

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

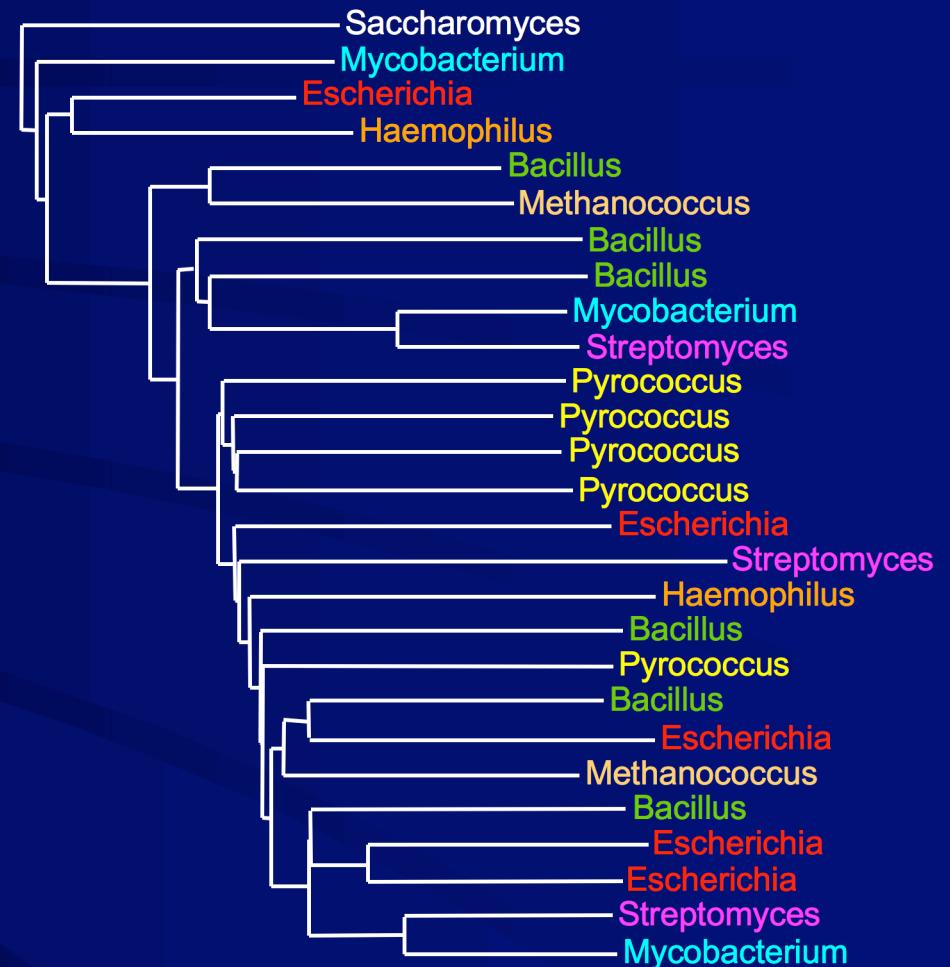


Étapes d'une phylogénie

- 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 dupliqués) ne reflète pas l'arbre des espèces.
- Il est nécessaire d'identifier 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 applications (*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

Click [here](#) for a description of the 2.0 version of BLAST

Choose the program to use and the database to search:

Program: Database:

Choix de la banque

The query sequence is [filtered](#) for low complexity regions by default

Enter your sequence in [FASTA](#) format:

```
>MyProtein
MAVTQTAQACDLVIFGAKGDLARRKLLPSLYQLEKAGQLNPDTRIICVGGRADWDKAAYTK
VVREALETFMKGTEIDEGLWDTLSARLDFCNLDVNNTAAFSRLGAMLDQKNRITINYFAMP
PSTFGAICKGLGEAKLNAKPARVVMEKPLGTSLATSQEINDQVGEYFEECQVYRIDHYLG
KETVLNLLALRFANSLFVNNWDNRTIDHVEITV
```

Séquence requête

Advanced options for the BLAST server:

Expect: Filter:

Descriptions: Alignments:

Genetic codes (blastx only):

Other advanced options:

Submit/Clear query:

Paramètres de BLAST

Filtrage des résultats

Filter your BLAST results by taxon, keyword or date

Select a taxon in the list or enter a taxon name [[Help](#)]

Taxon:

And/or enter a keyword [[Help](#)]

Keyword:

And/or enter a date (e.g. 1/Jun/93) [[Help](#)]

Date:

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
<input checked="" type="checkbox"/> G6PD_ECOLI 491 Glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49)...	428	e-119
<input checked="" type="checkbox"/> Q3Z2K3_SHISO 491 Glucose-6-phosphate dehydrogenase.	427	e-118
<input checked="" type="checkbox"/> Q8XCJ6_ECO57 491 Glucose-6-phosphate 1-dehydrogenase (G6PD).	426	e-118
<input checked="" type="checkbox"/> Q7UAF3_SHIFL 491 Glucose-6-phosphate dehydrogenase.	425	e-118
<input checked="" type="checkbox"/> Q8ZNW2_SALTY 491 Glucose-6-phosphate dehydrogenase (EC 1.1.1.49).	408	e-113
<input checked="" type="checkbox"/> Q8Z5X1_SALTI 491 Glucose 6-phosphate dehydrogenase (EC 1.1.1.49).	408	e-113
<input checked="" type="checkbox"/> Q5PN04_SALPA 491 Glucose 6-phosphate dehydrogenase.	408	e-113
<input checked="" type="checkbox"/> Q57NB4_SALCH 491 Glucose-6-phosphate dehydrogenase.	406	e-112
<input checked="" type="checkbox"/> Q6D4B4_ERWCT 491 Glucose-6-phosphate 1-dehydrogenase (G6PD).	367	e-100
<input checked="" type="checkbox"/> G6PD_ERWCH 491 Glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49)...	360	1e-98

Références croisées

Sequence list "list":

Edit Modify Cross Refs Retrieve Analyze View History

Query = [MyProtein](#)

[G6PD_ECOLI](#)

[Q3Z2K3_SHISO](#)

[Q8XCJ6_ECO57](#)

[Q7UAF3_SHIFL](#)

[Q8ZNW2_SALTY](#)

Sequence list "crosslist":

Edit Modify Cross Refs Retrieve Analyze View History

Sequences 1 to 20 of 40

Select sequences Display: 20 per page

Save selection Next Go to page 1 Clear Reset Select all:

1. [AE001165.BB0636](#) Location/Qualifiers (length = 1437)

FT CDS 4690..6126

2. [AE001292.ZWF](#) Location/Qualifiers (length = 1320)

FT CDS 3329..4648

3. [AE001609.ZWF](#) Location/Qualifiers (length = 1539)

FT CDS 8115..9653

4. [AE002211.CP0524](#) Location/Qualifiers (length = 1539)

FT CDS complement(24263..25801)

5. [AE002314.ZWF](#) Location/Qualifiers (length = 1524)

FT CDS 7013..8536

Alignements multiples

Sequence list "list":

Edit Modify Cross Refs Retrieve Analyze View History

Query = MyProtein

G6PD_ECOLI

alignment with width :

Q3Z2K3_SHISO

residues.

Q8XCJ6_ECO57

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

Q7UAF3_SHIFL

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

Q8ZNW2_SALTY

	550	560	570	580	590	600
<u>Q4NXP9_9DELT</u>	G-WRRVVIEKPFGRDLD <ins>SAVAL</ins> NRELASVVREEQIYRIDHYLGKETVQNLMVFRFANGIF					
<u>Q5E0A8_VIBF1</u>	G-WKRLII <ins>EKPF</ins> GYDLASALELGKQIHQYFEEHQIYRIDHYLGKETVQNLLVLRFSNVMF					
<u>Q8NQ63_CORGL</u>	A-WRRVII <ins>EKPF</ins> CHNL <ins>SAHEL</ins> NQLVNAVFPESSVFRIDHYLGKETVQNILALRFANQLF					
<u>Q6M517_CORGL</u>	A-WRRVII <ins>EKPF</ins> CHNL <ins>SAHEL</ins> NQLVNAVFPESSVFRIDHYLGKETVQNILALRFANQLF					
<u>Q8FT74_COREF</u>	S-WRRVII <ins>EKPF</ins> CNDLKTAHELNETVNAVFPEESVFRIDHYLGKETVQNILALRFANQLF					
<u>Q6NH41_CORDI</u>	S-WRRVIVEKPFCHDQE <ins>SARQLNNLINSVFP</ins> KE <ins>SFRIDHYLGKETVQNIMALRFANQLF</ins>					
<u>G6PD_MYCTU</u>	R-WSRVVIEKPF <ins>GHDLASARELNKAVNAVFPEEA</ins> VFRIDHYLGKETVQNILALRFANQLF					
<u>G6PD_MYCBO</u>	R-WSRVVIEKPF <ins>GHDLASARELNKAVNAVFPEEA</ins> VFRIDHYLGKETVQNILALRFANQLF					
<u>Q741B6_MYCPA</u>	R-WSRVVIEKPF <ins>GHDLASAQSLNKAVNAVFPEESVFRIDHYLGKETVRN</ins> ILALRFANQLF					
<u>Q4NEB6_9MICC</u>	K-WRRVVIEKPF <ins>GHDLASARQLNDIVESVFP</ins> DAVFRIDHYLGKETVQNILALRFANQLF					
<u>Q8G732_BIFLO</u>	A-WRRVII <ins>EKPF</ins> GHDLASAKELDSVVSEVFDPSSVFRIDHYLGKETVQNLLALRFANAMY					
<u>Q47ND1_THEFY</u>	S-WRRVVIEKPF <ins>GHDLASAQELNALVDDVFP</ins> HSVFRIDHYLGKETVQNLLALRFSNTLF					
<u>Q3WC06_9ACTO</u>	GGWRRVVIEKPF <ins>GHDLASARELN</ins> SLVDDVFTPDDVFRIDHYLGKETVQNLFALRFANTLF					
<u>Q4C2R2_CROWT</u>	K--HRIVIEKPF <ins>GKDLSAQVLNRVVQNVCKEEQVY</ins> RIDHYLGKETVQNLLVFRFANAIF					
<u>G6PD_SYNYY3</u>	K--SRIVIEKPF <ins>CRDLSSAQSLNRVVQSVCKENQVY</ins> RIDHYLGKETVQNLMVFRFANAIF					
<u>Q8DLF3_SYNEL</u>	K--HRLVIEKPF <ins>CRDLTSQALNEVVQSVCKEEQVY</ins> RIDHYLGKETVQNLMVFRFANAIF					
<u>Q7UVB9_RHOBA</u>	P--RRVII <ins>EKPF</ins> G <ins>TDLTAQRLNESIHN</ins> VREDQIYRIDHYLGKETVQNIFALRFANSIF					

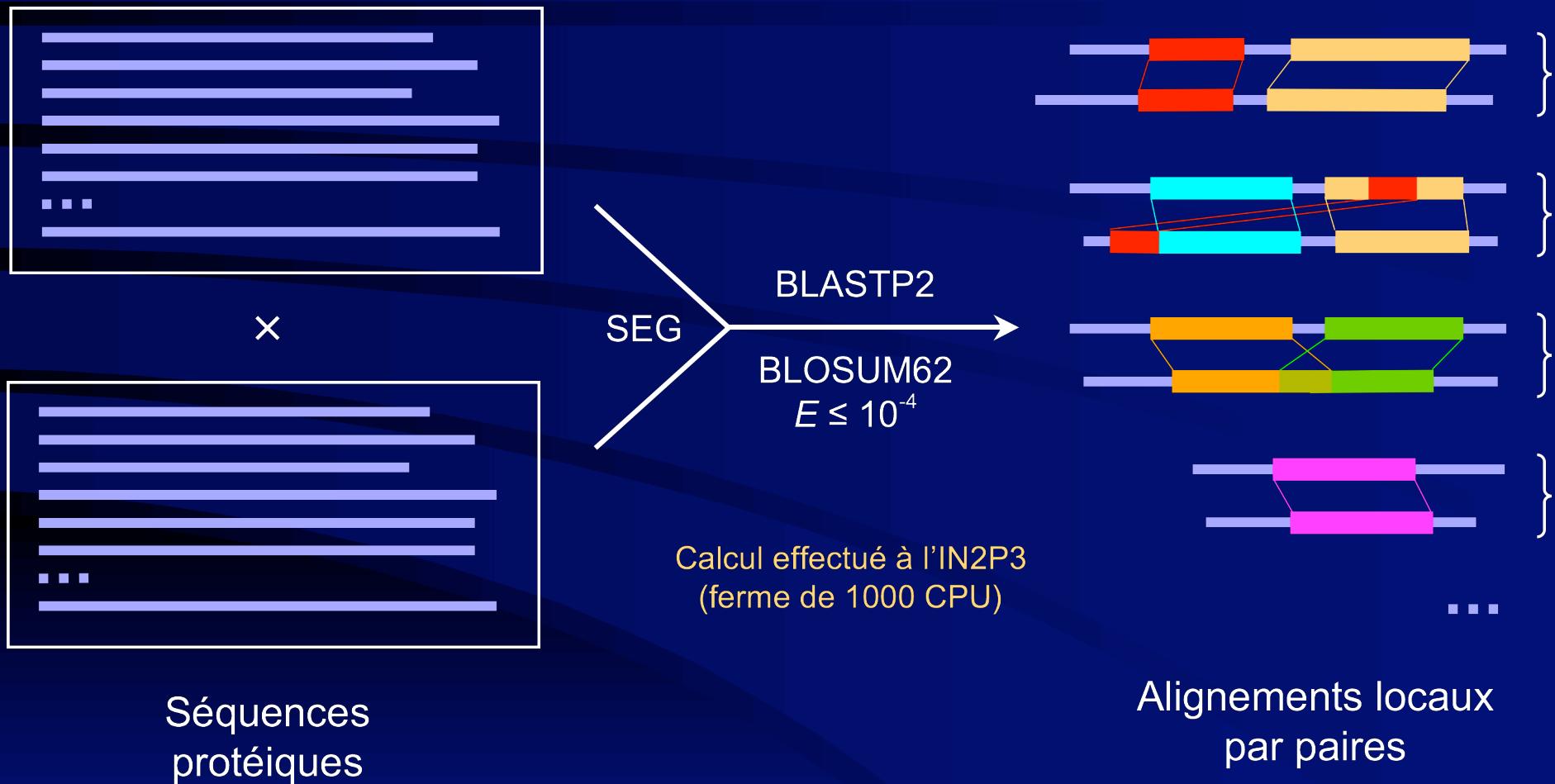
Banques de familles de gènes

- Automatisation des procédures :
 - Pour la recherche de similarités entre les protéines.
 - 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.

