Tree pattern matching

Applied to phylogenetic trees

Jean-François Dufayard
Laurent Duret
Guy Perrière
François Rechenmann

INRIA
RHÔNE-ALPES

UMR 5558
Biométrie, biologie
évolutive.

Introduction

- Comparative sequence analysis is a powerful approach to understand genome evolution and is widely used to predict the function of genes.

- This approach requires a phylogenetic analysis to distinguish orthologous and paralogous genes.

- To simplify such phylogenomic analyses, we have developped two databases of homologous genes: HOVERGEN (vertebrates), HOBACGEN (bacteria and archea):

    - ➢ Genes are classified into families (BLASTP).
    - ➢ Multiple alignments and phylogenetic trees are computed for each family.
    - ➢ Taxonomic data from NCBI.
    - ➢ Protein sequences from SwissProt and TrEMBL.
    - ➢ DNA sequences from EMBL.

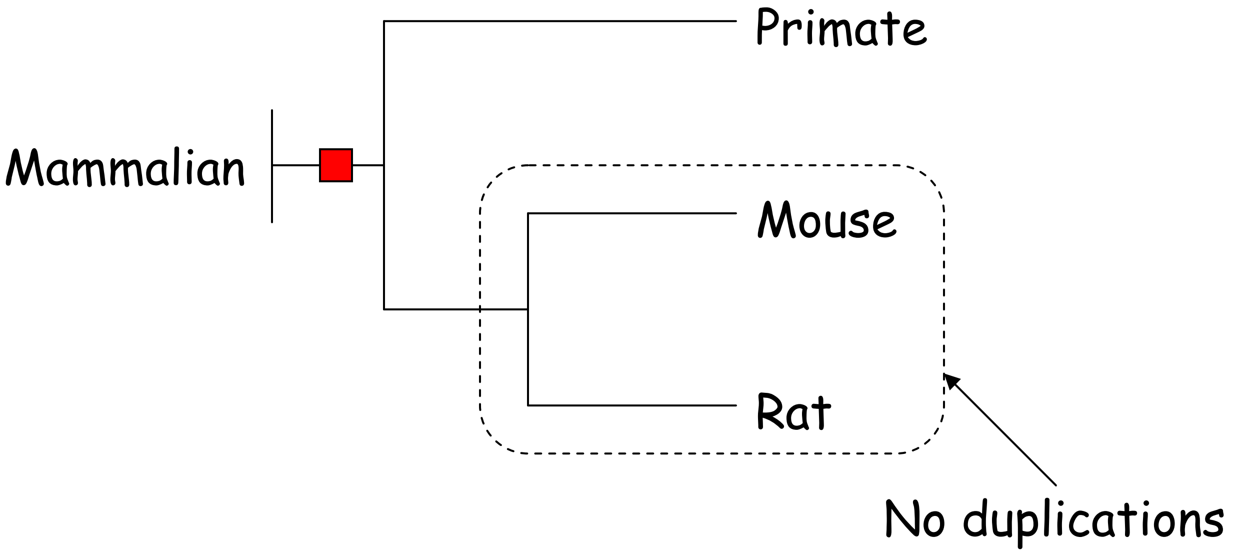
Requesting a phylogenetic tree database

Example

Find 3 genes:

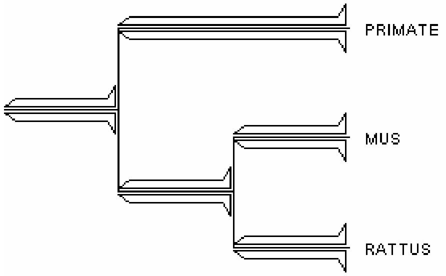
- 1 from primate
- 1 from mouse
- 1 from rat

- Primate gene must be paralogous to others
- Mouse gene and rat gene must be strictly orthologous
- Gene duplication must be posterior to mammalian divergence

