Bioinformatique: Phylogénie

Module E2M2 Avril 2009

Laurent Duret

Biométrie et Biologie Evolutive

UMR CNRS n° 5558

(Bat. Grégoire Mendel, 2eme étage)

Université Claude Bernard - Lyon 1

Phylogénie moléculaire

- Point de départ: un ensemble de séquences (ADN ou protéines) homologues alignées

- Résultat du processus: un arbre décrivant les relations évolutives entre les séquences étudiées = arbre généalogique des séquences = arbre phylogénétique

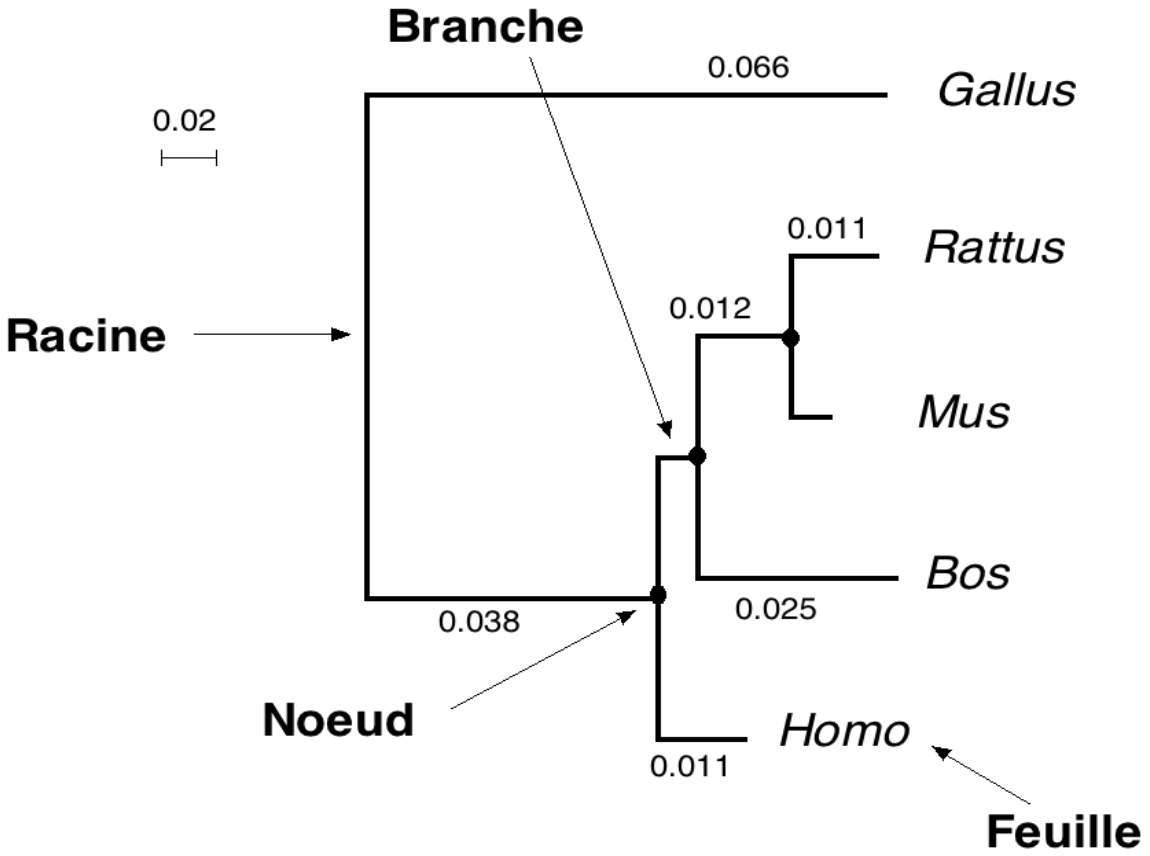
```
CLUSTAL W (1.74) multiple sequence alignment

Xenopus         ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGGTCCAAACAGCGTT---GGCTCTCTA
Gallus          ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCAACATGCAAATG
Bos             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACCCAAAACAGCACCAACGTGCAAATG
Homo            ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Mus             ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACTCAAAACAGCACCAACGTGCAAATG
Rattus          ATGCATCCGCCACCATGACCAGCGGGAGGTAGCTCTCAAAACAGCACCAACGTGCAAATG
                ******  **** *********  *  ***  *   * *** * *             *
```

Applications de la phylogénie moléculaire

- Taxonomie (classification des espèces)
- Identification d'échantillons (diagnostic, contrôle qualité, etc...)
- Détection de tranferts horizontaux
- Etude de l'évolution des gènes (duplications, acquisition de nouvelles fonctions, etc...)
- Epidémiologie (e.g. origine du HIV)
- ...

Arbre Phylogénétique

- Branche interne: entre 2 noeuds; branche externe: entre un noeud et une feuille
- Longueur des branches horizontales est proportionnelle à la distance évolutive entre séquences (nombre de substitution / site)
- Topologie = forme de l'arbre = ordre de branchement



Alignements et gaps

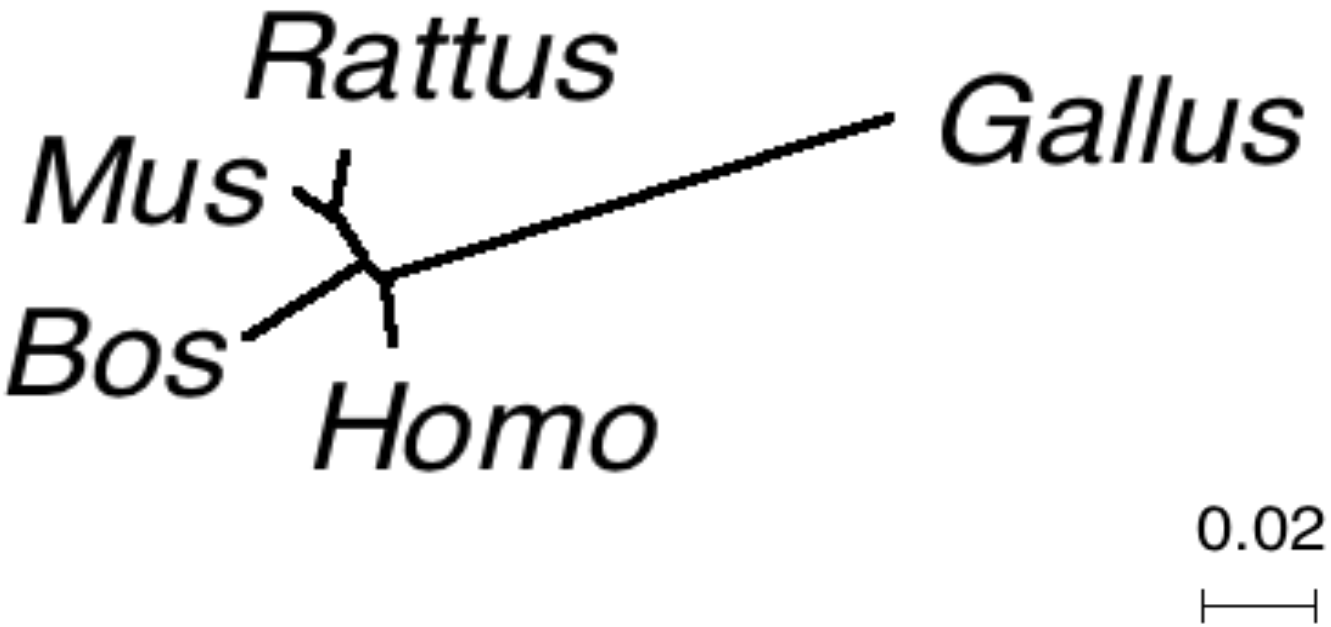
- La qualité de l'alignement est essentielle: chaque colonne de l'alignement (site) est supposée contenir des résidus (nucléotides, acides aminés) homologues, *i.e.* dérivant d'un ancêtre commun
  - les parties non-fiables de l'alignement doivent être omises de l'analyse phylogénétique
- La plupart des méthodes utilisées actuellement se basent uniquement sur l'analyse des substitutions; elles ne prennent pas en compte les indels (gaps)

```
Xenopus     ATGCATGGGCCAACATGACCAGGAGTTGGTGTCggtCCAAACAGCGTT---GGCTCTCTA
Gallus      ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCaacATGCAAATG
Bos         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Homo        ATGCATCCGCCACCATGACCAGCAGGAGGTAGCagtCAAAACAGCACCaacGTGCAAATG
Mus         ATGCATCCGCCACCATGACCAGCAGGAGGTAGCactCAAAACAGCACCaacGTGCAAATG
Rattus      ATGCATCCGCCACCATGACCAGCGGGAGGTAGCtctCAAAACAGCACCaacGTGCAAATG
                                             ---           ---
```

