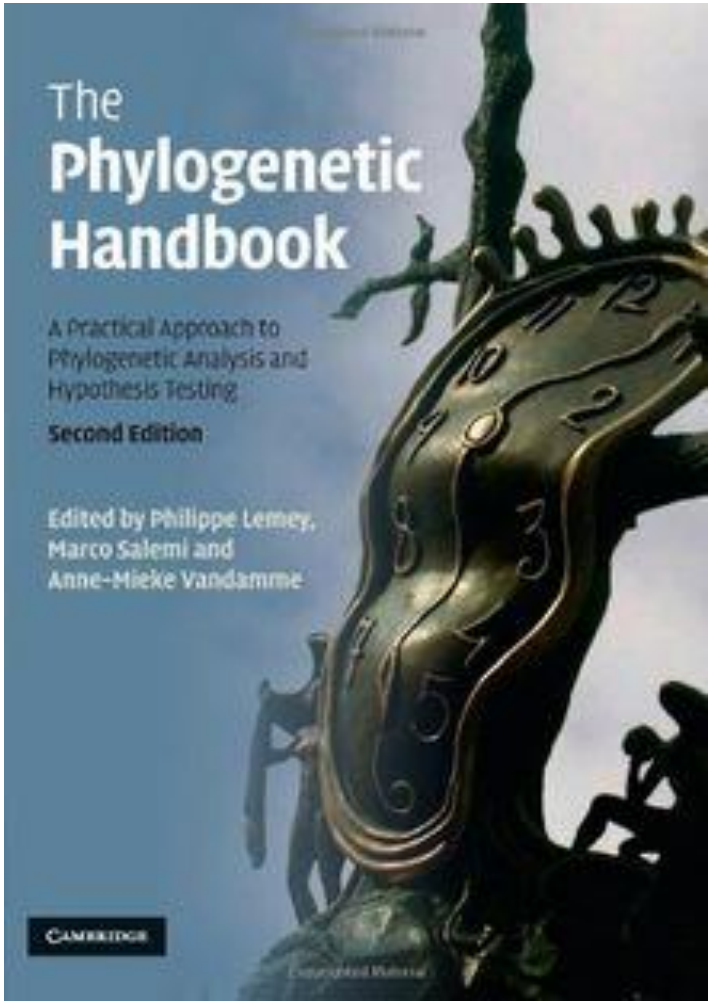
Notions élémentaires d'analyse phylogénétique moléculaire

Tamara Basta-Le Berre

tamara.basta-le-berre@u-psud.fr

Pour approfondir vos connaissances:



édition 2012

Plan du cours

1. Phylogénétique moléculaire : définition et applications

2. Le processus d'analyse phylogénétique moléculaire

→ Décider quel organisme et quelles séquences seront utilisés

→ Obtenir les séquences soit expérimentalement soit à partir d'une base de données

→ Assembler les séquences choisies dans un alignement multiple

→ Utiliser cet alignement pour générer les arbres phylogénétiques

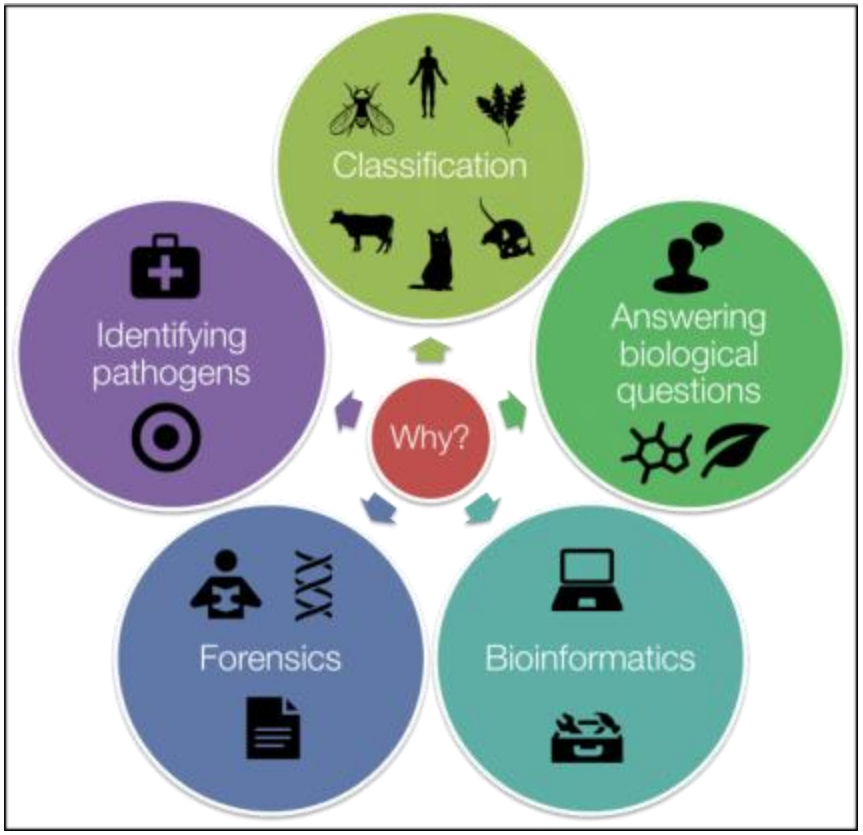
3. Algorithmes pour la phylogénie moléculaire

4. Classification cladistique

Phylogénétique moléculaire : définition et applications

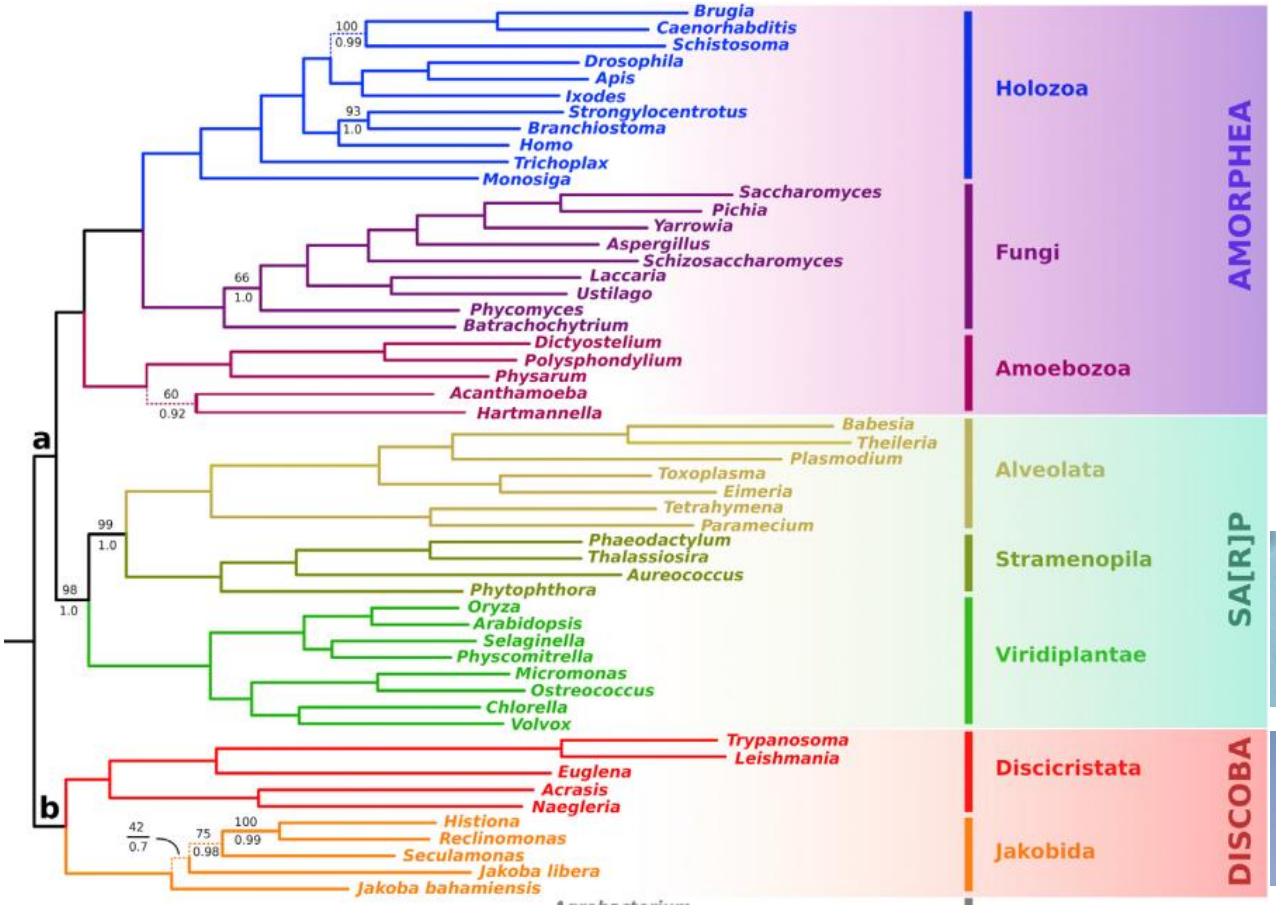
L'analyse phylogénétique moléculaire est l'activité qui consiste à reconstruire l'arbre phylogénétique d'un ensemble de séquences nucléotidiques ou protéiques d'intérêt.

Les applications sont fondamentales pour les études de biologie :



Phylogénétique moléculaire : définition et applications

* *Euglena* (protiste) a été utilisé par le passé (à tort) comme organisme model pour l'étude de la photosynthèse chez les plantes → compréhension de l'émergence des transitions majeurs en biologie (nouvel métabolisme).

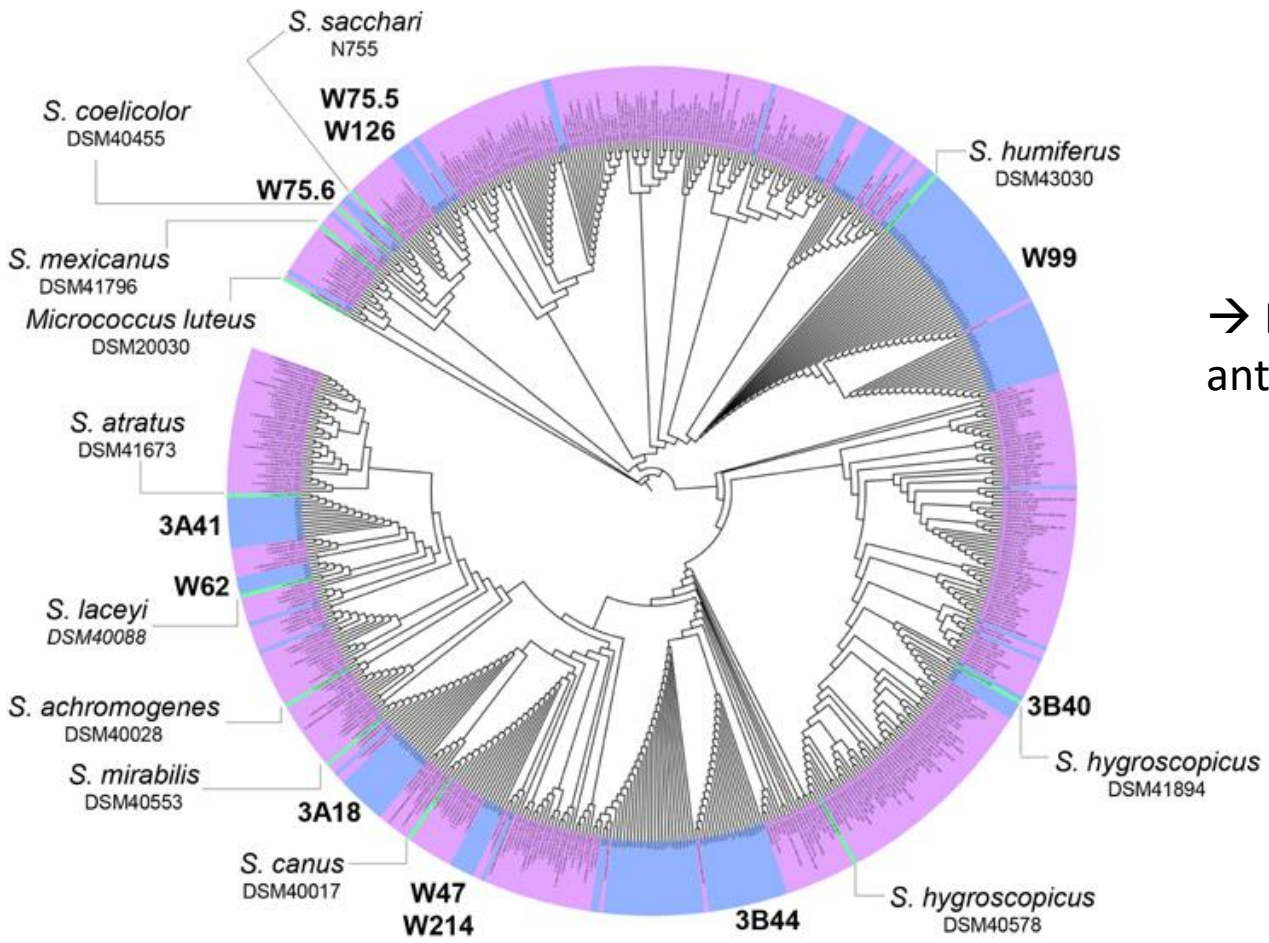


Chlamydomonas sp.(algue verte)

Euglena sp.

Phylogénétique moléculaire : déf applications

→ Isolation de 300 Actinobactéries (*Streptomyces* → production des métabolites secondaires)



frontiers in Microbiology
ORIGINAL RESEARCH
published: 09 October 2015
doi: 10.3389/fmicb.2015.01081

Diversity and functions of volatile organic compounds produced by *Streptomyces* from a disease-suppressive soil

*Viviane Cordovez[1,2], Victor J. Carrion[1], Desalegn W. Etalo[1], Roland Mumm[3,4], Hua Zhu[5], Gilles P. van Wezel[1,5] and Jos M. Raaijmakers[1,5]**

→ Recherche ciblé de nouveaux antibiotiques

Phylogénétique moléculaire : définition et applications

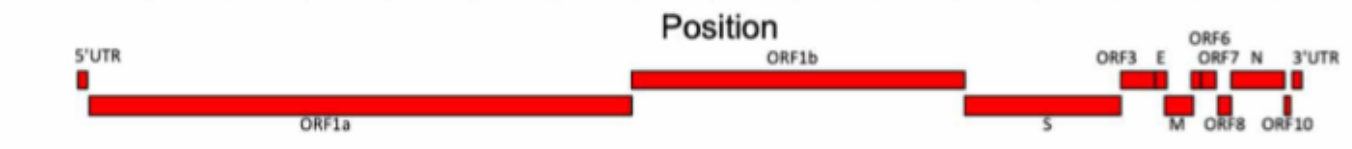
SCIENCE ADVANCES | RESEARCH ARTICLE

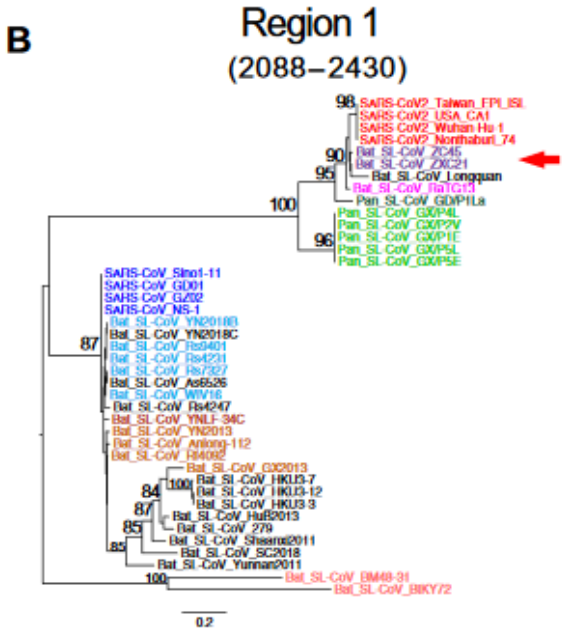
July, 2020

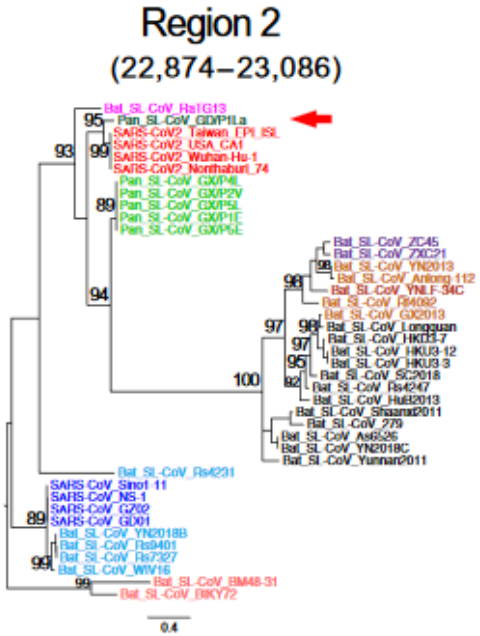
CORONAVIRUS

Emergence of SARS-CoV-2 through recombination and strong purifying selection

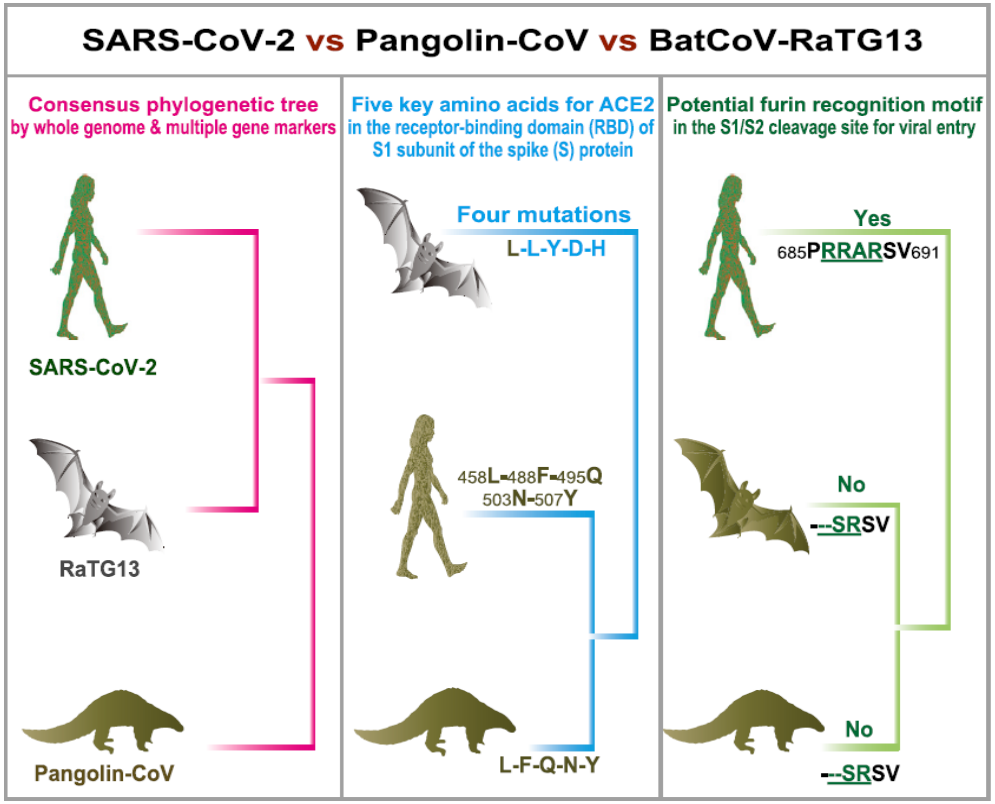
Xiaojun Li[1]*, Elena E. Giorgi[2]*, Manukumar Honnayakanahalli Marichannegowda[1], Brian Foley[2], Chuan Xiao[3], Xiang-Peng Kong[4], Yue Chen[1], S. Gnanakaran[2], Bette Korber[2,5], Feng Gao[1,6†]







Phylogénétique moléculaire : définition et applications



Understanding the origins of SARS-CoV-2 is critical for deterring future zoonosis, discovering new drugs, and developing a vaccine

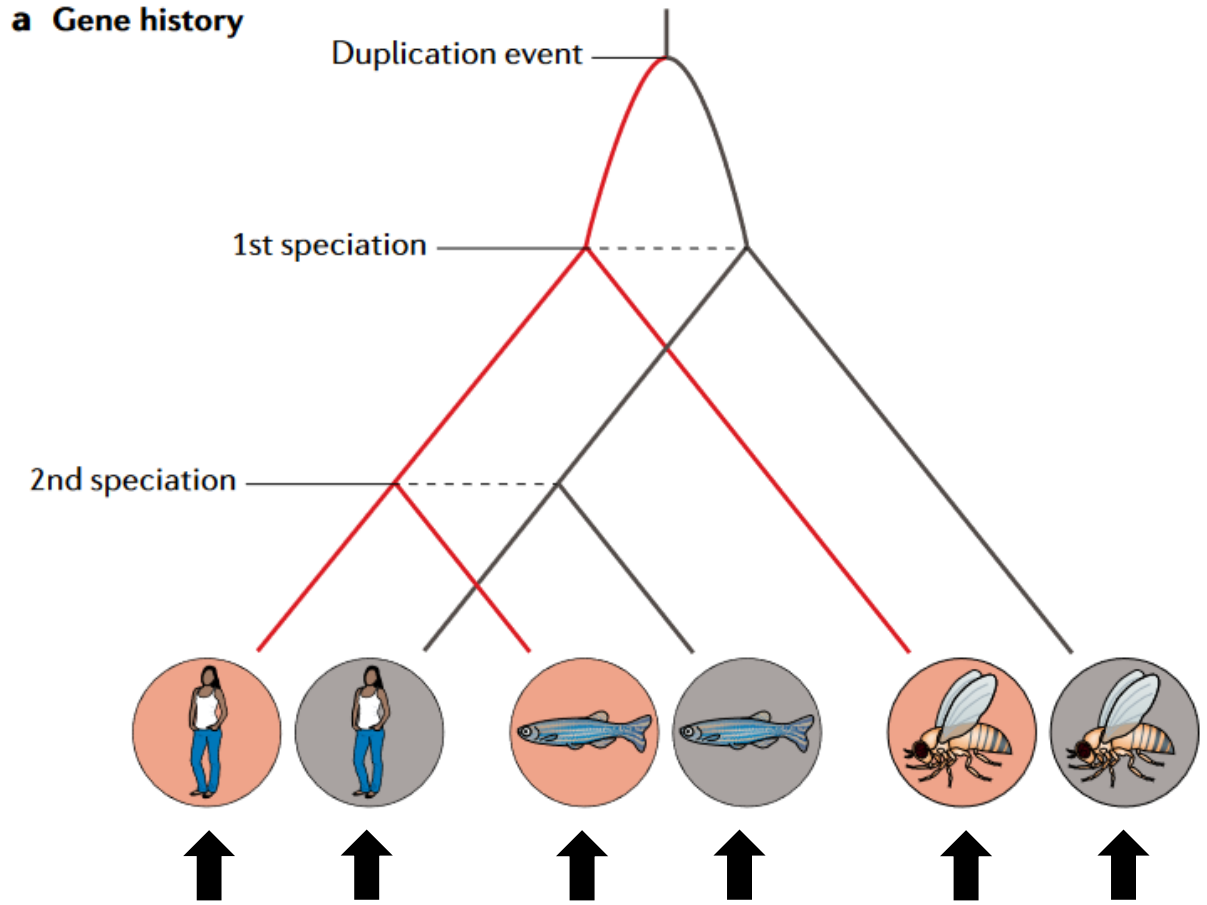
Zhang et al., 2020, Current Biology 30, 1346–1351

