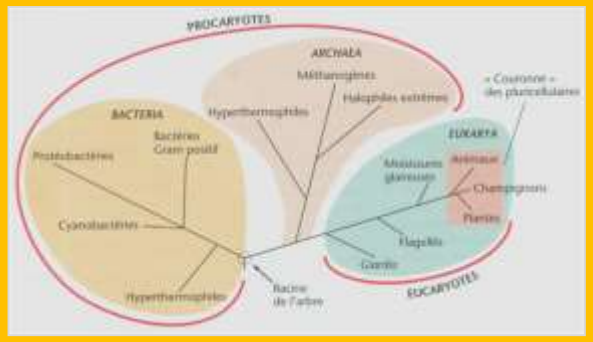
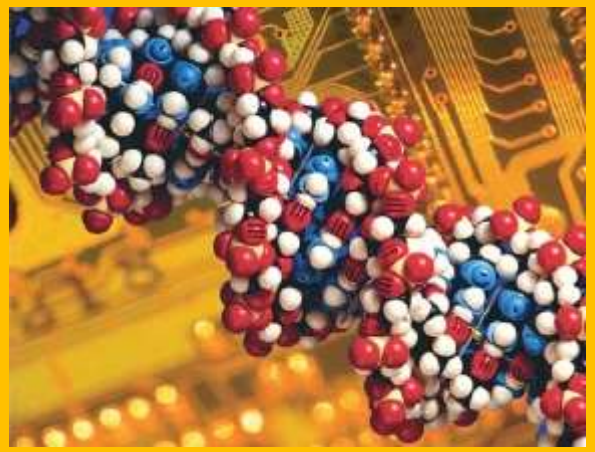
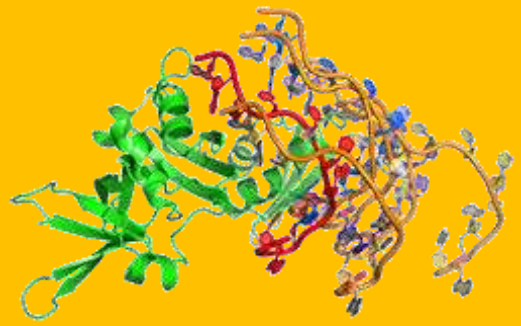
Université Frères Mentouri Constantine 1
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
1e année Master Biotechnologie alimentaire



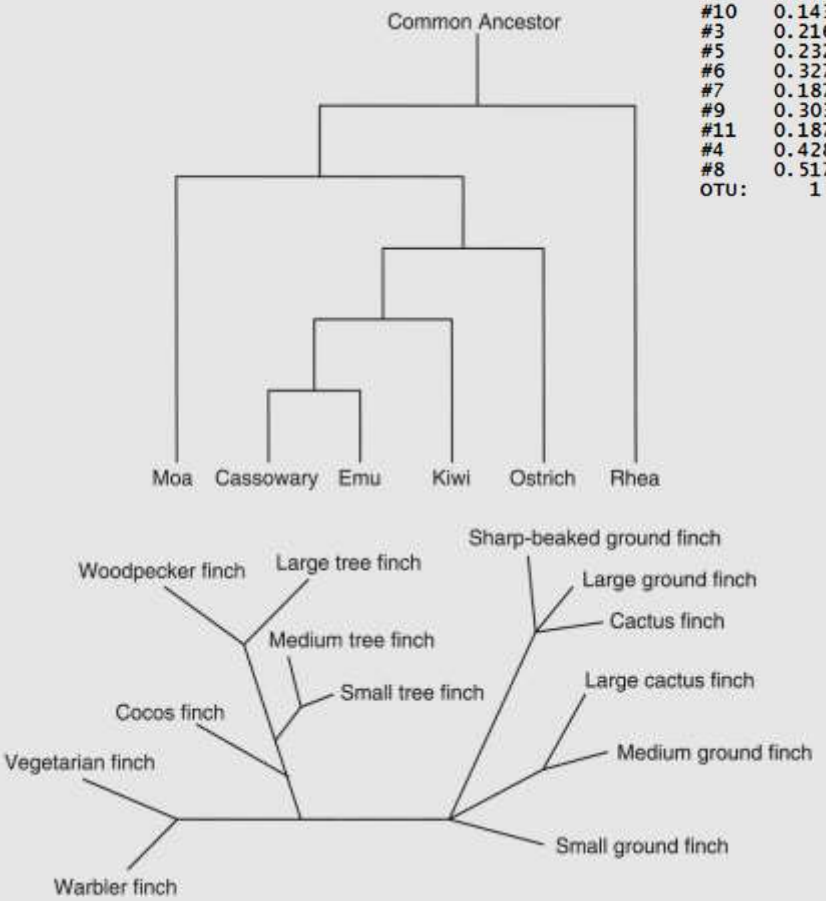
COURS DE BIOINFORMATIQUE

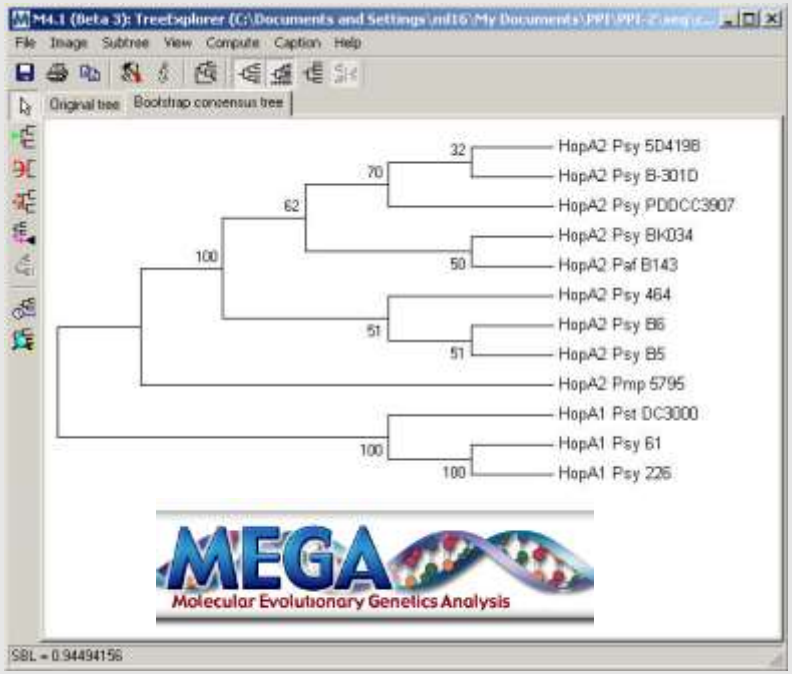




CHAPITRE III
PHYLOGÉNIE

```
 0.000  0.547  11 -LOG2(S(SM)) before clustering
#1     0.000
#2     0.187 0.000
#3     0.216 0.148 0.000
#4     0.428 0.298 0.389 0.000
#5     0.232 0.163 0.041 0.447 0.000
#6     0.327 0.252 0.148 0.346 0.120 0.000
#7     0.187 0.252 0.312 0.298 0.298 0.206 0.000
#8     0.517 0.382 0.292 0.546 0.259 0.259 0.489 0.000
#9     0.303 0.322 0.337 0.322 0.371 0.371 0.184 0.462 0.000
#10    0.143 0.120 0.263 0.346 0.206 0.346 0.252 0.489 0.322 0.000
#11    0.187 0.206 0.263 0.396 0.298 0.396 0.206 0.489 0.141 0.206 0.000
OTU:       1     2     3     4     5     6     7     8     9    10    11
 0.000  0.547  11 -LOG2(S(SM)) UNWEIGHTED AVERAGE LINKAGE
#1     0.000
#2     0.187 0.000
#10    0.143 0.120 0.000
#3     0.216 0.148 0.263 0.000
#5     0.232 0.163 0.206 0.041 0.000
#6     0.327 0.252 0.346 0.148 0.120 0.000
#7     0.187 0.252 0.252 0.312 0.298 0.206 0.000
#9     0.303 0.322 0.322 0.337 0.371 0.371 0.184 0.000
#11    0.187 0.206 0.206 0.263 0.298 0.396 0.206 0.141 0.000
#4     0.428 0.298 0.346 0.389 0.447 0.346 0.298 0.322 0.396 0.000
#8     0.517 0.382 0.489 0.292 0.259 0.259 0.489 0.462 0.489 0.546 0.000
OTU:       1     2    10     3     5     6     7     9    11     4     8
```



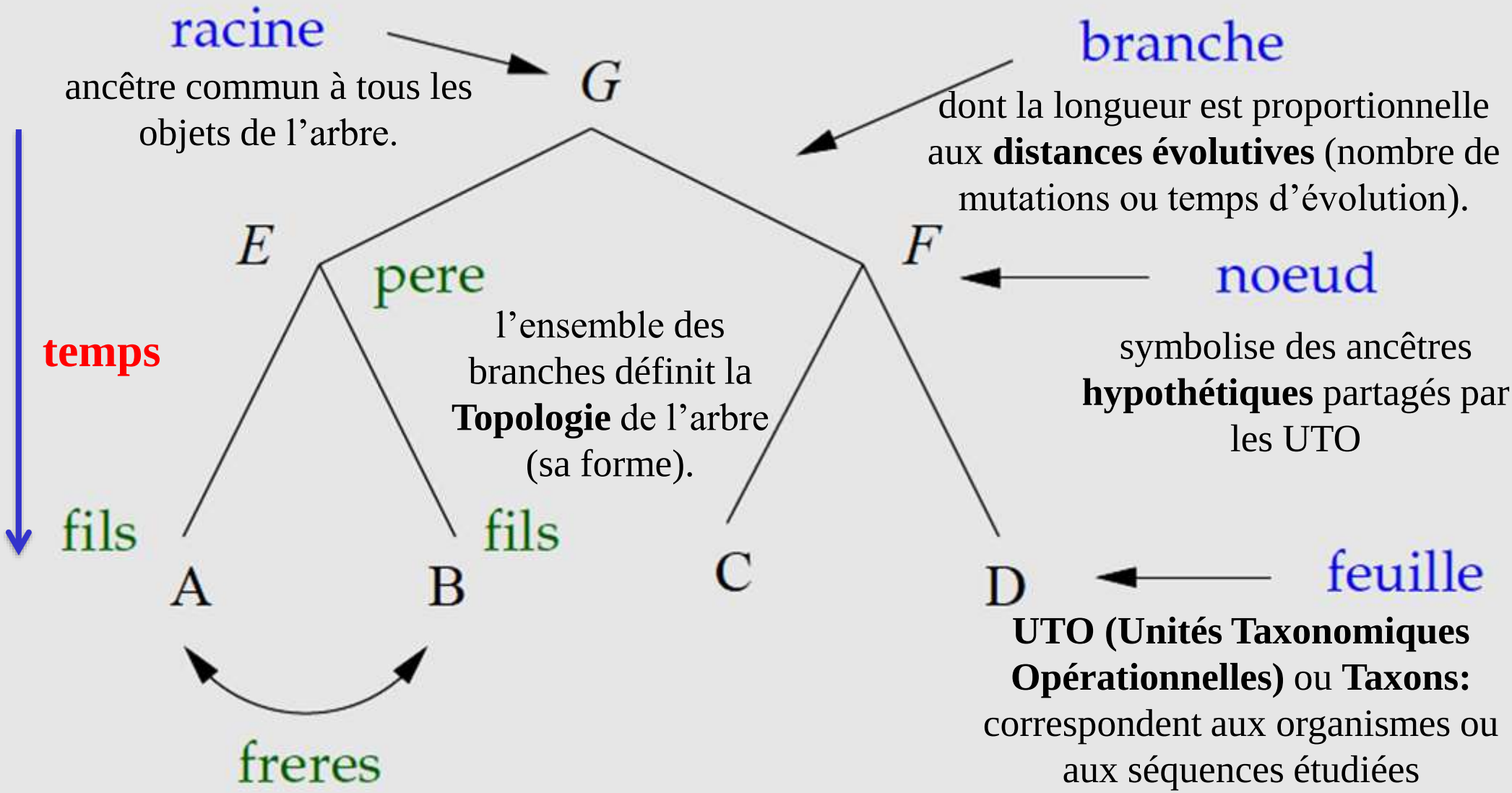


DÉFINITION

Phylogénie: Du grec ancien *phŷlon* (« tribu, race ») et *géneia* « qui engendre»

- La **phylogénie moléculaire** renseigne sur les changements survenus au niveau des séquences biologiques au cours du temps (évolution) ;
- La **phylogénie** est l'étude des relations de parentés entre différents êtres vivants en vue de comprendre l'évolution de ces organismes.
- La phylogénie trouve ses applications dans plusieurs domaines : systématique, génétique des populations, écologie, épidémiologie, phylogéographie, etc.
- Les relations mises en évidence par la phylogénie sont représentés **graphiquement** sous la forme d'**arbres phylogénétiques**.

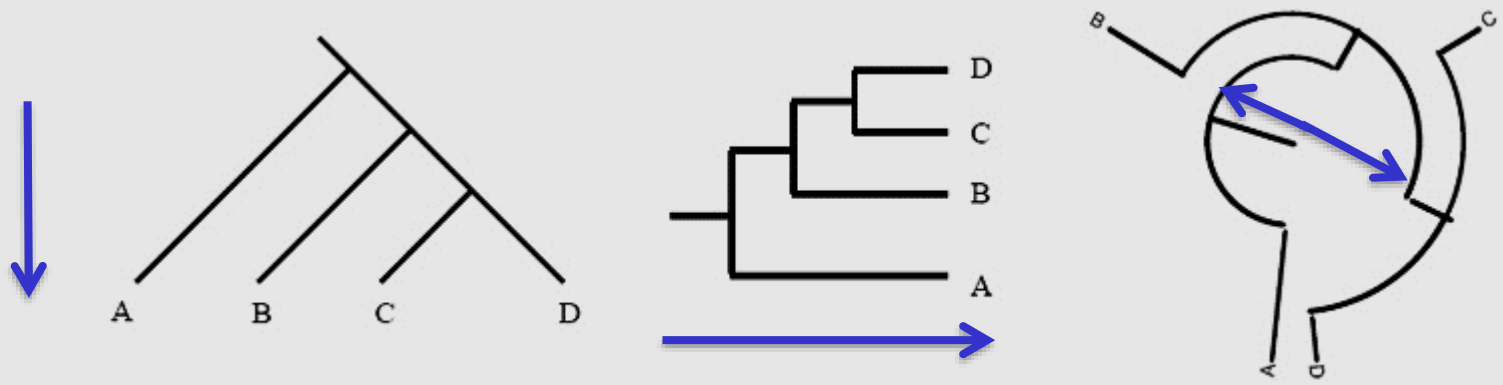
LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE



Le temps s'écoule de la racine vers les feuilles

LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE

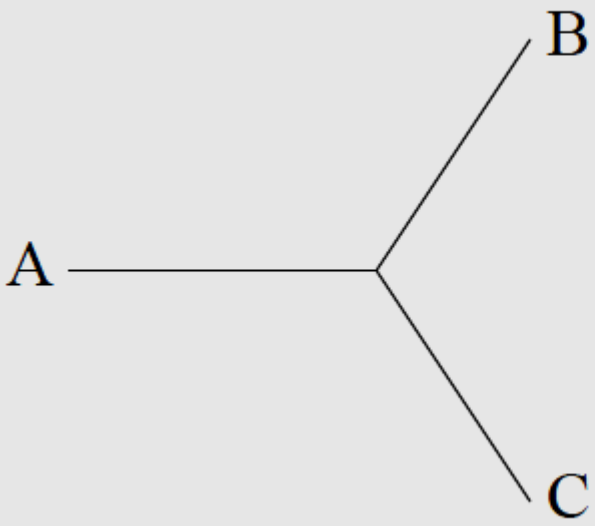
- ✓ Plusieurs styles de représentations géométriques des arbres phylogénétiques sont utilisés, avec des différences au niveau de la forme des branches et de la position de la racine dans l'arbre.

