Assemblage rationnel d'une collection de séquences

Guy Perrière

Pôle Bioinformatique Lyonnais
Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558

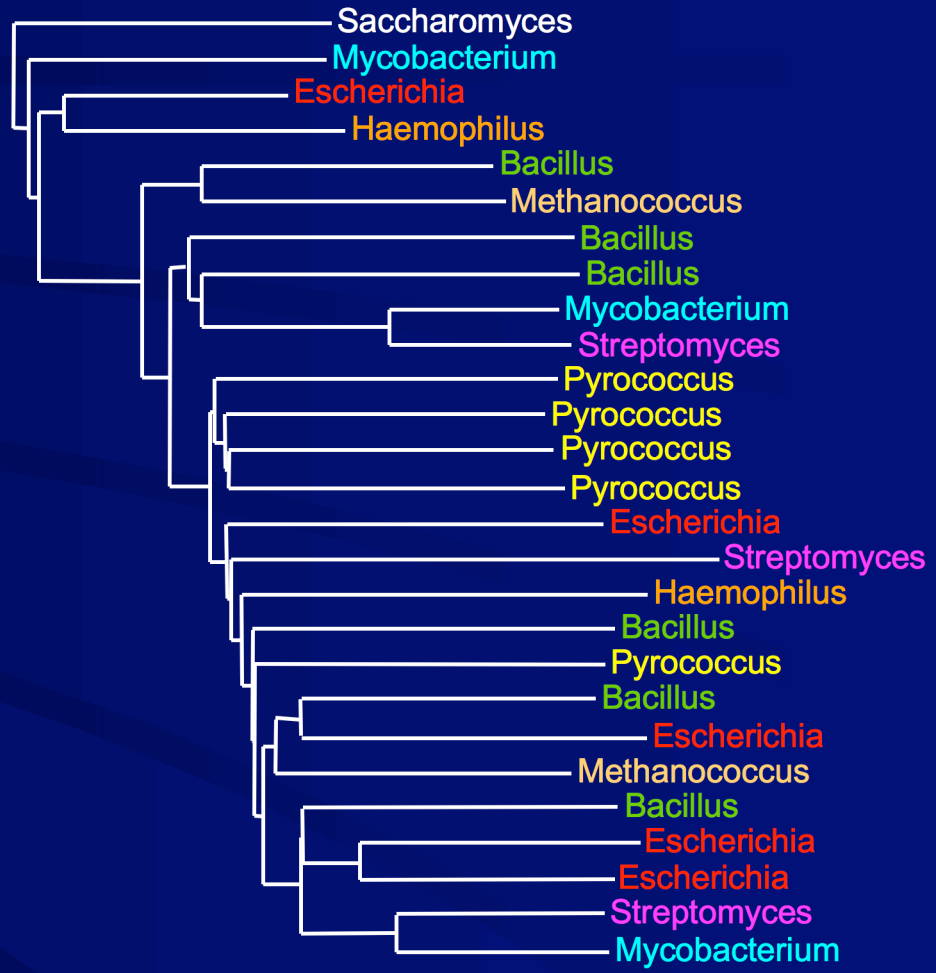
perriere@biomserv.univ-lyon1.fr

Étapes d'une phylogénie

- Récupération d'un ensemble de séquences homologues :
  - Recherche de similarités dans les banques à partir d'une séquence (FASTA, BLAST, Smith-Waterman).
- Alignement des séquences homologues :
  - Programmes d'alignements multiples (CLUSTAL, MUSCLE, DIALIGN).
- Construction de la phylogénie proprement dite :
  - Méthodes de reconstruction (parcimonie, distances, maximum de vraisemblance, méthodes bayesiennes).

Le problème des paralogues

- Les duplications de gènes sont nombreuses.
- Des duplications peuvent être suivies de pertes :
  - Existence de paralogies cachées.
- Un arbre fait avec des paralogues (gènes du-pliqués) ne reflète pas l'arbre des espèces.
- Il est nécessaire d'identi-fier les orthologues.



Aminotransférases pyridoxal-phosphate dépendantes (III)

Serveurs BLAST

- Il existe un grand nombre de serveurs permettant d'effectuer des recherches BLAST mais…
  - Toutes les options ne sont pas toujours accessibles.
  - Peu sont exhaustifs du point de vue des banques.
  - Tous ne permettent pas d'accéder à des banques mises à jour quotidiennement.
  - Les possibilités de filtrage pré- ou post-recherche sont rares et limitées.
  - Généralement pas de lien directs avec d'autres appli-cations (*e.g.*, alignements multiples).

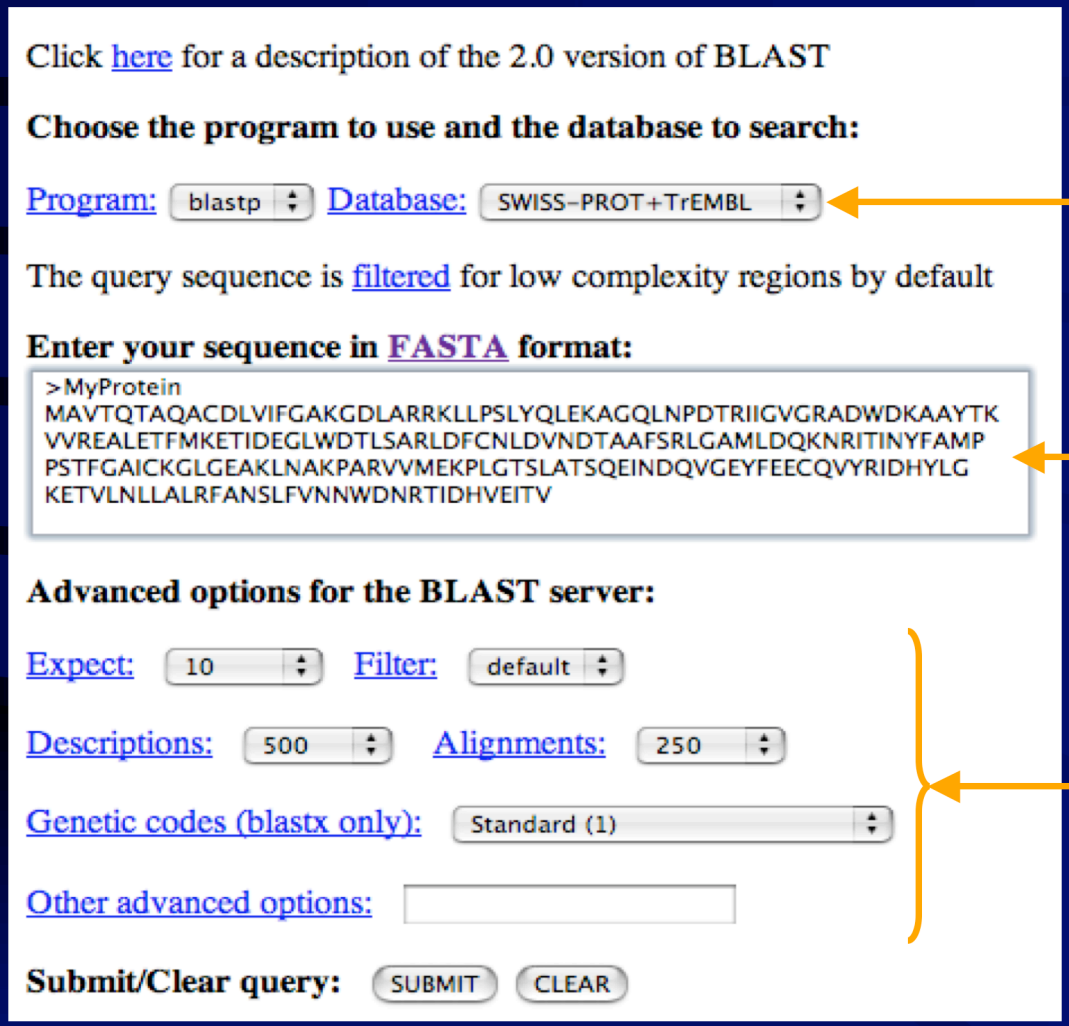
BLAST au NCBI

- Répond à (presque) toutes les questions précédentes.
- Est particulièrement rapide.
- Bénéficie d'une interface graphique de visualisation des résultats.
- Mais…
  - Est de ce fait très sollicité !
  - Toujours pas de tuyauterie vers l'étape suivante (construction d'un alignement multiple).

BLAST au PBIL

- Répond à (presque) toutes les questions précédentes.
- Bénéficie d'une tuyauterie vers des programmes d'alignement (et de prédiction de structures des protéines).
- Mais…
  - N'est pas toujours très rapide.
  - Ne possède pas l'interface de visualisation des similarités.

Paramètres généraux

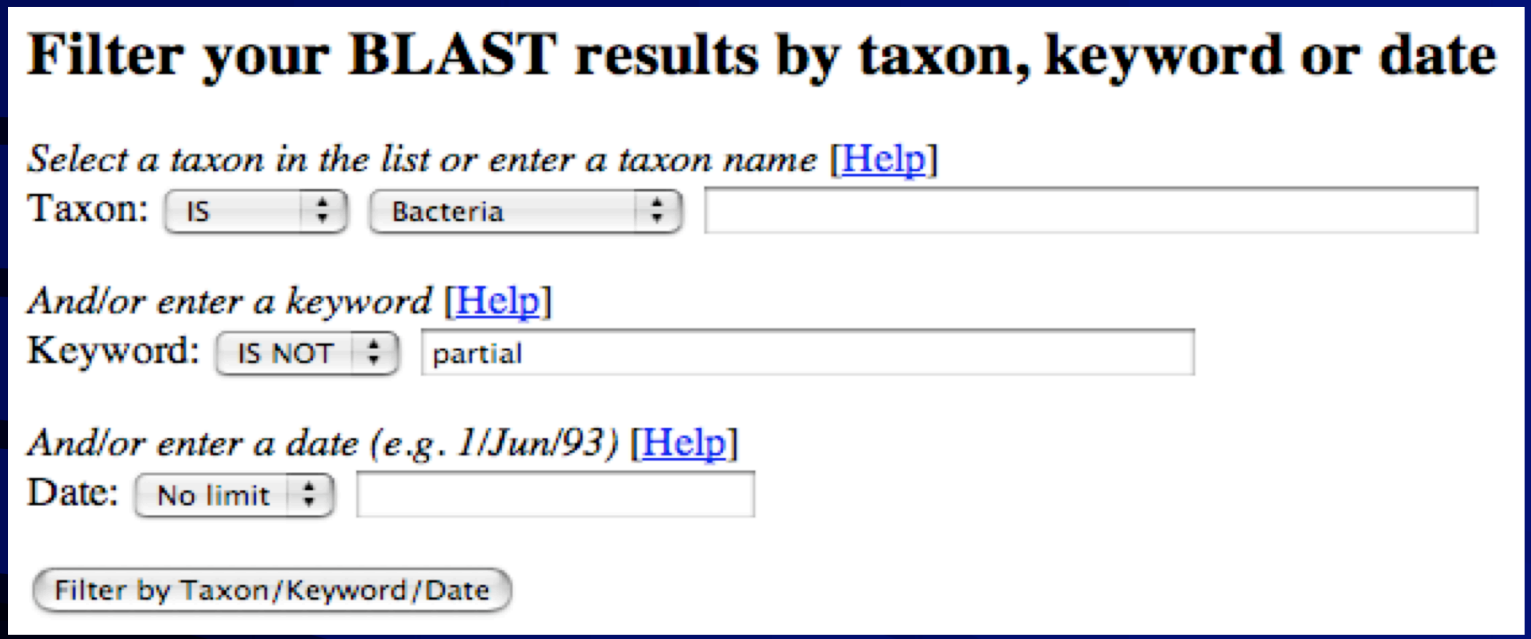


Choix de la banque

Séquence requête

Paramètres de BLAST

Filtrage des résultats



Sélection de séquences

Or select sequences in the list below

Get selected sequences in FASTA format or as a list for further analyses: FASTA

Query= MyProtein
(213 letters)

Database: sptrembl: Non-redundant SWISSPROT + TREMBL database (from UniProt Rel. 6 (SWISS-PROT 48 + TrEMBL 31): Last Updated: Nov 2, 2005)

2,470,253 sequences; 805,771,824 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value
G6PD_ECOLI 491 Glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49)..	428	e-119
Q3Z2K3_SHISO 491 Glucose-6-phosphate dehydrogenase.	427	e-118
Q8XCJ6_ECO57 491 Glucose-6-phosphate 1-dehydrogenase (G6PD).	426	e-118
Q7UAF3_SHIFL 491 Glucose-6-phosphate dehydrogenase.	425	e-118
Q8ZNW2_SALTY 491 Glucose-6-phosphate dehydrogenase (EC 1.1.1.49).	408	e-113
Q8Z5X1_SALTI 491 Glucose 6-phosphate dehydrogenase (EC 1.1.1.49).	408	e-113
Q5PN04_SALPA 491 Glucose 6-phosphate dehydrogenase.	408	e-113
Q57NB4_SALCH 491 Glucose-6-phosphate dehydrogenase.	406	e-112
Q6D4B4_ERWCT 491 Glucose-6-phosphate 1-dehydrogenase (G6PD).	367	e-100
G6PD_ERWCH 491 Glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49)..	360	1e-98

Références croisées

Alignements multiples

Sequence list "list":

Edit **Modify** **Cross Refs** **Retrieve** **Analyze** **View** **History**

Query = MyProtein

G6PD_ECOLI
Q3Z2K3_SHISO
Q8XCJ6_ECO57
Q7UAF3_SHIFL
Q8ZNW2_SALTY

SHOW alignment with width : 60

all residues.

Conservation level of 0 % (works with "using conservation level of" option in list above, negative value will hide).

View CLUSTALW in: [MPSA (Mac, UNIX) , About...] [AnTheProt (PC) , Download...] [HELP]

```
                                550       560       570       580       590       600
                                 |         |         |         |         |         |
Q4NXP9_9DELT    G-WRRVVIEKPFGRDLDSAVALNRELASVVREEQIYRIDHYLGKETVQNLMVFRFANGIF
Q5E0A8_VIBF1    G-WKRLIIEKPFGYDLASALELGKQIHQYFEEHQIYRIDHYLGKETVQNLLVLRFSNVMF
Q8NQ63_CORGL    A-WRRVIIEKPFGHNLESAHELNQLVNAVFPESSVFRIDHYLGKETVQNILALRFANQLF
Q6M517_CORGL    A-WRRVIIEKPFGHNLESAHELNQLVNAVFPESSVFRIDHYLGKETVQNILALRFANQLF
Q8FT74_COREF    S-WRRVIIEKPFGNDLKTAHELNETVNAVFPEESVFRIDHYLGKETVQNILALRFANQLF
Q6NH41_CORDI    S-WRRVIVEKPFGHDQESARQLNNLINSVFPEKSVFRIDHYLGKETVQNIMALRFANQLF
G6PD_MYCTU      R-WSRVVIEKPFGHDLASARELNKAVNAVFPEEAVFRIDHYLGKETVQNILALRFANQLF
G6PD_MYCBO      R-WSRVVIEKPFGHDLASARELNKAVNAVFPEEAVFRIDHYLGKETVQNILALRFANQLF
Q741B6_MYCPA    R-WSRVVIEKPFGHDLESAQSLNKAVNAVFPEESVFRIDHYLGKETVRNILALRFANQLF
Q4NEB6_9MICC    K-WRRVVIEKPFGHDLESARQLNDIVESVFPPDAVFRIDHYLGKETVQNILALRFANQLF
Q8G732_BIFLO    A-WRRVIIEKPFGHDLASAKELDSVVSEVFDPSSVFRIDHYLGKETVQNLLALRFANAMY
Q47ND1_THEFY    S-WRRVVIEKPFGHDLASAQELNALVDDVFPPHSVFRIDHYLGKETVQNLLALRFSNTLF
Q3WC06_9ACTO    GGWRRVVIEKPFGHDLESARELNSLVDDVFTPDDVFRIDHYLGKETVQNLFALRFANTLF
Q4C2R2_CROWT    K--HRIVIEKPFGKDLSSAQVLNRVVQNVCKEEQVYRIDHYLGKETVQNLLVFRFANAIF
G6PD_SYNY3      K--SRIVIEKPFGRDLSSAQSLNRVVQSVCKENQVYRIDHYLGKETVQNLMVFRFANAIF
Q8DLF3_SYNEL    K--HRLVIEKPFGRDLTSAQALNEVVQSVCKEEQVYRIDHYLGKETVQNLMVFRFANAIF
Q7UVB9_RHOBA    P--RRVIIEKPFGTDLKTAQRLNESIHNVFREDQIYRIDHYLGKETVQNIFALRFANSIF
```

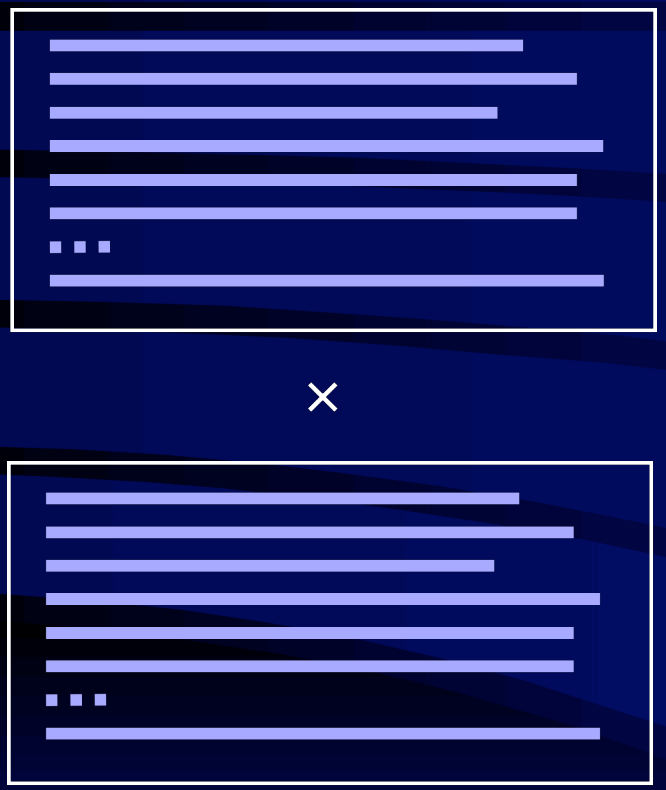
Banques de familles de gènes

- Automatisation des procédures :
  - Pour la recherche de similarités entre les pro-téines.
  - Pour le regroupement en familles de gènes homologues.
  - Pour le calcul des alignements et des arbres.
- Caractéristiques communes :
  - Séquences protéiques et nucléiques.
  - Alignements multiples et arbres.
  - Données taxonomiques.