Le processus d'analyse phylogénétique moléculaire

1. **Décider quel organisme et quelles séquences seront utilisés**
2. Obtenir les séquences soit expérimentalement soit à partir d'une base de données
3. Assembler les séquences choisies dans un alignement multiple
4. Utiliser cet alignement pour générer les arbres phylogénétiques

Choix d'organisme et des séquences

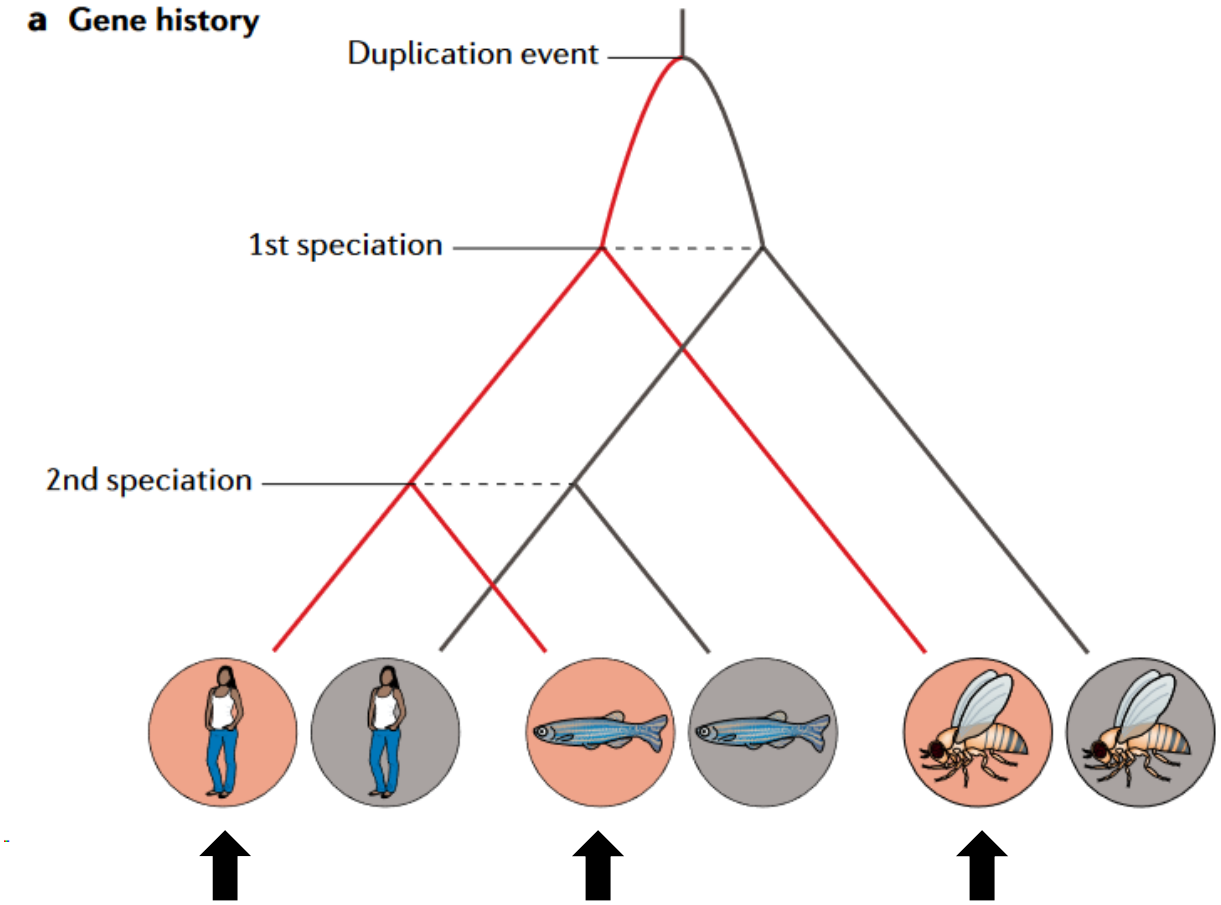
→ Gènes **homologues –** ayant un ancêtre en commun



HOMOLOGUES

Choix d'organisme et des séquences

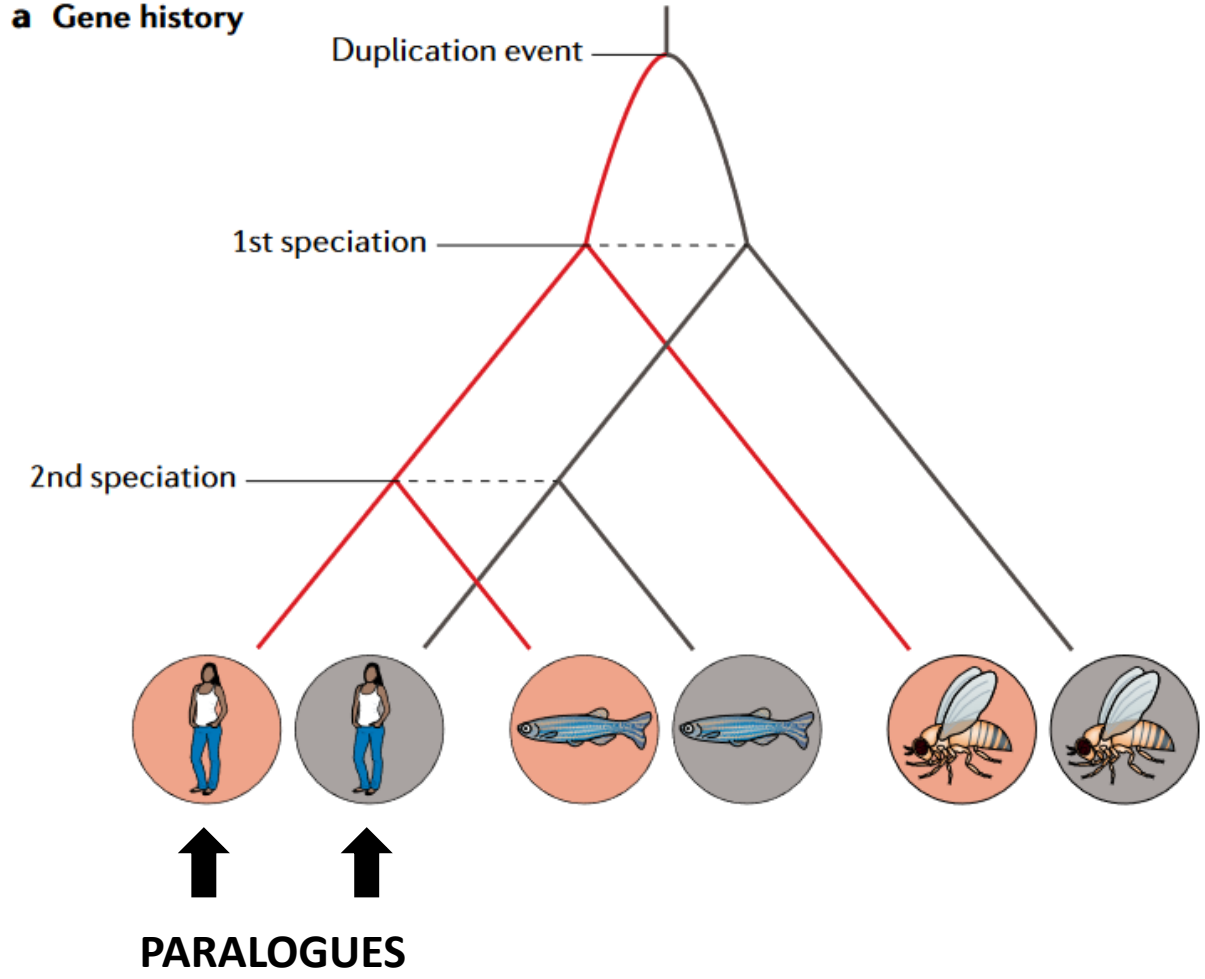
→ gènes **orhtologues –** les gènes homologues présents chez des espèces différentes



ORTHOLOGUES

Choix d'organisme et des séquences

→ Gènes **paralogues –** les gènes homologues ayant subit une duplication et que l'on retrouve au sein d'une même espèce



Choix d'organisme et des séquences

→ séquences **homologues** – ayant un ancêtre en commun

→ 2 séquences nucléotidiques non-homologues présentent 25%-50% d'identité

→ 2 séquences protéiques non-homologues présentent jusqu'au 20% d'identité

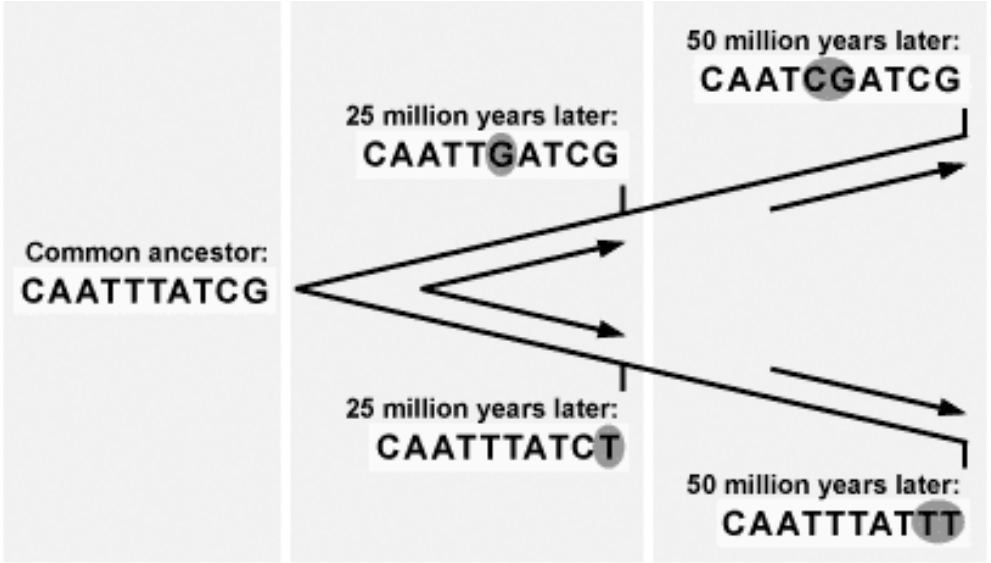
ATCGTCGTTTATGTCCCCAA	Gène 1	40 % identité de séquence → pas assez d'information pour en déduire la relation de parenté
ACCGTCGATAATGTGGGCTT	Gène 2	

ATGGTCGATAATGTCCCCAA	Gène 1	80% identité de séquence → ancêtre récent → séquences vraisemblablement homologues
ATCGTCGTTTATGTCCCCAA	Gène 2	

→ En général utiliser les séquences nucléotidiques avec > 60% identité ou les séquences protéiques avec > 25% d'identité

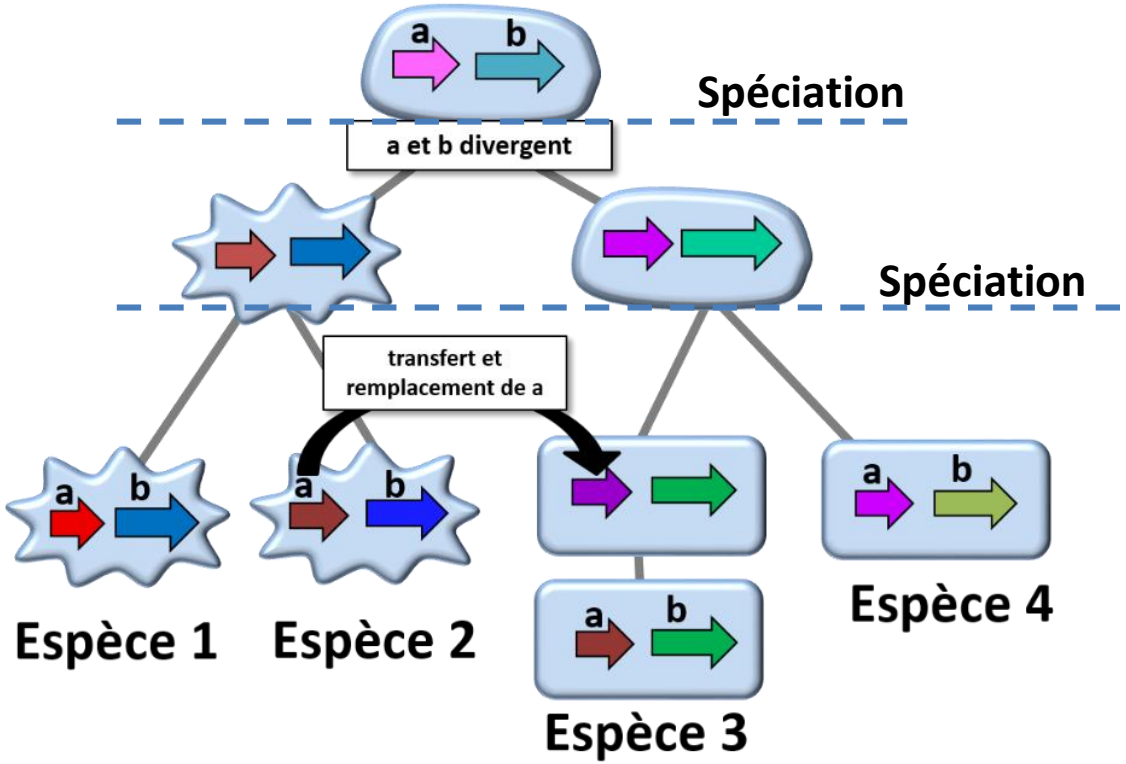
Choix d'organisme et des séquences

→ séquences qui se comportent comme une horloge moléculaire (le taux moyen de mutation sur des longues périodes est constant) → permet de mesurer quand dans le passé deux séquences ont divergé à partir d'un ancêtre commun (CM1)



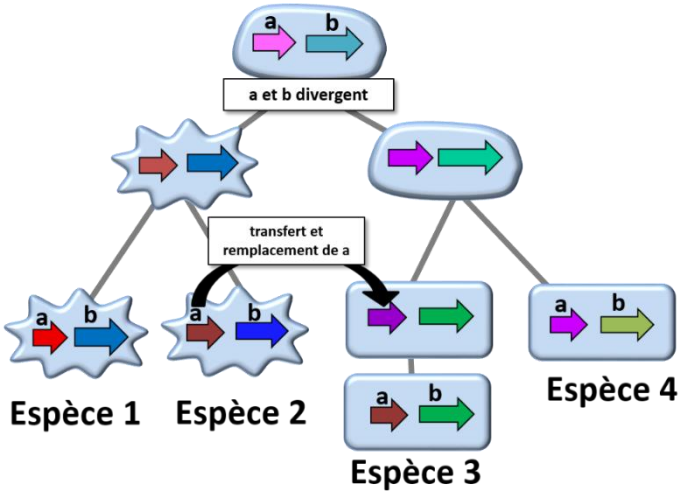
Choix d'organisme et des séquences

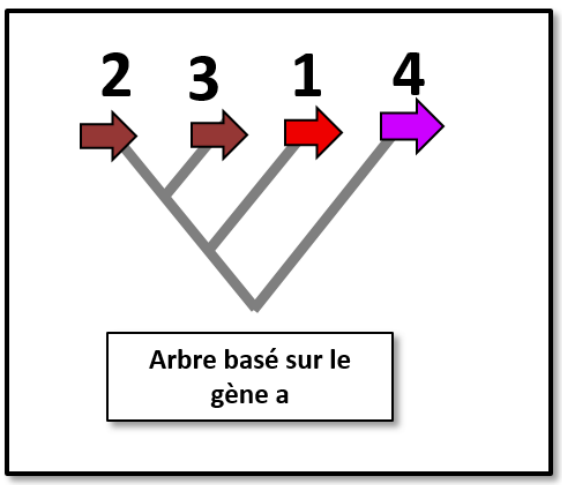
→ Pour étudier l'histoire évolutive des espèces les gènes utilisés doivent être des **orthologues** et doivent se transmettre verticalement au sein des lignées (absence des transferts latéraux)

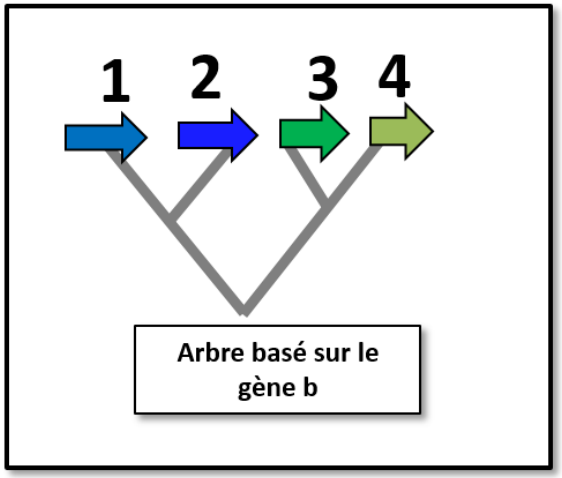


→ L'arbre obtenu à partir des séquences a = arbre obtenu à partir des séquences b?

Choix d'organisme et des séquences







Choix d'organisme et des séquences

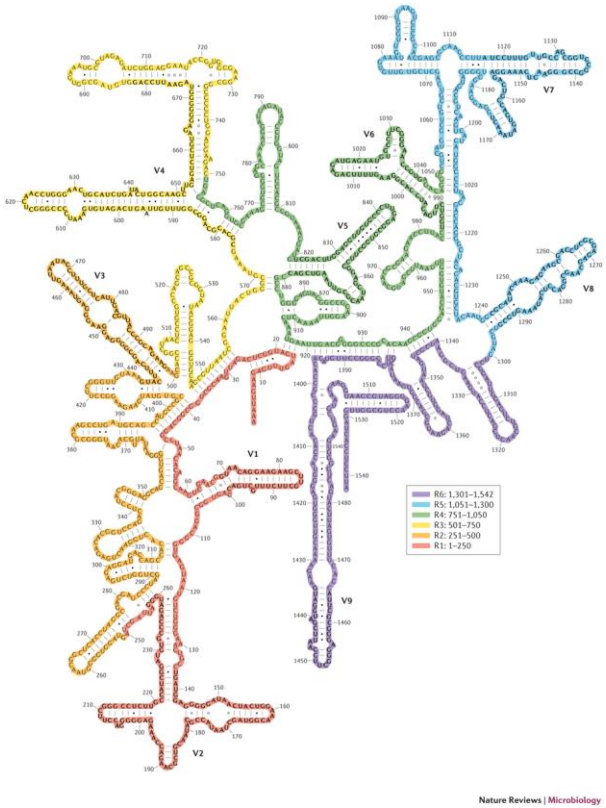
-la fonction de la séquence doit rester **conservé**

-La séquence doit comporter un **signal phylogénétique robuste** (être suffisamment longue pour comporter les informations statistiquement significatives, avoir un nombre suffisant des variations de séquence)

-l'absence du transfert horizontal des gènes – le gène doit être **hérité verticalement** de la cellule mère aux cellules filles

-si possible choisir le gène pour lequel un **grand nombre de séquences** est déjà disponible

Choix d'organisme et des séquences



« the gold standard »: le gène codant pour la 16S rRNA

-présent dans toutes les cellules

-a la même fonction dans toutes les cellules

-comporte 1500 – 2000 nucléotides – assez pour être statistiquement significatif

-comporte environ 50 hélices et 500 nucléotides qui évoluent indépendamment

-suffisamment conservé pour l'aligner facilement et précisément

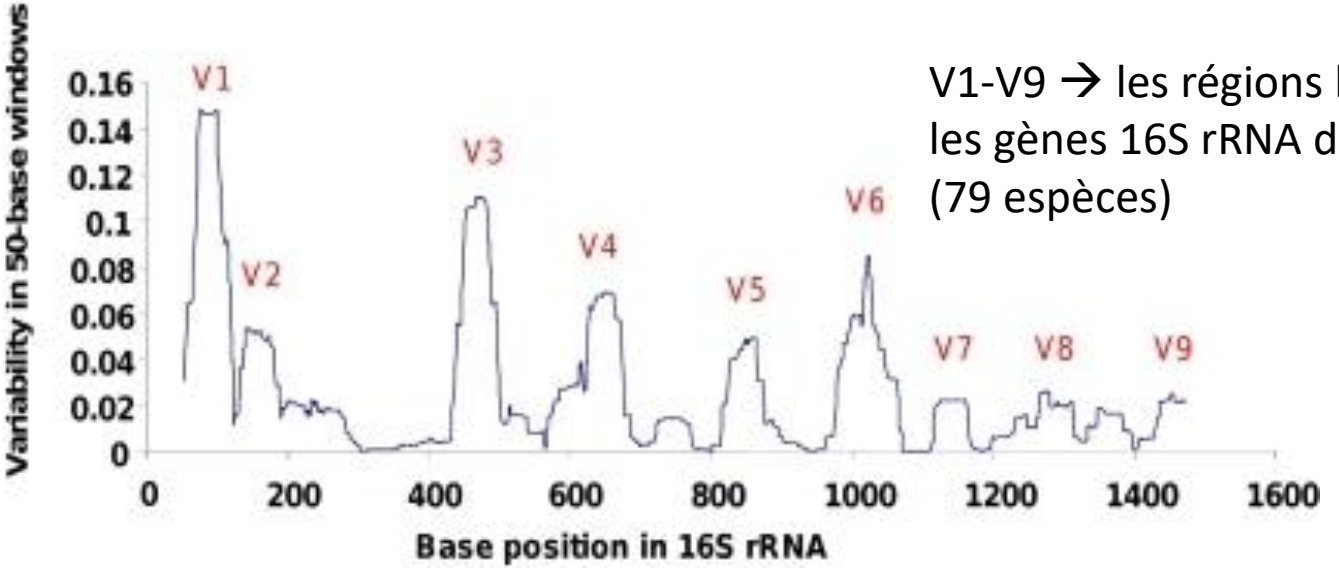
- ne se transmets pas horizontalement (très rare)

Choix d'organisme et des séquences

-comporte des régions qui évoluent *lentement* et d'autres qui évoluent *rapidement*

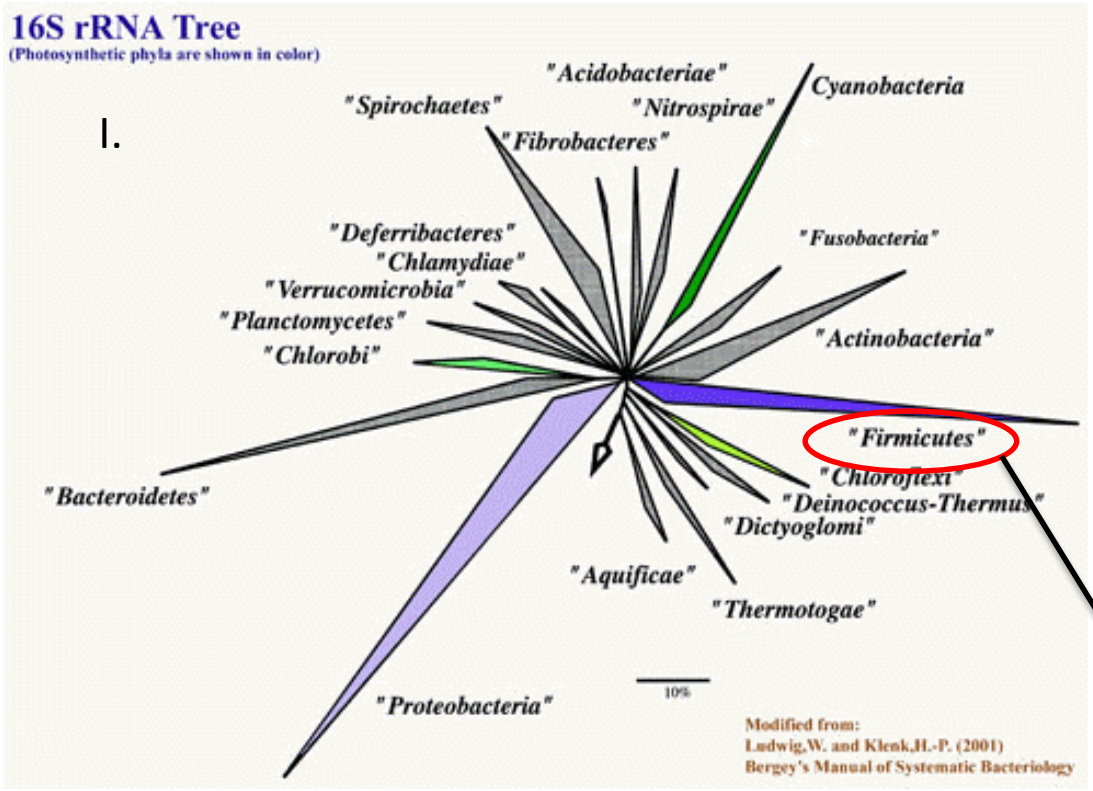
utile pour déterminer la phylogénie des espèces éloignées

utile pour déterminer la phylogénie des espèces proches



V1-V9 → les régions hypervariables parmi les gènes 16S rRNA des *Pseudomonas sp*. (79 espèces)

