

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider-BISKRA.

Faculté des sciences exactes et des
sciences de la nature et de la vie



Département d'informatique

Calcul à l'ADN et Bio-informatique

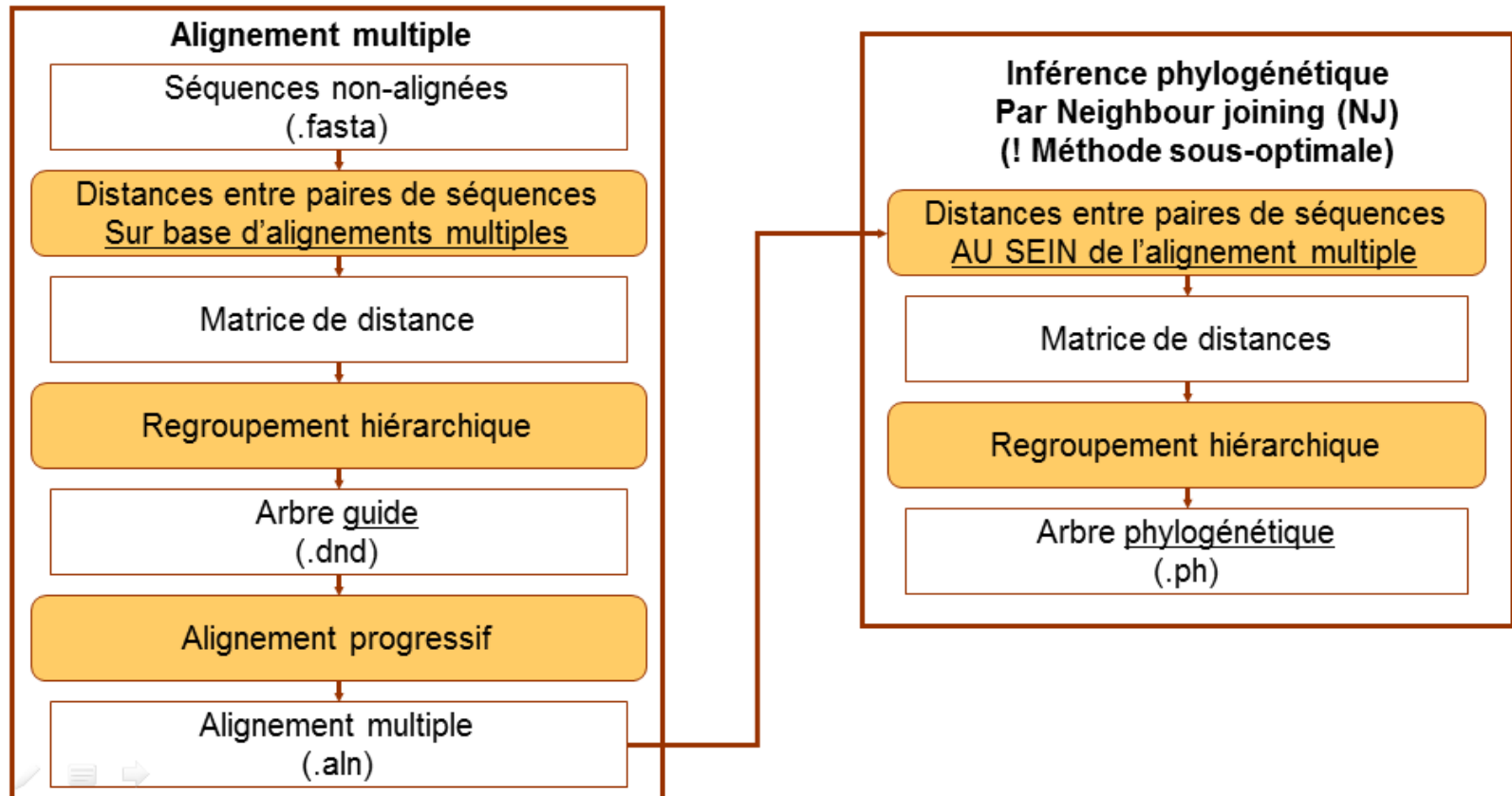
Réalisé par :

Dr. TBERMACINE Ahmed

CHAPITRE 4: PHYLOGÉNIE

2

INTRODUCTION



INTRODUCTION

- Nous plongerons dans l'art complexe de tracer les liens évolutifs entre les organismes à l'aide d'outils informatiques avancés.
- La phylogénie, telle une carte génétique du vivant, nous permet de comprendre les relations évolutives entre les espèces.
- Ce chapitre fournira un aperçu approfondi des méthodes et des algorithmes utilisés pour reconstruire des arbres phylogénétiques, mettant en lumière l'intersection entre la biologie et l'informatique

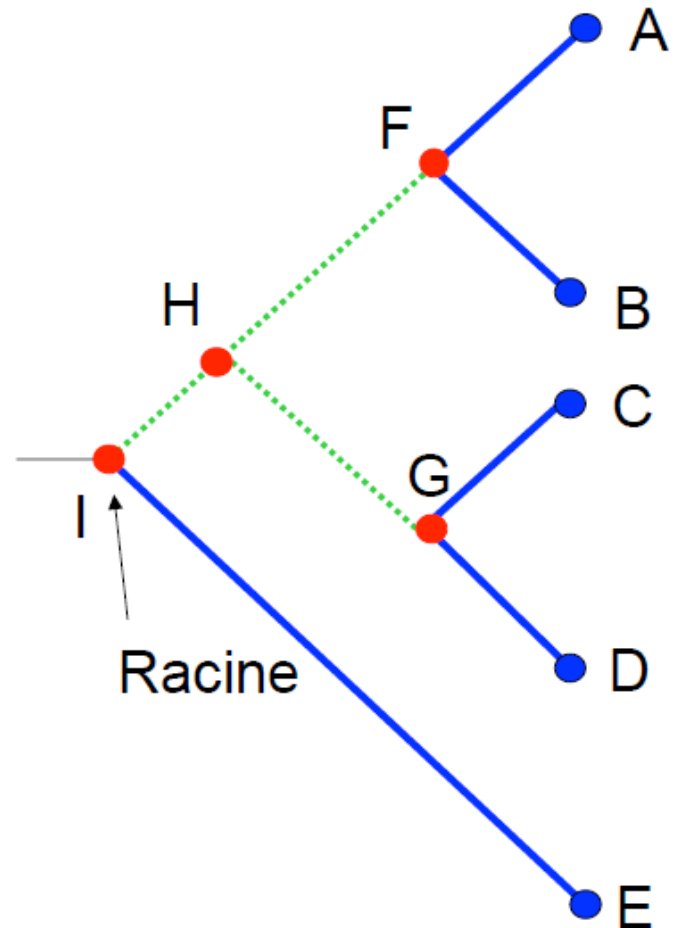
PHYLOGÉNIE

- La phylogénie est l'étude des relations de parentés entre différents organismes vivants en vue de comprendre leur évolution.
- Plus particulièrement, la phylogénie moléculaire renseigne sur les changements survenus au niveau des séquences biologiques (nucléiques et protéiques) au cours du temps « évolution ».

ARBRES PHYLOGÉNÉTIQUES

• Structure des arbres phylogénétiques

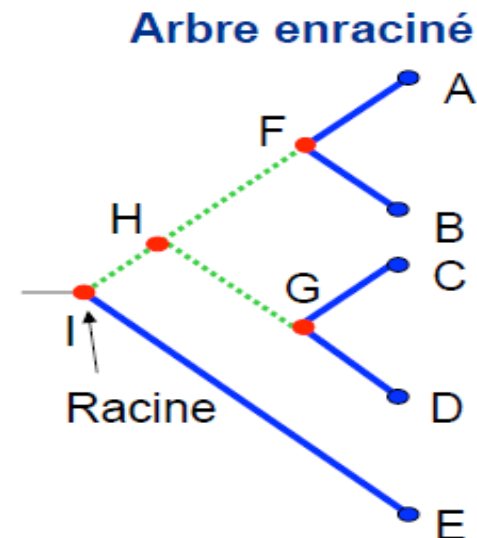
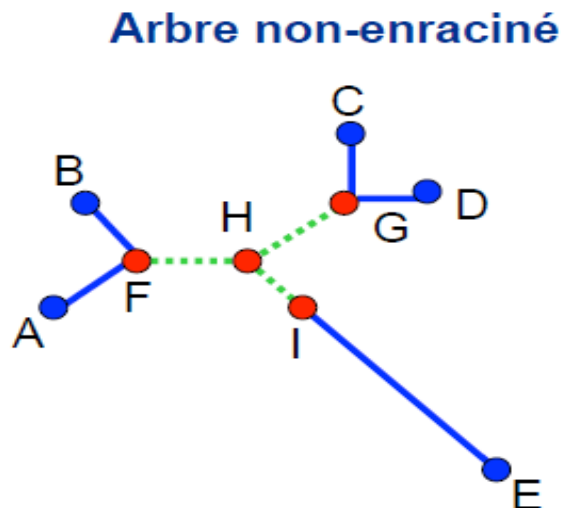
- Les relations évolutives entre les objets étudiés (espèces, organes, séquences) sont représentées par des arbres phylogénétiques
- Les arbres sont des graphes composés de *noeuds* et de *branches*
 - Noeuds = unités taxonomiques
 - Feuilles ou **OTU = Unités Taxonomiques Opérationnelles** (A, B, C, D, E), espèces pour laquelle on dispose d'échantillons (en phylogénie moléculaire, les échantillons consistent en séquences macromoléculaires).
 - Noeuds internes ou **HTU = Unités taxonomiques Hypothétiques** (F, G, H, I), correspondent aux espèces ancestrales, inférées à partir des OTU.
 - Branches = relations de parenté(ancêtre/descendants) entre unités taxonomiques
 - Branches internes
 - Branches externes
- On appelle **topologie** l'ensemble des branchements de l'arbre.