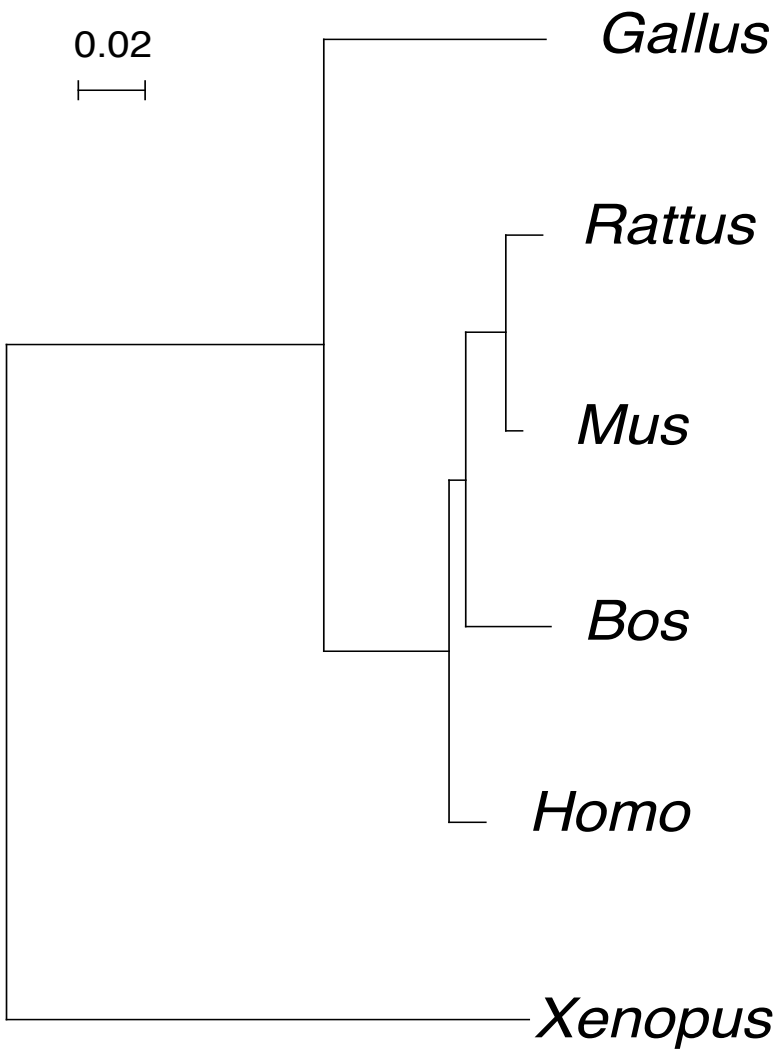
Arbres avec ou sans racine

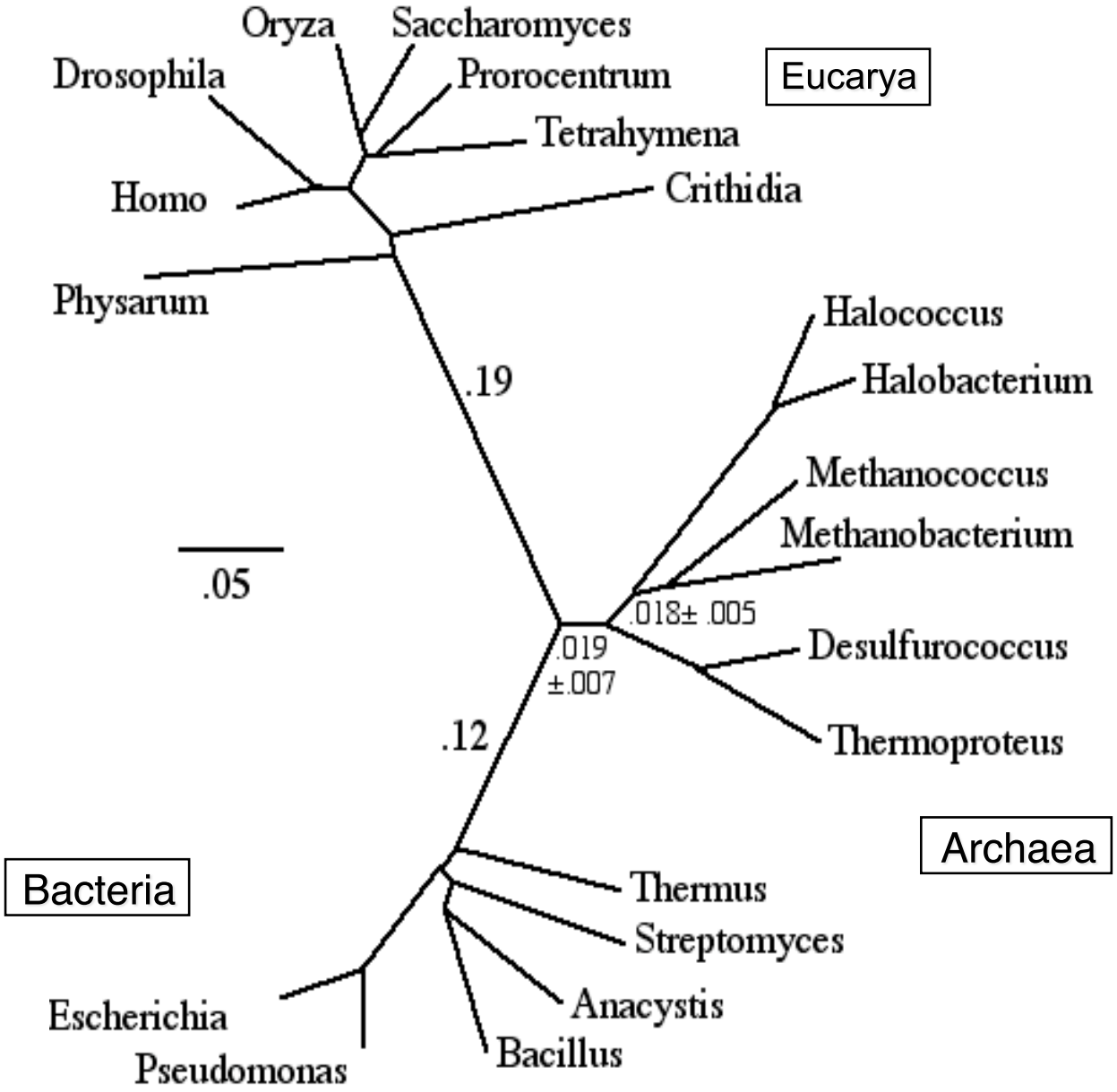
- La plupart des méthodes phylogénétiques produisent des arbres sans racine. C'est parce qu'elles détectent des différences entre séquences mais n'ont aucun moyen de savoir quelle direction ces changements ont suivi au cours du temps
- Deux façons d'enraciner un arbre:
  - Méthode du groupe externe: inclure dans l'analyse un groupe de séquences connues *a priori* comme externe au groupe étudié. La racine se trouvera sur la branche joignant le groupe externe aux autres séquences
  - Faire l'hypothèse de l'horloge moléculaire, *i.e.* que toutes les séquences ont évolué à la même vitesse depuis leur divergence de l'ancêtre commun. La racine se trouvera en un point équidistant de toutes les feuilles de l'arbre.

