http://pbil.univ-lyon1.fr/members/mgouy/cours/MasterSante&Pop/cours2010.pdf

Méthodes statistiques et informatiques en phylogénie moléculaire

Master 2 : « Santé et Populations»

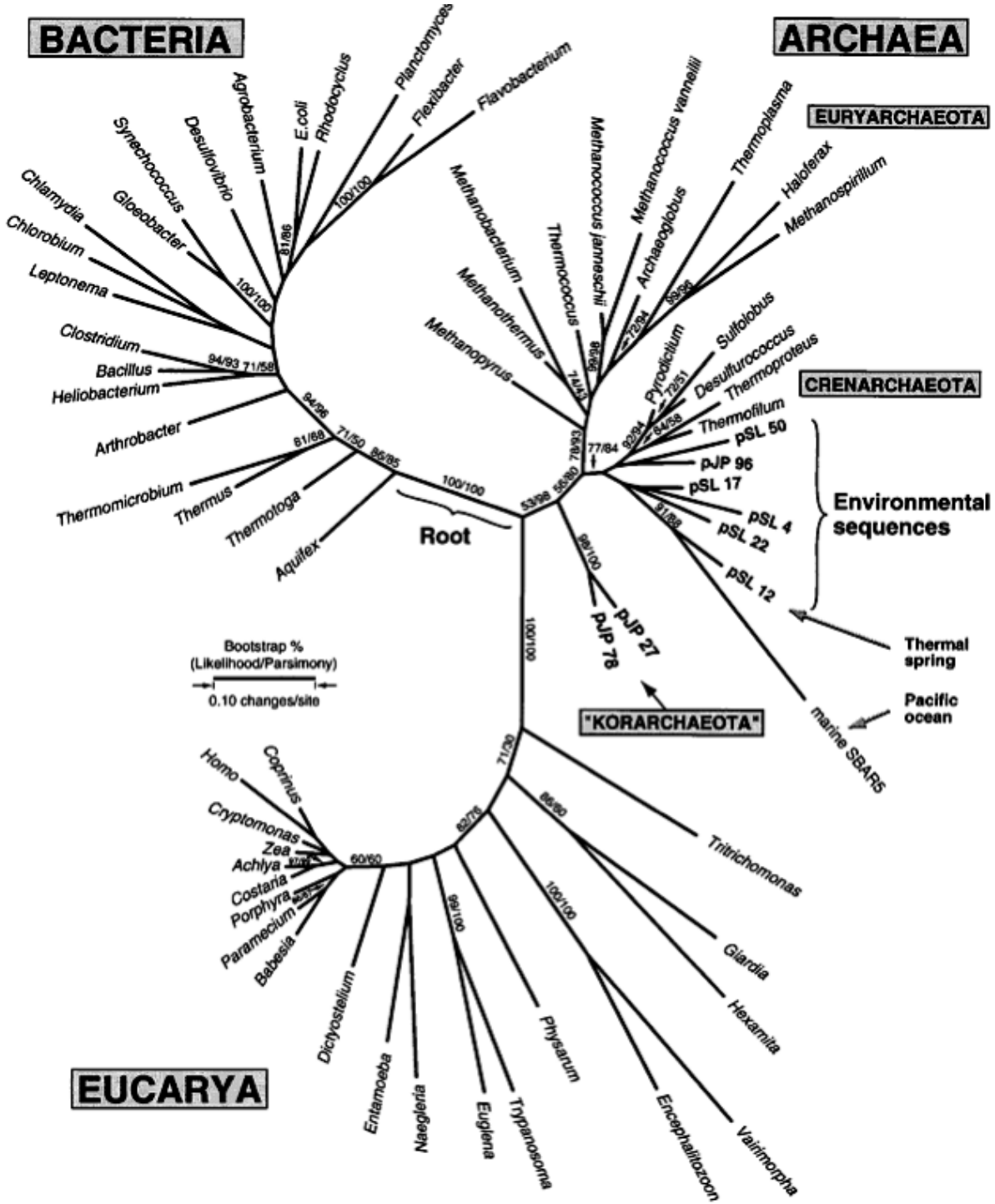
spécialité

« Biostatistique, Bioinformatique, Génome »

M2_6 : « Evolution moléculaire et phylogénie »

Manolo Gouy

Septembre 2010



La phylogénie universelle

ARNr petite sous-unité

Barns et al. (1996) PNAS 93:9188

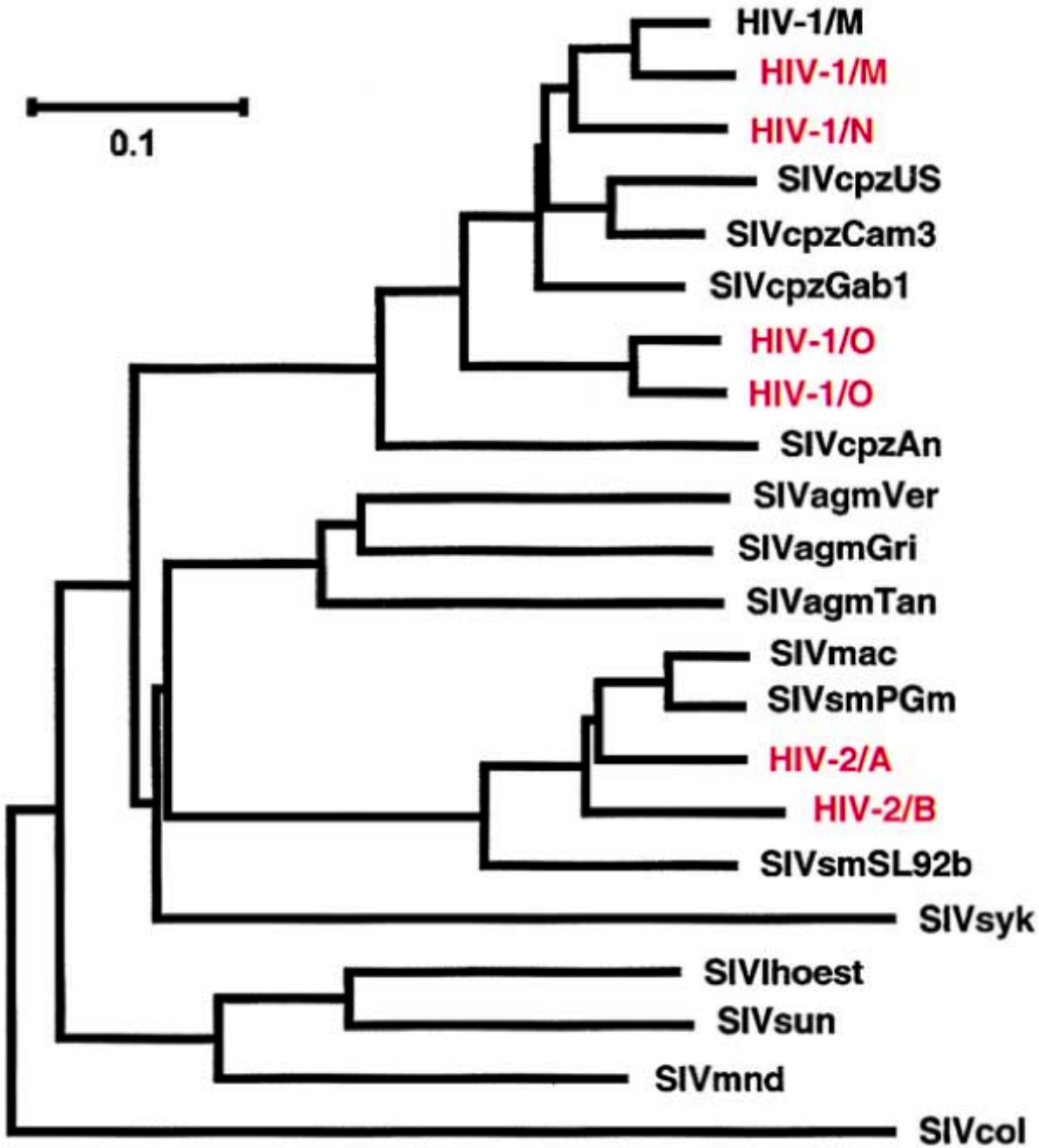


Figure 1. Evolution of AIDS Viruses

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

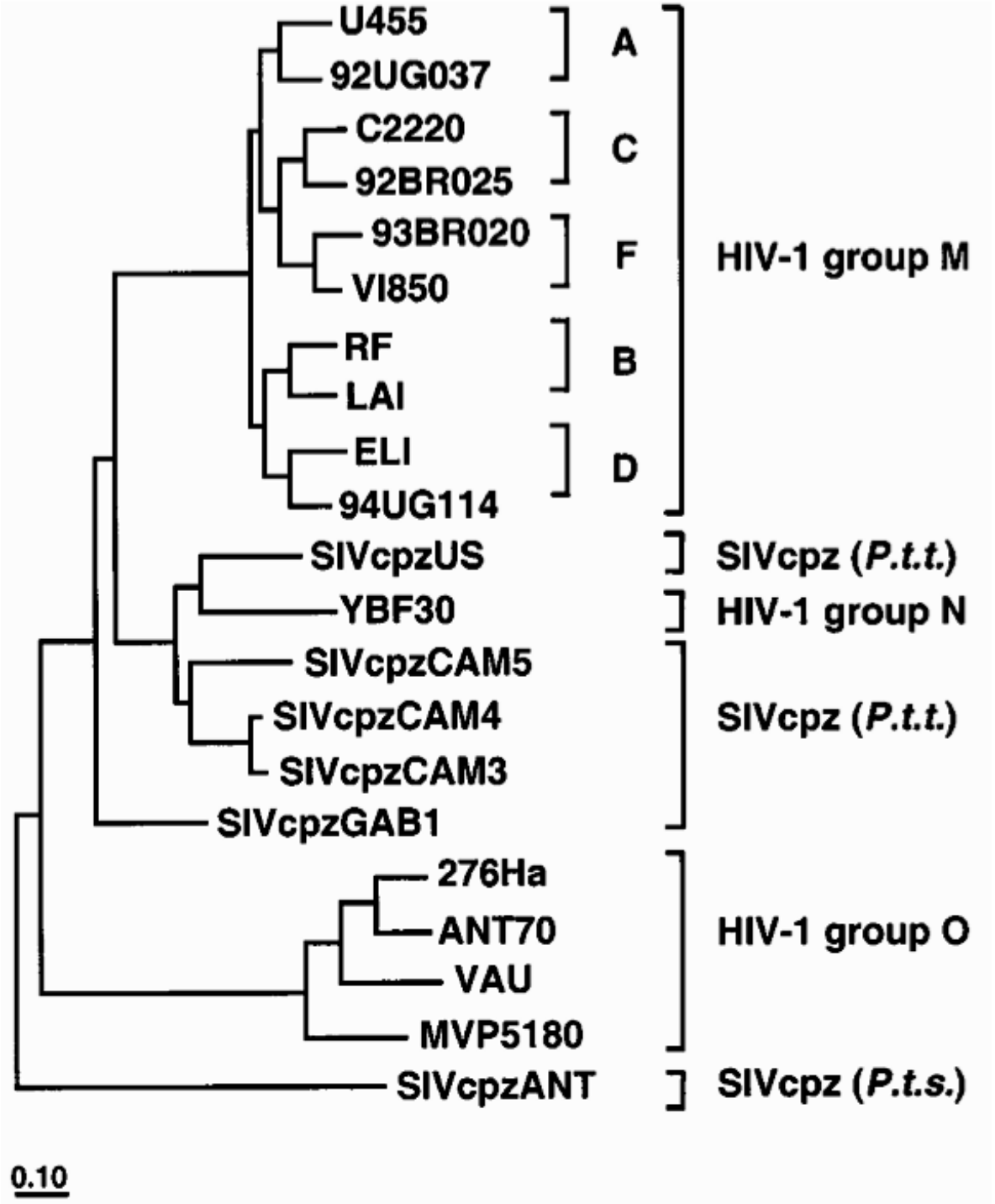


Fig. 2. Evolutionary relationships of members of the HIV-1/SIVcpz lineage based on maximum-likelihood phylogenetic analysis of full-length Env protein sequences (*52*). The three groups of HIV-1 (M, N, and O) are indicated by brackets at the right, as are five representative subtypes of the M group (A through F). The SIVcpz strains were isolated from either *P. t. troglodytes* (*P.t.t.*) or *P. t. schweinfurthii* (*P.t.s.*) animals. The scale bar indicates 0.1 amino acid replacement per site after correction for multiple hits (*52*).

Hahn et al. (2000)
Science 287:607

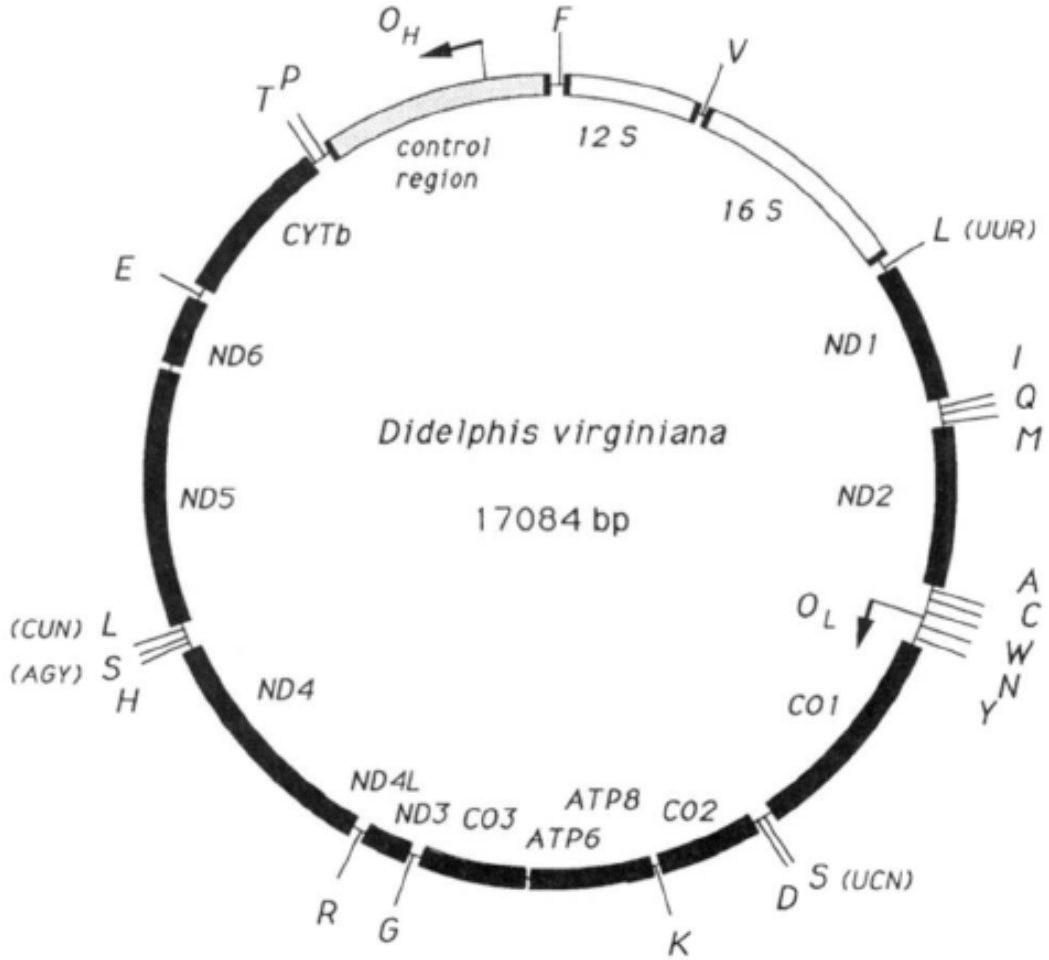
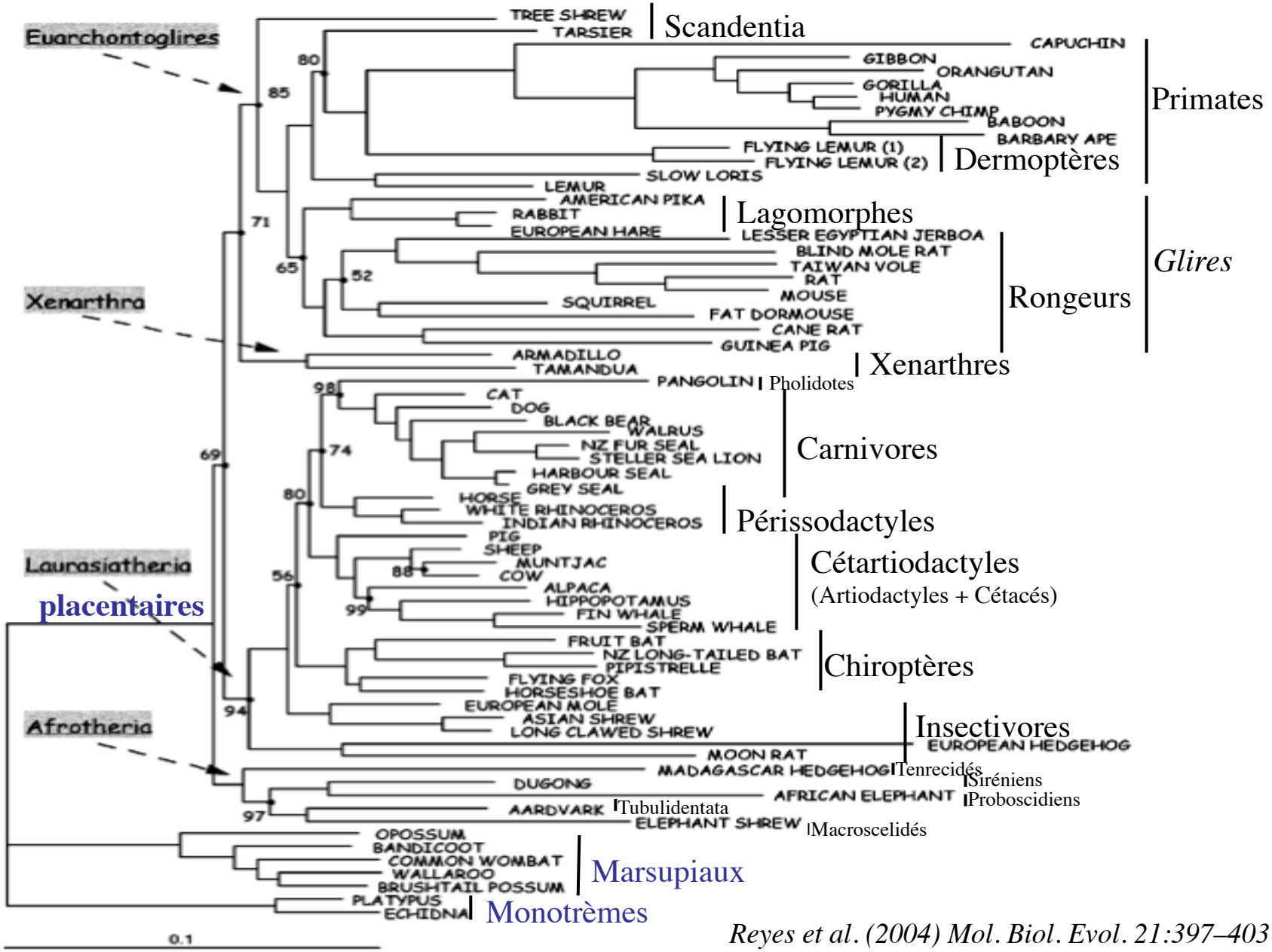


