

Introduction à la phylogénie

Marine Djaffardjy (marine.djaffarjdy@u-psud.fr)

Sarah Cohen Boulakia



Comprendre le monde,
construire l'avenir



Matériel de cours :
Théophile Sanchez

Déroulement des séances

- 09/03 :
 - Introduction à la phylogénie
 - TP Phylogénie
 - Introduction à Python
- 16/03 :
 - Début du projet

Les origines de la phylogénie

La classification du vivant

IV^e siècle av. J.-C. :

- *Histoire des plantes* de Théophraste
- *Histoire des animaux* d'Aristote

1735 : *Systema Naturæ* de Carl von Linné

Règne	<i>Animalia</i>
Embranchement	<i>Chordata</i>
Classe	<i>Mammalia</i>
Ordre	<i>Primates</i>
Famille	<i>Hominidae</i>
Genre	<i>Homo</i>
Espèce	<i>Sapiens</i>

Classification d'*Homo sapiens*

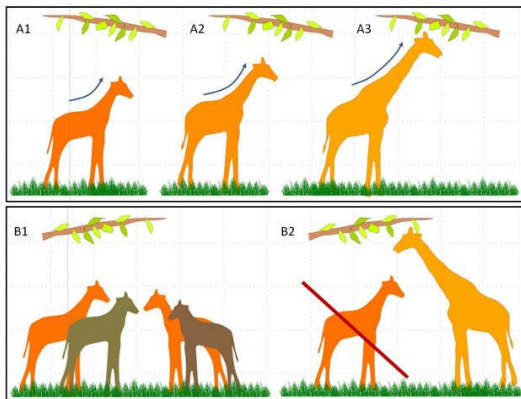


Les origines de la phylogénie

Comment les espèces évoluent-elles ?

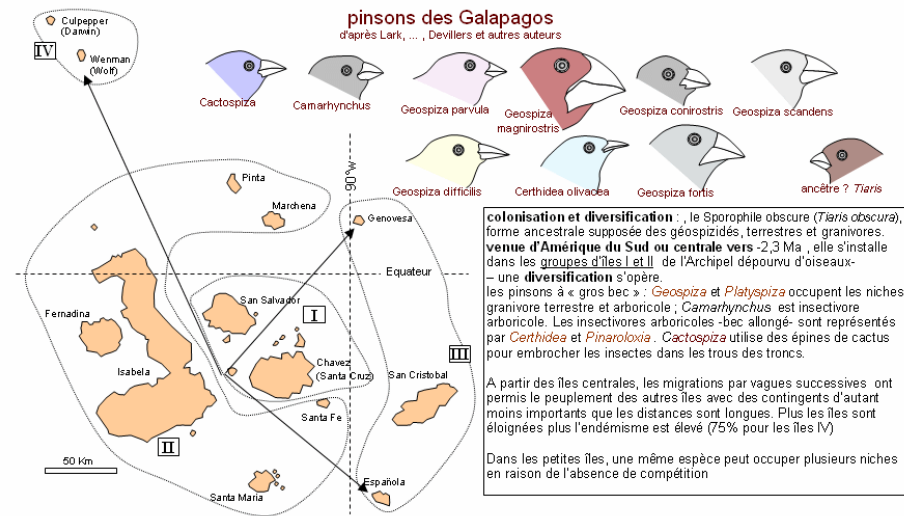
A : Transformisme selon Lamarck

B : Évolution selon Darwin



Bregliano 2017

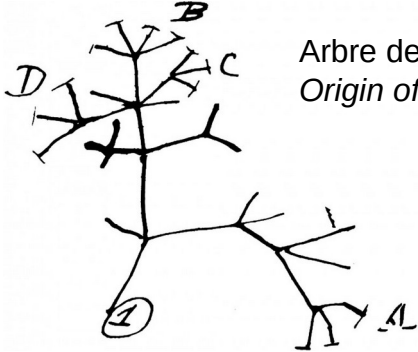
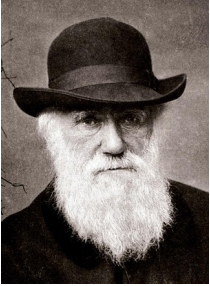
La seule classification valable est généalogique



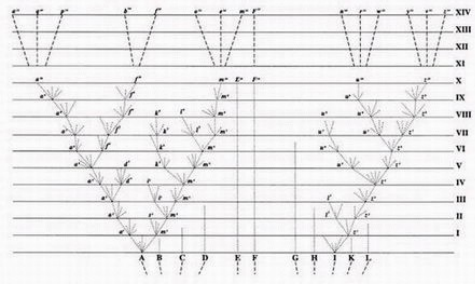
Académie de Dijon

Les origines de la phylogénie

Étude des relations de parentés entre les êtres vivants.



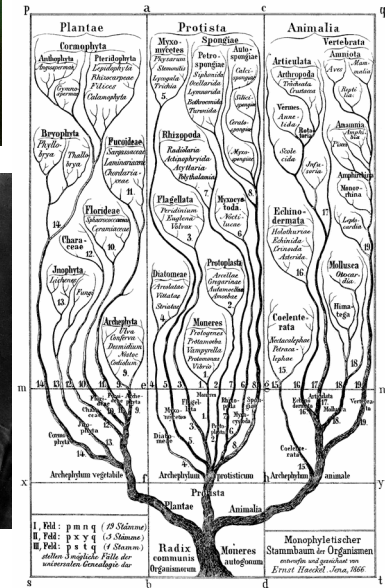
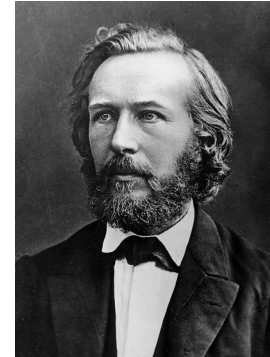
Arbre de la vie publié par Darwin dans *On the Origin of Species by Natural Selection* - 1859



