Automatic identification of large collections of protein-coding or rRNA sequences

Anne-Muriel Arigon[*,1], Guy Perrière[1] and Manolo Gouy[1]

[*] To whom correspondence should be addressed.
Tel: +33 426234475; Fax: +33 472431388; Email: arigon@biomserv.univ-lyon1.fr

[1] Université de Lyon; université Lyon 1; CNRS; UMR 5558
Laboratoire de Biométrie et Biologie Evolutive
43 boulevard du 11 novembre 1918
Villeurbanne F-69622, France.

Abstract:

The number of available genomic sequences is growing very fast, due to the development of massive sequencing techniques. Sequence identification is needed and contributes to the assessment of gene and species evolutionary relationships. Automated bioinformatics tools are thus necessary to carry out these identification operations in an accurate and fast way. We developed HoSeqI (Homologous Sequence Identification), a software environment allowing this kind of automated sequence identification using homologous gene family databases. HoSeqI is accessible through a Web interface (http://pbil.univ-lyon1.fr/software/HoSeqI/) allowing to identify one or several sequences and to visualize resulting alignments and phylogenetic trees. We also implemented another application, MultiHoSeqI, to quickly add a large set of sequences to a family database in order to identify them, to update the database, or to help automatic genome annotation. Lately, we developed an application, ChiSeqI (Chimeric Sequence Identification), to automate the processes of identification of bacterial 16S ribosomal RNA sequences and of detection of chimeric sequences.

Keywords:

Automatic identification, similarity, alignment, phylogeny, chimera.

Introduction

Identification is used in many fields, such as microbiology, medicine, environment. Sequence identification consists in the attribution of an unknown taxonomic unit to a taxonomic group of a pre-established classification. Thus, to identify a new taxon or a new sequence, it is necessary to find its nearest known taxon. In the medical field, methods of identification are used to detect and recognize micro-organisms implied in pathologies, which thus helps choosing the most suitable treatment. Identification can also be used in the agro-alimentary field as tools for food traceability. In other contexts such as identification of species or taxons from environmental organism molecular markers, the confrontation of a new sequence with a database, or sequence database update, the assignment of a new sequence to a collection is necessary. The number of available biological sequences increasing considerably with the development of massive sequencing techniques, it is necessary to rapidly classify these sequences into existing databases.

According to analyzed data, the approach used for identification differs and several tools exist. Identification tools often vary with the type of sequences and thus with the sequence databases for which they were developed. Several tools exist to make sequence identification and most of them are domain specific or data specific. For instance, some applications allow bacterial identification as BIBI (Bioinformatic Bacterial Identification) [1], PhyID/CD [2] or MicroSeq (Microbial identification System), others are specialized for the medical domain as RIDOM (Ribosomal Differenciation Of Medical Organisms) [3] or for the identification of Ribosomal RNA sequences as the RDP classifier (Ribosomal Database Project) [4], or TaxI [5] based on DNA barcodes.

We are interested in the homologous gene family databases HOVERGEN and

HOGENOM [6] developed in our group. In these databases, homologous sequences are clustered into families, *i.e.*, sequences of the same family share a common ancestor. Sequence alignments and phylogenetic trees for each family are also stored in these databases. Thus, these databases can be used for different purposes, among which phylogenetic analyses, and they allow the study of sequence evolutionary relationships. In order to build these family databases, several complex automated procedures are needed (similarity search, gene clustering, multiple alignment and tree computations). With the very fast growth of biological data, gene family database updates are time-consuming and tedious. Moreover, the addition of a single sequence to a given family from these databases can have many repercussions on the topology of the associated phylogenetic tree; these changes may be located near the introduced sequence, but they may also be located in deep nodes. In such case, the phylogenetic information brought by the whole family should be taken into account. Also, as HOVERGEN and HOGENOM contain large families, with several thousand sequences, powerful algorithms are required in order to manage large amount of sequences. Available identification tools, such as those presented previously, are developed to treat specific data and cannot be used effectively with large family databases. Thus, it is necessary to develop methods and bioinformatics tools (i) to carry out identification processes in a precise and rapid way, and (ii) to quickly add sequences to these databases without integrally updating them.

