http://pbil.univ-lyon1.fr/members/mgouy/cours/INSA/BIM2008.pdf

Introduction aux méthodes pour la phylogénie moléculaire

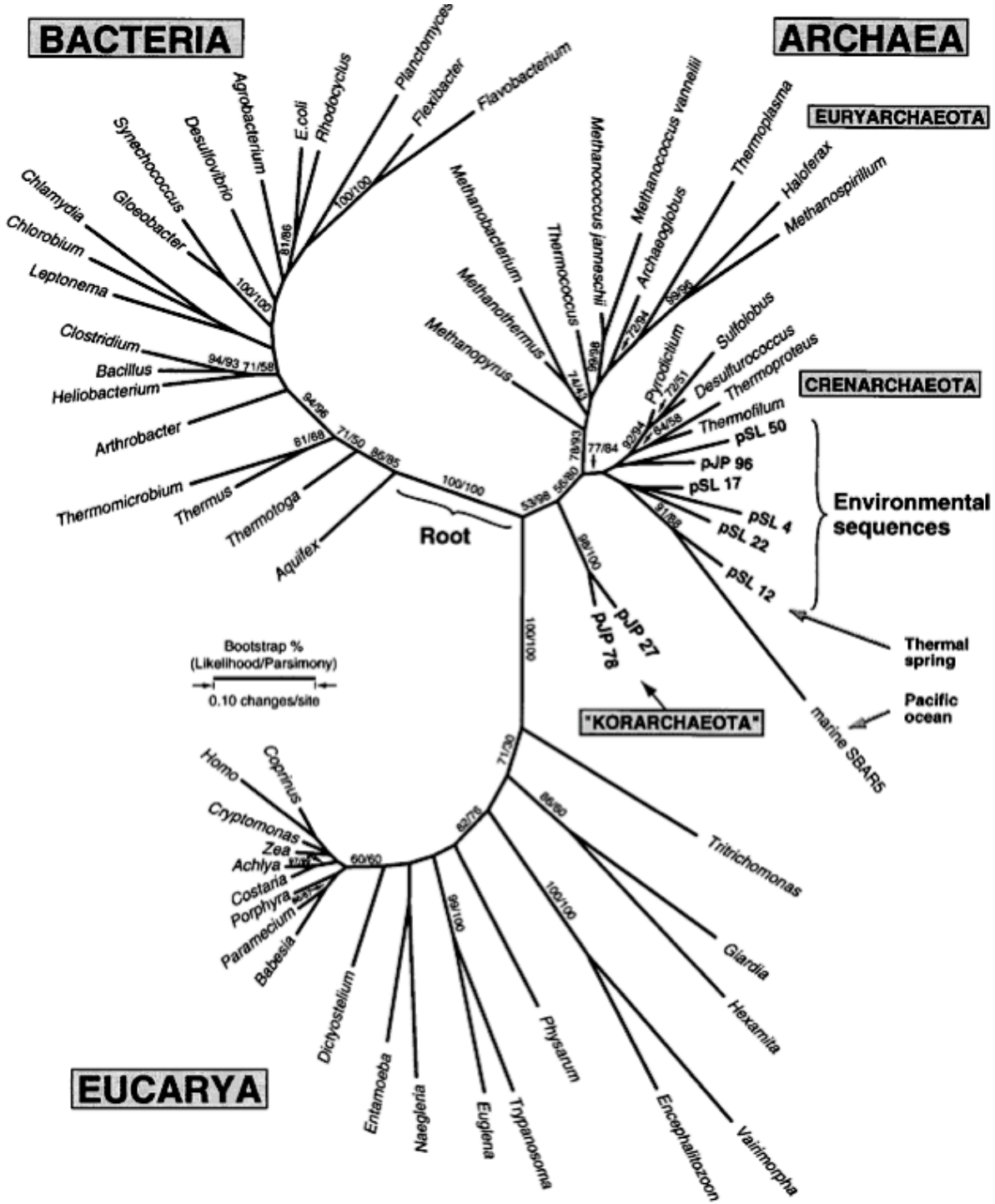
INSA de Lyon

« Bioinformatique et Modélisation »

4-BIM

Manolo Gouy

Avril 2008



La phylogénie universelle

ARNr petite sous-unité

Barns et al. (1996) PNAS 93:9188

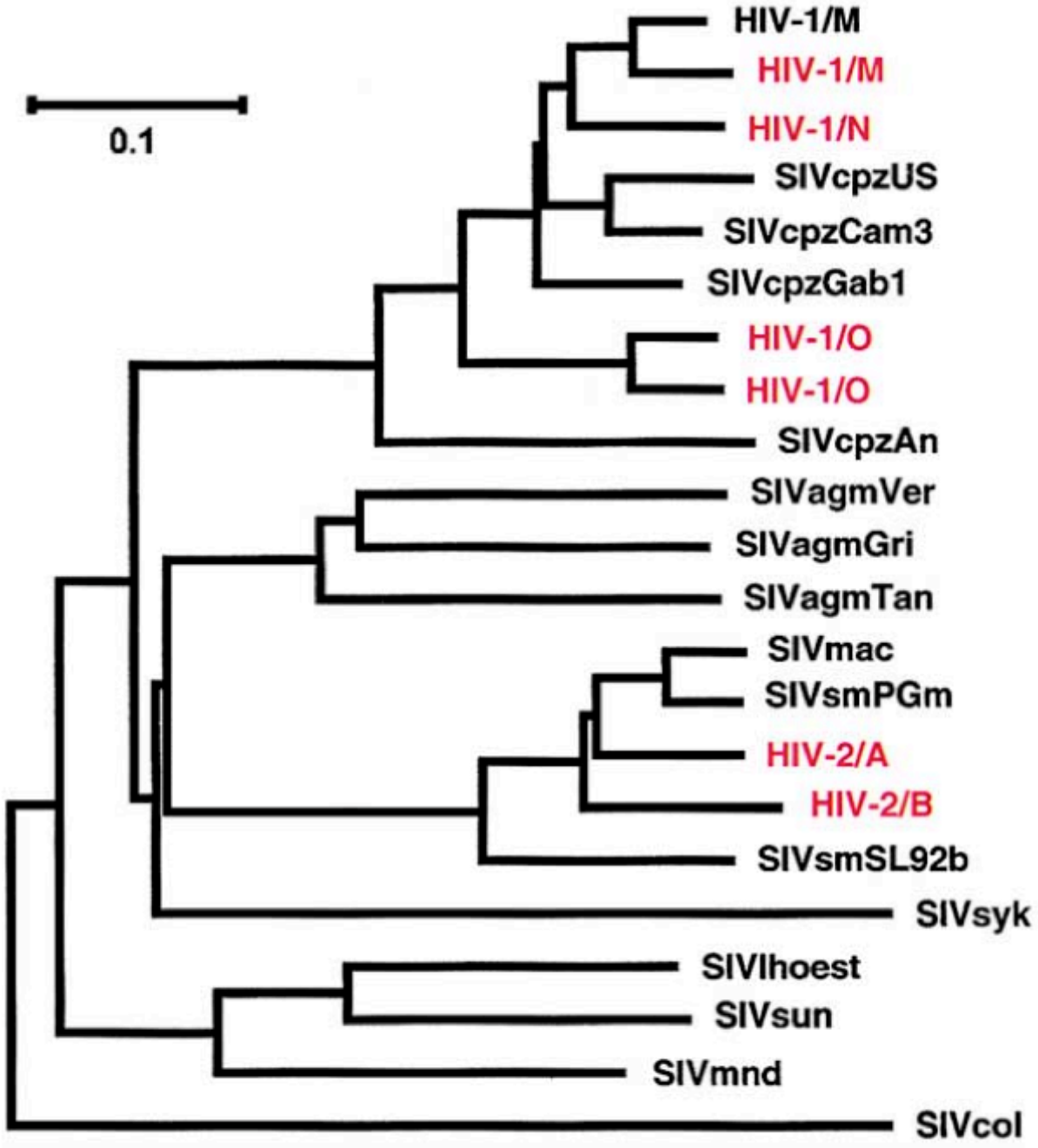


Figure 1. Evolution of AIDS Viruses

Origine du virus du SIDA

cpz: chimpanzé --> HIV-1
agm: singe vert africain
mac: macaque
sm: *Cercocebus atys* --> HIV-2
syk: *Cercopithecus albogularis*
lhoest: *C. lhoesti*
sun: *C. solatus*
mnd: *Mandrillus sphinx*
col: *Colobus guereza*

Sharp (2002) Cell 108:305

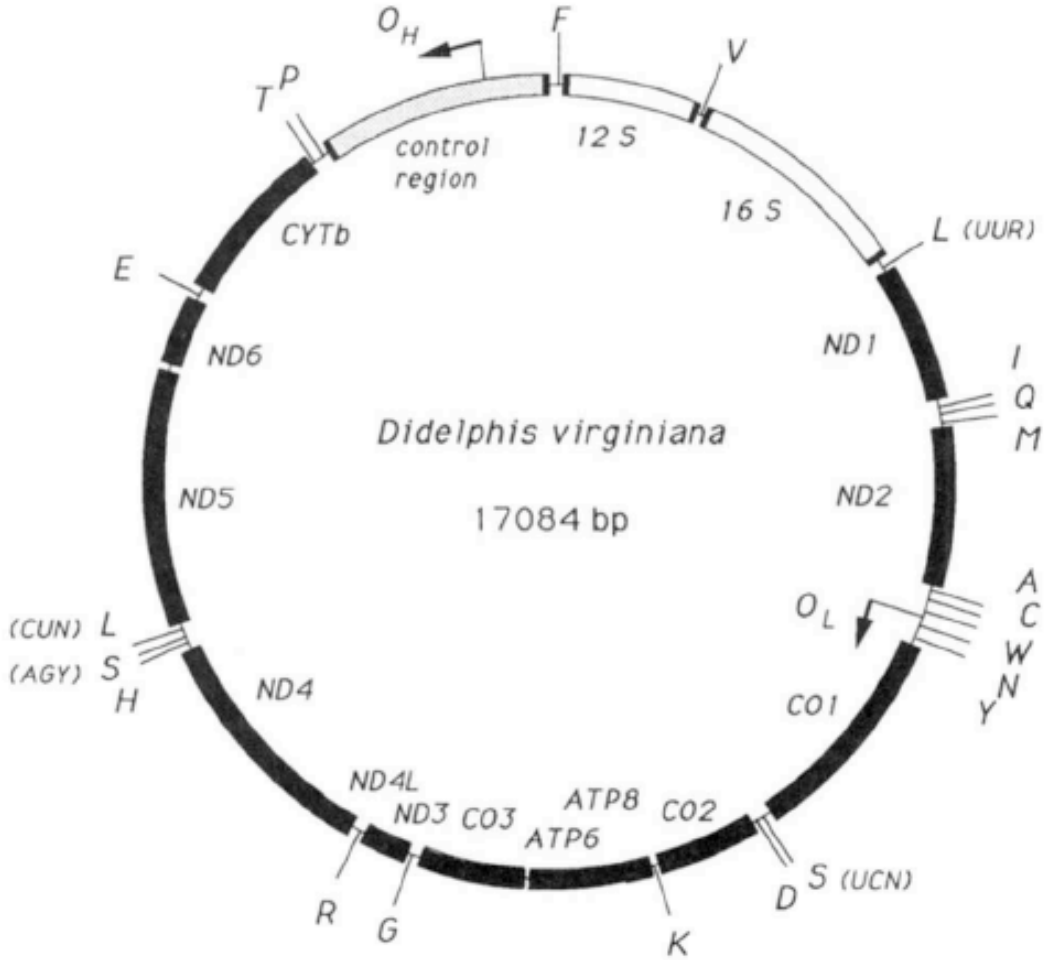
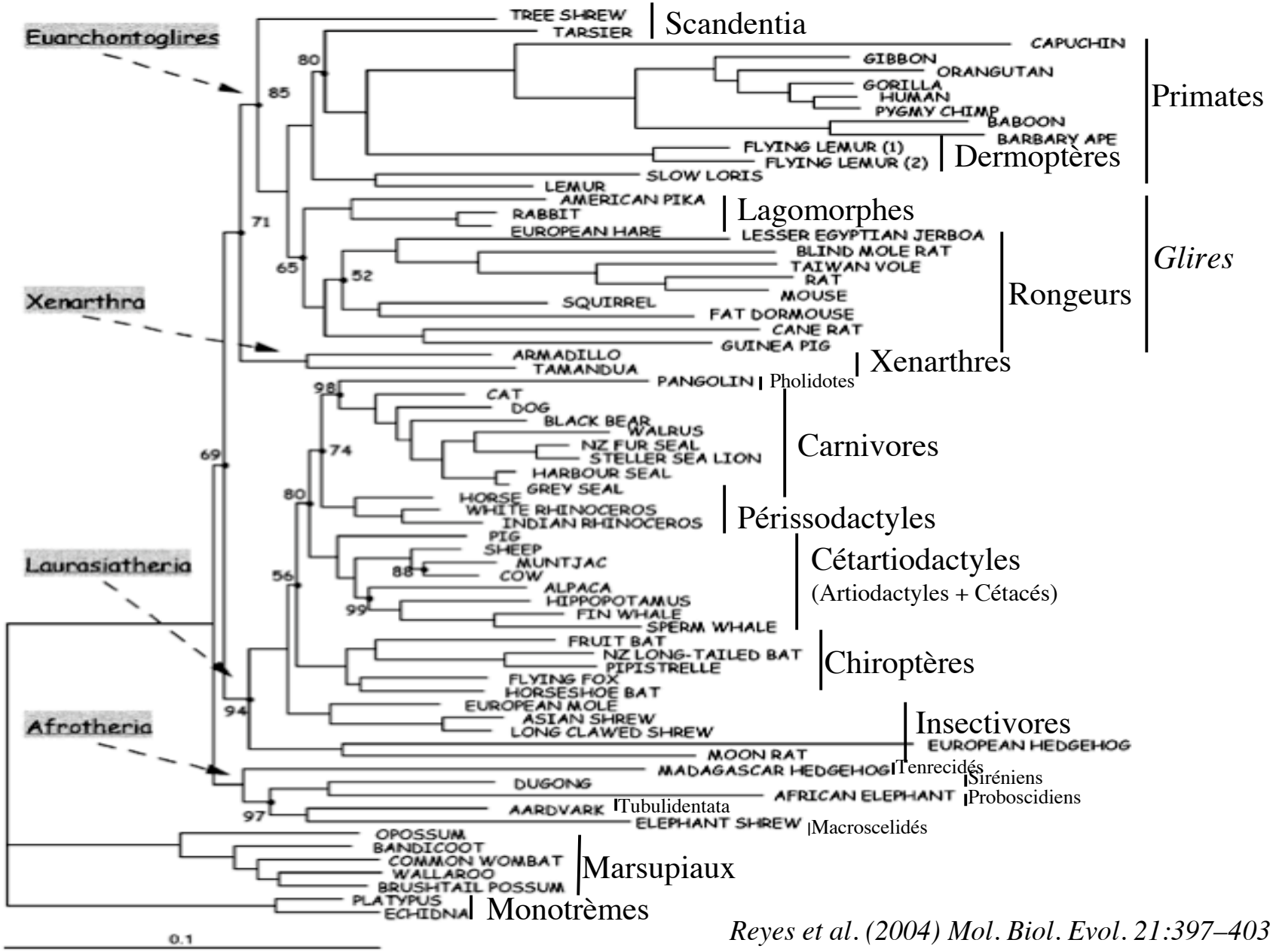


FIGURE 1.—Map of the *D. virginiana* mitochondrial DNA molecule. The location of origins of replication as well as the identity and arrangement of the various genes were determined by comparison of published mammalian sequences. Each tRNA is identified by its one-letter amino acid code. The tRNAs for serine and leucine are further identified by their codon family specificity. The *ATPase6* and *ATPase8* genes overlap by 46 nucleotides.

Janke et al. (1994)
Genetics 137:243



Reyes et al. (2004) Mol. Biol. Evol. 21:397–403

FIG. 1.—Phylogenetic tree of placental mammals reconstructed using the program MrBayes from mitochondrial H-stranded protein-coding genes using ungapped first and second codon positions with the exclusion of Leu synonymous sites. Posterior probabilities (PP) supporting the tree nodes are only reported when less than 100. Marsupialia and Monotremata were used as outgroups. The lengths of the branches are proportional to the number of nucleotide substitutions per site.

