

Recherche de motifs dans les arbres phylogénétiques



Jean-François Dufayard

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



Mini CV



-Licence d'informatique

-Réconciliation d'arbres phylogénétiques \longleftrightarrow -Maîtrise d'informatique

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

-Gestion de grandes familles de gènes homologues \longleftrightarrow -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)



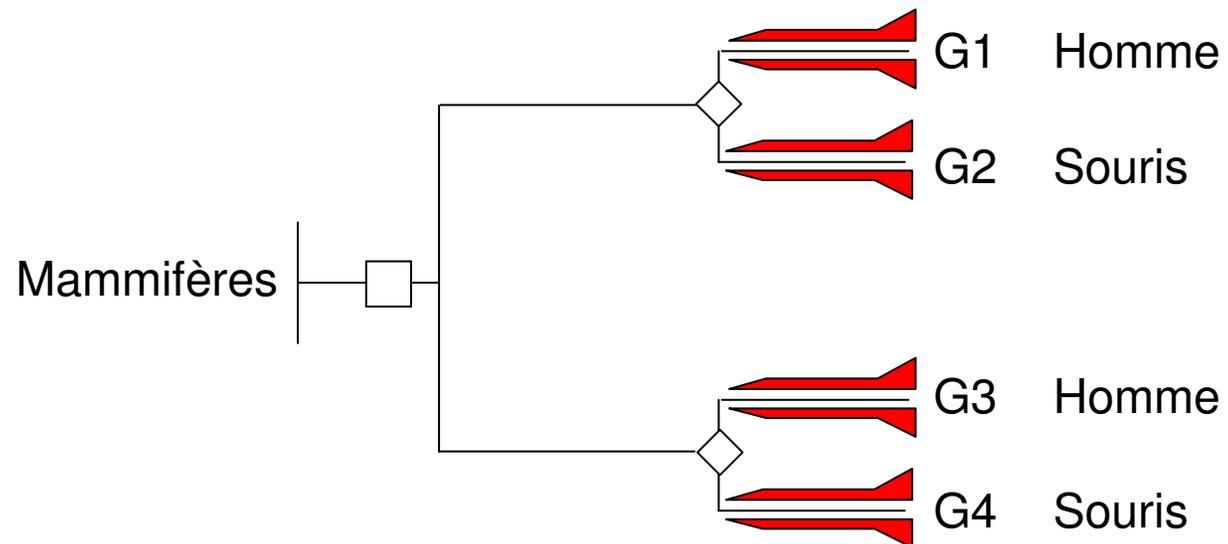
- Base de données d'arbres phylogénétiques:
HOVERGEN: 10 000 arbres issus de vertébrés.
HOBACGEN: 19 000 arbres issus de procaryotes.

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

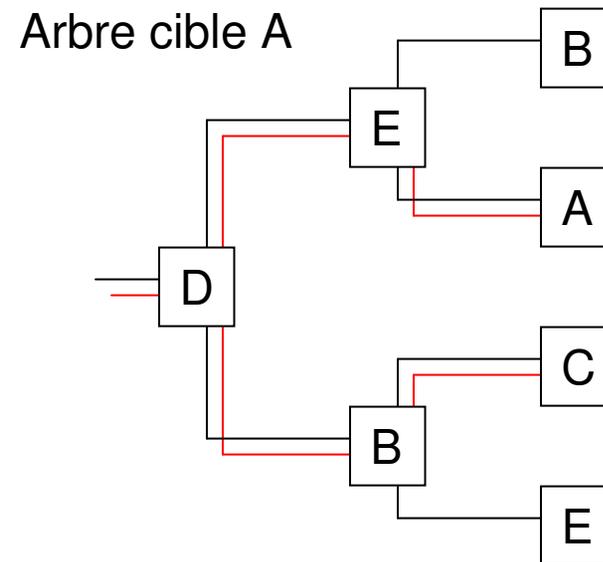
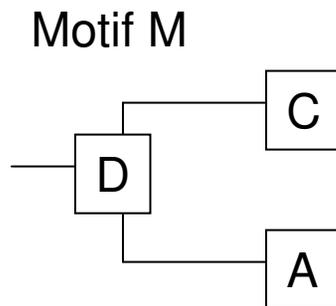


Recherche de motifs dans les arbres (1/2)

Recherche de motifs

- Ordonnés
- Non ordonnés

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



Recherche de motifs dans les arbres (2/2)