Les banques du PBIL

- Banques généralistes, regroupant des séquences provenant d'un grand nombre d'espèces :
  - HOVERGEN (séquences de vertébrés d'UniProt).
  - HOGENOM (protéomes complets d'UniProt).
  - HOMOLENS (protéomes complets d'Ensembl).
- Banques spécialisées, dédiées à des familles d'interêt :
  - HOPPSIGEN (pseudogènes humains et de la souris présent dans Ensembl).
  - NuReBase (récepteurs nucléaires de GenPept).
  - RTKdb (récepteurs à Tyrosine Kinase d'UniProt).

Recherche de similarités



×

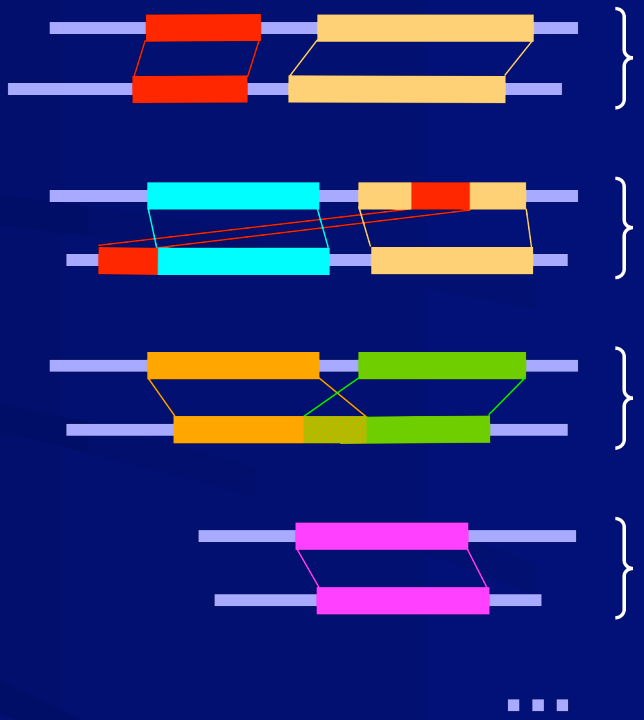
SEG

BLASTP2

BLOSUM62

$E \leq 10^{-4}$

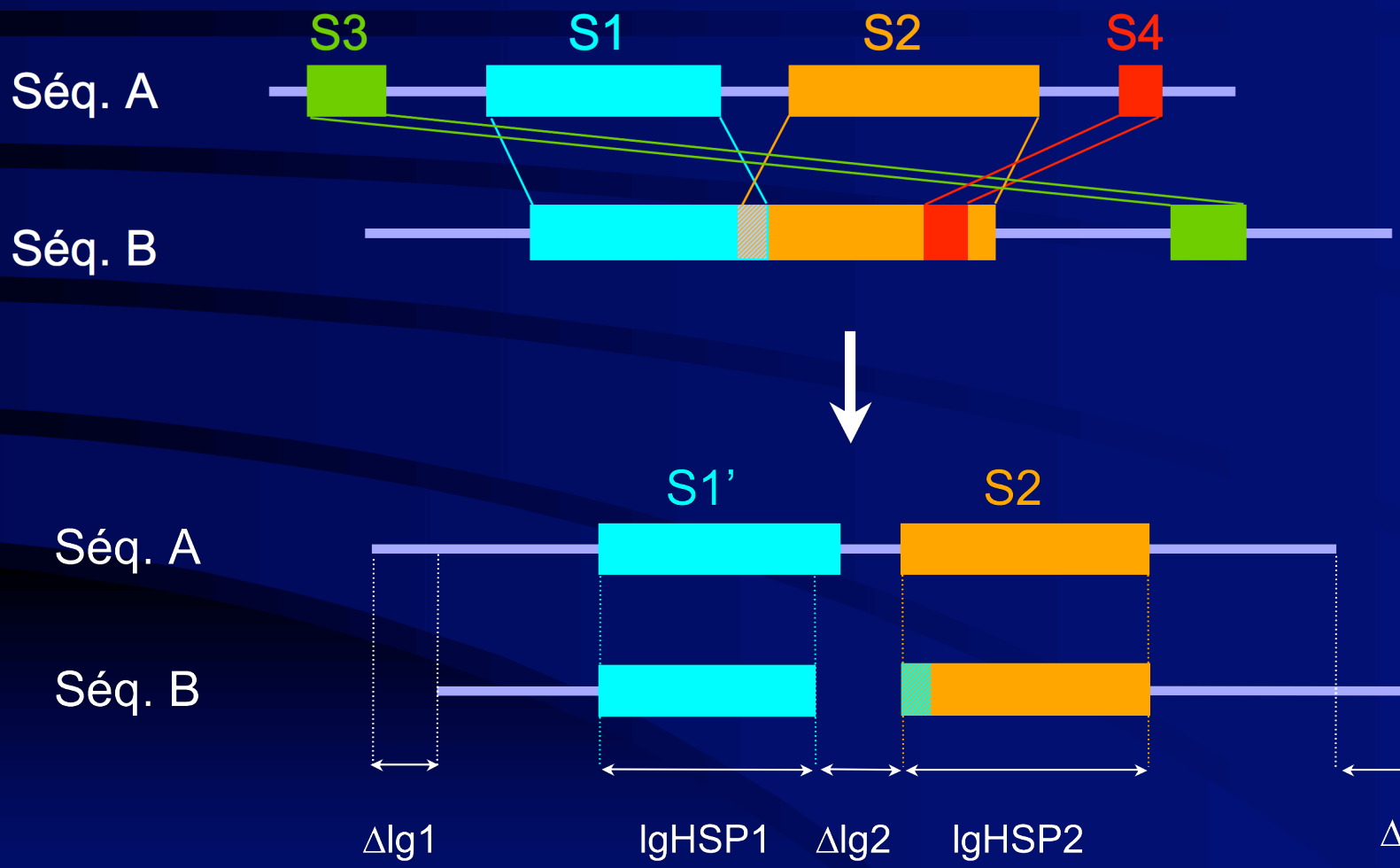
Calcul effectué à l'IN2P3
(ferme de 1000 CPU)



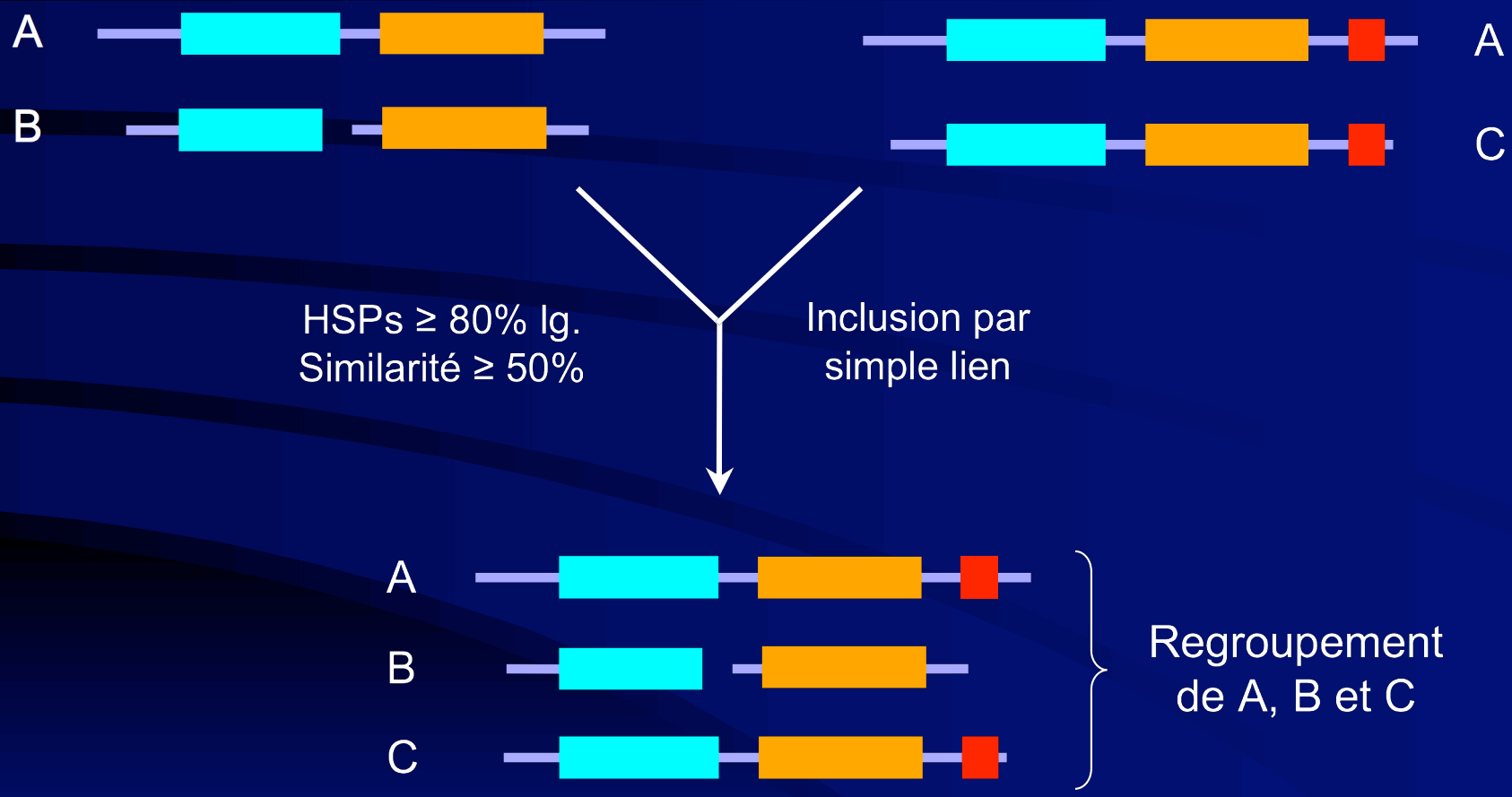
Séquences
protéiques

Alignements locaux
par paires

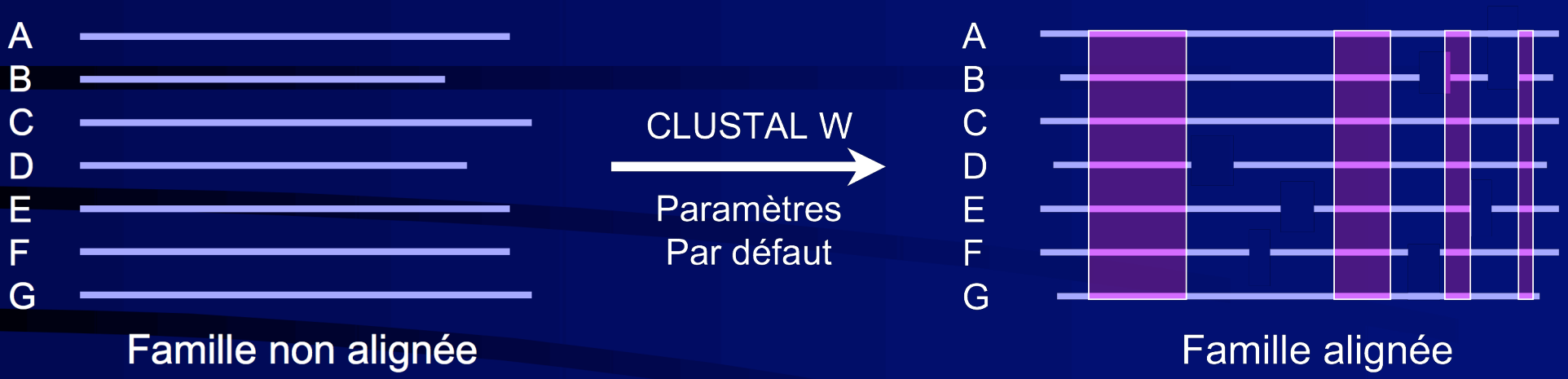
Sélection des segments



Assemblage des familles

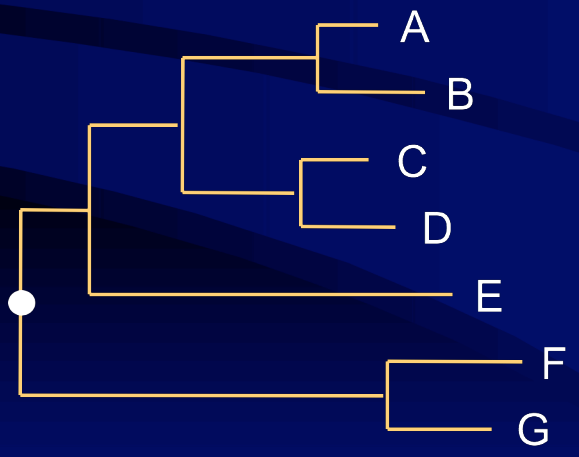


Alignements et arbres (1)



Famille non alignée

Famille alignée



BIONJ

Divergence
observée

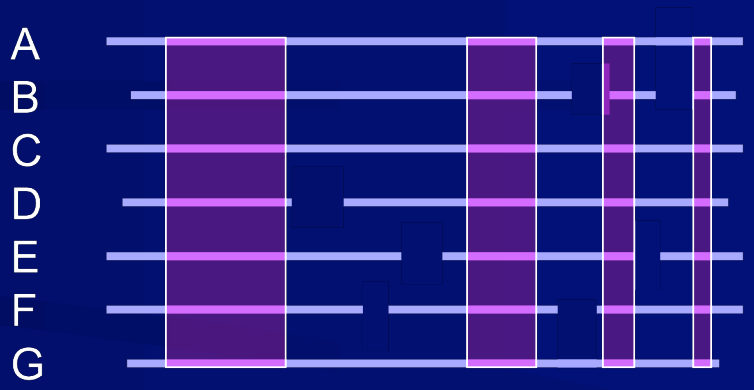
Arbre de la famille

Alignements et arbres (2)

Famille non alignée

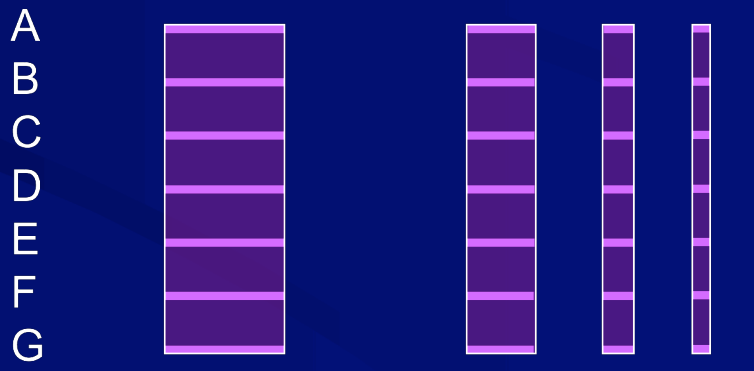
MUSCLE

Paramètres
Par défaut



Famille alignée

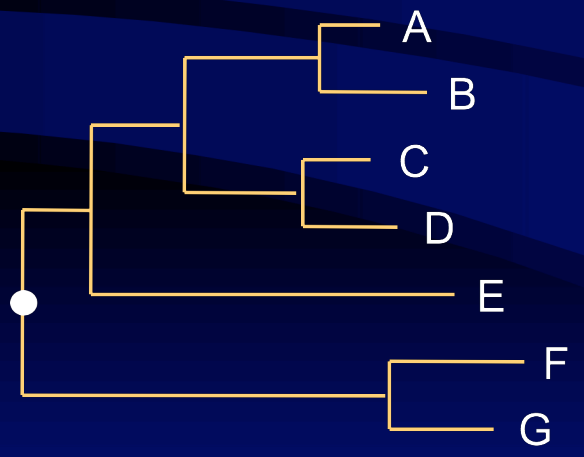
GBLOCKS



Blocs conservés

PHYML

JTT + Γ

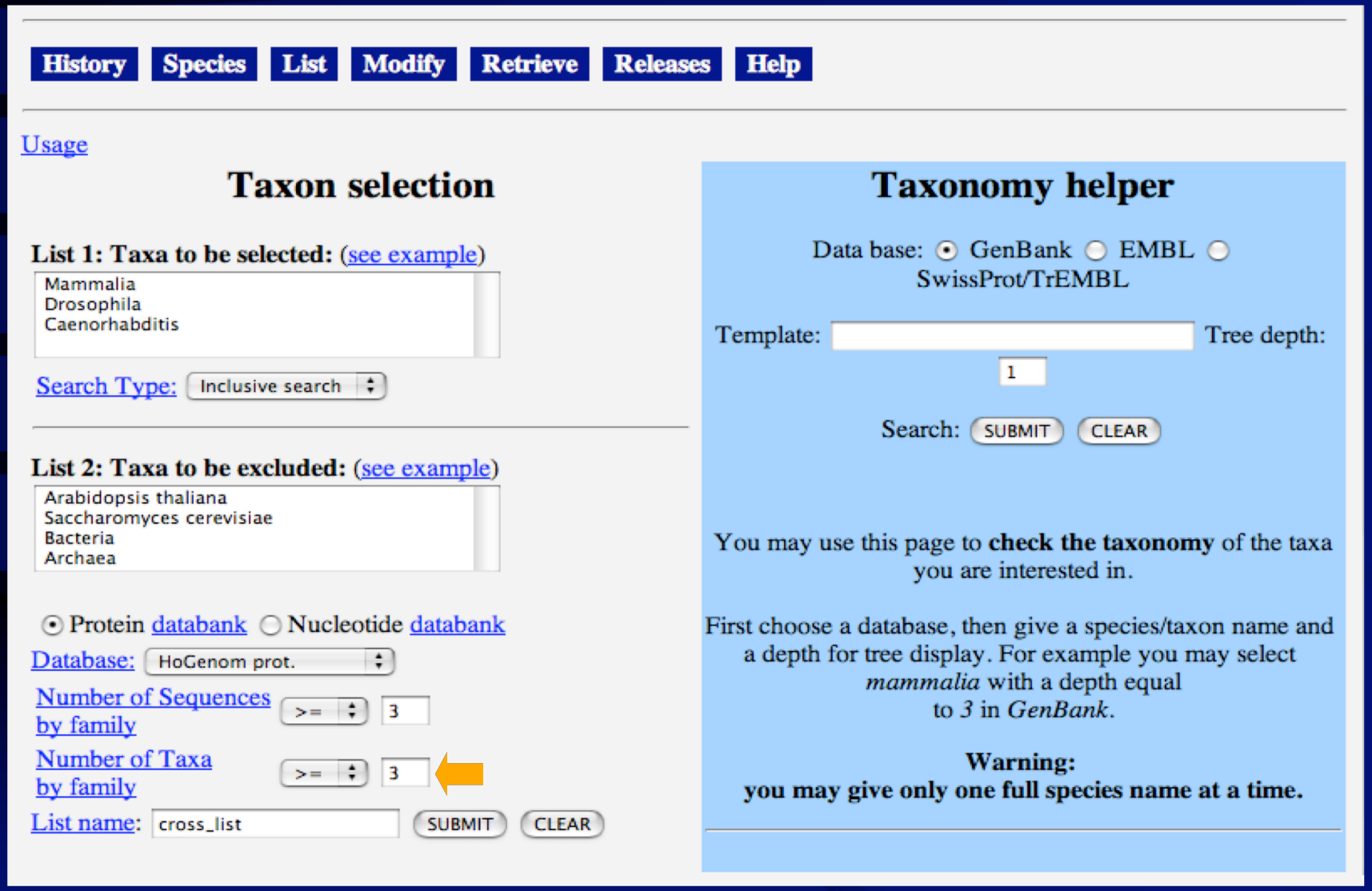


Arbre de la famille

HOGENOM 3.0

- 263 génomes complets.
- 954 455 séquences protéiques dont :
  - 7097 séquences partielles non classées.
- 227 950 séquences nucléiques.
- 81 475 familles dont :
  - 70 788 familles comprenant aux moins deux séquences.
  - 1148 familles contenant plus de 100 séquences.