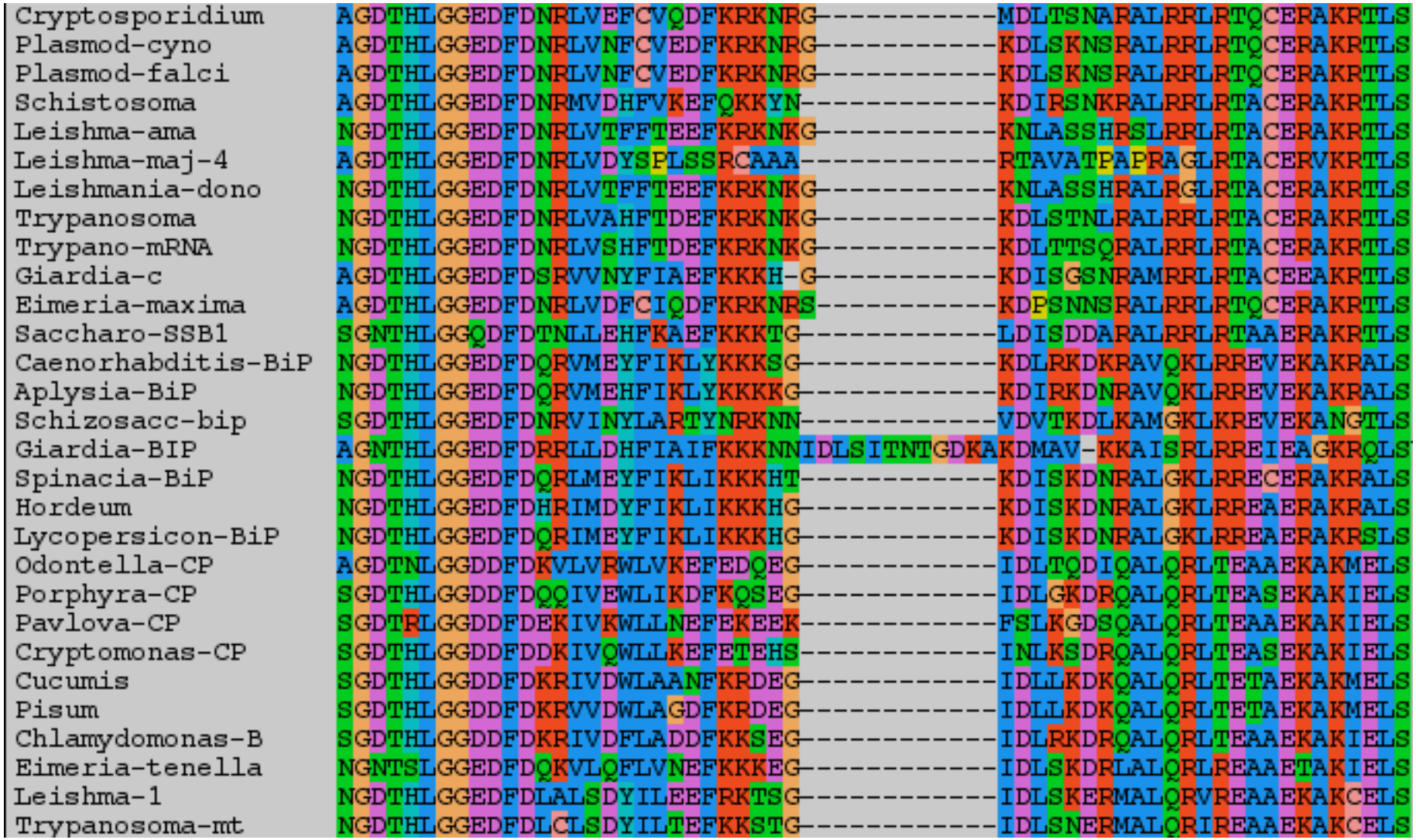
Algorithmes pour la Phylogénie Moléculaire

- Point de départ: un ensemble de séquences d'ADN ou de protéines homologues et alignées.
- Résultat final: un arbre décrivant les relations évolutives entre les séquences étudiées

