

# Génomique comparative et phylogénies bactériennes

Guy Perrière

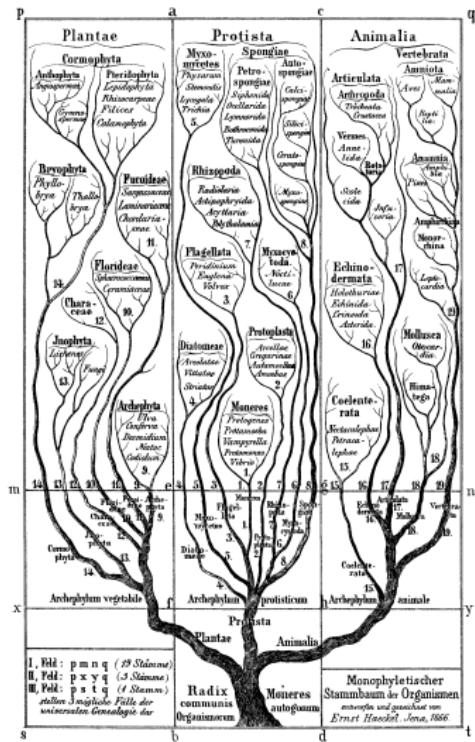
Laboratoire de Biométrie et Biologie Évolutive  
UMR CNRS n° 5558  
Université Claude Bernard – Lyon 1

18 octobre 2018

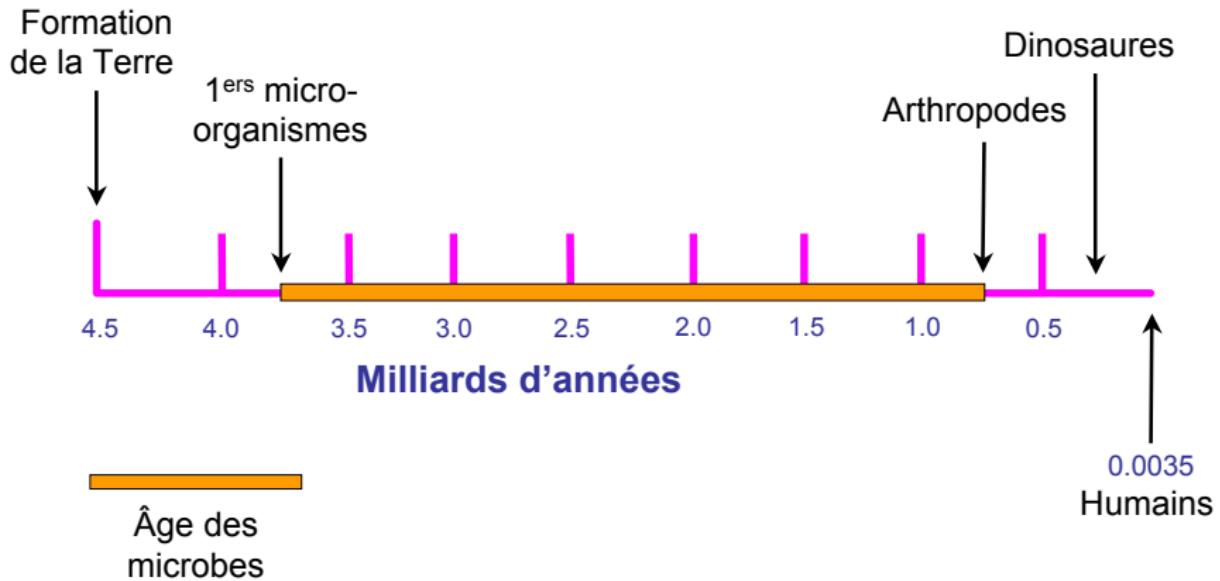
<http://pbil.univ-lyon1.fr/members/perriere/cours/EcoMi/>

# Classification phylogénétique

- Premier arbre « universel » publié en 1866 par Haeckel :
  - Subdivision du vivant en trois règnes :
  - Plantes.
  - Animaux.
  - Protistes.
- Position des organismes microscopiques unicellulaires ?
- Question qui se pose encore aujourd’hui.



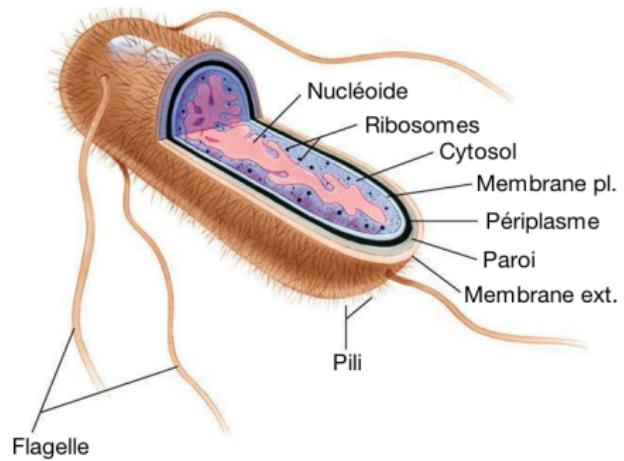
# Histoire de la vie sur Terre



# Les procaryotes

- Comprennent les bactéries et les archées :

- Organisation générale comparable :
    - Organismes unicellulaires dépourvus de noyau.
  - Différences au niveau :
    - Des mécanismes de réplication et de traduction.
    - Des phospholipides de la membrane plasmique.



# Comment les classer

- Critères morphologiques inopérants.
- Pendant longtemps, utilisation de critères biochimiques :
  - Coloration (Gram positives et Gram négatives).
  - Fonctions biochimiques (principe des galeries Api) :

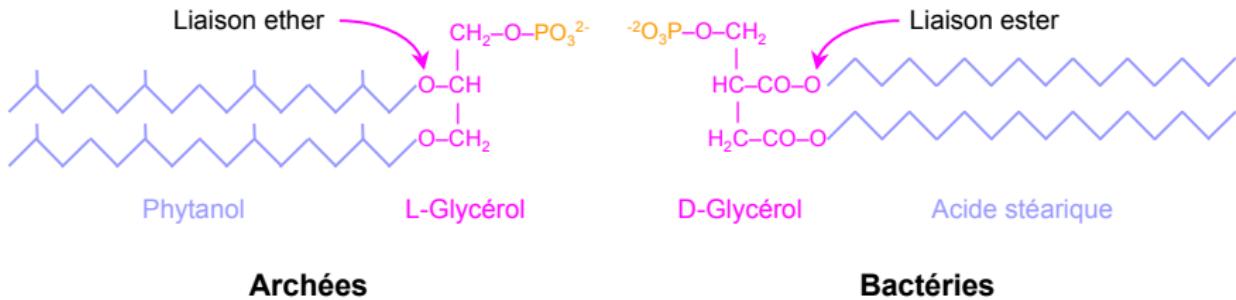


- Depuis 1977, utilisation des données de séquences :
  - Construction de phylogénies moléculaires.

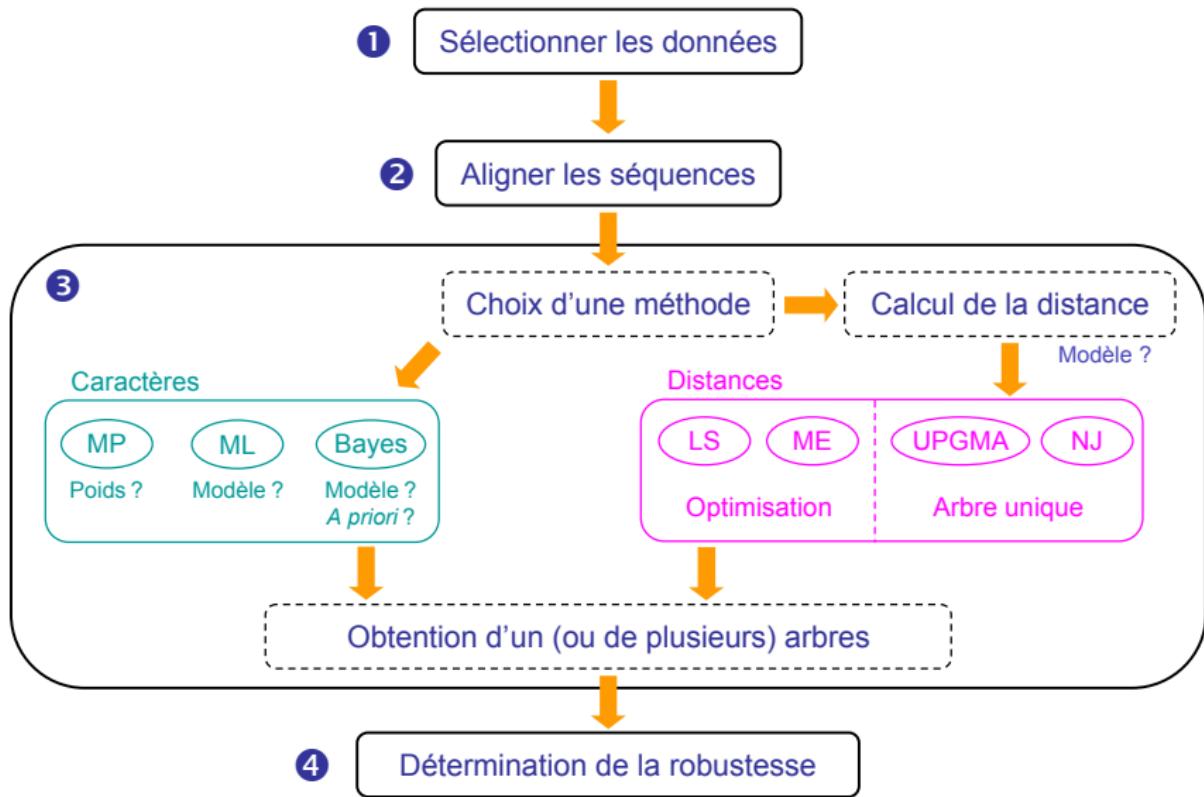
# Paroi des archées

## ■ Structure particulière :

- Pas de peptidoglycane.
- Diglycérides-phosphates membranaires spécifiques :
  - L-Glycérol au lieu de D-Glycérol.
  - Isoprènes au lieu d'acides gras.
  - Liaisons avec le Glycérol de type ether au lieu d'ester.



## Étapes d'une phylogénie



# Typologie des méthodes

- Méthodes fondées sur l'utilisation de caractères :
  - Maximum de parcimonie (*Maximum of Parsimony* – MP).
  - Maximum de vraisemblance (*Maximum Likelihood* – ML).
  - Approche bayésienne.
- Méthodes fondées sur des matrices de distances :
  - Classification ascendante hiérarchique au lien moyen (*Unweighted Pair-Group Method with Arithmetic means*, UPGMA).
  - Moindres carrés (*Least Squares* – LS).
  - Minimum d'évolution (*Minimum of Evolution* – ME).
  - *Neighbor-Joining* (NJ).

# Données utilisées

## ■ Point de départ :

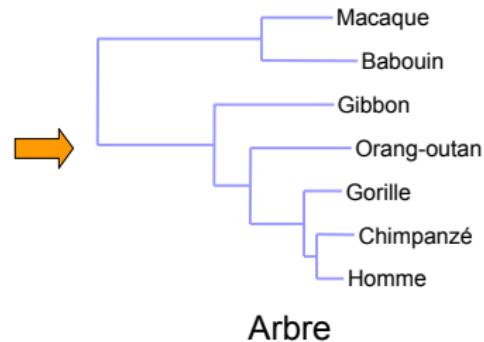
- Un ensemble de séquences *homologues* alignées.
- Chaque position dans l'alignement constitue un *site*.

## ■ Résultat obtenu :

- Un arbre décrivant les relations évolutives entre les séquences (*i.e.*, un arbre phylogénétique).

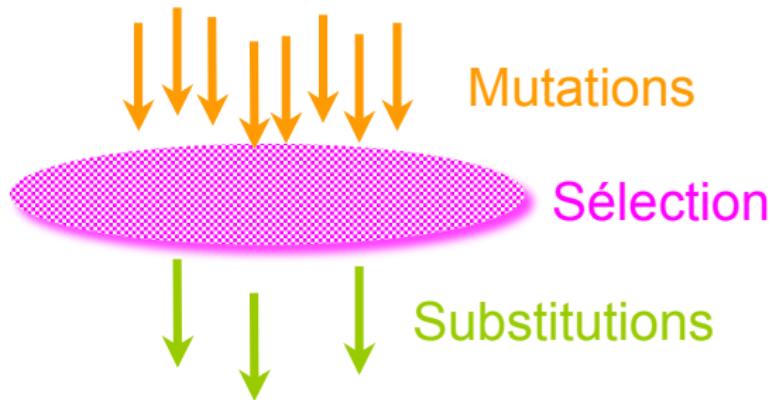
Gibbon	AAGCTTTACAGGTGCAACCGTCCCTCATTAATCGCCCACGGACTAACCTCTT
Orang	AAGCTTCACCGGCCAACCAACCTCATGATTGCCATGGACTCACATCCT
Gorille	AAGCTTCACCGGCCGCAAGTTGTTCTTATAATTGCCACGGACTTACATCAT
Homme	AAGCTTCACCGGCCGCAAGTCATTCCTCATTAATCGCCCACGGACTAACATCCT
Chimpanzé	AAGCTTCACCGGCCGCAATTATCCTCATTAATCGCCCACGGACTAACATCCT
Macaque	AAGCTTTCCGGCGCAACCATCCTTATGATCCTCACGGACTCACCTCTT
Babouin	AAGCTTCTCCGGTGCACCATCCTTATGATTGCCACGGACTCACCTCTT

Alignement



# Mutations et substitutions

- La grande majorité des *mutations* sont soit neutres (*i.e.*, n'ont aucun effet sur le phénotype), soit délétères :
  - Les mutations avantageuses sont très rares.
- Les *substitutions* correspondent aux mutations qui ont passé le crible de la sélection naturelle.



