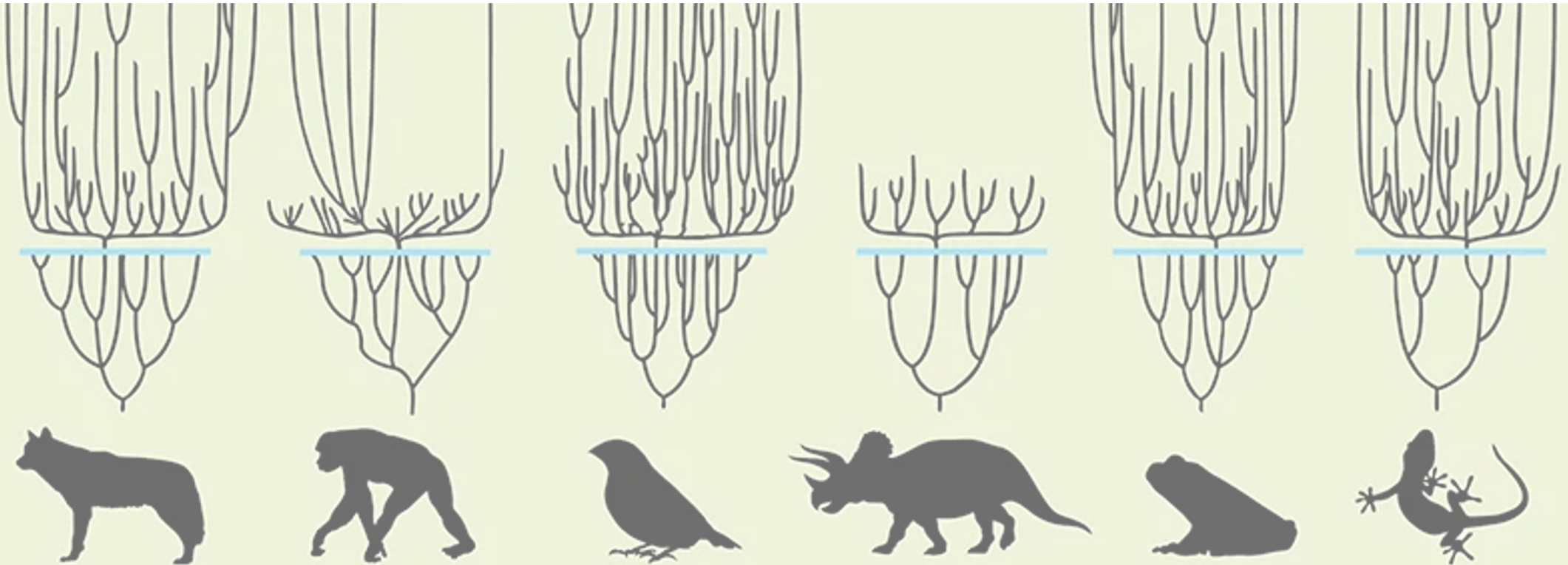


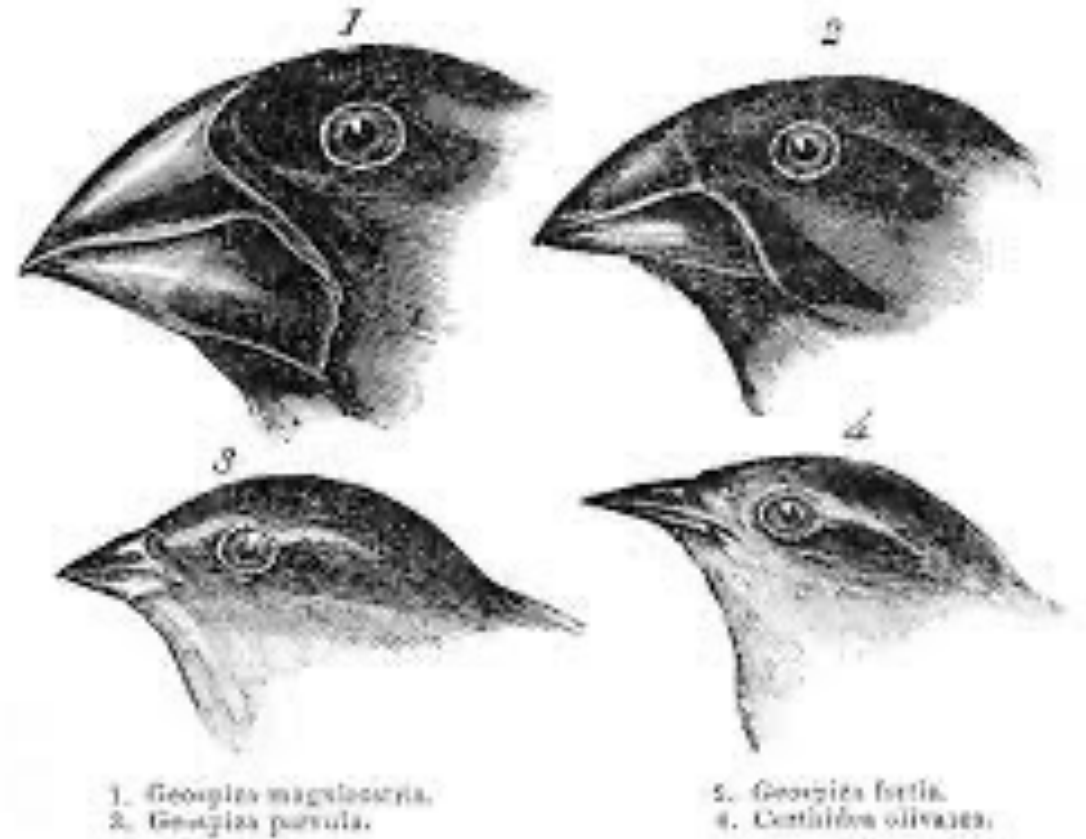
