

Bio-informatique appliquée

Construction des arbres phylogénétiques

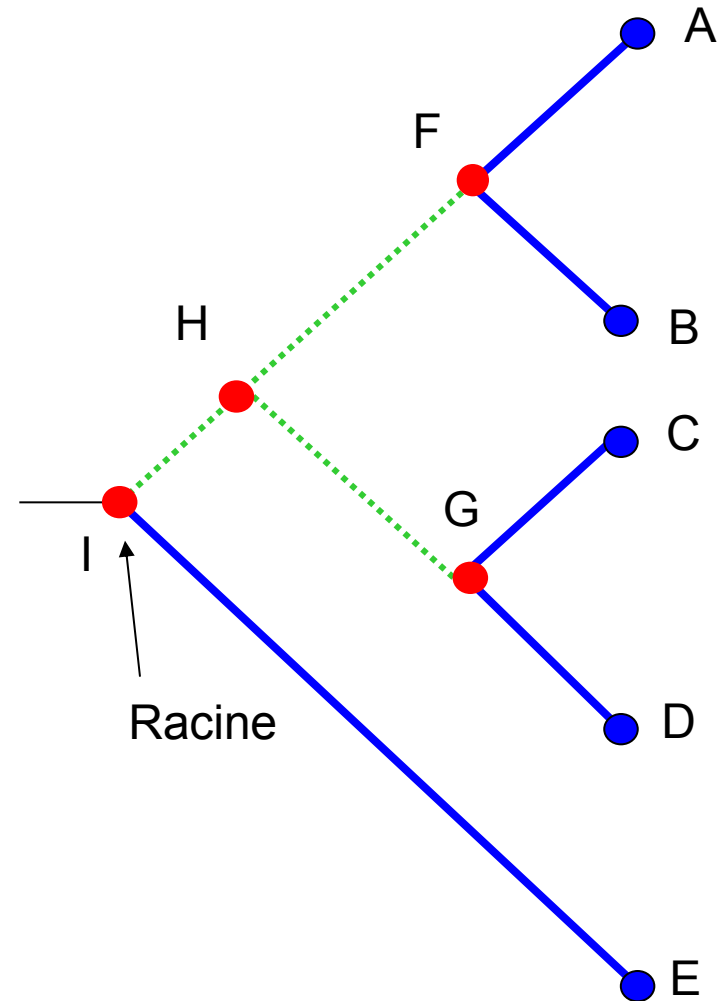
Emese Meglécz

Emese.Meglécz@imbe.fr

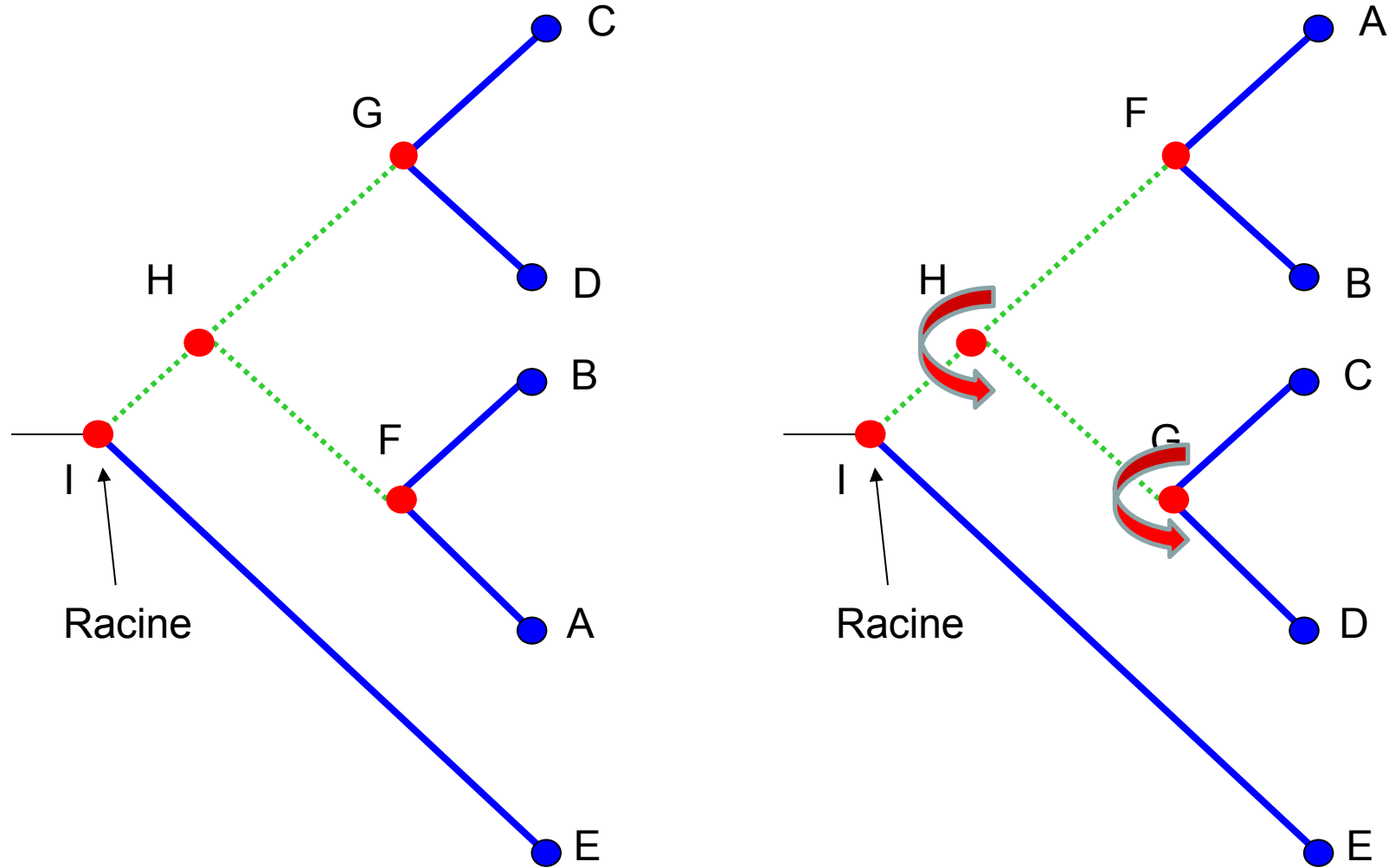
<http://www.imbe.fr/~emeglec/teach.html>

Structure des arbres

- Les relations évolutives entre les objets étudiés sont représentés par des arbres phylogénétiques
- Les arbres sont des graphes composés de
 - *noeuds* et de *branches*
 - noeuds = unités taxonomiques
 - Feuilles ou OTU = Unités Taxonomique Opérationnelles ou (A, B, C, D, E)
 - Noeuds internes ou HTU = Unités taxonomique Hypothétiques (F, G, H, I)
 - branches = relations de parentés (ancêtre /descendants) entre les unités taxonomiques
 - Branches internes
 - Branches externes
- l'ensemble des branchements de l'arbre =topologie de l'arbre

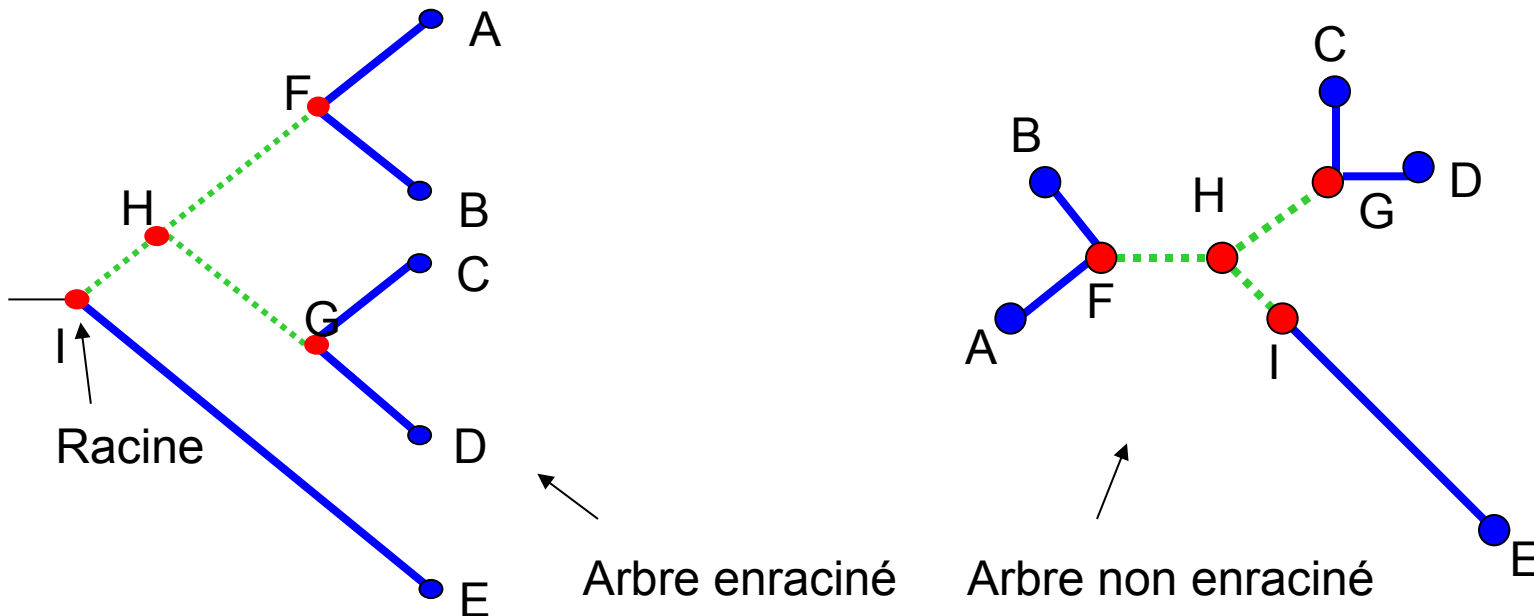


Topologies identiques (isomorphie)



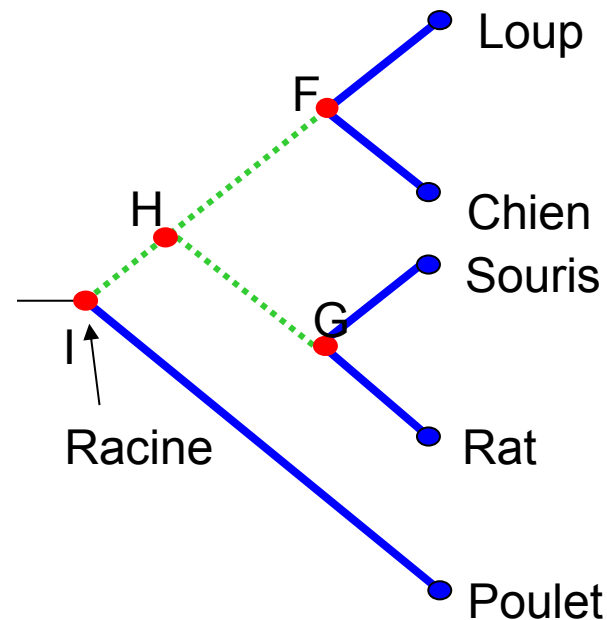
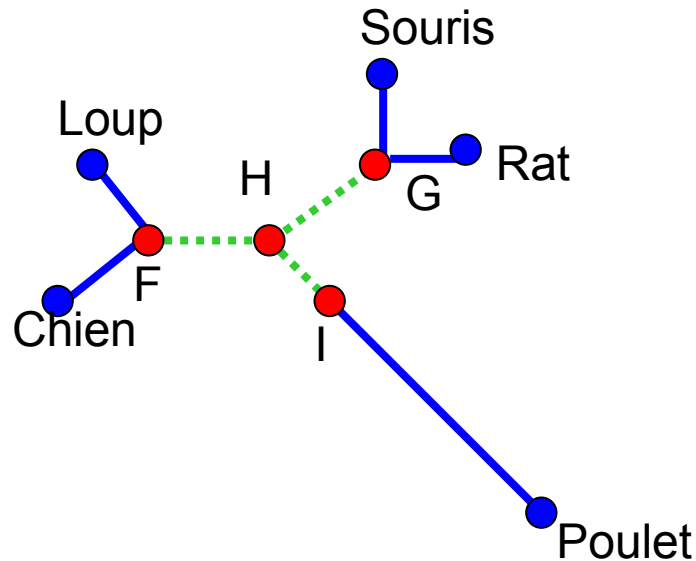
Pour évaluer la distance entre deux noeuds d'un arbre, il faut prendre en compte la longueur totale du chemin le plus court pour les rejoindre (somme des longueurs de branches).

Arbres enracinés vs Arbres non enracinés



- La racine symbolise le *dernier ancêtre commun* (i.e. le plus récent) de toutes les OTU (*Cenancestor* = *MRCA* (*Most Recent Common Ancestor*))
- La **racine** définit un chemin évolutif unique vers chaque feuille.
- Les arbres non enracinés ne sont pas réellement des arbres phylogénétiques car ils n'ont pas de dimension temporelle

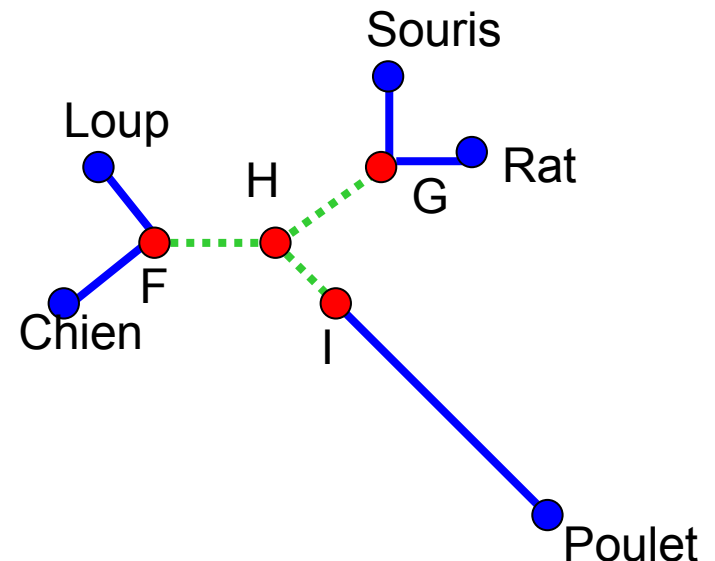
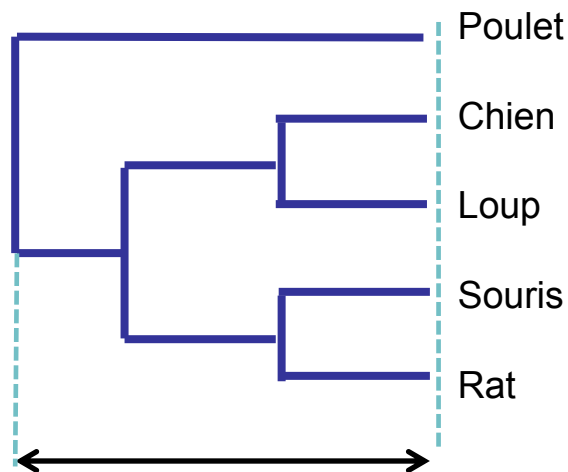
Comment enracer un arbre phylogénétique ?



