Recherche de motifs dans les arbres phylogénétiques

Jean-François Dufayard
Sous la direction de Laurent Duret, Guy Perrière et François Rechenmann

INRIA RHÔNE-ALPES

UMR 5558
Biométrie, biologie évolutive.

Mini CV

-Licence d'informatique

-Réconciliation d'arbres phylogénétiques ⟺ -Maîtrise d'informatique

-Recherche de motifs dans les arbres phylogénétiques ⟺ -DEA d'informatique

-Gestion de grandes familles de gènes homologues ⟺ -Thèse d'informatique (en cours)

Contexte

Bases de données de familles de gènes homologues

L'analyse comparative de séquences:

- Mécanismes d'évolution des génomes
- Prédiction de la fonction des gènes

➢ Analyse phylogénétique.

Bases de données de familles de gènes homologues: HOVERGEN (vertébrés) et HOBACGEN (procaryotes),

- Gènes classés en familles (BLASTP).
- Alignement multiple et arbre phylogénétique calculés pour chaque famille.
- Données taxonomiques issues du NCBI.
- Séquences protéiques issues de SwissProt et TrEMBL.
- Séquences nucléiques issues de EMBL.

Problématique

Problématique (1/2)

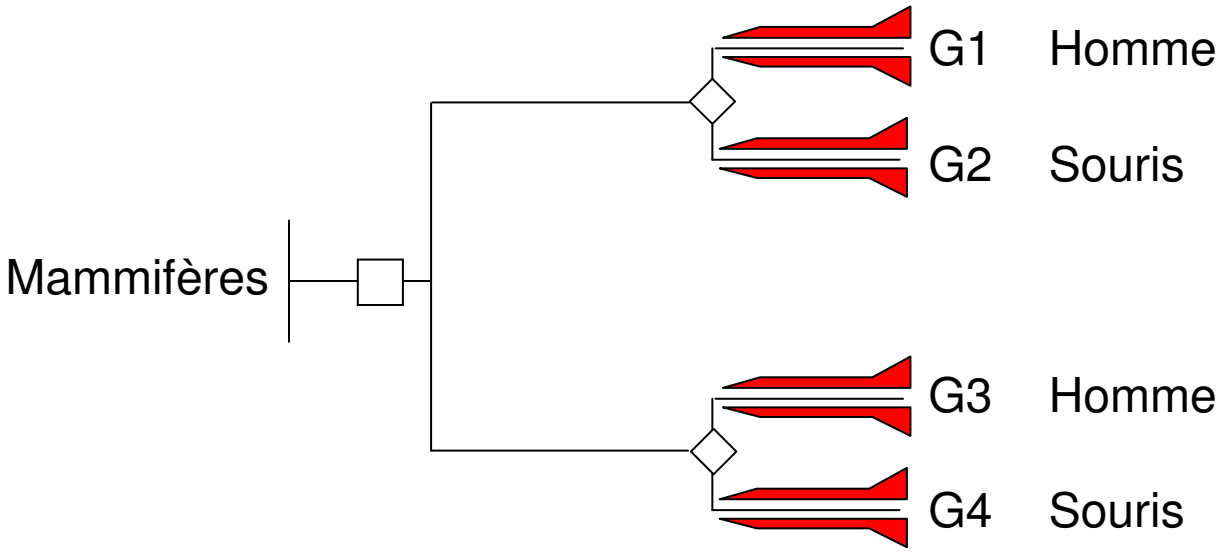
- Base de données d'arbres phylogénétiques:
  HOVERGEN: 10 000 arbres issus de vertébrés.
  HOBACGEN: 19 000 arbres issus de procaryotes.

Comment rechercher de l'information sur des critères phylogénétiques ?

Problématique (2/2)

Exemple:

- « *Je cherche 4 gènes homologues, sous forme de 2 paires de gènes hommes / souris. Les 2 paires de gènes doivent être des paires de gènes strictement orthologues. Les 2 paires de gènes doivent être paralogues entre-elles. Je souhaite que la duplication de gènes à l'origine de la paralogie soit postérieure à la première divergence des mammifères.* »



◇ Spéciation Pas de duplications □ Duplication

Point de vue informatique

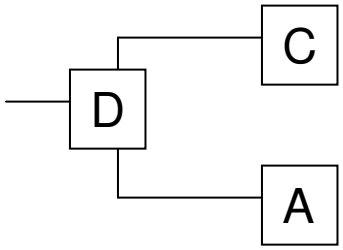
Recherche de motifs dans les arbres (1/2)