Two applications adapted to homologous gene family databases: HoSeqI and MultiHoSeqI

We have developed an application – HoSeqI (Homologous Sequence Identification) – and an other derived from the first – MultiHoSeqI. HoSeqI [7] is a web application (http://pbil.univ-lyon1.fr/software/HoSeqI/) that allows to automatically identify sequences in

large gene family databases. The identification process of an unknown sequence into these databases consists in (i) finding the homologous gene family to which this sequence belongs, using similarity search, (ii) aligning the analysed sequence with the whole family and (iii) reconstructing the phylogenetic tree of the family including the new sequence. HoSeqI proposes an interface allowing the user to submit his/her sequence and to choose the database. When the family of the studied sequence is determined, the user can obtain information on the selected family and choose between several multiple alignment and tree building programs. Then it is possible to visualize the resulting alignment and phylogenetic tree (figure 1).

The developments carried out for HoSeqI were also used to create another application – MultiHoSeqI – allowing to quickly add several thousand sequences to a homologous gene family database. MultiHoSeqI corresponds to a generalization of HoSeqI to *n* sequences. For each one of these *n* sequences, the application identifies the family to which it belongs, then for each family containing at least an added sequence, alignments are computed and phylogenetic trees are built including the new sequences. MultiHoSeqI is publicly usable on the HoSeqI web interface with a restriction on the number of sequences: the user can choose between the identification of only one sequence or of several sequences. If “several sequences” option is selected, the user submits his/her sequences, chooses the database, the multiple alignment and tree building programs. The process is then started on the server and the results are sent to the user by e-mail.

Use of MultiHoSeqI with sequences of bacterial genus *Frankia*

MultiHoSeqI has been used to add genes from several collections of protein sequences to the databases developed by the PBIL (Pôle BioInformatique Lyonnais): putative protein sequences from metagenomes and from completely sequenced bacterial genomes. In

collaboration with Philippe Normand (Laboratory of Soil Microbial Ecology, University of Lyon), Vincent Daubin and Simon Penel (Laboratory of Biometry and Evolutionary Biology, University of Lyon), this application was used to add predicted protein sequences from two completely sequenced genomes of the bacterial genus *Frankia* into the HOGENOM database in order to study the evolution of these genomes and to detect possible horizontal gene transfers to *Frankia*.

The bacterial genus *Frankia* belongs to the class Actinobacteria. Among Actinobacteria, there are in particular genus *Mycobacterium* (agents of tuberculosis and leprosy) and genus *Streptomyces* (soil bacteria at the origin of many antibiotics). Twelve species of *Frankia* are recognized today. These bacteria fix nitrogen and convert atmospheric N2 gas into ammonia, this in symbiosis with a large spectrum of dicot plants, called actinorhizal. These plants, with their symbiotic bacteria, are collectively responsible for approximately 15% of the biologically fixed nitrogen in the world. This association presents a major ecological interest and many actinorhizal plants are used by the pharmaceutical industry because of their large production of phenolic molecules of various activities (*e.g.*, antimicrobial, antioxydant, antiviral, anti-inflammatory drugs, antispasmodic, antitumor). So, it is interesting to search information about genes involved in symbiosis and to study how symbiosis evolved.

Three strains of *Frankia* were available when we made this study: HFPCcI3 (CcI3), EAN1pec (EAN) and ACN14a (ACN). The two first strains were sequenced in the DOE Joint Genome Institute in collaboration with D. Benson (University of Connecticut) and L. Tisa (University of New Hampshire). The third strain was sequenced in the Génoscope in collaboration with P. Normand. Genomes of these strains are circular and their size vary between 5.38 millions of base pairs (Mb) for CcI3, 7.50 Mb for ACN and 9.08 Mb for EAN.

