NBRF, proteins are first grouped in super-families on the basis of sequence similarity and functional relationships; super-families are then subdivided in families, subfamilies or entries on the basis of similarity level (respectively 50%, 80% and 95% similarity) (25). However, this classification is not adequate for our purpose since rapidly evolving orthologous genes could be classified in different families (for example human and hagfish insulins are only 45% similar) whereas highly conserved paralogous genes may belong to the same sub-family (alpha- and beta-actins are 94% similar). In HOVERGEN orthologous genes are gathered, while paralogous genes which duplication unambiguously predates vertebrate divergence are classified in distinct families.

Usage of HOVERGEN

HOVERGEN has been initially developed to facilitate the search of homologous vertebrate genes. Besides information on homology relationships, HOVERGEN includes data absent from GenBank, relevant to gene structure or regulation. Thus, HOVERGEN is particularly useful for comparative sequence analysis. As an example of complex use of HOVERGEN, suppose that we want to study ancient regulatory elements in 5′-non-coding regions of vertebrate genes. First, we use 'query' to select in the ACNUC sequence database homologous genes from different classes of vertebrates and for which at least 1000 nt have been sequenced in 5′ of the initiation codon. Then we use the graphical interface to visualize phylogenetic trees of corresponding gene families and select orthologous genes that fit the criteria that we had fixed for our comparative study. Visualization of the phylogenetic tree also permits to recognize complex evolutionary histories involving multiple gene losses or duplications. Thus, it is possible to eliminate from the study, genes for which orthology relationships are impossible to determine. Since the usage of HOVERGEN is quite simple, it is possible to rapidly change selection criteria to extend the study to, say, all genes with 5′-non-coding sequences that encompass a CpG island.

The range of application of HOVERGEN is not limited to comparative sequence analysis. In molecular phylogeny studies, it is essential to exclude paralogous genes to reconstruct trees of species. Furthermore, it is important to dispose of as many genes as possible in order to increase the reliability of phylogenetic trees. Therefore, HOVERGEN will be particularly useful to elucidate vertebrate phylogeny that remains obscure at many points.

After usage, HOVERGEN also appeared to be useful for the 'classical' molecular biologist who wants to have an overall view of what is known about a particular gene. Finally, HOVERGEN should be an interesting tool for teaching purpose.

PERSPECTIVES

HOVERGEN is operational and has been already used for various studies (article in preparation). A feed-back is expected from users in order to add new functionality or new data relevant to gene structure or expression.

The database structure of HOVERGEN could be adapted for the study of other taxons. Database construction programs are available and could be used with any set of sequences. The limiting step is the classification in families that, as previously discussed, necessitates a manual expertise. Interested persons should contact the authors.

AVAILABILITY

HOVERGEN data and software are freely available through anonymous ftp at 'biom3.univ-lyon1.fr'. Software is provided with on-line help. Present version runs on Sun workstations using SunOS (version 4.0 or more) or Solaris. The graphical interface has been written in C language using the SUIT interface constructing tool (26) which is compatible with many different systems. Therefore, HOVERGEN should be easily installed on other mini-computers. A color screen is recommended.

HOVERGEN is updated at each new GenBank release (bimonthly). HOVERGEN release 4 (from GenBank 80, December 1993) requires 400 megabytes (Mb) of hard disk. A demonstration version of only 4 Mb is also available on our ftp server. Help can be obtained via e-mail to: 'duret@biomserv.univ-lyon1.fr'.

ACKNOWLEDGEMENTS

We especially thank C.Gautier and G.Matassi for their suggestions and helpful discussions after usage of HOVERGEN. F.Dorkeld is gratefully acknowledged for his help in writing C programs.

REFERENCES

1. Green,P., Lipman,D., Hillier,L., Waterson,R., States,D. and Claverie,J.M. (1993) *Science*, **259**, 1711–1716.
2. Duret,L., Dorkeld,F. and Gautier,C. (1993) *Nucleic Acids Res.*, **21**, 2315–2322.
3. Mouchiroud,D. and Bernardi,G. (1993) *J. Mol. Evol.*, **37**, 109–116.
4. Collins,F. and Galas,D. (1993) *Science*, **262**, 43–46.
5. O'Huigin,C. and Li,W.H. (1992) *J. Mol. Evol.*, **35**, 377–384.
6. Goodman,M., Czelusniak,J., Koop,B.F., Tagle,D.A. and Slightom,J.L. (1987) *Cold Spring Harbor Symposia on Quantitative Biology*, **LII**, 875–890.
7. Burks,C., Cassidy,M., Cinkowsky,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.D., Keen,G.M., Kelley,T.A., Kelly,M., Kristofferson,D. and Ryals,J. (1991) *Nucleic Acid Res.*, **19**, 2221–2225.
8. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) *Genomics*, **13**, 1095–1107.
9. Bird,A. (1992) *Cell*, **70**, 5–8.
10. Mouchiroud,D., D'Onofrio,G., Aissani,B., Macaya,G., Gautier,C. and Bernardi,G. (1991) *Gene*, **100**, 181–187.
11. Li,W.H. and Sadler,A. (1991) *Genetics*, **129**, 513–523.
12. Krawetz,S.A. (1989) *Nucl Acids Res.*, **17**, 3951–3957.
13. Kristensen,T., Lopez,R. and Prydz,H. (1992) *DNA Sequence*, **2**, 343–346.
14. Li,W.H., Luo,C. and Wu,C. (1985) In MacIntyre,R.J. (ed.), Molecular Evolutionary Genetics. Plenum, Press, New York, pp.1–94.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16. Sokal,R.R. and Michener,C.D. (1958) *Univ. Kansas Sci. Bull*, **28**, 1409–1438.
17. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *CABIOS*, **8**, 189–191.
18. Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
19. Felsenstein,J. (1989)*Cladistics*, **5**, 164–166.
20. Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and Di Paola,G. (1985) *CABIOS*, **1**, 167–172.
21. Mouchiroud,D. and Gautier,C. (1990) *J. Mol. Evol.*, **31**, 81–91.
22. Wolfe,K.H. and Sharp,P.M. (1993) *J. Mol. Evol.*, **37**, 441–456.
23. Bloch,K.D., Friedrich,S.P., Lee,M.E., Eddy,R.L., Shows,T.B. and Quertermous,T. (1989) *J. Biol. Chem.*, **264**, 10851–10857.
24. Inoue,A., Yanagisawa,M., Kimura,S., Kasuya,Y., Miyauchi,T., Goto,K. and Masaki,T. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2863–2867.
25. Barker,W.C., George,D.G., and Hunt,L.T (1990) *Methods in Enzymol.*, **183**, 31–49.
26. Conway, M.J. (1992) SUIT: the Simple Interface Toolkit Version 2.3 Reference Manual, University of Virginia.