

# **Calcul intensif en Bioinformatique**

**Réalisations en génétique et  
génomique évolutives**

**Pôle Rhône Alpin de BioInformatique (PRABI)  
Laboratoire de Biologie et Biométrie Evolutives (LBBE)**

# Contexte

De plus en plus de génomes décryptés de plus en plus rapidement, chez  
les mammifères,  
les oiseaux,  
les poissons,  
les insectes,  
les plantes,  
des centaines de bactéries.

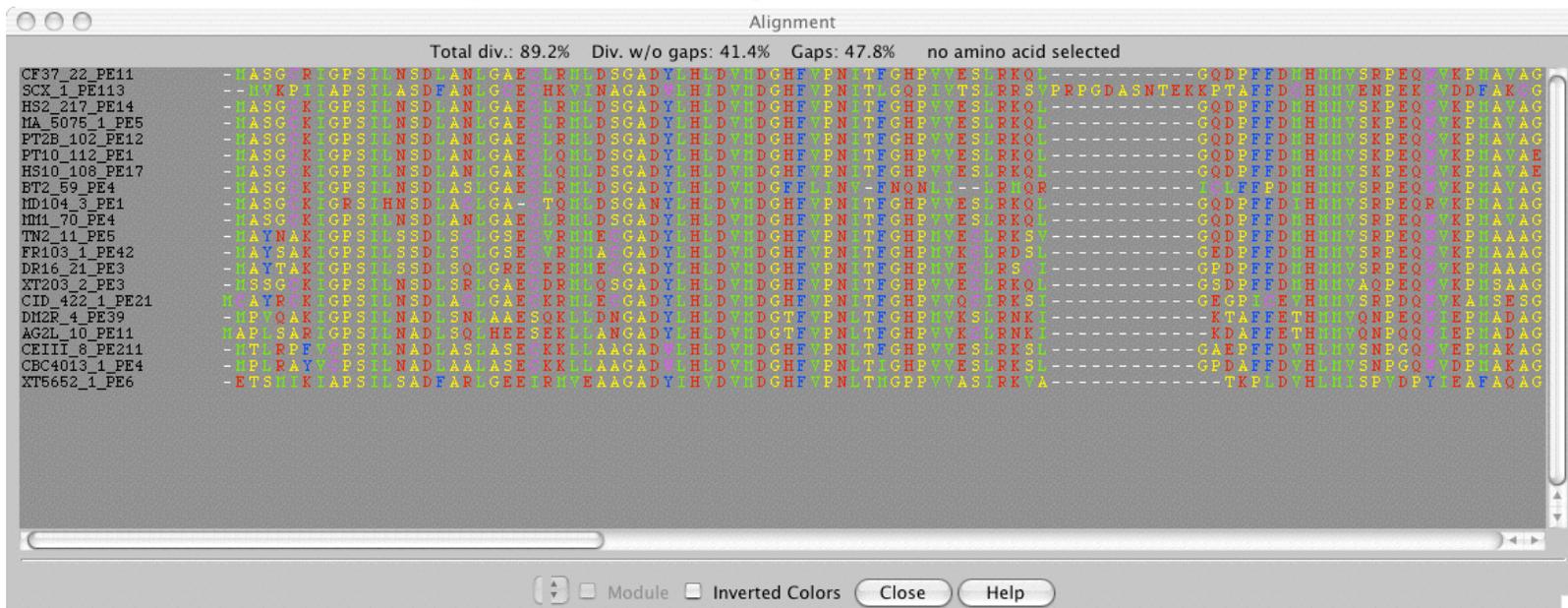
Il s'agit maintenant de *comprendre le contenu* de ces génomes, leur *structure*,  
leur *fonctionnement* et leur *origine*.

Le Laboratoire de Biométrie et Biologie Evolutive (LBBE) s'attache à décrire et  
comprendre les *mécanismes moléculaires* et les *forces évolutives* qui  
gouvernent l'organisation et l'évolution des génomes

# L'évolution des gènes et l'évolution des espèces

La construction, à partir de leur séquences alignées, d' « arbres phylogénétiques » de gènes, décrivant les relations évolutives entre les gènes, permet de retracer l'histoire des espèces à travers celle de leurs gènes.