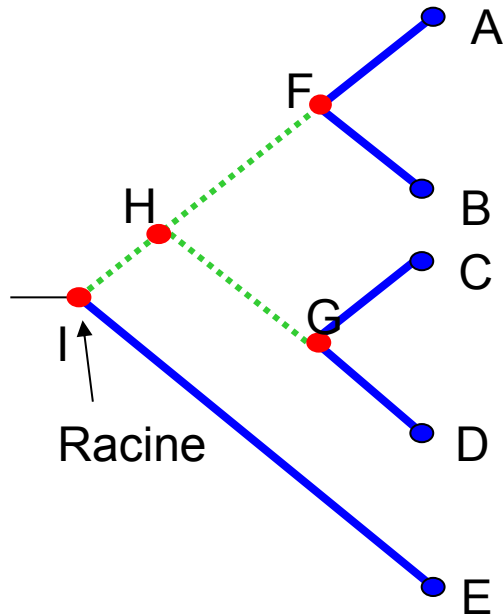
- Connaissance *a priori* du OTU le plus externe parmi les OTU étudiées
 - Exemple: chien, loup, souris, rat et poulet => **Groupe extérieur** est le poulet
- Sans connaissance *a priori* du OTU les plus externes parmi les OTU étudiées
 - Enracinement au poids moyen

Enracinement au poids moyen des arbres

- Hypothèse: Toutes les séquences évoluent à la même vitesse (i.e. hypothèse d'**horloge moléculaire**)
 - La même quantité d'évolution s'est produite dans chaque lignée évolutive depuis leur ancêtre commun à toutes
 - Les distances évolutives entre chaque feuille et la racine sont égales
 - La racine est placée au point de l'arbre équidistant de toutes les feuilles

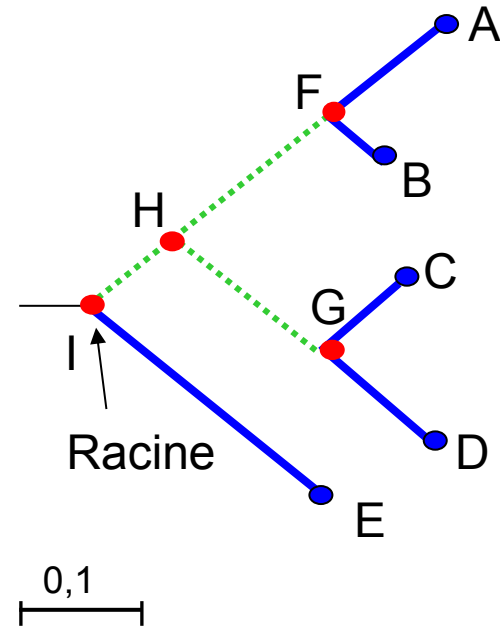


Échelle d'un arbre phylogénétique



Représentation sans échelle
(cladogramme)

- Les longueurs de branches ne sont pas proportionnelles au nombre de changements évolutifs. L'arbre représente uniquement l'ordre des branchements.



Représentation avec échelle
(phylogramme)

- Les longueurs de branches sont proportionnelles au nombre d'événements évolutifs (substitutions ou nombre de substitutions/sites)
- Echelle: nombre de substitutions ou nombre de substitution/sites

Cladistique, cladogrammes et clades

- **Cladistique**

- (du grec: klados = branche) classe les êtres vivants selon leurs relations de parenté, basé sur leurs caractères

- **Cladogramme**

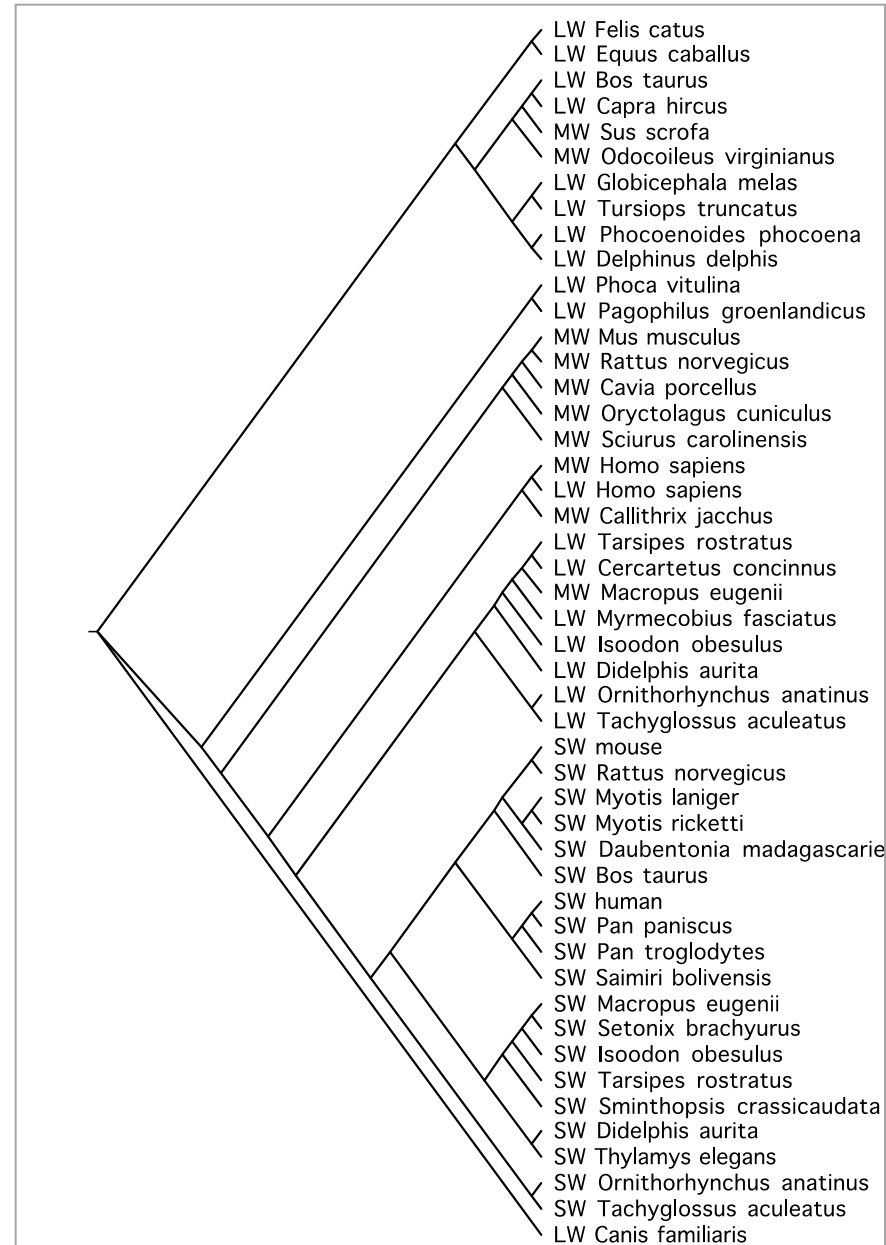
- Arbre, habituellement avec bifurcations, représentant un scénario évolutif des divergences entre espèces ou séquences.

- **Clade**

- Une branche de cladogramme avec un ancêtre commun et tout ses descendants.

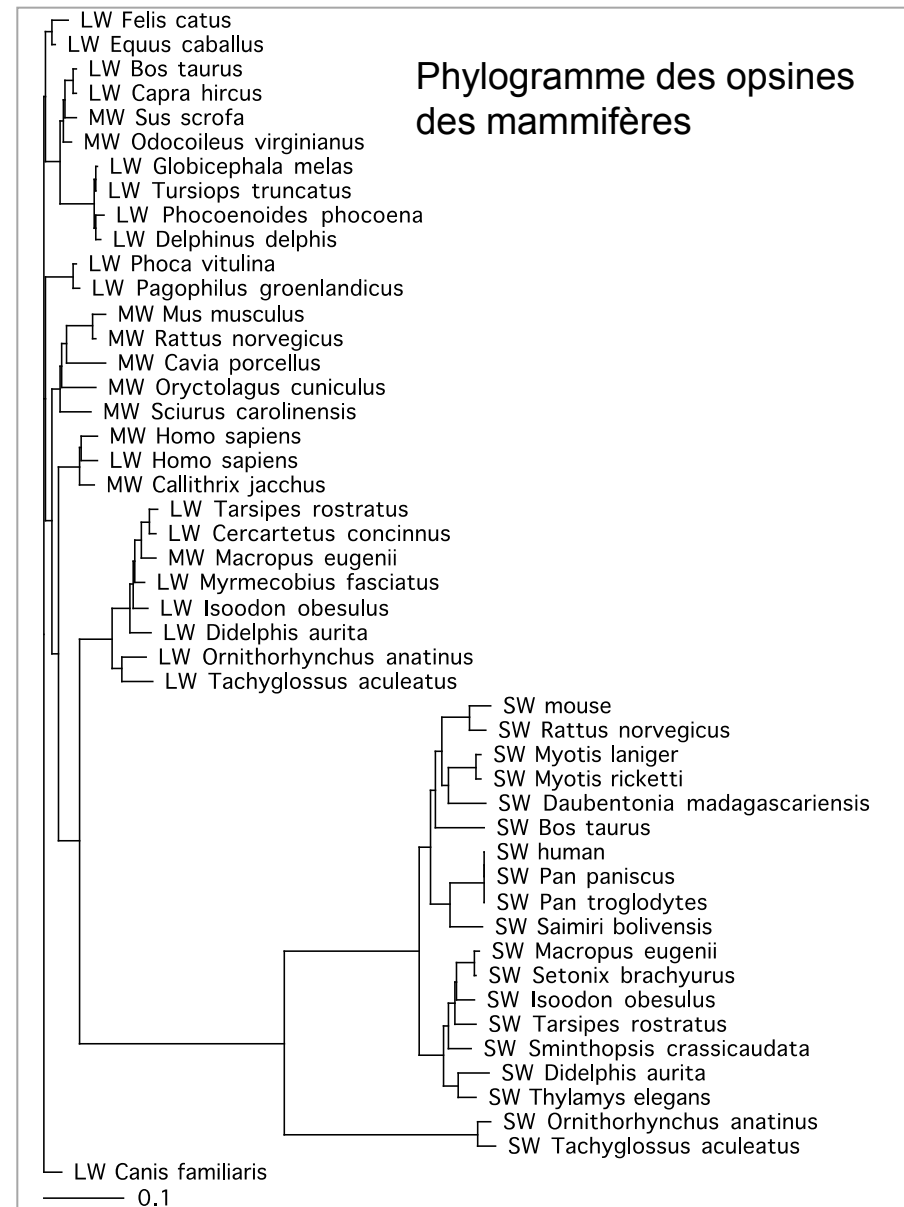
- **Note**

- La longueur de la branche ne reflète pas le temps ou le taux de divergence.
- Seule la topologie est informative, il n'y a pas d'échelle temporelle



Phylogramme

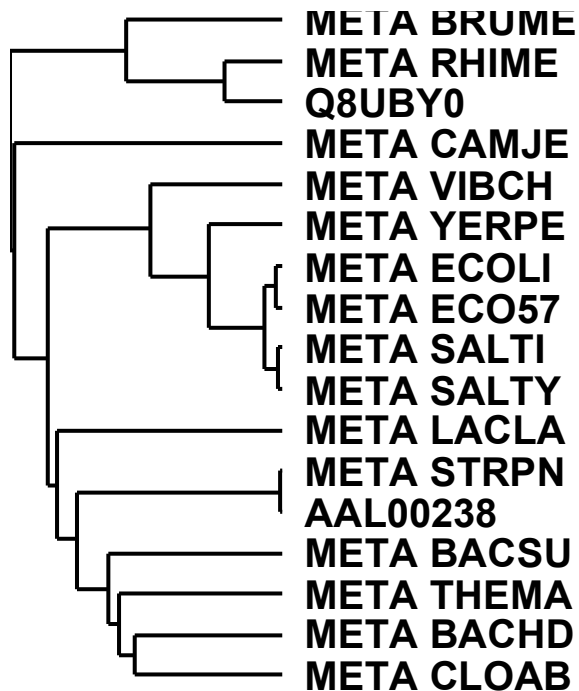
- **Phylogramme** : les longueurs des branches représentent les nombres d'événements évolutifs (e.g. mutations)
- **Notes:**
 - L'échelle relative est en bas
 - Arbre non-enraciné => la racine devrait être placée entre les opsines blues (SW) et les opsines rouge et vert (MW, LW)
 - La distance entre deux nœuds est la somme des branches entre eux.
 - La distance verticale a peu d'importance
 - $D_{(LW\ T.aculeatus - SW\ mouse)} \gg D_{(SW\ mouse - SW\ rattus)}$
 - Les longueurs des branches sont seulement des approximations des distances inférées.



Horloge moléculaire

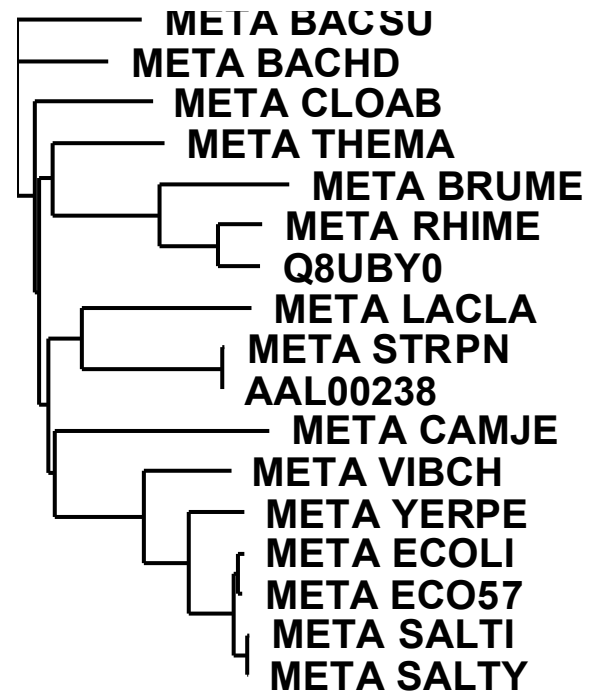
- **Chronogramme:** Longueur des branches représentent le temps de divergence.
 - Hypothèse de l'horloge moléculaire suppose que le taux d'évolution ne varie pas entre les branches. Tous les OTUs sont alignés verticalement sur l'arbre.
 - L'horloge moléculaire n'est pas toujours valide. Par exemple les paralogues peuvent avoir les taux de mutation fort différents car ils ne sont pas soumis à la même pression de sélection.

L'arbre avec hypothèse de l'horloge moléculaire (e.g. UPGMA)



_0.1

L'arbre sans l'horloge moléculaire (e.g. neighbour-joining)