ARBRES PHYLOGÉNÉNIQUES

❖ Arbres enracinés ou non enracinés

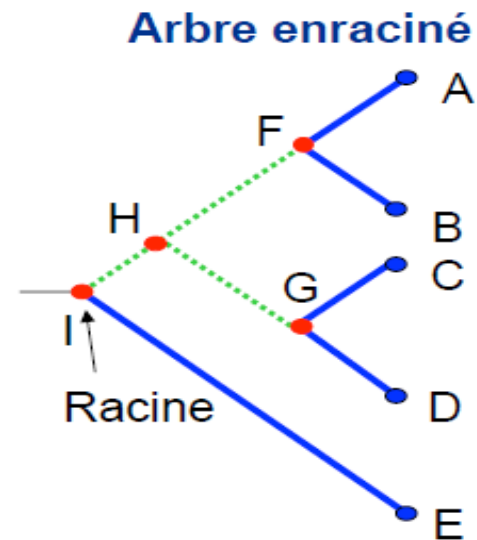
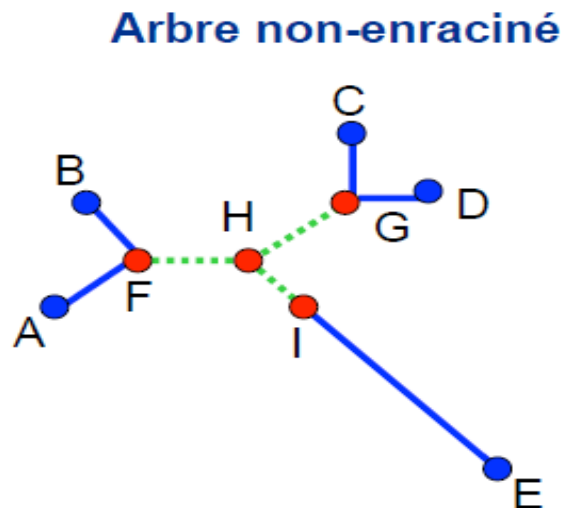
- Les Arbres enracinés illustrent les relations évolutives entre les taxons qui partagent un ancêtre commun.
- La **racine** représente l'ancêtre le plus récent partagé par tous les taxons de l'arbre. Il est généralement placé à la base de l'arbre et sert de point de référence pour interpréter les modèles de ramification et les relations évolutives entre les taxons.
- La directionnalité de l'évolution et la chronologie relative des événements de divergence sont révélées par des arbres enracinés.



ARBRES PHYLOGÉNIQUES

❖ Arbres enracinés ou non enracinés

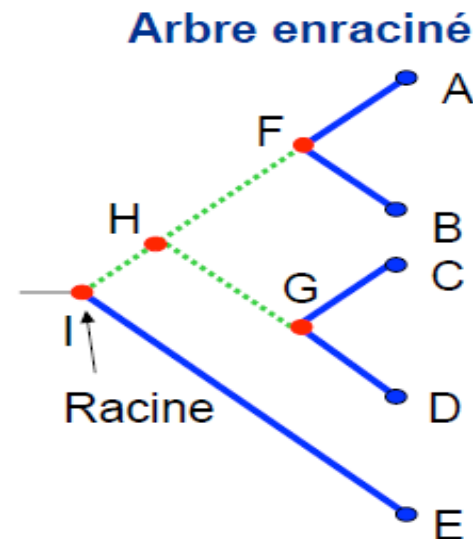
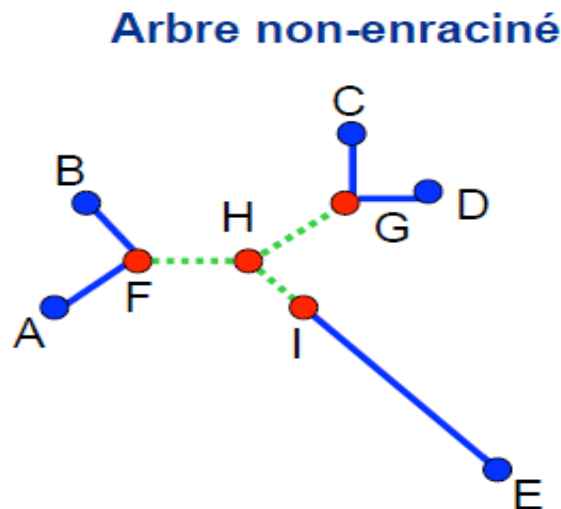
- **Les arbres non racinés** décrivent les relations évolutives entre les taxons qui n'ont pas de racine commune désignée.
- Au lieu d'une racine singulière, les arbres non racinés décrivent les relations taxonomiques comme une série de branches et de nœuds connectés.



ARBRES PHYLOGÉNIQUES

❖ Arbres enracinés ou non enracinés

- Les **arbres non enracinés** ne peuvent pas fournir d'informations sur la direction de l'évolution ou la chronologie relative des événements de divergence car ils n'ont pas de racine.
- Les arbres non racinés sont principalement utilisés pour visualiser les relations entre les taxons, mais ils peuvent également être utilisés pour déduire des modèles évolutifs et identifier des groupes de taxons étroitement liés.



ARBRES PHYLOGÉNIQUES

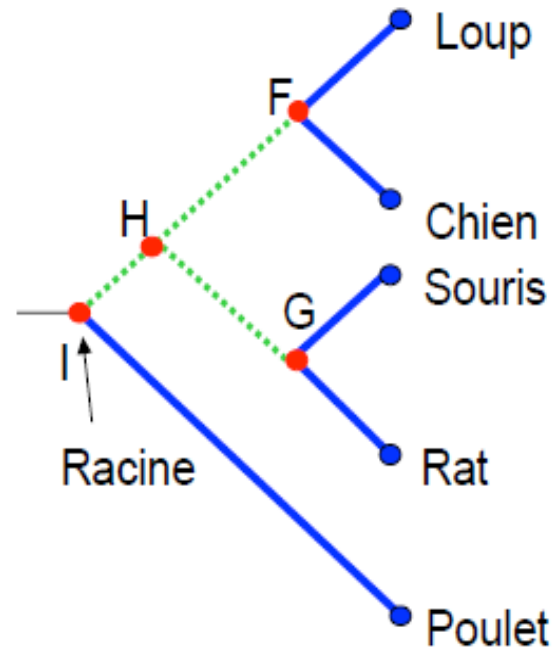
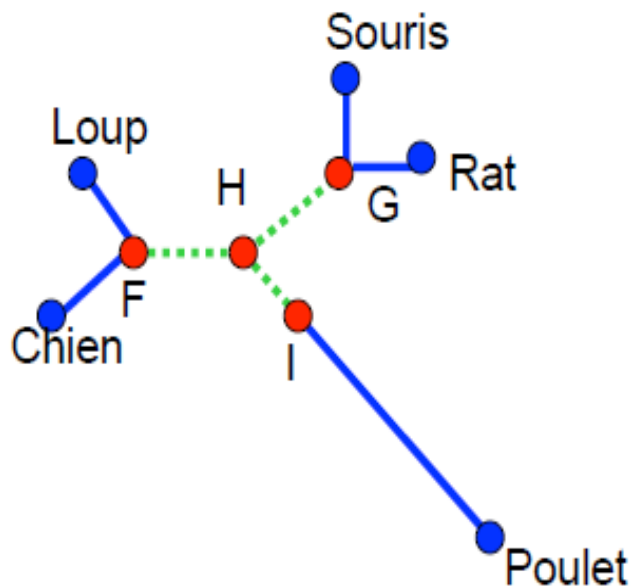
❖ Arbres enracinés ou non enracinés

- Les arbres phylogénétiques enracinés et non racinés offrent des avantages et des applications uniques.
- Lorsque le moment relatif des événements de divergence et la direction de l'évolution sont essentiels, les **arbres enracinés** sont fréquemment utilisés.
- Ils sont particulièrement précieux pour étudier l'histoire de l'évolution et en déduire des états ancestraux.
- En revanche, les **arbres non racinés** sont fréquemment utilisés pour l'analyse exploratoire, l'analyse de regroupement et la visualisation initiale des relations taxonomiques.
- Lorsque l'accent est mis sur les relations relatives entre les taxons par opposition à leur histoire évolutive absolue, ils sont particulièrement utiles.

ARBRES PHYLOGÉNÉTIQUES

❖ Comment enraciner un arbre phylogénétique ?

- Si on connaît à priori l'OTU le plus externe parmi les OTU étudiés, on le place en tant que groupe extérieur, ici par exemple c'est le poulet. Sans cette connaissance, on enracine au poids moyen

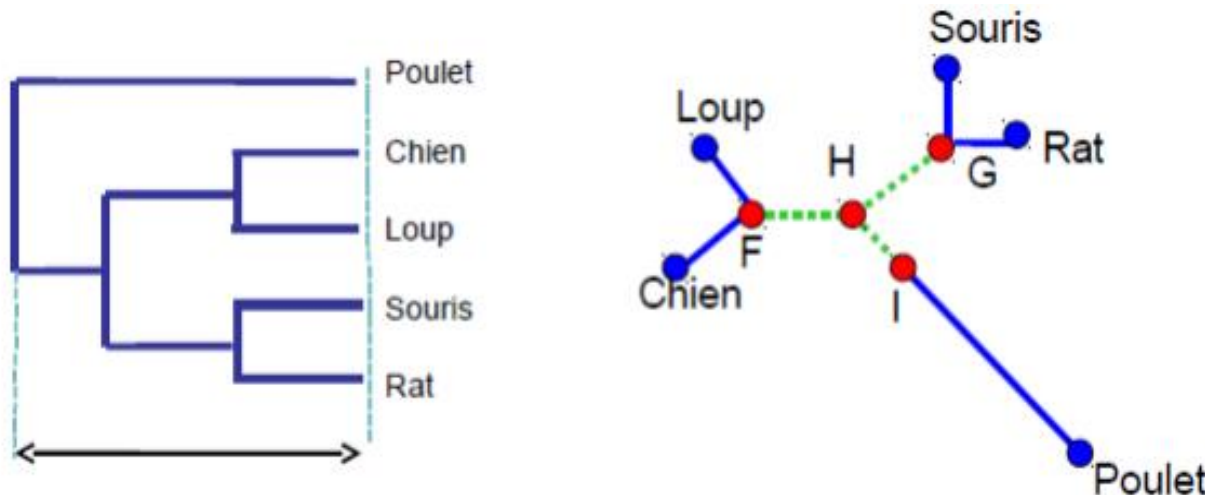


ARBRES PHYLOGÉNÉTIQUES

❖ Comment enraciner un arbre phylogénétique ?