## Matériel de base : l'alignement de séquences



The screenshot shows a window titled "Alignment" with the following statistics: Total div.: 89.2%, Div. w/o gaps: 41.4%, Gaps: 47.8%, and no amino acid selected. The window displays a multiple sequence alignment of 25 protein sequences, each starting with a hyphen and a sequence identifier. The sequences are color-coded by amino acid type: red for basic (K, R), green for acidic (D, E), blue for hydrophobic (L, I, V, F, Y, W, M), orange for polar (S, T, N, Q, R, G, P, A, D, S, T, E, K, R, Q, W, K, P, A, V, A, G), and purple for other (C, H, G, P, V, E, S, L, R, K, Q, W, K, P, A, V, A, G). The alignment is shown in a grid format with a scrollbar on the right and a status bar at the bottom with buttons for "Module", "Inverted Colors", "Close", and "Help".

```
CF37_22_PE11 -WASGGR GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
SCX_1_PE113 --WYKPI APSII ASDFAN GOECHK V NAGADW HDVDVDGHFYPN TFGQPVTSLRRS PRPGDASNTEKEPTAFFDCHHSVENDEKWDFFAKCG
HS2_217_PE14 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
MA_5075_1_PE5 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
PT2B_102_PE12 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
PT10_112_PE1 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
HS10_108_PE17 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
BT2_59_PE4 -WASGCK GPSTI NSDLAS GAECR HDGADY HEDVDGFFFLN -FNQNL -LR LOR ---GQDPFFD HHSVSRBEQWVKPFAVAG
MD104_3_PE1 -WASGCK GRSHNSDLAC GA-CTQ HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
MM1_70_PE4 -WASGCK GPSTI NSDLAN GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
TN2_11_PE5 -WYNAK GPSTI SSDLSC GSECYR HECGADY HEDVDGHFYPN TFGHPVECLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
FR103_1_PE42 -WYSAK GPSTI SSDLSC GSECYR HECGADY HEDVDGHFYPN TFGHPVECLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
DR16_21_PE3 -WYATAK GPSTI SSDLSC GRECYR HECGADY HEDVDGHFYPN TFGHPVECLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
XT203_2_PE3 -WSSGCK GPSTI NSDLNR GAECR HDGADY HEDVDGHFYPN TFGHPVESLRKQ ---GQDPFFD HHSVSRBEQWVKPFAVAG
CID_422_1_PE21 -WCAIRC GPSTI NSDLAC GAECR HECGADY HEDVDGHFYPN TFGHPVECLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
DN2R_4_PE39 -WYVQAK GPSTI NADLSN AAESQK HDGADY HEDVDGTFYPN TFGHPVEKSLRNL ---KTAFFE HHSVSRBEQWVKPFAVAG
AG2L_10_PE11 -WAPLSAR GPSTI NADLSQ HEESEK HDGADY HEDVDGTFYPN TFGHPVEKSLRNL ---KTAFFE HHSVSRBEQWVKPFAVAG
CEIII_8_PE211 -WTLRPE GPSTI NADLAS ASECKK HAAGADW HDVDVDGHFYPN TFGHPVESLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
CBC4013_1_PE4 -WPLRAY GPSTI NADLAA ASECKK HAAGADW HDVDVDGHFYPN TFGHPVESLRKS ---GQDPFFD HHSVSRBEQWVKPFAVAG
XT5652_1_PE6 -ETSMK APSII SADFAF CEERL HAAGADY HDVDVDGHFYPN TFGPPVASTIRK A ---TKPLD HHSVSRBEQWVKPFAVAG
```