# Homologie ou similarité ?

- La phylogénie moléculaire est fondée sur l'utilisation de séquences homologues :
  - Deux séquences sont dites homologues si et seulement si elles possèdent un ancêtre commun.
  - L'existence d'un ancêtre commun est inférée à partir de la similarité.
  - Seuil variable suivant les circonstances :
    - Similarité sans homologie (convergence, répétitions).
    - Homologie avec faible similarité (limitation à quelques positions clés dans les séquences).

# Famille des insulines

Chaîne B

Chaîne A

Q14641	<b>EILRGCGPRFGKHL</b> LSYCPMPEKTFTTTPGG...[x]58.....	<b>SGRHRFDPFCC</b> EVICDDGTSVKLCT
P51460	<b>REKLCGHHFVR</b> ALVRVCGGPRWSTEA.....[x]51.....	<b>AAATNPARYCCLSGC</b> TQQDLLTLCPY
P04808	<b>VIKLCGREL</b> VRAQIAICGMSTWS.....[x]109.....	<b>PYVALFEKCC</b> LIGCTKRSILAKYC
P26732	<b>VHTY</b> CGRHLARTLADLCWEAGVD.....[x]25.....	<b>GIVDECC</b> LRPCSVDVLLSYC
P26733	<b>ARTY</b> CGRHLADTLADLCF--GVE.....[x]23.....	<b>GVVDECC</b> FRPCTLDVDLLSYCG
P26735	<b>SQFY</b> CGDFLARTMSILCWPDMP.....[x]25.....	<b>GIVDECCY</b> RPCTTDVILKLYCDKQI
P26736	<b>GH</b> YICGRYLAYKMDLCLWRAGFE.....[x]25.....	<b>GI</b> ADECCQLQPCTNDVLLSYC
P15131	<b>VARY</b> CGEKLNSALKLVCRGNYNTMF.....[x]58.....	<b>GVFDECCR</b> KSCSISELQTYCGR
P07223	<b>RRG</b> VCGSALADLVDFACSSSNQPAMV.....[x]29.....	<b>QGTTNIVCECCMKPCTLSEL</b> RQYC
P25289	<b>PRG</b> ICGSNLAGFRAFICCSNQNQSPSMV.....[x]44.....	<b>QRTTNLVCECCF</b> NYCTPDVVRKYCY
P80090	<b>PRGLCGST</b> LANMVQWLCSTYTTSSKV.....[x]30.....	<b>ESRPSIVCECCFNQCTVQELLAYC</b>
P31241	<b>PRG</b> ICGSDLADLRAFICCSRRNQPAMV.....[x]44.....	<b>QRTTNLVCECCY</b> NVCTVDVFYEYCY
P91797	<b>PRGLCGNRL</b> ARAHANLCFLLRNTPDIFPR.....[x]86.....	<b>EVMAEP</b> SLVCDCCYNECSVRKLATYC
P22334	<b>AEY</b> LCGSTLADVLSFVCGNRGYNSQP.....[x]31.....	<b>GLVEECCY</b> NVCDYSQLESYCNPYS
P01308	<b>NQHLCGSH</b> ILVEALYLVCGERGFFYTPKT.....[x]35.....	<b>GIVEQCC</b> TSICSLYQLENYCN
P01343	<b>PETL</b> CGAELVDALQFVCGDRGFYF.....[x]12.....	<b>GIVDECC</b> FRSCDLRRLEMYCAPLK
P01344	<b>SETL</b> CGGELVDTLQFVCGDRGFYF.....[x]12.....	<b>GIVEECC</b> FRSCDLALLETYCATPA

\*\*

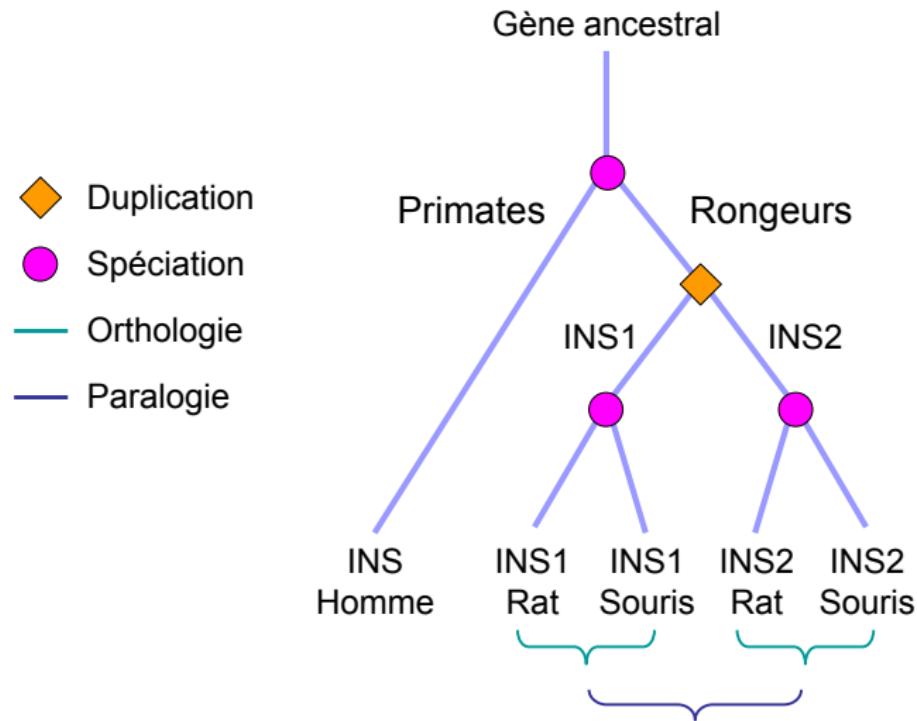
.\*

\*\*

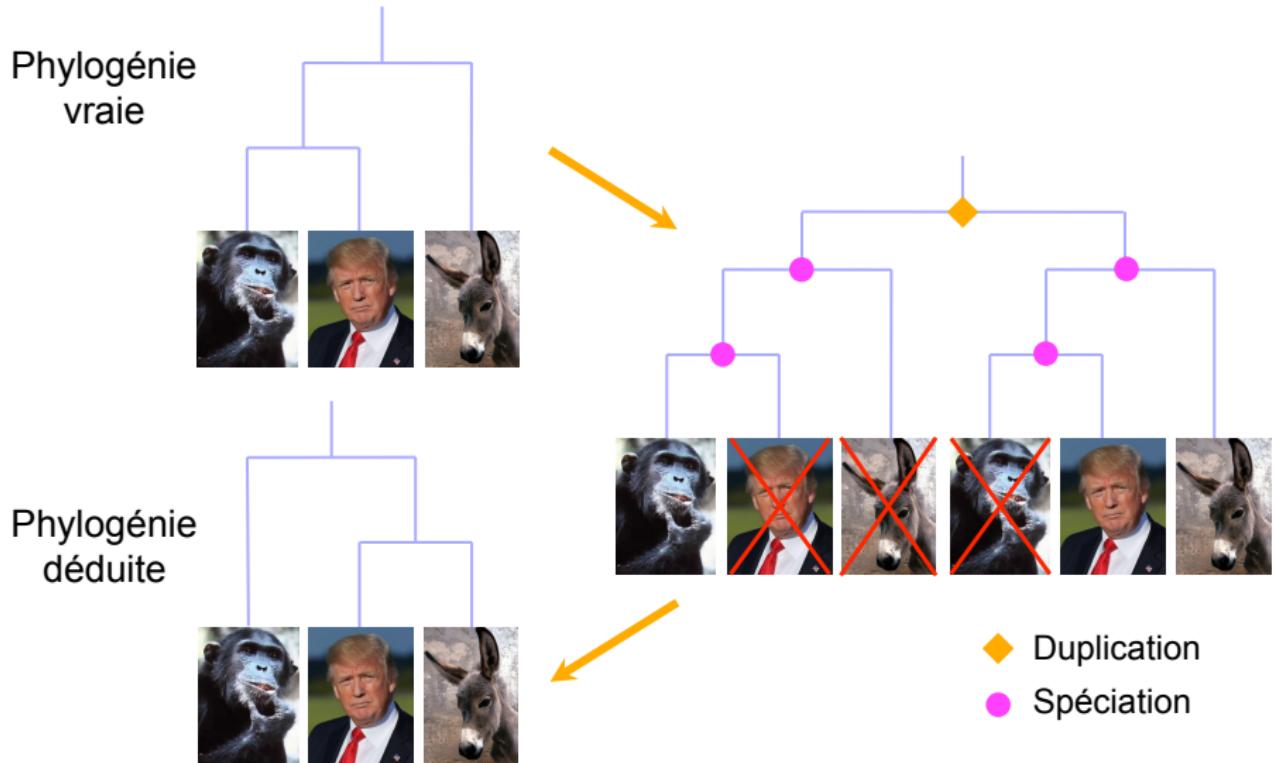
\*.

\*.

# Orthologues et paralogues



# Duplications et phylogénie

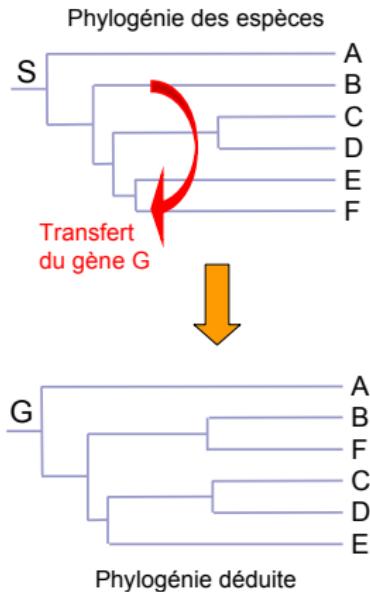


# Les paralogues sont fréquents

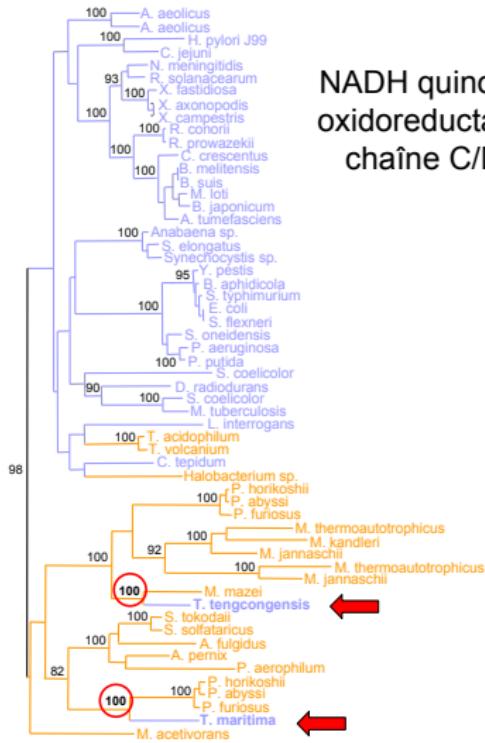
- Nombre très important même chez les organismes unicellulaires :
  - 30% des gènes d'*E. coli* K12.
  - 40% en moyenne chez les mammifères.
- Existence de duplications multiples :
  - Les relations d'orthologie sont souvent non bijectives.
- Divergences pouvant être importantes après duplication :
  - Difficulté à identifier de nombreux paralogues.

# Les transferts horizontaux

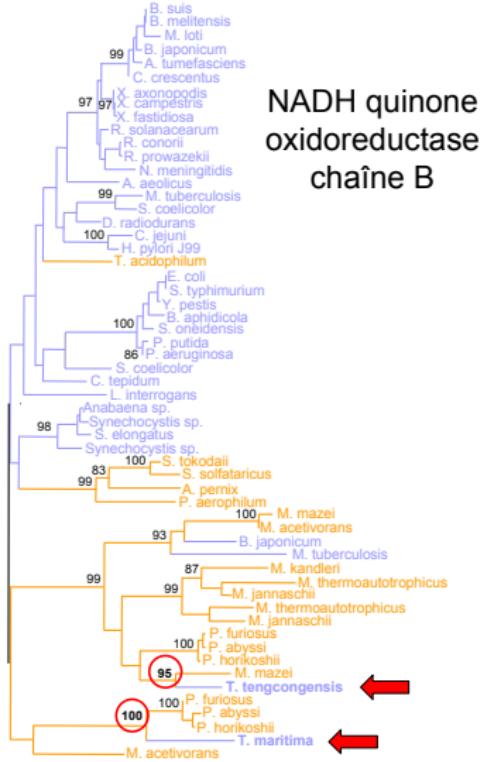
- Transmission de gènes entre taxons différents.
- Phénomènes supposés très fréquents chez les prokaryotes :
  - Implication de différents mécanismes :
    - Transformation.
    - Conjugaison.
    - Transduction.
  - 17.6% des gènes d'*E. coli* auraient été obtenus par transfert.