❖ Enracinement au poids moyen des arbres

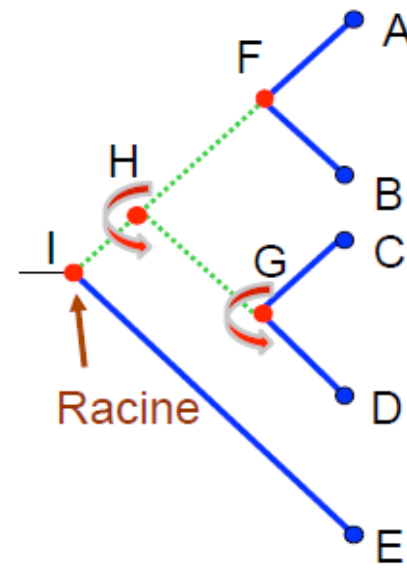
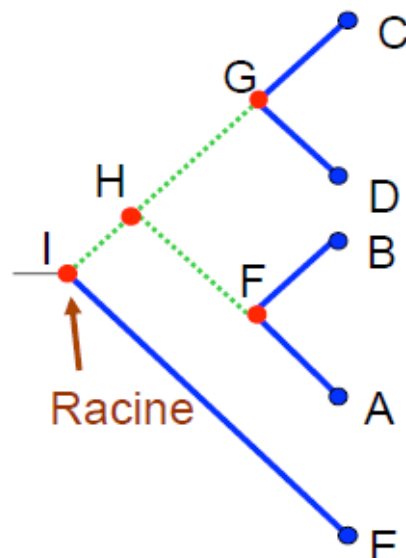
- On émet l'hypothèse que toutes les séquences évoluent à la même vitesse, c'est **l'hypothèse d'horloge moléculaire**.
- La même quantité d'évolution s'est produite dans chaque lignée évolutive depuis leur ancêtre commun à toutes.
- Les distances évolutives entre chaque feuille et la racine sont égales.
- La racine est placée au point de l'arbre équidistant de toutes les feuilles.



ARBRES PHYLOGÉNÉTIQUES

❖ L'isomorphisme des arbres phylogénétiques ?

- Il faut éviter le piège d'évaluer les distance entre feuilles sur base de leur proximité verticale.
- Les structures ci-dessous sont absolument identiques.
- Pourtant les feuilles B et D semblent voisines sur le graphe de gauche, et éloignées sur celui de droite.
- Pour évaluer la distance entre deux nœuds d'un arbre, il faut prendre en compte la longueur totale du chemin le plus court pour les rejoindre (somme des longueurs de branches).

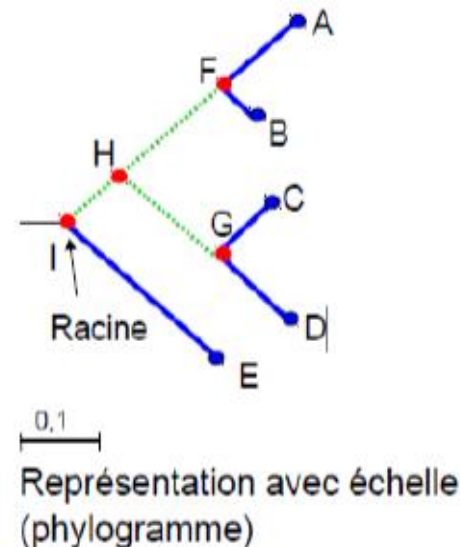
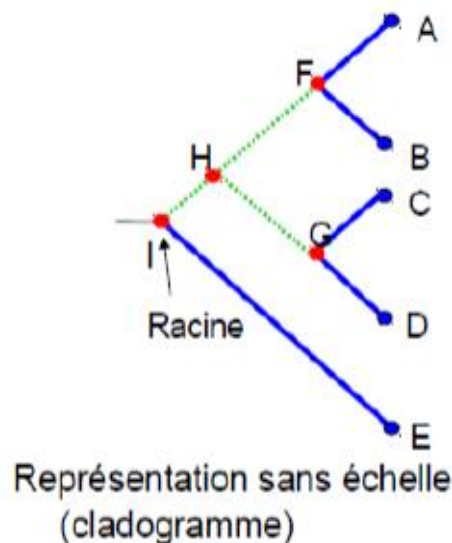


ARBRES PHYLOGÉNIQUES

❖ Différentes représentations graphiques pour les arbres

❑ Cladogramme

- Arbre représentant un scénario évolutif des divergences entre espèces ou séquences.
- les longueurs de branches ne sont pas proportionnelles au nombre de changements évolutifs.
- L'arbre représente uniquement l'ordre des branchements.



ARBRES PHYLOGÉNIQUES

❖ Définitions autour du Cladogramme

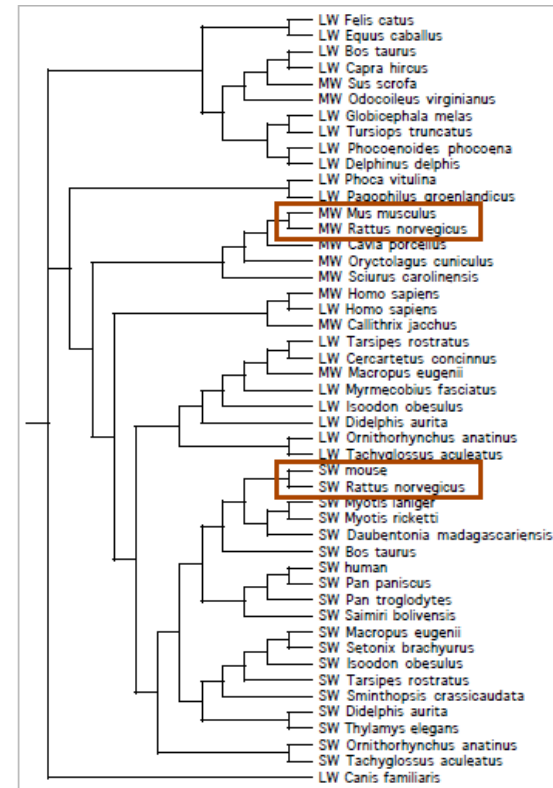
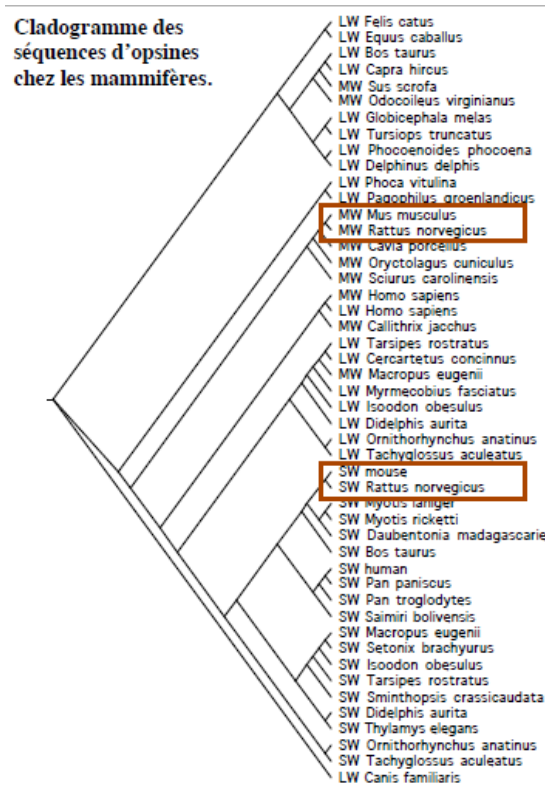
- **Cladistique**: classe les êtres vivants selon leurs relations de parenté, basé sur leurs caractères.
- **Clade**: une branche de cladogramme avec un ancêtre commun et tous ses descendants.
- **Attention**
 - la longueur de la branche ne reflète pas le temps ou le taux de divergence.
 - Seule la topologie est informative, **il n'y a pas d'échelle temporelle**.

ARBRES PHYLOGÉNÉNIQUES

❖ Définitions autour du Cladogramme

❑ Ceci est également un cladogramme

Quoique les branches soient rectangulaires, le dessin ci-contre représente une succession de divergences évolutives, sans considération pour l'échelle de temps ou les distances entre éléments.