Choix d'organisme et des séquences



-I. au départ choisir les séquences qui couvrent l'ensemble des phylums d'un domaine

-II. puis restreindre le choix à des séquences proches du phylum d'intérêt

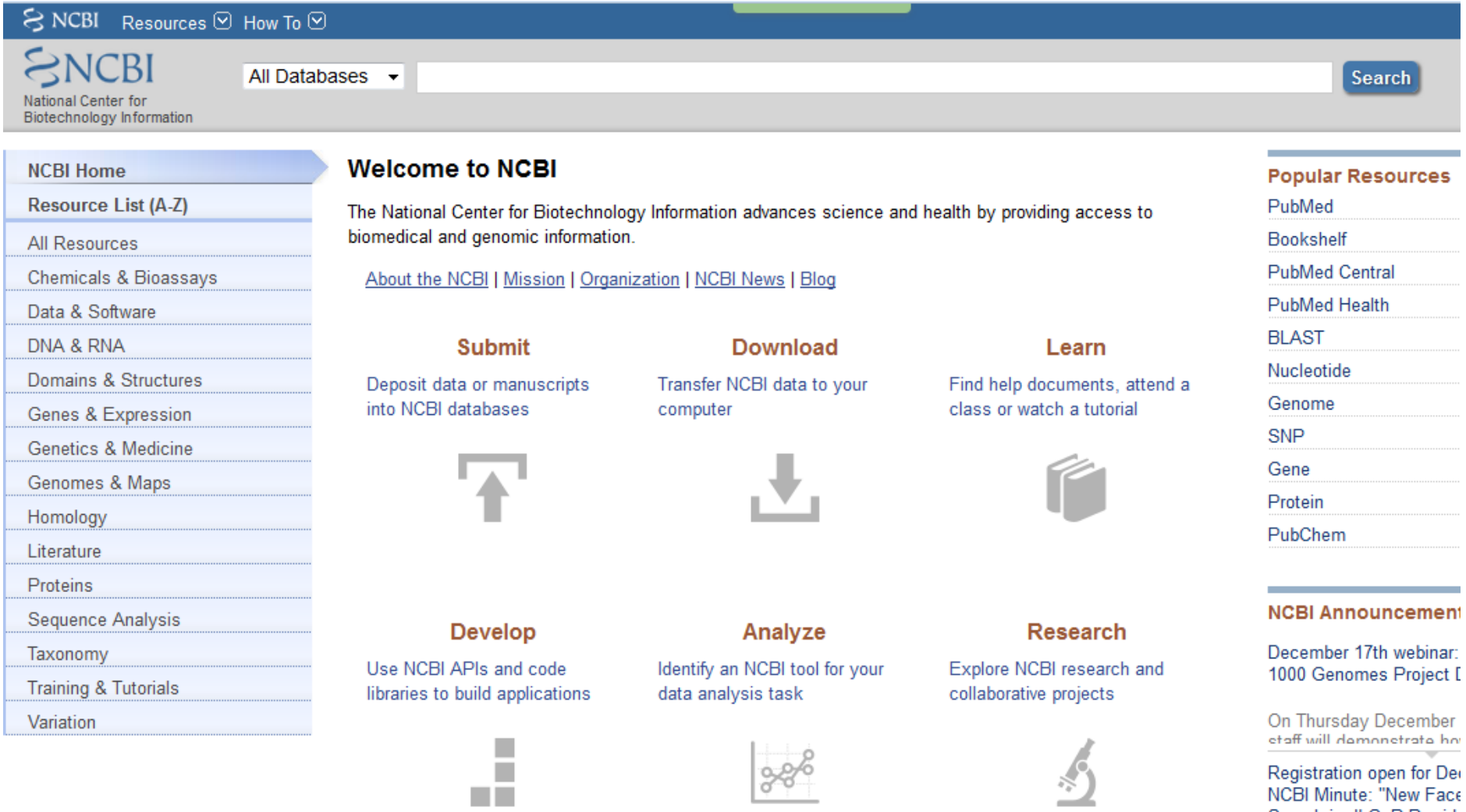
II. générer un arbre des firmicutes

Le processus d'analyse phylogénétique moléculaire

1. Décider quel organisme et quelles séquences seront utilisés
2. **Obtenir les séquences soit expérimentalement soit à partir d'une base de données**
3. Assembler les séquences choisies dans un alignement multiple
4. Utiliser cet alignement pour générer les arbres phylogénétiques

NCBI

http://www.ncbi.nlm.nih.gov/ → National Center for Biotechnology Information



TUTORIELS

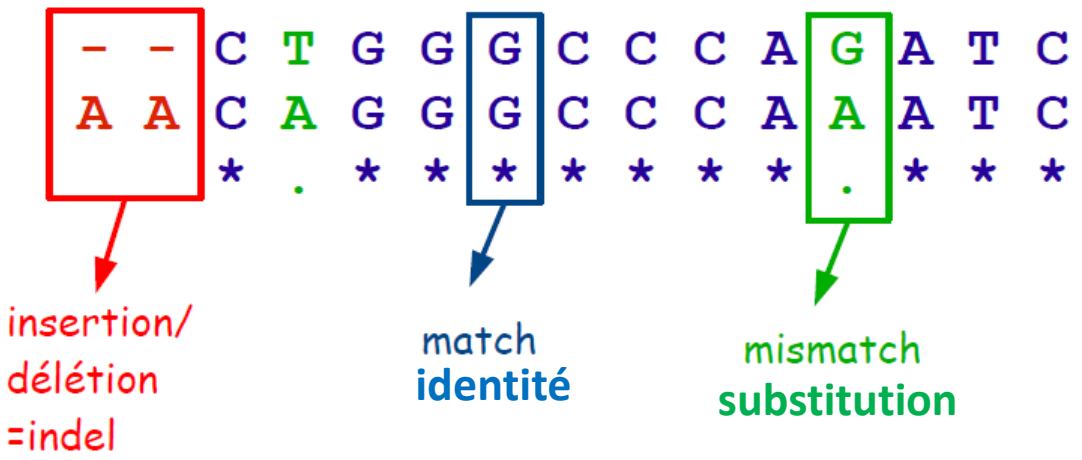
Recherche de séquences dans le NCBI
ORFinder
Recherche de séquences dans UniProt
Recherche de séquences homologues par BLAST
Alignement de séquences par Clustal Omega
Choix de sites: Gblocks
Construction d'arbres phylogénétiques avec IQ-Tree
Edition d'arbres avec iTOL
AlphaFold Protein Structure Database
ColabFold – utilisation d'AlphaFold 2
Foldseek – recherche des structures similaires dans le AFDB

Le processus d'analyse phylogénétique moléculaire

1. Décider quel organisme et quelles séquences seront utilisés
2. Obtenir les séquences soit expérimentalement soit à partir d'une base de données
3. **Assembler les séquences choisies dans un alignement multiple**
4. Utiliser cet alignement pour générer les arbres phylogénétiques

Un alignement qu'est que c'est?

→ écriture de deux (ou plus) séquences, l'une sur l'autre de façon à faire apparaitre des identités (ou similitudes)



Pour répondre à la question: "quelle est la **similarité entre 2 séquences ?**
et donc: est-ce que ces deux séquences sont **homologues ? » → MEME FONCTION**

tiré de http://biologie.univ-mrs.fr/upload/p202/Cours3_alignement.pdf

Comment quantifier la similarité de séquences?

Matrice de score

exemple: on postule qu'une identité vaut +2 et une substitution vaut -1

→ce postulat peut se résumer sous forme de matrice de scores:

	A	B	C	D	E
A	2	-1	-1	-1	-1
B	-1	2	-1	-1	-1
C	-1	-1	2	-1	-1
D	-1	-1	-1	2	-1
E	-1	-1	-1	-1	2

```
A  B  C  D  E
A  B  C  D  E
2  2  2  2  2
Score = 10
```

```
A  B  C  D  E
A  A  A  A  A
2 -1 -1 -1 -1
Score = -2
```

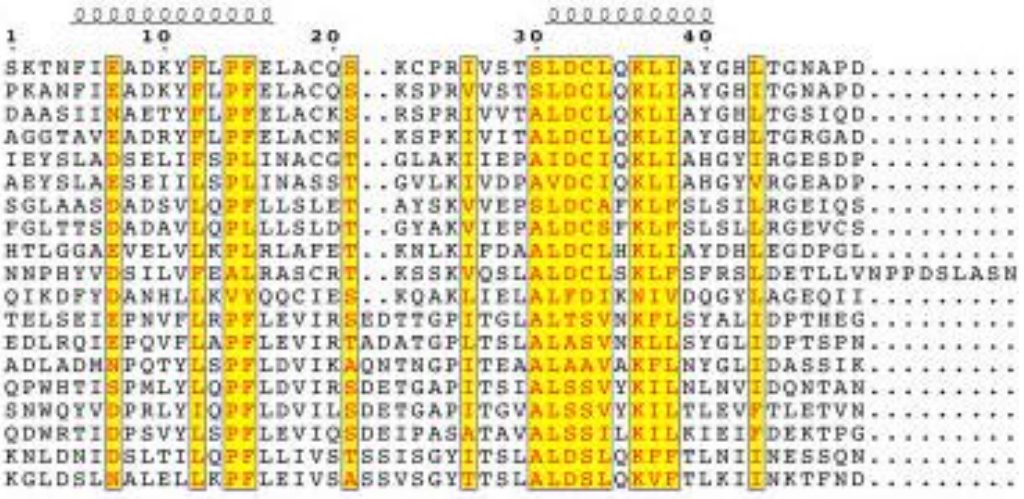
→quelles sont les critères pour déterminer les valeurs que l'on utilise pour les séquences biologiques?

Exemples de matrices de score

Série de matrices:
PAM (**P**oint **A**ccepted **M**utations)
BLOSUM (**BLO**cks of amino acid **SU**bstitution **M**atrix)

Objectif: quantifier la ressemblance entre séquences protéiques ne présentant pas (que) d'identités

Principe: alignement d'un grand nombre de séquences homologues et calcul pour chaque position de fréquences de substitution d'acides aminés



→ Les substitutions observés très fréquemment (ex. L → V) auront un score élevé

Exemples de matrices de score

BLOSUM62 : fréquences de substitution d'acides aminés dans des blocs de séquence conservés, sans insertion, présentant au maximum 62% d'identité de séquence.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Exemples de matrices de score

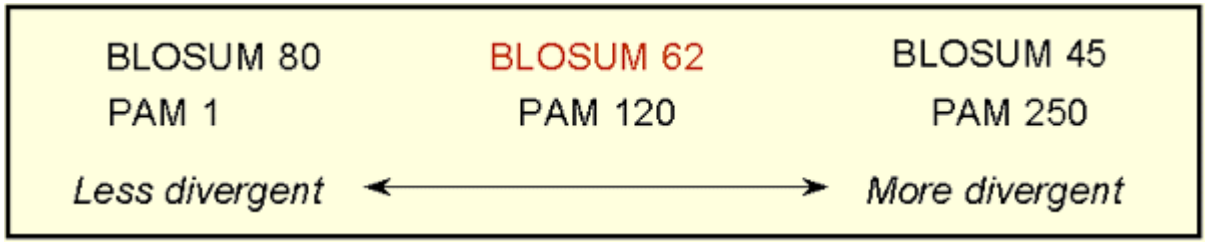
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

substitution neutre

substitution sur-représentée, probablement favorable

substitution sous-représentée, probablement défavorable

Quelle matrice choisir?



utiliser pour l'alignement des séquences très conservées - évolutivement très proches (haut % identité)

BLOSUM 62 matrice utilisée par défaut par nombreux algorithmes d'alignement (EMBOSS, BLAST)

utiliser pour l'alignement des séquences peu conservées

Comment peut on aligner deux séquences?

→alignement **LOCAL**:
i) cherche à repérer, dans deux séquences, uniquement des régions de plus grande similitude

EMBOSS – Matcher, Water, **BLAST**

Algorithme de Smith et Waterman, 1981:

i) n'importe quelle case de la matrice de comparaison peut être considérée comme point de départ

ii) tout score somme qui devient inférieur à zéro stoppe la progression du calcul

tiré de http://genet.univ-tours.fr/gen001400_fichiers/chap5/genach5ec10a.htm

Page de résultats - score

score est normalisé (bits), il permet de comparer 2 alignements, et de dire lequel est le meilleur → oui mais quel score = homologie?

```
>sp|Q92125.1|ANXA7_XENLA RecName: Full=Annexin A7; AltName: Full=Annexin-7; AltName: Full=Annexin
VII; AltName: Full=Synexin
Length=512

 GENE ID: 397854 LOC397854 | annexin VII [Xenopus laevis]
(10 or fewer PubMed links)

 Score = 40.0 bits (92),  Expect = 0.011, Method: Compositional matrix adjust.
 Identities = 45/128 (35%), Positives = 57/128 (44%), Gaps = 34/128 (26%)

Query  26   PKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHG----------------GGWGQ  67
            P PGG+  G   YPG  +PG     P GG G+G P G                  GG+G
Sbjct  67   PAPGGYPGGMPSYPG--APGFGA--PAGGQGYGAPPGAPAYGVPGYGGPGFNAPAGGYGA  122

Query  68   PHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGG  127
            P+ GG+G P  GG+G P GG  G GG     +++PS          GA   G + G + G
Sbjct  123  PNAGGFGVPPAGGYGSP-GGAPGYGG-----FSQPS------SQSYGAGGPGQMPGQMPG  170

Query  128  YMLGSAMS  135
             M G A S
Sbjct  171  QMPGQAPS  178
```

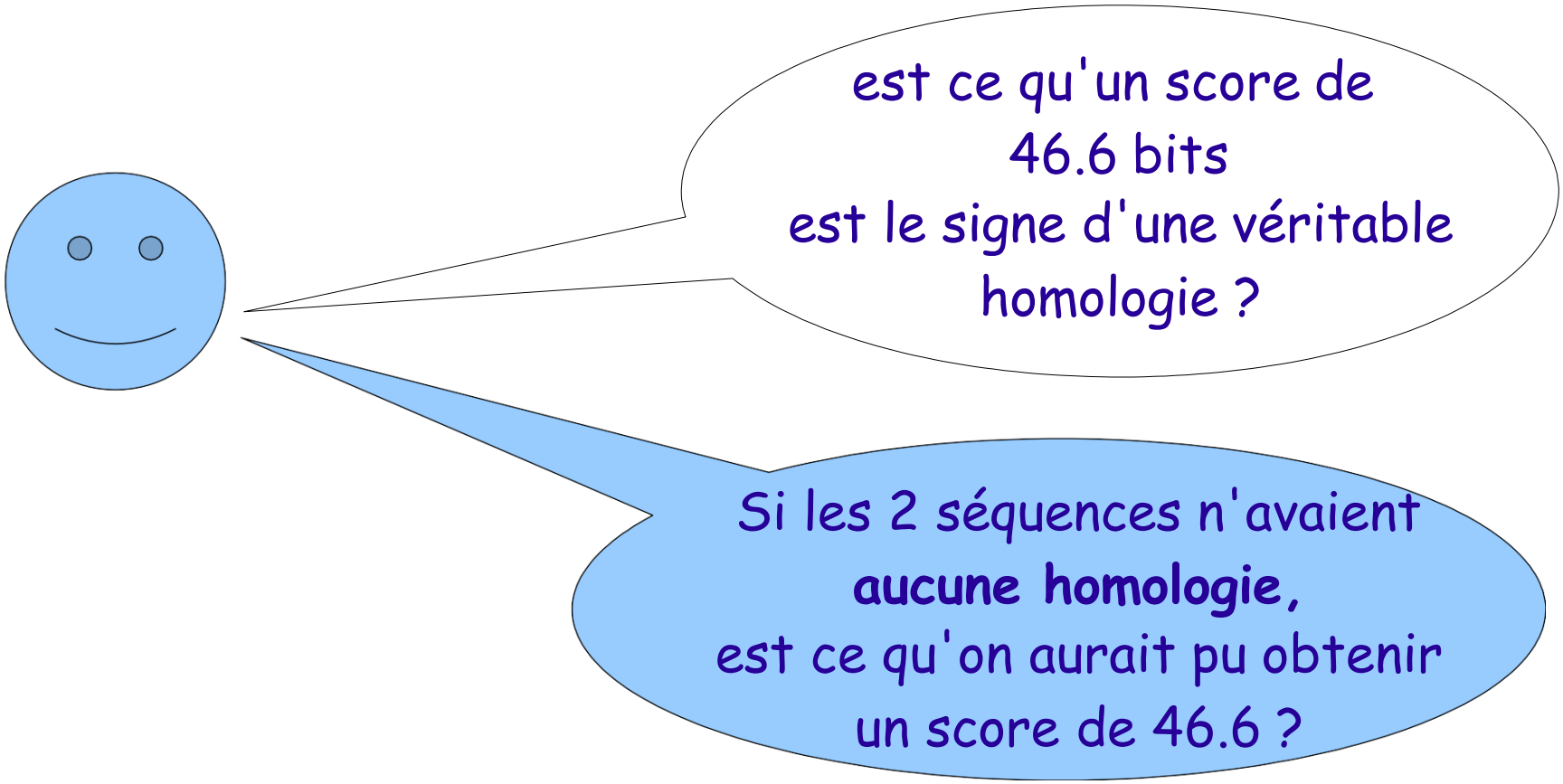
longueur: 128
identité: 35%
gap: 26%

```
>sp|Q92125.1|ANXA7_XENLA RecName: Full=Annexin A7; AltName: Full=Annexin-7; AltName: Full=Annexin
VII; AltName: Full=Synexin
Length=512

 GENE ID: 397854 LOC397854 | annexin VII [Xenopus laevis]
(10 or fewer PubMed links)

 Score = 34.7 bits (76),  Expect = 0.62, Method: Compositional matrix adjust.
 Identities = 34/86 (39%), Positives = 41/86 (47%), Gaps = 23/86 (26%)

Query  26   PKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHG----------------GGWGQ  67
            P PGG+  G   YPG  +PG     P GG G+G P G                  GG+G
Sbjct  67   PAPGGYPGGMPSYPG--APGFGA--PAGGQGYGAPPGAPAYGVPGYGGPGFNAPAGGYGA  122

Query  68   PHGGGWGQPHGGGWGQPHGGGWGQGG  93
            P+ GG+G P  GG+G P GG  G GG
Sbjct  123  PNAGGFGVPPAGGYGSP-GGAPGYGG  147
```

longueur: 86
identité: 39%
gap: 26%

tiré de http://biologie.univ-mrs.fr/upload/p202/Cours4_Blast.pdf

E-value - définition



tiré de http://biologie.univ-mrs.fr/upload/p202/Cours4_Blast.pdf

E-value - définition

E-value de X

=

on s'attendrait à trouver X alignements de score équivalent purement par chance contre une banque de données de taille équivalente

s = 46 > Evalue = 4e-4 : je m'attends à trouver en moyenne 0.0004 alignements de score 46 purement par hasard (si je blaste 2500 séquences aléatoires, j'en obtiendrai ~ 1)

s = 267 > Evalue = 1e-70: il faut que je blaste 1e70 séquences aléatoires avant de tomber au hasard sur un alignement de cette qualité ...

tiré de http://biologie.univ-mrs.fr/upload/p202/Cours4_Blast.pdf

E-value - définition

faux-positifs: on a un alignement, mais les séquences ne sont pas homologues

tiré de http://biologie.univ-mrs.fr/upload/p202/Cours4_Blast.pdf

Alignement multiple de séquences

→ En général, le biologiste dispose de multiples séquences d'intérêt

→ Besoin d'un alignement multiple pour identifier les résidus conservés (importants pour la fonction) et inférer un arbre phylogénétique

→ On recherche un alignement global, cependant inexploitable en pratique: pour 5 séquences de 500 AA mémoire requise est de 10 Po (10^{15} octets)

→ Alignements multiples utilisent toujours des heuristiques (la solution n'est pas toujours optimale et nécessite d'optimiser « à la main »)

Alignement multiple de séquences

→ Choix de programme d'alignement (https://www.ebi.ac.uk/jdispatcher/msa):

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch MView

T-Coffee

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.
Try it out at WebPRANK.

Alignement multiple de séquences

→ Choix de programme d'alignement en fonction du nombre de séquences leur degré de conservation et la qualité de l'alignement:

Tableau 5.6 – Quelques programmes d'alignement multiple.