Arbre dessiné par Darwin dans ses notes - 1837



Illustration issue de *Kunstformen der Natur* d'Haeckel

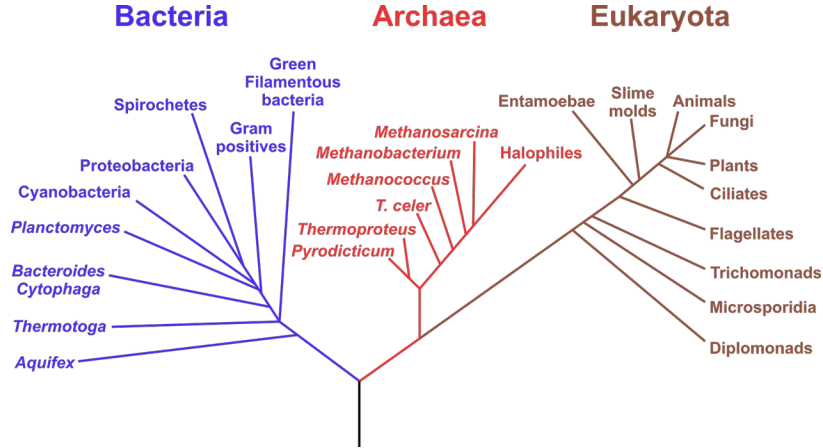


L'arbre de la vie de Haeckel - 1866

Phylogénie et généalogie

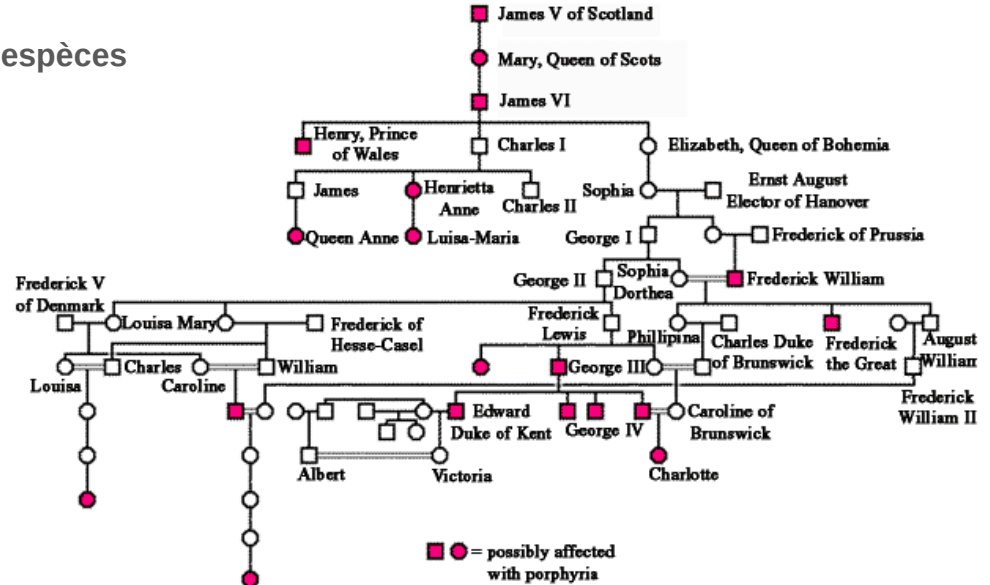
Phylogénie : qui est le plus proche de qui ?

- échelle des **individus**, des **populations** ou des **espèces**



Arbre phylogénétique du vivant

Généalogie : qui descend de qui ?

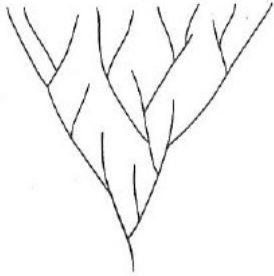


Arbre généalogique de la famille royale d'Angleterre

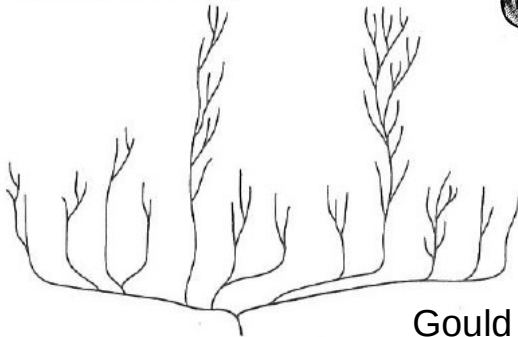
Remarques sur les arbres phylogénétiques

La représentation en cône fait souvent oublier que la grande majorité des espèces n'existent plus aujourd'hui.

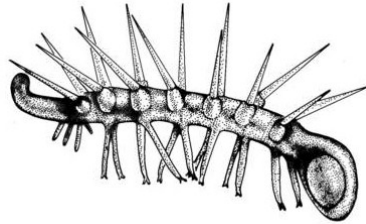
Cône de diversité croissante



Diversification et décimation



Gould



Hallucigenia

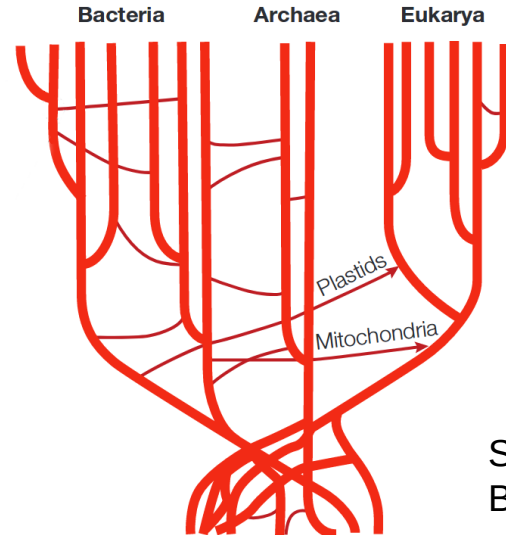
Deux branches d'un arbre peuvent être reliées en cas de transferts horizontaux ou d'hybridation.

OPEN ACCESS Freely available online

PLOS GENETICS

The Date of Interbreeding between Neandertals and Modern Humans

Sriram Sankararaman^{1,2*}, Nick Patterson², Heng Li², Svante Pääbo^{3*}, David Reich^{1,2*}



Smets et
Barkay

Common ancestral community of primitive cells

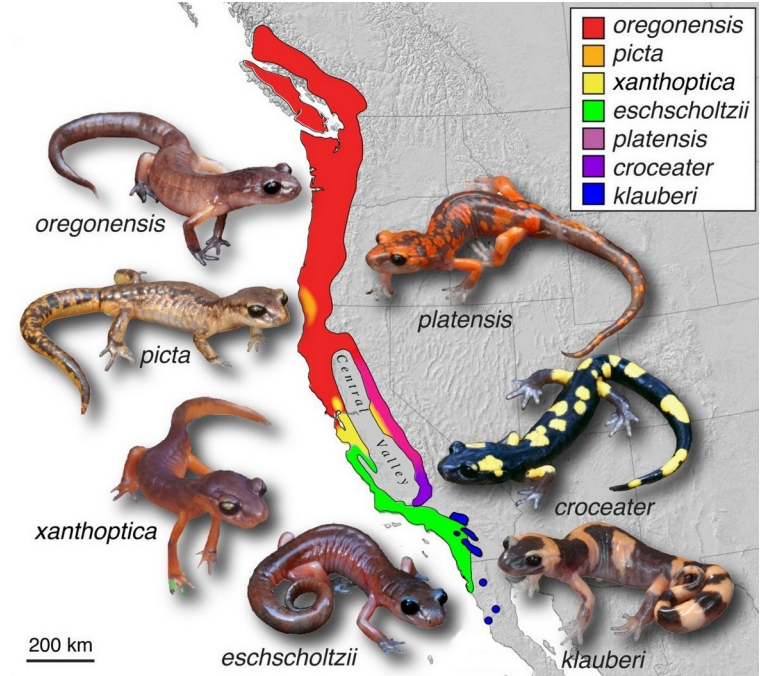
Les données

On peut utiliser plusieurs types de données en phylogénie :

- Les données phénotypiques (morphologie, protéines etc...)
- Les données moléculaires (séquences ADN, ARN et protéiques)