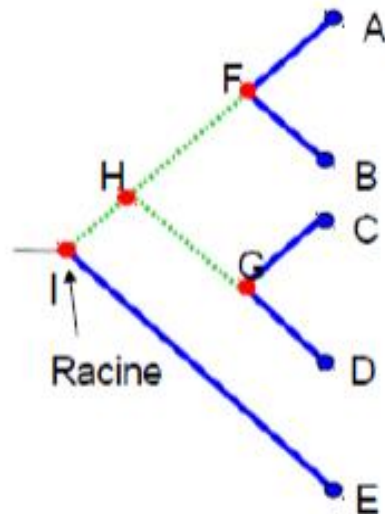


ARBRES PHYLOGÉNIQUES

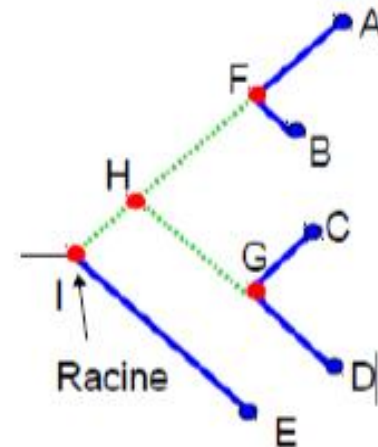
❖ Différentes représentations graphiques pour les arbres

□ Phylogramme:

- Arbre dont les longueurs des branches représentent le nombre d'évènements évolutifs.



Représentation sans échelle
(cladogramme)



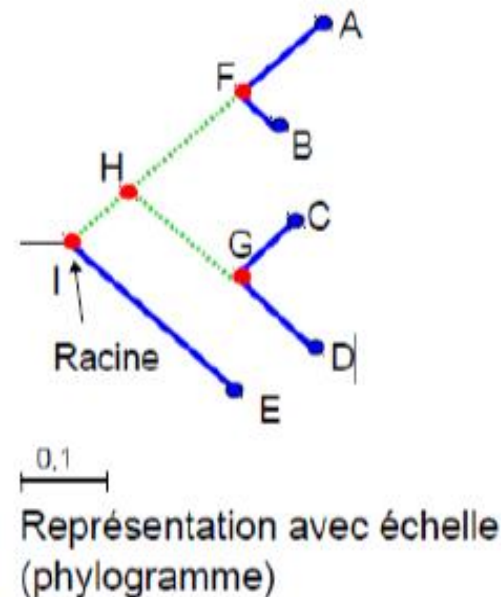
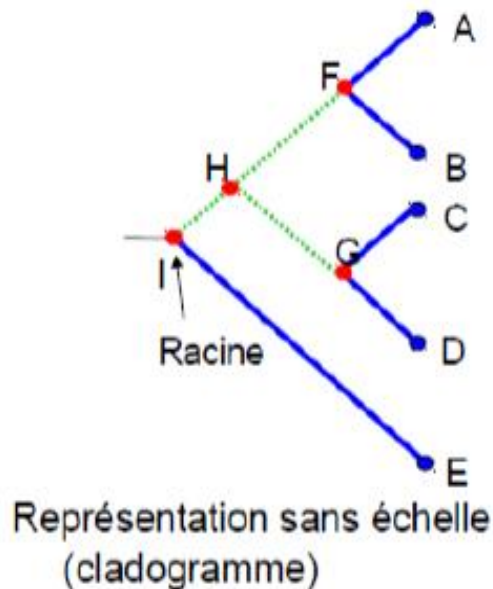
Représentation avec échelle
(phylogramme)

ARBRES PHYLOGÉNÉNIQUES

❖ Différentes représentations graphiques pour les arbres

❑ Phylogramme:

- L'échelle est en nombre de substitutions, ou en nombre de substitutions par sites.
- La distance entre deux nœuds est la somme des branches entre eux. Les longueurs des branches sont seulement des approximations des séquences inférées.



ARBRES PHYLOGÉNIQUES

❖ Différentes représentations graphiques pour les arbres

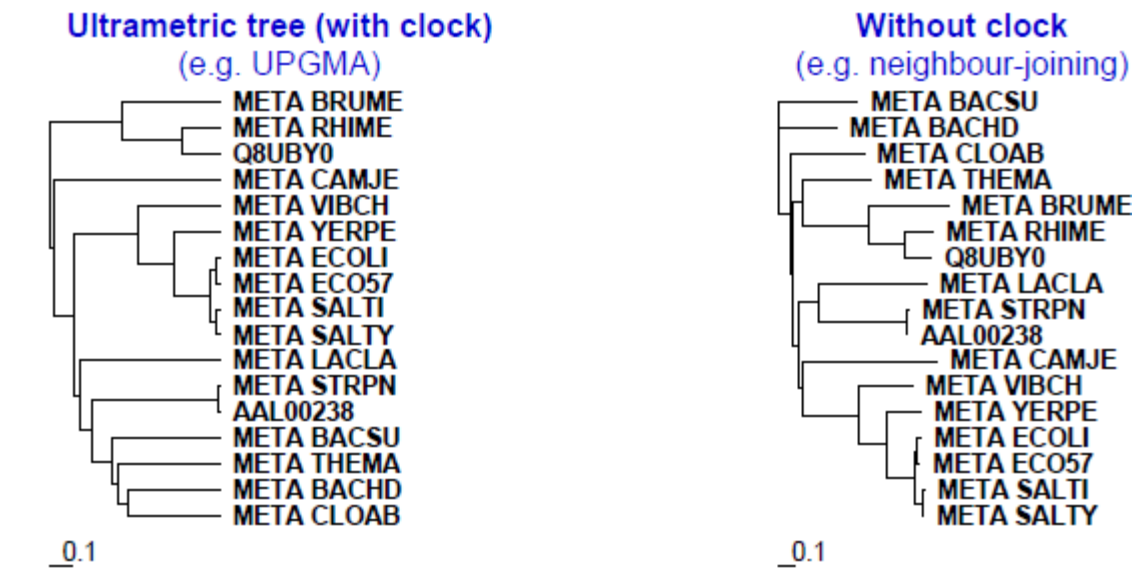
❑ Chronogramme:

- La longueur des branches représente le temps de divergence.
- L'hypothèse de l'horloge moléculaire suppose que le taux d'évolution ne varie pas entre les branches.
- Tous les OTU sont alignés verticalement sur l'arbre.
- L'horloge moléculaire n'est pas toujours valide
- Par exemple, les paralogues peuvent avoir un taux de mutation différent car ils ne sont pas soumis à la même pression de sélection.

ARBRES PHYLOGÉNIQUES

❖ Chronogramme versus Phylogramme

- L'arbre de droite est un Phylogramme: Les OTU ne sont pas alignées, car les différentes branches sont associées à des taux variables de mutations.
- L'arbre de gauche est un chronogramme: Longueurs de branches proportionnelles au temps



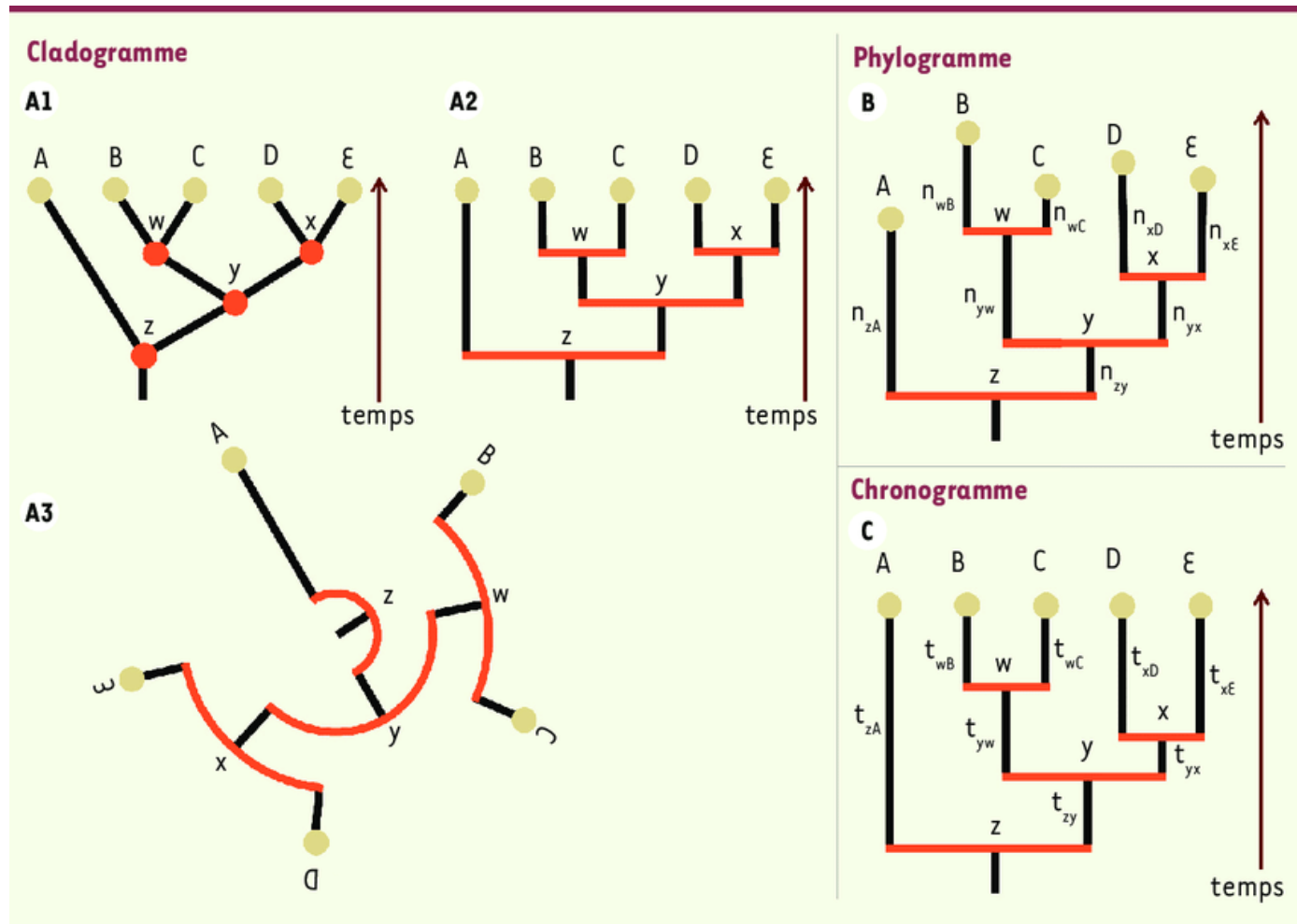
ARBRES PHYLOGÉNIQUES

❖ L'hypothèse de l'horloge moléculaire

- L'hypothèse de l'horloge moléculaire = les taux d'évolution ne varient pas entre branches.
- Cette hypothèse n'est généralement pas valide:
 - ✓ Dans certains cas, deux gènes divergent d'un ancêtre commun, mais l'un diverge plus vite que l'autre. Cette dernière situation est d'ailleurs plutôt commune: une duplication crée une certaine redondance, et l'une des copies peut évoluer plus vite que l'autre, qui conserve la fonction initiale (les mutations sont produites sur les deux copies, mais sont contre-sélectionnées sur l'une des deux).
 - ✓ Par ailleurs, pour un même gène, le taux de mutations peut varier selon les époques.