# L'évolution des gènes et l'évolution des espèces

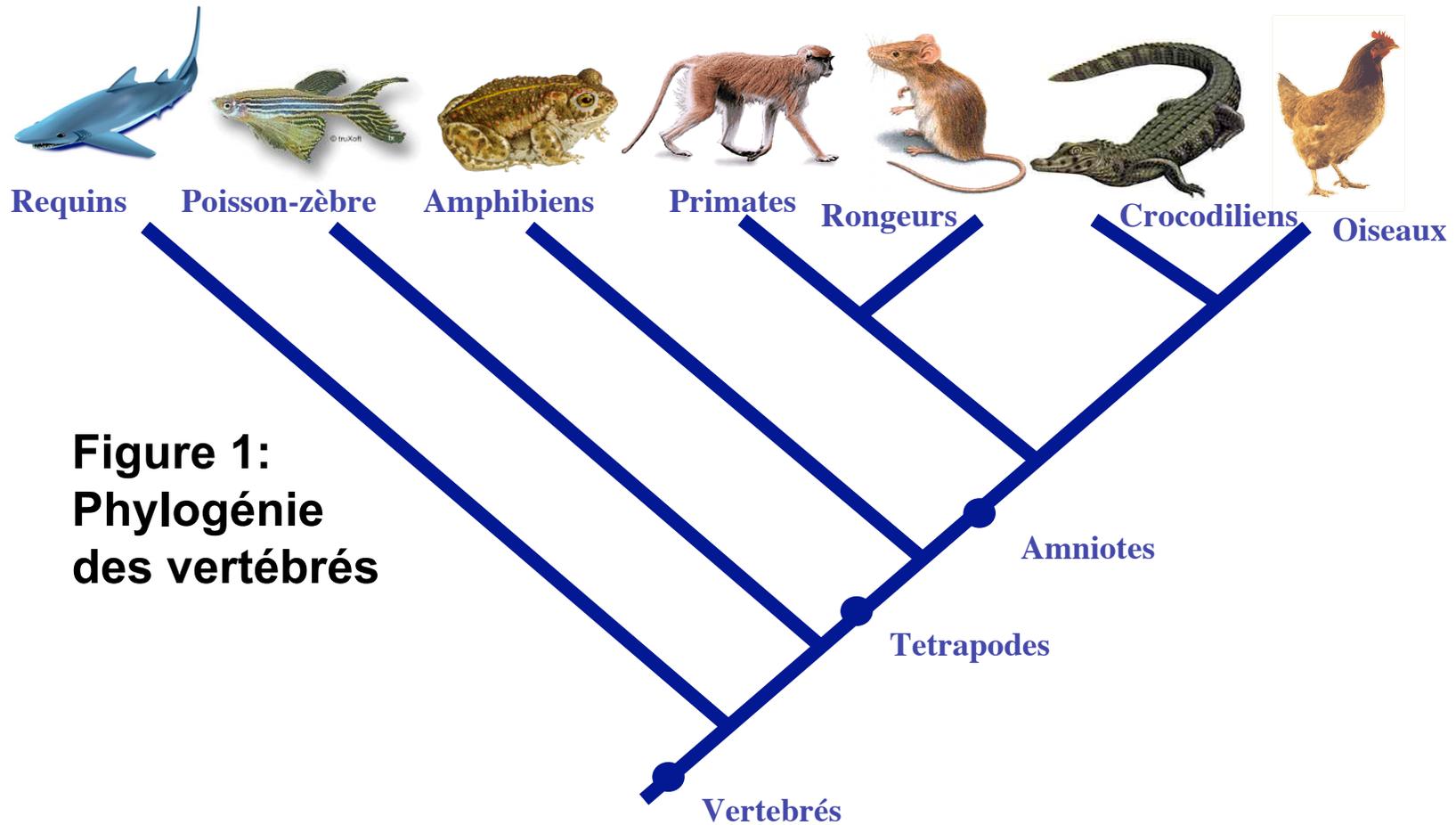


Figure 1:  
Phylogénie  
des vertébrés

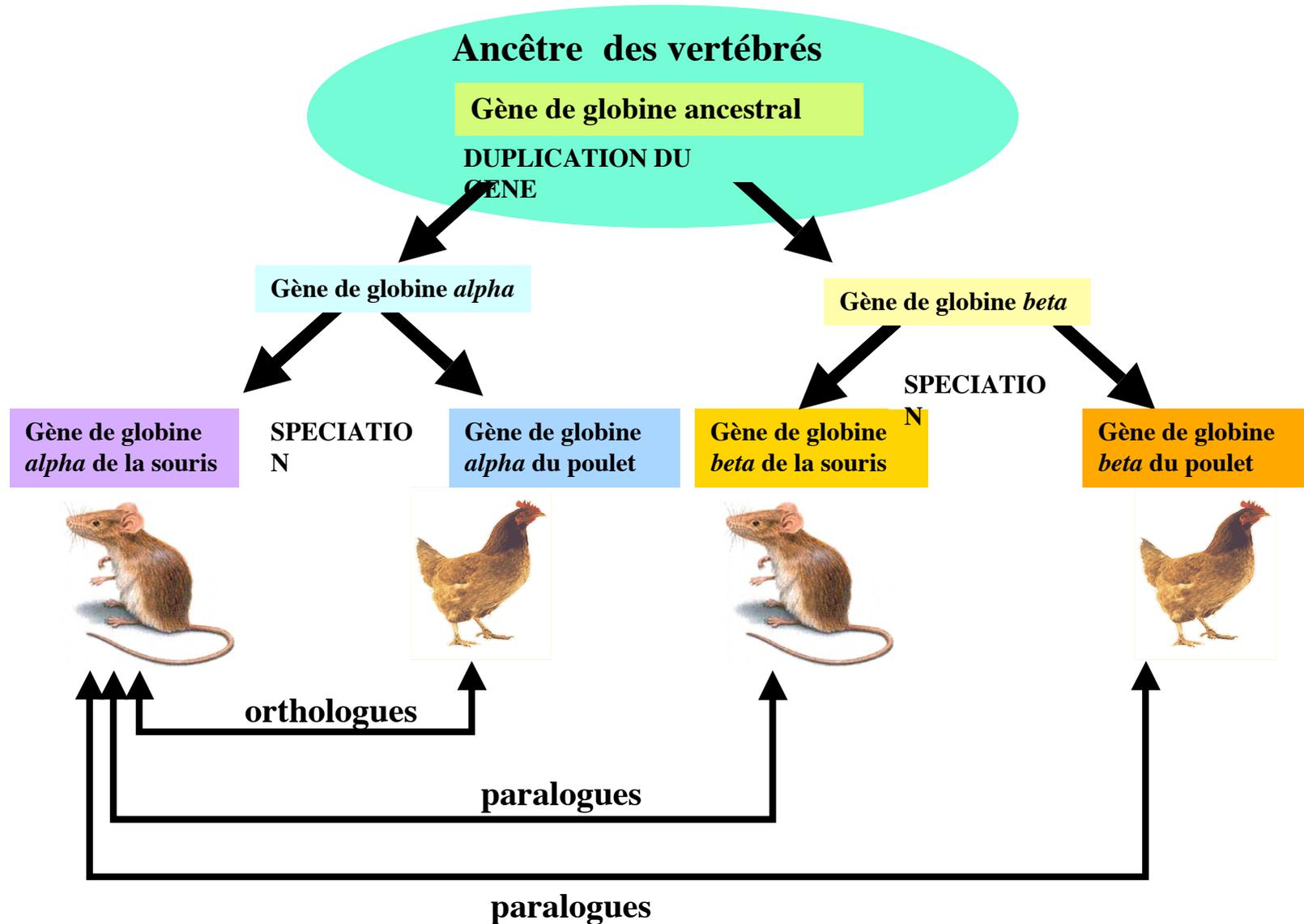
# L'évolution des gènes et l'évolution des espèces

## Relations d'homologie, d'orthologie et de paralogie

Deux gènes sont homologues lorsqu'ils descendent du même gène ancestral. Plus précisément, ils sont *paralogues* lorsqu'ils dérivent d'un événement de *duplication*, *orthologues* lorsqu'ils dérivent d'un événement de *spéciation*.