  = une généalogie de séquences
  = un arbre phylogénétique

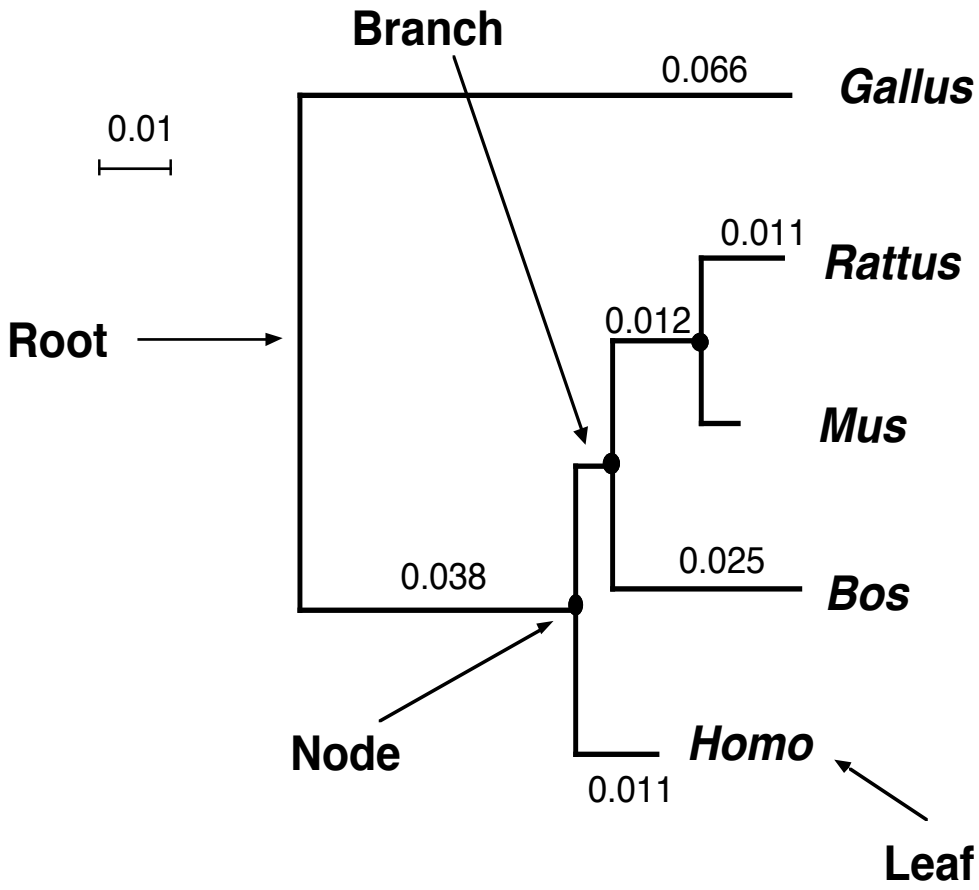
```
CLUSTAL W (1.74) multiple sequence alignment

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

Exemple d'alignement mutiple protéique (fragment)



Arbre Phylogénétique

- Branche Interne: entre 2 nœuds. Branche Externe: entre un nœud et une feuille
- Les longueurs des branches horizontales sont proportionnelles aux distances évolutives entre séquences ancestrales (unité = substitution / site).
- Topologie d'arbre = forme de l'arbre = ordre de branchement des nœuds



Alignement et Gaps

- La qualité de l'alignement est essentielle : chaque colonne de l'alignement (site) est supposée contenir des résidus homologues (nucléotides, acides aminés) qui dérivent d'un ancêtre commun.

  ==> Les parties non fiables de l'alignement doivent être omises du reste des analyses.

- La plupart des méthodes ne tiennent compte que des substitutions ; les gaps (événements d'insertion/délétion) ne sont pas utilisés.

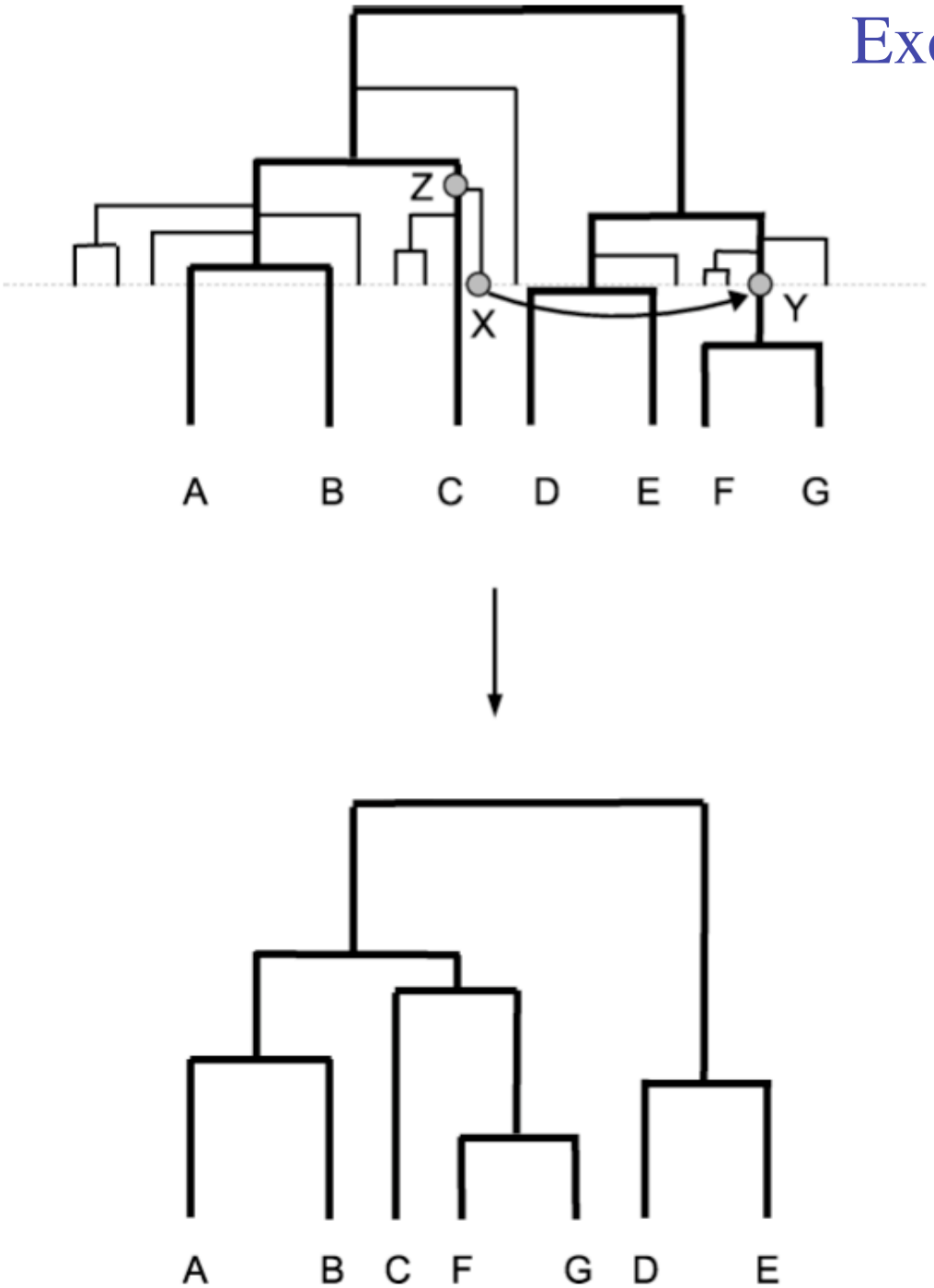
  ==> les sites contenant des gaps sont ignorés.

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

Arbre des gènes *vs*. Arbre des espèces

- L'histoire évolutive des gènes reproduit celle des espèces qui les porte, sauf si:
  - Transfert horizontal = transfert de gène entre espèces
  - Duplication génique : orthologie/ paralogie

Exemple de transfert horizontal



En gras: arbre de 7 espèces.

Un ancêtre des espèces F et G obtient son gène par transfert horizontal depuis un donneur, parent d'un ancêtre de l'espèce C.

L'arbre du gène diffère de l'arbre des espèces porteuses de ce gène.

Orthologie / Paralogie

speciation

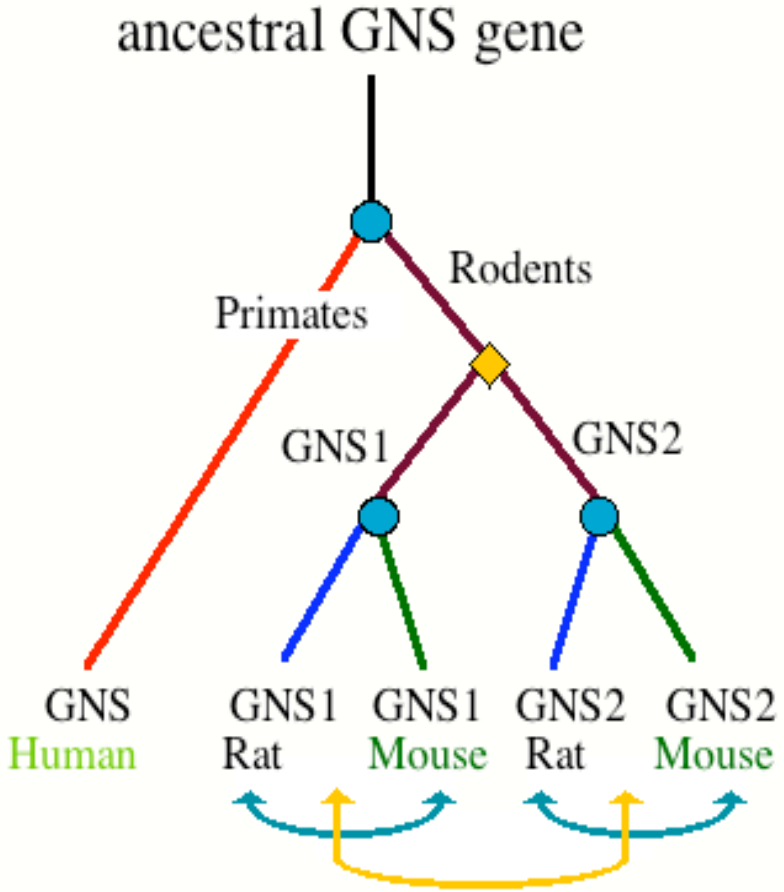
duplication

Homology : two genes are homologous iff they have a common ancestor.

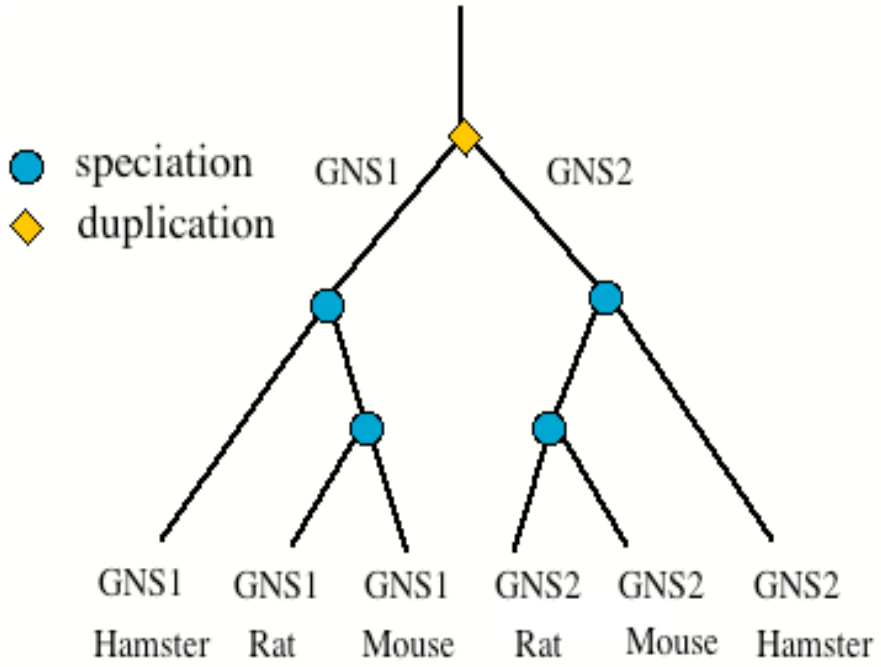
Orthology : two genes are orthologous iff they diverged following a speciation event.

Paralogy : two genes are paralogous iff they diverged following a duplication event.

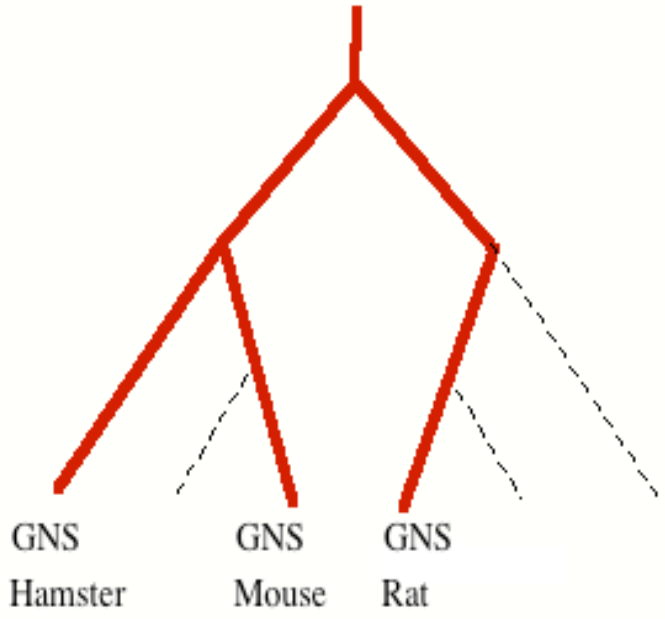
Orthology ≠ functional equivalence



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



true tree



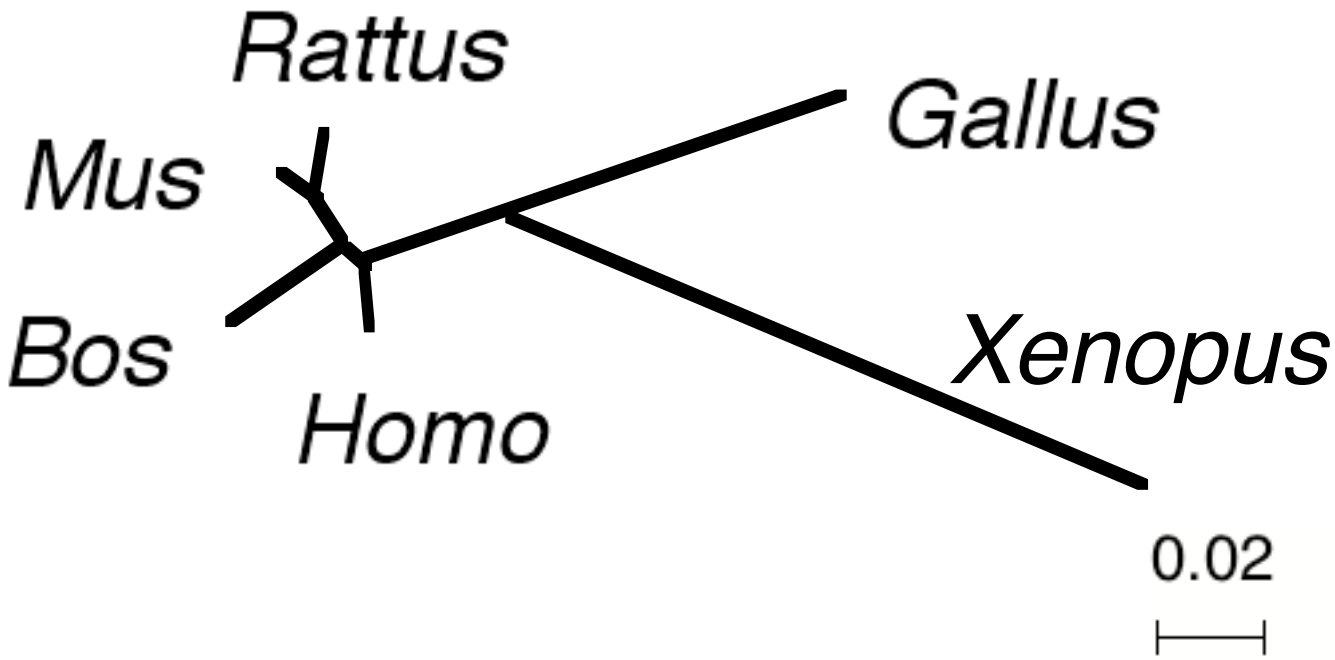
tree obtained with a partial sampling of homologous genes

!! Des pertes de gènes peuvent se produire au cours de l'évolution : même avec des séquences génomiques complètes, il peut être difficile de détecter la paralogie !!

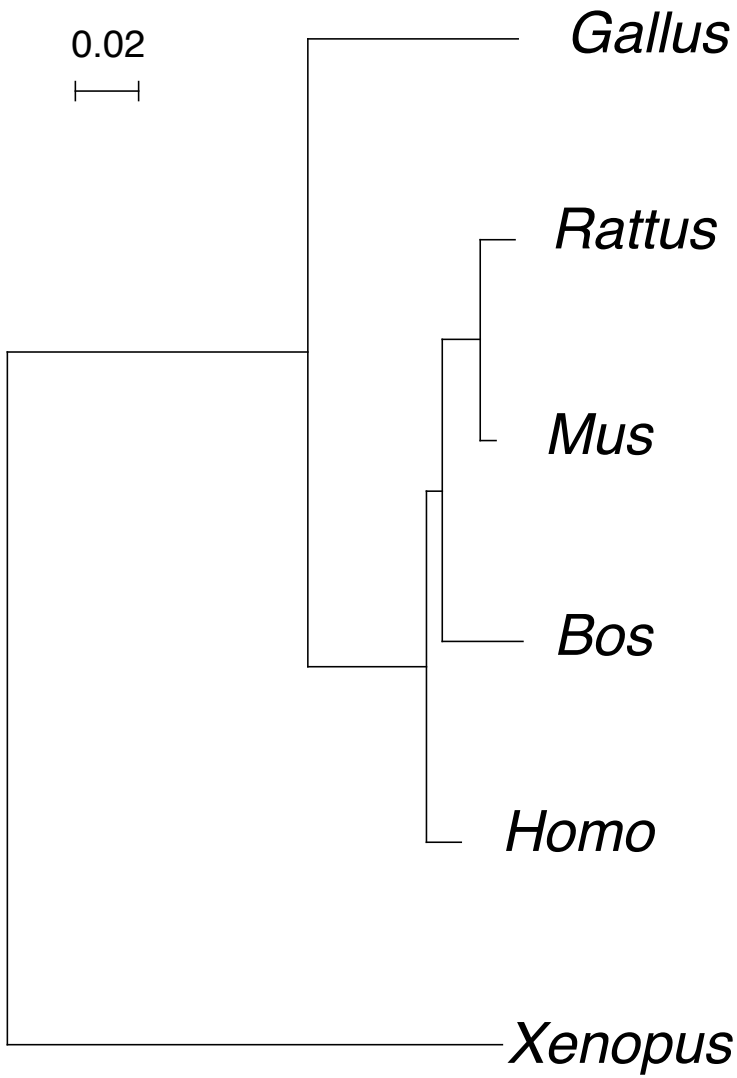
Arbres racinés et non-racinés

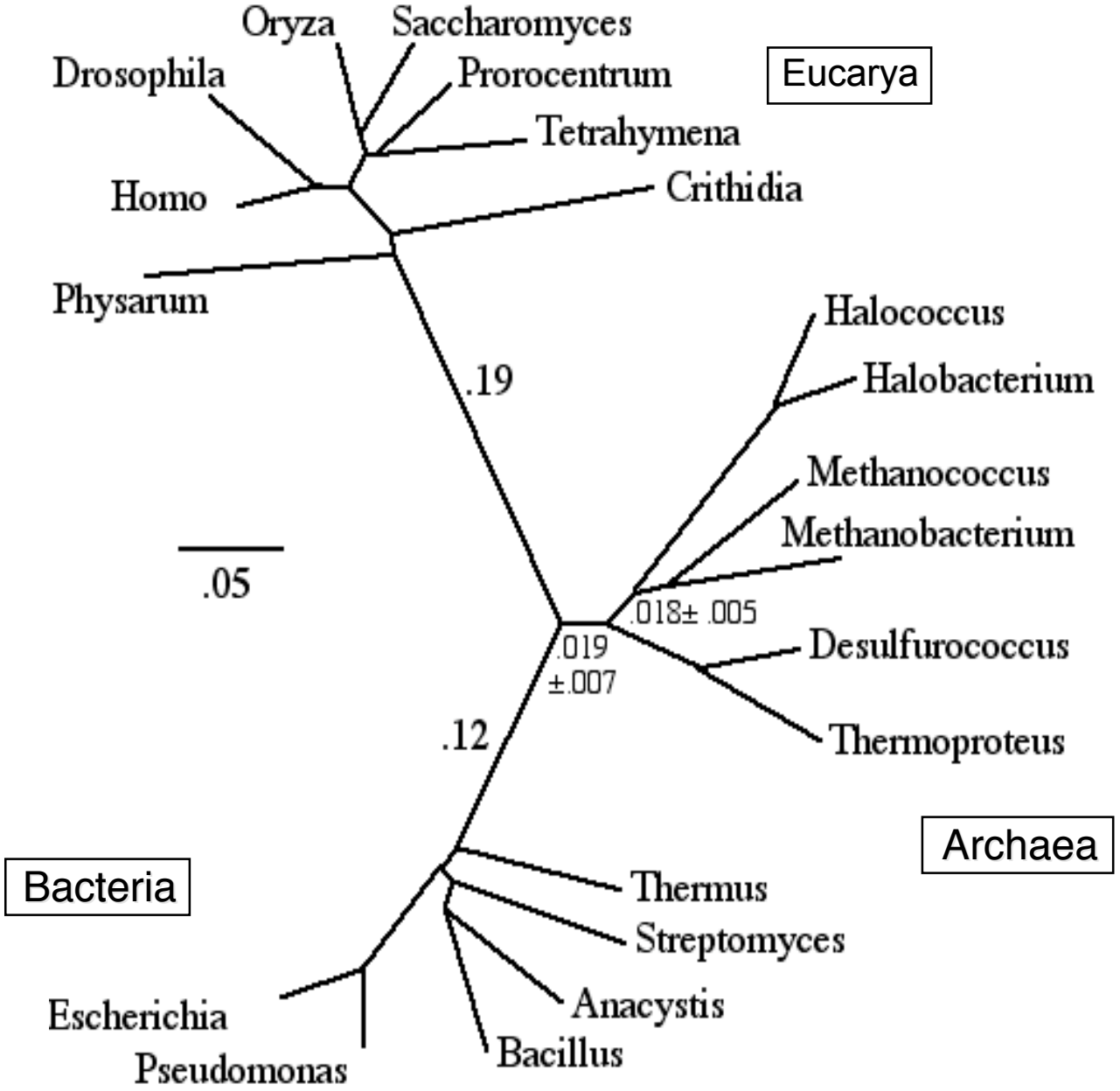
- La plupart des méthodes phylogénétiques produisent des arbres non racinés. La raison est que les méthodes détectent des différences entre séquences, sans avoir le moyen de les orienter temporellement.
- Deux façons d'enraciner un arbre non raciné:
  - Méthode du groupe externe : inclure dans l'analyse un groupe de séquences dont on sait *a priori* qu'elles sont externes au groupe étudié; la racine est sur la branche qui relie le groupe externe aux autres séquences.
  - Faire l'hypothèse de l'horloge moléculaire : toutes les lignées sont supposées évoluer à la même vitesse depuis leur divergence; la racine est au point de l'arbre équidistant de toutes ses feuilles.

Arbre non raciné



Arbre raciné



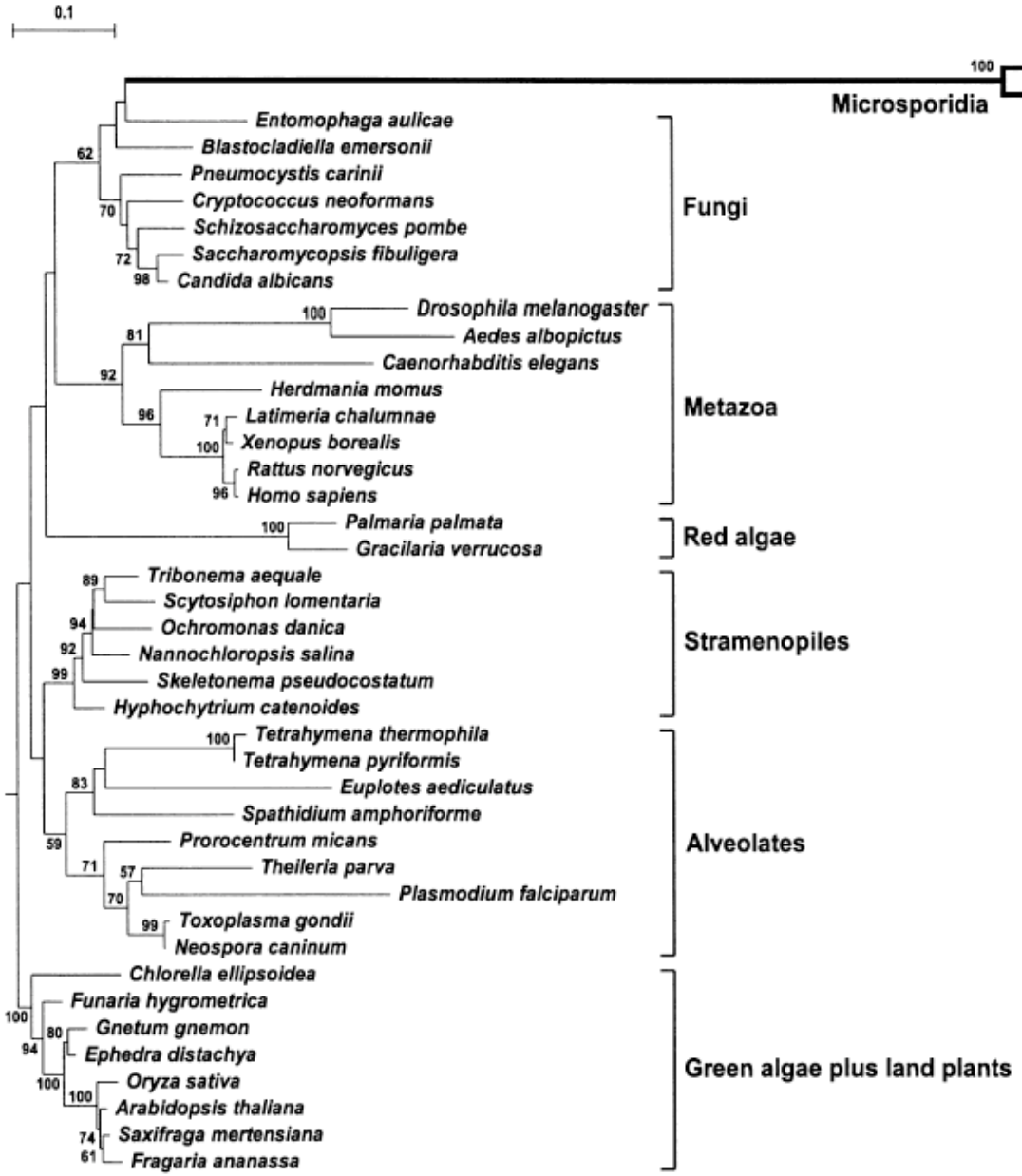


Phylogénie universelle

Déduite de la comparaison de séquences de SSU et LSU rRNA (2508 sites homologues) en utilisant la distance de Kimura à 2 paramètres et la méthode NJ.

L'absence de racine de cet arbre est exprimée par le graphisme circulaire.

Racinement par le centre: incorrect si fortes différences de vitesse entre lignées



Analyse en distance de 42 séquences de LSU rRNA de microsporidies et d'autres eucaryotes.

Van de Peer et al. (2000) Gene 246:1

Nombre de topologies d'arbres binaires non racinés 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	$\sim 2 \times 10^{20}$

Méthodes pour la reconstruction phylogénétique

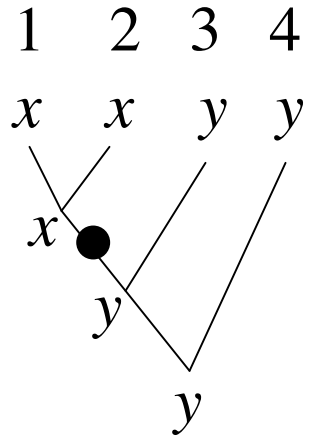
Trois familles principales de méthodes :

- Parcimonie
- Méthodes de distances
- Méthodes probabilistes (maximum de vraisemblance, *méthodes bayesiennes*)

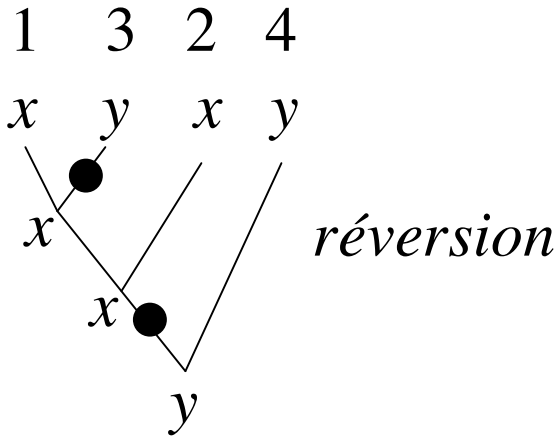
Pourquoi la parcimonie ?

Soit un caractère relevé dans 4 espèces $\{1, 2, 3, 4\}$ et présentant les états $\{x,x,y,y\}$.
Quelle histoire évolutive a pu conduire à cet état final ?

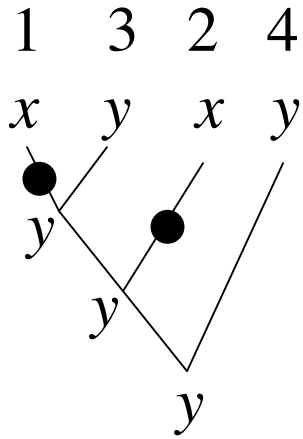
Egalité par ascendance commune:
deux espèces possèdent le même état de caractère car elles l'ont hérité sans le transformer de leur dernier ancêtre commun



Présence d'homoplasie: des états identiques sont observés bien qu'ils n'aient pas été hérités, inchangés, du dernier ancêtre.



réversion



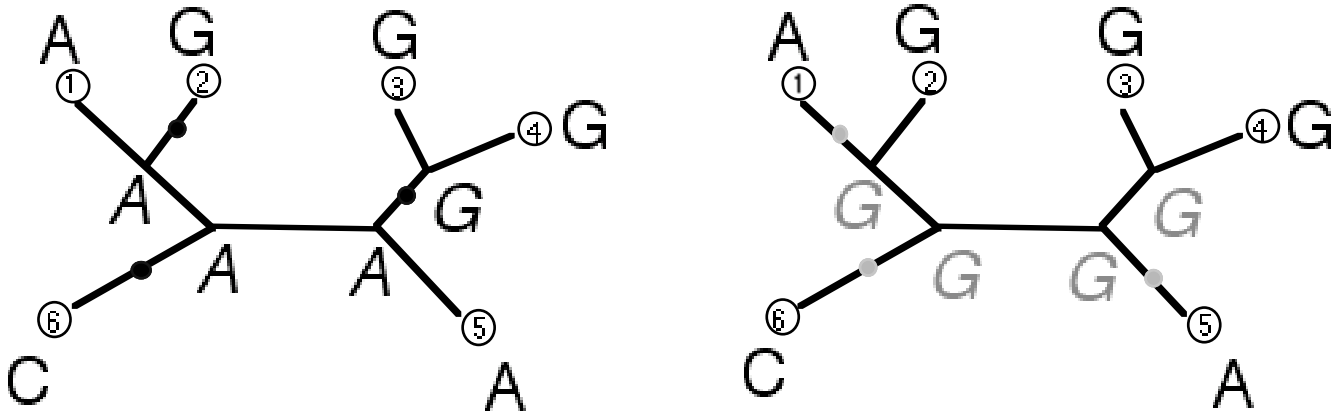
convergence

Les scénarios homoplasiques demandent plus de changements au cours de l'évolution. La parcimonie parie que convergences et réversions sont rares et recherche l'histoire qui demande le moins possible de changements.

Parcimonie (1)