Recherche de motifs

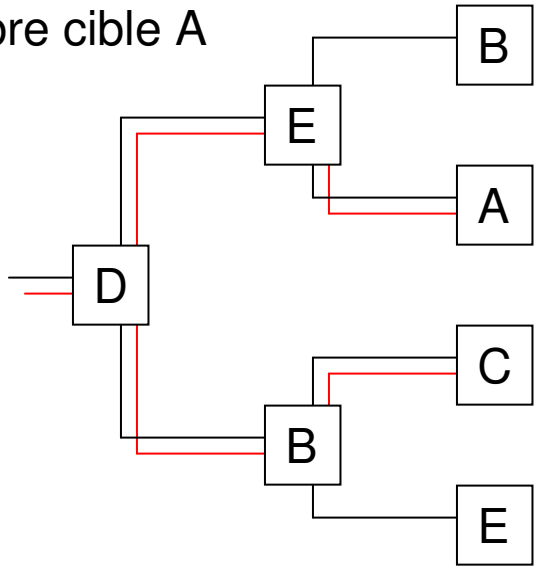
- Ordonnés
- Non ordonnés

La recherche de motifs non ordonnés est un problème NP-complet

Motif M



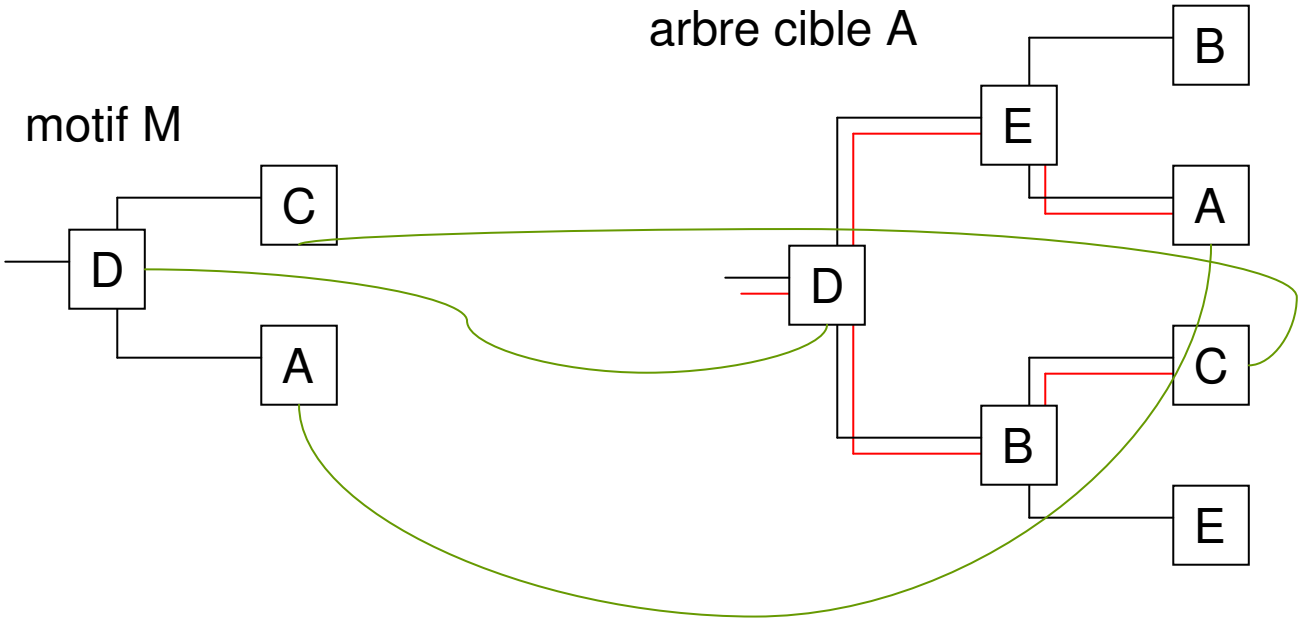
Arbre cible A



Recherche de motifs dans les arbres (2/2)

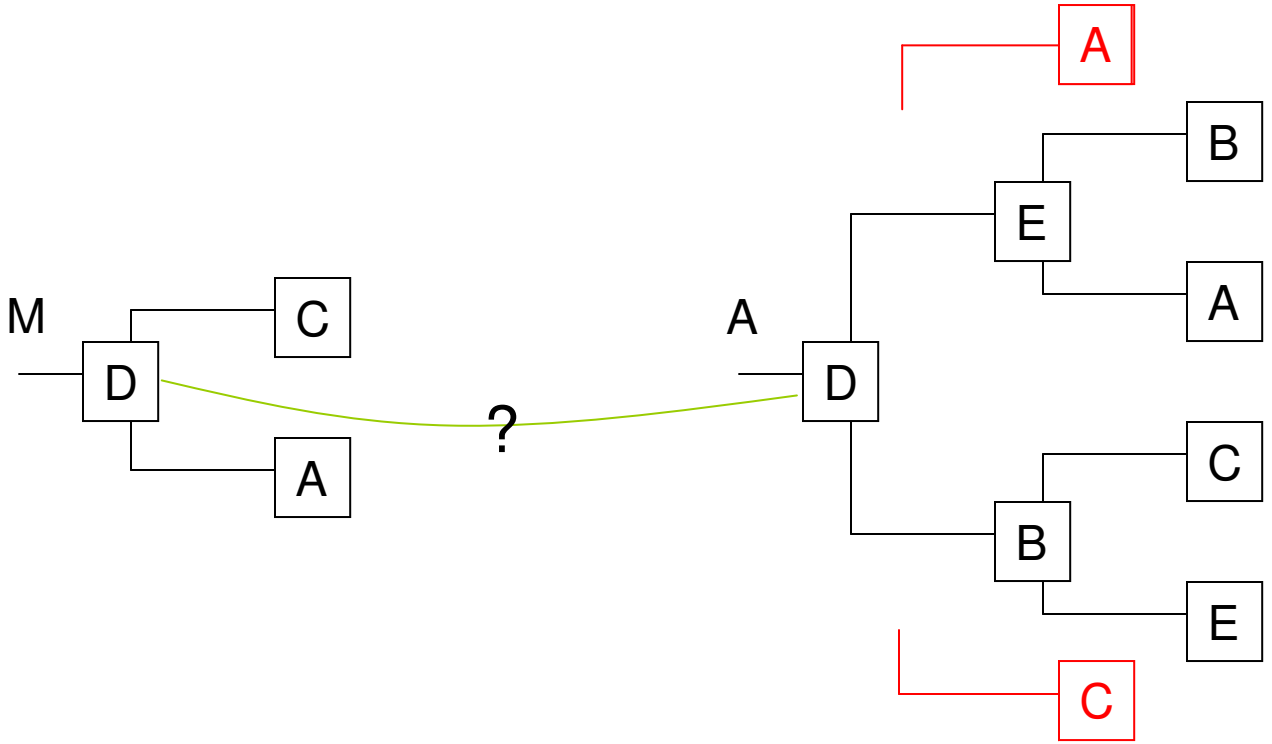
A = (V, E, racine(A))
M = (W, F, racine(M))
« M est un motif non ordonné de A » ⇔ ∃ f une fonction injective telle que :

i) $f : w \in W \rightarrow v \in V$
ii) $f(u) = f(v) \Leftrightarrow u = v$
iii) étiquette(u) = étiquette(f(u))
iv) « u est un ancêtre de v » ⇔ « f(u) est un ancêtre de f(v) »



Méthode de résolution dite « naïve »

