Phylogénies

Céline Brochier

04.91.1.64.75

Celine.brochier@up.univ-mrs.fr

http://http://194.57.197.233:800/

Notions fondamentales

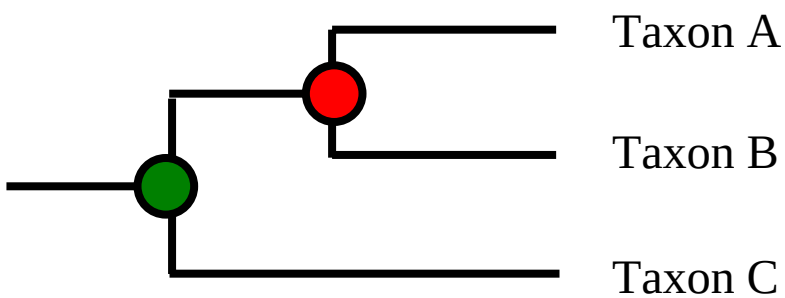
- Définitions:
  - Phylogénie : Etude des relations de parenté entre les organismes ou les taxons
  - Taxon : Rang taxonomique identifié, quelque en soit le niveau
    - Bos taurus est un taxon de rang spécifique
    - Bos est taxon de rang générique
    - Bovidae est un taxon de rang familial
    - Certiodactyla est un taxon de rang ordinal
- Hypothèse origine unique de la vie ⇔ tous les organismes sont plus ou moins fortement apparentés

Notions fondamentales

- Les relations entre les organismes sont représentés par des arbres

  ⇒Les feuilles représentent les taxons (UTO = unités taxonomiques opérationnelles)

  ⇒ Les nœuds symbolisent des ancêtres hypothétiques (UTH = unités taxonomiques hypothétiques)



  ⇒A est plus proche parent de B que de C, car A et B partagent un ancêtre commun exclusif qui n'est pas un ancêtre de C

  ⇒A, B e C partagent également un ancêtre commun qui est plus ancien que le dernier ancêtre commun de A et B

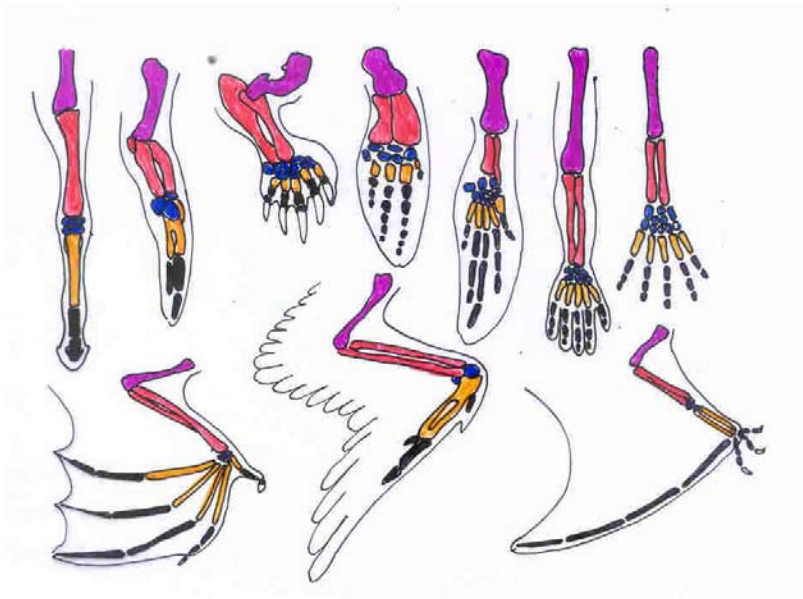
Construction de phylogénies

- Basée sur la comparaison de caractère homologues (i.e. ayant une origine évolutive commune)
- Caractère : tout attribut observable d'un organisme (écologique, morphologique, comportemental, moléculaire, physiologique…)
  - Ex : œil, hibernation, respiration cellulaire, séquence de l'ARNr 16S…
- État de caractère : forme particulière d'un caractère
  - Ex : Caractère : couleur de l'œil : état de caractère : bleu, marron, vert, absence d'œil
  - Caractère : position 50 du facteur d'élongation IF-5A : état de caractère : présence d'une lysine

Processus évolutifs : Homologie

- Homologie : « sont dit homologues des caractères (moléculaires, morphologiques …) ayant une origine évolutive commune mais pouvant avoir des formes ou des fonctions différentes »

Ex. Le membre des tétrapodes est constitué de plusieurs os organisés selon le même plan, mais pouvant présenter des formes différentes et des fonctions différentes (nage, course, vol…)



Processus évolutifs : Analogie

- Analogie ou similitude fonctionnelle : « sont dit analogues des caractères (moléculaires, morphologiques …) pouvant présenter des similitudes de fonction ou de forme, mais ayant une origine évolutive distincte »

  Ex. les ailes des insectes et celles des oiseaux sont deux structures qui accomplissent la même fonction, voler, mais les organes qui permettent cette activité sont différents entre eux, et par leur origine et par leur structure.

  Ex. Cristallines de l'œil sont dérivées de protéines d'origine évolutive différentes (chaperonines, lactate dehydrogenase B…)

  $\Rightarrow$ Recrutement secondaire indépendant de protéines ayant des caractéristiques physico-chimiques semblables

Construction de phylogénies

- Établissement de matrices de caractères qui décrivent pour chaque taxon, l'état de caractère de chaque caractère
  - Ex. Comparaison de séquences d'ARNr 18S chez les métazoaires

```
Taxon 1          AACGAAAAAAAAAAAAAAAAGAAAAAAAAGAAAACCACA
Taxon 2          ATACGCATAAAAAAAAAAACGAAAAAAAAGAAAACCACA
Taxon 3          ATAAGAATAAGAAAAAAAAACGAAAAAAAAGAAAACCACA
Taxon 4          TTATGCATAAGAAAAAAAAGCGAAATAAAAGAAAACCACA
Taxon 5          ATCCTAATAAGAAAACAAAGCGAAATAAAAGAAAACCACA
Taxon 6          ATAAACATAAGAAAACAAACCGAAATAAAAGAAAACCACA
Taxon 7          TTAAGAATAAGAAAACAAACCGATATAAAAGAAAACCACA
Taxon 8          AGGCGCATAAGAAAACAAACCGATATAAAAGAACACCACA
Taxon 9          AGCCACATAAGAAAACAAAACGATATAAGAGAACACCACA
Taxon 10         AGACGCATAAGAAAACAACACGATATAAGAGAACACCATA
                 12345678910…
```

- Caractères = sites homologues (1,2,3,4,5,6,7,8,9,10…)
- États de caractères = nature du nucléotide à la position considérée

Construction de phylogénies

- Choix d'une méthode de reconstruction
  - Maximum de parcimonie
  - Méthodes de distances
  - Maximum de vraisemblance
  - Méthodes bayésiennes…
- Données de la matrice de caractères => inférence d'un arbre phylogénétique qui modélise au mieux les données de la matrices

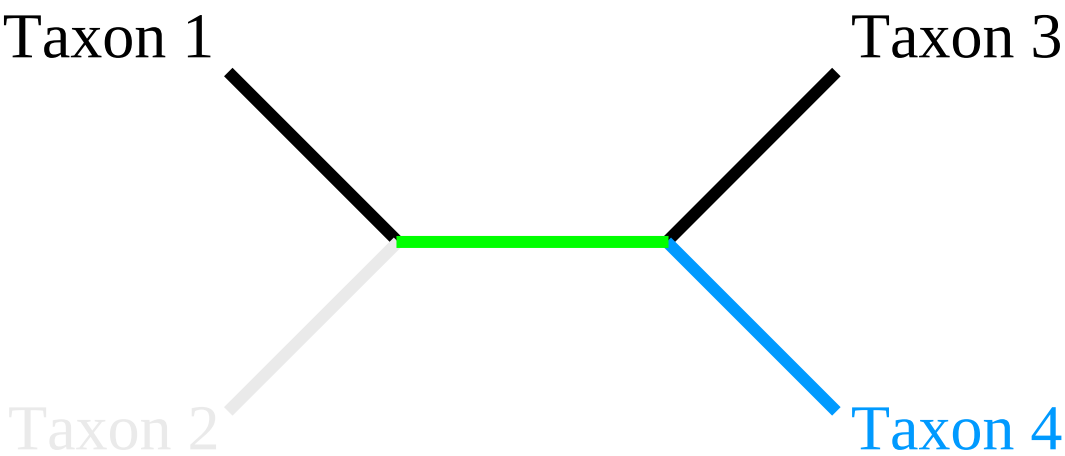
Construction de phylogénies

- Un arbre phylogénétique est une structure mathématique qui est utilisée pour modéliser l'histoire évolutive d'un groupe d'organismes.

- **LES PHYLOGENIES SONT DES HYPOTHESES !** elles ne peuvent pas être observées, elles ne peuvent être qu'inférées, parce qu'elles reflètent des événements évolutifs passés

Arbres non racinés

- Les arbres obtenus sont généralement non racinés

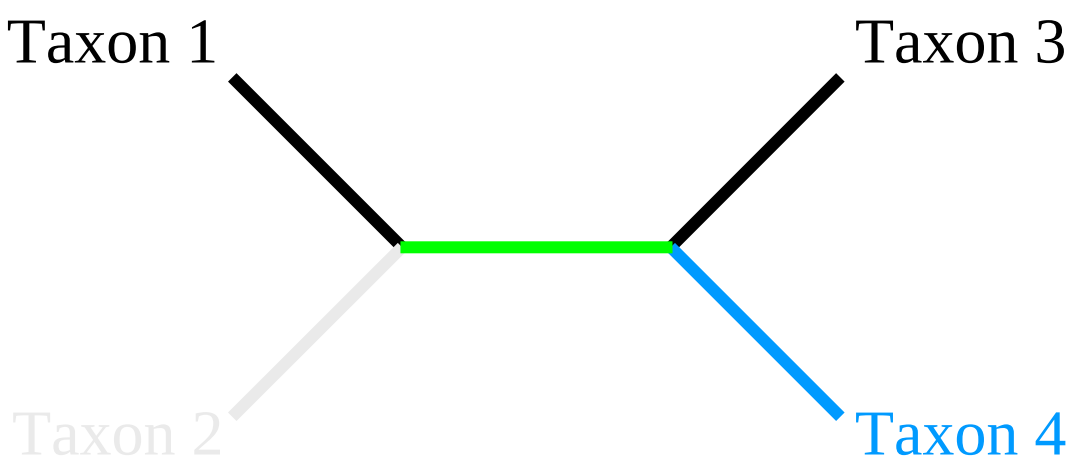


  - Les arbres non racinés ne permettent pas une orientation de l'arbre dans le temps ! => Pas d'indications sur les relations de parentés entre les taxons

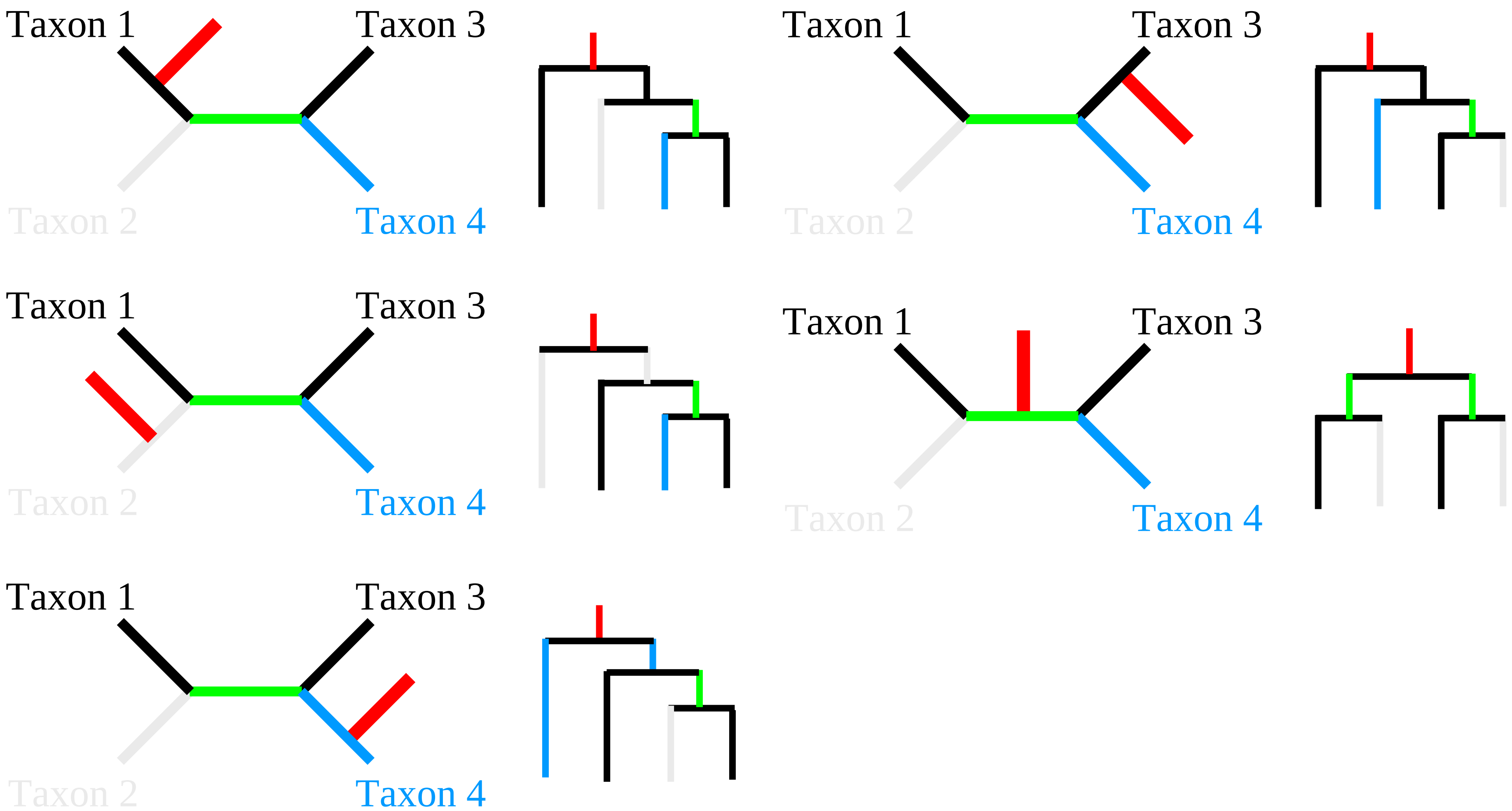
  $\Rightarrow$ Placement d'une racine

Raciner les phylogénies

- Placement de la racine dans un arbre à 4 taxons
  - 5 emplacements potentiels
  - Utilisation d'un groupe extérieur
  - Placement au poids moyen