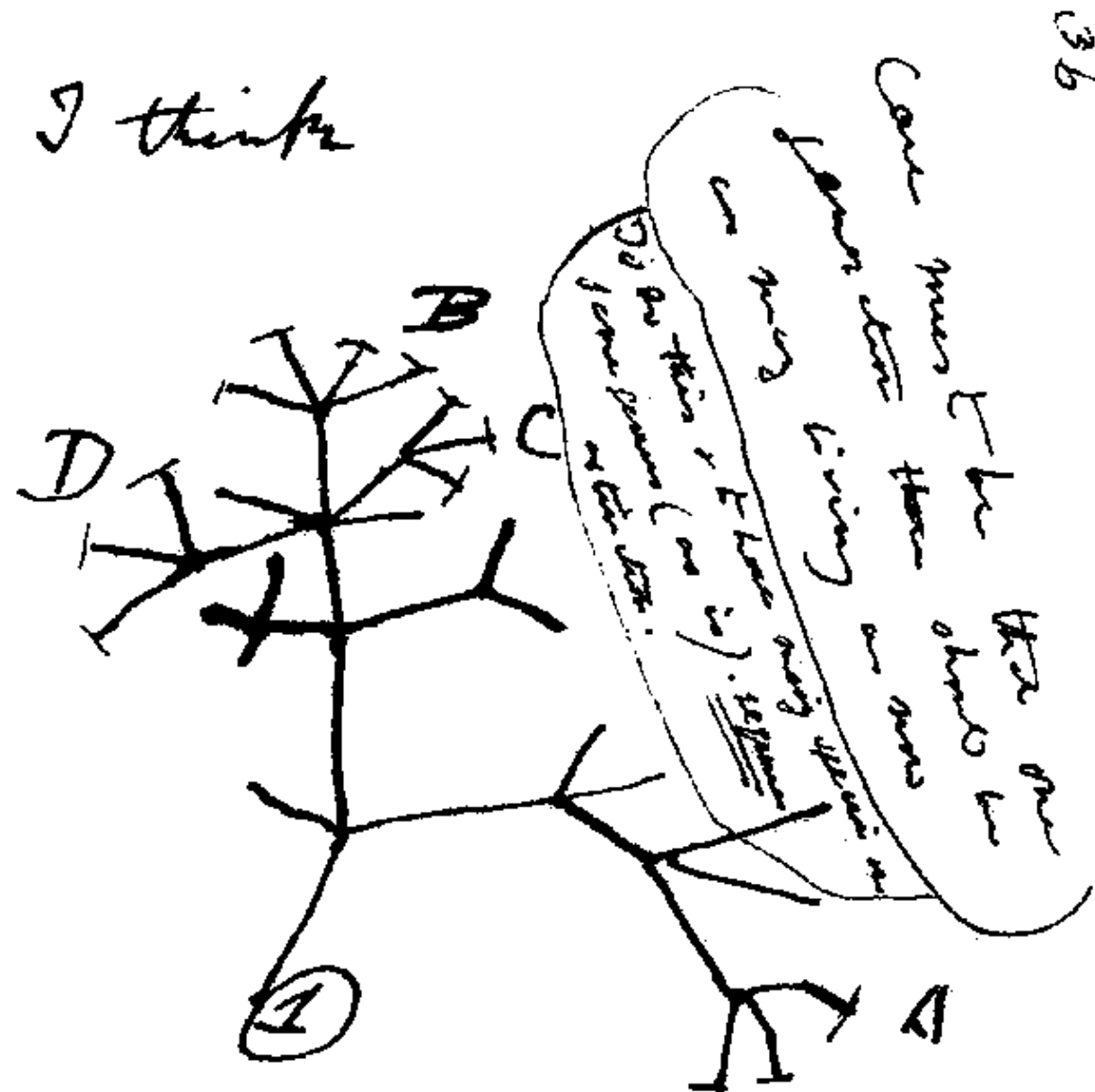
Introduction à la phylogénétique

Gwen Kettenburg