# Exemple de transfert

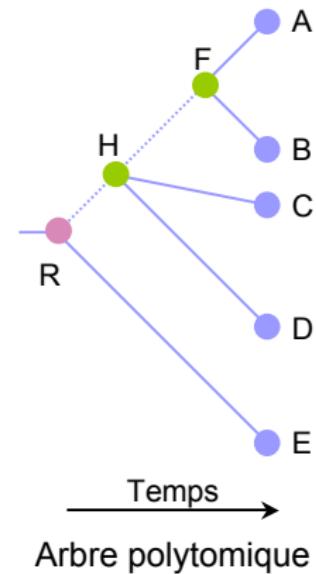
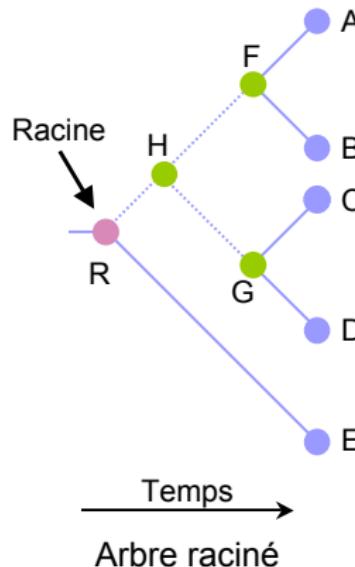
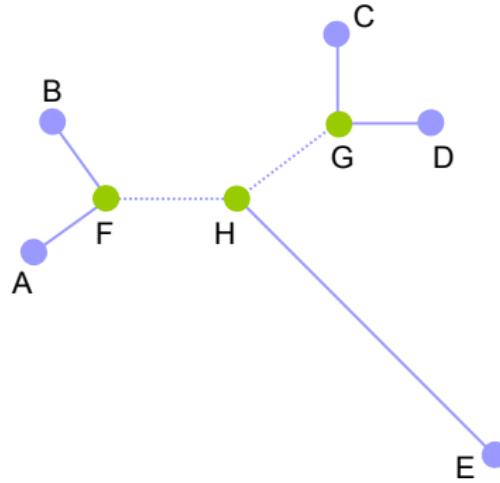


NADH quinone  
oxidoreductase  
chaîne C/D



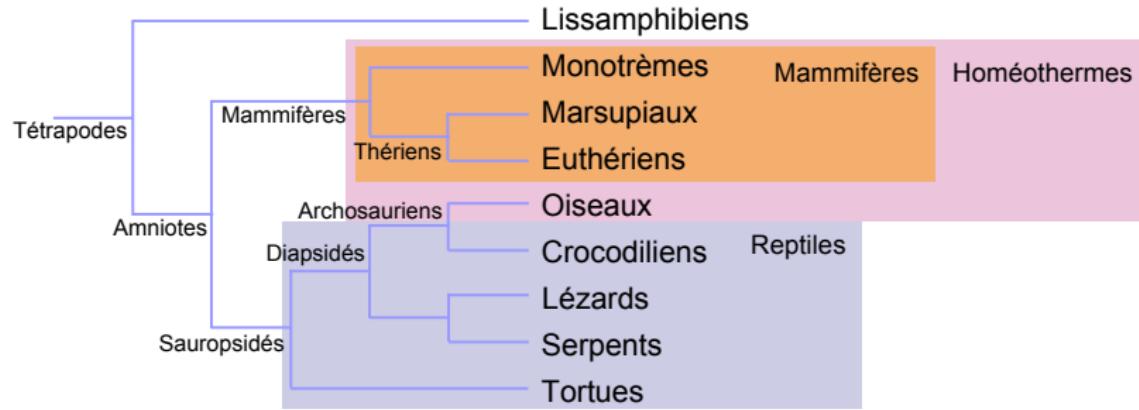
NADH quinone  
oxidoreductase  
chaîne B

# Typologie des arbres



- Unité Taxonomique Opérationnelle (UTO)
- Unité Taxonomique Hypothétique (UTH)

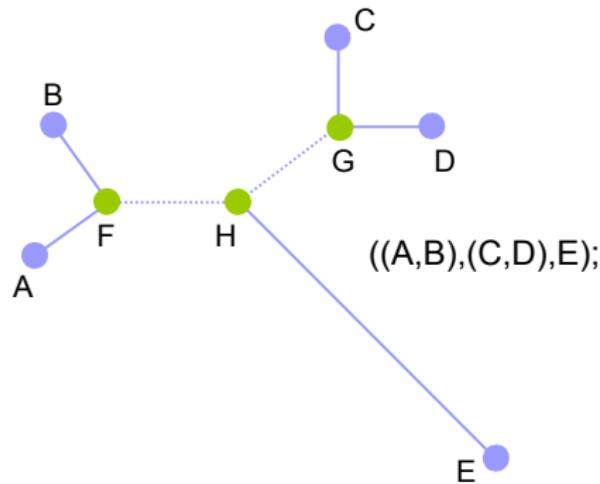
# Mono-, poly- et paraphylie



- Dans cette phylogénie des Tétrapodes :
  - Les Mammifères sont *monophylétiques*.
  - Les Homéothermes sont *polyphylétiques*.
  - Les Reptiles (au sens ancien du terme) sont *paraphylétiques*.

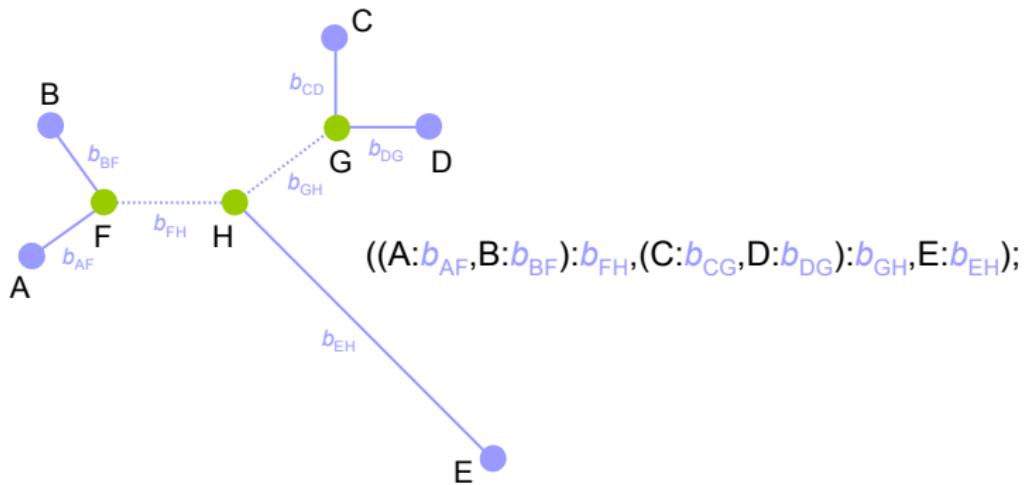
## Format Newick standard

- Les UTO (ou groupes d'UTO) descendants d'un même nœud sont placés entre parenthèses.
- Les UTO et groupes d'UTO sont séparés par des virgules.
- La fin de l'arbre est indiquée par un point virgule.



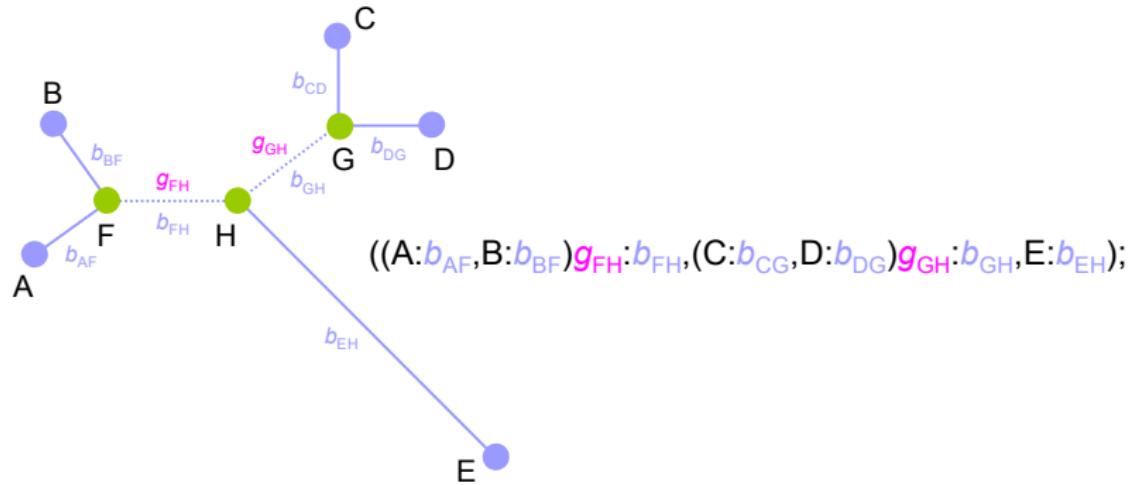
# Extensions courantes

- Longueurs des branches indiquées par leur valeur précédée de deux points.



# Extensions courantes

- Longueurs des branches indiquées par leur valeur précédée de deux points.
- Robustesse des branches internes indiquées par un nombre localisé après les parenthèses fermantes délimitant les groupes.



$((A:b_{AF}, B:b_{BF})g_{FH}:b_{FH}, (C:b_{CG}, D:b_{DG})g_{GH}:b_{GH}, E:b_{EH});$

# Nombre d'arbres racinés

- Soit  $B_r^{(n)}$  le nombre d'arbres racinés à  $n$  UTO :
- Pour construire un arbre raciné à  $n$  UTO, il suffit d'ajouter une UTO à un arbre raciné à  $n - 1$  UTO.
- Un arbre raciné à  $n - 1$  UTO possède  $n - 1$  branches terminales et  $n - 2$  branches internes, soit  $2n - 3$  branches au total.
- On en déduit la formule de récurrence :

$$\begin{aligned} B_r^{(n)} &= (2n - 3)B_r^{(n-1)} \\ &= (2n - 3) \times (2n - 5) \times \cdots \times 9 \times 7 \times 5 \times 3 \times 1 \end{aligned}$$

Il est ensuite facile de démontrer que :

$$B_r^{(n)} = \frac{(2n - 3)!}{2^{n-2}(n - 2)!}$$

# Nombre d'arbres non racinés

- Soit  $B_u^{(n)}$  le nombre d'arbres non racinés à  $n$  UTO.
- Pour construire un arbre non raciné à  $n$  UTO, il suffit d'ajouter une UTO à un arbre non raciné à  $n - 1$  UTO.
- Un arbre non raciné à  $n - 1$  UTO possède  $n - 1$  branches terminales et  $n - 4$  branches internes, soit  $2n - 5$  branches au total.
- On en déduit la formule de récurrence :

$$\begin{aligned} B_u^{(n)} &= (2n - 5)B_u^{(n-1)} \\ &= (2n - 5) \times (2n - 7) \times \cdots \times 9 \times 7 \times 5 \times 3 \times 1 \end{aligned}$$

De la même façon que précédemment, on en déduit que :

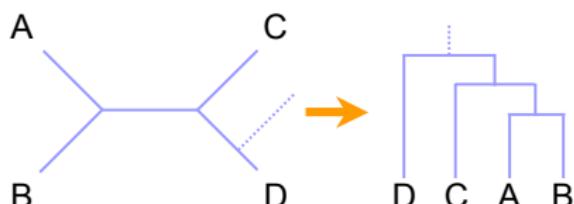
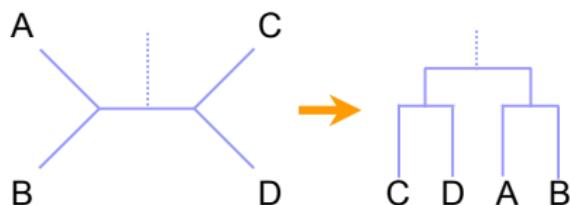
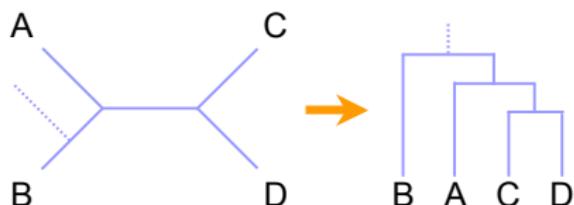
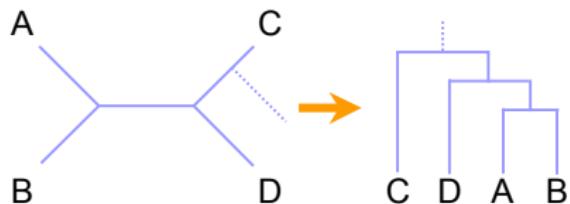
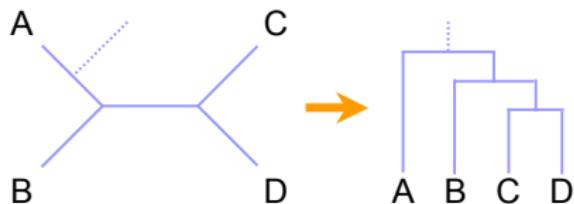
$$B_u^{(n)} = \frac{(2n - 5)!}{2^{n-3}(n - 3)!} = B_r^{(n-1)}$$

# L'arbre caché dans la forêt

- Le nombre d'arbres (racinés ou non) croît donc extrêmement rapidement :
  - Retrouver le bon arbre est pratiquement impossible dès que  $n \geq 12$ .