MultiHoSeqI was used to add sequences of genomes of the strains CcI3 (4557 sequences)

and EAN (7976 sequences) to a local version of the database HOGENOM (based on the version 2 of October 2004) containing sequences of the strain ACN of *Frankia*. Among the 12533 sequences that were analysed, 2450 (19,5%) sequences could not be identified. These sequences are orphan for which no recognizable homolog exits in the HOGENOM database. So MultiHoSeqI identified 10083 sequences of *Frankia* and computed alignments and phylogenetic trees of 4435 families containing at least one *Frankia* sequence.

By using the database thus updated, phylogenetic trees were analyzed in order to detect possible horizontal gene transfers to Frankia. Firstly, we carried out a pattern search using the tool FamFetch [8,9]: we searched genes of *Frankia* whose closest relatives were not from other Actinobacteria. Secondly, we re-built phylogenetic trees of families selected with the pattern search using Gblocks [10] to filter alignments and then PhyML to compute the trees. Possible horizontal gene transfers were detected if *Frankia* proteins were significantly clustered with sequences from distant species.

An example of a possible horizontal gene transfer to *Frankia* of the gene coding for prephenate dehydratase is shown in figure 2. Sequences from the three *Frankia* strains are localised within sequences of Alpha-Proteobacteria, a group phylogenetically distant from *Frankia*. So some ancestor of *Frankia* probably acquired this gene from an Alpha-Proteobacterial donor. It is interesting to notice that Alpha-Proteobacterial species of this tree are also nitrogen fixing bacteria.

An application adapted to 16S ribosomal RNA sequence databases: ChiSeqI

We are also interested in 16S ribosomal RNA databases, such as the American database, RDP [11] or the European database, Ribosomal RNA Database [12]. These databases contain

16S ribosomal RNA (rRNA) sequences which are commonly used for bacterial identification because these molecules are ubiquitous, abundant in cells and having a conserved structure. When sequences come from PCR amplification, chimeras, i.e. artefactual sequences produced by the experimental protocol and composed of several DNA fragments of distinct origins, can occur. Chimeras represent an important problem because they suggest the presence of non existing organisms. It is thus necessary to be able to detect them. The tasks of identification and chimera detection require the chaining of different programs (for similarity search, alignment and tree computation) that are sometimes complex to handle. When these tasks must be carried out on a large number of sequences, these two processes then become time-consuming. So, it is necessary to have automated tools making it possible to carry out these operations in a precise and fast way.

Several available tools allow to identify rRNA sequences, some we presented previously (BIBI, RDP classifier); others can detect chimera, such as the Robison-Cox et al. method [13], Check chimera [4], mglobalCHI package [14], Bellerophon [15], PhyID/CD, CCODE (Chimera and Cross-Over DEtection) [16], Mallard [17] or Pintail [18]. However these methods give various results. They are complementary and none is known to be the best. They allow to obtain information to help the user to determine if the analyzed sequence is chimeric and they do not allow automatic detection of chimera and identification of non-chimeric sequences in a single operation.

In the context of a collaboration with the Laboratory of Soil Microbial Ecology, we developed an application - called ChiSeqI (Chimeric Sequence Identification) - allowing to automate the processes: (i) of identification of non chimeric sequences using a database of 16S rRNA sequences; and (ii) of detection of chimeras among a set of 16S rRNA sequences.

In both cases, developed methods employ a phylogenetic approach. The database used for the identification was created from a data set of bacterial 16S rRNA sequences of cultivated strains (made available by Richard Christen, Laboratory of Virtual Biology, University of Nice). The algorithm developed for the application is divided into two parts: first, for each analyzed sequence, it determines if the sequence is not a chimera; second, it identifies the species to which the treated sequence belongs. This algorithm allows to detect and to discard chimeric sequences and then, it allows to identify the species to which the non chimeric sequences belong. But it does not analyse chimeras and does not determine break points.