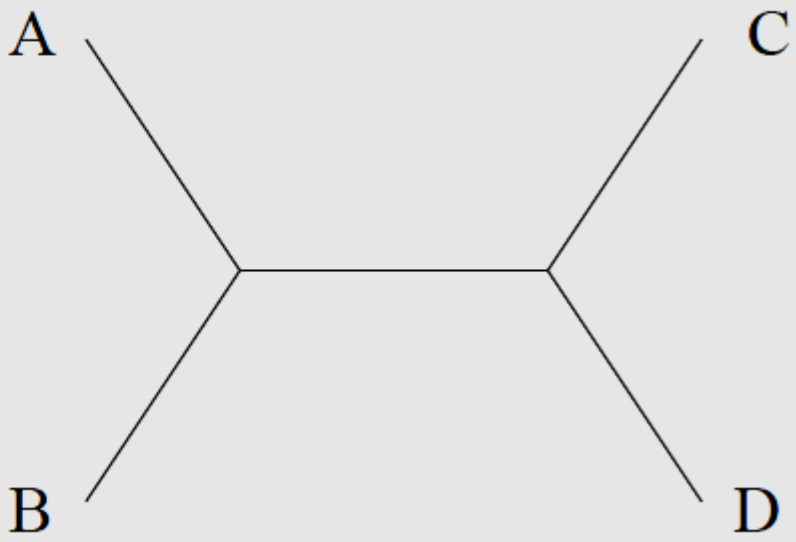


sens de lecture

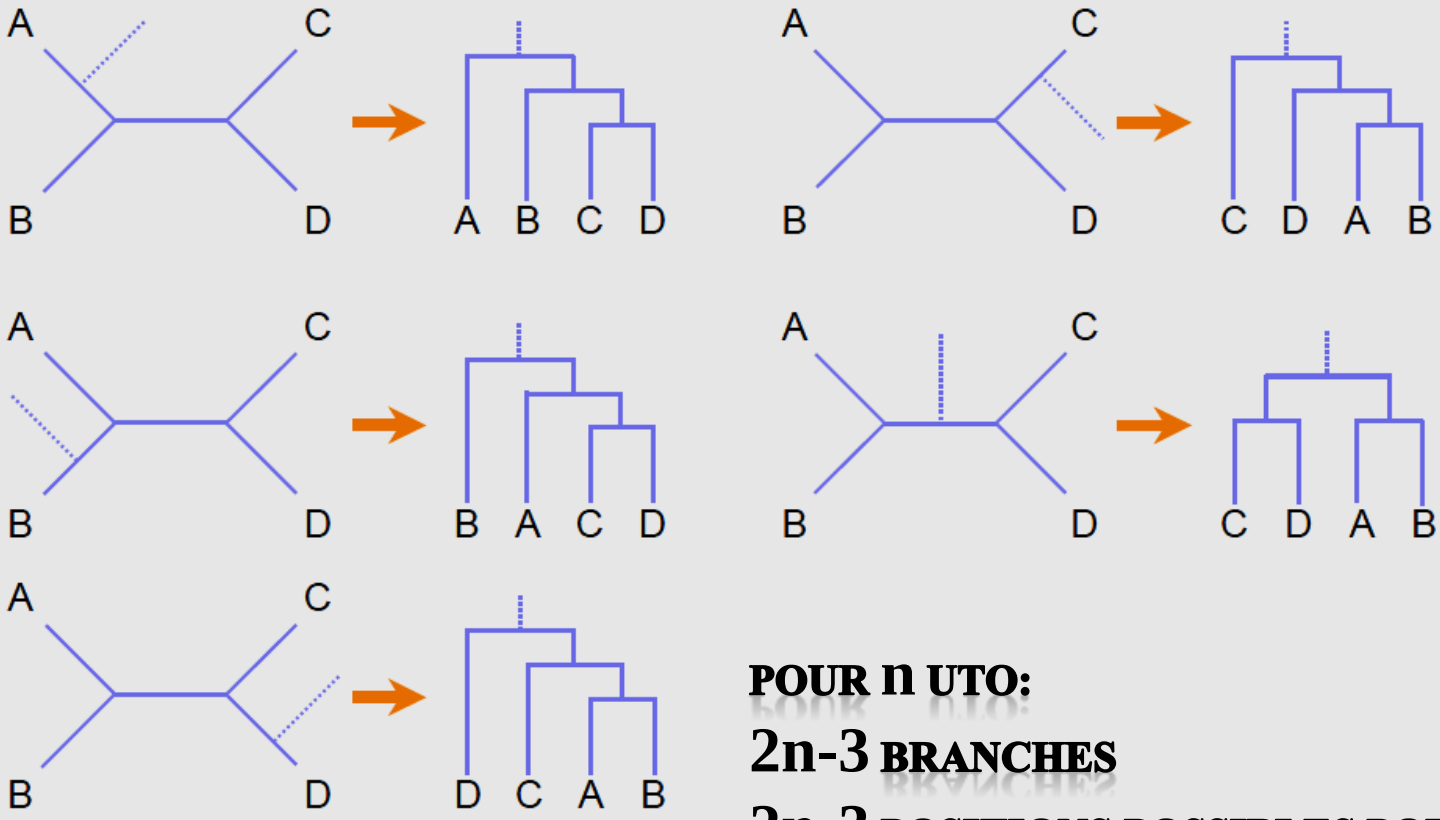
LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE

- ✓ Un arbre peut être **non-raciné**, dans ce cas c'est une représentation **intemporelle** ;
- ✓ La plus part des méthodes de phylogénie produisent des arbres **non-racinés** qui doivent être **racinés** par la suite.





✓ Le **racinement** d'un arbre phylogénétique est effectué par le positionnement d'une racine sur l'une des branche de cet arbre.



POUR n UTO:

2n-3 BRANCHES

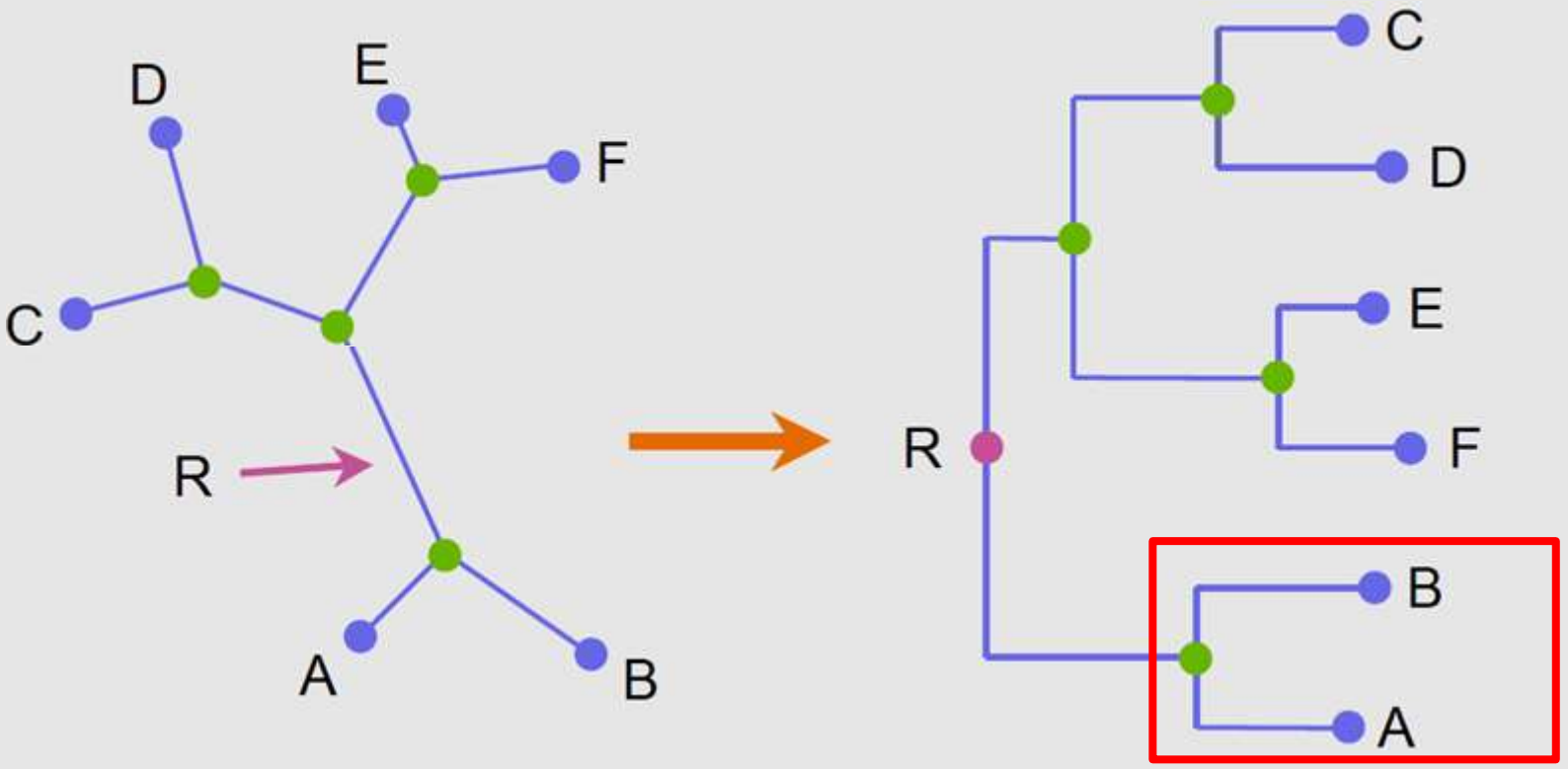
2n-3 POSITIONS POSSIBLES POUR LA RACINE

RACINEMENT D'UN ARBRE

- ✓ La méthode de racinement la plus utilisée est celle du **groupe externe** ;
- ✓ Elle consiste à choisir, en plus des séquences de l'étude, une ou plusieurs séquences (groupe externe) ne faisant pas partie du groupe étudié;
- ✓ le choix des séquences du groupe externe doit se faire de sorte à ce que ce dernier soit à **l'extérieur du groupe d'étude**, tout en étant le **plus proche possible**.

RACINEMENT D'UN ARBRE

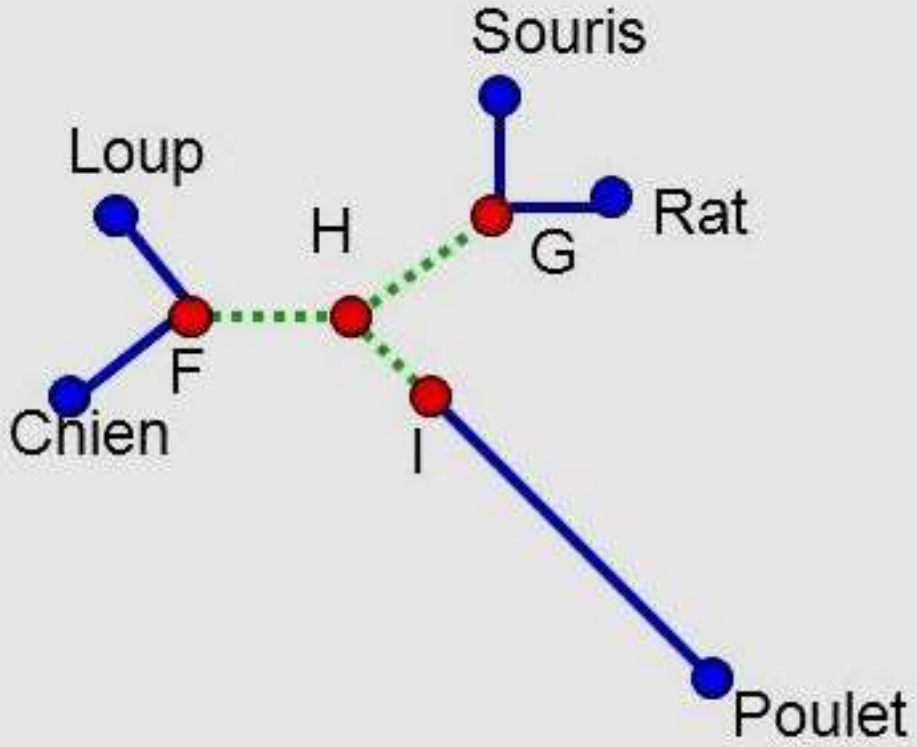
- ✓ Racinement par le groupe {A, B}, supposé extérieur aux organismes d'intérêt que sont C, D, E et F. La position de la racine sera la médiane de la somme des plus grandes branches reliant le groupe externe aux UTOs :