$n$	$B_r^{(n)}$	$B_u^{(n)}$
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025
15	$\approx 2.13 \times 10^{14}$	$\approx 7.91 \times 10^{12}$
20	$\approx 8.20 \times 10^{21}$	$\approx 2.22 \times 10^{20}$
30	$\approx 4.95 \times 10^{38}$	$\approx 8.69 \times 10^{36}$
50	$\approx 2.75 \times 10^{76}$	$\approx 2.84 \times 10^{74}$

# Position de la racine



Arbre non raciné à  $n$  UTO :

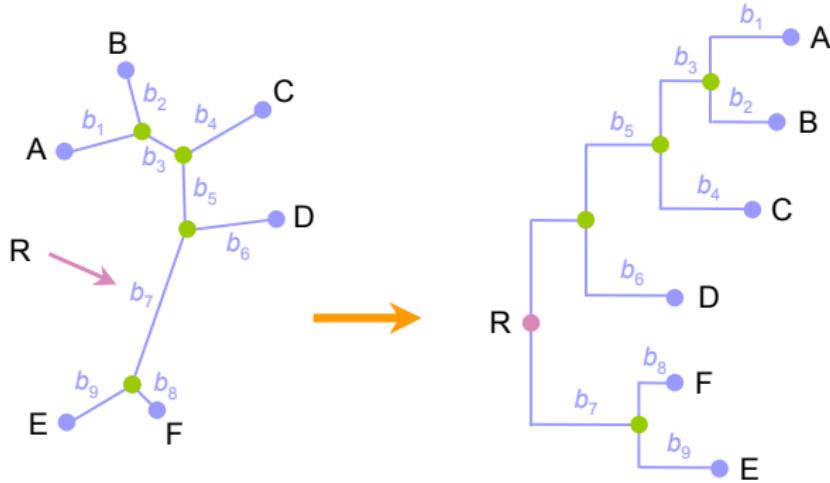
- $n$  branches externes.
- $n - 3$  branches internes.
- $2n - 3$  positions pour la racine.

# Racinement d'un arbre

- La plupart des méthodes produisent des arbres sans racine :
  - Pas d'estimation de la direction des changements au cours du temps.
- Plusieurs méthodes de racinement existent :
  - Au point moyen :
    - Hypothèse que toutes les séquences ont évolué à la même vitesse depuis leur divergence avec l'ancêtre commun.
  - À l'aide d'un groupe externe (*outgroup*) fixé *a priori* et connu comme étant extérieur aux taxons étudiés.
  - En utilisant un paralogue.

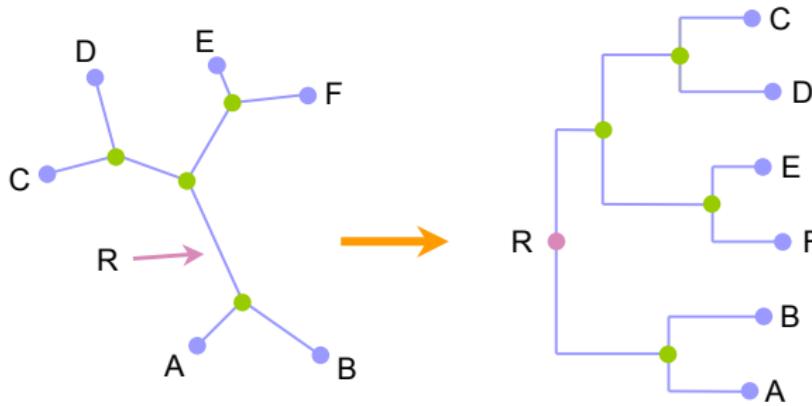
# Racinement au point moyen

- Détermination des deux UTO les plus distantes dans l'arbre :
  - Placement de la racine au milieu du chemin.
- Dans l'arbre ci-dessous, A et E sont les deux UTO les plus éloignées et le racinement au point moyen donne :



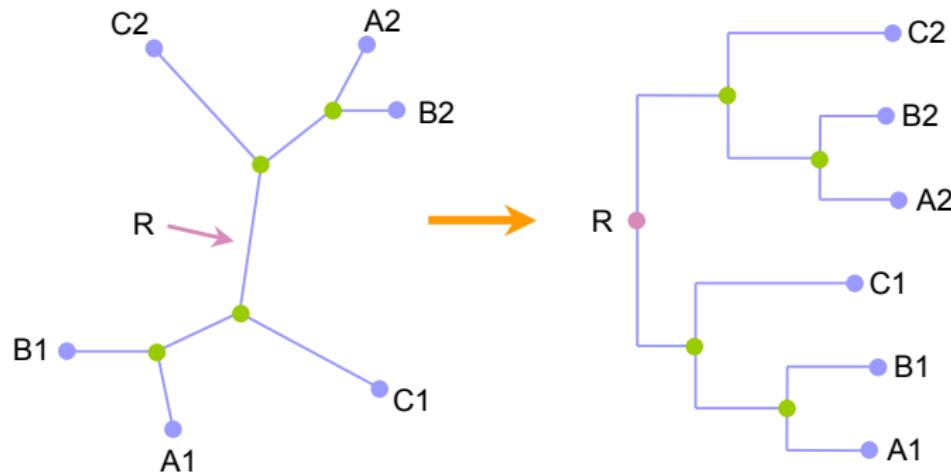
# Racinement par un groupe externe

- Choix du groupe externe :
  - Une espèce ou un groupe d'espèces monophylétique qui ne soit ni trop proche ni trop éloigné des organismes d'intérêt.
- Racinement par le groupe  $\{A, B\}$ , supposé extérieur aux organismes d'intérêt que sont C, D, E et F :



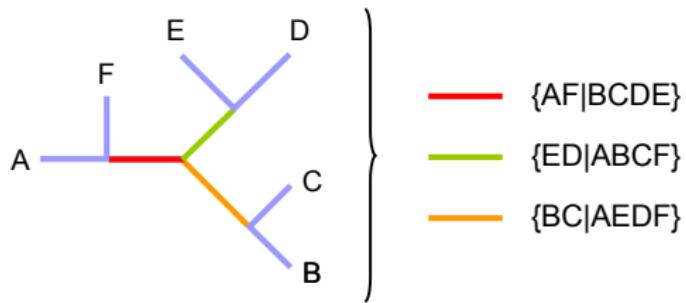
# Racinement par un parologue

- Duplication chez l'ancêtre commun à l'ensemble des organismes étudiés :
  - Racinement en utilisant une des deux copies paralogues.
  - Utilisé pour la construction de phylogénies « universelles » (*i.e.*, regroupant les trois domaines du vivant).



# Mesure de la fiabilité

- L'information véhiculée par un arbre réside entièrement dans ses branches internes :
  - La topologie  $\tau$  se déduit de l'ensemble des *bipartitions* définies par les branches internes :



- La fiabilité d'un arbre se ramène donc à la fiabilité de ses branches internes.

# Principe du *bootstrap*

1  
 ACGTACATAGTATAGCG..TCTAGTGGTACCGTATG  
 AGGTACATAGTATGG-G..TATACTGGTACCGTATG  
 ACGTAAAT-GTATAGAG..TCTAATGGTAC-GTATG  
 ACGTACATGGTATAGCG..ACTACTGGTACCGTATG

 $\ell$ 

Alignement de départ

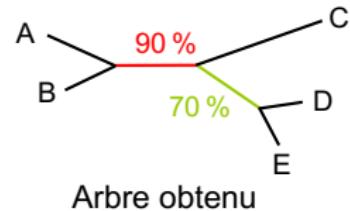
2 Échantillonnage aléatoire avec remise de  $\ell$  sites }  $B$  fois  $(B \geq 500)$

1  
 GATCAGTCATGTATAGG..TCTAGTGGTACGTATAT  
 TGAGAGTCATGTATGGT..GTATACTGGTACGTAAT  
 TGAC-GTAATGTATAGG..TCTAAATGGTACTCTAAT  
 TGACGGTCATGTATAGG..ACTACTGGTACGTATAT

 $\ell$ 

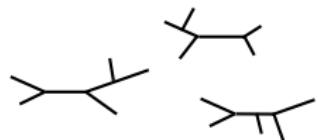
$B$  alignements rééchantillonnés

1  
 Construction de l'arbre



Construction de  $B$  arbres

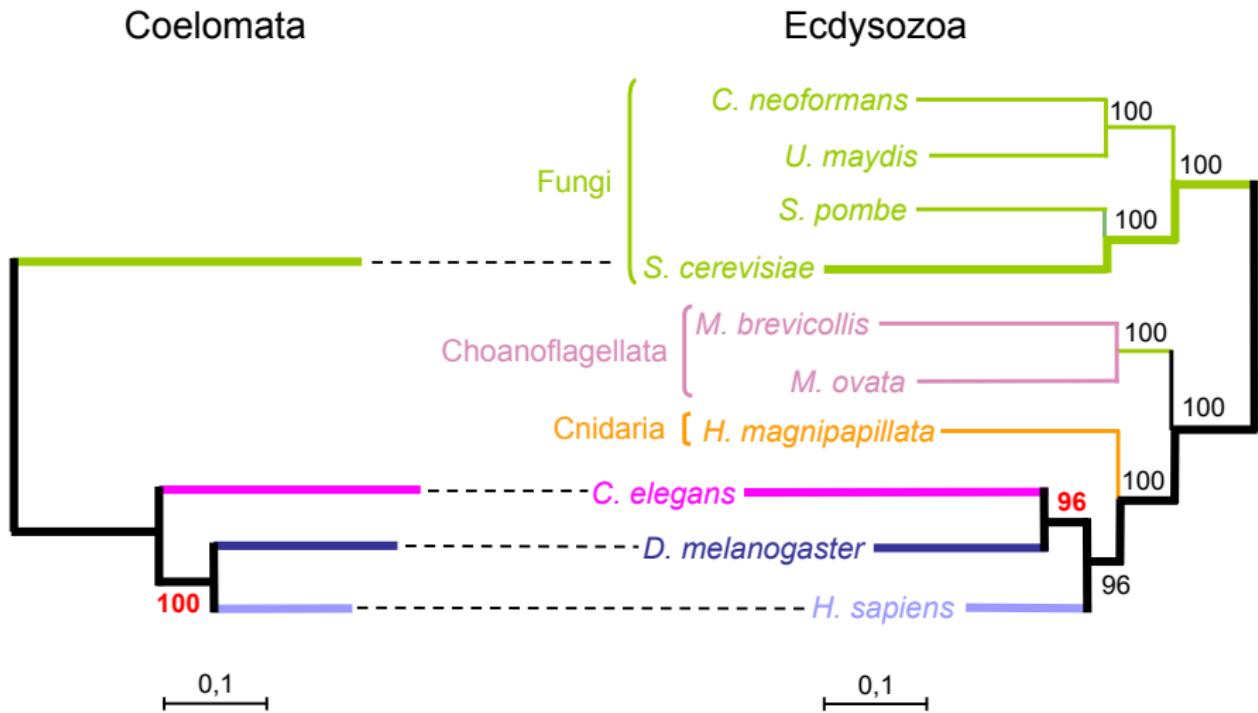
4 Pour chaque branche interne % des arbres « artificiels » contenant cette même branche



# Limitations et usage

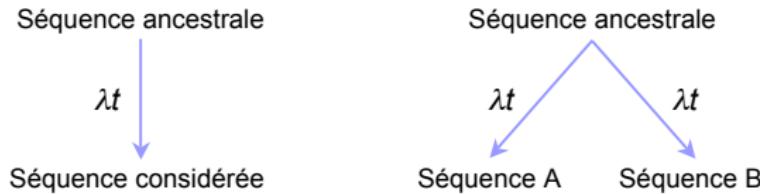
- Ne permet pas de déterminer si un arbre est vrai ou faux :
  - Un arbre faux peut avoir des branches soutenues par de fortes valeurs de *bootstrap*.
- Non-indépendance des observations (sites) :
  - Surestimation des scores faibles et sous-estimation des scores forts.
- En théorie, seuil en fonction d'un risque d'erreur fixé *a priori* :
  - En pratique, valeurs fluctuantes suivant les utilisateurs.
  - Seuils communément admis :
    - 100% : robustesse maximale.
    - 95-99% : très fort soutien par les données.
    - 90-94% : fort soutien par les données.
    - 80-89% : soutien modéré par les données.
    - < 80% : pas de soutien.

# Un exemple classique



# Distances évolutives

- Utilisées par toutes les méthodes de reconstruction sauf la parcimonie.
- Mesurent le nombre de substitutions produites sur les deux lignées depuis la divergence de l'ancêtre commun :
  - Sont rapportées à la longueur des séquences.
  - Sont exprimées en nombre de substitutions/site.