- Etape 1: Pour une topologie d'arbre donnée, et pour un site donné de l'alignement, calculer, à l'aide de l'algorithme de Fitch, le plus petit nombre total de changements dans tout l'arbre.

  Soit *d* ce nombre total de changements.



X: ancestral nucleotide

• : substitution event

Exemple: A ce site et pour cette forme d'arbre, au moins 3 substitutions sont nécessaires pour expliquer le pattern de nucléotides présent aux feuilles de l'arbre. Plusieurs scénarios distincts à 3 changements sont possibles.

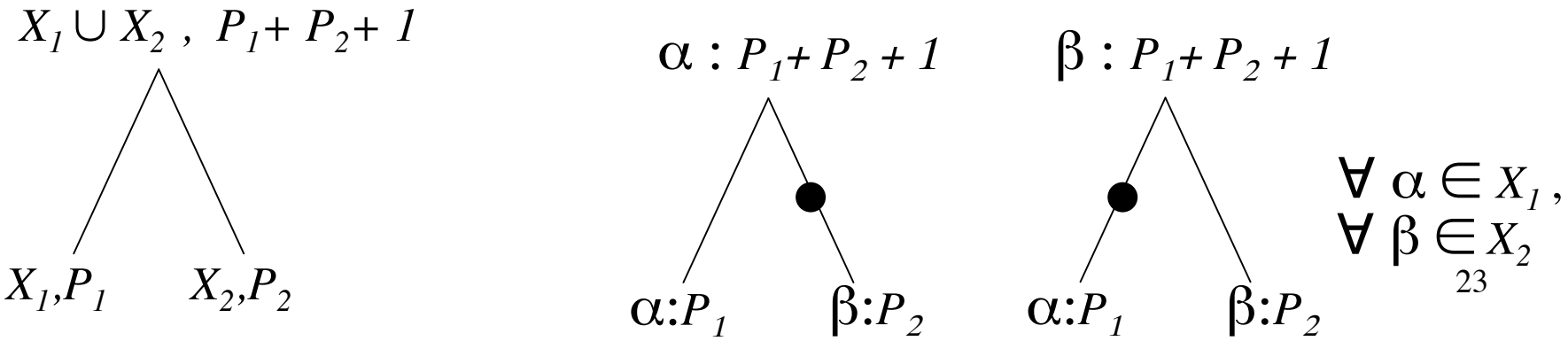
Algorithme de Fitch : calcul du nombre minimal de changements

Raciner arbitrairement l'arbre et calculer récursivement, à chaque nœud, deux objets:

- P: nombre minimal de changements dans le sous-arbre dont ce nœud est racine
- X: ensemble des résidus tous également possibles à ce nœud pour ce nbre minimal.

1er cas: $X_1 \cap X_2$ n'est pas vide

2ème cas: $X_1 \cap X_2$ est vide

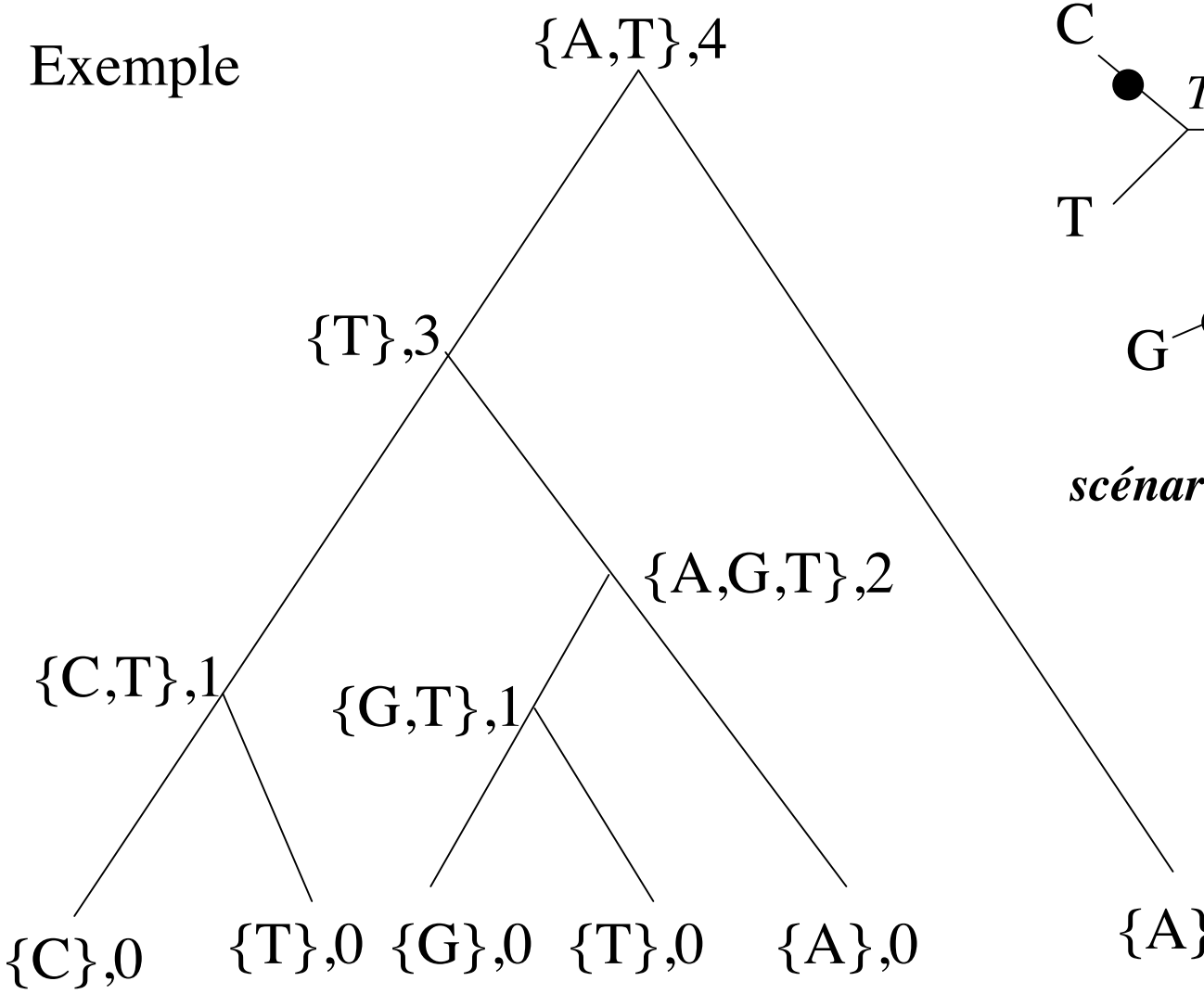


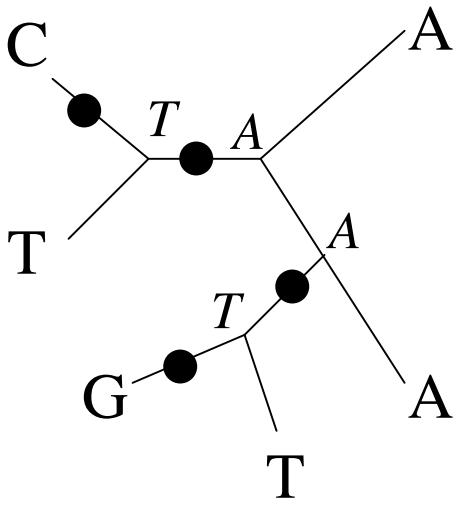
Algorithme de Fitch (suite)

Initialisation du calcul récursif aux feuilles de l'arbre

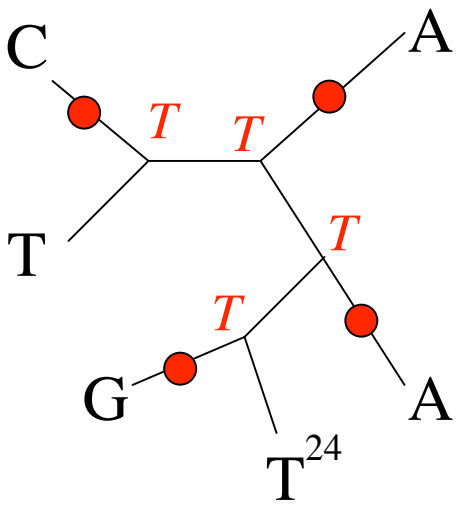
X={résidu présent à cette feuille}, P = 0

Exemple





scénario ancestral non unique !



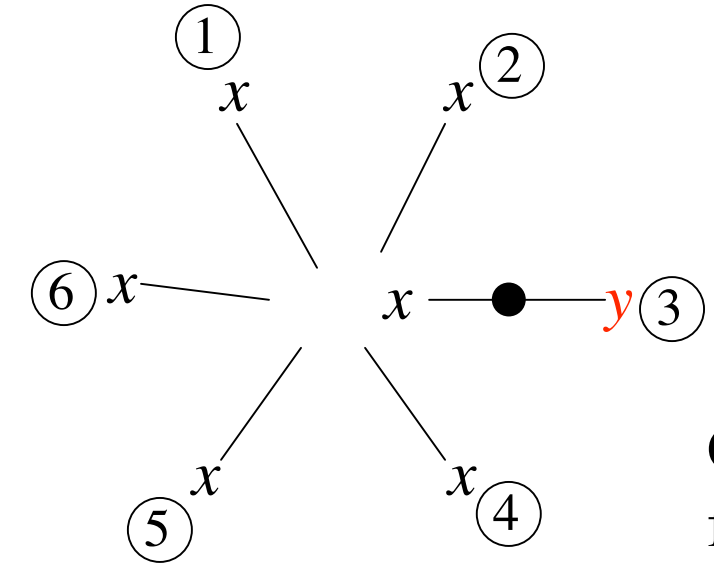
Parcimonie (2)

- Etape 2:
  - calculer *d* (étape 1) pour chaque site de l'alignement.
  - Sommer les valeurs *d* pour tous les sites.
  - Ceci donne la longueur *L* de l'arbre.

- Etape 3:
  - Calculer la valeur *L* (étape 2) pour toutes les formes d'arbre possibles.
  - Retenir l'arbre le plus court

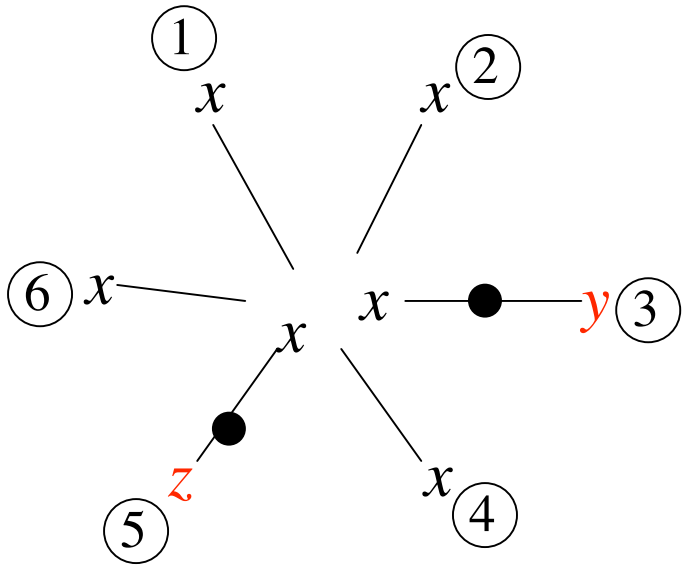
    = le (ou les) arbre(s) qui nécessite(nt) le plus petit nombre de changements
    = le (ou les) arbre(s) le(s) plus parcimonieux.

Parcimonie : sites informatifs



Quelle que soit la topologie choisie, ce site contribue 1 pas

Ces sites ne contiennent pas d'information favorisant certaines topologies d'arbre:ils sont non-informatifs. Un site est **informatif** si et seulement si au moins 2 états présents chacun au moins 2 fois.



Quelle que soit la topologie choisie, ce site contribue 2 pas

Quelques propriétés de la Parcimonie

- Conduit à des arbres sans racine.
- Algorithme et principe généraux (ADN, protéines, morphologie)
- La position des changements sur chaque branche n'est pas unique => la parcimonie ne permet pas de définir la longueur des branches de façon unique.
- Plusieurs arbres peuvent être également parcimonieux (même longueur, la plus petite de toutes).
- Le nombre d'arbres croit très vite avec le nombre de séquences traitées:

  La recherche de l'arbre le plus court doit être limitée à une fraction de l'ensemble de tous les arbres possibles => On n'a plus de certitude de trouver l'arbre le plus court.

Exemple d'heuristique d'exploration de l'espace des topologies (PHYLIP)

• Définir un ordre, arbitraire, des séquences.

• Débuter avec les 3 premières séquences et l'unique topologie possible; ajouter la séquence suivante dans toutes les positions possibles sur l'arbre courant; retenir la meilleure position.

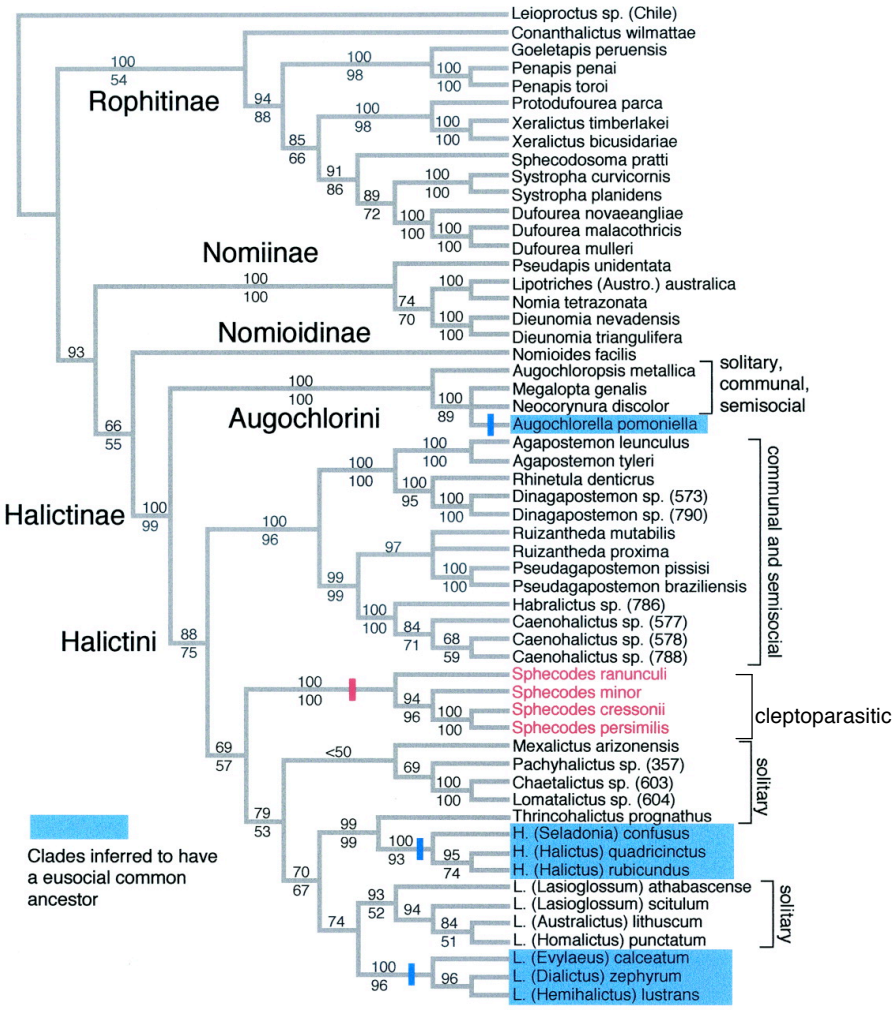
• Faire des réarrangements locaux: chaque branche interne définit 4 sous-arbres *a, b, c, d* et une topologie entre eux $\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}$; évaluer les alternatives $\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}$ et $\begin{smallmatrix} a & c \\ d & b \end{smallmatrix}$

• Recommencer tant qu'il reste des séquences à ajouter.

• Faire des réarrangements globaux: évaluer toutes les positions alternatives de chaque sous-arbre de l'arbre courant; s'arrêter quand aucune alternative n'améliore l'arbre courant.

Ceci transforme un calcul impossible (toutes les topologies) en un calcul assez rapide jusqu'à 20 ou 30 séquences. On répète souvent toute la recherche pour plusieurs ordres initiaux.

Evolution of sociality in a primitively eusocial lineage of bees



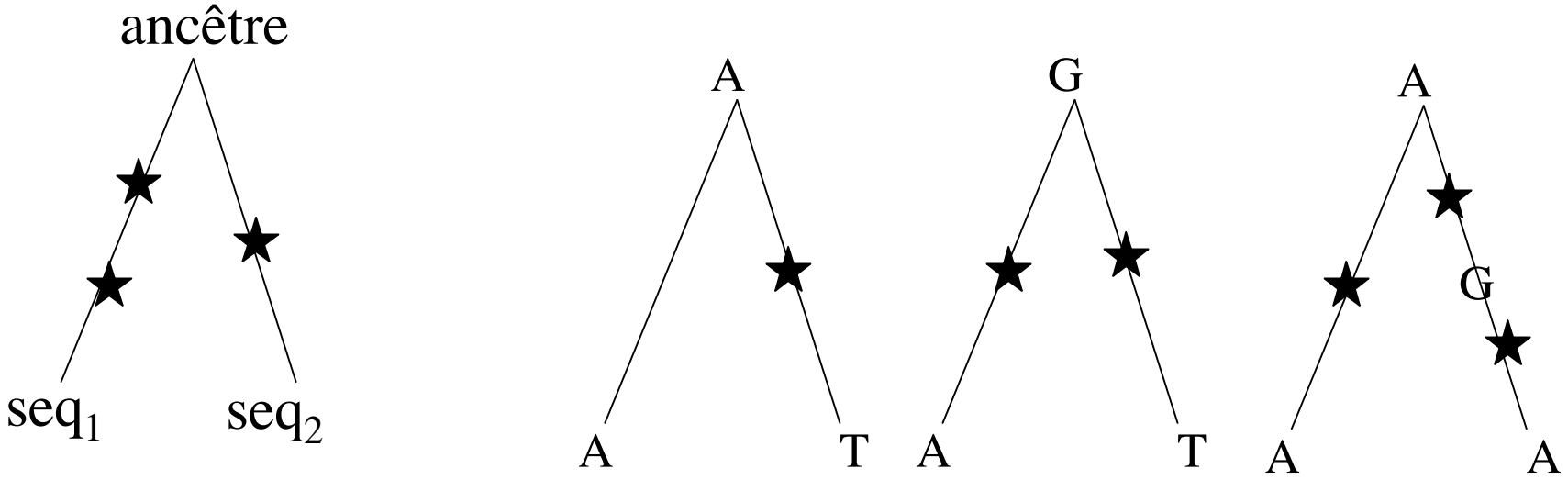
Phylogeny of the halictid subfamilies, tribes, and genera. Strict consensus of six trees based on equal weights parsimony analysis of the entire data set of three exons and two introns. Two regions within the introns were excluded because they could not be aligned unambiguously. Gaps coded as a fifth state or according to the methods described in ref. 23 yielded the same six trees. Bootstrap values above the nodes indicate bootstrap support based on the exons introns data set. Bootstrap values below the nodes indicate support based on an analysis of exons only. For the exons introns analysis the data set included 1,541 total aligned sites (619 parsimony-informative sites), the trees were 3,388 steps in length.

Advanced eusocial insects, such as ants, termites, and corbiculate bees, cannot provide insights into the earliest stages of eusocial evolution because eusociality in these taxa evolved long ago (in the Cretaceous) and close solitary relatives are no longer extant. In contrast, primitively eusocial insects, such as halictid bees, provide insights into the early stages of eusocial evolution because eusociality has arisen recently and repeatedly. I show that eusociality has arisen only three times within halictid bees.