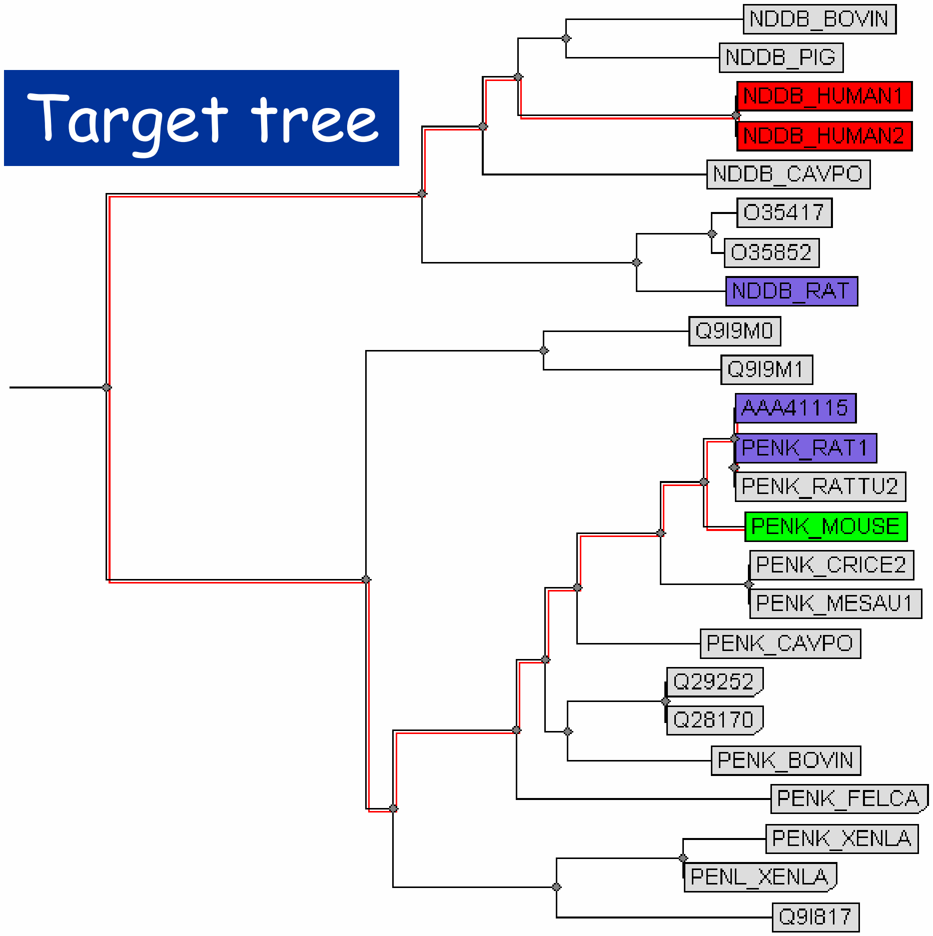


A simple pattern

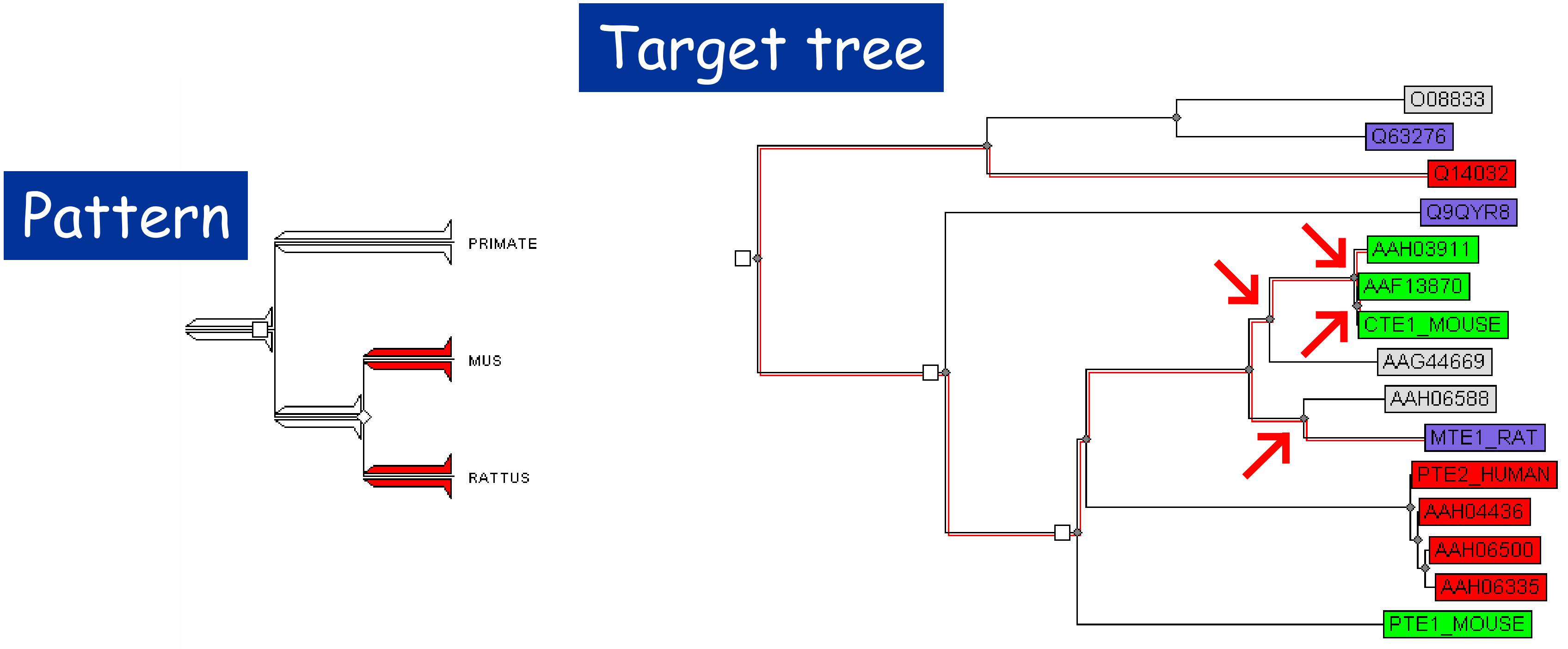
Pattern



Target tree

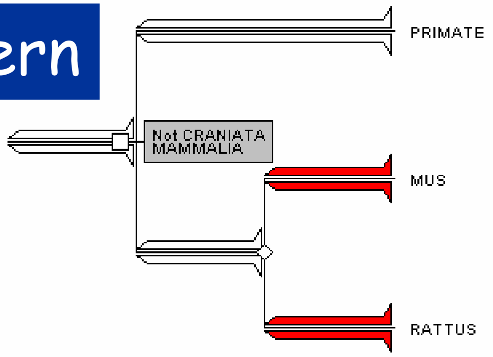


Speciation and duplication nodes

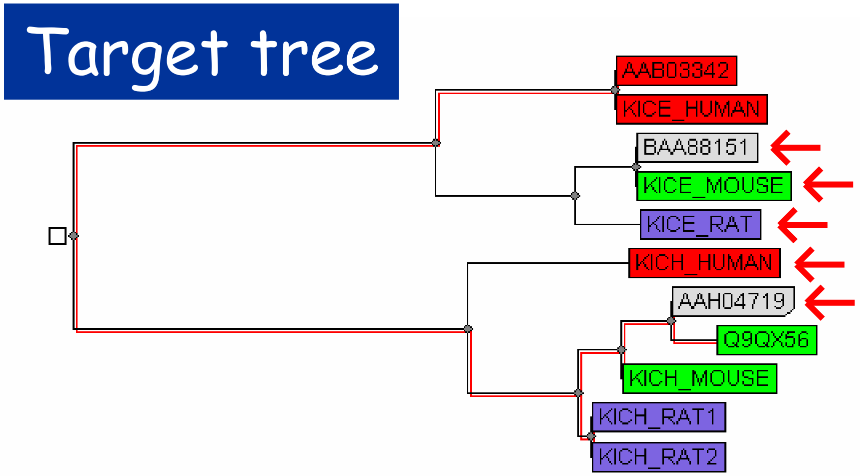


Constraints on subtrees

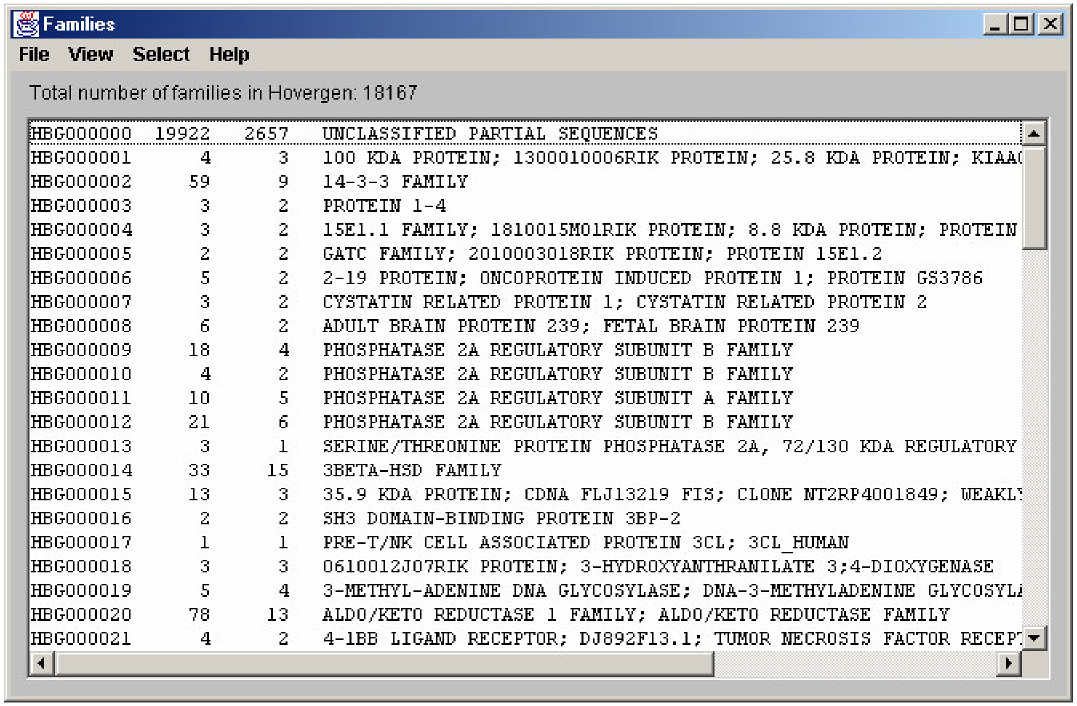
Pattern



Target tree



FAMFETCH

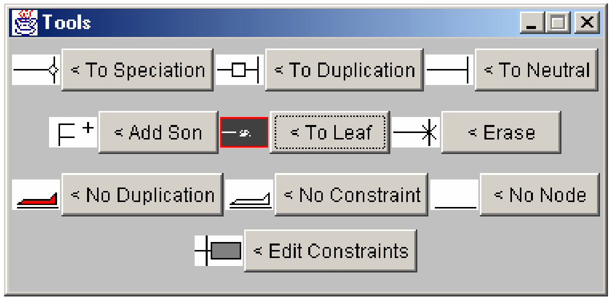


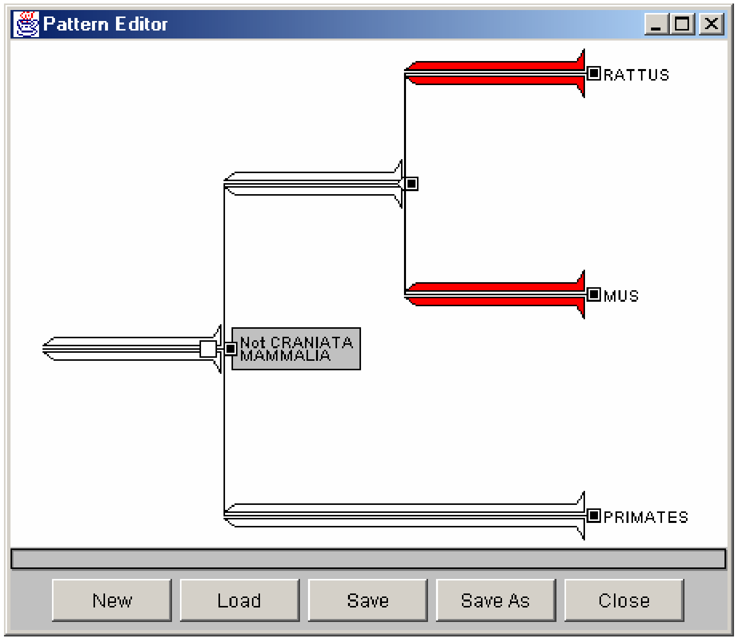
Total number of families in Hovergen: 18167

```
HBG000000  19922   2657   UNCLASSIFIED PARTIAL SEQUENCES
