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

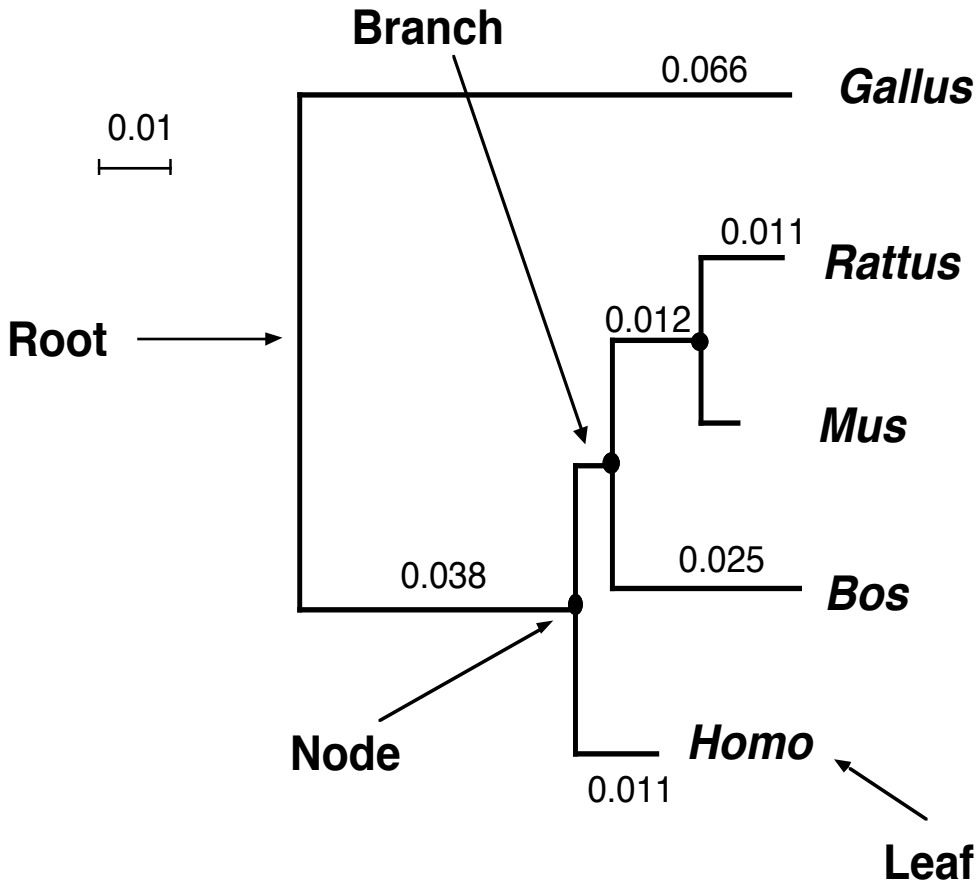
  = une généalogie de séquences
  = un arbre phylogénétique

```
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
```

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

Orthologie / Paralogie

speciation

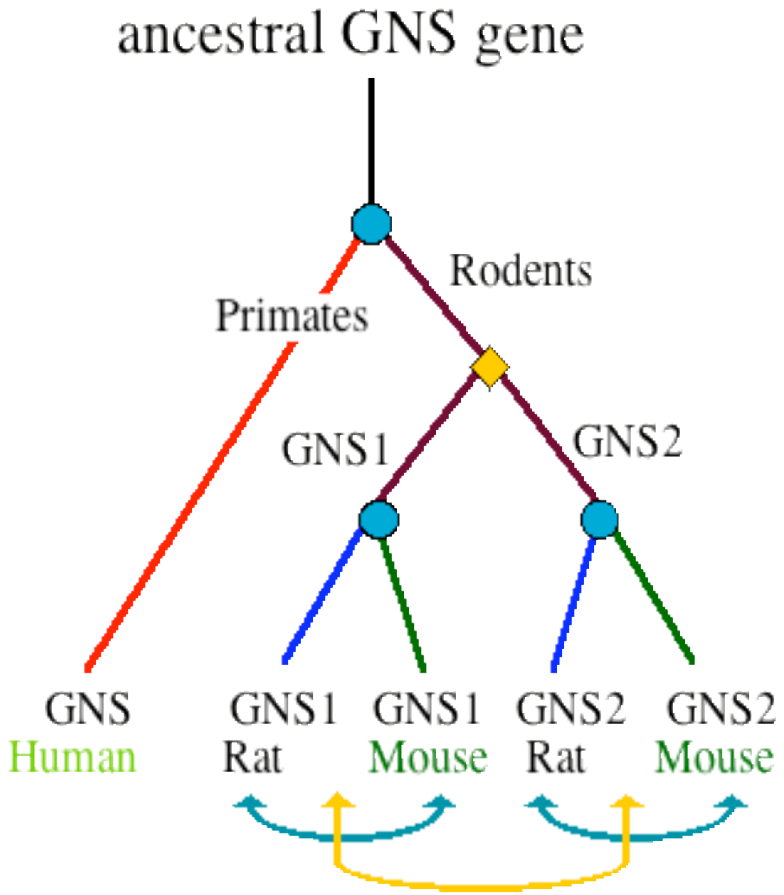
duplication

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

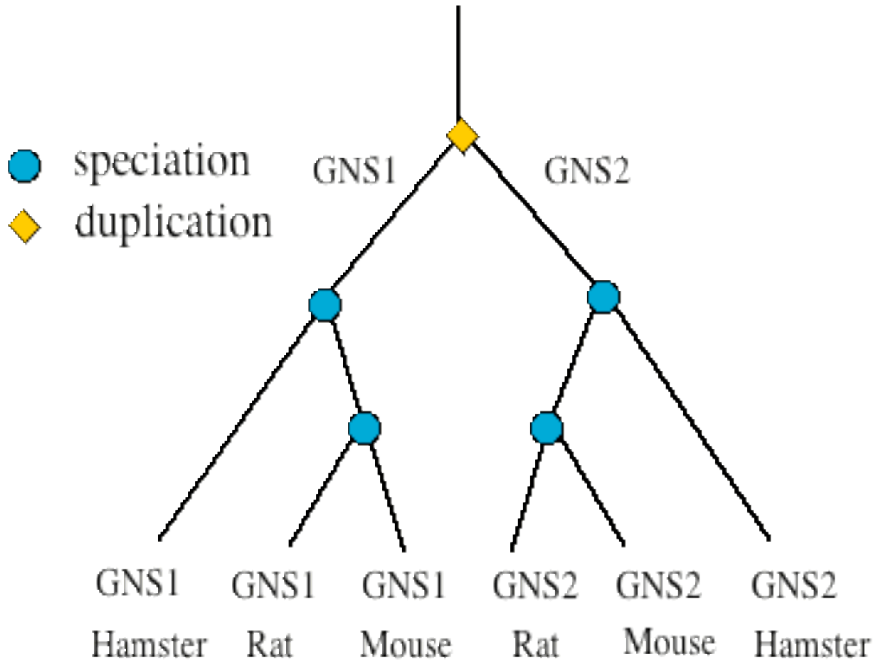
Orthology : two genes are orthologous iff they diverged following a speciation event.