Danforth, Bryan N. (2002) Proc. Natl. Acad. Sci. USA 99, 286-290

Distance évolutive entre 2 séquences

Définition: la distance évolutive entre 2 séquences est le nombre total de substitutions produites sur les 2 lignées depuis leur divergence divisé par le nombre de sites comparés. Elle s'exprime en substitutions/site.



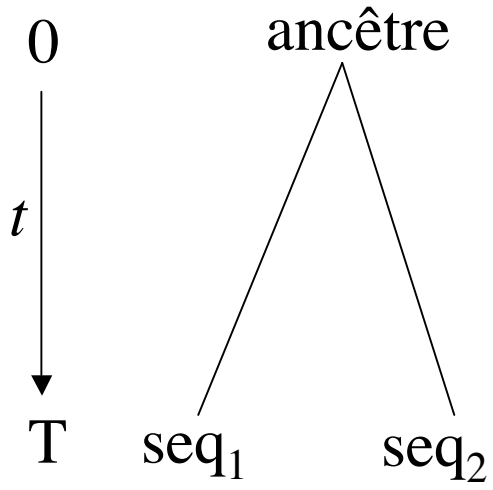
La distance évolutive n'est pas directement observable. Elle est supérieure à la divergence observée entre séquences.
Des hypothèses de régularité du processus évolutif permettent de l'estimer à partir de quantités observables.

Calcul de la distance évolutive entre 2 séquences selon le modèle de Jukes et Cantor.

Hypothèses:
- toutes les substitutions sont équiprobables
- tous les sites évoluent indépendamment selon la même loi
- la vitesse d'évolution est constante dans le temps

Donc ce modèle se résume à un unique paramètre

λ = proba subst. i --> j par unité de temps

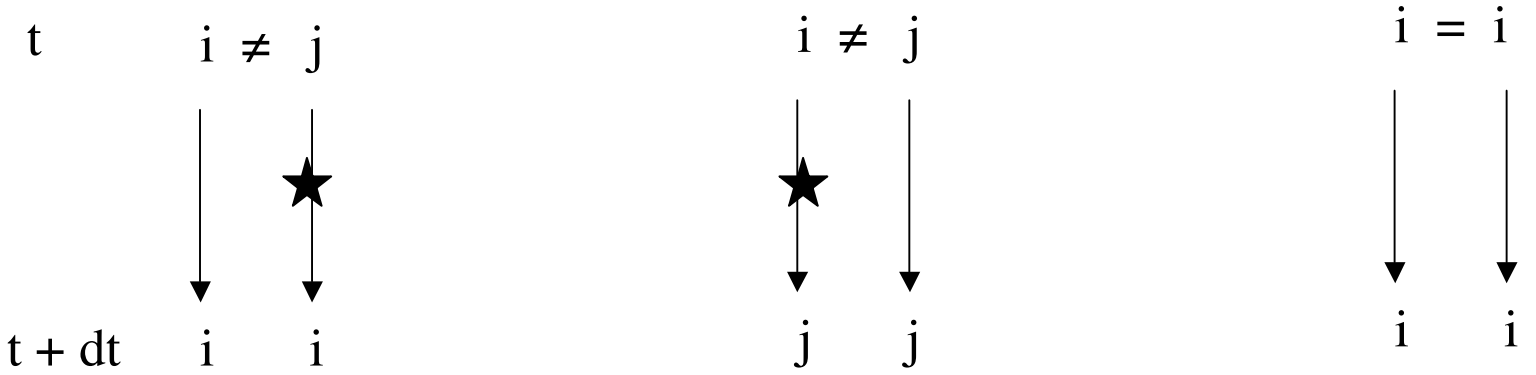


Sous ce modèle, la distance évolutive est :

$d(seq_1, seq_2) = 2 . 3\lambda . T = 6\lambda T$

Distance de Jukes & Cantor (suite)

On va calculer Q(t) = proba qu'un site contienne 2 bases identiques à l'instant t



$Q(t+dt) = (1 - Q(t))\,(1 - 3\lambda dt)\,\lambda dt + (1 - Q(t))\,\lambda dt\,(1 - 3\lambda dt) + Q(t)(1 - 3\lambda dt)^2$

$= Q(t) + 2\,\lambda dt - 8\,\lambda\, dt\, Q(t) - 6\lambda^2 dt^2 + 15\,\lambda^2\, dt^2\, Q(t)$

$Q'(t) = \lim_{dt \to 0} (Q(t+dt) - Q(t)) / dt = 2\,\lambda - 8\,\lambda\, Q(t)$ Equation différentielle 1er ordre

On a négligé les scénarios à ≥ 2 changements car $\propto dt^2$.

Solution: $Q(t) = 1 - (3/4)\,(1 - e^{-8\lambda t})$ avec condition initiale $Q(0) = 1$.

Distance de Jukes & Cantor (fin)

Soit P(t) = proba qu'un site homologue contienne 2 bases différentes

$$P(t) = 1 - Q(t) = (3/4)\,(1 - e^{-8\lambda t}) \qquad \text{[eq. 1]}$$

En inversant [1] on obtient:

$$8\lambda t = -\ln(1 - (4/3)\,P(t)) \qquad \text{[eq. 2]}$$

P(T) est estimable à partir de 2 séquences alignées :

P(T) = nbre de sites avec bases différentes / nbre de sites comparés

Sous le modèle J&C, la distance évolutive est

$$d = 6\lambda T$$
$$= -(6/8)\ln(1 - (4/3)\,P(T))$$

D'où finalement :

$$d = -(3/4)\ln(1 - (4/3)p)$$

avec

p = fraction observée de sites différents entre les séquences

ln, logarithme népérien

Effet de la correction de Jukes & Cantor

p	d
0.10	0.107
0.20	0.23
0.40	0.57
0.60	1.21
0.75	∞

Distance évolutive selon le modèle de Kimura à 2 paramètres

Hypothèses:
- toutes les transitions sont équiprobables
- toutes les transversions sont équiprobables
- tous les sites évoluent indépendamment selon la même loi
- la vitesse d'évolution est constante dans le temps

Ainsi, ce modèle s'exprime à l'aide de 2 paramètres, les taux de transitions et de transversions par unité de temps

$$d(seq1,seq2) = -(1/2)\ln[1 - 2P - Q] - (1/4)\ln[1 - 2Q]$$

où

Kimura (1980)
JMolEvol 16:111

P = fraction de sites qui présentent une transition
Q = fraction de sites qui présentent une transversion

Modélisation markovienne de l'évolution d'une séquence

On modélise l'évolution d'un site avec l'hypothèse:
il existe des taux de substitution i->j, par unité de temps, qui s'appliquent à tout instant de l'évolution.
Matrice des <u>taux instantanés de substitution</u>:

M =

↙	A	T	C	G
A	$-\lambda_A$	m_{TA}	m_{CA}	m_{GA}
T	m_{AT}	$-\lambda_T$	m_{CT}	m_{GT}
C	m_{AC}	m_{TC}	$-\lambda_C$	m_{GC}
G	m_{AG}	m_{TG}	m_{CG}	$-\lambda_G$

m_{ij} = probabilité substitution i-->j par unité de temps.
Les λ_i sont tels que somme des colonnes = 0
λ_i = proba. que i mute :

$$\lambda_i = \sum_{j \neq i} m_{ij}$$

Jukes & Cantor (1 param.)

↙	A	T	C	G
A	$-\lambda_A$	r	r	r
T	r	$-\lambda_T$	r	r
C	r	r	$-\lambda_C$	r
G	r	r	r	$-\lambda_G$

Eq. (1/4, 1/4, 1/4, 1/4)

Kimura à 2 paramètres

↙	A	T	C	G
A	$-\lambda_A$	r	r	α r
T	r	$-\lambda_T$	α r	r
C	r	α r	$-\lambda_C$	r
G	α r	r	r	$-\lambda_G$

Eq. (1/4, 1/4, 1/4, 1/4)

Tamura 92 (3 param.)

↙	A	T	C	G
A	$-\lambda_A$	$\frac{1-\theta}{2}r$	$\frac{1-\theta}{2}r$	$\alpha\frac{1-\theta}{2}r$
T	$\frac{1-\theta}{2}r$	$-\lambda_T$	$\alpha\frac{1-\theta}{2}r$	$\frac{1-\theta}{2}r$
C	$\frac{\theta}{2}r$	$\alpha\frac{\theta}{2}r$	$-\lambda_C$	$\frac{\theta}{2}r$
G	$\alpha\frac{\theta}{2}r$	$\frac{\theta}{2}r$	$\frac{\theta}{2}r$	$-\lambda_G$

Eq.((1-θ)/2, (1-θ)/2, θ/2, θ/2)

Felsenstein 84 (5 param.)

↙	A	T	C	G
A	$-\lambda_A$	$\beta\pi_A$	$\beta\pi_A$	$\alpha\frac{\pi_A}{\pi_R}+\beta\pi_A$
T	$\beta\pi_T$	$-\lambda_T$	$\alpha\frac{\pi_T}{\pi_Y}+\beta\pi_T$	$\beta\pi_T$
C	$\beta\pi_C$	$\alpha\frac{\pi_C}{\pi_Y}+\beta\pi_C$	$-\lambda_C$	$\beta\pi_C$
G	$\alpha\frac{\pi_G}{\pi_R}+\beta\pi_G$	$\beta\pi_G$	$\beta\pi_G$	$-\lambda_G$

Eq. (π_A, π_T, π_C, π_G)

$\pi_R = \pi_A + \pi_G$ $\pi_Y = \pi_C + \pi_T$

Fréquences d'équilibre d'un modèle de Markov

Chaque modèle évolutif possède une fréquence d'équilibre F_{eq}:

telle que $\frac{dF_{eq}(t)}{dt} = 0$ soit $MF_{eq} = 0$

[F_{eq} est le vecteur propre associé à la valeur propre 0 de M]