FIGURE 1.—Map of the *D. virginiana* mitochondrial DNA molecule. The location of origins of replication as well as the identity and arrangement of the various genes were determined by comparison of published mammalian sequences. Each tRNA is identified by its one-letter amino acid code. The tRNAs for 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

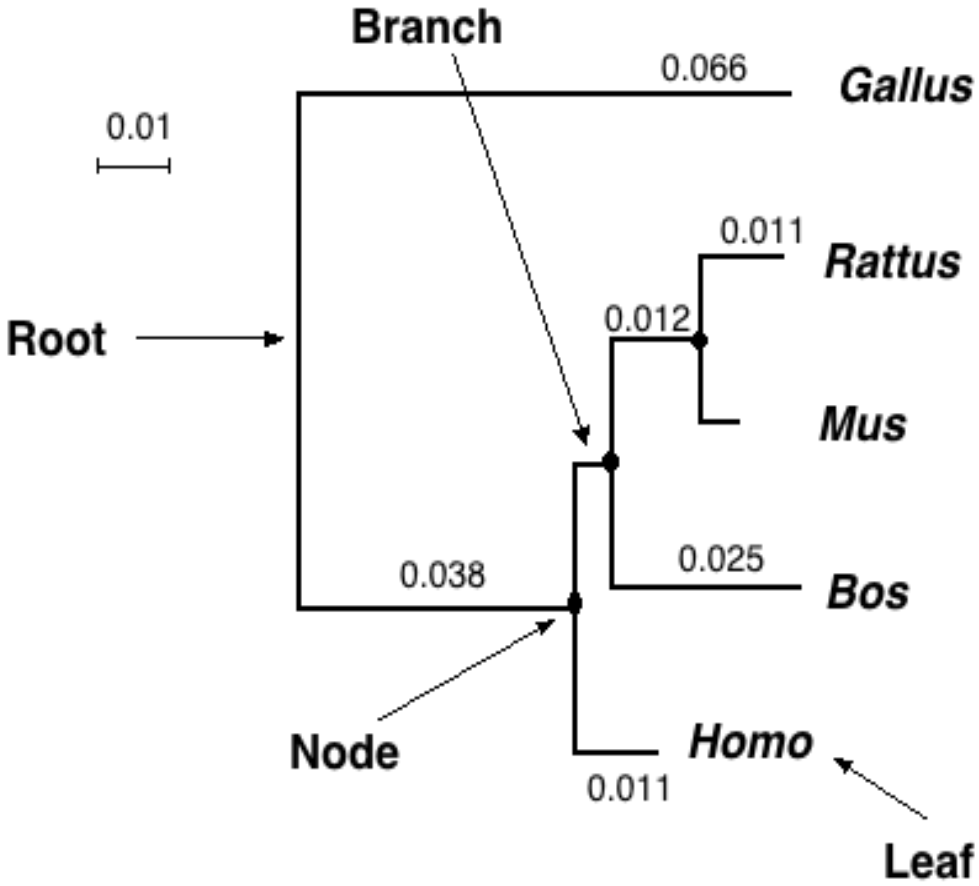
  = une généalogie de séquences
  = un arbre phylogénétique

```
CLUSTAL W (1.74) multiple sequence alignment

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

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.

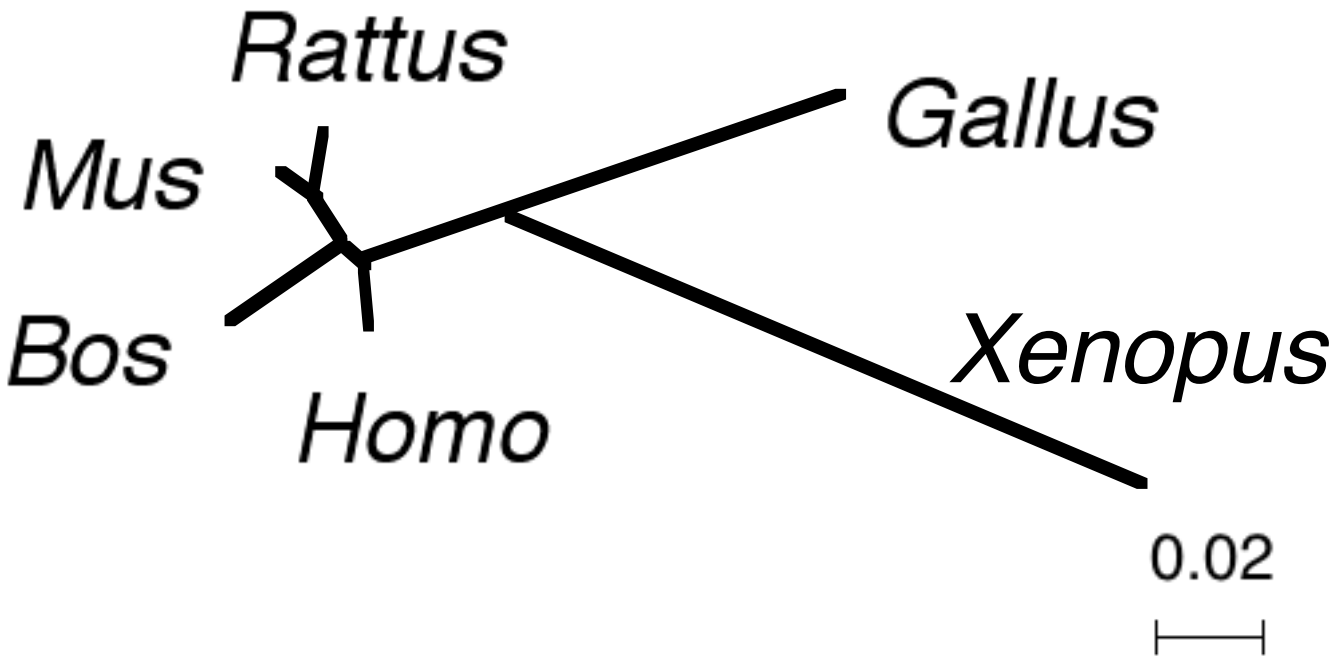
  ==> les sites contenant des gaps sont ignorés.

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

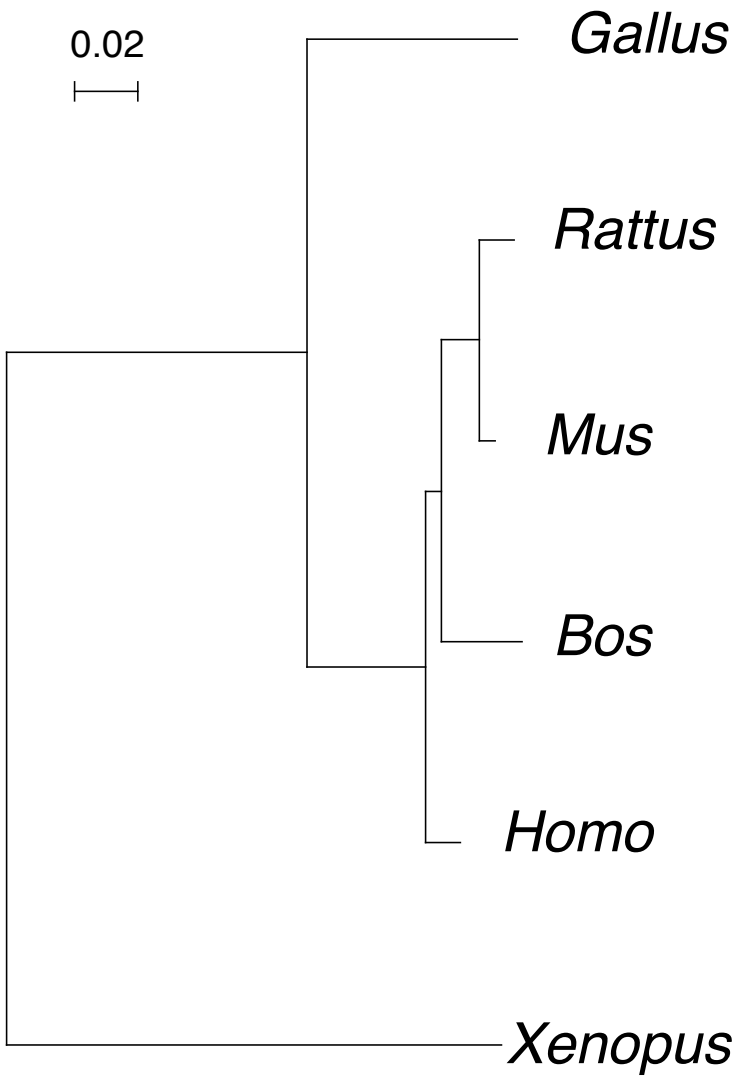
Arbres racinés et non-racinés

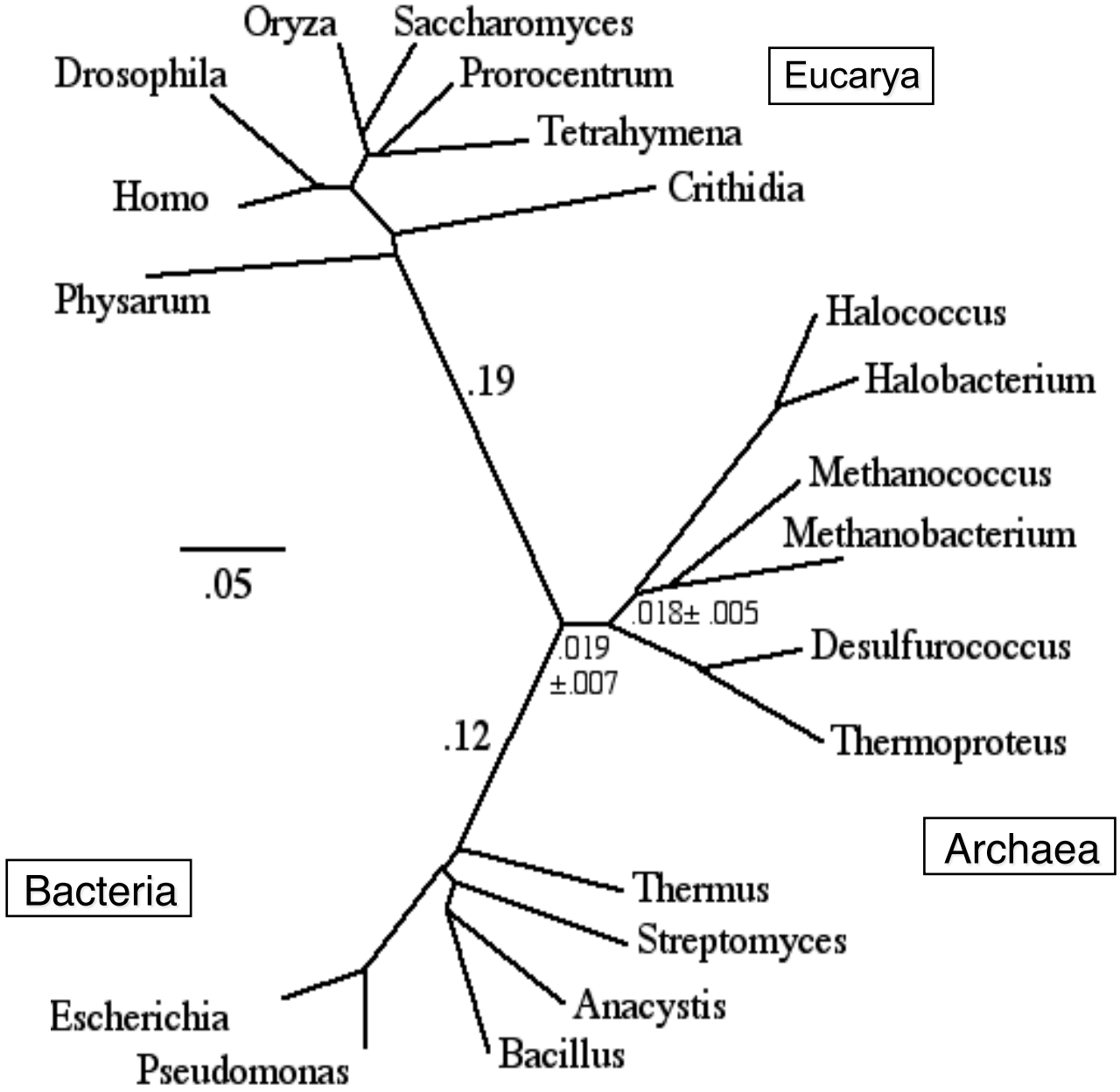
- La plupart des méthodes phylogénétiques produisent des arbres non racinés. La raison est que les méthodes détectent des différences entre séquences, sans avoir le moyen de les orienter temporellement.
- Deux façons d'enraciner un arbre non raciné:
  - Méthode du groupe externe : inclure dans l'analyse 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



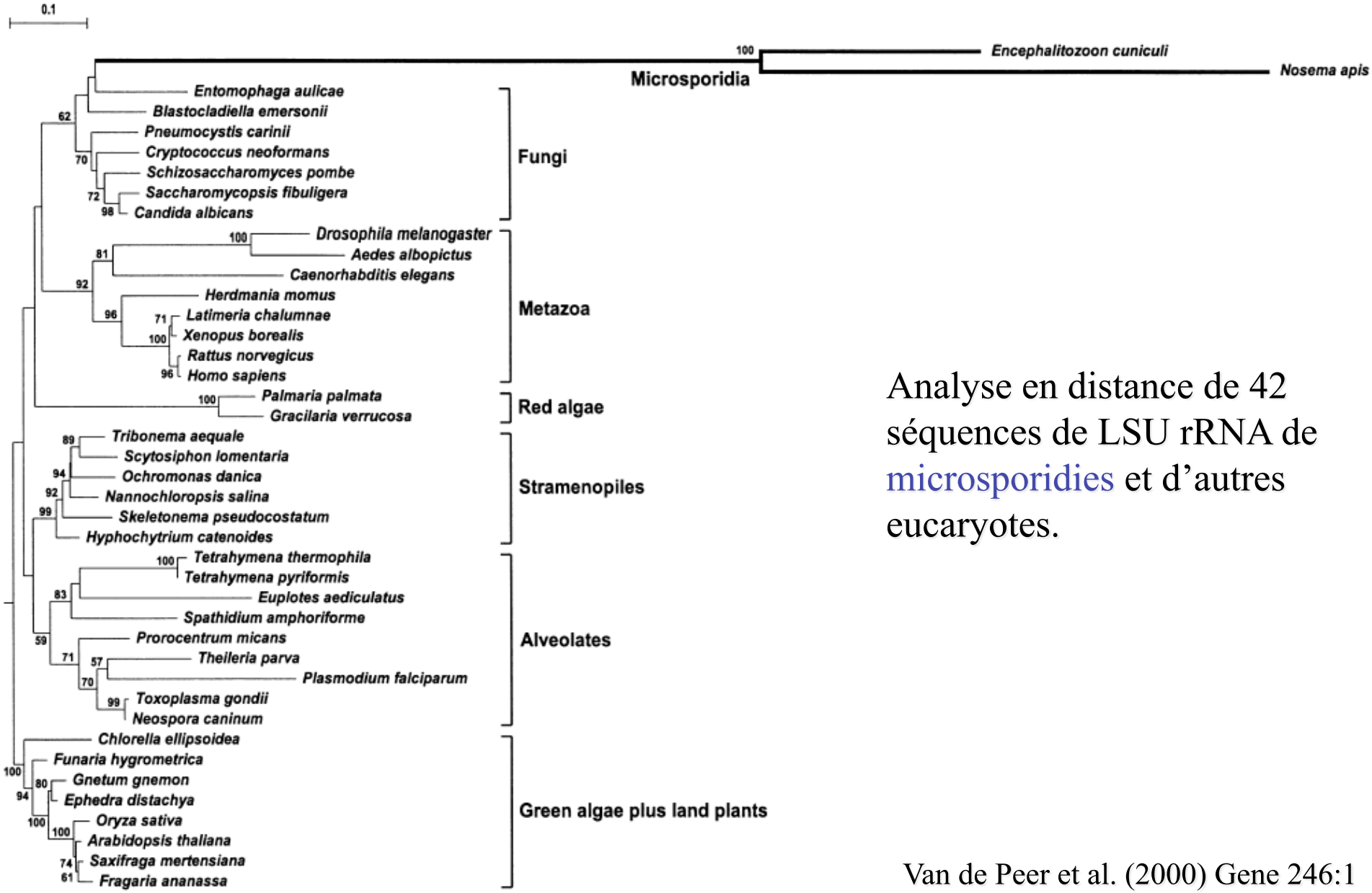


Phylogénie universelle

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

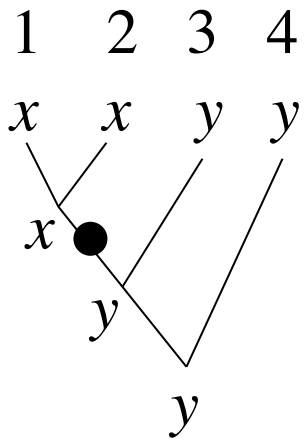
Quatre familles principales de méthodes :

- Parcimonie
- Méthodes de distances
- Maximum de vraisemblance
- Méthodes bayesiennes

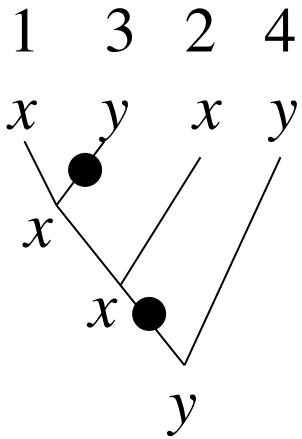
Pourquoi la parcimonie ?

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 ?

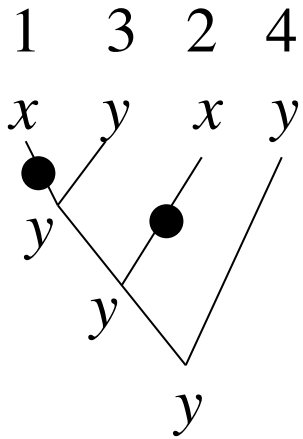
Egalité par ascendance commune:
deux espèces possèdent le même état de caractère car elles l'ont hérité sans le transformer de leur dernier ancêtre commun



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



réversion



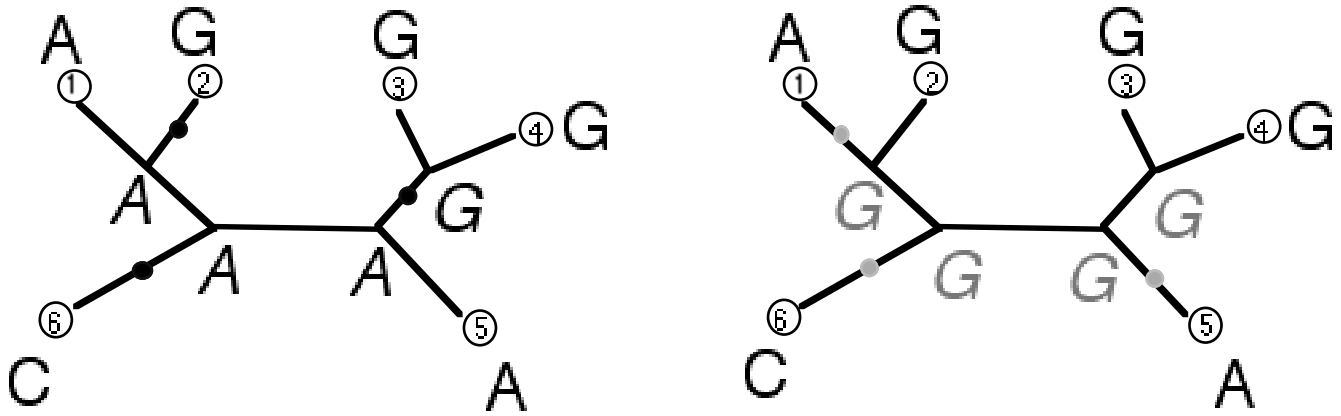
convergence

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

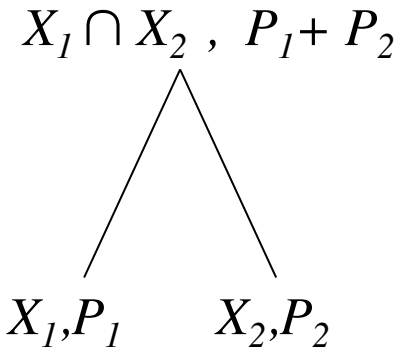
Exemple: A ce site et pour cette forme d'arbre, au moins 3 substitutions sont nécessaires pour expliquer le pattern de nucléotides présent aux feuilles de l'arbre. Plusieurs scénarios distincts à 3 changements sont possibles.

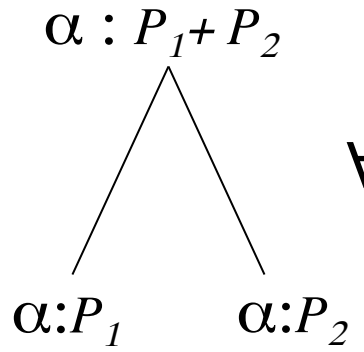
Algorithme de Fitch : calcul du nombre minimal de changements

Raciner arbitrairement l'arbre et calculer récursivement, à chaque nœud, deux objets:
- X: ensemble des résidus tous également possibles à ce nœud
- P: nombre minimal de changements dans le sous-arbre dont ce nœud est racine

---

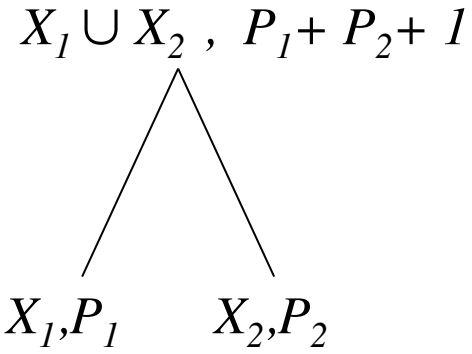
1er cas: $X_1 \cap X_2$ n'est pas vide

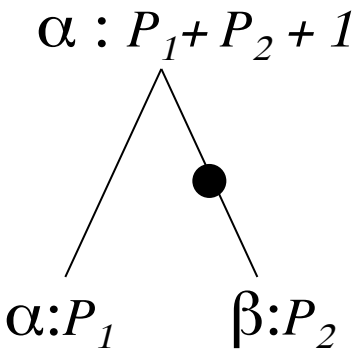


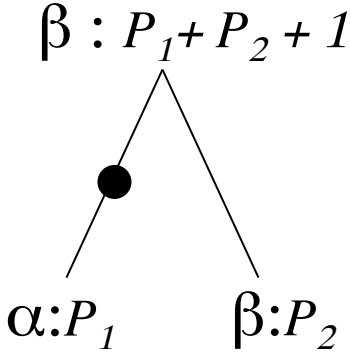


$\forall\ \alpha \in X_1 \cap X_2$

2ème cas: $X_1 \cap X_2$ est vide







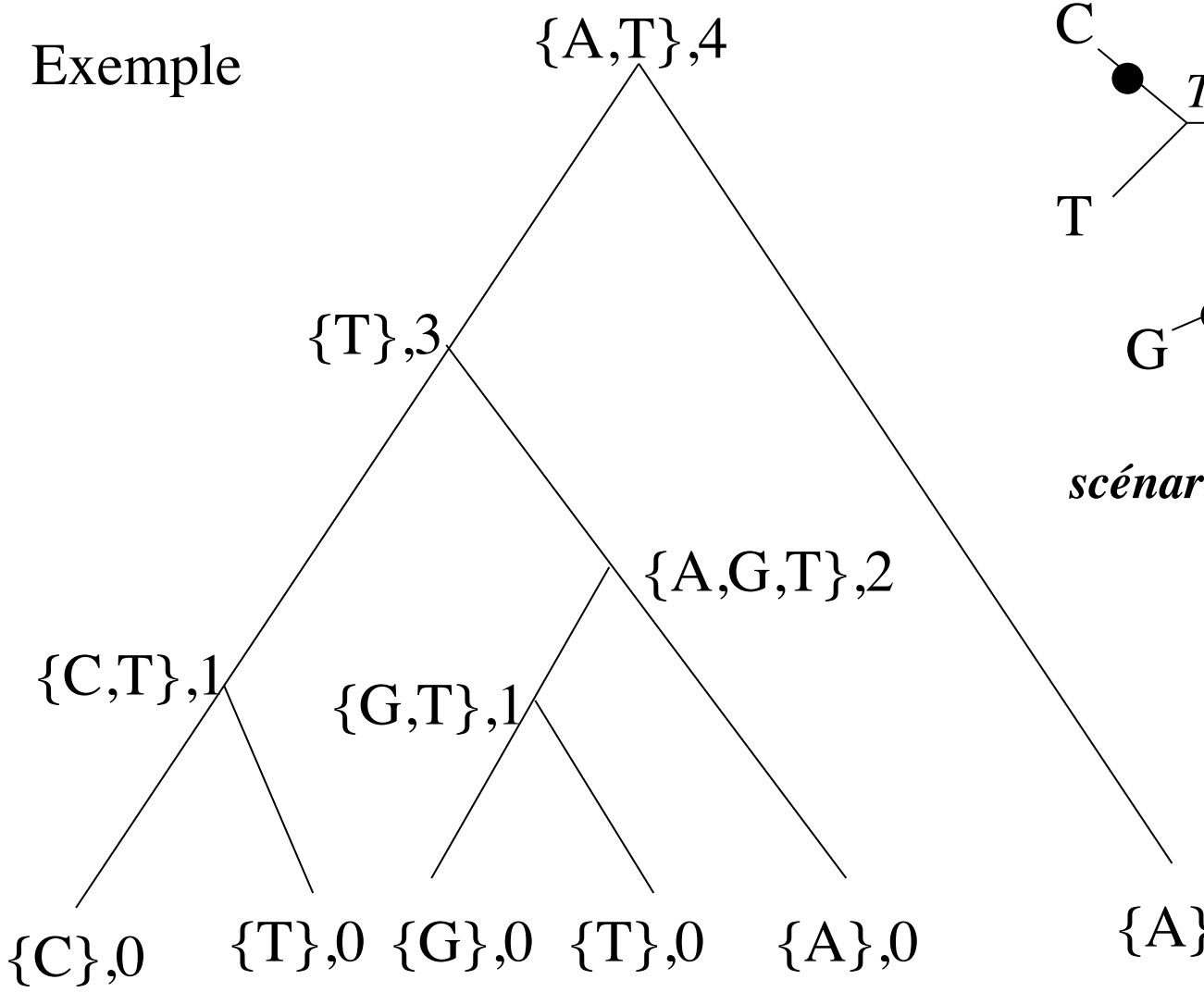
$\forall\ \alpha \in X_1$,
$\forall\ \beta \in X_2$

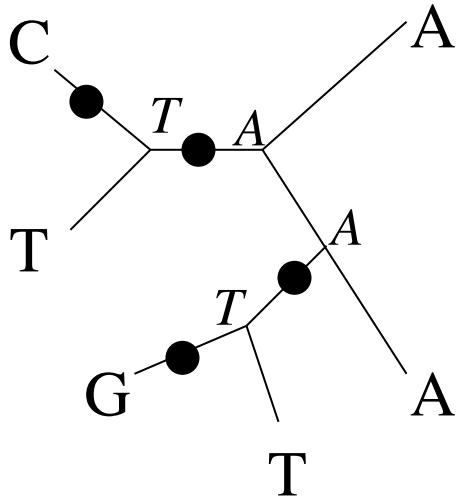
Algorithme de Fitch (suite)

Initialisation du calcul récursif aux feuilles de l'arbre

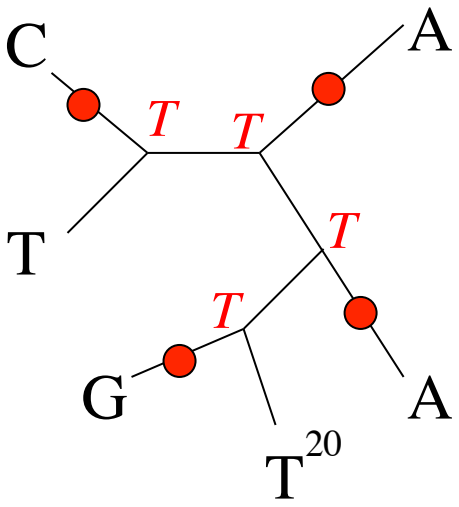
X={résidu présent à cette feuille}, P = 0

Exemple





scénario ancestral non unique !



Parcimonie (2)

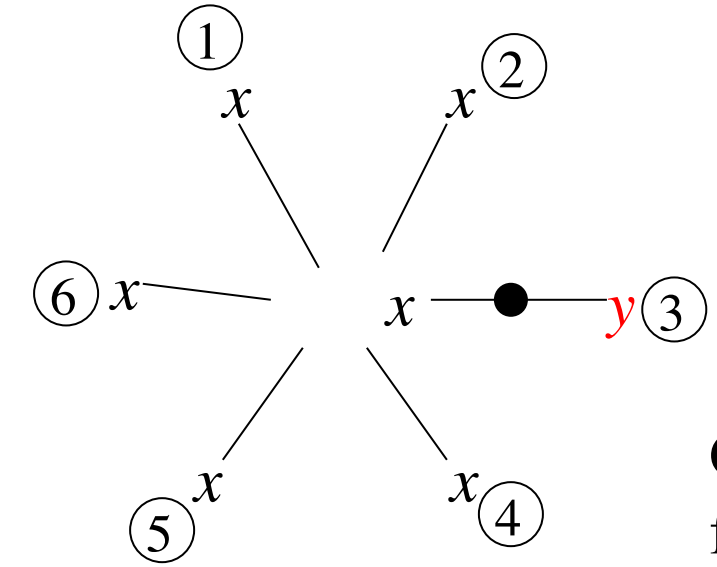
- Etape 2:
  - calculer *d* (étape 1) pour chaque site de l'alignement.
  - Sommer les valeurs *d* pour tous les sites.
  - Ceci donne la longueur *L* de l'arbre.

- 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

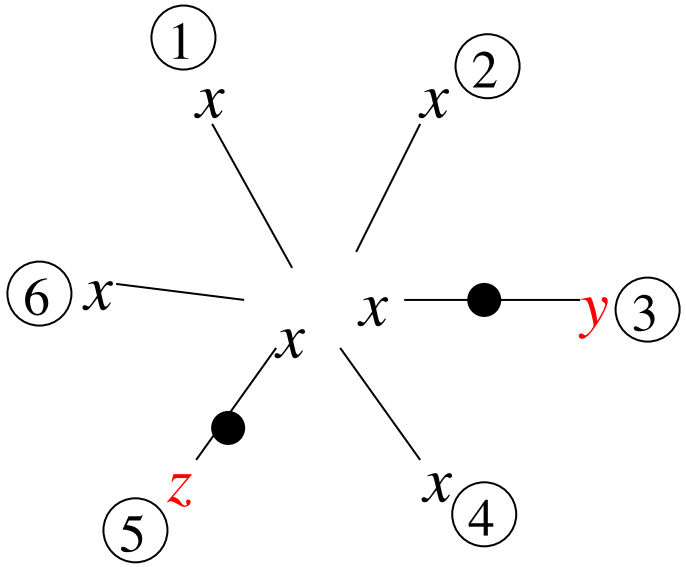
    = le (ou les) arbre(s) le(s) plus parcimonieux.

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.

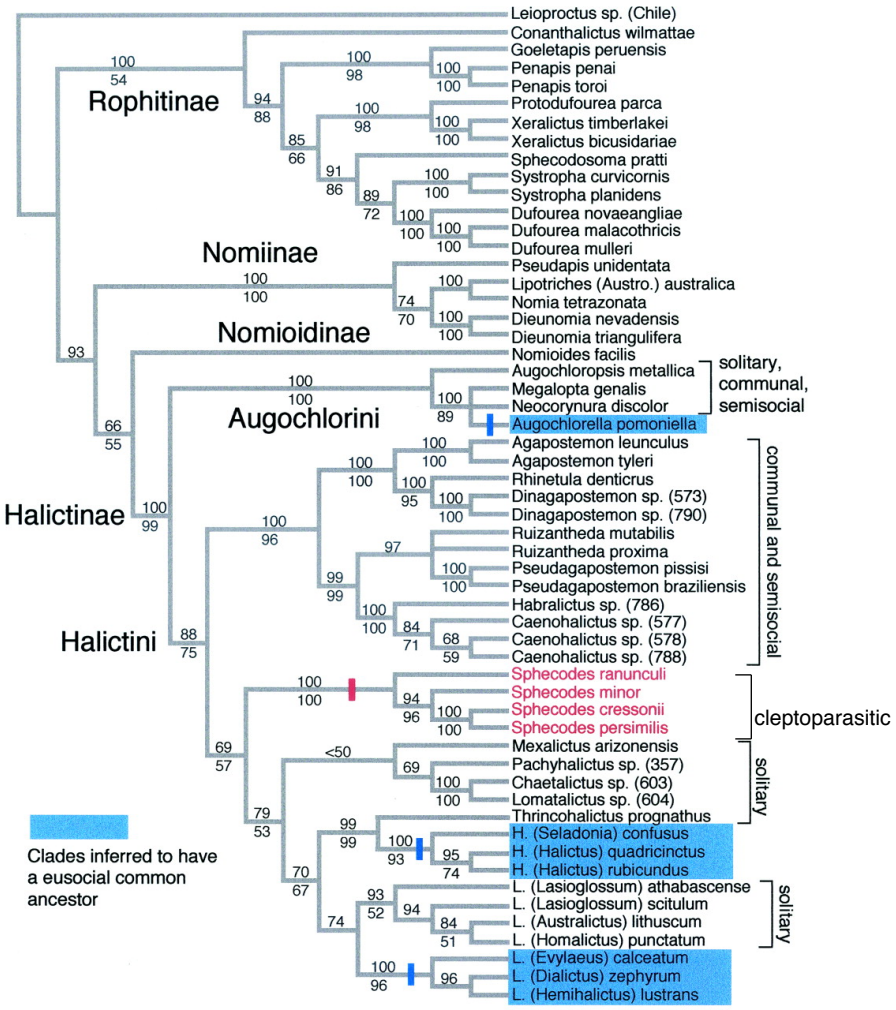