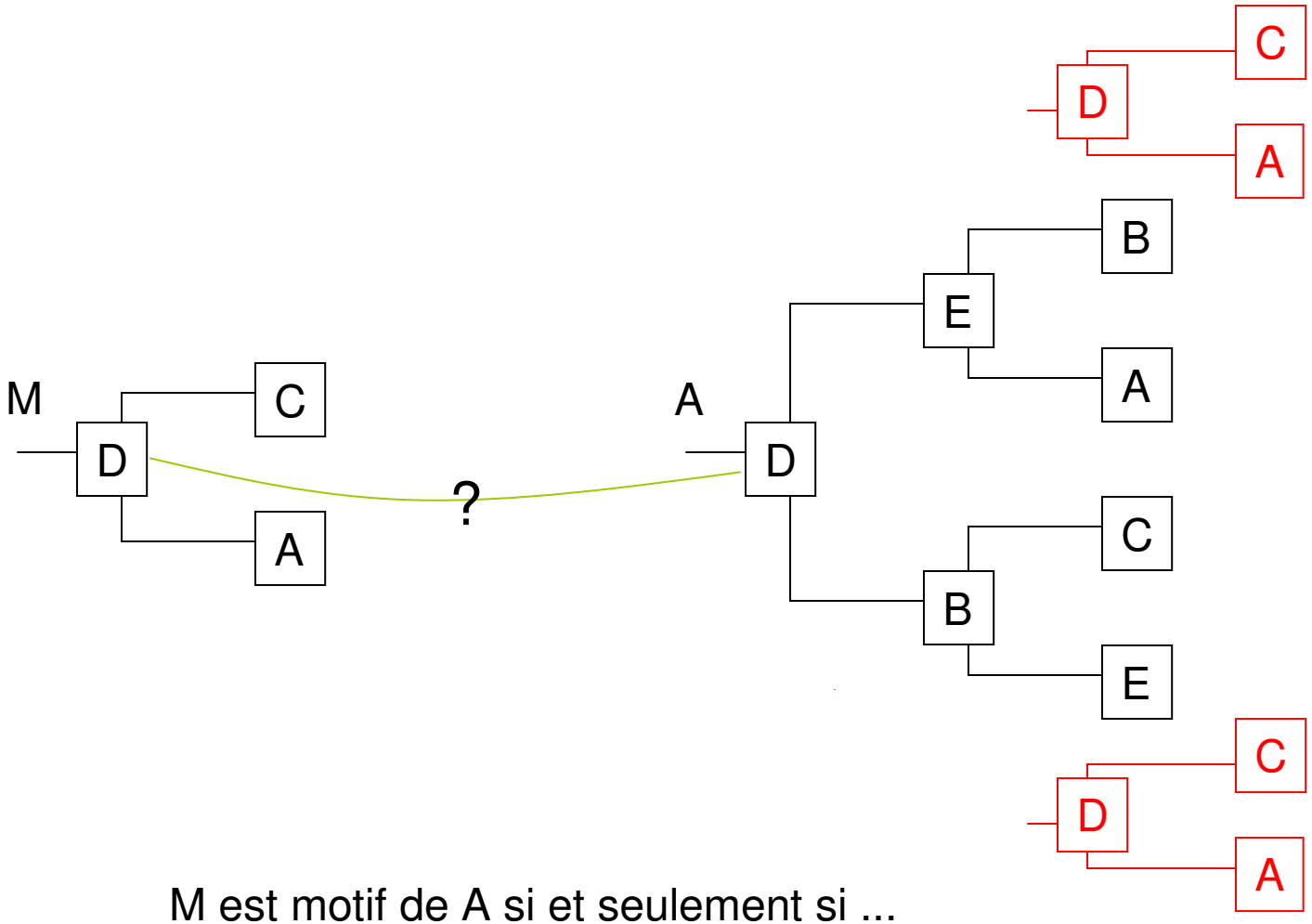
Développer l'ensemble des possibles



M est motif de A si et seulement si ...

Méthode de résolution dite « naïve »