Arbre non-raciné



Arbre raciné





Phylogenie Universelle

Déduite de l'alignement d'ARNr (2508 sites homologues) méthode NJ, distance Kimura's 2-paramètre.

Nombre d'arbres sans racine possibles pour *n* taxa

$$N_{arbres} = 3.5.7...(2n-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

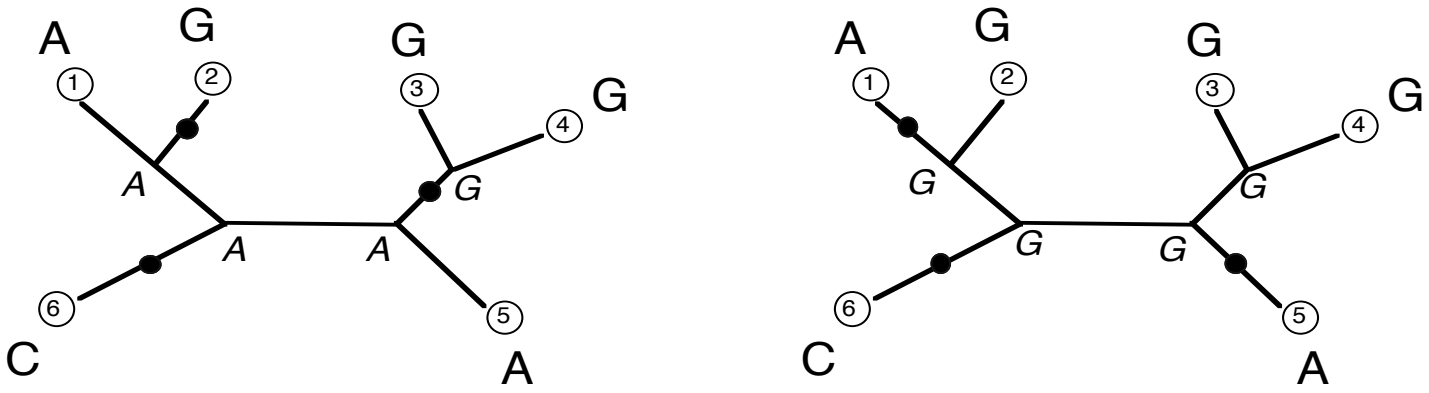
n	N_{arbres}
4	3
5	15
6	105
7	945
...	...
10	2,027,025
...	...
20	~ 2.10^{20}

Méthodes de reconstruction phylogénétique

- Trois principales familles de méthodes:
  - Parcimonie
  - Méthode de distance
  - Méthode du maximum de vraisemblance

Parcimonie (1)

- Etape 1: pour une forme d ’arbre donnée, et pour un site donné de l ’alignement, déterminer quels résidus placés aux nœuds de l ’arbre demandent le moins de changements totaux possible dans tout l ’arbre à ce site. Soit *d* ce nombre minimal de changements (substitutions).



X : nucléotide ancestral ● : événement de substitution

Exemple: à ce site, dans cet arbre, il faut au minimum 3 substitutions pour expliquer les résidus observés. Plusieurs scénarios à 3 changements sont possibles.

Parcimonie (2)

- Etape 2:
  - Calculer *d* (= étape 1) pour chaque site de l 'alignement
  - Sommer les nombres *d* sur tous les sites de l 'alignement
  - Ceci donne la longueur de l 'arbre (*L*)

- Etape 3:
  - Calculer *L* (= étape 2) pour toutes les formes d 'arbres possibles
  - Retenir le(s) arbre(s) le(s) plus court(s) = arbre(s) qui nécessite(nt) le moins de changements = arbre(s) le(s) plus parcimonieux

Quelques propriétés de la parcimonie

- Plusieurs arbres peuvent être également parcimonieux (même longueur, la plus courte possible).
- Le placement des substitutions sur chaque branche n 'est pas défini de manière unique => il n 'y a pas de longueur de branche définie de manière unique en parcimonie
- Le nombre d 'arbres à évaluer croît très vite avec le nombre de séquences traitées:
  - ⇨ La parcimonie peut être très longue en calcul
  - ⇨ La recherche doit souvent être limitée à une partie de l 'ensemble de tous les arbres (heuristique) => il n 'y a pas de certitude mathématique d 'avoir obtenu l 'arbre le plus court

Construction d ’arbres phylogénétiques par méthodes de distance

- Principe général:

Alignement de séquences

⬇ (1)

Matrice de distances évolutives entre paires

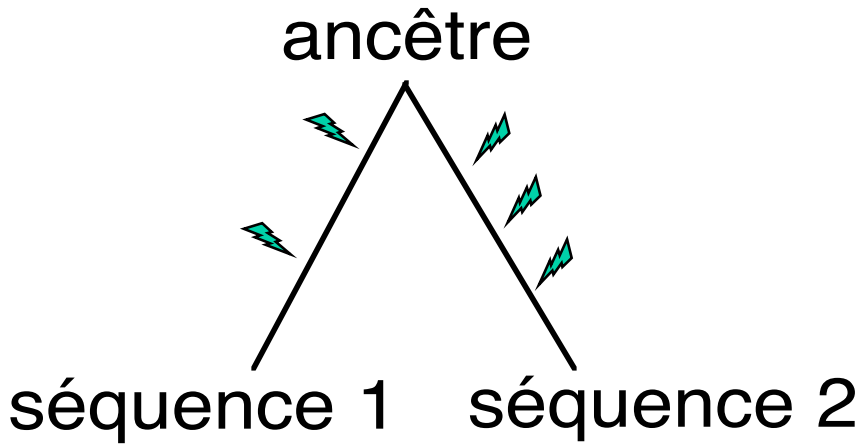
⬇ (2)

Arbre (non raciné)

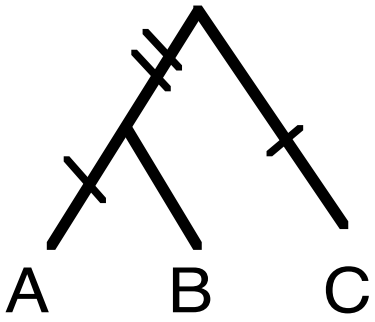
- (1) Mesures de distances évolutives
- (2) Calcul d’arbre à partir d ’une matrice de distance

Distances évolutives

- Elles mesurent le nombre total de substitutions produites sur les deux lignées depuis la divergence de l ’ancêtre commun
- Rapporté à la longueur des séquences
- Exprimées en substitutions / site



Arbre phylogénétique / matrice de distances évolutives entre paires



	A	B	C
A	0		
B	1	0	
C	4	3	0

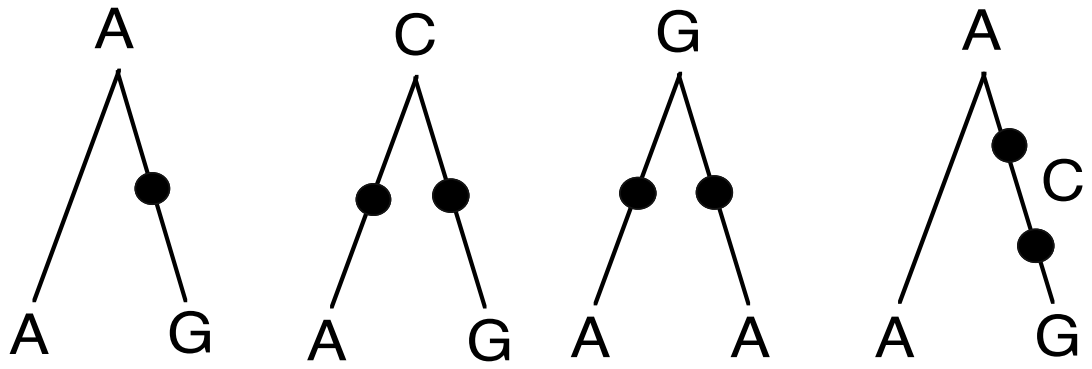
arbre

matrice de distances

Mesure des distances évolutives (1):