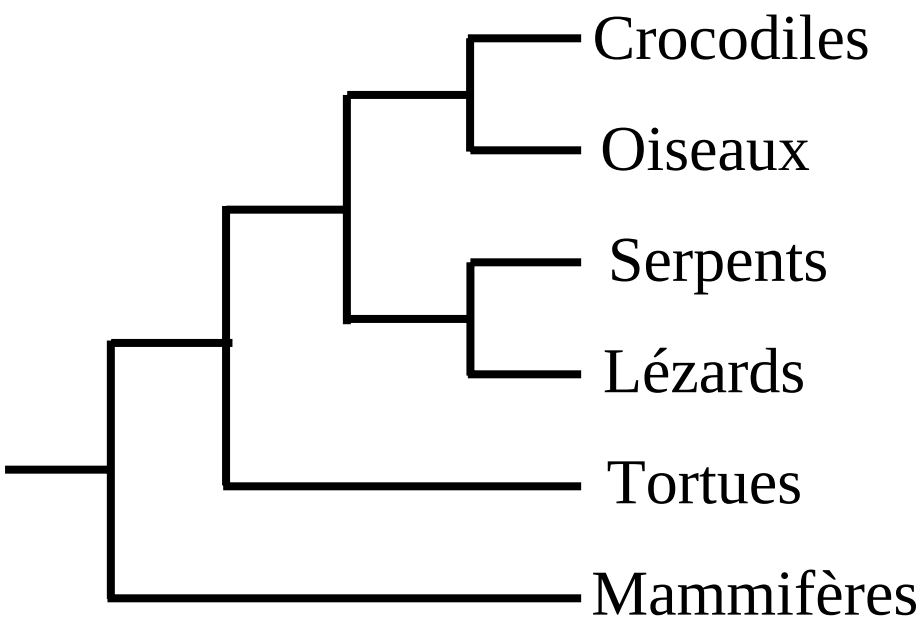


Raciner les phylogénies



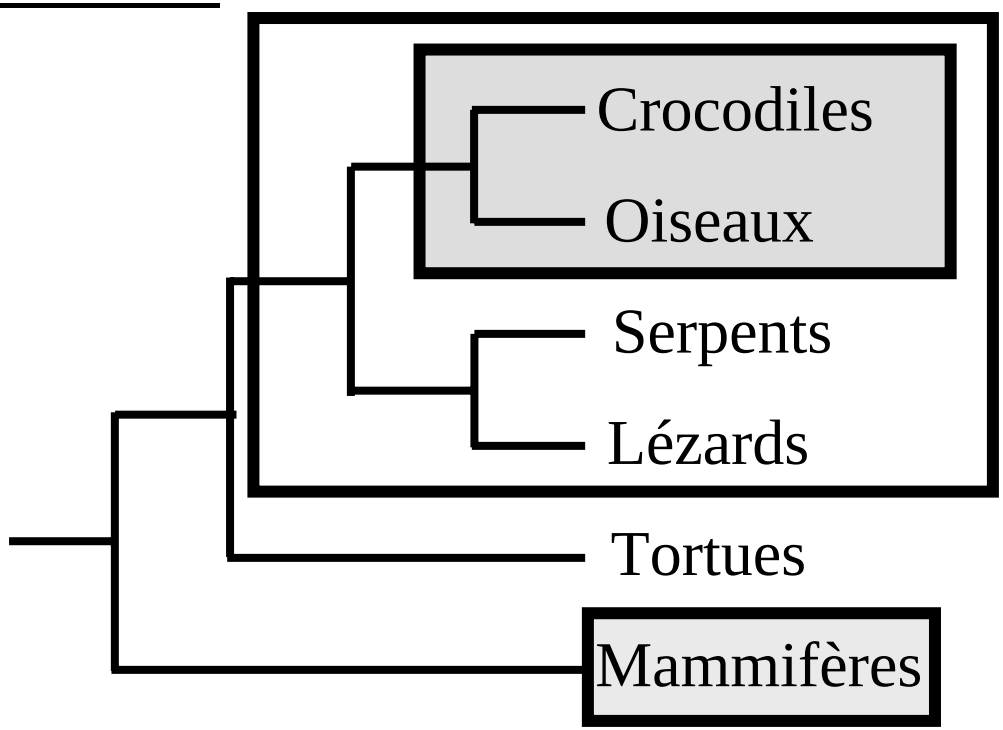
Notion de groupe en phylogénie

- Ex. Phylogénie des amniotes



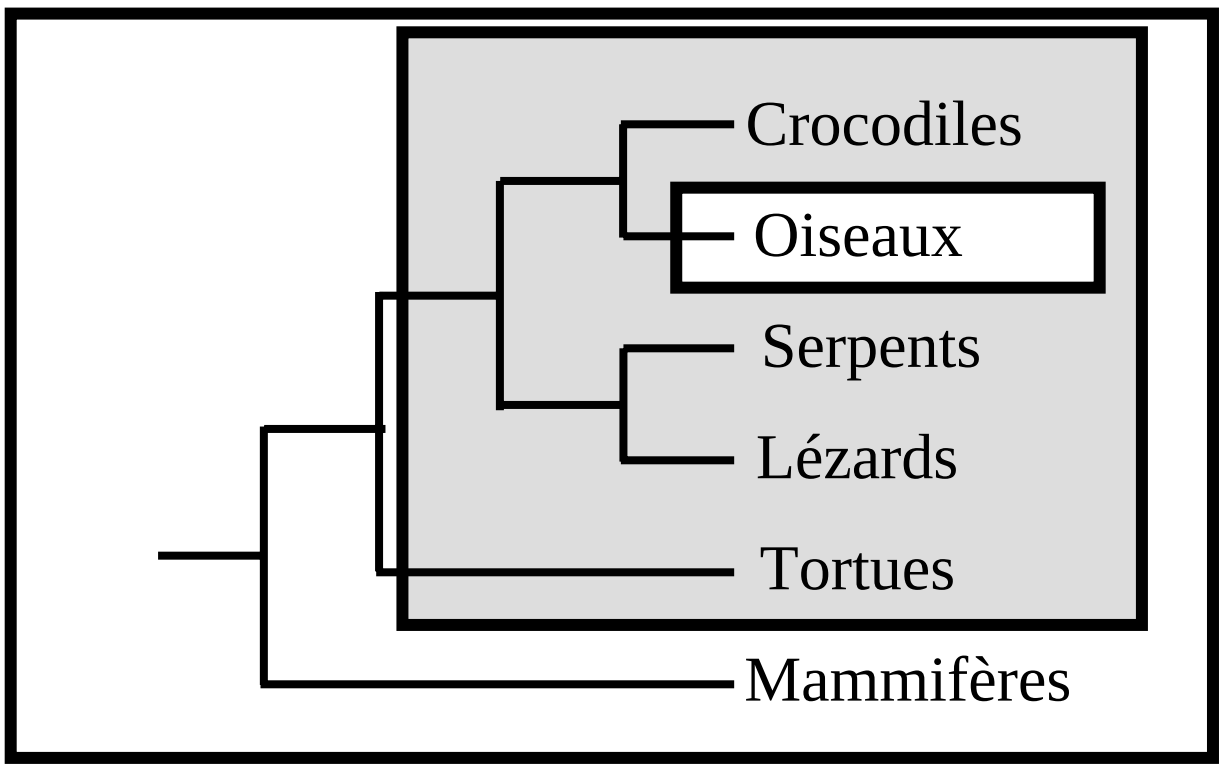
Groupes Monophylétiques

- Définition : groupes incluant un ancêtre et la totalité de ses descendants



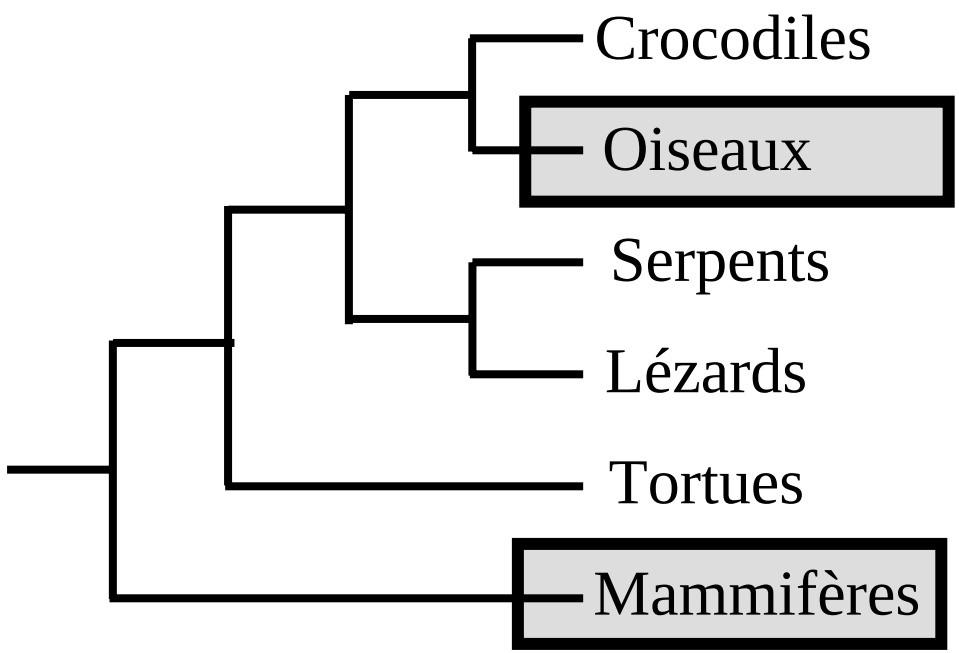
Groupes Paraphylétiques

- Définition : groupes incluant un ancêtre et une partie de ses descendants



Groupes Polyphylétiques

- Définition : groupes ayant des origines évolutives distinctes



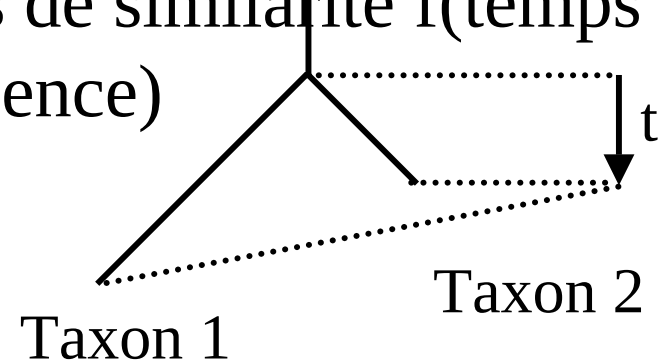
Groupe des homéothermes

Phylogénies moléculaires

- Basées sur la comparaison de séquences moléculaires homologues
  - $\Rightarrow$ Origine évolutive commune
  - $\Rightarrow$ Descendent d'un même ancêtre
  - $\Rightarrow$ Sont apparentées
- Phylogénies des gènes $\Leftrightarrow$ Phylogénies des taxons ?

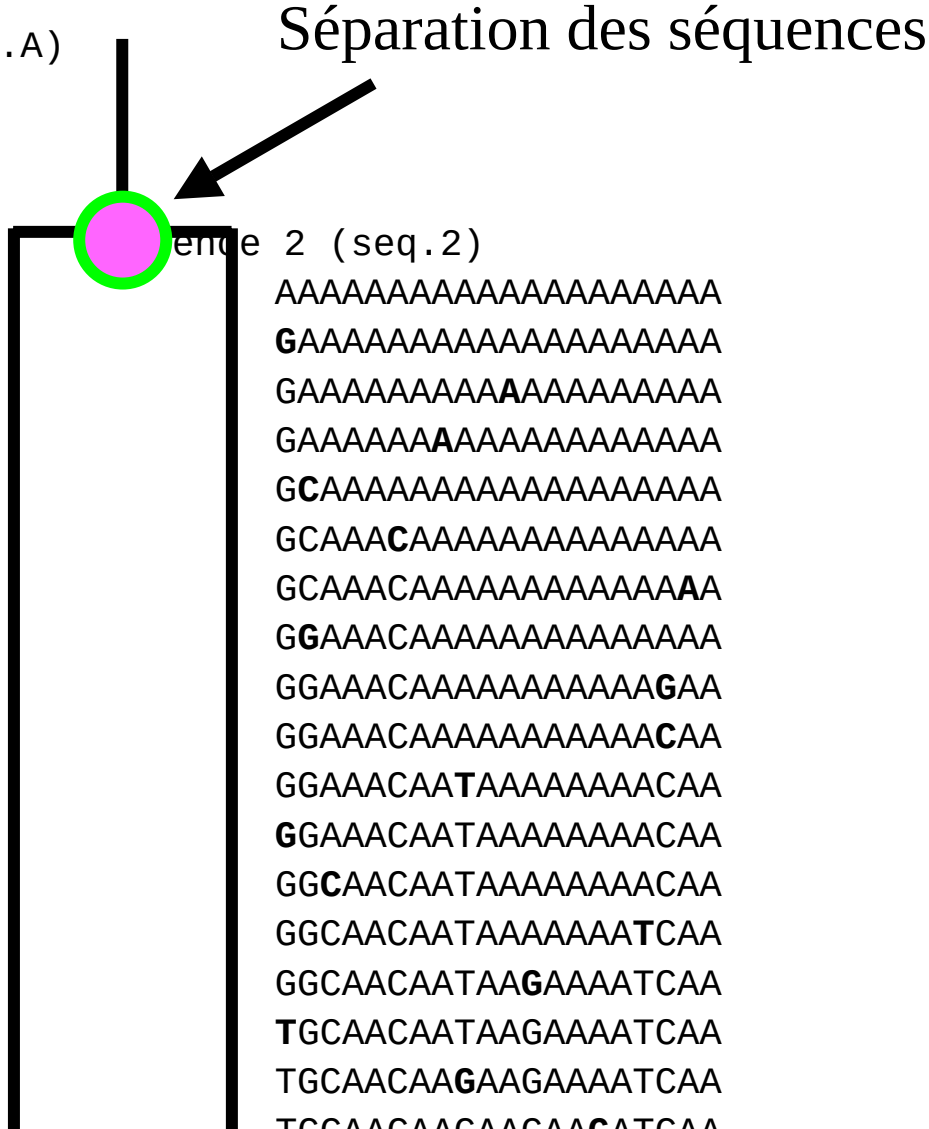
Critère d'homologie

- Pas de critère absolu
- L'homologie va être définie par la similarité observée entre les séquences
  - Deux séquences homologues partagent une origine évolutive commune => présentent un degrés de similarité f(temps écoulé depuis leur divergence)



Évolution des séquences

Séquence initiale (seq.A)
AAAAAAAAAAAAAAAAAAAA



	Séquence 1 (seq.1)	Séquence 2 (seq.2)
t=0	AAAAAAAAAAAAAAAAAAAA	AAAAAAAAAAAAAAAAAAAA
t=1	AAAAAAAAAAAAA**C**AAAAAA	**G**AAAAAAAAAAAAAAAAAA
t=2	AAAAAAAAA**G**AAACAAAAAA	GAAAAAAAAA**A**AAAAAAAAA
t=3	AAAAA**G**AAAGAAACAAAAAA	GAAAAAA**A**AAAAAAAAAAA
t=4	A**C**AAAGAAAGAAACAAAAAA	G**C**AAAAAAAAAAAAAAAAAA
t=5	ACAAAGAAAGAAACA**G**AAAA	GCAAA**C**AAAAAAAAAAAAAA
t=6	ACAAA**T**AAAGAAACAGAAAA	GCAAACAAAAAAAAAAAA**A**A
t=7	**C**CAAATAAAGAAACAGAAAA	G**G**AAACAAAAAAAAAAAAAA
t=8	CC**C**AATAAAGAAACAGAAAA	GGAAACAAAAAAAAAAA**G**AA
t=9	CCCAATAAA**T**AAACAGAAAA	GGAAACAAAAAAAAAAA**C**AA
t=10	CCCAAT**A**AATAAACAGAAAA	GGAAACAA**T**AAAAAAAACAA
t=11	CCCAATA**G**ATAAACAGAAAA	**G**GAAACAATAAAAAAAACAA
t=12	CCCAATAGATAA**G**CAGAAAA	GG**C**AACAATAAAAAAAACAA
t=13	CCC**A**ATAGATAAGCAGAAAA	GGCAACAATAAAAAAA**T**CAA
t=14	CCCAATA**A**ATAAGCAGAAAA	GGCAACAATAA**G**AAAATCAA
t=15	CCCA**T**TAAATAAGCAGAAAA	**T**GCAACAATAAGAAAATCAA
t=16	CCCATTAAATAAGCAG**A**AAA	TGCAACAA**G**AAGAAAATCAA
t=17	CCCATTAAATA**G**GCAGAAAA	TGCAACAAGAAGAA**G**ATCAA

Séquences biologiques : Homologie ou similarité ?

- Deux séquences sont dites homologues si elles possèdent un ancêtre commun

- L'existence d'un ancêtre commun est inférée à partir de la similarité

- Seuil pour les protéines :30 % d'identité sur une longueur de 100 AA $\Rightarrow$ homologie entre les séquences

Séquences biologiques : Similarité sans homologie

- La similarité n'est pas toujours due à de l'homologie :
  - Convergence ou simple hasard pour de courtes séquences (quelques résidus)
  - Biais de composition en aa ou nucléotides (taux de GC…)
  - Existence de régions de faible complexité (*e.g.*, cas de la fibroïne $[GSGAGA]_n$) :
    - Présentes dans 40 % des protéines.
    - Peuvent représenter jusqu'à 15 % du total des résidus (Ala, Gly, Pro, Ser, Glu et Gln)

Séquences biologiques : Homologie sans similarité

- Deux séquences peuvent être homologues sans que leur similarité soit forte :

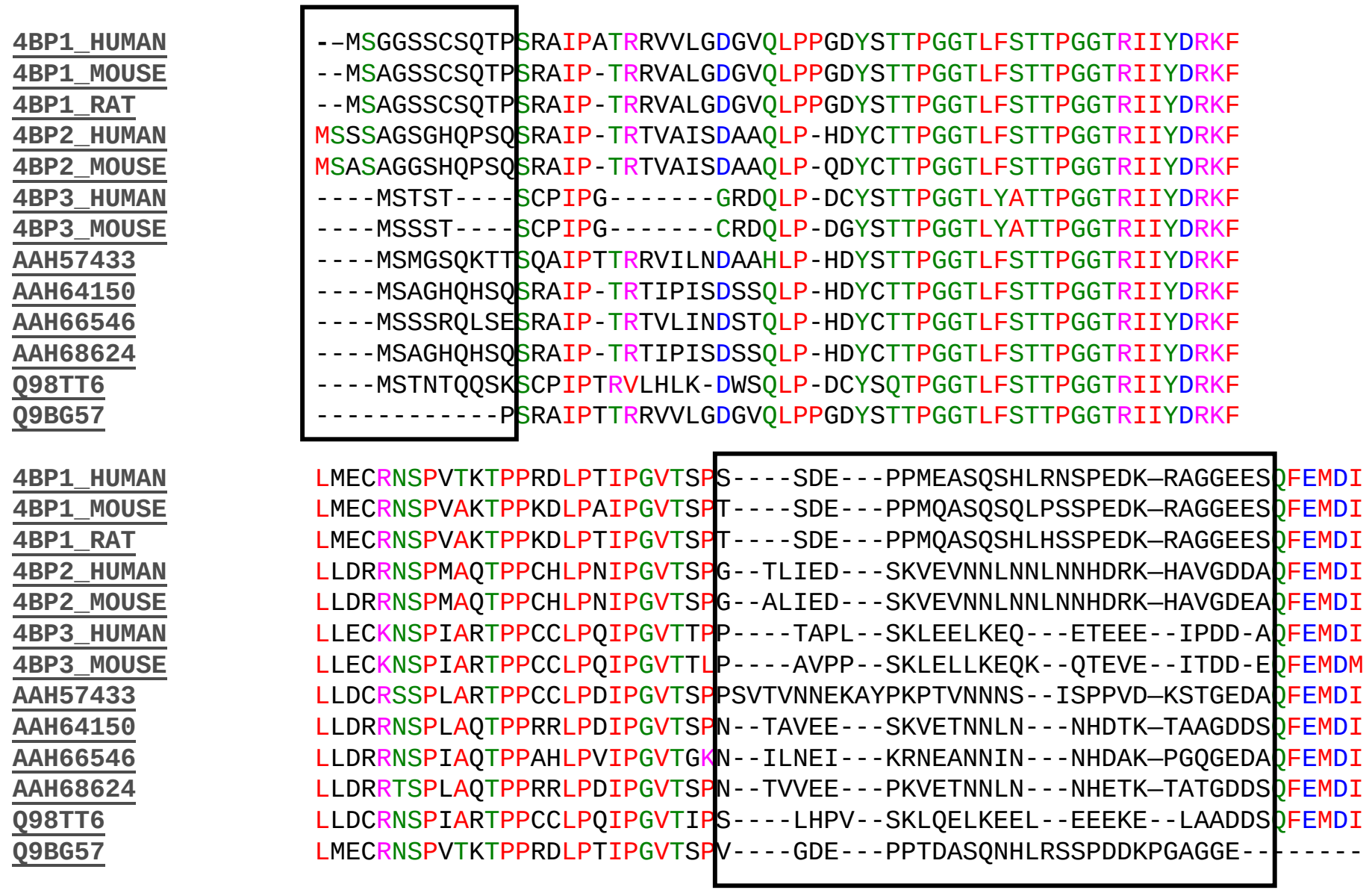
```
ACP_KLEAE  ---MEMKIDALAGTLESSDVMVRIGPAAQPGIQLEIDSIVKQEFGAAIQQVVRETLAQLG
ACP_ECOLI  STIEERVKKIIGEQLGVKQEEVTDN--ASFVEDLGADSLDTVELVMALEEEFDTEIPDEE
               *   :  :  *   :  *     *    :*  **:   *   *:::  :  :::

ACP_KLEAE  VKECDNVQLARVQAAALRWQQ
ACP_ECOLI  AEKITTVQAAIDYINGHQA--
            ::   ** *     : :
```

La similarité entre ces protéines est faible mais les données fonctionnelles et biochimiques montrent qu'elles sont homologues.