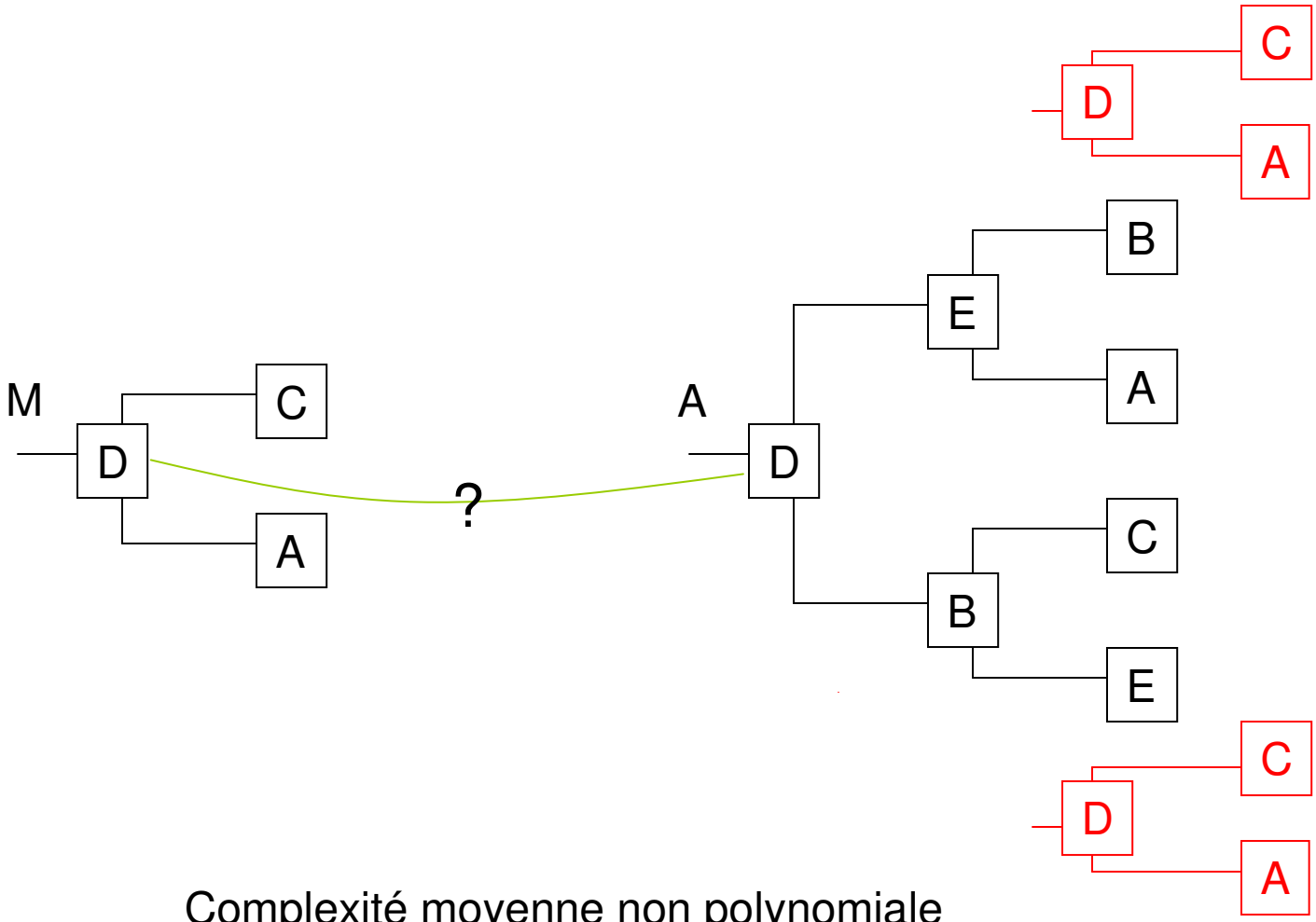
Développer l'ensemble des possibles



M est motif de A si et seulement si ...

Méthode de résolution dite « naïve »

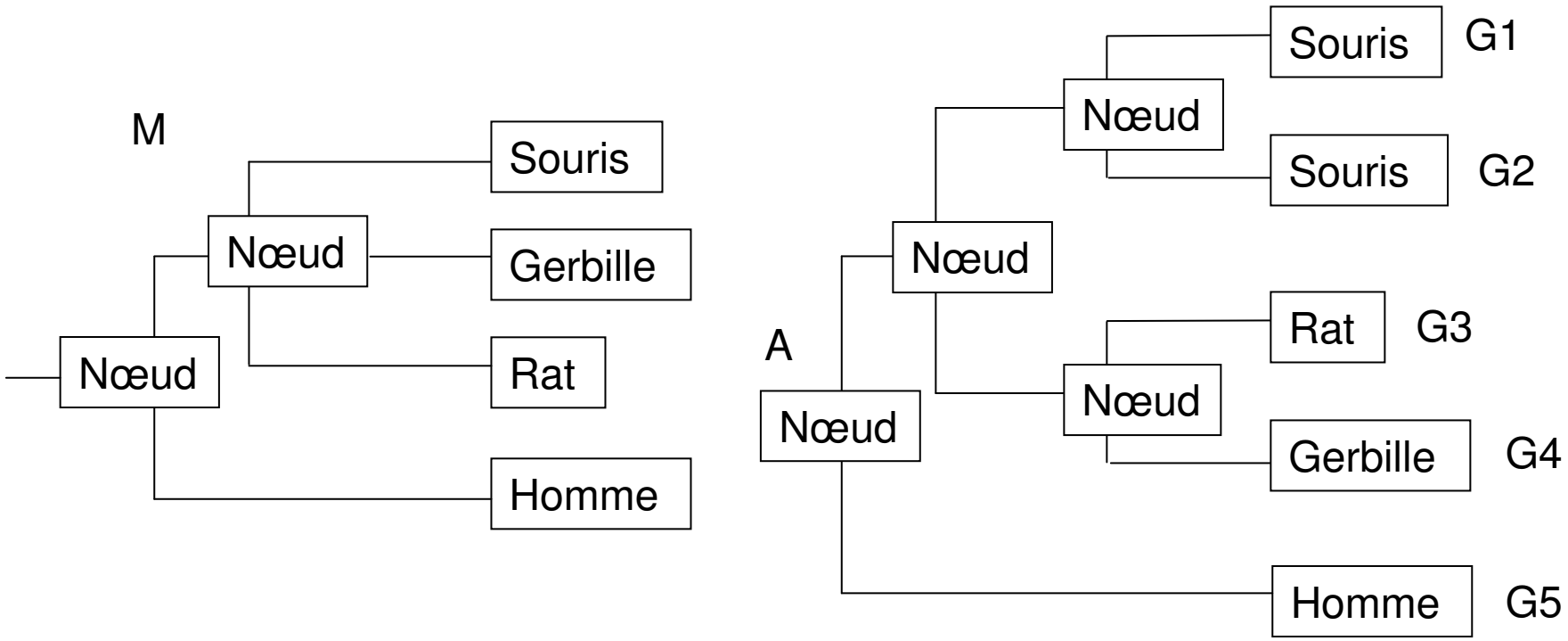
Développer l'ensemble des possibles



Complexité moyenne non polynomiale

Point de vue applicatif

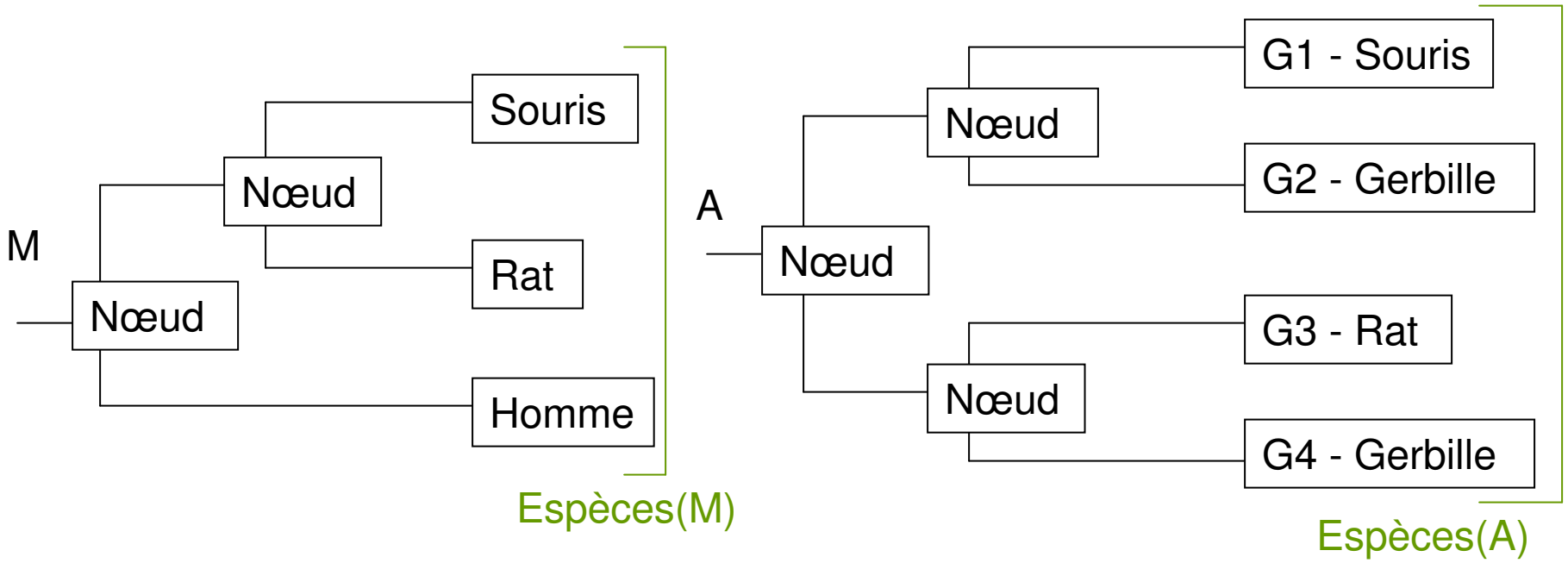
Transposition aux arbres phylogénétiques (1/2)