	Rapidité	Séquences proches	Séquences éloignées	Qualité
Multalin	++	+++	+	++
Clustal W	+	++	++	+++
Muscle	+++	+++	+	+++
MAFFT	++	++	+	+++
T-Coffee	+	+	+++	+++
DIALIGN	+	+	+++	+

Le processus d'analyse phylogénétique moléculaire

1. Décider quel organisme et quelles séquences seront utilisés

2. Obtenir les séquences soit expérimentalement soit à partir d'une base de données

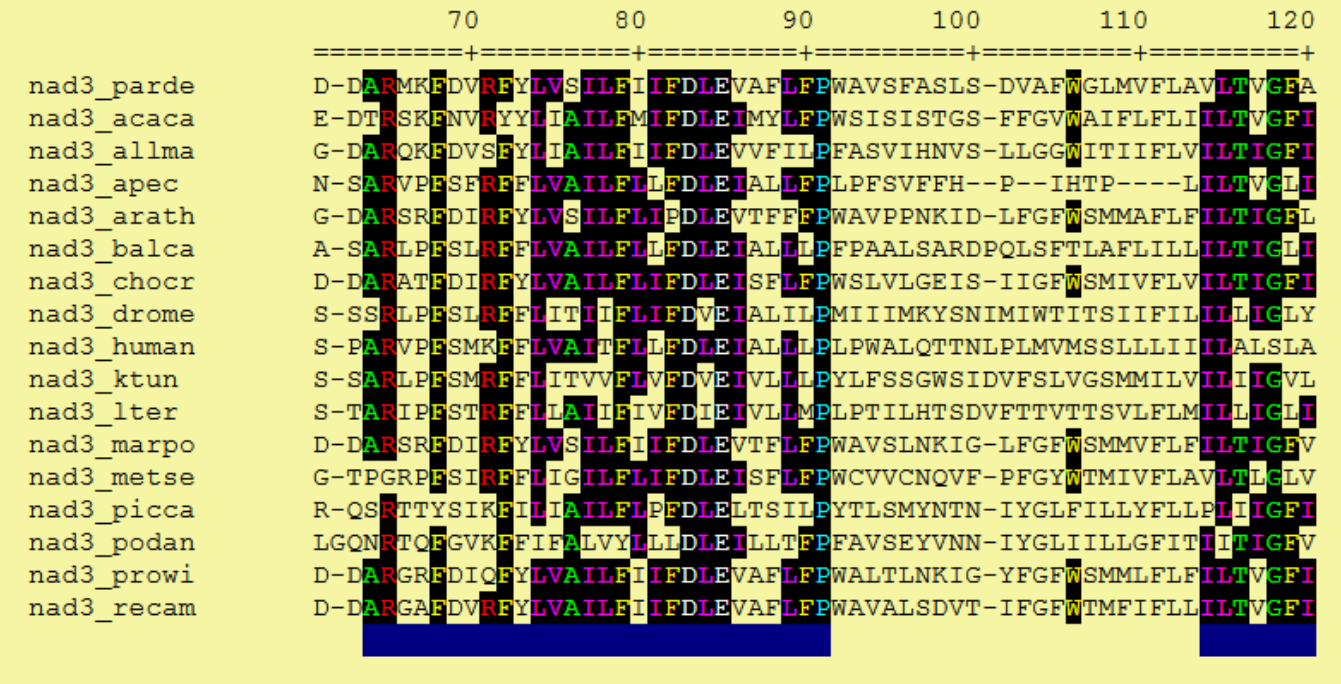
3. Assembler les séquences choisies dans un alignement multiple
   →**extraire les positions pertinentes (GBlocks)**

4. Utiliser cet alignement pour générer les arbres phylogénétiques

Choix des sites dans un alignement

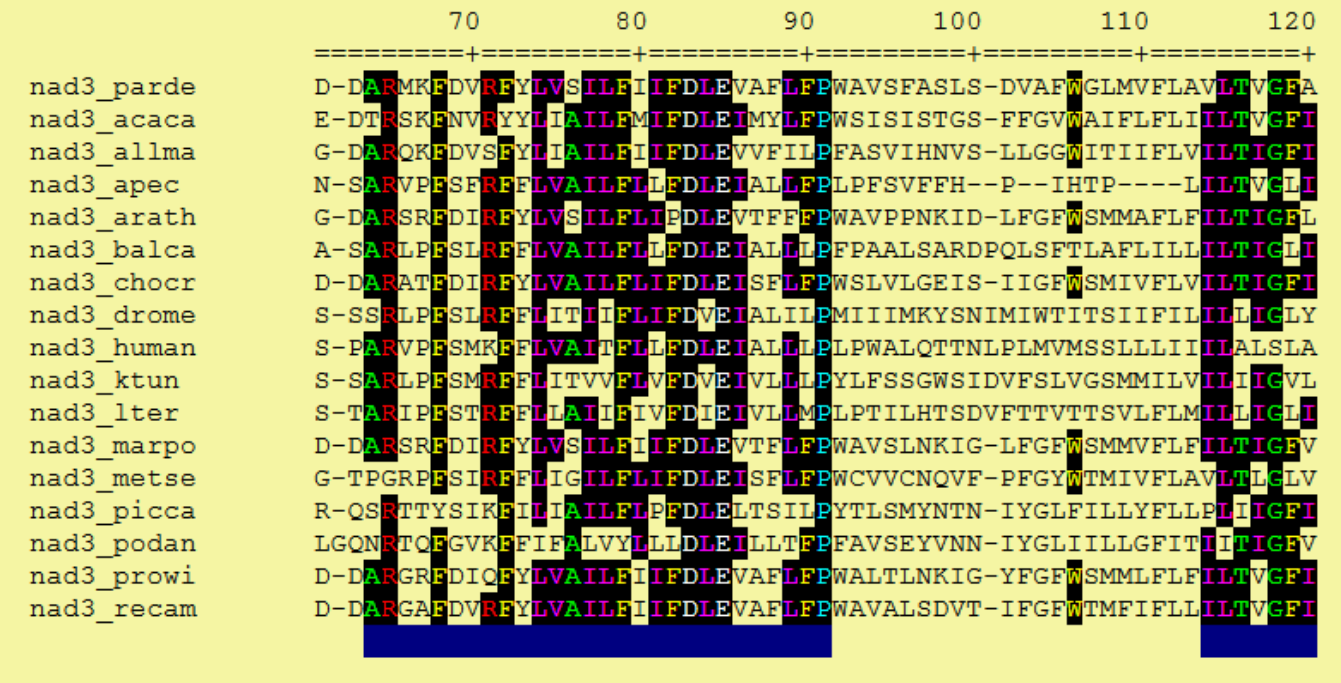
→ Pour construire un arbre phylogénétique il est nécessaire d'utiliser les sites composés de résidus homologues

→ Les alignements comportent toujours des régions ou l'alignement (détermination de l'homologie) est incertain → souhaitable de les éliminer



Choix des sites dans un alignement

→ Gblocks: classe les sites en conservé, non-conservé et très conservés, en fonction de la variabilité pour chaque position



→ Elimine les sites non-conservés et les blocs trop courts

Le processus d'analyse phylogénétique moléculaire

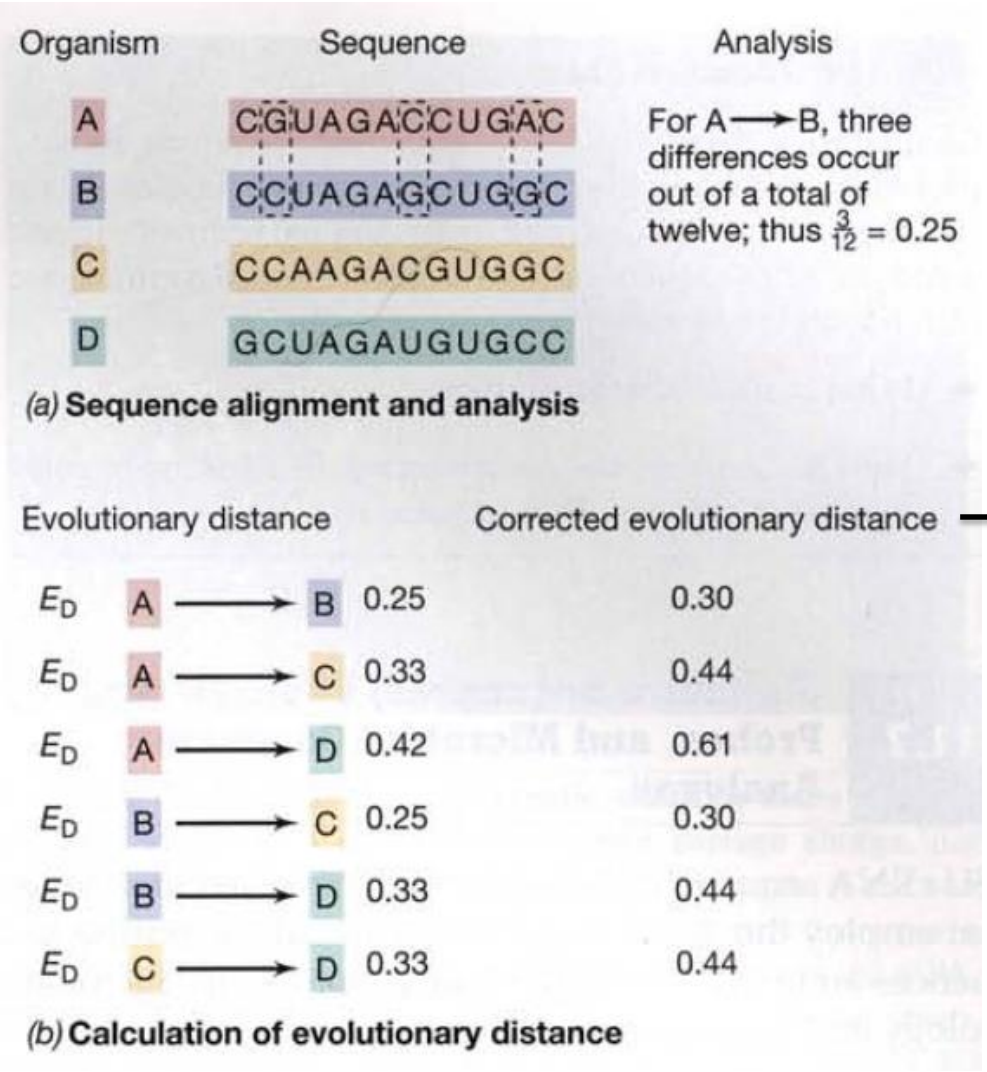
1. Décider quel organisme et quelles séquences seront utilisés

2. Obtenir les séquences soit expérimentalement soit à partir d'une base de données

3. Assembler les séquences choisies dans un alignement multiple →extraire les positions pertinentes (GBlocks)

4. **Utiliser cet alignement pour générer les arbres phylogénétiques**

Calcul de la distance évolutive



Organism	Sequence	Analysis
A	CGUAGACCUGAC	For A → B, three differences occur out of a total of twelve; thus $\frac{3}{12} = 0.25$
B	CCUAGAGCUGGC	
C	CCAAGACGUGGC	
D	GCUAGAUGUGCC	

(a) **Sequence alignment and analysis**

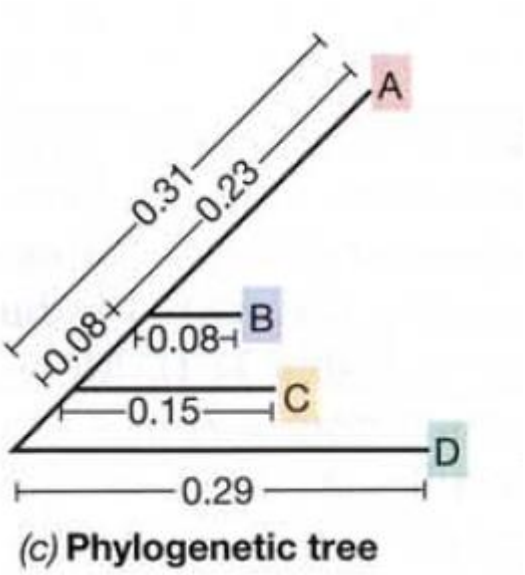
Evolutionary distance			Corrected evolutionary distance
E_D	A → B	0.25	0.30
E_D	A → C	0.33	0.44
E_D	A → D	0.42	0.61
E_D	B → C	0.25	0.30
E_D	B → D	0.33	0.44
E_D	C → D	0.33	0.44

(b) **Calculation of evolutionary distance**

→ correction selon un modèle évolutif: plus qu'un changement à pu se produire à une position donnée

Calcul de la distance évolutive

Construction de l'arbre phylogénétique: traitement informatique des données de *distances* donnant le meilleur ajustement (cf. Algorithmes pour la phylogénie moléculaire)



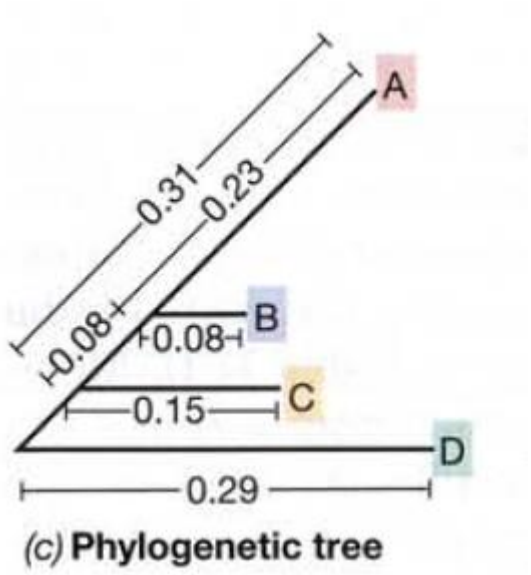
(c) **Phylogenetic tree**

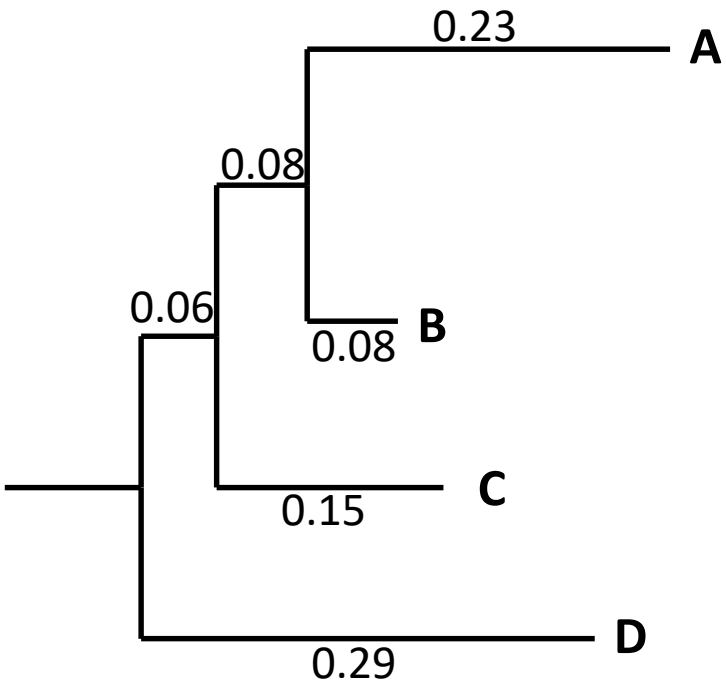
Evolutionary distance			Corrected evolutionary distance
E_D	A → B	0.25	0.30
E_D	A → C	0.33	0.44
E_D	A → D	0.42	0.61
E_D	B → C	0.25	0.30
E_D	B → D	0.33	0.44
E_D	C → D	0.33	0.44

(b) **Calculation of evolutionary distance**

Calcul de la distance évolutive

Construction de l'arbre phylogénétique: traitement informatique des données de distances donnant le meilleur ajustement

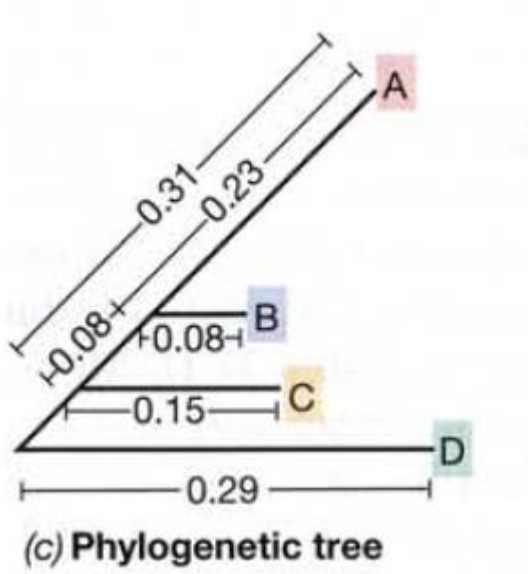


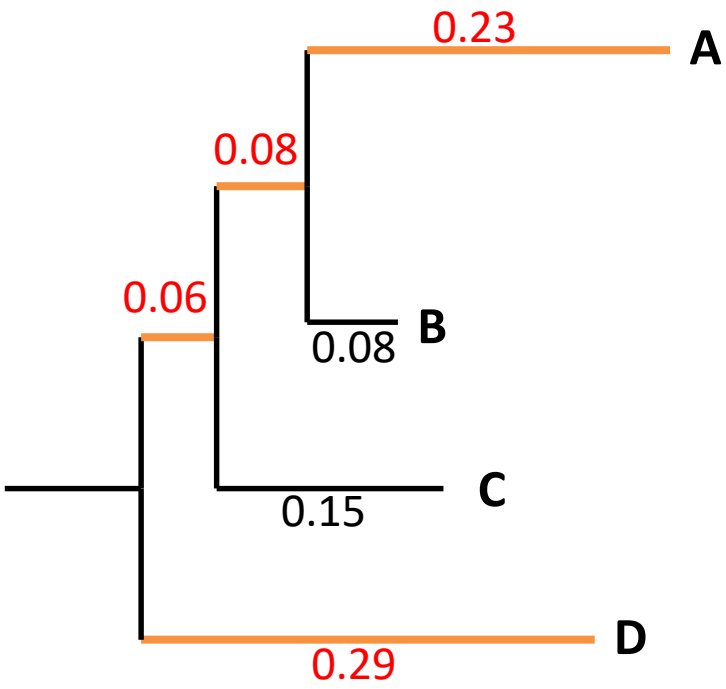


Quelle est la distance évolutive entre les espèces A et D?

Calcul de la distance évolutive

Construction de l'arbre phylogénétique: traitement informatique des données de distances donnant le meilleur ajustement





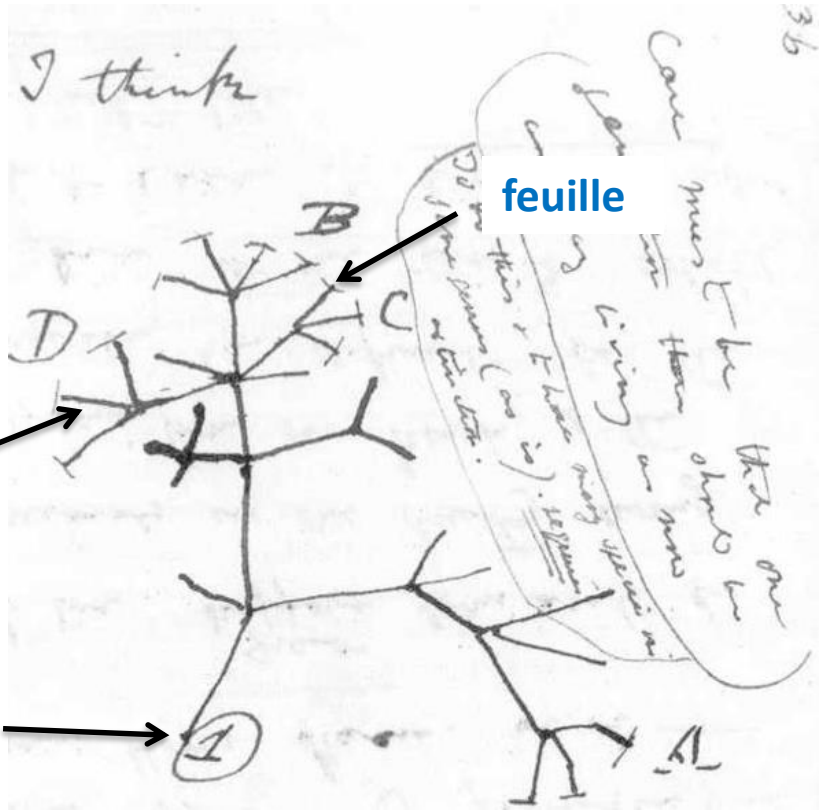
Quelle est la distance évolutive entre les séquences A et D?
(0.23 + 0.08 + 0.06) + 0.29 = 0.66

Comment lire un arbre phylogénétique

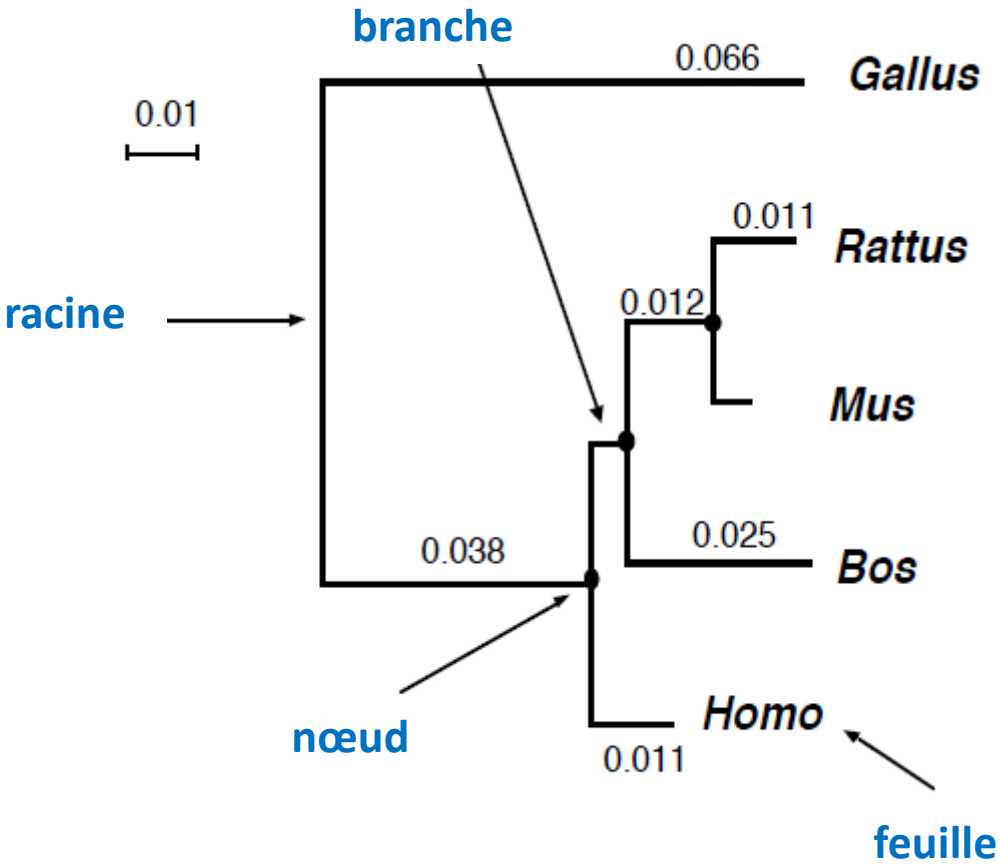
« Fortunately, one can interpret trees and use them for organizing knowledge of biodiversity without knowing the details of phylogenetic inference.
The reverse is, however, not true. **One cannot really understand phylogenetics if one is not clear what an evolutionary tree is.**"

Baum et al. , 2005 Science

-l'idée centrale de Darwin: malgré leur diversité les organismes modernes sont tous les descendants d'un ancêtre commun universel