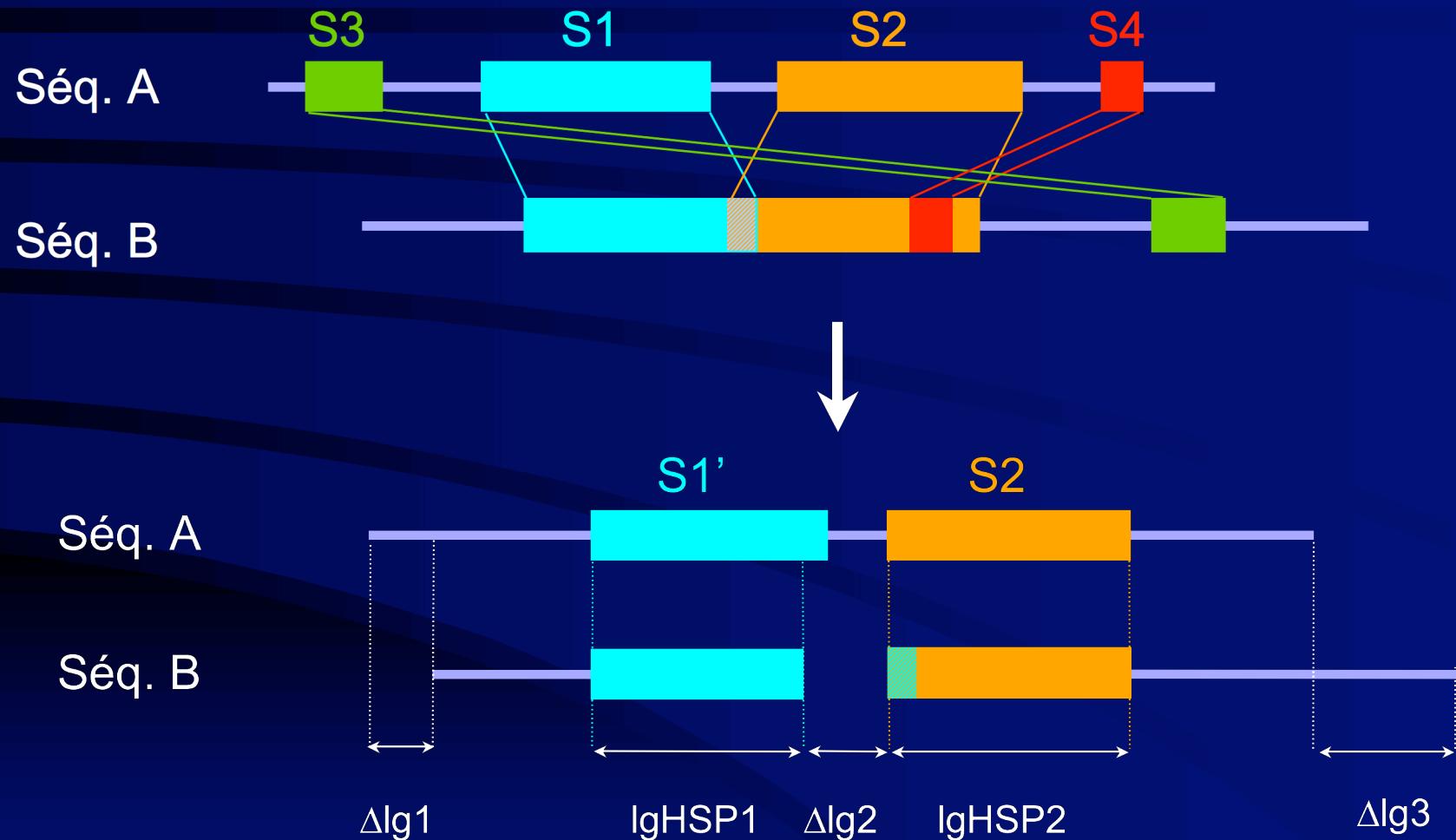
Les banques du PBIL

- 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'intérêt :
 - HOPPSIGEN (pseudogènes humains et de la souris présent dans Ensembl).
 - NuReBase (récepteurs nucléaires de GenPept).
 - RTKdb (récepteurs à Tyrosine Kinase d'UniProt).

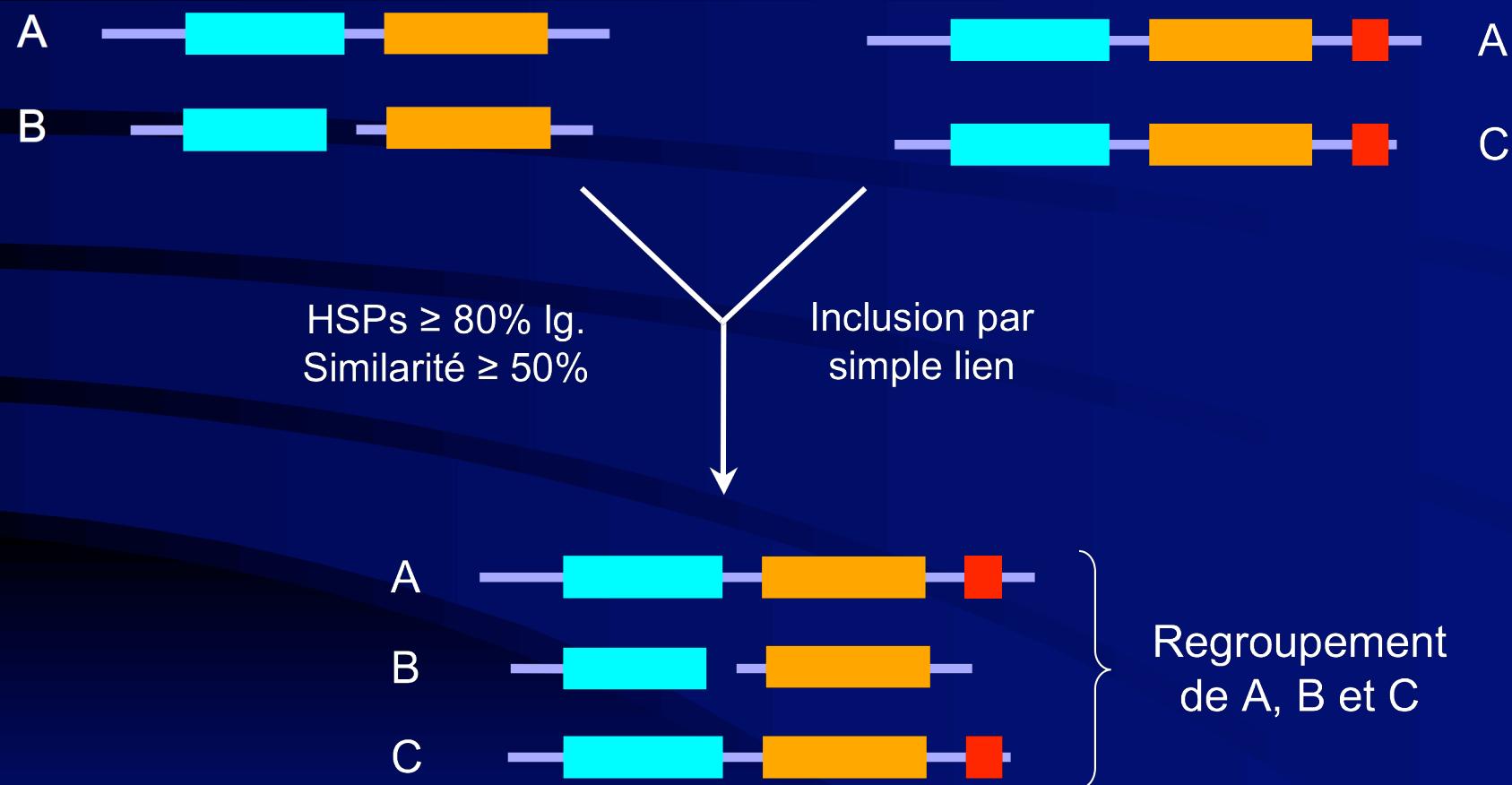
Recherche de similarités



Sélection des segments



Assemblage des familles

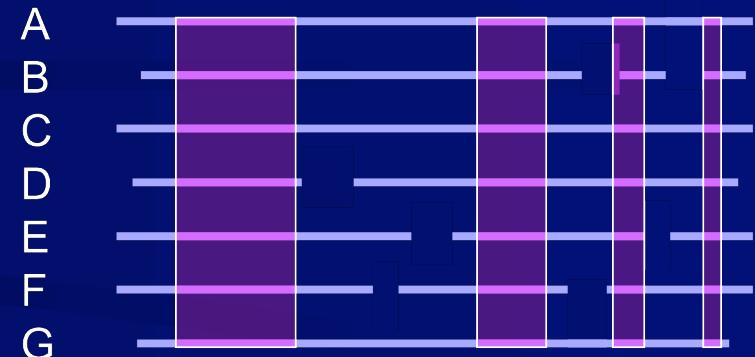


Alignements et arbres (1)

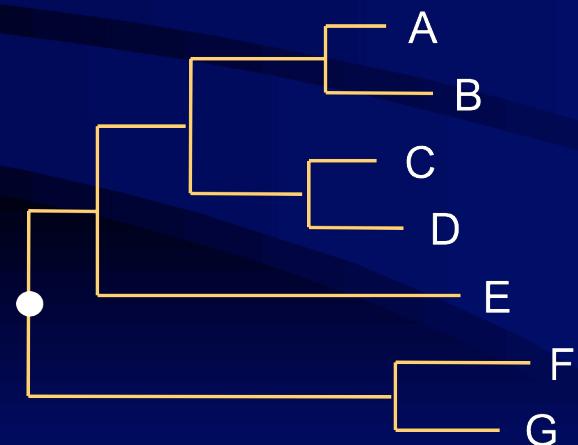


Famille non alignée

CLUSTAL W
Paramètres
Par défaut



Famille alignée



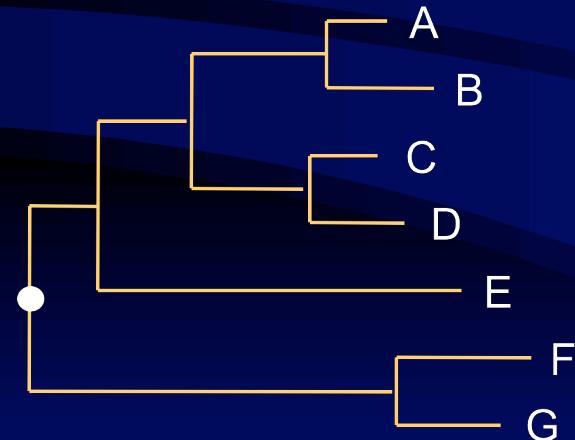
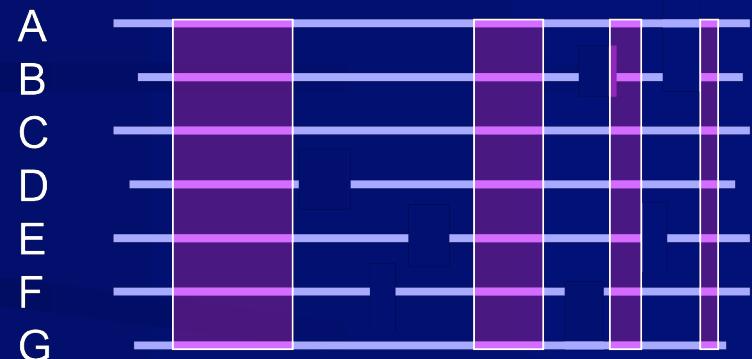
Arbre de la famille

BIONJ
Divergence
observée

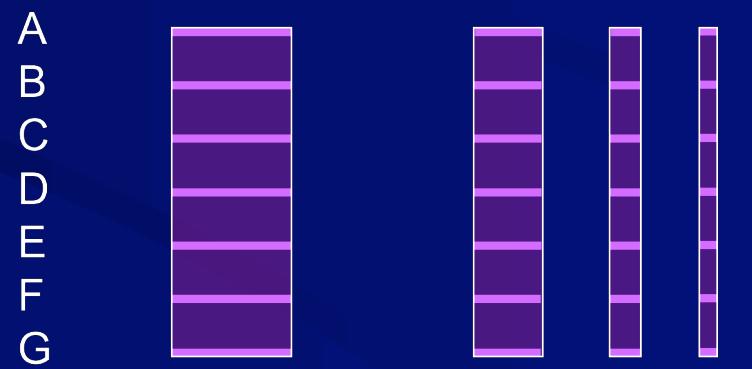
Alignements et arbres (2)



MUSCLE
Paramètres
Par défaut



PHYML
JTT + Γ



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

History **Species** **List** **Modify** **Retrieve** **Releases** **Help**

[Usage](#)

Taxon selection

List 1: Taxa to be selected: ([see example](#))

Mammalia
Drosophila
Caenorhabditis

[Search Type:](#)

List 2: Taxa to be excluded: ([see example](#))

Arabidopsis thaliana
Saccharomyces cerevisiae
Bacteria
Archaea

Protein [databank](#) Nucleotide [databank](#)

Database:

[Number of Sequences by family](#) 

[Number of Taxa by family](#)

List name:

Taxonomy helper

Data base: GenBank EMBL SwissProt/TrEMBL

Template: Tree depth:

Search:

You may use this page to **check the taxonomy** of the taxa you are interested in.

