

Phylogénie

Dr. Bousbaa Fatima Zohra

f.bousbaaf@cu-aflou.edu.dz
Centre universitaire d'Aflou
Institut des sciences
Département d'Informatique

2024-2025

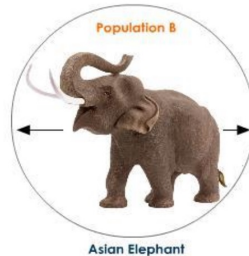
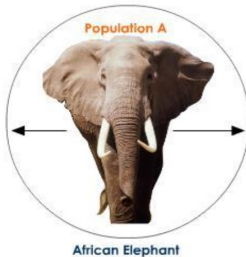


Plan

- 1 Introduction
- 2 Construction des arbres phylogénétiques
- 3 Algorithmes pour la phylogénie moléculaire
 - Méthode de vraisemblance
 - Méthode de parcimonie
 - Méthode de distance
 - Distance simple
 - Distance évolutive
 - Modèle UPGMA

Phylogénie : À quoi ça sert?

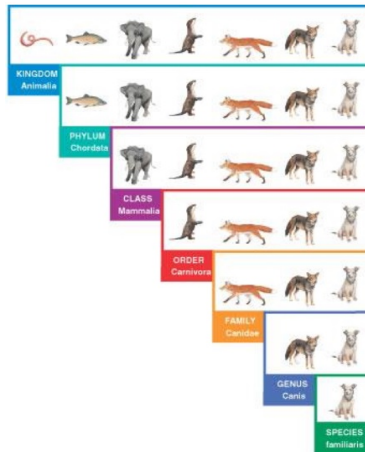
- L'isolement d'une population et l'adaptation à son environnement peut entraîner la création d'une nouvelle espèce.



- Histoire évolutive de familles de gènes :
 - Analyse des duplications et des pertes de gènes.
 - Histoire évolutive des organismes les portant.
- Epidémiologie.

Phylogénie : À quoi ça sert?

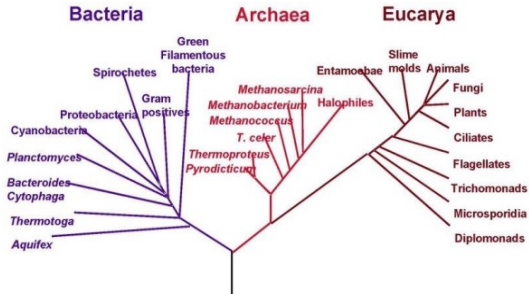
- Étude des relations d'évolution entre des groupes d'organismes (espèces, populations). Basée sur la notion d' "héritage".
- Taxonomie: Science qui consiste à classer, identifier et nommer les organismes. Basée sur des caractéristiques communes, différentes du reste de la diversité biologique.



Phylogénie : À quoi ça sert?

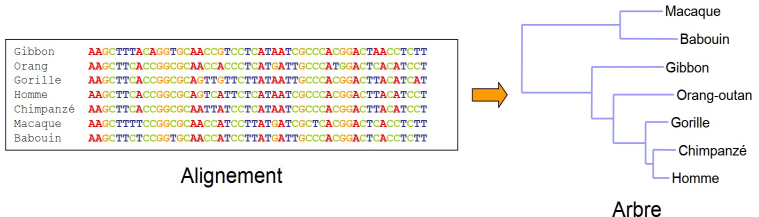
- Premier objectif des études phylogénétiques: Reconstruire l'arbre de vie de toutes les espèces vivantes à partir des données génétiques observées.

Phylogenetic Tree of Life





Phylogénie : Données/Résultats

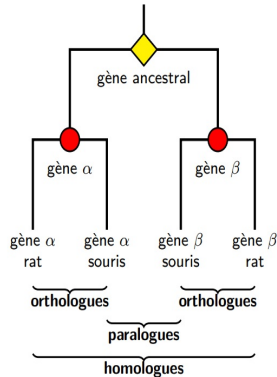
- Données
 - Un ensemble de séquences **homologues** alignées.
 - Chaque position dans l'alignement constitue un site.
- Résultats
 - Un arbre décrivant les relations évolutives entre les séquences (i.e., un arbre phylogénétique).