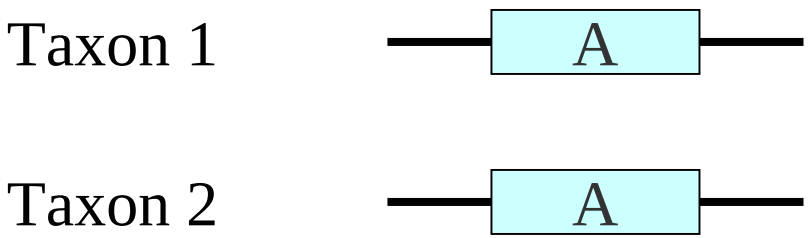
Construction d'une phylogénie moléculaire

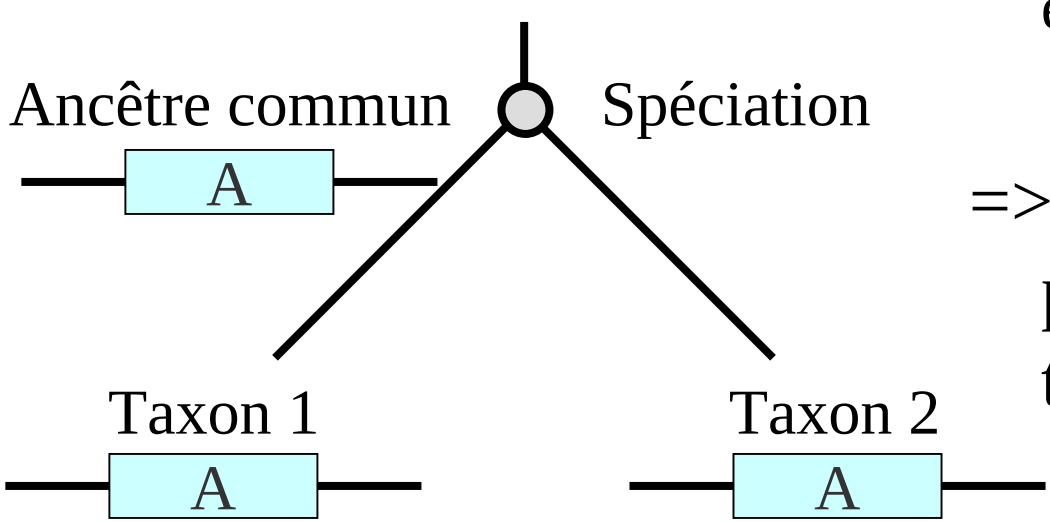
- Récupération des séquences homologues dans les banques de données (Swiss-Prot, Genbank…)
- Alignement des séquences (clustalw)
- Vérification manuelle de l'alignement et élimination des régions où l'homologie des positions n'est pas sure
- Construction de la phylogénie
- Interprétation ⇔ identification des orthologues, homologues, xénologues pour faire des hypothèses sur l'évolution des gènes

Elimination des régions où l'homologie des sites est douteuse



Séquences biologiques : Orthologie

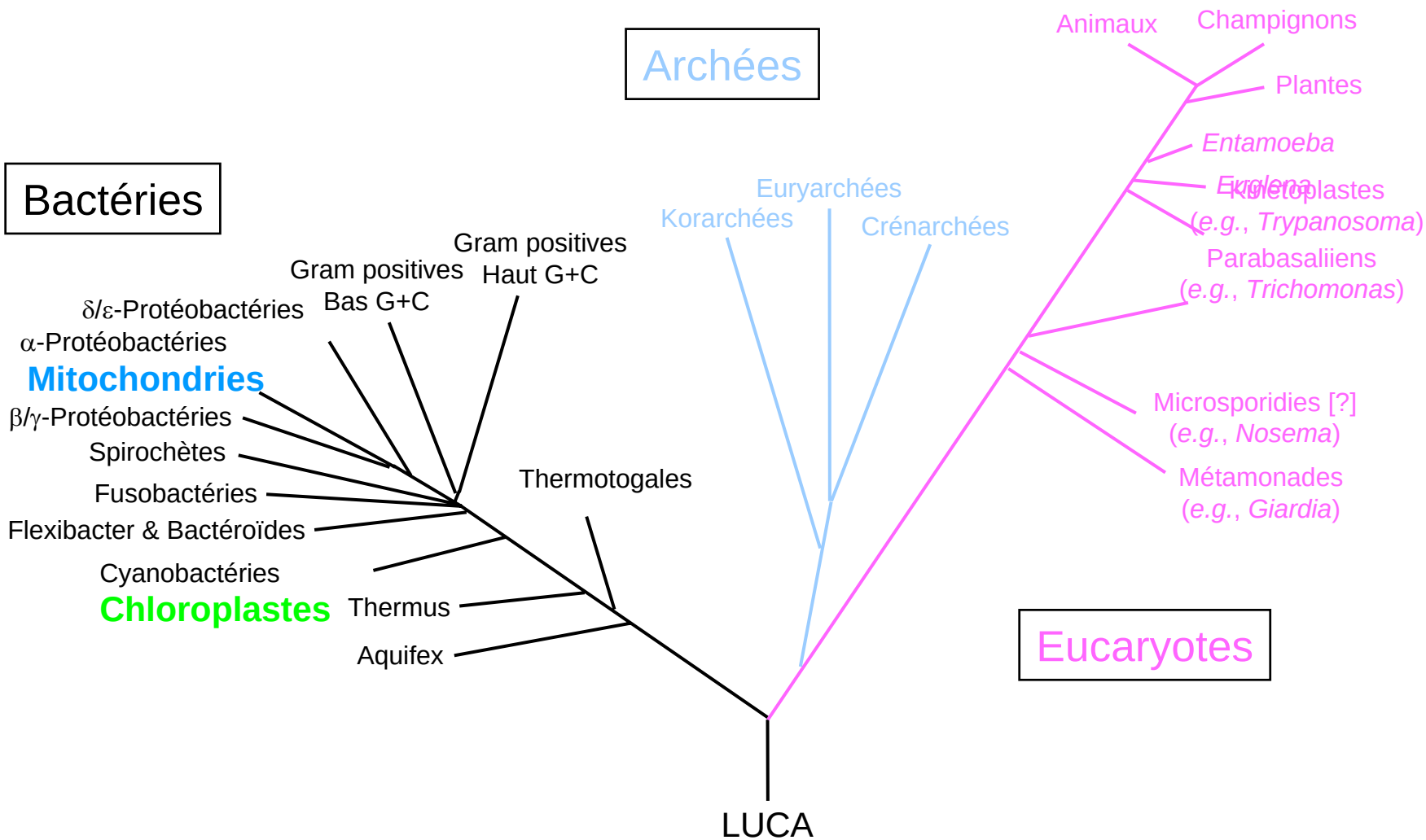




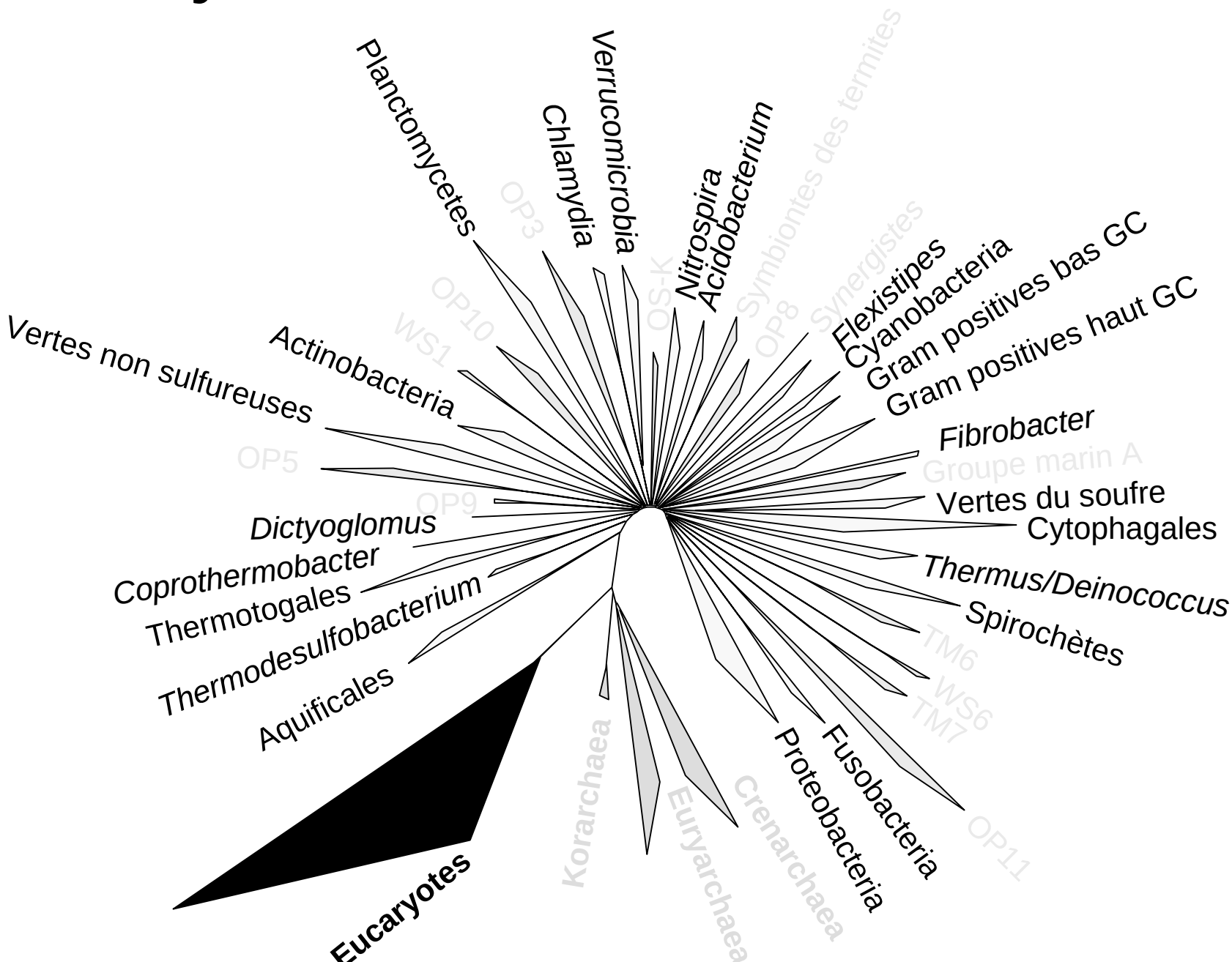
- Définition «gènes présents dans des organismes différents, ayant évolué à partir d'un même gène ancestral suite à des événements de spéciation»

=> Étude des relations de parentés entre les taxons

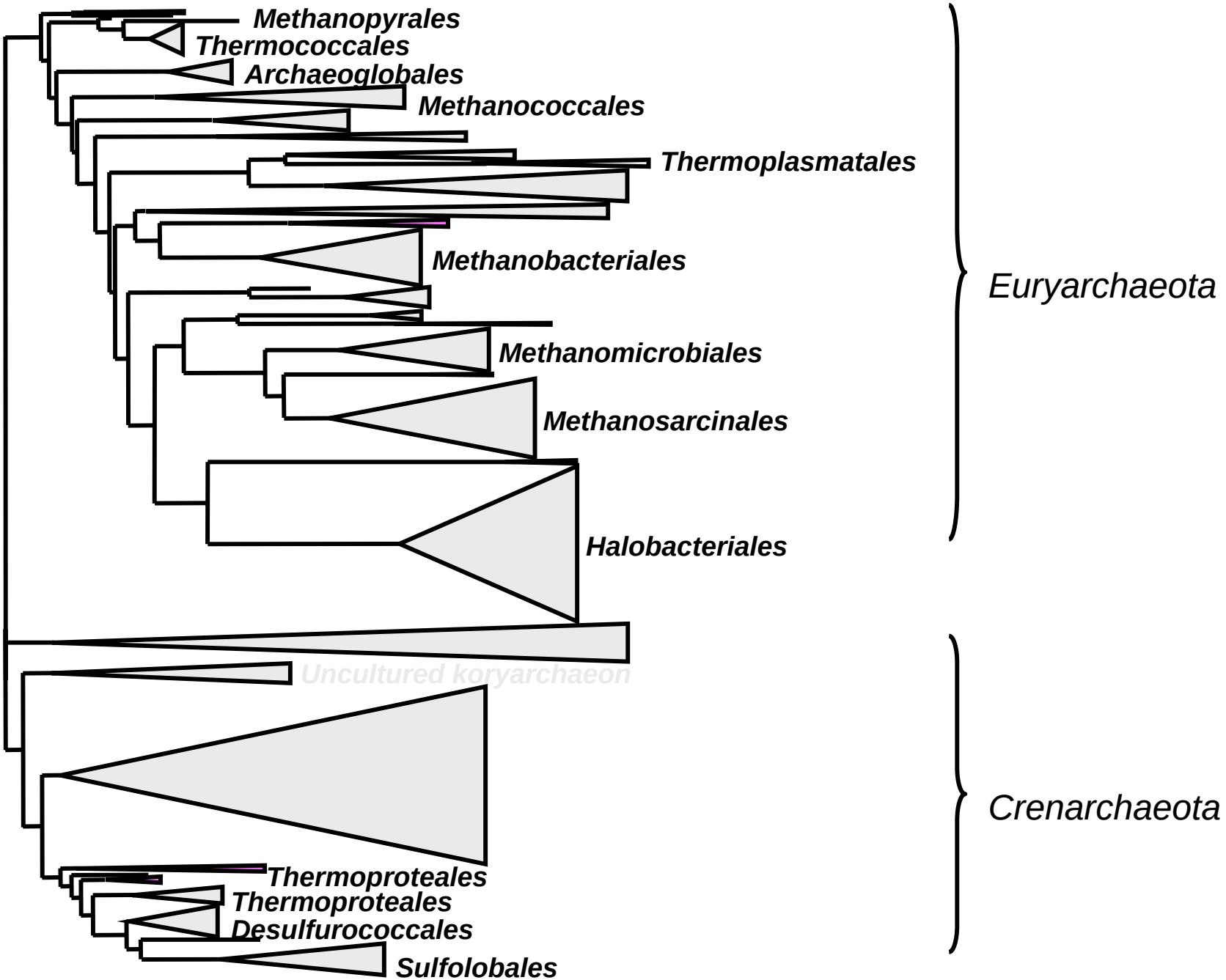
L’arbre de universel du vivant (ARNr 16S)