First choose a database, then give a species/taxon name and a depth for tree display. For example you may select *mammalia* with a depth equal to 3 in *GenBank*.

Warning:
you may give only one full species name at a time.

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		GENE FAMILY HBG191019	
HBG208967	120		Number of sequences	16
HBG174990	40		Number of taxons	5
HBG130840	27		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

Sequence list "HBG191019_lst"

History View Modify Retrieve Analyze

Number of sequences Matching sequences: 16

Number of taxons

Definition

A180_HUMAN Clathrin coat assembly protein AP180 (Clathrin coat associated protein AP180) (91 kDa synaptosomal-associated protein).

A180_MOUSE Clathrin coat assembly protein AP180 (Clathrin coat associated protein AP180) (91 kDa synaptosomal-associated protein) (Phosphoprotein Fl-20).

A180_RAT Clathrin coat assembly protein AP180 (Clathrin coat associated protein AP180) (91 kDa synaptosomal-associated protein).

Nucleotide Sequences Ref



Données sur les familles

	Keyword	Number of sequences	Percentage	Link to all associated sequences
Number of sequences	HBG191019	16	[100%]	Sequences related to HBG191019 in Hogenprot
Number of taxons	DOMAIN	14	[87%]	Sequences related to DOMAIN in Hogenprot
Definition	HYPOTHETICAL PROTEIN	8	[50%]	Sequences related to HYPOTHETICAL PROTEIN in Hogenprot
Nucleotide Sequences	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

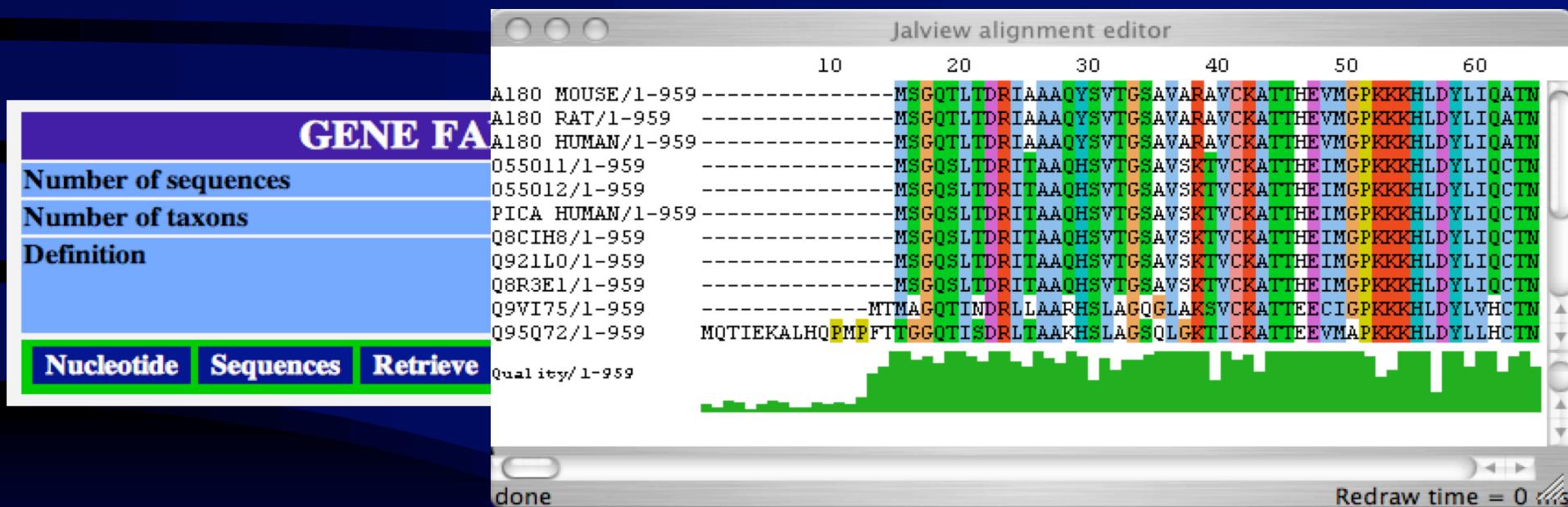
Données sur les familles

[View alignment with JalView](#) [Download alignment](#)

Threshold for coloring = 80% of identity ([100%](#) [90%](#) [80%](#) [70%](#) [60%](#) [50%](#) [Clustalw consensus](#))

GENE F		
Nucleotide	Sequences	Retrieval
A180_MOUSE	-----MSGQTLTDRIAAQYSVTGSAVARAVCKATTHEV MGPKKKHLDYL	
A180_RAT	-----MSGQTLTDRIAAQYSVTGSAVARAVCKATTHEV MGPKKKHLDYL	
A180_HUMAN	-----MSGQTLTDRIAAQYSVTGSAVARAVCKATTHEV MGPKKKHLDYL	
055011	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
055012	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
PICA_HUMAN	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
Q8CIH8	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
Q921L0	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
Q8R3E1	-----MSGQSLTDRITAAQHSVTGSAVS KTVCKATTHEIMGPKKKHLDYL	
Q9VI75	-----MTMAGQTINDRLLAARHSLAGQGLAKSVCKATTEECIG PKKKHLDYL	
Q95072	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL	
Q9XZI6	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL	
Q9UA05	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL	
Q9XZI7	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL	
Q95073	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL	
Q95075	MQTIEKALHQ PMPFTTGGQTISDRLTA A KHSLAGSQLGKTICKATTEEVMA PKKKHLDYL .*:.*:.*: *:.*:.*: . :.*:*****.* :*****	