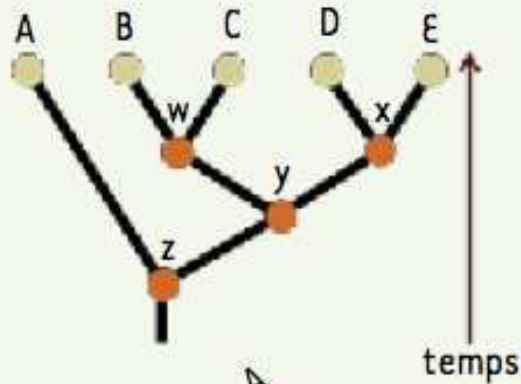


_0.1

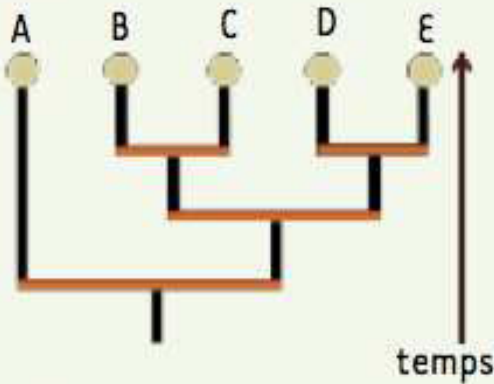
Résumé : représentations arborescentes

Cladogramme

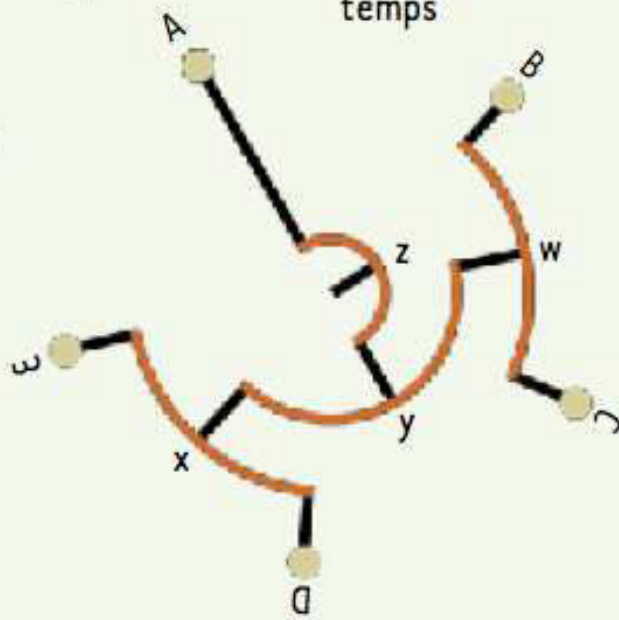
A 1



A 2

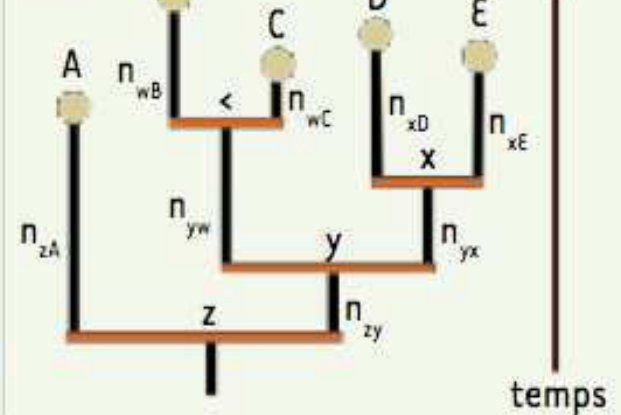


A 3



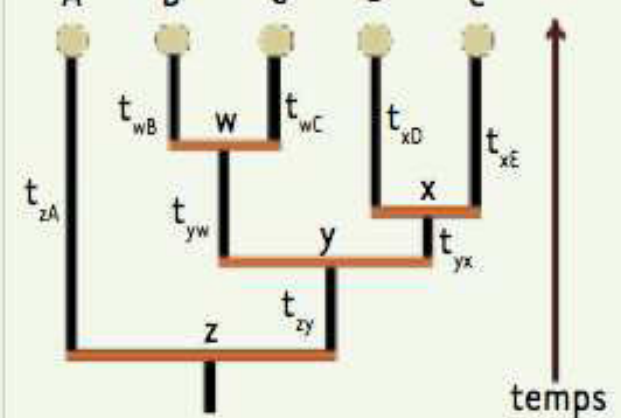
Phylogramme

B



Chronogramme

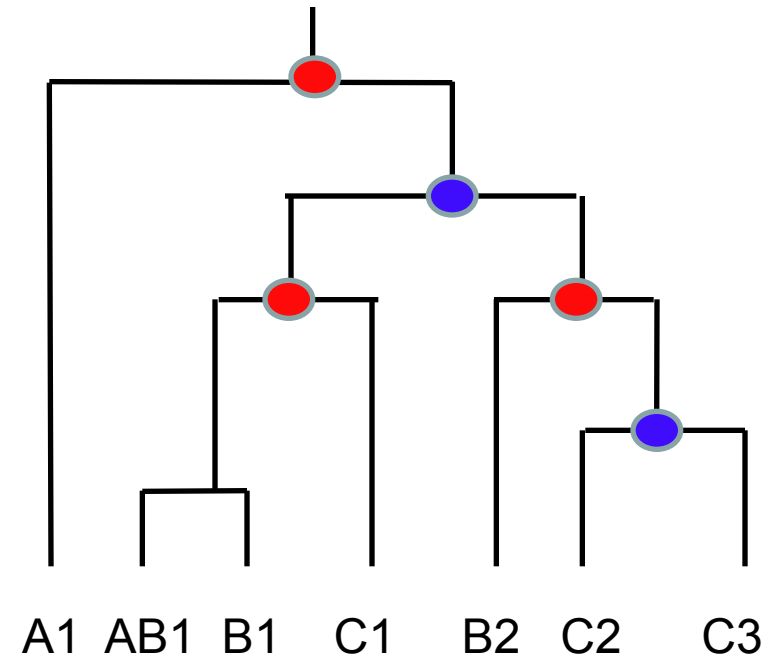
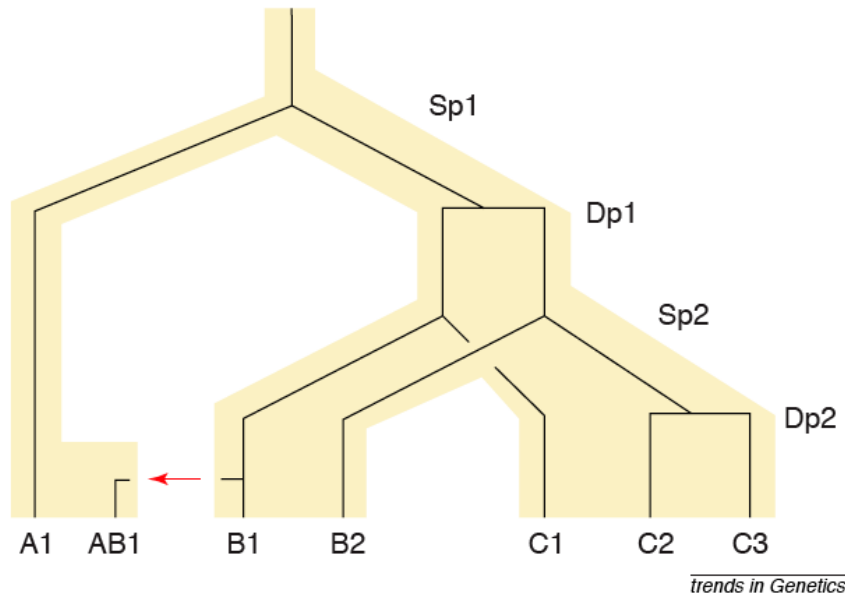
C



Arbre des gènes vs. arbre des espèces



- L'arbre des espèces représente les relations évolutives entre espèces.
- L'arbre des molécules représente l'histoire évolutive des molécules apparentés (gènes, protéines).
- L'arbre des espèces peut être inféré à partir des molécules, mais attention aux
 - Paralogie (duplications des gènes).
 - Xénologie (transfères horizontaux).

Ortologie/Paralogie



	A1	AB1	B1	B2	C1	C2	C3
A1		X	O	O	O	O	O
AB1	X		X	X	X	X	X
B1	O	X		P	O	P	P
B2	O	X	P		P	O	O
C1	O	X	O	P		P	P
C2	O	X	P	O	P		P
C3	O	X	P	O	P	P	

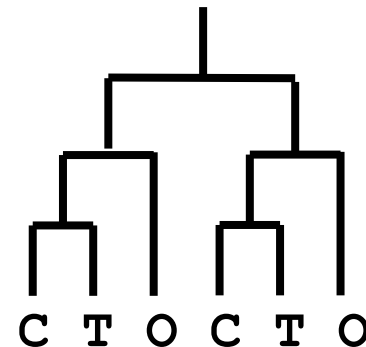
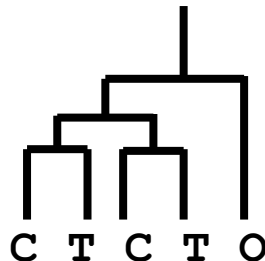
A, B, C représentent les espèces
1, 2, 3 les copies des gènes

-  Spéciation
-  Duplication

Arbre des gènes vs. arbre des espèces

- BLASTp de protéine F1PLD4 récepteur olfactif du chien contre la banque non-redondant des protéines à l'NCBI
 - Max target sequences = 1000
- L'espèce le plus éloigné du chien est l'ornithorynque (*Ornithorhynchus anatinus*) parmi les espèces qui ont eu un hit.
- Le dernier hit (e-valeur la plus élevée) provient du *Chrysochloris asiatica* (Taupe dorée du Cap; taupe natif de l'Afrique de Sud)
- Quel arbre correspond à ces observations ?

C: Chien
T: Taupe
O: Ornithorynque



Arbre des gènes vs. arbre des espèces

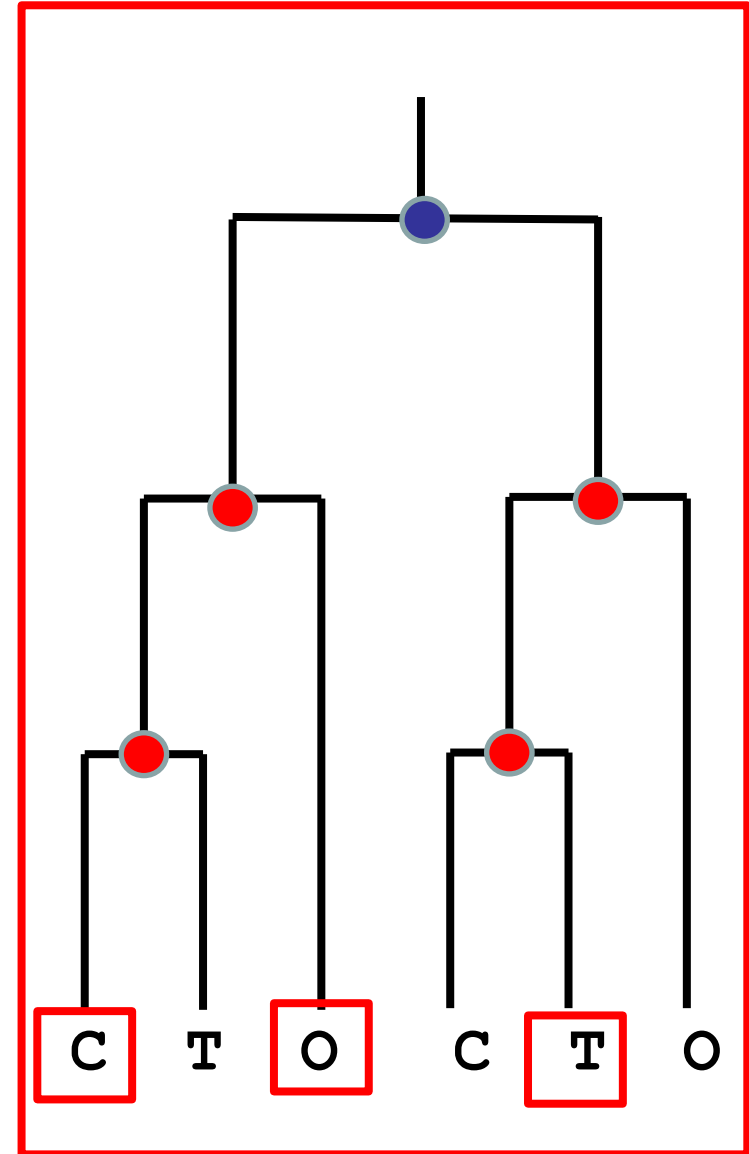
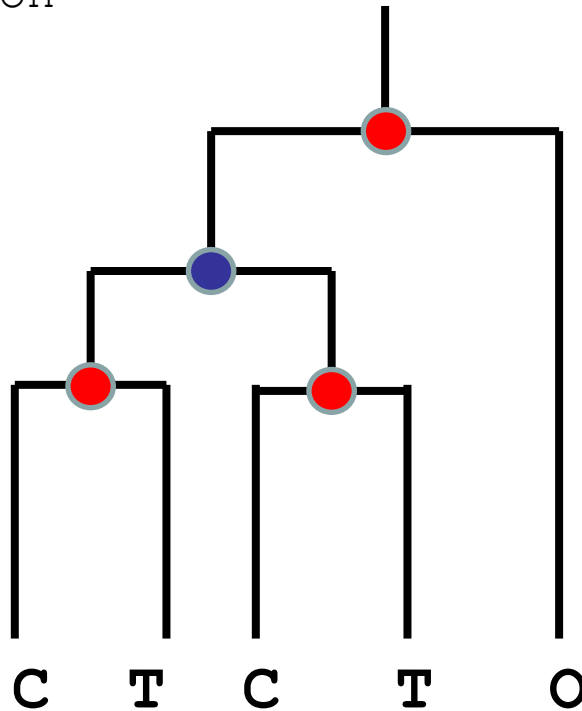
C: Chien