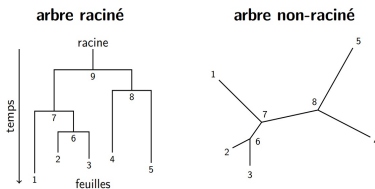
- Un arbre de phylogénie est également utilisé pour représenter l'évolution commune d'une famille de gènes, ou de virus comme le HIV ou l'influenza.

Phylogénie : Homologie

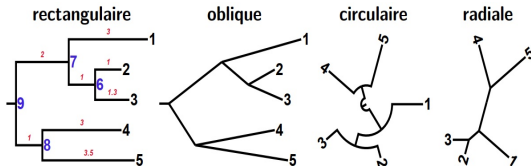
- Deux séquences sont dites *homologues* lorsqu'elles possèdent un ancêtre commun
- Les événements de spéciation  donnent des *orthologues*
- Les événements de duplication  donnent des *paralogues*



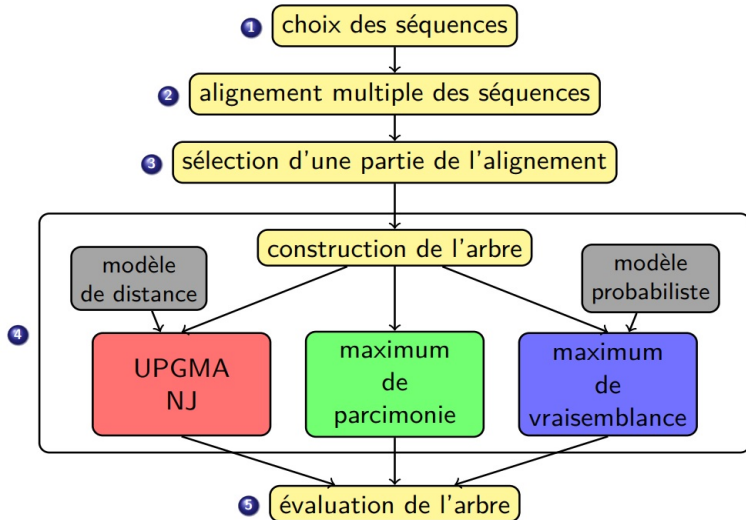
Arbre : Codage et représentations



- Codage : correspondance entre arbres et parenthèses imbriquées
(format Newick) : $((1,(2,3)),(4,5));$
 $((1,(2,3)6)7,(4,5)8)9;$
 $((1:3,(2:1,3:1.3)6:1)7:2,(4:3,5:3.5)8:1)9;$
- Représentation :



Construction des arbres phylogénétiques



Alignement multiple des séquences

- Toutes les approches phylogénétiques moléculaires commencent par un alignement multiple des séquences.

1

beta	MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLGAFSDGLA
delta	MVHLTPEEKTAVNALWGKVN--VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAMGNPKVKAHGKKVLGAFSDGLA
epsilon	MVHFTAEKAAVTSLSWKMN--VEEAGGEALGRLLVVYPWTQRFFDSFGNLSASPAILGNPKVKAHGKKVLTSFGDAIK
gamma	MGHFTAEEDKATITSLWGKVN--VEDAGGETLGRLLVVYPWTQRFFDSFGNLSASAIMGNPKVKAHGKKVLTSFGDAIK
theta	-MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAAPPATKTYFSHL-DLSP-----GSSQVRAGQKVADALSLAVE
alpha	-MVLSPADKTNVKAAMGVGAHAGEYGAELERMFSLFPTTKTYFPHF-DLSH-----GSAQVKGHGKKVADALTNAVA
zeta	-MSLTKTERTIIIVSMWAKISTQADTIGTETLERLFLSLHPQTKTYFPHF-DLHP-----GSAQLRAHGSKVVAAGDAVK
myoglobin	-MGLSDGEWQLVLNVWVKVEADIPGHGQEVLRIRLFKGHPETLEKFKFKHLKSEDEMKASEDLKKHGATVLTALGGILK