Données sur les familles

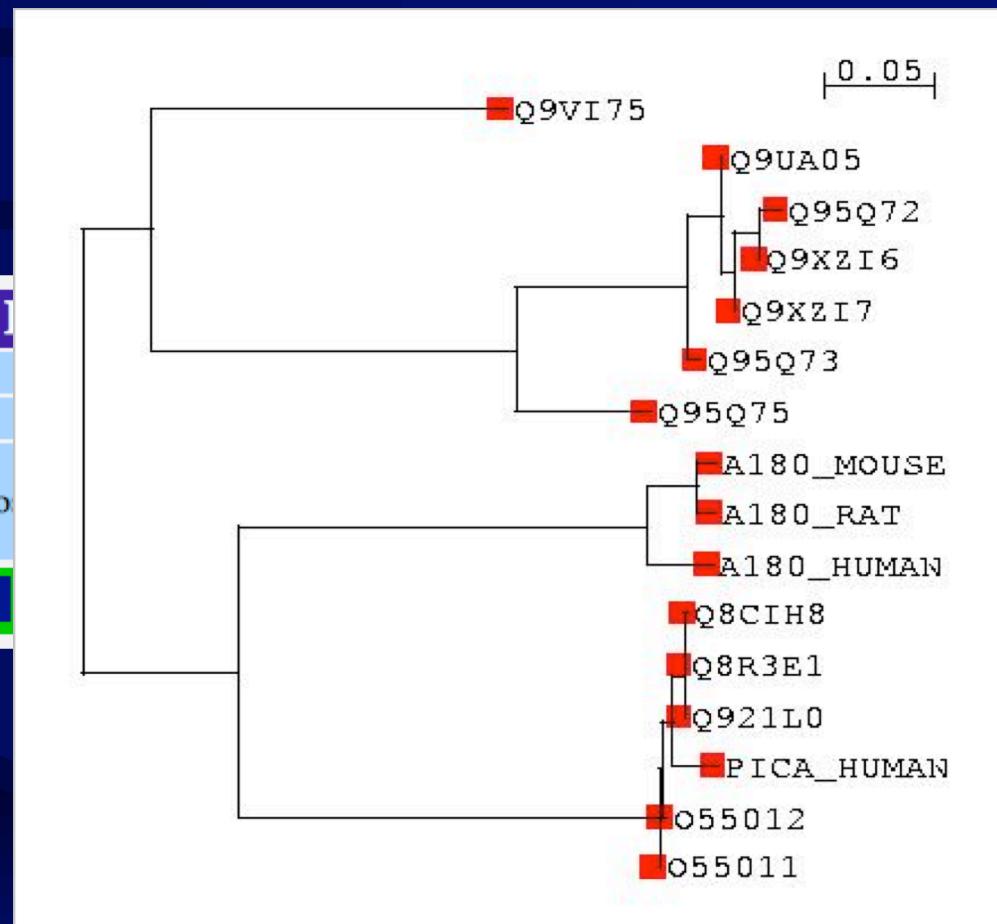


Données sur les familles

GENE FAMILY ID: 10000000000000000000000000000000

Number of sequences	Number of taxons	Definition
1	1	Phosphatidylinositol-3-O-

Nucleotide Sequences Retrieve Species



Données sur les familles

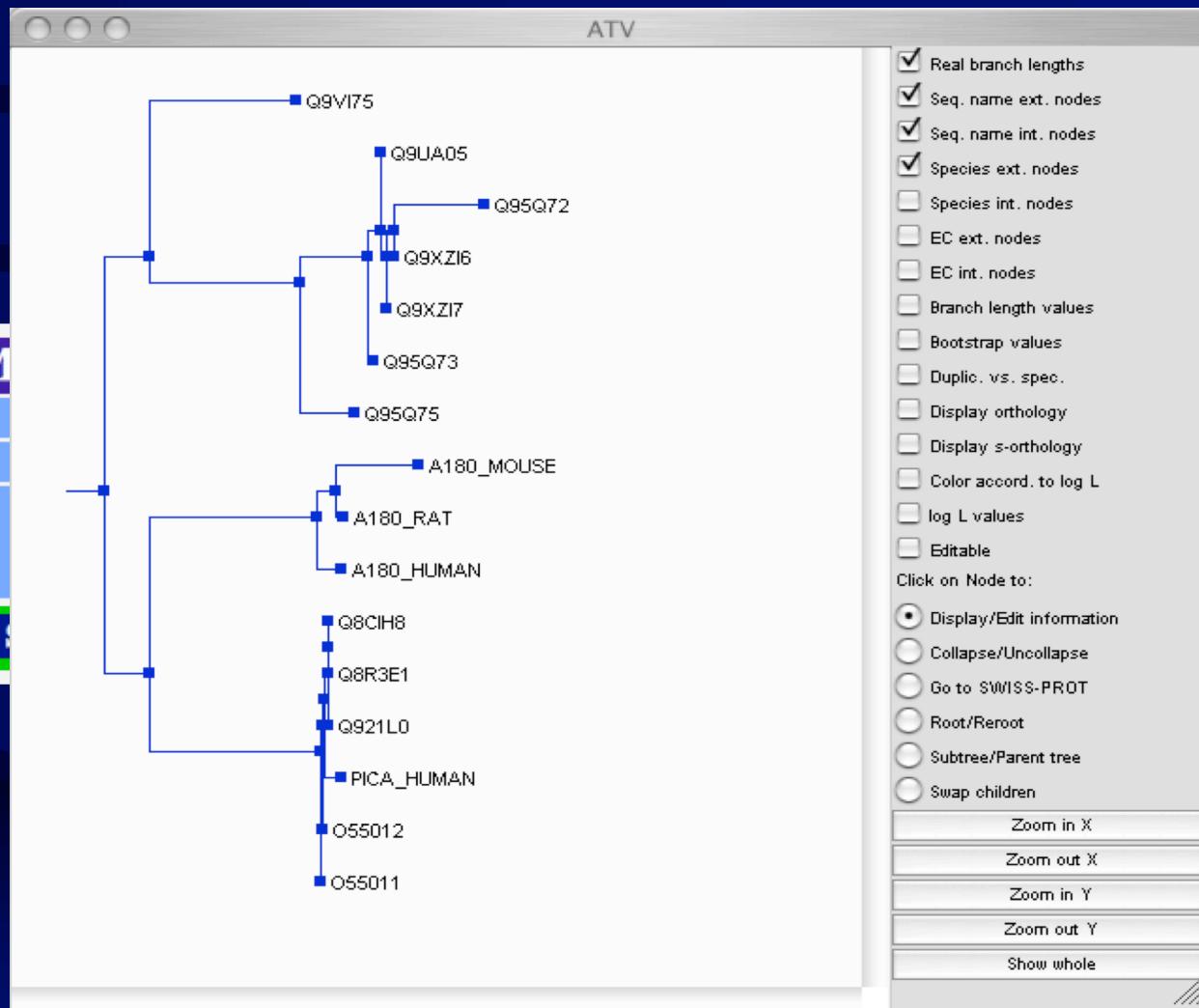
GENE FAM

Number of sequences

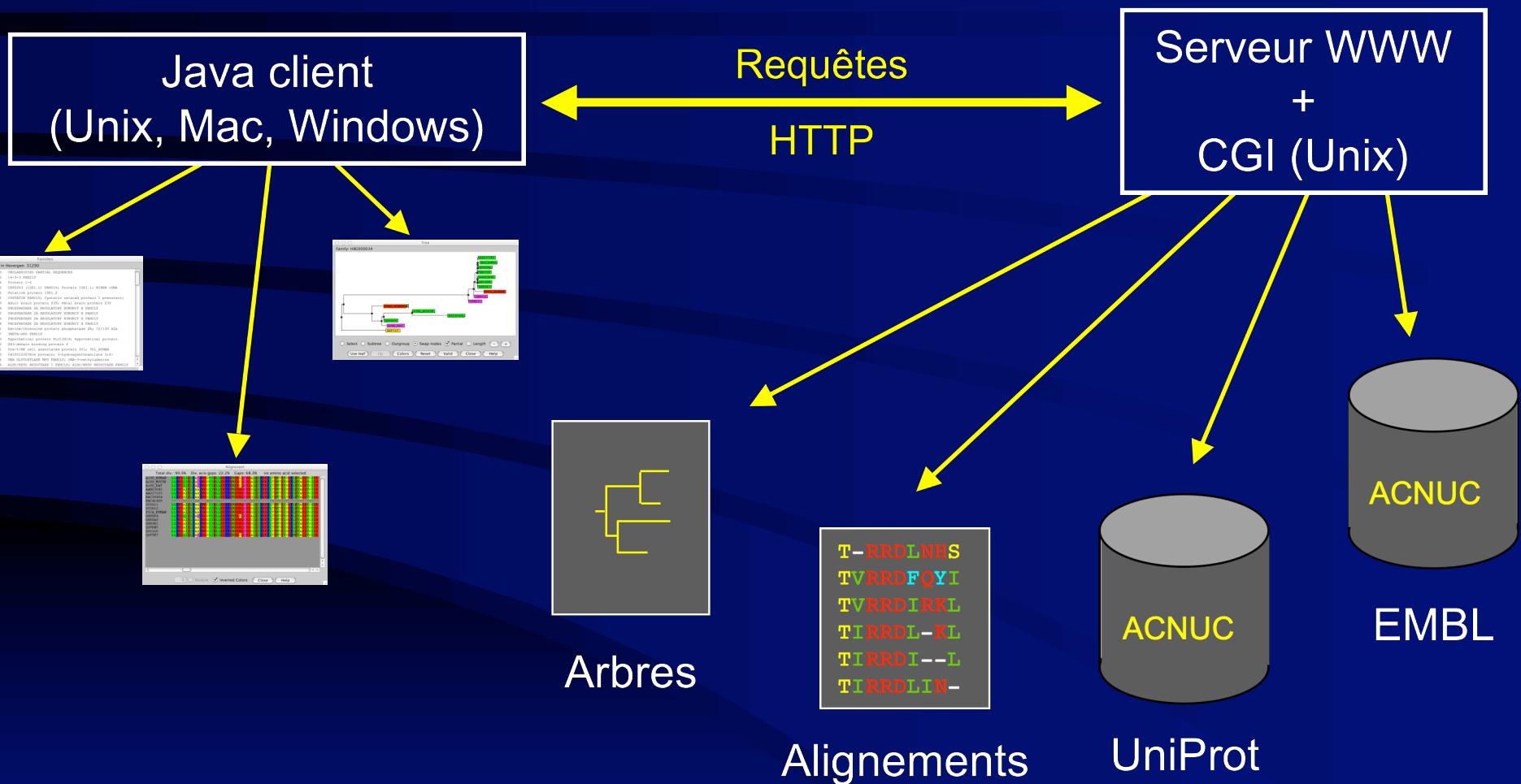
Number of taxons

Definition

Nucleotide Sequences Retrieve



Organisation client/serveur



L'interface FamFetch

Families			
Total number of families in Hovergen: 31290			
HBG000000	26749	2803	UNCLASSIFIED PARTIAL SEQUENCES
HBG000002	58	10	14-3-3 FAMILY
HBG000003	6	4	Protein 1-4
HBG000004	3	2	UPF0203 (15E1.1) FAMILY; Protein 15E1.1; RIKEN cDNA
HBG000005	4	2	Putative protein 15E1.2
HBG000007	3	2	CYSTATIN FAMILY; Cystatin related protein 1 precursor;
HBG000008	10	3	Adult brain protein 239; Fetal brain protein 239
HBG000009	30	6	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000010	2	2	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000011	18	6	PHOSPHATASE 2A REGULATORY SUBUNIT A FAMILY
HBG000012	31	8	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000013	1	1	Serine/threonine protein phosphatase 2A; 72/130 kDa
HBG000014	43	17	3BETA-HSD FAMILY
HBG000015	9	3	Hypothetical protein FLJ13219; Hypothetical protein
HBG000016	3	2	SH3-domain binding protein 2
HBG000017	1	1	Pre-T/NK cell associated protein 3CL; 3CL_HUMAN
HBG000018	3	3	0610012J07Rik protein; 3-hydroxyanthranilate 3;4-
HBG000019	5	4	DNA GLYCOSYLASE MPG FAMILY; DNA-3-methyladenine
HBG000020	83	16	ALDO/KETO REDUCTASE 1 FAMILY; ALDO/KETO REDUCTASE FAMILY

L'interface FamFetch

Families

Total number of families in Hovergen: 31290

				UNCLASSIFIED PARTIAL SEQUENCES
HBG000000	26749	2803		
HBG000002	58	10	14-3-3	
HBG000003	6	4		Protein
HBG000004	3	2		UPF020
HBG000005	4	2		Putati
HBG000007	3	2		CYSTAI
HBG000008	10	3		Adult
HBG000009	30	6		PHOSPH
HBG000010	2	2		PHOSPH
HBG000011	18	6		PHOSPH
HBG000012	31	8		PHOSPH
HBG000013	1	1		Serine
HBG000014	43	17		3BETA-
HBG000015	9	3		Hypoth
HBG000016	3	2		SH3-dc
HBG000017	1	1		Pre-T/
HBG000018	3	3		061001
HBG000019	5	4		DNA GL
HBG000020	83	16		ALDO/K

Tree

Family: HBG000034

The tree diagram illustrates the evolutionary relationships between the sequences listed in the table. The root node is on the left, and the tree branches to the right. Nodes are labeled with sequence identifiers and species names. Some nodes are highlighted with colored boxes: green for most, red for A180_HUMAN, yellow for Q9PTK7, and purple for O55011 and O55012. The tree structure shows a complex branching pattern with multiple sister groups.

Select Subtree Outgroup Swap nodes Partial Length - +

Use leaf Up Colors Reset Valid Close Help

L'interface FamFetch

Families

Total number of families in Hovergen: 31290

				UNCLASSIFIED PARTIAL SEQUENCES	
HBG000000	26749	2803			
HBG000002	58	10	14-3-3		
HBG000003	6	4		Protein	
HBG000004	3	2		UPF020	
HBG000005	4	2		Putati	
HBG000007	3	2		CYSTAI	
HBG000008	10	3	Adult		
HBG000009	30	6		PHOSPH	
HBG000010	2	2		PHOSPH	
HBG000011	18	6		PHOSPH	
HBG000012	31	8		PHOSPH	
HBG000013	1	1	Serine		
HBG000014	43	17	3BETA-		
HBG000015	9	3	Hypoth		
HBG000016	3	2	SH3-dc		
HBG000017	1	1	Pre-T/		
HBG000018	3	3	061001		
HBG000019	5	4	DNA GL		
HBG000020	83	16	ALDO/K		

A180_HUMAN

ID A180_HUMAN STANDARD; PRT; 907 AA.

AC O60641; Q9NTY7;

DT 15-JUN-2002 (Rel. 41, Created)

DT 15-JUN-2002 (Rel. 41, Last sequence update)

DT 15-JUN-2002 (Rel. 41, Last annotation update)

DE Clathrin coat assembly protein AP180 (Clathrin coat associated protein AP180) (91 kDa synaptosomal-associated protein).

GN SNAP91 OR KIAA0656.

OS Homo sapiens (Human).

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

OX NCBI_TaxID=9606;

RN [1]

RP SEQUENCE FROM N.A.

RC TISSUE=BRAIN;

RX MEDLINE=98403880; PubMed=9734811;

RA Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,

RA Kotani H., Nomura N., Ohara O.;

RT "Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro.";

RL DNA Res. 5:169-176(1998).

RN [2]

[Close](#) [Help](#)

L'interface FamFetch

Families

Total number of families in Hovergen: 31290

			UNCLASSIFIED PARTIAL SEQUENCES
HBG000000	26749	2803	
HBG000002	58	10	14-3-3
HBG000003	6	4	Protein
HBG000004	3	2	UPF020
HBG000005	4	2	Putati
HBG000007	3	2	CYSTAI
HBG000008	10	3	Adult
HBG000009	30	6	PHOSPH
HBG000010	2	2	PHOSPH
HBG000011	18	6	PHOSPH
HBG000012	31	8	PHOSPH
HBG000013	1	1	Serine
HBG000014	43	17	3BETA-
HBG000015	9	3	Hypoth
HBG000016	3	2	SH3-dc
HBG000017	1	1	Pre-T/
HBG000018	3	3	061001
HBG000019	5	4	DNA GL
HBG000020	83	16	ALDO/K

Total div.: 90.5% Div. w/o gaps: 22.2% Gaps: 68.3% no amino acid selected

Alignment

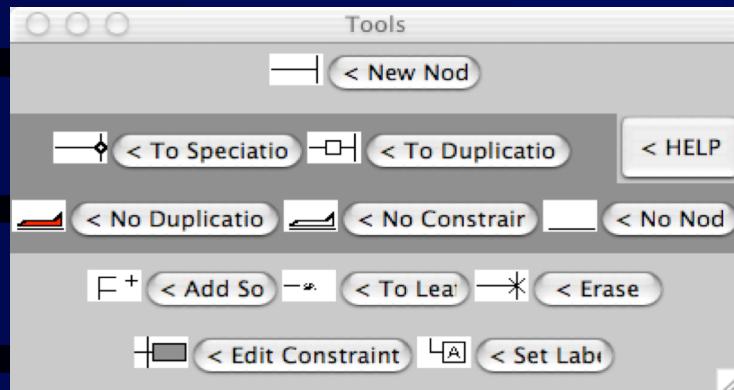
A180_HUMAN: LLFKDLIKLFACYNDGVINNLEKFFEMKKGQCKDALEIYKRFLLTRMTRVSEFLKVAEQVGIDK
A180_MOUSE: LLFKDLIKLFACYNDGVINNLEKFFEMKKGQCKDALEIYKRFLLTRMTRVSEFLKVAEQVGIDK
A180_RAT: LLFKDLIKLFACYNDGVINNLEKFFEMKKGQCKDALEIYKRFLLTRMTRVSEFLKVADEVVGIDK
AAH23843: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
AAO17153: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
BAC39454: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
BAC41439: -
055011: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
055012: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
PICA_HUMAN: LLFKDAIRLFAAYHEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
Q8K0D4: LLFKDLIKLFACYNDGVINNLEKFFEMKKGQCKDALEIYKRFLLTRMTRVSEFLKVAEQVGIDK
Q8R0A9: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
Q8R3E1: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
Q8VDN5: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
Q921L0: LLFKDAIRLFAAYNEGIIMNLEKYFDMMKKNQCKEGLDIYKKFLTRMTRISEFLKVAEQVGIDR
Q9PTK7: LLFKDLIKLFACYNDGVINNLEKFFEMKKGQCKDALEIYKRFLLTRMTRVSEFMKVAEQVGIDK