Requêtes sur les familles



Affichage des familles

Family list "cross_list"

History | View

Family	Sequences	Species	Definition
HBG114039	10	5	UPF0046 FAMILY
HBG127257	124	5	FAMILY 1 OF G-PROTEIN COUPLED RECEPTORS. STRONGEST TO THE OTHER 5HT-1 SUBTYPE RECEPTORS FAMILY 1 OF
HBG127994	87	5	LIGAND-GATED IONIC CHANNEL FAMILY
HBG112858	8	4	CG32549-PD; CG32549 protein Cytosolic purine 5'-nucleotidase Hypothetical 64.3 kDa protein Hypoth
HBG215638	48	5	NUCLEAR HORMONE RECEPTOR FAMILY. NR3 SUBFAMILY NUCLEAR HORMONE RECEPTOR FAMILY. NR2 SUBFAMILY
HBG191019	16	5	Clathrin coat assembly protein AP180 Phosphatidylinositol-binding clathrin assembly protein
HBG112887	6		
HBG208967	120		
HBG174990	40		
HBG130840	27		

GENE FAMILY HBG191019

Number of sequences	16
Number of taxons	5
Definition	Clathrin coat assembly protein AP180 Phosphatidylinositol-binding clathrin assembly protein

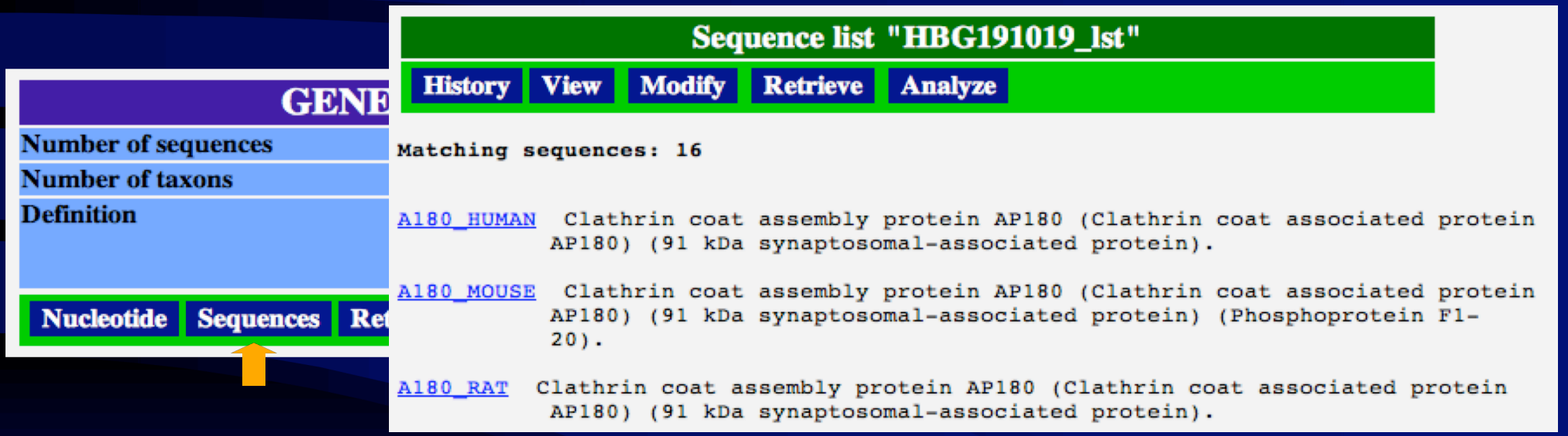
Nucleotide | Sequences | Retrieve | Species | Keywords | Alignment | Tree

Données sur les familles

GENE FAMILY HBG191019	
Number of sequences	16
Number of taxons	5
Definition	Clathrin coat assembly protein AP180 Phosphatidylinositol-binding clathrin assembly protein

Nucleotide | Sequences | Retrieve | Species | Keywords | Alignment | Tree

Données sur les familles

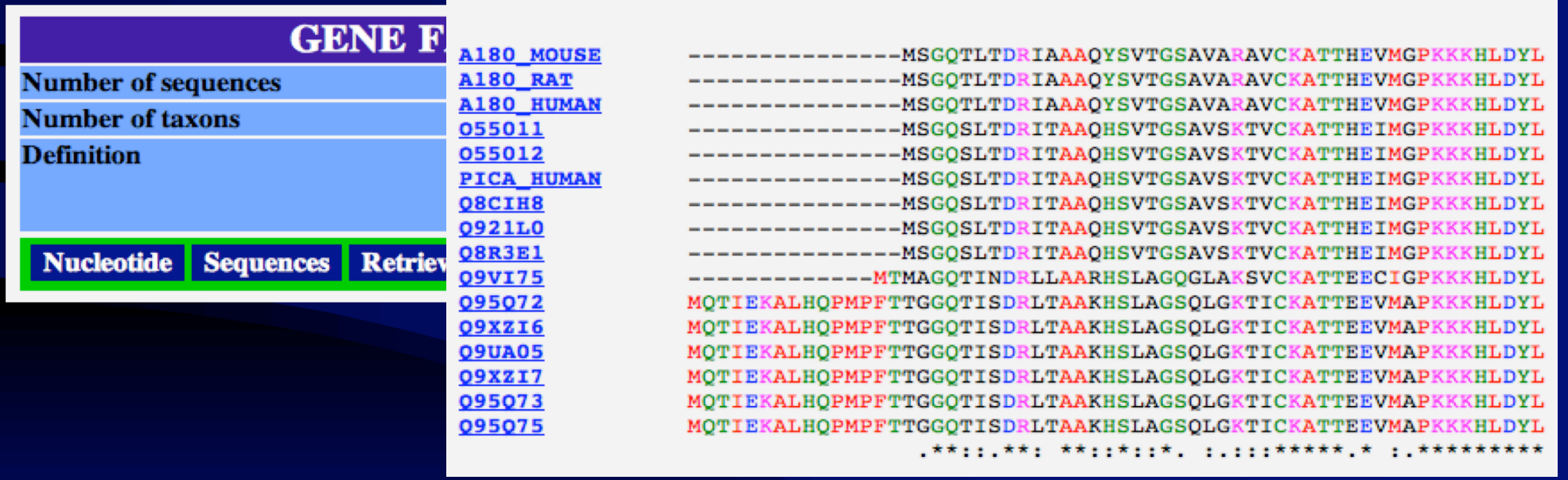


Données sur les familles

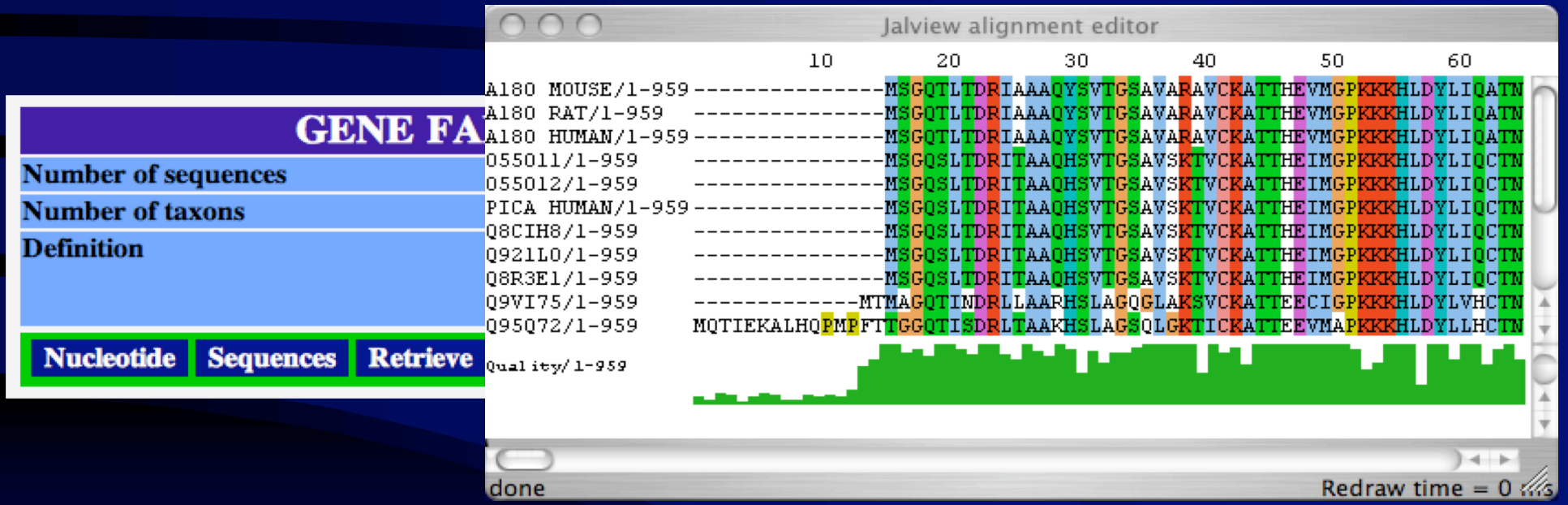
Number of sequences
Number of taxons
Definition
Nucleotide Sequen

Keyword	Number of sequences	Percentage	Link to all associated sequences
HBG191019	16	[100%]	Sequences related to HBG191019 in Hogenprot
DOMAIN	14	[87%]	Sequences related to DOMAIN in Hogenprot
HYPOTHETICAL PROTEIN	8	[50%]	Sequences related to HYPOTHETICAL PROTEIN in Hogenprot
TREMBL-RELEASE 23	7	[43%]	Sequences related to TREMBL-RELEASE 23 in Hogenprot
C32E8.10	6	[37%]	Sequences related to C32E8.10 in Hogenprot
SP-RELEASE 41	4	[25%]	Sequences related to SP-RELEASE 41 in Hogenprot
TREMBL-RELEASE 24	4	[25%]	Sequences related to TREMBL-RELEASE 24 in Hogenprot
VARSPLIC	3	[18%]	Sequences related to VARSPLIC in Hogenprot

Données sur les familles



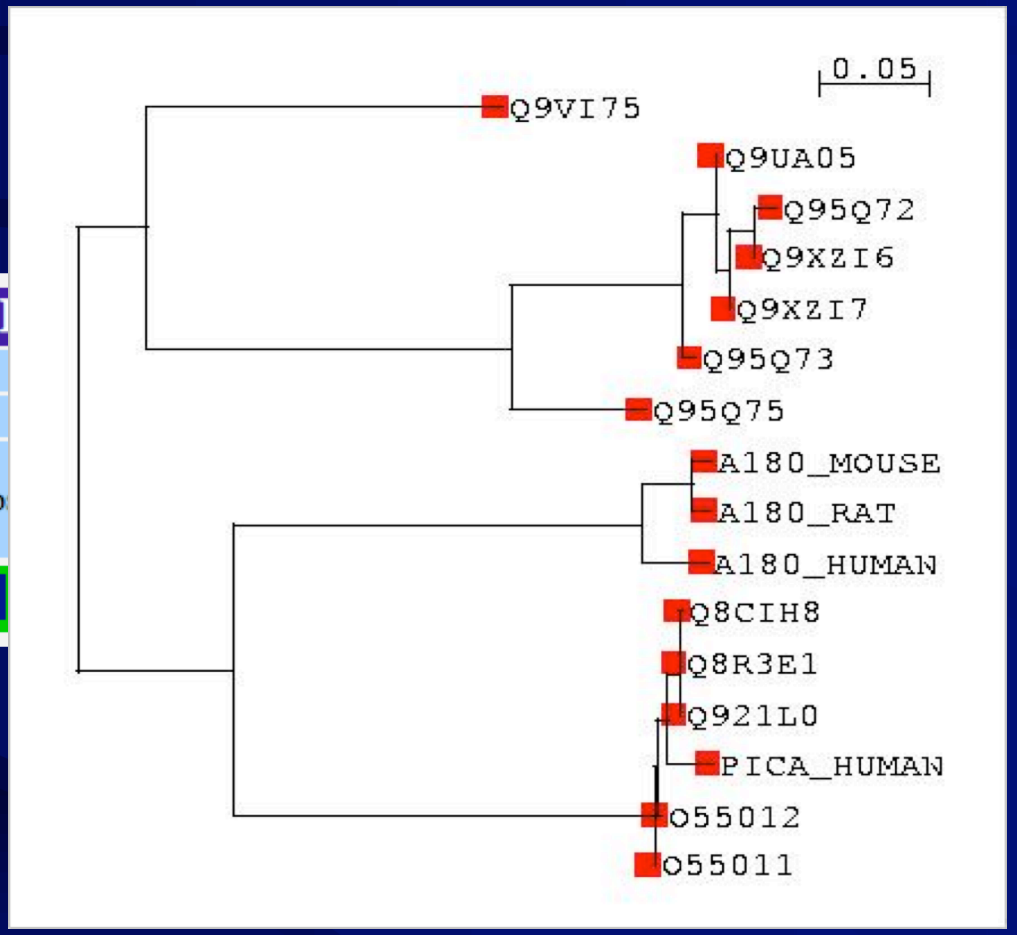
Données sur les familles



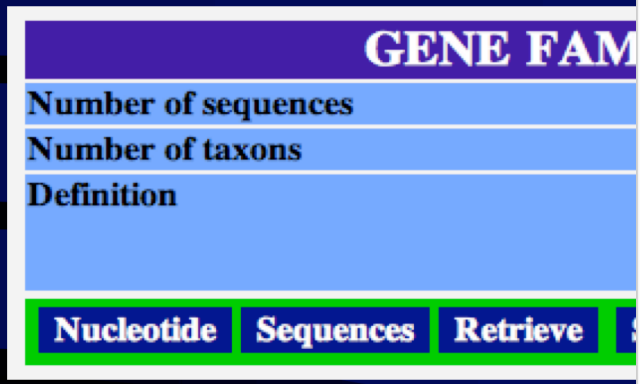
Données sur les familles

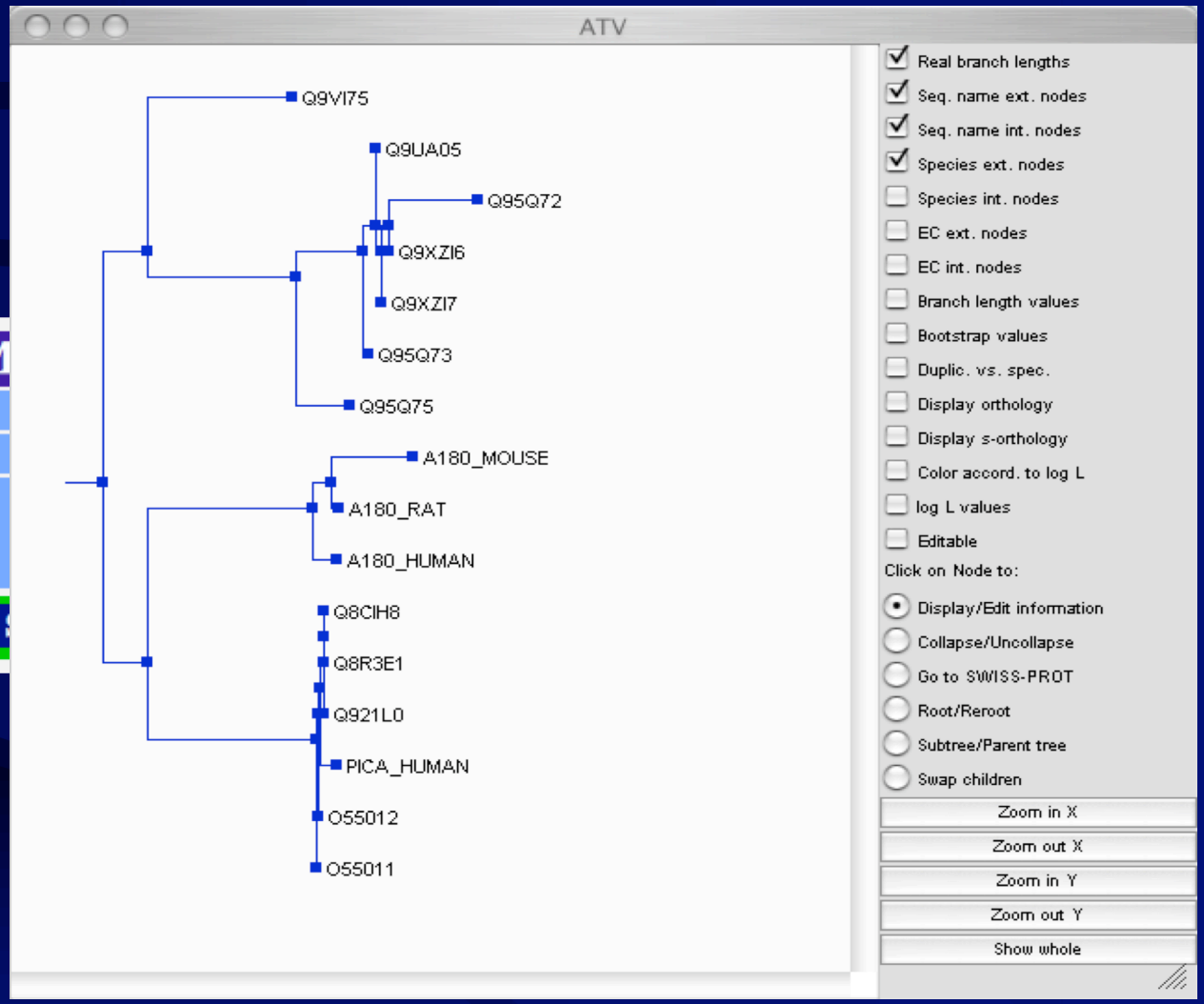
GENE FAMILY	
Number of sequences	
Number of taxons	
Definition	Pho