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

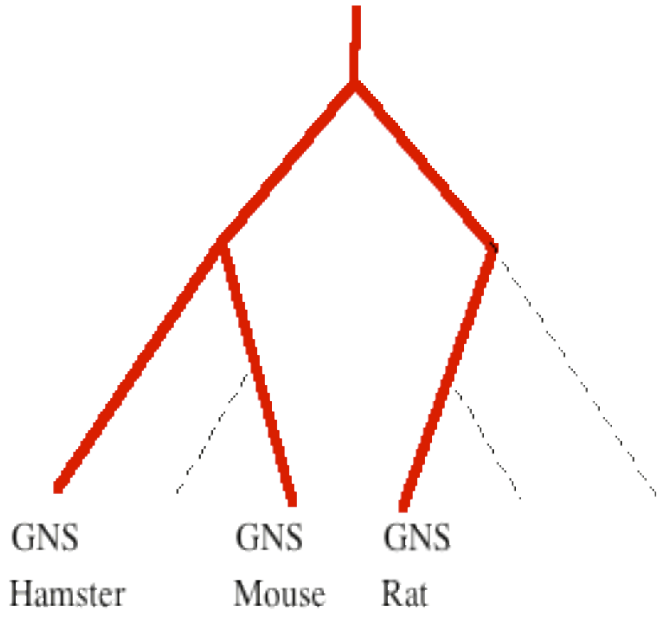
Orthology ≠ functional equivalence



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



true tree



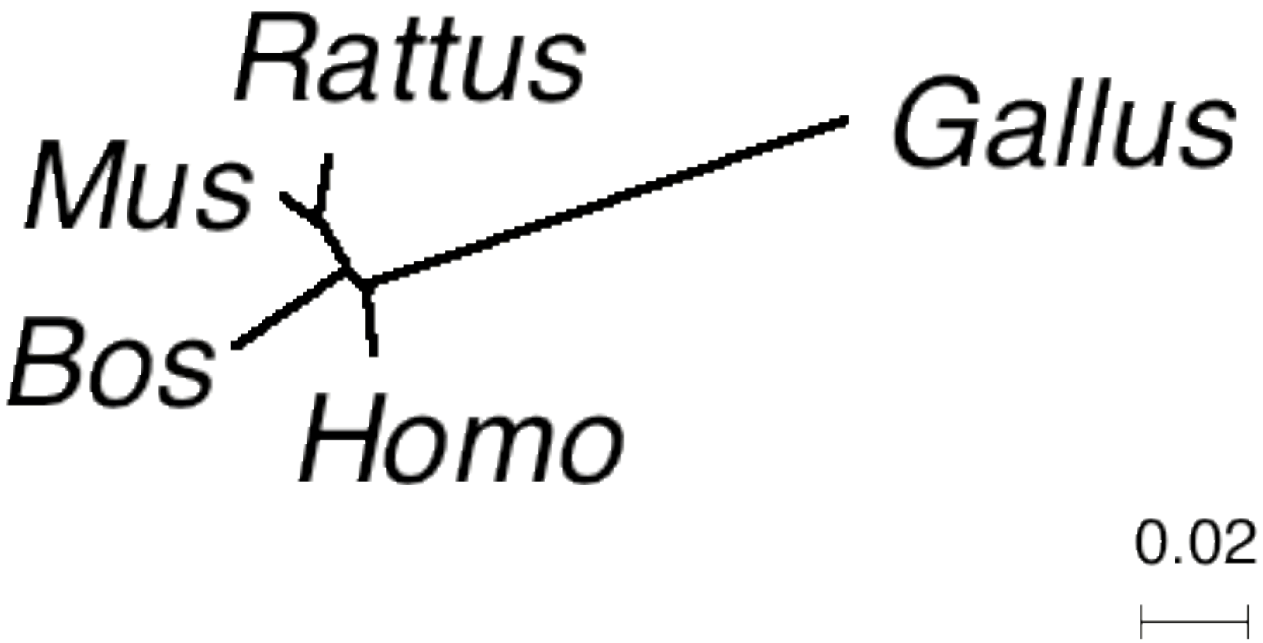
tree obtained with a partial sampling of homologous genes

!! 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 !!

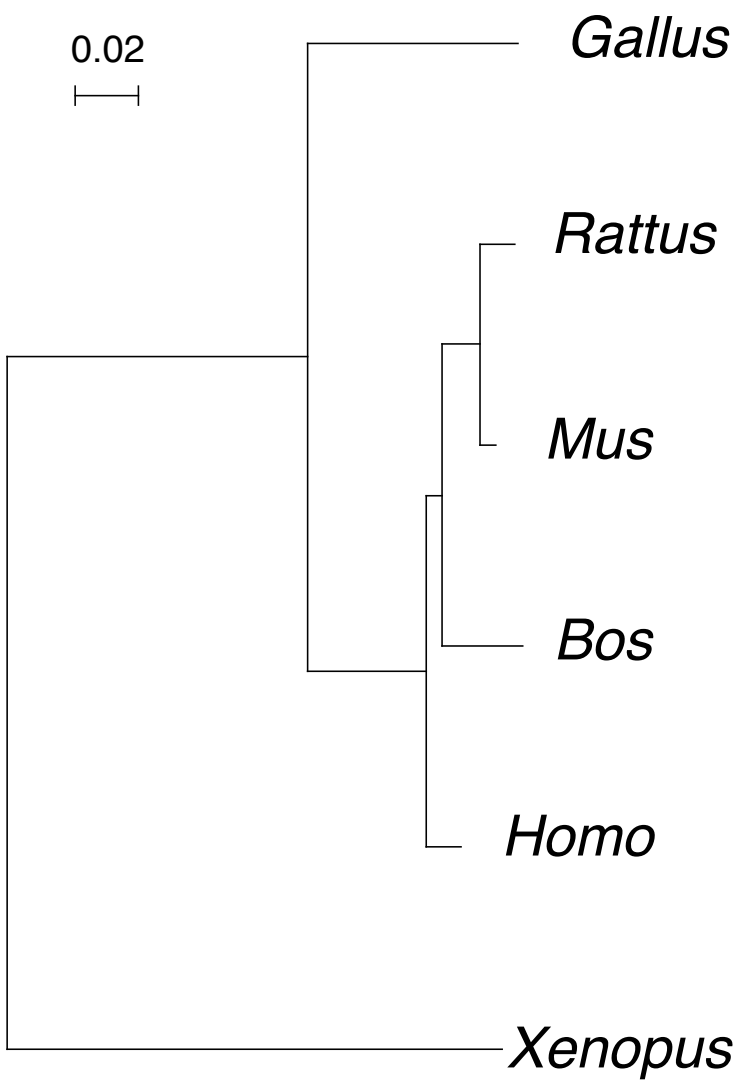
Arbres racinés et non-racinés

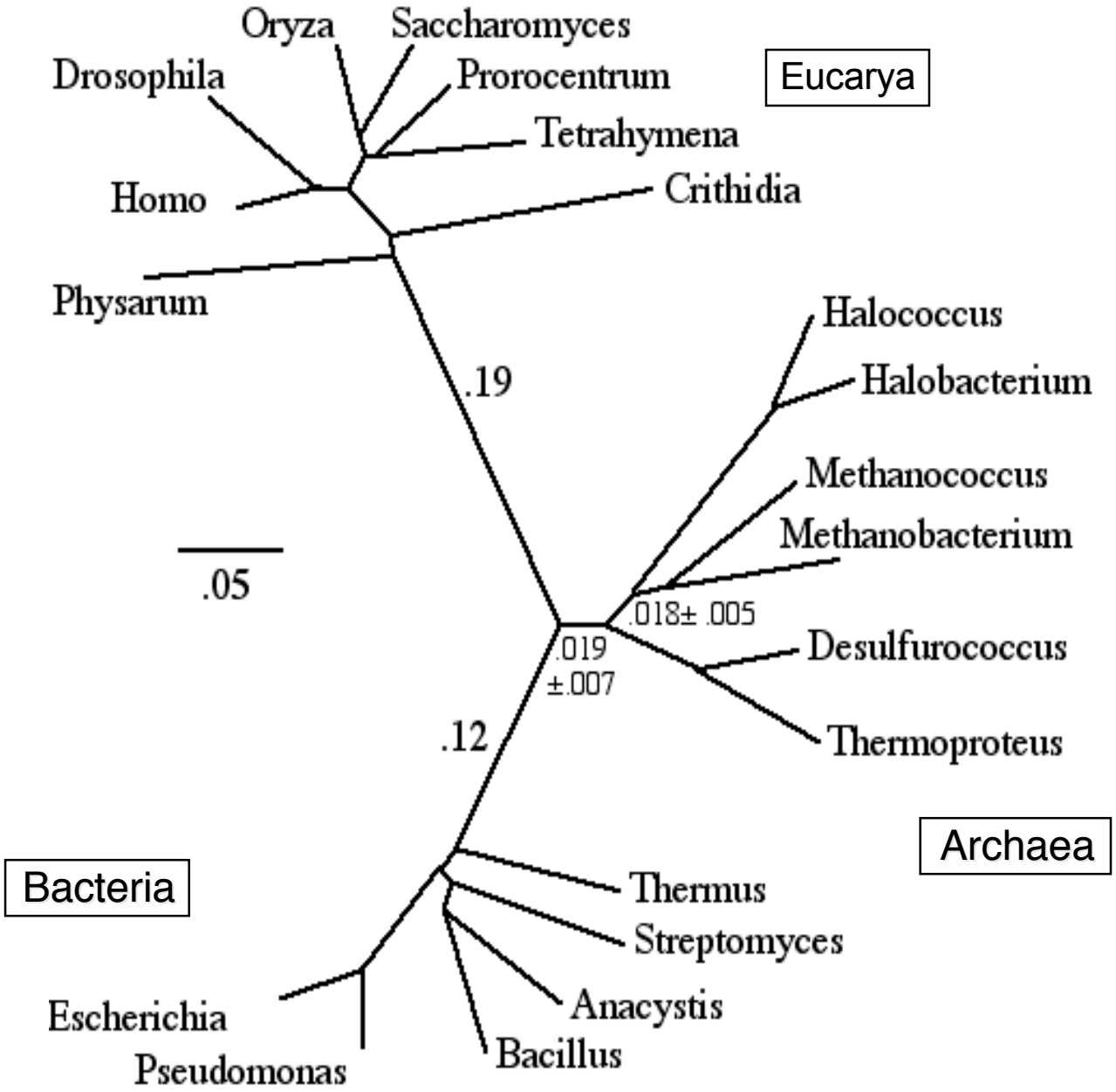
- La plupart des méthodes phylogénétiques produisent des arbres non racinés. La raison est que les méthodes détectent des différences entre séquences, sans avoir le moyen de les orienter temporellement.
- Deux façons d'enraciner un arbre non raciné:
  - Méthode du groupe externe : inclure dans l'analyse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