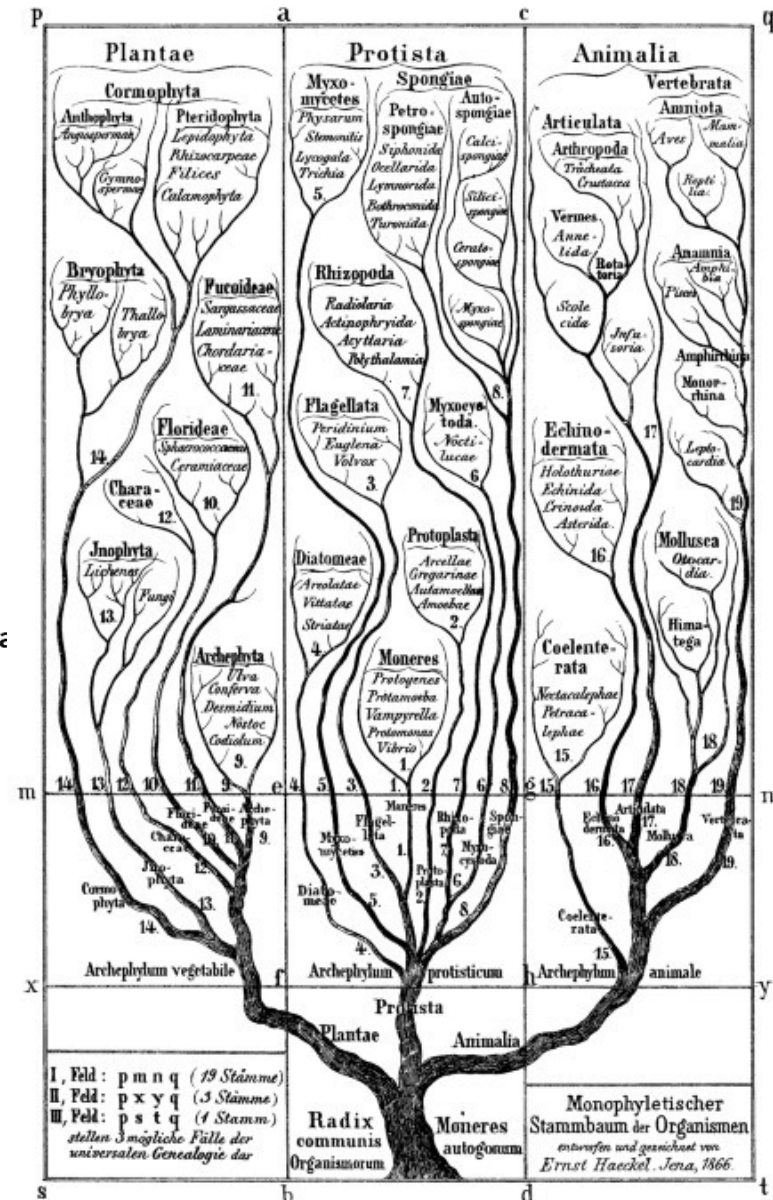
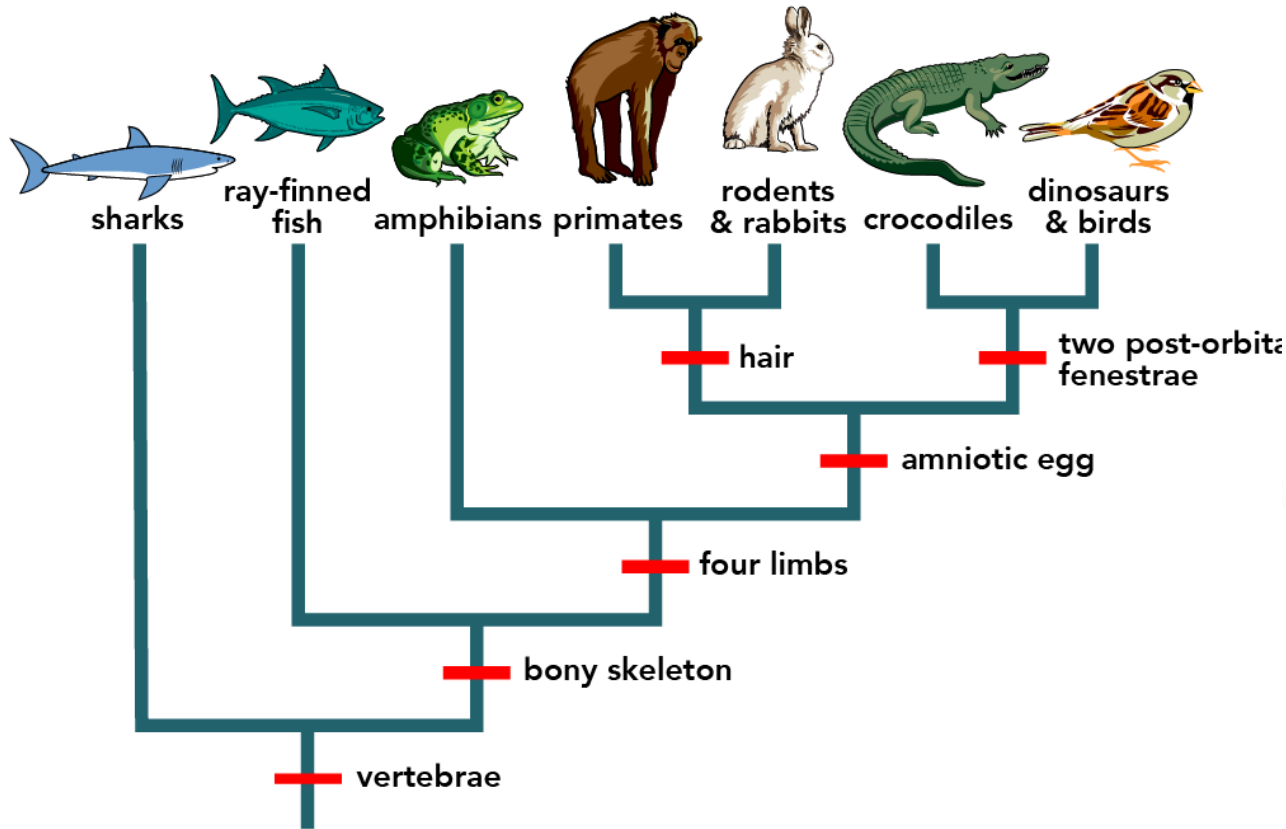
Diapositives adaptées de Richard Ree et Andrew Hipp, Université de Chicago



Qu'est-ce qu'une phylogénie ?