# Divergence observée

- Appelée  $p$  (ou  $p$ -distance), c'est l'estimation la plus simple de la distance entre deux séquences :

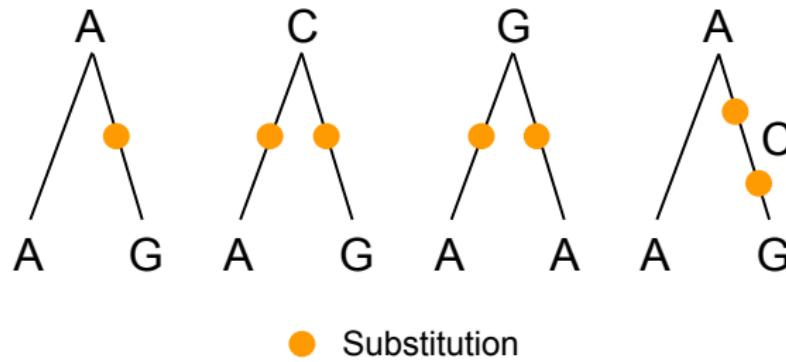
$$p = n/\ell$$

avec  $n$  le nombre total de substitutions et  $\ell$  le nombre de sites homologues comparés.

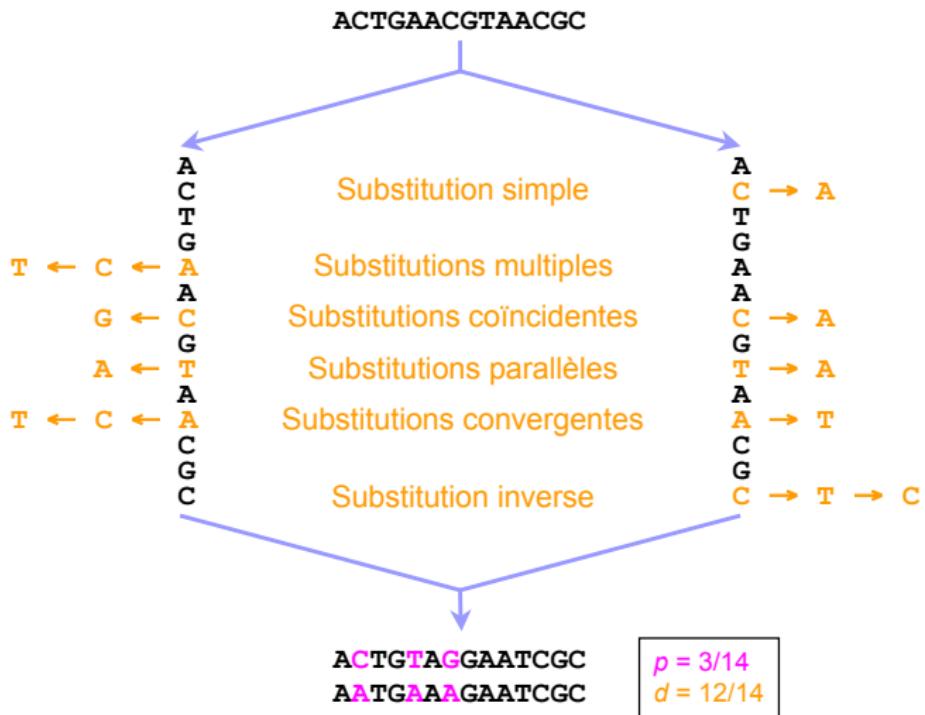
- Variation pour deux séquences de composition homogène :
  - Pour l'ADN :  $0 \leq p \leq 0.75$ .
  - Pour les protéines :  $0 \leq p \leq 0.95$ .

# Substitutions multiples

- La distance évolutive réelle ( $d$ ) est généralement supérieure à la divergence observée ( $p$ ).
- En faisant des hypothèses sur la nature du processus évolutif, il est possible d'estimer  $d$  à partir de  $p$ .

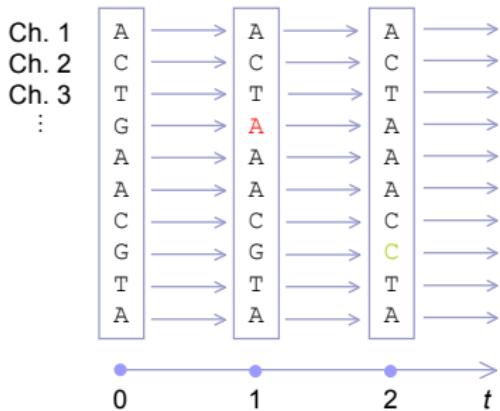


# Types de substitutions



# Modélisation markovienne de l'évolution

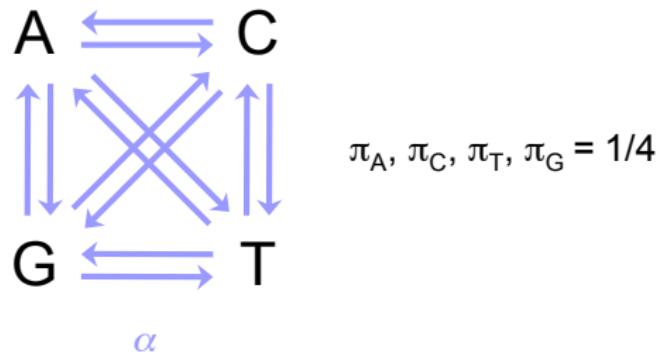
- Utilisée pour les séquences nucléotidiques et protéiques.
- Hypothèse que les phénomènes de substitution suivent un processus *markovien* :
  - Détermination des probabilités de substitution :
  - Calcul de la distance évolutive  $d$ .
- Propriétés fondamentales :
  - Indépendance des sites.
  - Tous les sites évoluent suivant le même processus.



Évolution des sites d'une séquence d'ADN selon un processus markovien

# Modèle de Jukes et Cantor

- Premier modèle d'évolution moléculaire pour les séquences nucléotidiques (1969).
- Toutes les substitutions sont équiprobables :
  - Taux de substitution instantané  $\alpha$  pour chaque nucléotide.
  - Calcul d'une seule probabilité de substitution.



# Propriétés

- Distance évolutive inférée :

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

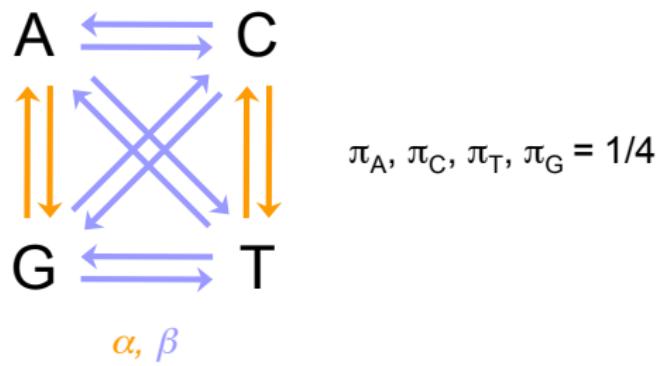
avec  $p$  la divergence observée.

- Approprié uniquement pour des séquences peu divergentes ( $p \leq 0.1$ ).
- Calcul impossible si les séquences sont trop divergentes :

$$p \rightarrow 0.75 \Rightarrow d \rightarrow \infty$$

# Modèle de Kimura à deux paramètres

- Deuxième modèle « historique », proposé par Kimura (1980).
- Distingue les transitions des transversions :
  - Définition de deux taux instantanés ( $\alpha$  et  $\beta$ ).
  - Calcul de deux probabilités de substitution.



# Propriétés

- Distance évolutive inférée :

$$d = -\frac{1}{2} \ln(1 - 2r - v) - \frac{1}{4} \ln(1 - 2v)$$

avec  $r$  et  $v$  resp. les fréquences observées des transitions et des transversions.

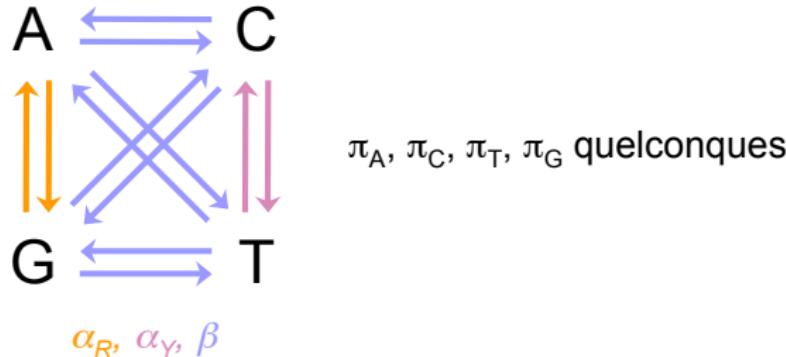
- Calcul impossible si les séquences sont trop divergentes :

$$v \rightarrow 0.5 \Rightarrow d \rightarrow \infty$$

$$r \rightarrow (1 - v)/2 \Rightarrow d \rightarrow \infty$$

# Modèle de Tamura et Nei

- Introduction des fréquences des bases du jeu de données ( $\pi_A$ ,  $\pi_C$ ,  $\pi_T$ ,  $\pi_G$ ) comme paramètres du modèle.
- Distingue les transitions entre purines de celles entre pyrimidines et les transversions :
  - Définition de trois taux instantanés ( $\alpha_R$ ,  $\alpha_Y$  et  $\beta$ ).
  - Calcul de trois probabilités de substitution.



# Distance évolutive inférée

- Formule de Tamura et Nei pour le calcul du nombre de substitutions :

$$d = \frac{2\pi_T\pi_C}{\pi_Y}(a_1 - \pi_R b) + \frac{2\pi_A\pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y\pi_R b$$

avec :

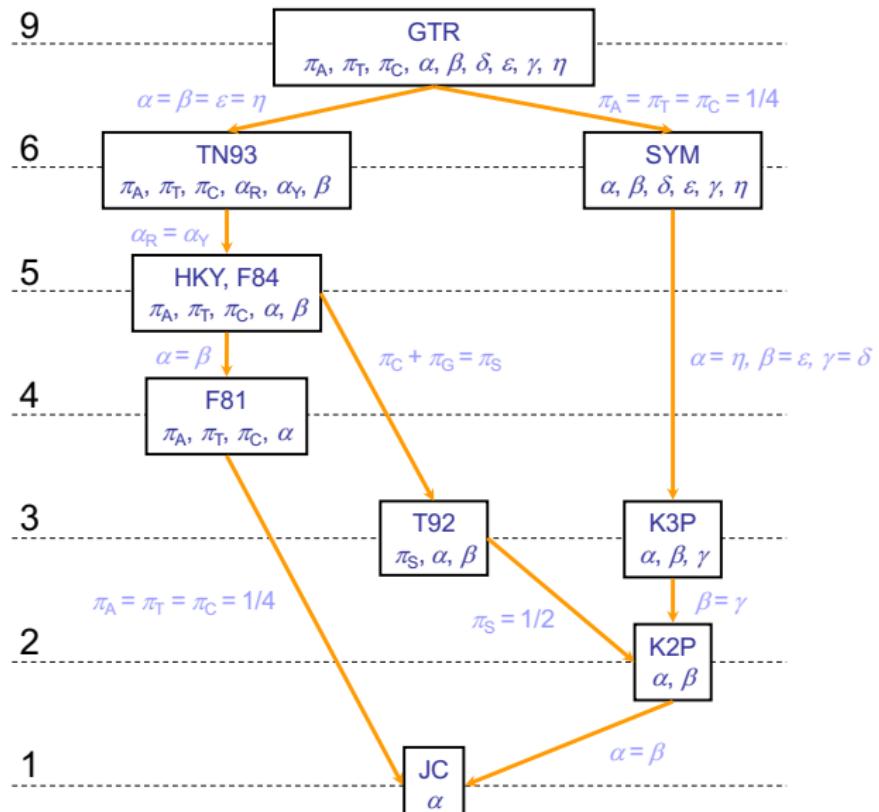
$$\begin{cases} a_1 = -\ln \left( 1 - \frac{\pi_Y}{2\pi_T\pi_C} r_Y - \frac{1}{2\pi_Y} v \right) \\ a_2 = -\ln \left( 1 - \frac{\pi_R}{2\pi_A\pi_G} r_R - \frac{1}{2\pi_R} v \right) \\ b = -\ln \left( 1 - \frac{1}{2\pi_R\pi_Y} v \right) \end{cases}$$

et  $r_R$ ,  $r_Y$  et  $v$ , les fréquences observées des deux types de transitions et des transversions entre deux séquences.

# Propriétés

- Modèle le plus complexe pour lequel il existe une solution analytique au calcul des probabilités de substitution.
- Limites d'utilisation :
  - Logarithmes indéfinis pour les valeurs de  $a_1$ ,  $a_2$  ou  $b$  si séquences trop divergentes (valeurs négatives).
  - Le modèle n'est pas utilisable pour des séquences présentant plus de 50% de transversions.

## Imbrication des modèles



# Utilité des modèles complexes

- Modélisent mieux l'évolution des séquences :
  - Plus proches de la réalité biologique.
- Séquences trop courtes :
  - Erreurs d'échantillonnage (valeurs de  $d < 0$ ).
  - Variance importante.
- Séquences trop divergentes :
  - Méthodes à plus de quatre paramètres fréquemment inapplicables.
- Séquences peu divergentes :
  - Toutes les méthodes donnent des résultats comparables.