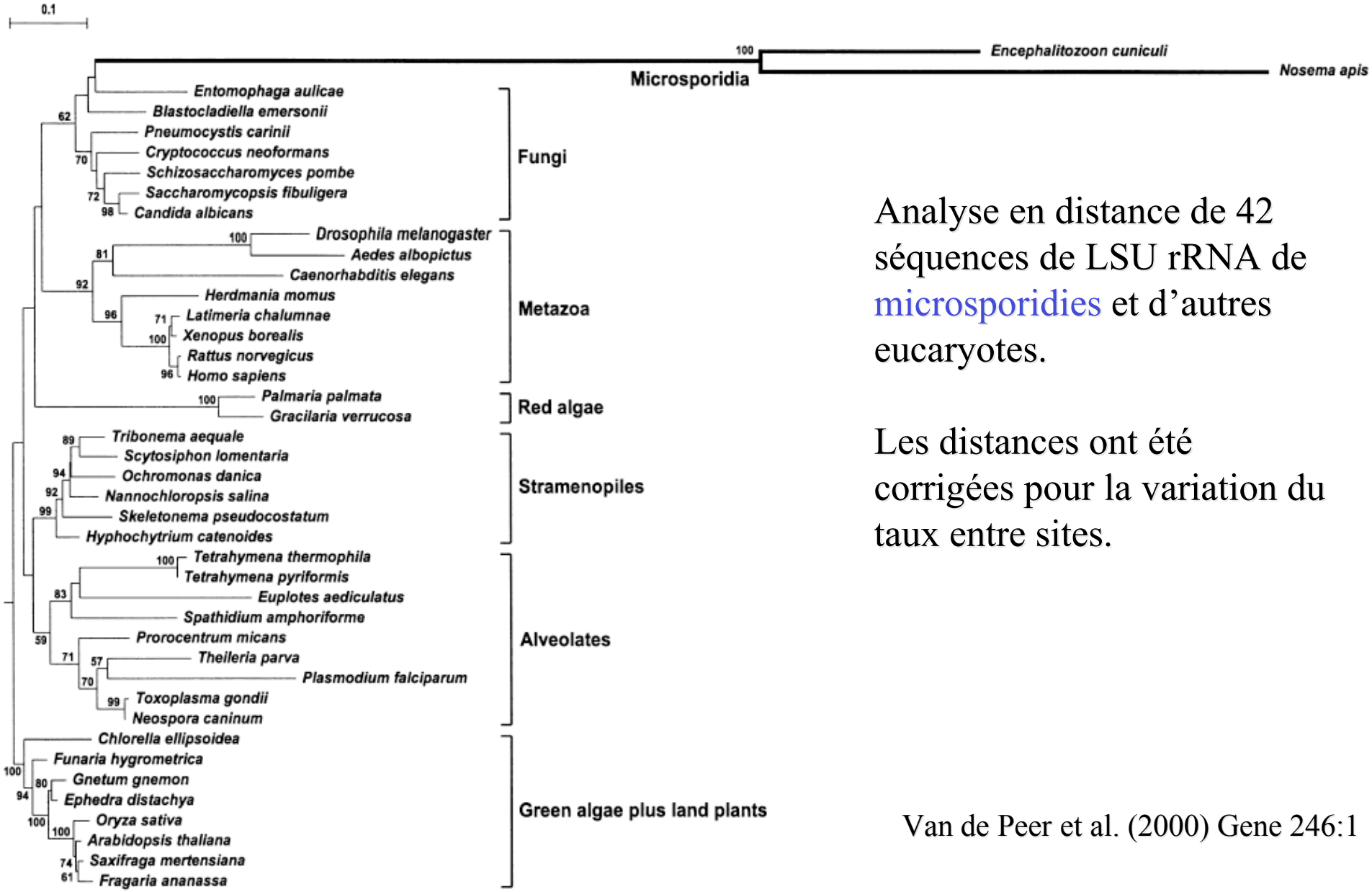


Phylogénie universelle

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.

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

Nombre de topologies d'arbre possibles pour n taxa

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

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

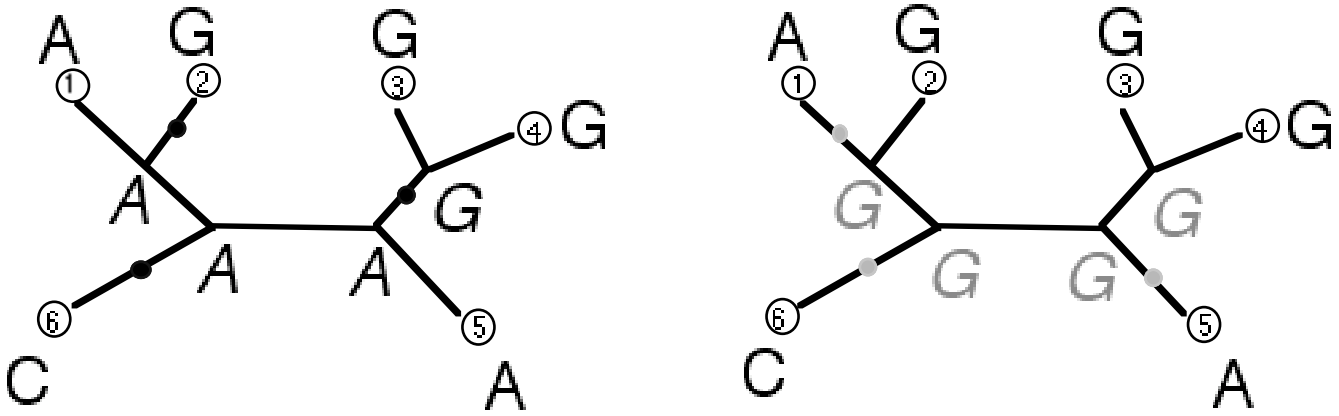
Quatre familles principales de méthodes :

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

Parcimonie (1)

- Etape 1: Pour une topologie d'arbre donnée, et pour un site donné de l'alignement, déterminer quels résidus ancestraux (aux feuilles de l'arbre) nécessitent 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.

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'abre 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.

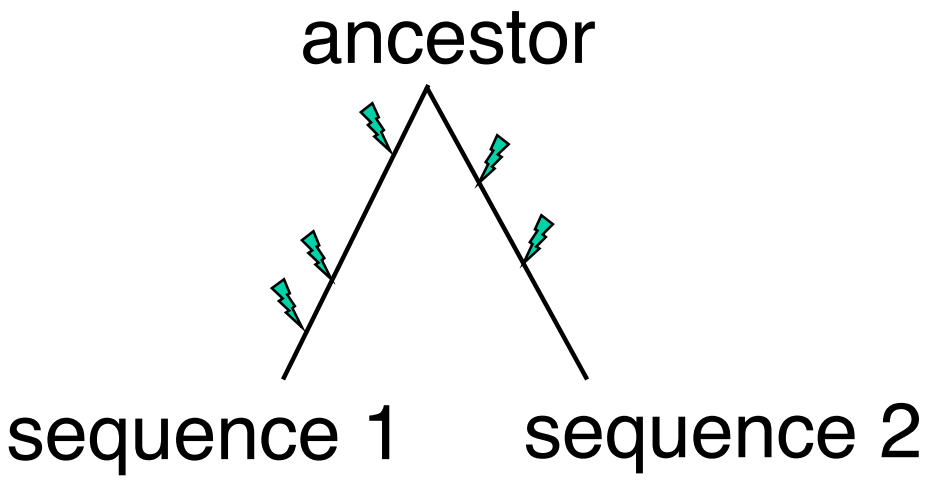
Quelques propriétés de la Parcimonie

- Plusieurs arbres peuvent être également parcimonieux (même longueur, la plus petite de toutes).
- 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.
- 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.

Distance Evolutive

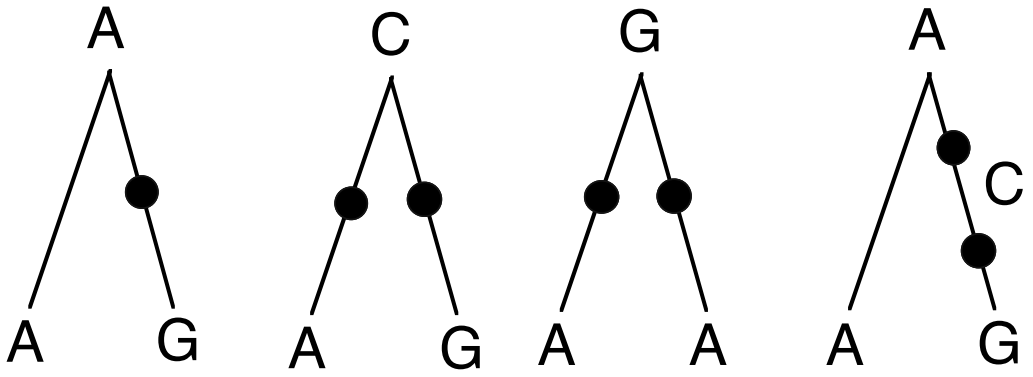
- Elle mesure le nombre de substitutions produites sur les 2 lignées depuyis leur divergence.
- Divisée par la longueur des séquences.
- Exprimée en substitutions / site



Quantification des distances évolutives (1):

Le problème des changements multiples ou cachés

- *d* (vraie distance évolutive) ≥ fraction de différences observées (*p*)



- $d = p$ + changements cachés
- Au prix d'hypothèses de régularité du processus évolutif, il devient possible d'estimer *d* à partir des différences observées entre les séquences.

Quantification des distances évolutives (2):

distance de Jukes & Cantor (DNA)

- Hypothèses du modèle (Jukes & Cantor, 1969) :
  (a) Tous les sites évoluent indépendamment et selon le même processus.
  (b) Toutes les substitutions ont la même probabilité.
  (c) Les probabilités de substitutions sont constantes dans le temps.

- Quantification de la distance évolutive (d) en fonction de la fraction de différences observées (p):

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

p	d
0,10	0,11
0,20	0,23
0,40	0,57
0,60	1,21
0,75	+ ∞

Quantification des distances évolutives (2):

distance de Kimura à 2 paramètres (DNA)

- Hypothèses du modèle :

(a) Tous les sites évoluent indépendamment et selon le même processus.

(b) Il existe deux taux de substitution :

Un pour les transitions, un pour les transversions.