Si une séquence évolue longtemps avec des probabilités de substitution constantes, elle atteindra une composition en bases d'équilibre $F_{eq} = (\pi_A, \pi_T, \pi_C, \pi_G)$ qui restera inchangée.

Modélisation de la variation du taux d'évolution entre sites

Densité $f(r)$ de la distribution gamma:

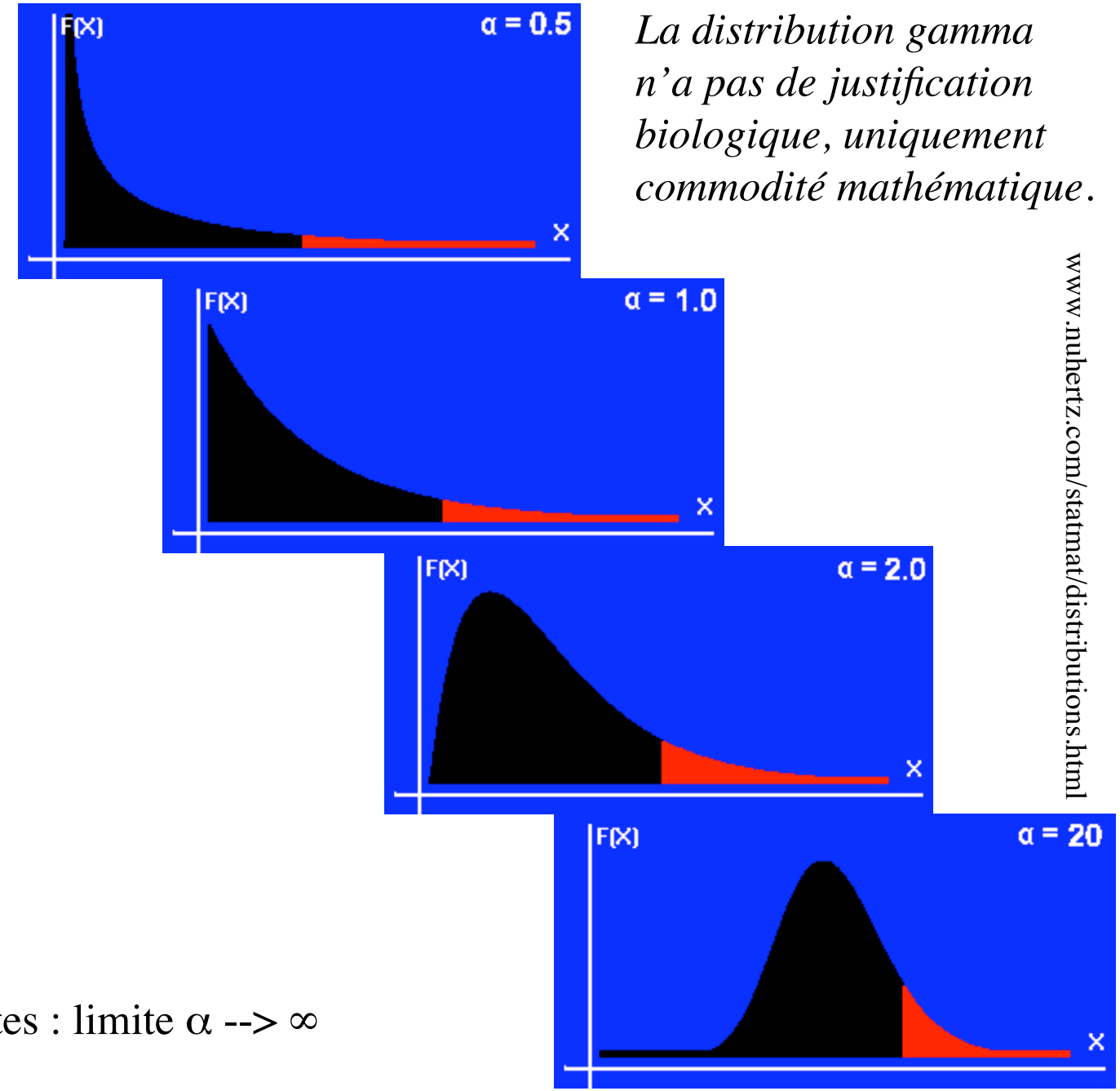
$$f(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$$

α: paramètre de forme
β: paramètre d'échelle

moyenne: αβ
variance: $\alpha\beta^2$

En phylogénie, utilisée pour modéliser la distribution des taux d'évolution entre sites avec $\beta=1/\alpha$ pour avoir
moyenne = 1
variance = $1/\alpha$

Pas de variation entre sites : limite α --> ∞



La distribution gamma n'a pas de justification biologique, uniquement commodité mathématique.

www.nuhertz.com/statmat/distributions.html

Calcul de distance évolutive avec variation du taux d'évolution entre sites (1)

Jukes & Cantor (1 param.)

↙	A	T	C	G
A	$-\lambda_A$	r	r	r
T	r	$-\lambda_T$	r	r
C	r	r	$-\lambda_C$	r
G	r	r	r	$-\lambda_G$

Chaque site évolue selon Jukes & Cantor avec un taux r qui lui varie entre les sites selon la distribution gamma.

1. Sous le modèle de Jukes & Cantor

Chaque site a la probabilité $f_\alpha(r)$ que $m_{i->j} = \lambda r$

où f_α est la densité de la distribution gamma de paramètre α et de moyenne 1,

λ taux moyen de substitution par base et par unité de temps.

$$d(seq_1, seq_2) = \frac{3}{4}\alpha\left[(1 - 4P/3)^{-1/\alpha} - 1\right]$$

avec P = fraction observée de sites différents entre les 2 séquences.

Calcul de distance évolutive avec variation du taux d'évolution entre sites (2)

Kimura à 2 paramètres

↙	A	T	C	G
A	$-\lambda_A$	r	r	κ r
T	r	$-\lambda_T$	κ r	r
C	r	κ r	$-\lambda_C$	r
G	κ r	r	r	$-\lambda_G$

Chaque site évolue selon Kimura 2P avec un taux r, qui lui varie entre les sites selon la distribution gamma, et un rapport transition/transversion κ qui est le même pour tous les sites.

2. Sous le modèle de Kimura à 2 paramètres

Chaque site a la probabilité $f_\alpha(r)$ que $m_{i->j} = \kappa\,\lambda\,r$ (i->j transition), λ r (i->j transvers.)

où f_α est la densité de la distribution gamma de paramètre α et de moyenne 1

λ taux moyen de transversion par base et par unité de temps

$$d(seq_1, seq_2) = \frac{\alpha}{4}\left[2\left(1-2P-Q\right)^{-1/\alpha} + \left(1-2Q\right)^{-1/\alpha} - 3\right]$$

Jin & Nei (1990) MBE 7:82

avec P,Q = fraction observée de sites avec transitions et transversions entre les 2 séquences

Construction d'arbres phylogénétiques par méthodes de distances

Principe général :

Alignement de séquences

⬇ (1)

Matrice des distances évolutives entre paires de séquences

⬇ (2)

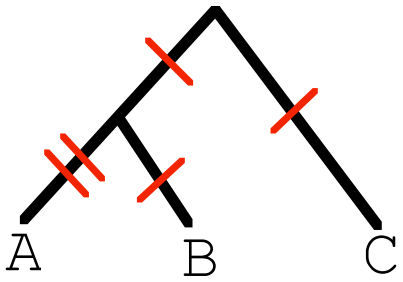
Arbre (non raciné)

(1) Mesure des distances évolutives.

(2) Calcul d'un arbre à partir des distances.

Correspondance entre arbres et matrices de distance

- Tout arbre phylogénétique induit une matrice de distances entre paires de séquences
- Une matrice de distances « parfaite » correspond à un unique arbre phylogénétique



tree

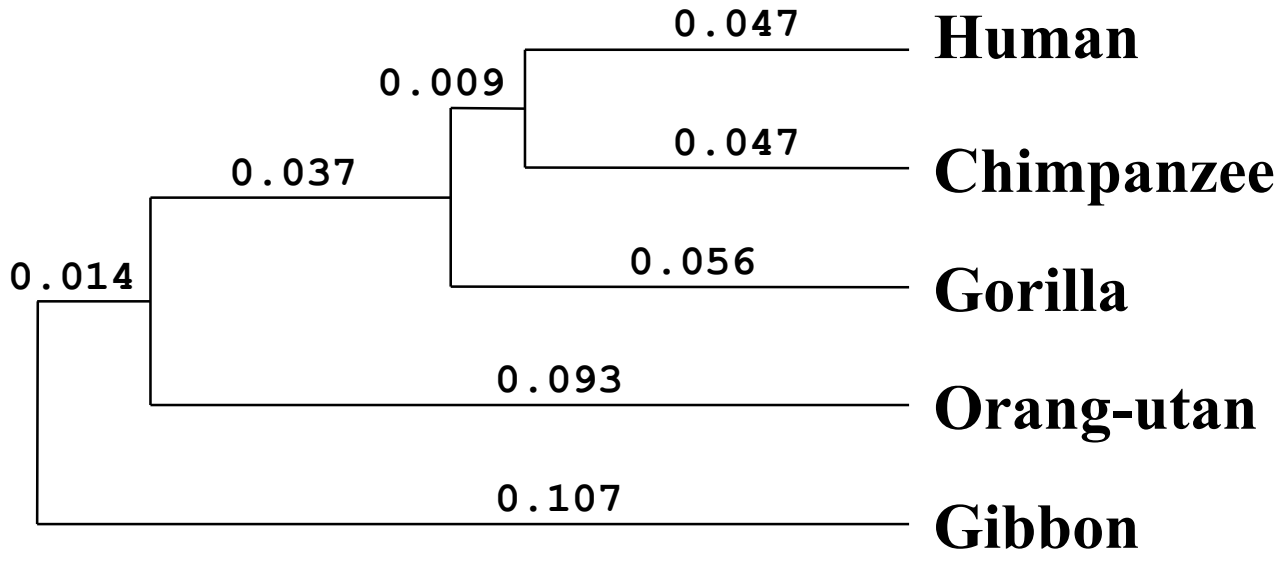
	A	B	C
A	**0**		
B	**3**	**0**	
C	**4**	**3**	**0**

distance matrix

Une (mauvaise) méthode : UPGMA

	Human	Chimpanzee	Gorilla	Orang-utan	Gibbon
Human	-	0.088	0.103	0.160	0.181
Chimpanzee	0.094	-	0.106	0.170	0.189
Gorilla	0.111	0.115	-	0.166	0.189
Orang-utan	0.180	0.194	0.188	-	0.188
Gibbon	0.207	0.218	0.218	0.216	-

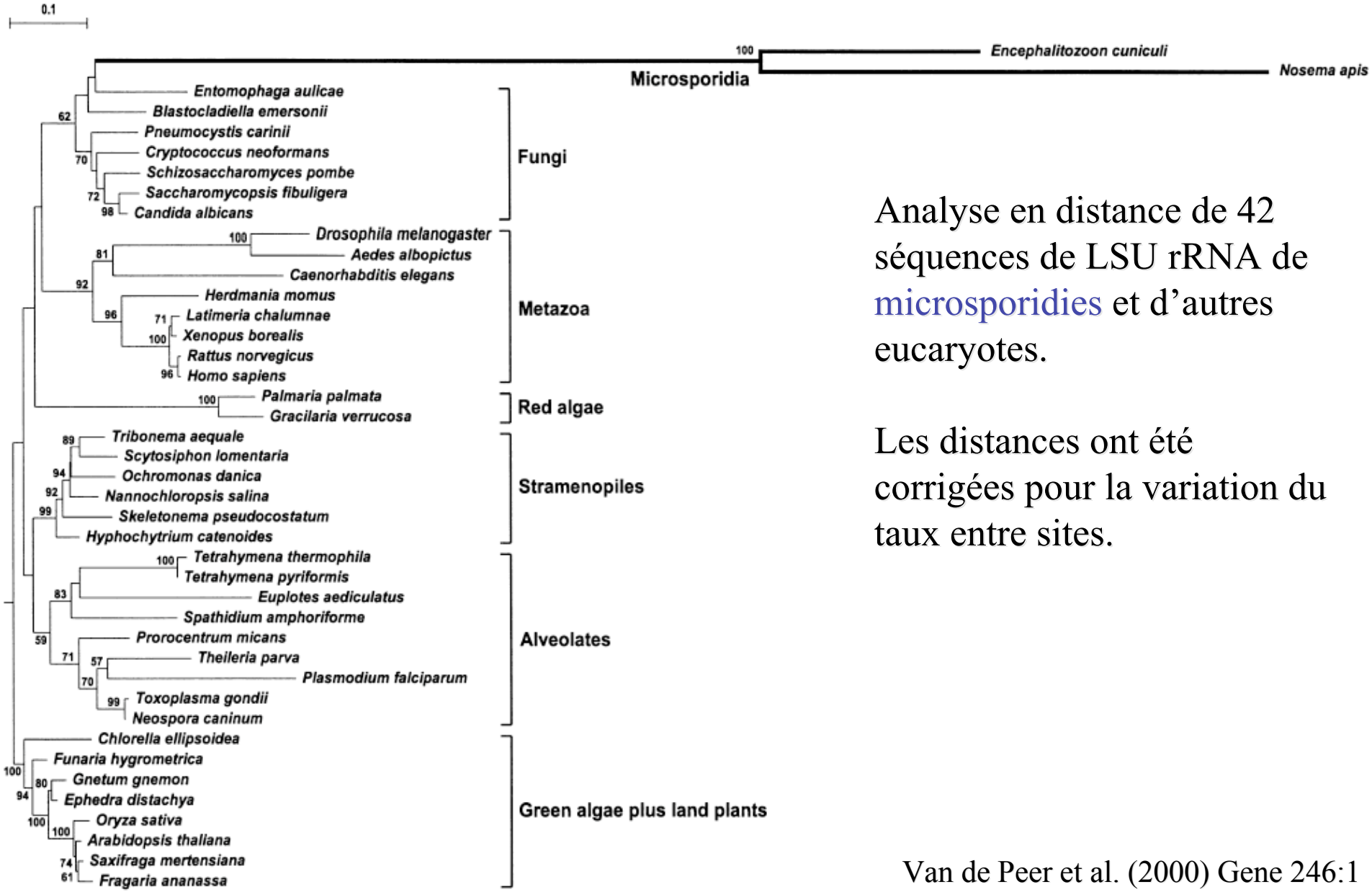
Proportion de différences (*p*) (au dessus de la diagonale) et distances de Kimura à 2 paramètres (*d*) (au dessous) pour un fragment d'ADN mitochondrial (895 pb).



Arbre UPGMA résultant

d(Gibbon,[Human+Chimp]) = 1/2 [d(Gibbon,Human) + d(Gibbon,Chimp)]

Exemple extrême de taux d'évolution variable entre lignées



Analyse en distance de 42 séquences de LSU rRNA de microsporidies et d'autres eucaryotes.

Les distances ont été corrigées pour la variation du taux entre sites.

Van de Peer et al. (2000) Gene 246:1

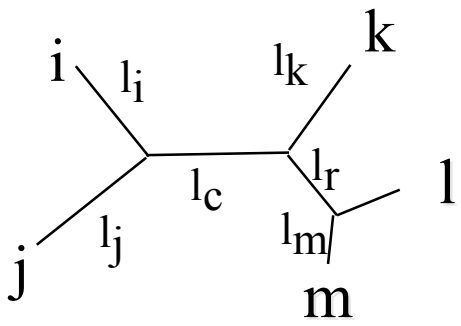
UPGMA : propriétés

- UPGMA produit un arbre raciné et des longueurs de branches.
- C'est une méthode très rapide.
- Mais UPGMA échoue si le taux d'évolution varie entre lignées.
- UPGMA n'aurait pas détecté l'origine évolutive des microsporidies parmi les champignons.

==> besoin de méthodes insensibles aux variations du taux d'évolution.

Matrice de distance -> arbre

A chaque arbre on peut associer une distance δ entre séquences :



$\delta(i,m) = l_i + l_c + l_r + l_m$

d(i,m) = distance mesurée entre les séqs i et m

Il est possible de calculer les valeurs des longueurs des branches qui optimisent la ressemblance entre δ et la distance évolutive d :

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

Solution générale de ce problème:
Rzhetsky & Nei (1993) MBE 10:1073

Il est alors possible de calculer la longueur totale de l'arbre :

S = sum of all branch lengths

forme d'arbre ==> «meilleures» longueurs des branches ==> longueur totale de l'arbre

La Méthode d' Evolution Minimale

- Pour toutes les formes d'arbre possibles :