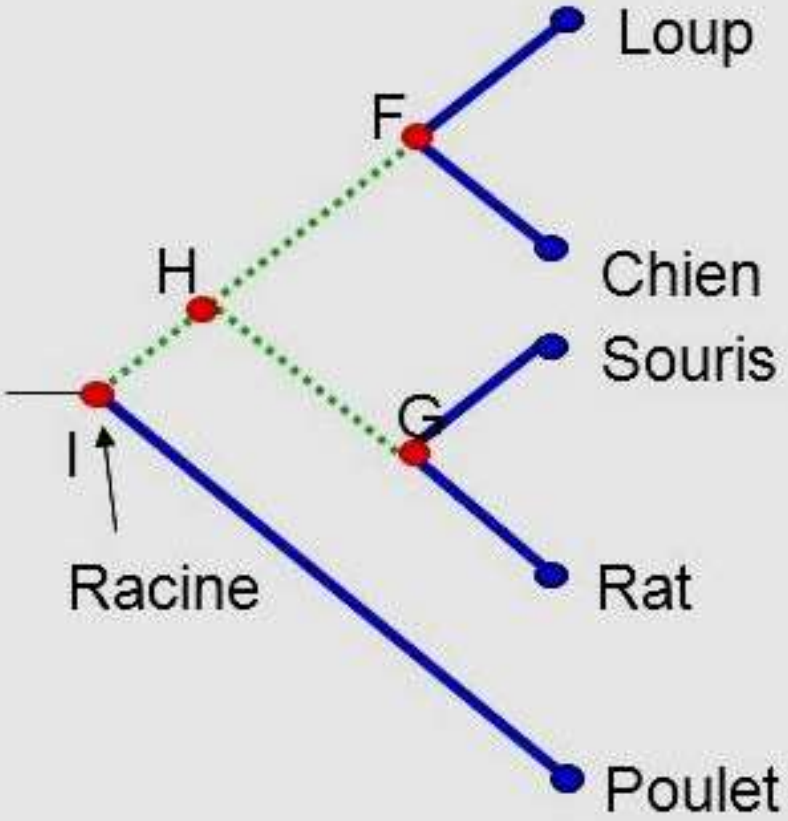


EXEMPLE DE RACINEMENT D'UN ARBRE PAR LA MÉTHODE DU GROUPE EXTERNE

LA STRUCTURE D'UN ARBRE PHYLOGÉNÉTIQUE

RACINEMENT D'UN ARBRE





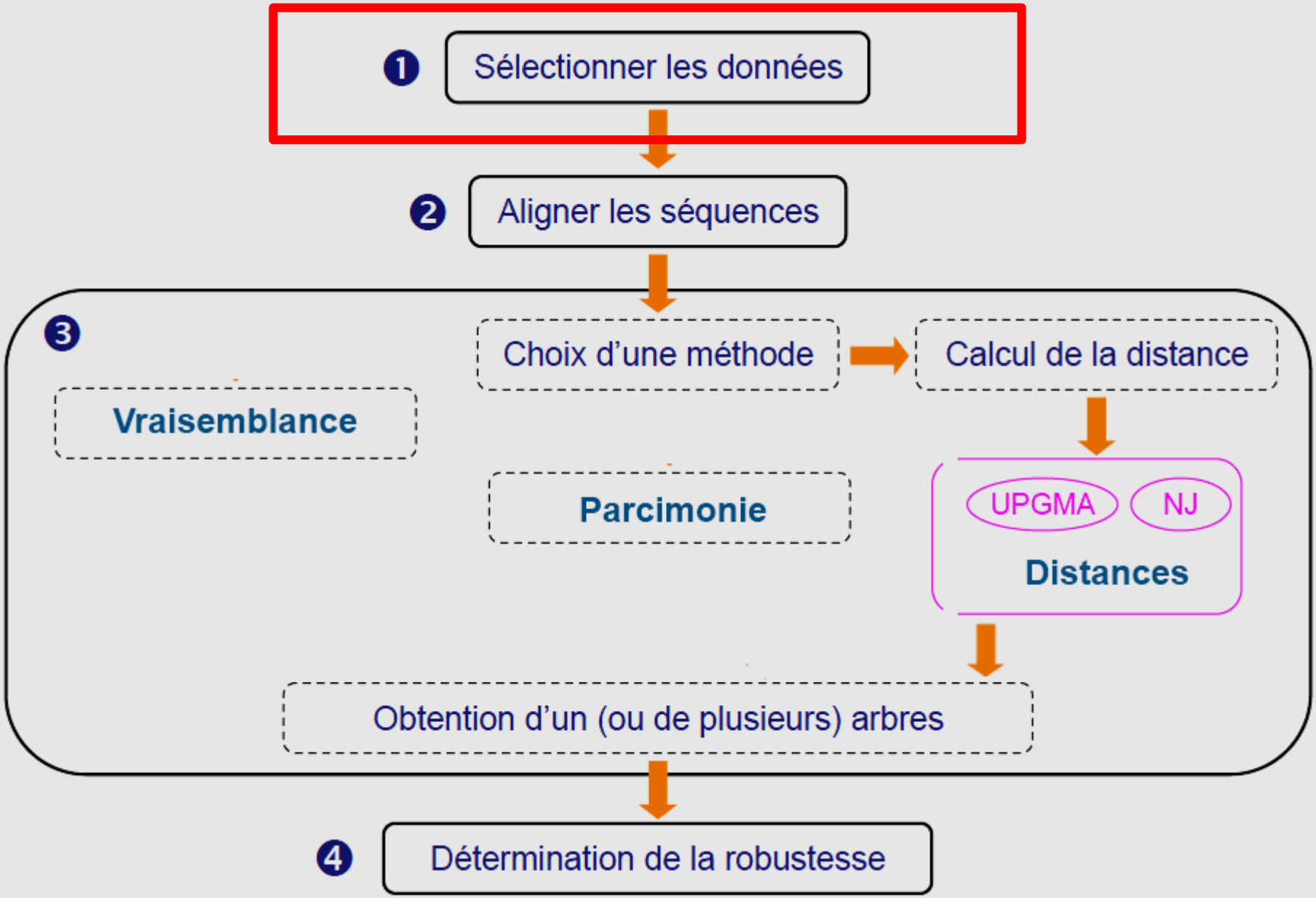
Connaissance *a priori* du OTU le plus externe parmi les OTU étudiées

Exemple: chien, loup, souris, rat et poulet

=> **Groupe extérieur** est le poulet

EXEMPLE DE RACINEMENT D'UN ARBRE PAR LA MÉTHODE DU GROUPE EXTERNE

ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE



DONNÉS DE LA PHYLOGÉNIE

✓ Les données utilisées pour la construction des arbres phylogénétiques sont des séquences **homologues alignées**, elles sont classées en deux groupes distincts :

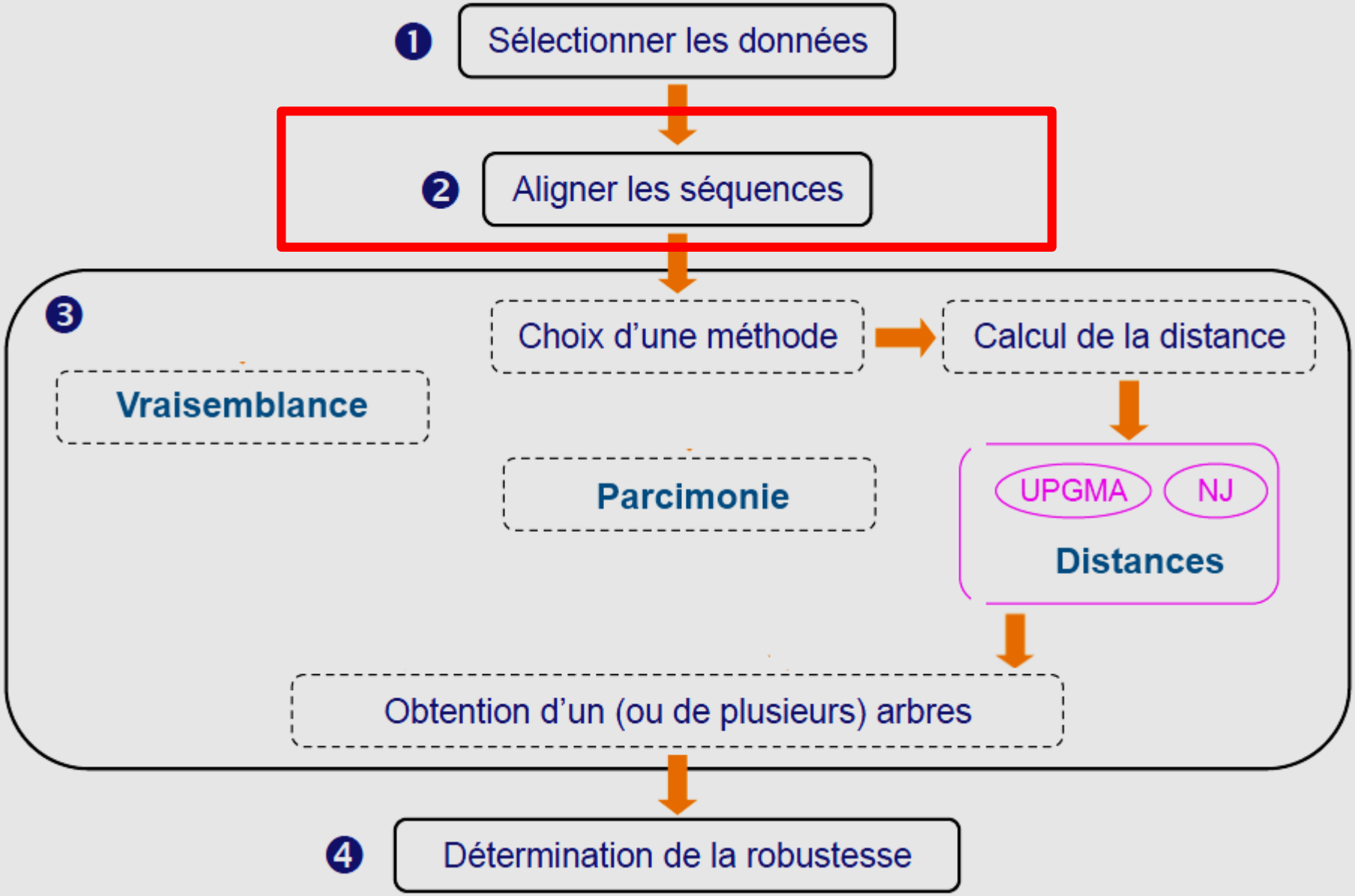
1. LES DONNÉES PHÉNOTYPIQUES

comprennent des caractères observables (aux différents états : morphologiques, biochimiques et physiologiques) mesurables par un langage binaire (de type présence d'un caractère donné / absence de ce même caractère).

2. LES DONNÉES MOLÉCULAIRES

Elles sont soit des **séquences nucléiques** telles que les séquences de gènes, ou des fragments particuliers du génome : gènes d'ARNm, d'ARNr, gènes ménagers, RFLP *(Restriction fragment length polymorphism*), Microsatellites, ITS (*Internal Transcribed Spacer*), transposons, ou des **séquences de protéines** enzymatiques ou de structure : protéines mitochondriales, actine, etc. On parle de **marqueurs moléculaires**. Un marqueur moléculaire est donc un indicateur de la variabilité génétique dans le temps.

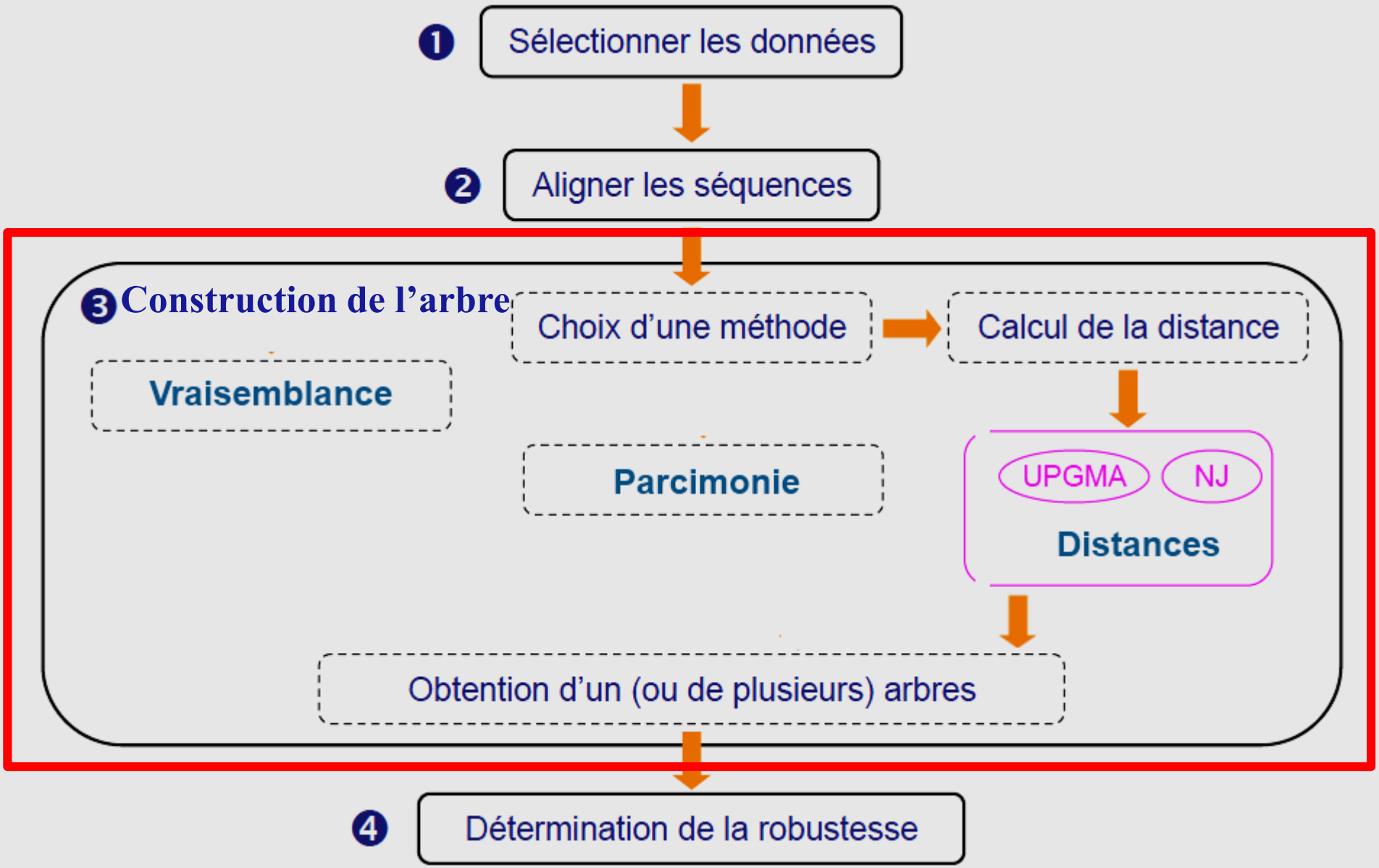
ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE



ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE

- ✓ Toutes les méthodes d'alignements multiples sont utilisables en phylogénie, mais se sont les méthodes d'alignements multiples progressifs qui sont privilégiées en raison de leur rapidité d'exécution et de la taille importante des données généralement analysées pour la construction des arbres phylogénétiques ;
- ✓ Exemple de méthodes employées : **CLUSTAL, Coffee**, **Muscle**...

ÉTAPES DE CONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE



MÉTHODES DE CONSTRUCTION DES ARBRES

Trois types de méthodes sont utilisées:

- ✓ **Méthodes de distances**: fondées sur le calcul des distances entre les séquences. Utilisées surtout pour des séquences de forte similitude **(plus rapides, les plus utilisées)** ;
- ✓ **Méthodes de parcimonie:** reposent sur la recherche d'un arbre comprenant le minimum de changements évolutifs (le moins de mutations possibles) en calculant un **score de parcimonie** ;
- ✓ **Méthodes probabilistes:** recherchent l'arbre optimal en attribuant une probabilité à chaque changement dans les séquences. Incluent les méthodes du maximum de vraisemblance et de l'Inférence bayésienne.