# Modèles protéiques simples

- Adaptation du modèle de Jukes et Cantor aux acides aminés :

$$d = -\frac{19}{20} \ln \left( 1 - \frac{20}{19} p \right)$$

- Modèle de Poisson (hypothèse de l'absence de réversions) :

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

- Approximation de Kimura pour le modèle PAM :

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

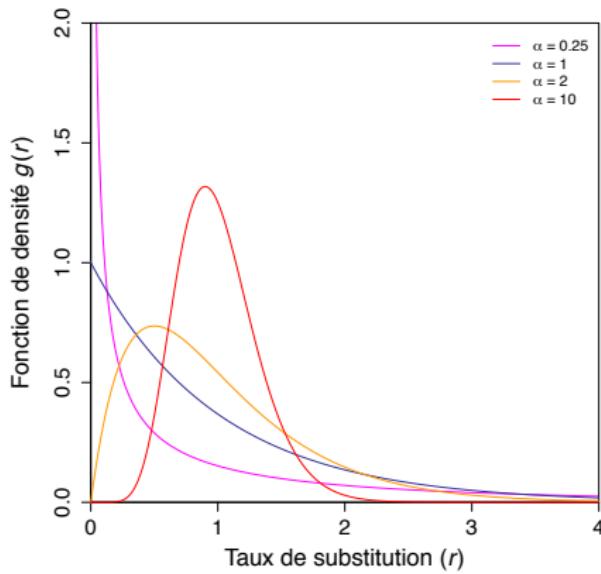
- Hypothèses sous-jacentes très simplificatrices !

# Modèles protéiques empiriques

- Pas d'expression analytique pour le calcul des probabilités de substitution.
- Estimation des valeurs à partir des fréquences observées sur des grands ensembles de séquences alignées :
  - Inférence par maximum de parcimonie :
    - PAM (*Point Accepted Mutation*, Dayhoff *et al.*, 1978).
    - JTT (Jones, Taylor et Thornton, 1992).
  - Inférence par maximum de vraisemblance :
    - WAG (Whelan et Goldman, 2001).
    - LG (Le et Gascuel, 2008).
- Calcul de la distance évolutive par approximation numérique.

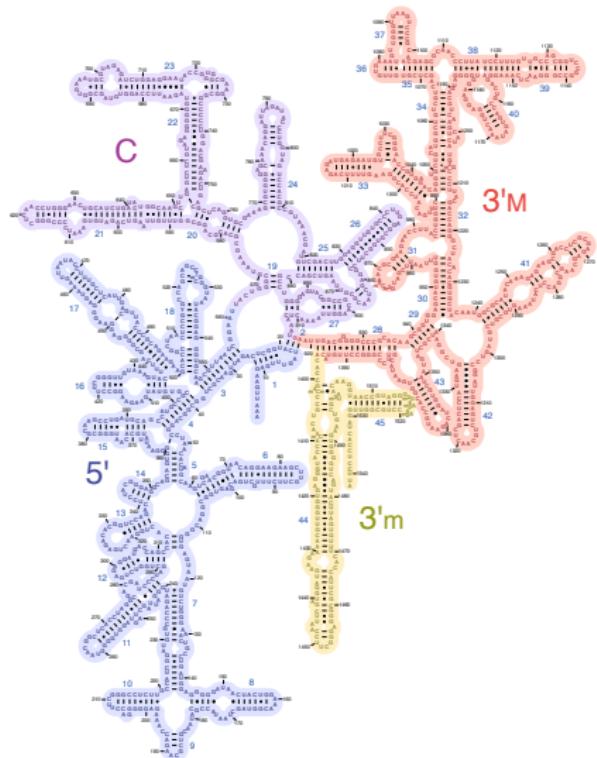
# Correction par la loi Gamma

- Hypothèse des modèles standards :
  - Tous les sites évoluent à la même vitesse :
    - Existence de nombreux contre-exemples (*e.g.*, ARNr).
- Introduction d'un facteur correctif  $r$  :
  - Modélisation par une distribution  $\Gamma$  :
    - Détermination du paramètre de forme ( $\alpha$ ) de cette loi.



# Contraintes structurales

- La structure secondaire d'un ARNr est indispensable à sa fonction :
  - Conservation au cours de l'évolution.
  - Conséquences au niveau des séquences :
    - Vitesse peu élevée au niveau des régions appariées.
    - Vitesse plus importante au niveau des boucles.

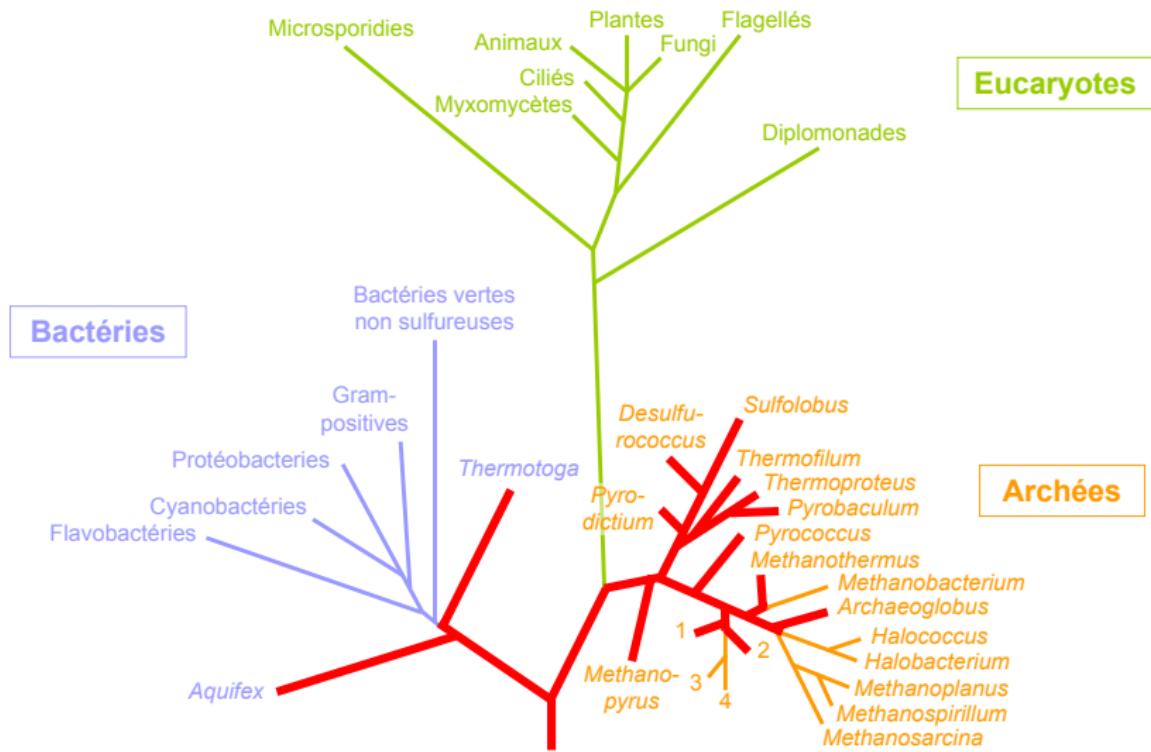


# Choix d'un marqueur

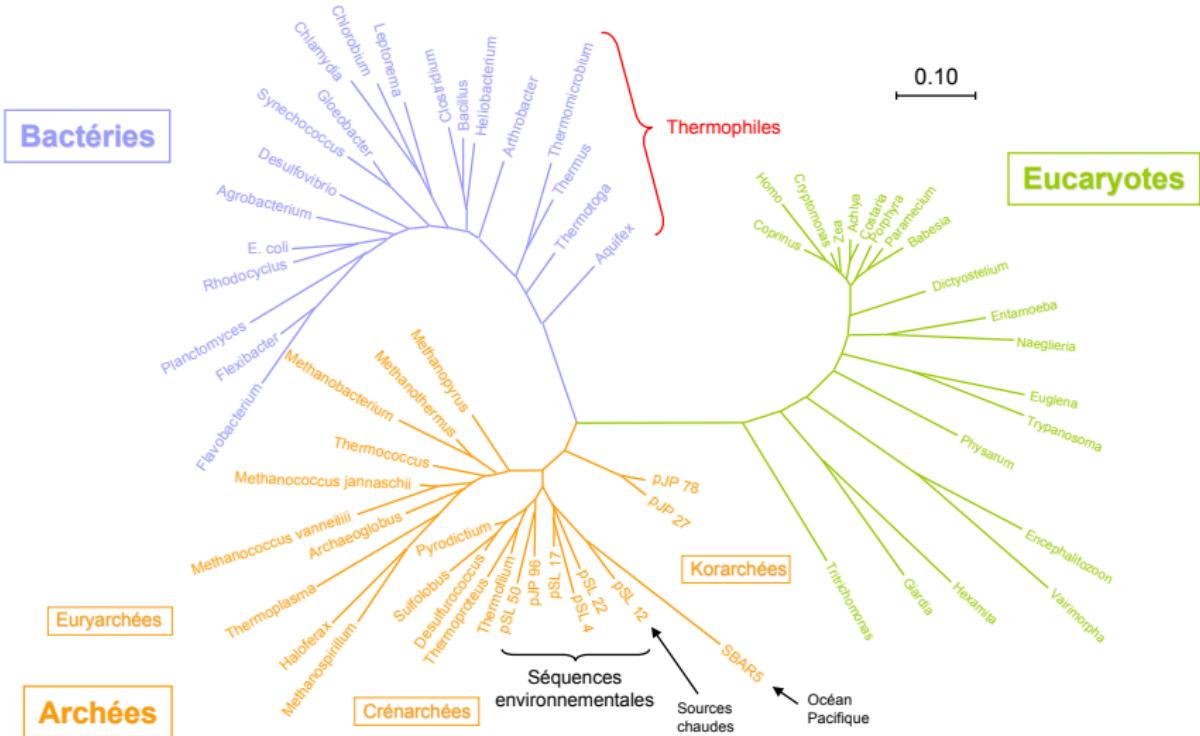
- Gènes évoluant très lentement (temps de divergence entre  $10^7$  et  $10^9$  ans).
- Présents dans l'ensemble des organismes étudiés.
- Pas de transferts !
- Seul un très petit nombre de gènes répondent aux critères précédents :
  - ARNr de la petite (16S/18S) sous-unité du ribosome.
  - Protéines *heat-shock* (e.g., Hsp70).
  - Facteurs d'elongation de la traduction.
  - Protéines ribosomiques.

# Séquences disponibles

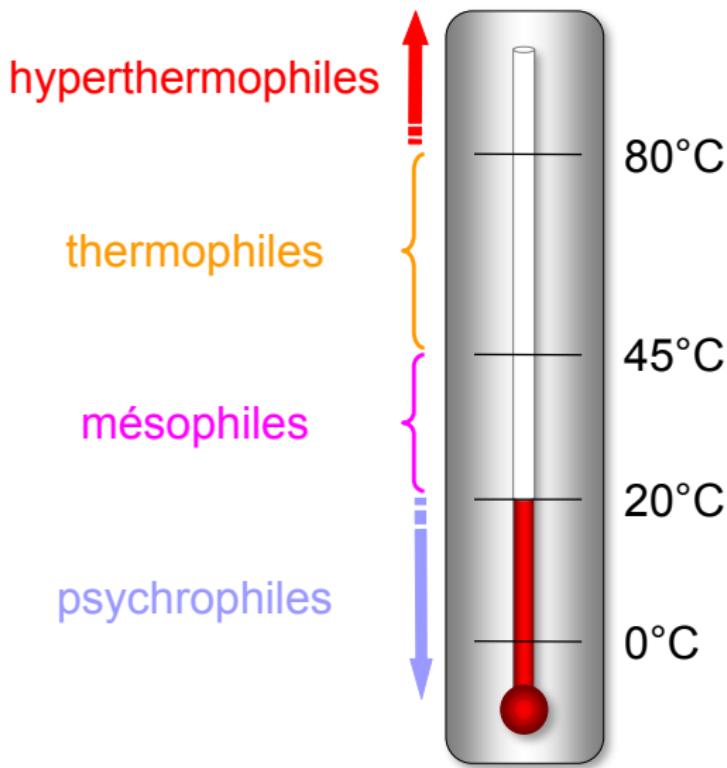
- 8 738 430 séquences d'ARNr 16S/18S dans GenBank (16 octobre 2018).
- Banques de données contenant des séquences directement utilisables pour la reconstruction phylogénétique :
  - *Ribosomal Database Project* :  
<http://rdp.cme.msu.edu/>
  - *SILVA rRNA Database Project* :  
<http://www.arb-silva.de/>
  - *Greengenes 16S rRNA gene database* :  
<http://greengenes.lbl.gov/>

Arbre de Stetter *et al.* (1996)

Arbre de Barns *et al.* (1996)