In order to identify the species to which belongs a sequence, it is necessary to determine the closest relatives of the database. Firstly, the set of the most similar sequences is selected. Then these sequences are aligned, including the analyzed sequence, and finally, the corresponding phylogenetic tree is built. Various alignment methods are proposed to the user: CLUSTAL W [19], MULTALIN [20], MABIOS [21], MENTALIGN [22] and MUSCLE [23], and for tree building: QuickTree [24], FastME [25], BIONJ [26] and PhyML [27]. The species to which the studied sequence belongs, is considered to be close (even identical) to the species of the nearest sequence in the phylogenetic tree.

The principle used for chimera detection is also based on closest neighbours. A chimeric sequence corresponds to the fusion of several sequence segments. Thus, if a chimera is compared to a set of sequences, one expects to find several phylogenetically distant matches, which obtain good alignment scores with various parts of the chimeric sequence. In the developed detection method, the analyzed sequence is cut in two halves of equal length. Then, for each half-sequence, the set of the most similar sequences is identified in the 16S rRNA sequence database and the redundancy between the two sets is eliminated. The alignment of these two sets of homologous sequences is then computed, and the two half-sequences are added to this alignment. If the two half-sequences are distant in the corresponding

phylogenetic tree, *i.e.* if the phylogenetic distance calculated between the nearest nodes to the half-sequences is higher than a threshold, then the two parts of the studied sequence come from different organisms. The sequence is then considered as a chimera.

Here is an example of chimera detection. The analyzed sequence is a 16S rRNA sequence of an unknown proteobacterium (Genbank accession number: X68474). ChiSeqI was used to determine if this sequence is chimeric. After the similarity search with the two half-sequences, the alignment and the phylogenetic tree (figure 3) were computed. On this tree, the two half-sequences are circled. The first half-sequence is close to a group of Acetobacteraceae and the second is close to a group of Hyphomicrobiaceae. So the analyzed sequence seems to be a chimera between at least these two distinct organisms.

Conclusion

We have presented here three applications allowing to rapidly and automatically identify genomic sequences. Firstly, HoSeqI and MultiHosSeqI are adapted to homologous sequence databases. Via a web interface, HoSeqI determines homologous gene families to which the series of query sequences belong and proposes to visualize alignments and phylogenetic trees of these families, including analyzed sequences. HoSeqI thus contributes to the study of the evolutionary background of new sequences. MultiHosSeqI is used to add many sequences to a homologous family database. This program can be used to quickly identify a large amount of sequences, but also to make an immediate update of a database or to annotate sequences. This application was used to add sequences of *Frankia* genomes to the HOGENOM database, which allowed us to detect potential horizontal gene transfers to *Frankia*. Finaly, ChiSeqI is adapted to the 16S rRNA sequence databases and proposes a module to detect chimera. The usefulness of ChiSeqI is thus to automatically identify a set of 16S rRNA sequences by

determining, for each sequence, the species to which it belongs and its taxonomy, after having discarded chimeras.

References

[1] Devulder G., Perrière G., Baty F. and Flandrois J. P. (2003) BIBI, a bioinformatic bacterial identification tool. J. Clin. Microbiol., 41, 1785–1787.

[2] Flandrois J. P., Mignard S., Dantony E., Gouy M. and Devulder G. (2005). Génération et visualisation de la phylogénie des Bacteria pour l'étude des incohérences taxinomie-phylogénie. In Proceedings of Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), G. Perrière, A. Guénoche, C. Geourjon, Ed. , p. 277–285, Lyon.

[3] Harmsen D., Dostal S., Roth A., Niemann S., Rothganger J., Sammeth M., Albert J., Frosch M. and Richter E. (2003). RIDOM : comprehensive and public sequence database for identification of Mycobacterium species. BMC Infect Dis, 3, 26.

[4] Cole J. R., Chai B., Farris R. J., Wang Q., Kulam S. A., McGarrell D. M., Garrity G. M. and Tiedje J. M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res., 33, D294–D296.

[5] Steinke D., Vences M., Salzburger W. and Meyer A. (2005). TaxI : a software tool for DNA barcoding using distance methods. Philos Trans R Soc Lond B Biol Sci, 360(1462), 1975–80.