Comment lire un arbre phylogénétique



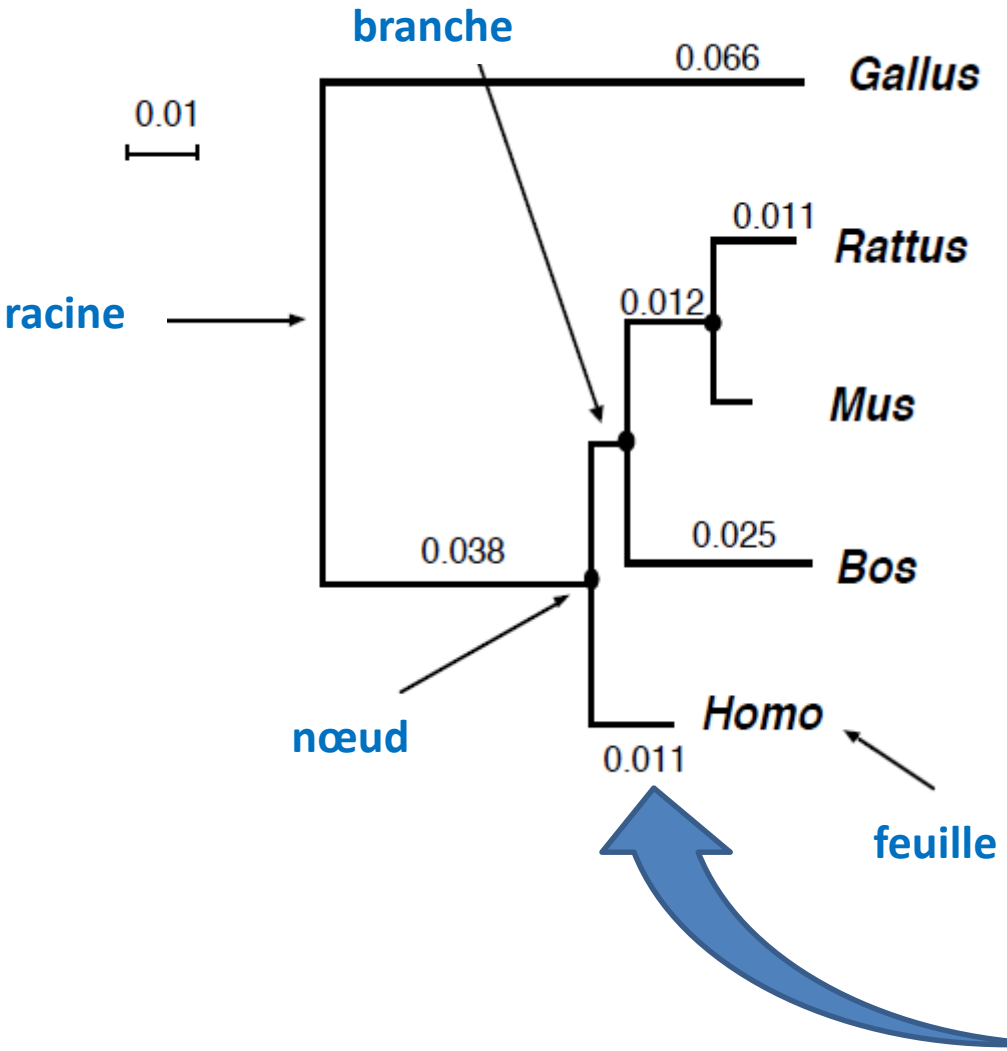
feuille: les extrémités des branches, ce sont les organismes modernes ou leurs macromolécules (protéines, ADN, ARN)

nœud: le dernier ancêtre commun de tous les organismes reliés par ce nœud

branche: relie les nœuds, représente le chemin évolutif parcouru entre l'ancêtre (nœud) et la feuille; la longueur de branches correspond au temps (si *échelle* = temps) ou la quantité de changement évolutif (si *échelle* = distance évolutive)

racine: c'est la « base » de l'arbre – l'ancêtre commun de tous les organismes/molécules dans l'arbre

Comment lire un arbre phylogénétique



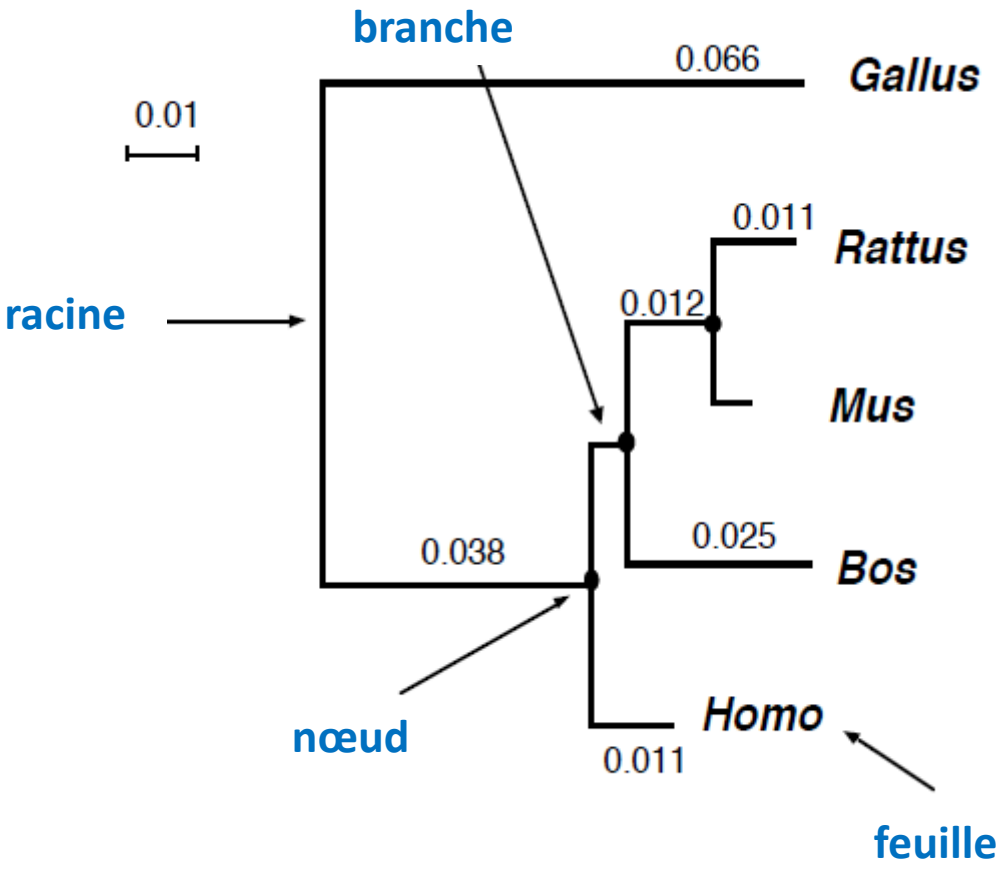
feuille: les extrémités des branches, ce sont les organismes modernes ou leurs macromolécules (protéines, ADN, ARN)

nœud: le dernier ancêtre commun de tous les organismes reliés par ce nœud

branche: relie les nœuds, représente le chemin évolutif parcouru entre l'ancêtre (nœud) et la feuille; la longueur de branches correspond au temps (si *échelle* = temps) ou la quantité de changement évolutif (si *échelle* = distance évolutive)

Utilisé plus fréquemment

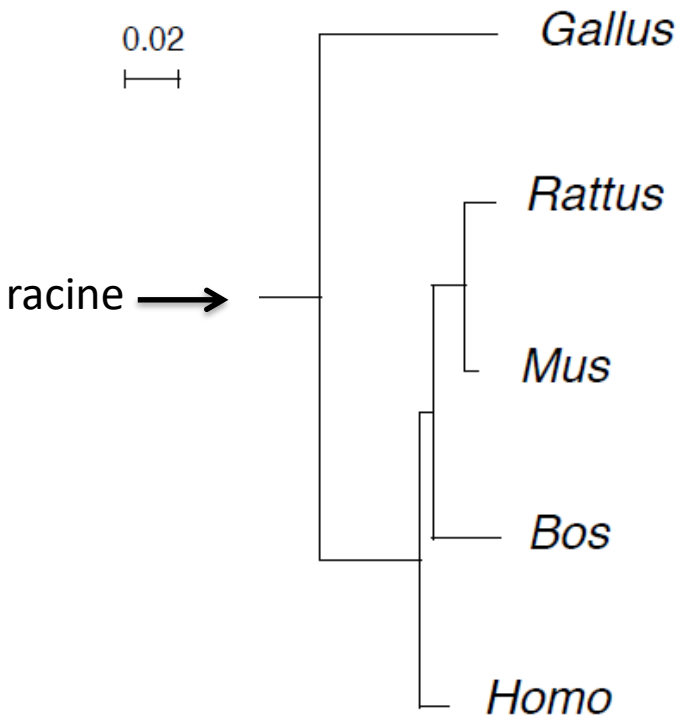
Comment lire un arbre phylogénétique



La longueur des branches horizontales est proportionnelle à la quantité d'évolution entre les séquences et leurs ancêtres (unité = substitution / site).

Comment lire un arbre phylogénétique

Arbre raciné: la racine (l'ancêtre commun de toutes les feuilles dans l'arbre) est le point le plus ancien et détermine le sens d'écoulement du temps sur toutes les branches

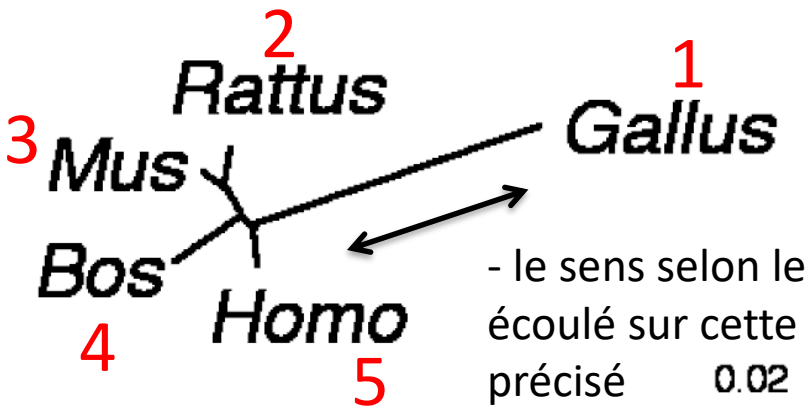


La plupart des méthodes produisent des arbres non enracinés car elles détectent des différences entre séquences mais n'ont aucun moyen d'orienter temporellement ces différences.

Comment lire un arbre phylogénétique

Arbres non racinés: le sens d'écoulement évolutif n'est pas précisé

- graphisme circulaire



- le sens selon lequel le temps c'est écoulé sur cette branche n'est pas précisé

- la racine pourrait être placé par exemple sur la branche du poulet (Gallus) ou de la souris (Mus) → cette dernière possibilité est erronée !

Comment lire un arbre phylogénétique

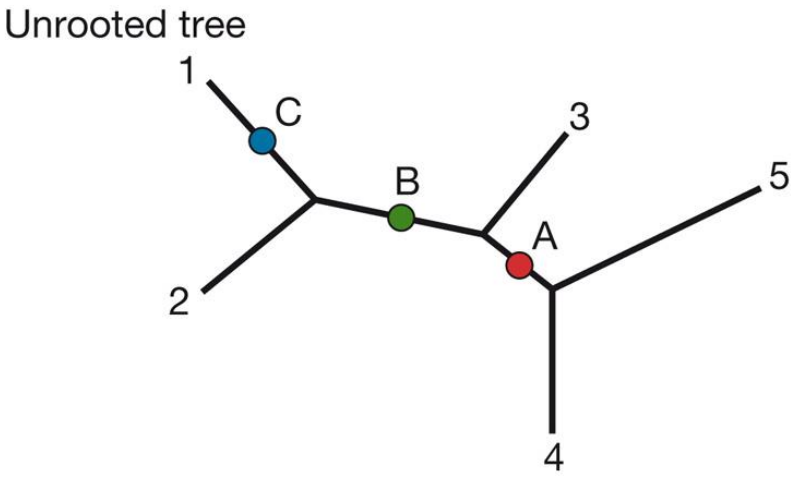
1 Gallus (oiseau)

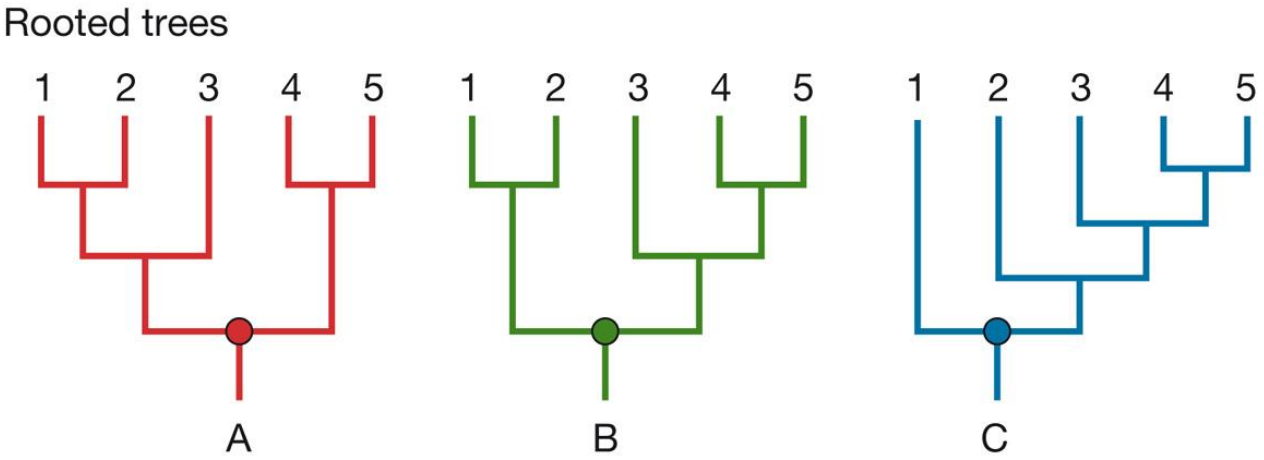
2 Rattus (mammifère)

3 Mus (mammifère)

4 Bos (mammifère)

5 Homo (mammifère)

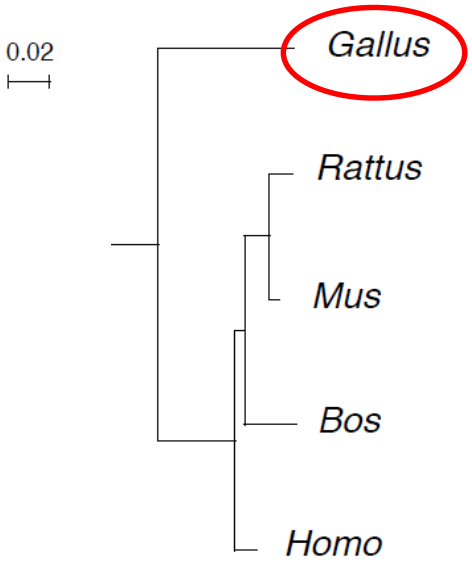


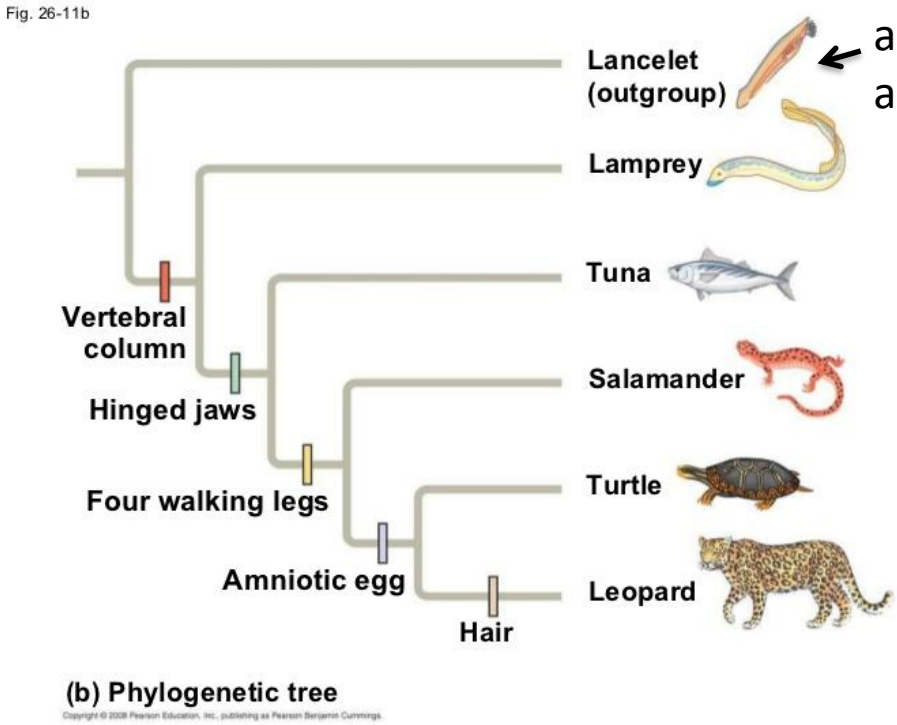


-la racine C est la seule qui décrit correctement l'évolution de ces organismes

Comment lire un arbre phylogénétique

Pour enraciner un arbre on peut utiliser un **groupe externe:**
- inclure dans les séquences analysés des séquences provenant d'organismes externes au groupe d'intérêt





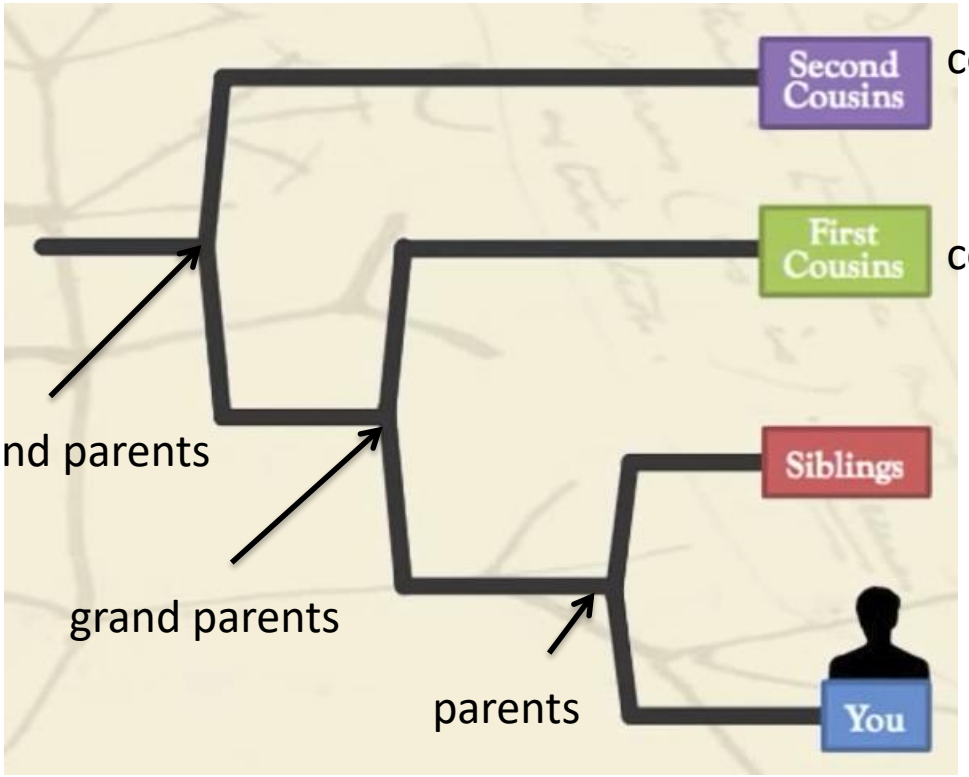
amphioxiformes: absence de vertèbre

Léopard est plus proche phylogénétiquement de la tortue que du salamandre – comment on déduit cela?

Comment lire un arbre phylogénétique

Qu'est ce que cela vaut dire: les espèces A et B sont plus apparentés que les espèces A et C? Dans un arbre phylogénétique, qui est plus proche de qui ?

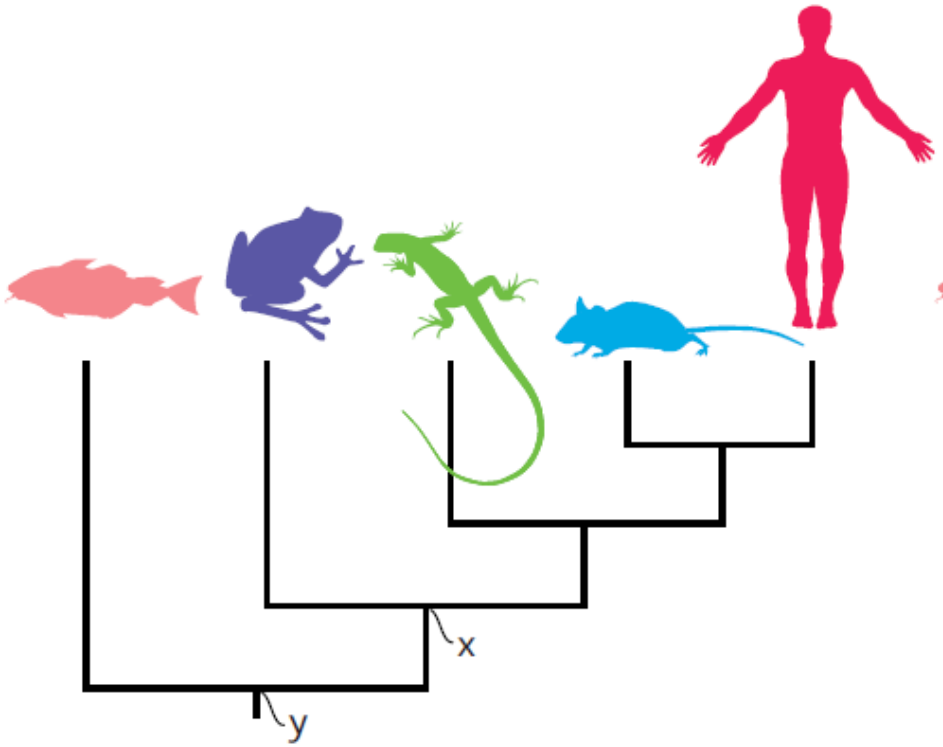
-les liens de parenté plus ou moins proches sont déterminés en fonction du dernier ancêtre commun: nous sommes plus proches de notre cousin germain que de notre cousin au 6-eme degré car notre dernier ancêtre commun (grand parents) a vécu une génération plutôt

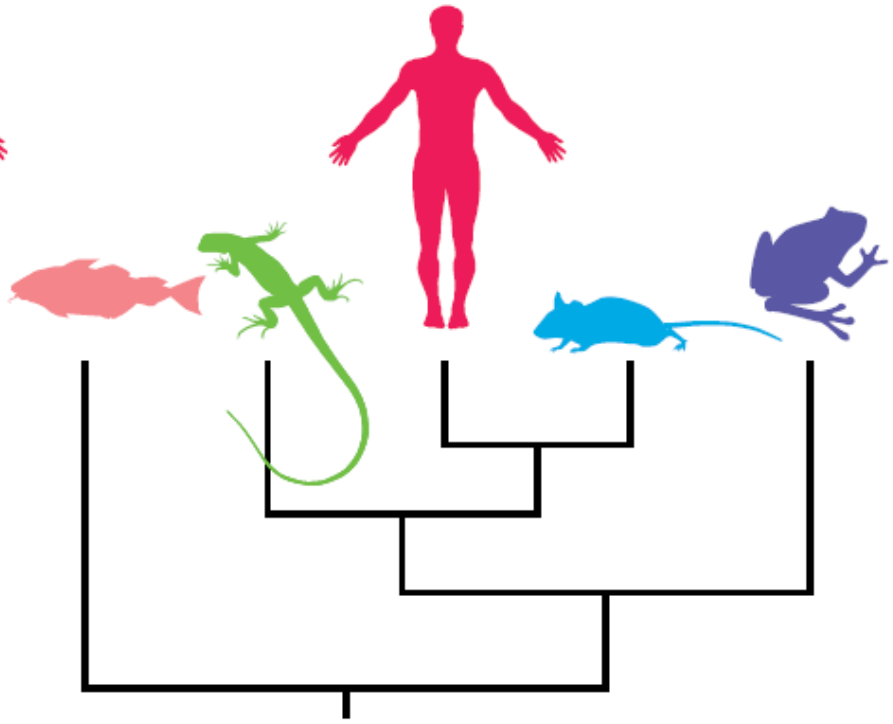


cousin 6-eme degré

cousin germain

Comment lire un arbre phylogénétique





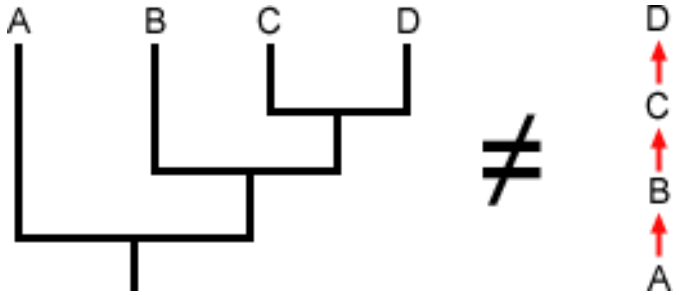
La grenouille est elle plus proche parent du poisson ou de l'homme?

Et pour cette arbre?