## Température optimale de croissance



# Principaux taxons bactériens I

Division	Subdivision	Genres représentatifs
Actinobactéries		<i>Actinomyces, Streptomyces</i>
Armatimonadètes	Armatimonadales	<i>Armatimonas</i>
	Chthonomonadètes	<i>Chthonomonas</i>
	Fimbriimonales	<i>Fimbriimonas</i>
Aquificales		<i>Aquifex, Hydrogenobacter</i>
Bactéries vertes non sulfureuses	Chloroflexales	<i>Chloroflexus, Herpetosiphon</i>
	Déhalococcoidètes	<i>Dehalogenimonas</i>
	Thermomicrobiales	<i>Thermomicrobium</i>
Bactéroidètes/Chlorobi	Bactéroidètes	<i>Aquimarine, Flavobacterium</i>
	Chlorobiales	<i>Chlorobium, Chloroherpeton</i>
	Ignavibactériales	<i>Ignavibacterium, Melioribacter</i>
Chlamydiales/Verrucomicrobiales	Chlamydiales	<i>Chlamydia, Chlamydophila</i>
	Lentisphaérales	<i>Lentisphaera</i>
	Verrucomicrobiales	<i>Roseibacillus, Verrucomicrobium</i>
Cyanobactéries	Chroococcales	<i>Cyanobacterium, Synechocystis</i>
	Nostocales	<i>Brasilonema, Nostoc</i>
	Oscillatoriiales	<i>Microcoleus, Oscillatoria</i>
	Pleurocapsales	<i>Dermocarpa, Xenococcus</i>
	Stigonématales	<i>Fischerella, Stigonema</i>
Déinococcus/Thermus	Déinococcales	<i>Deinococcus</i>
	Thermophiles	<i>Thermus</i>

# Principaux taxons bactériens II

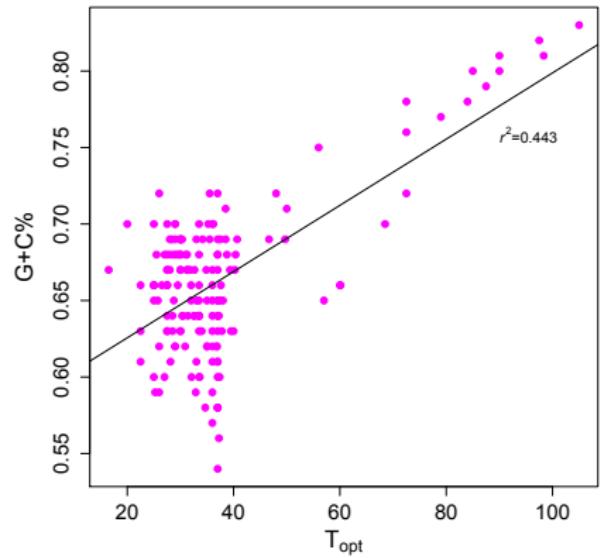
Fibrobactères/Acidobactéries	Acidobactériales Fibrobactériales	<i>Acidobacterium, Bryocella</i> <i>Fibrobacter</i>
Firmicutes	Bacilli Clostridiales Sélénomonadales Thermolithobactériales	<i>Bacillus, Staphylococcus</i> <i>Clostridium, Thermoanaerobacter</i> <i>Anaerovibrio, Dialister, Megasphaera</i> <i>Thermolithobacter</i>
Fusobactéries	Fusobactériales	<i>Fusobacterium, Leptotrichia</i>
Planctomycètes	Planctomycétales Phycisphaérales	<i>Isosphaera, Pirellula</i> <i>Phycisphaera</i>
Protéobactéries	$\alpha$ -Protéobactéries $\beta$ -Protéobactéries $\delta$ -Protéobactéries $\epsilon$ -Protéobactéries $\gamma$ -Protéobactéries	<i>Agrobacterium, Rickettsia</i> <i>Neisseria, Ralstonia</i> <i>Myxobacterium</i> <i>Helicobacter, Campylobacter</i> <i>Escherichia, Buchnera, Pseudomonas</i>
Spirochètes	Spirochètales Leptospirales	<i>Treponema, Borrelia</i> <i>Leptonema, Leptospira</i>
Ténérictes	Acholéplasmatales Anaeroplasmatales Entomoplasmatales Mycoplasmatales	<i>Acholeplasma</i> <i>Anaeroplasma</i> <i>Mesoplasma, Spiroplasma</i> <i>Mycoplasma, Ureaplasma</i>
Thermodésulfobactéries	Thermodésulfobactériales	<i>Thermodesulfobacterium</i>
Thermotogales		<i>Thermotoga, Geotoga, Thermopallium</i>

# Principaux taxons archéens

Division	Subdivision	Genres représentatifs
Crénarchées	Thermoprotéales	<i>Thermoproteus, Pyrobaculum, Thermofilum</i>
	Sulfolobales	<i>Sulfolobus, Acidianus</i>
	Désulfurococcales	<i>Aeropyrum, Desulfurococcus</i>
	Cénarchéales	<i>Cenarchaeum</i>
	Caldisphériales	<i>Caldisphaera</i>
Euryarchées	Méthanobactériales	<i>Methanobacterium, Methanothermobacter</i>
	Méthanococcales	<i>Methanococcus, Methanothermococcus</i>
	Halobactériales	<i>Halobacterium, Halococcus</i>
	Thermoplasmatales	<i>Thermoplasma, Ferroplasma</i>
	Thermococcales	<i>Pyrococcus, Thermococcus</i>
	Archéoglobales	<i>Archaeoglobus</i>
	Méthanopyrales	<i>Methanopyrus</i>
	Méthanomicrobales	<i>Methanogenium</i>
	Méthanosarcinales	<i>Methanosarcina, Methanococcoides</i>
Thaumarchées		<i>Cenarchaeum, Nitrosoarchaeum, Nitrososphaera</i>
Nanoarchées [?]		<i>Nanoarchaeum</i>
Korarchées [?]		<i>Korarchaeum</i>
Lokiarchées [?]		<i>Lokiarchaeum</i>

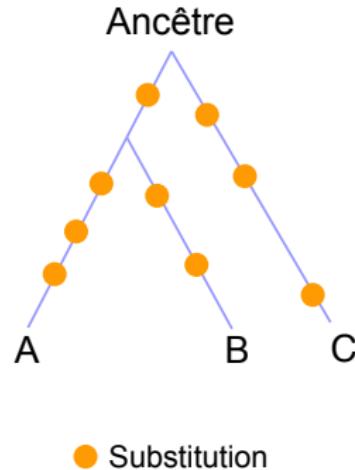
# Le biais hyperthermophile

- La structure secondaire des ARNr est indispensable à leur fonction.
- Les régions appariées sont d'autant plus stables qu'elles sont riches en G+C :
  - Trois liaisons H au lieu de deux.
  - Enrichissement en G+C chez tous les hyperthermophiles :
    - Regroupement des hyperthermophiles entre eux.

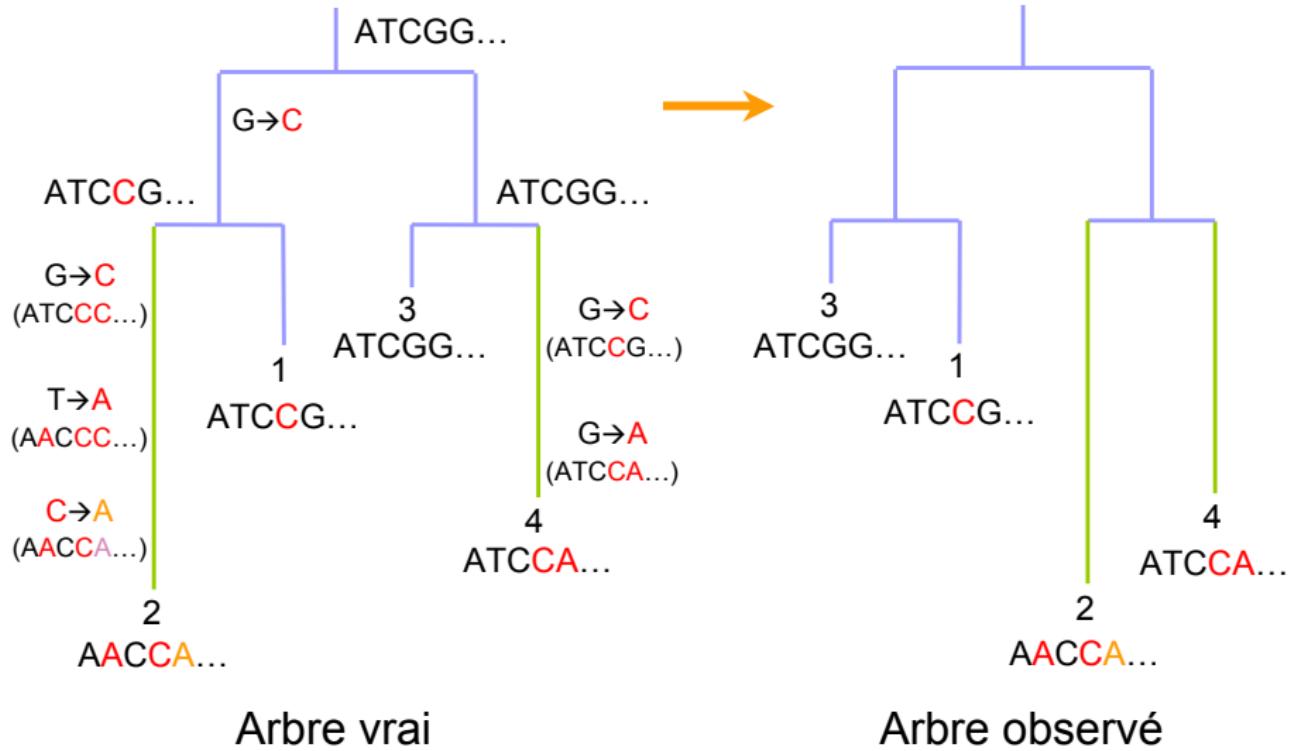


# Phénomène de saturation

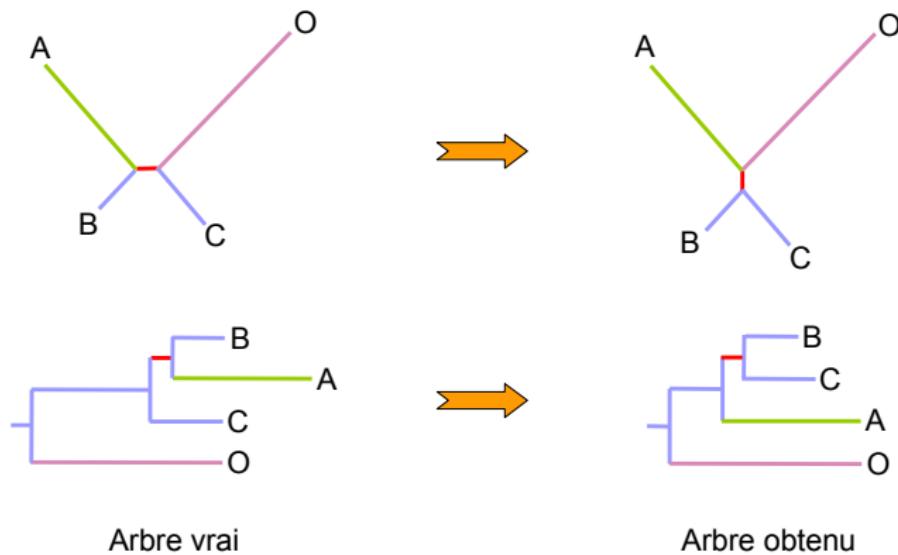
- Trop de substitutions depuis la divergence avec l'ancêtre commun :
  - Perte du signal phylogénétique.
  - Impossibilité de reconstruire l'arbre vrai quels que soient :
    - Le modèle.
    - La méthode.



# Attraction des longues branches



# Effet de l'*outgroup*



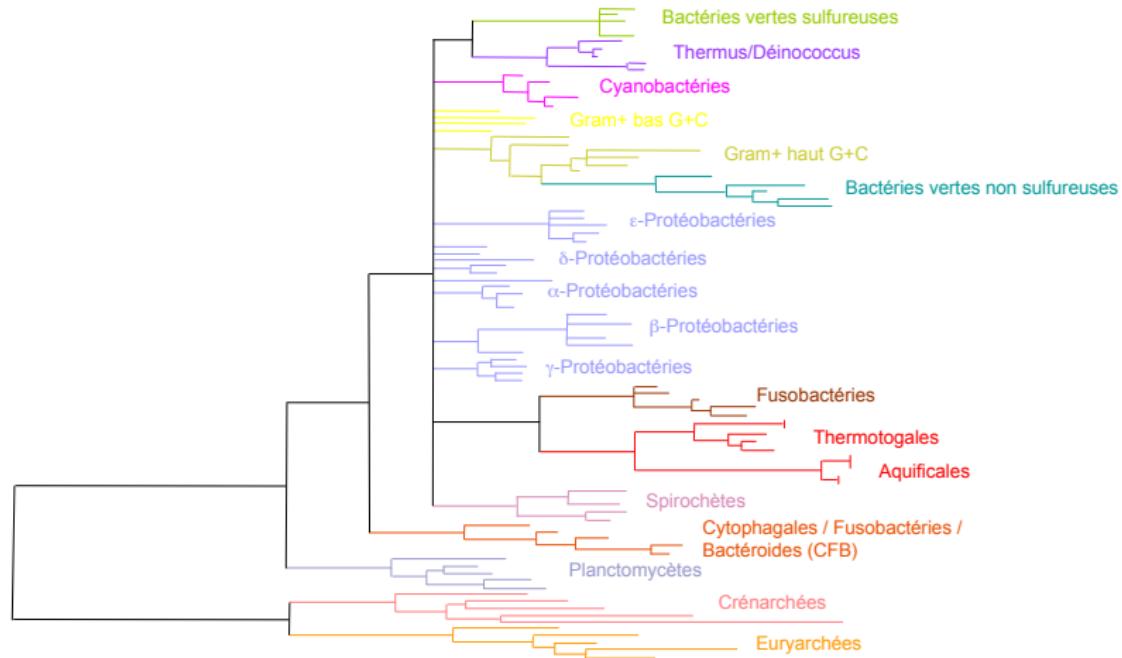
Attraction de la longue branche de la séquence A par la longue branche de l'*outgroup* (O)

# Sélection des sites

- Consiste à n'utiliser que les sites évoluant lentement (*i.e.*, non saturés) pour construire une phylogénie :
  - Sélection manuelle avec un éditeur (*e.g.*, SeaView).
  - Utilisation de programmes de filtrage des alignements (*e.g.*, Gblocks, trimAl) :
    - Paramètres à utiliser ?