- Les feuilles sont étiquetées par un nom d'espèce.
- Les nœuds sont étiquetés par une étiquette unique « Nœud ».
- Les arbres doivent être racinés.

- Les arbres phylogénétiques sont un cas particulier: L'ensemble des étiquettes aux feuilles est disjoint de l'ensemble des étiquettes aux nœuds.

Transposition aux arbres phylogénétiques (2/2)

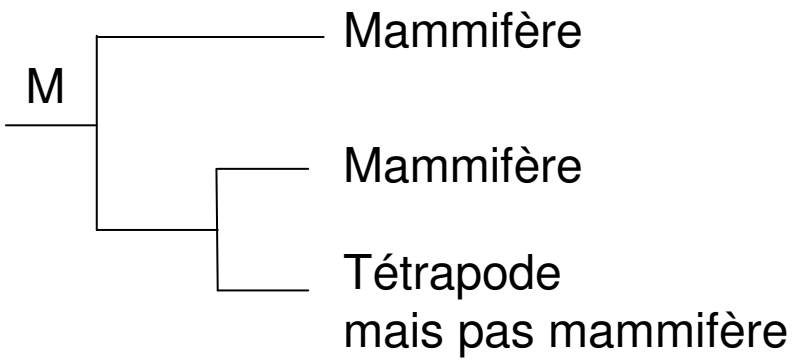


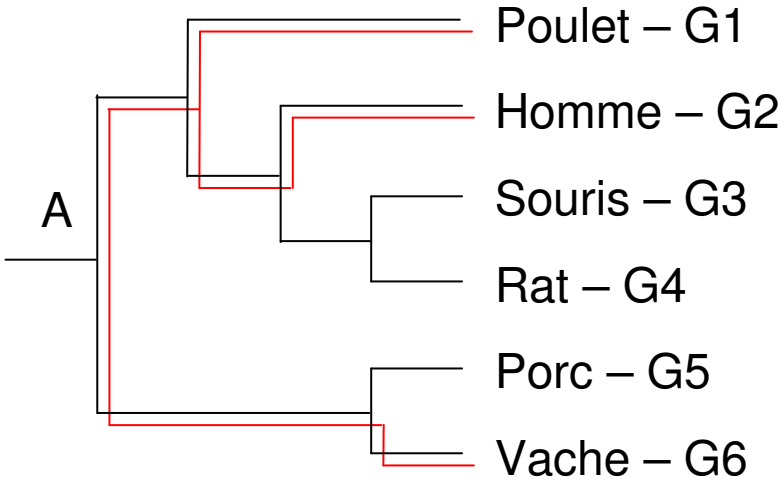
Soit espèce(X) la fonction rendant l ’ensemble des espèces aux feuilles de l ’arbre X :
« M est motif de A » $\Rightarrow$ espèces(M) $\subseteq$ espèces(A)

Améliorations

Introduction de différents niveaux taxonomiques

Les feuilles du motif peuvent être étiquetées avec des taxons de haut niveau.

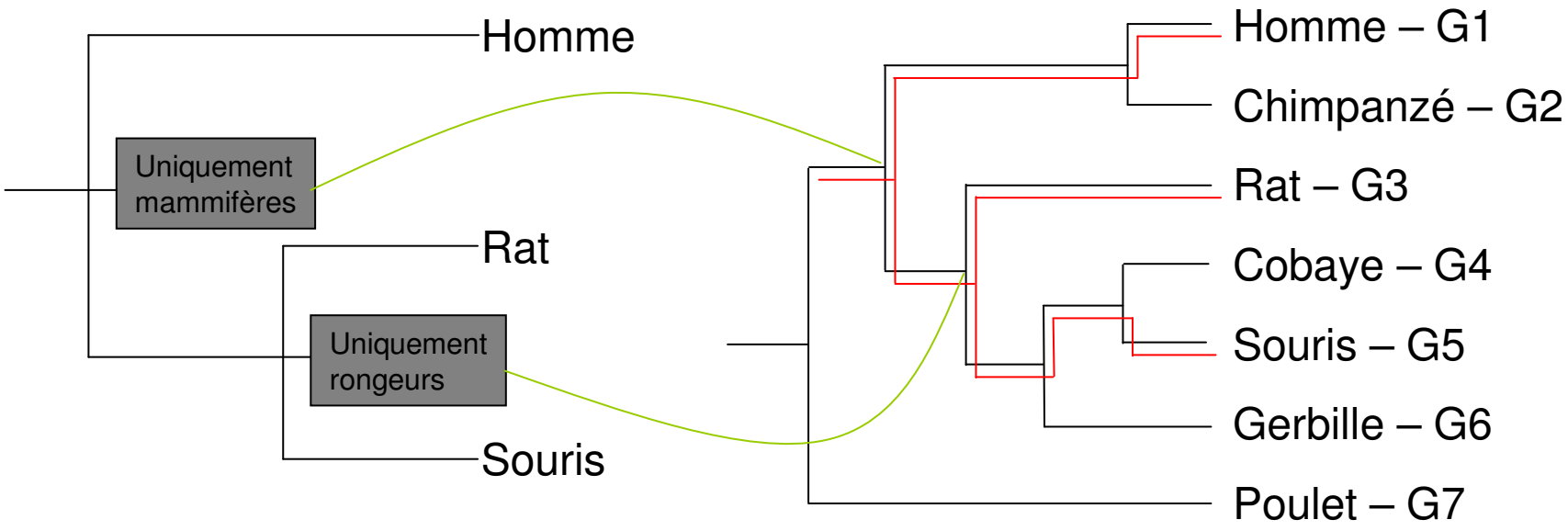




Objectif: retrouver des paralogies antérieures à l'apparition des mammifères.