T: Taupe

O: Ornithorynque

● Duplication

● Spéciation



Arbre des gènes vs. arbre des espèces

C: Chien

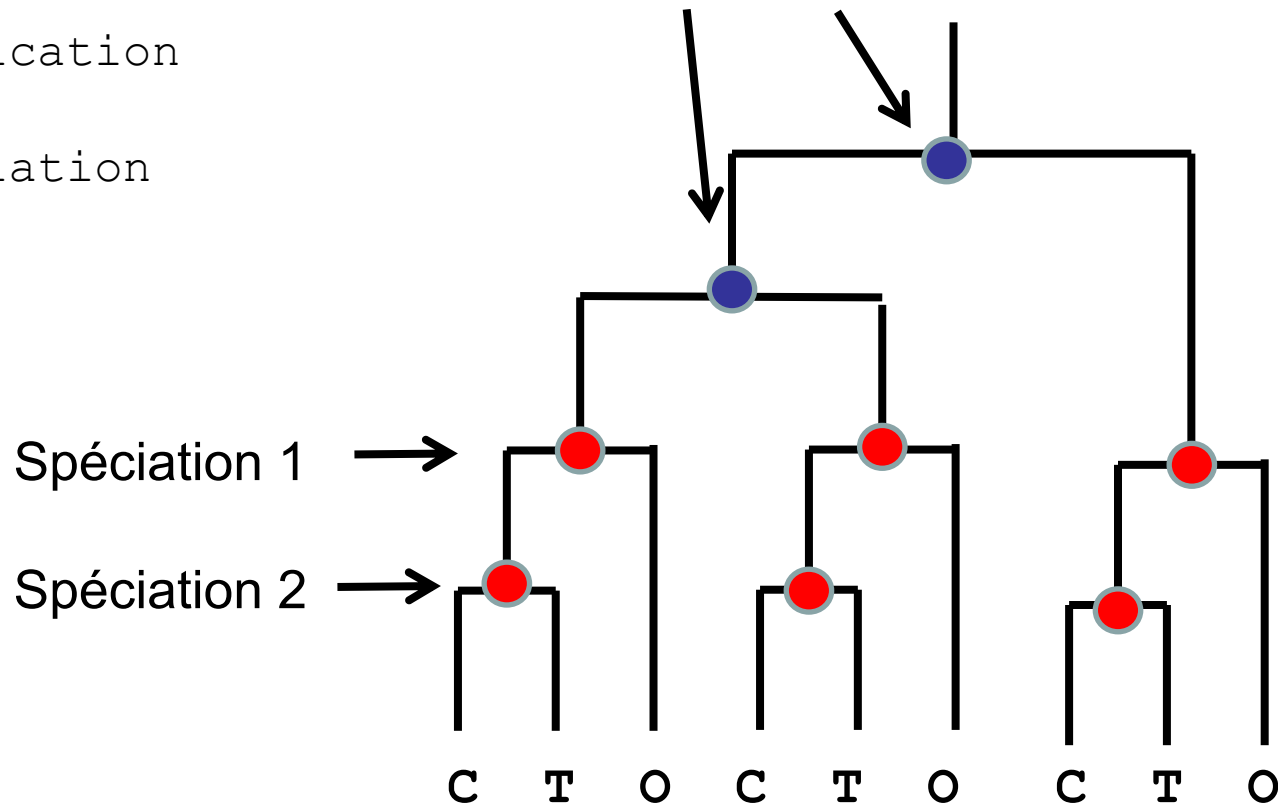
T: Taupe

O: Ornithorynque

● Duplication

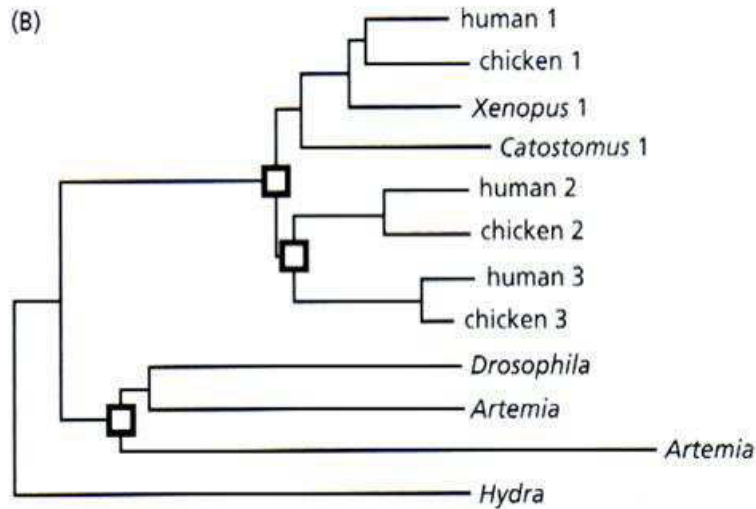
● Spéciation

Suite des duplications

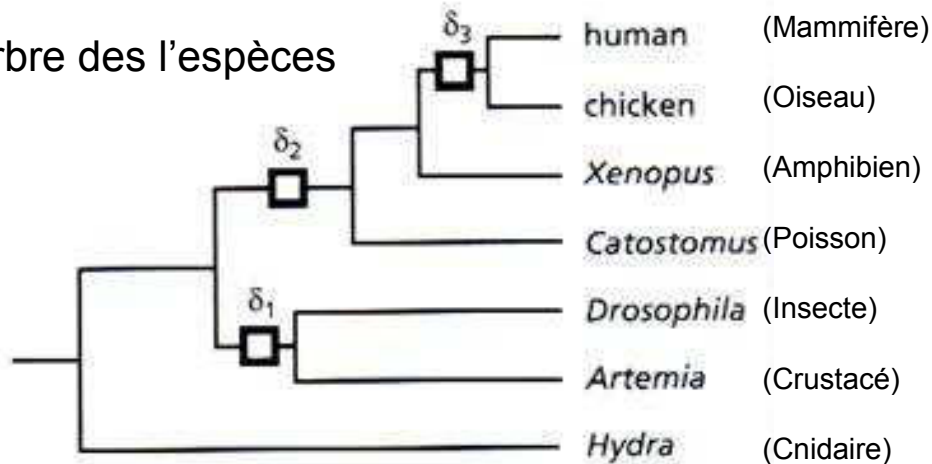


Réconciliation des arbres

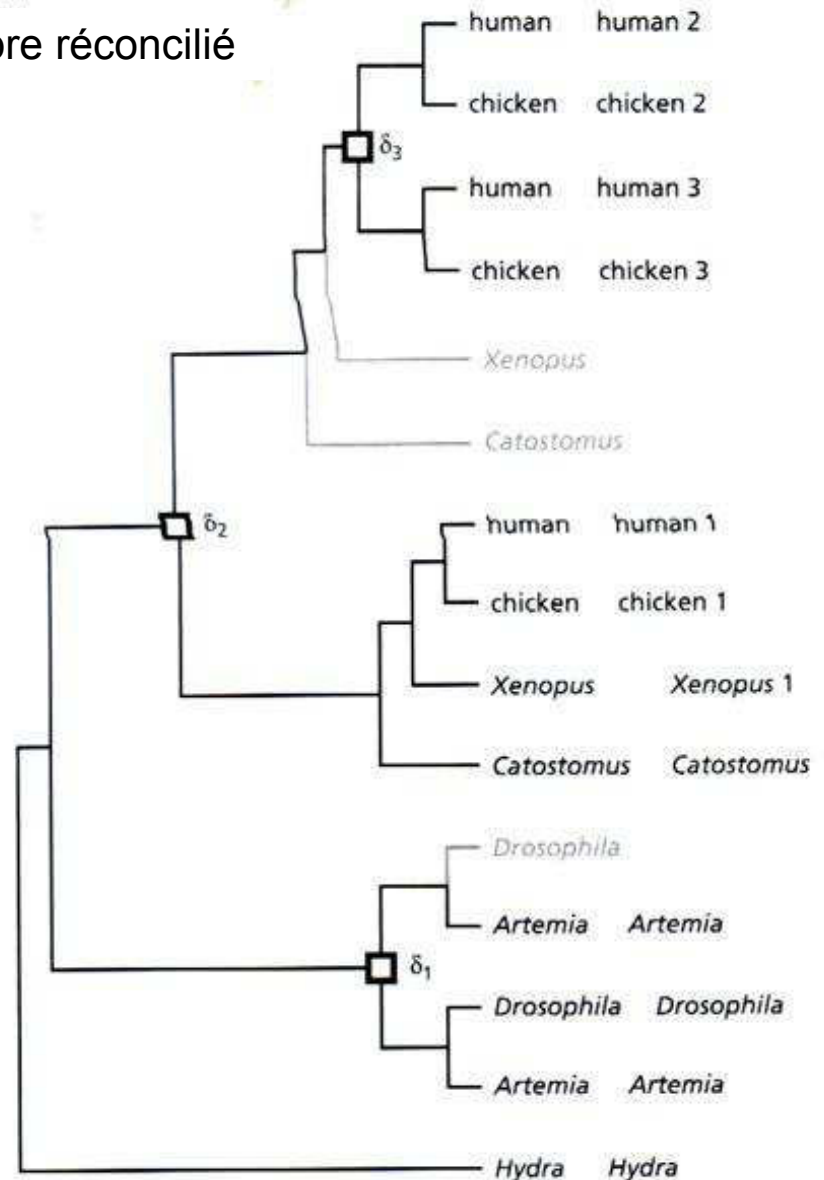
L'arbre des molécules



L'arbre des espèces



L'arbre réconcilié



Méthodes de construction des arbres phylogénétiques

Combien d'arbres ?

Nombre d'arbres enracinés

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

Nombre d'arbres non enracinés

$$N_U = \frac{(2n - 5)!}{2^{n-3}(n - 3)!}$$

Nombre d'arbres non enracinés
pour n OTU = nombre des arbres
enracinés pour n-1 OTU

n	Nb Rooted trees	Nb unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10 395	945
8	135 135	10 395
9	2 027 025	135 135
10	3,45E+07	2 027 025
11	6,55E+08	3,45E+07
12	1,37E+10	6,55E+08
13	3,16E+11	1,37E+10
14	7,91E+12	3,16E+11
15	2,13E+14	7,91E+12
16	6,19E+15	2,13E+14
17	1,92E+17	6,19E+15
18	6,33E+18	1,92E+17
19	2,22E+20	6,33E+18
20	8,20E+21	2,22E+20

Arbres vrais et arbres inférés

- Parmi tous les arbres possibles un seul représente la véritable histoire évolutive = ARBRE VRAI
- Le (ou les) arbre(s) obtenu(s) à partir d'un JDD particulier et une méthode de reconstruction est appelé ARBRE INFERE

Caractères et états de caractères

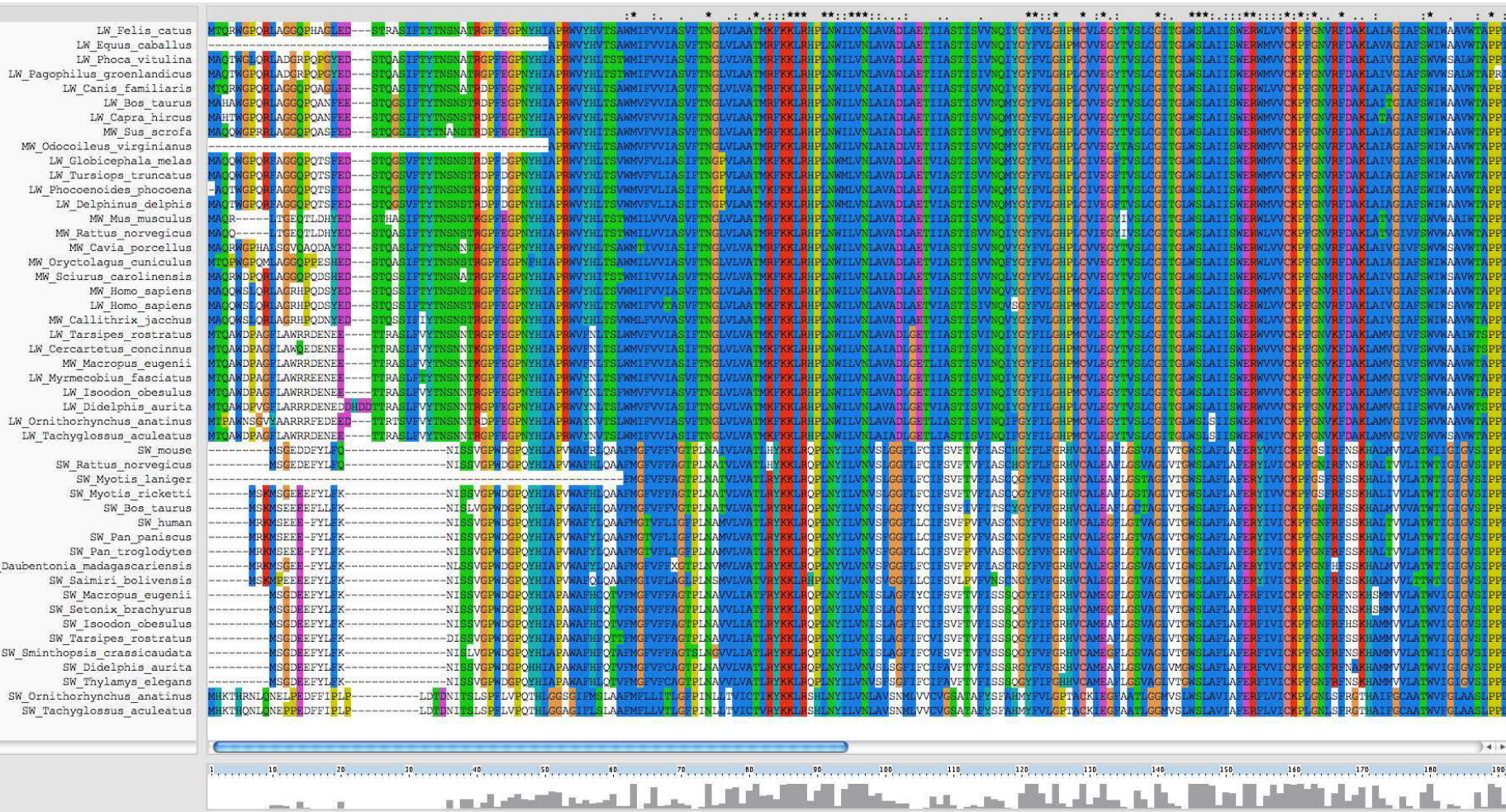
- *Caractère* = caractéristique observable d'un organisme (quantitative ou qualitative)
- *État de caractère* = forme particulière d'un caractère dans une OTU particulière (variable continue ou discrète)

Exemples:

- Caractère: Taille, Pos. 68 CYTB
- État de caractères: 1,68 cm, Alanine

Alignement multiples - Opsines

- Première étape de construction des arbres phylogénétiques: Alignement multiple
- Exemple: 50 opsines chez les mammifères.
- 2 groupes clairs:
 - Opsines rouges et verte (LW, MW)
 - Opsine bleue (SW)

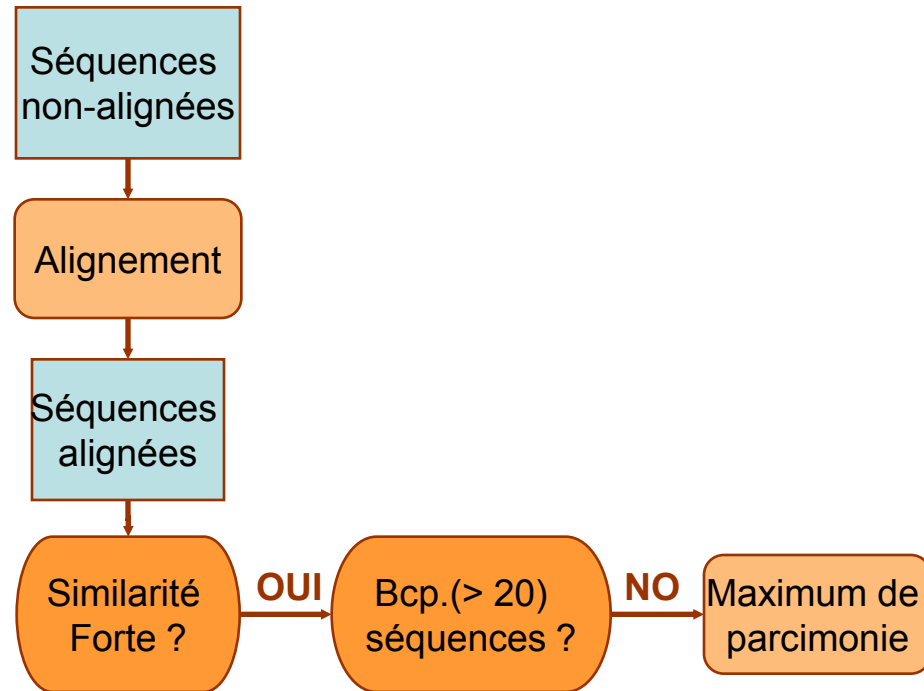


Méthodes

- Méthodes cladistiques
 - Basée sur l'étude des états de caractères (nucléotide ou acide aminé présent à une position, présence ou absence d'une insertion/délétion...)
 - Maximum de parcimonie
- Méthodes de distances (méthodes phénétiques)
 - Basées sur des mesures de distances (e.g. nombre de substitutions par site)
 - UPGMA, NJ, minimum d'évolution, moindres carrés...
- Méthodes statistiques
 - Basée sur l'étude des états de caractères et sur des distances
 - Maximum de vraisemblance
 - Méthodes bayésiennes

Choix des méthodes de construction des arbres phylogénétiques

- Approches alternatives
 - Maximum de parcimonie
 - Distance
 - Méthodes statistiques



Maximum de parcimonie - Principe

- Principe:
 - *Identifier la topologie T qui implique le plus petit nombre de changements évolutifs suffisant pour rendre compte des différences observées entre les OTU étudiées.*
 - Utilise des états de caractères discrets => L'arbre le plus parcimonieux => plus court chemin conduisant aux états de caractères observés
- Algorithme
 - Construction de tous les arbre possibles
 - Pour tous les site de l'alignement (caractère), on compte le nombre de substitutions nécessaire pour expliquer chaque arbre
 - On retient l'arbre qui nécessite le plus petit nombre de substitutions au total (en tenant compte de tous les sites)
- Caractéristique des arbres obtenus
 - Solutions multiples => plusieurs arbres avec le même nombre minimum de changements peuvent être obtenus
 - Le longueur des branches ne reflète par la distance évolutive (arbre sans échelle = cladogramme)
 - Arbres non enracinés

Maximum de parcimonie - Méthode

Matrice de caractères

		Sites								
		1	2	3	4	5	6	7	8	9
Séquences	A	A	A	G	A	G	T	T	C	A
	B	A	G	C	C	G	T	T	C	T
	C	A	G	A	T	A	T	C	C	A
	D	A	G	A	G	A	T	C	C	T

Maximum de parcimonie - Méthode

Déterminer toutes les topologies possibles

4 UTO => 3 arbres non racinés

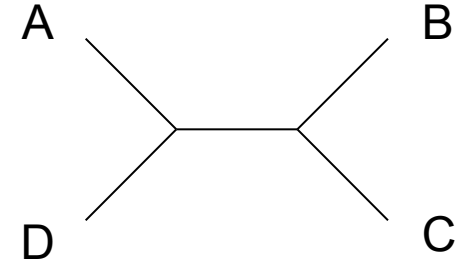
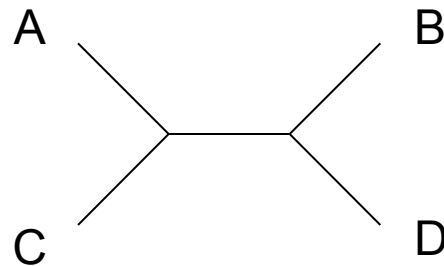
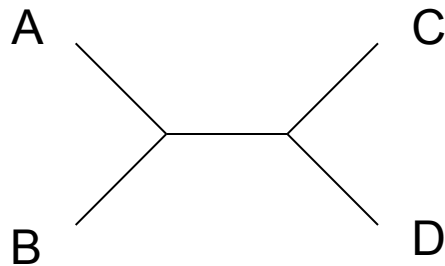
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Maximum de parcimonie - Méthode

Déterminer toutes les topologies possibles

4 UTO => 3 arbres non racinés

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



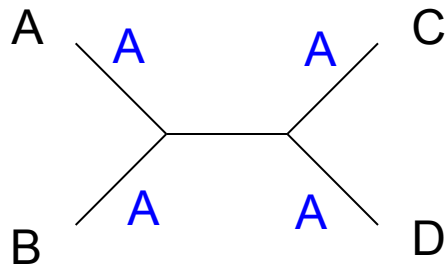
Maximum de parcimonie - Méthode

Étude du caractère n°1

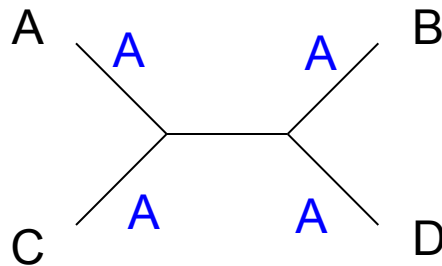
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère constant (même état de caractère à tous les sites)

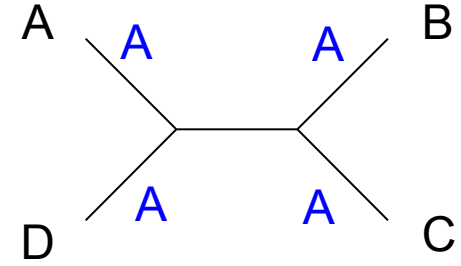
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 0