$A = (V, E, \text{racine}(A))$

$M = (W, F, \text{racine}(M))$

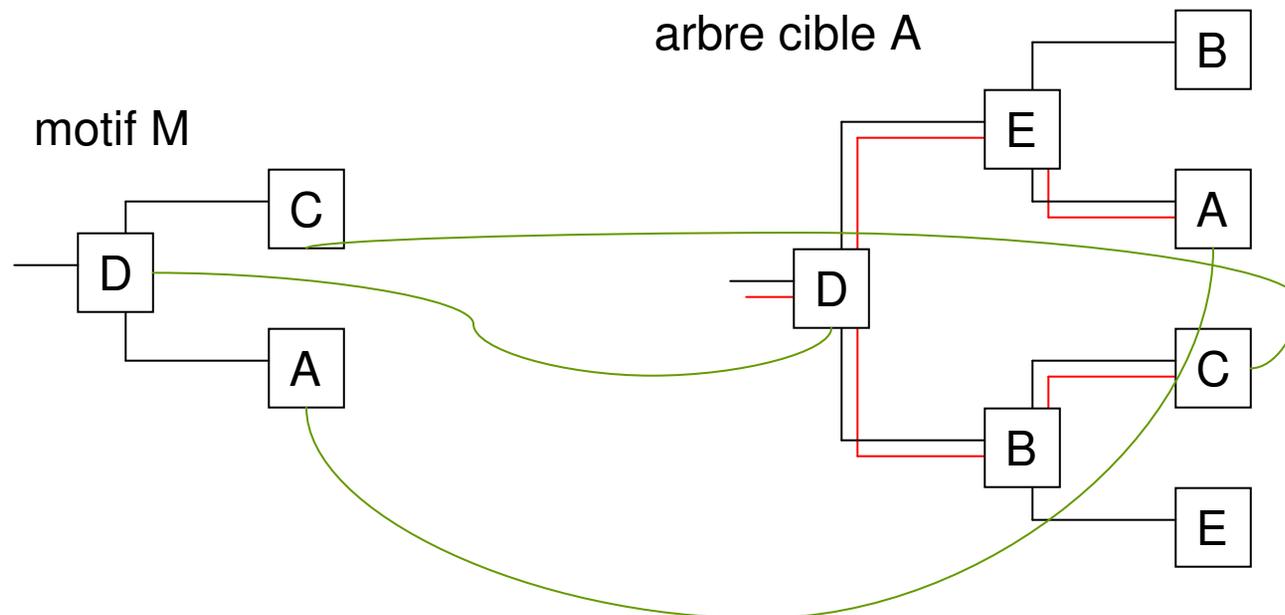
« M est un motif non ordonné de A » $\Leftrightarrow \exists f$ une fonction injective telle que :

i) $f : w \in W \rightarrow v \in V$

ii) $f(u) = f(v) \Leftrightarrow u = v$

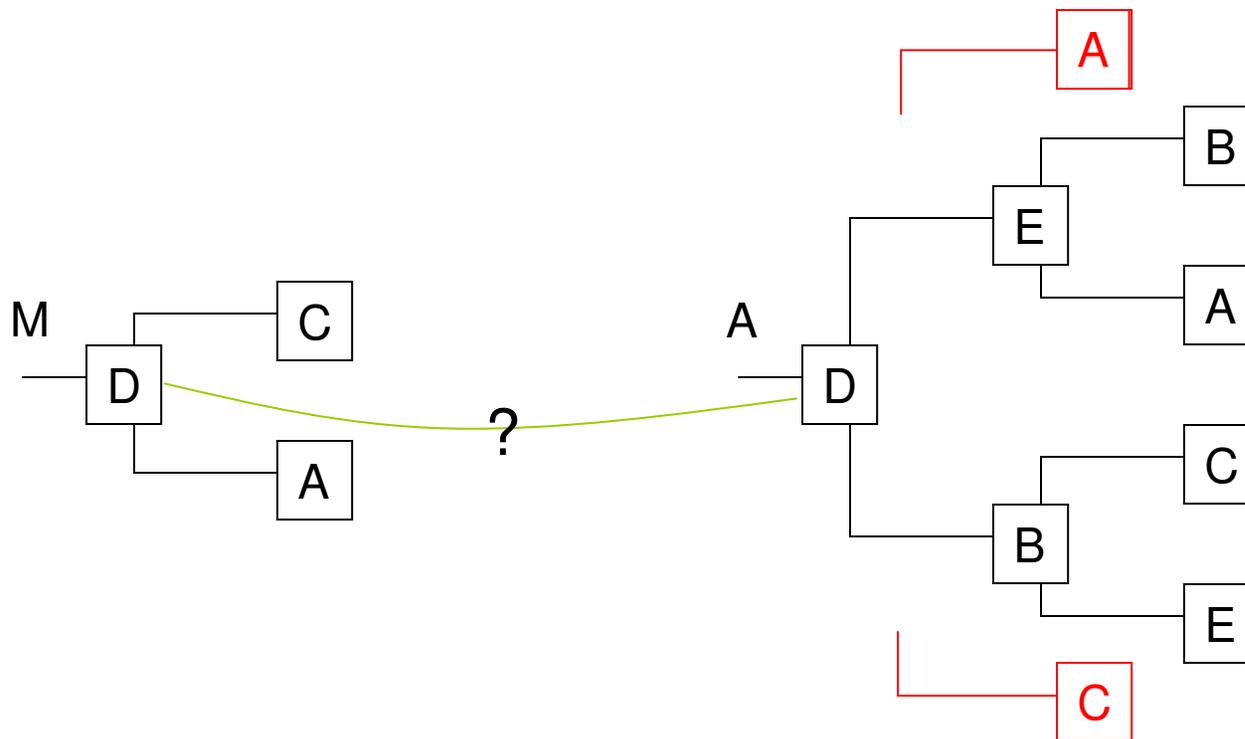
iii) $\text{étiquette}(u) = \text{étiquette}(f(u))$

iv) « u est un ancêtre de v » \Leftrightarrow « f(u) est un ancêtre de f(v) »



Méthode de résolution dite « naïve »