Contraintes sur les nœuds du motif

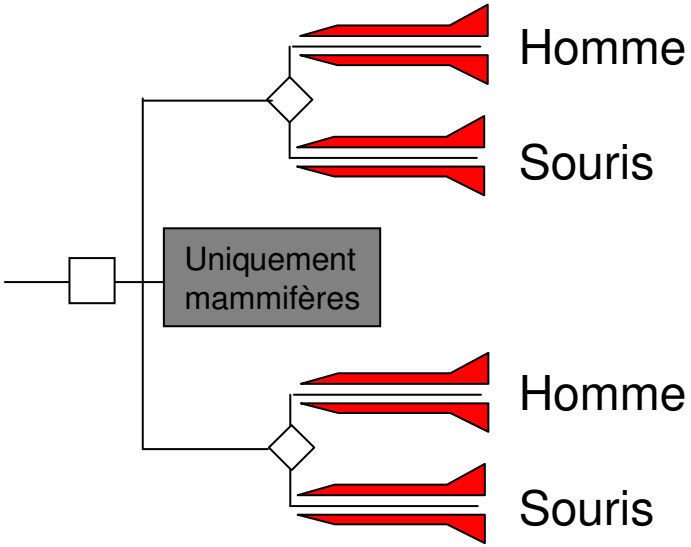
Contrôler quels taxons sont présents sous certains nœuds du motif.

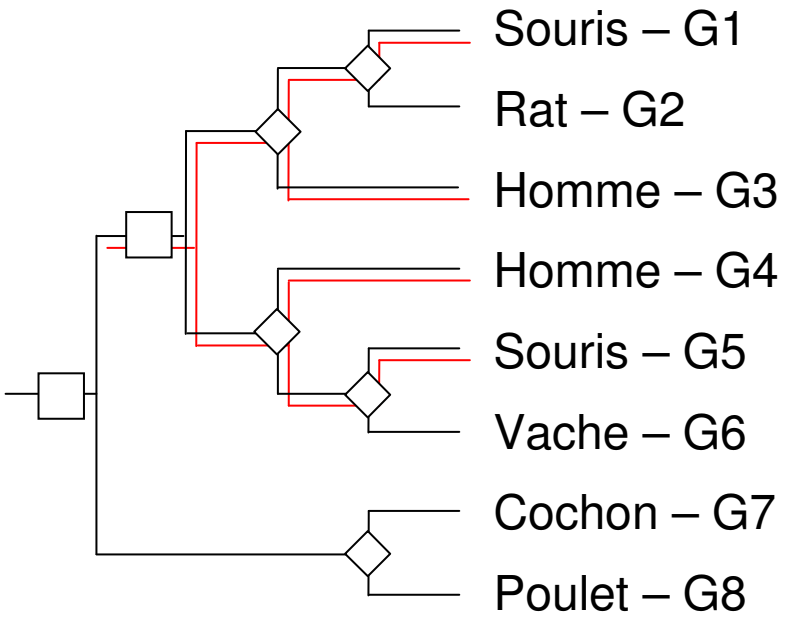


Objectif: Eviter de sélectionner certaines paralogies.

Nœuds de duplications et de spéciations

La base HOVERGEN a été entièrement annotée par des nœuds de duplications.