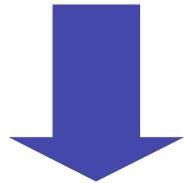
Identifier si des gènes sont paralogues ou orthologues est une étape essentielle dans l'étude de ces gènes.

# L'évolution des gènes et l'évolution des espèces



# **Thématiques PRABI/LBBE à forte demande en temps de calcul**

- **Phylogénie Moléculaire & Evolution**



- **Analyse et comparaison de séquence à grande échelle**
- **Algorithmique**
- **Statistiques**

# Les bases de données de familles de gènes homologues

**But** Comparer des séquences homologues entre diverses espèces.

**Contenu** Familles de séquences homologues

Chaque famille est associée :

à un alignement de séquence

à un arbre phylogénétique dans lequel les événements de duplication ou spéciation sont mis en évidence.

**Les principales bases sont**

HOGENOM génomes complets des bactéries, des archées et de certains eukaryotes,

HOMOLENS génomes complets animaux,

HOVERGEN gènes de vertébrés.

Ces bases de données sont actualisées régulièrement et mises à disposition de l'ensemble de la communauté scientifique

# Les bases de données de familles de gènes homologues

## **Structure**

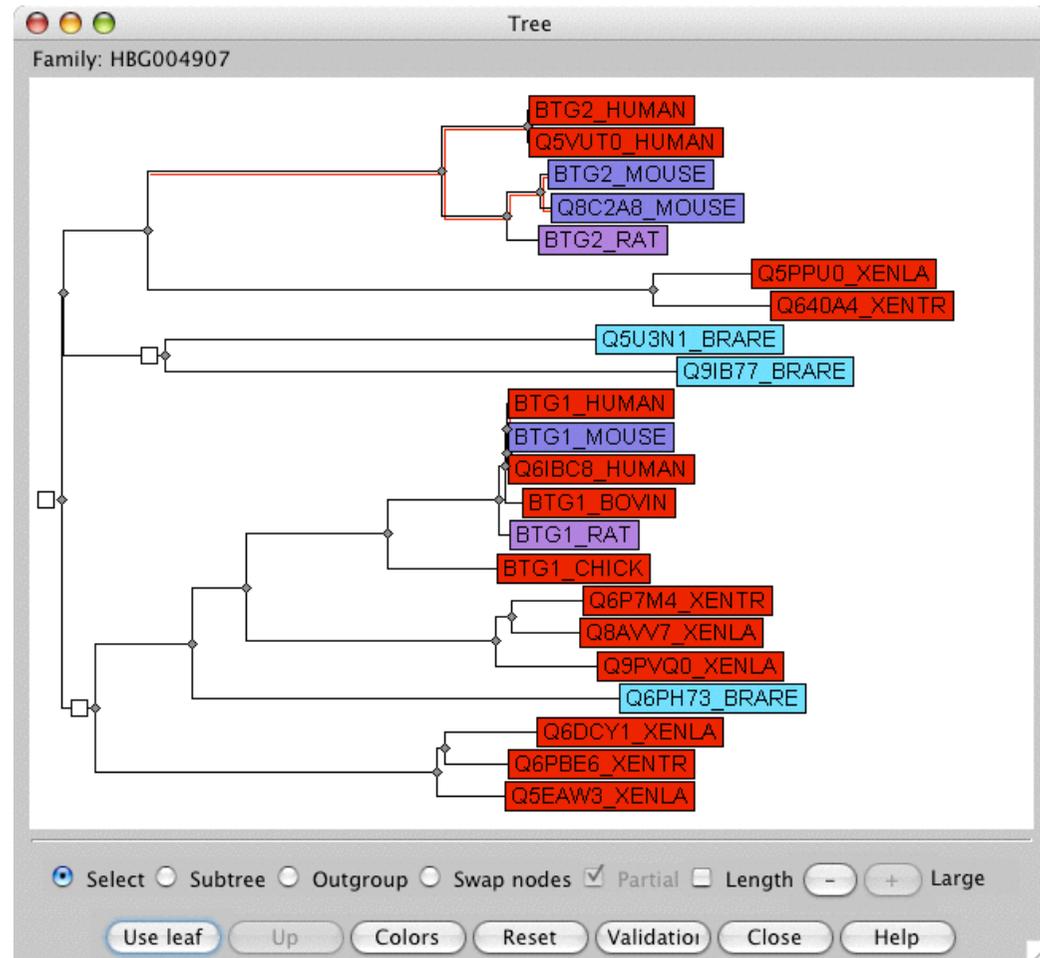
Les bases sont structurées sous le système ACNUC développé au LBBE. Chaque base comprend l'ensemble des séquences et des annotations des gènes et de leur protéines associées, l'information taxonomique, fonctionnelle, bibliographique, ainsi que les alignements et les arbres phylogénétiques de chaque famille.