Module Inverted Colors

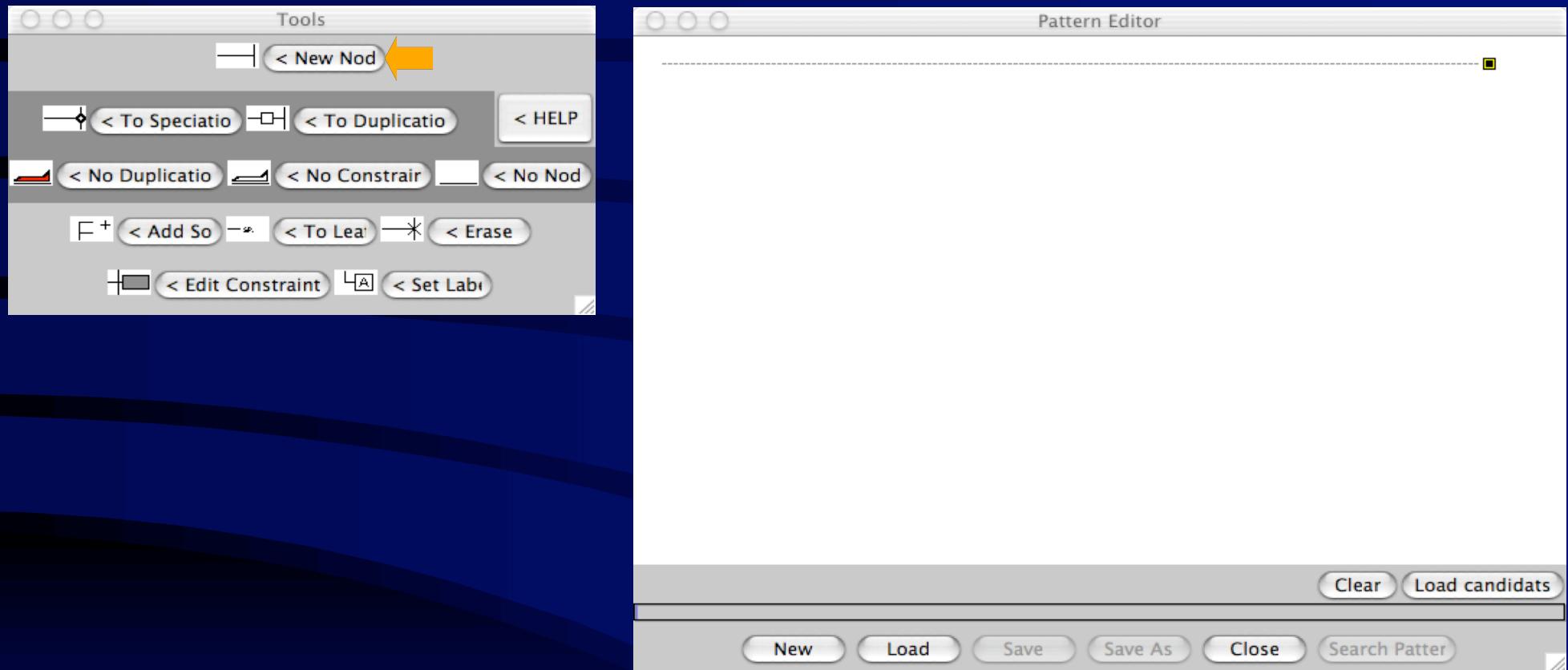
Recherche de motifs

- 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.