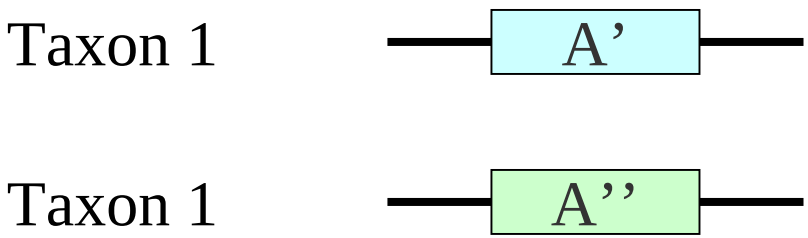
L’incroyable diversité des Eubactéries…



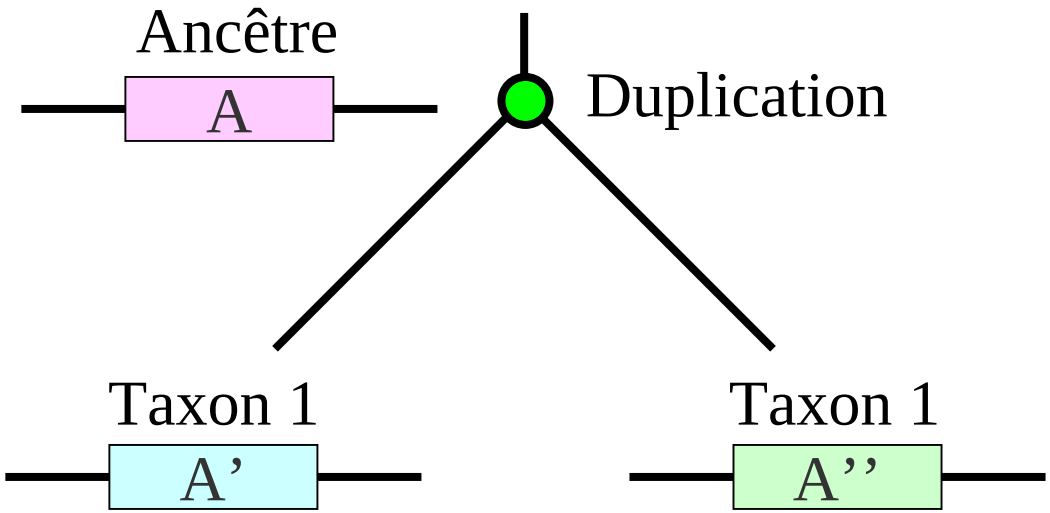
et des Archébactéries…



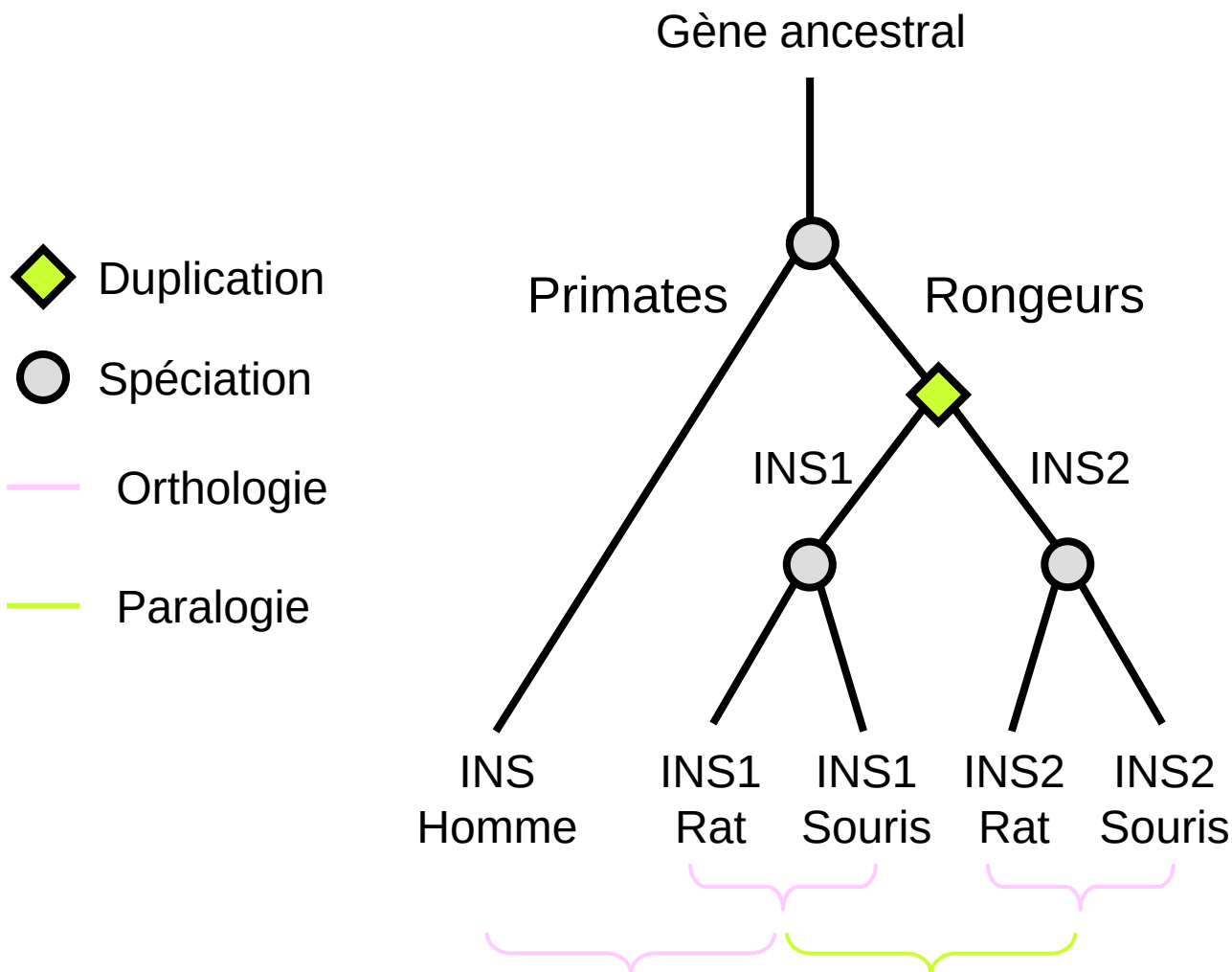
Séquences biologiques : Paralogie



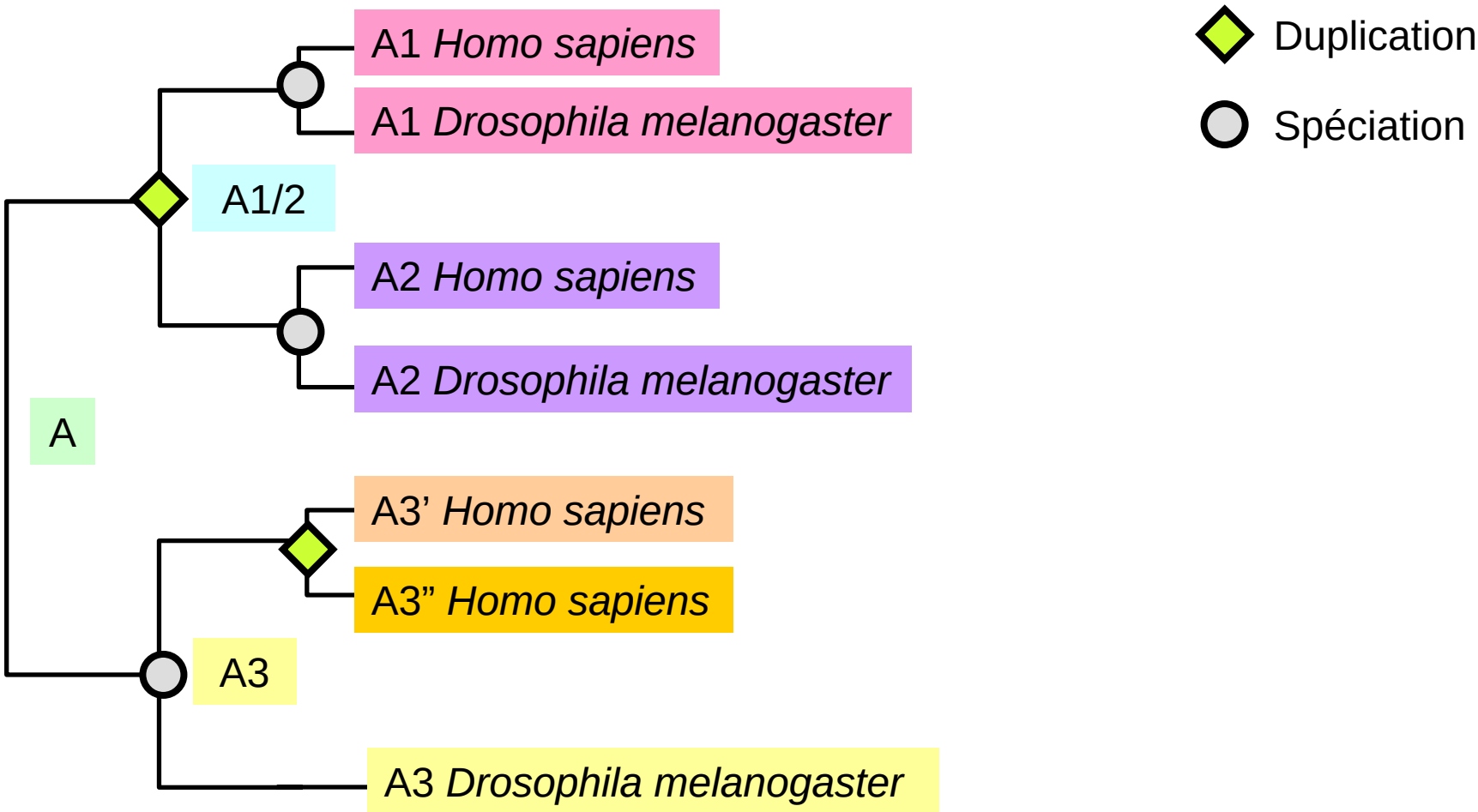
- Définition « gènes issus d'événements de duplication au sein d'un génome»



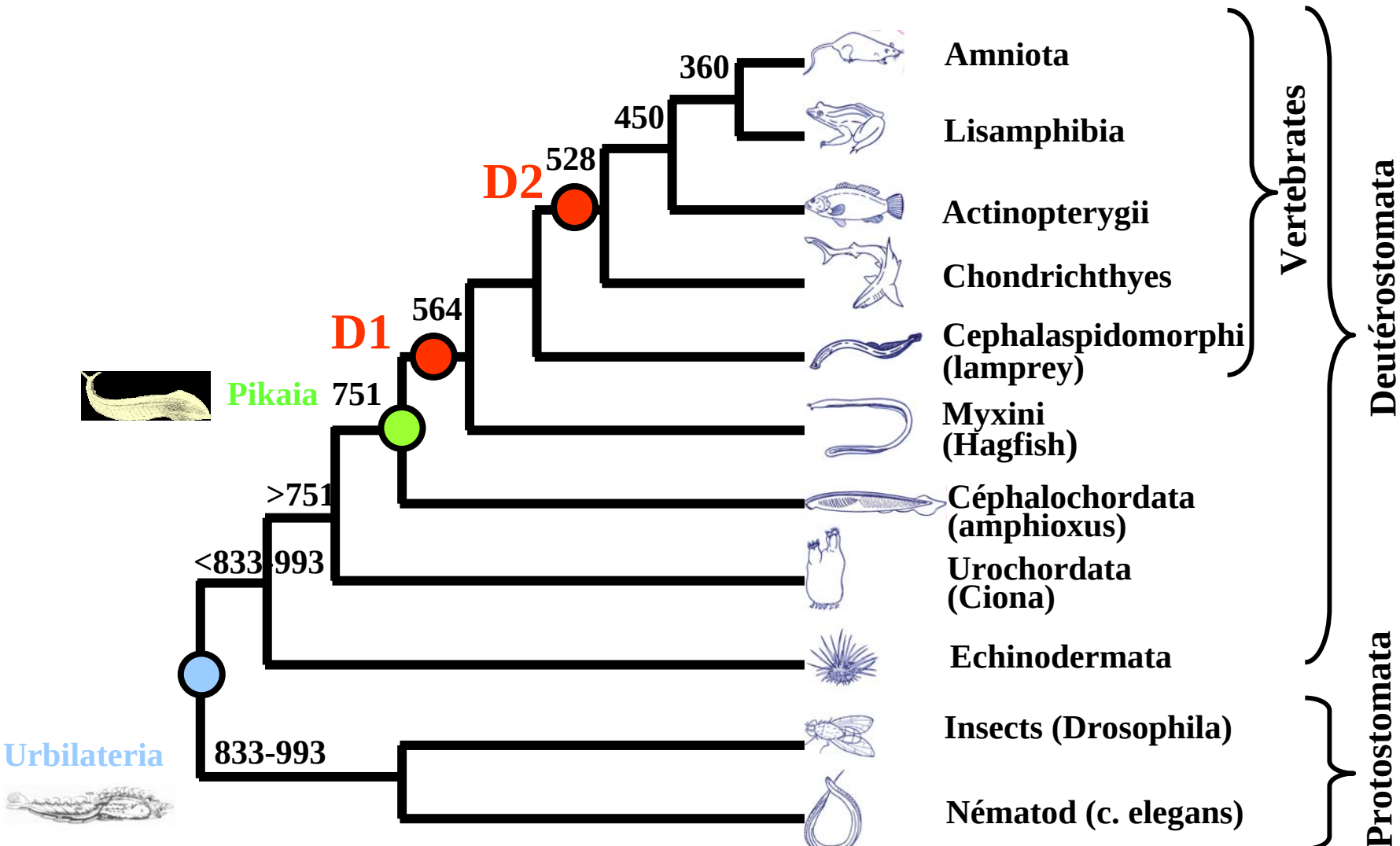
Séquences biologiques : Orthologues et paralogues