Transitions : G <–>A ou C <–>T Transversions : autres changements

↲	A	T	C	G
A	-	4.4	6.5	20.7
T	4.7	-	21.0	7.2
C	5.0	8.2	-	5.3
G	9.4	3.3	4.2	-

Taux de substitutions relatifs observés sur 13 pseudogènes de mammifères

Transitions
Transversions

Li et coll. 1984

(c) Les probabilités de substitutions sont constantes dans le temps.

Quantification des distances évolutives (3):

distance de Kimura à 2 paramètres (DNA)

- Quantification de la distance évolutive (d) en fonction des différences observées (p: transitions, q: transversions):

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

Kimura (1980) J. Mol. Evol. 16:111

Quantification des distances évolutives(4):

PAM and Kimura's distances (proteins)

- Hypothèses du modèle (Dayhoff, 1979) :

  (a) Tous les sites évoluent indépendamment et selon le même processus.

  (b) Chaque type de remplacement d'acide aminé a une probabilité donnée, empirique (matrices PAM ou Blosum) :
  Un grand nombre de séquences protéiques très similaires ont été compilées, et les probabilités de remplacement de chaque a.a. par chaque a.a. ont été mesurées.

  (c) Le processus de remplacement des acides aminés est à taux constant.

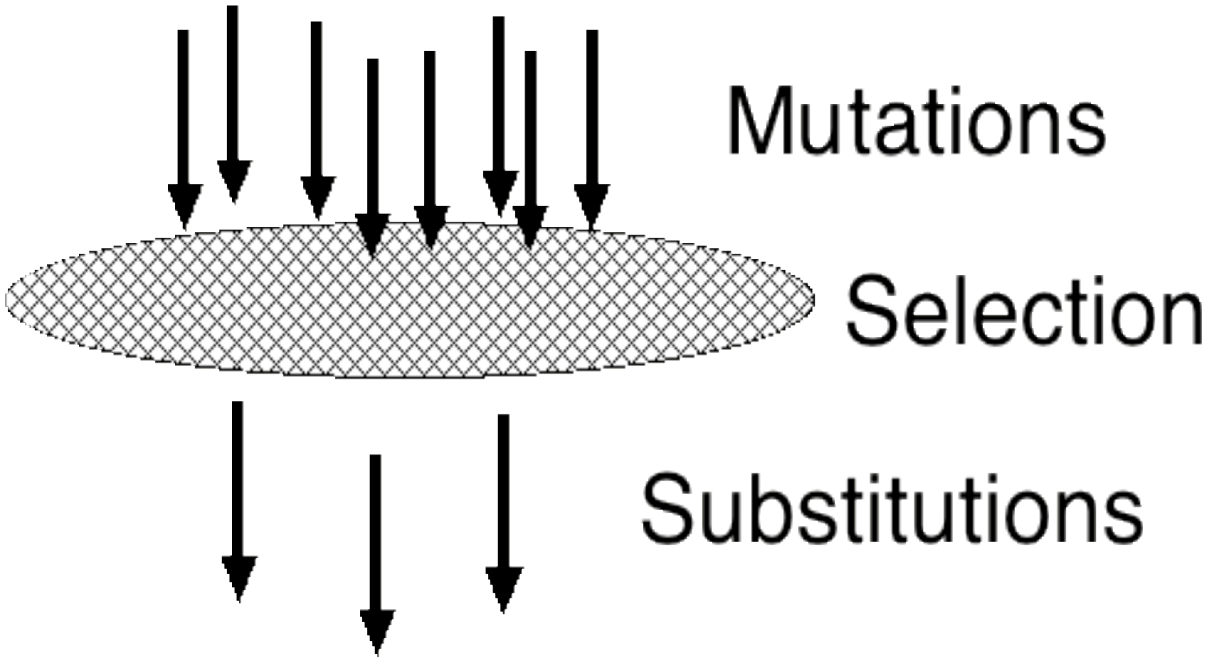
- Quantification de la distance évolutive (d) :
  Le nombre de remplacements le plus compatible avec le pattern observé de changements étant données les probas des remplacements individuels (distance PAM)

- Approximation empirique de Kimura:

$$d = - \ln(1 - p - 0.2\, p^2)$$

(Kimura, 1983) où p = fraction de différences observées

Taux de Substitution = f (mutation, sélection)



NB: la majorité des mutations sont soit neutres (i.e. sans effet phénotypique), soit délétères. Les mutations avantageuses sont rares.

Quantification des distances évolutives (5):

Distances synonymes and non-synonymes (ADN codant): Ka, Ks

- Hypothèse des modèles précédents :

  (a) Tous les sites évoluent indépendamment et selon le même processus.

- Problème: dans les gènes protéiques, il y a deux classes de sites avec des taux évolutifs très différents
  - Les substitutions non-synonymes (changent l' a.a.): lentes
  - Les substitutions synonymes (ne changent pas l' a.a.): rapides
- Solution: calculer deux distances évolutives
  - Ka = distance non-synonyme

    = nbre substitutions non-synonymes / nbre de sites non-synonymes

  - Ks = distance synonyme

    = nbre substitutions synonymes / nbre de sites synonymes

Le code génétique

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

Quantification des distances évolutives (6):

Calcul de Ka et Ks

Les détails de la méthode sont complexes. Brièvement :

- Répartir les sites des gènes comparés en 3 catégories :
  I: non dégénérés, II: partiellement dégénérés, III: complètement dégénérés
- Calculer le nombre de sites non-synonymes = I + 2/3 II
- Calculer le nombre de sites synonymes = III + 1/3 II
- Calculer le nombre de changements synonymes et non-synonymes
- Calculer, selon la méthode de Kimura à 2 paramètres, Ka et Ks

Li, Wu & Luo (1985) Mol.Biol.Evol. 2:150

Ka et Ks : exemple

# sites	observed diffs.	J & C	K2P	K_A	K_S
10254	0.077	0.082	0.082	0.035	0.228

Urotrophin de rat (AJ002967) et souris (Y12229)

Souvent, l'une de ces deux situations se produit :

- Séquences évolutivement proches : Ks est informatif, Ka ne l'est pas.
- Séquences évolutivement distantes : Ks est saturé , Ka est informatif.

Détection de sélection par le rapport Ka/Ks

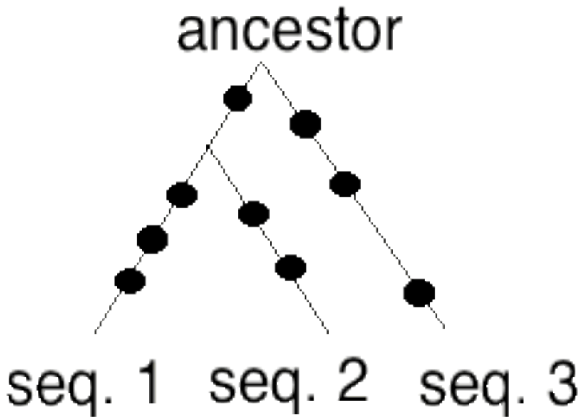
- Ka/Ks ≈ 1 ; neutralité
  exemple : pseudogènes

- Ka/Ks < 1 ; sélection purificatrice
  cas le plus fréquent pour les gènes protéiques

- Ka/Ks > 1 ; sélection positive -
  cas rare indicateur d'adaptation au niveau moléculaire

Saturation: perte du signal phylogénétique

- Quand les séquences homologues comparées ont subi trop de substitutions depuis leur divergence, il est impossible de reconstituer leur histoire phylogénétique, quelle que soit la méthode employée.



- NB: avec les méthodes de distance, le phénomène de saturation peut se manifester par l'impossibilité mathématique de calculer *d*.

  Exemple: Jukes-Cantor: $p \rightarrow 0{,}75 \Rightarrow d \rightarrow \infty$

- NB: souvent la saturation n'est pas détectable

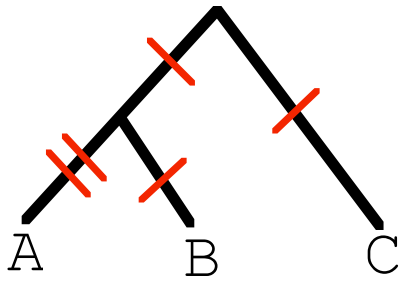
Quantification des distances évolutives (7):

Autres mesures de distance

- Plusieurs autres modèles , plus réalistes, du processus évolutif au niveau moléculaire ont été proposés
  - composition en bases biaisées (Tajima & Nei).
  - Prise en compte de la variation du taux d'évolution entre sites.
  - etc ...