• Faire des réarrangements locaux: chaque branche interne définit 4 sous-arbres *a, b, c, d* et une topologie entre eux $\begin{smallmatrix} a & c \\ & \rangle\!-\!\langle \\ b & d \end{smallmatrix}$; évaluer les alternatives $\begin{smallmatrix} a & b \\ & \rangle\!-\!\langle \\ c & d \end{smallmatrix}$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

• 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

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 M des taux instantanés de substitution:

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)

Evolution de la probabilité de présence d'une base :

$$A(t+dt) = A(t)[1- \overbrace{(m_{AT}+ m_{AC}+ m_{AG})}^{\lambda_A}dt] + T(t)\, m_{TA}dt + C(t)\, m_{CA}dt + G(t)\, m_{GA}dt$$
$$T(t+dt) = T(t)[1- (m_{TA}+ m_{TC}+ m_{TG})dt] + A(t)\, m_{AT}dt + C(t)\, m_{CT}dt + G(t)\, m_{GT}dt$$
etc…

En écriture matricielle F(t):vecteur des probabilités des 4 bases au temps t:

$$F(t+dt) = F(t) + MF(t)dt \text{ donc } \frac{dF(t)}{dt} = MF(t) \text{ donc } F(t) = e^{Mt}F(0)$$

On obtient la matrice P des <u>probabilités conditionnelles de changement</u>:
$P(t) = e^{Mt} = \Lambda^{-1} e^{\Delta t} \Lambda$ (si $M = \Lambda^{-1} \Delta \Lambda$ avec Δ diagonale)

ancêtre: i ---------- t unités de temps ----------> j : descendant

$P_{ij}(t)$ = proba j en t sachant i en 0

$$A(t) = A(0)P_{AA}(t)+ T(0)P_{TA}(t)+ C(0)P_{CA}(t)+ G(0)P_{GA}(t)$$

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.

Jukes & Cantor (1 param.)

M =

↙	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

M =

↙	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.)

M =

↙	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.)

M =

↙	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)

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

Réversibilité des modèles évolutifs de Markov

Définition : $\forall i,j \;\; \pi_j m_{ji} = \pi_i m_{ij} \quad <=> \quad \forall i,j,t \;\; \pi_j P_{ji}(t) = \pi_i P_{ij}(t)$

(à l'équilibre des fréquences des bases)

L'observation de l'état d'une séquence aux deux extrémités d'une branche n'indique rien sur le sens de l'évolution.

SeqA ⟷ SeqB

Seule contrainte: réversibilité

GeneralTimeReversible (9 param.)

M =

↙	A	T	C	G
A	$-\lambda_A$	$a\pi_A$	$b\pi_A$	$c\pi_A$
T	$a\pi_T$	$-\lambda_T$	$d\pi_T$	$e\pi_T$
C	$b\pi_C$	$d\pi_C$	$-\lambda_C$	$f\pi_C$
G	$c\pi_G$	$e\pi_G$	$f\pi_G$	$-\lambda_G$

10 symboles mais 9 paramètres car $\pi_A + \pi_T + \pi_C + \pi_G = 1$

Eq. (π_A, π_T, π_C, π_G)

Tamura & Nei 93 (6 paramètres)

M =

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

Eq. (π_A, π_T, π_C, π_G)

C'est le modèle réversible le plus général dont on connaisse l'expression analytique des probabilités conditionnelles de changement, P(t).

[Tamura & Nei (1993) *MolBiolEvol* 10:512]

Tamura & Nei 1993: probabilités conditionnelles de changement P(t)

$P =$

↙	A	T	C	G
A	$\pi_A + \frac{\pi_A \pi_Y}{\pi_R} E + \frac{\pi_G}{\pi_R} E_R$	$\pi_A (1 - E)$	$\pi_A (1 - E)$	$\pi_A + \frac{\pi_A \pi_Y}{\pi_R} E - \frac{\pi_A}{\pi_R} E_R$
T	$\pi_T (1 - E)$	$\pi_T + \frac{\pi_T \pi_R}{\pi_Y} E + \frac{\pi_C}{\pi_Y} E_Y$	$\pi_T + \frac{\pi_T \pi_R}{\pi_Y} E - \frac{\pi_T}{\pi_Y} E_Y$	$\pi_T (1 - E)$
C	$\pi_C (1 - E)$	$\pi_C + \frac{\pi_C \pi_R}{\pi_Y} E - \frac{\pi_C}{\pi_Y} E_Y$	$\pi_C + \frac{\pi_C \pi_R}{\pi_Y} E + \frac{\pi_T}{\pi_Y} E_Y$	$\pi_C (1 - E)$
G	$\pi_G + \frac{\pi_G \pi_Y}{\pi_R} E - \frac{\pi_G}{\pi_R} E_R$	$\pi_G (1 - E)$	$\pi_G (1 - E)$	$\pi_G + \frac{\pi_G \pi_Y}{\pi_R} E + \frac{\pi_A}{\pi_R} E_R$

avec $\pi_R = \pi_A + \pi_G$; $\pi_Y = \pi_T + \pi_C$; $E = e^{-\beta t}$, $E_R = e^{-(\alpha_R + \beta)t}$, $E_Y = e^{-(\alpha_Y + \beta)t}$

GTR : 9 paramètres

π_A, π_T, π_C, a, b, c, d, e, f

$a = b = e = f$

Tamura & Nei 93 : 6 paramètres

$\pi_A, \pi_T, \pi_C, \beta, \alpha_R, \alpha_Y$

$\alpha_R = \alpha_Y$

Felsenstein 84 : 5 paramètres

$\pi_A, \pi_T, \pi_C, \beta, \alpha$

$\pi_A = \pi_T$

$\pi_C = \pi_G$

Tamura 92 : 3 paramètres