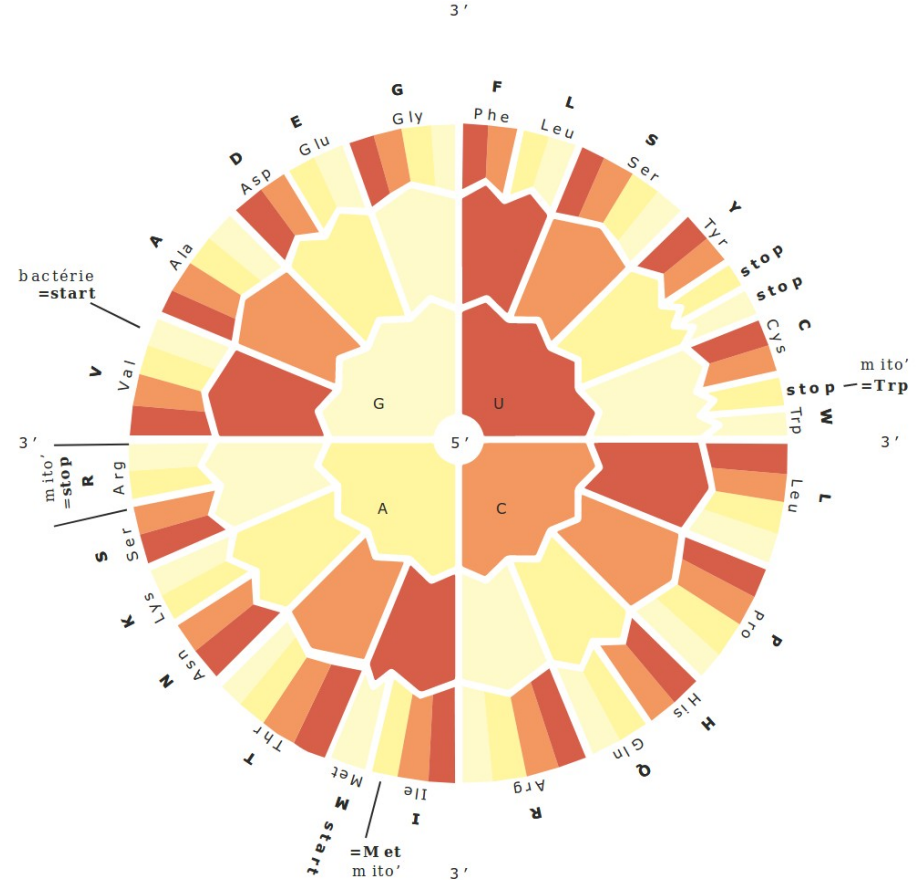
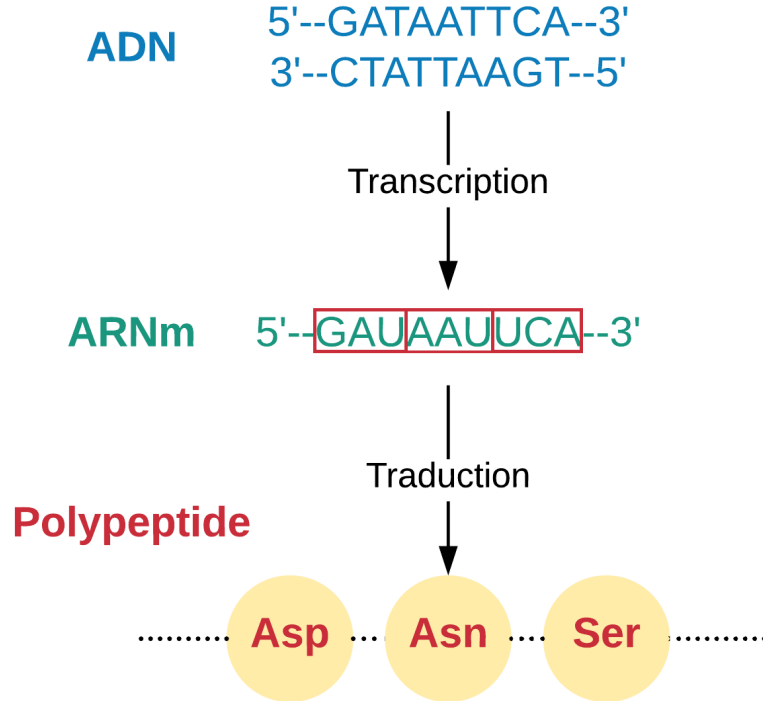
Mais aussi :

- Les données spatiales (répartition des organismes)
- Les données culturelles (notamment pour les populations humaines)
- ...



Stebbins

La synthèse des protéines



Les mutations

Les mutations ponctuelles :

- **substitution** : remplacement d'un nucléotides par un autre
- **insertion** : ajout d'un ou plusieurs nucléotides
- **délétion** : perte d'un ou plusieurs nucléotides

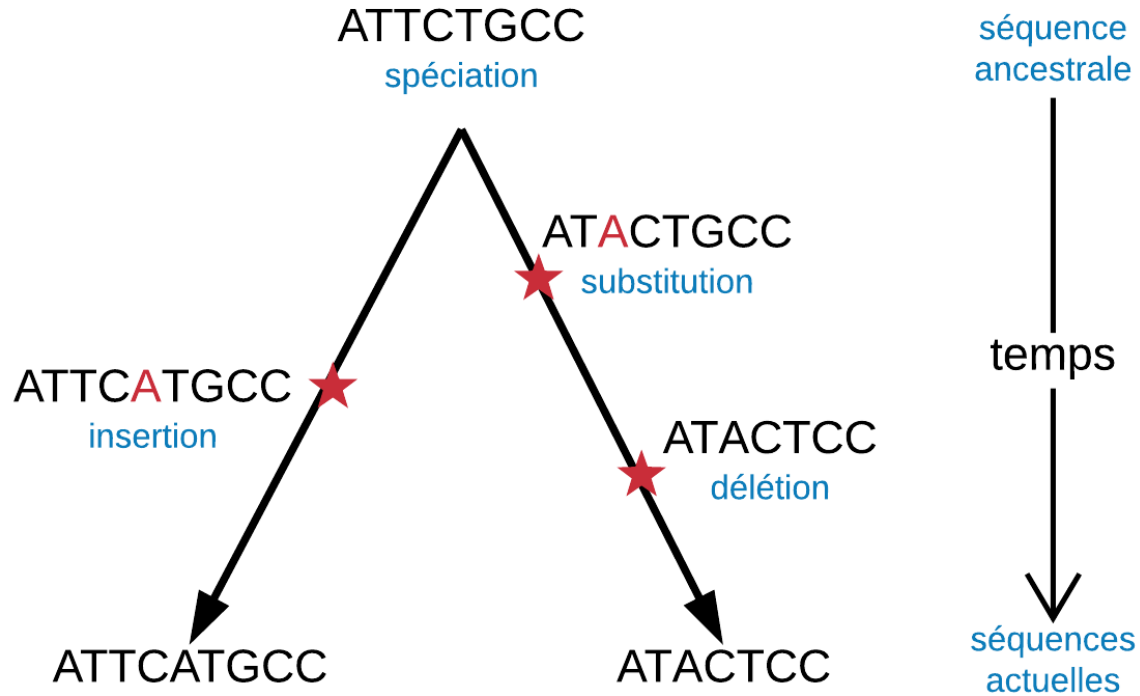
Une insertion ou une délétion crée un décalage du cadre de lecture et peut modifier complètement le peptide traduit. Elles sont donc plus rare que les substitutions.

Une mutation est **silencieuse** si elle ne modifie pas la protéine traduite.

Les mutations chromosomiques :

- **duplication**, inversion, translocation, délétion, insertion

Évolution des séquences



Alignement

L'alignement comme représentation de l'histoire évolutive des séquences :



Les insertions/délétions sont regroupées sous le terme **indel**. Elles introduisent des **gaps** dans l'alignement.

Les substitutions introduisent des **mismatches** dans l'alignement.

Alignement

Si on veut aligner les séquences (A T T C A T G C C) et (A T A C T C C), il existe de nombreuses possibilités :

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

5 matches
2 mismatches
2 gap

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

6 matches
1 mismatches
2 gap

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

5 matches
1 mismatches
4 gap

Alignement

On introduit un système de score :

match = +2
mismatch = -1
gap = -2

Score de similarité = \sum scores élémentaires

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

5 matches
2 mismatches
2 gap

Score = 4

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

6 matches
1 mismatches
2 gap

Score = 7

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

5 matches
1 mismatches
4 gap

Score = 1

Lacroix

Alignement global vs local

Alignement **global** :

- On prend deux séquences et on cherche le meilleur alignement sur l'ensemble des deux séquences.
- Exemple : comparer deux séquences d'ADN entre deux individus proches.