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

Construction d'arbres phylogénétiques par méthodes de distances

Principe général :

- Alignement de séquences

⬇ (1)

- Matrice des distances évolutives entre paires de séquences

⬇ (2)

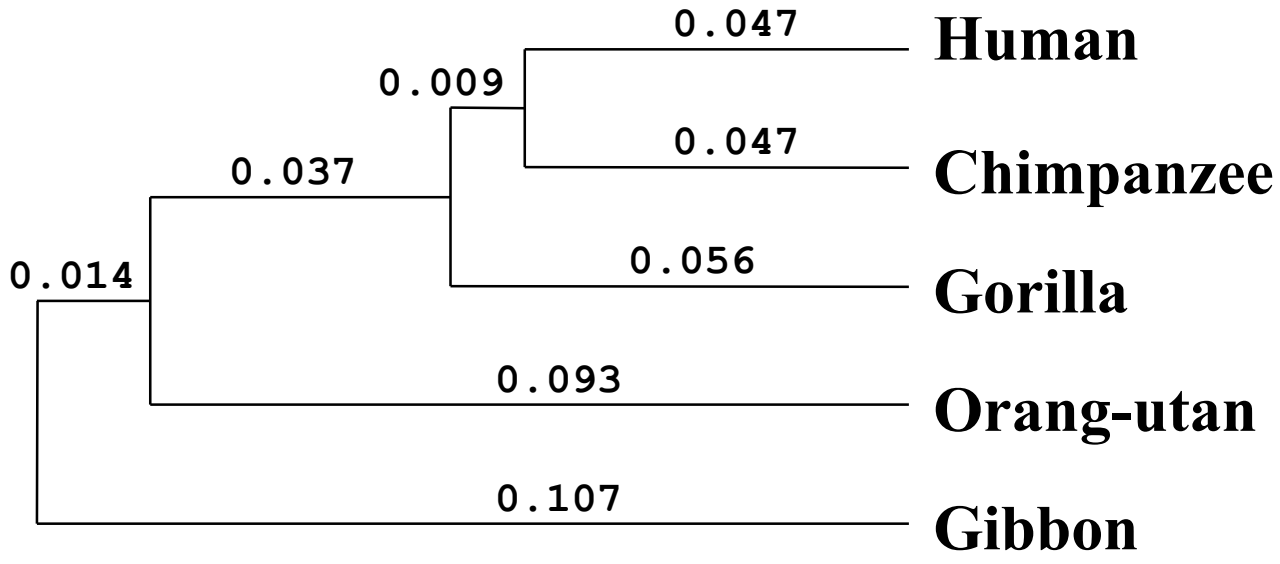
- Arbre (non raciné)

- (1) Measure des distances évolutives.
- (2) Calcul d'un arbre à partir des distances.

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	-

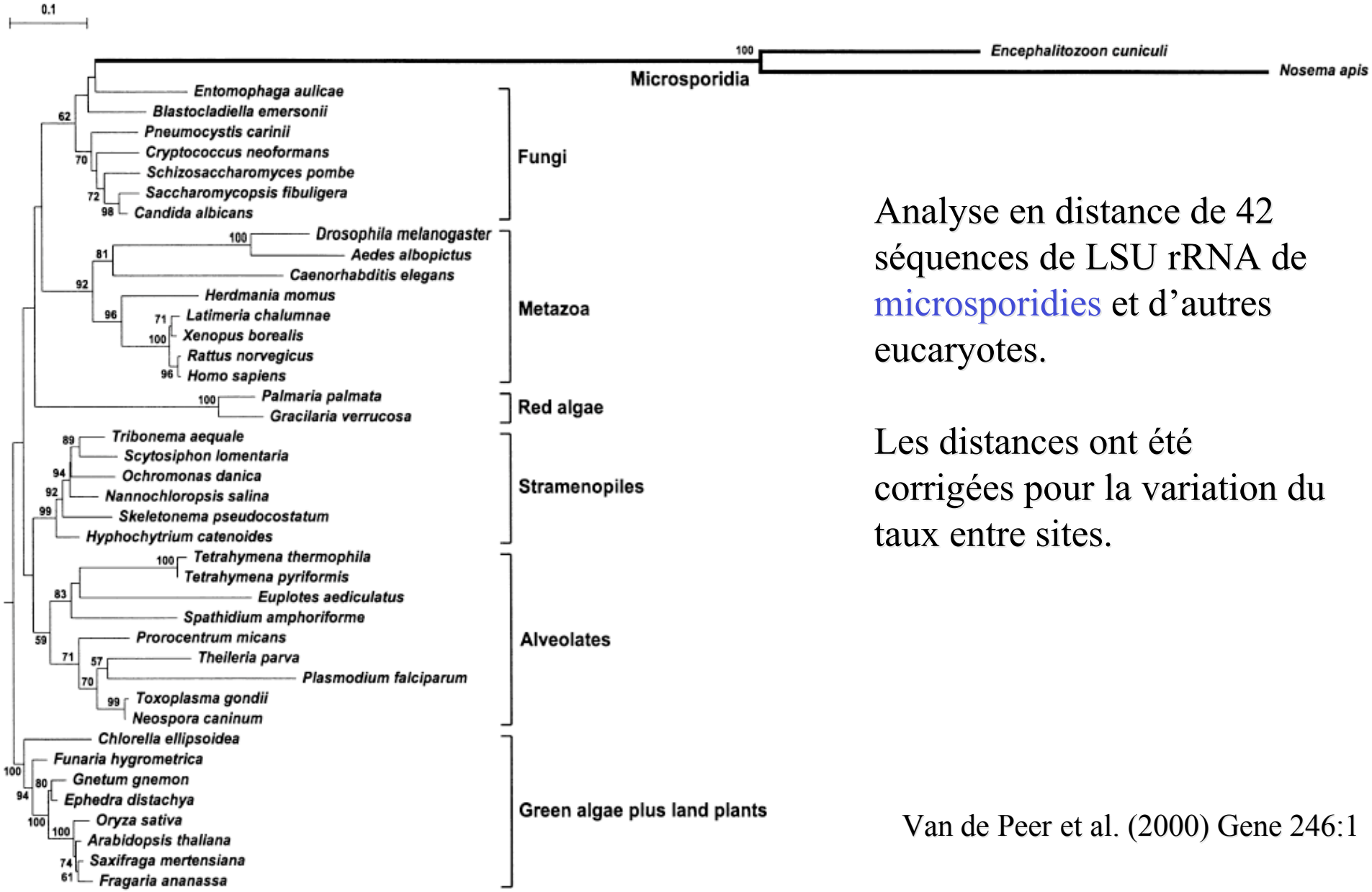
Proportion de différences (*p*) (au dessus de la diagonale) et distances de Kimura à 2 paramètres (*d*) (au dessous) pour un fragment d'ADN mitochondrial (895 pb).



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

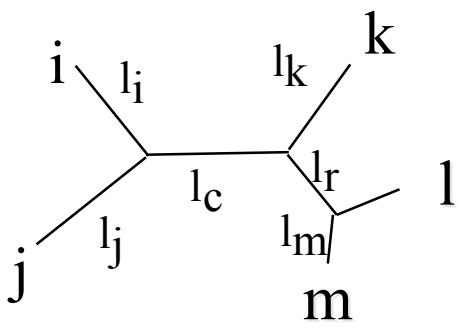
UPGMA : propriétés

- 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 (1):

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

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

S = sum of all branch lengths

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

Matrice de distance -> arbre (2):

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.

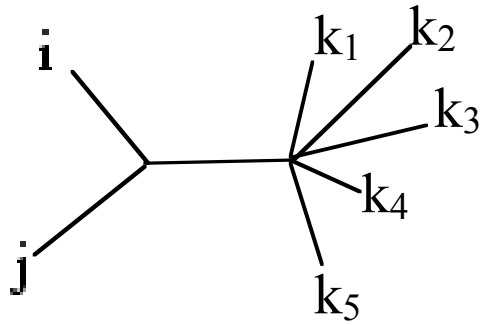
=> une méthode approchée (heuristique) est nécessaire.

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

Matrice de distance -> arbre (3):

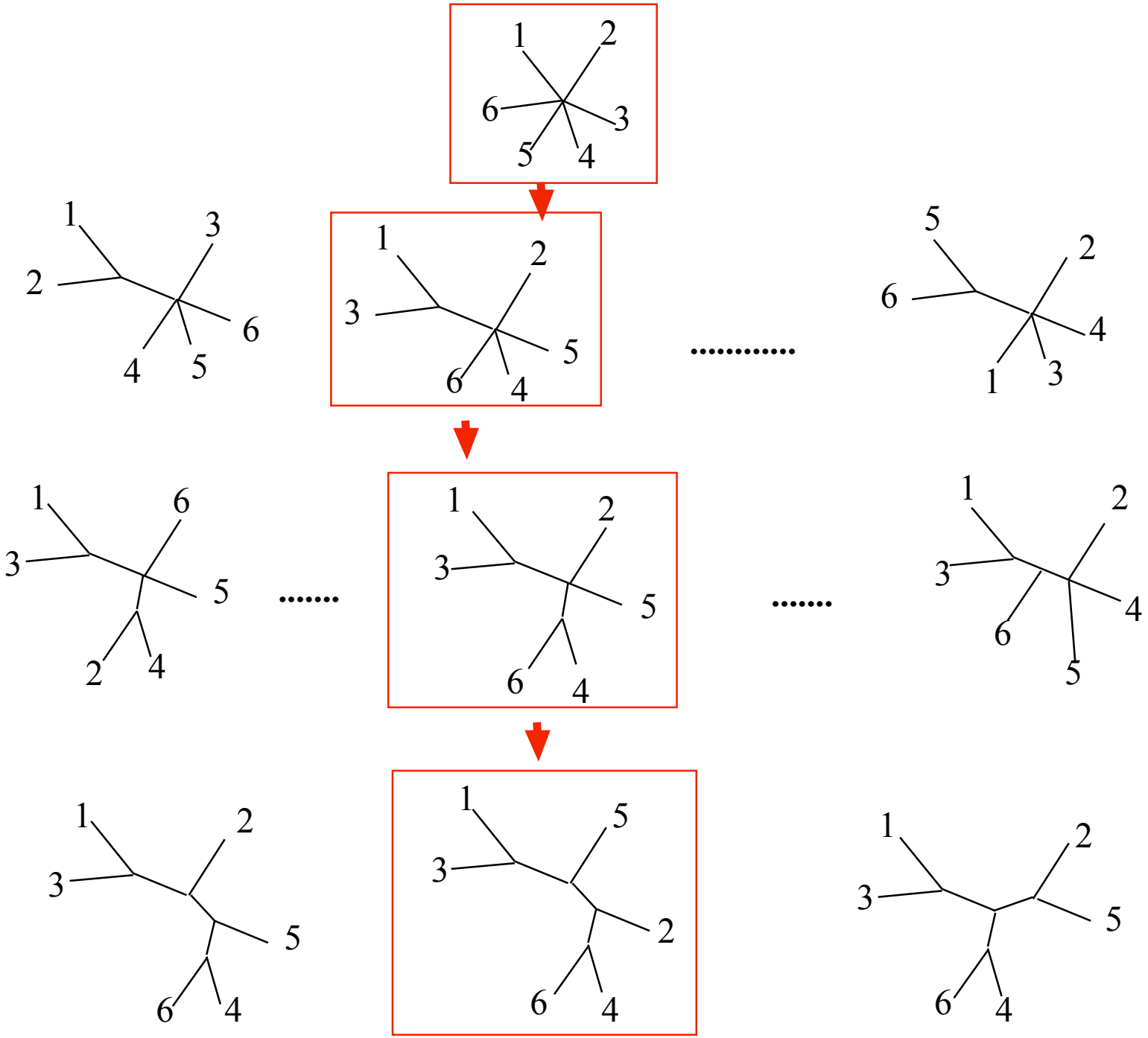
Neighbor-Joining : algorithme

- 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.