θ, α, r

$\theta = 1/2$

Kimura à 2 paramètres

α, r

$\alpha = 1$

Jukes & Cantor

r

Imbrications de ces modèles

On sait donc exprimer analytiquement la matrice P des probabilités conditionnelles pour 5 de ces 6 ces modèles.

Pour GTR, il faut diagonaliser numériquement la matrice M pour obtenir P, les paramètres étant donnés.

Tamura 1992: probabilités conditionnelles de changement P(t)

Transformer l'expression du modèle plus général TN93 avec:

$$\beta = r; \quad \alpha_R = \alpha_Y = \frac{(\alpha - 1)}{2} r; \quad \pi_A = \pi_T = \frac{(1-\theta)}{2}; \quad \pi_C = \pi_G = \frac{\theta}{2}; \quad \pi_R = \pi_Y = \frac{1}{2}$$

P =

↙	A	T	C	G
A	$\frac{(1-\theta)}{2}(1+e^{-rt}) + \theta e^{-\frac{\alpha+1}{2}rt}$	$\frac{(1-\theta)}{2}(1-e^{-rt})$	$\frac{(1-\theta)}{2}(1-e^{-rt})$	$\frac{(1-\theta)}{2}(1+e^{-rt}) - (1-\theta)e^{-\frac{\alpha+1}{2}rt}$
T	$\frac{(1-\theta)}{2}(1-e^{-rt})$	$\frac{(1-\theta)}{2}(1+e^{-rt}) + \theta e^{-\frac{\alpha+1}{2}rt}$	$\frac{(1-\theta)}{2}(1+e^{-rt}) - (1-\theta)e^{-\frac{\alpha+1}{2}rt}$	$\frac{(1-\theta)}{2}(1-e^{-rt})$
C	$\frac{\theta}{2}(1-e^{-rt})$	$\frac{\theta}{2}(1+e^{-rt}) - \theta e^{-\frac{\alpha+1}{2}rt}$	$\frac{\theta}{2}(1+e^{-rt}) + (1-\theta)e^{-\frac{\alpha+1}{2}rt}$	$\frac{\theta}{2}(1-e^{-rt})$
G	$\frac{\theta}{2}(1+e^{-rt}) - \theta e^{-\frac{\alpha+1}{2}rt}$	$\frac{\theta}{2}(1-e^{-rt})$	$\frac{\theta}{2}(1-e^{-rt})$	$\frac{\theta}{2}(1+e^{-rt}) + (1-\theta)e^{-\frac{\alpha+1}{2}rt}$

[Galtier & Gouy (1998) MolBiolEvol 15:871]

Les durées d'évolution ne sont pas estimables

Certains paramètres du modèle n'ont de sens qu'à un coefficient d'échelle près puisque seuls des produits *paramètre* . *t* apparaissent dans les termes de la matrice *P(t)*.

On reparamétrise la matrice *P* en 2 types de paramètres:

- *r*, un paramètre de **longueur** qui fixe l'ordre de grandeur des produits *paramètre* . *t*
- un ou des paramètres **qualitatifs** qui fixent les valeurs relatives des termes les uns par rapport aux autres.

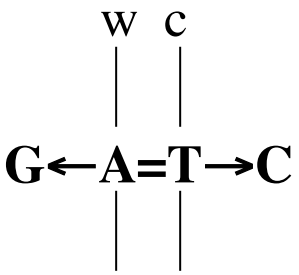
Exemple, pour le modèle de Tamura & Nei 93, on remplace ses paramètres t, β, α_R, α_Y par $r = \beta t$, α_R / β , α_Y / β

$$P_{AG}(t) = \pi_G + \frac{\pi_G \pi_Y}{\pi_R} E - \frac{\pi_G}{\pi_R} E_R \quad \text{avec} \quad E = e^{-\beta t},\ E_R = e^{-(\alpha_R + \beta)t}$$

$$E = e^{-r}\ ,\ E_R = e^{-(\frac{\alpha_R}{\beta}+1)r}\ ,\ E_Y = e^{-(\frac{\alpha_Y}{\beta}+1)r}$$

Un modèle de Markov non réversible

Hypothèse: les deux brins d'ADN sont répliqués dans les mêmes conditions.



$m_{AG} = w_{AG} + c_{TC}$, plus généralement, $m_{ij} = w_{ij} + c_{\bar{i}\bar{j}}$

Symétrie des brins vis à vis réplication : $w_{ij} = c_{ij}$

$\forall i,j \;\; m_{ij} = w_{ij} + c_{\bar{i}\bar{j}} = c_{ij} + w_{\bar{i}\bar{j}} = m_{\bar{i}\bar{j}}$

Lobry & Sueoka 95 (6 param.)

M =

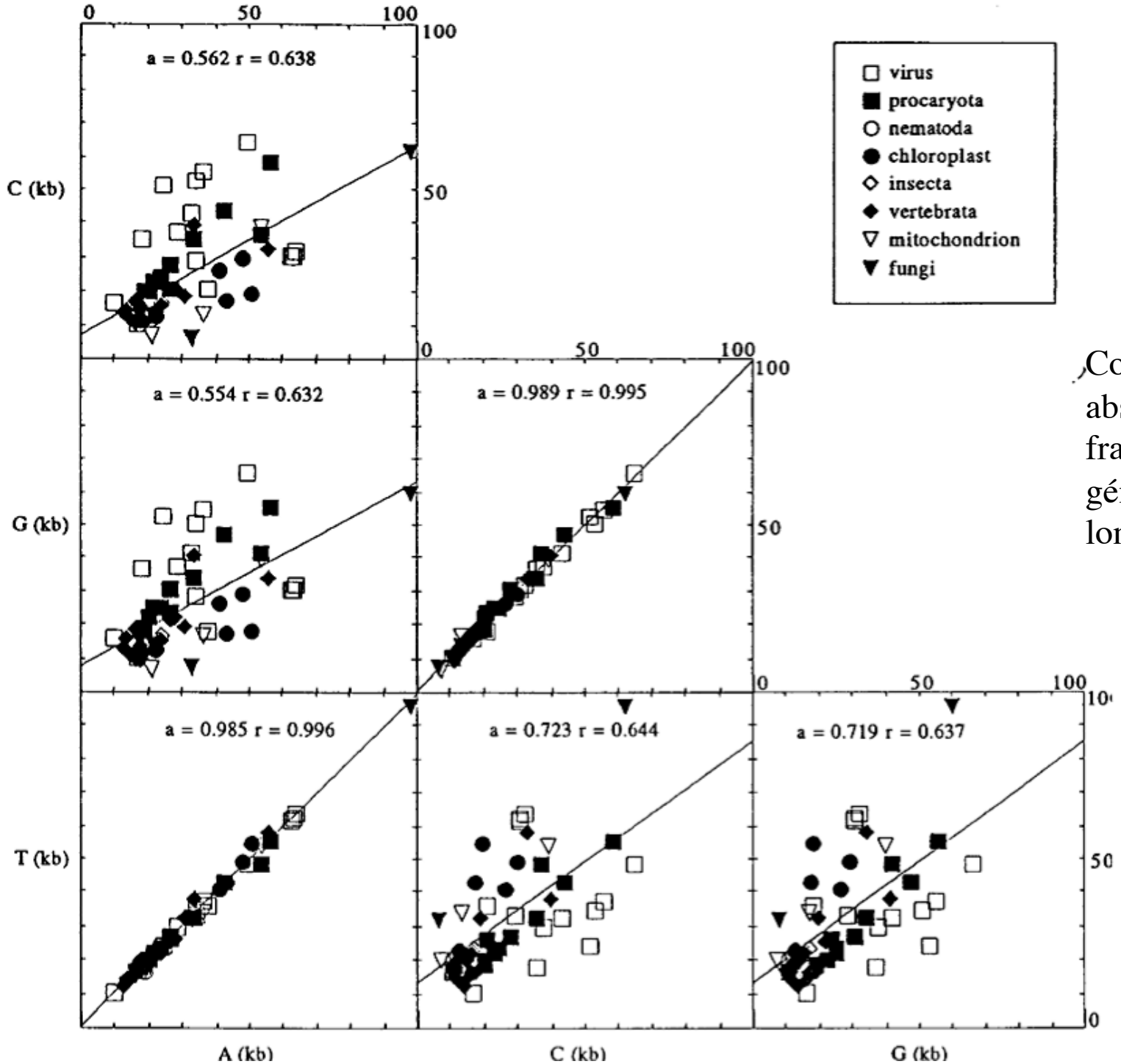
↙	A	T	C	G
A	$-\lambda_A$	a	d	b
T	a	$-\lambda_T$	b	d
C	e	c	$-\lambda_C$	f
G	c	e	f	$-\lambda_G$

Eq. $(u/2v, u/2v, (v-u)/2v, (v-u)/2v)$

$u = b+d$; $v = b+c+d+e$

Non réversible : $\pi_A m_{AC} \neq \pi_C m_{CA}$

à l'équilibre : [A]=[T] et [C]=[G]



Composition absolue en bases de fragments génomiques de longueur > 50 kb

Lobry 1995
JMolEvol 40:326

Modèle markovien: longueur d'une branche

Longueur d'une branche = nbre attendu de subst. le long de la branche

à l'équilibre, $l = \sum_i \pi_i \lambda_i t$, où $t =$ durée de la branche

Exemple, pour le modèle de Felsenstein84:

$$\lambda_A = \alpha \frac{\pi_G}{\pi_R} + \beta(1 - \pi_A),\ \lambda_T = \alpha \frac{\pi_C}{\pi_Y} + \beta(1 - \pi_T),\ \text{etc...}$$

donc:

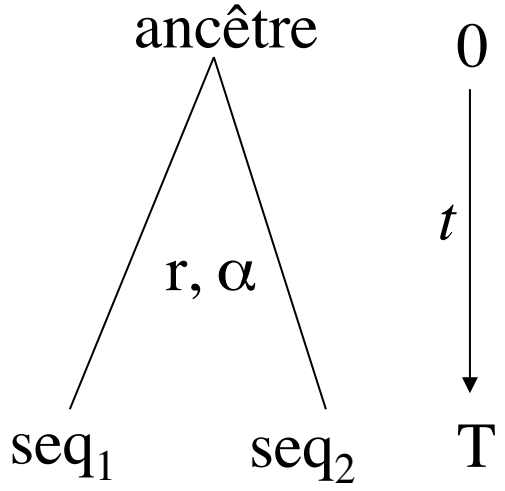
$$l = \left[2\alpha \left(\frac{\pi_A \pi_G}{\pi_R} + \frac{\pi_T \pi_C}{\pi_Y} \right) + \beta \left(1 - \sum_i \pi_i^2 \right) \right] t \quad \text{paramétrisation } t, \alpha, \beta, \pi_i$$