Nb CE= 0



Nb CE= 0

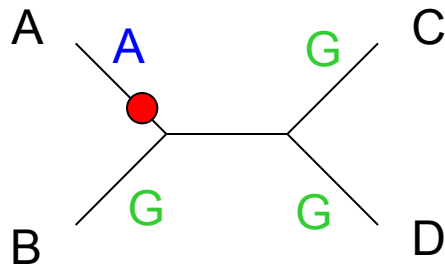
Maximum de parcimonie - Méthode

Étude du caractère n°2

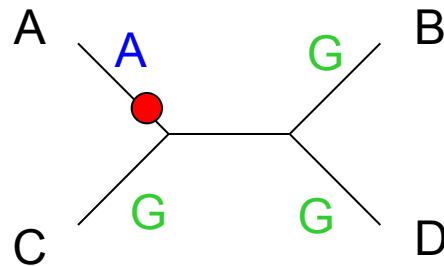
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable mais non informatif

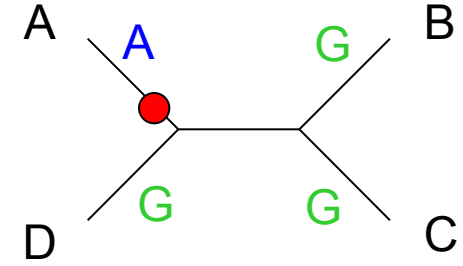
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 1



Nb CE= 1

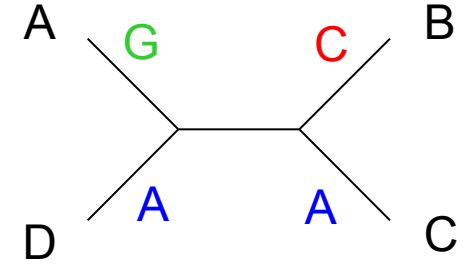
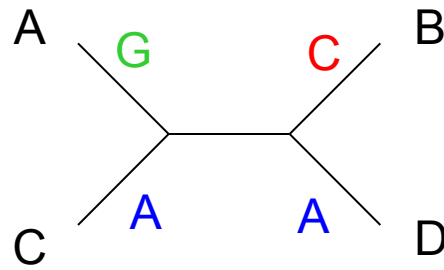
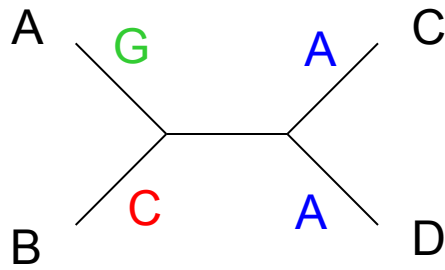


Nb CE= 1

Maximum de parcimonie - Méthode

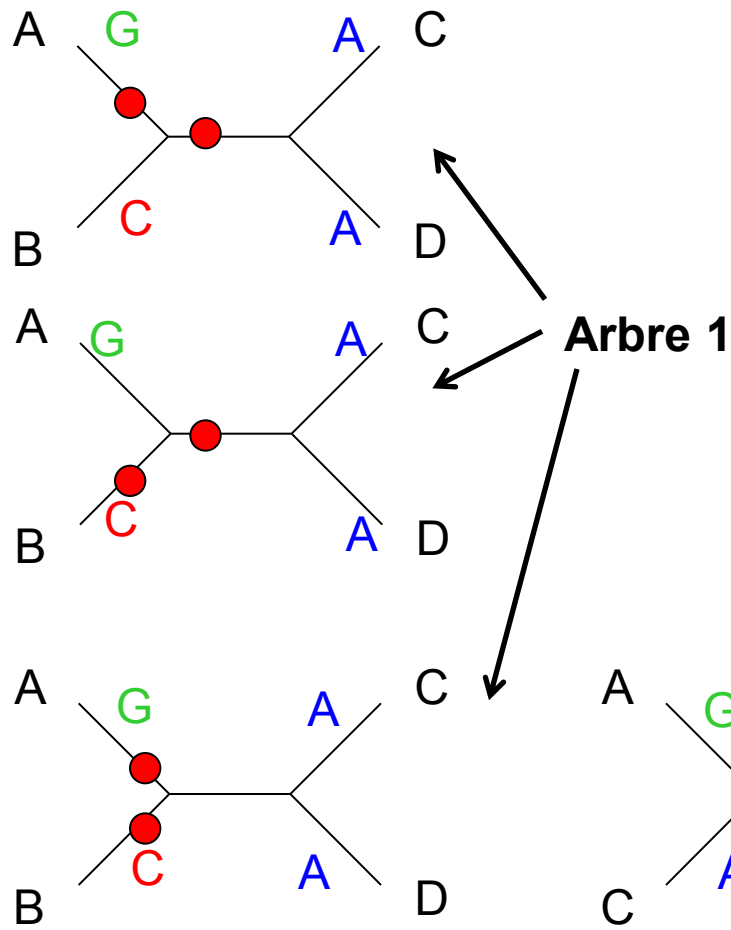
Étude du caractère n°3

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

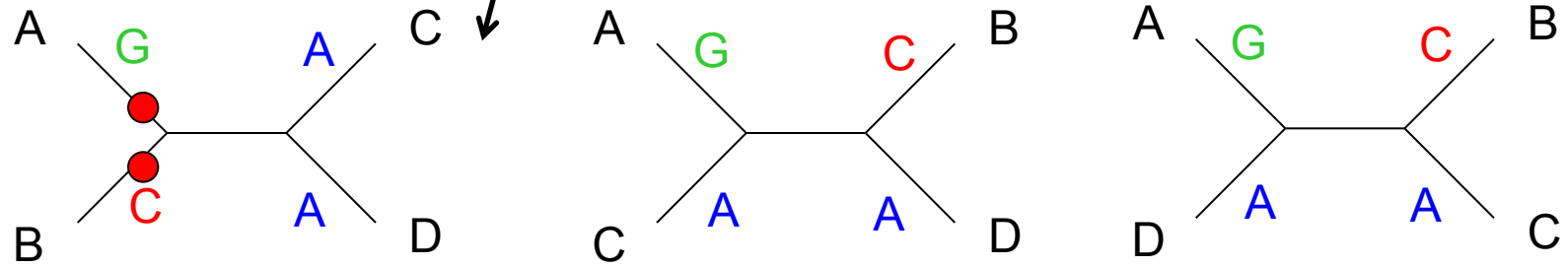


Maximum de parcimonie - Méthode

Étude du caractère n°3



	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



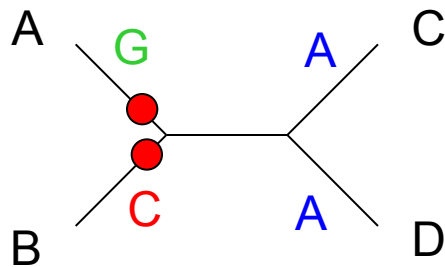
Maximum de parcimonie - Méthode

Étude du caractère n°3

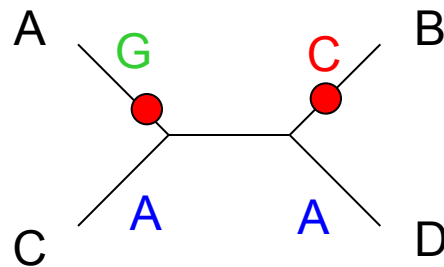
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable mais non informatif

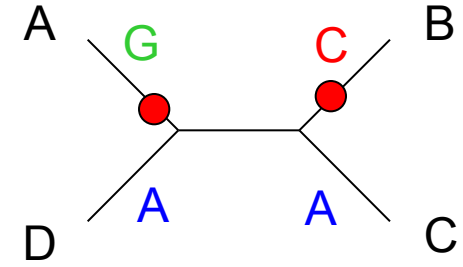
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 2



Nb CE= 2



Nb CE= 2

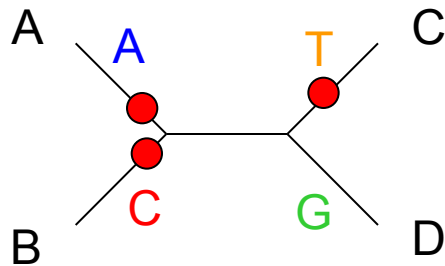
Maximum de parcimonie - Méthode

Étude du caractère n°4

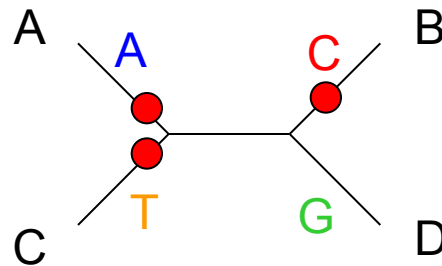
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable mais non informatif

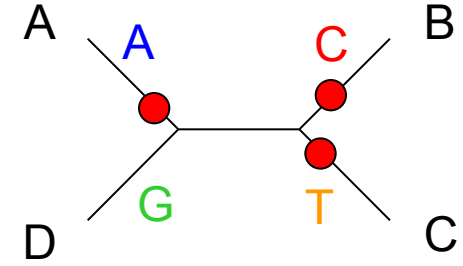
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 3



Nb CE= 3

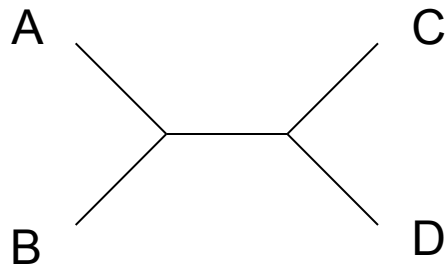


Nb CE= 3

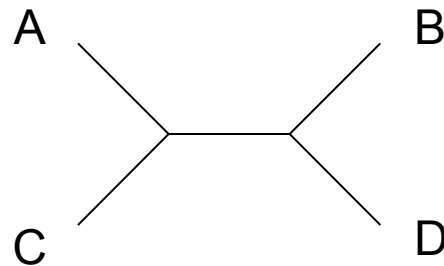
Maximum de parcimonie - Méthode

Étude du caractère n°5

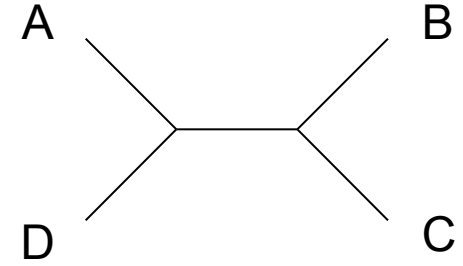
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



Nb CE= ?



Nb CE= ?



Nb CE= ?

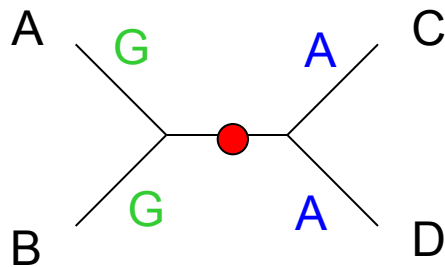
Maximum de parcimonie - Méthode

Étude du caractère n°5

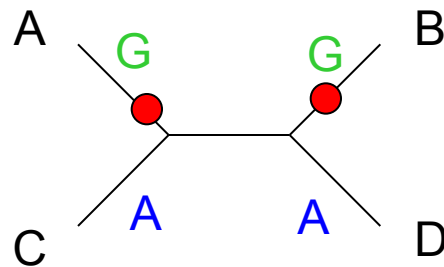
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable et **informatif** (au moins 2 états de caractère sont partagés par au moins 2 OTU)

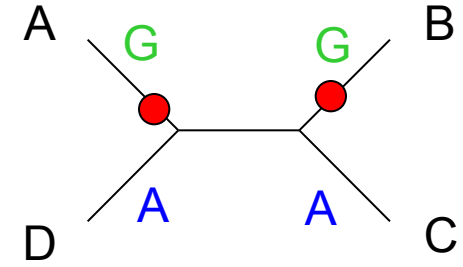
Caractère favorisant la première topologie par rapport aux deux autres



Nb CE= 1



Nb CE= 2



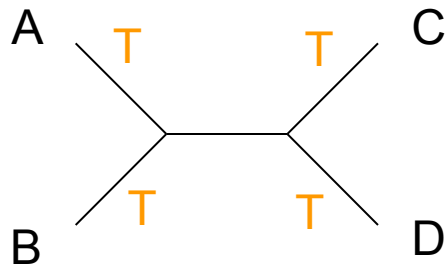
Nb CE= 2

Maximum de parcimonie - Méthode

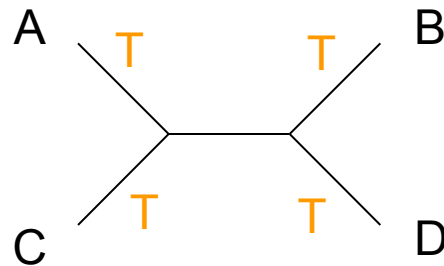
Étude du caractère n°6

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

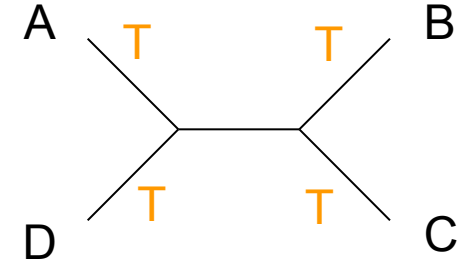
Caractère constant (même état de caractère chez tous les OTUs)
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 0



Nb CE= 0



Nb CE= 0

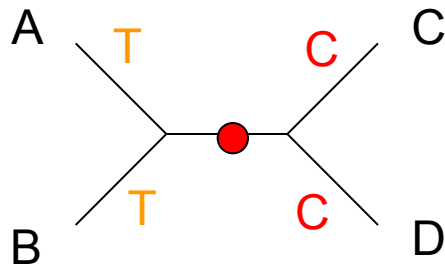
Maximum de parcimonie - Méthode

Étude du caractère n°7

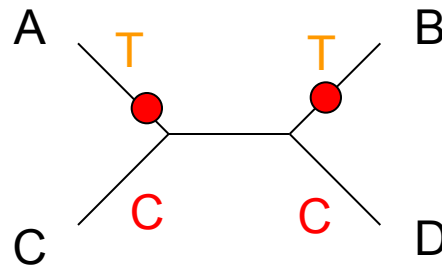
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable et **informatif**

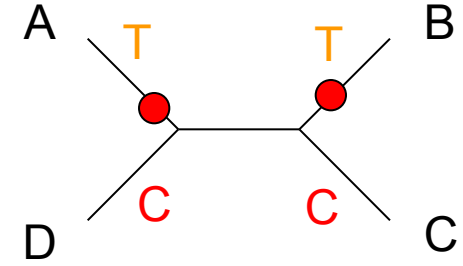
Caractère favorisant la première topologie par rapport aux deux autres



Nb CE= 1



Nb CE= 2



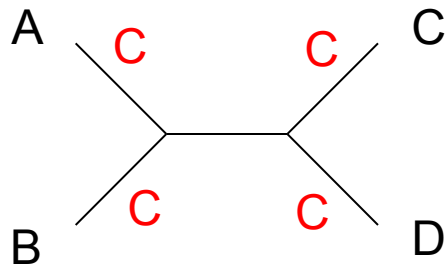
Nb CE= 2

Maximum de parcimonie - Méthode

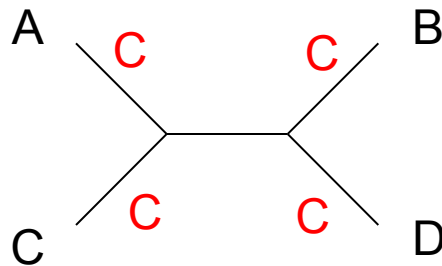
Étude du caractère n°8

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

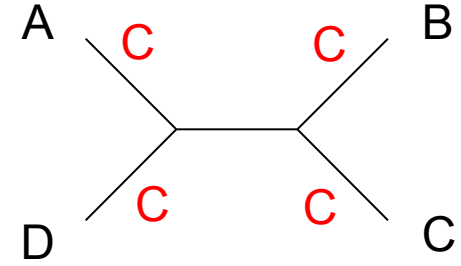
Caractère constant (même état de caractère à tous les OTUs)
Caractère ne favorisant aucune topologie par rapport à une autre



Nb CE= 0



Nb CE= 0

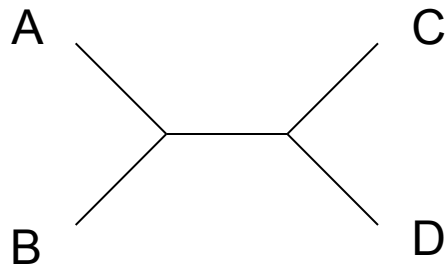


Nb CE= 0

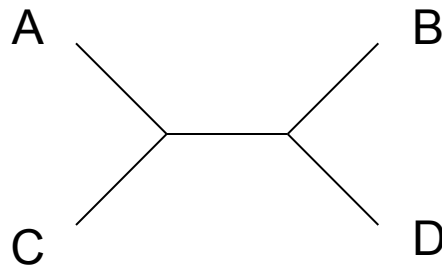
Maximum de parcimonie - Méthode

Étude du caractère n°9

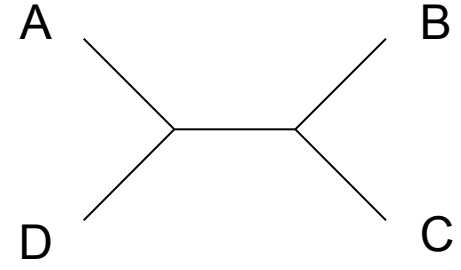
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



Nb CE= ?