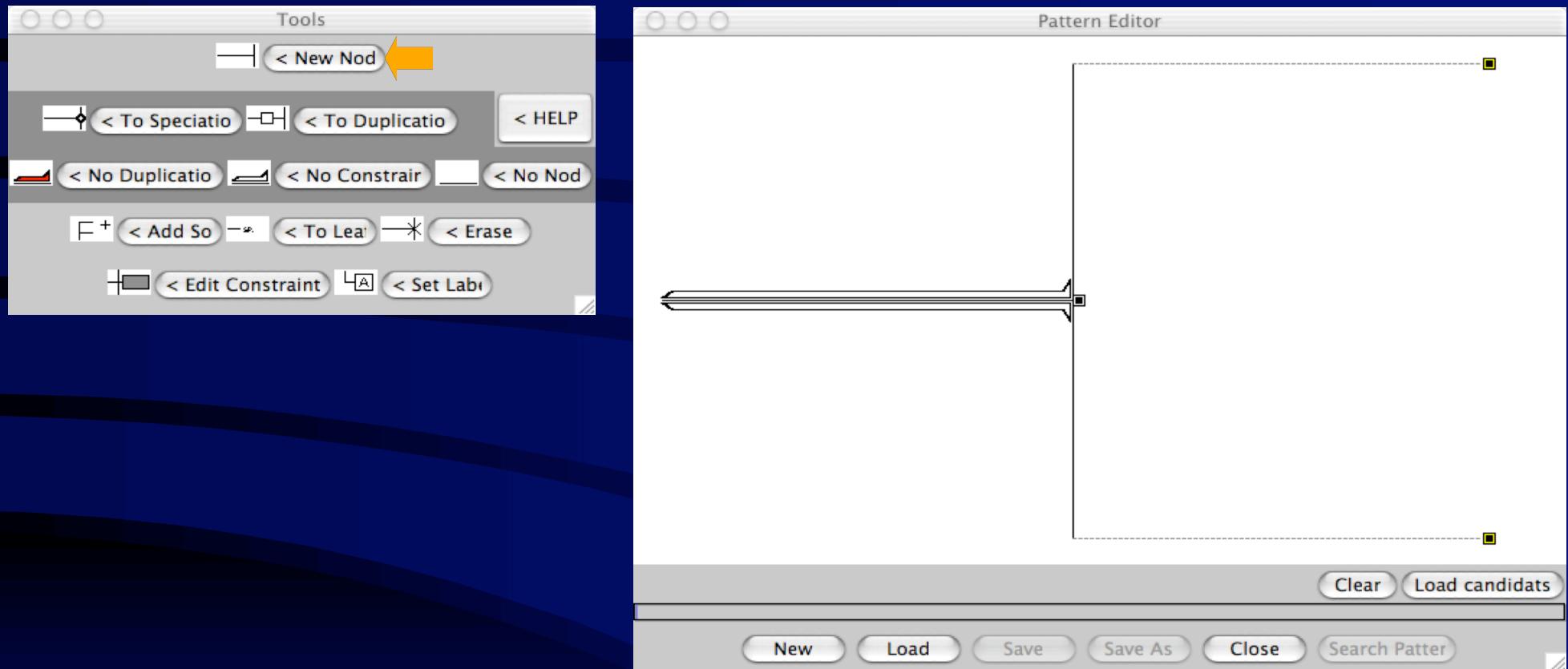
Construction d'une requête



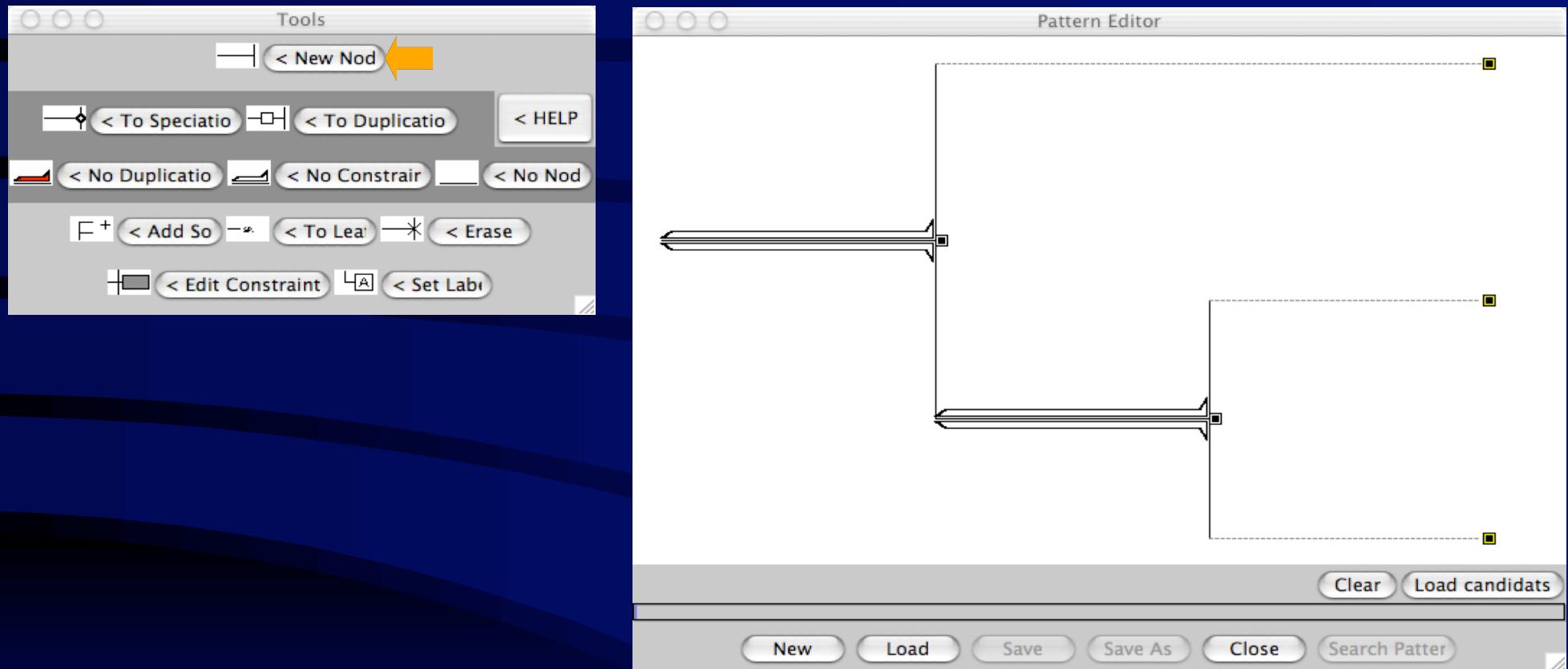
Construction d'une requête



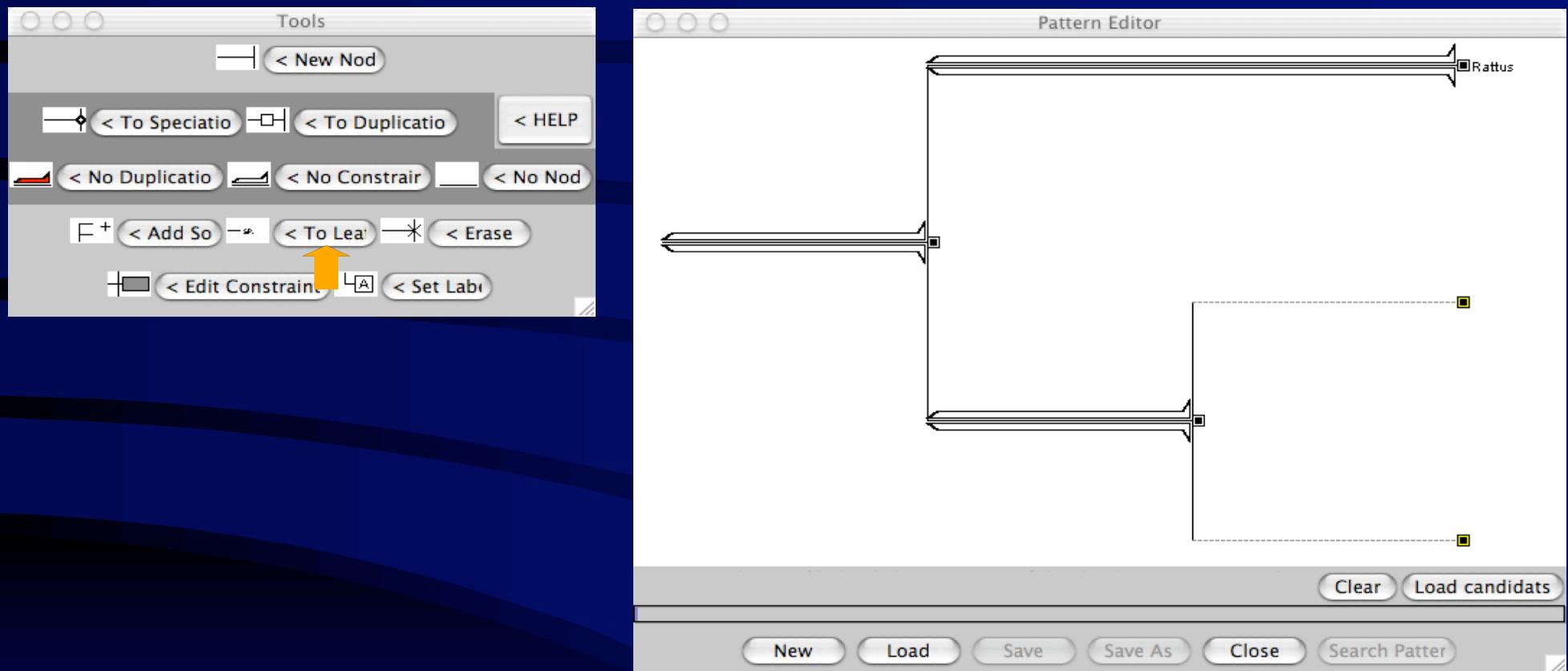
Construction d'une requête



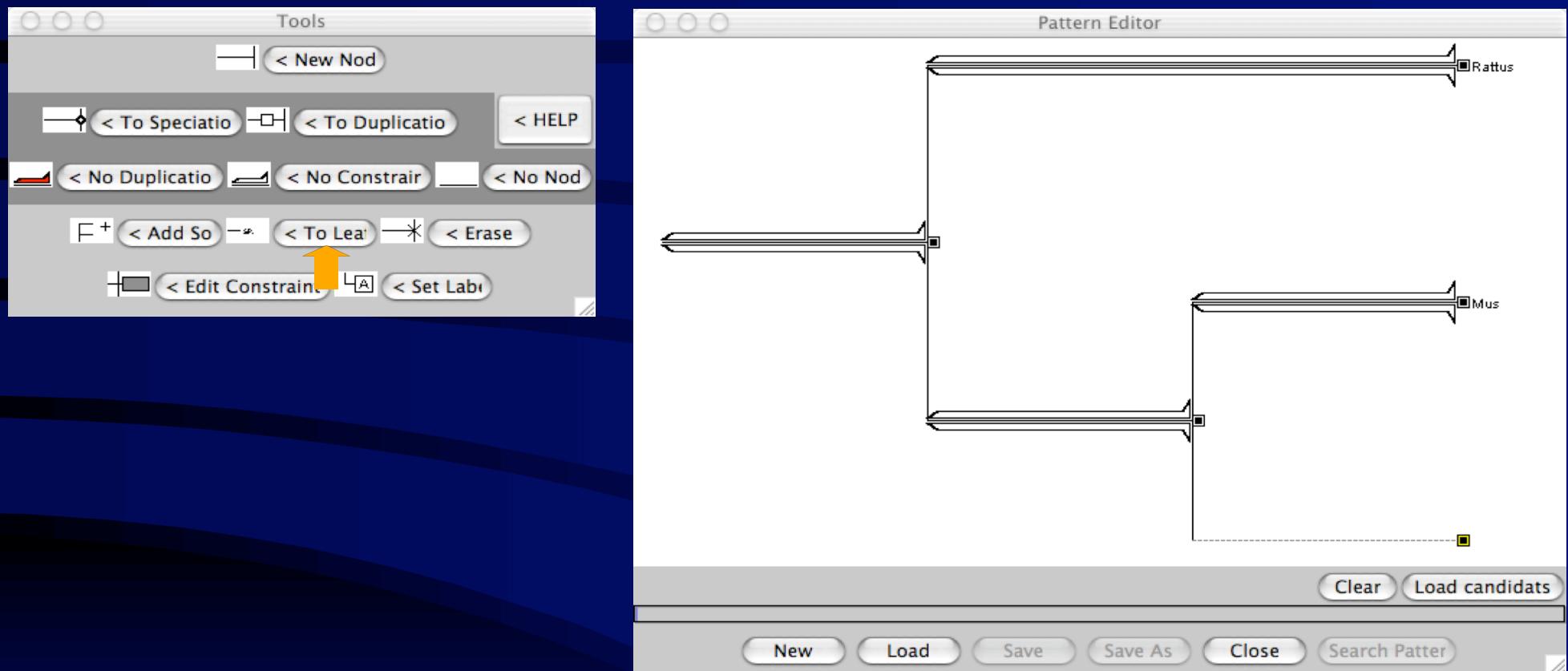
Construction d'une requête



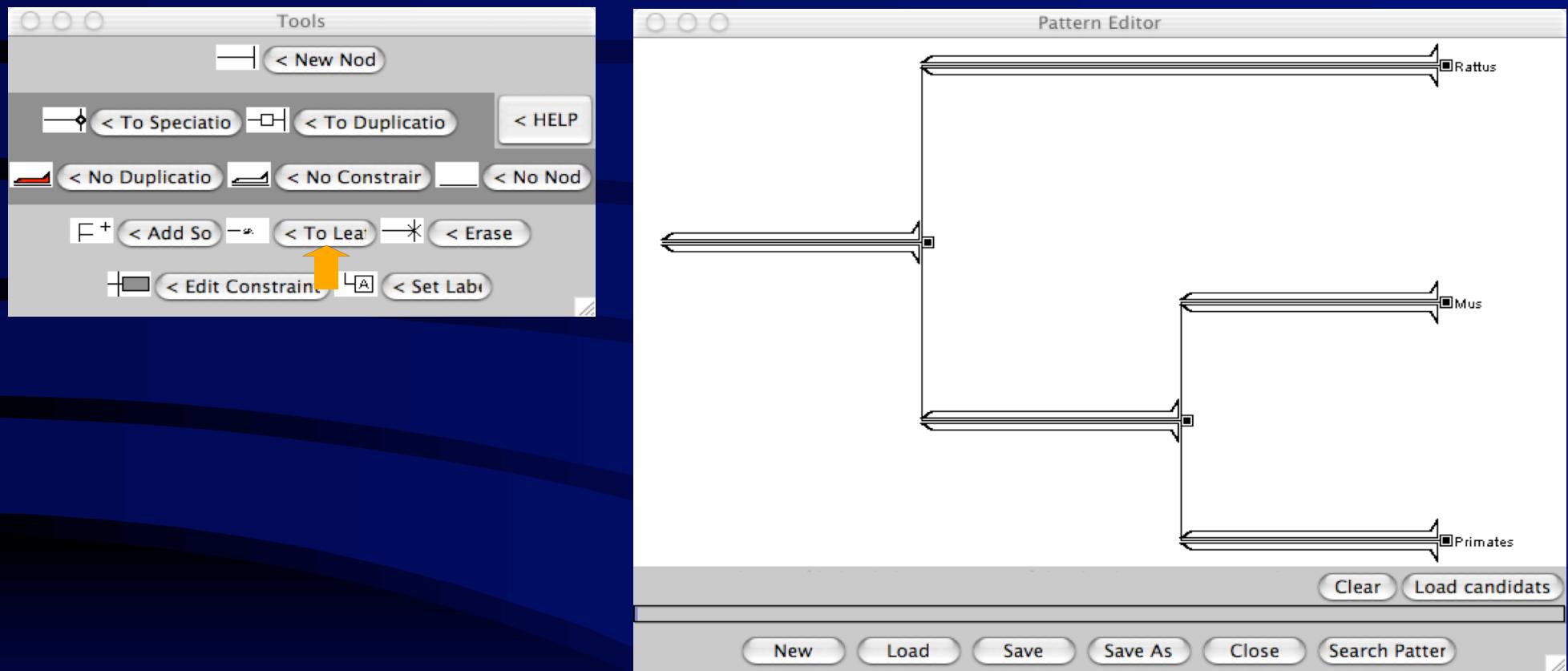
Construction d'une requête



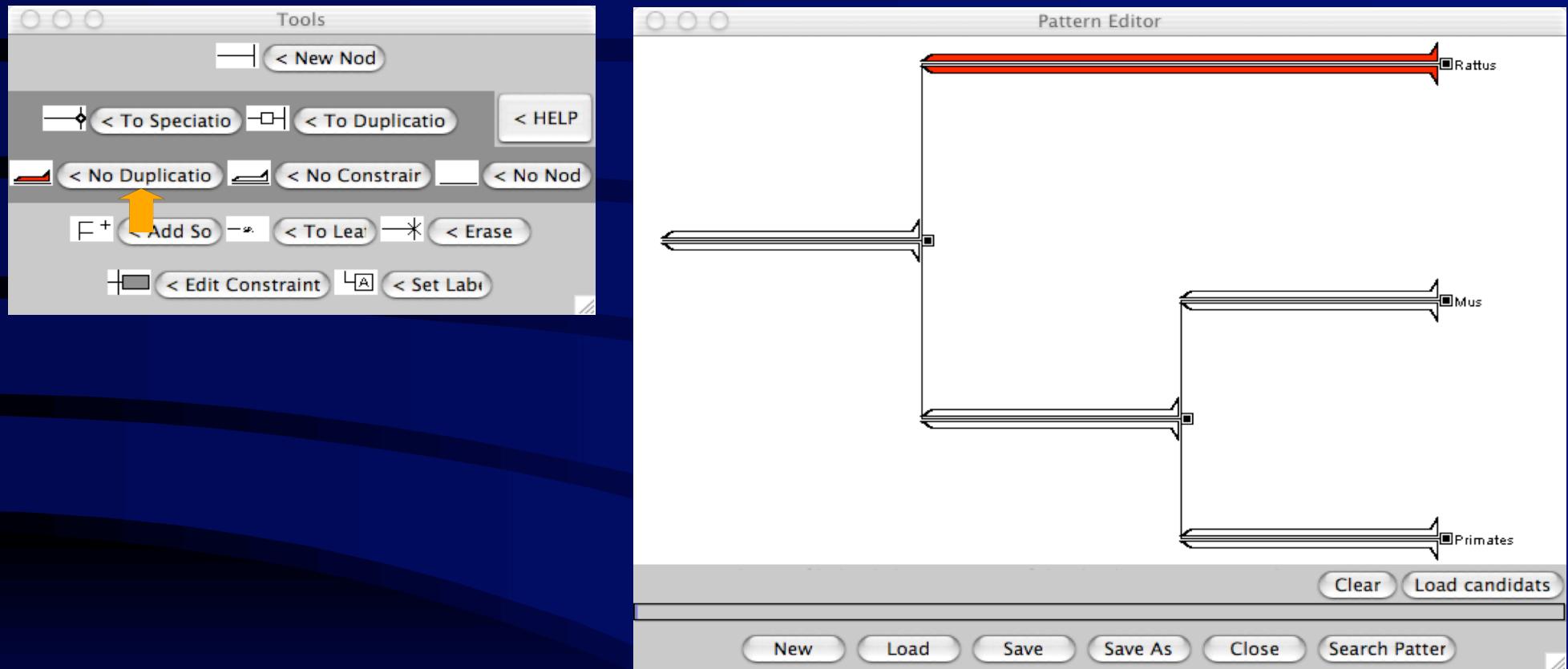
Construction d'une requête



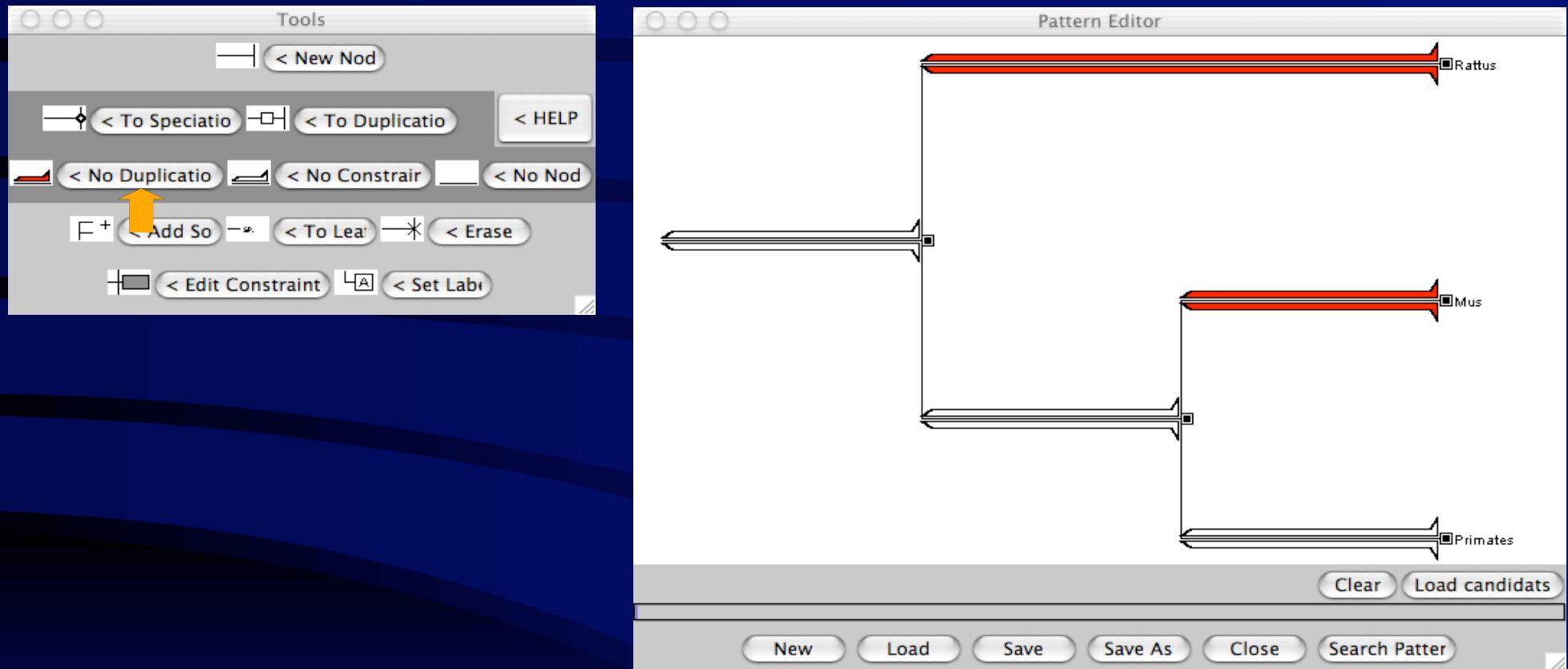
Construction d'une requête



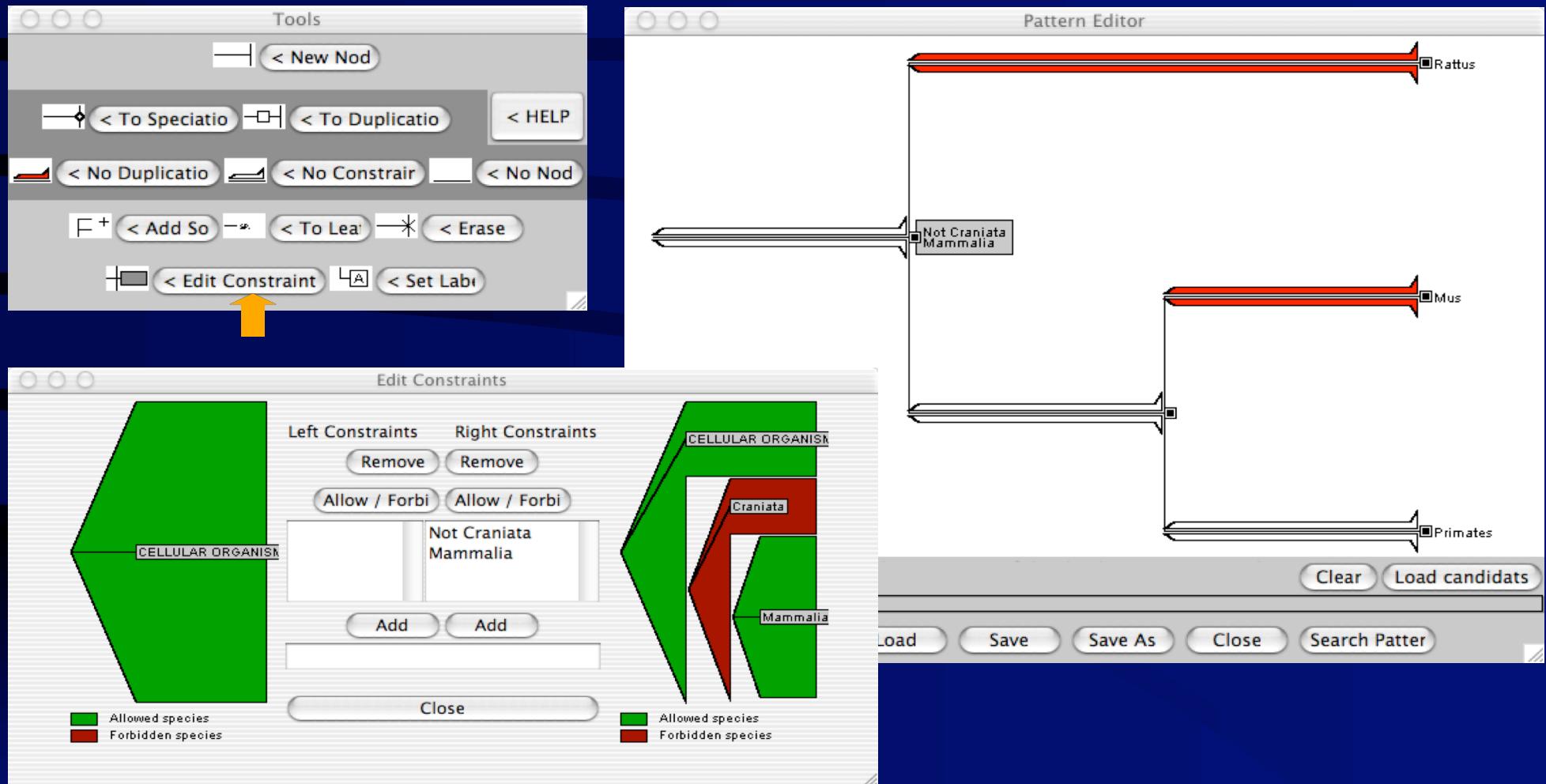
Construction d'une requête



Construction d'une requête



Construction d'une requête



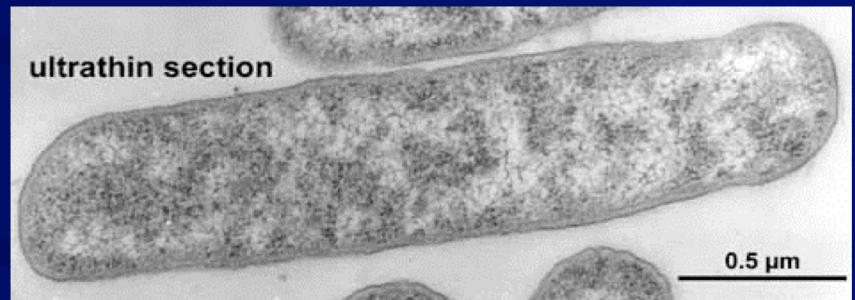
Exemple d'analyse