$$l = \left[2\frac{\alpha}{\beta} \left(\frac{\pi_A \pi_G}{\pi_R} + \frac{\pi_T \pi_C}{\pi_Y} \right) + \left(1 - \sum_i \pi_i^2 \right) \right] r \quad \text{paramétrisation } \alpha/\beta, r, \pi_i$$

l est proportionnelle au paramètre *r*

Calcul de la distance évolutive entre 2 séquences selon le modèle de Kimura à 2 paramètres.

Définition: la distance évolutive entre 2 séquences est le nombre total de substitutions produites sur les 2 lignées depuis leur divergence.



x ----> TR αr

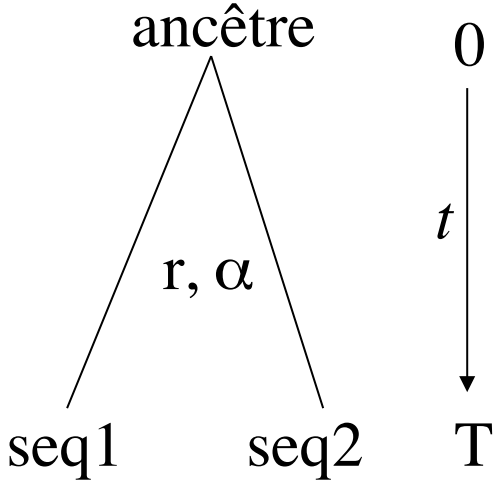
----> TV_1 r

----> TV_2 r

distance(seq1,seq2) :

$d = 2\,r\,(\alpha+2)\,T = 2\,\lambda\,T$

$\lambda = \alpha r + r + r = r(\alpha+2)$ proba d'un chgt qcq



P(t)= proba site occupé par transition
Q(t)=proba site occupé par transversion

t : 1-P(t)-Q(t) P(t) Q(t)

↓ ↓ TV_{12} ↓ TV_{12} ↓ *TR ou 0*

t+dt: Q(t+dt) Q(t+dt) Q(t+dt)

$Q(t+dt) = 2(1-P(t)-Q(t))\ 2rdt\ (1-\lambda dt) +$ identité en t et une transv. qcq

$2\ P(t)\ 2rdt\ (1-\lambda dt) +$ P en t et une transversion qcq

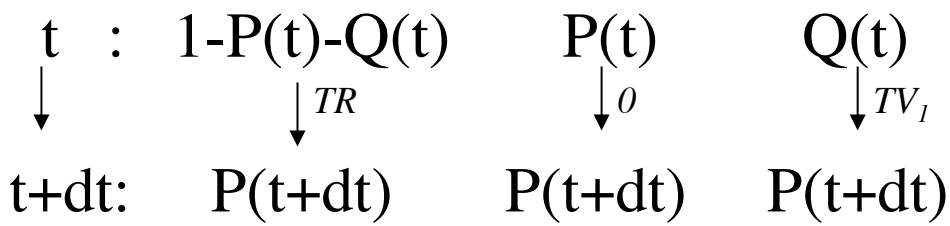
$2Q(t)\alpha rdt\ (1-\lambda dt) +$ Q en t et une transition

$Q(t)\ (1-\lambda dt)^2$ aucun changement

$Q(t+dt) = Q(t) + 4rdt - 8rQ(t)dt + Kdt^2$ où K constant ou borné

Donc $Q'(t) = 4r - 8rQ(t)$

Donc $Q(t) = 1/2 - (e^{-8rt})/2$ car $Q(0) = 0$



$P(t+dt) = 2(1-P(t)-Q(t))\,(1-\lambda dt)\,\alpha r dt +$ identité en t et une transition

$P(t)\,(1-\lambda dt)^2 +$ P en t et aucun changement

$2\,Q(t)\,(1-\lambda dt)\,rdt$ Q en t et une transversion unique

d'où

$P'(t) = 2\alpha r - 4r(\alpha + 1)P(t) + 2r(1 - \alpha)Q(t)$

$= - 4r(\alpha + 1)P(t) + r(\alpha - 1)e^{-8rt} + r(\alpha + 1)$

Solution avec condition initiale P(0)= 0 :

$$P(t) = (1/4)e^{-8rt} - (1/2)e^{-4r(1+\alpha)t} + 1/4$$

Donc, en fin de divergence, instant T, on a

$$Q(T) = (1/2) - (1/2)e^{-8rT} \qquad [1]$$

et

$$P(T) = (1/4)e^{-8rT} - (1/2)e^{-4r(1+\alpha)T} + 1/4 \qquad [2]$$

en inversant [1] : $4rT = -(1/2)\ln[1 - 2Q(T)]$
et en reportant dans [2] on obtient:

$$P(T) = (1/2) - (1/2)Q(T) - (1/2)e^{-4\alpha rT}(1-2Q(T))^{1/2} \qquad [3]$$

en inversant [3] :

$$2\alpha rT = -(1/2)\ln[1-2P(T)-Q(T)] + (1/4)\ln[1- 2Q(T)] \qquad [4]$$

on avait

$d(seq1,seq2) = 2r(\alpha+2)T = 2\,\alpha\,rT + 4rT$

donc

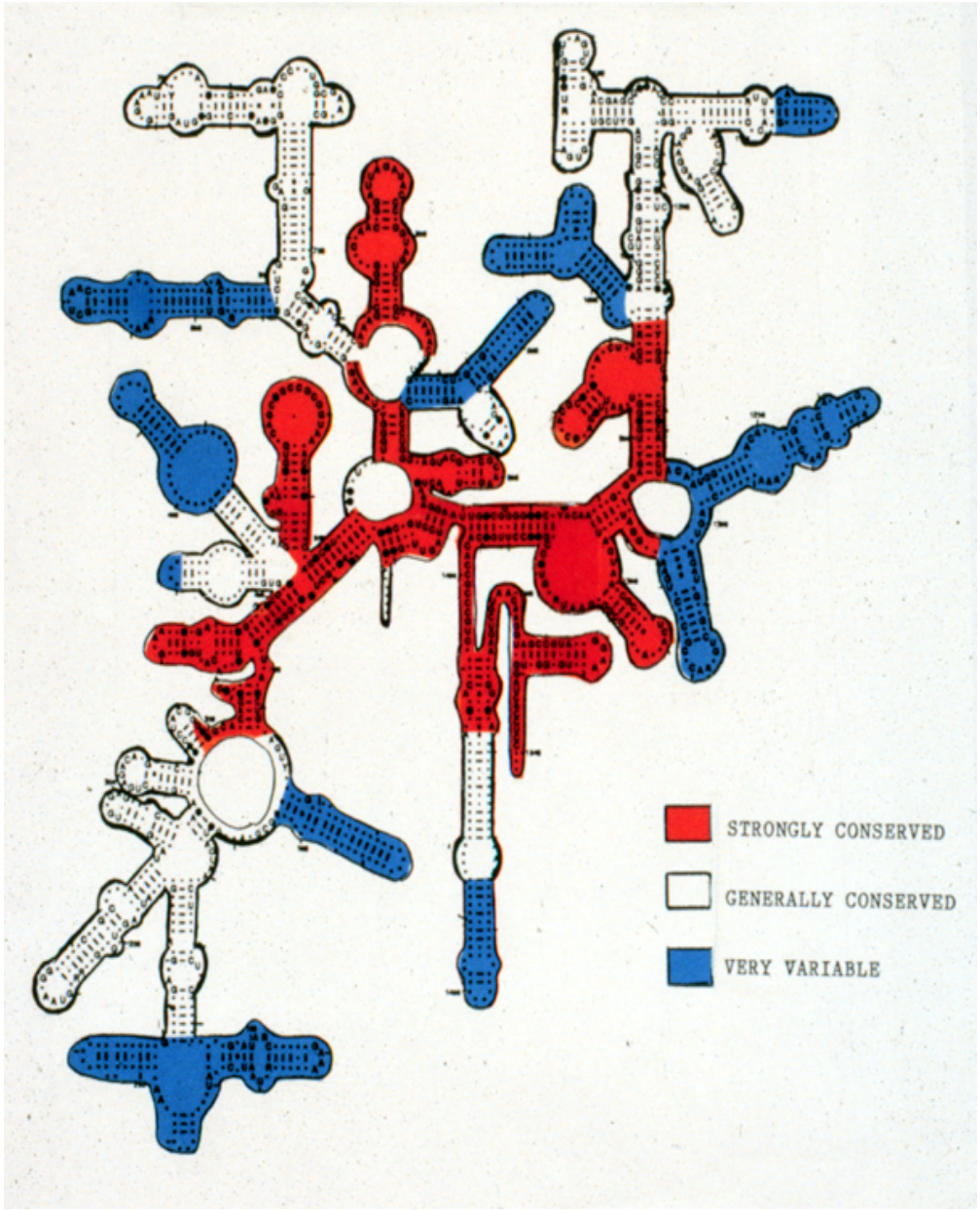
$$d(seq1,seq2) = -\,(1/2)\ln[1 - 2P - Q] - (1/4)\ln[1 - 2Q]$$

où

Kimura (1980)
JMolEvol 16:111

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

Variation de la vitesse d'évolution entre sites



Small subunit ribosomal RNA (18S or 16S)

Modélisation de la variation du taux d'évolution entre sites

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

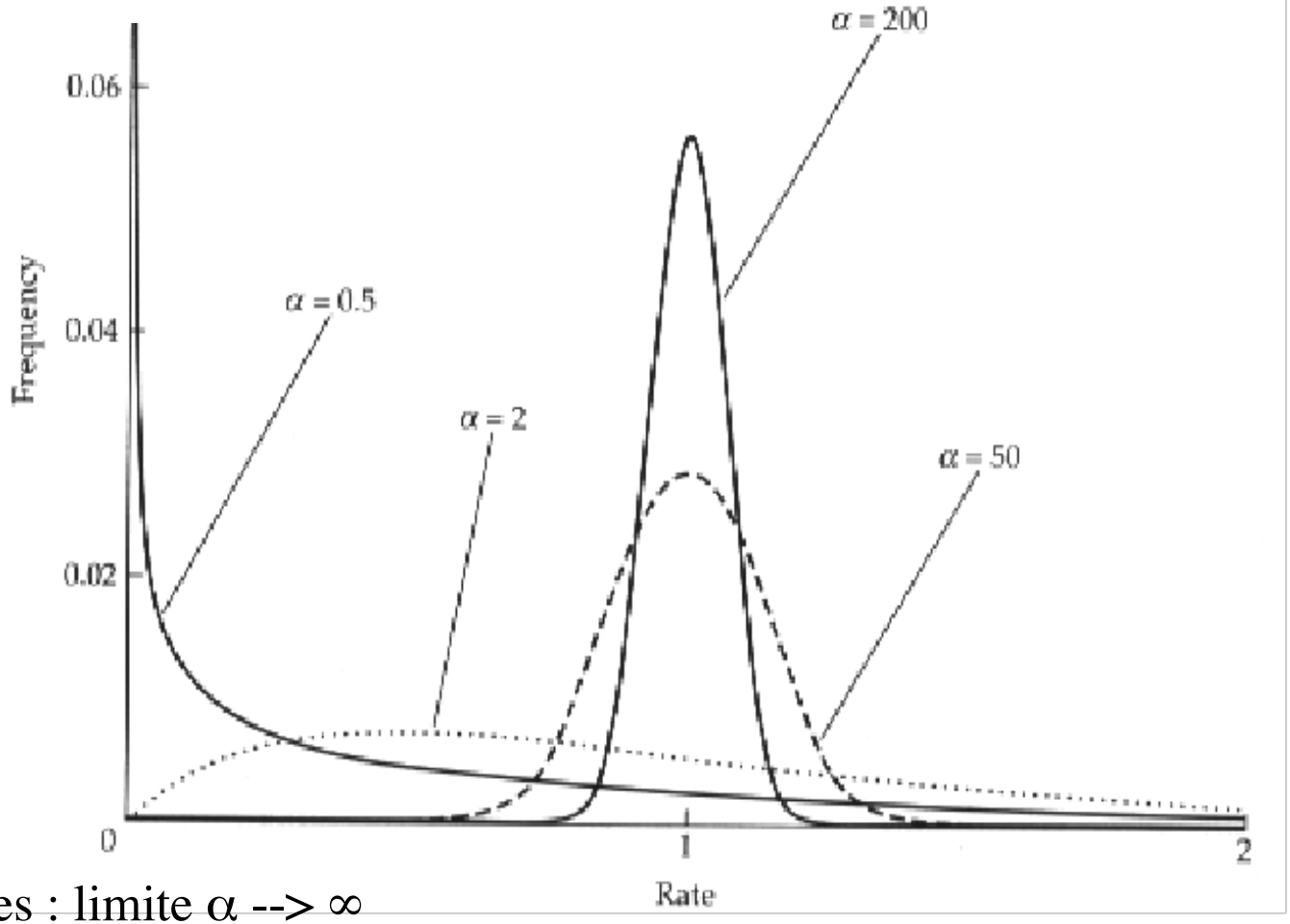
$$f(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$$

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 β=1/α pour avoir
moyenne = 1
variance = 1/α

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



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

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

1. Sous le modèle de Jukes & Cantor

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

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

m 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

2. Sous le modèle de Kimura à 2 paramètres

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

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

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

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

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)

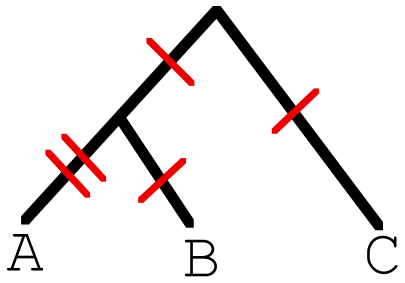
Arbre (non raciné)

(1) Mesure des distances évolutives.

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

Correspondence 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

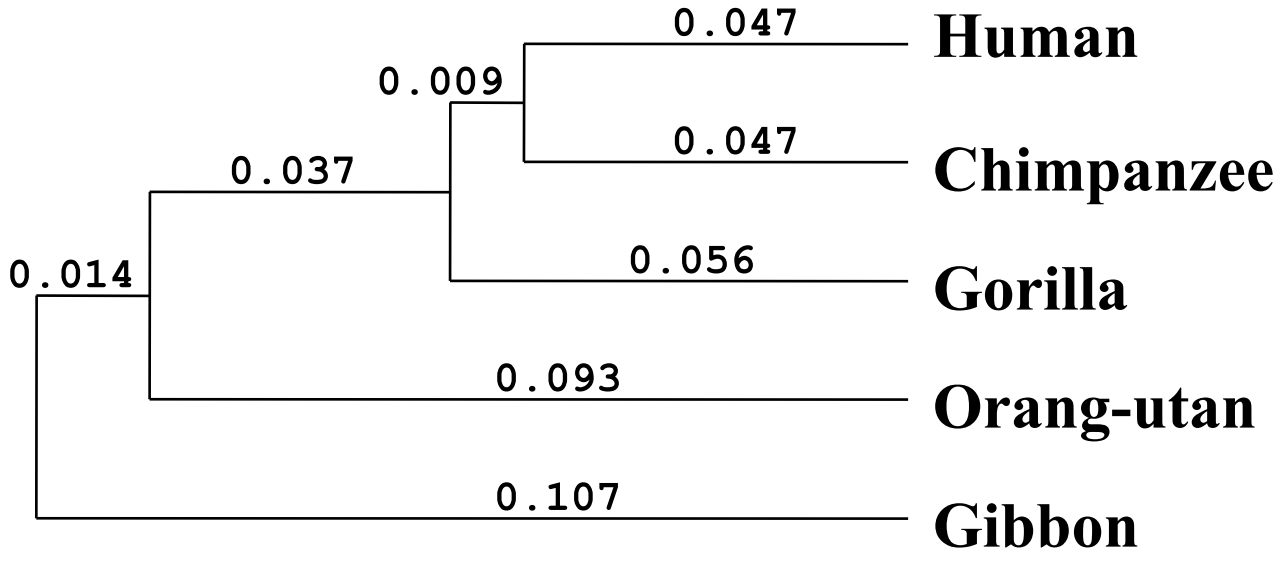
	A	B	C
A	**0**		
B	**3**	**0**	