Nb CE= ?



Nb CE= ?

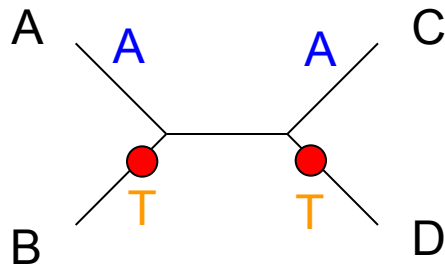
Maximum de parcimonie - Méthode

Étude du caractère n°9

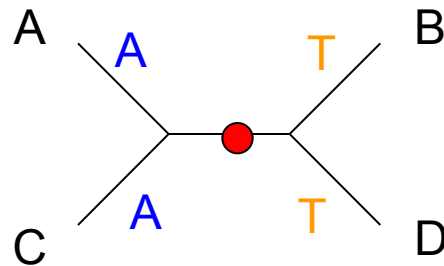
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable et **informatif**

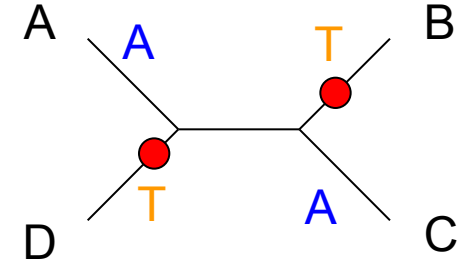
Caractère favorisant la deuxième topologie par rapport aux deux autres



Nb CE= 2



Nb CE= 1



Nb CE= 2

Maximum de parcimonie - Méthode

Bilan:

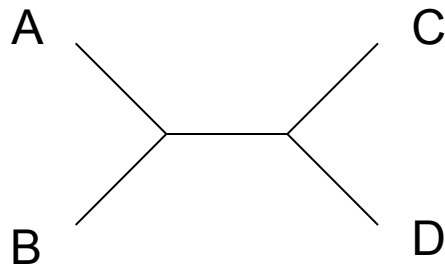
$$T1 = 0+1+2+3+1+0+1+0+2=10$$

$$T2 = 0+1+2+3+2+0+2+0+1=11$$

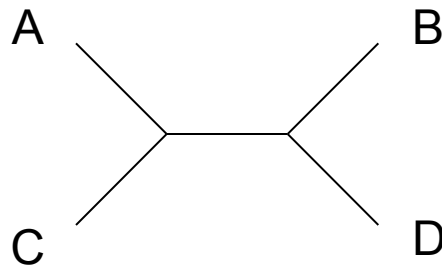
$$T3 = 0+1+2+3+2+0+2+0+2=12$$

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

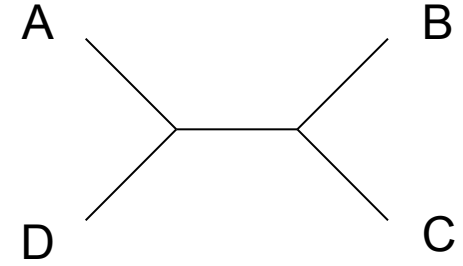
L'arbre le plus parcimonieux = arbre 1



Nb CE= 10



Nb CE= 11



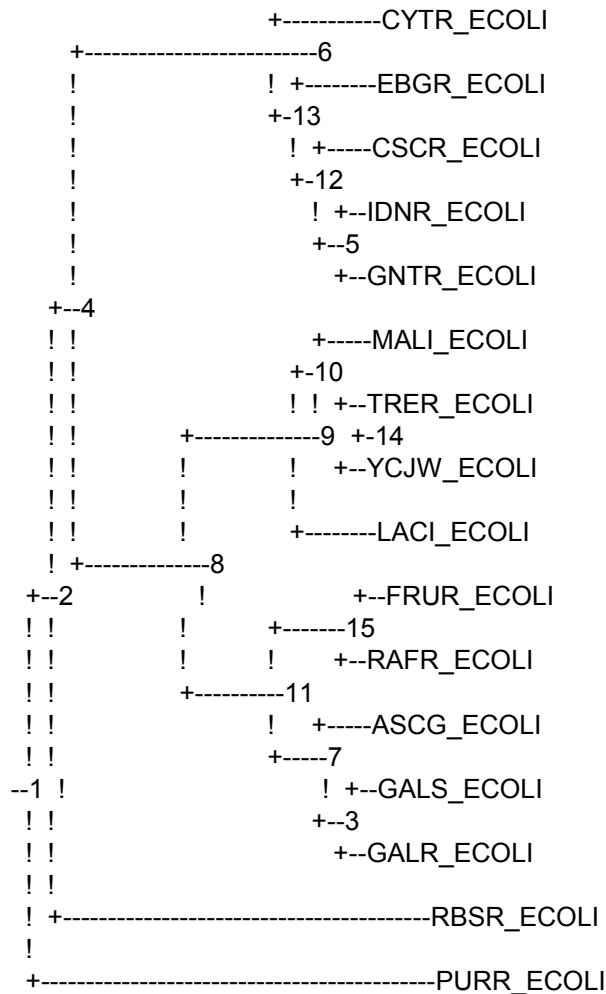
Nb CE= 12

Maximum de parcimonie - Classification des sites

- Caractères **invariants** si toutes les OTU possèdent le même état de caractères pour un site donné
- Caractères **variables**
 - **Non informatif** si les états de caractères à ce site ne favorisent aucune topologie parmi l'ensemble des topologies possibles
 - **Informatif** si les états de caractères à ce site favorise une (ou plusieurs) topologie(s) parmi l'ensemble des topologies possibles

Un site est informatif s'il présent au moins deux états de caractères chacun partagés par au moins deux séquences.

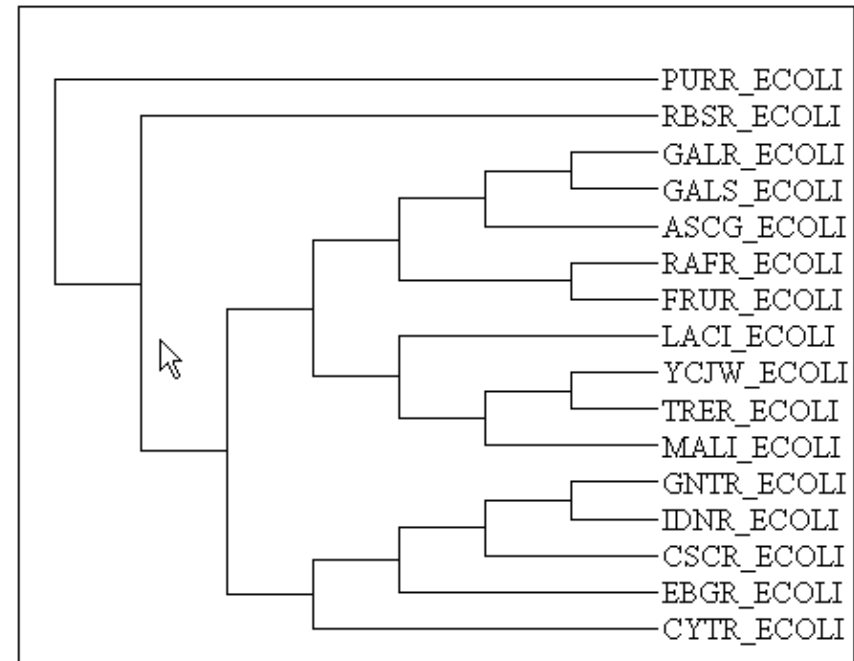
Maximum de parcimonie - Exemple



remember: this is an unrooted tree!

requires a total of 4095.000

- *Protéines de E.coli* contenant le domaine lacI-type HTH
 - Arbre sans échelle, non enraciné
 - Gauche: représentation format texte (*protpars* output)
 - Bas: Visualisation par *njplot* (integer dans *ClustalX*)

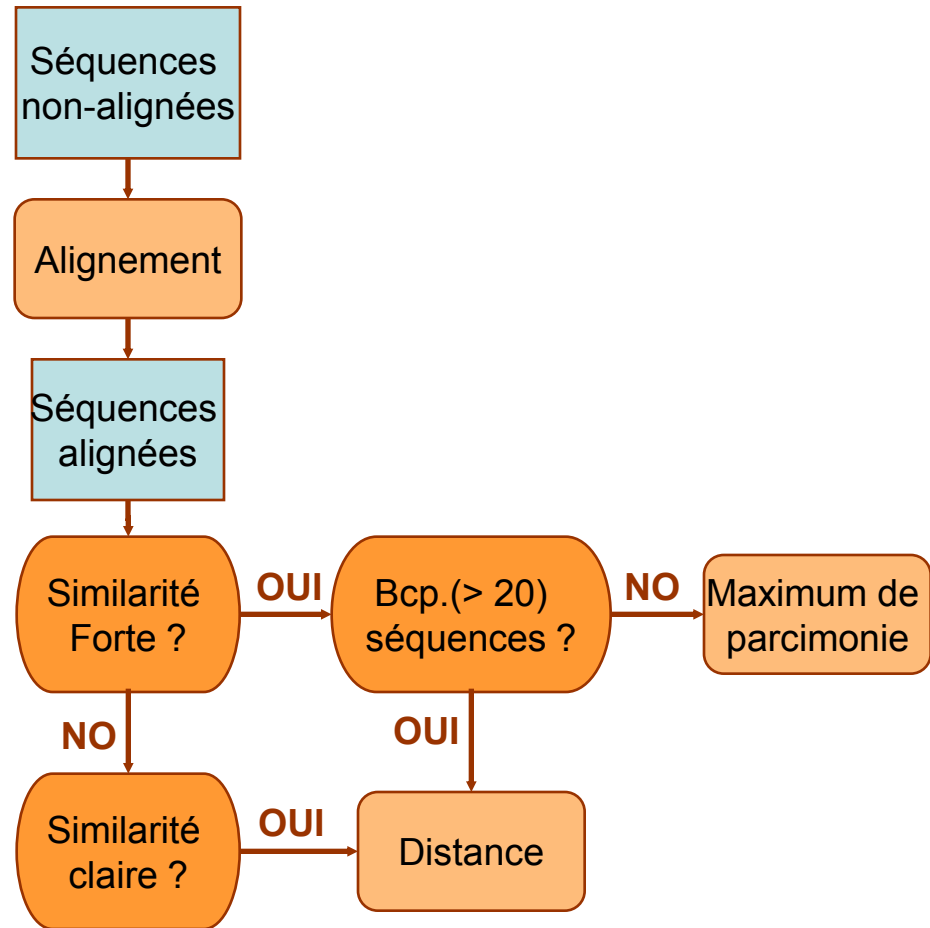


Maximum de parcimonie - Désavantages

- Le nombre d'arbres augmente exponentiellement avec le nombre d'OTUs (séquences).
- Hypothèse de l'horloge moléculaire => suppose que toutes les branches ont évolué avec la même vitesse.
- Fonctionne seulement avec les protéines très conservées.

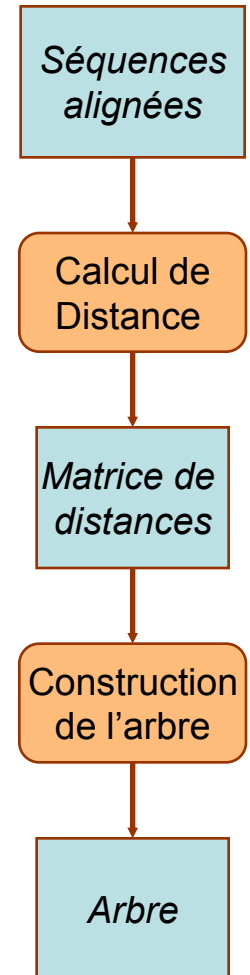
Choix des méthodes de construction des arbres phylogénétiques

- Approches alternatives
 - Maximum de parcimonie
 - Distance
 - Méthodes statistiques



Méthodes de Distance

- Alignement multiple
- Calcul de distance entre chaque paire des séquences
- Construction de l'arbre qui correspond le plus possible à la matrice de distances
 - La longueur des branches devrait correspondre aux distances, mais généralement on ne peut pas trouver un arbre où les longueurs des branches correspondent parfaitement avec la matrice de distances.
 - Arbres enracinés ou non-enracinés
- Il existe plusieurs méthodes de construction de l'arbre basées sur la distance.
 - Fitch-Margoliah
 - Neighbour-Joining
 - UPGMA



Méthodes de distances - Principe général

- **Calcul de toutes les distances évolutives (D_{ij})** séparant chaque paire d'UTO \Rightarrow Élaboration d'une matrice de distances à partir d'un alignement
- **Reconstruction d'un arbre phylogénique** dont les longueurs de branches (d_{ij}) *représentent* au mieux les distances évolutives de la matrice (D_{ij})

Calcul des distances entre deux séquences d'acides nucléiques

- Alignement des séquences
- **p-distance**: distance observée
 - s : nombre de substitutions observées entre deux séquences alignées
 - n : nombre de sites alignés
 - $p = s/n$

	1	2	3	4	5	6	7	8	9	10
A	A	A	G	A	G	T	T	C	A	A
B	A	G	C	C	G	T	T	C	T	A
C	A	G	A	T	A	T	C	C	A	A
D	A	G	A	G	A	T	C	C	T	A

p	A	B	C	D
A	0	0,4	0,6	0,6
B		0	0,5	0,5
C			0	0,2
D				0

Calcul des distances entre deux séquences d'acides nucléiques

- Distance p sous-estime les distances évolutives, quand les séquences sont éloignées (substitution multiples)
- Modèle de Jukes et Cantor
 - tous les sites évoluent indépendamment et selon le même processus
 - toutes les substitutions sont équiprobables
 - $d = -\frac{3}{4} \log(1 - \frac{4}{3}p)$
- Kimura à 2 paramètres
 - tous les sites évoluent indépendamment et selon le même processus
 - Les taux de substitution des transitions (p) et des transversions (q) sont différents
 - $d = -\frac{1}{2} \log[(1 - 2p - q)(1 - 2q)^{1/2}]$

Méthodes de distances - Principe général

- **Calcul de toutes les distances évolutives (D_{ij})** séparant chaque paire d'UTO \Rightarrow Élaboration d'une matrice de distances à partir d'un alignement
- **Reconstruction d'un arbre phylogénique** dont les longueurs de branches (d_{ij}) *représentent* au mieux les distances évolutives de la matrice (D_{ij})

Méthodes de distances

Calcul des arbres à partir de matrice de distances

- **Algorithme itératif de clustering** (par exemple UPGMA)
 - Regroupe les séquences par ordre de distance dans la matrice
 - Produit un arbre enraciné
 - Points faibles:
 - Repose sur l'hypothèse d'horloge moléculaire
 - Les longues branches (correspondant parfois à des évolutions rapides) sont considérées comme outgroups.
- **Neighbour-Joining (NJ)**
 - Minimise la somme des longueurs de branches de l'arbre résultant.
 - Ne repose pas sur une hypothèse d'horloge moléculaire
 - Retourne un arbre non-enraciné
 - Approprié quant certaines des séquences évoluent plus vite que d'autres.
- Méthode de **Fitch-Margoliah**
 - Minimise la somme des carrés de différences entre distances de la matrice et distances dans l'arbre

Reconstruction d'un arbre phylogénique

Algorithme itératif de clustering

- Algorithme itératif de clustering: *création à chaque étape d'un nouveau cluster regroupant deux clusters proches*
 1. Assigner chaque objet à un cluster séparé.
 2. Identifier la paire de clusters les plus proches, et les regrouper en un seul.
 3. Répéter la seconde étape jusqu'à ce qu'il ne reste qu'un seul cluster.
- Il existe plusieurs possibilités pour définir la distance entre deux groupes.
 - Liaison simple (single linkage): distance entre groupes A et B est la distance entre les plus proches de leurs éléments respectifs.
 - Liaison moyenne (average linkage): distance moyenne entre tous les objets des deux groupes (=UPGMA, Unweighted Pair-Group Method by arithmetic Averaging).
 - Liaison complète (complete linkage): distance entre les éléments les plus éloignés des groupes A et B.