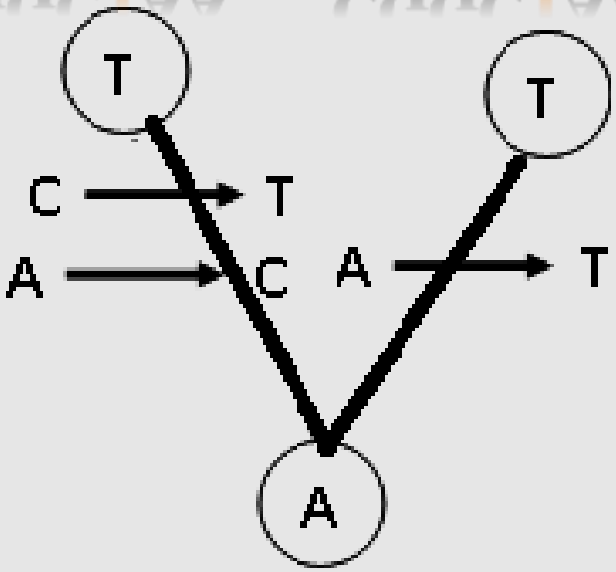
MÉTHODES DE DISTANCES

NOTION DE DISTANCE ÉVOLUTIVE *d*

- ✓ La distance évolutive (notée *d*) entre deux séquences est définie comme le nombre moyen de substitutions par site s'étant produites depuis que ces séquences ont divergé de leur ancêtre commun ;
- ✓ La distance observée est toujours inférieure à la distance évolutive **D<*d***
- ✓ L'unité de la distance évolutive est le nombre total de substitutions par site et est rapportée à la longueur des deux séquences alignées ;
- ✓ L'estimation des distances évolutives est à la base des méthodes de distances.

AGGCTAA **CGGCTAA**

Distance observée D
D=1-S



Distance évolutive d

CGGCAAA

Une distance évolutive prend en considération les divergences évolutives

CORRECTION DE LA DISTANCE

distance observée < la distance évolutive

Il existe plusieurs méthodes qui corrigent la distance, c'est-à-dire qui rapprochent les deux valeurs D et *d* (modèles de Jukes-Cantor ; Kimura ; Tamura-Nei) :

Exemple:

LE MODÈLE DE JUKES-CANTOR (POUR LES SÉQUENCES NUCLÉIQUES):

suppose que les quatre nucléotides ont les mêmes fréquences p(A) = p(C) = p(G) = p(T) et que leurs substitutions sont équiprobables et que la probabilité de transition et la probabilité de transversion sont égales

$$d = -3/4 \ln (1-4/3D)$$

Alignement de séquences

Mesures de distances évolutives

Matrice de distances évolutives entre paires de séquences

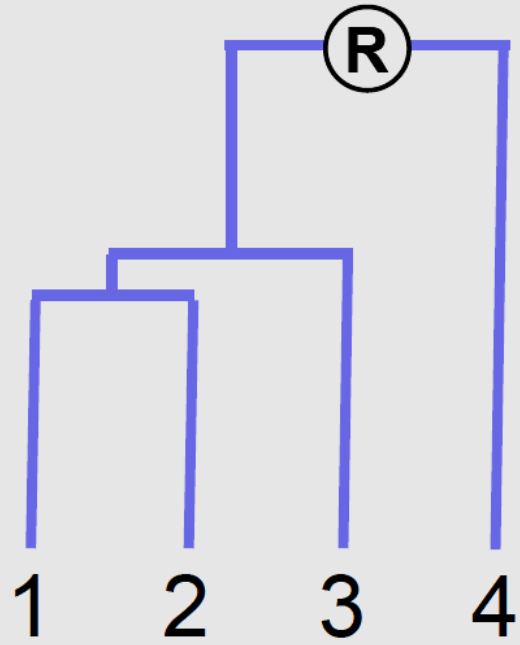
Calcul de l'arbre à partir de la matrice

Arbre

- ✓Parmi ces méthodes, la méthode de la classification ascendante hiérarchique (UPGMA) et la méthode du *Neighbor-Joining* (NJ), sont des exemples très connus, les deux méthodes construisent un **arbre unique**.

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE (*Unweight Pair Group Method with Arithmetic mean ou UPGMA*) (Sneath et Sokal, 1973):

- ✓ Méthode la plus simple d'un point de vue algorithmique ;
- ✓ Tire son nom du fait que la construction de l'arbre démarre à partir des feuilles ;
- ✓ Produit des arbres racinés ;
- ✓ Les distances arborées générées par cette méthode sont dites **ultramétriques**: les longueurs des chemins allant de la racine à n'importe quelle feuille sont égales, on considère que tous les UTO évoluent à la même vitesse (théorie de l'**horloge moléculaire**) ;



ALGORITHME

- ✓UPGMA utilise un algorithme de **clustérisation séquentielle** car elle construit l'arbre pas à pas, au fur et à mesure que les **clades** sont définis. Un **clade** est une paire d'individus très similaires regroupés ensemble, l'arbre final est appelé **cladogramme** ;
- ✓L'algorithme UPGMA repose sur quatre principales étapes :

MÉTHODES DE DISTANCES

UPGMA

EXEMPLE: le gène codant pour un ARNt mitochondrial a été séquencé chez 5 espèces d'Hominidés.

```
> Homme (L00016)
AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATTCTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGGACTTCAAACTCT
ACTCCCACTAATAGCTTTTTGATGACTTCTAGCAAGCCTCGCTAACCTCGCCTTACCCCCCACTATTAACCTACTGGGAGAACTCTCTGTGCTAGTAACCACGTTCTCCTGATCAAATATCACTCTCCTACT
TACAGGACTCAACATACTAGTCACAGCCCTATACTCCCTCTACATATTTACCACAACACAATGGGGCTCACTCACCCACCACATTAACAACATAAAACCCTCATTCACACGAGAAAACACCCTCATGTTCAT
ACACCTATCCCCATTCTCCTCCTATCCCTCAACCCCGACATCATTACCGGGTTTCCTCTTGTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACCG
AGAAAGCTCACAAGAACTGCTAACTCATGCCCCCATGTCTGACAACATGGCTTTCTCAACTTTTAAAGGATAACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA
ACCATGCACACTACTATAACCACCCTAACCCTGACTTCCCTAATTCCCCCCATCCTTACCACCCTCGTTAACCCTAACAAAAAAAACTCATACCCCCATTATGTAAAATCCATTGTCGCATCCACCTTTATT
ATCAGTCTCTTCCCCACAACAATATTCATGTGCCTAGACCAAGAAGTTATTATCTCGAACTGACACTGAGCCACAACCCAAACAACCCAGCTCTCCCTAAGCTT

> Chimpanzée (V00672)
AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATTCTGCCTAGCAAACTCAAATTATGAACGCACCCACAGTCGCATCATAATTCTCTCCCAAGGACTTCAAACTCT
ACTCCCACTAATAGCCTTTTGATGACTCCTAGCAAGCCTCGCTAACCTCGCCCTACCCCCTACCATTAATCTCCTAGGGGAACTCTCCGTGCTAGTAACCTCATTCTCCTGATCAAATACCACTCTCCTACT
CACAGGATTCAACATACTAATCACAGCCCTGTACTCCCTCTACATGTTTACCACAACACAATGAGGCTCACTCACCCACCACATTAATAACATAAAGCCCTCATTCACACGAGAAAATACTCTCATATTTTT
ACACCTATCCCCATCCTCCTTCTATCCCTCAATCCTGATATCATCACTGGATTCACCTCCTGTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTCACGACCCCTTATTTACCG
AGAAAGCTTATAAGAACTGCTAATTCATATCCCCATGCCTGACAACATGGCTTTCTCAACTTTTAAAGGATAACAGCCATCCGTTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA
ACCATGTATACTACCATAACCACCTTAACCCTAACTCCCTTAATTCTCCCCATCCTCACCACCCTCATTAACCCTAACAAAAAAAACTCATATCCCCATTATGTGAAATCCATTATCGCGTCCACCTTTATC
ATTAGCCTTTTCCCCACAACAATATTCATATGCCTAGACCAAGAAGCTATTATCTCAAACTGGCACTGAGCAACAACCCAAACAACCCAGCTCTCCCTAAGCTT

> Orang-outan (V00675)
AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATCCTCTCTCAAGGCCTTCAAACTCT
ACTCCCCCTAATAGCCCTCTGATGACTTCTAGCAAGCCTCACTAACCTTGCCCTACCACCCACCATCAACCTTCTAGGAGAACTCTCCGTACTAATAGCCATATTCTCTTGATCTAACATCACCATCCTACT
AACAGGACTCAACATACTAATCACAACCCTATACTCTCTCTATATATTCACCACAACACAACGAGGTACACCCACACACCACATCAACAACATAAAACCTTCTTTCACACGCGAAAATACCCTCATGCTCAT
ACACCTATCCCCCATCCTCCTCTTATCCCTCAACCCCAGCATCATCGCTGGGTTCGCCTACTGTAAATATAGTTTAACCAAAACATTAGATTGTGAATCTAATAATAGGGCCCCACAACCCCTTATTTACCG
AGAAAGCTCACAAGAACTGCTAACTCTCACTCCATGTGTGACAACATGGCTTTCTCAGCTTTTAAAGGATAACAGCTATCCCTTGGTCTTAGGATCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAACAG
CCATGTTTACCACCATAACTGCCCTCACCTTAACTTCCCTAATCCCCCCCATTACCGCTACCCTCATTAACCCCAACAAAAAAAACCCATACCCCCACTATGTAAAAACGGCCATCGCATCCGCCTTTACTA
TCAGCCTTATCCCAACAACAATATTTATCTGCCTAGGACAAGAAACCATCGTCACAAACTGATGCTGAACAACCACCCAGACACTACAACTCTCACTAAGCTT

> Gorilla (V00658)
AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATTCTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGGACTCCAAACCCT
ACTCCCACTAATAGCCCTTTGATGACTTCTGGCAAGCCTCGCCAACCTCGCCTTACCCCCCACCATTAACCTACTAGGAGAGCTCTCCGTACTAGTAACCACATTCTCCTGATCAAATACCACCCTTTACT
TACAGGATCTAACATACTAATCACAGCCCTGTACTCCCTTTATATATTTACCACAACACAATGAGGCCCACTCACACACCACATCACCAACATAAAACCCTCATTTACACGAGAAAACATCCTCATATTCAT
GCACCTATCCCCATCCTCCTCTATCCCTCAACCCCGATATTATCACCGGGTTCACCTCCTGTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGATAACAGAGGCTCACAACCCCTTATTTACCG
AGAAAGCTCGTAAGAGCTGCTAACTCATACCCCCGTGCTTGACAACATGGCTTTCTCAACTTTTAAAGGATAACAGCTATCCATTGGTCTTAGGACCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA
ACTATGTACGCTACCATAACCACCTTAGCCCTAACTTCCTTAATTCCCCCTATCCTTACCACCTTCATCAATCCTAACAAAAAAAGCTCATACCCCCATTACGTAAAATCTATCGTCGCATCCACCTTTATC
ATCAGCCTCTTCCCCACAACAATATTTCTATGCCTAGACCAAGAAGCTATTATCTCAAGCTGACACTGAGCAACAACCCAAACAATTCAACTCTCCCTAAGCTT

> Gibbon (V00659)
AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATTCTGCCTTGCAAACTCAAACTACGAACGAACTCACAGCCGCATCATAATCCTATCTCGAGGGCTCCAAGCCTT
ACTCCCACTGATAGCCTTCTGATGACTCGCAGCAAGCCTCGCTAACCTCGCCCTACCCCCCACTATTAACCTCCTAGGTGAACTCTTCGTACTAATGGCCTCCTTCTCCTGGGCAAACACTACTATTACACT
CACCGGGCTCAACGTACTAATCACGGCCCTATACTCCCTTTACATATTTATCATACACACAACGAGGCACACTTACACACCACATTAAAACACATAAACCCTCACTCACACGAGAAAACATATTAATACTTAT
GCACCTCTTCCCCCTCCTCCTCCTAACCCTCAACCCTAACATCATTACTGGCTTTACTCCCTGTAAACATAGTTTAATCAAAACATTAGATTGTGAATCTAACAATAGAGGCTCGAAACCTCTTGCTTACCG
AGAAAGCCCACAAGAACTGCTAACTCACTATCCCATGTATGACAACATGGCTTTCTCAACTTTTAAAGGATAACAGCTATCCATTGGTCTTAGGACCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA
GCAATGTACACCACCATAGCCATTCTAACGCTAACCTCCCTAATTCCCCCCATTACAGCCACCCTTATTAACCCCAATAAAAAGAACTTATACCCGCACTACGTAAAAATGACCATTGCCTCTACCTTTATA
ATCAGCCTATTTCCCACAATAATATTCATGTGCACAGACCAAGAAACCATTATTTCAAACTGACACTGAACTGCAACCCAAACGCTAGAACTCTCCCTAAGCTT
```

MÉTHODES DE DISTANCES

UPGMA

La séquence est très conservée : sur 895 pb il y a une seule déletion chez l'Orang-Outan et