## **Accès aux bases et interrogation**

Les bases sont accessibles par l'intermédiaire d'applications de type serveur client, via un site web, et via des sockets et des API en divers langages.

# Les bases de données de familles de gènes homologues

Un exemple d'interface:  
Le logiciel **FamFetch**  
(application client/serveur)



# Calcul intensif pour la construction et la mise à jour des bases de données de familles de gènes homologues

*Il faut noter que les calculs nécessaires à la création et à la mise à jour des bases sont effectués plusieurs fois par an, ceci pour les différentes releases des bases développées.*

## Recherche de similarité

Les séquences des gènes sont associées en familles sur la base de leur *similarité*.

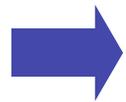


Comparer deux à deux tous les gènes d'un ensemble de génomes, soit pour la dernière release de HOGENOM environ *1 million de séquences* de 600 caractères en moyenne. La recherche de similarité entre les séquences est effectuée en utilisant un algorithme complexe (BLAST) sur la base d'une matrice de substitution (ou unitaire).

*La parallélisation de ces calculs au CCIN2P3 nous permet de mener à bien cette étape en une durée raisonnable.*

# Calcul intensif pour la construction et la mise à jour des bases de données de familles de gènes homologues

## Calcul d'alignements

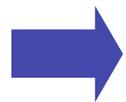


Les séquences de chaque famille sont alignées entre elle, afin de pouvoir en retirer l'information phylogénétique.

Le calcul d'un alignement est consommateur en temps de calcul, surtout lorsque la famille contient plusieurs dizaines de séquences, et il s'agit ici de calculer des dizaines de milliers d'alignements à chaque release de la base de données.

*Le CCIN2P3 permet de calculer en parallèle des centaines d'alignements.*

## Calcul d'arbres phylogénétiques



A partir de chaque alignement, un arbre phylogénétique est calculé. Le calcul d'un arbre peut être très rapide ou au contraire demander un temps de calcul massif selon la fiabilité désirée. Le bootstrap, par exemple, multiplie le temps de calcul d'un arbre par un facteur de 500 à 1000.

*Ici encore la parallélisation massive des calculs au CCIN2P3 est nécessaire*

# Autres développements

*La recherche de similarité de séquences, les calculs d'alignements et de phylogénies sont des outils bioinformatiques majeurs massivement utilisés en biologie.*

## Phylogénie

- Calculs de plusieurs millions d'arbres phylogénétiques pour détecter les transferts de gènes entre bactéries hyperthermophiles et les archées (Alexandra Calteau)
- Maximum de vraisemblance, développement, simulation (Bastien Bousseau)
- Chaînes de Markov, simulation d'évolution, 400 génomes à analyser (Anamaria Necsulea)
- Des milliers d'arbres de gènes pour reconstruire l'histoire du vivant (Sophie Abby)