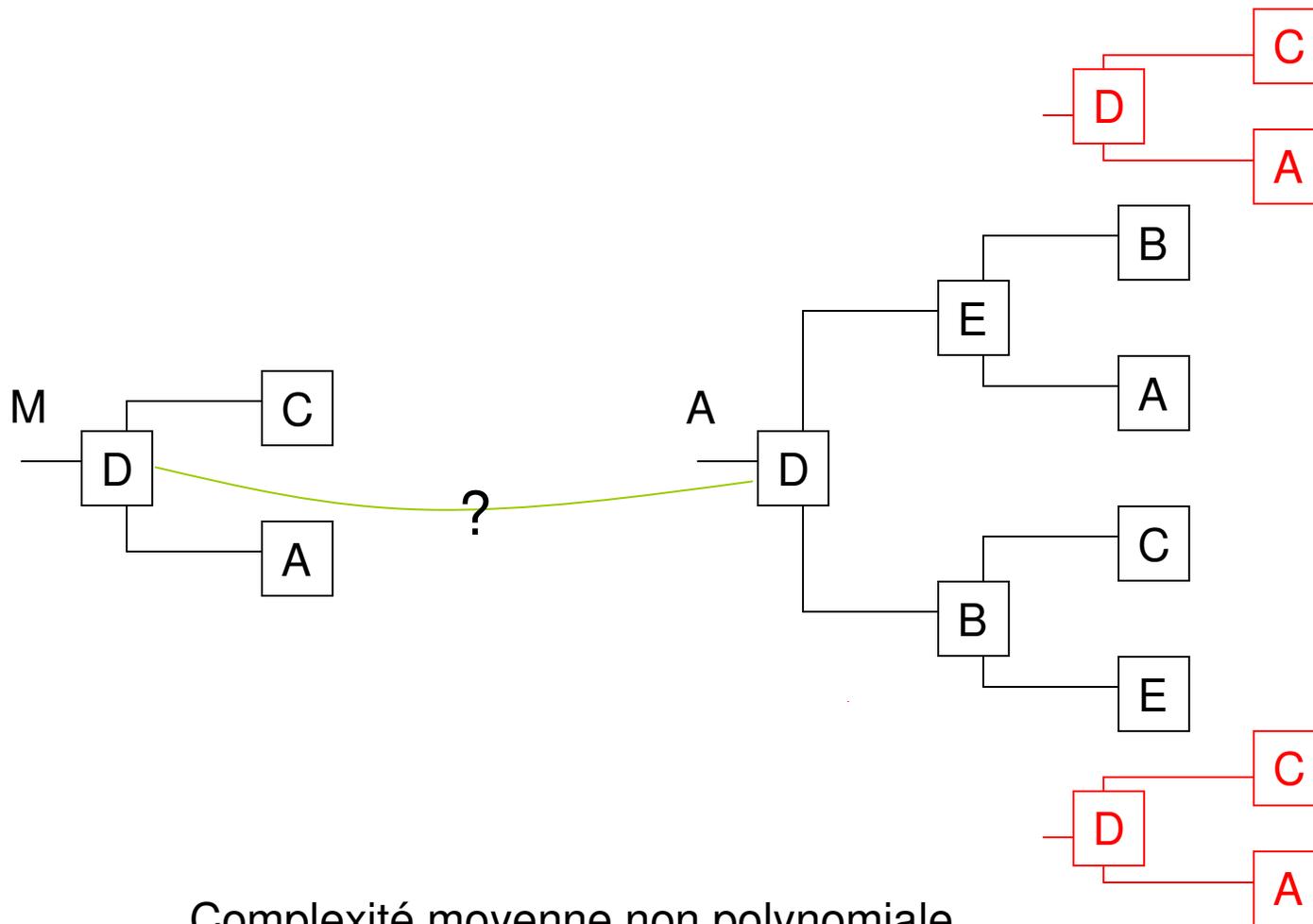
Développer l'ensemble des possibles



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

Méthode de résolution dite « naïve »