- 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 1 tant que $N \geq 4$.

Saitou & Nei (1987) Mol.Biol.Evol. 4:406


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

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)

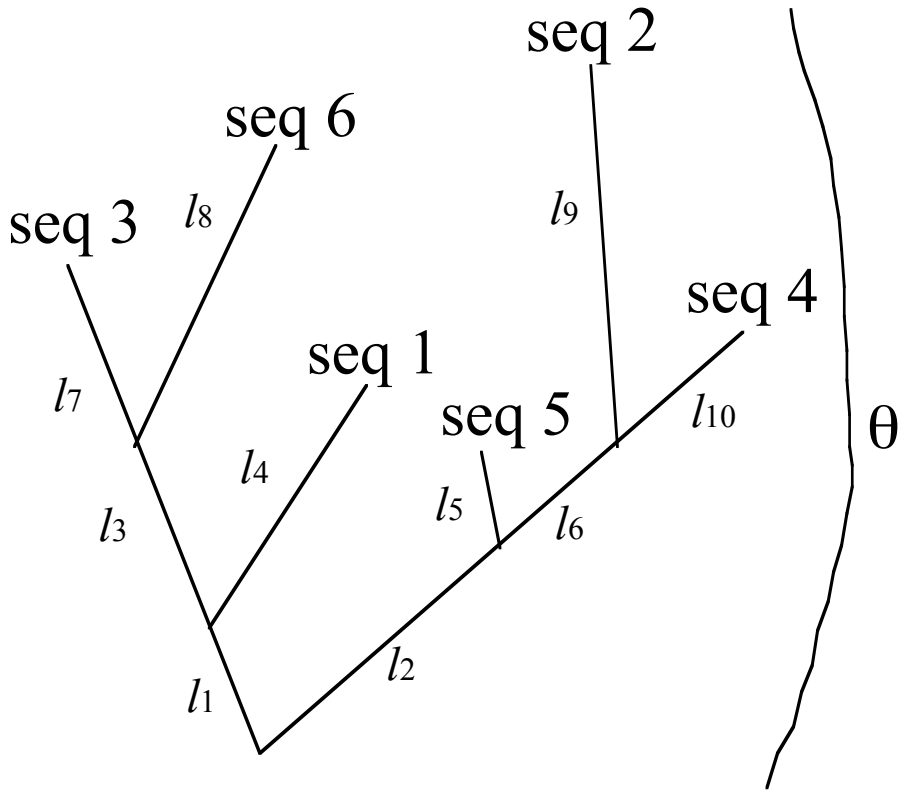
- Hypothèses
  - Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques.
  - Les sites évoluent indépendamment les uns des autres.
  - Les sites évoluent selon la même loi (on peut aussi modéliser la variation des taux entre sites par une loi gamma).

Méthode du Maximum de vraisemblance(2)

Modèle probabiliste de l'évolution de séquences

l_i, longueur de la branche i = nbre attendu de subst. par site le long de la branche

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



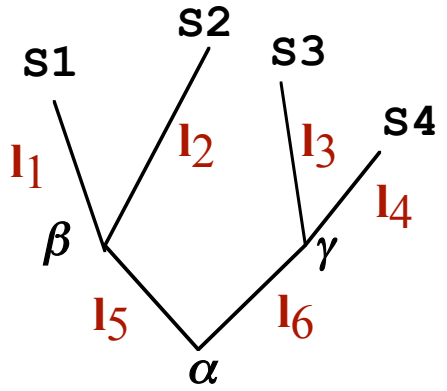
En somme, il faut savoir calculer

$\text{Proba}_{\textit{branche } i}(x \rightarrow y)$

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

Algorithme du maximum de vraisemblance (1)

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



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

α, β, γ : bases ancestrales inconnues et variables

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

P(S1, S2, S3, S4) =

$$\Sigma_\alpha \Sigma_\beta \Sigma_\gamma \, P(\alpha)\, P_{l5}(\alpha,\beta)\, P_{l6}(\alpha,\gamma)\, P_{l1}(\beta,S1)\, P_{l2}(\beta,S2)\, P_{l3}(\gamma,S3)\, P_{l4}(\gamma,S4)$$

où P(S7) est estimée les fréquences moyennes des bases dans les séquences.

Algorithme du maximum de vraisemblance(2)

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

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

- Etape 3: calculer les longueurs des branches *l1, l2, …, l6* et les valeurs du paramètre θ qui correspondent à la valeur maximale de P(Sq1, Sq2, Sq3, Sq4). C'est la vraisemblance de l'arbre.

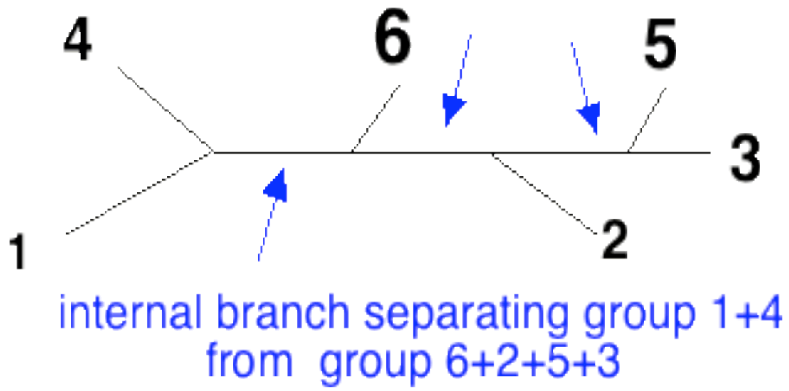
- Etape 4: calculer la vraisemblance de tous les arbres possibles. Retenir l'arbre associé à la plus haute vraisemblance.

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.

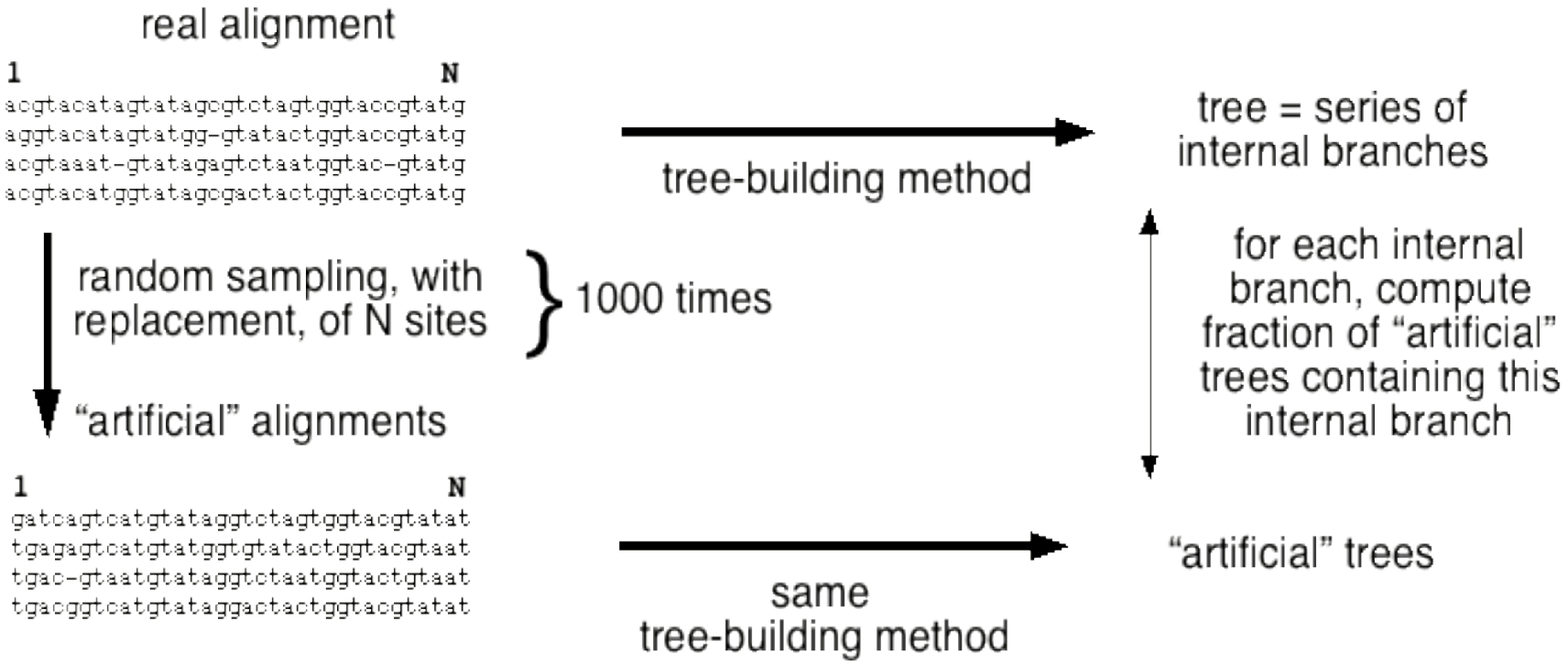
Fiabilité des arbres phylogénétiques: le bootstrap