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

[7] Arigon A.M., Perrière G. and Gouy M. (2006) HoSeqI: automated homologous sequence identification in gene family databases. Bioinformatics. 2006 Jul 15;22(14):1786-7.

[8] Perrière G., Duret L. and Gouy M. (2000). HOBACGEN : database system for

comparative genomics in bacteria. Genome Res, 10(3), 379–85.

[9] Dufayard J. F., Duret L., Penel S., Gouy M., Rechenmann F. and Perrière G. (2005) Tree pattern matching in phylogenetic trees : automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics, 21(11), 2596–603.

[10] Castresana J (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol, 17(4), 540–52.

[11] Cole J.R., Chai B., Farris R.J., Wang Q., Kulam-Syed-Mohideen A.S., McGarrell D.M., Bandela A.M., Cardenas E., Garrity G.M., Tiedje J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res., 35 : D169-D72.

[12] Wuyts J., Perriere G. and Van De Peer Y. (2004) The European ribosomal RNA database. Nucleic Acids Res., 32, D101–D103.

[13] Robison-Cox J. F., Bateson M. M. and Ward D. M. (1995). Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. Appl Environ Microbiol, 61(4), 1240–5.

[14] Komatsoulis G. A. and Waterman M. S. (1997). A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations. Appl Environ Microbiol, 63(6), 2338–46.

[15] Huber T., Faulkner G. and Hugenholtz P. (2004). Bellerophon : a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics, 20(14), 2317–9.

[16] Gonzalez J. M., Zimmermann J. and Saiz-Jimenez C. (2005). Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics, 21(3), 333–7.

[17] Ashelford K. E., Chuzhanova N.A., Fry J.C., Jones A. J. and Weightman A.J. (2005) "At least one in twenty 16S rRNA sequence records currently held in public repositories estimated to contain substantial anomalies." Applied and Environmental Microbiology,

(12): 7724-7736.

[18] Ashelford K. E., Chuzhanova N.A., Fry J.C., Jones A. J. and Weightman A.J. (2006) "New Screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras." Applied and Environmental Microbiology, 72(9): 5734-5741.

[19] Thompson J. D., Higgins D. G. and Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res., 22, 4673–4680.

[20] Corpet F. (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res., 16, 10881–10890.

[21] Abdeddaïm S. (1997) Fast and sound two-step algorithms for multiple alignment of nucleic sequences. Int. J. Artif. Intell. Tools, 6, 179–192.

[22] Dufayard J.F. (2004) Incremental algorithms for the alignment and the phylogeny of large homologous sequence families. Ph.D. Thesis, Joseph Fourier University, Grenoble, France.

[23] Edgar R.C. (2004) MUSCLE: amultiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5, 113.

[24] Howe K., Bateman A. and Durbin R. (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics, 18, 1546–1547.

[25] Desper R. and Gascuel O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol., 19, 687–705.

[26] 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.

[27] Guindon S. and Gascuel O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol., 52, 696–704.

Legends of figures:

Figure 1.

Screenshot of HoSeqI interface showing an example of sequence identification. Information on the family to which the analyzed sequence belongs and the resulting multiple sequence alignment and phylogenetic tree can be visualized. The query sequence name is boxed in the tree and is at the top line in the alignment.

Figure 2.

Example of possible horizontal gene transfer to *Frankia.*

Figure 3.

Phylogenetic tree obtained using ChiSeqI in order to determine if the analyzed sequence, named X68474 and circled in the tree, is chimeric.

Figure 1.