281 sites variables

```
Gibbon          ACCCTCACTCACACGAGAAAACATATTAATACTTATGCACCTCTTCCCCCTCCTCCTCCT   420
Orang-outan     ACCTTCTTTCACACGCGAAAATACCCTCATGCTCATACACCTATCCCCCATCCTCCTCTT   420
Gorilla         ACCCTCATTTACACGAGAAAACATCCTCATATTCATGCACCTATCCCCCATCCTCCTCCT   420
Homme           ACCCTCATTCACACGAGAAAACACCCTCATGTTCATACACCTATCCCCCATTCTCCTCCT   420
Chimpanzée      GCCCTCATTCACACGAGAAAATACTCTCATATTTTTACACCTATCCCCCATCCTCCTTCT   420
                 ** **  * ***** ***** *   * **  *  * ***** * **** * *****  *

Gibbon          AACCCTCAACCCTAACATCATTACTGGCTTTACTCCCTGTAAACATAGTTTAATCAAAAC   480
Orang-outan     ATCCCTCAACCCCAGCATCATCGCTGGGTTCGCCTACTGTAAATATAGTTTAACCAAAAC   480
Gorilla         ATCCCTCAACCCCGATATTATCACCGGGTTCACCTCCTGTAAATATAGTTTAACCAAAAC   480
Homme           ATCCCTCAACCCCGACATCATTACCGGGTTTTCCTCTTGTAAATATAGTTTAACCAAAAC   480
Chimpanzée      ATCCCTCAATCCTGATATCATCACTGGATTCACCTCCTGTAAATATAGTTTAACCAAAAC   480
                * ******* **    ** **  * ** **  *    ****** ********* ******

Gibbon          ATTAGATTGTGAATCTAACAATAGAGGCTCGAAACCTCTTGCTTACCGAGAAAGCCCACA   540
Orang-outan     ATTAGATTGTGAATCTAATAATAGGGCCCCACAACCCCTTATTTACCGAGAAAGCTCACA   540
Gorilla         ATCAGATTGTGAATCTGATAACAGAGGCTCACAACCCCTTATTTACCGAGAAAGCTCGTA   540
Homme           ATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACCGAGAAAGCTCACA   540
Chimpanzée      ATCAGATTGTGAATCTGACAACAGAGGCTCACGACCCCTTATTTACCGAGAAAGCTTATA   540
                ** ************* * ** ** * *     *** ***  *************    *

Gibbon          AGAACTGCTAACTCACTATCCCATGTATGACAACATGGCTTTCTCAACTTTTAAAGGATA   600
Orang-outan     AGAACTGCTAACTCTCA-CTCCATGTGTGACAACATGGCTTTCTCAGCTTTTAAAGGATA   599
Gorilla         AGAGCTGCTAACTCATACCCCCGTGCTTGACAACATGGCTTTCTCAACTTTTAAAGGATA   600
Homme           AGAACTGCTAACTCATGCCCCCATGTCTGACAACATGGCTTTCTCAACTTTTAAAGGATA   600
Chimpanzée      AGAACTGCTAATTCATATCCCCATGCCTGACAACATGGCTTTCTCAACTTTTAAAGGATA   600
                *** ******* **      ** **  ******************* *************

Gibbon          ACAGCTATCCATTGGTCTTAGGACCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA   660
Orang-outan     ACAGCTATCCCTTGGTCTTAGGATCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAACA   659
Gorilla         ACAGCTATCCATTGGTCTTAGGACCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA   660
Homme           ACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA   660
Chimpanzée      ACAGCCATCCGTTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA   660
                ***** **** ***********  ********************************** *
```

La matrice de distances (exprimée) entre les cinq taxons est donnée par la relation :

$$D_{ij}=d_{ij}=\textit{nombre}\ \textbf{de substitutions}\ \textit{entre i et j/Taille de l'alignement}$$

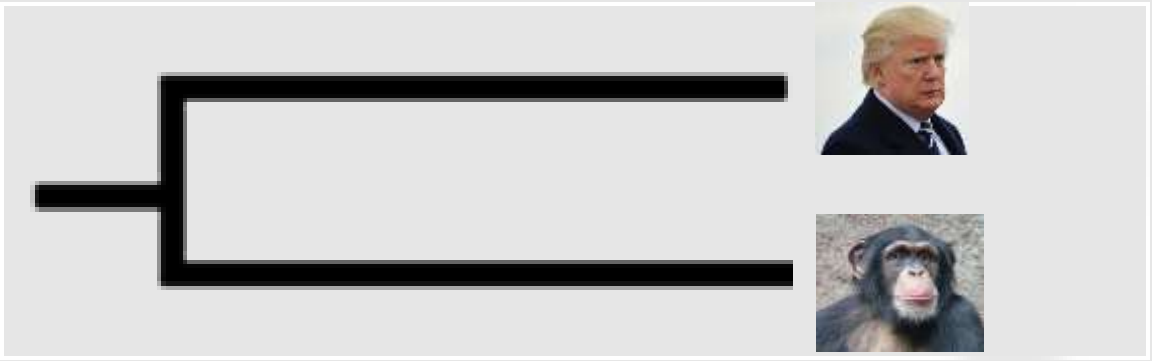
	Homme	Chimpanzée	Gorille	Orang-Outan
Chimpanzée	0,092			
Gorille	0,106	0,111		
Orang-Outan	0,177	0,193	0,188	
Gibbon	0,207	0,218	0,218	0,219

ÉTAPE 1: Identifier les deux UTO ayant la plus petite distance d_{ij}

Homme et Chimpanzé vont donc constituer le premier clade de l'arbre

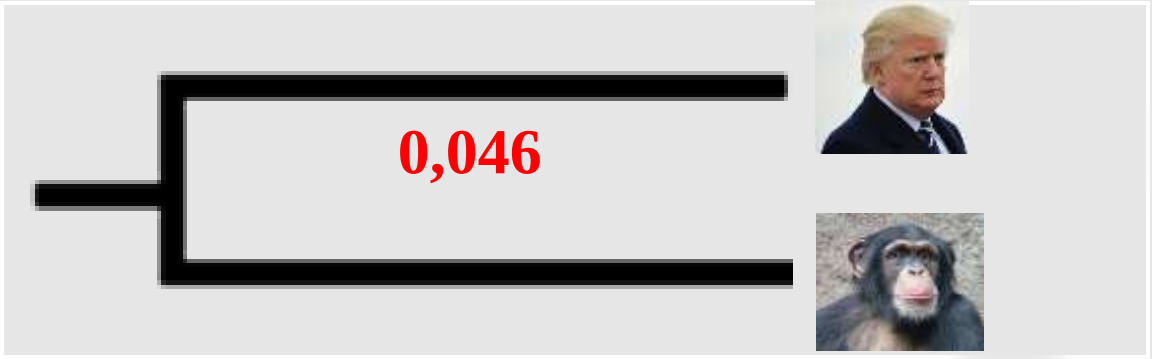
ÉTAPE 1: Identifier les deux UTO ayant la plus petite distance d_{ij}

Homme et Chimpanzé vont donc constituer le premier clade de l'arbre



ÉTAPE 2: Mettre ces deux UTO à égale distance du nœud formant le clade *ij*. La longueur de la branche du clade est d_{ij} / 2

la distance qui sépare Homme de Chimpanzé ; soit 0,092 / 2 = 0,046



ÉTAPE 3:

Construire un nouvel ensemble en considérant le clade *ij* comme étant un individu à part qu'il faut comparer avec le reste des individus en calculant une nouvelle matrice de dimensions n-1 et m-1.

On suppose que Homme et Chimpanzé forment un seul individu *U*. Donc l'ensemble des individus sera E = {*U*, Gorille, Orang-Outan, Gibbon}.

A cette étape, nous devons calculer les distances qui séparent le taxon U du reste des taxons qui sont Taxon1, Taxon2, Taxon5:

- La distance entre U et Gorille est : *d* (*U*, Gorille) = (*d* (Homme, Gorille) + *d* (Chimpanzé, Gorille))/2

$$d\,(U, \text{Gorille}) = ((0{,}106) + (0{,}111)) / 2 = 0{,}108$$

- La distance entre U et Taxon2 est : *d* (*U*, Orang-O.)= (*d* (Homme, Orang-O.)+ *d* (Chimp., Orang-O.))/2

$$d\,(U, \text{Orang-O.}) = ((0{,}177) + (0{,}193)) / 2 = 0{,}185$$

La distance entre U et Taxon5 est : *d* (*U*, Gibbon)=(*d* (Homme, Gibbon)+ *d* (Chimpanzé, Gibbon))/2

$$d\,(U, \text{Gibbon}) = ((0{,}207) + (0{,}218)) / 2 = 0{,}213$$

ÉTAPE 3:

On suppose que Taxon3 et Taxon4 forment un seul individu *U*. Donc l'ensemble des individus sera E = {*U*, Gorille, Orang-Outan, Gibbon}.

Nouvelle matrice de distances:

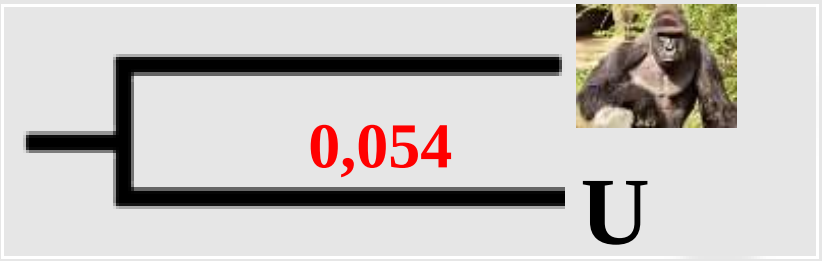
	U	Gorille	Orang-Outan
Gorille	0,108		
Orang-Outan	0,185	0,188	
Gibbon	0,213	0,218	0,219

ÉTAPE 4: Recommencer à partir de l'étape 1:

A partir de cette nouvelle matrice:

la distance qui sépare U du Gorille

0,108 / 2 = 0,054



ÉTAPE 4: Recommencer à partir de l'étape 1:

A partir de cette nouvelle matrice:

Gorille et U forment le clade *W*, donc l'ensemble des individus sera E = {*W*, Ourang-Outan, Gibbon}.

La distance entre Ourang-Outan et *W* est donnée par la relation :

$$d(\text{Ourang-Outan}, W) = (d(\text{Ourang-Outan, Gorille}) + d(\text{Ourang-Outan, U}))/2$$

$$d\ (\text{Ourang-Outan,W}) = ((0{,}188) + (0{,}185)) / 2 = 0{,}1865$$

La distance entre Taxon5 et W est donnée par la relation :

$$d\ (\text{Gibbon}, W) = (d(\text{Gibbon, Gorille}) + d(\text{Gibbon, U}))/2$$

$$d\ (\text{Gibbon}, W) = ((0{,}218) + (0{,}213)) / 2 = 0{,}2155$$

un nouveau cycle va donner la matrice de distances suivantes:

	W	Orang-Outan
Orang-Outan	0,1865 ←	
Gibbon	0,2155	0,219

Le nouveau clade sera:

la distance qui sépare W de Orang-Outan

0,1865 / 2 = 0,09325

MÉTHODES DE DISTANCES

UPGMA

Les taxons W et Ourang-Outan forment un seul individu. On lui donne le nom *Z*. Donc l'ensemble des individus sera E = {*Z*, Gibbon}.

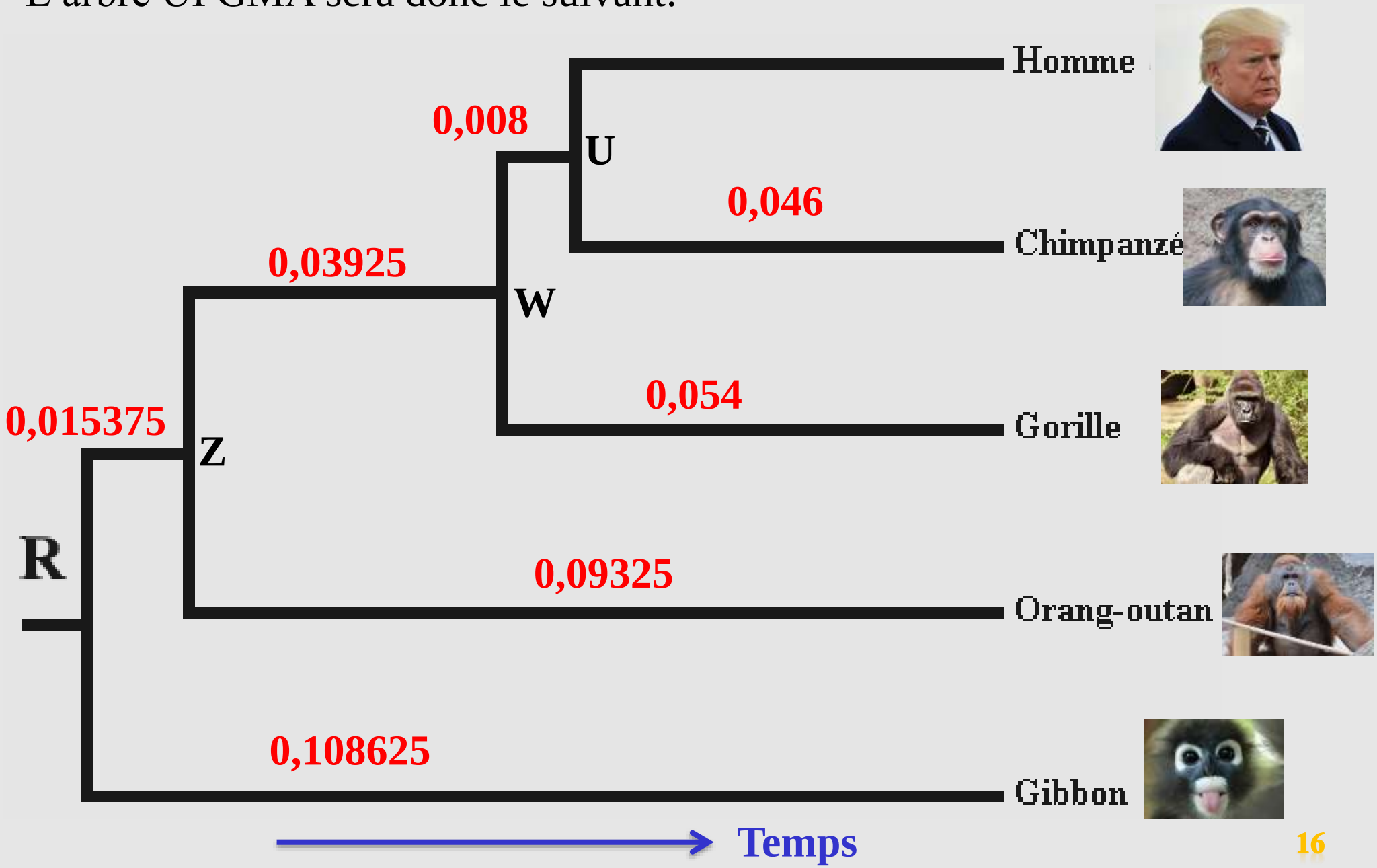
La distance entre ces deux OTUs est : $d(Z, \text{Gibbon})=(d(\text{W}, \text{Gibbon})+ d(\text{Ourang}, \text{Gibbon}))/2$