Pas de duplications

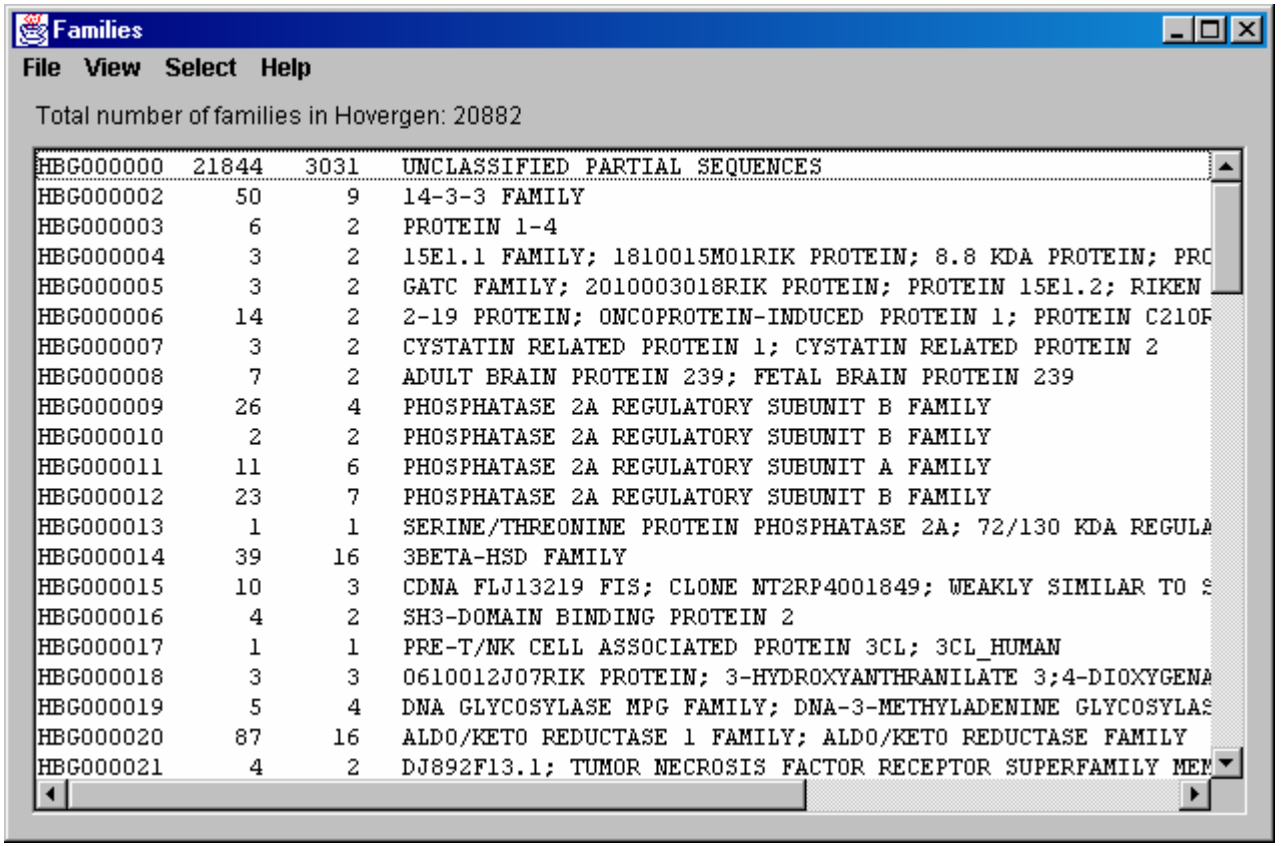
Spéciation

Duplication

Implémentation

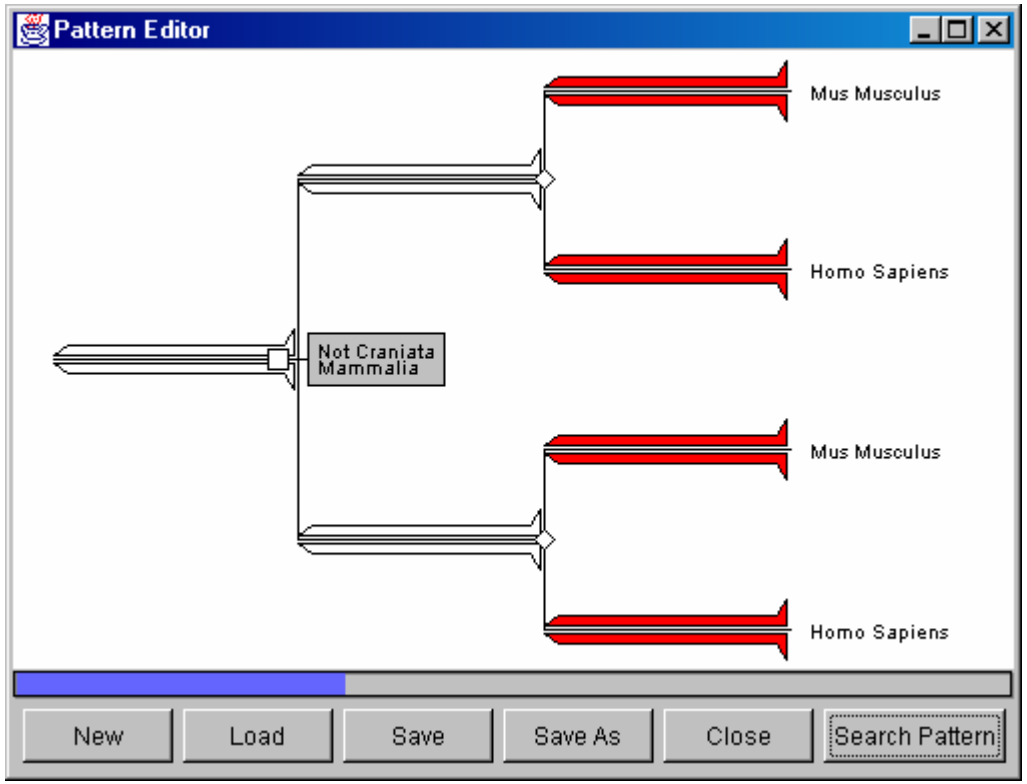
Intégration dans FamFetch (1/4)

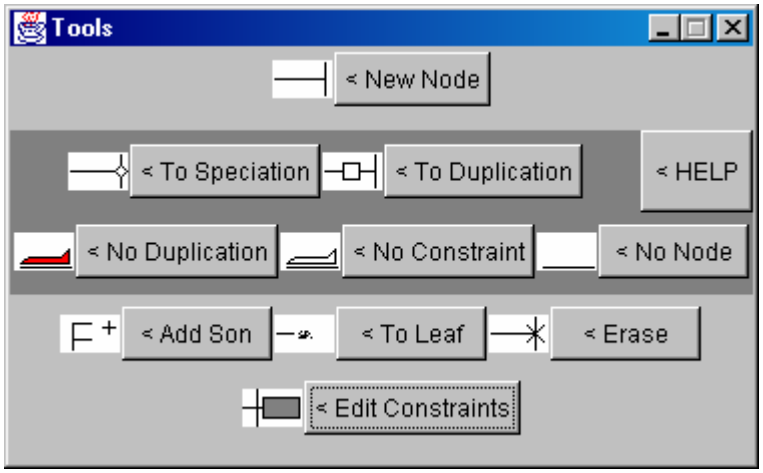
FamFetch: Interface d'accès aux bases de données développées au Pôle Bio-Informatique Lyonnais. (HOBACGEN, HOVERGEN, ACNUC…)