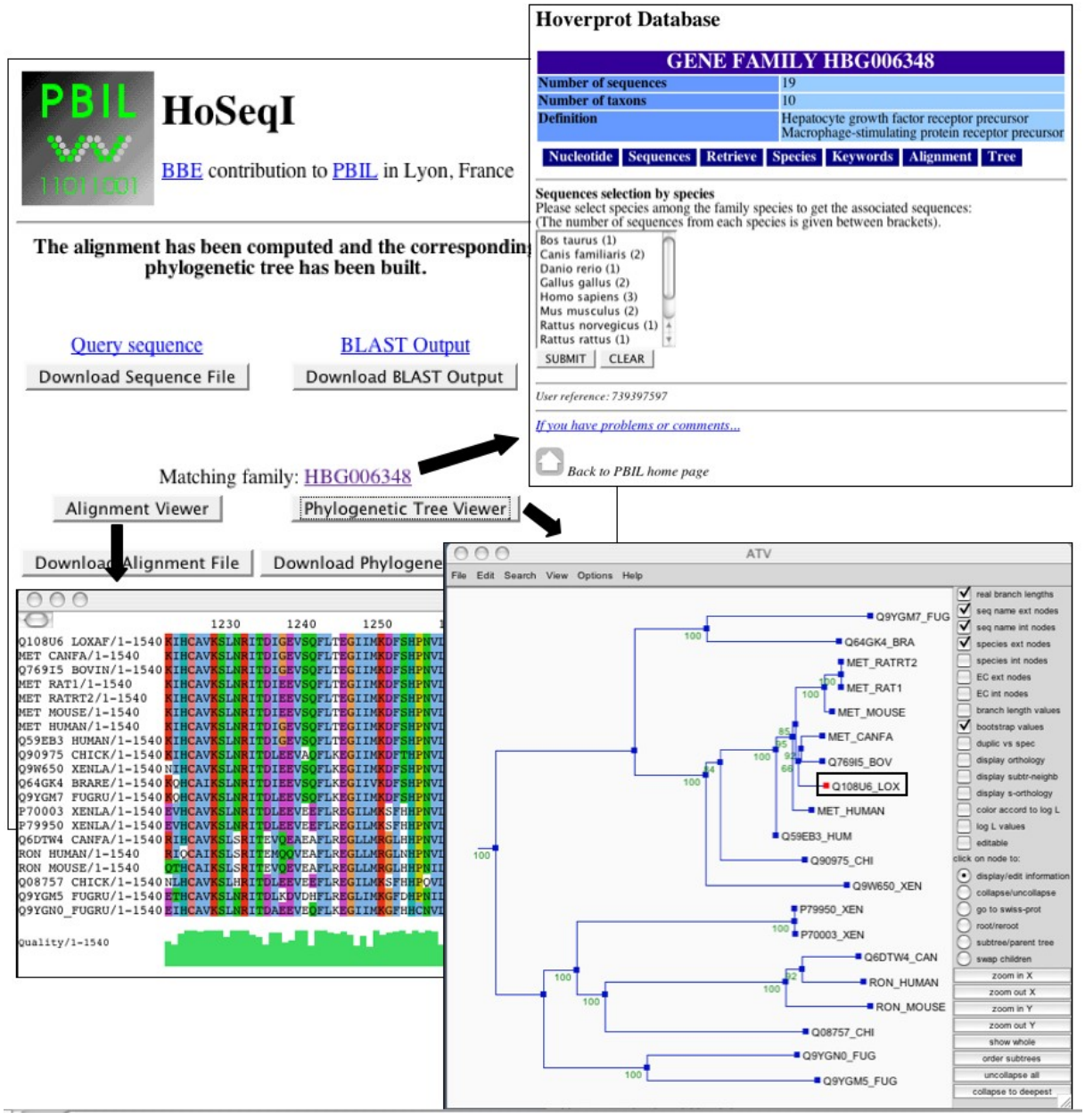

HoSeqI
BBE contribution to PBIL in Lyon, France
The alignment has been computed and the correspondin
phylogenetic tree has been built.
Query sequence
Download Sequence File
BLAST Output
Download BLAST Output
Matching family: HBG006348
Alignment Viewer
Phylogenetic Tree Viewer
Download Alignment File
Download Phylogene
Hoverprot Database
GENE FAMILY HBG006348
Number of sequences 19
Number of taxons 10
Definition Hepatocyte growth factor receptor precursor
Macrophage-stimulating protein receptor precursor
Nucleotide
Sequences
Retrieve
Species
Keywords
Alignment
Tree
Sequences selection by species
Please select species among the family species to get the associated sequences:
(The number of sequences from each species is given between brackets).
Bos taurus (1)
Canis familiaris (2)
Danio rerio (1)
Gallus gallus (2)
Homo sapiens (3)
Mus musculus (2)
Rattus norvegicus (1)
Rattus rattus (1)
SUBMIT
CLEAR
User reference: 739397597
If you have problems or comments...
Back to PBIL home page
ATV
File Edit Search View Options Help
Q9YGM7_FUG
Q64GK4_BRA
MET_RATRT2
MET_RAT1
MET_MOUSE
MET_CANFA
Q769I5_BOV
Q108U6_LOX
MET_HUMAN
Q59EB3_HUM
Q90975_CHI
Q9W650_XEN
P79950_XEN
P70003_XEN
Q6DTW4_CAN
RON_HUMAN
RON_MOUSE
Q08757_CHI
Q9YGN0_FUG
Q9YGM5_FUG

Figure 2

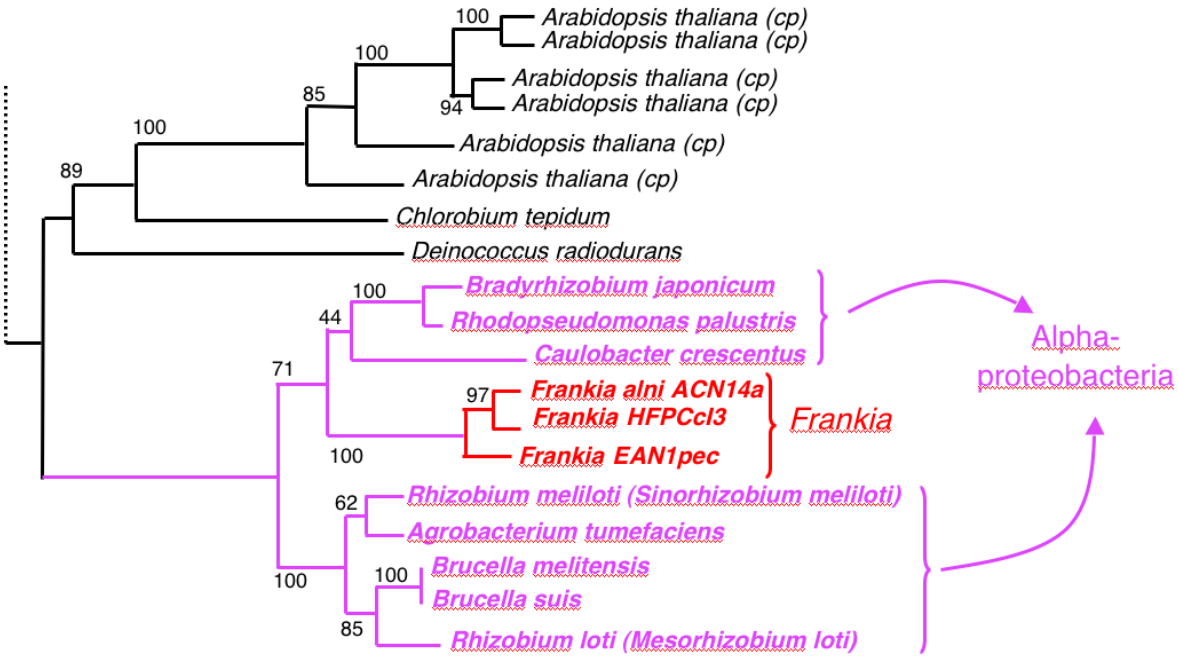
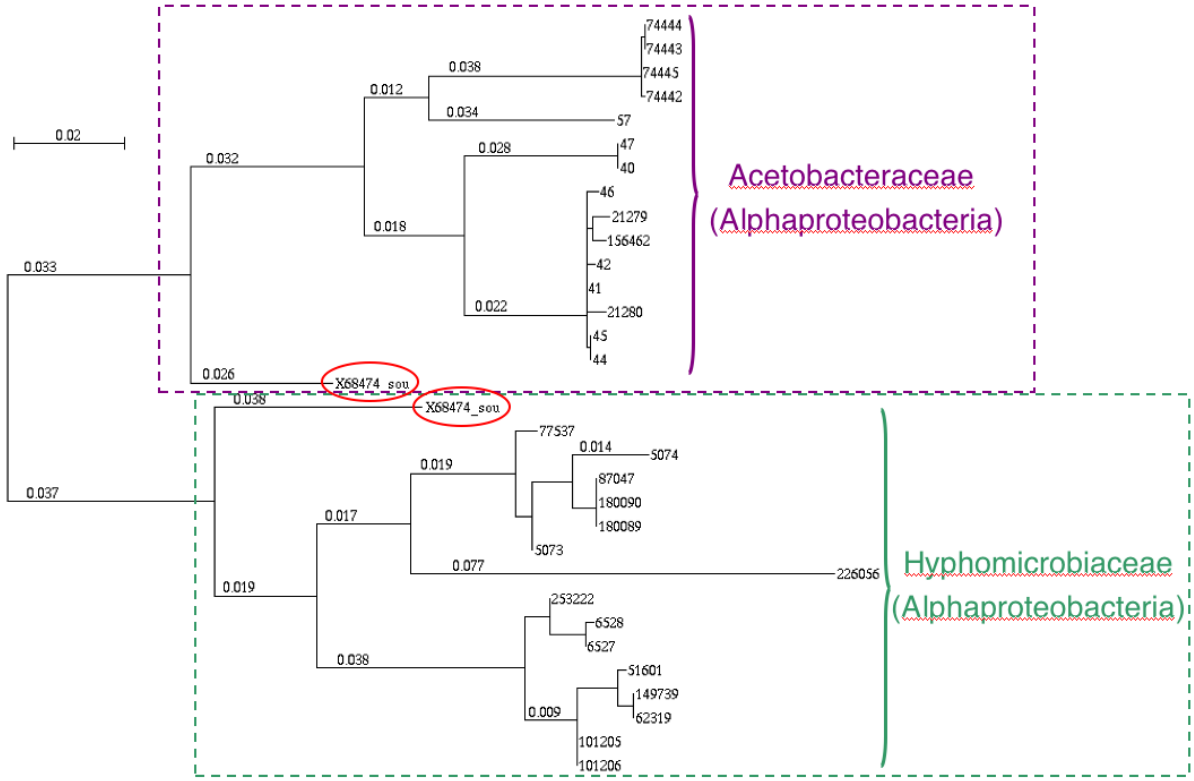