Nucleotide | Sequences | Retrieve | Species

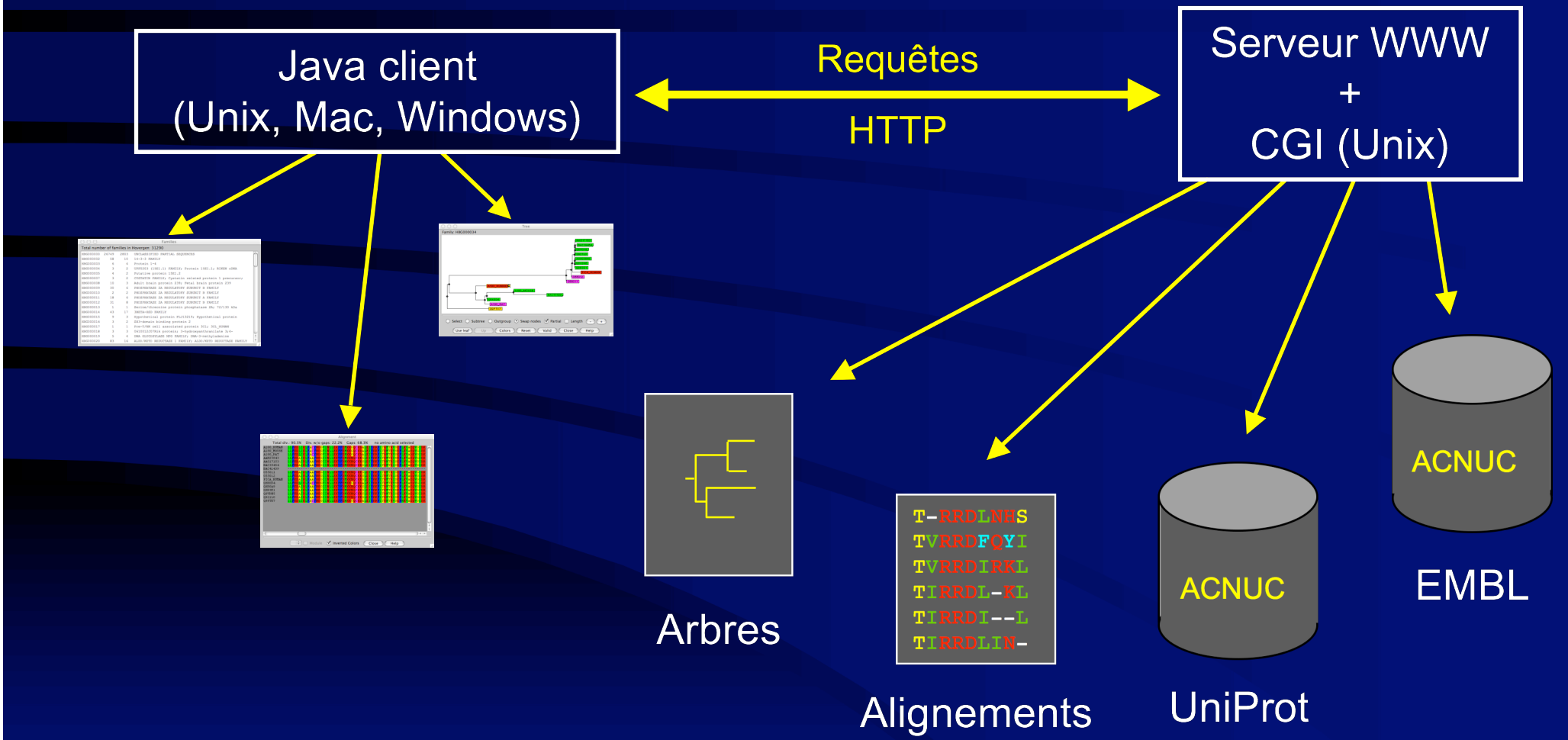


Données sur les familles

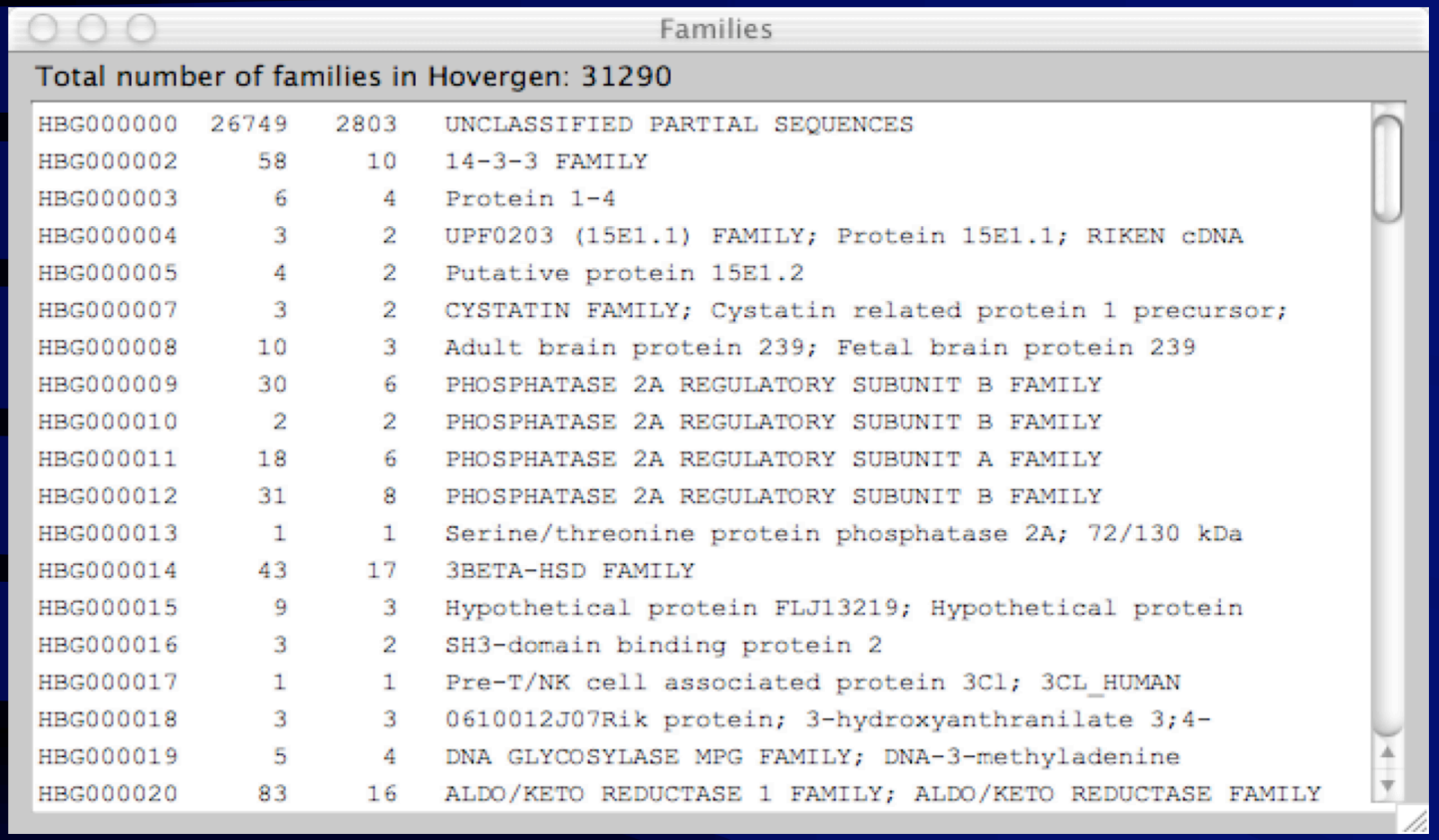




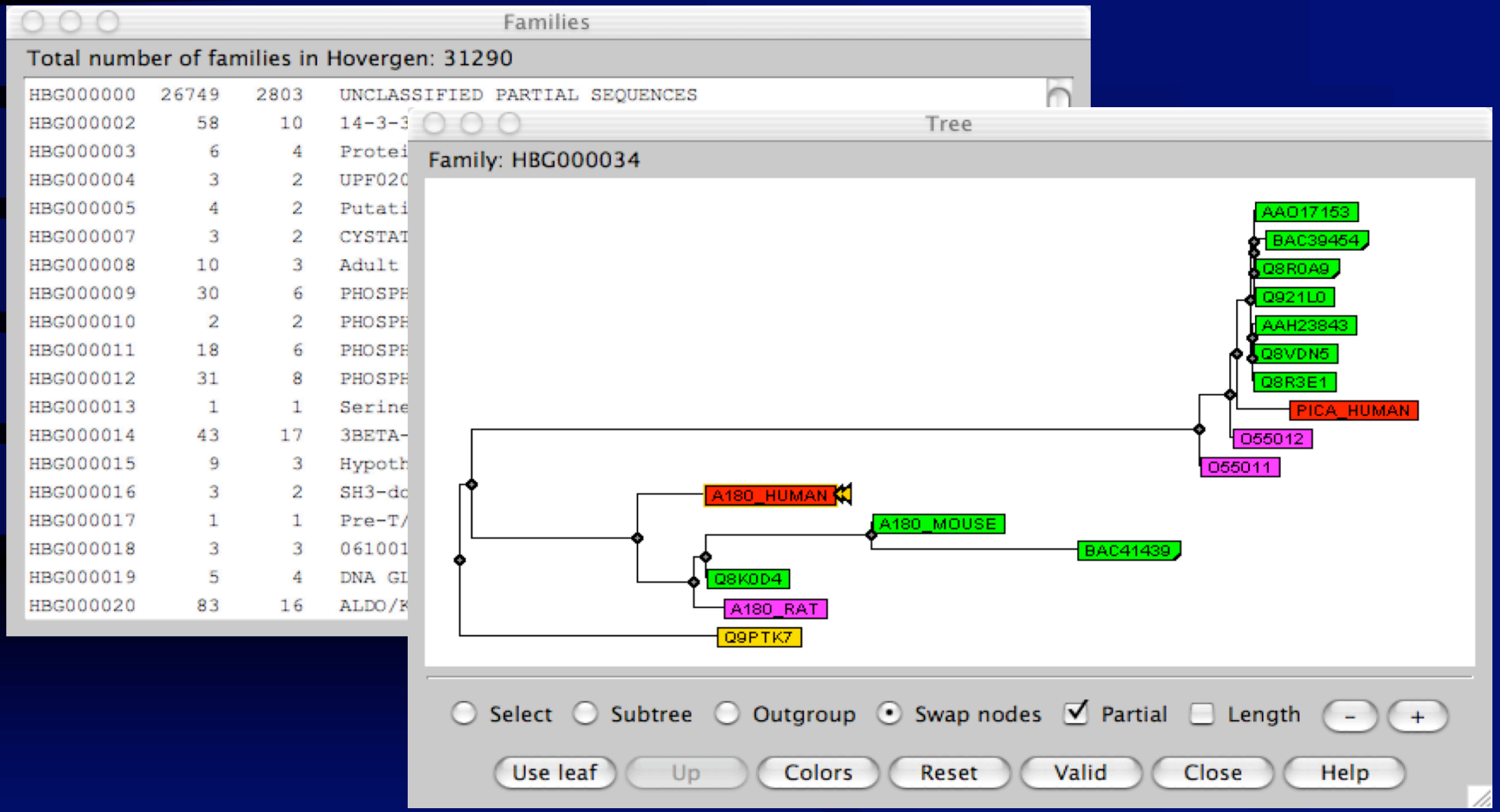
Organisation client/serveur



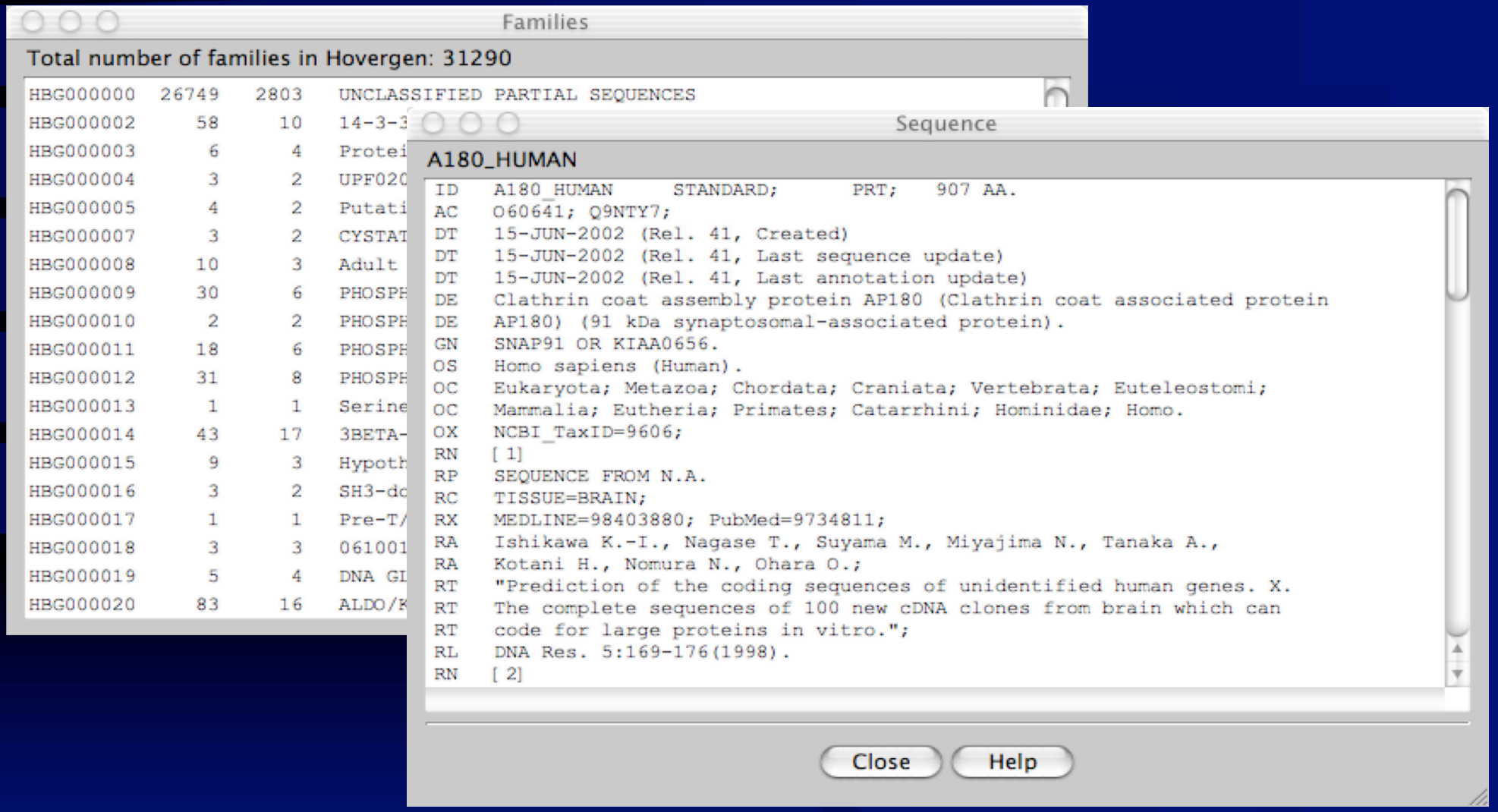
L'interface FamFetch



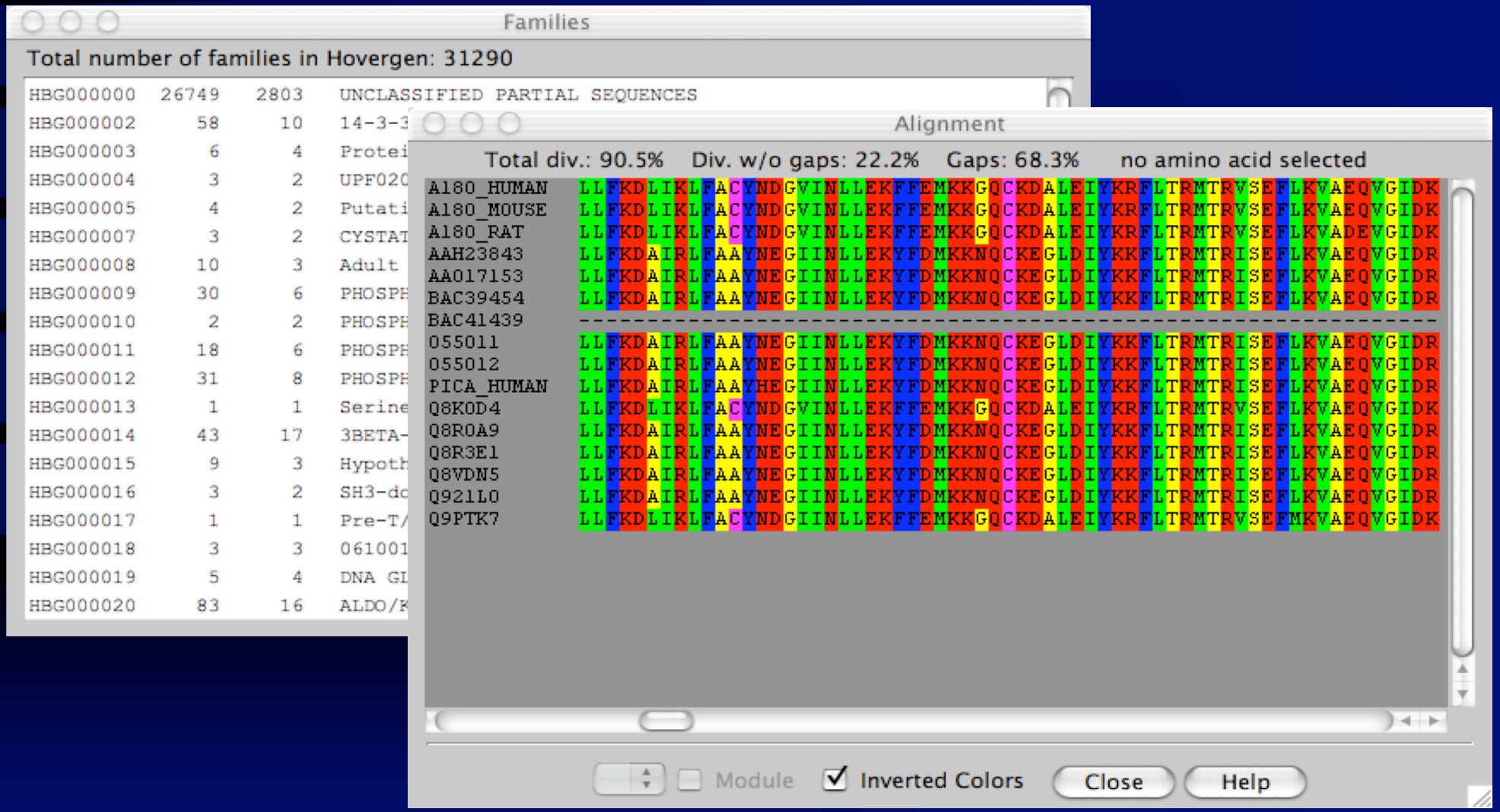
L'interface FamFetch



L'interface FamFetch



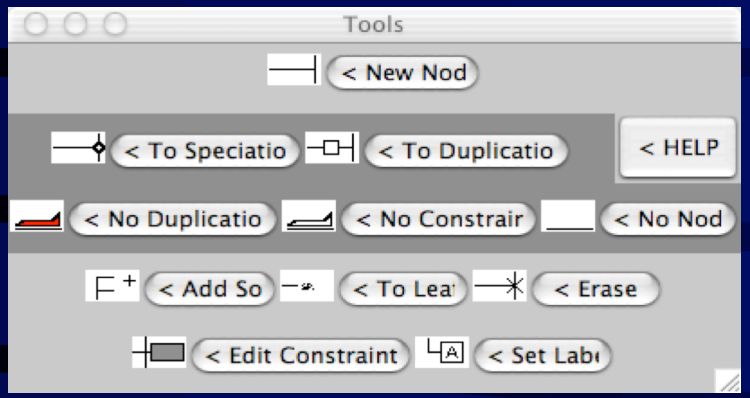
L'interface FamFetch