Séquences biologiques : Orthologues et paralogues



Duplications massives de génomes

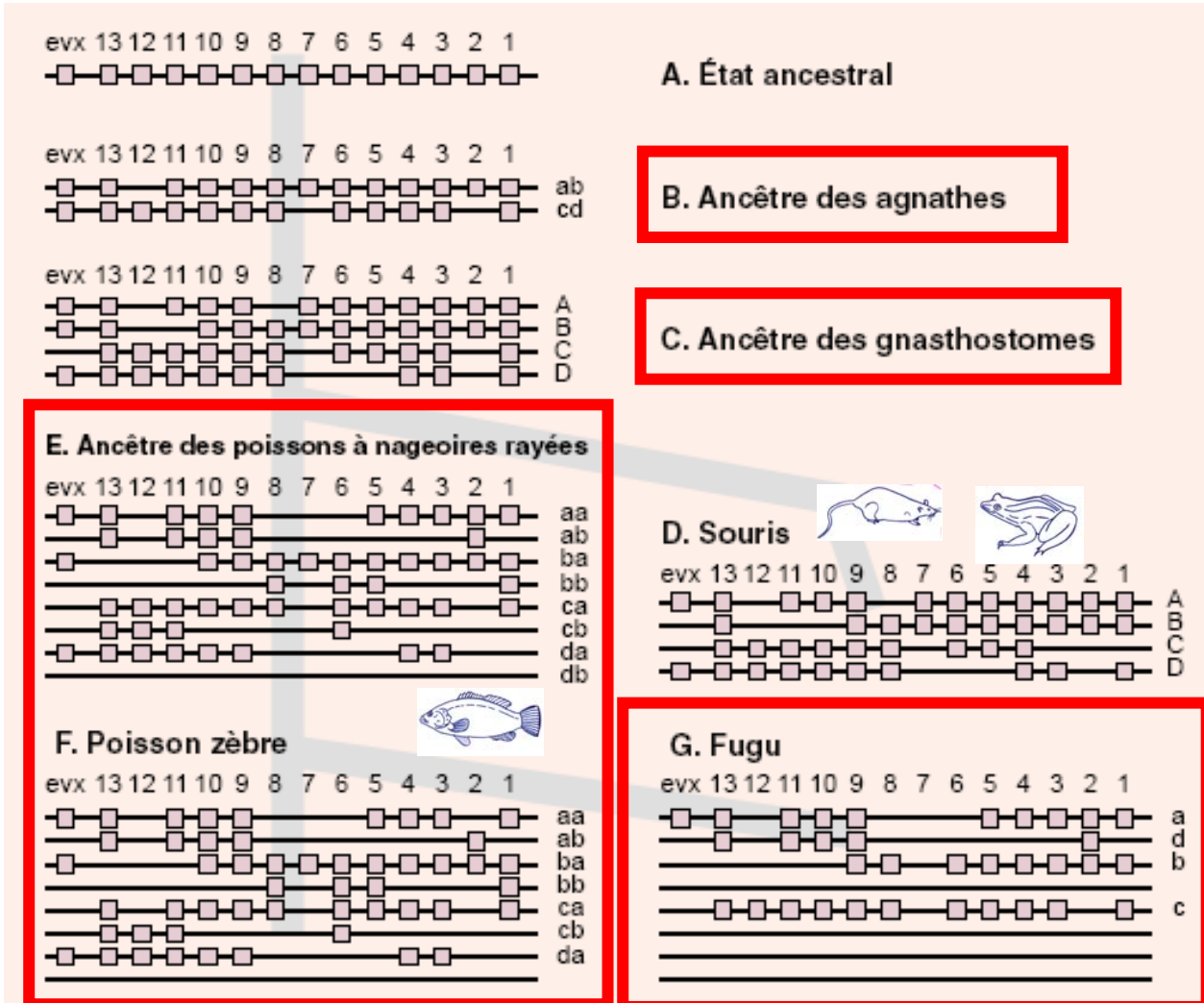


Clusters de gènes hox chez les chordés

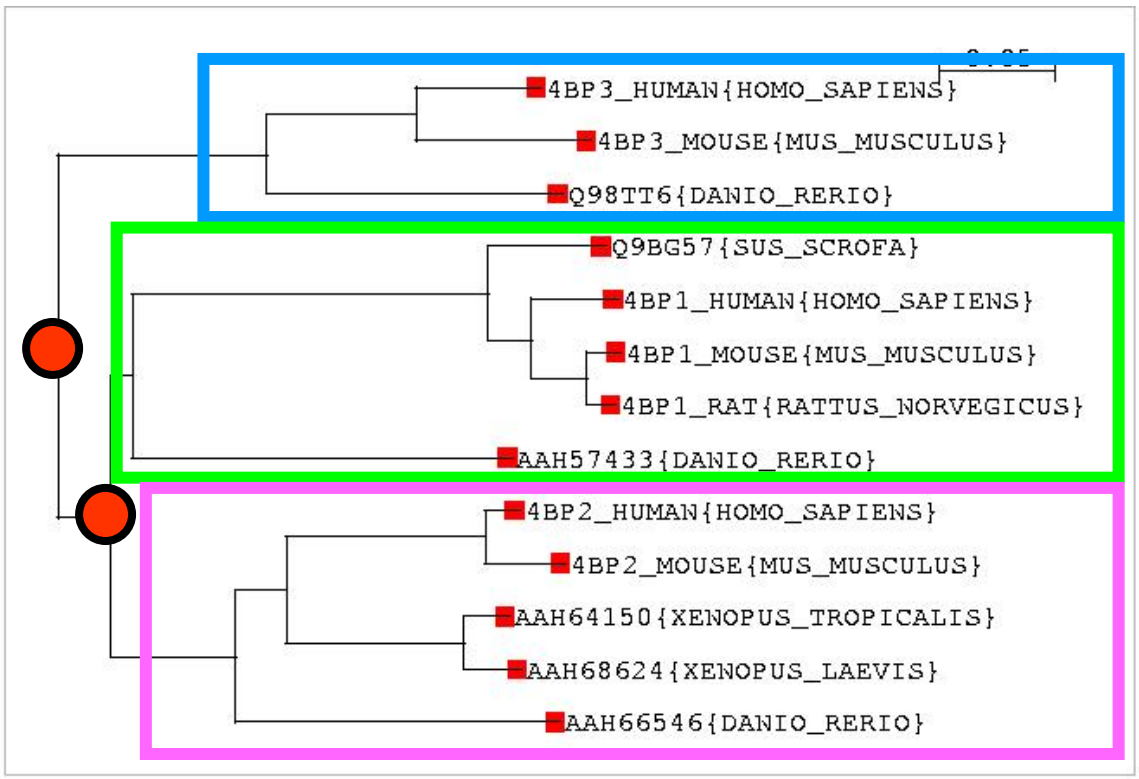
1ère duplication

2ième duplication

Duplications supplémentaires chez certains poissons à nageoires rayonnées

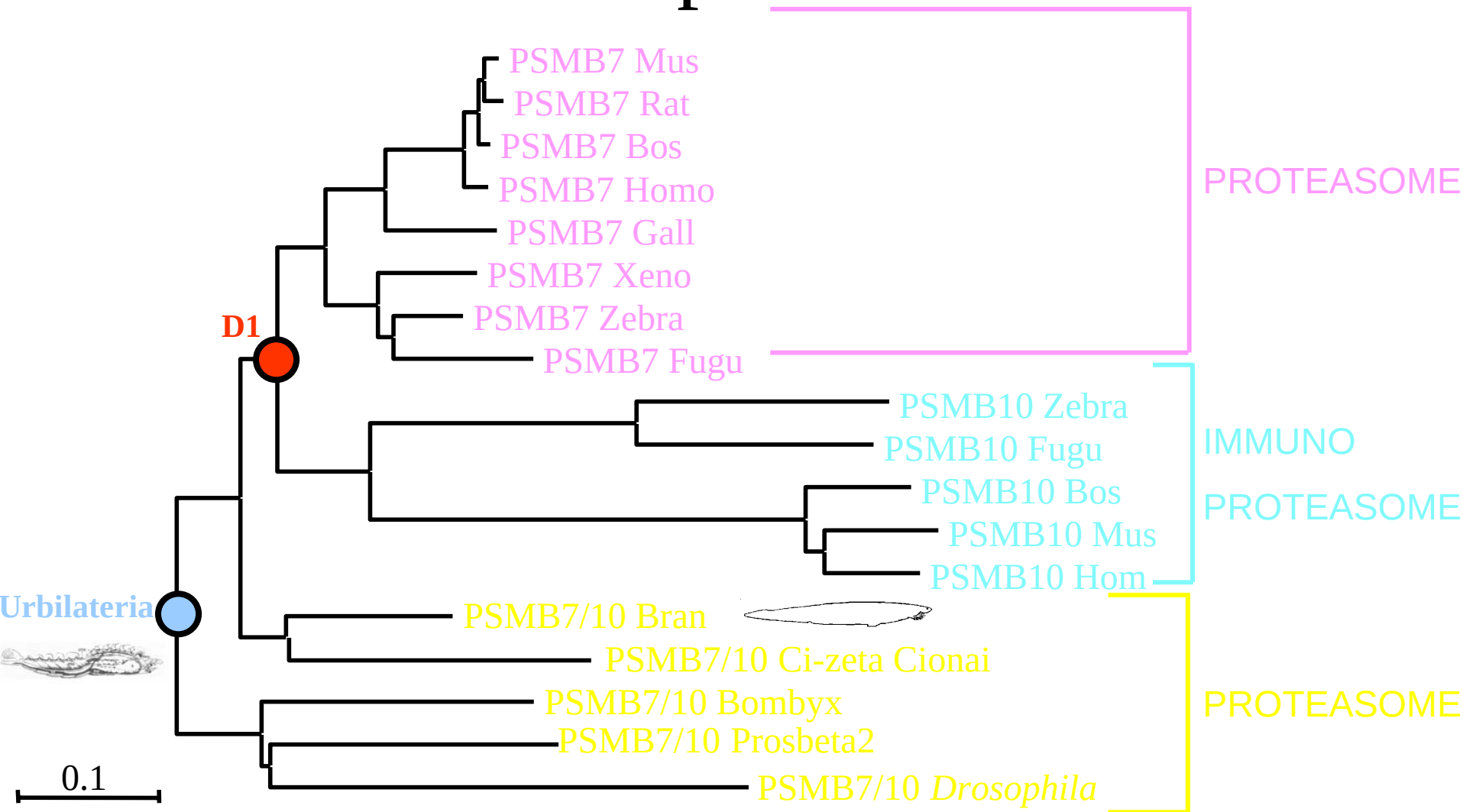


Exemple de duplications

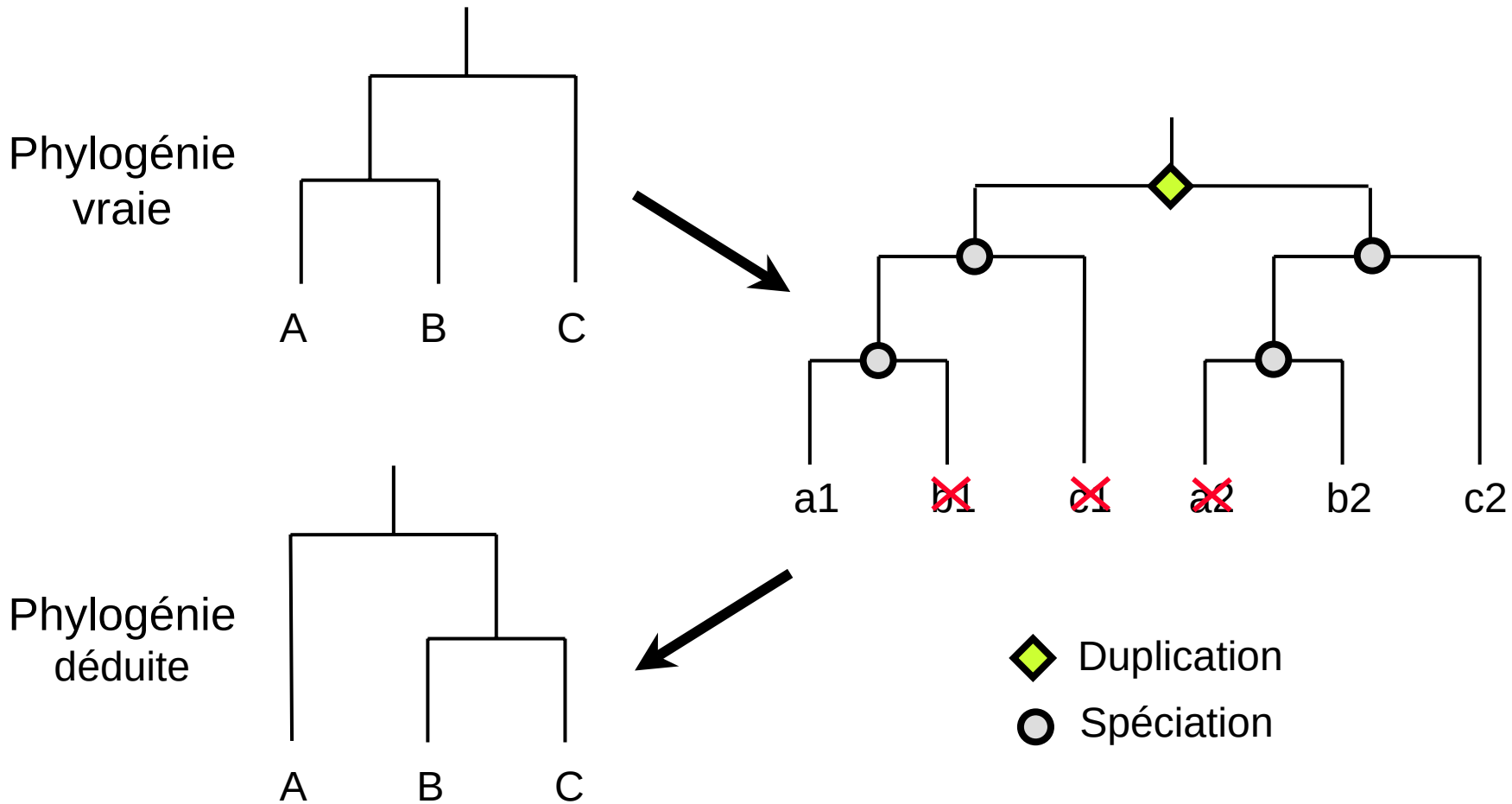


Metazoan translation initiation factor 4E binding protein 1

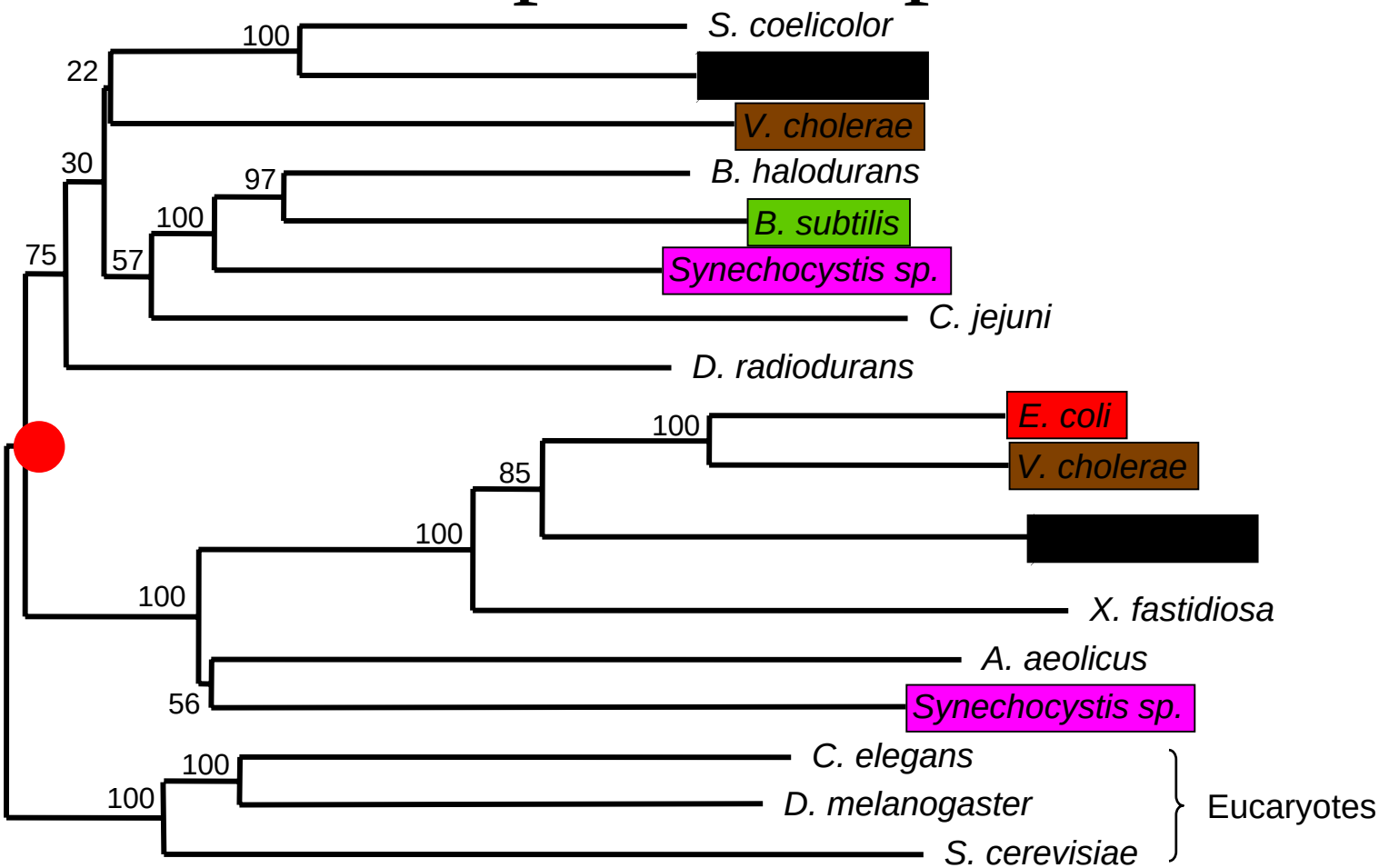
Gènes du Protéasome / Immunoprotéasome



Séquences biologiques : Paralogues et phylogénies

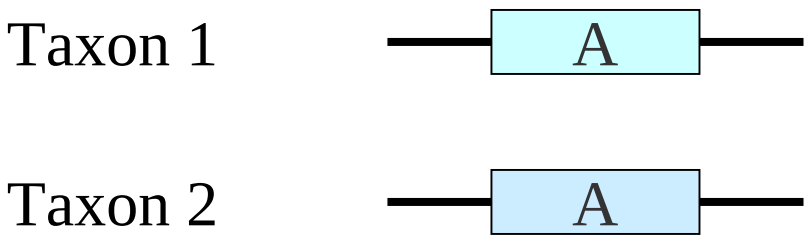


Séquences biologiques : Un exemple de duplication

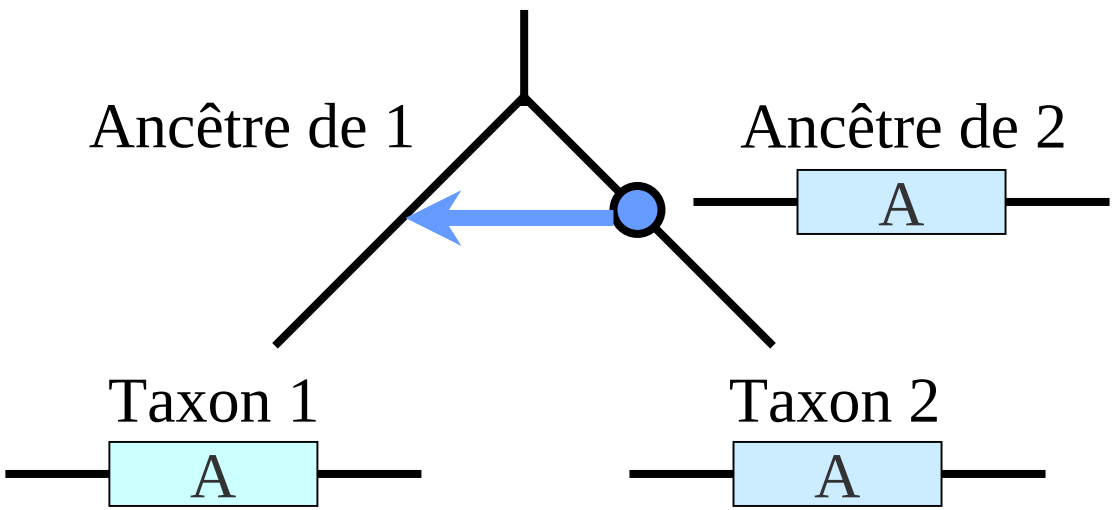


Grande sous-unité de la glutamate synthase

Séquences biologiques : Xénologie

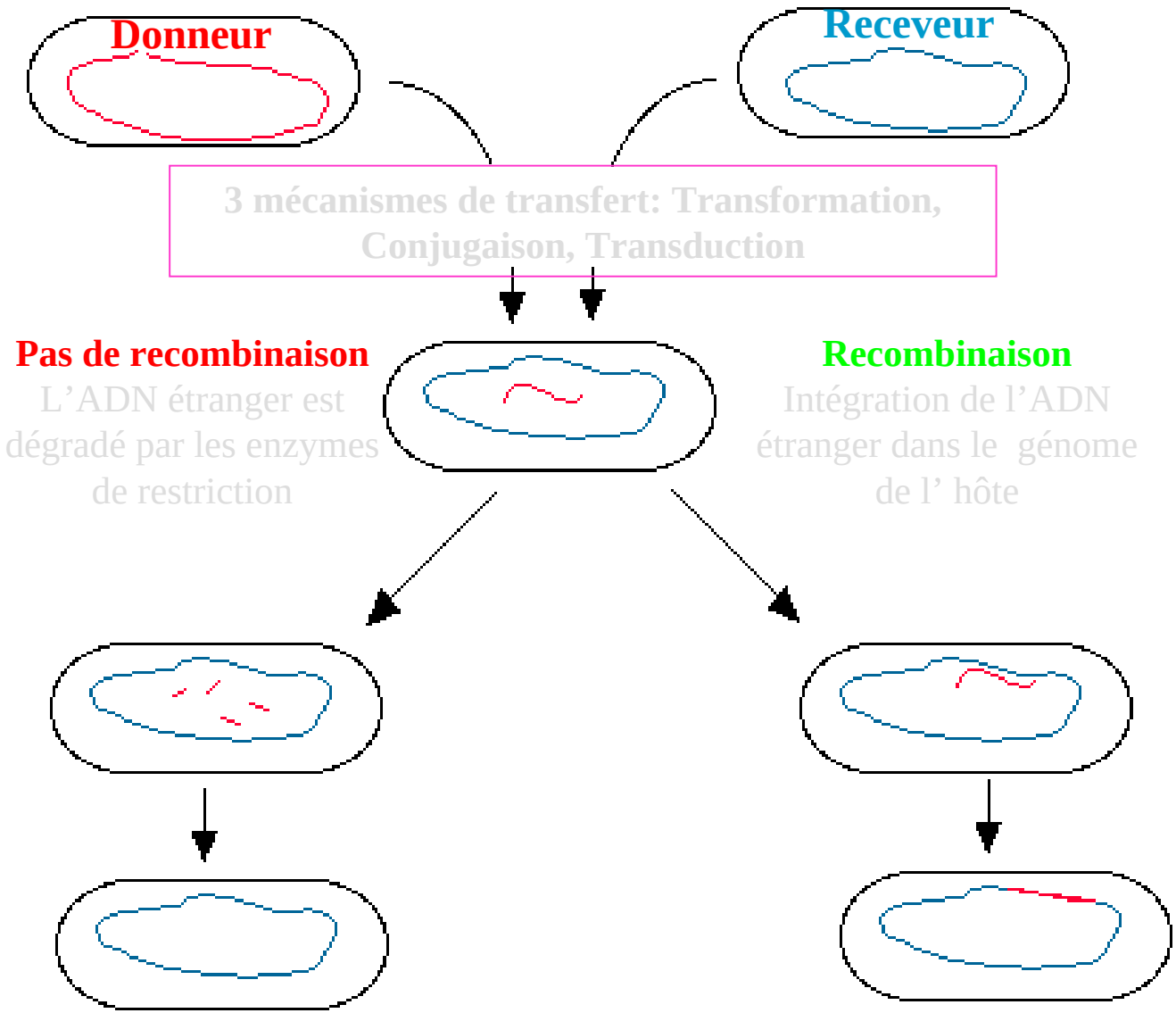


- Définition « gènes ayant été acquis par transfert horizontal »