Alignement **local** :

- On cherche un très bon alignement, mais on accepte qu'il ne concerne qu'une sous-partie des deux séquences.
- Exemple : comparer les séquences d'ADN de l'Homme et de la mouche

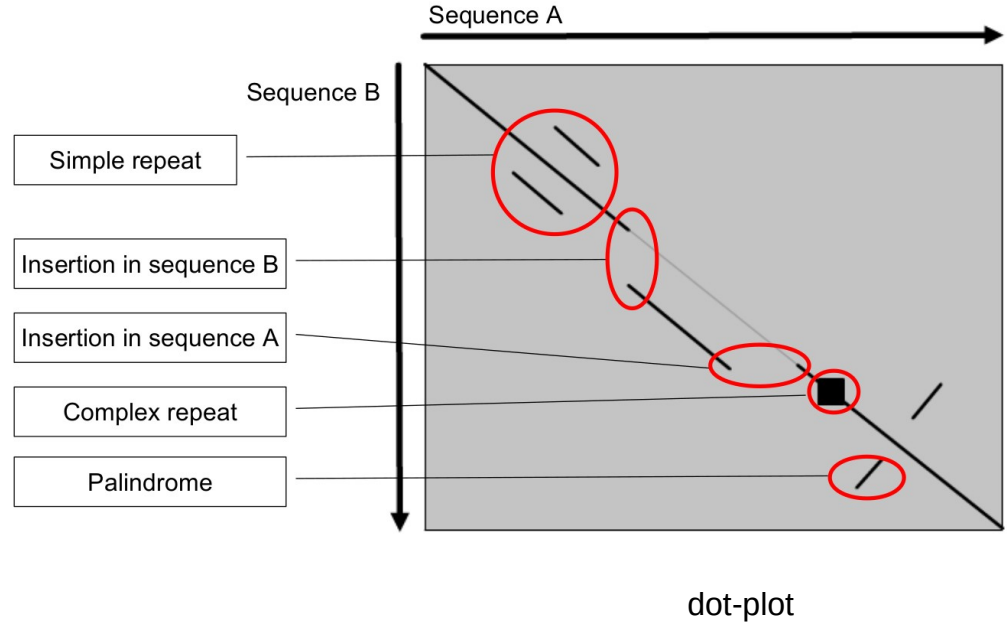
Liste d'algorithmes d'alignements

Alignement par paire :

- Needleman-Wunsch
- Smith-Waterman : pour l'alignement local
- Dot-plot
- ...

Alignement multiple :

- Clustal
- Bowtie
- BWA
- Muscle
- ...



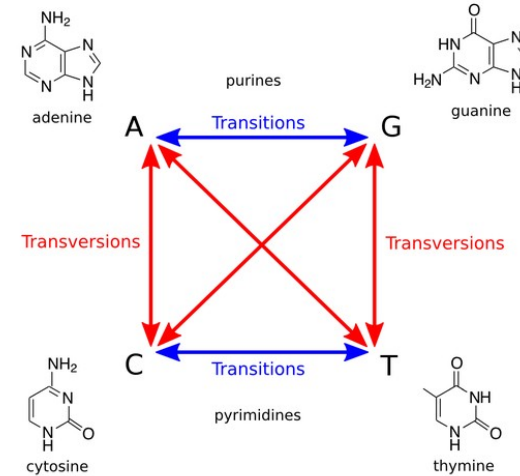
Blast pour la recherche de séquences similaires dans une BD

Needleman-Wunsch

Needleman-Wunsch est un algorithme d'alignement local souvent utilisé pour aligner deux séquences. La première étape est de choisir une matrice de substitution qui associe un score pour chaque substitutions. Il faut aussi choisir une pénalité en cas d'indel. Ici, on va comparer des séquences d'ADN, mais il est possible d'appliquer le même algorithme sur des séquences peptidiques en utilisant des matrices de substitutions adaptées (PAM et BLOSUM).

On choisit une pénalité de *gap* de -1. La matrice suivante pénalise moins les transitions que les transversions car les transitions sont plus fréquentes à cause des propriétés chimiques des bases azotées :

	A	T	G	C
A	3	0	1	0
T	0	3	0	1
G	1	0	3	0
C	0	1	0	3




Needleman-Wunsch

La première matrice est la matrice de score M, la seconde est la matrice de *traceback* T. “l” correspond à *left* et “u” à *up* :

		A	T	T	C	A	T	G	C	C
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1									
T	-2									
A	-3									
C	-4									
T	-5									
C	-6									
C	-7									

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

Needleman-Wunsch

		A	T	T	C	A	T	G	C	C
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	3	2							
T	-2									
A	-3									
C	-4									
T	-5									
C	-6									
C	-7									

On remplit la matrice en suivant la formule :

$$M_{ij} = \max \begin{cases} M_{i-1j-1} + s(A_i, B_j) \\ M_{i-1j} + s(A_i, gap) \\ M_{ij-1} + s(B_j, gap) \end{cases}$$

avec la fonction s qui calcule le score entre deux nucléotides à partir de la matrice de substitution.

Needleman-Wunsch

		A	T	T	C	A	T	G	C	C
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	3	2	1	0	-1	-2	-3	-4	-5
T	-2	2	6	5	4	3	2	1	0	-1
A	-3	1	5	6	5	7	6	5	4	3
C	-4	0	4	6	9	8	8	7	8	7
T	-5	-1	3	7	8	9	11	10	9	9
C	-6	-2	2	6	10	9	10	11	13	12
C	-7	-3	1	5	9	10	10	10	14	16

On remplit la matrice en suivant la formule :

$$M_{ij} = \max \begin{cases} M_{i-1j-1} + s(A_i, B_j) \\ M_{i-1j} + s(A_i, gap) \\ M_{ij-1} + s(B_j, gap) \end{cases}$$

avec la fonction s qui calcule le score entre deux nucléotides à partir de la matrice de substitution.

Needleman-Wunsch

La matrice de *traceback* T est remplie en fonction de quelle case a été choisie pour calculer M_{ij} (M_{i-1j-1} , M_{i-1j} ou M_{ij-1}). “l” correspond à *left*, “u” à *up* et “d” à *diagonal*:

$$M_{ij} = \max \begin{cases} M_{i-1j-1} + s(A_i, B_j) \\ M_{i-1j} + s(A_i, gap) \\ M_{ij-1} + s(B_j, gap) \end{cases}$$

		A	T	T	C	A	T	G	C	C
		l	l	l	l	l	l	l	l	l
A	u	d	l	l	l	d	d	l	l	l
T	u	u	d	d	l	l	l	l	l	l
A	u	d	u	d	d	d	d	l	l	l
C	u	u	u	d	d	l	d	l	d	d
T	u	u	d	d	u	d	d	l	l	d
C	u	u	u	u	d	l	d	d	d	d
C	u	u	u	u	d	d	d	d	d	d

Needleman-Wunsch

On peut maintenant procéder au *traceback* en partant de la dernière case. Si on rencontre un “l”, on ajoute un *gap* dans la séquence A. Si on rencontre un “u”, on ajoute un *gap* dans la séquence B.

On obtient finalement l'alignement suivant:

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

		A	T	T	C	A	T	G	C	C
		l	l	l	l	l	l	l	l	l
A	u	d	l	l	l	d	d	l	l	l
T	u	u	d	d	l	l	l	l	l	l
A	u	d	u	d	d	d	d	l	l	l
C	u	u	u	d	d	l	d	l	d	d
T	u	u	d	d	u	d	d	l	l	d
C	u	u	u	u	d	l	d	d	d	d
C	u	u	u	u	d	d	d	d	d	d

Les arbres phylogénétiques

Noeuds : Unités Taxonomiques (UT)

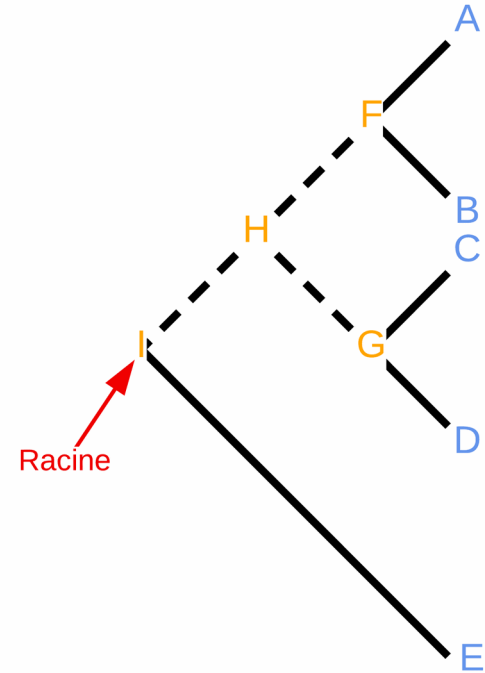
- Opérationnelles (UTO) : A, B, C, D, E
- Hypothétiques (UTH) : F, G, H, I