Développer l'ensemble des possibles

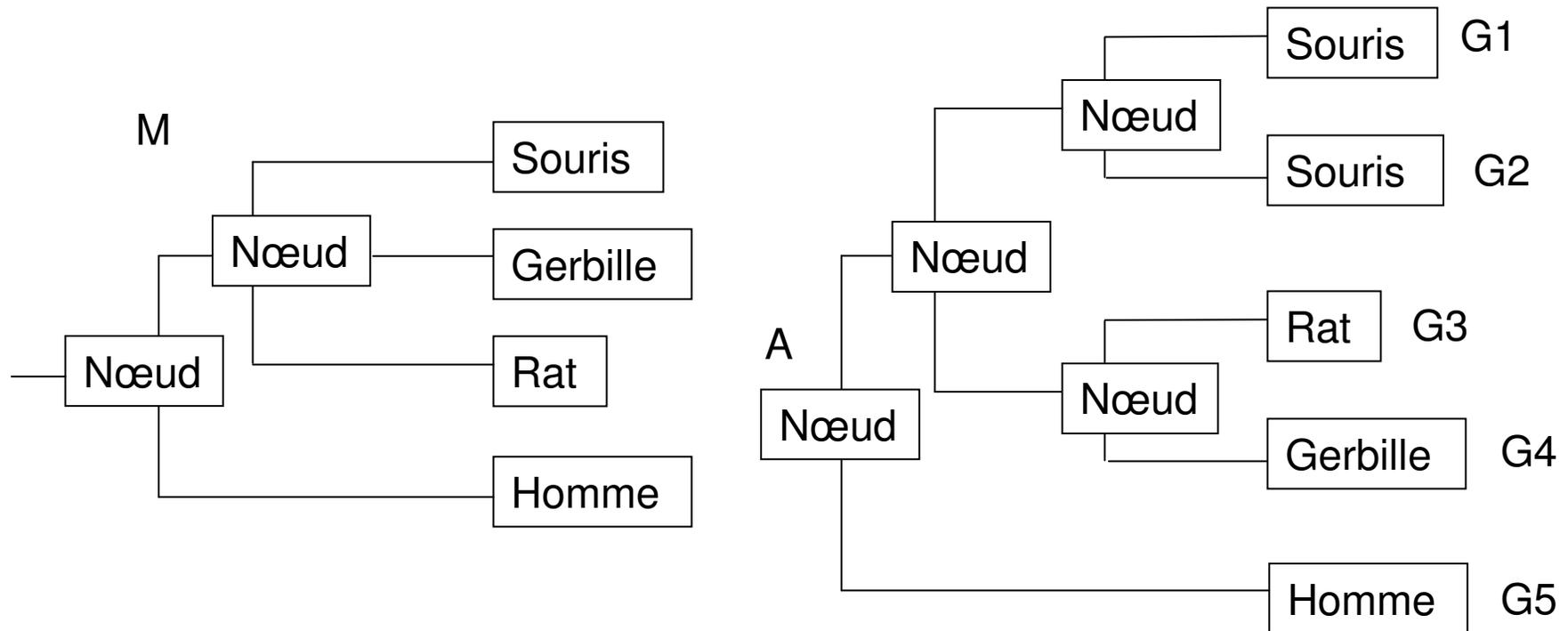


Complexité moyenne non polynomiale

Point de vue applicatif



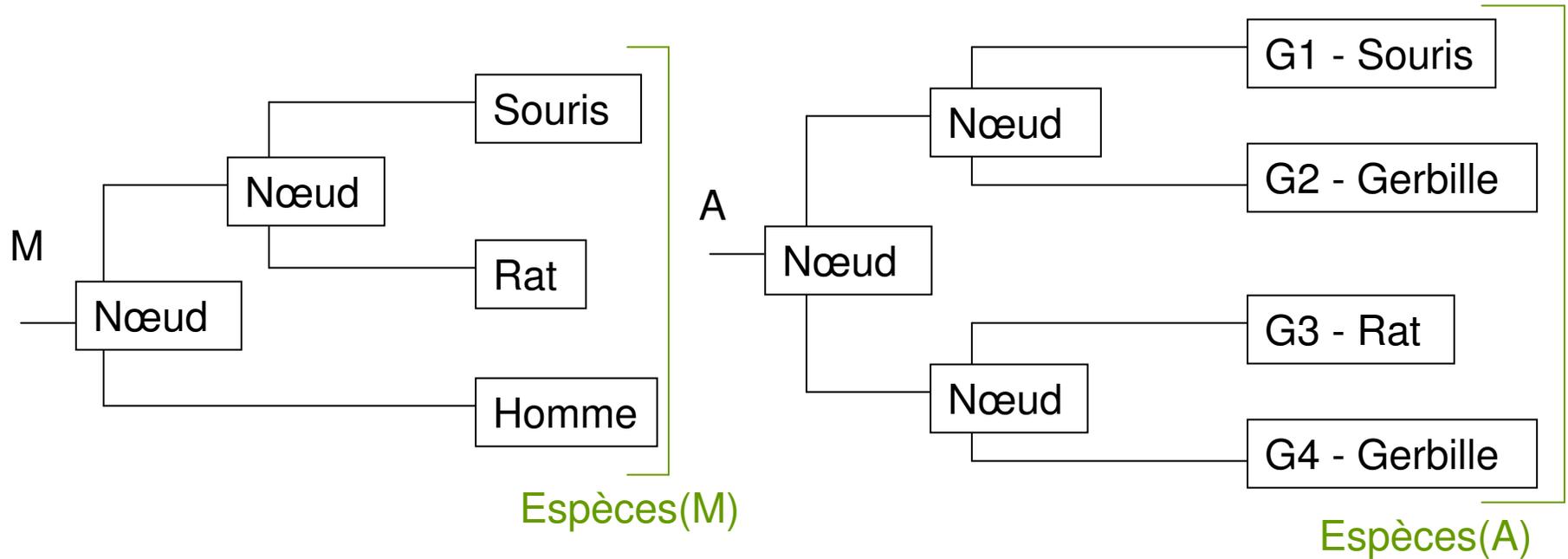
Transposition aux arbres phylogénétiques (1/2)



- 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 :

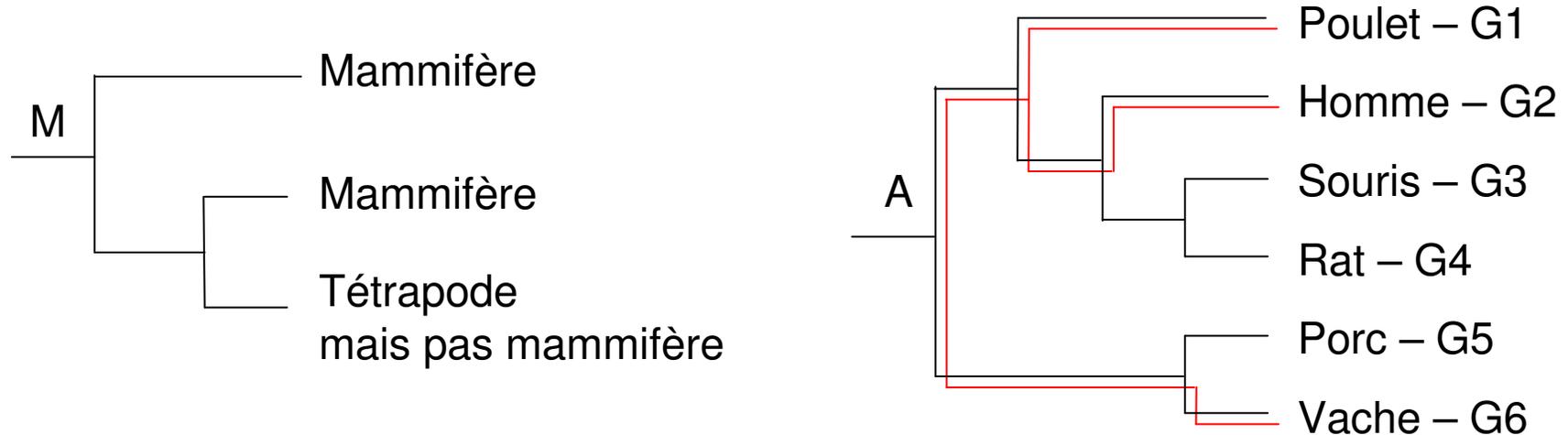
« M est motif de A » \Rightarrow espèces(M) \subseteq espèces(A)