Problème des substitutions multiples ou cachées

- *D* (distance évolutive réelle) ≥ fréquence des différences observées (*p*)



- $D = p$ + différences cachées
- En faisant des hypothèses sur la nature du processus évolutif, on peut estimer *D* à partir des différences observées entre deux séquences. *D* estimée: *d*

Mesure des distances évolutives (2):

Distance de Jukes et Cantor (ADN)

- Hypothèses du modèle:
  - (a) Tous les sites évoluent indépendemment et selon le même processus
  - (b) Toutes les substitutions sont équiprobales
- Mesure de la distance évolutive (d) en fonction de la fréquence des différences observées (p):

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

$$V(d) = \frac{9p(1-p)}{(3-4p)^2 N}$$

N = nombre de sites comparés

p	d
0,10	0,11
0,20	0,23
0,40	0,57
0,60	1,21

Mesure des distances évolutives (3):

Distance de Poisson (protéines)

- Hypothèses du modèle:
  - (a) Tous les sites évoluent indépendemment et selon le même processus
  - (b) Toutes les substitutions sont équiprobales
- Mesure de la distance évolutive (d) en fonction de la fréquence des différences observées (p):

$$d = - \ln(1 - p)$$

- !! Les hypothèses des modèles Jukes-Cantor et Poisson sont très simplificatrices !!

Mesure des distances évolutives (4):

Distance de Kimura à deux paramètres (ADN)

- Hypothèses du modèle:
  - (a) Tous les sites évoluent indépendemment et selon le même processus
  - (b) Le taux de substitutions de type transition est différent du taux de substitutions de type transversion
- Mesure de la distance évolutive (d) en fonction de la fréquence des différences observées (p: transitions, q: transversions):

$$d = -\frac{1}{2}\ln[(1-2p-q)\sqrt{1-2q}\,]$$

Mesure des distances évolutives (5):

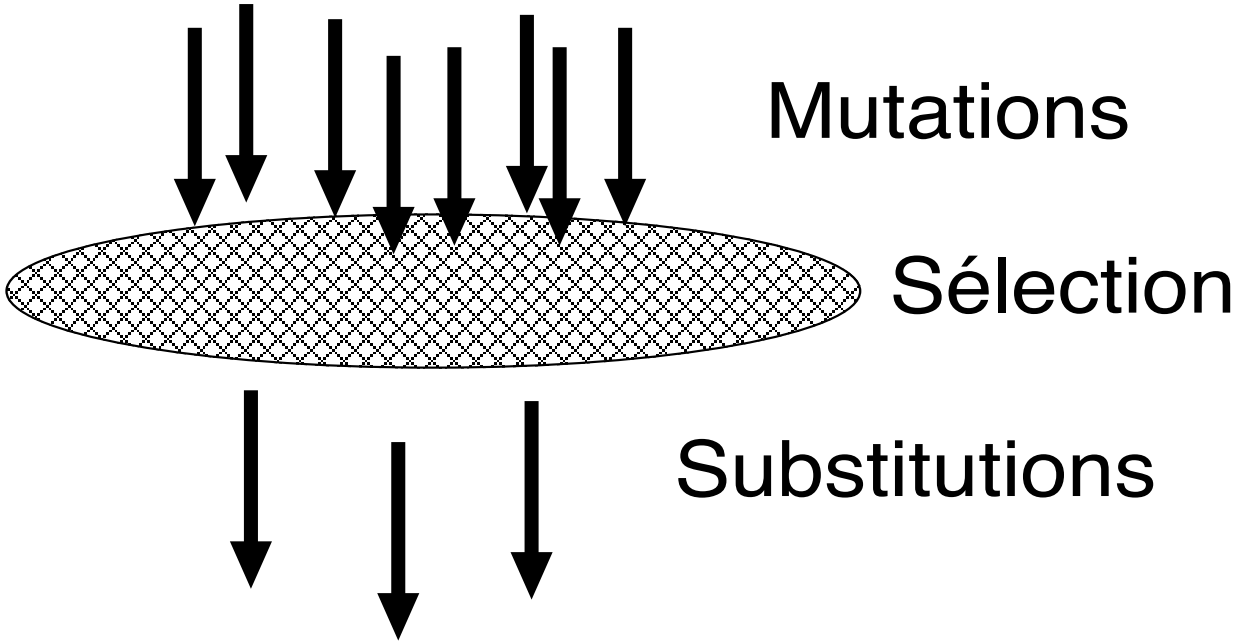
Distance synonymes et non-synonymes (ADN codant): Ka, Ks

- Hypothèse des modèles précédents:
  - (a) Tous les sites évoluent indépendemment et selon le même processus
- Problème: dans les gènes protéiques, il existe deux classes de sites qui ont des vitesses évolutives très différentes:
  - Substitutions non-synonymes (changent l 'a.a.): lentes
  - Substitutions synonymes (ne changent pas l 'a.a.): rapides
- Solution: calculer deux distances évolutives
  - Ka = distance non-synonyme
  - Ka = nbre substitutions non-synonymes / nbre sites non-synonymes
  - Ks = distance synonyme
  - Ks = nbre substitutions synonymes / nbre sites synonymes

Code génétique

TTT	**Phe**	**TCT**	**Ser**	**TAT**	**Tyr**	**TGT**	**Cys**
TTC	**Phe**	**TCC**	**Ser**	**TAC**	**Tyr**	**TGC**	**Cys**
TTA	**Leu**	**TCA**	**Ser**	**TAA**	**∅**	**TGA**	**∅**
TTG	**Leu**	**TCG**	**Ser**	**TAG**	**∅**	**TGG**	**Trp**
CTT	**Leu**	**CCT**	**Pro**	**CAT**	**His**	**CGT**	**Arg**
CTC	**Leu**	**CCC**	**Pro**	**CAC**	**His**	**CGC**	**Arg**
CTA	**Leu**	**CCA**	**Pro**	**CAA**	**Gln**	**CGA**	**Arg**
CTG	**Leu**	**CCG**	**Pro**	**CAG**	**Gln**	**CGG**	**Arg**
ATT	**Ile**	**ACT**	**Thr**	**AAT**	**Asn**	**AGT**	**Ser**
ATC	**Ile**	**ACC**	**Thr**	**AAC**	**Asn**	**AGC**	**Ser**
ATA	**Ile**	**ACA**	**Thr**	**AAA**	**Lys**	**AGA**	**Arg**
ATG	**Met**	**ACG**	**Thr**	**AAG**	**Lys**	**AGG**	**Arg**
GTT	**Val**	**GCT**	**Ala**	**GAT**	**Asp**	**GGT**	**Gly**
GTC	**Val**	**GCC**	**Ala**	**GAC**	**Asp**	**GGC**	**Gly**
GTA	**Val**	**GCA**	**Ala**	**GAA**	**Glu**	**GGA**	**Gly**
GTG	**Val**	**GCG**	**Ala**	**GAG**	**Glu**	**GGG**	**Gly**

Taux de substitution = f(mutation, sélection)



NB: la grande majorité des mutations sont soit neutres (i.e. n ’ont aucun effet sur le phénotype), soit délétères. Les mutations avantageuses sont très rares.