  - Calculer sa longueur totale, S
- Choisir l'arbre dont la longueur S est minimale.

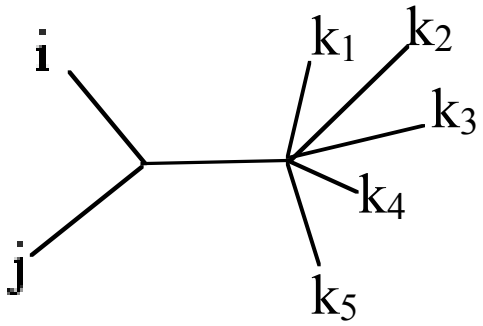
Problème: cette méthode n'est pas réalisable en pratique avec plus de ~ 25 séquences.

=> une méthode approchée (heuristique) est nécessaire.

=> ***Neighbor-Joining*** est une heuristique de "Evolution Minimale"

Neighbor-Joining : algorithme

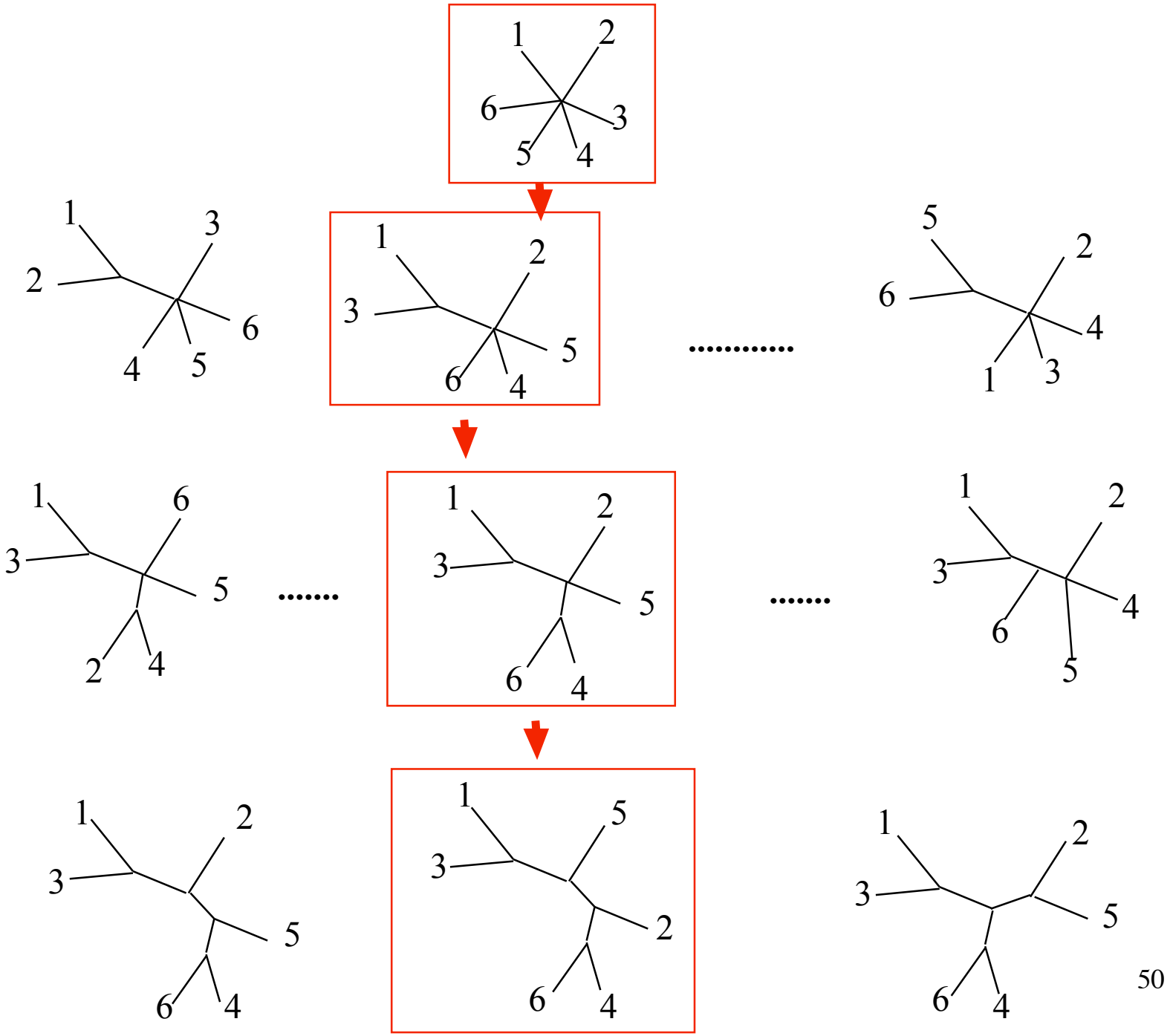
- Etape 1: Utiliser les distances *d* mesurées entre les *N* séquences
- Etape 2: Pour <u>toute</u> paire *i* et *j*: considérer la topologie en étoile suivante, et calculer $S_{i,j}$, somme des “meilleures” longueurs de branches.



$$S_{ij} = \frac{1}{2(N-2)} \sum_{\substack{k \neq i \\ k \neq j}} \left(d_{ik} + d_{jk} \right) + \frac{1}{2} d_{ij} + \frac{1}{N-2} \sum_{\substack{k<l \\ k \neq i, k \neq j \\ l \neq i, l \neq j}} d_{kl}$$

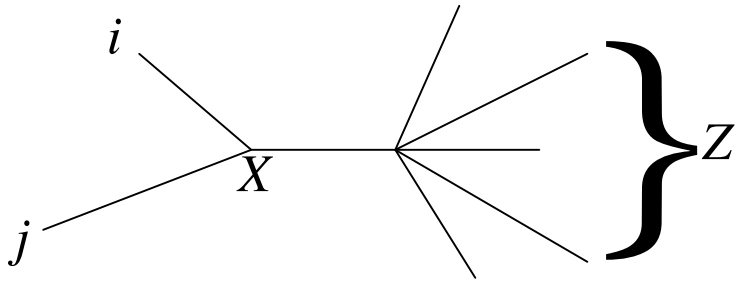
- Etape 3: Retenir la paire (*i*,*j*) de valeur $S_{i,j}$ minimale. Grouper *i* et *j* dans l’arbre.
- Etape 4: Calculer de nouvelles distances *d* entre *N*-1 objets: la paire (*i*,*j*) et les *N*-2 autres séquences : $d_{(i,j),k} = (d_{i,k} + d_{j,k}) / 2$
- Etape 5: retourner à l’étape 2 tant que $N \geq 4$.

Saitou & Nei (1987) Mol.Biol.Evol. 4:406 ; Studier & Keppler (1988) Mol.Biol.Evol. 5:729


1 2 6 3 5 4
1 3 2 6 4 5
1 2 3 5 6 4
5 2 6 4 1 3
1 6 3 5 2 4
1 2 3 5 6 4
1 2 3 4 6 5
1 2 3 5 6 4
1 5 3 2 6 4
1 2 3 5 6 4

Neighbor-Joining: calcul des longueurs des branches

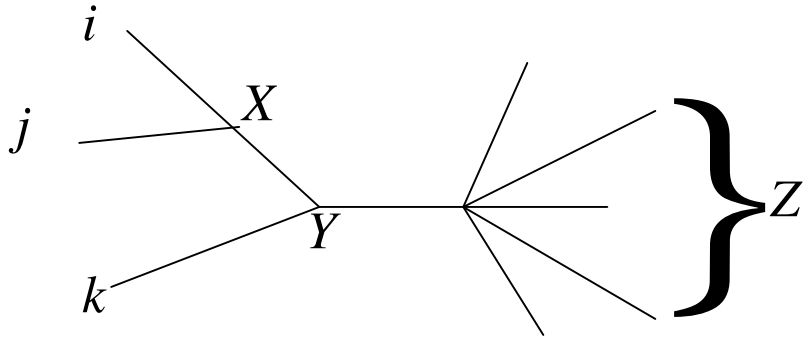
Branches périphériques



$$L_{iX} = \left(d_{ij} + d_{iZ} - d_{jZ}\right)/2$$

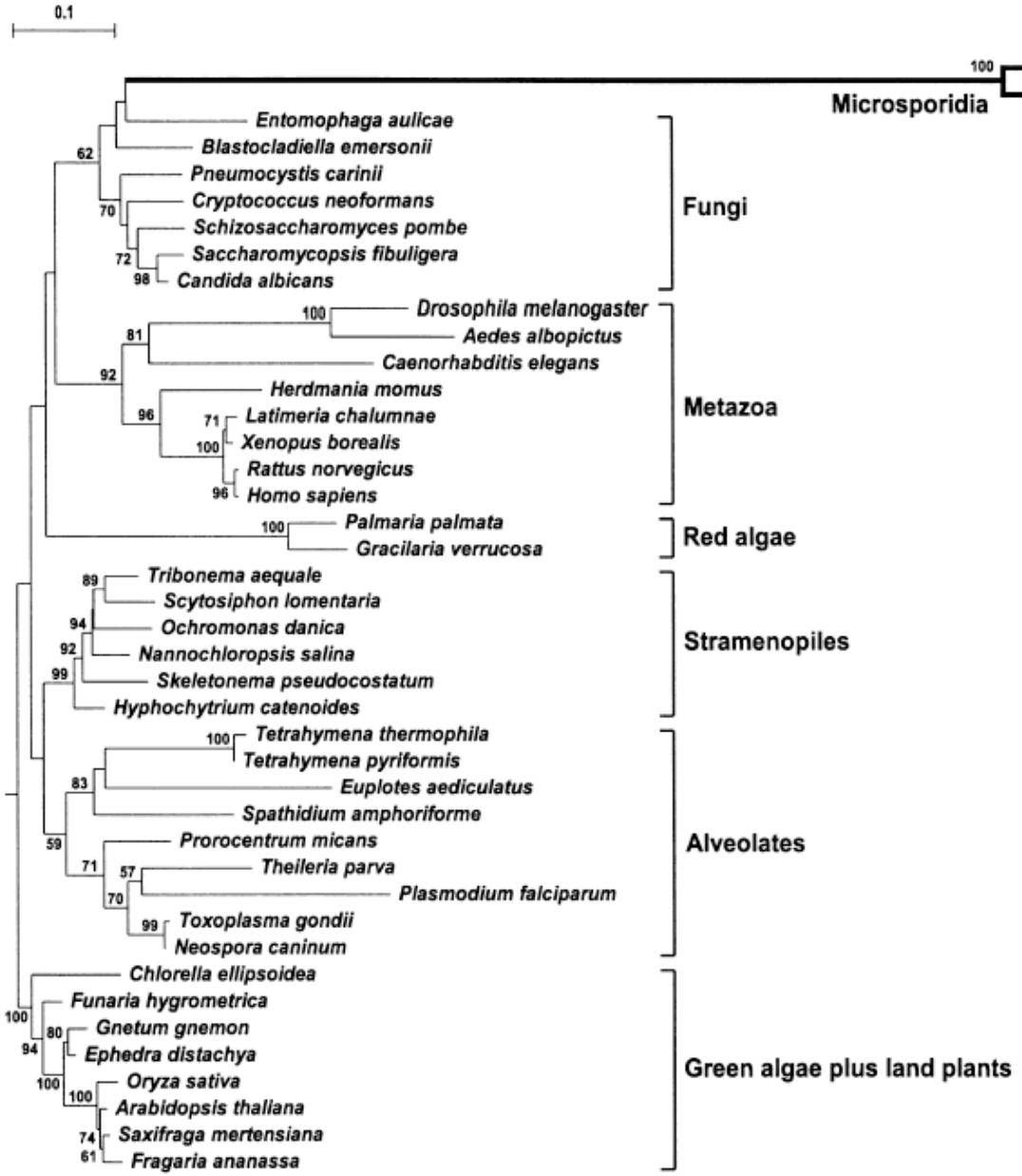
avec $d_{iZ} = \frac{1}{N-2} \sum_{\substack{k \neq i \\ k \neq j}} d_{ik}$

Branches internes



$$L_{XY} = L_{(ij)Y} - d_{ij}/2$$

Exemple d'arbre construit par Neighbor-Joining



Analyse en distance de 42 séquences de LSU rRNA de microsporidies et d'autres eucaryotes.

Les distances ont été corrigées pour la variation du taux entre sites.

Van de Peer et al. (2000) Gene 246:1

Matrice de distance -> arbre (4):

La 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 longueur totale est minimale).
- En ce sens, NJ est très similaire à la parcimonie car les longueurs de branches représentent des substitutions.
- NJ produit des arbres non racinés, qui doivent être racinés par un groupe externe.
- NJ trouve l'arbre vrai si les distances sont « arborées », même si les taux varient entre lignées. Ainsi NJ est très performant si on l'applique sur des distances bien estimées.

Méthode du Maximum de vraisemblance (1)

(programmes fastDNAml, PAUP*, PROML, PROTML, PhyML)

- Hypothèses
  - Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
  - Les sites évoluent indépendamment les uns des autres.
  - Les sites évoluent selon la même loi (on peut affaiblir cette hypothèse).
  - Les taux de substitution ne changent pas au cours du temps le long d'une branche. Ils peuvent varier entre branches.

On va présenter cette méthode dans le cas simple du modèle de Jukes & Cantor

Méthode du Maximum de vraisemblance (2)

Application du modèle de Jukes & Cantor à une branche évolutive

ancêtre $\longrightarrow$ descendant

taux de subst. λ pendant t unités de temps

Nbre de subst. attendu sur la branche: $3\lambda t$

On travaille avec $l = 3\lambda t$ = 'longueur' de la branche

On peut montrer que :

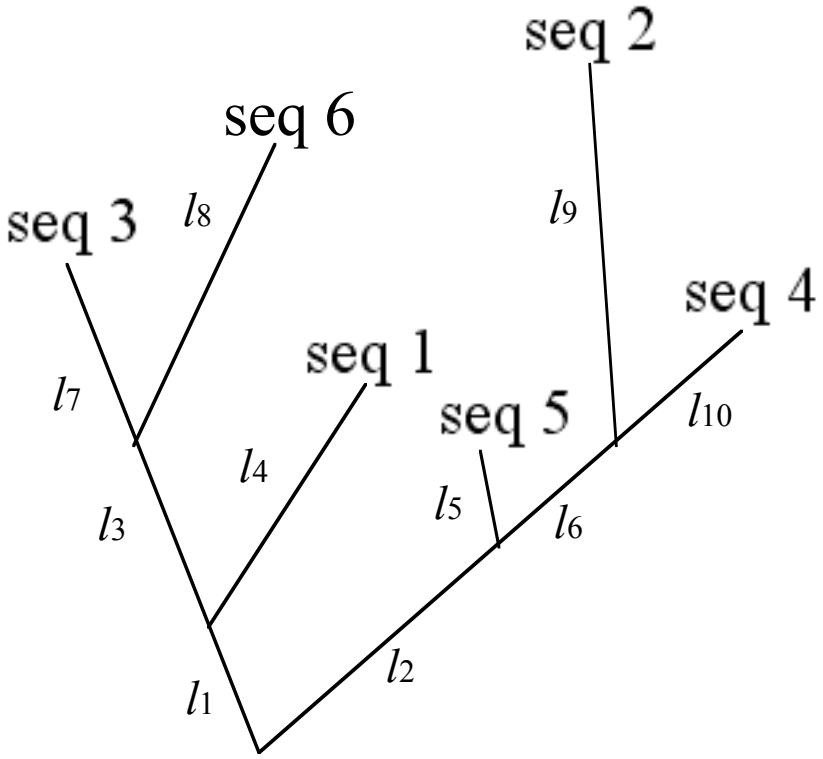
$$\text{Proba}(\text{desc} = j \mid \text{anc} = i) = (1/4)(\ 1 - e^{-l/3}\ ) \qquad \forall\, j \neq i$$

$$\text{Proba}(\text{desc} = i \mid \text{anc} = i) = (1/4)(\ 1 + 3\, e^{-l/3}\ ) \qquad \forall\, i$$

Méthode du Maximum de vraisemblance(3)

Modèle probabiliste de
l'évolution de séquences

l_b, longueur de la branche b = nbre attendu de substitutions par site le long de la branche



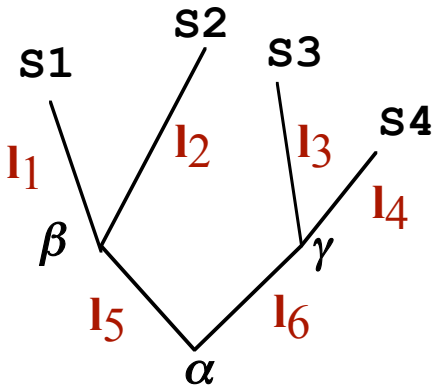
On sait calculer

$P_{branche\ b}(y \text{ en fin} \mid x \text{ en début})$

pour toutes bases x & y, toute branche b

Algorithme du maximum de vraisemblance (1)

- Etape 1: Pour une forme d'arbre racinée donnée, pour un site donné *y*, et pour un jeu de valeurs des longueurs de branches donné, on calcule la probabilité que le pattern de nucléotides observés à ce site ait évolué le long de cet arbre.



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

α, β, γ : bases ancestrales inconnues et variables

l1, *l2*, …, *l6*: longueurs des branches données

$$L(y) = \sum_{\alpha}\sum_{\beta}\sum_{\gamma} P_{anc}(\alpha)\, P_{l5}(\alpha,\beta)\, P_{l6}(\alpha,\gamma)\, P_{l1}(\beta,S1)\, P_{l2}(\beta,S2)\, P_{l3}(\gamma,S3)\, P_{l4}(\gamma,S4)$$

où $P_{anc}(S7)$ est estimée par les fréquences moyennes des bases dans les séquences.

Algorithme du maximum de vraisemblance(2)

Calcul général de la vraisemblance d'un site

$$L(y) = \sum_{i \in B} P_{anc}(r = i) L^{r,i}(y)$$

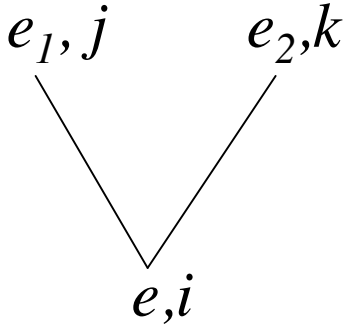
avec y : site; $B = \{A,C,G,T\}$; r : racine; P_{anc} : proba ancestrales des bases;

$L^{e,i}(y)$: vraisemblance au noeud e de l'arbre conditionnelle à base i à ce noeud

Définition récursive de $L^{e,i}(y)$