C	**4**	**3**	**0**

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	-

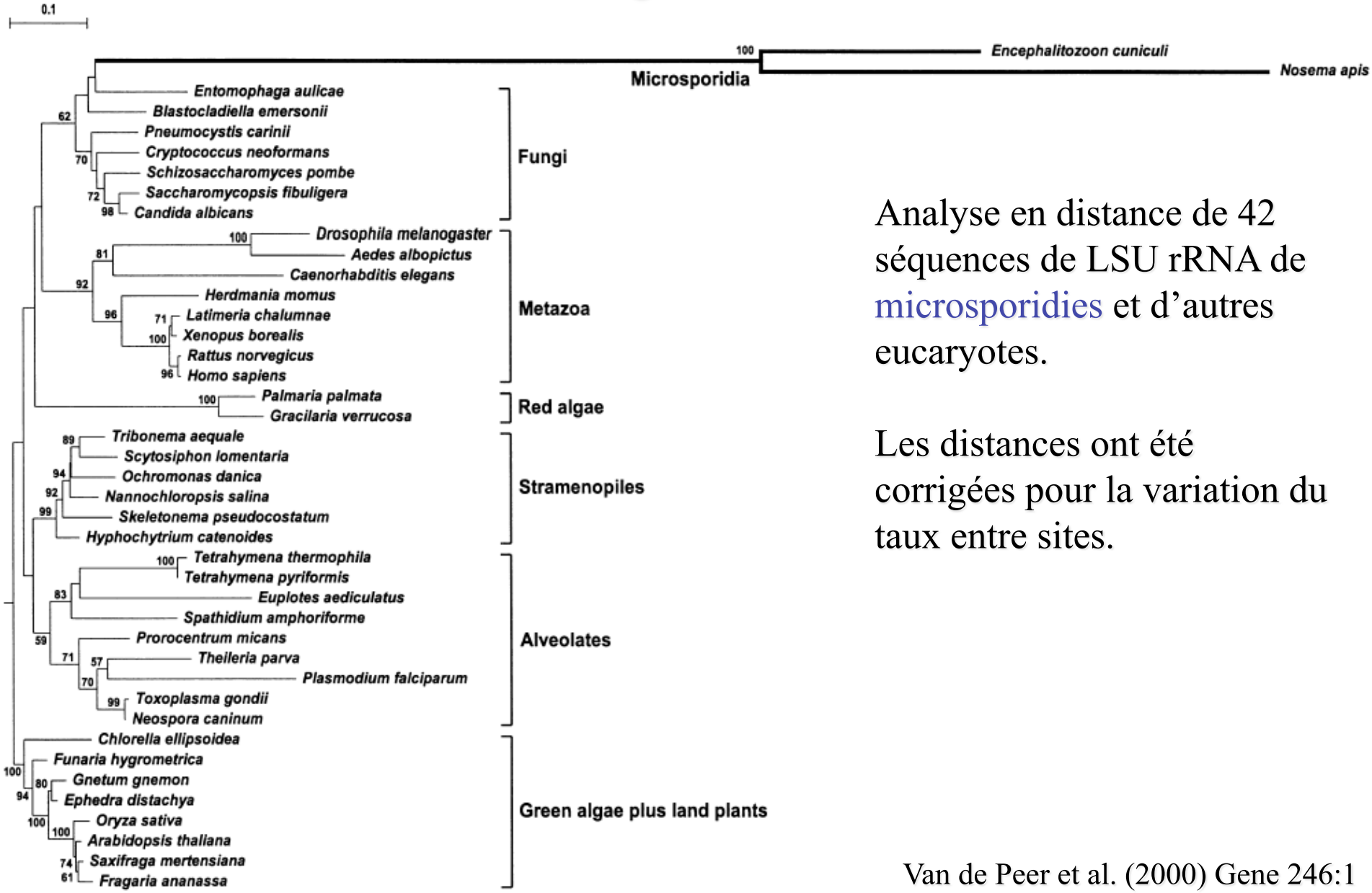
Proportion de différences (*p*) (au dessus de la diagonale) et distances de Kimura à 2 paramètres (*d*) (au dessous) pour un fragment d'ADN mitochondrial (895 pb).



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

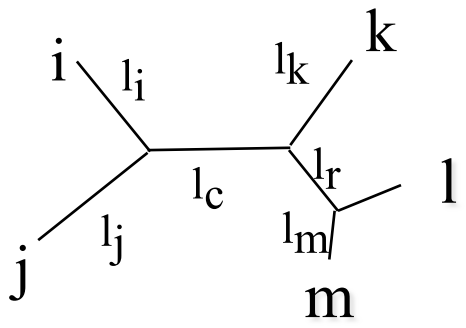
UPGMA : propriétés

- 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.

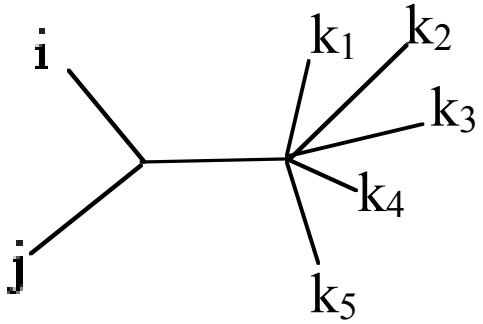
Problème: cette méthode n'est pas réalisable en pratique avec plus de ~ 25 séquences.

=> 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.



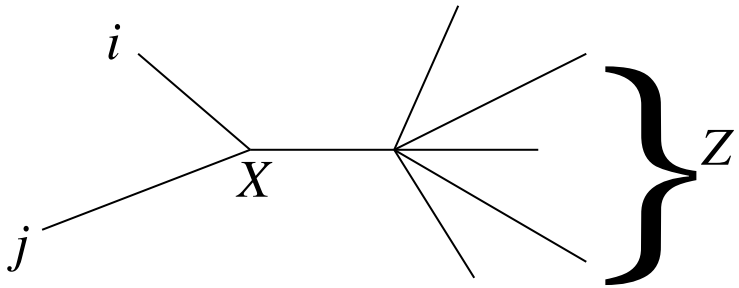
$$S_{ij} = \frac{1}{2(N-2)} \sum_{\substack{k \neq i \\ k \neq j}} \left(d_{ik} + d_{jk} \right) + \frac{1}{2} d_{ij} + \frac{1}{N-2} \sum_{\substack{k<l \\ k \neq i, k \neq j \\ l \neq i, l \neq j}} d_{kl}$$

- 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

Neighbor-Joining: calcul des longueurs des branches

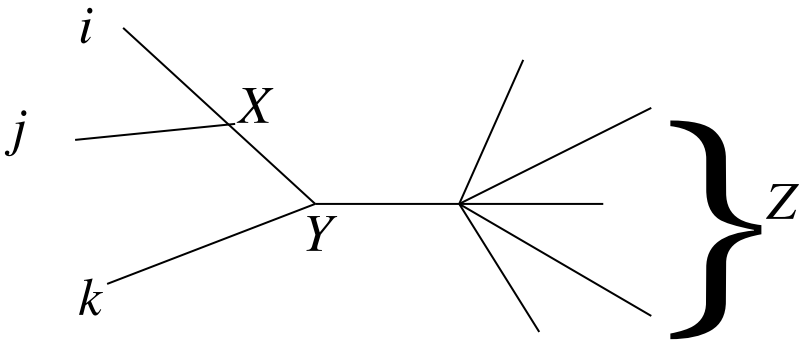
Branches périphériques



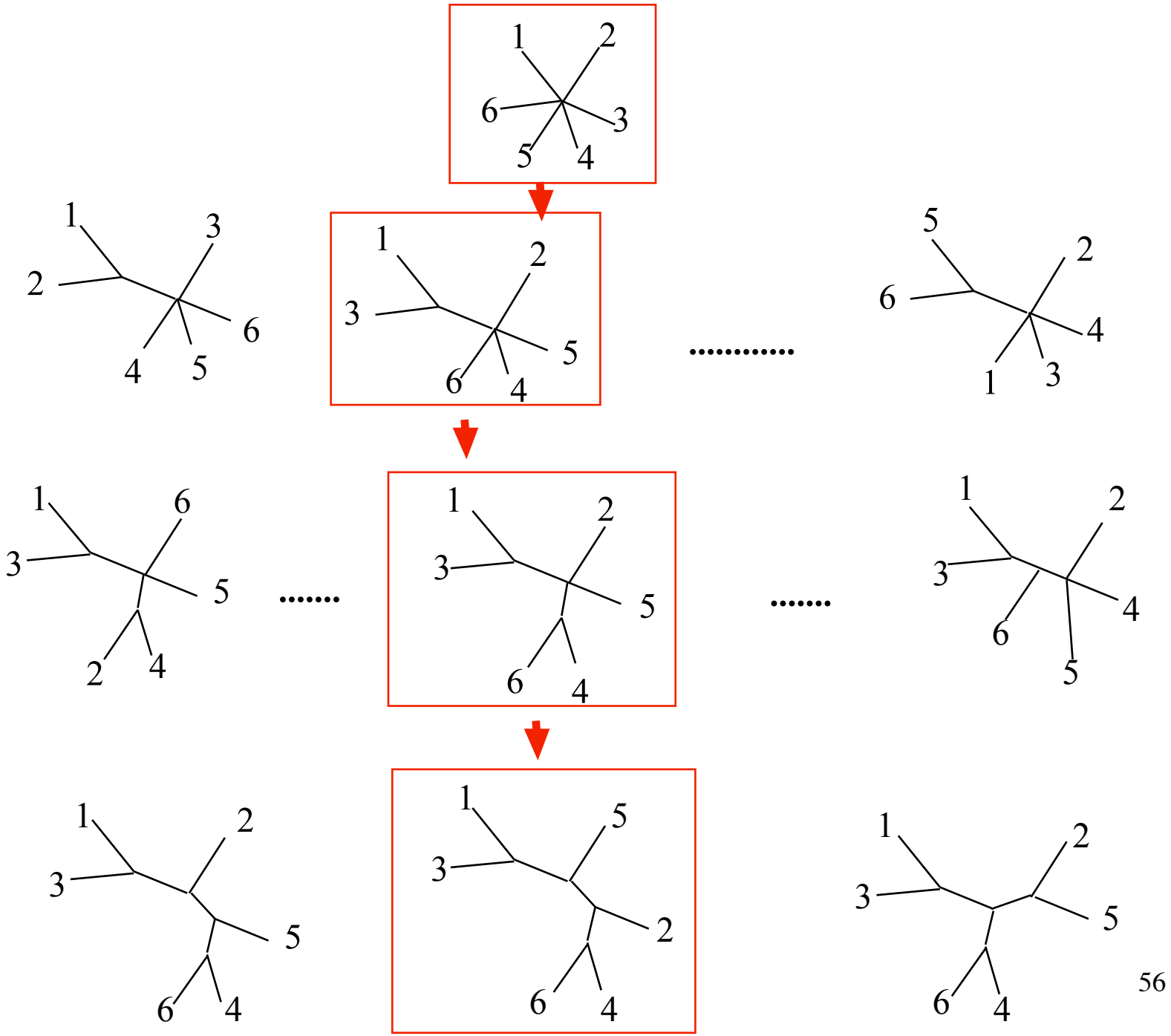
$$L_{iX} = \left(d_{ij} + d_{iZ} - d_{jZ}\right)/2$$

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

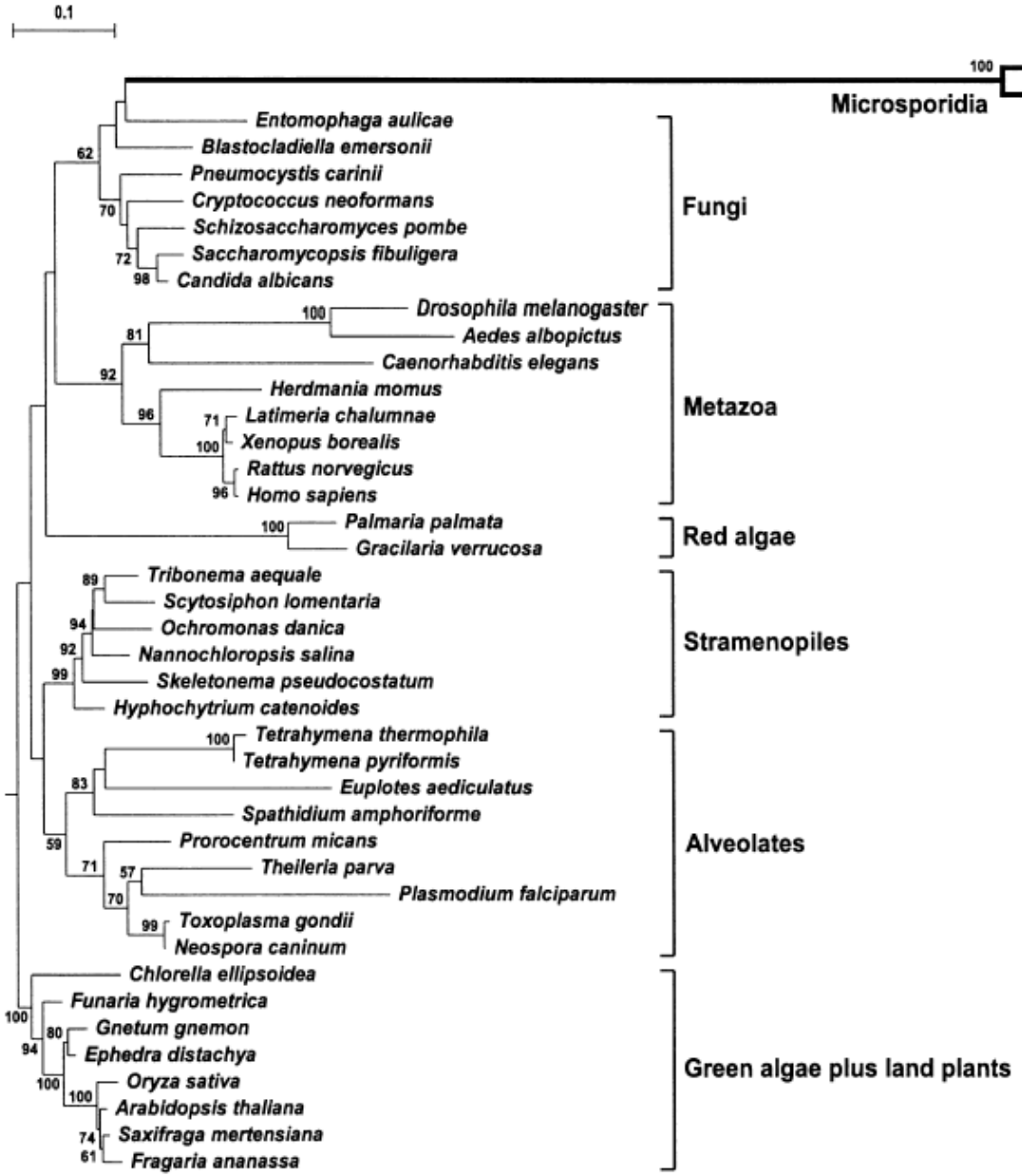
Branches internes



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


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

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)

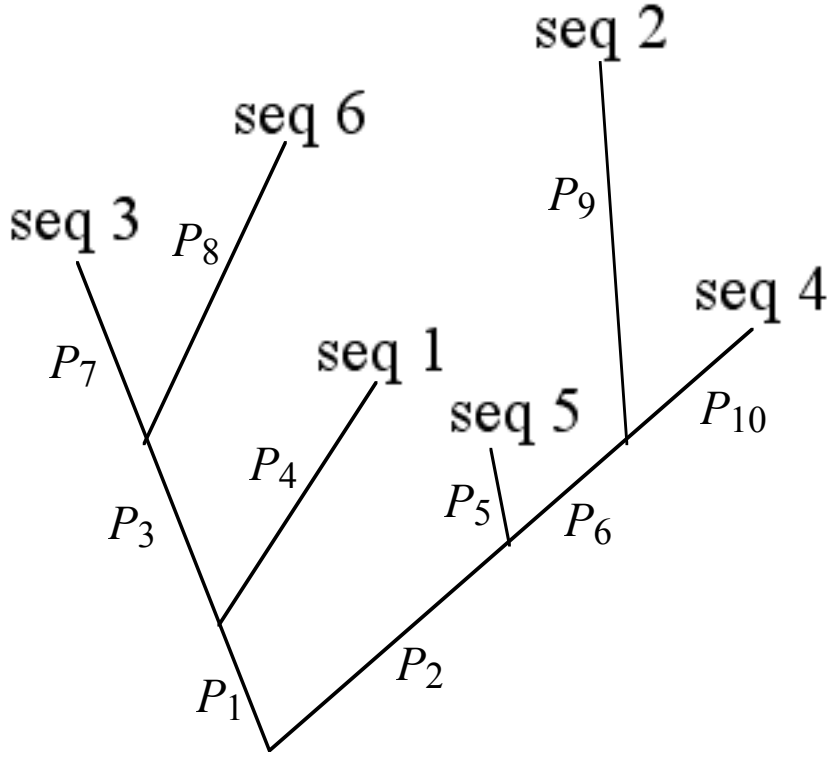
- Hypothèses
  - Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
  - Les sites évoluent indépendamment les uns des autres.
  - Les sites évoluent selon la même loi (on peut aussi modéliser la variation des taux entre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

Méthode du Maximum de vraisemblance(2)

Modèle probabiliste de l'évolution de séquences

Chaque branche est modélisée par un modèle de Markov distinct.

Matrice P_b : probabilités conditionnelles de substitution le long de la branche b



En général, on suppose que les matrices P_b ne diffèrent que par leur paramètre de longueur, r_b, et partagent leurs paramètres qualitatifs, θ.
Longueur d'une branche :

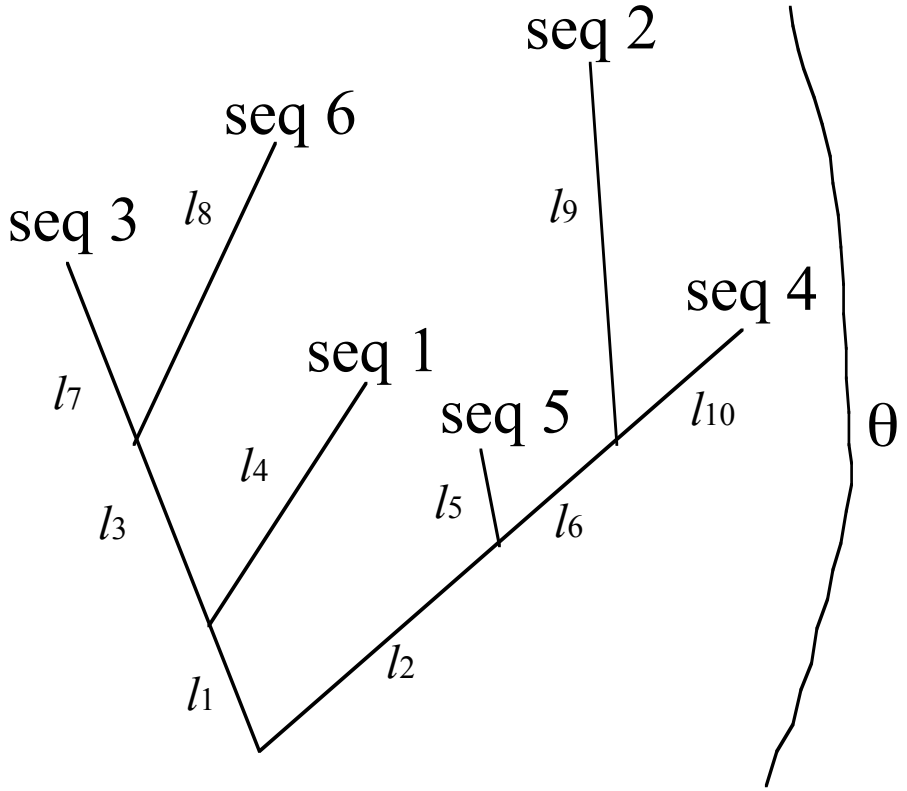
l_b = nbre attendu de subst. sur la branche $b \propto r_b$

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

θ, taux relatifs des substitutions (e.g., transition/transversion, biais G+C, fréquences d'équilibre)



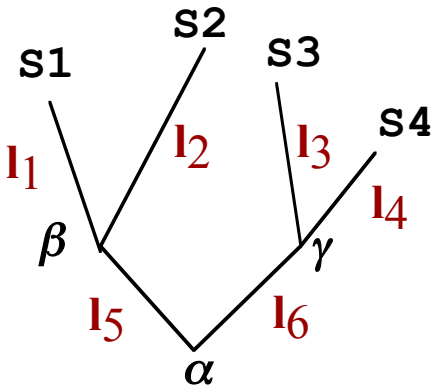
On sait calculer

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

pour toutes bases x & y, toute branche b, toutes valeurs θ

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

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

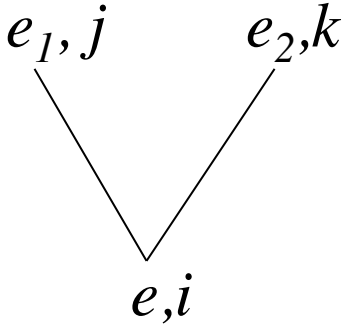
avec y : site; $B = \{A,C,G,T\}$; r : racine; P_{anc} : proba ancestrales des bases;

$L^{e,i}(y)$: 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)

Si le modèle utilisé est réversible, homogène (les paramètres qualitatifs ne varient pas entre branches), et stationnaire (à l'équilibre des fréquences de bases),

alors

la **vraisemblance est indépendante de la racine** de l'arbre.

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

stationnarité

$$= \sum_{i\in B}\sum_{j\in B}\sum_{k\in B} \pi_i P_{l_1}(j \mid i)L^{e_1,j}(y)P_{l_2}(k \mid i)L^{e_2,k}(y)$$

réversibilité

$$= \sum_{i\in B}\sum_{j\in B}\sum_{k\in B} \pi_j P_{l_1}(i \mid j)L^{e_1,j}(y)P_{l_2}(k \mid i)L^{e_2,k}(y)$$

$$= \sum_{j\in B}\sum_{k\in B} \pi_j L^{e_1,j}(y)L^{e_2,k}(y)\sum_{i\in B} P_{l_1}(i \mid j)P_{l_2}(k \mid i)$$

homogénéité

$$= \sum_{j\in B}\sum_{k\in B} \pi_j L^{e_1,j}(y)L^{e_2,k}(y)P_{l_1+l_2}(k \mid j)$$

Algorithme du maximum de vraisemblance(4)

- 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.

Maximum de vraisemblance : vitesse d'évolution variable entre sites