ARBRES PHYLOGÉNÉTIQUES

❖ Différentes représentations graphiques pour les arbres



ARBRES PHYLOGÉNÉNIQUES

- **L'inférence d'un arbre phylogénétique à partir d'un alignement multiple**
- Nombre d'arbres possibles
- Le nombre d'arbres possibles augmente de façon vertigineuse en fonction du nombre d'éléments terminaux (qu'ils représentent des molécules ou des espèces).
- Nombre d'arbres enracinés : $N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$
- Nombre d'arbres non enracinés : $N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$
- Le nombre d'arbres non enracinés pour n OTU est égal au nombre d'arbres enracinés pour $n-1$ OTU.
- Parmi tous les arbres possibles, un seul représente la véritable histoire évolutive (C'est l'arbre vrai). Puisqu'on ne dispose pas a priori de cet arbre, on doit l'inférer à partir des éléments actuels (les unités taxonomiques opérationnelles, UTO). Le(s) arbre(s) obtenu(s) sont appelés arbres inférés.

ARBRES PHYLOGÉNÉNIQUES

❖ L'inférence d'un arbre phylogénétique à partir d'un alignement multiple

n	Nb arbres enracinés	Nb arbres non-enracinés
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

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

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

ARBRES PHYLOGÉNIQUES

○ L'inférence d'un arbre phylogénétique à partir d'un alignement multiple

▣ Remarque

- Un caractère est une caractéristique observable d'un organisme (quantitative ou qualitative).
- Un état de caractère est une forme particulière d'un caractère dans une OTU particulière (variable continue ou discrète).

○ Exemples

- Caractère: Taille de la patte postérieure gauche. Etat du caractère: 1,68 cm.
- Caractère: acide aminé à la position 68 du produit du gène CYTB. Etat de ce caractère: alanine.

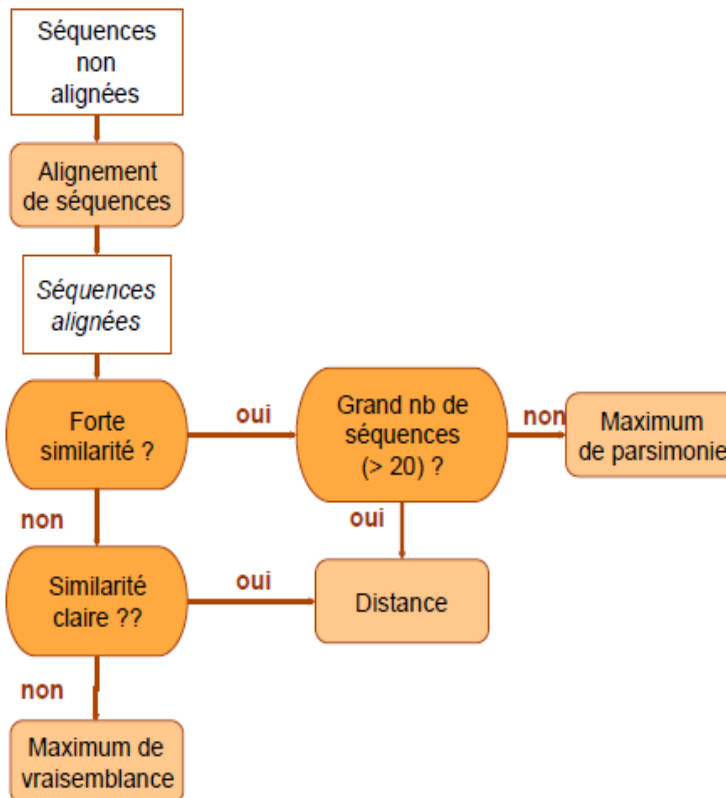
ARBRES PHYLOGÉNIQUES

- ❖ **L'inférence d'un arbre phylogénétique à partir d'un alignement multiple**
- ❖ **Méthodes d'inférence phylogénique**
 - Après avoir construit l'alignement multiple, on cherche à déterminer l'arbre phylogénétique correspondant en utilisant l'une des méthodes suivantes:
 - a) **Méthode cladistique:** Basée sur l'étude des états de caractère (Maximum de parcimonie)
 - b) **Méthode de distance:** Basée sur des mesures de distances (UGPMA, NJ, minimum d'évolution, moindres carrés...)
 - c) **Méthode statistique:** Basée sur l'étude des états de caractères et sur des distances (Maximum de vraisemblance, Méthodes bayésiennes).

ARBRES PHYLOGÉNÉTIQUES

- ❖ L'inférence d'un arbre phylogénétique à partir d'un alignement multiple
- ❖ Méthodes d'inférence phylogénétique

- Approches alternatives
 - Maximum de parcimonie
 - Distance
 - Maximum de vraisemblance



ARBRES PHYLOGÉNIQUES

❖ Méthodes d'inférence phylogénique

❖ **Maximum de parcimonie**

- Approche de construction d'arbres phylogénétiques qui cherche à trouver la topologie d'arbre nécessitant le moins de changements évolutifs pour expliquer les données observées.
- Cette méthode peut être appliquée à la fois aux arbres racinés et non racinés

ARBRES PHYLOGÉNIQUES

❖ Méthodes d'inférence phylogénique

❖ Maximum de parcimonie

○ Arbres enracinés:

L'objectif de la parcimonie maximale est de trouver l'arrangement des taxonomies qui minimise le nombre total de changements évolutifs nécessaires le long des branches.

○ Arbres Non enracinés:

La parcimonie maximale peut également être appliquée aux arbres non racinés, cherchant la topologie d'arbre nécessitant le moins de changements sans point de départ spécifique.

ARBRES PHYLOGÉNIQUES

❖ Méthodes d'inférence phylogénique

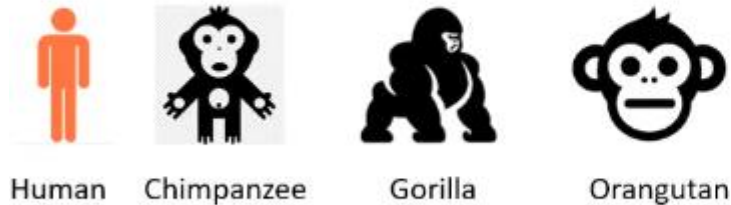
❖ Maximum de parcimonie

○ Note

- Caractère **invariant**: toutes les OTU possèdent le même état de caractères pour un site donné
- Caractère **variable**:
 - 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.