## Analyse de génomes complets

- Chaînes de Markov : détection d'isochores (Christelle Melo de Lima)
- MegaBlast : recherche de retropseudogènes, (Adel Khelifi)
- Analyse statistique R (ade4, seqinr) (Leonora Palmeira)

# Gains en temps de calculs

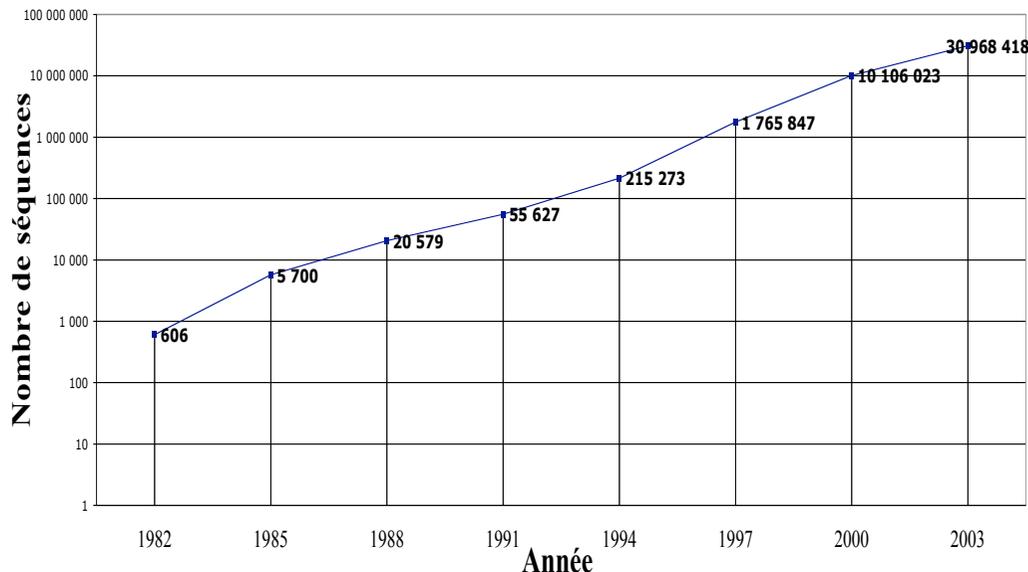
<b><i>Type de calcul</i></b>	<b><i>Durée estimée du calcul au LBBE</i></b>	<b><i>Durée du calcul au CCIN2P3</i></b>
Similarité de séquences pour construire la banque Hogenom	1 an	3 jours
Recherche de « pseudo gènes » dans le génome humain	plus de un an	moins d'1 mois
Prédiction d' « isochores » (zones plus ou moins riches en gènes) dans le génome humain	3 ans	10 jours
Calculs de plusieurs millions d'arbres phylogénétiques pour détecter les transferts de gènes entre espèces.	plusieurs années	quelques mois

# Perspectives

**Le temps de calcul était devenu un facteur limitant pour une grande partie des activités du LBBE.**

L'accès aux ressources de calcul du CCIN2P3 nous permet de

- faire face à l'augmentation exponentielle du volume de données biologiques et donc
  - de poursuivre la mise à jour des bases de données que nous mettons à disposition de la communauté
  - de maintenir nos capacités d'analyses de séquences.
- envisager des analyses qui n'étaient pas réalisables avec les moyens propres au laboratoire. Ainsi, de nouvelles perspectives de recherche sont désormais ouvertes.



**Evolution du nombre de séquences dans la banque GenBank**

# Participants

**Pascal Calvat**      **User Support IN2P3**

**Laurent Duret**  
**Vicent Daubin**  
**Christian Gautier**  
**Manolo Gouy**  
**Simon Penel**  
**Guy Perrière**

**Sophie Abby**  
**Bastien Boussau**  
**Alexandra Calteau**  
**Adel Khelifi**  
**Christelle Melo de Lima**  
**Anamaria Necuslea**  
**Leonor Palmeira**