100
Arabidopsis thaliana (cp)
Arabidopsis thaliana (cp)
100
Arabidopsis thaliana (cp)
85
94
Arabidopsis thaliana (cp)
100
Arabidopsis thaliana (cp)
89
Chlorobium tepidum
Deinococcus radiodurans
100
Bradyrhizobium japonicum
44
Rhodopseudomonas palustris
Caulobacter crescentus
71
97
Frankia alni ACN14a
Frankia HFPCcl3
Frankia
100
Frankia EAN1pec
62
Rhizobium meliloti (Sinorhizobium meliloti)
Agrobacterium tumefaciens
100
100
Brucella melitensis
Brucella suis
85
Rhizobium loti (Mesorhizobium loti)
Alpha-
proteobacteria

Figure 3


0.02
0.033
0.037
0.032
0.012
0.038
0.034
0.028
0.018
0.022
0.026
0.038
74444
74443
74445
74442
57
47
40
46
21279
156462
42
41
21280
45
44
X68474_sou
X68474_sou
77537
0.014
5074
0.019
87047
180090
180089
0.017
5073
0.077
226056
0.019
253222
6528
6527
0.038
51601
149739
0.009
62319
101205
101206
Acetobacteraceae
(Alphaproteobacteria)
Hyphomicrobiaceae
(Alphaproteobacteria)