Améliorations



Introduction de différents niveaux taxonomiques

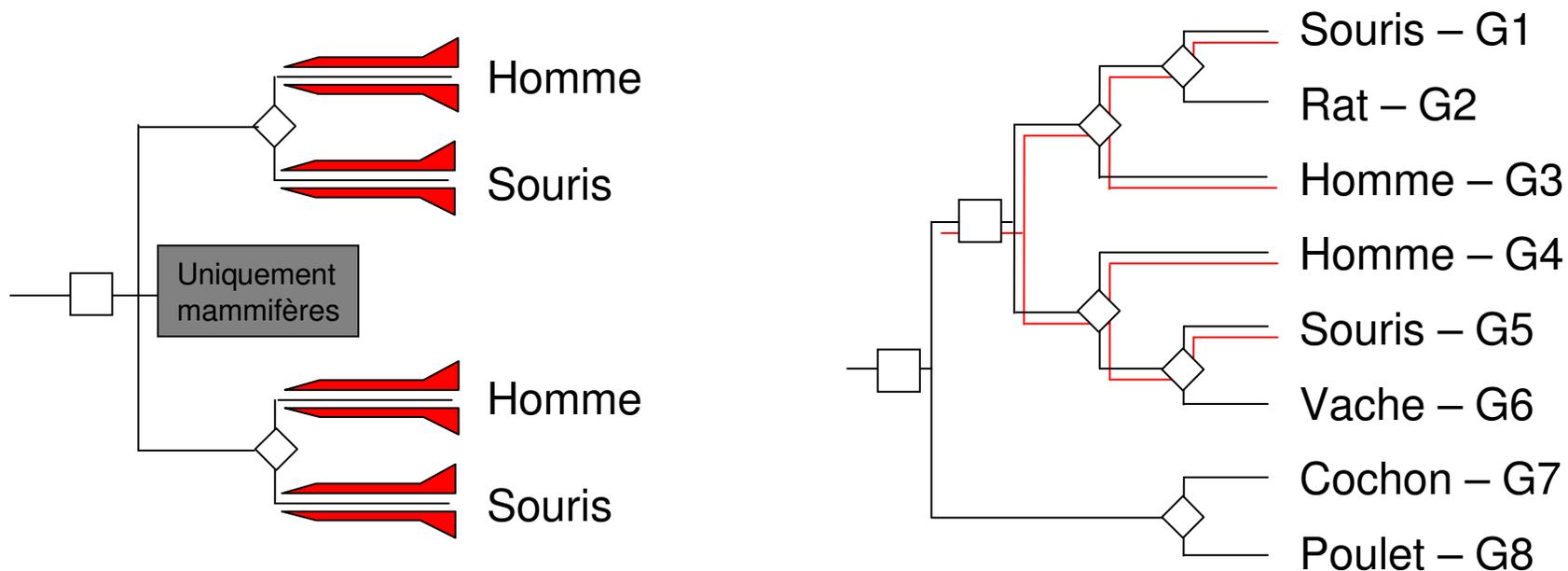
Les feuilles du motif peuvent être étiquetées avec des taxons de haut niveau.