Branches :

- Internes : succession d'organismes reliant deux UTH
- Externes : succession d'organismes reliant les UTH aux UTO

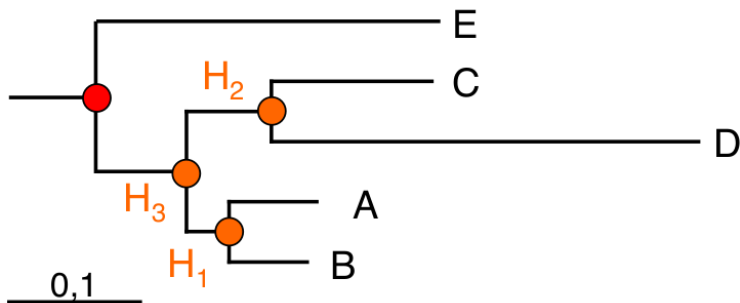
Topologie : Forme de l'arbre

Racine : Ancêtre commun le plus récent de tous les UTO



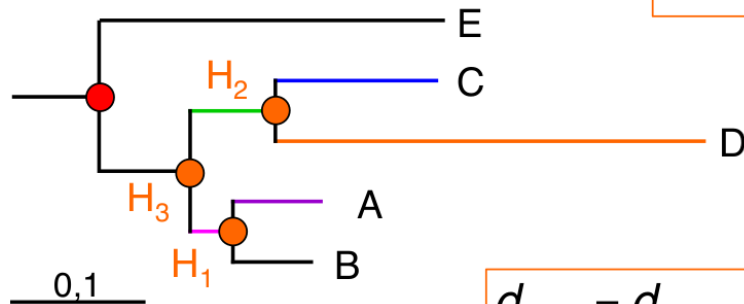
Adapté de Brochier-Armanet
(frangun.org)

Apparentement et similarité



- B est plus apparenté à A qu'à C, D ou E
- B est apparenté de manière égale à C et D
- B est plus apparenté à C ou D qu'à E
- C est plus apparenté à D qu'à A, B ou E
- C est plus apparenté A ou B qu'à E
- E est aussi apparenté à A, B, C ou D
- A, B, C, D et E sont apparentés de manière égale à leur ancêtre commun

Apparemment et similarité



... mais C est plus similaire à A qu'à D

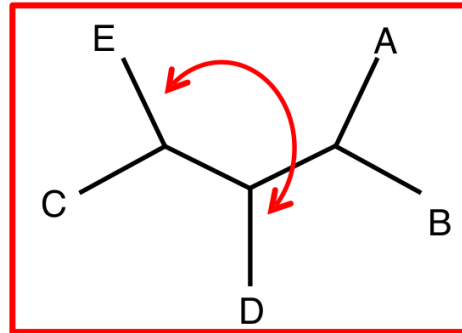
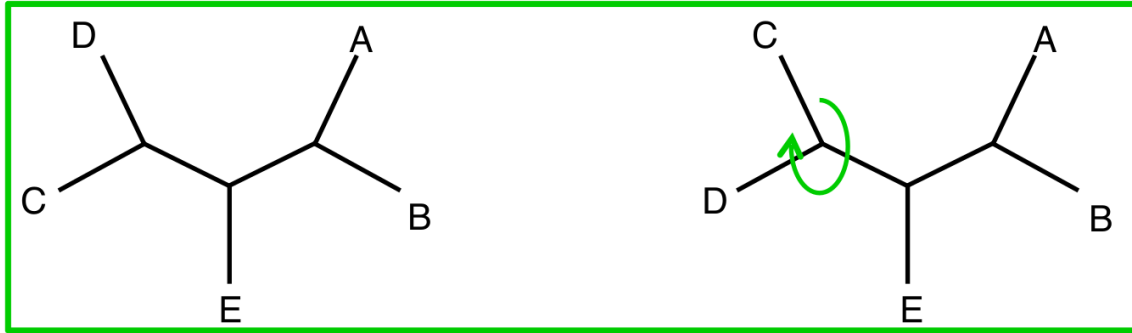


$$d_{C-D} = d_{C-H_2} + d_{D-H_2} > d_{C-A} = d_{C-H_2} + d_{H_2-H_3} + d_{H_3-H_1} + d_{H_1-A}$$

- B est plus apparenté à A qu'à C, D ou E
- B est apparenté de manière égale à C et D
- B est plus apparenté à C ou D qu'à E
- C est plus apparenté à D qu'à A, B ou E
- C est plus apparenté A ou B qu'à E
- E est aussi apparenté à A, B, C ou D
- A, B, C, D et E sont apparentés de manière égale à leur ancêtre commun

Topologie

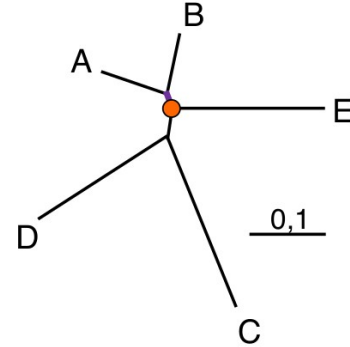
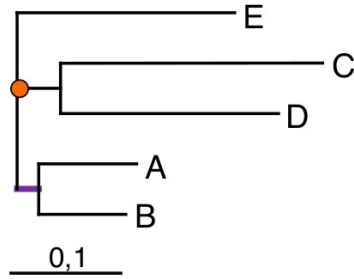
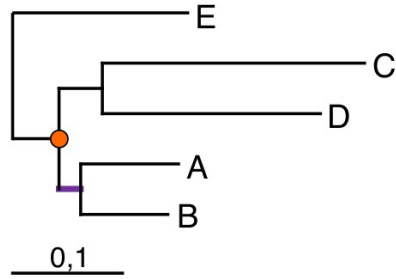
Deux branches sœurs peuvent pivoter librement autour du nœud qui les connecte.



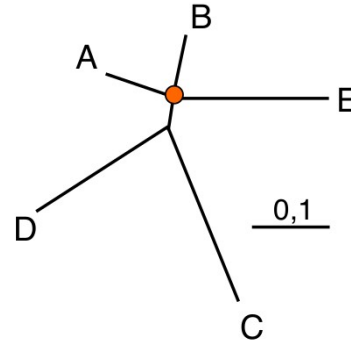
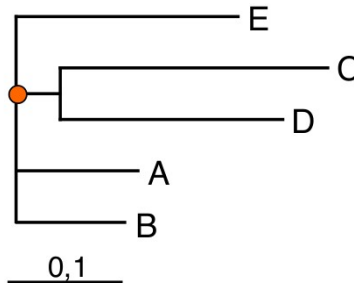
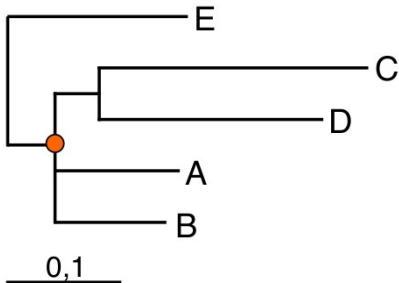
Cet arbre est différent
des deux précédents