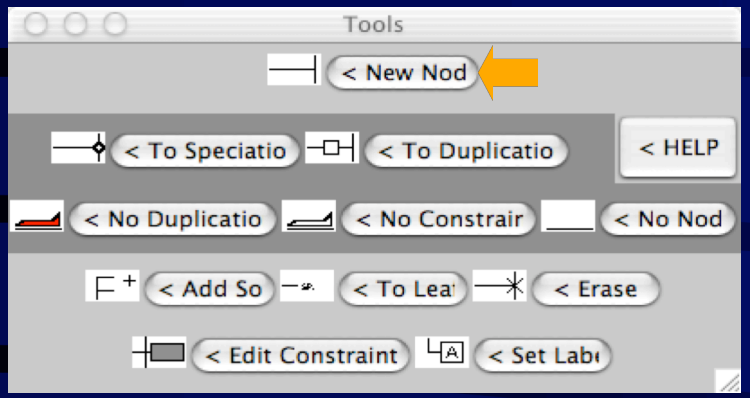
Recherche de motifs

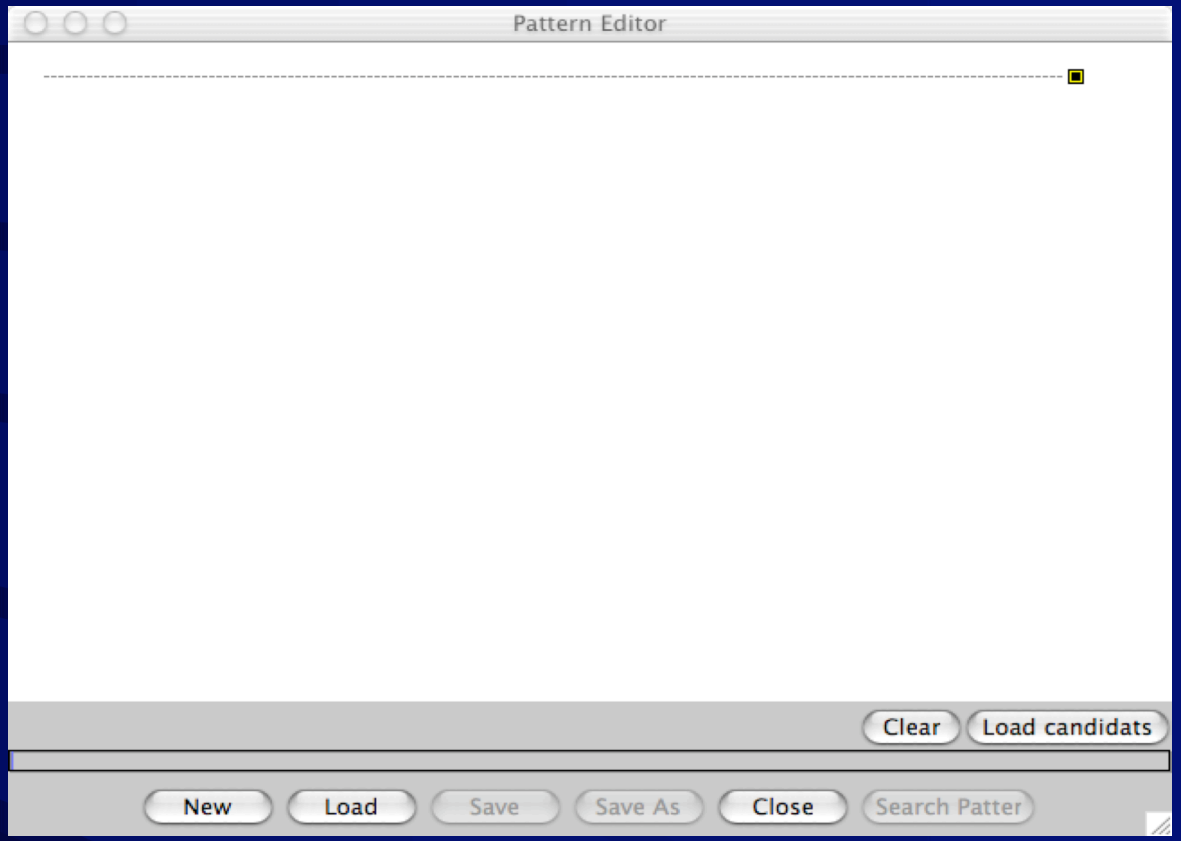
- Récupération des familles pour lesquelles l'arbre présente un motif donné :
  - Implementation d'un algorithme de recherche de motifs d'arbre non ordonnés.
  - Introduction de contraintes sur les motifs :
    - Présence ou absence de taxons dans certaines parties de l'arbre.
    - Interdiction de la présence de nœuds de duplication :
      - ✓ Recherche d'orthologues.