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

Nœuds de duplications et de spéciations

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



 Pas de duplications

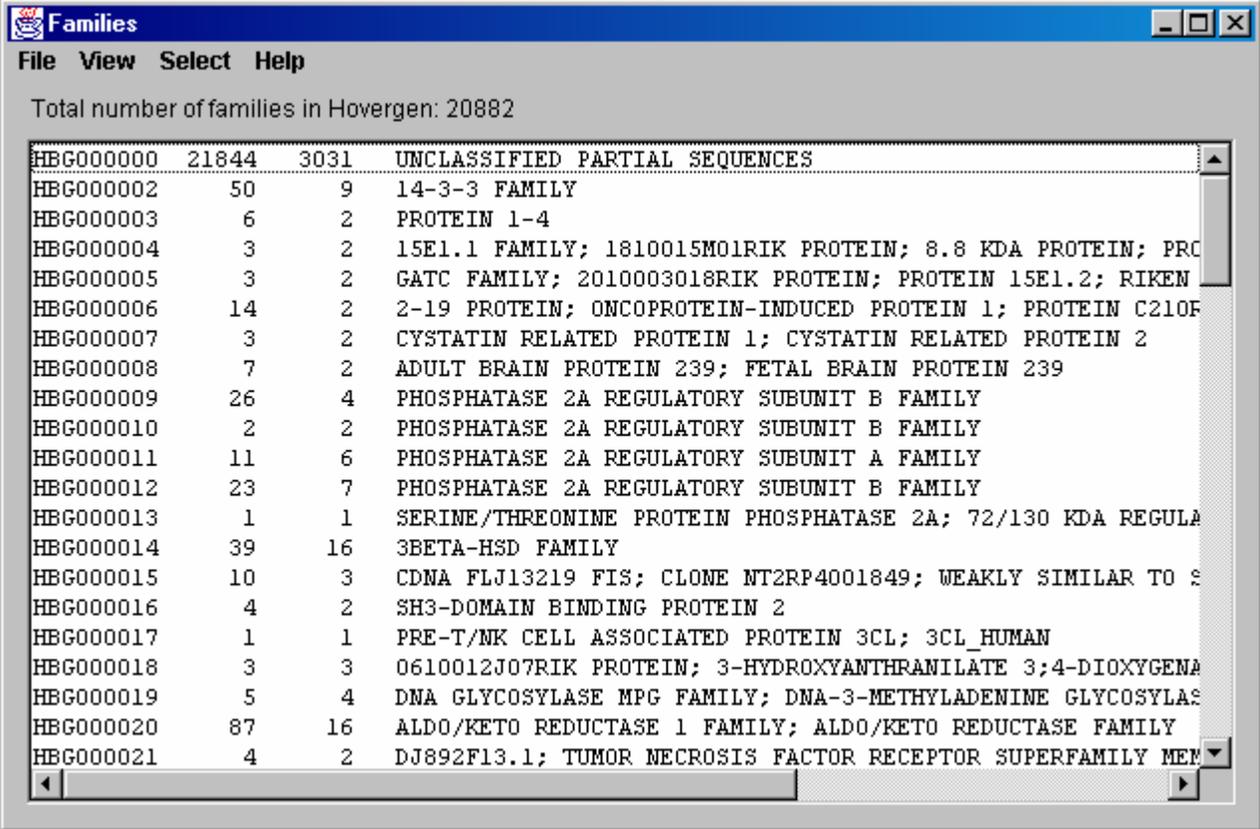
 Spéciation  Duplication

Implémentation



Intégration dans FamFetch (1/4)

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

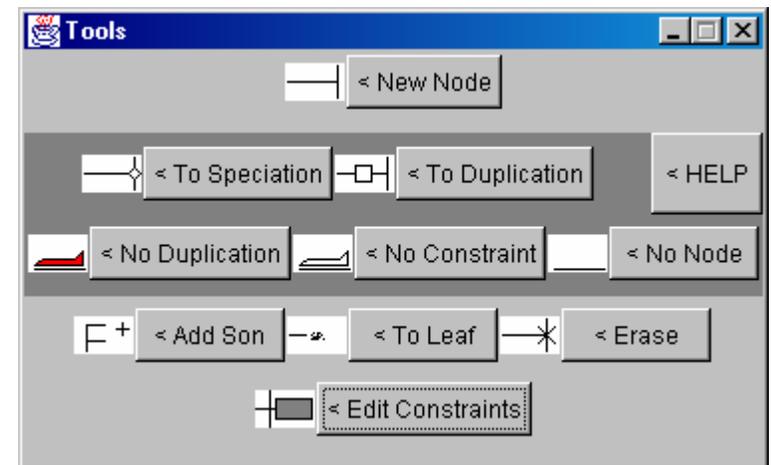
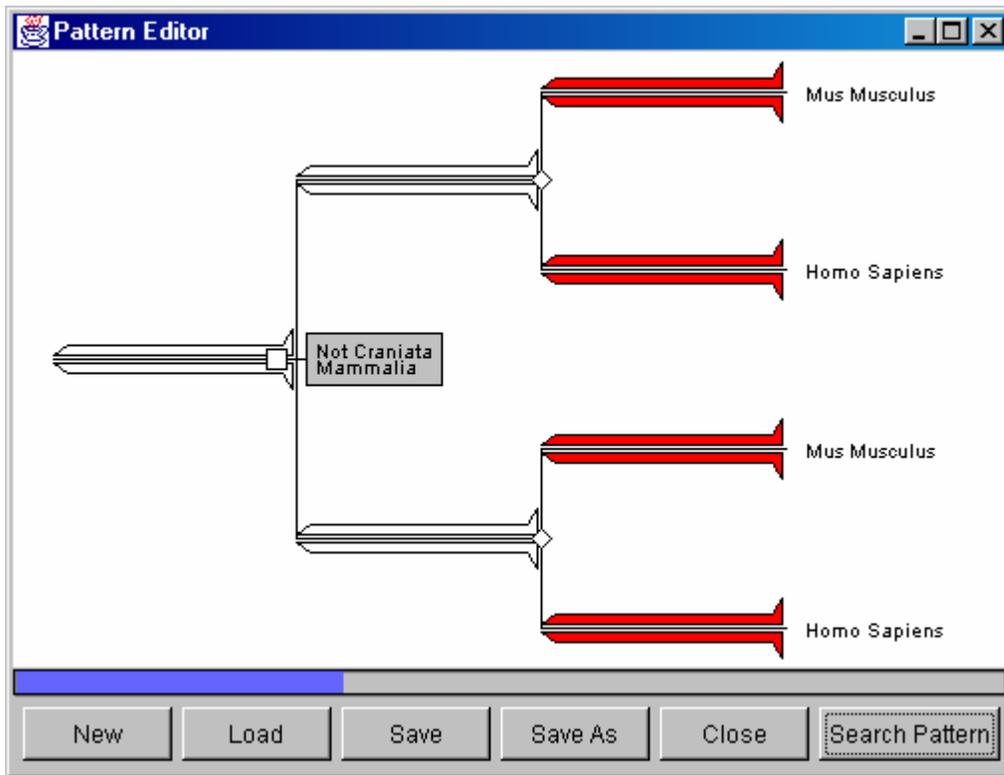


The screenshot shows a window titled 'Families' with a menu bar (File, View, Select, Help) and a status bar indicating 'Total number of families in Hovergen: 20882'. The main content is a list of families with columns for family ID, two counts, and a description.