si e est un noeud interne : soient e_1 et e_1 ses 2 descendants

$$L^{e,i}(y) = \sum_{j \in B} \sum_{k \in B} P(e_1 = j \mid e = i) L^{e_1,j}(y) P(e_2 = k \mid e = i) L^{e_2,k}(y)$$



si e est une feuille :

$$L^{e,i}(y) = \begin{cases} 1 \text{ si } i \text{ est la base au site } y \text{ de la sequence } e \\ 0 \text{ sinon} \end{cases}$$

Algorithme du maximum de vraisemblance(3)

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

$$L = \prod_{\text{sites } y} L(y)$$

  C'est la vraisemblance du modèle. En pratique on calcule $log(L) = \Sigma log(L(y))$

- Etape 3: calculer les longueurs des branches $l1, l2, \ldots, l6$ et les valeurs du paramètre θ qui correspondent à la valeur maximale de L.
- Etape 4: calculer la vraisemblance de tous les arbres possibles. Retenir l'arbre associé à la plus haute vraisemblance.

Maximum de vraisemblance : propriétés

- C'est la méthode la mieux justifiée au plan théorique.
- Des expériences de simulation de séquences ont montré que cette méthode est supérieure aux autres dans la plupart des cas.
- Mais c'est une méthode très lourde en calculs.
- Il est presque toujours impossible d'évaluer tous les arbres possibles car ils sont trop nombreux. Une exploration partielle de l'ensemble des arbres est réalisée.

fastDNAml: une implémentation du principe du maximum de vraisemblance en phylogénie moléculaire [Olsen et coll. (1994) Comput Appl Biosci. 10:41-48]

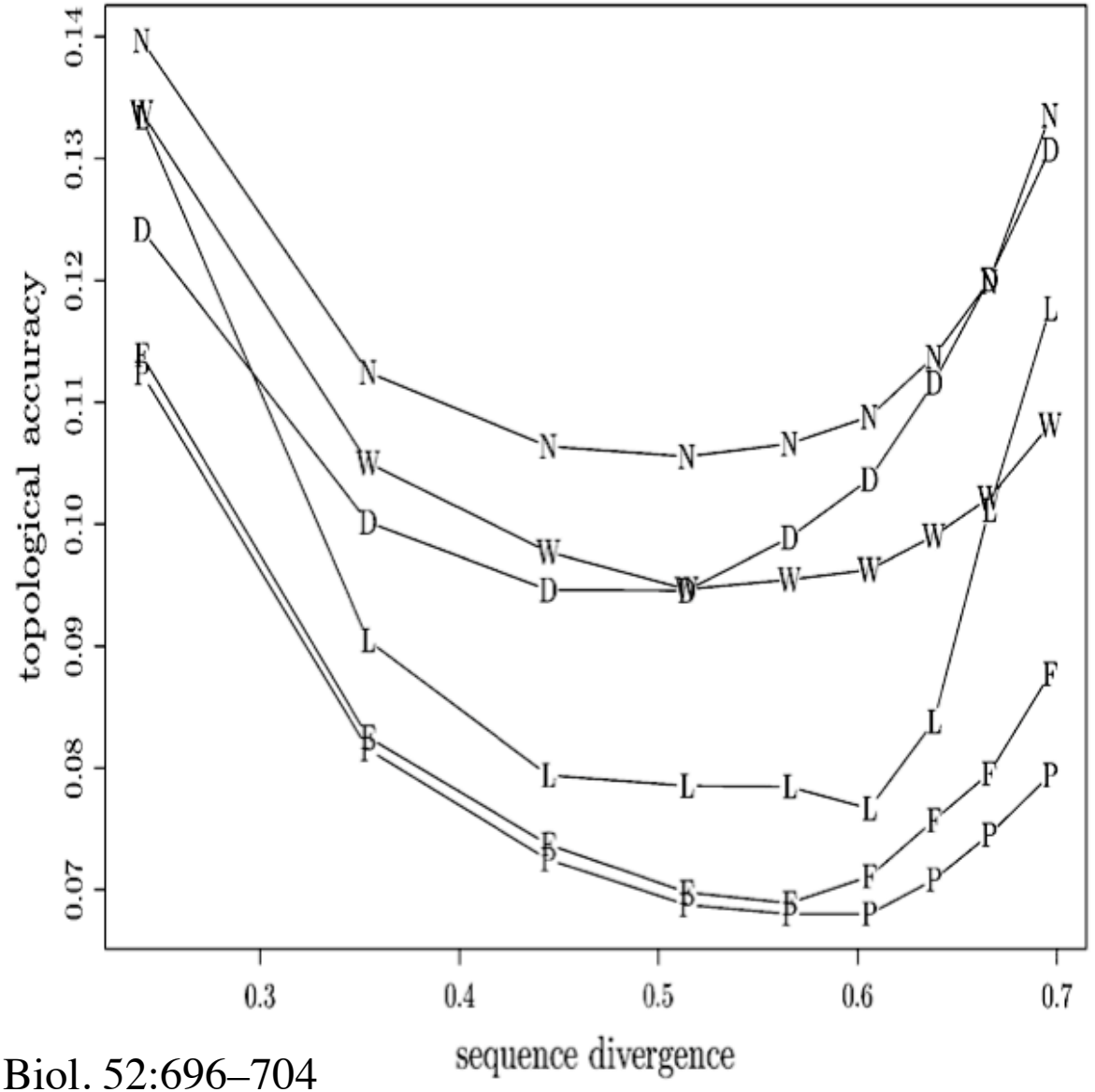
Le modèle utilisé est Felsenstein84 à 5 paramètres
π_A, π_T, π_C: fixés aux fréquences moyennes des bases dans les séqs
α : fixé *a priori* par l'utilisateur
$r=\beta t$: paramètre de longueur, estimé au max. de vraisemblance pour chaque branche

Exploration des topologies d'arbres:
chaque sous-arbre de l'arbre candidat est déplacé d'un certain nombre de noeuds, la vraisemblance de cette topologie candidate est calculée; ceci continue jusqu'à ce qu'aucun déplacement n'améliore la vraisemblance.
L'utilisateur limite le nombre maximum de noeuds franchis.

Comparaison des performances des méthodes par expériences de simulation de séquences et d'arbres

P, PHYML
F, fastDNAml
L, NJML
D, DNAPARS
N, NJ

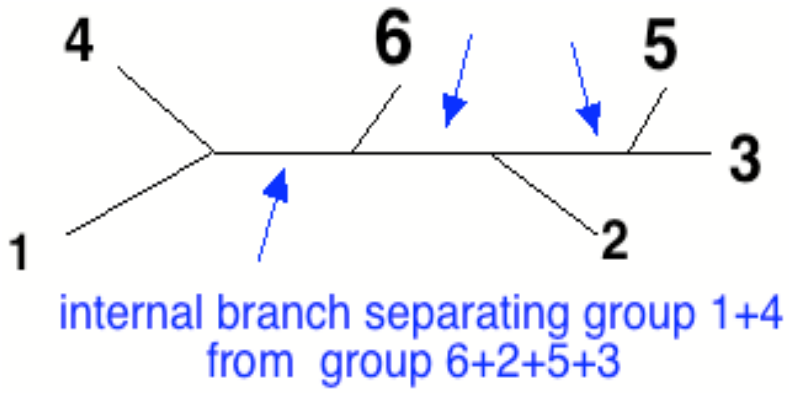
5000 arbres aléatoires
40 taxons, 500 bases
pas d'horloge moléculaire
Niveau de divergence variable
K2P, $\alpha = 2$



Guindon & Gascuel (2003) Syst. Biol. 52:696–704

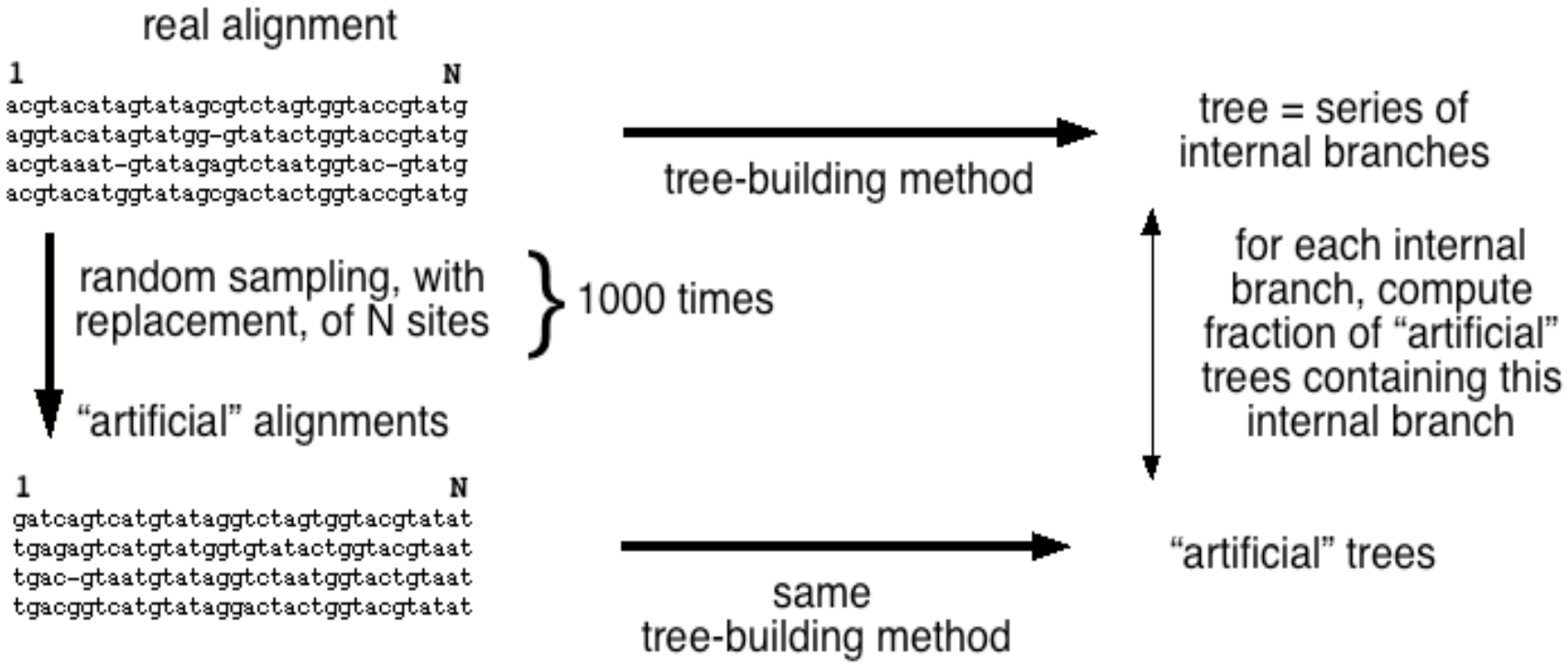
Fiabilité des arbres phylogénétiques: le bootstrap

- L'information phylogénétique contenue dans un arbre non raciné réside entièrement dans ses branches internes.



- La forme de l'arbre est déterminée par la liste des branches internes.
- Evaluer la fiabilité d'un arbre = évaluer celle de chaque branche interne.

Procédure de bootstrap (1)



Le soutien de chaque branche interne est exprimé en pourcentage de réplications.

Procédure de bootstrap (2)

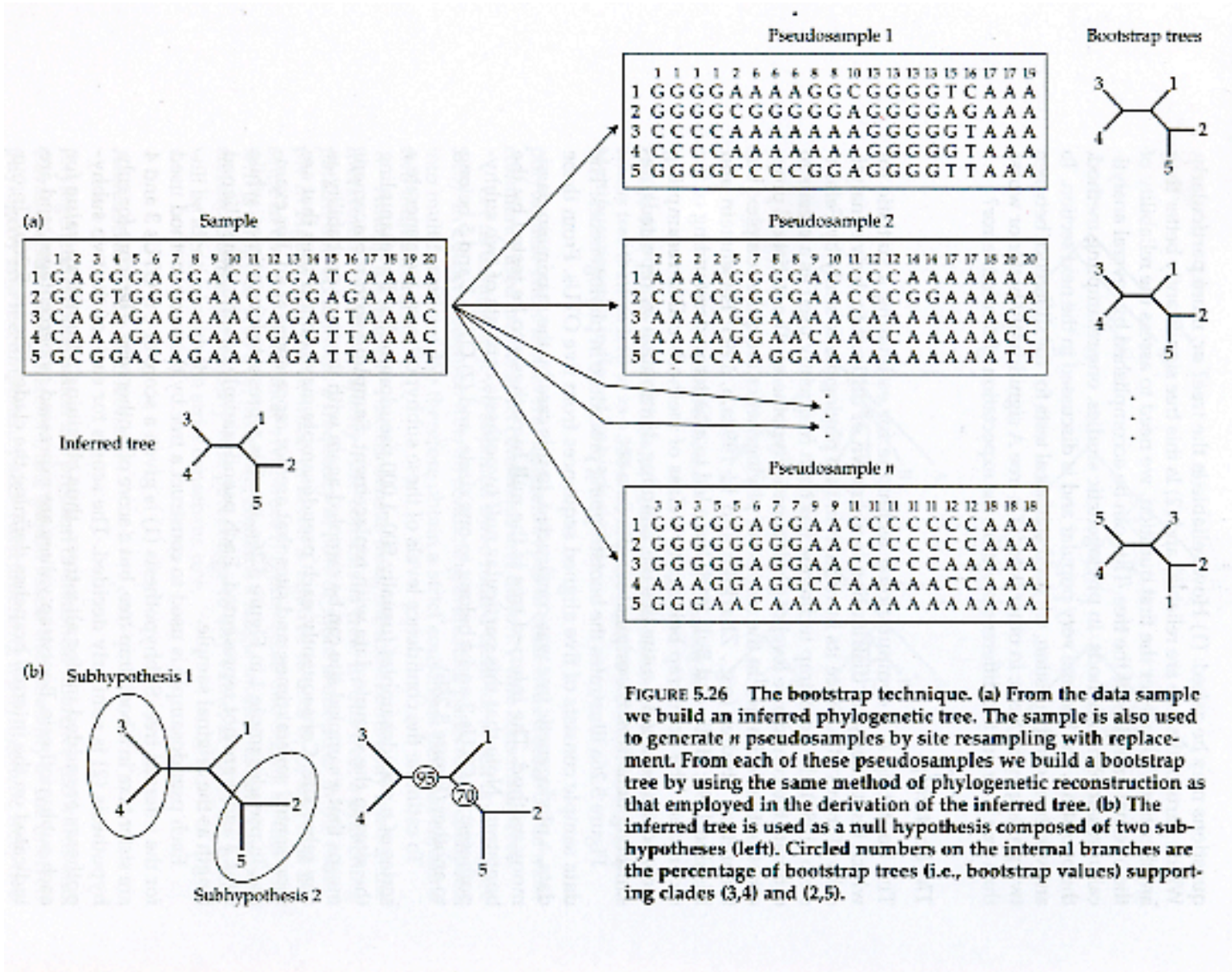
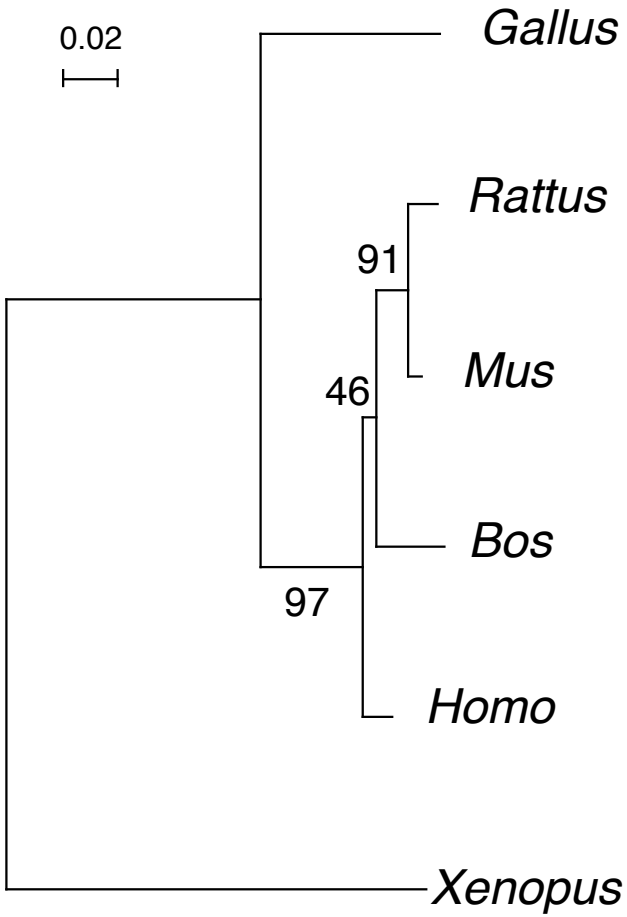


FIGURE 5.26 The bootstrap technique. (a) From the data sample we build an inferred phylogenetic tree. The sample is also used to generate *n* pseudosamples by site resampling with replacement. From each of these pseudosamples we build a bootstrap tree by using the same method of phylogenetic reconstruction as that employed in the derivation of the inferred tree. (b) The inferred tree is used as a null hypothesis composed of two subhypotheses (left). Circled numbers on the internal branches are the percentage of bootstrap trees (i.e., bootstrap values) supporting clades (3,4) and (2,5).

Arbre "bootstrappé”



Procédure de bootstrap : propriétés

- Les branches internes soutenues par ≥ 90% des réplications sont statistiquement significatives.
- La procédure de bootstrap détecte si les séquences sont suffisamment longues pour soutenir un nœud donné.
- La procédure de bootstrap n'aide pas à déterminer si la méthode de construction d'arbre est bonne. Un arbre faux peut avoir un score de bootstrap de 100 % pour chacune de ses branches !

Comparaison des temps d'exécution de divers algorithmes de phylogénie

TABLE 1. Average run times for various methods. The computing times were measured on a 1.8-GHz (1 Go RAM) PC with Linux. For PHYML, the number in parentheses is the average number of refinement stages.

	Simulations		Real data	
Method	40 taxa (500 bp)	100 taxa (500 bp)	218 taxa (4,182 bp)	500 taxa (1,428 bp)
DNADIST+ NJ/BIONJ	0.3 sec	2.3 sec	50 sec	2 min, 19 sec
DNADIST+ Weighbor	1.5 sec	22 sec	4 min, 52 sec	58 min, 40 sec
DNAPARS	0.5 sec	6 sec	4 min, 4 sec	13 min, 12 sec
PAUP*	3 min, 21 sec	1 hr, 4 min		
PAUP*+ NJ	1 min, 10 sec	22 min	10 hr, 50 min	
MrBayes	2 min, 6 sec	32 min, 37 sec		
fastDNAml	1 min, 13 sec	26 min, 31 sec		
NJML	15 sec	6 min, 4 sec		
MetaPIGA	21 sec	3 min, 27 sec	4 hr, 45 min	9 hr, 4 min
MetaPIGA+ NJ	6 sec	23 sec	1 hr, 40 min	3 hr
PHYML	2.7 sec (6.4)	12 sec (8.3)	8 min, 13 sec (15)	11 min, 59 sec (13)

distance < parcimonie ~ PHYML << bayesien < MV classique
NJ DNAPARS PHYML MrBayes fastDNAml,PAUP*

Guindon & Gascuel (2003) Syst. Biol. 52:696–704