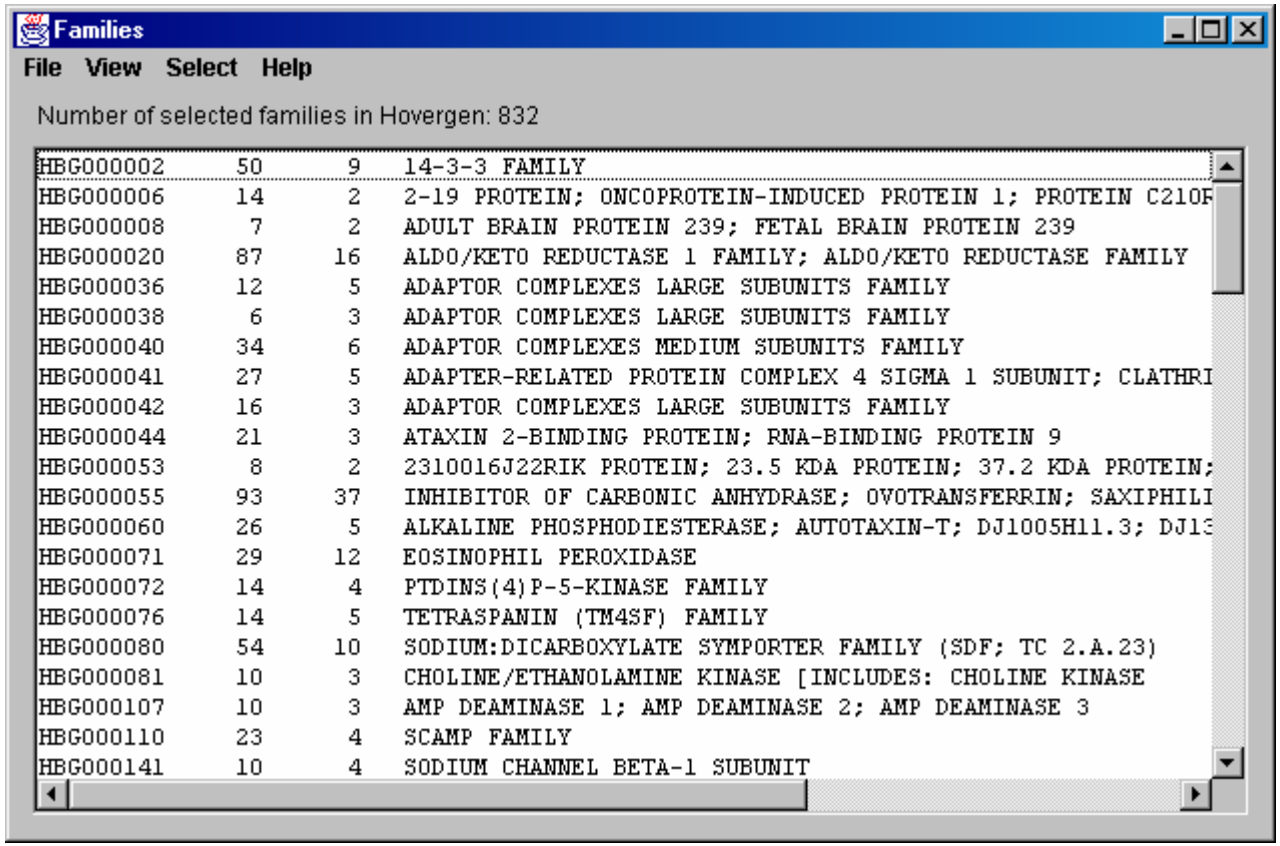
Racinement des arbres par la méthode du point central, ou par réconciliation d'arbres (HOVERGEN).

Intégration dans FamFetch (2/4)

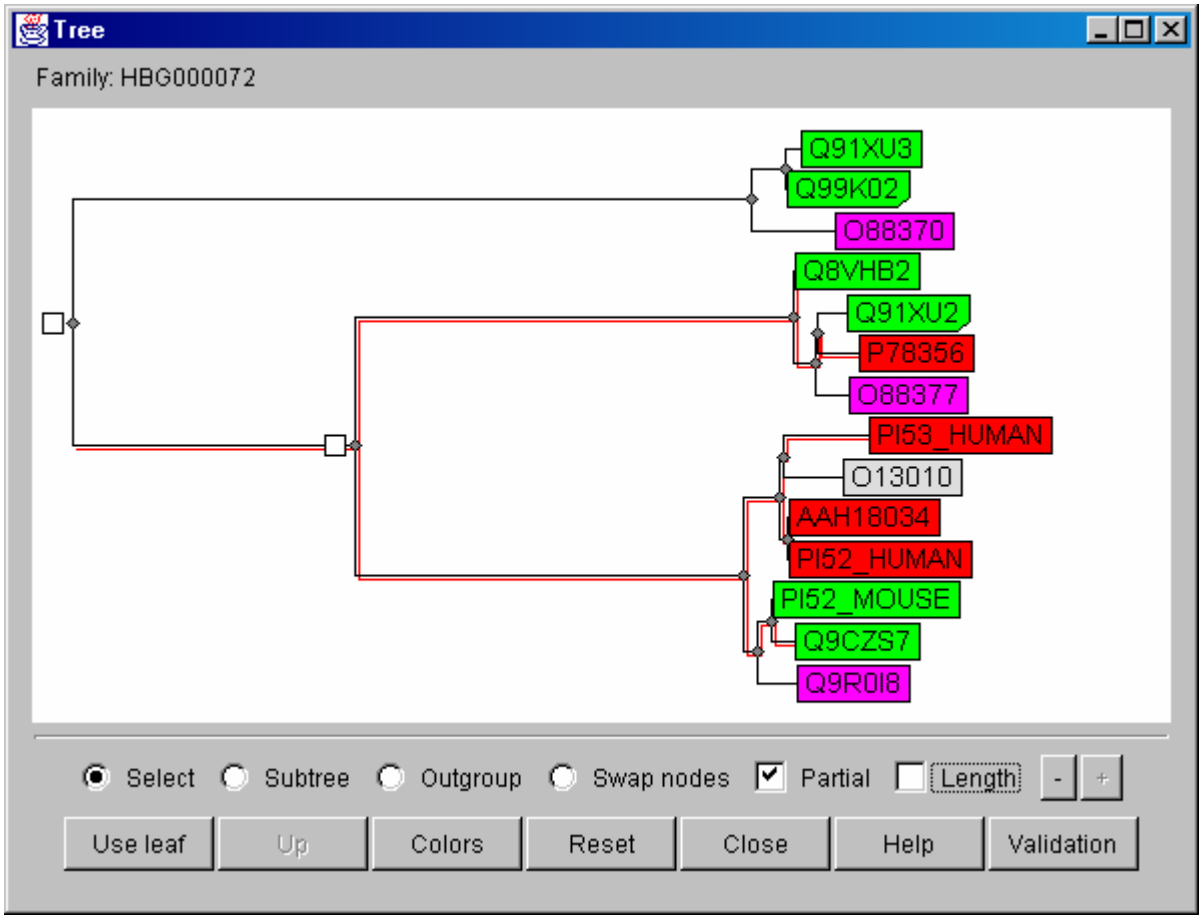




Intégration dans FamFetch (3/4)



Intégration dans FamFetch (4/4)



Pour conclure …

Conclusion / Discussion

• Système permettant de résoudre des requêtes sur une base d'arbres, prenant en compte des critères phylogénétiques.

• Implémentation compatible avec une utilisation interactive.

• Système concurrencé par la recherche à la main.

• Sensible à la qualité des arbres.

• Nécessite des arbres enracinés.