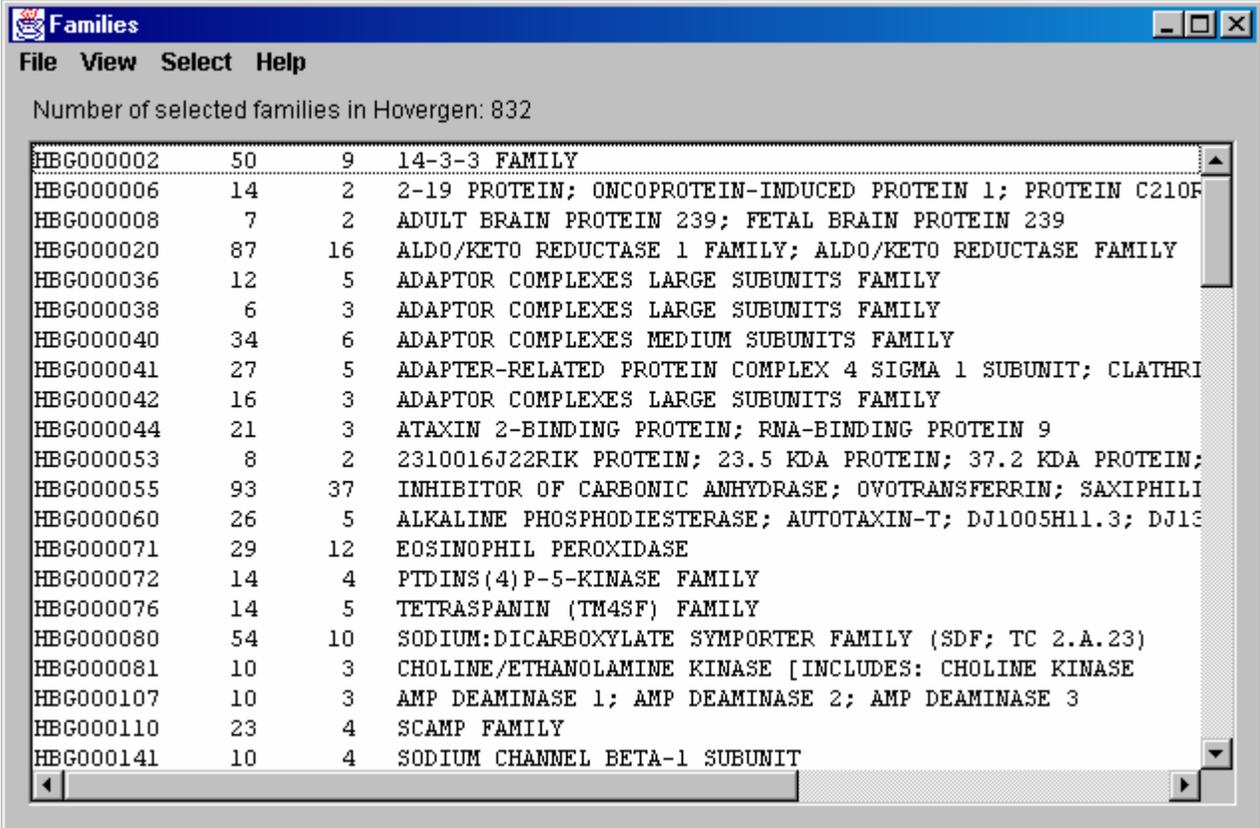
Family ID	Count 1	Count 2	Description
HBG000000	21844	3031	UNCLASSIFIED PARTIAL SEQUENCES
HBG000002	50	9	14-3-3 FAMILY
HBG000003	6	2	PROTEIN 1-4
HBG000004	3	2	15E1.1 FAMILY; 1810015M01RIK PROTEIN; 8.8 KDA PROTEIN; PRO
HBG000005	3	2	GATC FAMILY; 2010003018RIK PROTEIN; PROTEIN 15E1.2; RIKEN
HBG000006	14	2	2-19 PROTEIN; ONCOPROTEIN-INDUCED PROTEIN 1; PROTEIN C210F
HBG000007	3	2	CYSTATIN RELATED PROTEIN 1; CYSTATIN RELATED PROTEIN 2
HBG000008	7	2	ADULT BRAIN PROTEIN 239; FETAL BRAIN PROTEIN 239
HBG000009	26	4	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000010	2	2	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000011	11	6	PHOSPHATASE 2A REGULATORY SUBUNIT A FAMILY
HBG000012	23	7	PHOSPHATASE 2A REGULATORY SUBUNIT B FAMILY
HBG000013	1	1	SERINE/THREONINE PROTEIN PHOSPHATASE 2A; 72/130 KDA REGULA
HBG000014	39	16	3BETA-HSD FAMILY
HBG000015	10	3	CDNA FLJ13219 FIS; CLONE NT2RP4001849; WEAKLY SIMILAR TO S
HBG000016	4	2	SH3-DOMAIN BINDING PROTEIN 2
HBG000017	1	1	PRE-T/MK CELL ASSOCIATED PROTEIN 3CL; 3CL_HUMAN
HBG000018	3	3	0610012J07RIK PROTEIN; 3-HYDROXYANTHRAMILATE 3;4-DIOXYGENA
HBG000019	5	4	DNA GLYCOSYLASE MPG FAMILY; DNA-3-METHYLADENINE GLYCOSYLAS
HBG000020	87	16	ALDO/KETO REDUCTASE 1 FAMILY; ALDO/KETO REDUCTASE FAMILY
HBG000021	4	2	DJ892F13.1; TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY MEM

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

Intégration dans FamFetch (2/4)