Escherichia	CAGTAGCGGCGAGCGAACGGGGAGCAGGCCAGAGCC	TGAATCAGTGTGTGTGTTAGT
Pseudomonas	TAGTAGTGGCGAGCGAACGGGGATTAGCCCTTAAGCTTCATTGATTTTAGCG	-----
Rhodobacter	TAGTAGTGGCGAGCGAACGCCAGGGAGCCGT	GAGAACCGAGTG-----
Bacillus	CAGTAGCGGCGA-CGAACACGGGATCAGCCCAAACCAAGAGGCTTGCCCTGTGGTT	
Micrococcus	TAGTAGTGGCGAGCGAACGCCAGGGATGGGCT-AAACCGTATGTGTGATAACCCGGCA	
Streptomyces	CAGTAGTGGCGAGCGAACCCGGATGAGGCC-AAACCGTATACGTGTGAGACCCGGCA	
Pirellula	CAGTAGCGGCGAGCGAACAGCGAAATAGCCC-AAACCGTGGGATTTCCCTCACGGGG	
Anacystis	CAGTAGCGGCGAGCGAACGGGGAGCAGCCT-AAACCGAAACTCCACGGAGTTGGGGT	
Ruminobacter	AAGTAGTGGCGAACGAACGGGAAGCAGCCC-AAAAGTTGTATAAGTCATAGTT	-----
Leptospira	CAGTAGCGGCGAGCGAACGCCAGGGAGCCT-AAACCGCTGCCCTACGTTACAGATCTA	
Thermus	CAGTAGCGGCGAGCGAACAGGGAGCCT-AAACCGTCCGGCTTGTCGGGGCGGGG	

## Arbre de Brochier et Philippe (2002)



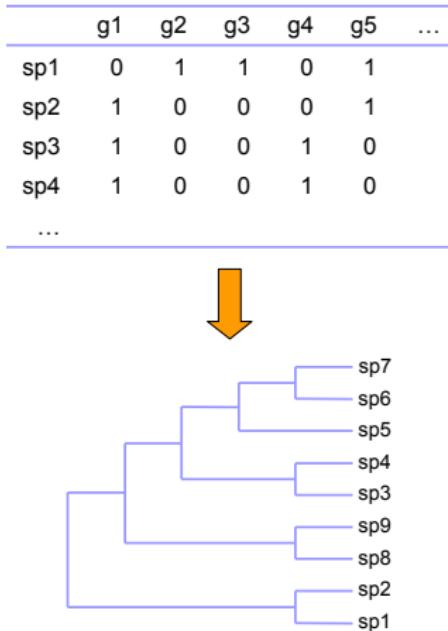
Utilisation des sites de l'ARNr évoluant lentement

# La phylogénomique

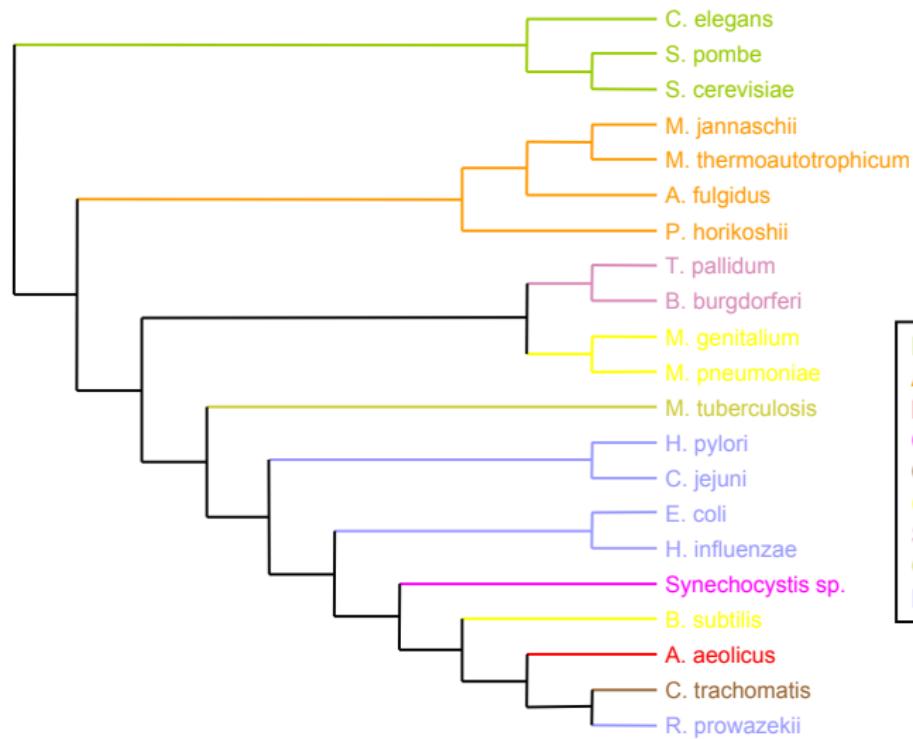
- Difficultés de résoudre les phylogénies au moyen d'un seul gène si les espèces sont très divergentes.
- Utilisation de l'information portée par un grand nombre de gènes, provenant de génomes complets :
  - Codage en présence/absence.
  - Concaténation d'orthologues présents dans de nombreuses espèces.
  - Construction de super-arbres intégrant l'information de nombreuses familles.

# Codage en présence/absence

- Approche rarement utilisée.
- Détermination de tous les orthologues présents ou non dans un ensemble d'espèces.
- Construction d'une matrice binaire.
- Calcul de l'arbre par maximum de parcimonie.



## Arbre de Tekaia (1999)



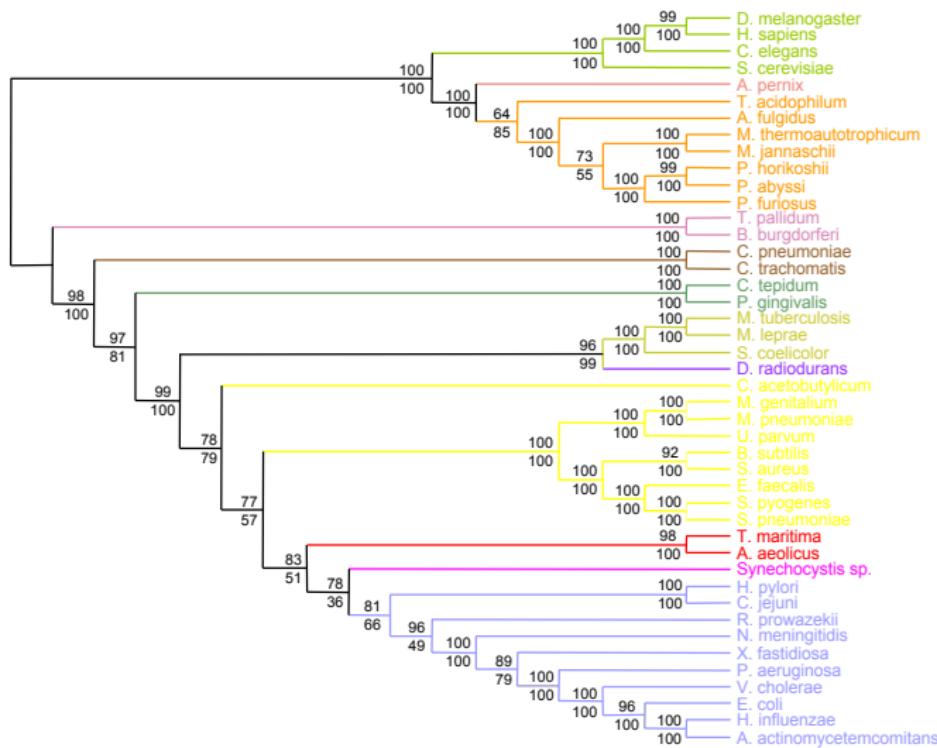
**Eucaryotes**  
**Archées**  
**Hyperthermophiles**  
**Cyanobactéries**  
**Chlamydiales**  
**Gram + bas G+C**  
**Spirochètes**  
**Gram + haut G+C**  
**Protéobactéries**

# Problèmes et limitations

- Pas de prise en compte de l'information phylogénétique contenue dans les gènes.
- Faible nombre de gènes en commun entre les différents organismes considérés :
  - Possibilités limitées de construire un arbre comprenant beaucoup de taxons.
  - Polyphyylie de certains grands groupes taxonomiques (*e.g.*, Gram+ bas G+C) jamais retrouvée dans d'autres études.
  - Topologie très sensible à l'ajout ou au retrait de certains taxons.

# Concaténation d'orthologues

- Méthode désormais la plus couramment utilisée en phylogénétique.
- Première utilisation pour la phylogénie des procaryotes par Brown *et al.* (2001) :
  - Utilisation de 23 gènes retrouvés dans 45 espèces (Bactéries, Archées et Eucaryotes) :
    - Incertitude quant à l'absence de transferts pour neuf d'entre eux.
  - Pas de prise en compte des différences de vitesses d'évolution entre/à l'intérieur des gènes :
    - Probables biais d'attraction des longues branches.

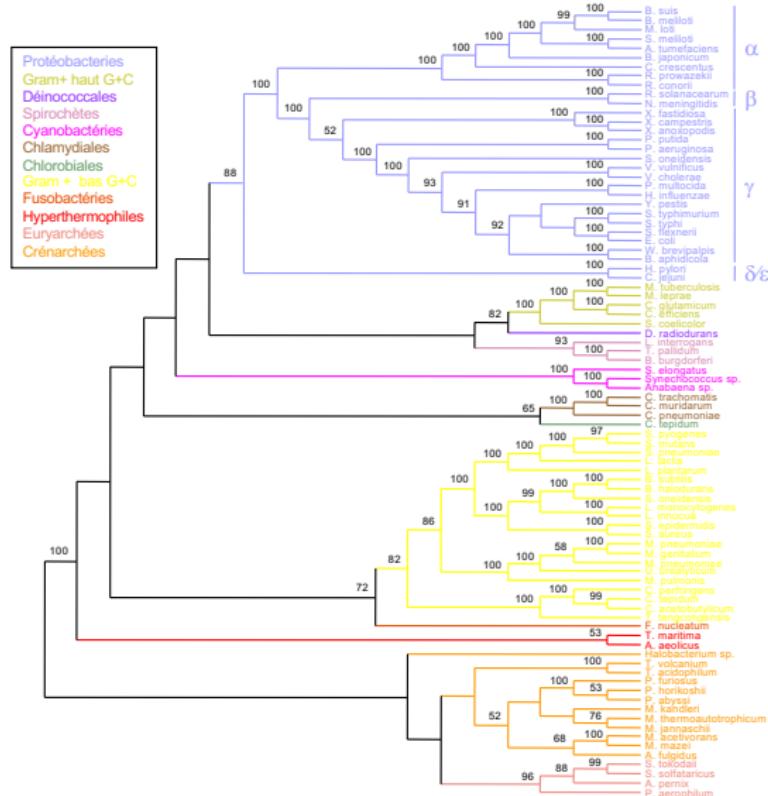
Arbre de Brown *et al.* (2001)

**Eucaryotes**  
**Crénarchées**  
**Euryarchées**  
**Spirochétées**  
**Chlamydiales**  
**Bactéries vertes**  
**sulfureuses**  
**Gram+ haut G+C**  
**Déinococcales**  
**Gram+ bas G+C**  
**Hyperthermophiles**  
**Cyanobactéries**  
**Protéobactéries**

# Approche par super-arbres

- Incorporation de l'information et du support statistique de centaines d'arbres de gènes :
  - Sélection d'un ensemble de gènes orthologues.
  - Construction des arbres individuels.
  - Sélection d'un sous-ensemble d'arbres congruents entre eux :
    - Élimination des paralogies et des transferts.
  - Intégration de l'information apportée :
    - Utilisation de l'information portée par les branches internes de l'intégralité des arbres.
    - Nombreuses méthodes disponibles (*e.g.*, MRP, PRM, SDM).

Arbre de Calteau *et al.* (2004)



## Que retenir ?

- Monophylie de chacun des trois grands domaines du vivant.
- Monophylie des grands groupes bactériens et archéens initialement définis par Woese (1977).
- Monophylie des différentes subdivisions des Protéobactéries.
- Polyphylie des bactéries Gram+.
- Positionnement variable de la plupart des grandes divisions taxonomiques les unes par rapport aux autres.