Construction d’une requête

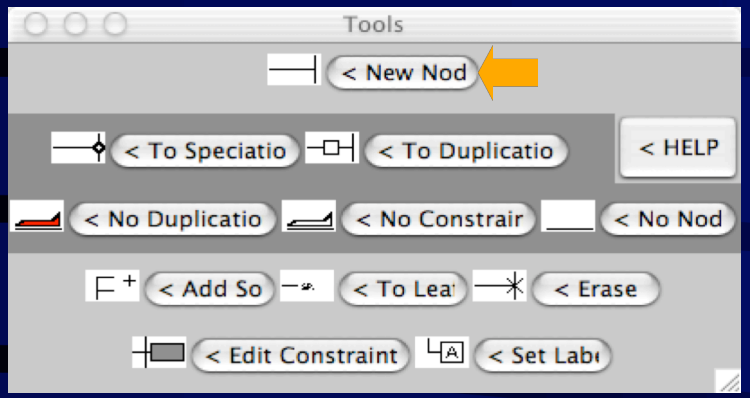


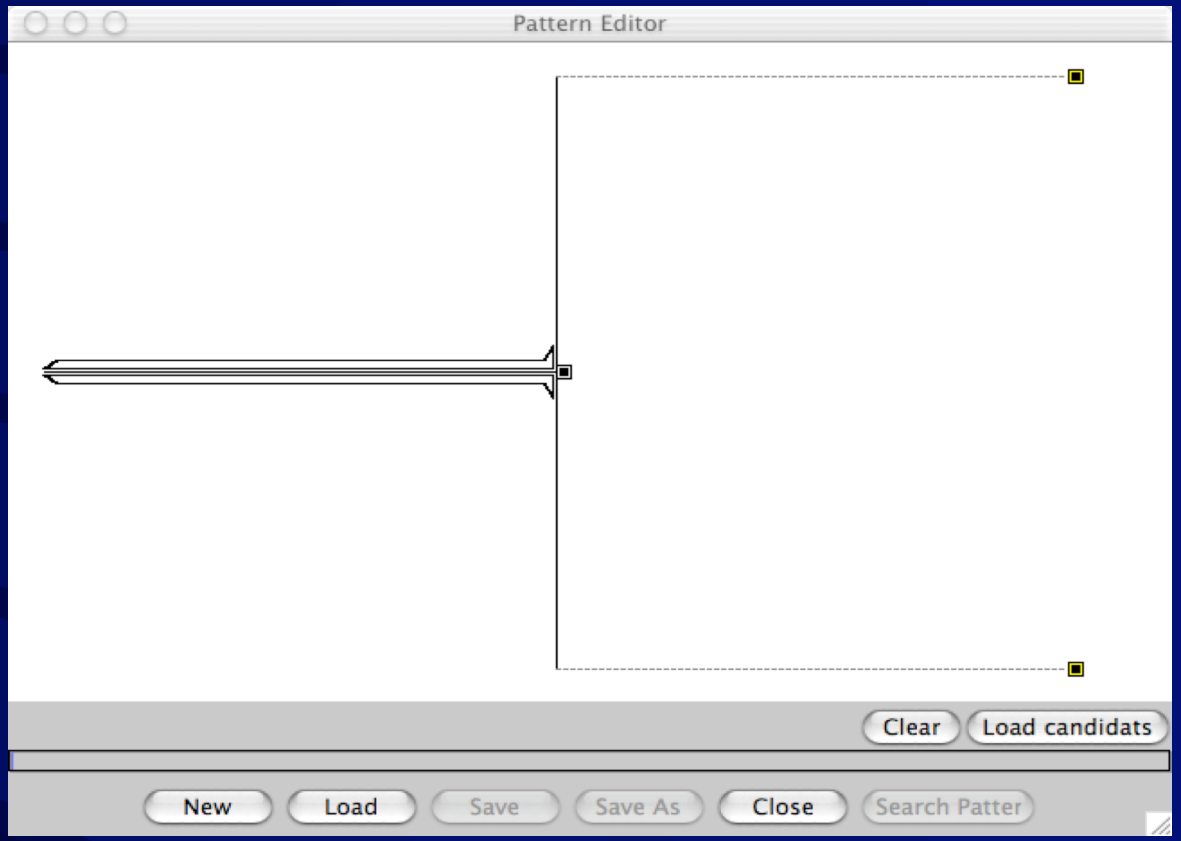
Construction d'une requête



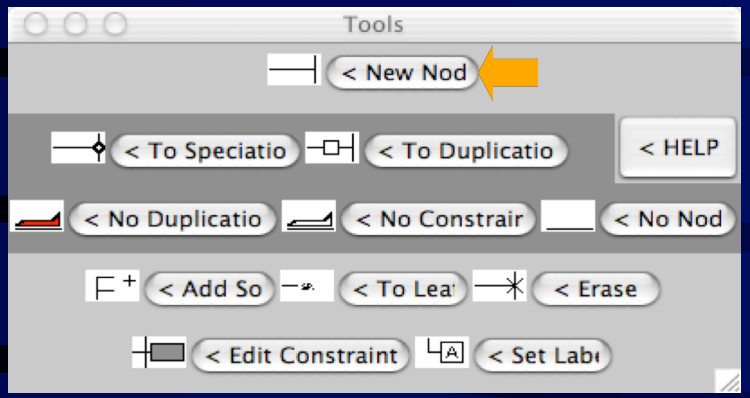


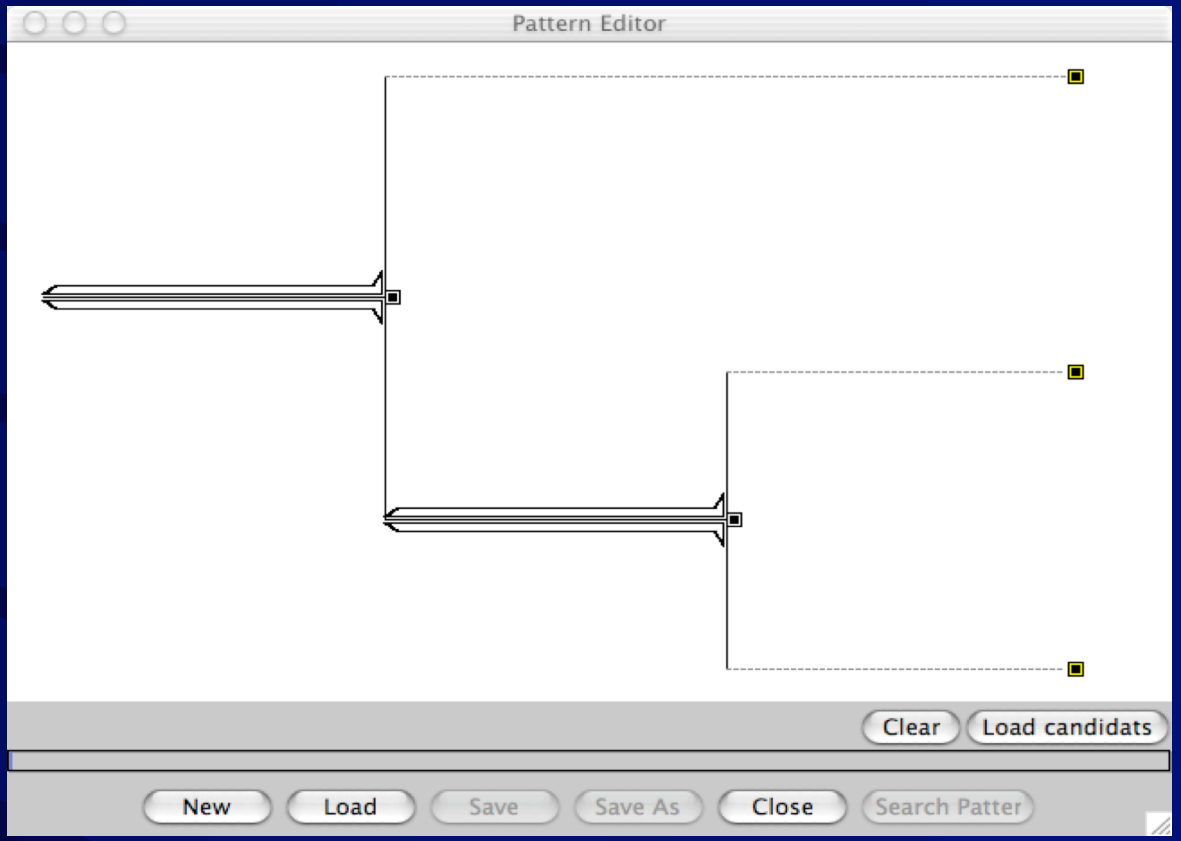
Construction d'une requête



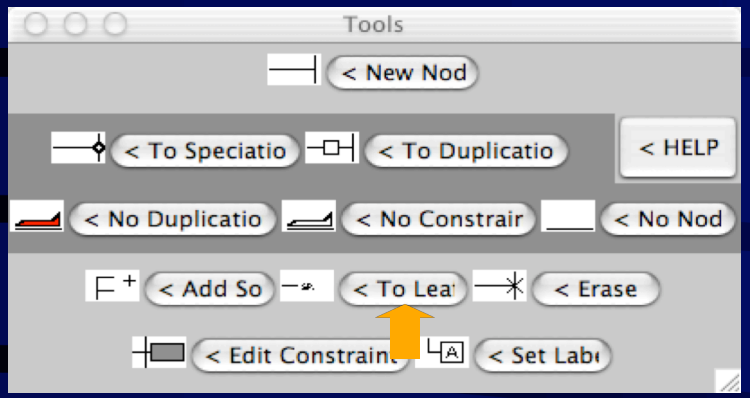


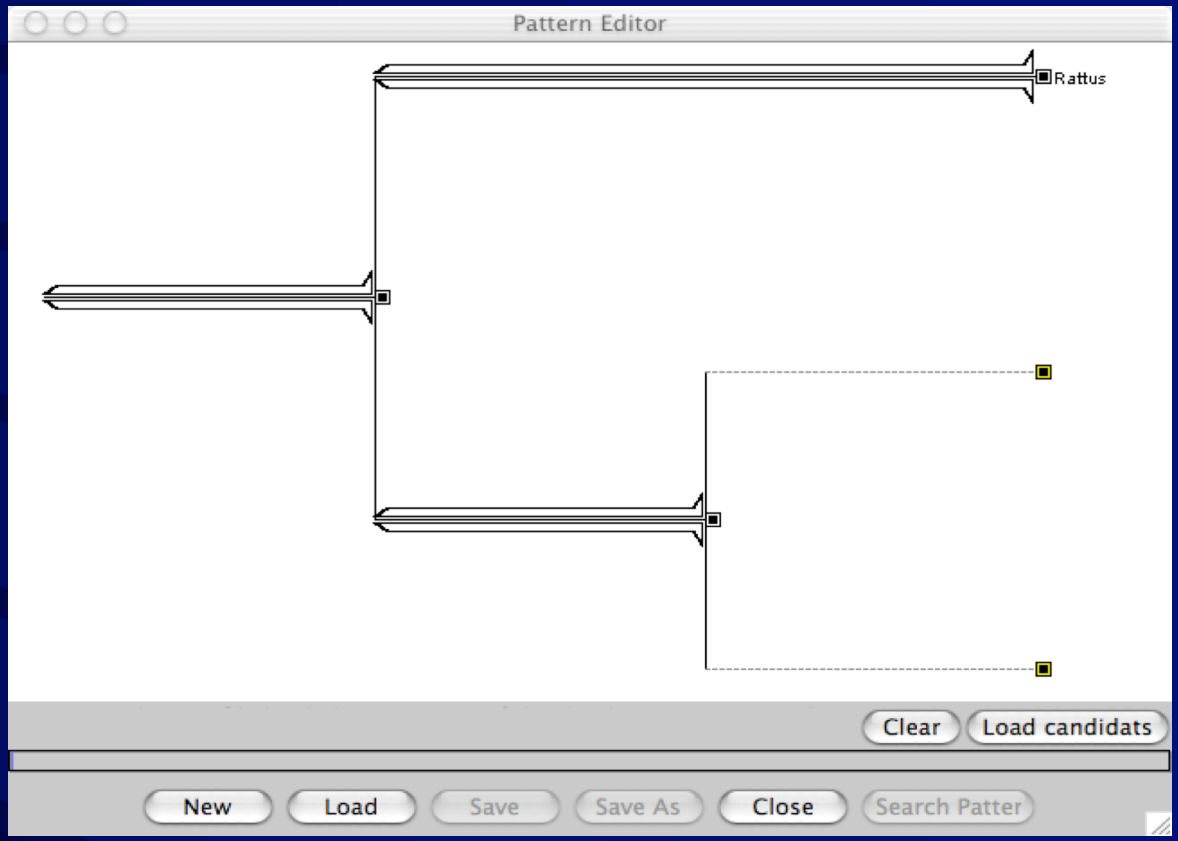
Construction d'une requête



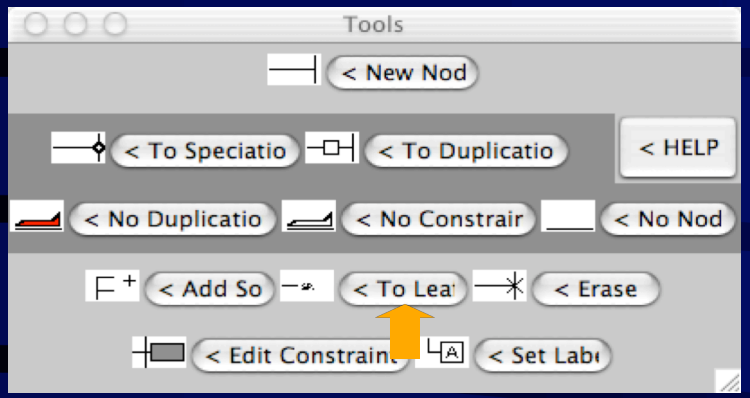


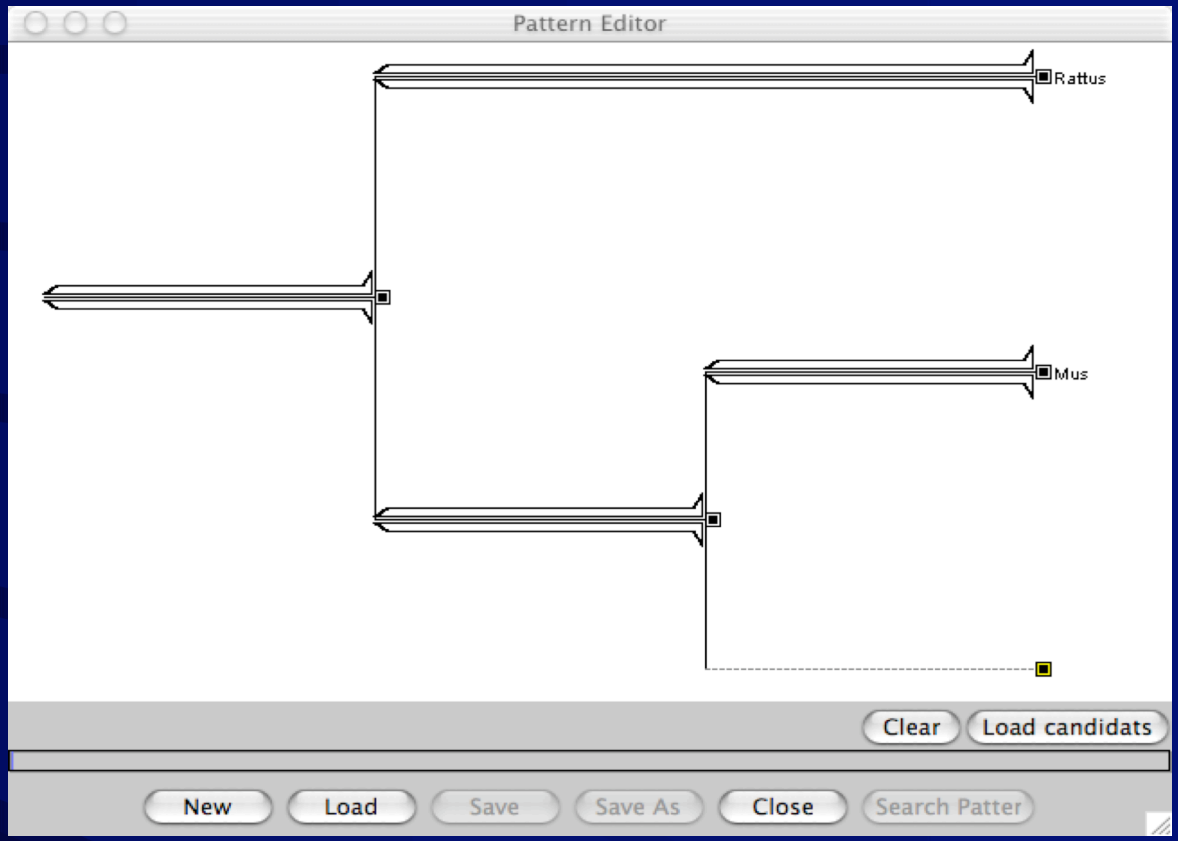
Construction d'une requête



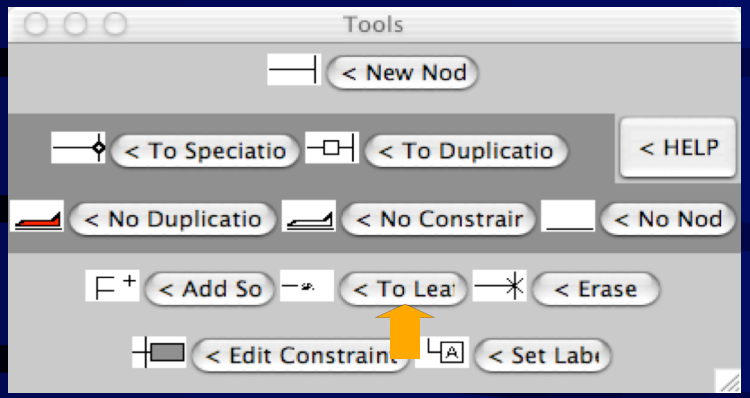


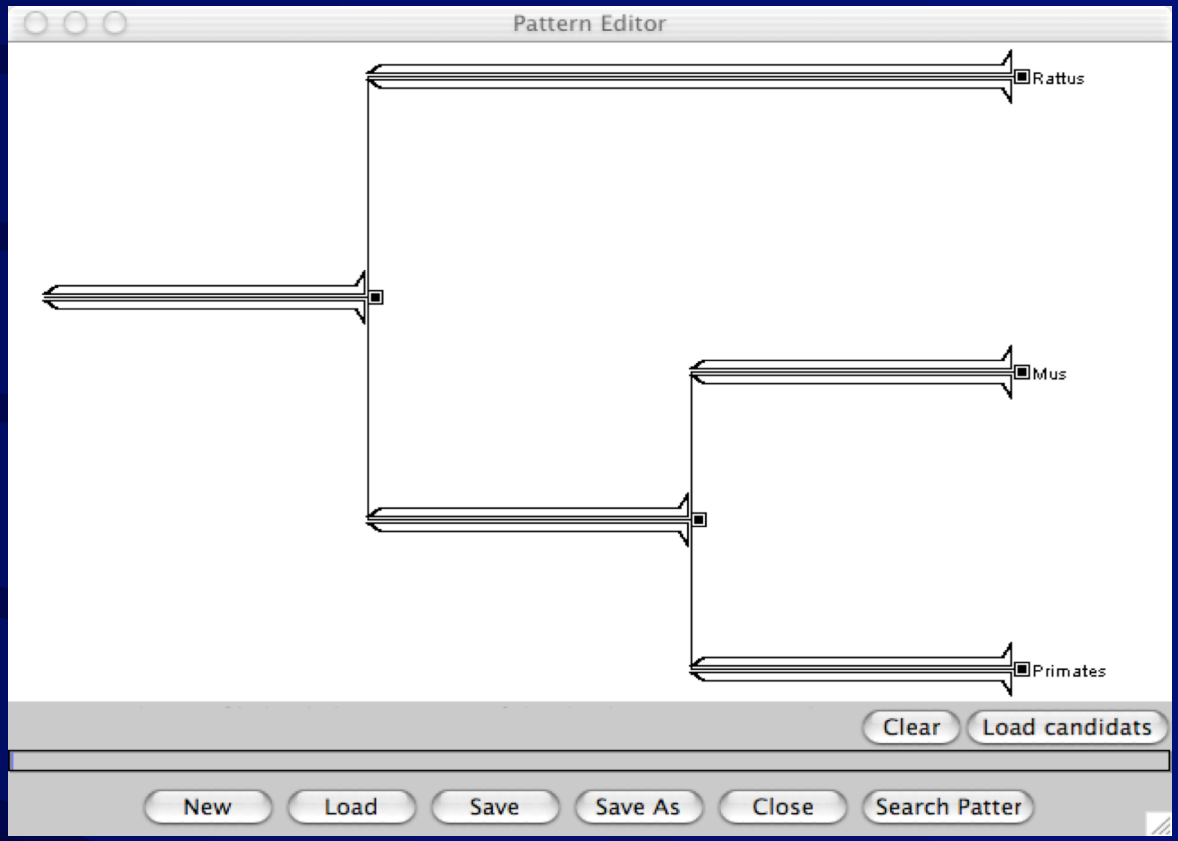
Construction d'une requête



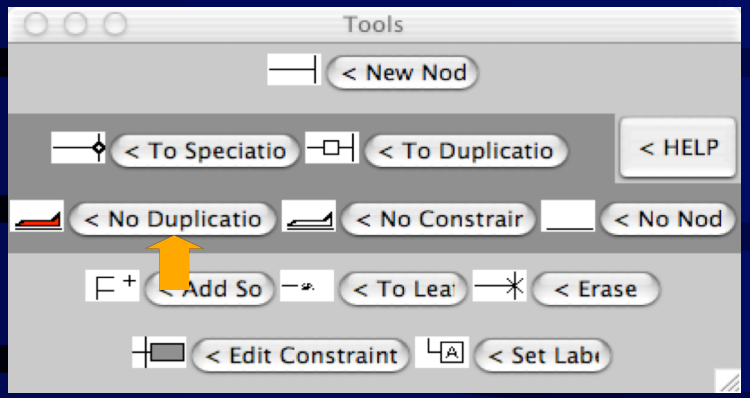


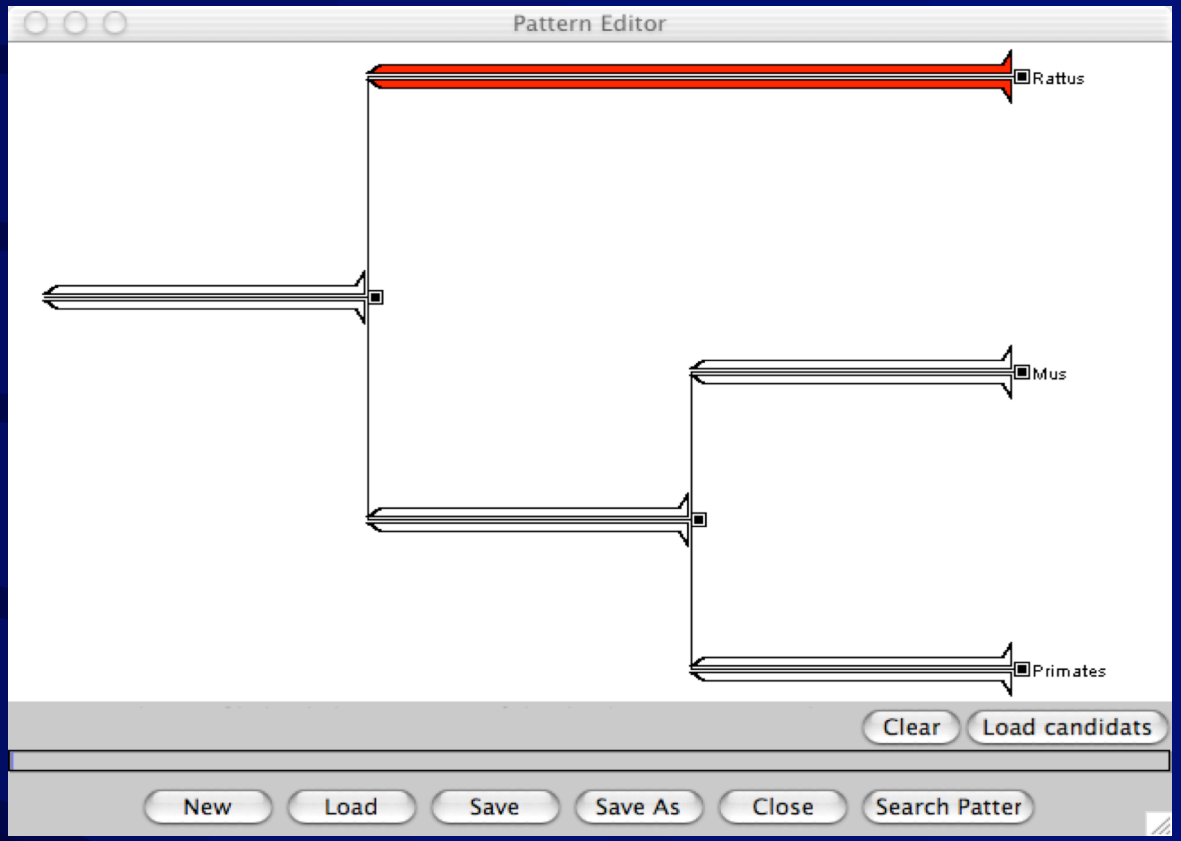
Construction d'une requête



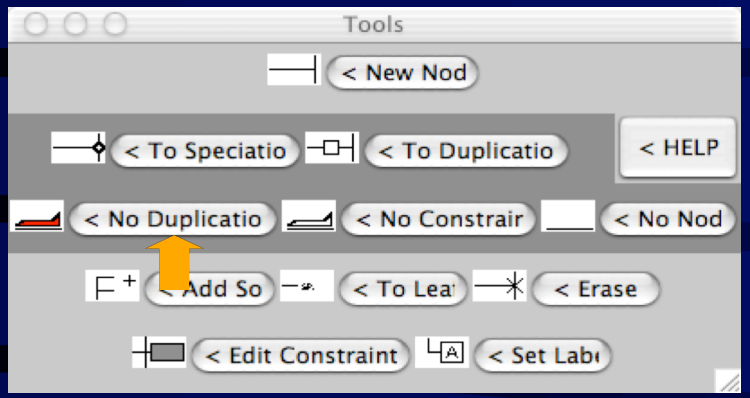


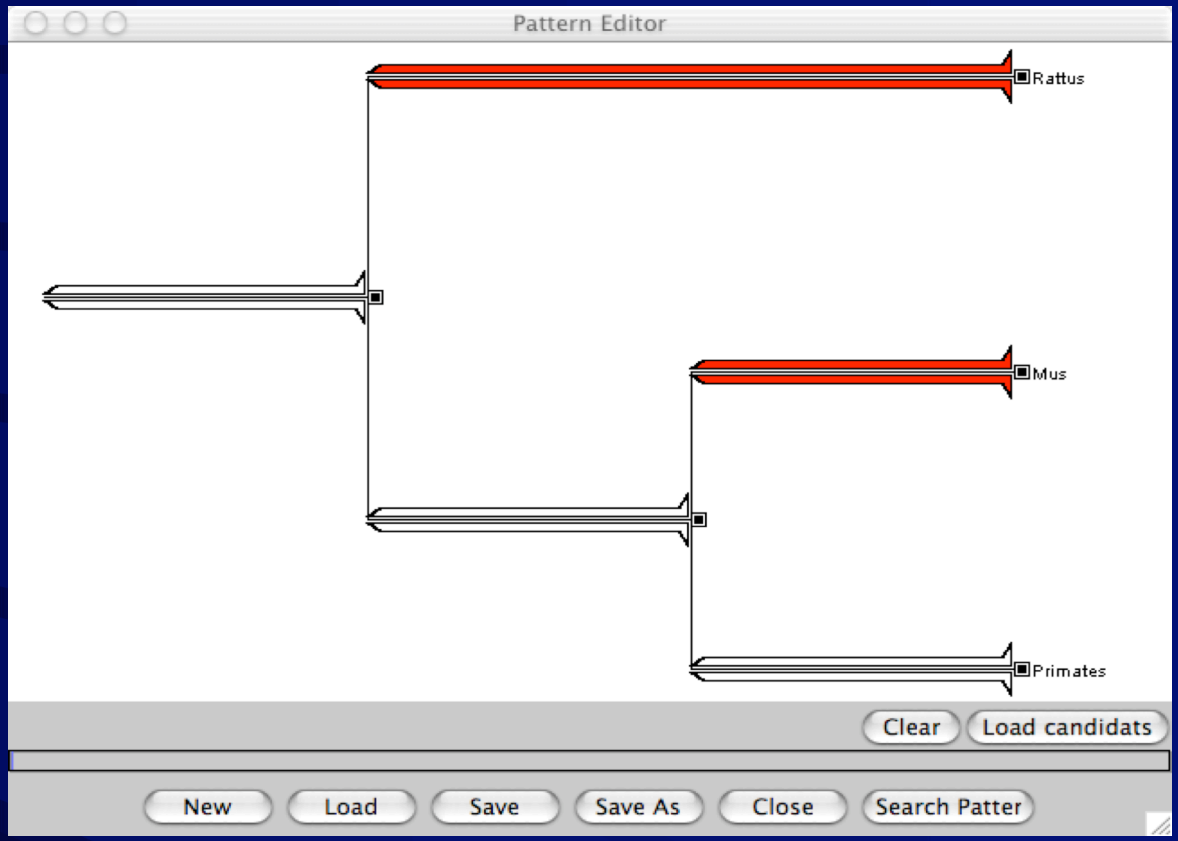
Construction d'une requête



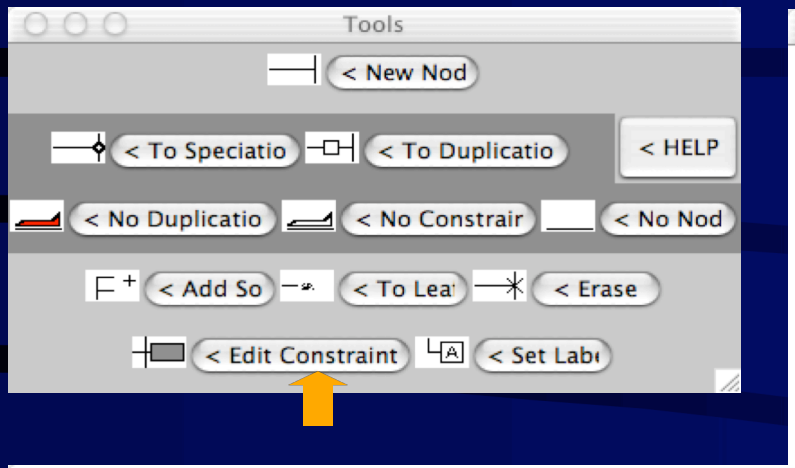


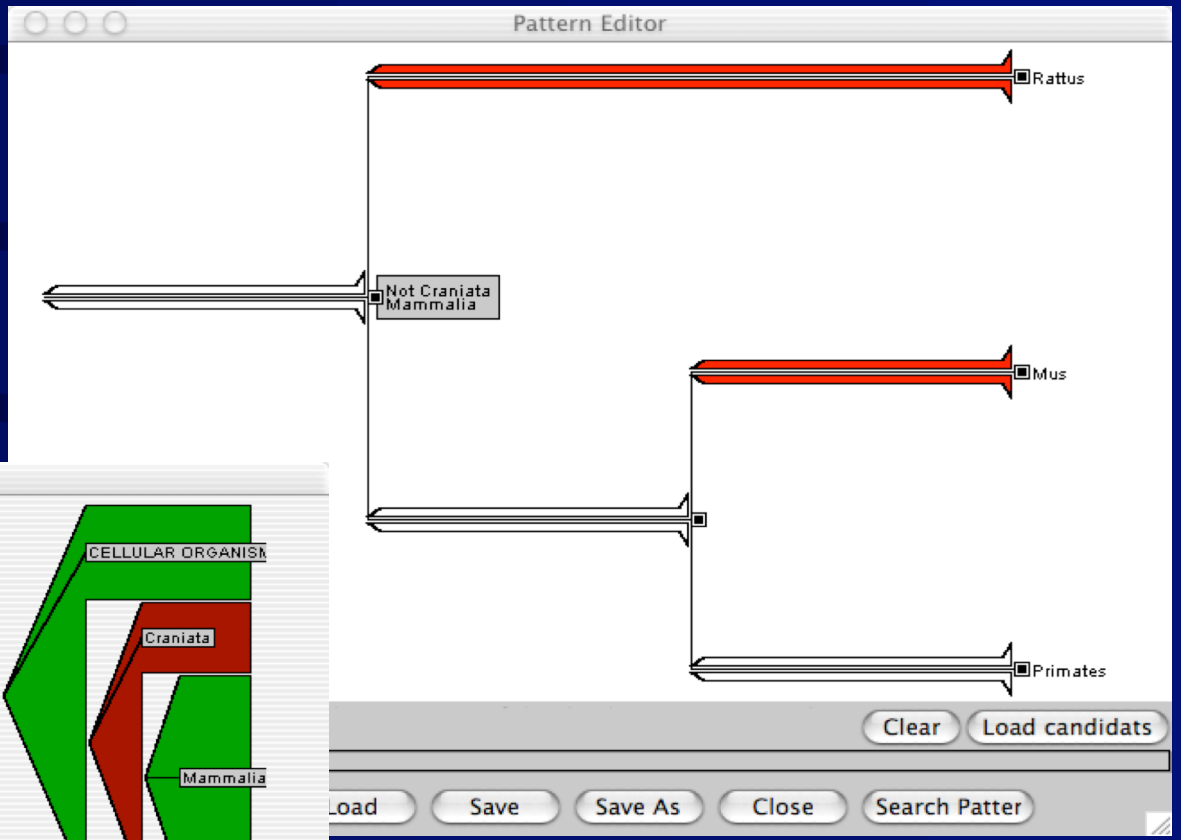
Construction d'une requête

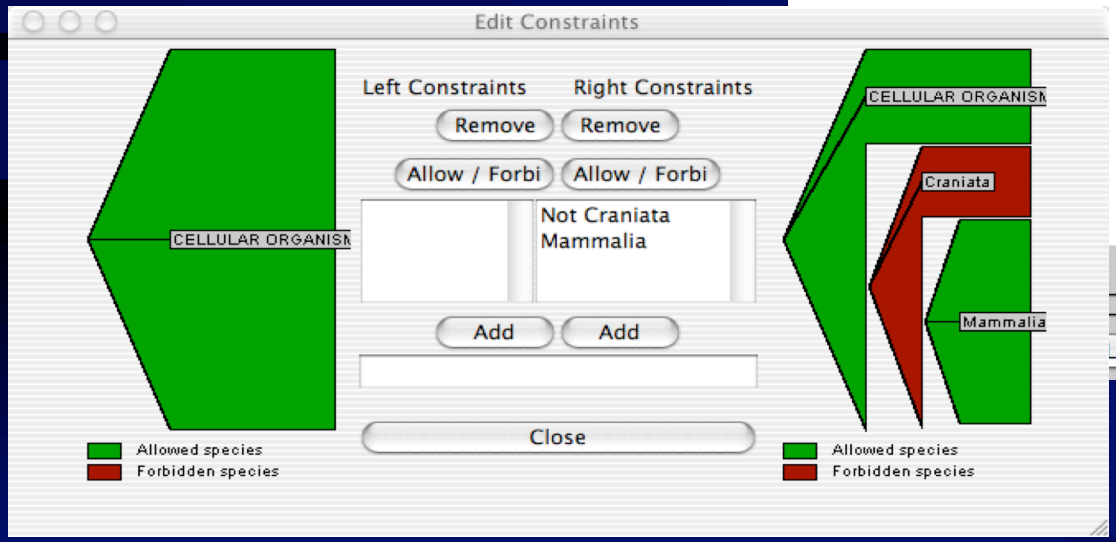




Construction d'une requête





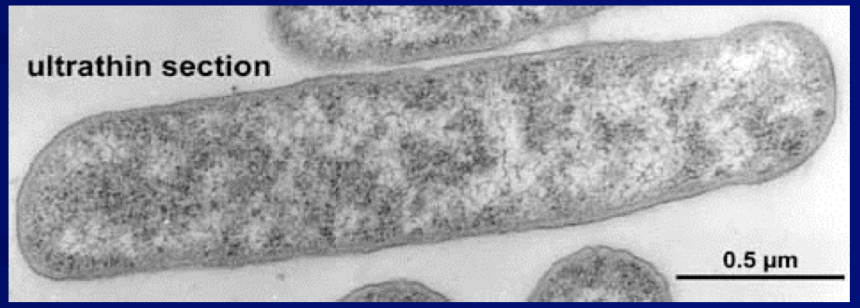


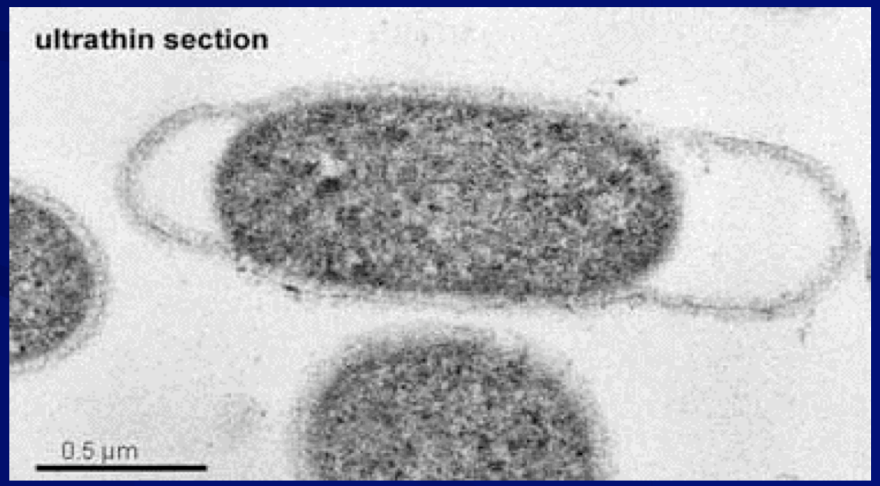
Exemple d'analyse