On peut ajouter l'hypothèse que les vitesses d'évolution des sites varient selon la distribution gamma f_α de paramètre de forme α et de moyenne 1. La vraisemblance du site *y* devient

$$L(y) = \int_0^\infty f_\alpha(u) L(y,u) du$$

où *L(y,u)* est la vraisemblance calculée comme plus haut en multipliant par *u* tous les paramètres de longueur du modèle probabiliste.

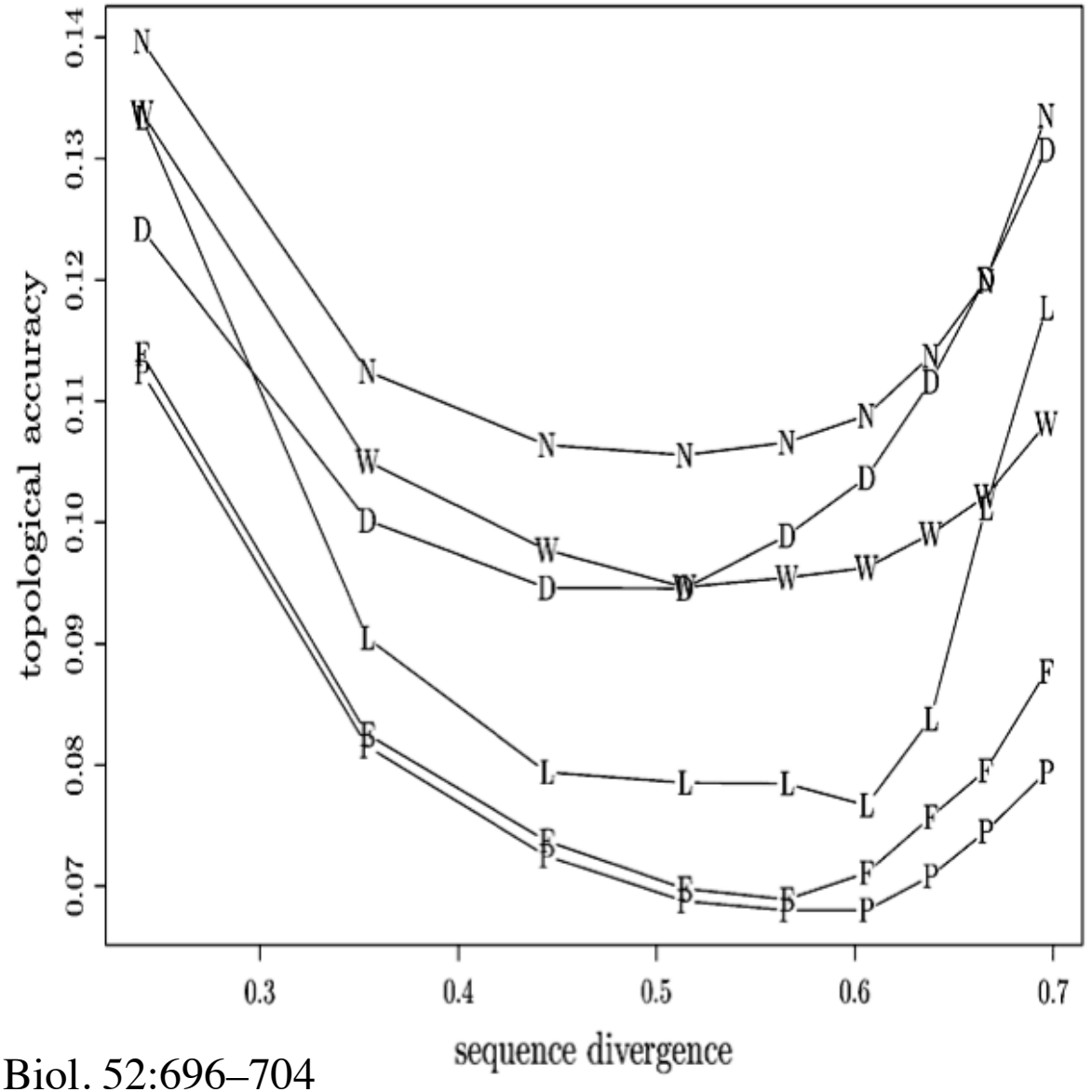
Pour obtenir quelque chose de calculable, on discrétise la distribution gamma en k quartiles représentés par leur moyenne w_j:

$$L(y) \approx \sum_{j=1}^{k} \frac{1}{k} L(y, w_j)$$

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

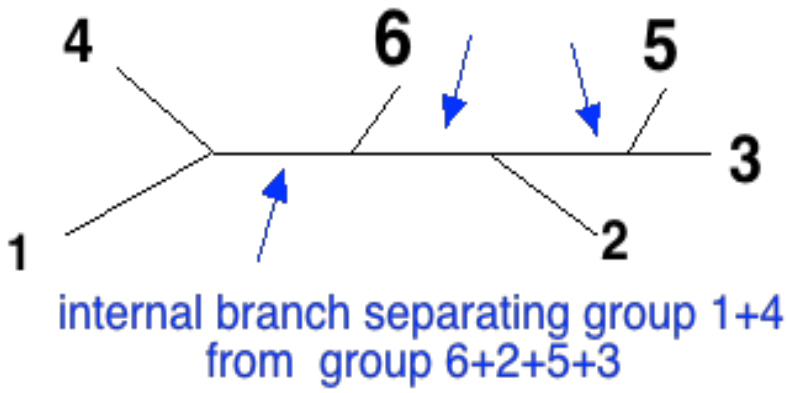
fastDNAml: une implémentation du principe du maximum de vraisemblance en phylogénie moléculaire [Olsen et coll. (1994) Comput Appl Biosci. 10:41-48]

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.

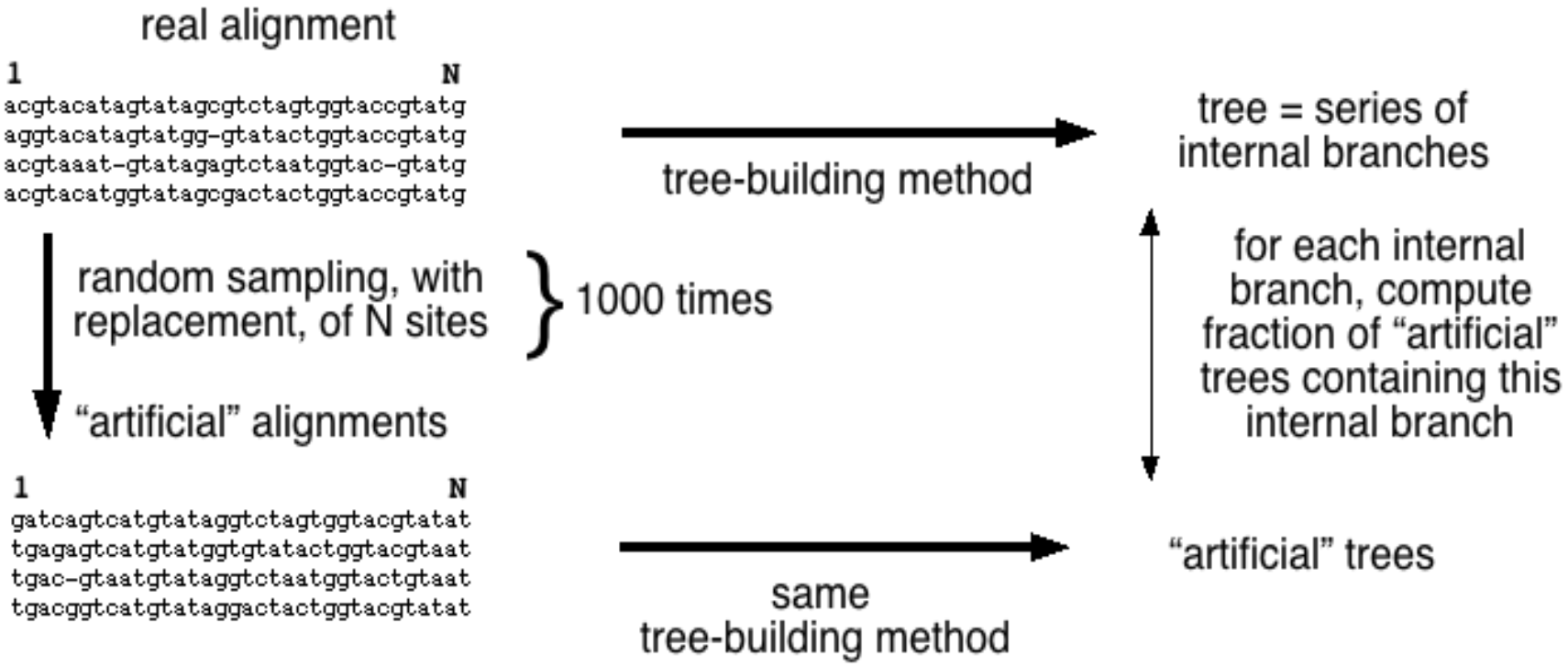
Fiabilité des arbres phylogénétiques: le bootstrap

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



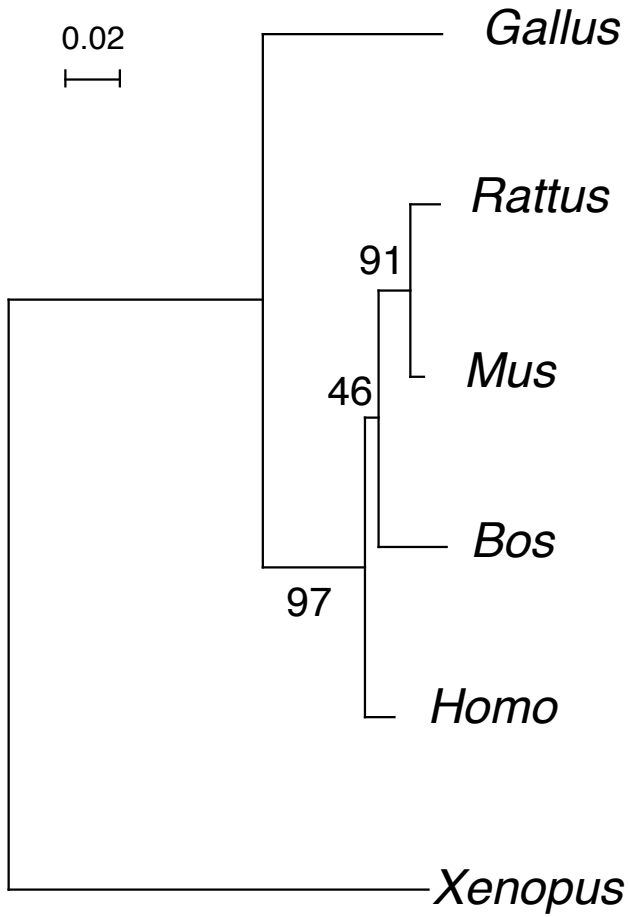
- La forme de l'arbre est déterminée par la liste des branches internes.
- Evaluer la fiabilité d'un arbre = évaluer celle de chaque branche interne.

Procédure de bootstrap



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

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 !

Approche Bayesienne en phylogénie moléculaire

Probabilité conditionnelle : $\Pr(A \mid B) = \dfrac{\Pr(A \cap B)}{\Pr(B)}$

Théorème de Bayes

$$\underbrace{\Pr(H \mid D)}_{a\ posteriori} = \frac{\overbrace{\Pr(D \mid H)}^{vraisemblance}\ \overbrace{\Pr(H)}^{a\ priori}}{\sum_h \Pr(D \mid h)\Pr(h)}$$

H : hypothèse; D : données

Remarque: si la distribution *a priori* est uniforme, le maximum de vraisemblance est équivalent au mode de la distribution *a posteriori*.

En phylogénie moléculaire, cela donne

H : paramètres du modèle et leurs distributions *a priori*

- topologie de l'arbre (τ) : *a priori* $f_{topo}(\tau)$
- longueurs des branches (ν) : *a priori* $f_b(\nu)$
- paramètres du processus de substitution (θ) : *a priori* $f_p(\theta)$
- variabilité des vitesses entre sites (α de la distribution gamma) : *a priori* $f_g(\alpha)$

$$\Pr(\tau \mid D) = \frac{\iiint L(D \mid \tau,\nu,\theta,\alpha) f_b(\nu) f_p(\nu) f_g(\alpha) d\nu d\theta d\alpha \; f_{topo}(\tau)}{\sum_t \iiint L(D \mid t,\nu,\theta,\alpha) f_b(\nu) f_p(\nu) f_g(\alpha) d\nu d\theta d\alpha \; f_{topo}(t)}$$

Ces quantités ne sont pas calculables directement !

Solution: utiliser une méthode de type « **Markov Chain Monte Carlo** » (MCMC), l'algorithme de Metropolis et coll. (1953).

Objectif : échantillonner une distribution statistique $f(\lambda)$, sans avoir besoin de calculer les valeurs $f(\lambda)$, mais uniquement des rapports de probabilités $f(\lambda_1)/ f(\lambda_2)$.

On va fabriquer, par chaîne de Markov, une suite de valeurs de nos paramètres (τ,ν,θ,α) dont la distribution à l'équilibre (loin dans la suite) est égale à la distribution conjointe *a posteriori* des paramètres.

On se donne des probas de changement d'état $\Pr(\lambda_2 | \lambda_1)$, pour toute paire λ_1 , λ_2 (l'essentiel est que tout passage $\lambda_1 \rightarrow \lambda_2$ soit possible, même en plusieurs étapes).

1. On part de valeurs initiales quelconques $\lambda_1 = (\tau_1, \nu_1, \theta_1, \alpha_1)$.
2. On choisi un autre état $\lambda_2 = (\tau_2, \nu_2, \theta_2, \alpha_2)$
3. Calculer

$$r = \underbrace{\frac{L(D \mid \tau_2, \nu_2, \theta_2, \alpha_2)}{L(D \mid \tau_1, \nu_1, \theta_1, \alpha_1)}}_{\text{rapport de vraisemblance}} \cdot \overbrace{\frac{\Pr(\tau_2, \nu_2, \theta_2, \alpha_2)}{\Pr(\tau_1, \nu_1, \theta_1, \alpha_1)}}^{\text{rapport des } a \text{ priori}} \cdot \underbrace{\frac{\Pr(\tau_1, \nu_1, \theta_1, \alpha_1 \mid \tau_2, \nu_2, \theta_2, \alpha_2)}{\Pr(\tau_2, \nu_2, \theta_2, \alpha_2 \mid \tau_1, \nu_1, \theta_1, \alpha_1)}}_{\text{rapport des propositions}}$$