ARBRES PHYLOGÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple

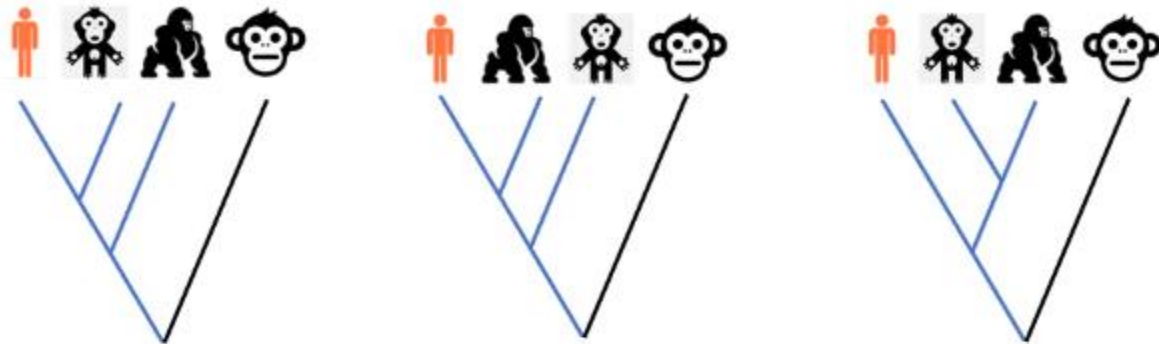


Gene X

	T	T	A	G	C	T	A	C	T
	C	T	A	G	C	T	C	C	C
	C	T	G	G	C	C	A	C	T
	C	T	G	G	A	C	C	C	T

ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple

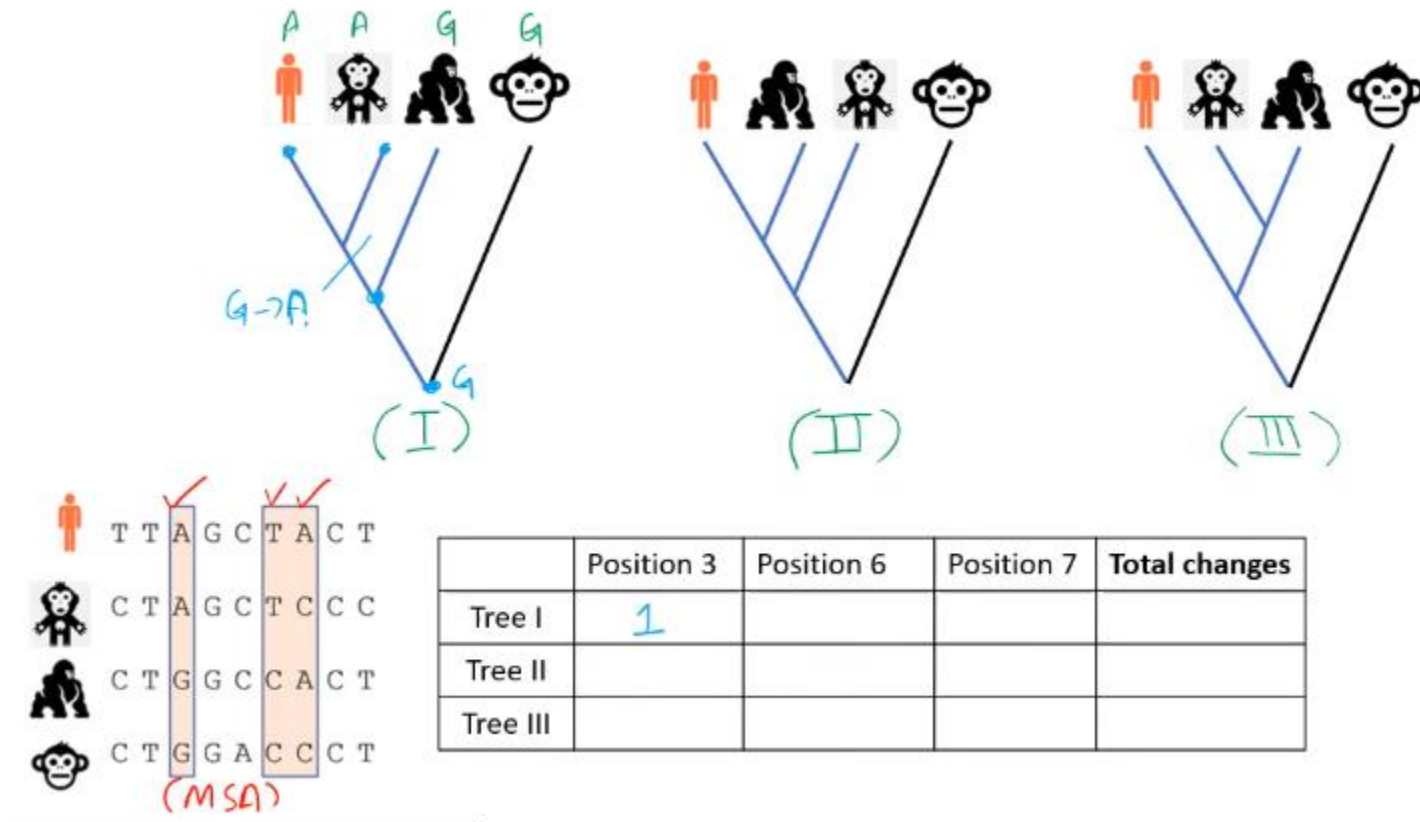


	T	T	A	G	C	T	A	C	T
	C	T	A	G	C	T	C	C	C
	C	T	G	G	C	C	A	C	T
	C	T	G	G	A	C	C	C	T

	Position 3	Position 6	Position 7	Total changes
Tree I				
Tree II				
Tree III				

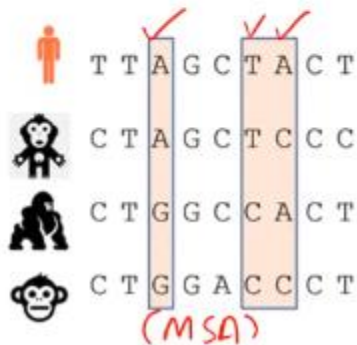
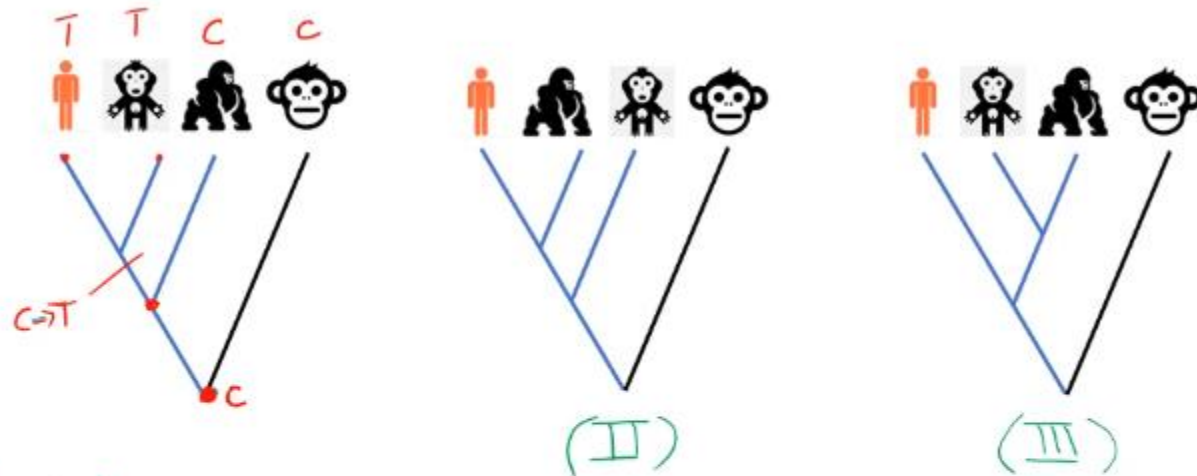
ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



ARBRES PHYLOGÉNÉNIQUES

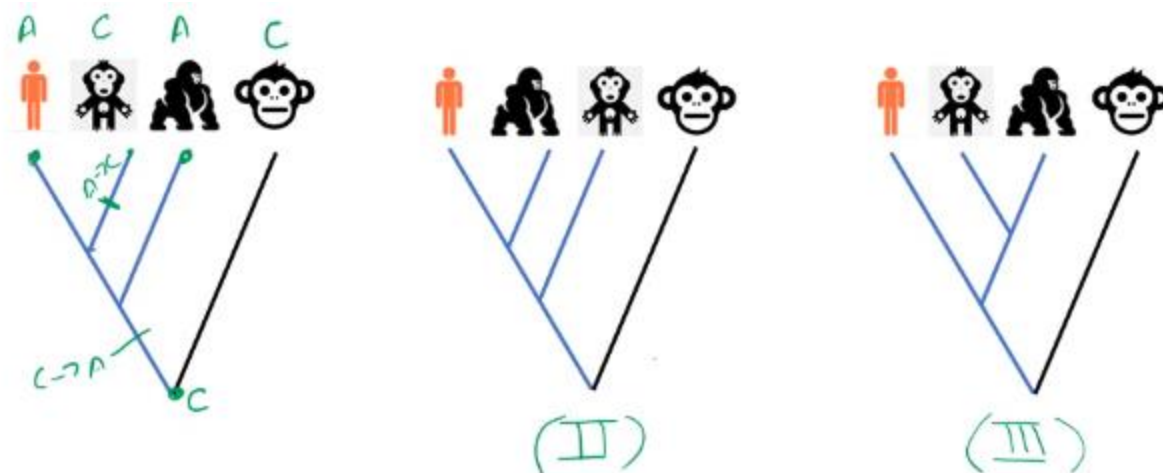
- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



	Position 3	Position 6	Position 7	Total changes
Tree I	1	1		
Tree II				
Tree III				

ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



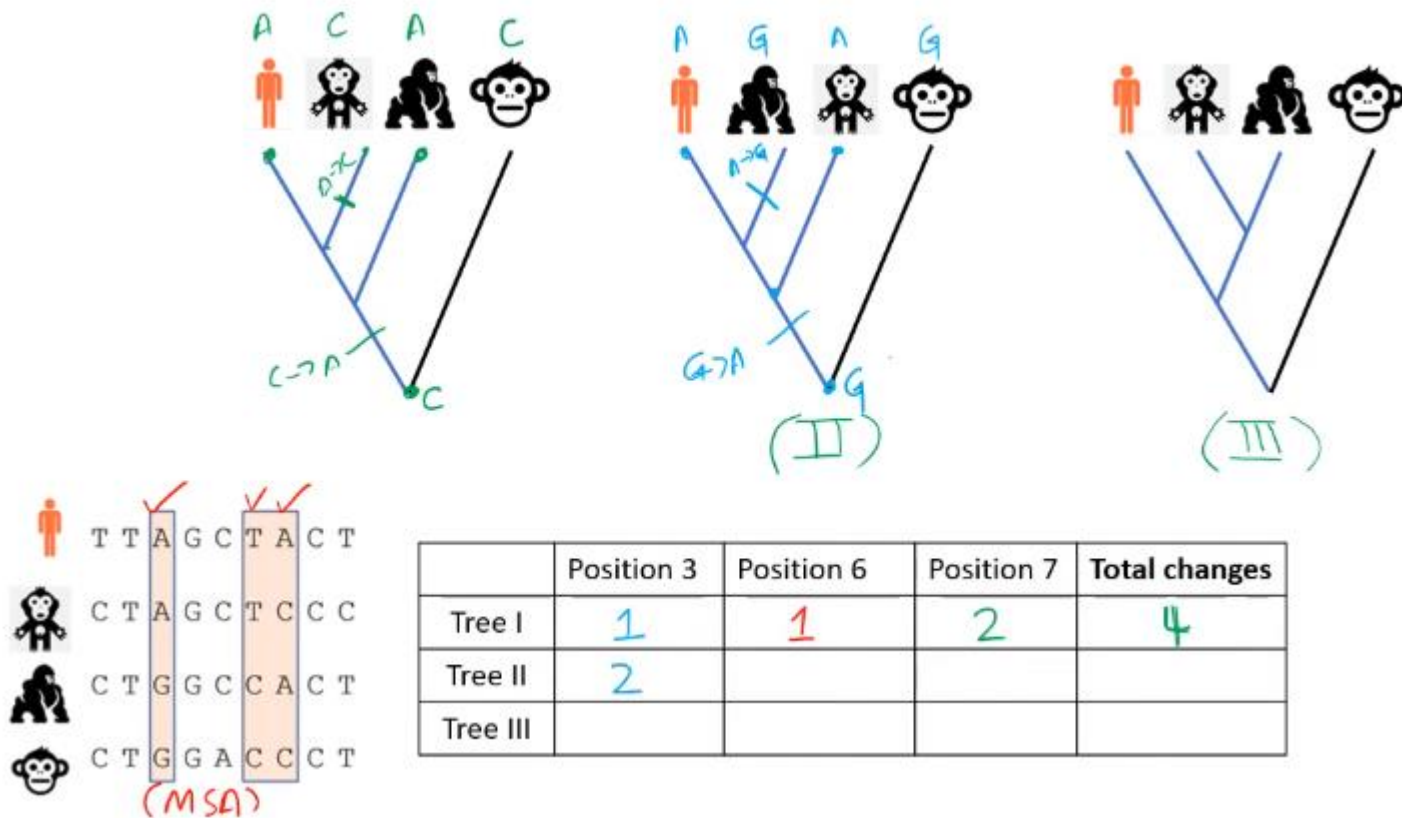
Human	T	T	A	G	C	T	A	C	T
Chimpanzee	C	T	A	G	C	T	C	C	C
Gorilla	C	T	G	G	C	C	A	C	T
Orangutan	C	T	G	G	A	C	C	C	T