- Question de l'origine de l'hyperthermophilie chez les procaryotes :
  - Caractère ancestral.
  - Acquisition indépendante dans plusieurs lignées.
  - Acquisition par l'intermédiaire de transferts horizontaux.
- Analyse de l'ensemble des arbres d'HOGENOM (100 ≥ nb. taxons ≥ 4) contenant des séquences de ces organismes :
  - Étude des incongruences phylogénétiques.

Bactéries hyperthermophiles

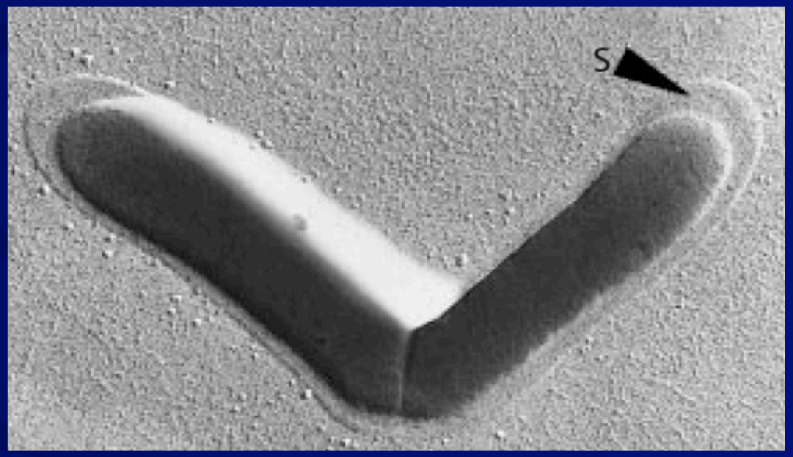
- *Aquifex aeolicus* :
  - Aérobie.
  - T_{opt} = 90°C.
  - Isolée dans des sources chaudes dans le parc de Yellowstone.
- *Thermotoga maritima* :
  - Anaérobie.
  - T_{opt} = 80°C.
  - Vit dans les sédiments marins proches des chemi-nées géothermiques.

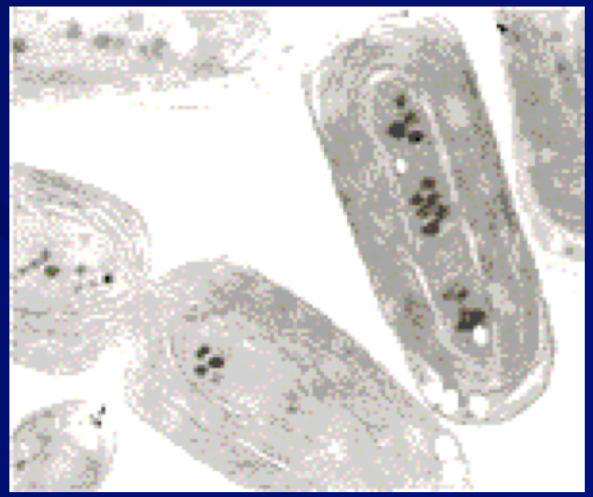




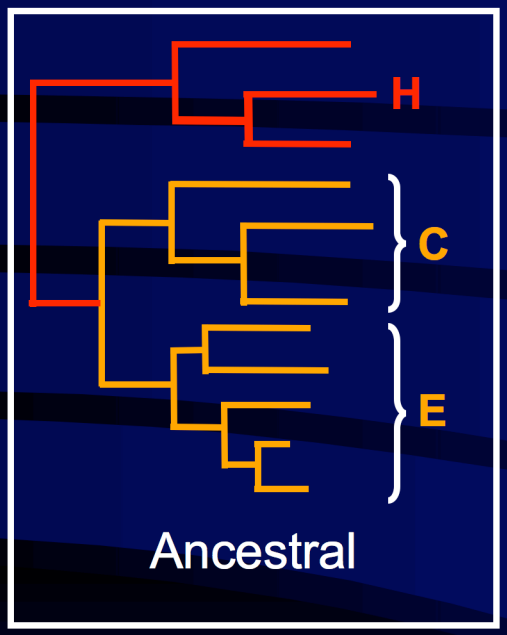
Contrôles

- Bactérie « quasi-hyperthermo-phile » :
  - *Thermoanaerobacter tengcongensis* :
    - Clostridiale (Firmicute).
    - T_{opt} = 75°C (50-80°C).
    - Reverse gyrase.
    - Isolée dans des sources chaudes en Chine.
- Bactérie thermophile :
  - *Thermosynechococcus elongatus* :
    - Cyanobactérie.
    - T_{opt} = 55°C.
    - Isolée dans des sources chaudes au Japon.