Arbres résolus et arbres multifurqués

Arbres résolus



Arbres multifurqués

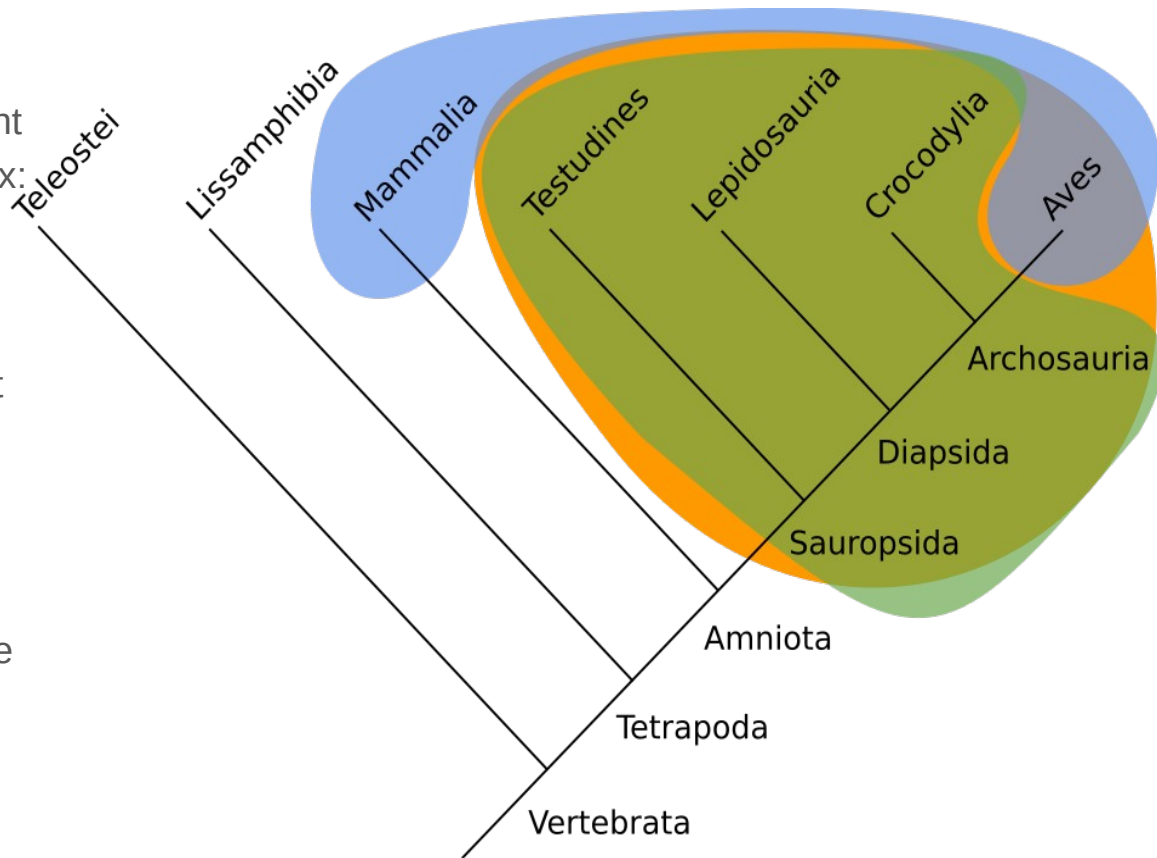


Groupe mono-, para- et poly-phylétiques

Monophylétique : Groupe contenant une espèce et **tous** ses descendants (ex: oiseaux et reptiles qui forment les sauropsides)

Paraphylétique : Groupe contenant une espèce et ses descendants **sauf quelques uns** (ex: les reptiles)

Polyphylétique : Groupe contenant plusieurs espèces mais pas leur ancêtre commun (ex: animaux à sang chaud)

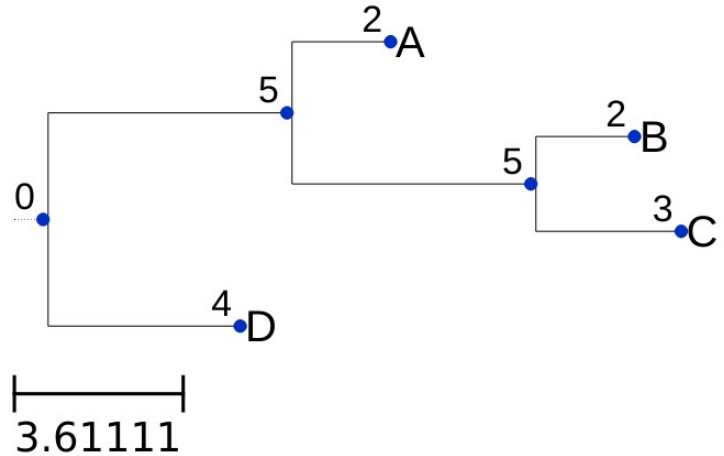


Le format Newick

Les groupes ayant la même racine sont écrit entre parenthèses et séparés par des virgules. Un groupe peut être soit une feuille de l'arbre, soit un autre groupe. La longueur de la branche de chaque groupe est écrite après un double point et l'arbre est terminé par un point virgule.

Exemple :

`((A:2,(B:2,C:3):5):5,D:4);`



UPGMA

L'algorithme UPGMA se base sur une matrice de distances entre les séquences. On fait l'hypothèse d'**horloge moléculaire** (arbre ultramétrique). Exemple avec la matrices de distance des séquences A, B, C et D :

	A	B	C	D
A				
B	4			
C	8	8		
D	2	4	8	

UPGMA

À chaque itération, les séquences avec la distance la plus faible sont regroupées puis une nouvelle matrice de distances est calculée avec le nouveau groupe selon la formule :

	(A,D)	B	C
(A,D)			
B	4		
C	8	8	

$$d_{ij,k} = \frac{n_i d_{ik}}{n_i + n_j} + \frac{n_j d_{jk}}{n_i + n_j}$$

Avec **i** et **j** les deux membres du groupe nouvellement formé et **k** les groupes restant. **n** est le nombre d'UTO.

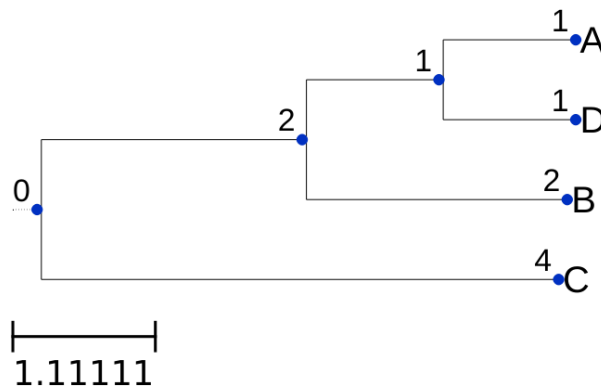
UPGMA

	A	B	C	D
A				
B	4			
C	8	8		
D	2	4	8	

	(A,D)	B	C
(A,D)			
B	4		
C	8	8	

	((A,D),B)	C
((A,D),B)		
C	8	

On obtient l'arbre $((A:1,D:1):1,B:2):2,C:4$;

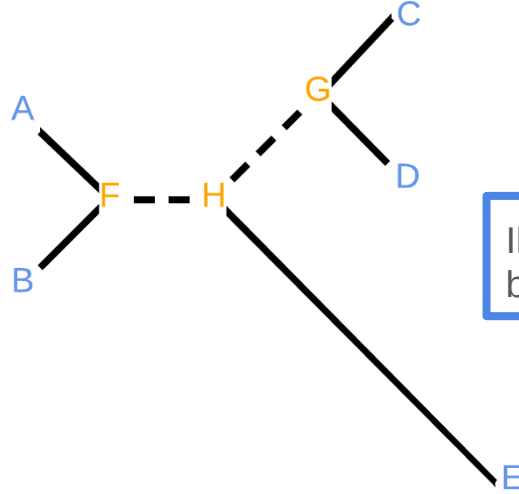
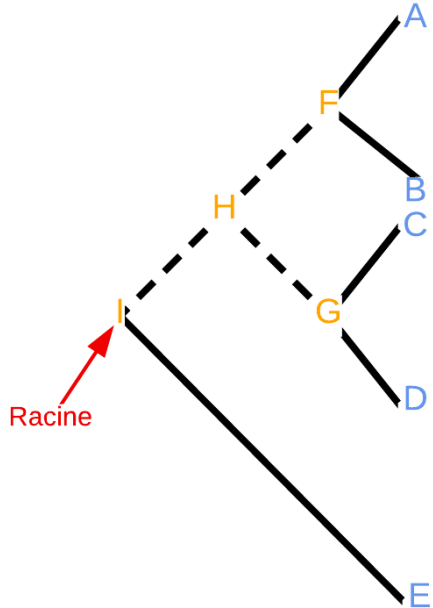


Les autres méthodes

- Parcimonie (algorithme de Fitch, TNT...)
- Distance (UPGMA, WPGMA, Neighbour Joining...)
- Maximum de vraisemblance (PhyML...)
- Inférence bayésienne

Arbres racinés et arbres non racinés

- La racine représente l'ancêtre commun le plus récent à tous les UTO
- Sans racine il n'est pas possible de déterminer les relations de parenté entre les UTO



Il y a autant de racines possibles que de branches dans un arbre non raciné

Enraciner un arbre phylogénétique

La majorité des méthodes de reconstruction phylogénétique produisent des arbres non racinés, car elles n'intègrent pas de dimension temporelle.