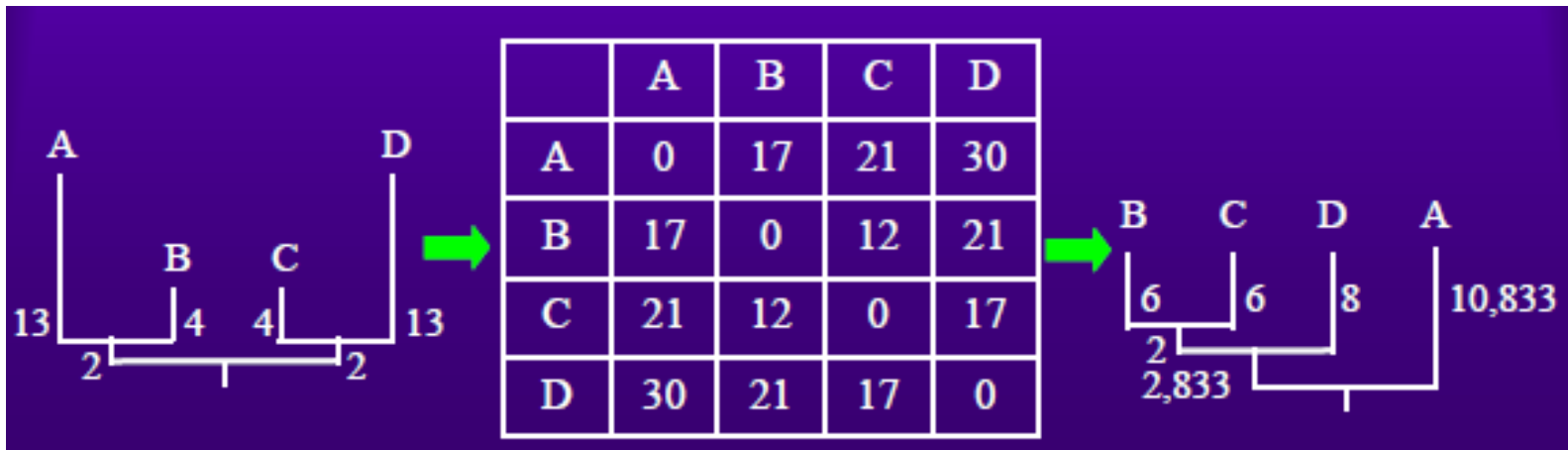
UPGMA

(Unweighted pair-group method with arithmetic means)

- Condition d'application
 - Hypothèse d'horloge moléculaire: constance des taux d'évolution le long des lignées
- Caractéristiques des arbres obtenus
 - Ils sont enracinés
 - Les longueurs des branches allant de la racine à n'importe quelle feuille sont égales
- Avantages de l'algorithme:
 - Rapidité & simplicité

Conclusions sur l'UPGMA

- Critiques:
 - Hypothèse de l'égalité des taux d'évolution entre les lignées.
 - Résultats faux si les distances de la matrice n'obéissent pas au critère d'horloge moléculaire
 - N'est presque plus utilisé
- Peut être réaliste si on étudie des espèces très proches

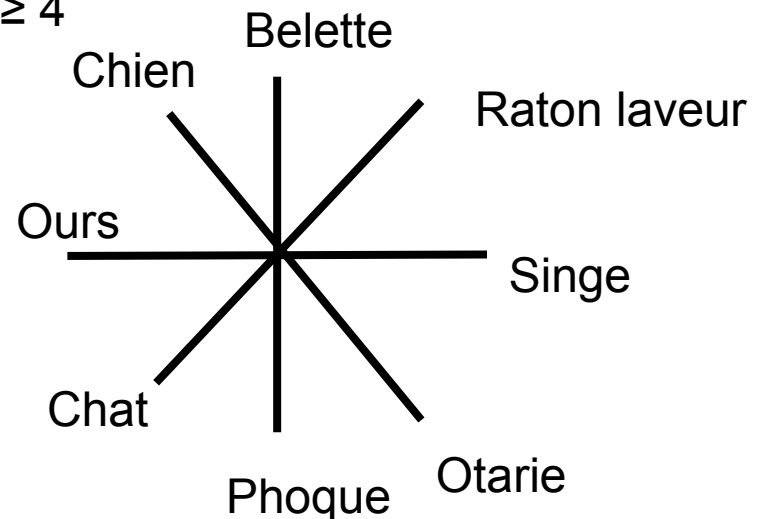


Neighbour joining (NJ) - Méthode

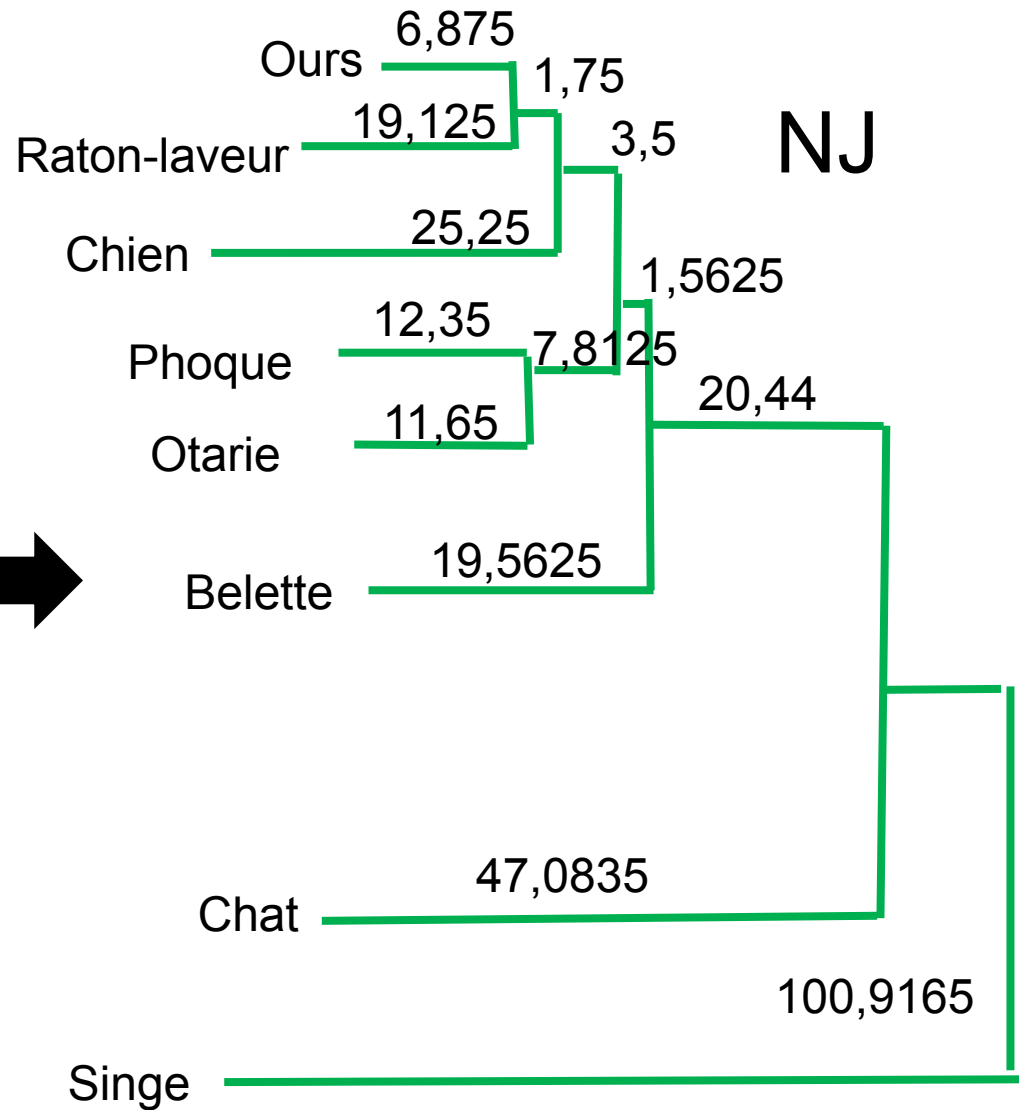
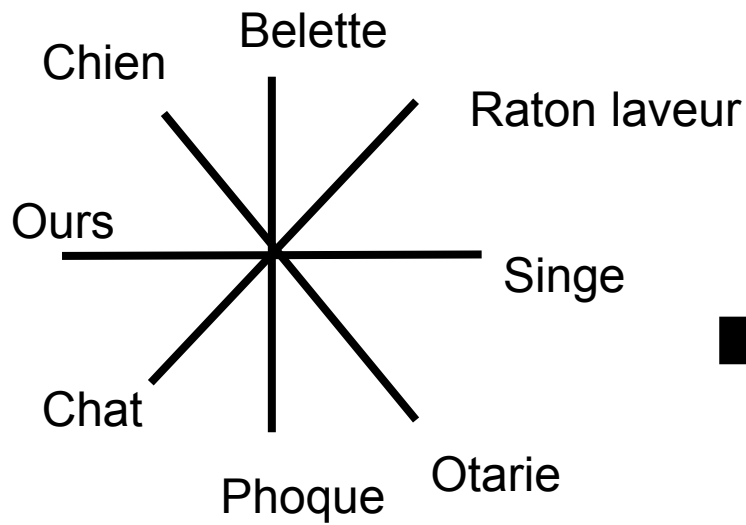
- Développé par Saitou et Nei (1987) est une approximation de l'algorithme pour trouver l'arbre le plus court (minimum évolution)
- Avantages
 - Rapidité => permet de travailler avec un très grand nombre de taxons (plusieurs centaines)
 - Bonne approximation de la méthode du minimum d'évolution
 - Retrouve l'arbre vrai si la matrice de distances est un reflet exact d'un arbre
- Conditions d'application
 - Les taux d'évolution ne sont pas les mêmes dans toutes les lignées
 - Les arbres ne sont pas enracinés
- Principe:
 - A chaque étape, rechercher le couple d'UTO qui minimise la longueur totale de l'arbre

Neighbour joining (NJ) - Algorithme

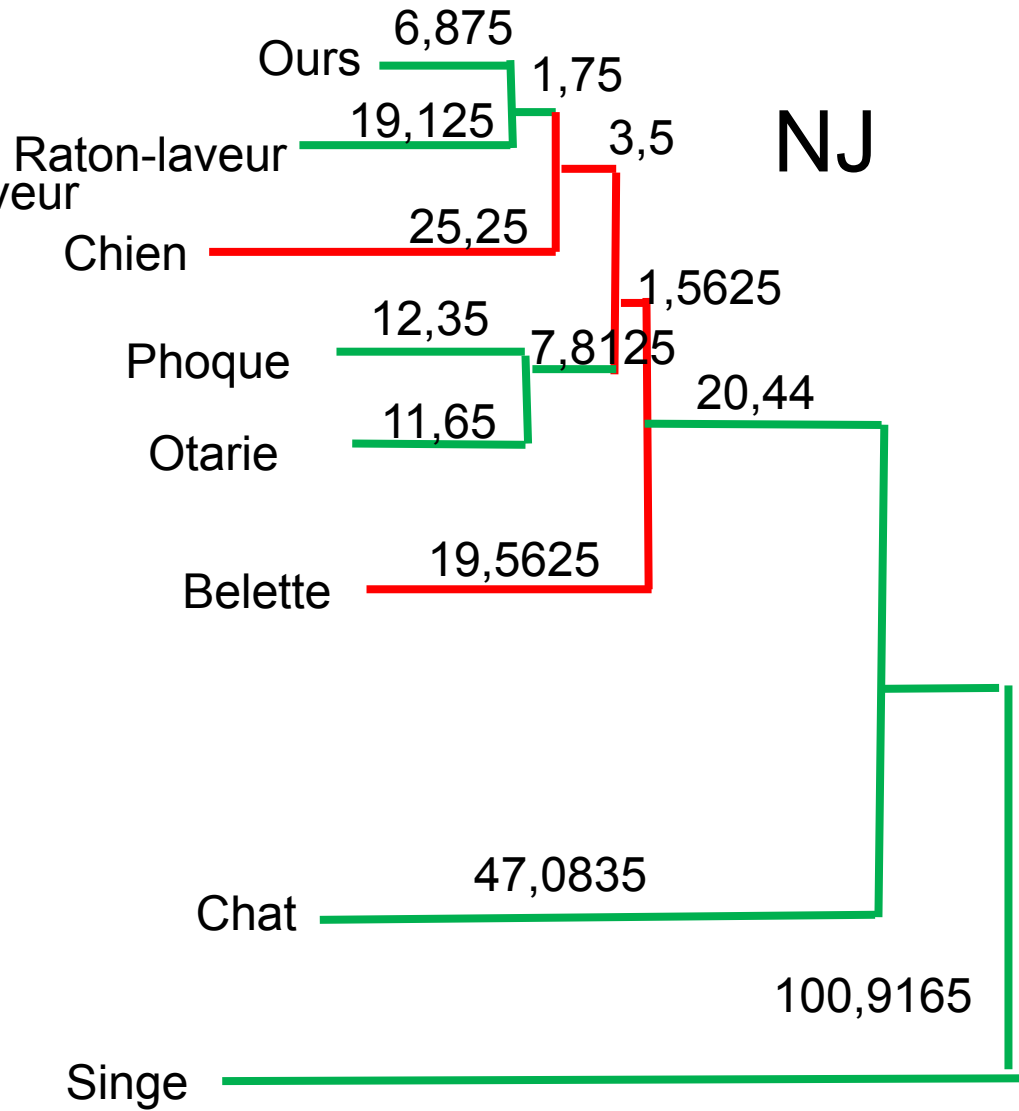
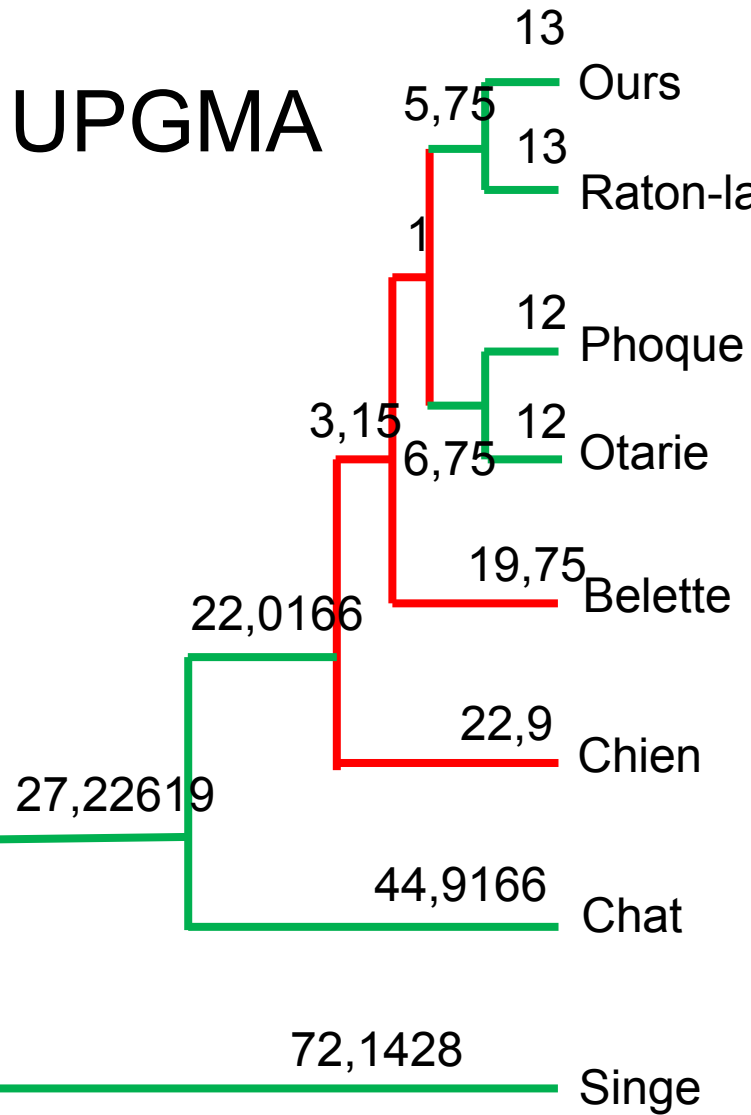
- Principe général:
 - **Point de départ = topologie en étoile**
 - Étape 1 : Pour toutes les paires i,j possibles, calculer $S_{i,j}$ la longueur de l'arbre obtenu
 - Étape 2 : Retenir la paire i,j générant la plus petite valeur $S_{i,j}$; grouper i et j dans l'arbre
 - Étape 3 : Calculer les nouvelles distances d entre les $N-1$ séquences
 - Étape 4 : Retourner à l'étape 1 si $N \geq 4$