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



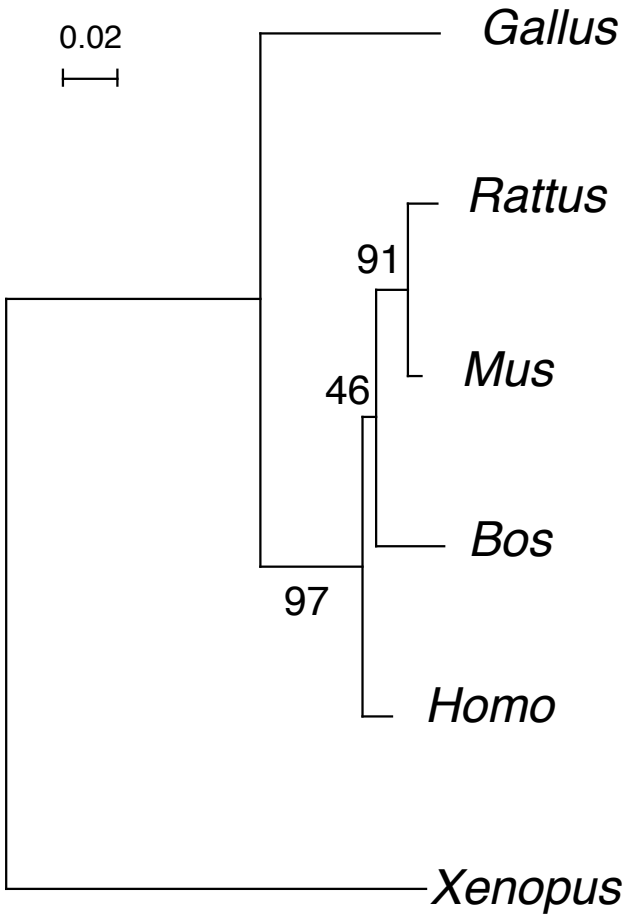
- La forme de l'arbre est déterminée par la liste des branches internes.
- Evaluer la fiabilité d'un arbre = évaluer celle de chaque branche interne.

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 !

Inférence bayesienne d'arbres phylogénétiques

But : calculer la *probabilité postérieure* de toutes les topologies (= formes) d'arbres, étant donné l'alignement de séquences.

$$\Pr(\tau|X) \propto \iint_{v,\theta} \Pr(X|\tau,v,\theta) \;.\; \Pr_{prior}(v,\theta)\,dv\,d\theta$$

likelihood of tree + parameters — *prior probability of parameter values*

τ: topologie d'arbre

X: séquences alignées

v: longueurs des branches de l'arbre

θ: paramètres du modèle de substitution (ex: ratio transit/transv)

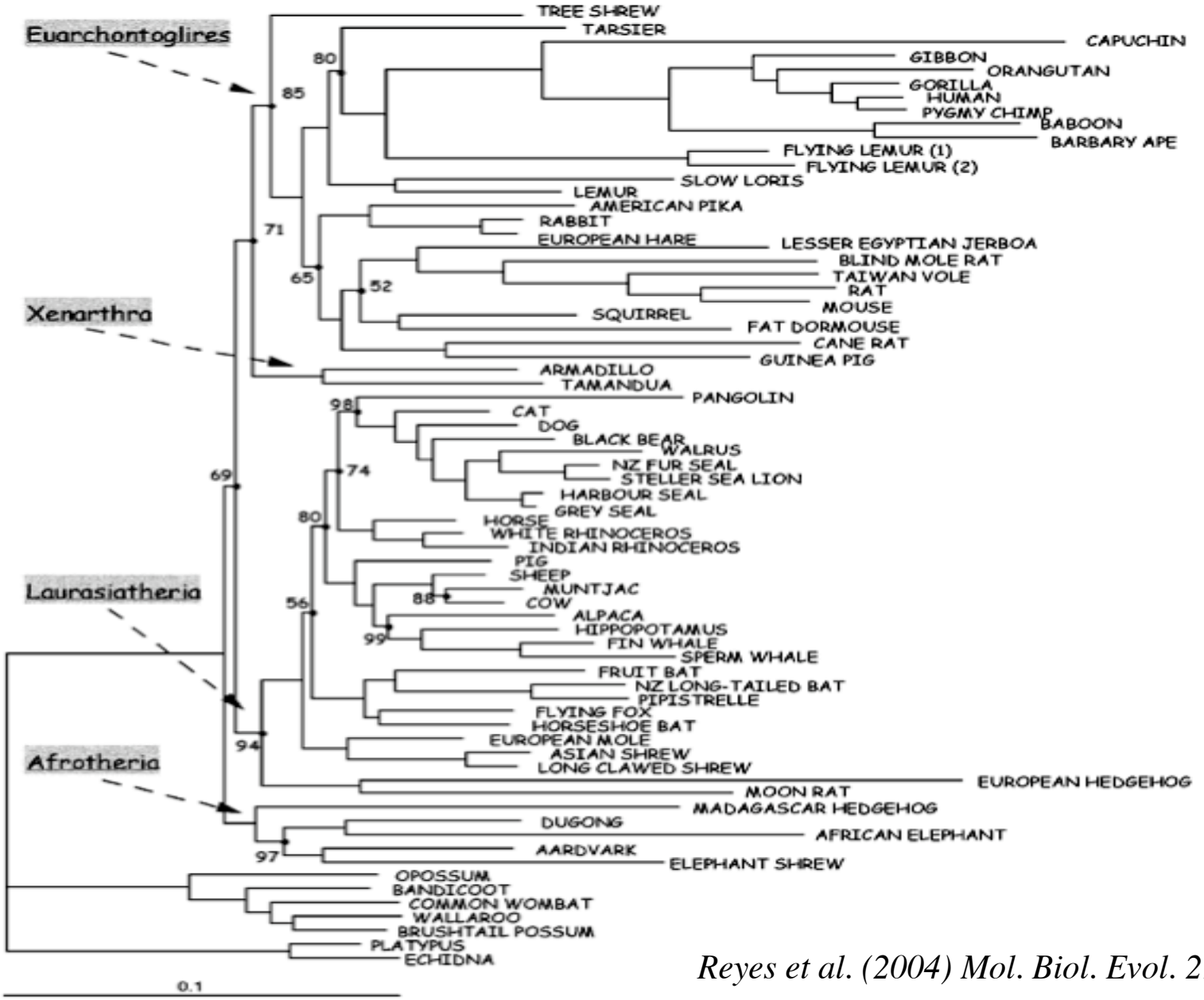
Le calcul direct de $Pr(\tau \mid X)$ est impossible en général.

Une méthode informatique appelée
Metropolis-coupled Markov chain Monte Carlo [MC^3]
permet de générer un échantillon aléatoire de la distribution postérieure des arbres $Pr(\tau \mid X)$.

(exemple: générer un échantillon aléatoire de 10 000 arbres)

Résultat:
- On retient l'arbre ayant la plus forte probabilité (celui trouvé le plus souvent dans l'échantillon).

- On calcule la *probabilité postérieure* de chaque clade de l'arbre: la fraction des arbres échantillonnés contenant ce clade.