HBG000001      4      3   100 KDA PROTEIN; 1300010006RIK PROTEIN; 25.8 KDA PROTEIN; KIAA(
HBG000002     59      9   14-3-3 FAMILY
HBG000003      3      2   PROTEIN 1-4
HBG000004      3      2   15E1.1 FAMILY; 1810015M01RIK PROTEIN; 8.8 KDA PROTEIN; PROTEIN
HBG000005      2      2   GATC FAMILY; 2010003018RIK PROTEIN; PROTEIN 15E1.2
HBG000006      5      2   2-19 PROTEIN; ONCOPROTEIN INDUCED PROTEIN 1; PROTEIN GS3786
HBG000007      3      2   CYSTATIN RELATED PROTEIN 1; CYSTATIN RELATED PROTEIN 2
HBG000008      6      2   ADULT BRAIN PROTEIN 239; FETAL BRAIN PROTEIN 239
HBG000009     18      4   PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000010      4      2   PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000011     10      5   PHOSPHATASE 2A REGULATORY SUBUNIT A FAMILY
HBG000012     21      6   PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000013      3      1   SERINE/THREONINE PROTEIN PHOSPHATASE 2A, 72/130 KDA REGULATORY
HBG000014     33     15   3BETA-HSD FAMILY
HBG000015     13      3   35.9 KDA PROTEIN; CDNA FLJ13219 FIS; CLONE NT2RP4001849; WEAKLY
HBG000016      2      2   SH3 DOMAIN-BINDING PROTEIN 3BP-2
HBG000017      1      1   PRE-T/NK CELL ASSOCIATED PROTEIN 3CL; 3CL_HUMAN
HBG000018      3      3   0610012J07RIK PROTEIN; 3-HYDROXYANTHRANILATE 3;4-DIOXYGENASE
HBG000019      5      4   3-METHYL-ADENINE DNA GLYCOSYLASE; DNA-3-METHYLADENINE GLYCOSYLA
HBG000020     78     13   ALDO/KETO REDUCTASE 1 FAMILY; ALDO/KETO REDUCTASE FAMILY
HBG000021      4      2   4-1BB LIGAND RECEPTOR; DJ892F13.1; TUMOR NECROSIS FACTOR RECEPT
```