Neighbour joining (NJ)

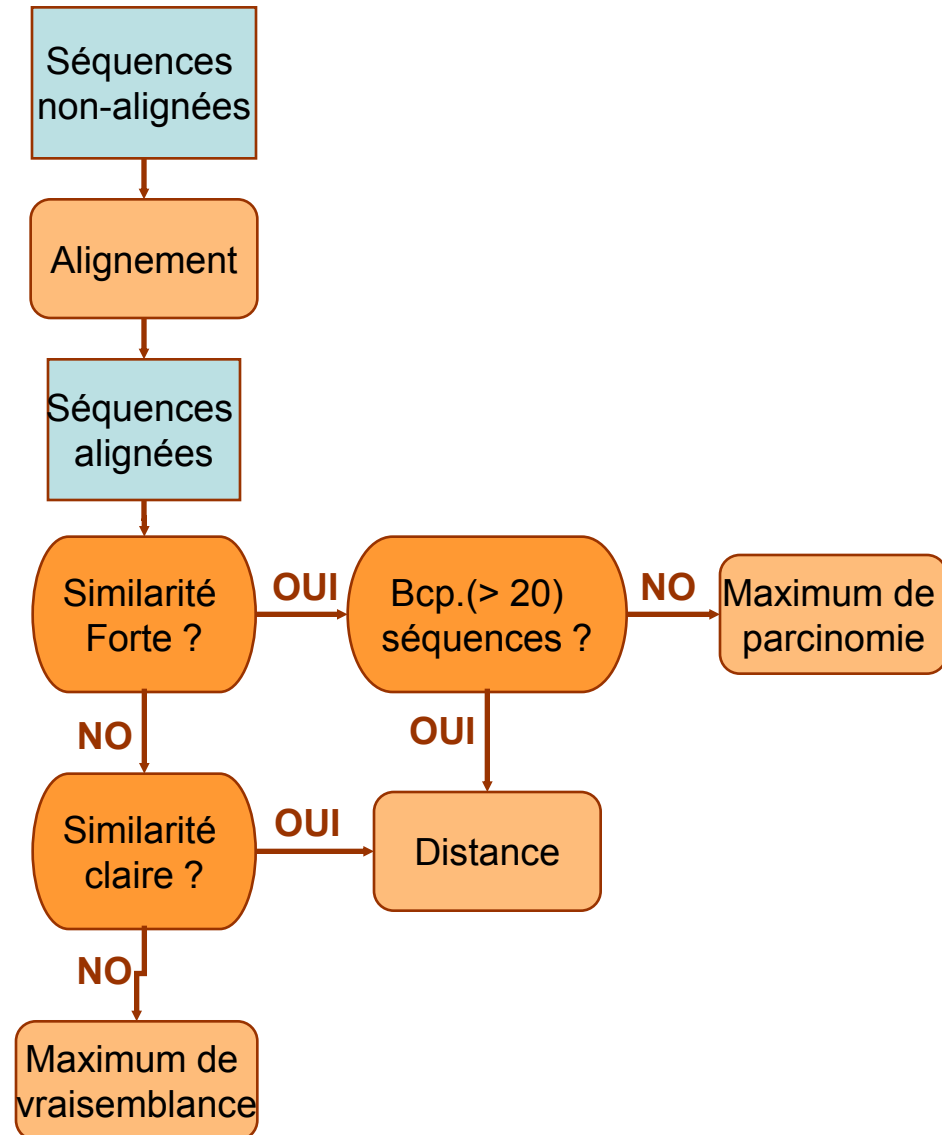


Comparaison UPGMA - NJ



Choix des méthodes de construction des arbres phylogénétiques

- Approches alternatives
 - Maximum de parcimonie
 - Distance
 - Méthodes statistiques

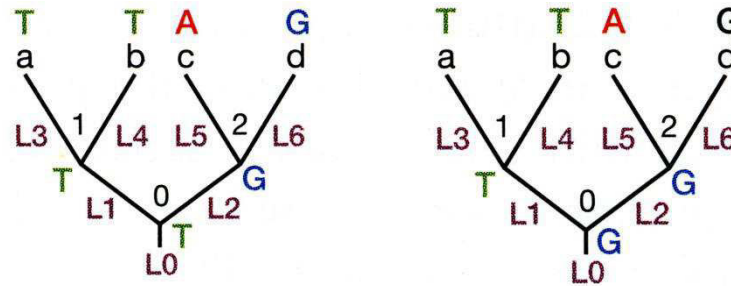


Maximum de vraisemblance

- Comme le méthode de maximum de parcimonie
 - Analyse chaque colonne de l'alignement
 - Analyse chaque arbre non-enraciné
- Pour chaque arbre et chaque colonne (=site = caractère)
 - Examine tout les combinaison des états de caractères pour chaque nœud interne

A. Sequences

	*
sequence a	A C G C G T T G G G
sequence b	A C G C G T T G G G
sequence c	A C G C A A T G A A
sequence d	A C A C A G G G A A



- Calcule de vraisemblance pour chaque combinaison des états des caractères
 $L(\text{site1}, \text{Arbre1}, \text{comb1}) = L_0 \times L_1 \times L_2 \times L_3 \times L_4 \times L_5 \times L_6$
 - Vraisemblance d'un arbre pour un caractère (colonne) est la somme de vraisemblance de chaque combinaison des états des caractères
 $L(\text{site1}, \text{Arbre1}) = L(\text{comb1}) + L(\text{comb2}) + \dots L(\text{comb64})$
- Le vraisemblance globale est calculée pour chaque arbre en ajoutant de vraisemblance de l'arbre en question à chaque site.
 $L(\text{Arbre1}) = L(\text{site1}) + L(\text{site2}) + \dots L(\text{siteN})$

Maximum de vraisemblance

- Avantages:
 - La vitesse d'évolution peut varier
 - Entre lignées (branches)
 - Entre sites
 - Peut utiliser différents modèles d'évolution (Jukes-Cantor, Kimura à deux)
 - Peut traiter les séquences avec des différentes compositions de bases
- Désavantages:
 - Temps de calcul longs => modification pour faire une approche heuristique
 - Arbre non-enraciné

Estimation de la robustesse des arbres Bootstrap

- En phylogénie, un arbre est un estimateur des données dont on dispose
 - Idée = estimer la *variabilité de l'arbre* (ou d'une partie de l'arbre = branches) en changeant les caractères
 - Si un arbre est robuste *i.e.* fortement soutenu par les données alors sa variabilité sera faible
 - Si un arbre est peu robuste alors il aura une grande variabilité

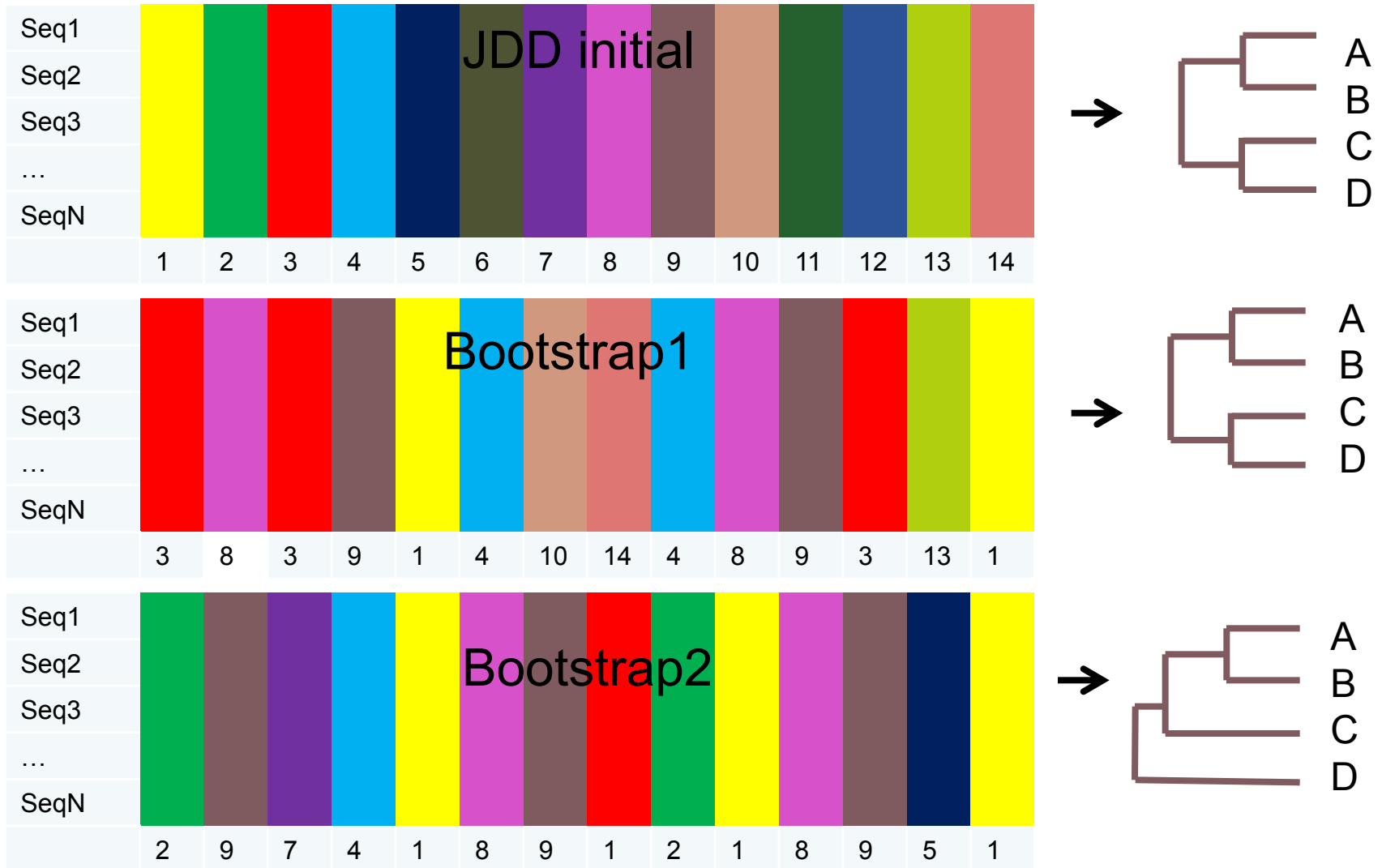
Le Bootstrap

- Principe

On estime les phylogénies obtenues à partir d'un certain nombre de ré-échantillonnages de même taille que notre jeu de données initial

- On réalise X tirages **avec remise** de n caractères parmi n caractère au sein du JDD initial
- Construction d'une nouvelle matrice de caractères de même taille (nombre de séquences et de sites) que le JDD initial
- Pour chaque tirage on calcule la phylogénie correspondante par la même méthode
- Pour chaque nœud, comptage des nombres des simulations où le nœud est soutenu.

Bootstrap



Interprétation du Bootstrap

Une valeur de bootstrap de 100% \neq un nœud vrai

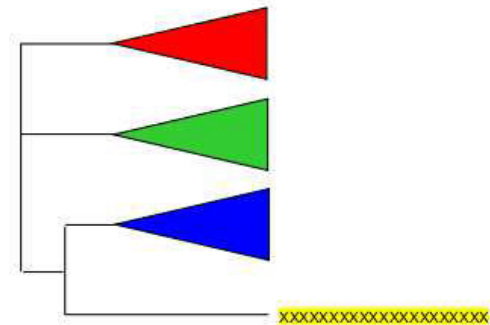
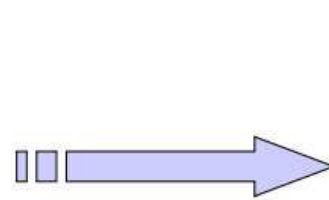
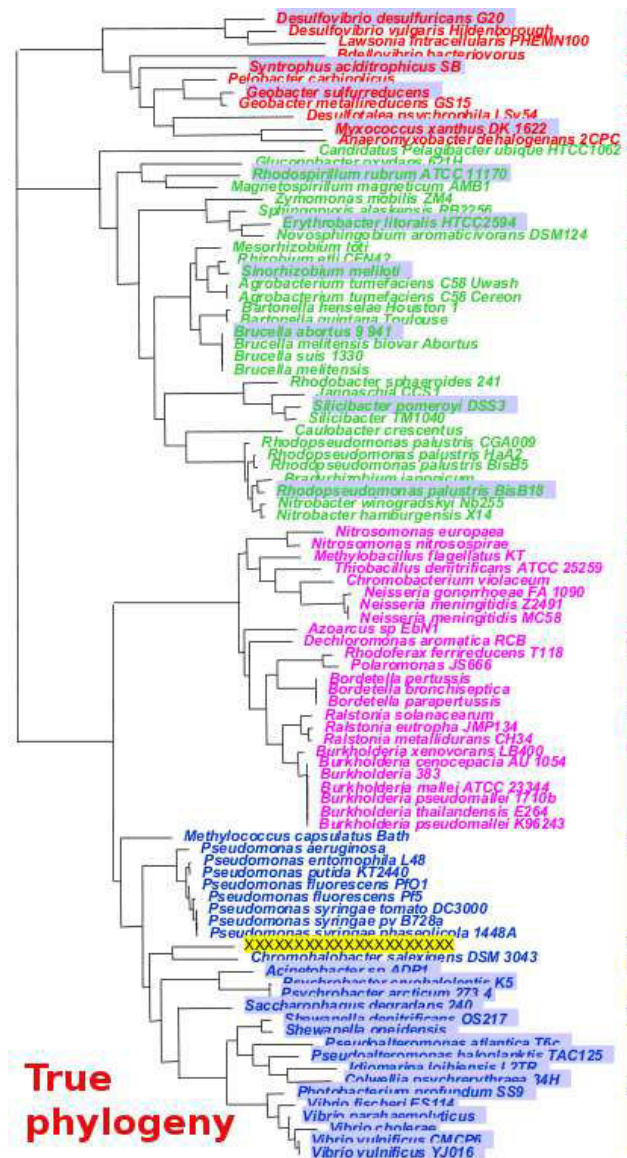
Une valeur de bootstrap de 100% = un nœud ROBUSTE

Robustesse : Les données soutiennent fort le nœud

Causes de l'incongruence/problèmes rencontrés en phylogénie moléculaire

- Problèmes d'échantillonnages
 - Séquences trop courtes => effets stochastiques
 - Échantillonnage taxonomique trop réduit
- Problèmes liés à la divergence des séquences
 - Séquences pas assez variables
 - Séquences trop divergentes => saturation
 - Séquences présentant des taux d'évolution hétérogènes (Attraction des longues branches)
- Transferts horizontaux

Causes de l'incongruence/problèmes rencontrés en phylogénie moléculaire



Inferred phylogeny

Échantillonnage
taxonomique trop réduit

Remarque:

L'arbre guide n'est pas un arbre phylogénétique

- L'alignement multiple progressif dépend de l'arbre guide
 - L'arbre guide est basée sur les **alignements par paires**
 - C'est une approximation de distance entre les paires des séquences et n'est pas la distance évolutive
- L'arbre phylogénétique est construite sur base de **l'alignement multiple**
 - L'arbre tente de décrire les distances évolutives entre les séquences

Bibliographie

- W. Mount. Bioinformatics: Sequence and Genome Analysis. (2004) pp. 692. <http://www.bioinformaticsonline.org/> (Code BU: 572.86 MOU)
- Perrière et Brochier-Armanet: Concepts et méthodes en phylogénie moléculaire, 2010, Springer (BU:570.11 PER)

Cours basée sur les cours de Céline Brochier-Armanet et Jacques van Helden