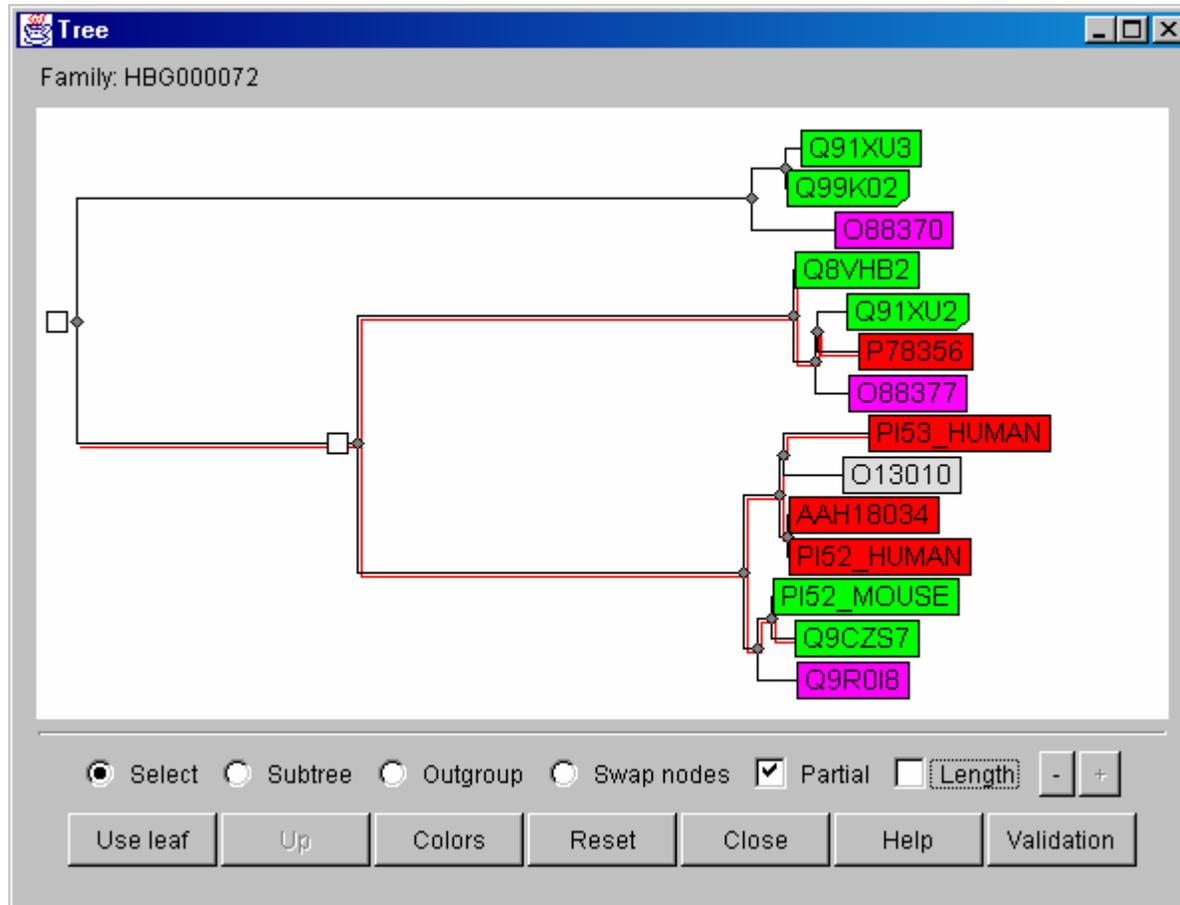
Intégration dans FamFetch (3/4)



Number of selected families in Hovergen: 832

HBG000002	50	9	14-3-3 FAMILY
HBG000006	14	2	2-19 PROTEIN; ONCOPROTEIN-INDUCED PROTEIN 1; PROTEIN C210F
HBG000008	7	2	ADULT BRAIN PROTEIN 239; FETAL BRAIN PROTEIN 239
HBG000020	87	16	ALDO/KETO REDUCTASE 1 FAMILY; ALDO/KETO REDUCTASE FAMILY
HBG000036	12	5	ADAPTOR COMPLEXES LARGE SUBUNITS FAMILY
HBG000038	6	3	ADAPTOR COMPLEXES LARGE SUBUNITS FAMILY
HBG000040	34	6	ADAPTOR COMPLEXES MEDIUM SUBUNITS FAMILY
HBG000041	27	5	ADAPTER-RELATED PROTEIN COMPLEX 4 SIGMA 1 SUBUNIT; CLATHRI
HBG000042	16	3	ADAPTOR COMPLEXES LARGE SUBUNITS FAMILY
HBG000044	21	3	ATAXIN 2-BINDING PROTEIN; RNA-BINDING PROTEIN 9
HBG000053	8	2	2310016J22RIK PROTEIN; 23.5 KDA PROTEIN; 37.2 KDA PROTEIN;
HBG000055	93	37	INHIBITOR OF CARBONIC ANHYDRASE; OVOTRANSFERRIN; SAXIPHILI
HBG000060	26	5	ALKALINE PHOSPHODIESTERASE; AUTOTAXIN-T; DJ1005H11.3; DJ13
HBG000071	29	12	EOSINOPHIL PEROXIDASE
HBG000072	14	4	PTDINS(4)P-5-KINASE FAMILY
HBG000076	14	5	TETRASPANIN (TM4SF) FAMILY
HBG000080	54	10	SODIUM:DICARBOXYLATE SYMPORTER FAMILY (SDF; TC 2.A.23)
HBG000081	10	3	CHOLINE/ETHANOLAMINE KINASE [INCLUDES: CHOLINE KINASE
HBG000107	10	3	AMP DEAMINASE 1; AMP DEAMINASE 2; AMP DEAMINASE 3
HBG000110	23	4	SCAMP FAMILY
HBG000141	10	4	SODIUM CHANNEL BETA-1 SUBUNIT

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.