L'enracinement se fait donc indépendamment de la méthode choisie.

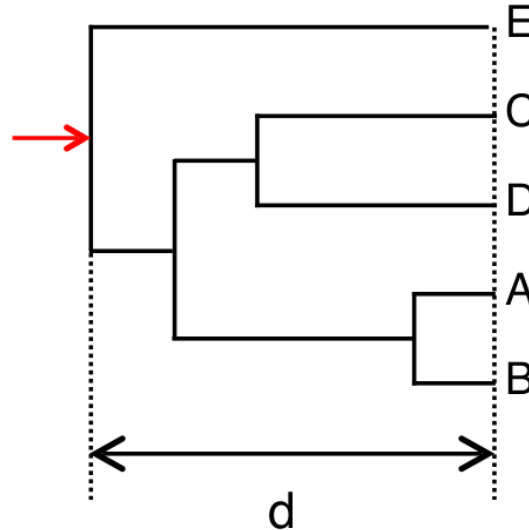
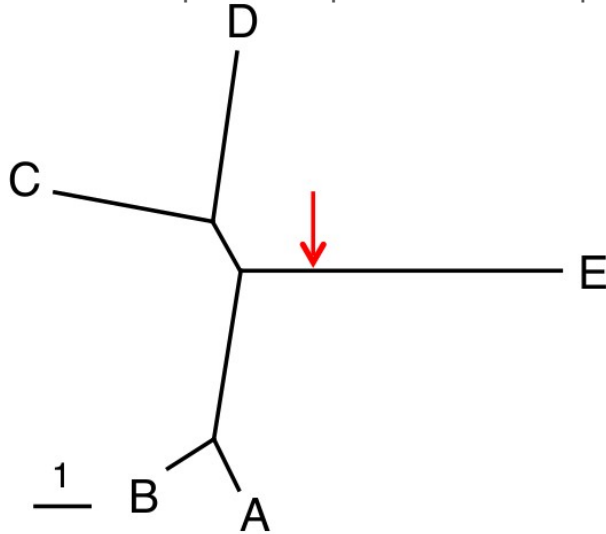
Deux approches :

- Enracinement au poids moyen
- Enracinement par un groupe extérieur

Enracinement au poids moyen

Hypothèse: Toutes les séquences évoluent à la même vitesse (i.e. l'hypothèse d'horloge moléculaire)

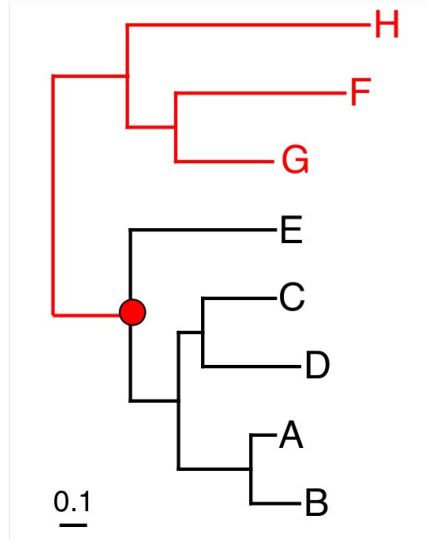
- La même quantité d'évolution s'est produite dans chaque lignée évolutive depuis leur ancêtre commun à toutes
- Les distances évolutives entre chaque feuille et la racine sont égales
- La racine est placée au point de l'arbre équidistant de toutes les feuilles



Enracinement avec un groupe extérieur

Pré-requis: inclure dans l'analyse un groupe de séquences homologues aux séquences analysées mais dont on sait a priori qu'elles sont extérieures aux séquences analysées

- La racine est définie par le nœud reliant le groupe extérieur aux séquences étudiées



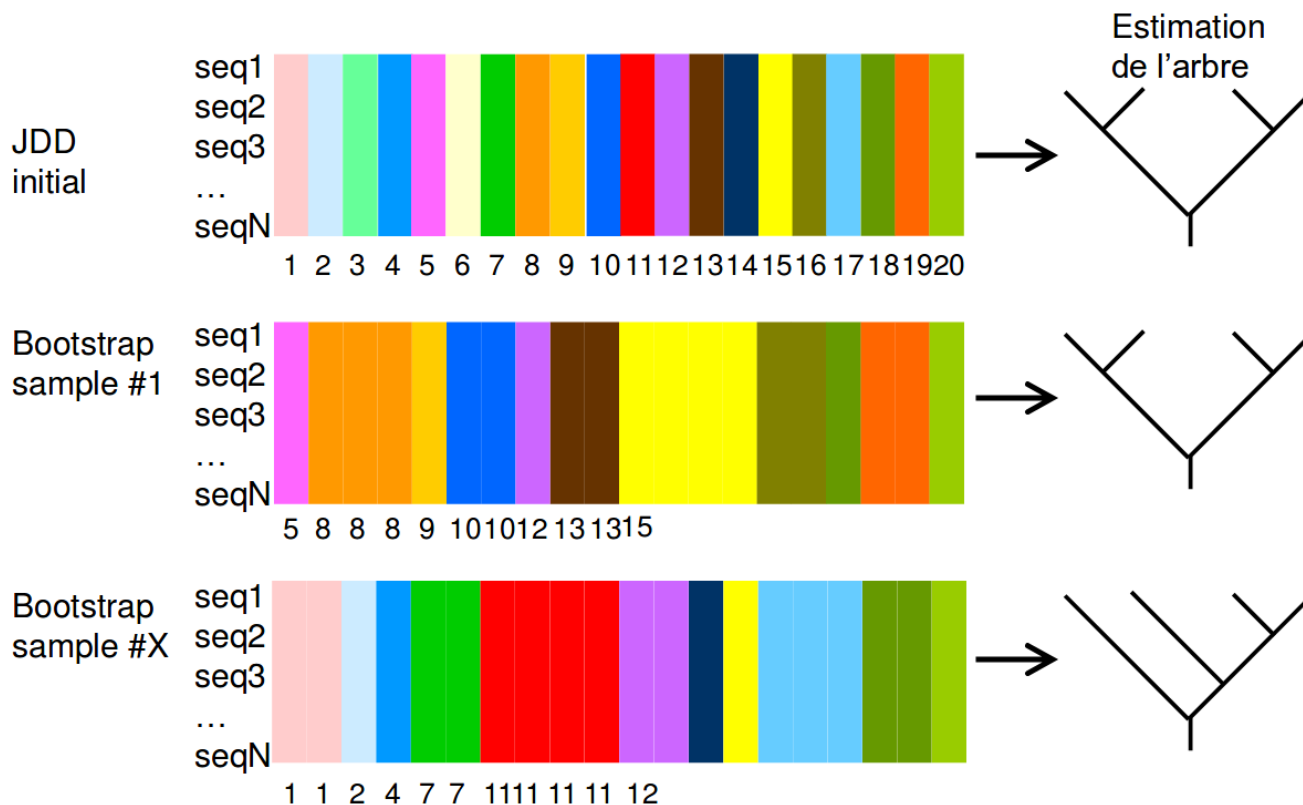
Bootstrapping

Le bootstrap permet d'évaluer la robustesse d'un arbre, c'est à dire la force avec laquelle les données (i.e l'alignement) soutiennent les regroupements observés. Un arbre robuste n'est pas très perturbé si on perturbe les données.