Comment lire un arbre phylogénétique

Quelques idées fausses:

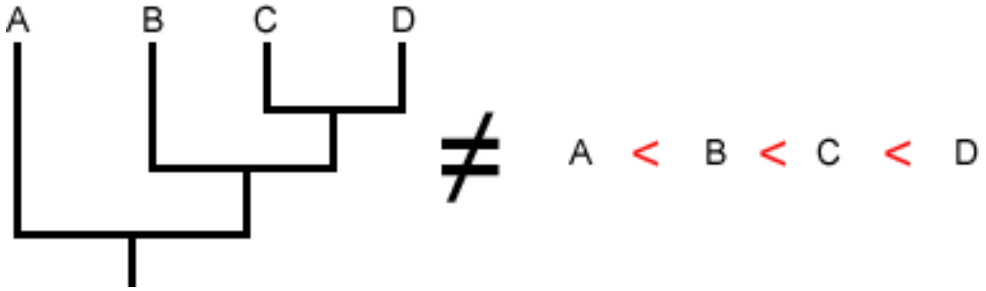
1. Les relation de parenté établies au cours de l'évolution produisent ressemblent à un arbre et non pas une échelle.



wrong

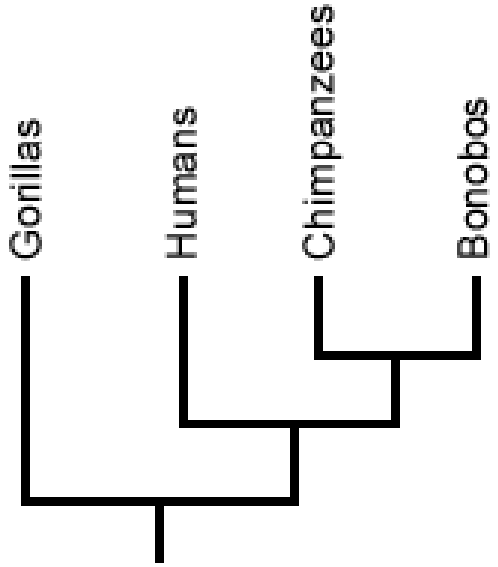
right

2. Les relation d'organismes supérieurs/inferieurs n'existent pas



Comment lire un arbre phylogénétique

3. Les humains ne sont pas les descendants des singes!



Les humains et les chimpanzés ont un ancêtre commun qui n'était ni un chimpanzé ni un homme.

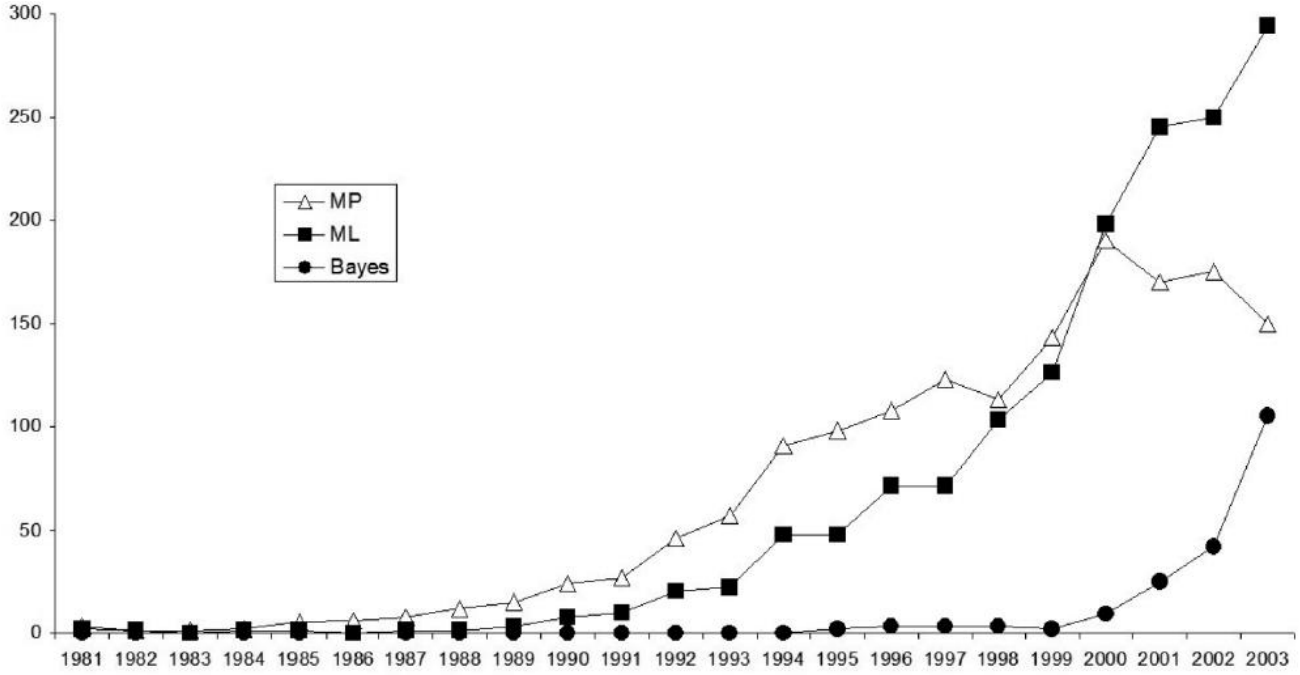
Les humains ne sont pas plus « évolués » que les autres êtres vivants!

Algorithmes pour la phylogénie moléculaire

→ Sont nombreux!

Algorithmes pour la phylogénie moléculaire

→ Parcimonie
→ Méthodes de distance
→ Maximum de vraisemblance
→ L'approche Bayésienne



Algorithmes pour la phylogénie moléculaire

LES MÉTHODES PROBABILISTES EN PHYLOGÉNIE MOLÉCULAIRE

(1) Les modèles d'évolution des séquences et le maximum de vraisemblance

Frédéric DELSUC et Emmanuel J. P. DOUZERY

Laboratoire de Paléontologie, Phylogénie et Paléobiologie,
Institut des Sciences de l'Évolution de Montpellier (ISEM), UMR 5554-CNRS,
Université Montpellier II, Montpellier, France

delsuc@isem.univ-montp2.fr

(2) L'approche bayésienne

Frédéric DELSUC et Emmanuel J. P. DOUZERY

Laboratoire de Paléontologie, Phylogénie et Paléobiologie,
Institut des Sciences de l'Évolution de Montpellier (ISEM), UMR 5554-CNRS,
Université Montpellier II, Montpellier, France

E-mail : douzery@isem.univ-montp2.fr

Algorithmes pour la phylogénie moléculaire

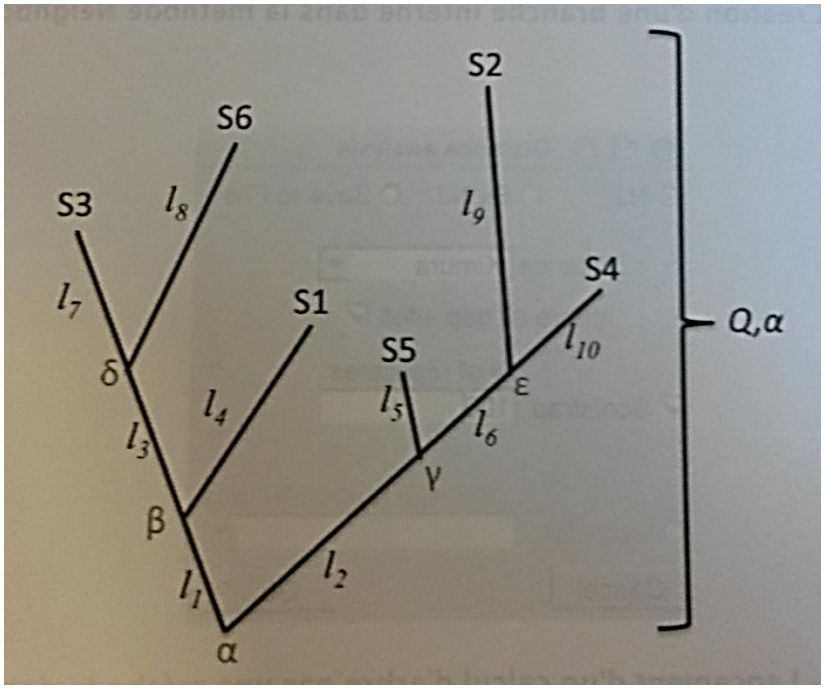
Consiste à probabiliser entièrement le processus évolutif (définir les probabilités de tous les évènements évolutifs possibles depuis n'importe quelle séquence ancestrale jusqu'aux feuilles), et à trouver quel est le scénario évolutif (topologie d'arbre et la longueur de ces branches) , qui a la plus forte probabilité d'avoir donné la naissance aux séquences analysées.

Solidement ancré dans les mathématiques et statistiques garantissant (*à condition que les données ont évolué selon les hypothèses probabilistes et que leur nombre tend vers l'infini*) que l'arbre trouvé sera l'arbre vrai des séquences analysées

Algorithmes pour la phylogénie moléculaire

LE MAXIMUM DE VRAISEMBLANCE

Objectif: calculer la vraisemblance du modèle, c'est-à-dire *la probabilité des données pour les valeurs des paramètres du modèle*



Les paramètres:

(i) Un arbre enraciné arbitrairement et ses longueurs de branches (l = nr. subst./site, lettres grecques = résidus ancestraux inconnus, toutes les combinaisons de valeurs sont envisagées)

(ii) Matrice de taux Q* commune à toutes les branches

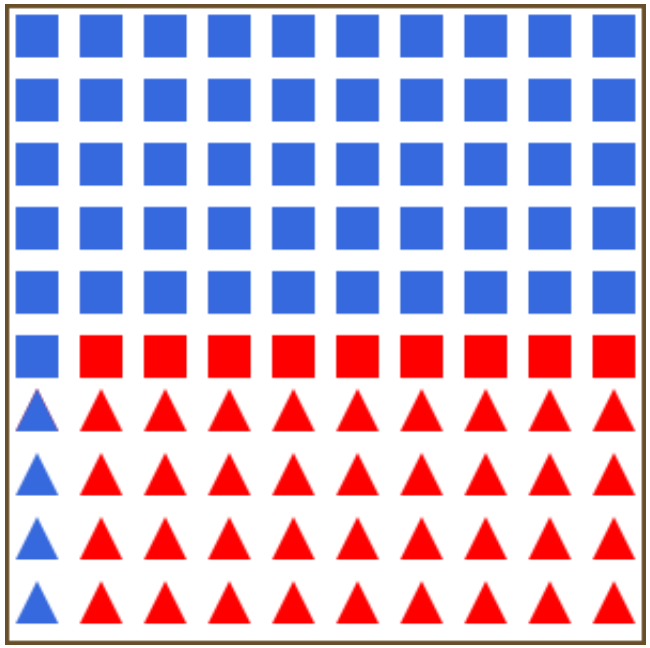
(iii) Valeur α qui détermine la variation de taux d'évolution entre sites

* Permet de modéliser le processus d'évolution d'un site d'une séquence

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

100 objets
60 carrés
9 carrés rouges



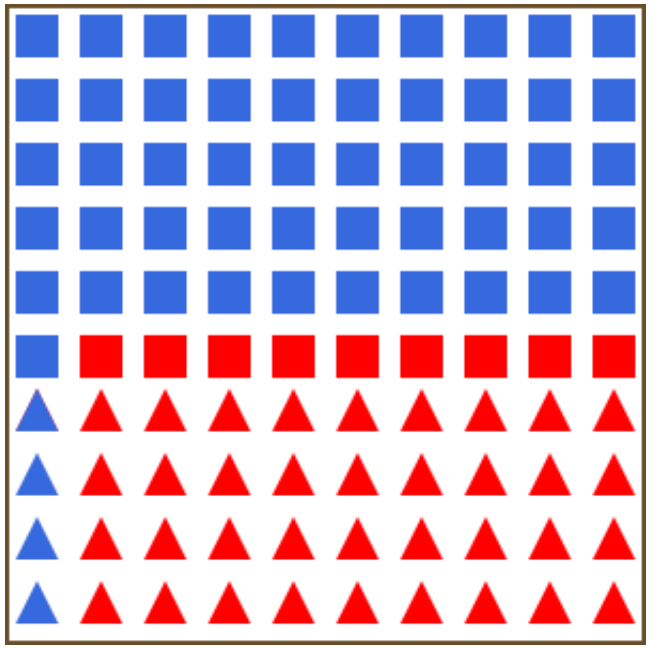
→ Probabilité de tirer un carré? P(Carré)=?

https://sciencetonnante.wordpress.com/2012/10/08/les-probabilites-conditionnelles-bayes-level-1/
https://sciencetonnante.wordpress.com/2012/10/15/linference-bayesienne-bayes-level-2/

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

100 objets
60 carrés
9 carrés rouges

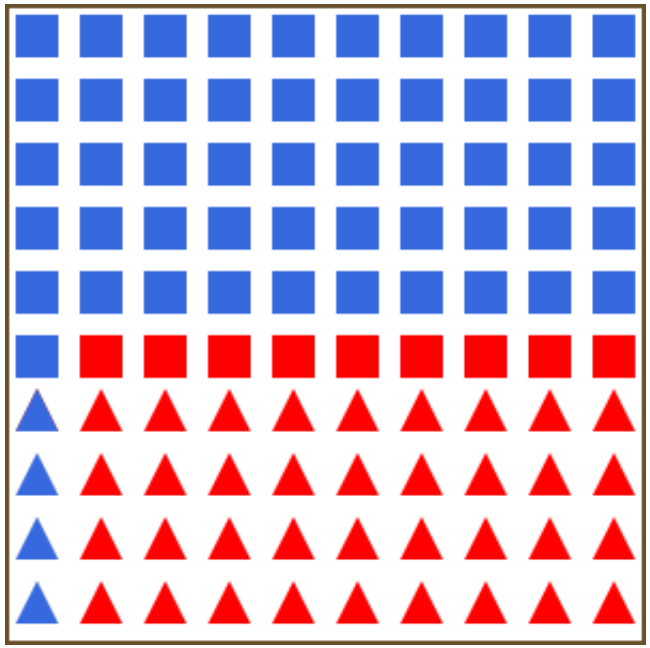


→ Probabilité de tirer un carré? P(Carré)= 60/100 = 60%

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

100 objets
60 carrés
9 carrés rouges

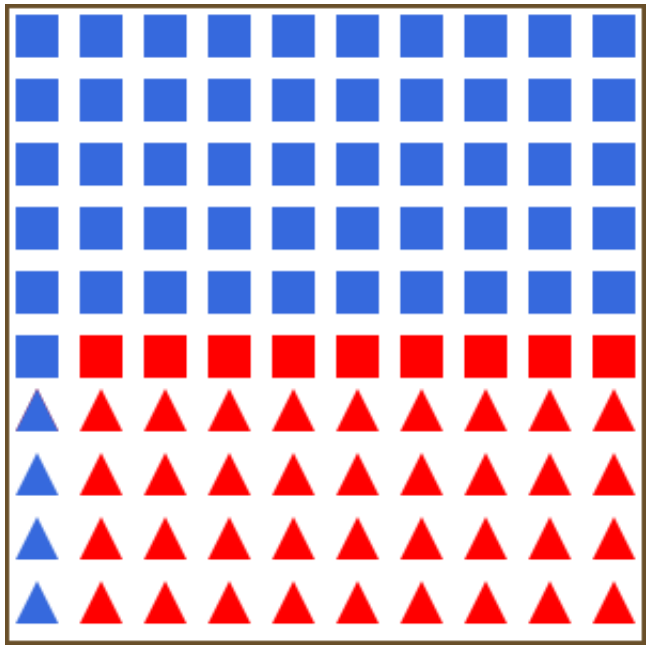


→ Probabilité de tirer un carré sachant qu'il est rouge? P (Carré | Rouge) = ?

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

100 objets
60 carrés
9 carrés rouges
36 triangles rouges



→ Probabilité de tirer un carré sachant qu'il est rouge? On considère uniquement les objets rouges (Carré | Rouge) = 9/45 = 20 %

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

la probabilité que l'objet soit un carré est fortement affectée par le fait de savoir qu'il est rouge. La « probabilité que l'objet soit carré »

$$P(\text{Carré}) = 60\%$$

n'est pas la même que la « probabilité que l'objet soit carré *sachant qu'il est rouge*

$$P(\text{Carré} \mid \text{Rouge}) = 9/45 = 20\ \%$$

ni la même que la « probabilité que l'objet soit rouge *sachant qu'il est carré*

$$P(\text{Rouge} \mid \text{Carré}) = 9/60 = 15\%$$

Algorithmes pour la phylogénie moléculaire

→ la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier:

la probabilité que l'objet soit un carré est fortement affectée par le fait de savoir qu'il est rouge. La « probabilité que l'objet soit carré »

$$P(\text{Carré}) = 60\%$$

n'est pas la même que la « probabilité que l'objet soit carré *sachant qu'il est rouge*

$$P(\text{Carré} \mid \text{Rouge}) = 9/45 = 20\ \%$$

ni la même que la « probabilité que l'objet soit rouge *sachant qu'il est carré*

$$P(\text{Rouge} \mid \text{Carré}) = 9/60 = 15\%$$

→ **comment connaître P(Rouge|Carré) si on connait P(Carré|Rouge)?**

Algorithmes pour la phylogénie moléculaire

→ **comment connaître P(Rouge|Carré) si on connait P(Carré|Rouge)?**

Thomas Bayes
(1701 – 1761)

→ Théorème de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(Carré | Rouge) = P(Rouge| Carré) x P(Carré) / P(Rouge)

= 9/60 x 60/100 / 45/100 = 0.09/0.45 = 0.2 = **20 %**

→ Probabilité de tirer un carré sachant qu'il est rouge? On considère uniquement les objets rouges (Carré | Rouge) = 9/45 = **20 %**

Algorithmes pour la phylogénie moléculaire

APPROCHE BAYESIENNE

Thomas Bayes
(1701 – 1761)

→ formule de Bayes : permet de calculer les *probabilités à posteriori* d'une hypothèse en se basant sur **les données connues**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

→ Utilisé en médecine (Quelle est la probabilité d'avoir la maladie si je suis dépisté positif?), identification des messages spam (Quelle est la probabilité que le message est un spam si il contient le mot viagra?)

→ Si la **qualité des données** est bonne le théorème de Bayes donnera des résultats fiables (« garbage in - garbage out »)

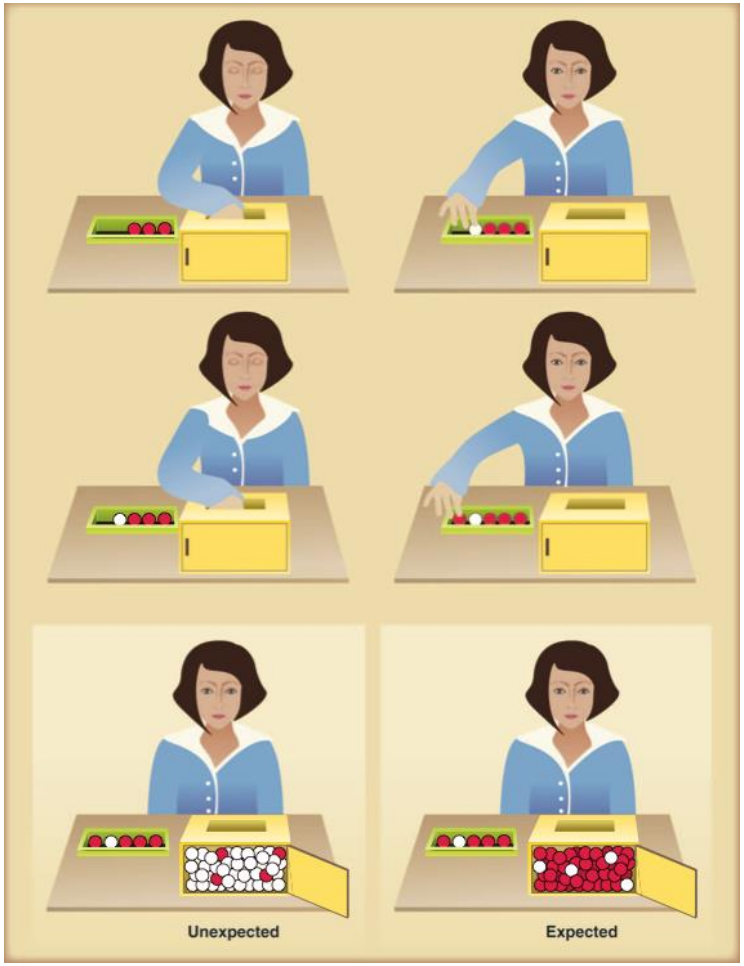
Algorithmes pour la phylogénie moléculaire

→ notre cerveau fait des raisonnements bayésiens inconscients

Test réalisé sur des enfants de 8 (!) mois:

Capacité d'émettre une hypothèse sur le contenu de la boite à partir de l'échantillon qu'on leur avait présenté

→ Les enfants sont capables de réaliser des inductions bayésiennes



Algorithmes pour la phylogénie moléculaire

L'approche Bayésienne appliqué à la phylogénie moléculaire:

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

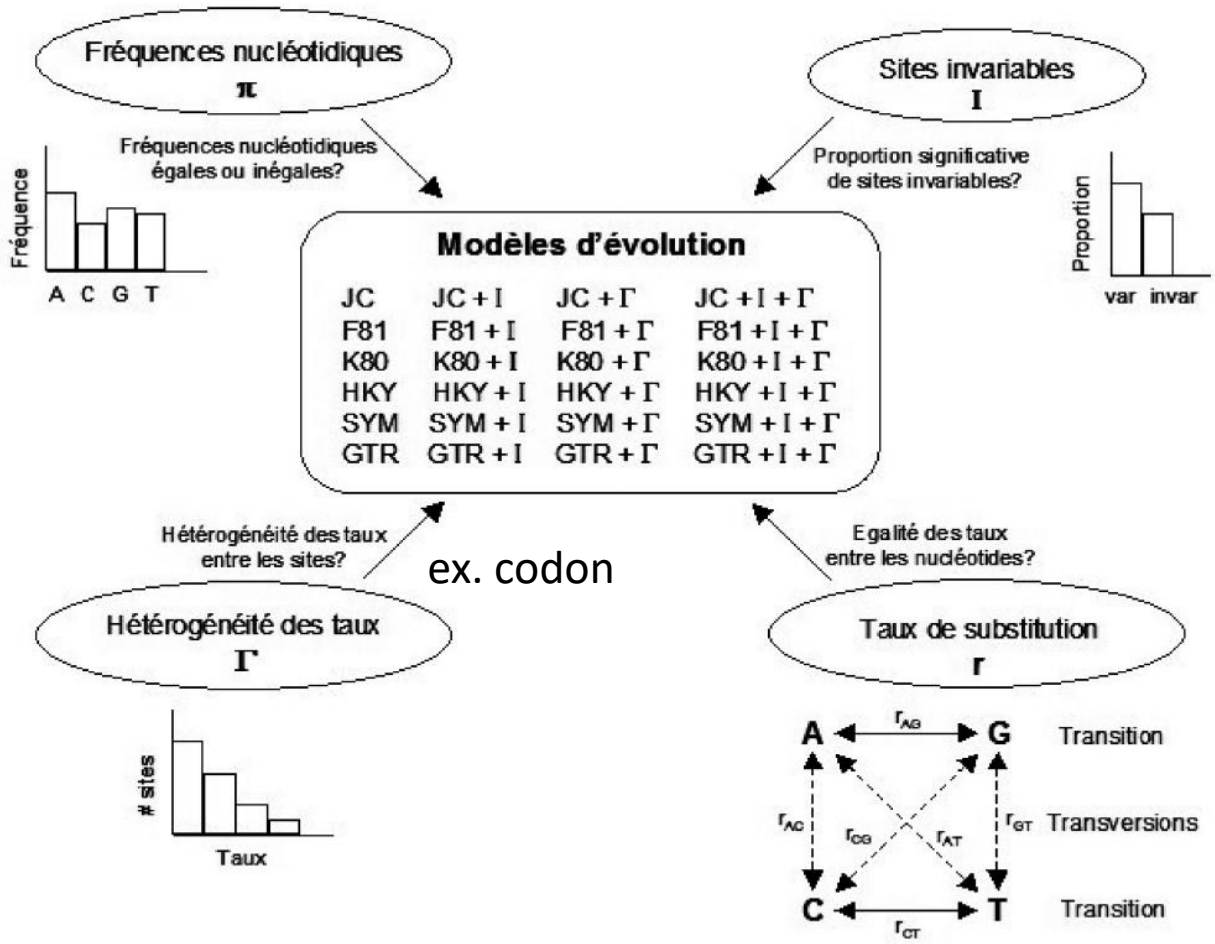
[Data] = données = alignement de séquences
[Tree] = arbre = topologie de l'arbre y compris la longueur de branches (*évolution des séquences*)