80

beta	HLDNLKGTFTATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
delta	HLDNLKGTFTQLSELHCDKLHVDPENFRLLGNVLVCVLAHFGKEFTPPQMAAYQKVVAGVANALAHKYH-----
epsilon	NMDNLKGTFAKLSELHCDKLHVDPENFKLLGNVMVILATHTFGKEFTPEVQAAYQKLVSAVAIALAHKYH-----
gamma	HLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVAIHFGKEFTPEVQASWQKMTAVASALSSRYH-----
theta	RLDDLPHALSALSHLHACQLRVDPASFPQLLGHCLLVTLARHYPGDFSPALQASLDKFLSHVISALVSEYR-----
alpha	HVDDMPNALSALSDLHAHKLVDVFNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----
zeta	SIDDIGGALSSELHAYILRVDPVNFKLLSHCLLVTLAARFPADFTAEAAHAWDKFLSVVSVLTKYR-----
myoglobin	KKGHHEAEIKPLAQSHATKKKIPVKYLEFISECIIQVLQSKHPGDFGADAQGMNKALELFRKDMASNYKELGFGQ

- Les zones de faible similarité sont ignorées.

Méthode de vraisemblance

- Le but de la méthode de maximum de vraisemblance est d'identifier un grand nombre de scénarios évolutifs possibles c'est-à-dire de trouver les valeurs des paramètres qui maximisent la probabilité d'observer les séquences.
- C'est une méthode très lourde en calculs.
- Il est presque toujours impossible d'évaluer tous les arbres possibles car ils sont trop nombreux.

Méthode de parcimonie

Méthode générale

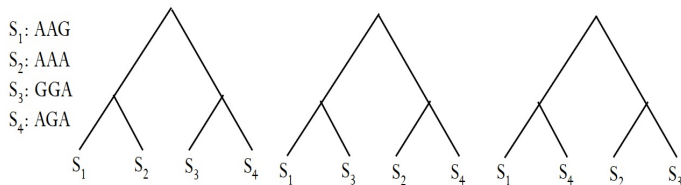
- Considérer toutes les topologies d'arbres possibles sur un ensemble de feuilles.
- Calculer un poids pour chaque arbre.
- Sélectionner un arbre de poids minimal.

Méthode de parcimonie

Pondération d'un arbre

Affecter des séquences aux noeuds internes de telle sorte à minimiser le poids total de l'arbre (somme des distances des branches).

Exemple

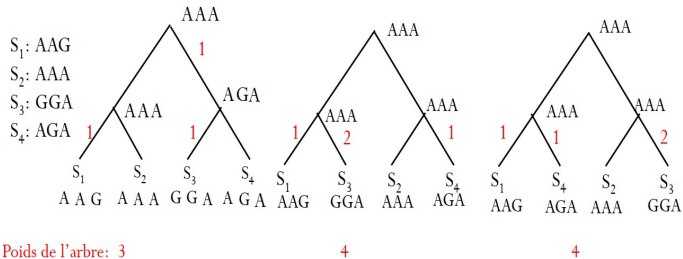


Méthode de parcimonie

Pondération d'un arbre

Affecter des séquences aux noeuds internes de telle sorte à minimiser le poids total de l'arbre (somme des distances des branches).

Exemple

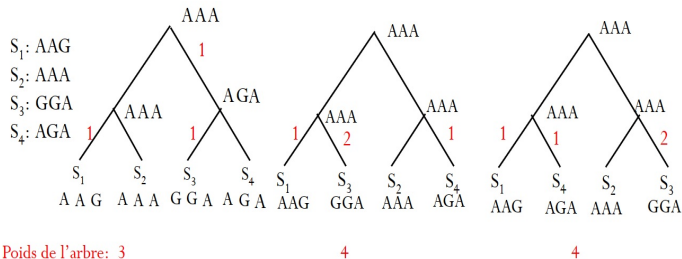


Méthode de parcimonie

Pondération d'un arbre

Affecter des séquences aux noeuds internes de telle sorte à minimiser le poids total de l'arbre (somme des distances des branches).