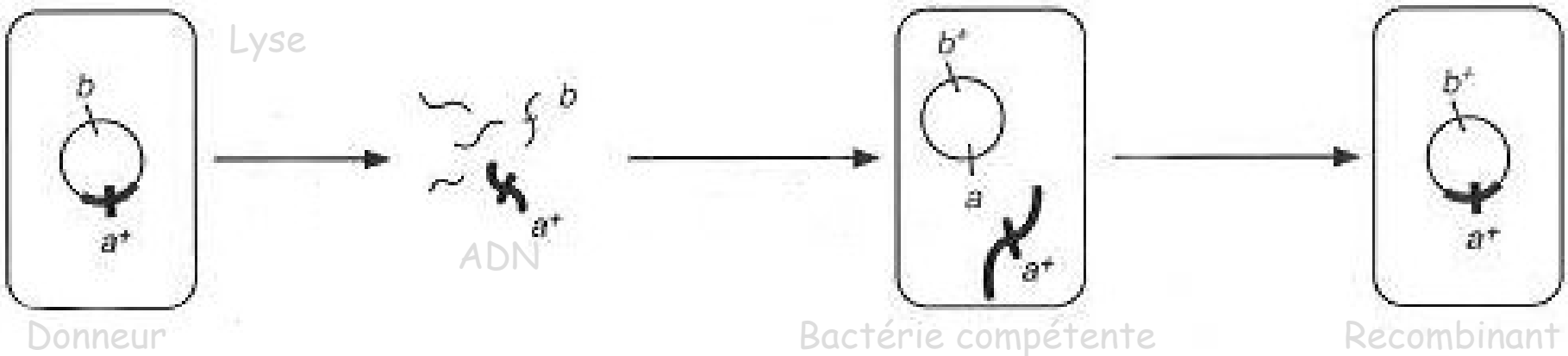


Transfert du gène A d'un ancêtre de 2 vers un ancêtre de A

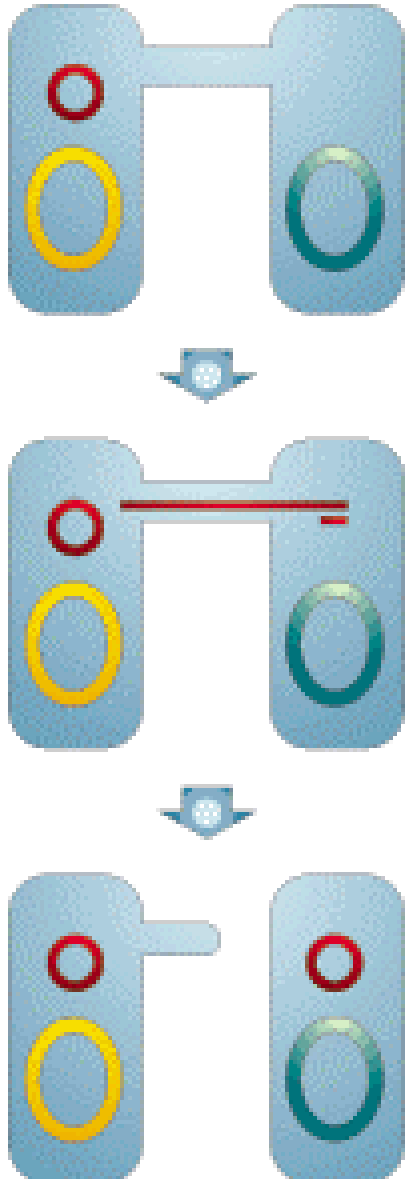
Transferts de gènes



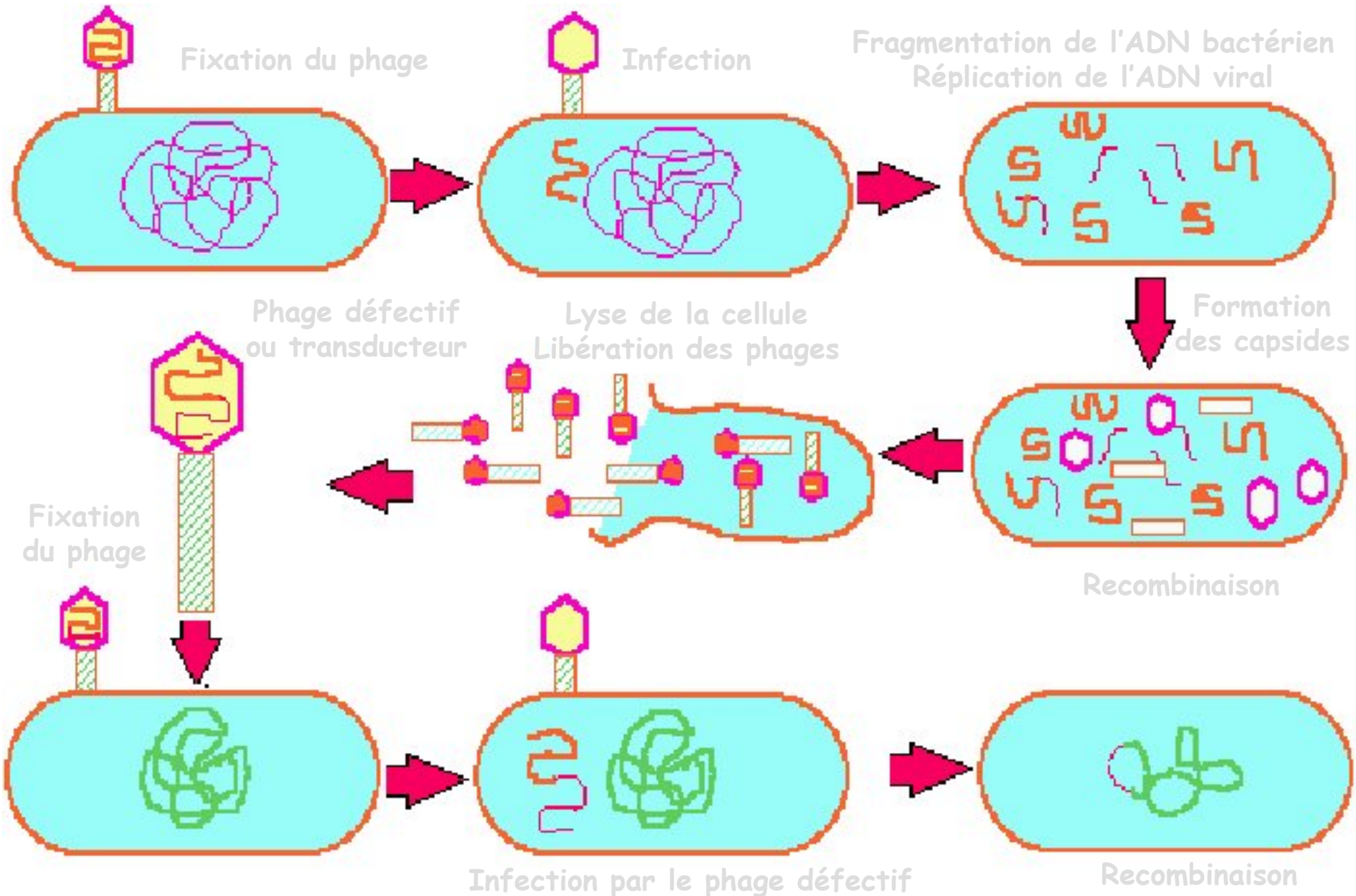
La transformation - Expériences de Griffith (1928)



La conjugaison (Lederberg et Tatum 1946)



La transduction

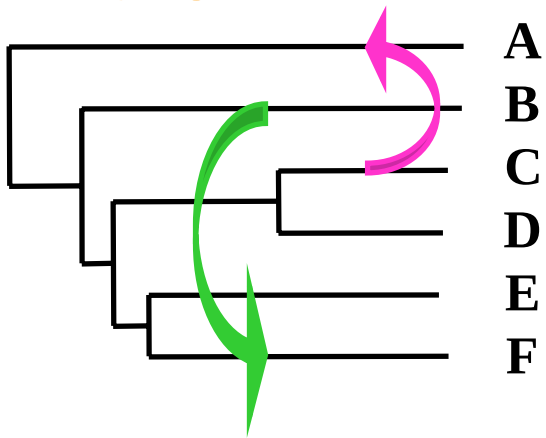


Quantification des transferts horizontaux

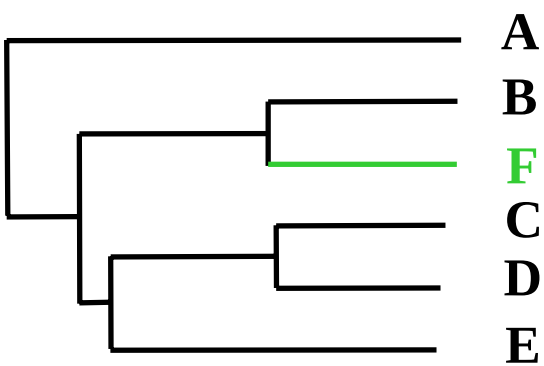
- Comparaison du contenu de génomes proches
  - 20% du génome de *Salmonella* est d'origine exogène récente (Lan et Reeves 1996)
  - Très peu de transferts chez *Chlamydia* ou *Rickettsia* (Kalman et al. 1999, Ogata et al. 2001)
- Recherche de similitudes par BLAST
  - 24% des gènes de *Thermotoga maritima* et 16% des gènes d'Aquifex aeolicus seraient d'origine archébactérienne (Aravind et al. 1998, Nelson et al. 1999)
- Biais de composition
  - 18% du génome d'*Escherichia coli* K-12 aurait été acquis par transfert depuis la divergence avec Salmonella (Lawrence et al. 1998)