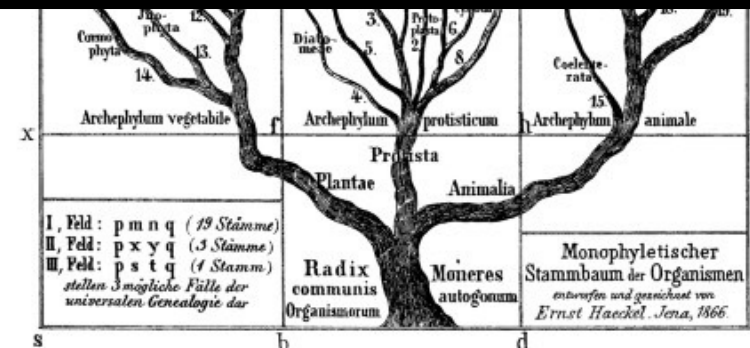
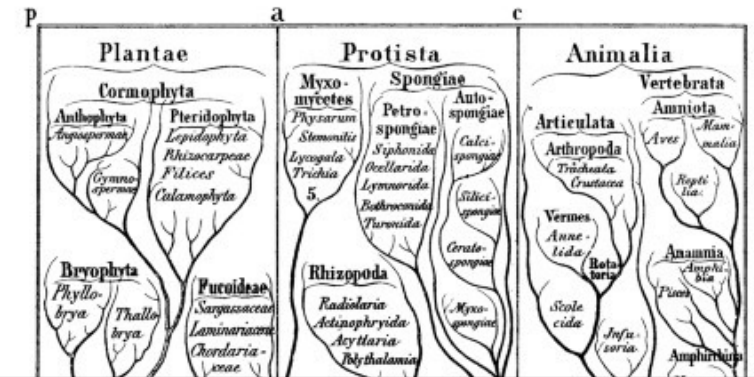


Qu'est-ce qu'une phylogénie ?



Qu'est-ce qu'une phylogénie ?

“Un arbre phylogénétique, ou phylogénie, est un diagramme qui représente les lignes de descendance évolutive de différentes espèces, organismes ou gènes à partir d'un ancêtre commun.”