Find each occurrence of the pattern in the database

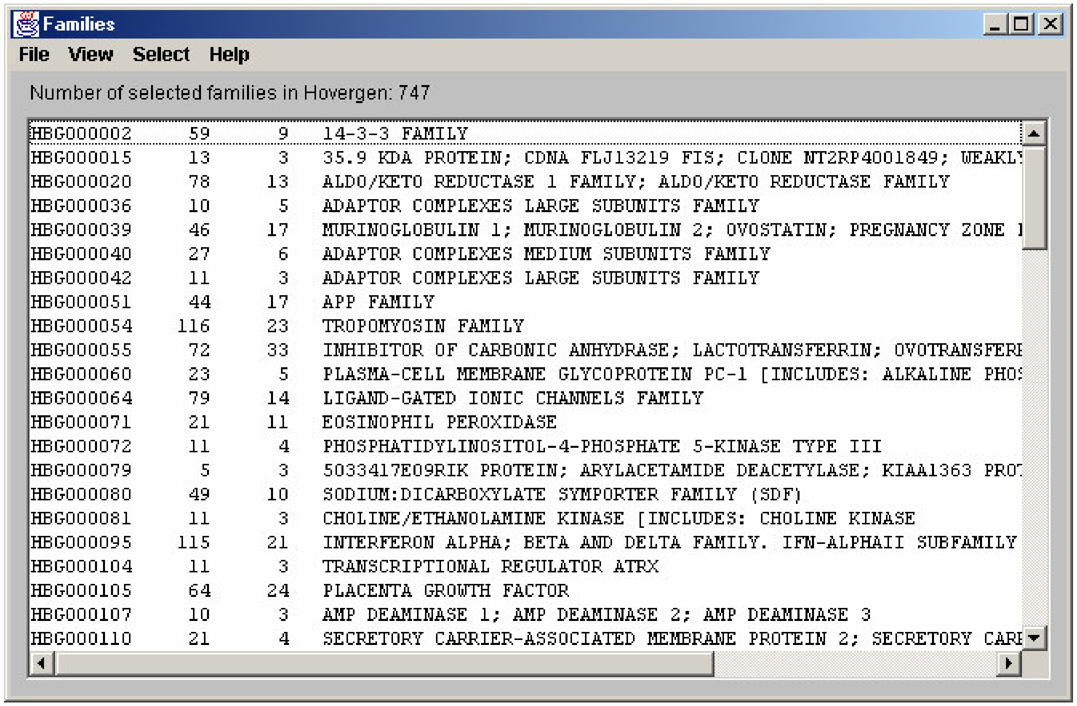
FAMFETCH





Find each occurrence of the pattern in the database

FAMFETCH

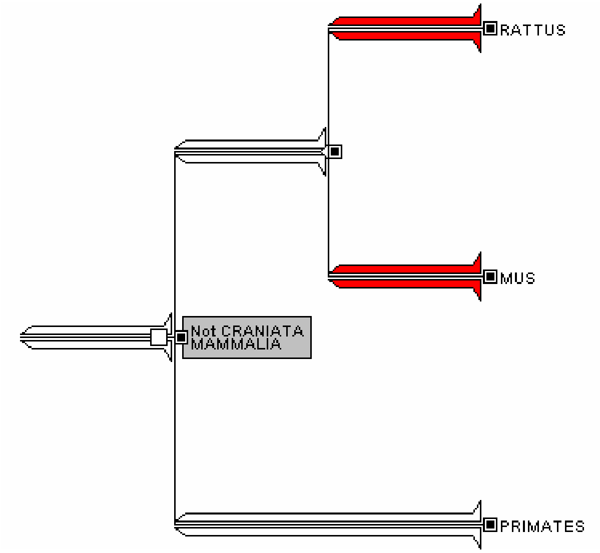


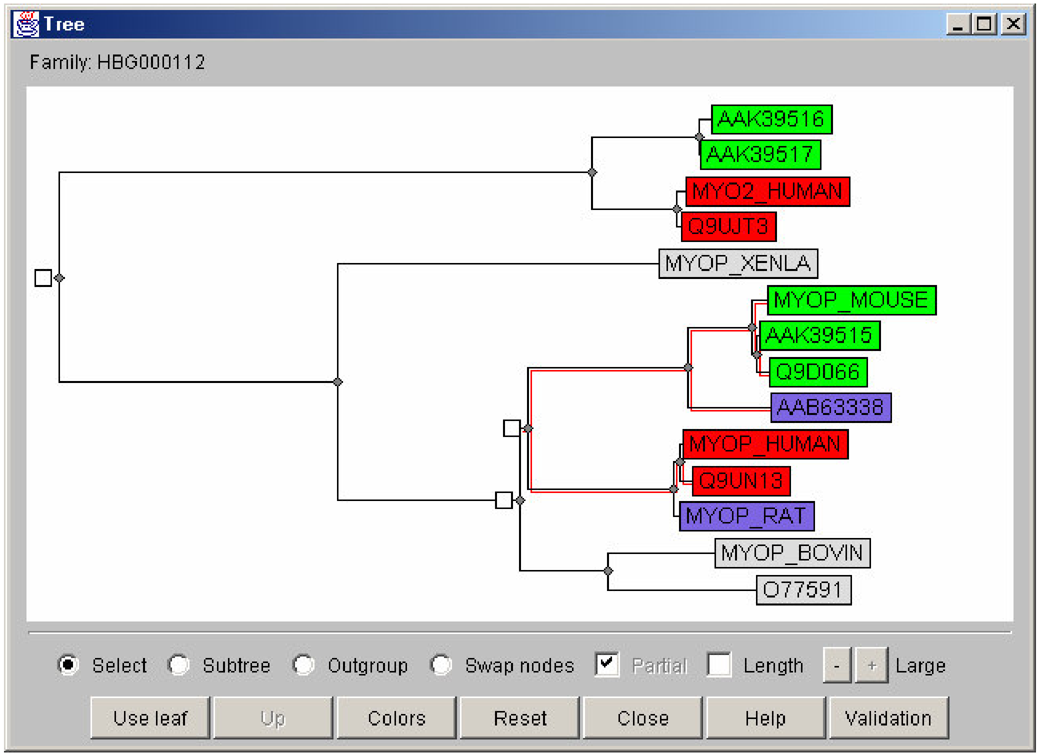
Number of selected families in Hovergen: 747

```
HBG000002      59        9    14-3-3 FAMILY
HBG000015      13        3    35.9 KDA PROTEIN; CDNA FLJ13219 FIS; CLONE NT2RP4001849; WEAKLY
HBG000020      78       13    ALDO/KETO REDUCTASE 1 FAMILY; ALDO/KETO REDUCTASE FAMILY
HBG000036      10        5    ADAPTOR COMPLEXES LARGE SUBUNITS FAMILY
HBG000039      46       17    MURINOGLOBULIN 1; MURINOGLOBULIN 2; OVOSTATIN; PREGNANCY ZONE I
HBG000040      27        6    ADAPTOR COMPLEXES MEDIUM SUBUNITS FAMILY
HBG000042      11        3    ADAPTOR COMPLEXES LARGE SUBUNITS FAMILY
HBG000051      44       17    APP FAMILY
HBG000054     116       23    TROPOMYOSIN FAMILY
HBG000055      72       33    INHIBITOR OF CARBONIC ANHYDRASE; LACTOTRANSFERRIN; OVOTRANSFERR
HBG000060      23        5    PLASMA-CELL MEMBRANE GLYCOPROTEIN PC-1 [INCLUDES: ALKALINE PHOS
HBG000064      79       14    LIGAND-GATED IONIC CHANNELS FAMILY
HBG000071      21       11    EOSINOPHIL PEROXIDASE