Mesure des distances évolutives (6):

Calcul de Ka et Ks

- Le détail de la méthode est fort complexe. En gros:
  - Séparer tous les sites des deux gènes protéiques comparés en trois catégories: I: non dégénéré, II: partiellement dégénéré, III: complètement dégénéré
  - Calculer le nombre de sites non-synonymes = I + 2/3 II
  - Calculer le nombre de sites synonymes = III + 1/3 II
  - Compter le nombre de changements synonymes et non-synonymes
  - Calculer, selon la méthode de Kimura, Ka et Ks
- On se trouve fréquemment dans l ’une de ces deux situations:
  - Séquences évolutivement peu distantes: Ks est informatif, Ka ne l ’est pas
  - Séquences évolutivement très distantes: Ks est saturé , Ka est informatif

Mesure des distances évolutives (7):

Autres mesures de distances

- Il existe d ’autres modèles évolutifs encore plus réalistes:
  - prise en compte des biais de composition en base (Tajima & Nei)
  - prise en compte des variabilités de taux de substitution entre sites
  - etc ...

Construction d ’arbres phylogénétiques par méthodes de distance

- Principe général:

Alignement de séquences

⬇ (1)

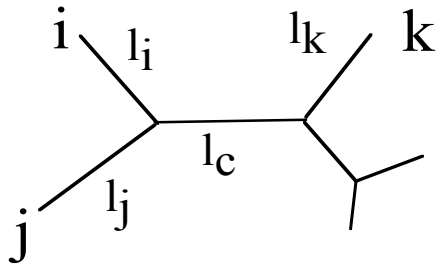
Matrice de distances évolutives entre paires

⬇ (2)

Arbre (non raciné)

- (1) Mesures de distances évolutives
- **(2) Calcul d’arbre à partir d ’une matrice de distance**

Matrice de distance -> arbre (1): # Préambule



- Considérons l'arbre suivant:
- Considérons deux ensembles de distances entre paires de séquences:
  - $d_{x,y}$ = distance de x à y mesurée sur les séquences
  - $\delta_{x,y}$ = distance de x à y déduite de l 'arbre ci-dessus :

    $\delta_{i,j} = l_i + l_j$ $\qquad$ $\delta_{i,k} = l_i + l_c + l_k$
- Il est possible de calculer les longueurs des branches (l_i, l_j, l_c, etc.) telles que les distances δ correspondent "le mieux possible" aux distances *d*. "le mieux possible" signifie que Δ est minimal :

$$\Delta = \sum_{1 \le x < y \le n} (d_{x,y} - \delta_{x,y})^2$$

- On peut alors calculer la longueur de l'arbre:

$$S = l_i + l_j + l_c + \ldots + l_k$$

Matrice de distance -> arbre (2):

Méthode d ’évolution minimale

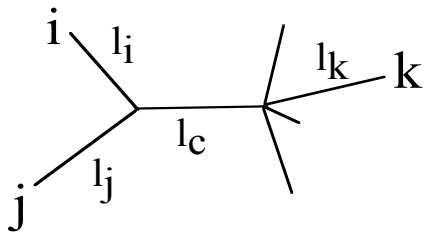
- Etape 1: pour une forme d'arbre donnée, on calcule les longueurs de branches telles que Δ est minimal. On calcule S.
- Etape 2: on répète l'étape 1 pour toutes les formes d'arbre possibles. On garde l'arbre avec S minimal.

- Problème: méthode très couteuse en calcul. Inutilisable en pratique avec plus de 20 séquences => méthode approximative (heuristique). Exemple: Neighbor Joining (Saitou & Nei 1987)

Matrice de distance -> arbre (3):

Méthode Neighbor Joining: algorithme

- On part d'une topologie en étoile et on construit progressivement un arbre selon:
  - Etape 1: Partir avec les distances *d* mesurées entre les *N* séquences
  - Etape 2: Pour tous les choix *i* et *j*: considérer la forme d'arbre ci-dessous, et calculer $S_{i,j}$



  - Etape 3: Choisir la paire (*i,j*) de plus faible valeur $S_{i,j}$. Grouper *i* et *j* dans l'arbre.

Matrice de distance -> arbre (4):

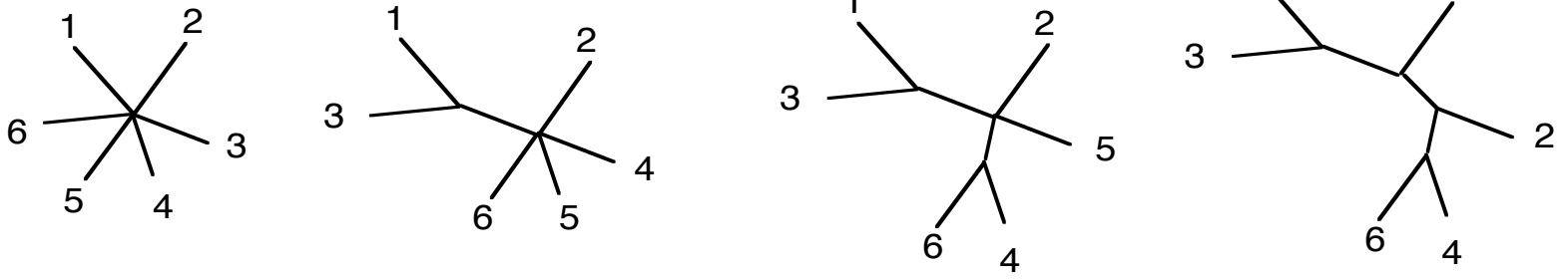
Méthode Neighbor Joining: algorithme (suite)

- Etape 4: Calculer de nouvelles distances d entre N-1 objets: la paire (i,j) et les N-2 séquences restantes.

$$d_{(i,j)k} = (d_{i,k} + d_{j,k}) / 2$$

- Etape 5: Retourner à l'étape 1 tant que $N \geq 4$. Quand $N=3$, on a obtenu un arbre (non-raciné)

- Exemple



Matrice de distance -> arbre (5):

Méthode Neighbor Joining (NJ): propriétés

- NJ est une méthode rapide, même pour des centaines de séquences
- L'arbre NJ est une approximation de l'arbre d'évolution minimale (celui dont la somme des longueurs de branches est minimale)
- En ce sens l'arbre NJ est similaire à la parcimonie puisque les longueurs de branche représentent des substitutions

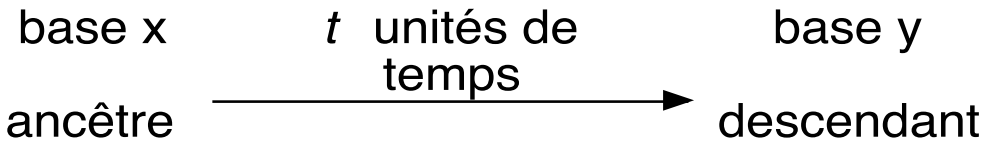
Méthodes de maximum de vraisemblance
(programme Phyml)

- Point de départ
  - On dispose d'un modèle (le plus réaliste possible) d'évolution des séquences (*Cf* les modèles utilisés pour les méthodes de distance)
  - Ce modèle fournit une expression mathématique décrivant les probabilités de substitution d'un résidu par un autre au cours du temps
  - Les paramètres de ce modèle ne sont pas connus
- Principe
  - Rechercher les paramètres du modèle (topologie de l'arbre, longueur des branches, …) qui maximisent la vraisemblance des données (i.e. de l'alignement)

Méthodes de maximum de vraisemblance

Exemple simple: modèle de substitution à un paramètre:
v = probabilité qu'une base change par unité de temps
(phyml utilise un modèle plus élaboré)

- Considérons l'évolution le long d'une branche de l'arbre:



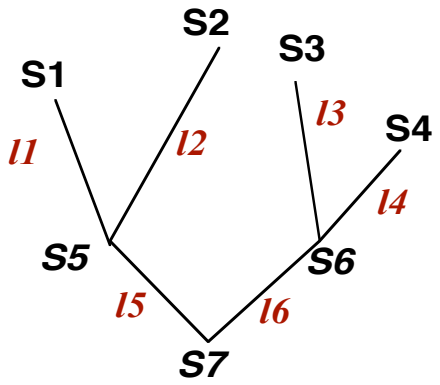
- Notre modèle probabiliste permet de calculer la probabilité de la substitution x → y le long de cette branche

$$P(x,y) = \begin{cases} \frac{3}{4}e^{-\frac{4}{3}l} + \frac{1}{4} & \text{si } x = y \\ -\frac{1}{4}e^{-\frac{4}{3}l} + \frac{1}{4} & \text{si } x \neq y \end{cases} \qquad \text{avec } l = 3\,vt$$

- La quantité $l = 3vt$ est le nombre moyen de substitution / site le long de cette branche, *i.e.* la longueur de la branche

Algorithme du maximum de vraisemblance (1)

- Etape 1: Considérons un arbre raciné donné, un site donné, et un jeu donné de longueurs des branches. Calculons la probabilité pour que les bases observées à ce site aient évolué le long de cet arbre.