- 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'HOMOGENOM ($100 \geq$ nb. taxons ≥ 4) contenant des séquences de ces organismes :
 - Étude des incongruences phylogénétiques.

Bactéries hyperthermophiles

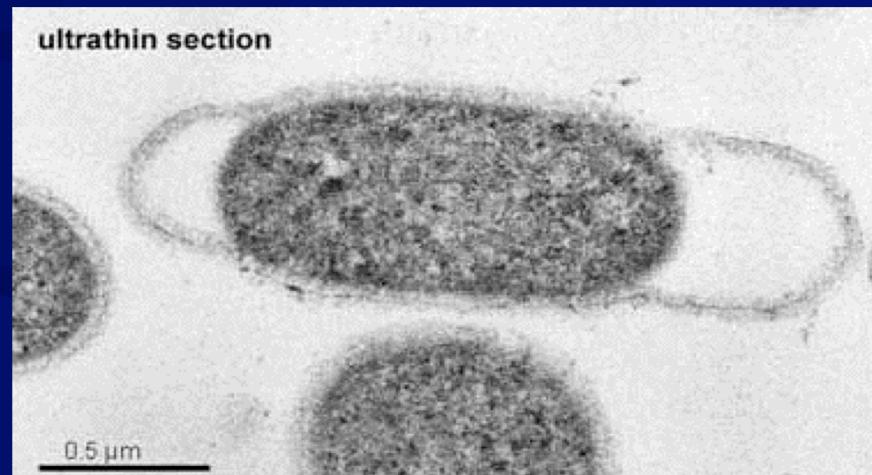
- *Aquifex aeolicus* :

- Aérobie.
- $T_{opt} = 90^\circ\text{C}$.
- Isolée dans des sources chaudes dans le parc de Yellowstone.



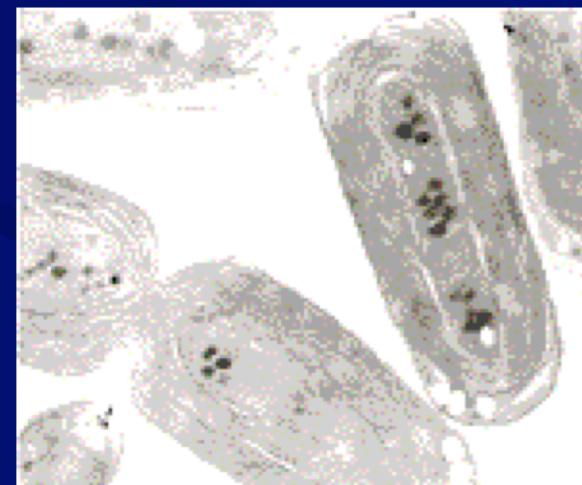
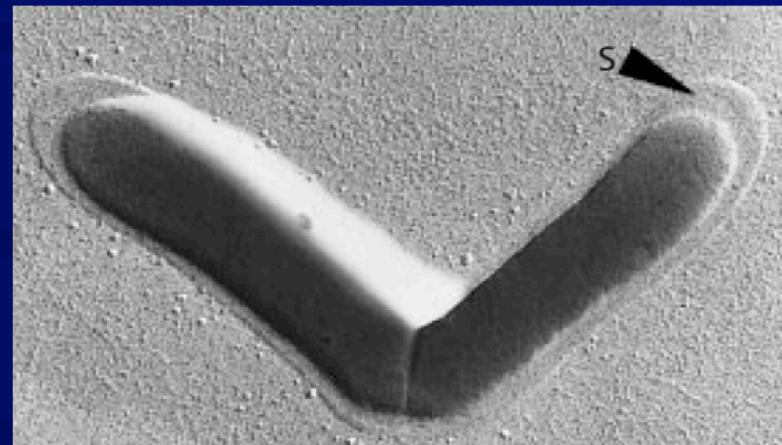
- *Thermotoga maritima* :

- Anaérobie.
- $T_{opt} = 80^\circ\text{C}$.
- Vit dans les sédiments marins proches des cheminées géothermiques.

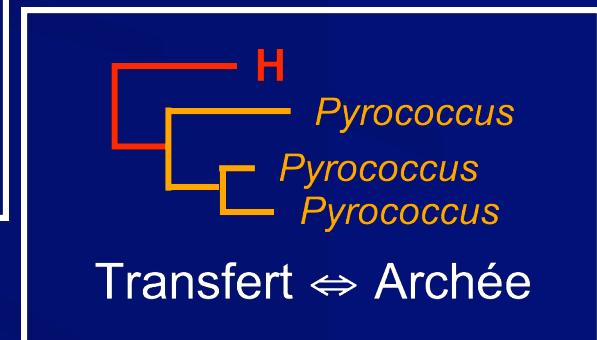
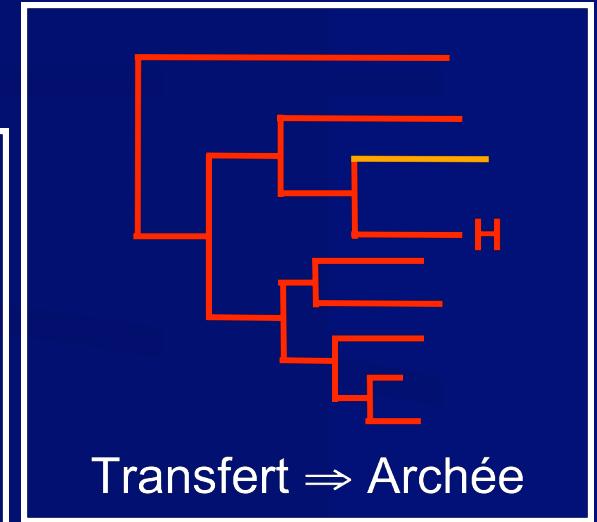
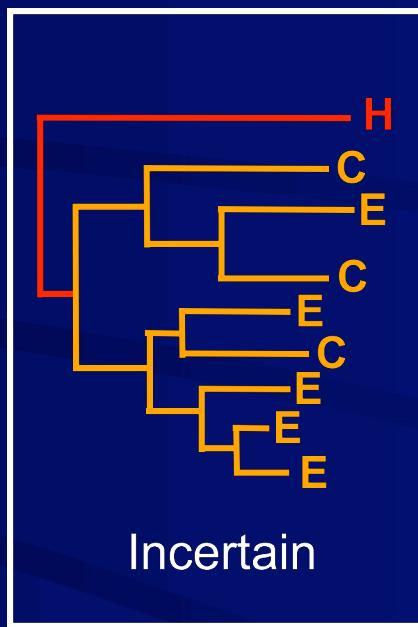
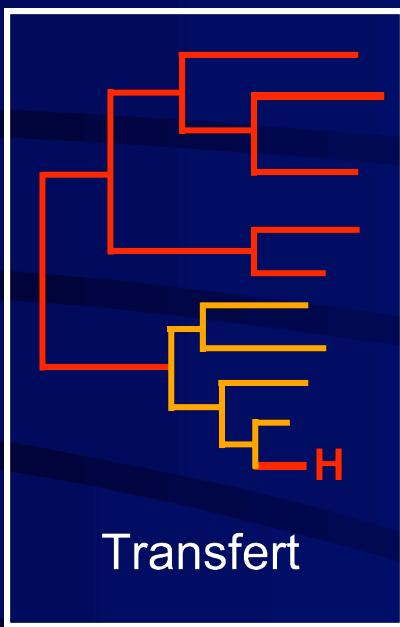
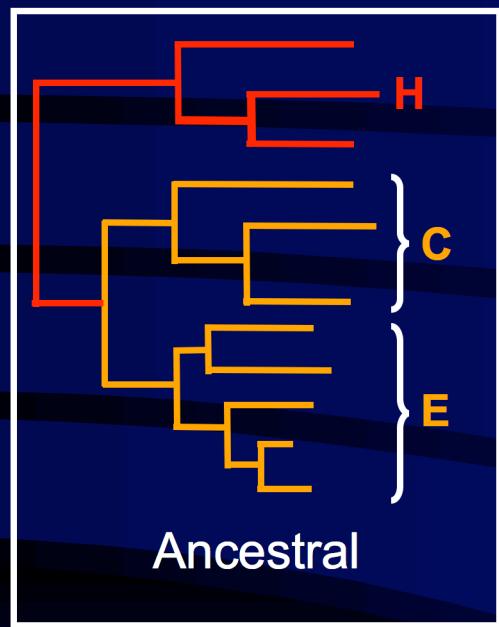