$$d\,(Z, \text{Gibbon}) = ((0{,}2155) + (0{,}219)) / 2 = 0{,}21725$$

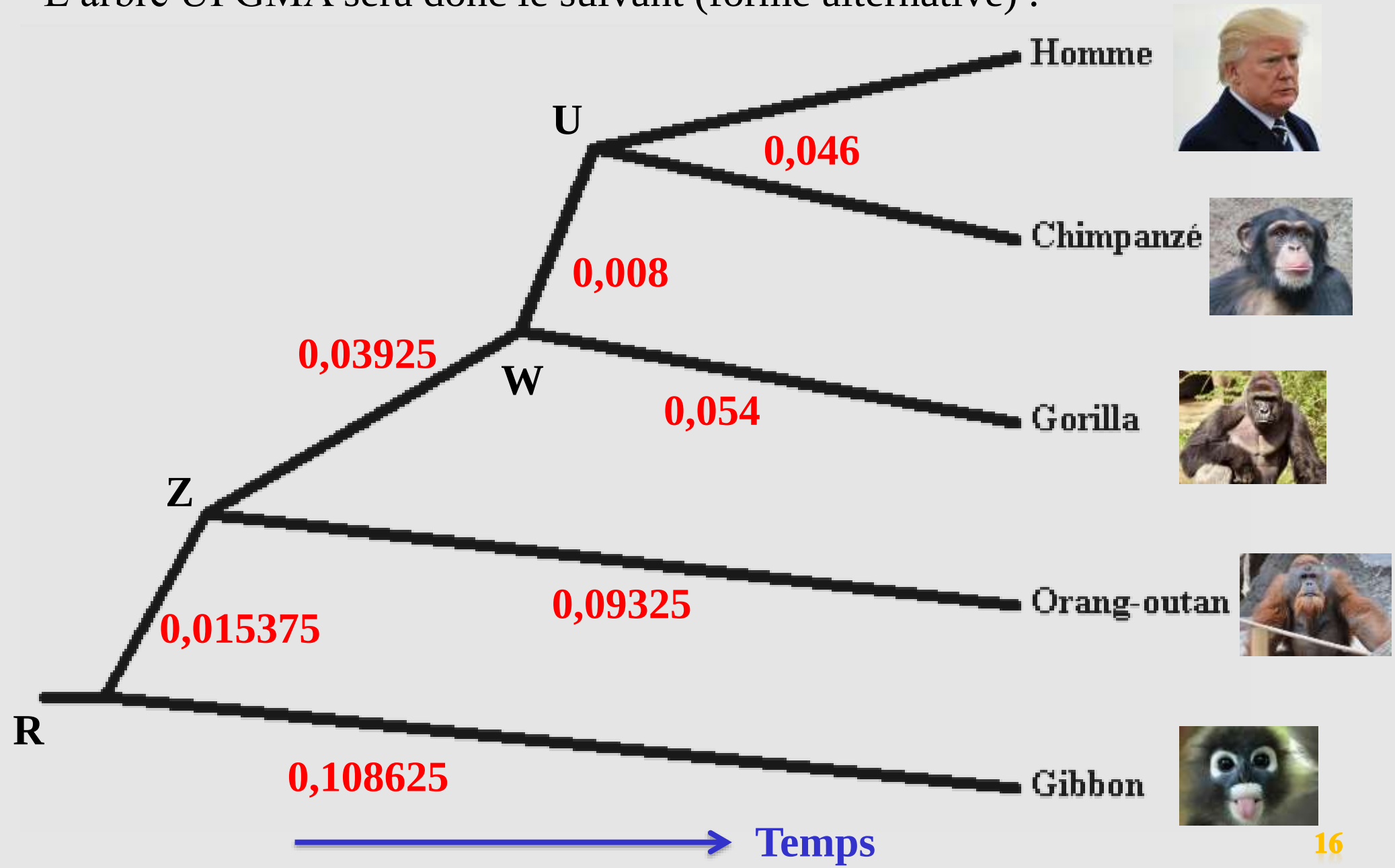
Et donc la taille des branches portant les UTO = $d\,(Z, \text{Gibbon})/2 = 0{,}21725/2 = 0{,}108625$

Le clade de ces deux OTUs est :

L'arbre UPGMA sera donc le suivant:



L'arbre UPGMA sera donc le suivant (forme alternative) :



NEIGHBOR-JOINING (Saitou et Nei, 1987):

- ✓ Méthode rapide, la plus utilisée;
- ✓ Le principe de NJ consiste en le calcul des longueurs des branches de l'arbre de sorte qu'elles soient les plus petites possibles
- ✓ Les distances arborées générées par cette méthode sont dites **additives**: les longueurs des chemins allant de la racine à n'importe quelle feuille ne sont pas égales, ce qui stipule que les caractères des UTO évoluent indépendamment les uns des autres (taux de mutations variables) ;

ALGORITHME

- ✓ NJ utilise un algorithme de clustérisation séquentielle;
- ✓ A chaque étape de clustérisation, NJ préfère les UTO qui réduisent la longueur des branches de l'arbre en choisissant les UTO voisins (**neighbors**);

✓ Pour construire un arbre à n UTO, l'algorithme du NJ suit les étapes suivantes:

1. Construire la matrice de distances entre les n UTO ;

2. Calcul de la divergence « r » entre chaque UTO *i* et les autres *k* UTO en suivant la formule:

$$r_i = \sum d_{ik}$$

3. Construction d'une nouvelle matrice en se basant sur le calcul de la longueur de l'arbre $S_{ij} = (n\text{-}2)\, d_{ij} - r_i - r_j$;

4. Le premier nœud U sera constitué par la paire d'UTO *ij* ayant la plus faible S et la distance arborée *b* de chaque branche est calculé par la formule:

$$b_{iu} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_i - r_j\right]$$

$$b_{ju} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_j - r_i\right]$$

5. Calculer les distances entre U et les *k* UTO restants:

$$d_{uk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)$$

6. Reprendre la première étape avec n-1 UTO (U remplace i et j).

Exemple précédent :

ÉTAPE 1: matrice de distance d_{ij}

	Homme	Chimpanzée	Gorille	Orang-Outan
Chimpanzée	0,092			
Gorille	0,106	0,111		
Orang-Outan	0,177	0,193	0,188	
Gibbon	0,207	0,218	0,218	0,219

✓ Pour construire un arbre à n UTO, l'algorithme du NJ suit les étapes suivantes:

1. Construire la matrice de distances entre les n UTO ;

2. Calcul de la divergence « r » entre chaque UTO *i* et les autres *k* UTO en suivant la formule:

$$\mathbf{r_i} = \sum \boldsymbol{d_{ik}}$$

3. Construction d'une nouvelle matrice en se basant sur le calcul de la longueur de l'arbre $S_{ij} = (n\text{-}2)\, d_{ij} - r_i - r_j$;

4. Le premier nœud U sera constitué par la paire d'UTO *ij* ayant la plus faible S et la distance arborée *b* de chaque branche est calculé par la formule:

$$b_{iu} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_i - r_j\right]$$

$$b_{ju} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_j - r_i\right]$$

5. Calculer les distances entre U et les *k* UTO restants:

$$d_{uk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)$$

6. Reprendre la première étape avec n-1 UTO (U remplace i et j).

ÉTAPE 2: calcul des r_i

$$r_{Homme} = d_{HC} + d_{HG} + d_{HO} + d_{HGi}$$

$$r_{Homme} = 0{,}092 + 0{,}106 + 0{,}177 + 0{,}207 = 0{,}582$$

$$r_{Chimpanzé} = d_{CH} + d_{CG} + d_{CO} + d_{Cgi} = 0{,}614$$

$$r_{Gorille} = d_{GH} + d_{GC} + d_{GO} + d_{GGi} = 0{,}623$$

$$r_{Orang\text{-}Outan} = d_{OH} + d_{OC} + d_{OG} + d_{OGi} = 0{,}777$$

$$r_{Gibbon} = d_{GiH} + d_{GiC} + d_{GiG} + d_{GiO} = 0{,}862$$

✓ Pour construire un arbre à n UTO, l'algorithme du NJ suit les étapes suivantes:

1. Construire la matrice de distances entre les n UTO ;
2. Calcul de la divergence « r » entre chaque UTO *i* et les autres *k* UTO en suivant la formule:

$$r_i = \sum d_{ik}$$

3. Construction d'une nouvelle matrice en se basant sur le calcul de la longueur de l'arbre $S_{ij} = (n-2)\, d_{ij} - r_i - r_j$;

4. Le premier nœud U sera constitué par la paire d'UTO *ij* ayant la plus faible S et la distance arborée *b* de chaque branche est calculé par la formule:

$$b_{iu} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_i - r_j\right]$$

$$b_{ju} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_j - r_i\right]$$

5. Calculer les distances entre U et les *k* UTO restants:

$$d_{uk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)$$

6. Reprendre la première étape avec n-1 UTO (U remplace i et j).

ÉTAPE 3: calcul de la matrice S_{ij} pour n = 5

$$S_{HC} = (n-2)\, d_{HC} - r_H - r_c = (5-2)\, 0{,}092 - 0{,}582 - 0{,}614$$

$$= -\,0{,}920$$

	Homme	Chimpanzée	Gorille	Orang-Outan
Chimpanzée	- 0,920			
Gorille	- 0,887	- 0,904		
Orang-Outan	- 0,828	- 0,812	- 0,836	
Gibbon	- 0,823	- 0,822	- 0,831	- 0,982

✓ Pour construire un arbre à n UTO, l'algorithme du NJ suit les étapes suivantes:

1. Construire la matrice de distances entre les n UTO ;
2. Calcul de la divergence « r » entre chaque UTO *i* et les autres *k* UTO en suivant la formule:

$$r_i = \sum d_{ik}$$

3. Construction d'une nouvelle matrice en se basant sur le calcul de la longueur de l'arbre $S_{ij} = (n-2)\, d_{ij} - r_i - r_j$;

4. Le premier nœud U sera constitué par la paire d'UTO *ij* ayant la plus faible S et la distance arborée *b* de chaque branche est calculé par la formule:

$$\boldsymbol{b_{iu} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_i - r_j\right]}$$

$$\boldsymbol{b_{ju} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_j - r_i\right]}$$

5. Calculer les distances entre U et les *k* UTO restants:

$$d_{uk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)$$

6. Reprendre la première étape avec n-1 UTO (U remplace i et j).

ÉTAPE 4: le premier nœud U est constitué par la paire ayant la valeur S minimale, c'est-à-dire Orang-Outang/Gibbon

	Homme	Chimpanzée	Gorille	Orang-Outan
Chimpanzée	- 0,920			
Gorille	- 0,887	- 0,904		
Orang-Outan	- 0,828	- 0,812	- 0,836	
Gibbon	- 0,823	- 0,822	- 0,831	- 0,982

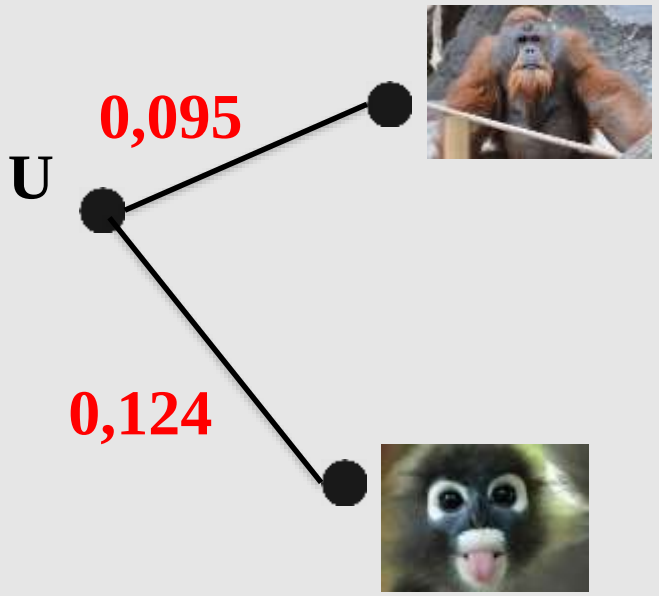
$$b_{OU} = \frac{1}{2\,(n-2)}[(n-2)d_{OGi} + r_O - r_{Gi}] = 0{,}095$$

$$b_{GiU} = \frac{1}{2\,(n-2)}[(n-2)d_{OGi} + r_{Gi} - r_O] = 0{,}124$$

ÉTAPE 4: le premier nœud U est constitué par la paire ayant la valeur S minimale, c'est-à-dire Orang-Outang/Gibbon

$$b_{OU} = \frac{1}{2\,(n-2)}[(n-2)d_{OGi} + r_O - r_{Gi}] = 0{,}095$$

$$b_{GiU} = \frac{1}{2\,(n-2)}[(n-2)d_{OGi} + r_{Gi} - r_O] = 0{,}124$$



✓ Pour construire un arbre à n UTO, l'algorithme du NJ suit les étapes suivantes:

1. Construire la matrice de distances entre les n UTO ;
2. Calcule de la divergence « r » entre chaque UTO *i* et les autres *k* UTO en suivant la formule:

$$r_i = \sum d_{ik}$$

3. Construction d'une nouvelle matrice en se basant sur le calcul de la longueur de l'arbre $S_{ij} = (n-2)\, d_{ij} - r_i - r_j$;

4. Le premier nœud U sera constitué par la paire d'UTO *ij* ayant la plus faible S et la distance arborée *b* de chaque branche est calculé par la formule:

$$b_{iu} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_i - r_j\right]$$

$$b_{ju} = \frac{1}{2\,(n-2)}\left[(n-2)d_{ij} + r_j - r_i\right]$$

5. Calculer les distances entre U et les *k* UTO restants:

$$\boldsymbol{d_{uk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)}$$

6. Reprendre la première étape avec n-1 UTO (U remplace i et j).

ÉTAPE 5: calcul de d_{UH}, d_{UC}, et d_{UG}

$d_{UH} = ½ (d_{GiH} + d_{OH} - d_{GiO})$

$= 0,083$

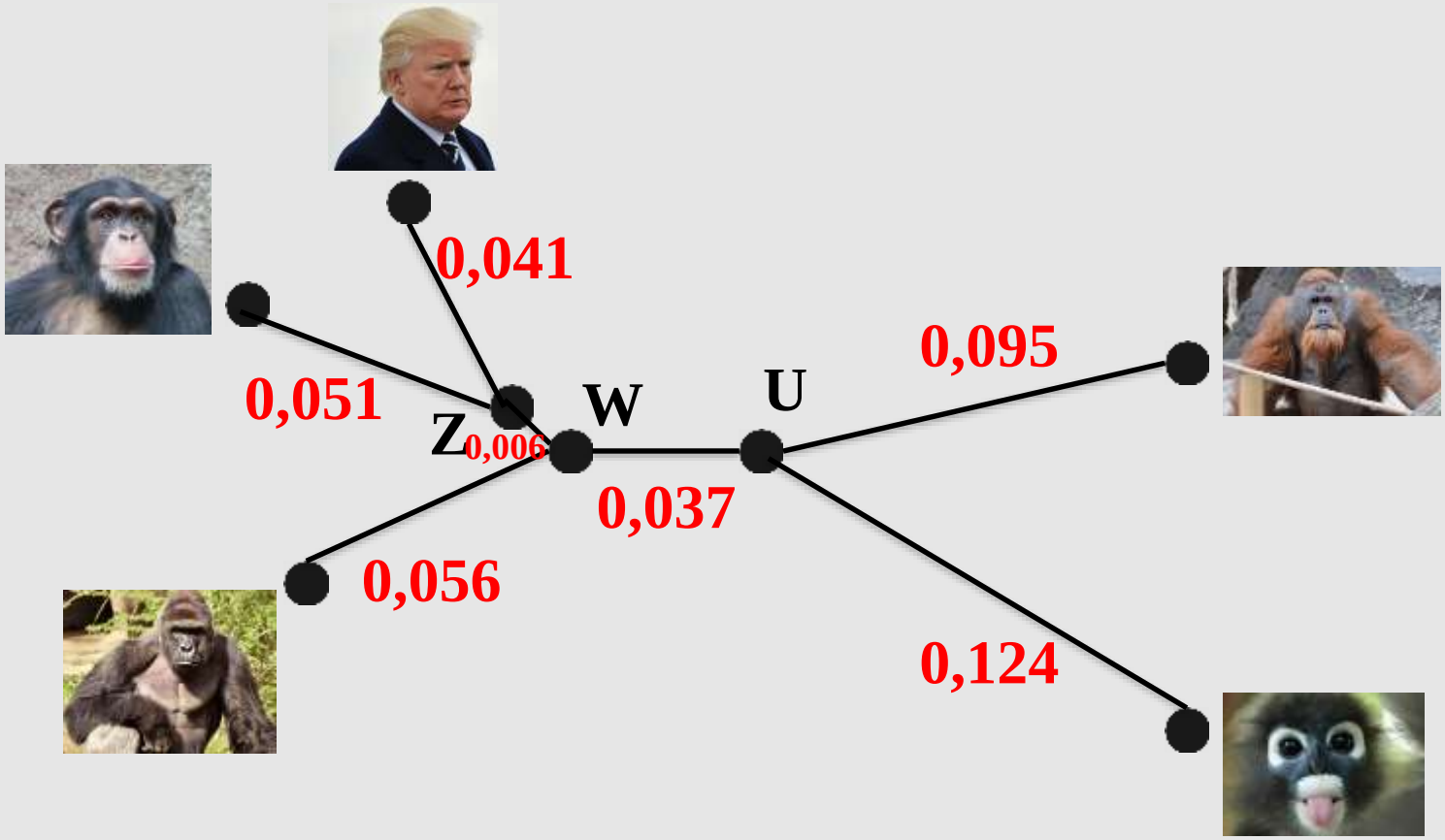
$d_{UC} = 0,096$

$d_{UG} = 0,094$

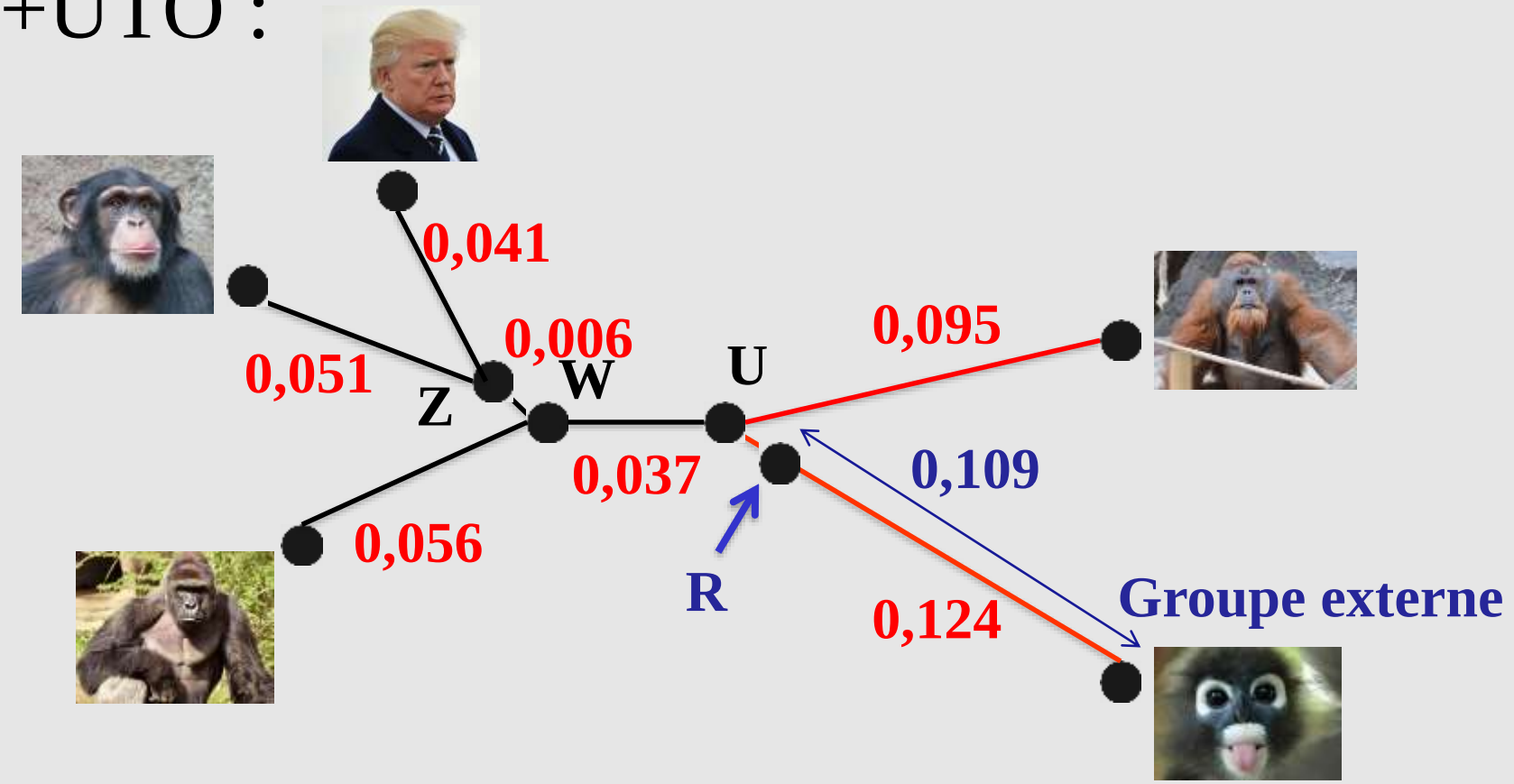
ÉTAPE 6: reprendre un nouveau cycle avec E = {U, Homme, Chimpanzé, Gorille}, n = 4

	Homme	Chimpanzée	Gorille
Chimpanzée	0,092		
Gorille	0,106	0,111	
U	0,083	0,096	0,094

✓NJ produit des arbres non-racinés, qui doivent être racinés par un groupe externe ;

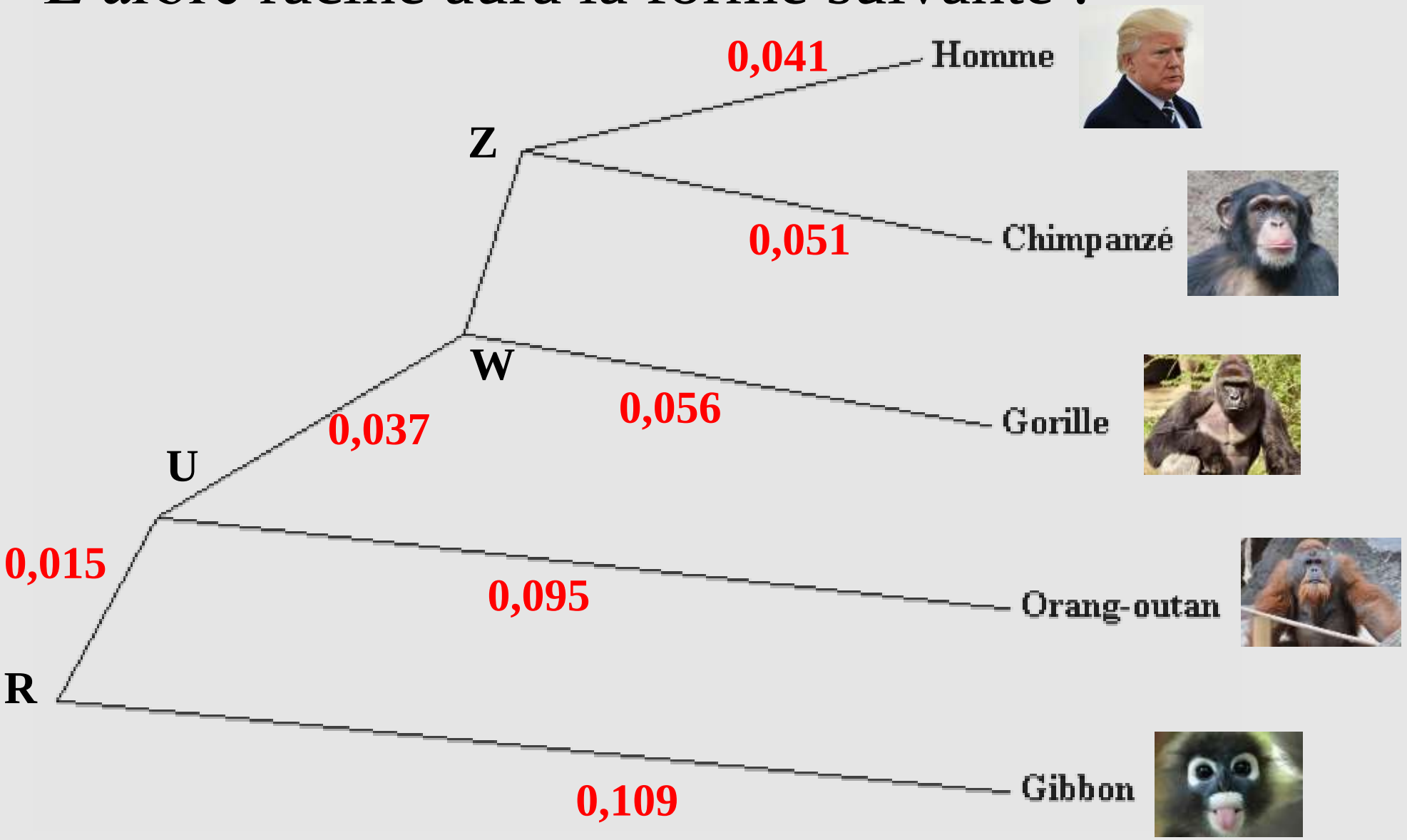


✓La position de la racine sera la médiane de la somme des plus grandes branches groupe externe+UTO :

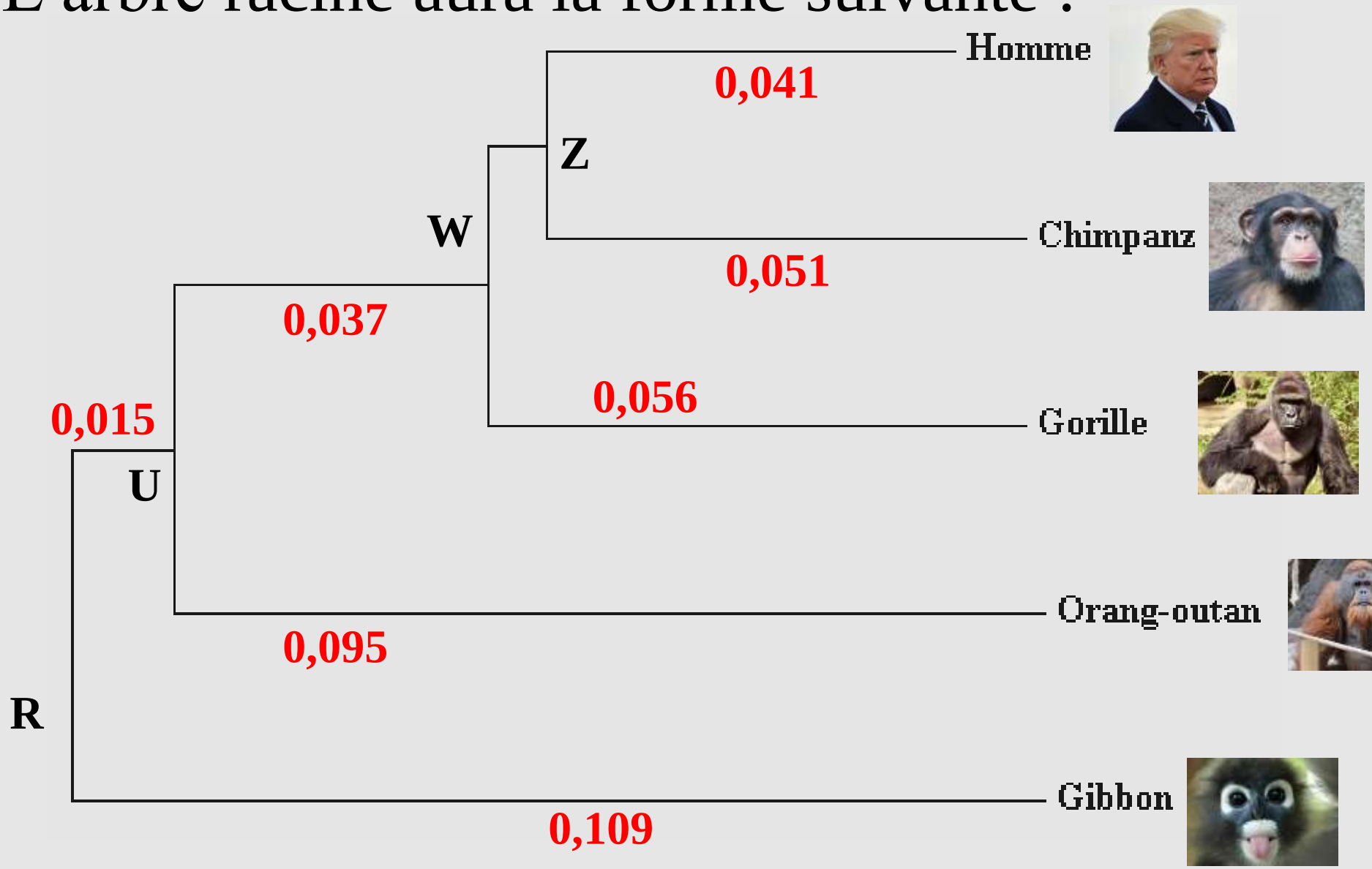


R = (0,124+0,095)/2= 0,109

✓L'arbre raciné aura la forme suivante :