→ *La méthode bayésienne produit une collection d'arbres dont l'information phylogénétique peut être résumée en calculant le consensus majoritaire.*

→ *La fréquence avec laquelle les différents nœuds apparaissent dans les arbres visités représente leur probabilité postérieure associée*

Algorithmes pour la phylogénie moléculaire

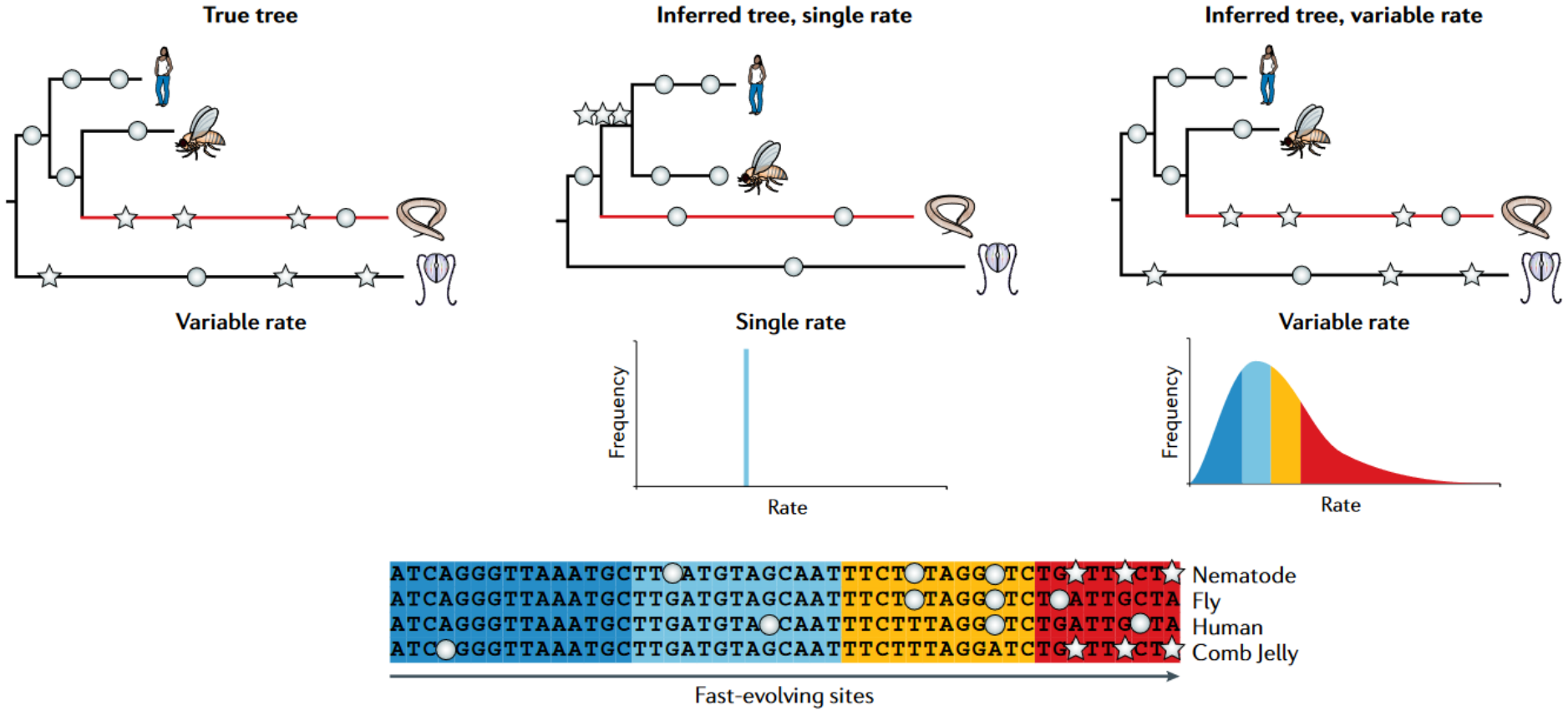
→ pour être statistiquement robustes et performantes, les méthodes probabilistes nécessitent l'incorporation de *modèles* qui décrivent de la façon la plus réaliste possible les processus biologiques *d'évolution des séquences*



r (transition) > r (transversion)

Algorithmes pour la phylogénie moléculaire

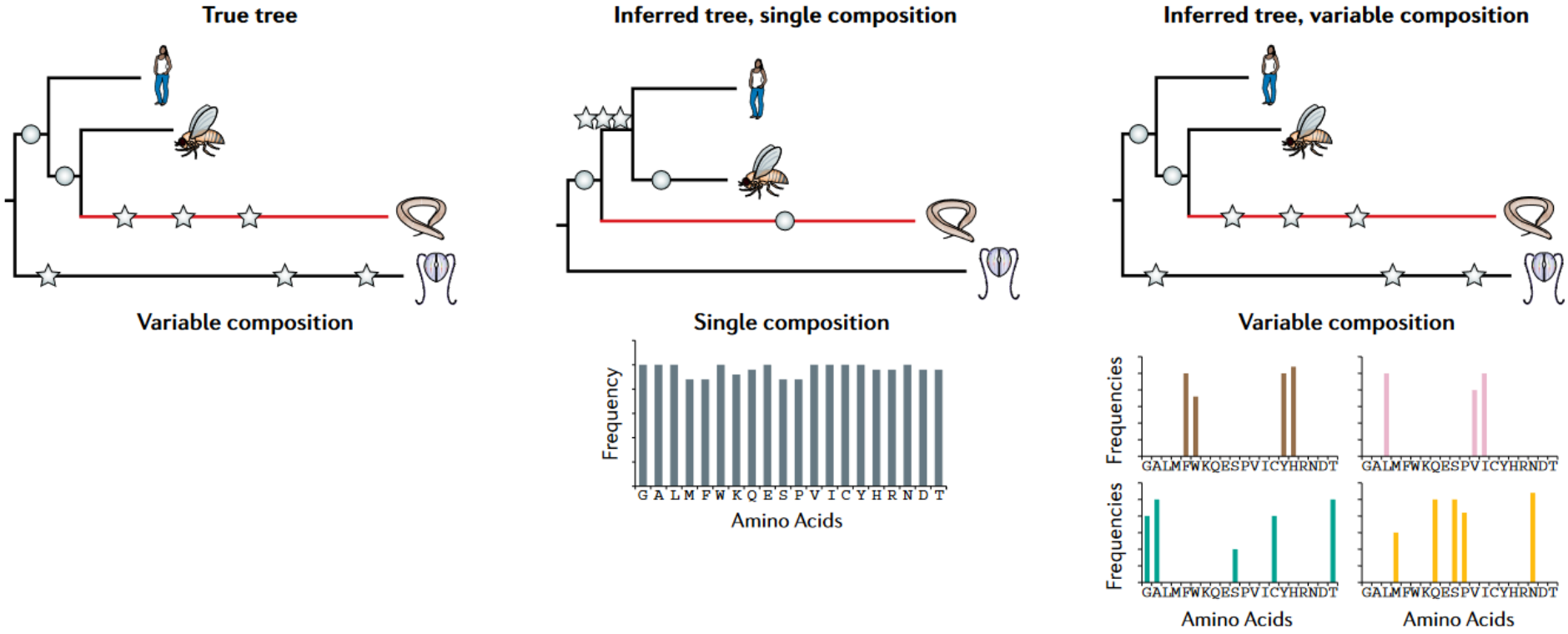
Hétérogénéité des taux de substitution entre les sites



Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428-444. doi: 10.1038/s41576-020-0233-0. Epub 2020 May 18. PMID: 32424311.

Algorithmes pour la phylogénie moléculaire

Hétérogénéité de composition entre les sites



Algorithmes pour la phylogénie moléculaire

→les méthodes probabilistes sont robustes vis-à-vis du non-respect de leur hypothèse de base mais par conséquence très dépendantes du modèle d'évolution choisi

A.

Complexité du modèle (- → +)

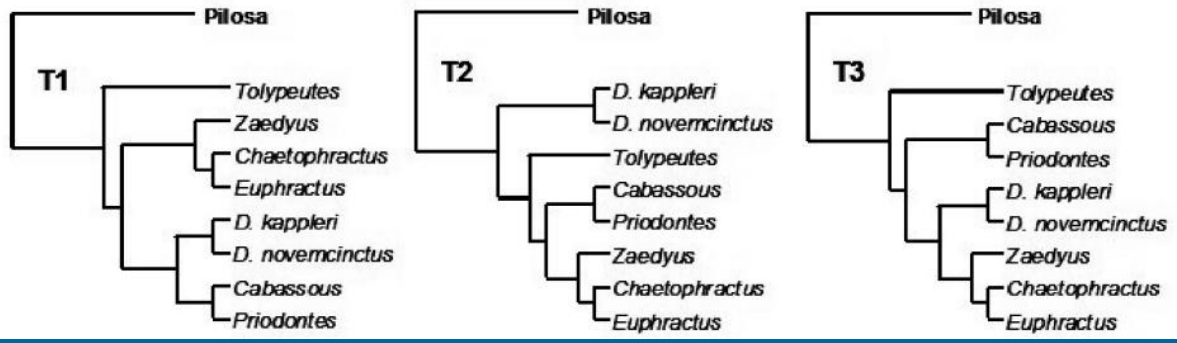
Complexité du modèle (- → +)

	+I	+Γ_8	+I +Γ_8
JC69	-4954,46 (T1)	-4762,20 (T1)	-4722,89 (T1)	-4722,79 (T1)
F81	-4920,87 (T1)	-4725,38 (T1)	-4680,57 (T1)	-4680,52 (T1)
K80	-4772,73 (T2)	-4565,29 (T2)	-4515,29 (T1)	-4515,05 (T1)
HKY85	-4721,72 (T2)	-4505,24 (T2)	-4433,33 (T3)	-4433,28 (T3)
TN93ef	-4701,20 (T2)	-4499,26 (T2)	-4457,37 (T1)	-4457,19 (T1)
TN93	-4670,37 (T2)	-4667,27 (T2)	-4413,94 (T1)	-4413,91 (T1)
K81	-4771,03 (T2)	-4563,07 (T2)	-4512,60 (T1)	-4512,33 (T1)
K81uf	-4720,69 (T2)	-4503,85 (T2)	-4431,75 (T3)	-4431,68 (T3)
TIMef	-4745,65 (T2)	-4545,29 (T2)	-4495,33 (T1)	-4495,19 (T1)
TIM	-4698,19 (T2)	-4490,40 (T2)	-4424,72 (T1)	-4424,71 (T1)
TVMef	-4726,08 (T2)	-4516,96 (T2)	-4476,06 (T1)	-4475,76 (T1)
TVM	-4693,69 (T2)	-4482,63 (T2)	-4426,06 (T3)	-4426,00 (T3)
SYM	-4701,20 (T2)	-4499,26 (T2)	-4457,37 (T1)	-4457,19 (T1)
GTR	-4670,37 (T2)	-4467,27 (T2)	**-4413,94 (T1)**	-4413,91 (T1)

→ selon le modèle choisi - trois topologies différentes

→ l'approche bayésienne *est sensible à l'inadéquation du modèle utilisé pour décrire l'évolution des séquences*

B.



Algorithmes pour la phylogénie moléculaire

→ Le principe de bootstrap:

Rééchantillonnage numérique par **tirage aléatoire avec remplacement** des sites nucléotidiques à introduire dans le nouveau jeu de données:

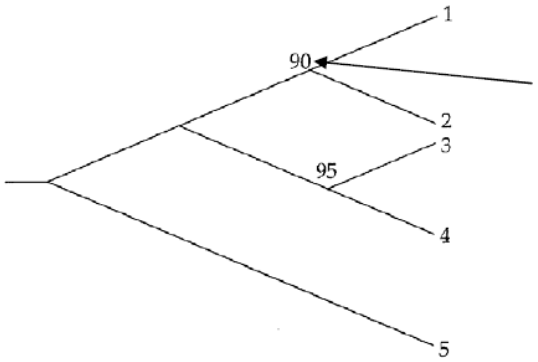
Jeu de données réel:

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

Jeu de données rééchantillonné:
(tirage aléatoire des sites 6,1,6,…)

1	AGATACTC…
2	AGATGCTT…
3	AAACACTC…
4	AAACACCC…
5	AGACATCC…

Un total de 100 jeux de données et donc 100 arbres sont obtenus dont 90 → clade (12) et 95 → clade (34)

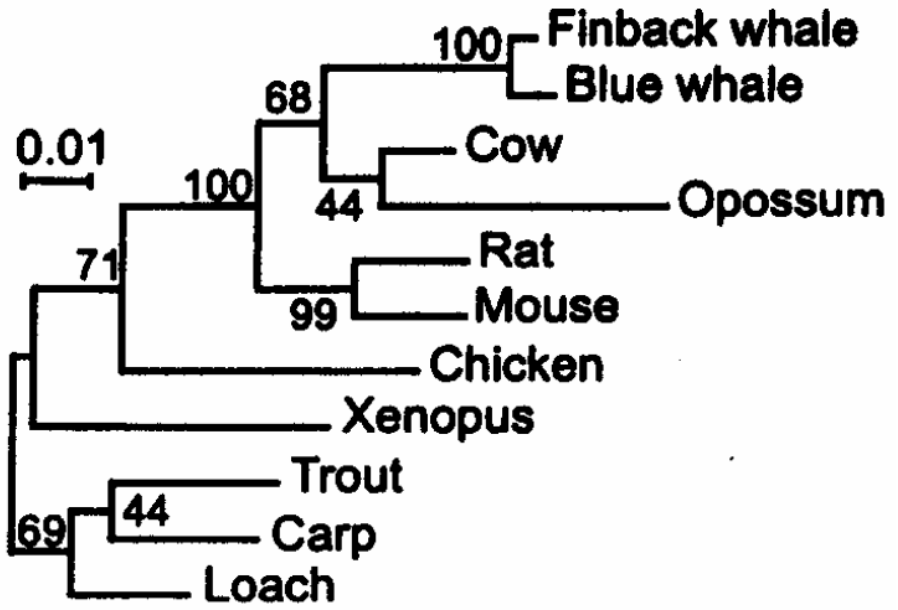


"proportion de bootstrap" = niveau de confiance

http://gepv.univ-lille1.fr/downloads/enseignements/M1-S8/M1-S8-Vekemans-Chap3-Phylo_molec.pdf

Algorithmes pour la phylogénie moléculaire

Interprétation des valeurs de **proportion de bootstrap** pour une branche donnée d'un arbre phylogénétique: ≈ probabilité de confiance que la longueur de la branche soit supérieure à zéro (la branche est dite "significative" si >95% ou 99%)

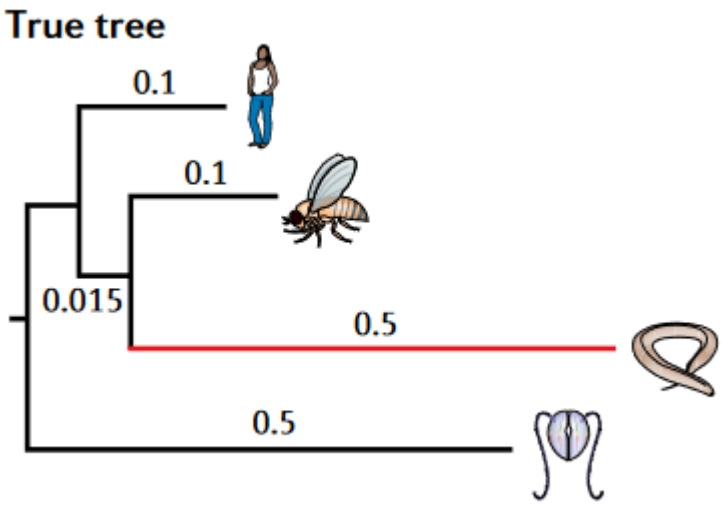


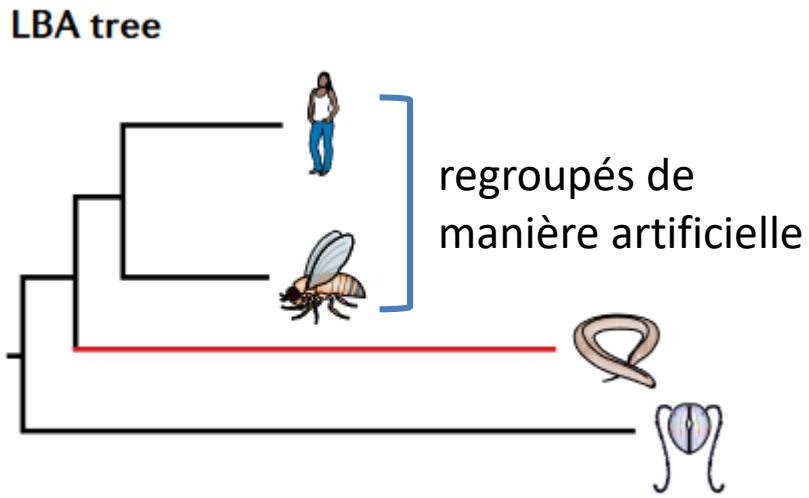
ex. arbre consensus sur 500 bootstraps avec méthode ML sur données de séquences de la cytochrome oxidase (gène mitochondrial)

(Nei & Kumar, 2000)

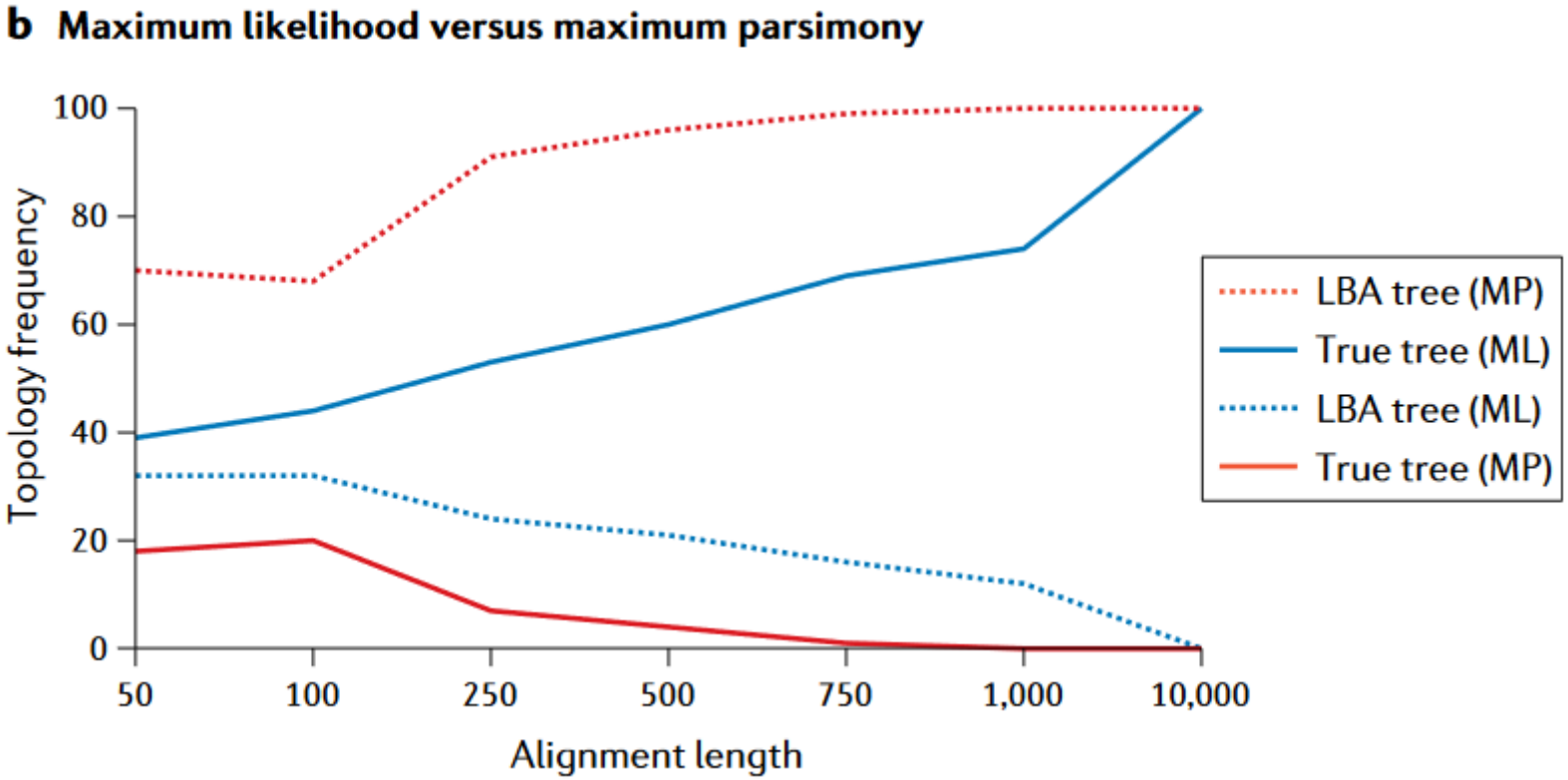
Algorithmes pour la phylogénie moléculaire

→ Les taux hétérogène de substitution au sein des différentes espèces peut conduire à **l'attraction de longues branches** un artéfact connu de l'analyse phylogénétique

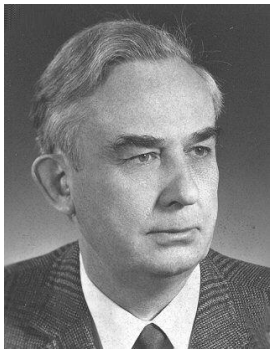




Algorithmes pour la phylogénie moléculaire



Classification cladistique



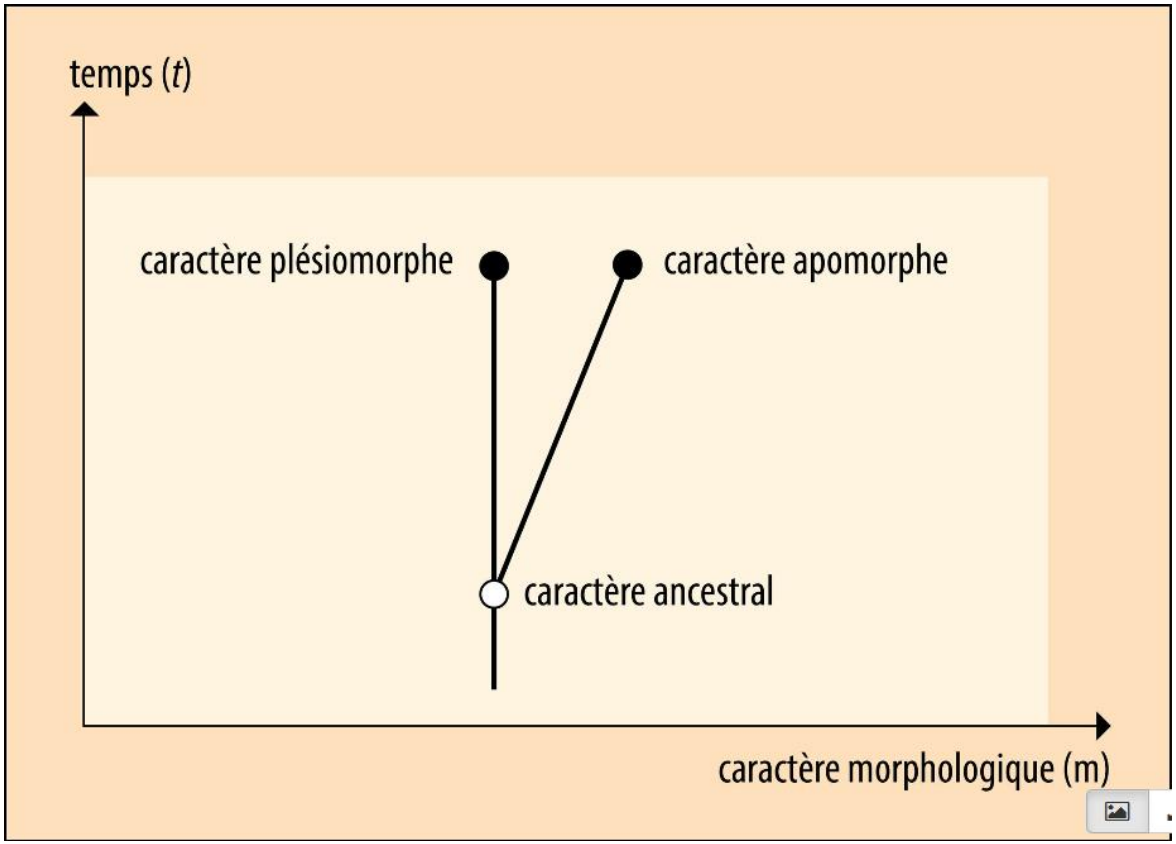
Willi Hennig
(1913-1976)

→ Entomologiste Allemand, revolutionne au milieu du XX siècle la classification du vivant en jetant les bases de la **cladistique**

1950, Grundzüge einer Theorie der *Phylogenetischen* Systematik

→ la distinction – pour un caractère donné – entre les états **plésiomorphes** (primitifs ou généraux) et **apomorphes** (évolués ou particuliers).