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

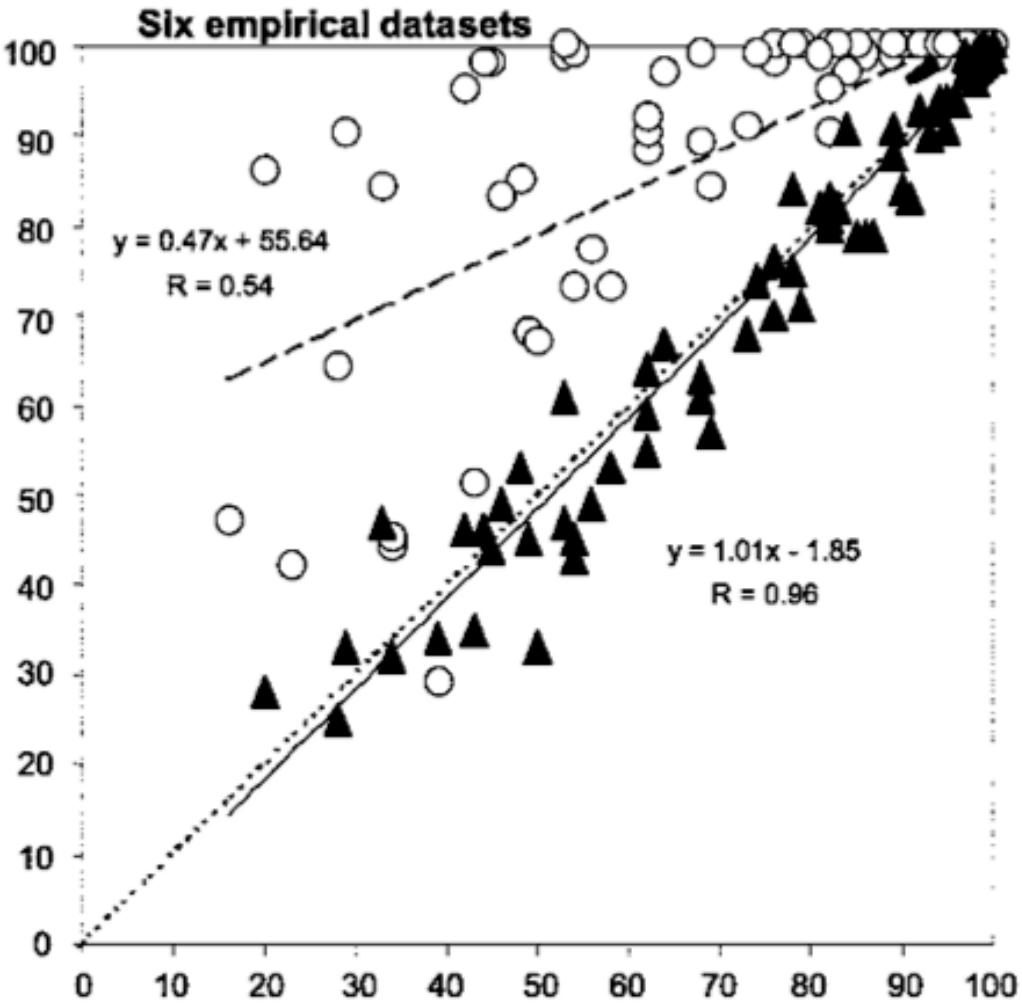
FIG. 1.—Phylogenetic tree of placental mammals reconstructed using the program MrBayes from mitochondrial H-stranded protein-coding genes using ungapped first and second codon positions with the exclusion of Leu synonymous sites. Posterior probabilities (PP) supporting the tree nodes are only reported 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.

Surestimation bayesienne du soutient des clades ?

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

Bayesian Posterior probability

Bootstrapped Posterior probability



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

Ainsi,

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

Conclusion d'expériences de simulation :

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

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

Du nouveau dans l'approche au maximum de vraisemblance

PHYML : a Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood

Guindon & Gascuel (2003) Syst. Biol. 52(5):696–704

ML recherche les valeurs des paramètres quantitatifs (ex: longueurs des branches) et qualitatifs (forme de l'arbre) qui maximisent la probabilité que les séquences observées aient évolué.

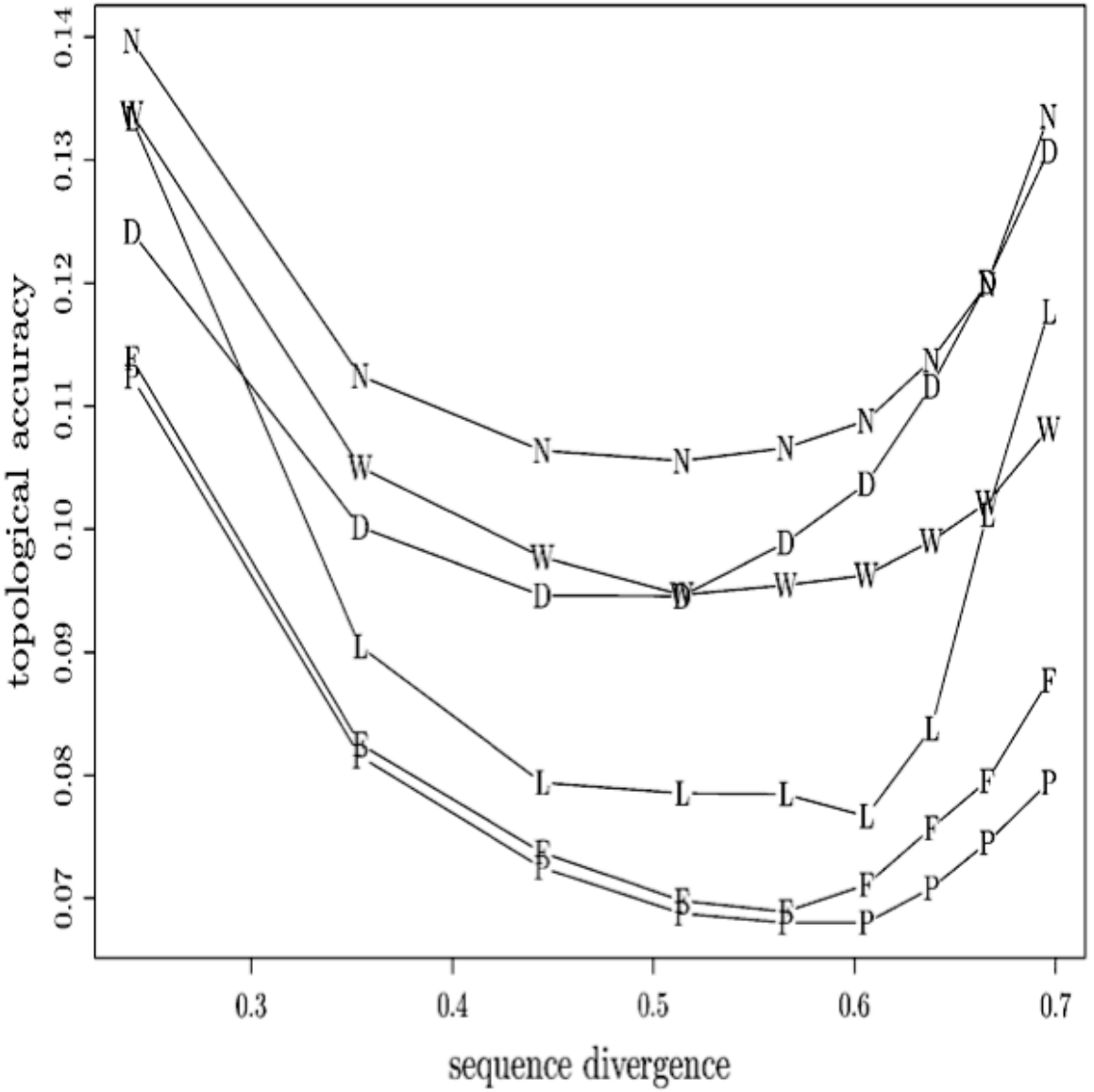
PHYML ajuste topologie et longueurs des branches simultanément.

Parcequ'un faible nombre d'iterations suffisent pour atteindre un optimum, PHYML est un algorithme rapide et précis.

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, $\alpha = 2$



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*

WWW resources for molecular phylogeny (1)

- Compilations
  - A list of sites and resources:
    `http://www.ucmp.berkeley.edu/subway/phylogen.html`
  - An extensive list of phylogeny programs
    `http://evolution.genetics.washington.edu/phylip/software.html`
- Databases of rRNA sequences and associated software
  - The rRNA WWW Server - Antwerp, Belgium.
    `http://rrna.uia.ac.be`
  - The Ribosomal Database Project - Michigan State University
    `http://rdp.cme.msu.edu/html/`

WWW resources for molecular phylogeny (2)

- **Database similarity searches (Blast) :**

`http://www.ncbi.nlm.nih.gov/BLAST/`
`http://www.infobiogen.fr/services/menuserv.html`
`http://bioweb.pasteur.fr/seqanal/blast/intro-fr.html`
`http://pbil.univ-lyon1.fr/BLAST/blast.html`

- **Multiple sequence alignment**

⇨ClustalX : multiple sequence alignment with a graphical interface (for all types of computers).
`http://www.ebi.ac.uk/FTP/index.html` and go to 'software'

⇨Web interface to ClustalW algorithm for proteins:
`http://pbil.univ-lyon1.fr/` and press "**clustal**"

WWW resources for molecular phylogeny (3)

- Sequence alignment editor
  - SEAVIEW : for windows and unix
    `http://pbil.univ-lyon1.fr/software/seaview.html`
- Programs for molecular phylogeny
  - PHYLIP : an extensive package of programs for all platforms
    `http://evolution.genetics.washington.edu/phylip.html`
  - CLUSTALX : beyond alignment, it also performs NJ
  - PAUP* : a very performing commercial package
    `http://paup.csit.fsu.edu/index.html`
  - PHYLO_WIN : a graphical interface, for unix only
    `http://pbil.univ-lyon1.fr/software/phylowin.html`
  - MrBayes : Bayesian phylogenetic analysis
    `http://morphbank.ebc.uu.se/mrbayes/`
  - PHYML : fast maximum likelihood tree building
    `http://www.lirmm.fr/~guindon/phyml.html`
  - WWW-interface at Institut Pasteur, Paris
    `http://bioweb.pasteur.fr/seqanal/phylogeny`

WWW resources for molecular phylogeny (4)

- Tree drawing

  NJPLOT (for all platforms)

  `http://pbil.univ-lyon1.fr/software/njplot.html`

- Lecture notes of molecular systematics

  `http://www.bioinf.org/molsys/lectures.html`

WWW resources for molecular phylogeny (5)

- Books
  - Laboratory techniques
    Molecular Systematics (2nd edition), Hillis, Moritz & Mable eds.; Sinauer, 1996.
  - Molecular evolution
    Fundamentals of molecular evolution (2nd edition); Graur & Li; Sinauer, 2000.
  - Evolution in general
    Evolution (2nd edition); M. Ridley; Blackwell, 1996.