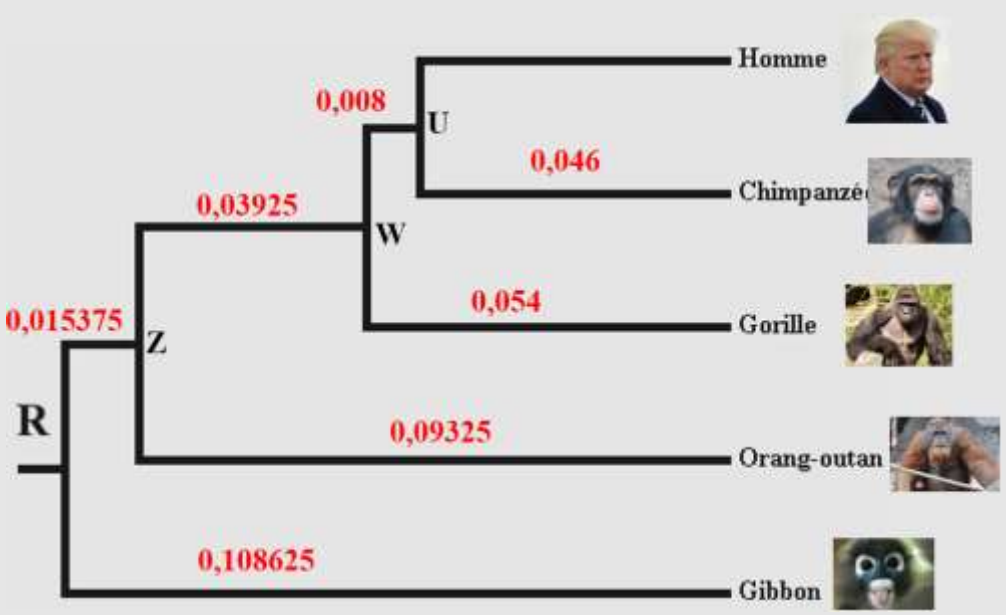


✓L'arbre raciné aura la forme suivante :

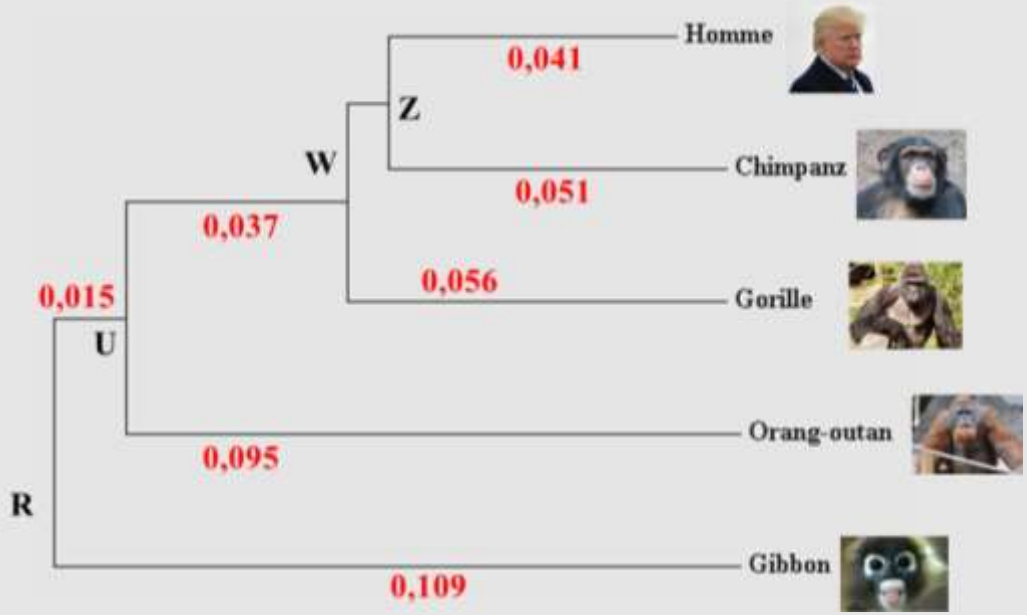


MÉTHODES DE DISTANCES

UPGMA VS NJ

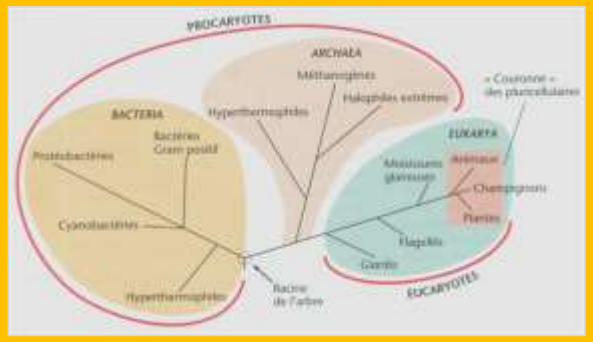


UPGMA

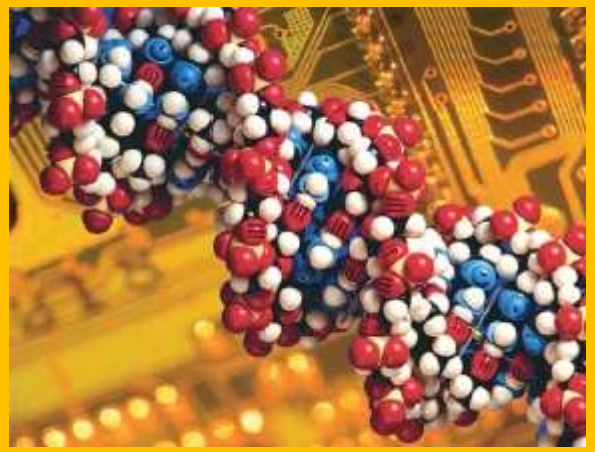


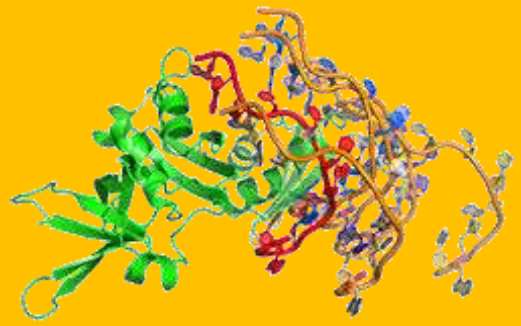
NJ

Université Frères Mentouri Constantine 1
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
1e année Master Biotechnologie alimentaire
2018-2019

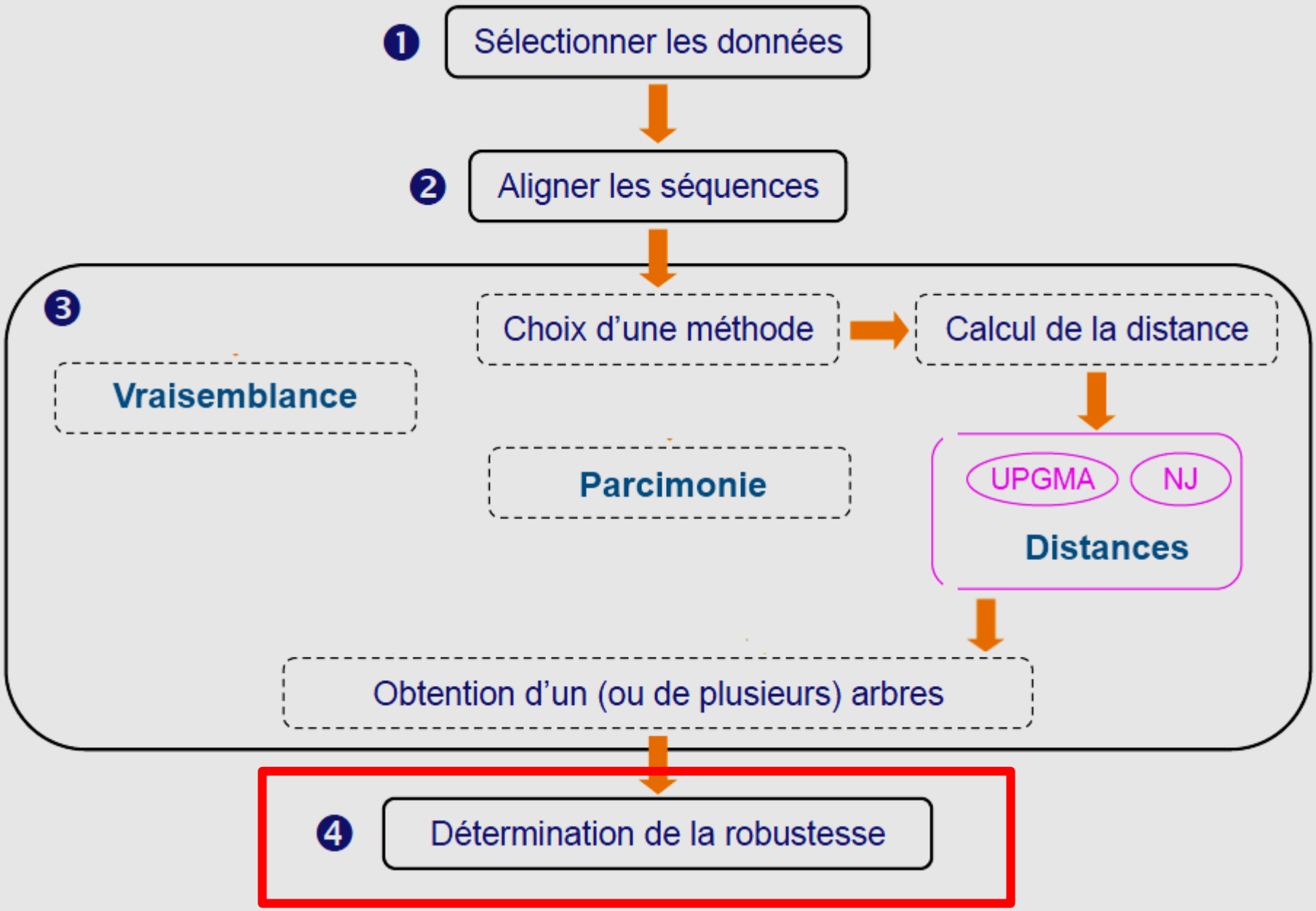


COURS DE BIOINFORMATIQUE





MÉTHODES DE DISTANCES



ROBUSTESSE D'UN ARBRE

- ✓La construction des arbres phylogénétiques est basée sur des hypothèses. Cela induit des erreurs au niveau topologique et conduit à des erreurs d'interprétation qui nécessitent d'être rectifiées.
- ✓La méthode du **bootstrap** est la plus utilisée pour vérifier la **robustesse** de l'arbre obtenu avec telle ou telle méthode de construction.

FIABILITÉ D'UN ARBRE

PRINCIPE DE LA MÉTHODE DU BOOTSTRAP

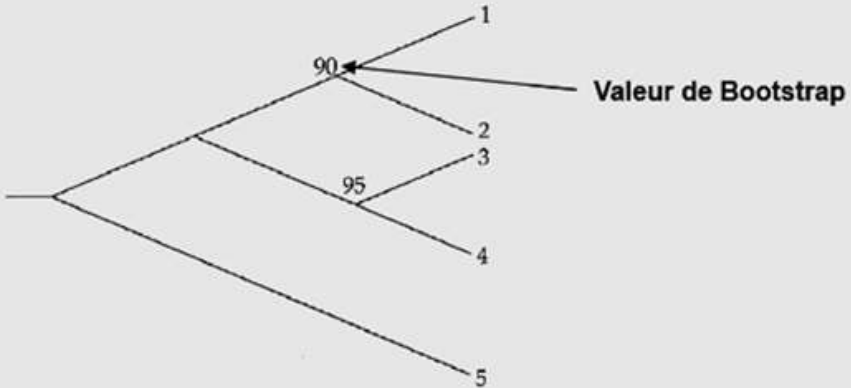
- ✓ On suppose que les caractères évoluent de manière indépendante;
- ✓ On crée un nouvel alignement en changeant pour chaque colonne (site ou position pour une séquence) de manière aléatoire l'état ;
- ✓ On compare l'arbre obtenu (artificiel) avec l'original. On répète au moins 1000 fois l'opération et on calcule pour chaque nœud le pourcentage où il est trouvé dans la même topologie que dans l'arbre original (c'est la **valeur de bootstrap**).

Alignement original

Taxa	Sequence
1	GCAGTACT…
2	GTAGTACT…
3	ACAATACC…
4	ACAACACT…
5	GCGGCATT…

1 permutation (position 1 et 8) = 1 arbre artificiel

	1	8
1	T	CAGTAC G..
2	T	TAGTAC G..
3	C	CAATAC A..
4	T	CAACAC A..
5	T	CGGCAT G..



FIABILITÉ D'UN ARBRE

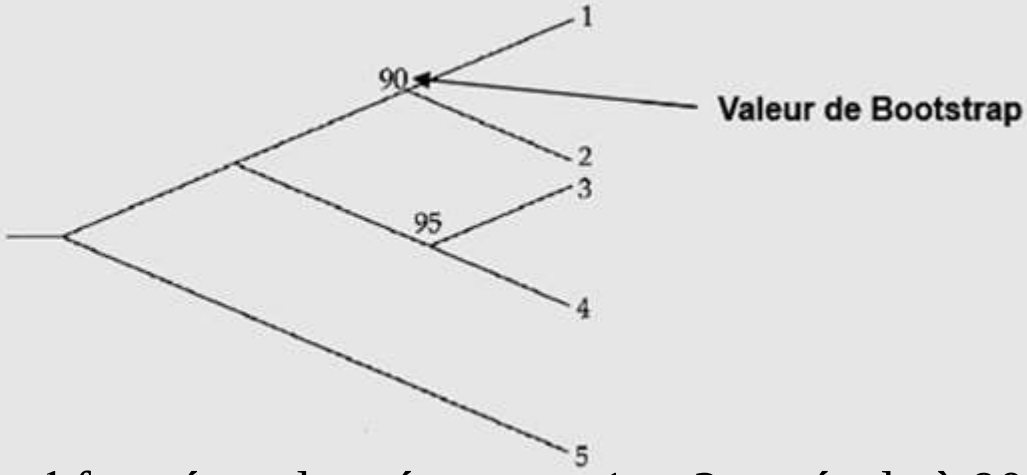
EXEMPLE

Alignement original

Taxa	Sequence
1	GCAGTACT…
2	GTAGTACT…
3	ACAATACC…
4	ACAACACT…
5	GCGGCATT…

1 permutation (position 1 et 8) = 1 arbre artificiel

	1	8
1	T	C AGTAC G..
2	T	T AGTAC G..
3	C	C AATAC A…
4	T	C AACAC A…
5	T	C GGCAT G..



- ✓ Dans cet arbre, la valeur de bootstrap du nœud formé par les séquences 1 et 2 est égale à 90 %. Cela veut dire que sur 100 arbres construits par permutations aléatoires, on retrouve ce nœud ou clade (séquence 1, séquence 2) 90 fois.

- ✓ Les branches de l'arbre original doivent être soutenues par les arbres artificiels par des valeurs de bootstrap ≥ 95% (parfois ≥90% ou ≥ 99% selon la similarité entre les séquences) pour qu'elles soient significatives.

OUTILS DE LA PHYLOGÉNIE

Pour réaliser une étude phylogénétique, de nombreux programmes informatiques sont disponibles. On peut citer par exemple les portail Mobyle (institut Pasteur) et Phylogeny, les programmes MEGA, Phylowin, PHYLOVIZ, Philip, PAUP, ARB, Bionumerics, etc.