(MSA)

	Position 3	Position 6	Position 7	Total changes
Tree I	1	1	2	4
Tree II				
Tree III				

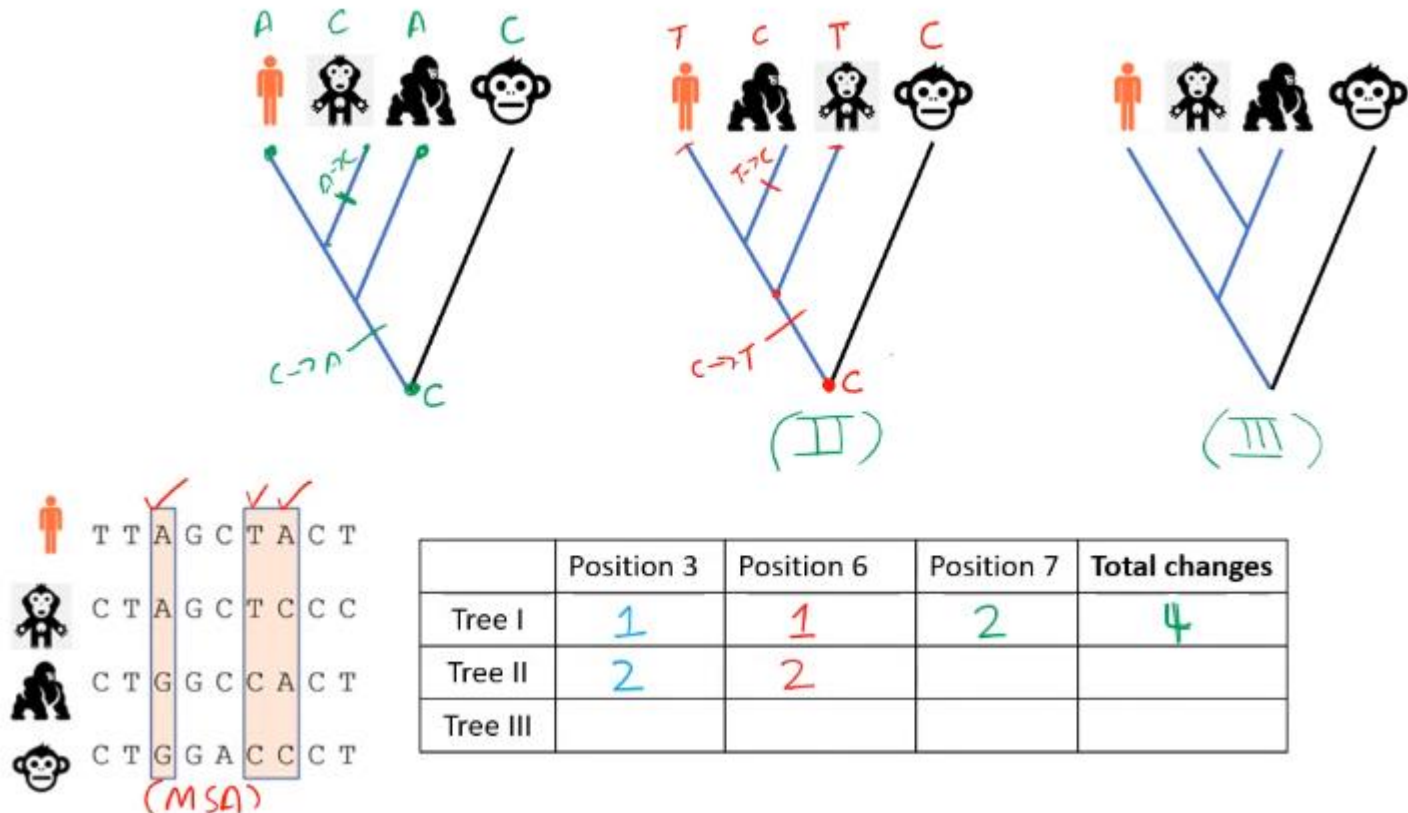
ARBRES PHYLOGÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



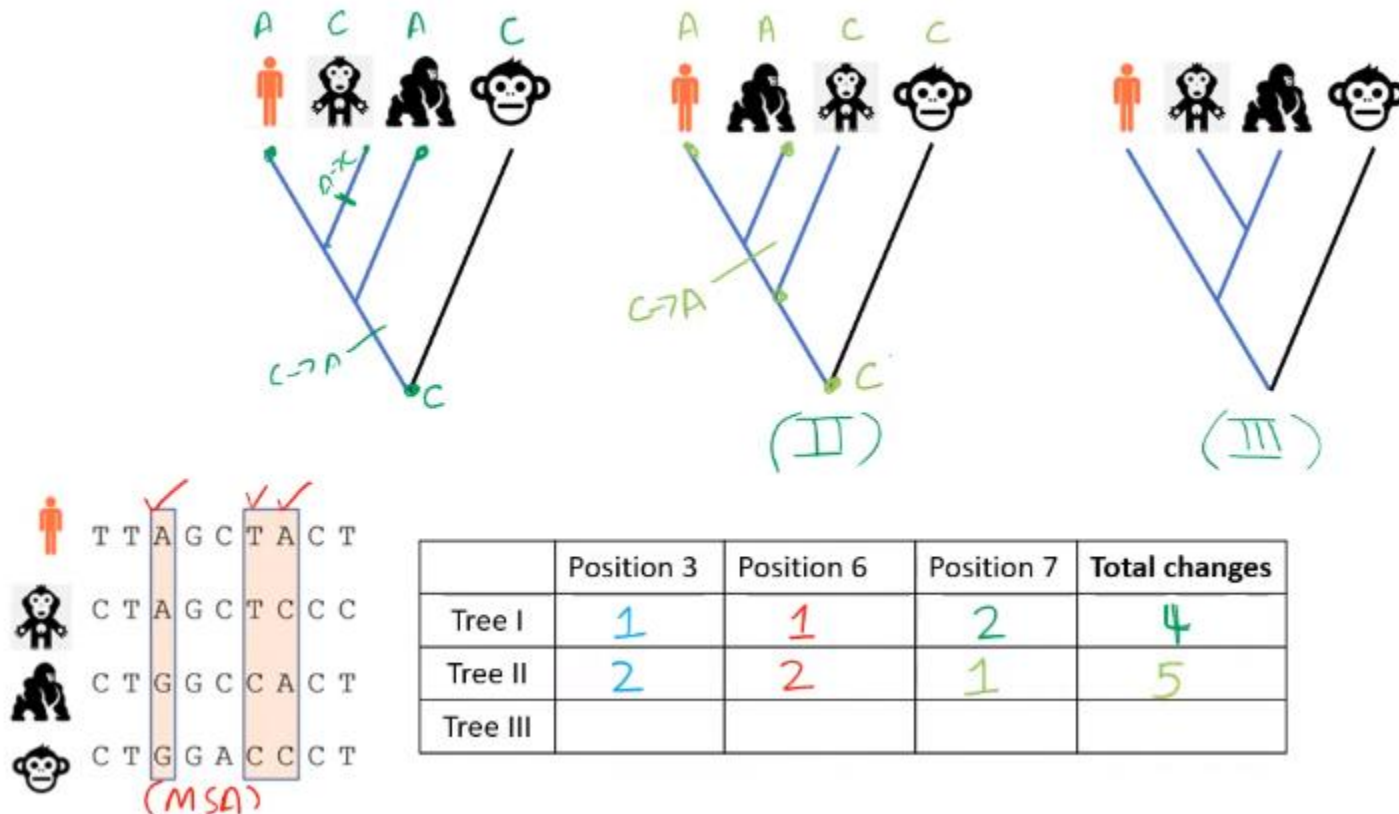
ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