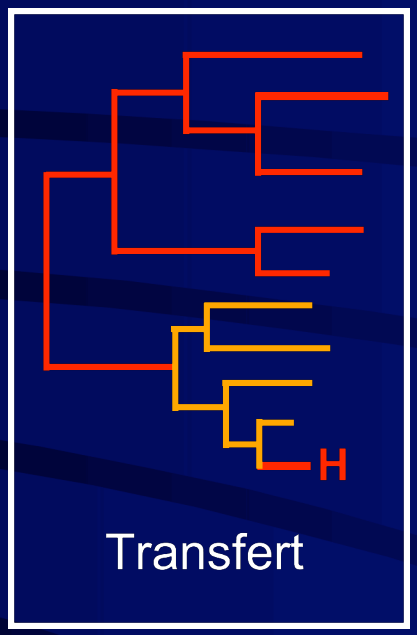




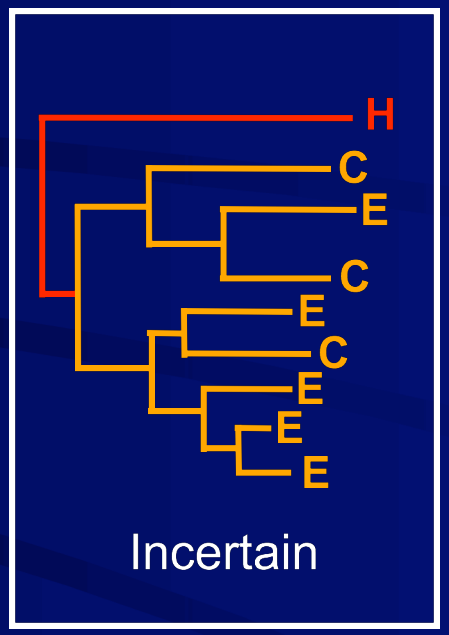
Types d'incongruences



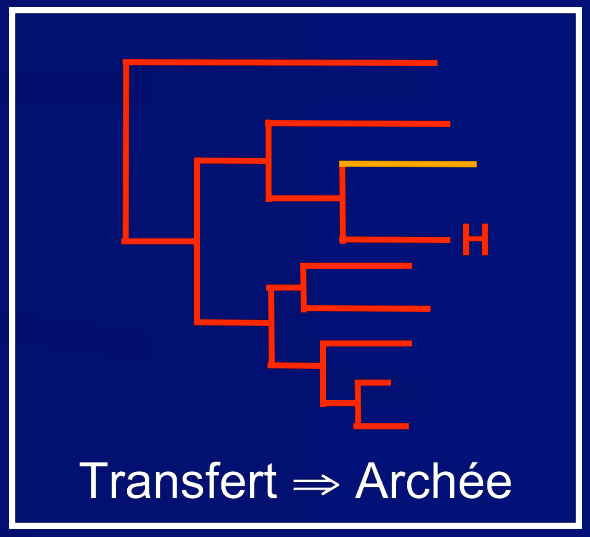
Ancestral



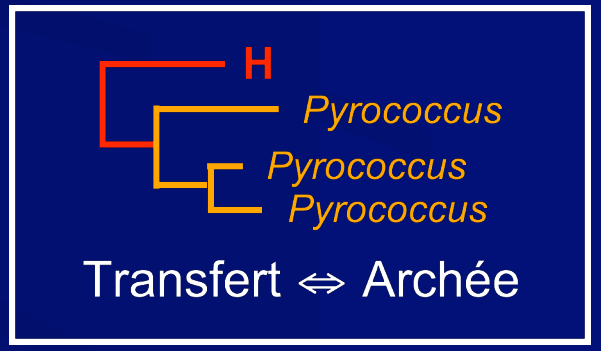
Transfert



Incertain



Transfert ⇒ Archée



Transfert ⇔ Archée

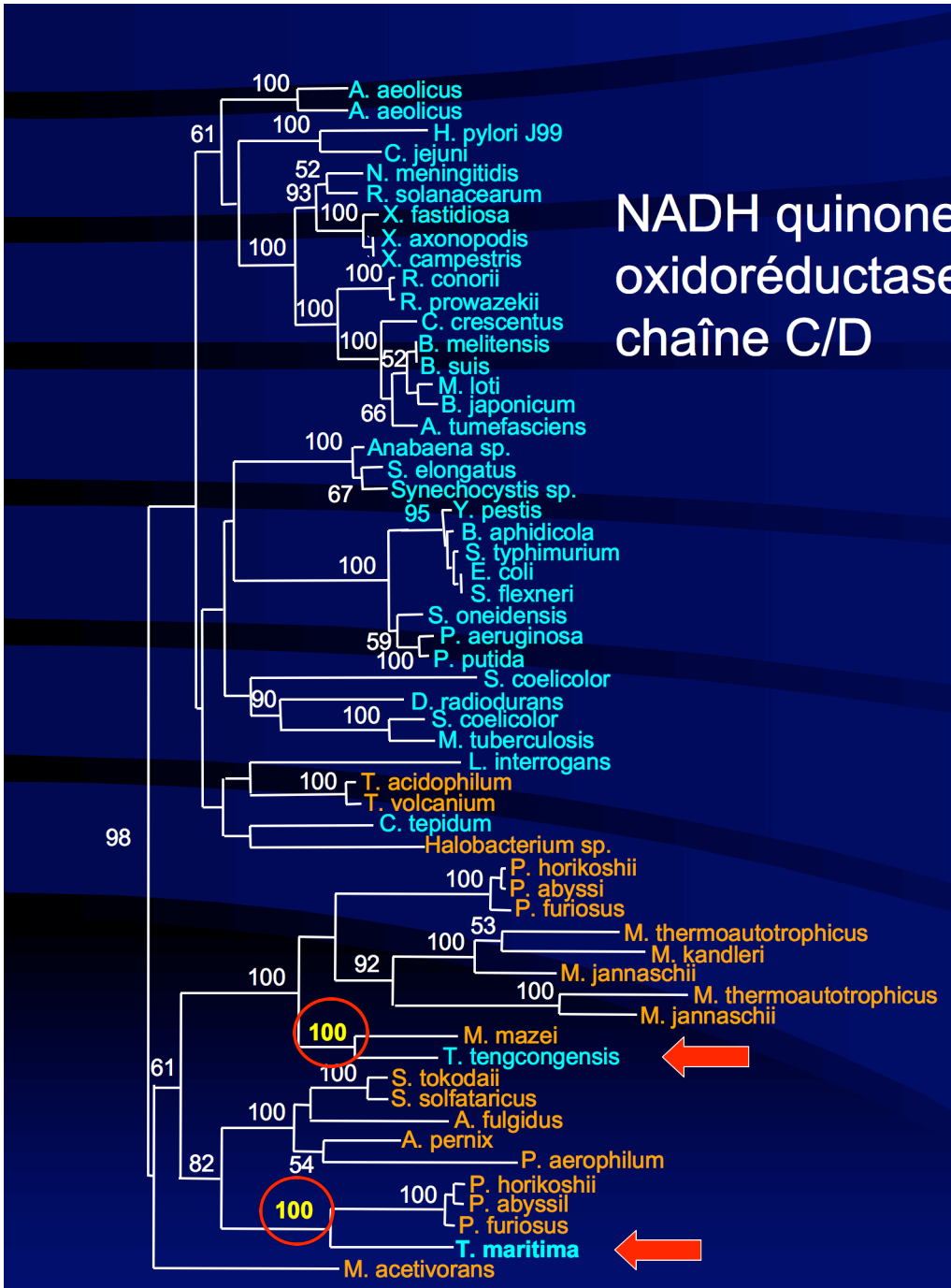
H : Bactérie hyperthermophile

E : Euryarchées C : Crénarchées

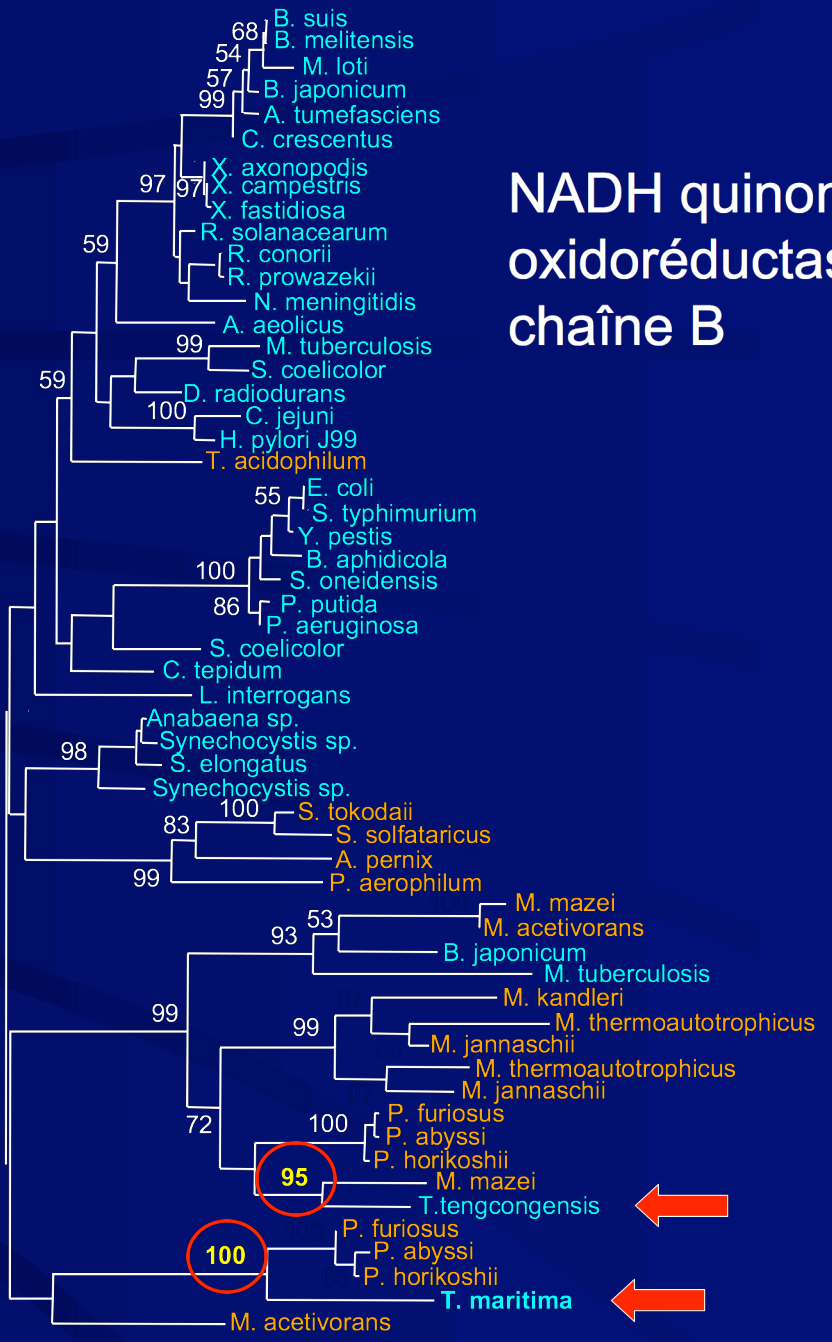
Bactéries Archées

Exemples de transferts

- Gène de la reverse gyrase.
- Trois gènes codant pour des sous-unités de la glutamate synthase chez *Thermotoga* et *Thermo-anaerobacter*.
- Sept gènes consécutifs codant pour des hétéro-disulfides réductases et des protéines hypothé-tiques d'*Aquifex*.
- Deux opérons (*mbx* et *ech*) codant pour des hydro-génases chez *Thermotoga* et *Thermoanaerobacter*.

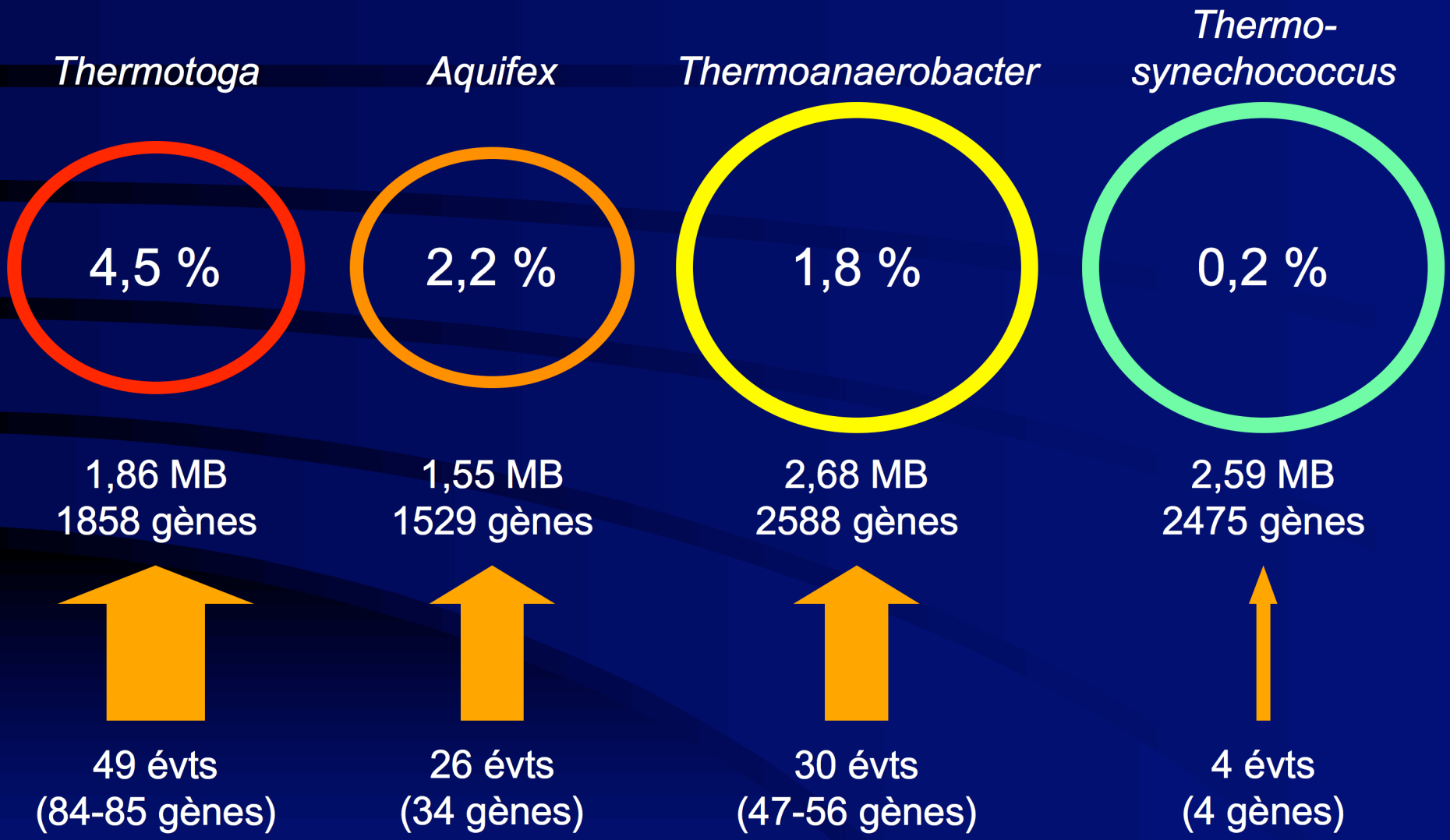


NADH quinone oxidoréductase chaîne C/D

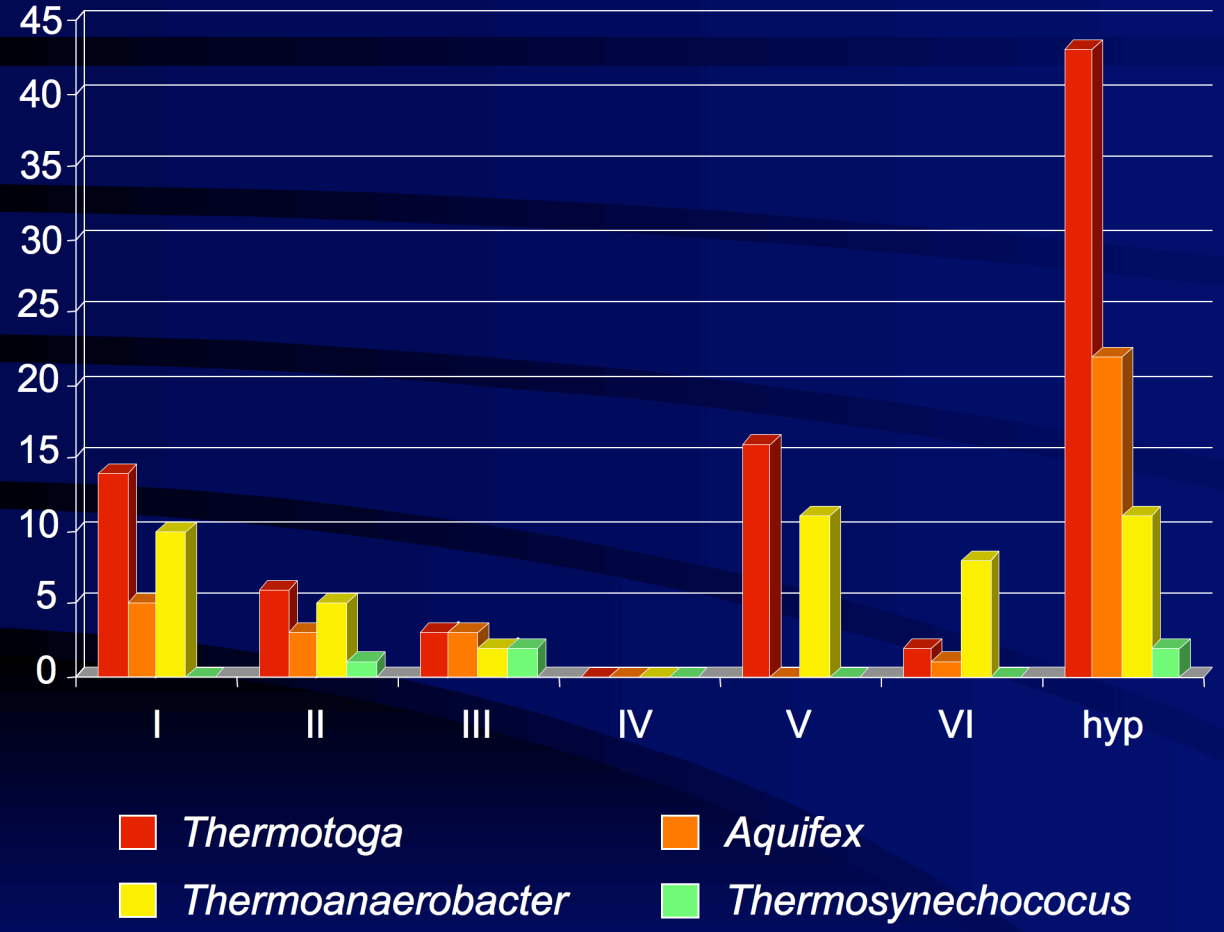


NADH quinone oxidoréductase chaîne B

Nombre de transferts détectés



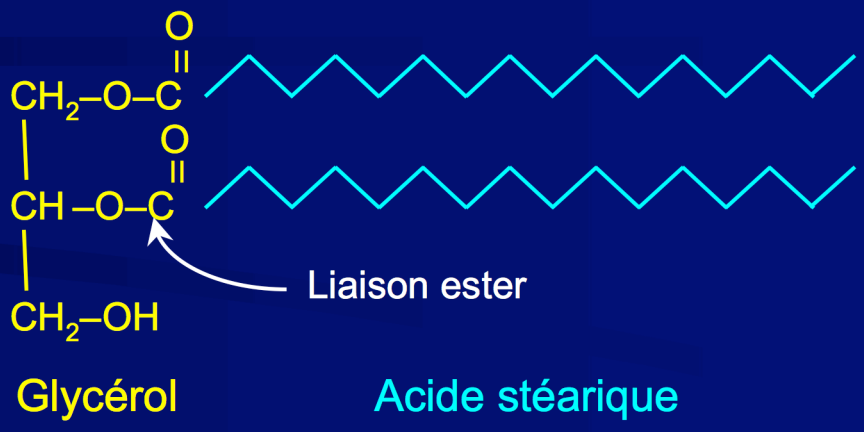
Classification fonctionnelle



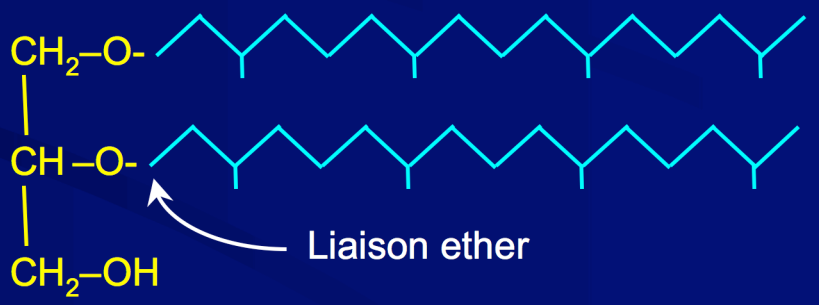
I	métabolisme intermédiaire
II	biosynthèse de petites molécules
III	biosynthèse de macro-molécules
IV	structure cellulaire
V	processus cellulaires
VI	autres fonctions
Hyp	protéines hypothétiques

La paroi des archées

- Structure particulière de la paroi :
  - Pas de peptidoglycane :
    - Dans certains cas, présence d'un pseudo-peptidoglycane.
  - Lipides membranaires possèdant des branchements.
  - Liaisons de type ether avec le glycérol.



Diglycéride bactérien



Diglycéride archéen

Conclusions

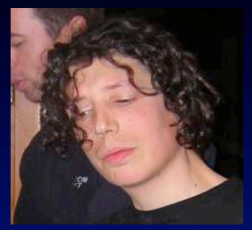
- Distribution phylogénétique multiple des hyper-thermophiles (Aquificales, Thermotogales, mais aussi Clostridiales) :
  - Plusieurs apparitions indépendantes de l'hyperthermo-philie (transferts ou acquisition multiple du caractère).
- Les bactéries hyperthermophiles ont subi plus de transferts depuis les archées que le témoin thermophile :
  - En majorité les transferts portent sur des gènes hypothétiques :
    - Congruence avec de nombreuses autres études.

Remerciements

Alexandra Calteau
Laboratoire de Biométrie et Biologie Évolutive

Laurent Duret
Laboratoire de Biométrie et Biologie Évolutive

Manolo Gouy
Laboratoire de Biométrie et Biologie Évolutive



Simon Penel
Laboratoire de Biométrie et Biologie Évolutive