Incongruence causée par les transferts horizontaux

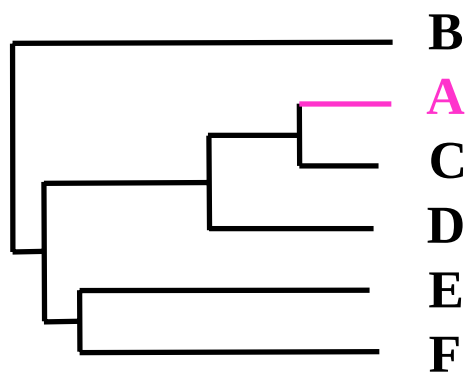
Phylogénie des espèces



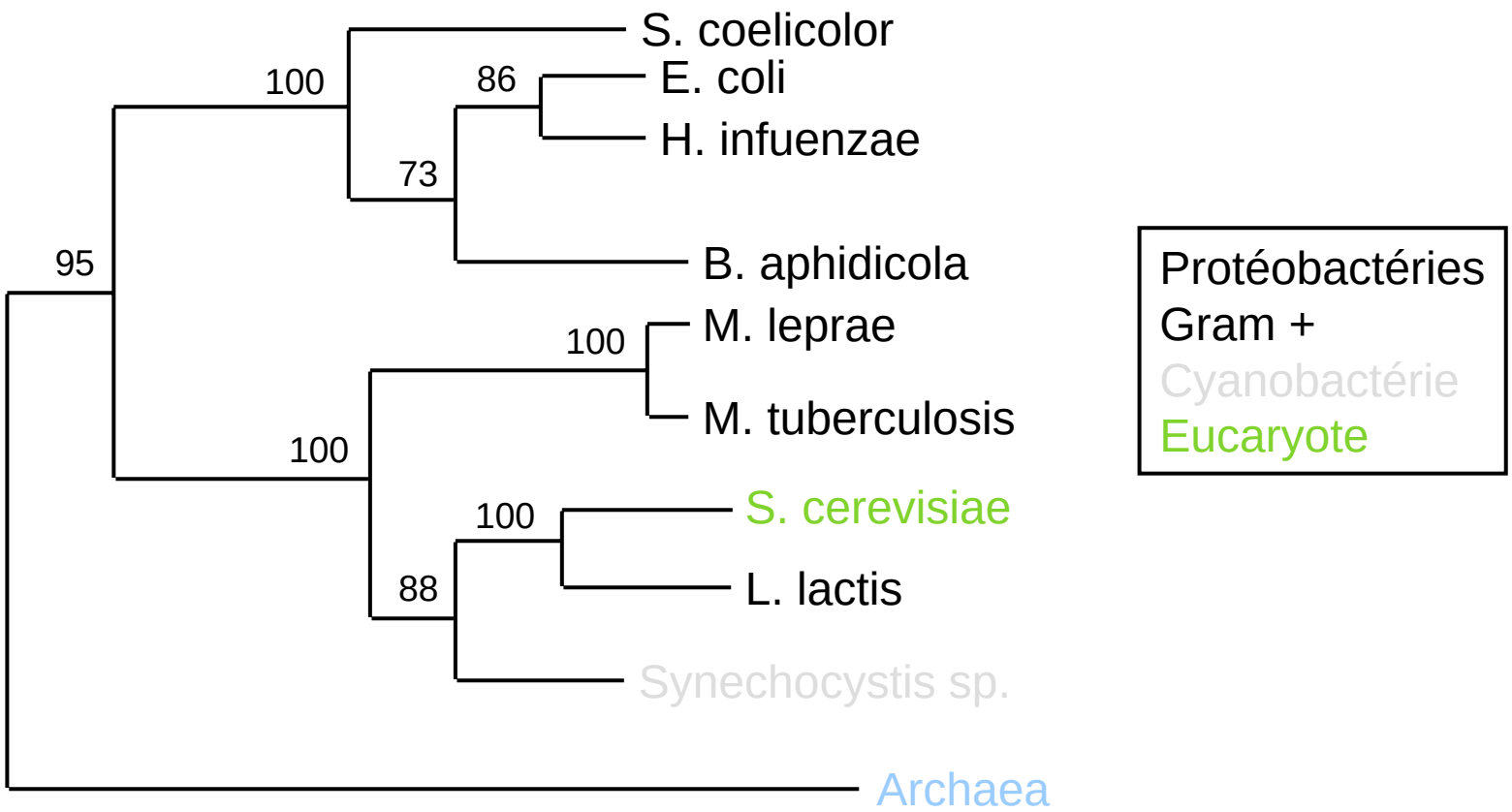
Phylogénie basée sur le gène 1



Phylogénie basée sur le gène 2



Cas du gène *ilvD*

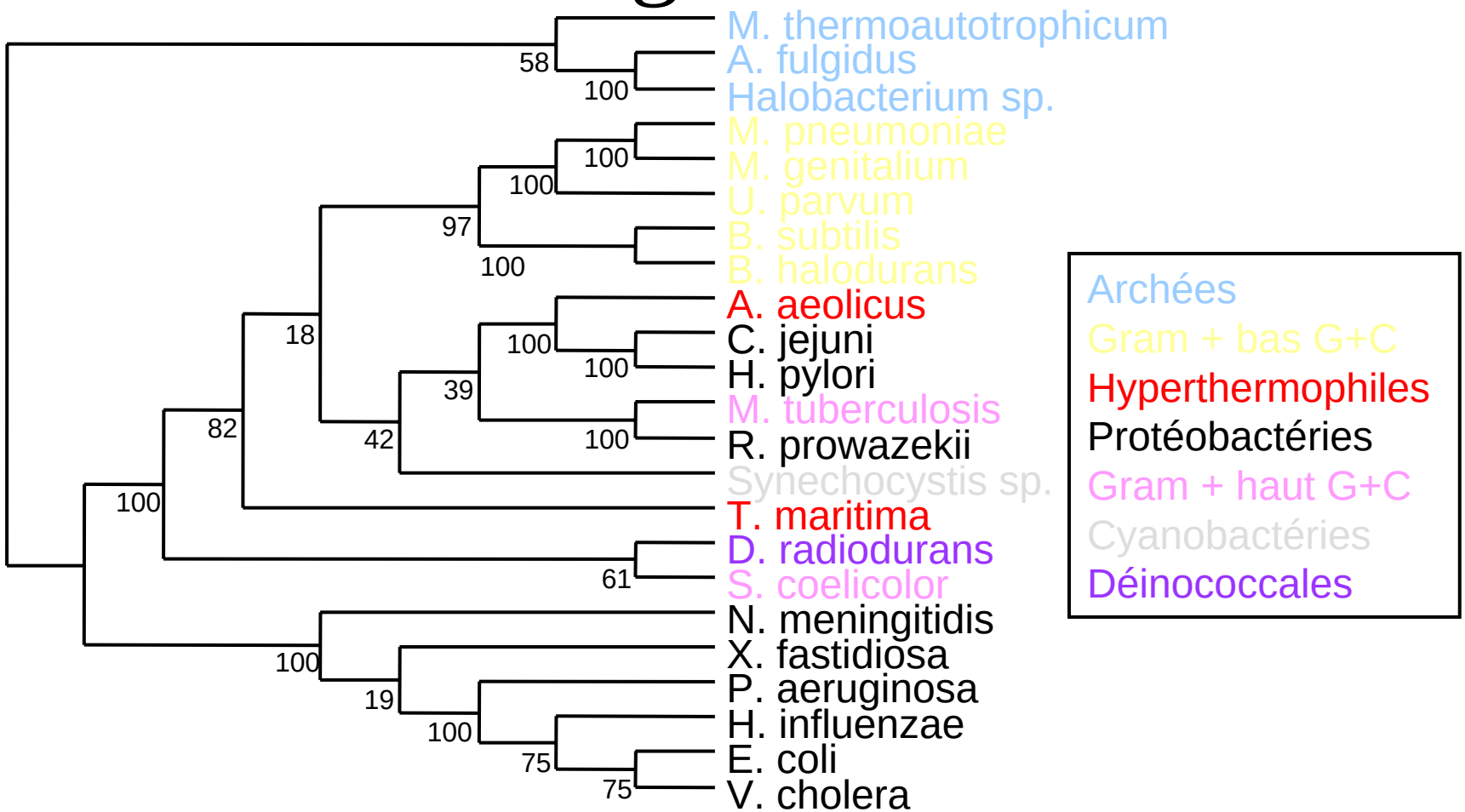


Famille de la Dihydroxy-acid deshydratase

Alignement des gènes *ilvD*

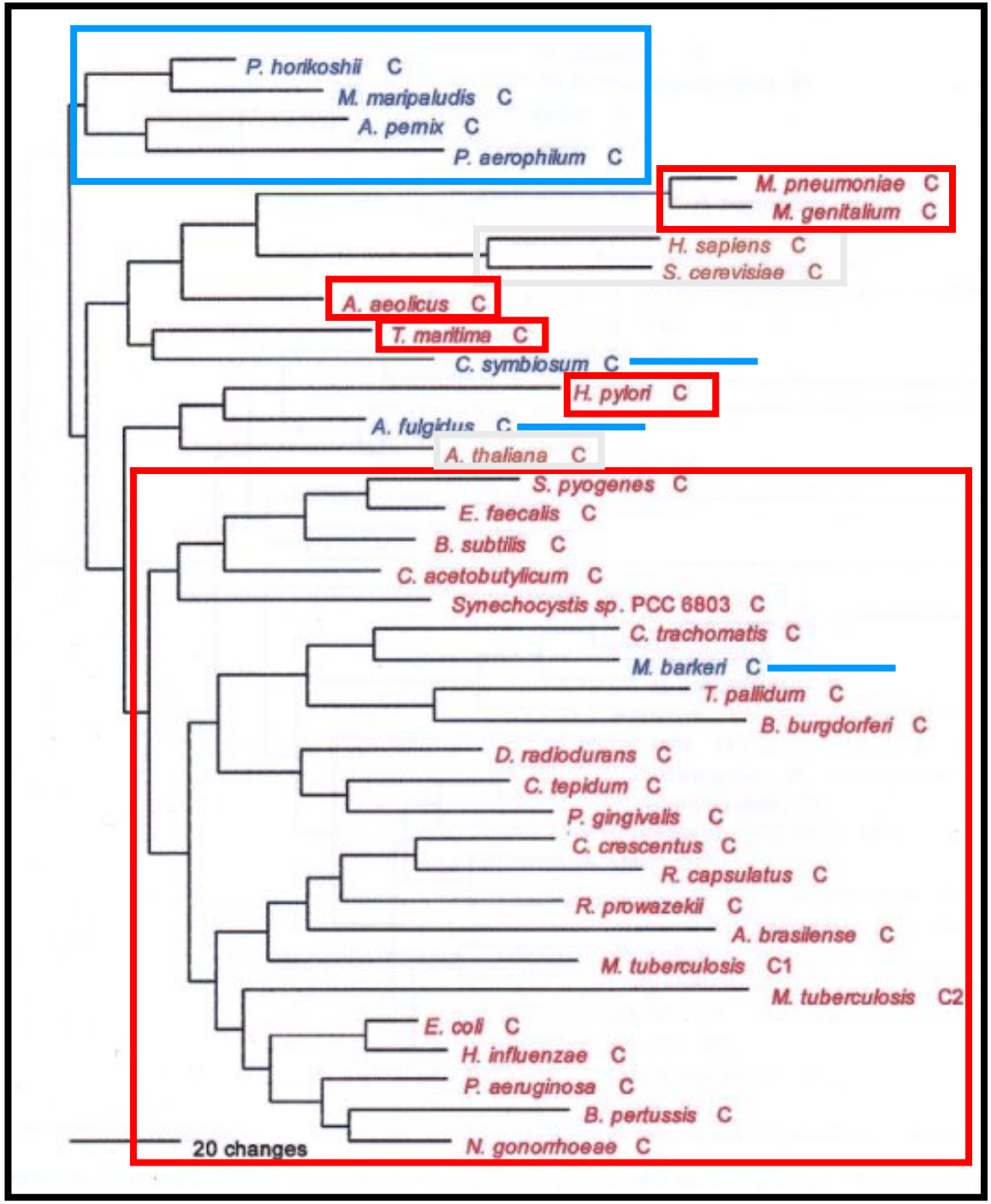
```
M. tuberculosis     GMSLPGSAAPPATDRRRDGFARRSGQAVVELLRR-------GITARDILTKEAFENA/…/
M. leprae           GMSLPGSAAPPATDRRRDGFARRSGQAVIELLRR-------GITARDILTKEAFENA/…/
L. lactis           GMSLPYSSSNPAVSQEKQEECDDIGLAIKNLLEK-------DIKPSDIMTKEAFENA/…/
S. cerevisiae       GLTIPNSSSFPAVSKEKLAECDNIGEYIKKTMEL-------GILPRDILTKEAFENA/…/
Synechocystis sp.   GMSLPYSSTMAAVDGEKADSTEESAKVLVEAIKK-------QILPSQILTRKAFENA/…/
H. influenzae       GLSLPGNGSMLATHADRKELFLKAGRQIVELCKRYYEQDDASVLPRSIGTFDAFENA/…/
E. coli             GLSQPGNGSLLATHADRKQLFLNAGKRIVELTKRYYEQNDESALPRNIASKAAFENA/…/
B. aphidicola       GLSLPGNGTLLATHIDRKKLFKKSARNIVKITKDYYLNNNKNVLPRNIANKESFENA/…/
S. coelicolor       GLSLPGNGSVLATHTARKTLYENAARTVLDLTRRYYEQDDDSVLPRNIATPAAFGNA/…/
                    *:: *  .   *    :     . :  : .  .           :  *    :* **

M. tuberculosis     AENLAAITPPD-----------------------------------PDGK-VLRALANPI
M. leprae           AENLASIAPPD-----------------------------------PDGQ-VIRTLHNPI
L. lactis           AENVETALDLD-----------------------------------FDSQDIMRPLKNPI
S. cerevisiae       AERAKKAPSLP-----------------------------------EGQEIIKPLSHPI
Synechocystis sp.   AEVLADIPDQP-----------------------------------PAGQDVIHSWDDPV
H. influenzae       GEQLDQYDIIR-NQDEELHKFFRAGPAGIRTTQAFSQDCRWDTVDNDRVNGCIRNKENAI
E. coli             PQTLEQYDVML-TQDDAVKNMFRAGPAGIRTTQAFSQDCRWDTLDDDRANGCIRSLEHAY
B. aphidicola       EKTLKKYDILS-TKNKNVIKMFHAGPGGNRTIKPFSQNYRWNKLDKDRVNGCIRSHENAY
S. coelicolor       ADWLKTWDVRGGSPSKEAVELWHAAPGCVRSAEAFSQSERWDTLDEDAEGGCIRSVEHAY
                     .                                                 :.   .:
```

Cas du gène *metG*



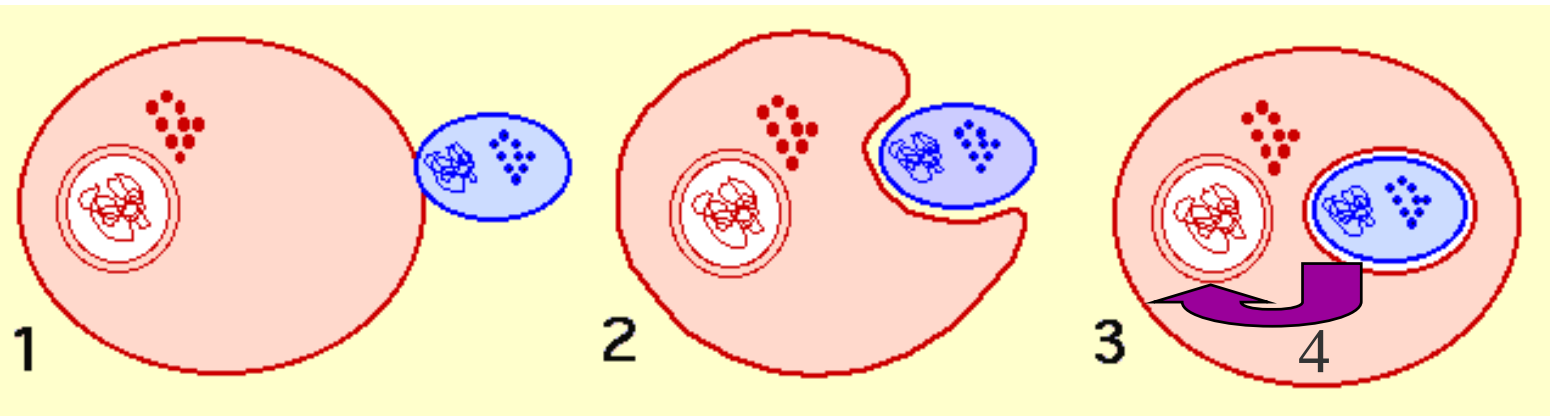
Famille de la Méthionyl-tRNA synthétase

Cystéine ARNt synthétase



L’origine des mitochondries

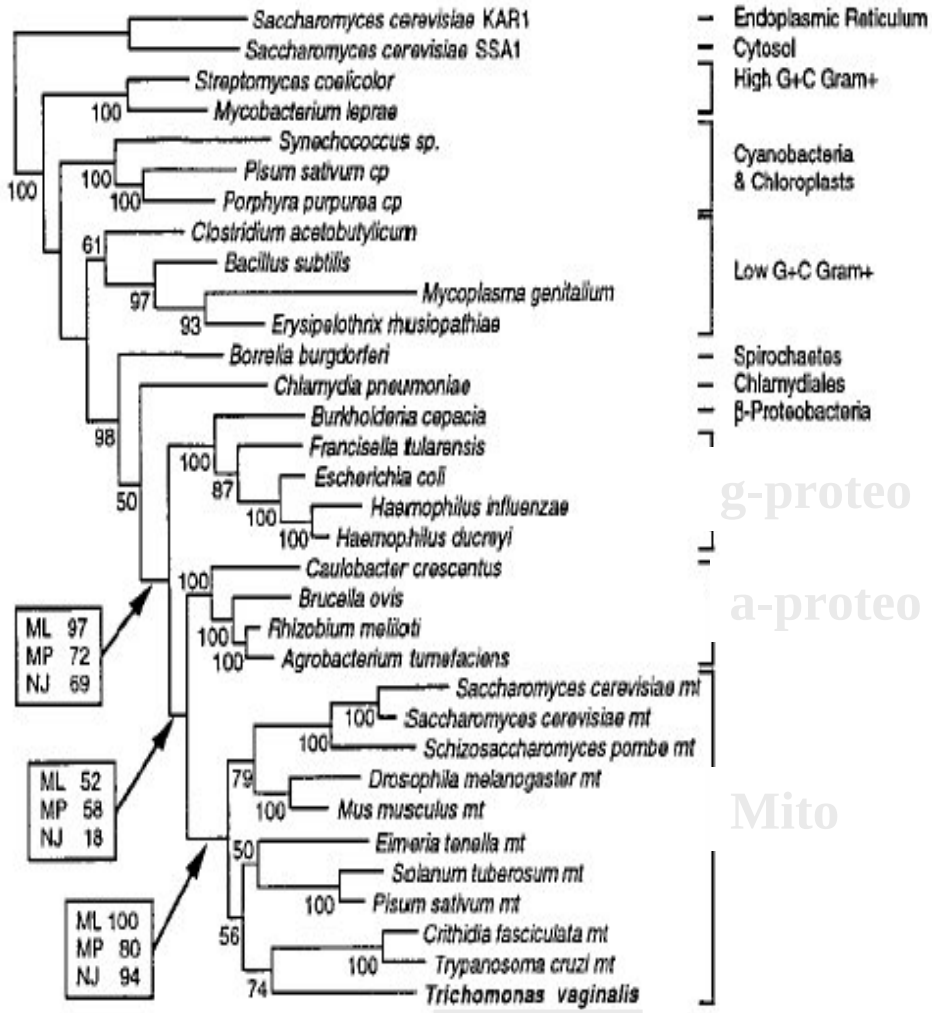
- Les études réalisées à partir des séquences des génomes de mitochondries penchent en faveur d'un seul événement d’endosymbiose qui a probablement eu lieu très tôt dans l'évolution de la cellule Eucaryote
- Plus précisément, les séquences d'ARNr 16S de mitochondries se placent à l'intérieur des a-Proteobacteria, à proximité des Rickettsiales
- Cependant, l'identité de la Protéobactérie la plus proche de celle qui fut à l'origine de toutes les mitochondries reste encore incertaine



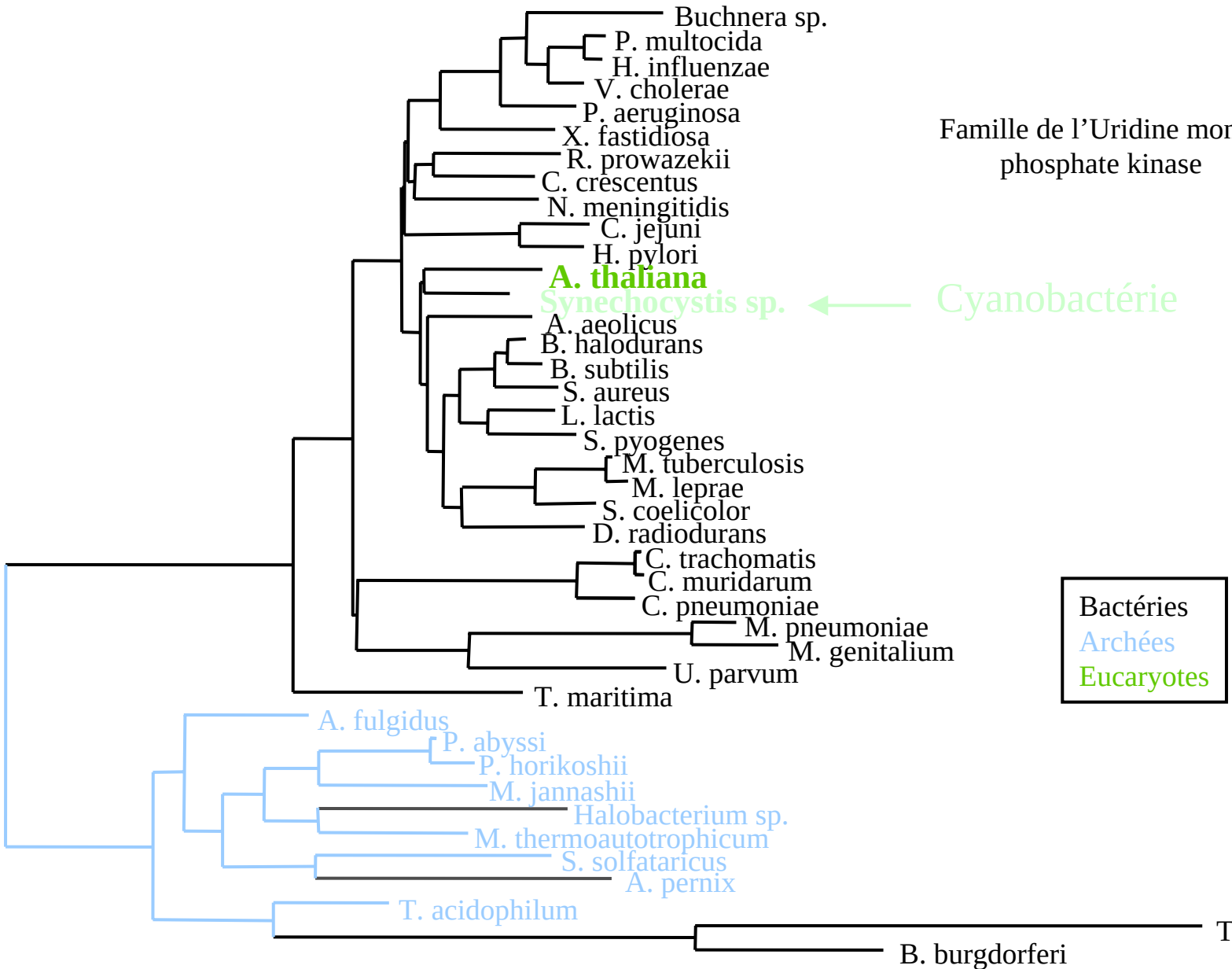
(1) et (2) Phagocytose d’une α-Protéobactérie par une cellule eucaryote
(3) Maintenance de l’α-Protéobactérie dans la cellule eucaryote et établissement d’une relation symbiotique
(4) Réduction du génome de l’α-Protéobactérie par perte de gènes et transfert de gènes vers le noyau

Transfert de gènes des mitochondries vers le noyau

- Le gène codant pour la protéine HSP70 est situé dans le noyau des eucaryotes
- Les phylogénies moléculaires les placent à proximité des a-protéobactéries
  - TF de gènes des mitochondries vers les noyaux
- *Trichomonas vaginalis*