Exemple



- Pas conseillée pour la construction phylogénétique à partir des séquences.

Distance simple

- La **p-distance** est l'estimation la plus simple de la distance entre deux séquences :

$$p = \frac{n}{l}$$

avec n le nombre de substitutions et l le nombre de sites.

Séquence	1	2	3	4	5	6	7	8	9	10
I	A	T	A	T	A	C	G	T	A	T
II	A	T	G	T	A	C	G	T	A	T
III	G	T	A	-	A	C	G	T	G	C
IV	G	C	G	T	A	T	G	C	A	C

Matrice de distances

	I	II	III	IV
I	-	0.1	0.4	0.6
II		-	0.5	0.5
III			-	0.6
IV				-

Distance évolutive

- La distance évolutive d (**d-distance**) est supérieure à la distance observée p .

Séquence1 : GAAAAG

Séquence2 : ATGAAG

substitution	Séquence 1	Séquence 2	p	d
Simple	G	G \rightarrow A	1	1
Multiples	A	A \rightarrow C \rightarrow T	1	2
Coïncidentes	T \rightarrow A	T \rightarrow G	1	2
Parallèles	T \rightarrow A	T \rightarrow A	0	1
Convergentes	C \rightarrow G \rightarrow A	C \rightarrow A	0	3
Inverse	G \rightarrow T \rightarrow G	G	0	2

$$p = \frac{3}{6} = 50\%$$

$$d = \frac{11}{6} = 183\%$$

Modèle de distance

Construction de l'arbre

- Agglomération des taxons depuis les paires les plus proches jusqu'aux plus éloignées, (ex : UPGMA, NJ).

Modèle UPGMA

Unweighted Pair Group Method of Arithmetic averages

- On sélectionne la distance la plus courte (ici par exemple d_{12}).
- On agglomère 1 et 2 en un seul taxon, et on calcule les nouvelles distances :

$$dn_{12} = \frac{dn_1 + dn_2}{2}$$

Taxon	1	2	3
2	3		
3	5	6	
4	6	7	7

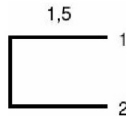
Modèle UPGMA

Unweighted Pair Group Method of Arithmetic averages

- On sélectionne la distance la plus courte (ici par exemple d_{12}).
- On agglomère 1 et 2 en un seul taxon, et on calcule les nouvelles distances :

$$dn_{12} = \frac{dn_1 + dn_2}{2}$$

Taxon	1+2	3
3	5,5	
4	6,5	7

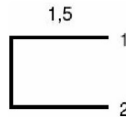


Modèle UPGMA

Unweighted Pair Group Method of Arithmetic averages

- On sélectionne à nouveau la distance la plus courte (ici par exemple $d_{(12)3}$)

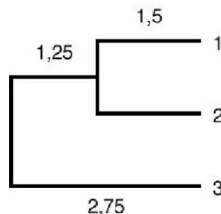
Taxon	1+2	3
3	5,5	
4	6,5	7



Modèle UPGMA

Unweighted Pair Group Method of Arithmetic averages

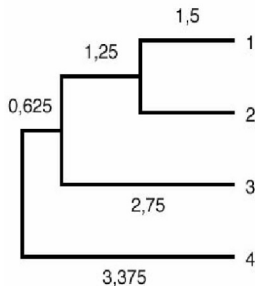
Taxon	12	3
3	5,5	
4	6,5	7



- Ainsi de suite jusqu'à l'obtention de l'arbre final (ici $d_{(123)4} = 6,75$)

Modèle UPGMA

Unweighted Pair Group Method of Arithmetic averages



Avantages de la méthode de distances

Avantages

- Rapidité des méthodes d'agglomération (essentiel avec de grands nombres de taxons).
- Elle peut être appliquée sur n'importe quel type de distances évolutives.
- Méthode performante car elle est équilibré entre rapidité et efficacité.