HBG000072      11        4    PHOSPHATIDYLINOSITOL-4-PHOSPHATE 5-KINASE TYPE III
HBG000079       5        3    5033417E09RIK PROTEIN; ARYLACETAMIDE DEACETYLASE; KIAA1363 PROT
HBG000080      49       10    SODIUM:DICARBOXYLATE SYMPORTER FAMILY (SDF)
HBG000081      11        3    CHOLINE/ETHANOLAMINE KINASE [INCLUDES: CHOLINE KINASE
HBG000095     115       21    INTERFERON ALPHA; BETA AND DELTA FAMILY. IFN-ALPHAII SUBFAMILY
HBG000104      11        3    TRANSCRIPTIONAL REGULATOR ATRX
HBG000105      64       24    PLACENTA GROWTH FACTOR
HBG000107      10        3    AMP DEAMINASE 1; AMP DEAMINASE 2; AMP DEAMINASE 3
HBG000110      21        4    SECRETORY CARRIER-ASSOCIATED MEMBRANE PROTEIN 2; SECRETORY CARR
```

Find each occurrence of the pattern in the database

FAMFETCH





Tree

Family: HBG000112

● Select ○ Subtree ○ Outgroup ○ Swap nodes ☑ Partial □ Length

Find each occurrence of the pattern in the database

CONCLUSION

- This system allows to request HOVERGEN and HOBACGEN using phylogenetic criteria

- 9800 trees in HOVERGEN, or 11500 tree in HOBACGEN can be explored with FAMFETCH, which remains an interactive application.

- Available in june 2002 as a new release of FAMFETCH:
  http://pbil.univ-lyon1.fr/hobacgen/client.html