Cas du gène *pyrH*

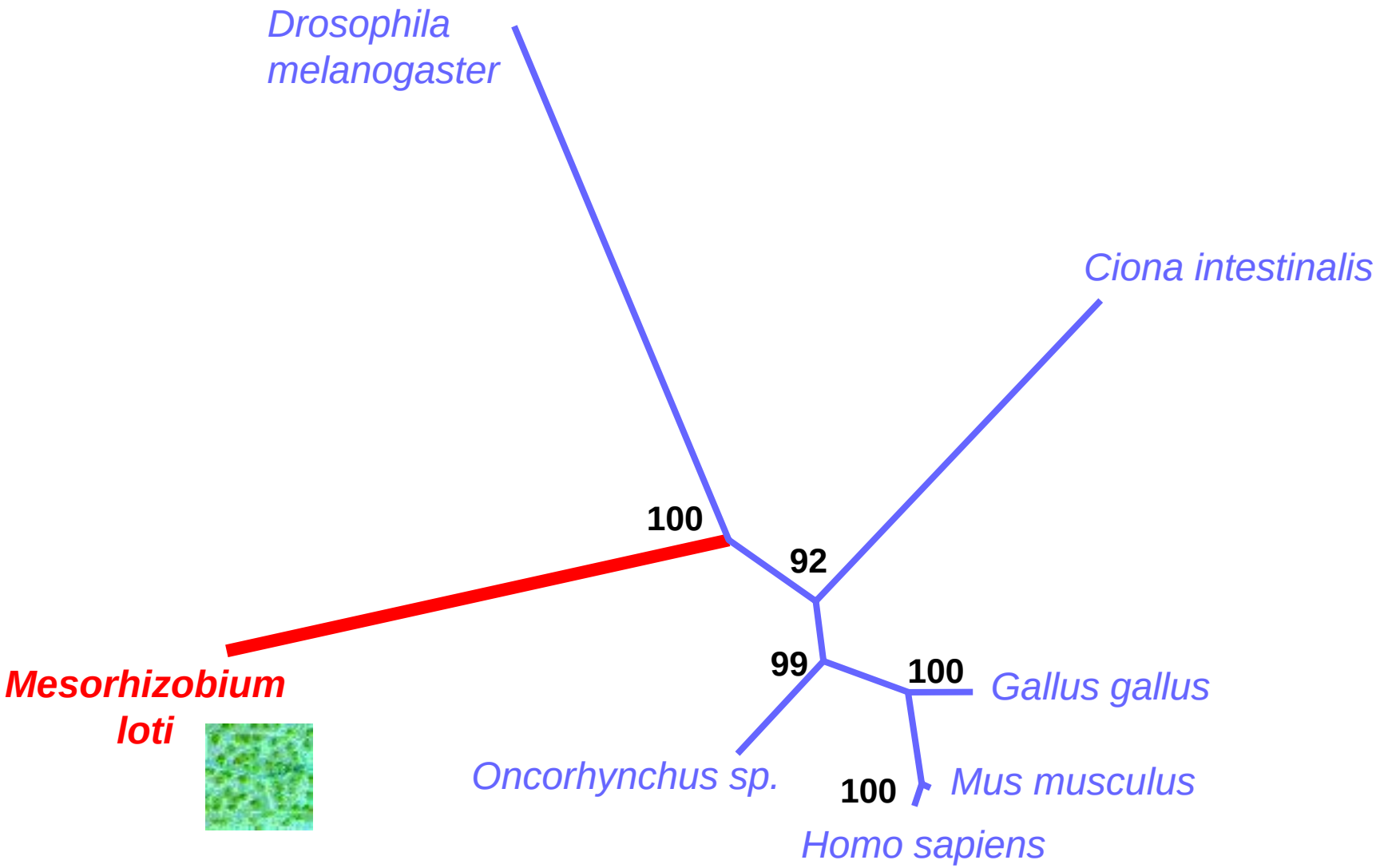


Famille de l'Uridine mono-phosphate kinase

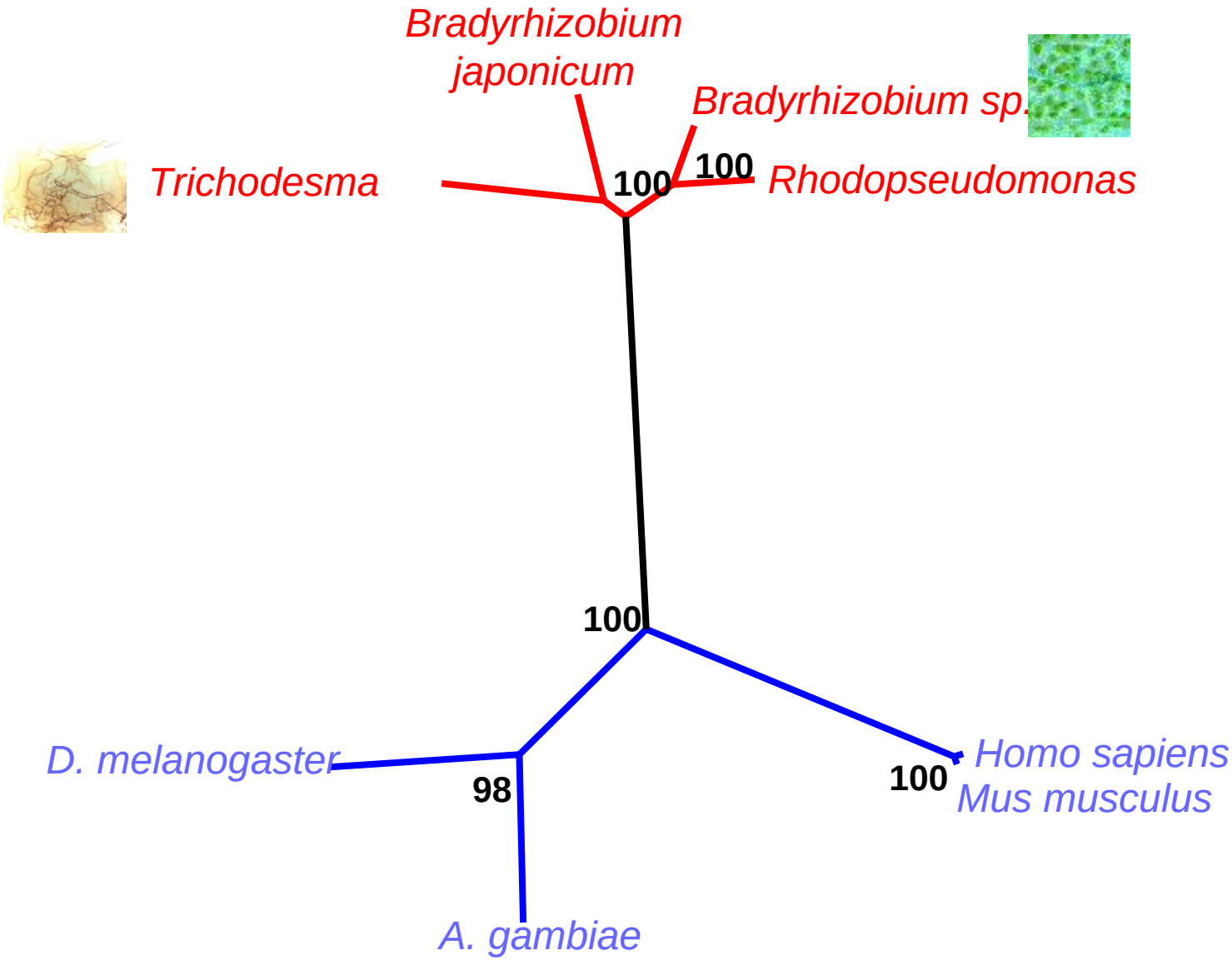
Cyanobactérie

Bactéries
Archées
Eucaryotes

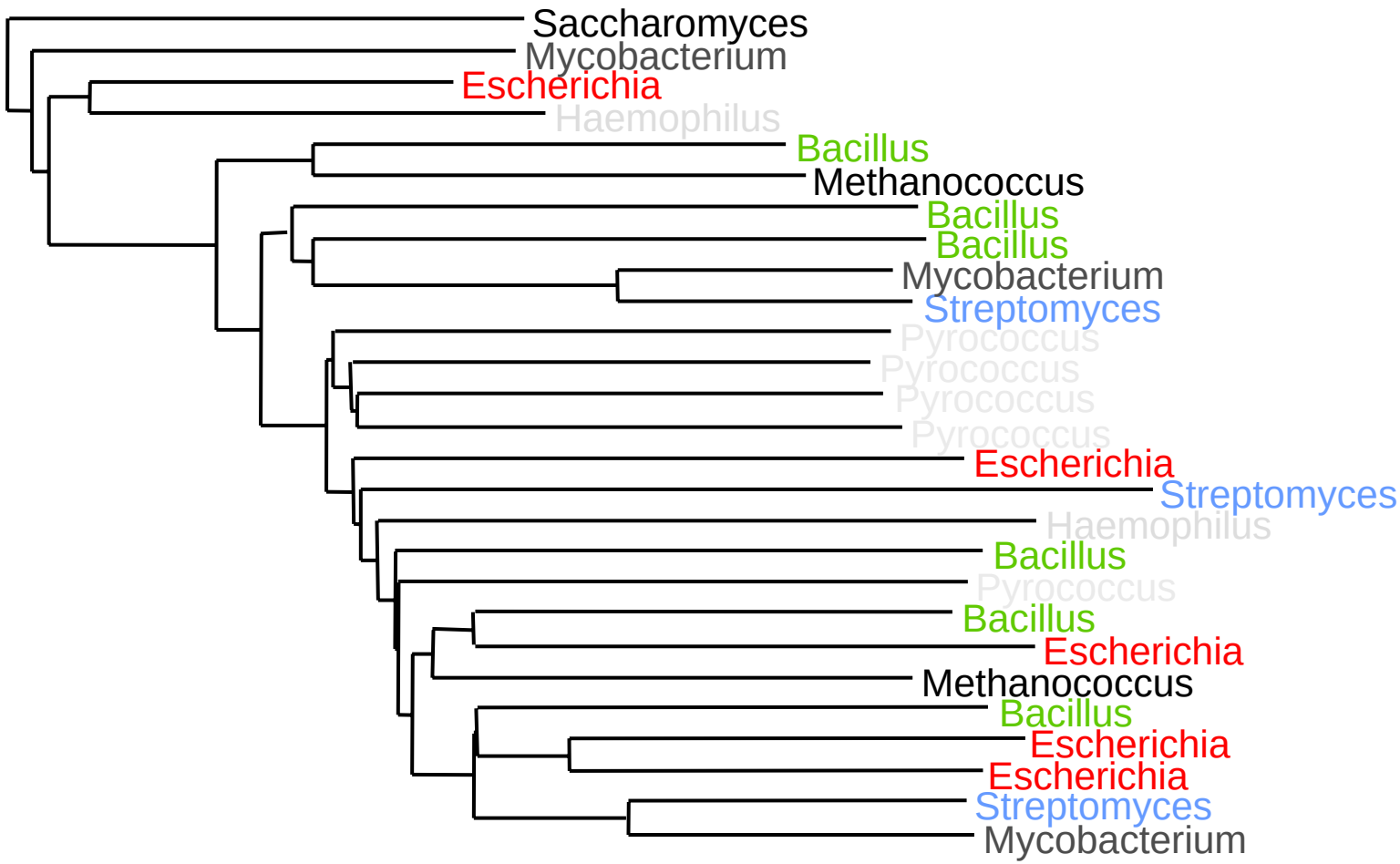
Cas de LUMA (gène codant une protéine du pore nucléaire)



Cas de NURIM (gène codant une protéine du pore nucléaire)



Séquences biologiques : Duplications multiples + transferts



Aminotransférases pyridoxal-phosphate dépendantes (III)

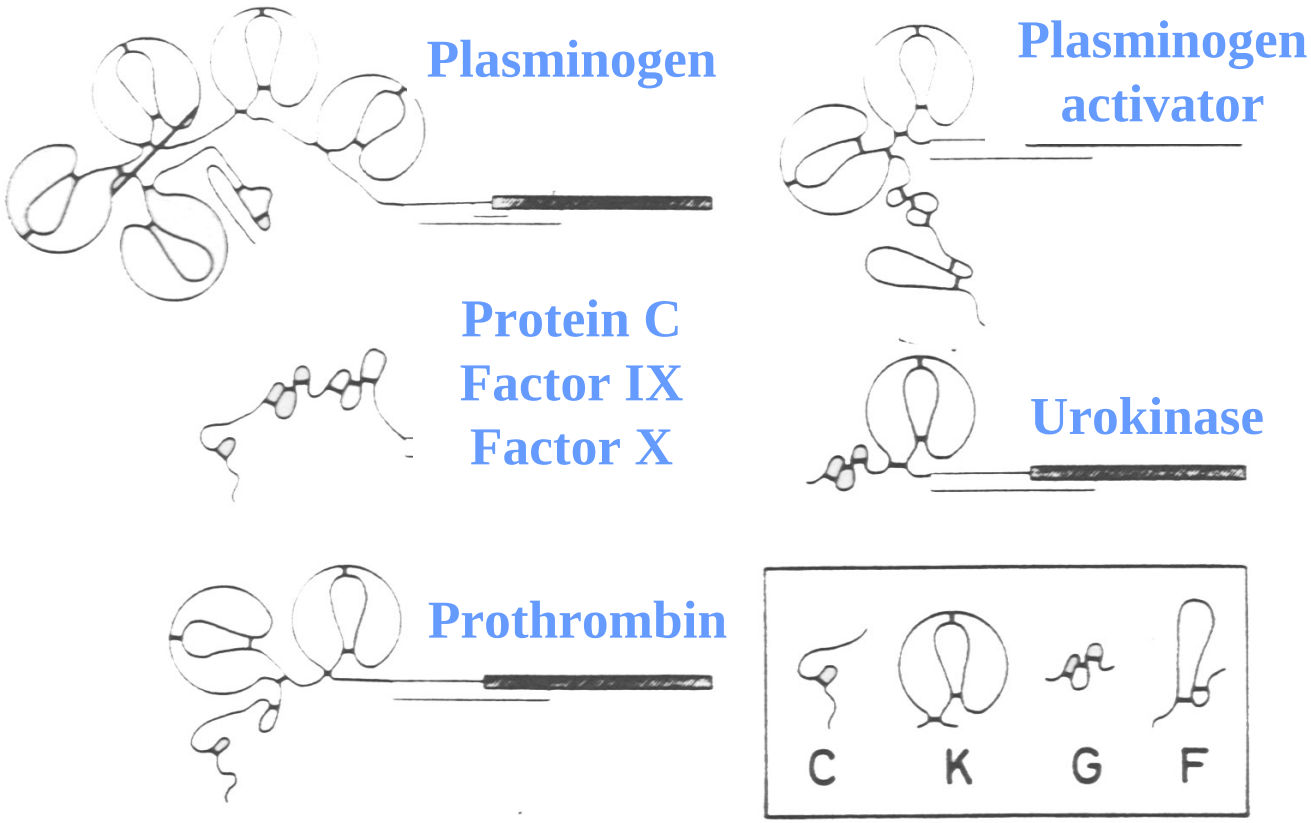
Séquences biologiques : Structure modulaire des protéines

- Les protéines sont composées de domaines fonctionnels

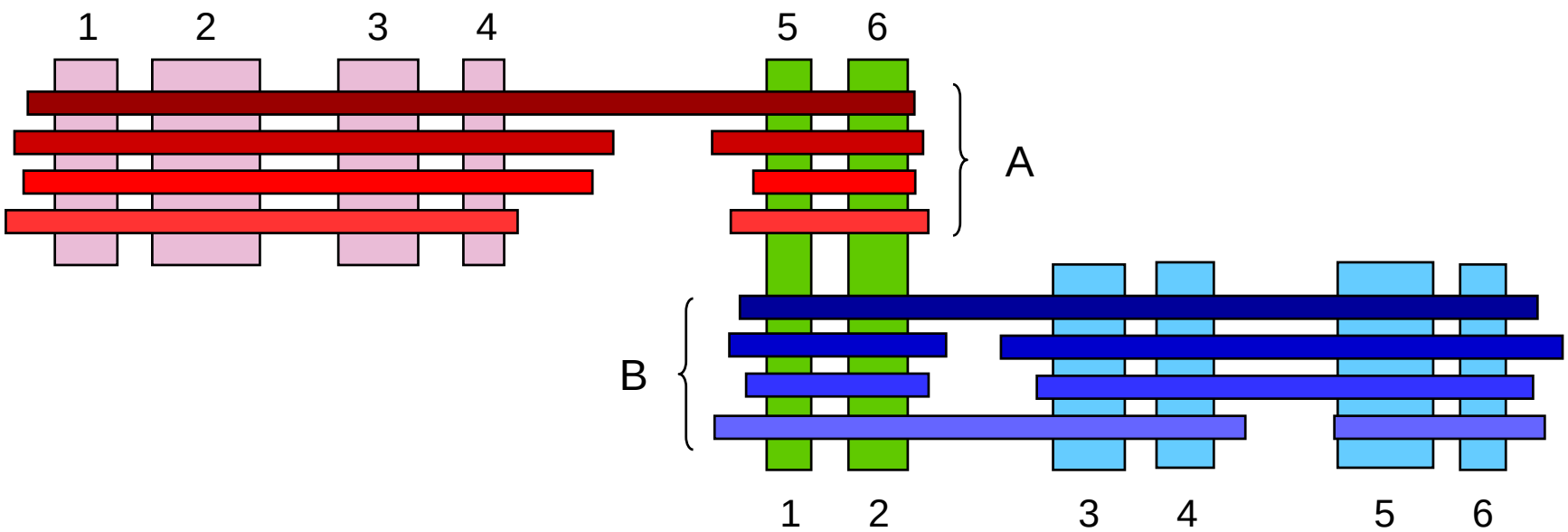
  => Ajout ou perte de domaines au cours de l'évolution

  => Création de protéines chimères ayant plusieurs ancêtres non apparentés

Séquences biologiques : Structure modulaire des protéines



Le problème des domaines



Les familles A & B ne sont connectées
que par deux domaines communs