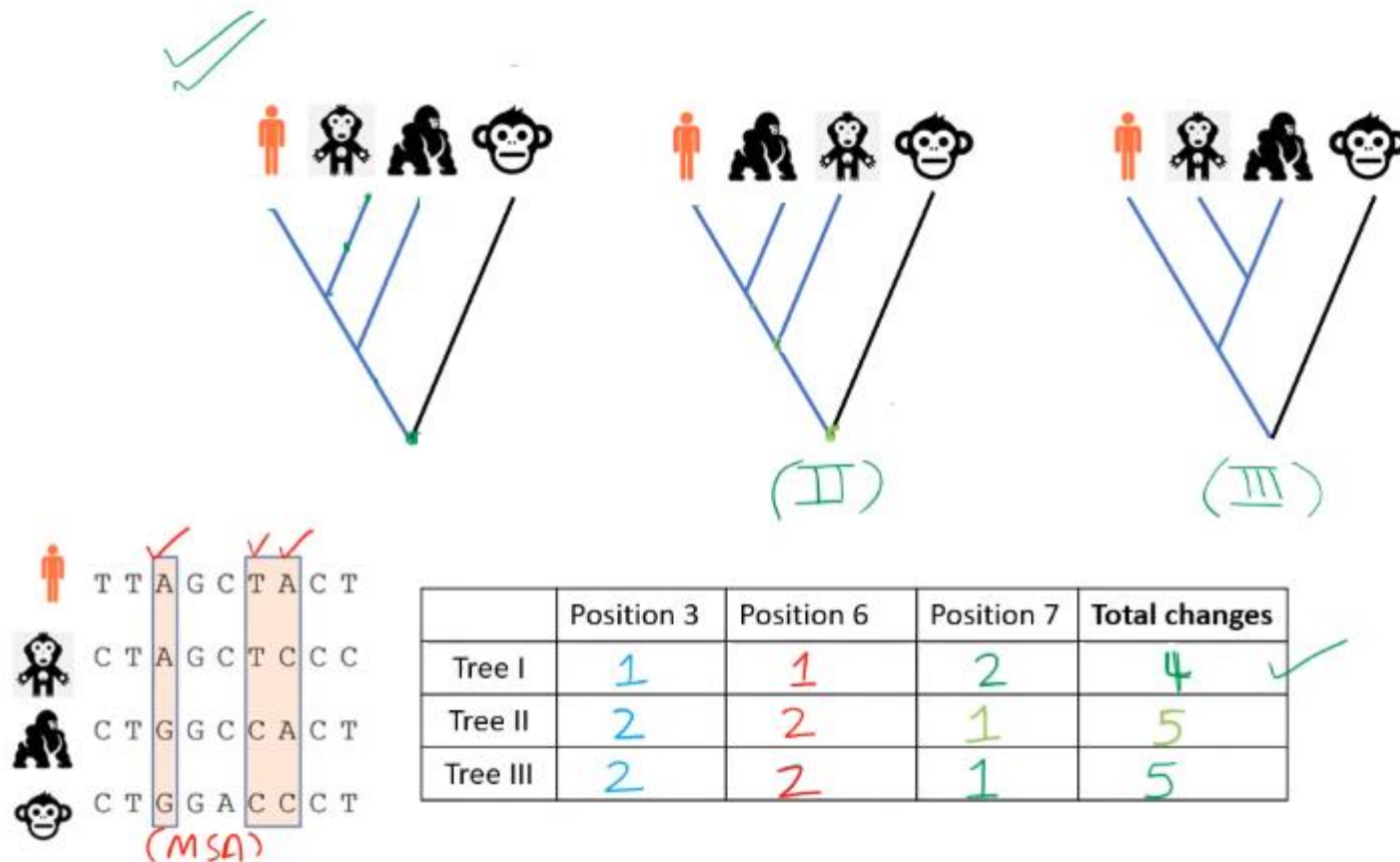
ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



ARBRES PHYLOGÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres enracinés)
 - Exemple



ARBRES PHYLOGÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres Non enracinés)
 - Exemple

Finding the most parsimonious tree:

123456789

Seq 1 TTAGCTACT

Seq 2 CTGGCCACT

Seq 3 CTAGCTCCC

Seq 4 CTGGACCCT

	1	2	3	4	5	6	7	8	9	Total
Tree 1										
Tree 2										
Tree 3										

ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres Non enracinés)
 - Exemple

Finding the most parsimonious tree:

123456789

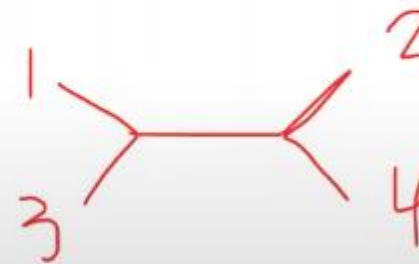
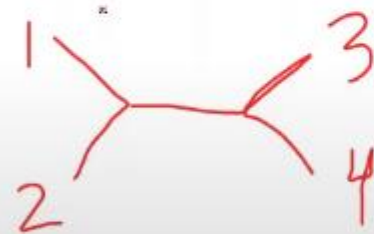
Seq 1 TTAGCTACT

Seq 2 CTGGCCACT

Seq 3 CTAGCTCCC

Seq 4 CTGGACCCT

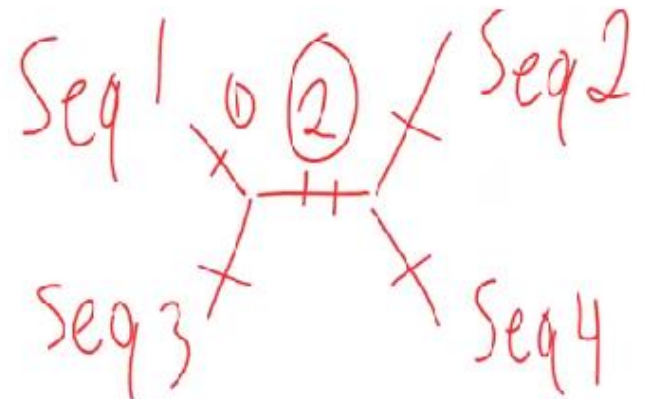
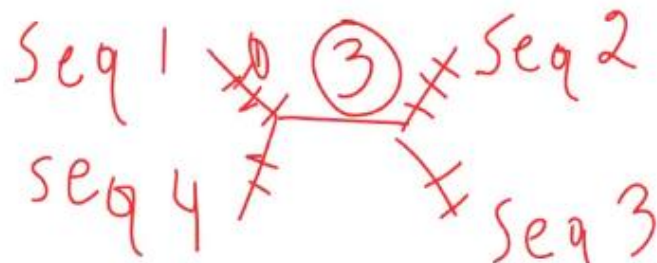
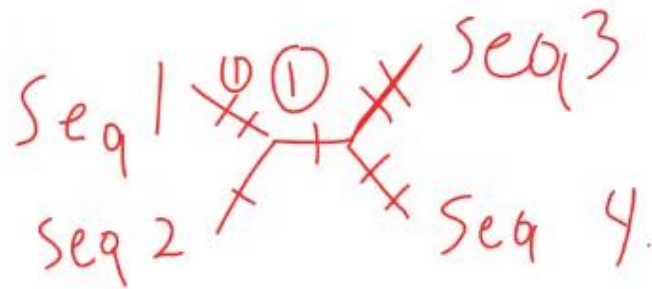
	1	2	3	4	5	6	7	8	9	Total
Tree 1										
Tree 2										
Tree 3										



ARBRES PHYLOGÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ **Maximum de parcimonie** (arbres Non enracinés)
- Exemple

123456789										
Seq 1 TTAGCTACT										
Seq 2 CTGGCCACT										
Seq 3 CTAGCTCCC										
Seq 4 CTGGACCCT										
	1	2	3	4	5	6	7	8	9	Sum
Tree 1	1	0	2	0	1	2	1	0	1	8
Tree 2	1	0	1	0	1	1	2	0	1	7
Tree 3	1	0	2	0	1	2	2	0	1	9



ARBRES PHYLOGÉNIQUES

❖ Méthodes d'inférence phylogénique

❖ Neighbour-Joining

- Une méthode de classification ascendante pour la création d'arbres phylogénétiques.
- Généralement utilisée pour les arbres à partir de données de séquences d'ADN ou de séquences protéiques,
- Elle nécessite la connaissance de la distance entre chaque paire de taxons pour former l'arbre

1. Calculer la matrice Q en basant sur la matrice de distance

$$Q(i, j) = (n - 2)d(i, j) - \sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k)$$

Tandis que $d(i, j)$ c'est le distance entre les deux taxons i et j .

2. Trouver une paire de taxons i et j (tandis que $i \neq j$) pour lequel $Q(i, j)$ a la valeur la plus petite. Ces taxons sont reliés à un nœud nouvellement créé, qui est connecté au nœud central.

3. Calculer la distance entre chacun des taxons de la paire et ce nouveau nœud :

$$\delta(i, u) = \frac{1}{2}d(i, j) + \frac{1}{2(n-2)} \left[\sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k) \right]$$

Et :

$$\delta(j, u) = d(i, j) - \delta(i, u)$$

i et j sont les paires de taxons et u le nœud nouvellement créé.

4. Calculer la distance entre chacun des autres taxons (les taxons qui ne sont pas pris en considération à l'étape précédente) et le nouveau nœud :

$$d(u, k) = \frac{1}{2} [d(i, k) + d(j, k) - d(i, j)]$$

Où u est le nouveau nœud, k est le nœud pour lequel nous voulons calculer la distance et i et j sont les membres de la paire qui viennent d'être jointe.

5. Relancer l'algorithme en remplaçant les deux taxons i et j « the pair of joined neighbors » par le nouveau nœud et en utilisant les distances calculées à l'étape précédente.

ARBRES PHYLOGÉNIQUES

❖ Méthodes d'inférence phylogénique

❖ Neighbour-Joining

○ Exemple

- Supposons que nous avons cinq taxons (a, b, c, d, e) et la matrice de distance D suivante:

	a	b	c	d	e
a	0	5	9	9	8
b		0	10	10	9
c			0	8	7
d				0	3
e					0

ARBRES PHYLOGÉNÉNIQUES

- ❖ Méthodes d'inférence phylogénique
- ❖ Neighbour-Joining
 - Exemple (solution)