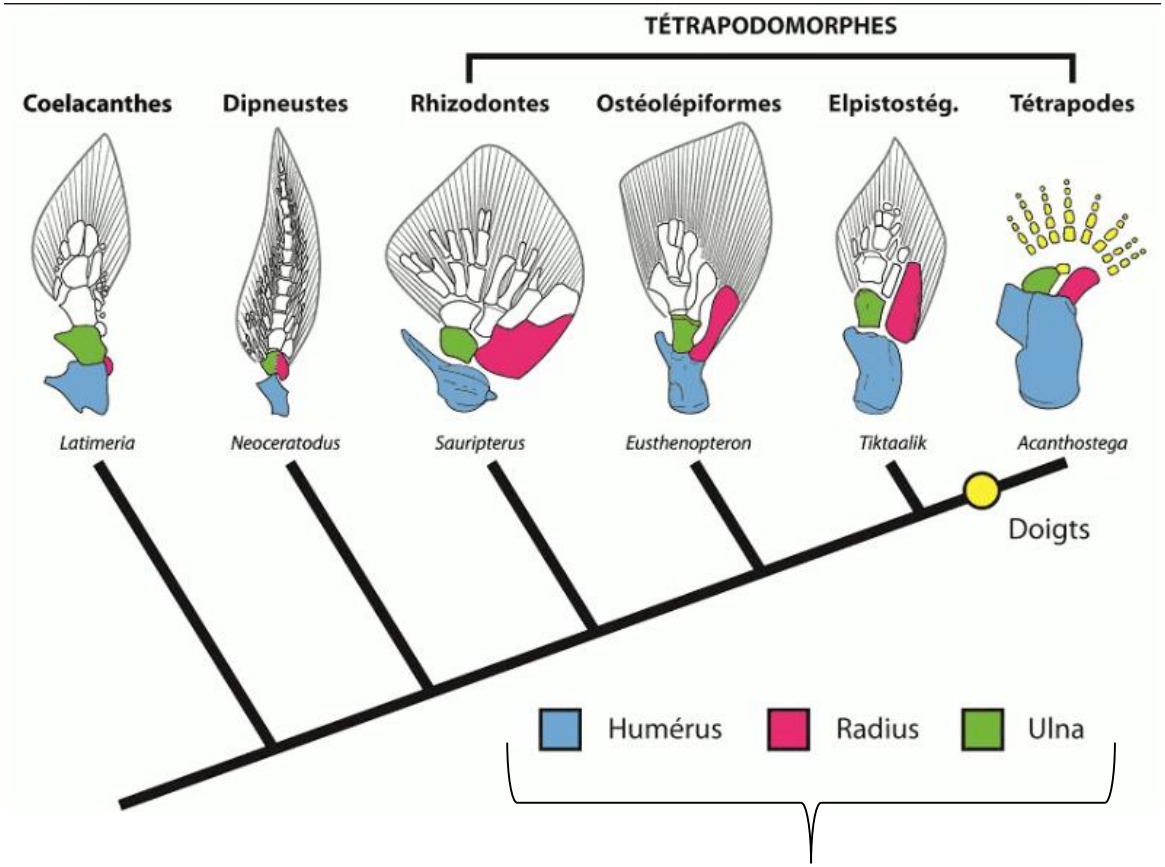
Classification cladistique



→ le caractère se transforme au cours du temps (apomorphe) , soit il perdure sans changer (plésiomorphe)

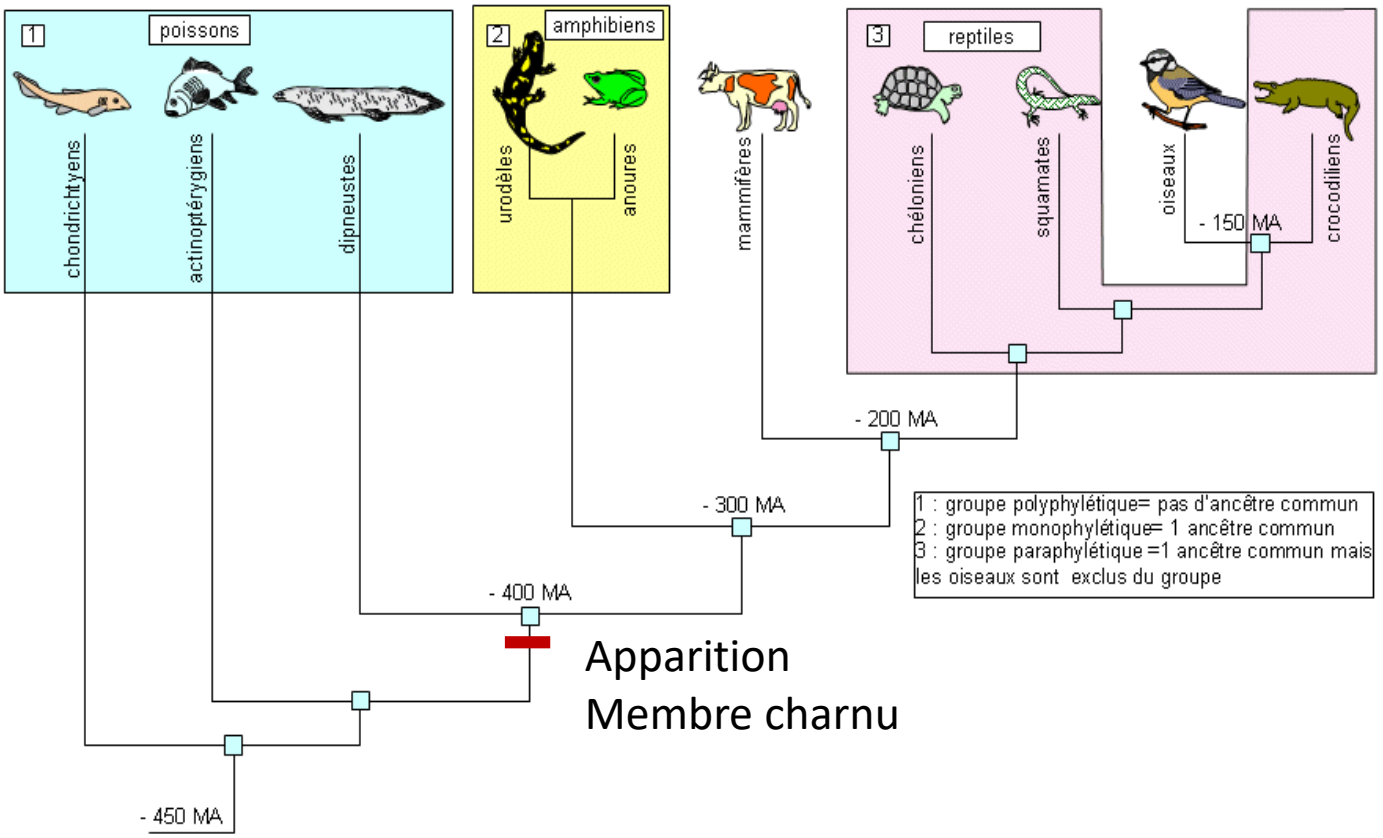
Classification cladistique

Exemple de membre charnu



Caractères synapomorphes

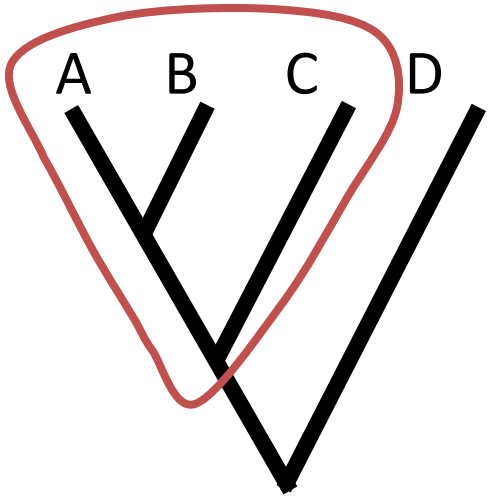
Classification cladistique



Classification cladistique

→ Les relations entre les espèces ou les taxons sont indiquées par les états apomorphes *qu'ils partagent* = ***synapomorphies***

→ Ces relations sont illustrées par un arbre dichotomique ou **cladogramme**

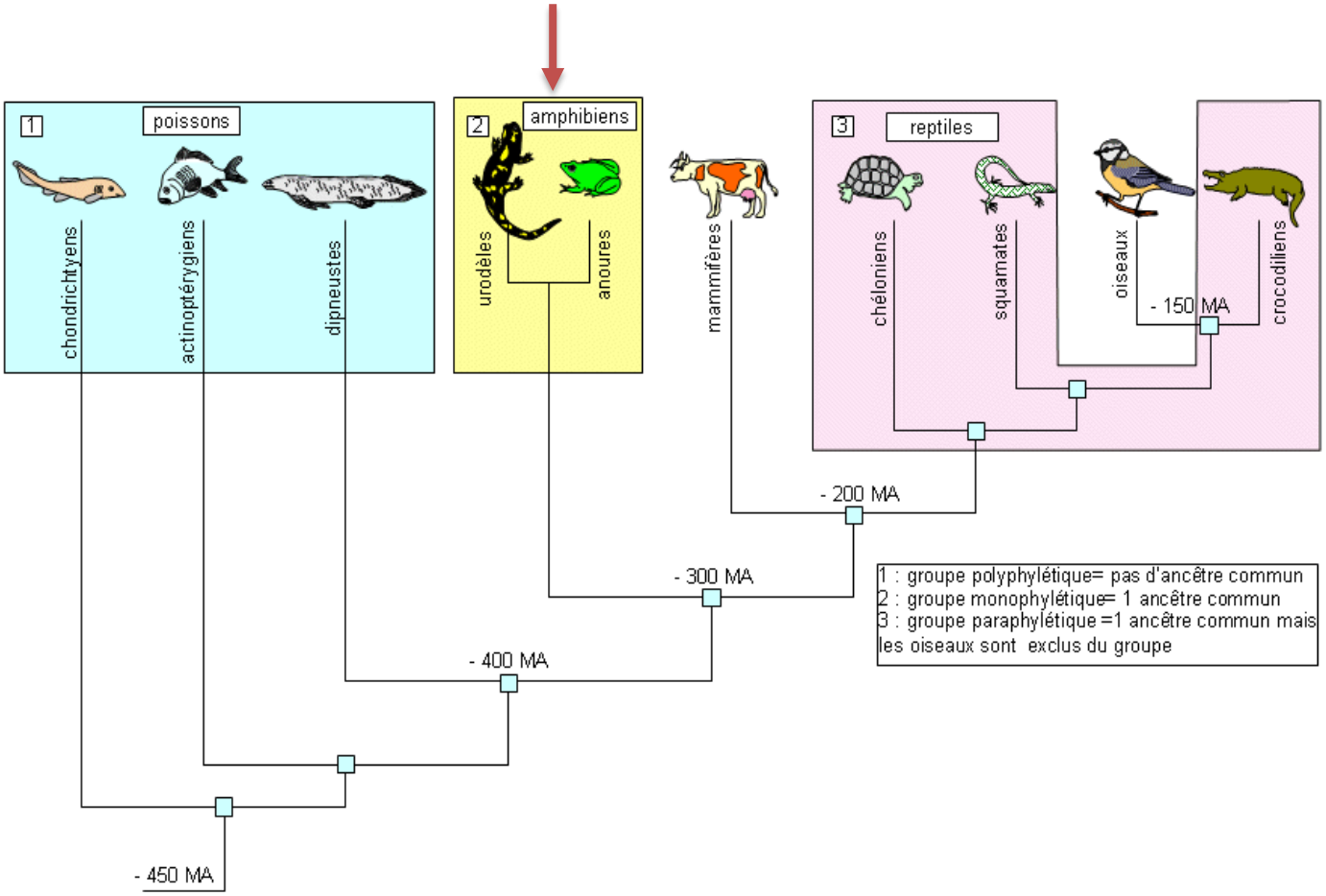


Clade: groupe *monophylétique* d'organismes comprenant tous les descendants d'un ancêtre commun

Uniquement les groupes monophylétiques (clades) sont considérés par la classification cladistique – à des différents niveaux taxonomiques (domaine, règne, ordre, famille etc.)

Classification cladistique

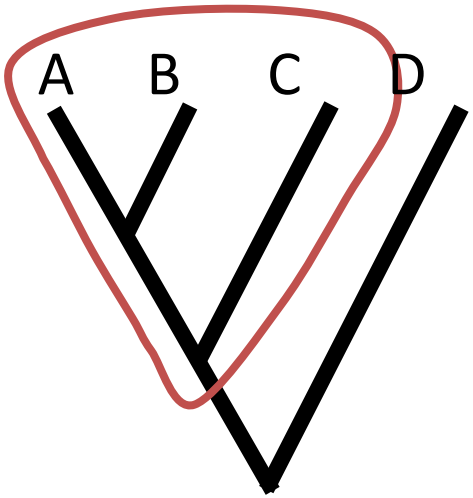
Amphibiens – groupe monophylétique



Classification cladistique

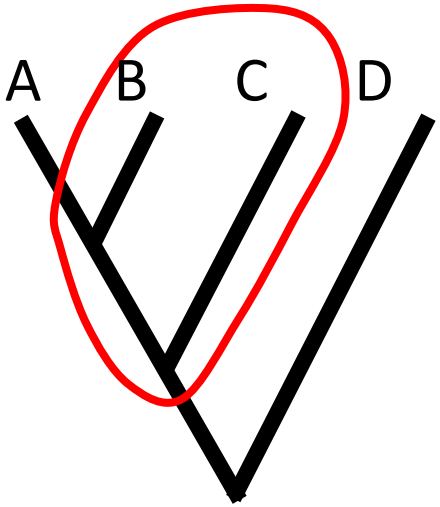
Monophyletic group

Includes an ancestor all of its descendants



Paraphyletic group

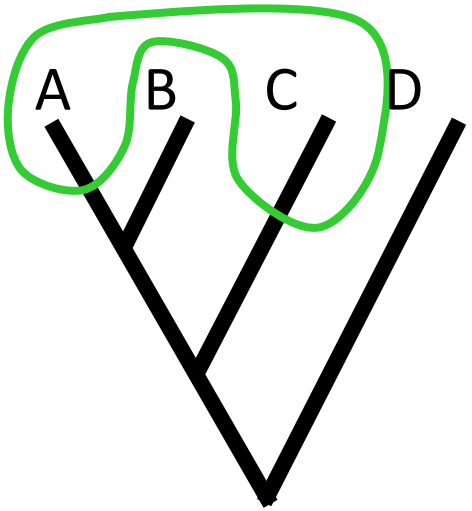
Includes ancestor and some, but not all of its descendants



Taxon A is highly derived and looks very different from B, C, and ancestor

Polyphyletic group

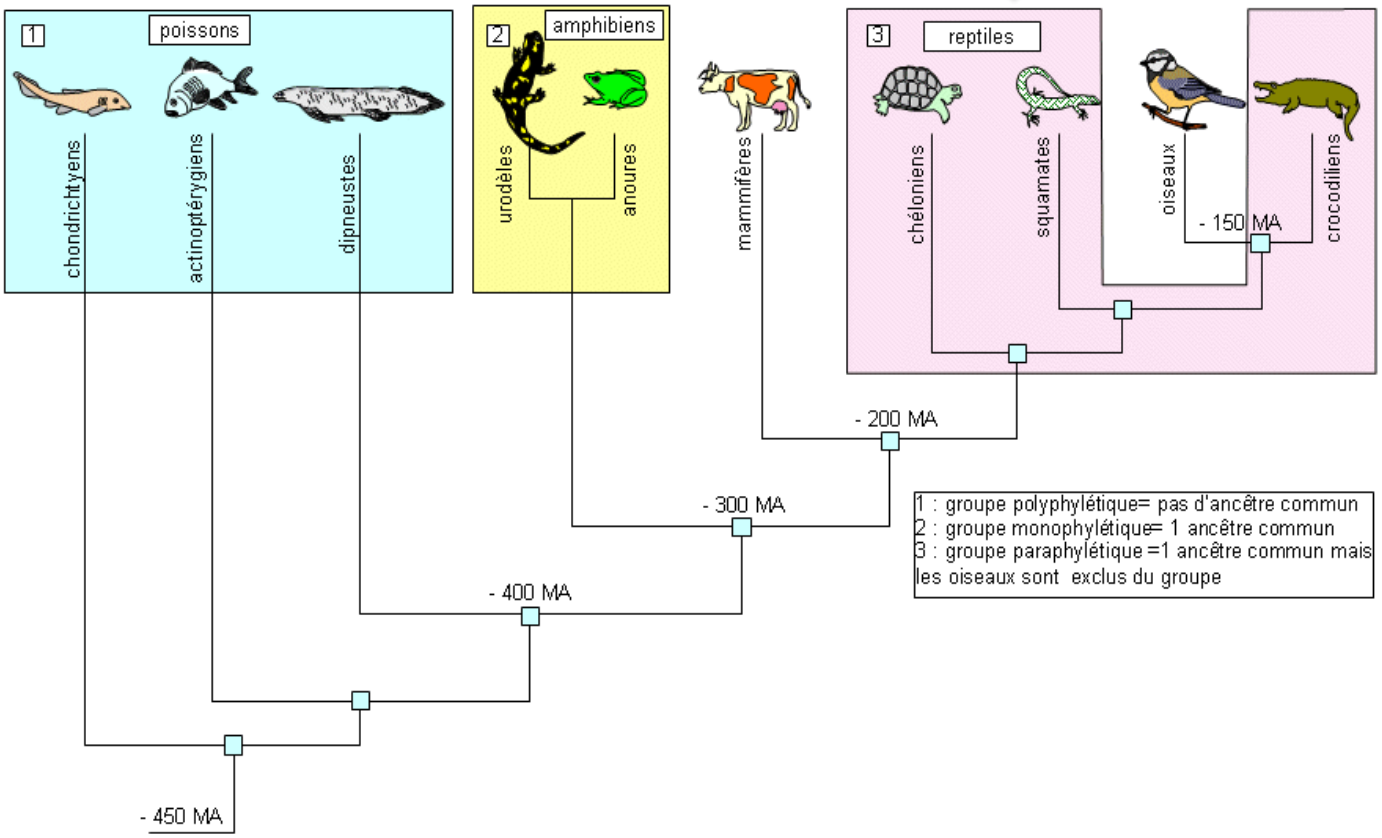
Includes two convergent descendants but not their common ancestor



Taxon A and C share similar traits through convergent evolution

Classification cladistique

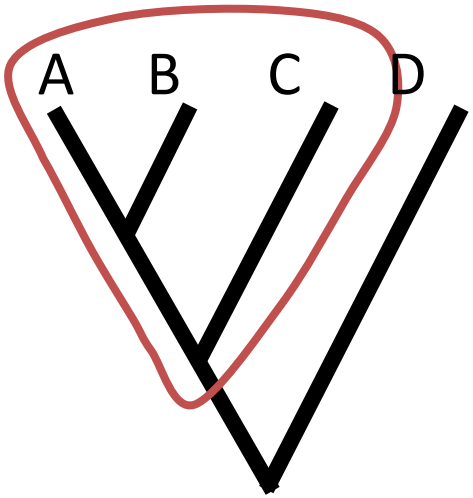
Reptiles – groupe paraphylétique



Classification cladistique

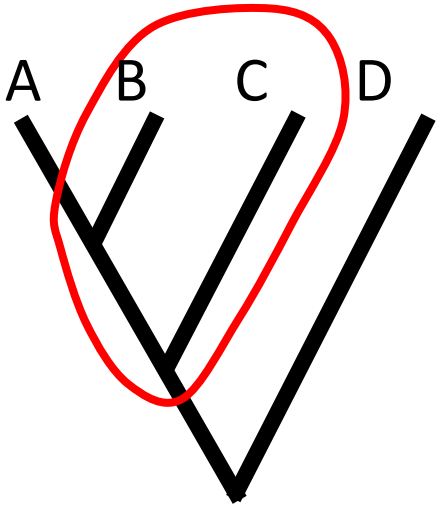
Monophyletic group

Includes an ancestor all of its descendants



Paraphyletic group

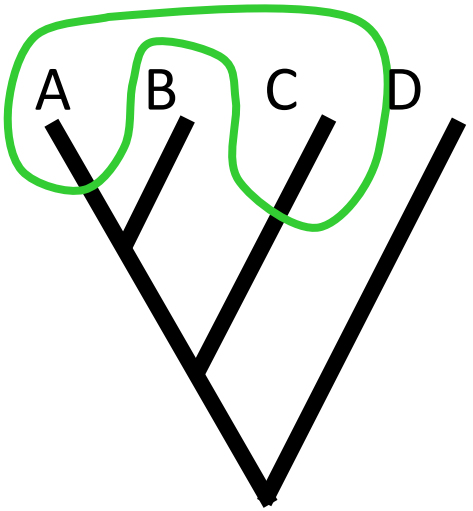
Includes ancestor and some, but not all of its descendants



Taxon A is highly derived and looks very different from B, C, and ancestor

Polyphyletic group

Includes two convergent descendants but not their common ancestor



Taxon A and C share similar traits through convergent evolution

Classification cladistique

Poissons – groupe polyphylétique