S1, S2, S3, S4: bases observées au site dans seq. 1, 2, 3, 4

S5, S6, S7: bases ancestrales inconnues et variables

l1, l2, ..., l6: longueur des branches données

P(S1, S2, S3, S4)=

$\Sigma_{S7}\Sigma_{S5}\Sigma_{S6} P(S7)\ P_{l5}(S7,S5)\ P_{l6}(S7,S6)\ P_{l1}(S5,S1)\ P_{l2}(S5,S2)\ P_{l3}(S6,S3)\ P_{l4}(S6,S4)$

Où P(S7) est estimée par les fréq. moyennes des bases dans les séquences.

Algorithme du maximum de vraisemblance (2)

- Etape 2: Calculons la probabilité que les séquences entières aient évolué:

  $P(Sq1, Sq2, Sq3, Sq4)=\Pi_{\text{tous les sites}} P(S1, S2, S3, S4)$

- Etape 2: Calculons les longueurs *l1, l2, ..., l6* qui donnent la plus haute valeur $P(Sq1, Sq2, Sq3, Sq4)$. C'est la *vraisemblance* de l'arbre.

- Etape 3: Calculons la vraisemblance de tous les arbres possibles. L'arbre prédit par la méthode est celui ayant la plus forte vraisemblance.

Maximum de vraisemblance: propriétés

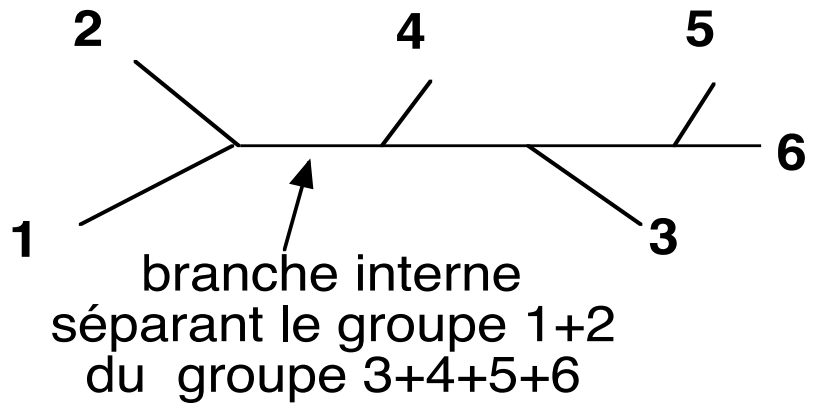
- C'est la méthode la mieux justifiée du point de vue théorique

- Il a été démontré que cette méthode marche mieux que toutes les autres dans la plupart des cas
- Il est presque toujours impossible d'évaluer tous les arbres possibles car il y en a trop. On fait alors une exploration partielle de l'ensemble des arbres (heuristique). On perd la certitude mathématique d'avoir trouvé l'arbre le plus vraisemblable.

- Plus lent que méthodes de distance, mais suffisamment rapide pour être utilisé (phyml: http://atgc.lirmm.fr/phyml/)
- Importance du choix du modèle

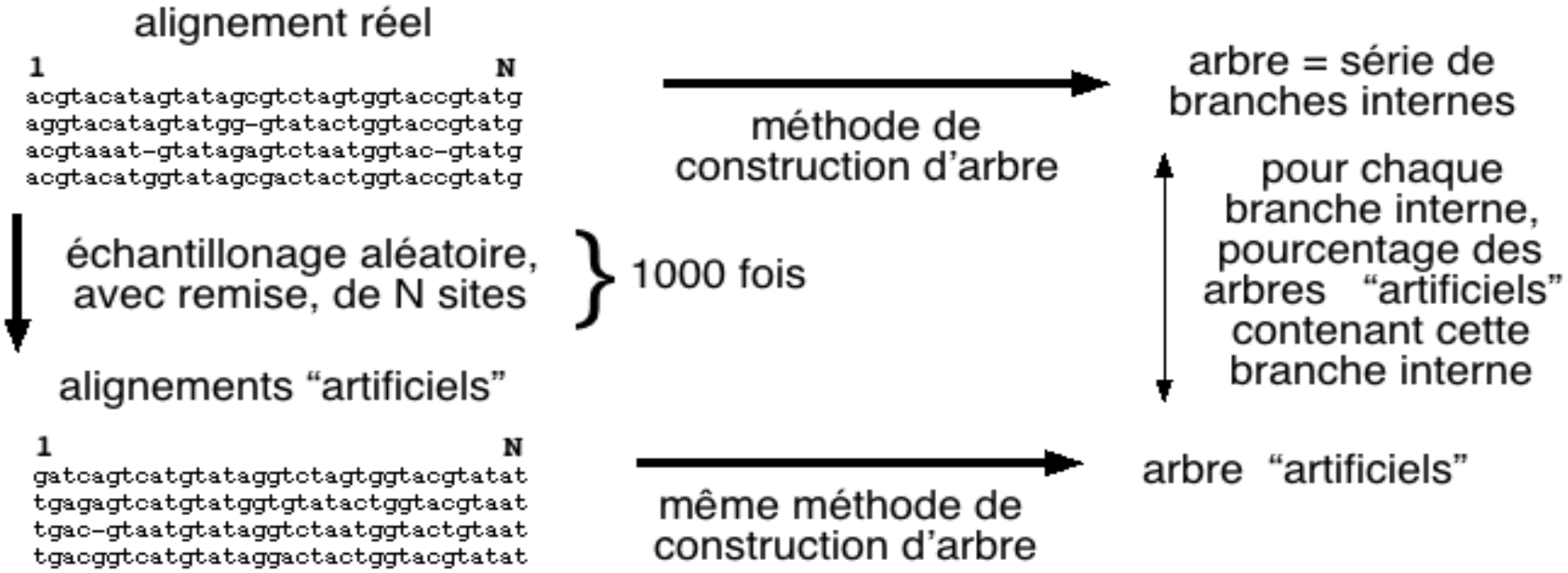
Fiabilité des arbres phylogénétiques: le bootstrap

- L'information phylogénétique véhiculée par un arbre sans racine réside entièrement dans ses branches internes