Déroulement :

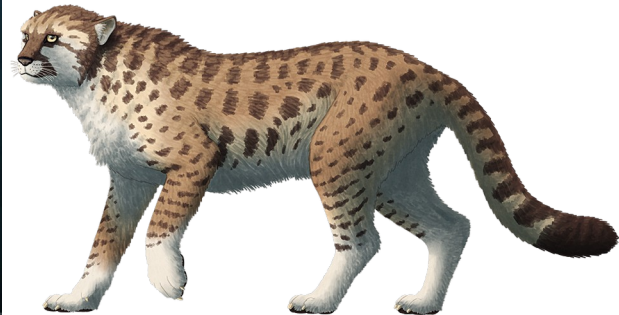
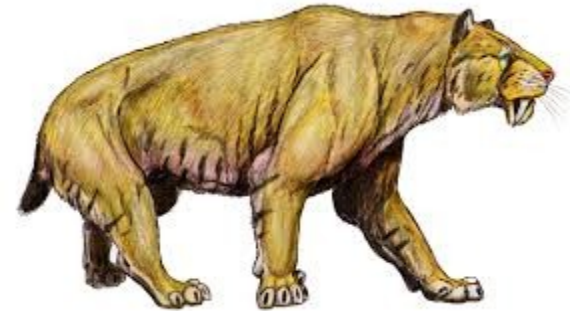
- On réalise X tirages avec remise de n sites parmi les n sites contenus dans l'alignement initial
- On calcul la phylogénie pour chacun des tirages
- La robustesse de chaque branche de l'arbre initial peut être estimée par le nombre de fois où cette même branche est retrouvée dans les répliquats de Bootstrap

Bootstrapping

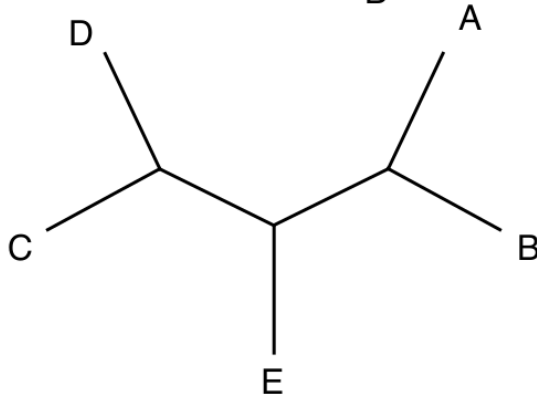
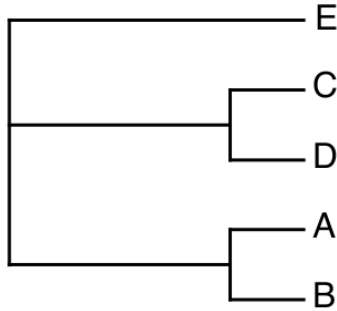


Projet

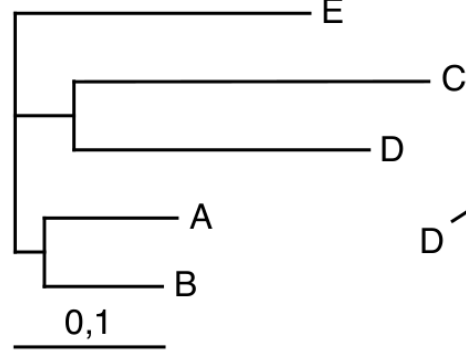
- Objectif : Reconstruire la phylogénie de trois espèces de félins disparues, le smilodon (tigre à dents de sabre), l'homoterium et le miracinonyx.
- Matériel disponible :
 - les séquences partielles de la protéines cytochrome b de ces 3 espèces
 - L'algorithme "Blast" et la base de données du NCBI
 - Python pour l'implémentation des algorithmes



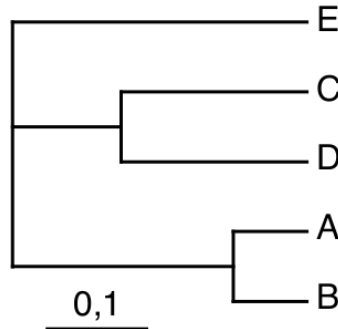
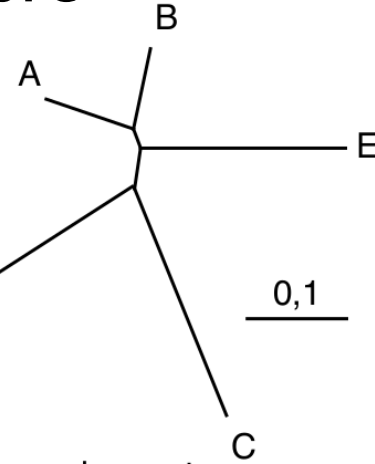
Longueurs des branches d'un arbre



Cladogrammes: la longueur des branches est arbitraire et ne reflète pas la distance évolutive séparant les séquences



Phylogrammes: la longueur des branches est proportionnelle à la distance évolutive entre les séquences (nb substitutions / site)

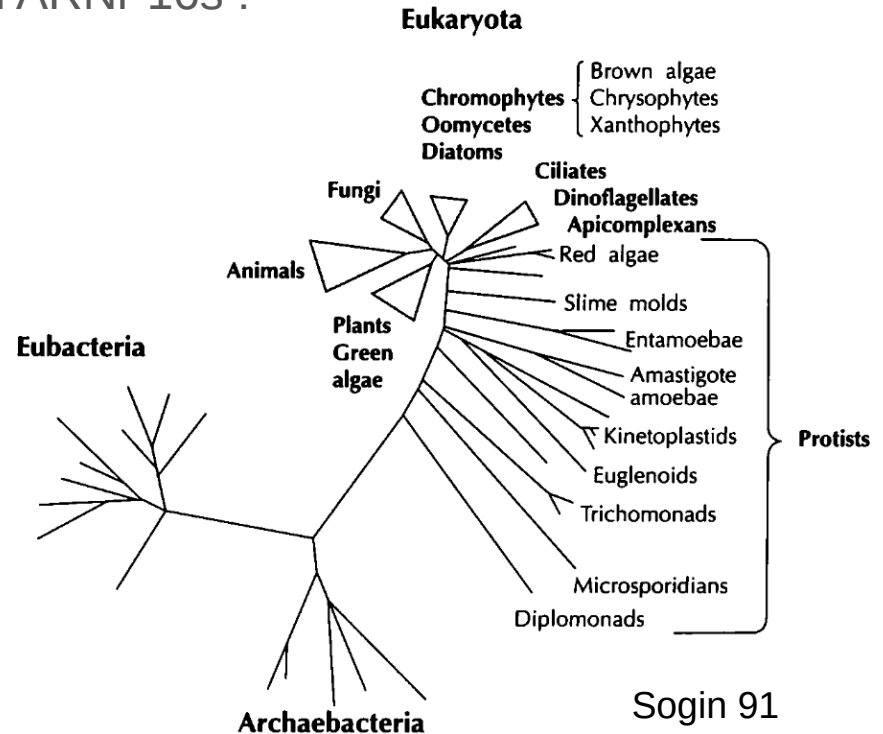


Arbres ultramétriques: la longueur des branches représente un % de divergence (phénogrammes) ou le temps (chronogrammes)

La phylogénie moléculaire

Arbre du vivant basé sur la séquence de l'ARNr 16s :

C'est un arbre **non-raciné**.



TP Phylogénie

3 séries de questions :

- [Série 1](#)
- [Série 2](#)
- [Série 3](#)

TP Python

Fichier tp_python.ipynb sur ecampus

Ouvrir jupyter avec `jupyter notebook --ip=127.0.0.1'`