Bioinformatical sequence identification from sequence family databases

Anne-Muriel Arigon, Guy Perrière & Manolo Gouy
Laboratoire de Biométrie et Biologie Evolutive - UMR CNRS 5558 - Université Claude Bernard - Lyon 1 - France
arigon@biomserv.univ-lyon1.fr

Motivation

The classification of a new sequence compared to a sequence collection is useful for:

- species or taxon identification from molecular markers of environmental organisms,
- confrontation of a new sequence to those of a database,
- sequence database update.

⇒ Need of **powerful bioinformatics tools** in order to automate the identification tasks.

Current available tools such as BIBI (Bioinformatic Bacterial Identification - Devulder et al., J. Clin. Microbiol, 41:1785-1787, 2003) cannot be used effectively with databases such as HOVERGEN (Homologous Vertebrate Genes Database) or HOGENOM (Homologous Sequences from Complete Genomes Database) (Perrière et al., Genome Res., 10:379-385, 2000) in which homologous protein gene sequences are clustered into families because:

- the whole family has to be taken into account to add a sequence to a family (comparison should not be restricted to the most similar sequences),
- large families have to be managed (several thousand sequences).

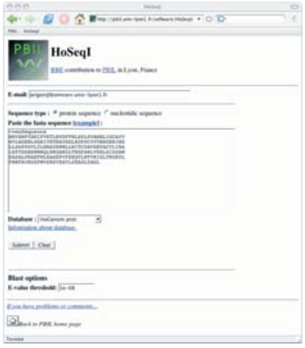
Objectives

Building of a complete environment allowing **the automatic identification of homologous sequences** and their classification inside large sequence databases.

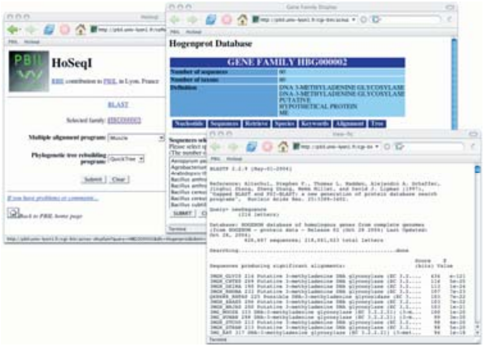
Results

Implementation of a Web application - called **HoSeqI (Homologous Sequence Identification) -** developed in HTML-PHP integrating:

- specifically developed algorithms,
- publicly available programs for similarity search, multiple alignment and phylogeny,
- interface with simple use.

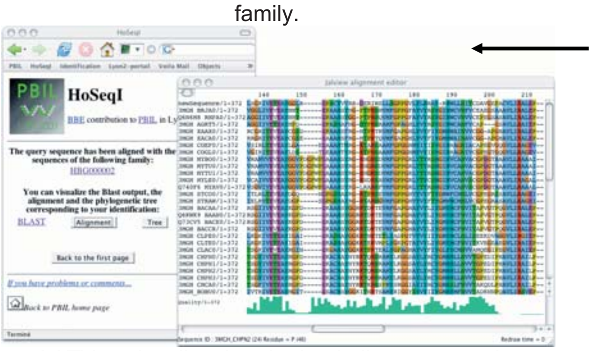


1st stage:
Find the gene family to which the query sequence belongs.

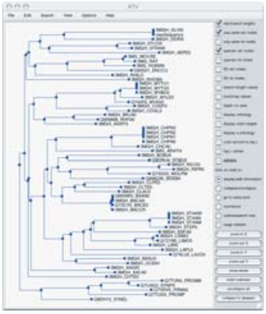


Links to BLAST output and information about proposed families.

2nd stage:
Align the query sequence with sequences of previously selected family.


HoSeqI
The query sequence has been aligned with the sequences of the following family:
HBG000002
BLAST
Alignment
Tree
Back to the first page

3rd stage:
Rebuild phylogenetic tree of the selected family by including the query sequence.



Use of Java applets Jalview and ATV to visualize the obtained alignments and phylogenetic trees.

Details on algorithms

1st stage: search of families to which the query sequence belongs

- BLAST: comparison of the query sequence with the database chosen by the user.
- BLAST results analysis: determination of families and BLAST scores of the most similar sequences of BLAST output.
- For each identified family, calculation of a weighted average of the scores.
- Selection of families with the highest average score and families with non-overlapping matches.

2nd stage: alignment

- Fast and powerful multiple alignment programs allowing:
  - to align a large number of sequences,
  - to gradually add a sequence to existing alignment.
- Various lists of programs according to the selected family: CLUSTAL W, MULTALIN (Corpet, Nucleic Acids Res., 16:10881-10890, 1988), MERITALIGN (Dufayard, PhD Thesis, 2004), MUSCLE (Edgar, Nucleic Acids Res., 32(5):1792-97, 2004), MABIOS (Abdeddaim, Int. J. Artif. Intell. Tools, 6:179–192, 1997).

3rd stage: phylogenetic tree building

- Fast phylogenetic tree building program allowing to locate approximately the query sequence in the selected family: QUICKTREE (Howe et al., Bioinformatics, 18(11):1546-1547, 2002).
- QUICKTREE phylogenetic tree is rooted at its midpoint.