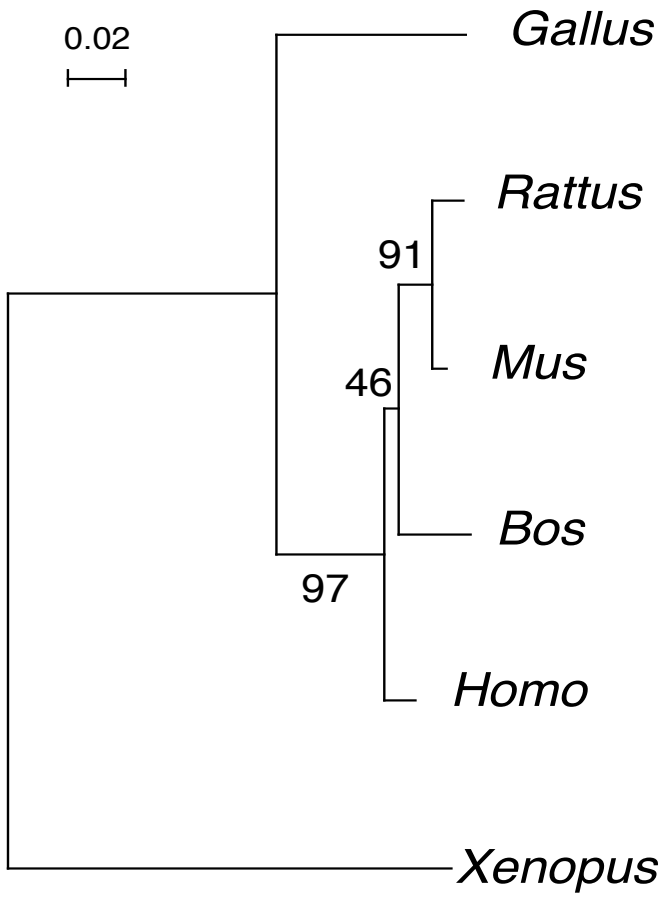
- La forme d'un arbre peut être déduite de la liste de ses branches internes.
- Tester la fiabilité d'un arbre = tester la fiabilité des branches internes

Procédure de bootstrap



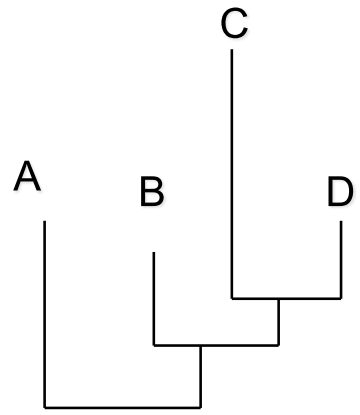
- Les branches internes soutenues par ≥90% des réplications sont considérées comme significatives
- Le bootstrap teste uniquement si la longueur des séquences étudiées était suffisante pour soutenir un nœud particulier

Arbre "bootstrapé"

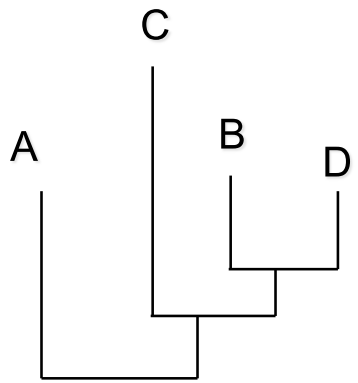


Causes d'erreurs dans les arbres phylogénétiques

- Alignement erroné: *ne garder que les parties fiables de l'alignement*
- Saturation (perte du signal phylogénétique): *éliminer les sites à évolution trop rapide*
- Attraction des longues branches: *augmenter l'échantillonage taxonomique*



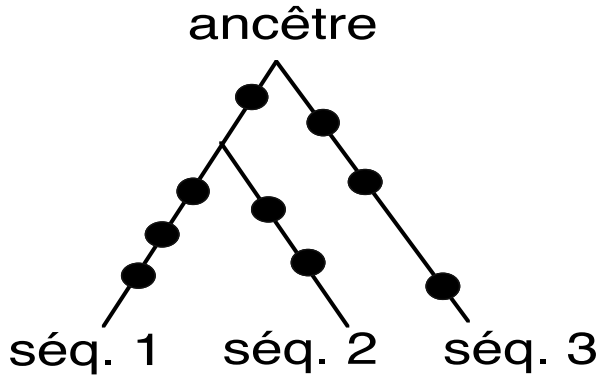
Arbre vrai



Arbre calculé

Saturation: perte du signal phylogénétique

- Lorsque les séquences homologues que l ’on compare ont subi trop de substitutions depuis leur divergence, il est impossible de déterminer l ’arbre phylogénétique (quelle que soit la méthode utilisée)



- NB: pour les méthodes de distance, ce phénomène de saturation se traduit par l ’impossibilité de calculer *d*. Exemple: Jukes-Cantor: $p \rightarrow 0{,}75 \Rightarrow d \rightarrow \infty$ et $V(d) \rightarrow \infty$
- NB: le phénomène de saturation n ’est pas toujours détectable

Pour faire un arbre ...

- Rechercher les séquences homologues (e.g. BLAST)
- Aligner les séquences (e.g. MUSCLE), vérifier et, éventuellement, éditer l'alignement (e.g. SEAVIEW)
- Déterminer les parties fiables de l'alignement sur lesquelles on se basera pour construire l'arbre
- Attention au problème de saturation (e.g. sites synonymes dans une région codante) => choix des sites que l'on va utiliser pour construire l'arbre (e.g. sites synonymes vs. non- synonymes)
- Choix de la méthode de reconstruction (compromis vitesse/fiabilité)
- Choix des paramètres (distance, modèle max. vraisemblance)
- Evaluation de la fiabilité des branches (bootstrap)
- Interprétation de l'arbre

Arbre des gènes / arbre des espèces

- L'histoire évolutive des gènes reflète celle des espèces qui les portent, sauf si:
  - Transfert horizontal = transfert de gènes entre espèces (*e.g.* bactéries, mitochondries)
  - Duplication de gènes: orthologie/ paralogie