4. Remplacer λ_1 par λ_2 si $r \geq 1$ ou si un tirage aléatoire selon une distribution uniforme sur [0,1] rend $\leq r$.
5. Retourner en 2.

Résultat: après une phase initiale (dite « burn-in »), les $(\tau, \nu, \theta, \alpha)$ sont distribués selon la distribution *a posteriori*.

MrBayes : un outil bayésien de phylogénie moléculaire

[Huelsenbeck & Ronquist (2001) Bioinformatics 17:754]

Les détails qu'il faut préciser dans une implémentation MCMC:

- Quels sont les paramètres (τ,ν,θ,α) et leurs distributions *a priori* ?
- Quelles sont les valeurs des rapports de proposition (probabilité passage $\lambda_1 \rightarrow \lambda_2$)?

θ Choix entre 3 modèles évolutifs: J&C (1 p.), K2P (2 p.), GTR (9 p.)

1. J&C : tout est dans la longueur des branches (paramètre ν).
2. K2P *a priori :* beta(x,y) telle que x/y = TR/TV attendu; variance diminue quand y augmente (par défaut: beta(1,1)).
3. GTR *a priori* : les 6 taux de modèle (A<->C, A<->G, A<->T, C<->G, C<->T, G<->T) suivent chacun une distribution gamma dont les moyennes relatives sont fixées (par défaut:tous égaux); les 3 paramètres de fréquences d'équilibre sont fixés aux fréquences moyennes des séquences.

α Distribution de la variabilité entre sites: choix entre absence de variabilité, distribution gamma de paramètre α discrétisée, avec ou sans sites invariables. *A priori*: α suit une distribution uniforme [min , max] ou une distribution exponentielle (par défaut Unif [0.1 , 50]); fraction de sites invariables Unif[0,1].

ν Distribution des longueurs des branches.

1. Cas non raciné
   - Unif [min, max]
   - Exponentielle(λ) (par défaut exp(10))
2. Cas raciné avec horloge moléculaire

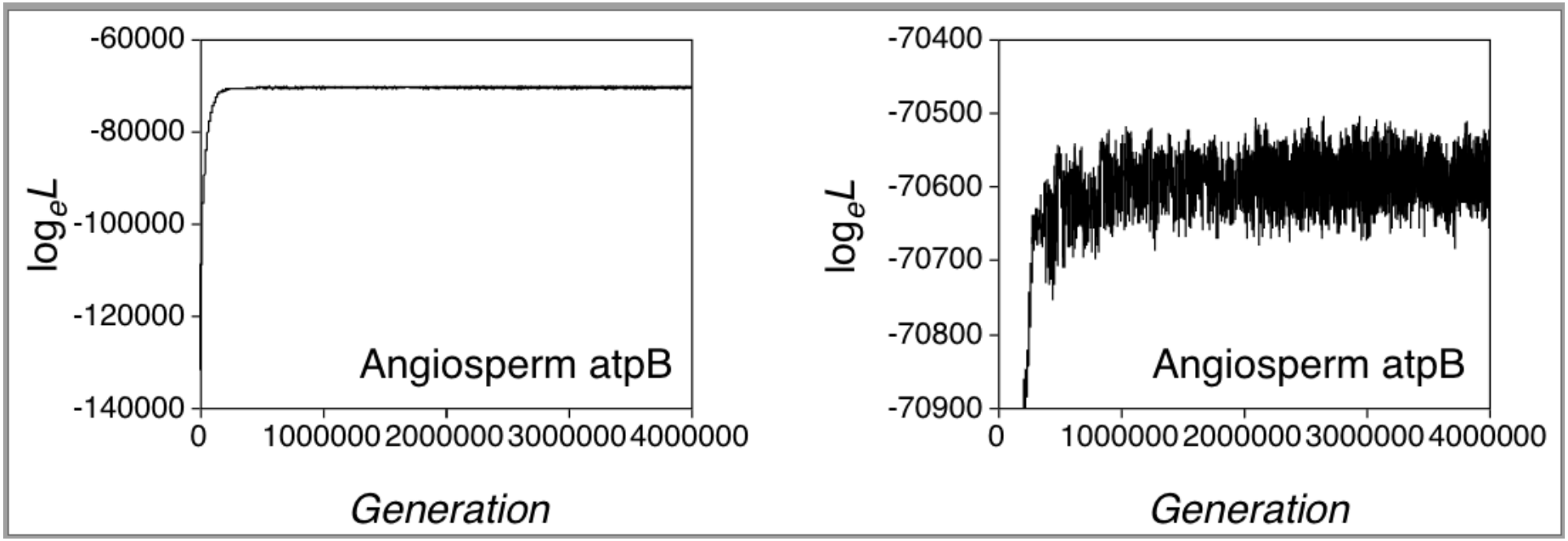
(voir mode d'emploi du programme)

τ Topologies

1. Distribution uniforme (par défaut)
2. Expression de contraintes topologiques (mode d'emploi obscur)

Rapports de proposition (probabilité passage $\lambda_1 -> \lambda_2$) : je n'ai pas trouvé de détail.

Exemple de chaîne de Markov échantillonnant l'espace des paramètres



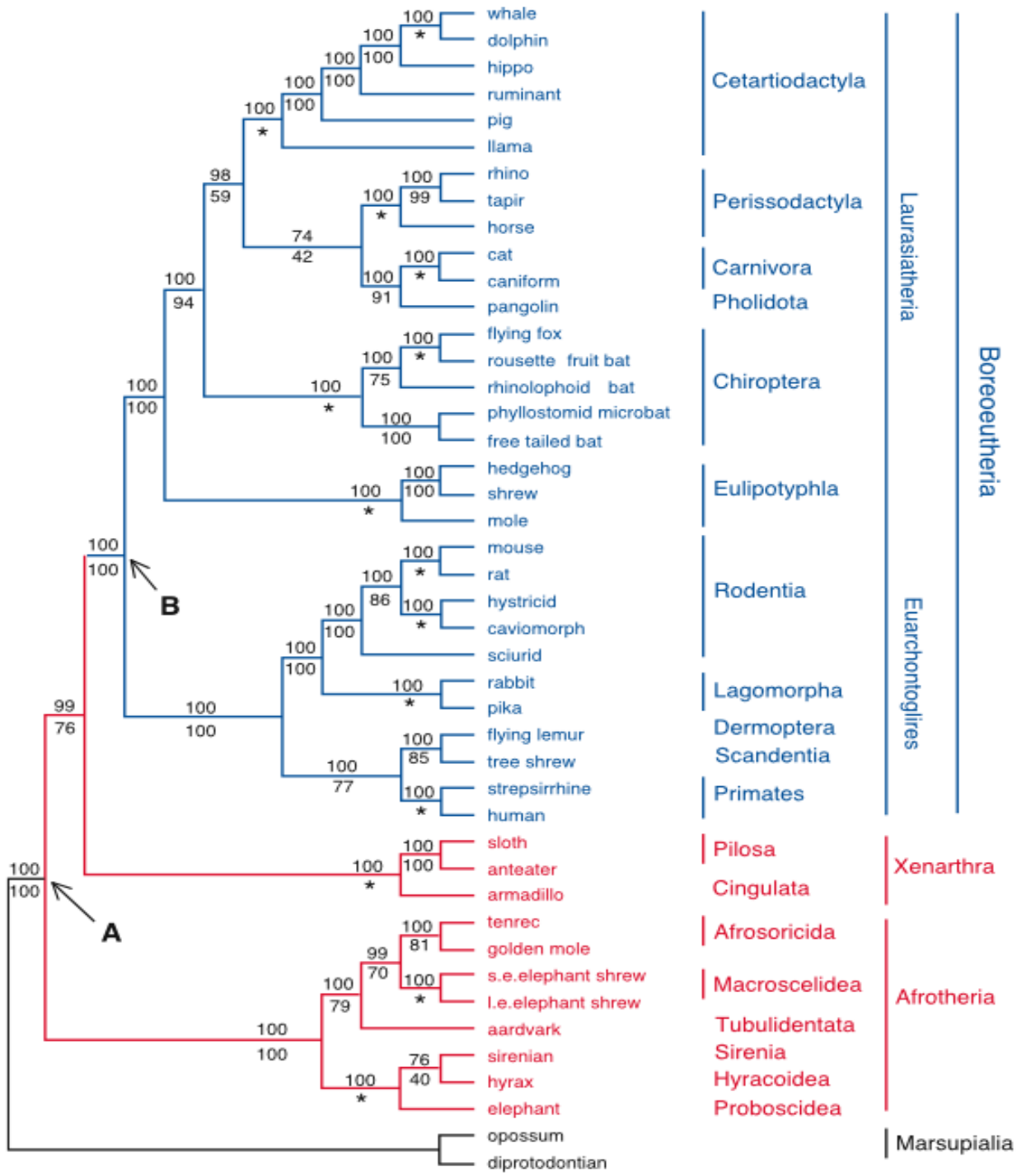
Les graphiques donnent le log de la probabilité d'observer les données.

Résultats du programme MrBayes

On tire un grand échantillon (10^6) dans l'espace des paramètres, dont on ignore le début (10^4), et qu'on sous-échantillonne (un terme sur 100 est retenu).

On obtient alors $\approx 10^4$ arbres.

Expression des résultats:
• liste de toutes les bipartitions trouvées et de leur probabilité *a posteriori*
• arbre consensus à 50 %, et probabilité *a posteriori* de ses clades
• arbre consensus majoritaire, et probabilité *a posteriori* de ses clades



Analyse bayesienne de la phylogénie des mammifères placentaires.

Arbre sans longueur de branches.

au dessus des branches: proba postérieure des clades;
au dessous: soutien de bootstrap par ML.

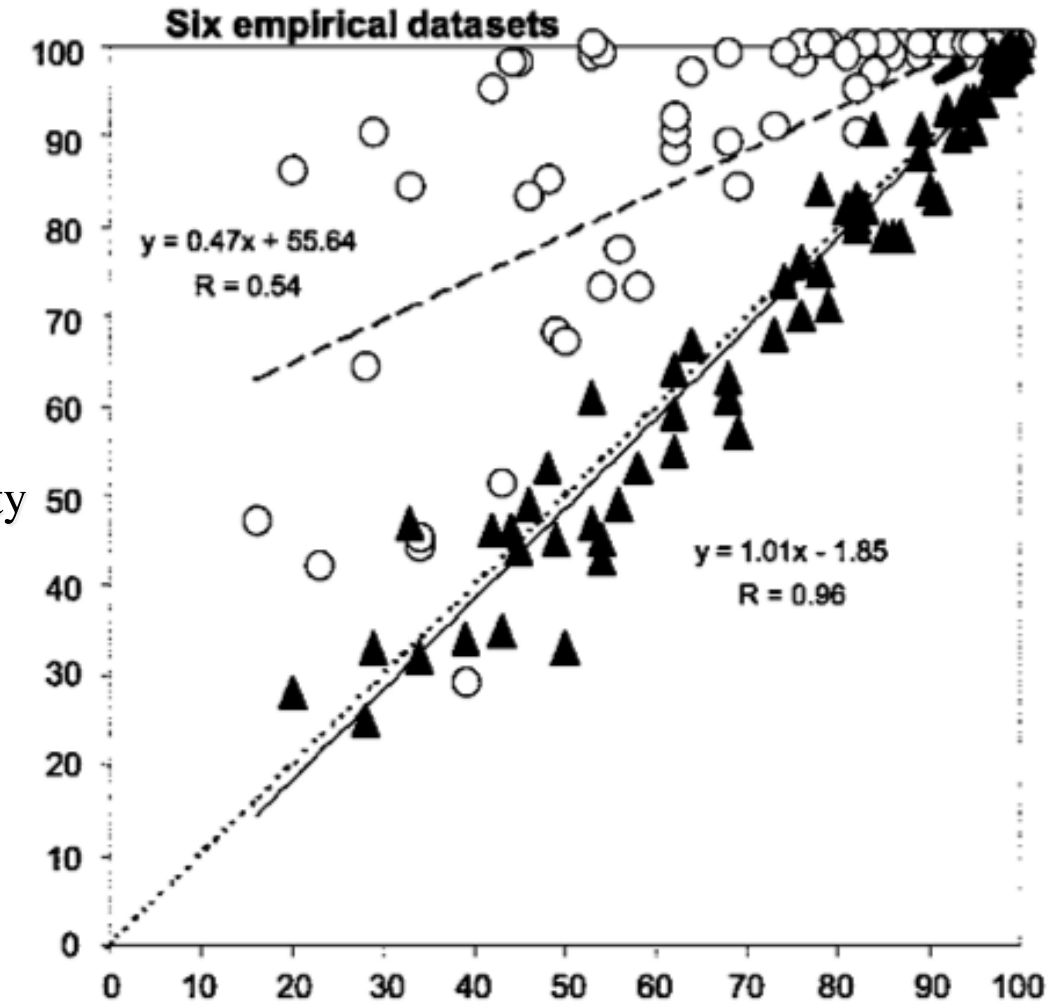
Murphy et al. (2001) Science 294:2348

Surestimation bayesienne du soutien des clades ?

Le soutien bayesien des clades est très supérieur au soutien par bootstrap

○ Bayesian Posterior probability

▲ Bootstrapped Posterior probability



Douady et al. (2003) Mol. Biol. Evol. 20:248–254

Ainsi,

Le soutien bayesien des clades est élevé
Le soutien de bootstrap des clades est faible
Lequel est le plus proche de la valeur statistique exacte?

Conclusion d'expériences de simulation :

- Quand l'évolution des séquences suit exactement les hypothèses du modèle, le soutien bayesien est correct et le soutien par bootstrap est pessimiste.

- L'inférence bayesienne est sensible à de faibles écarts entre modèle et réalité du processus évolutif et devient optimiste.

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