Contrôles

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



Types d'incongruences



H : Bactérie hyperthermophile

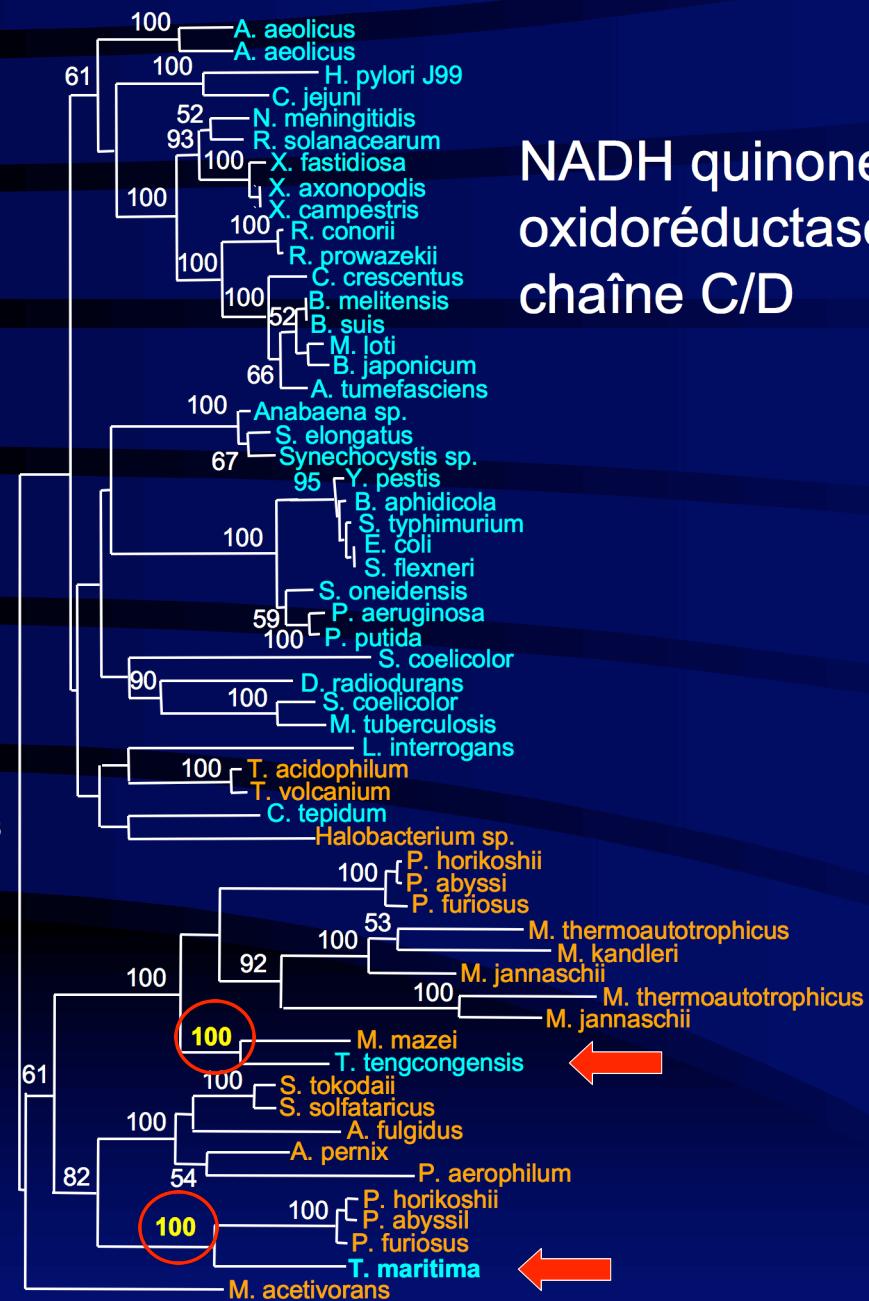
E : Euryarchées C : Crénarchées

■ Bactéries

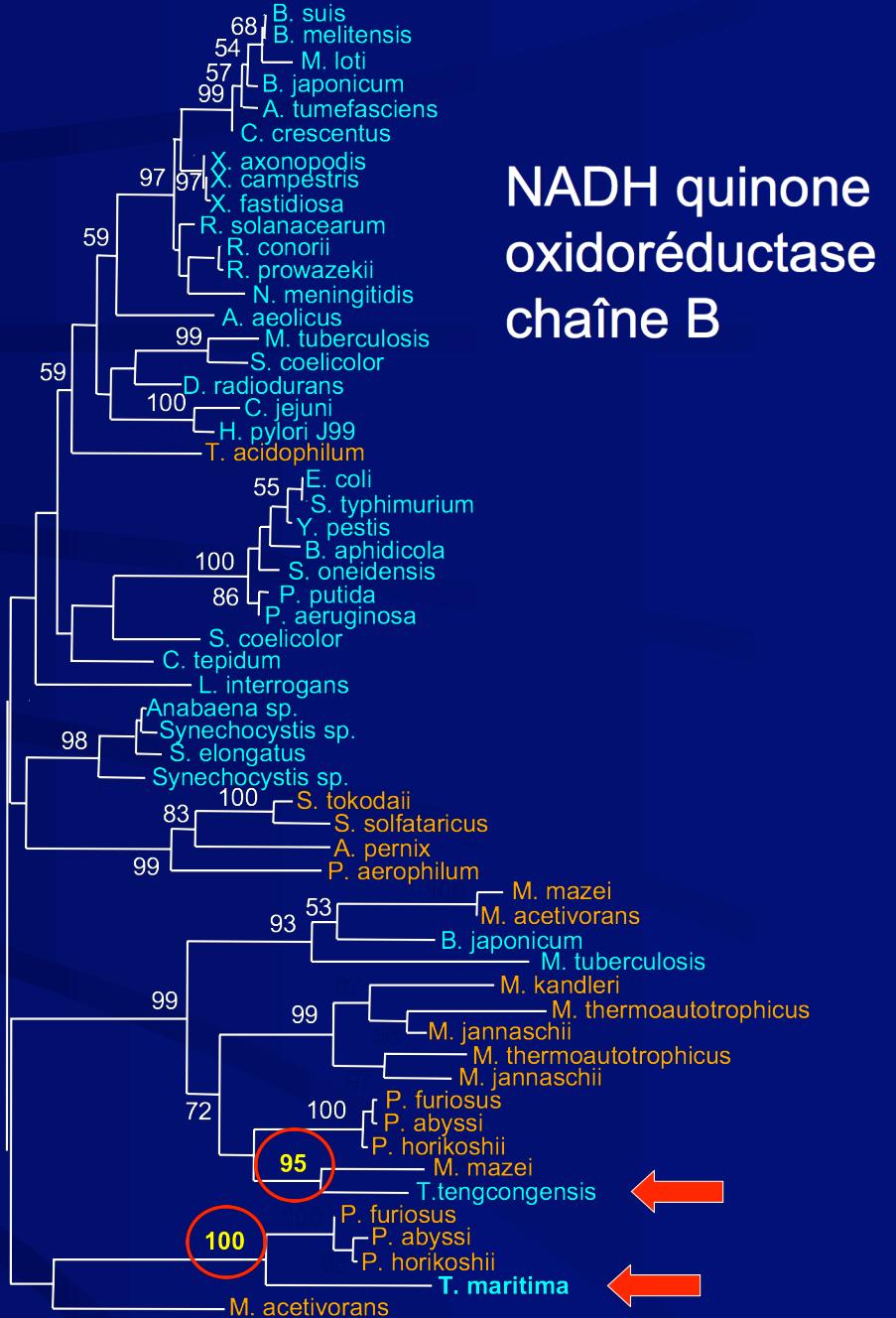
■ Archées

Exemples de transferts

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



NADH quinone oxidoreductase chaîne C/D



NADH quinone oxidoreductase chaîne B

Nombre de transferts détectés

Thermotoga



1,86 MB
1858 gènes



49 évts
(84-85 gènes)

Aquifex



1,55 MB
1529 gènes



26 évts
(34 gènes)

Thermoanaerobacter



2,68 MB
2588 gènes



30 évts
(47-56 gènes)

Thermo-synechococcus

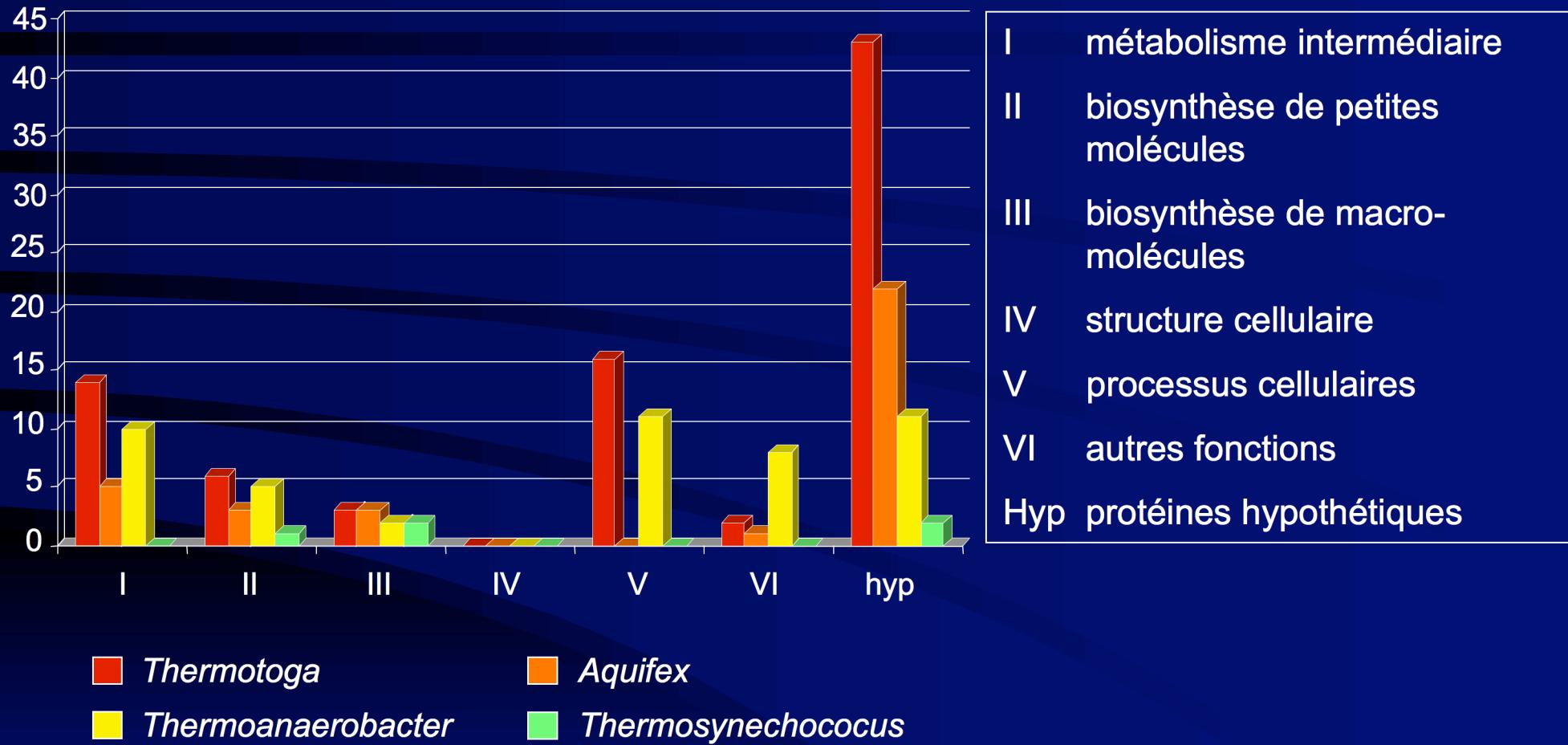


2,59 MB
2475 gènes



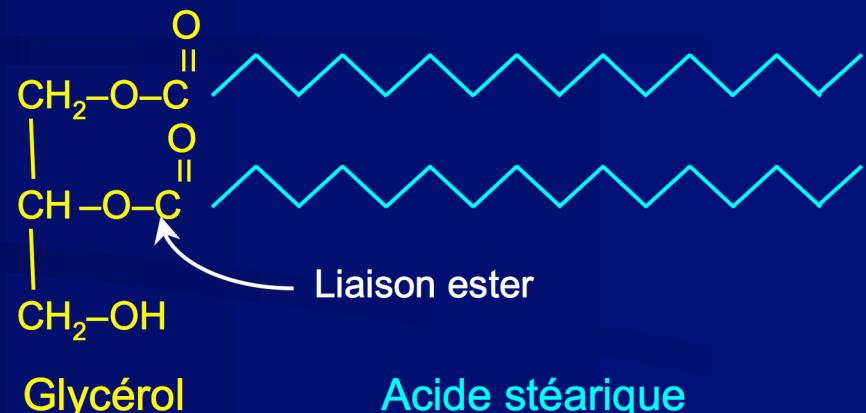
4 évts
(4 gènes)

Classification fonctionnelle



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 hyperthermophiles (Aquificales, Thermotogales, mais aussi Clostridiales) :
 - Plusieurs apparitions indépendantes de l'hyperthermo-phylie (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