Orthologie/Paralogie

spéciation

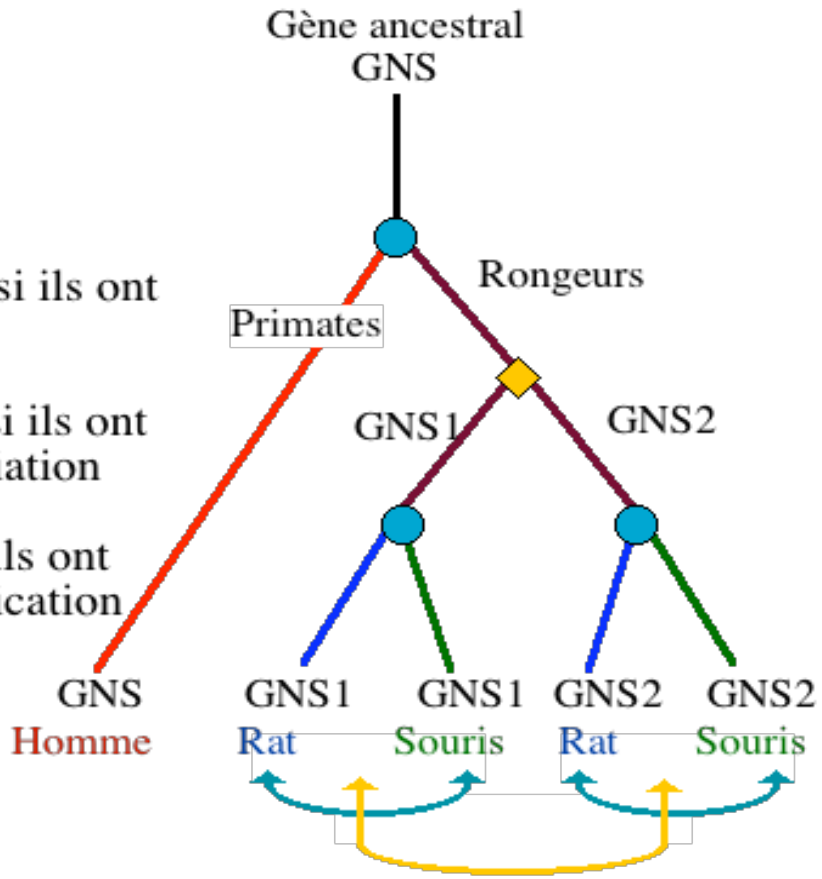
duplication

Homologie: deux gènes sont homologues si ils ont un ancêtre commun

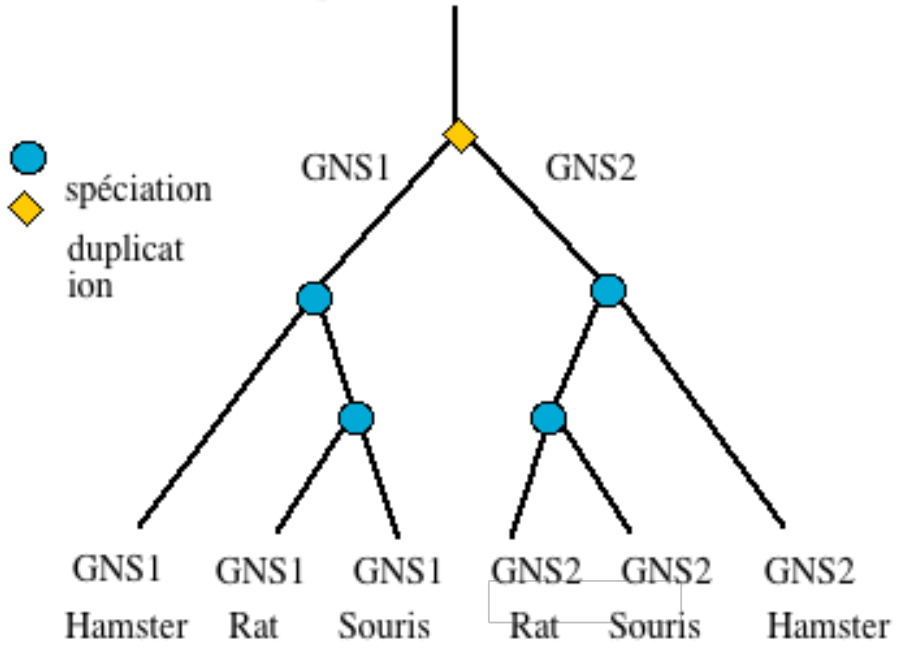
Orthologie: deux gènes sont orthologues si ils ont divergé à la suite d'un évènement de spéciation

Paralogie: deux gènes sont paralogues si ils ont divergé à la suite d'un évènement de duplication

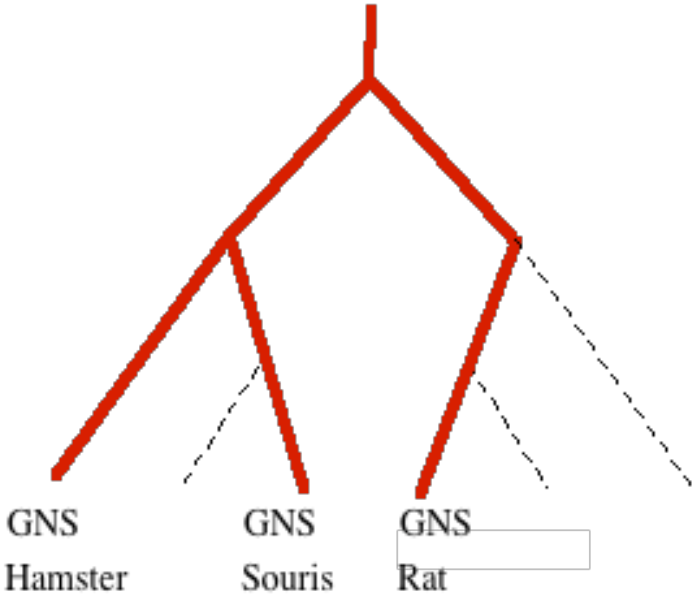
Orthologie ≠ équivalence fonctionnelle



Reconstruction de la phylogénie des espèces: artéfacts dus à la paralogie



Arbre vrai

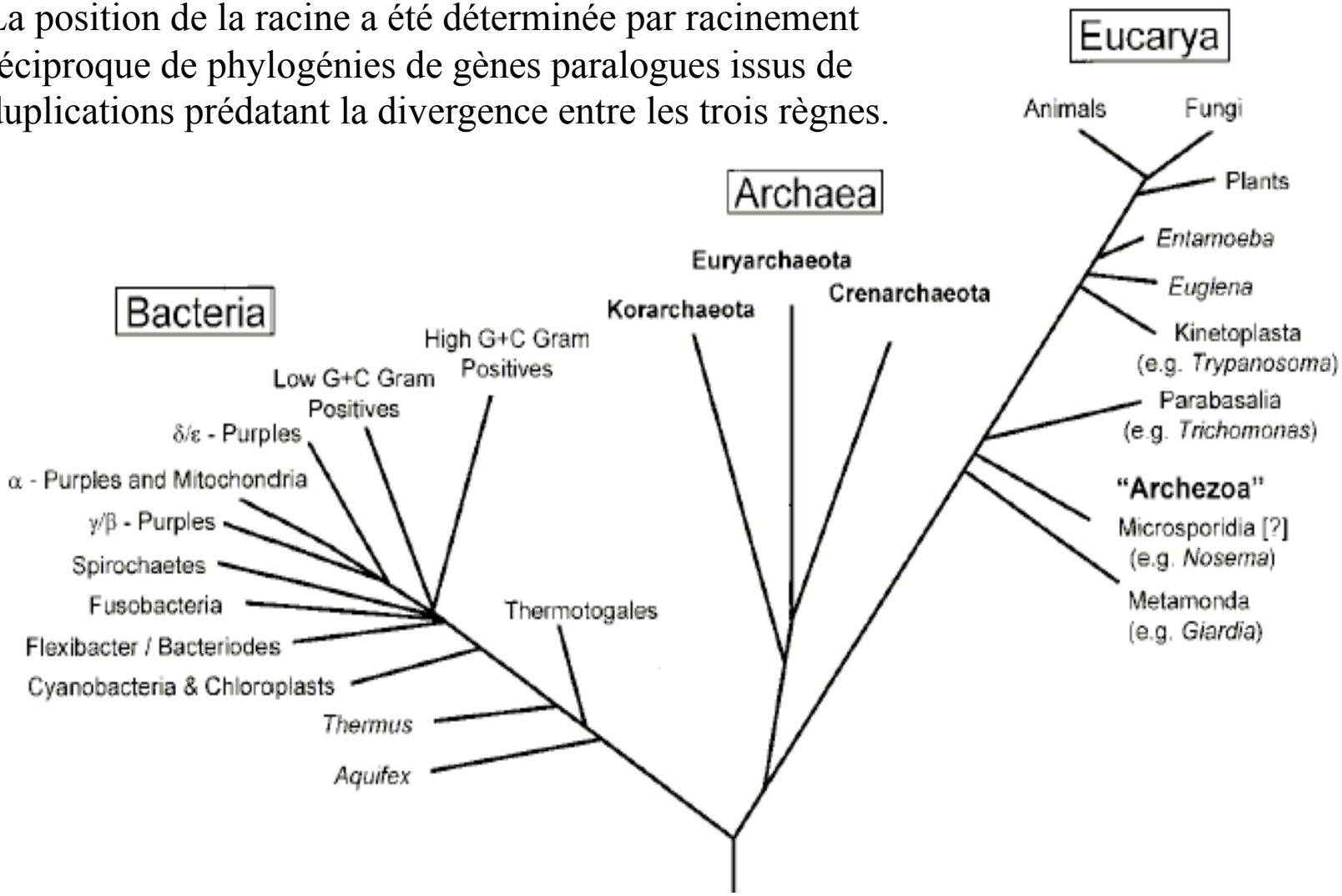


Arbre obtenu avec un échantillongage incomplet de gènes homologues

!! Des pertes de gènes peuvent se produire au cours de l'évolution: même avec la séquence complète d'un génome il n'est pas toujours possible de détecter de la paralogie !!

Phylogenie universelle (2)

La position de la racine a été déterminée par racinement réciproque de phylogénies de gènes paralogues issus de duplications prédatant la divergence entre les trois règnes.



Brown & Doolittle (1997) Microbiol.Mol.Biol.Rev. 61:456-502