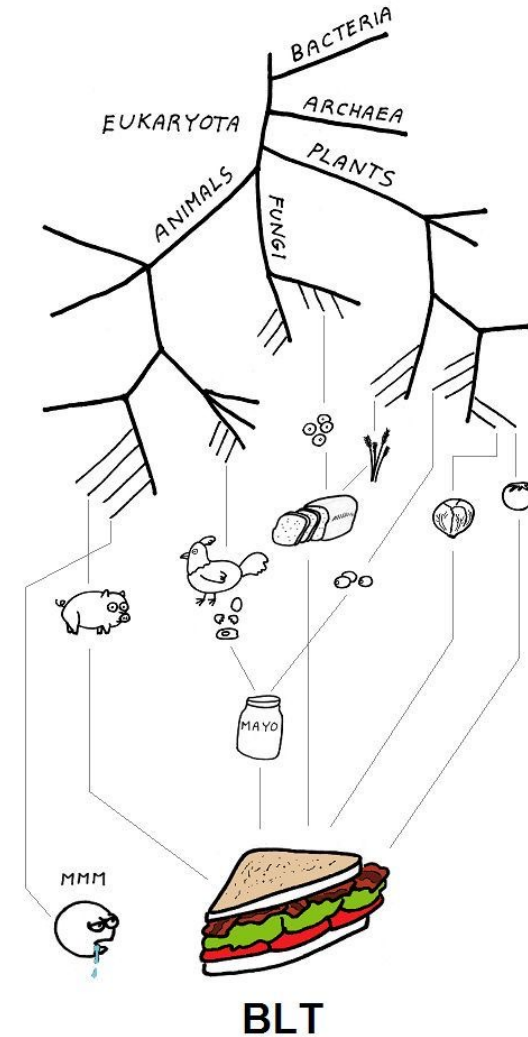
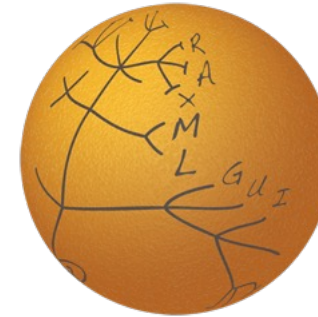


Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

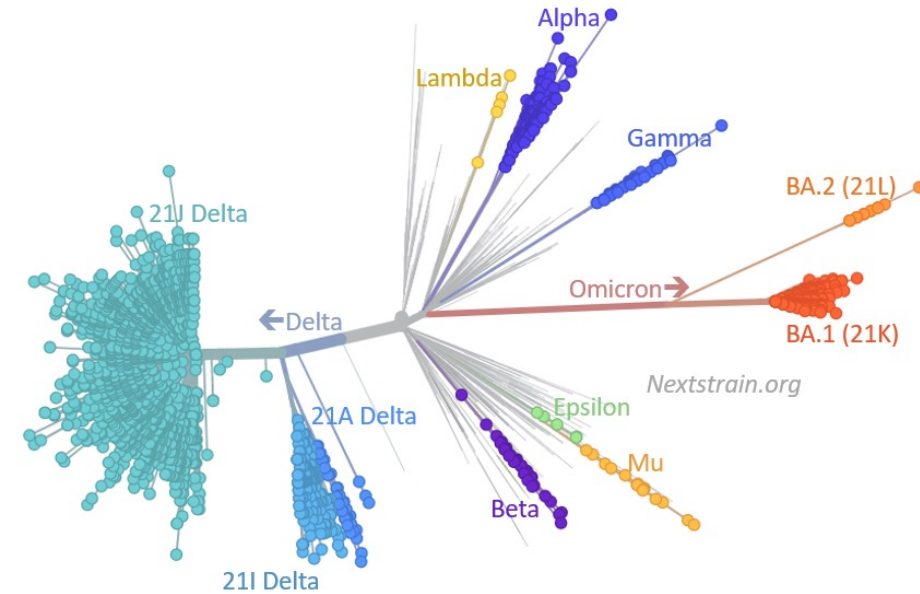
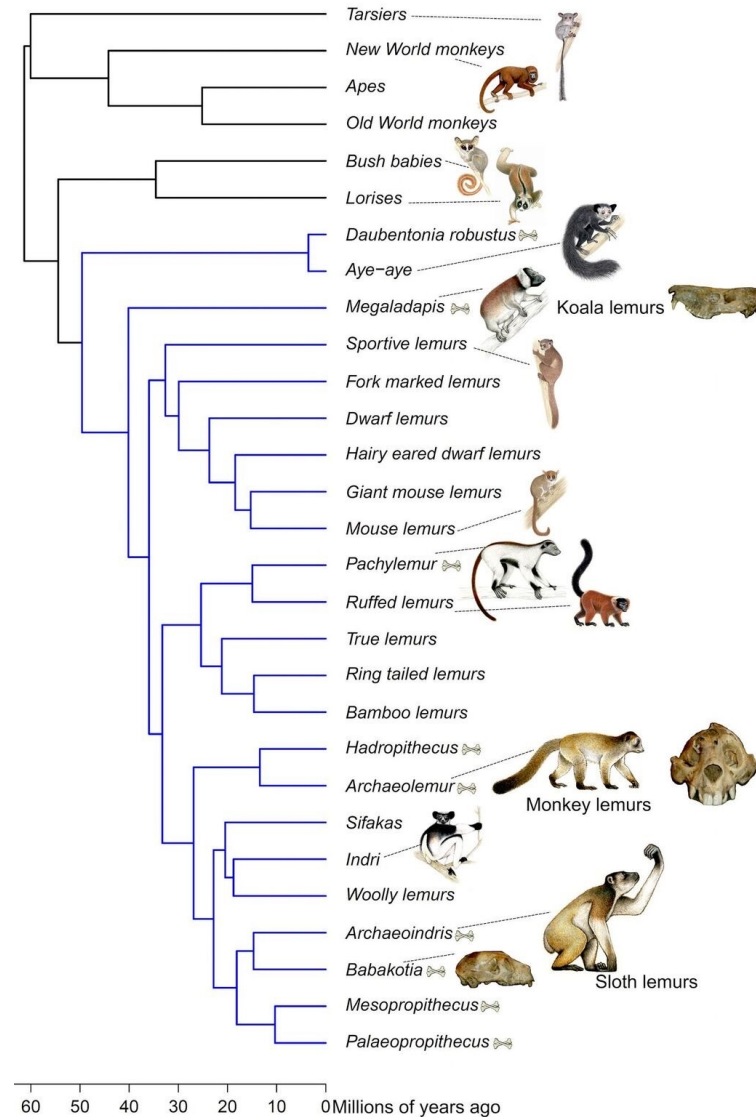
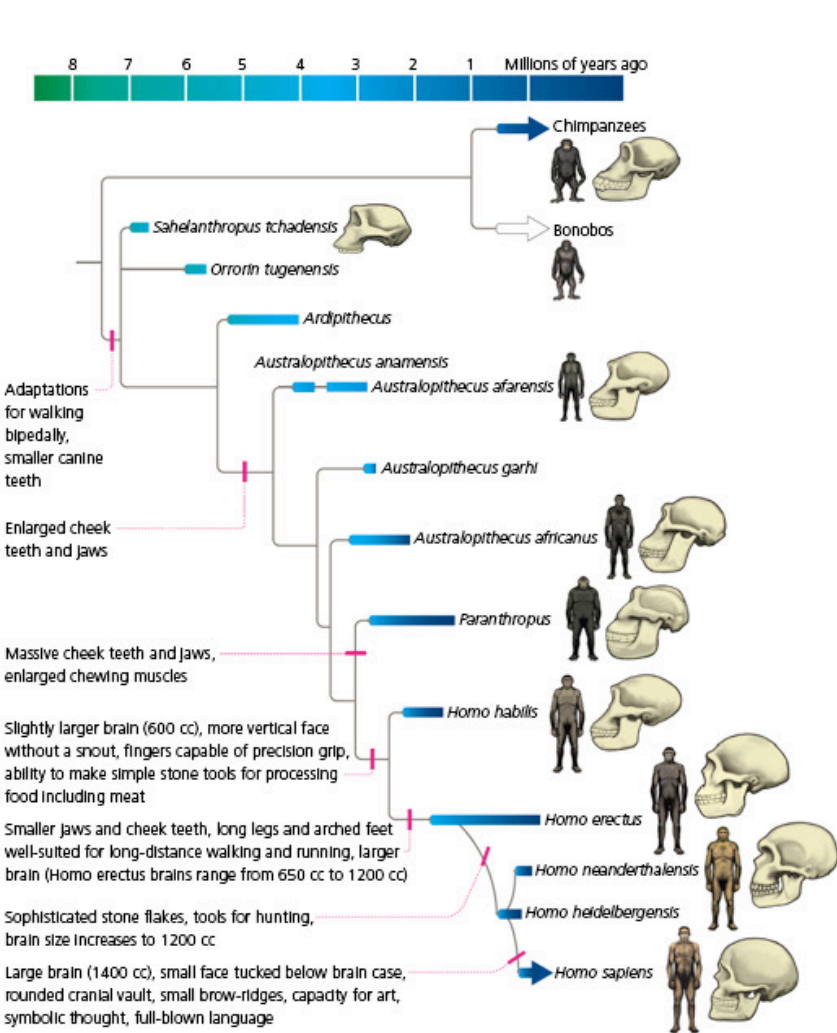
Objectifs:

- Volet cours magistral
 - Apprenez les bases de ce qu'est une phylogénie
 - Apprendre à lire les phylogénies
 - Bases de la modélisation phylogénétique
- Composant didacticiel
 - Apprenez à créer un arbre phylogénétique à partir de données de séquençage
 - Utilisation de séquences de protéines du cytochrome B du lémurien dans le logiciel RAxML
 - Modifier et visualiser l'arborescence dans R et FigTree

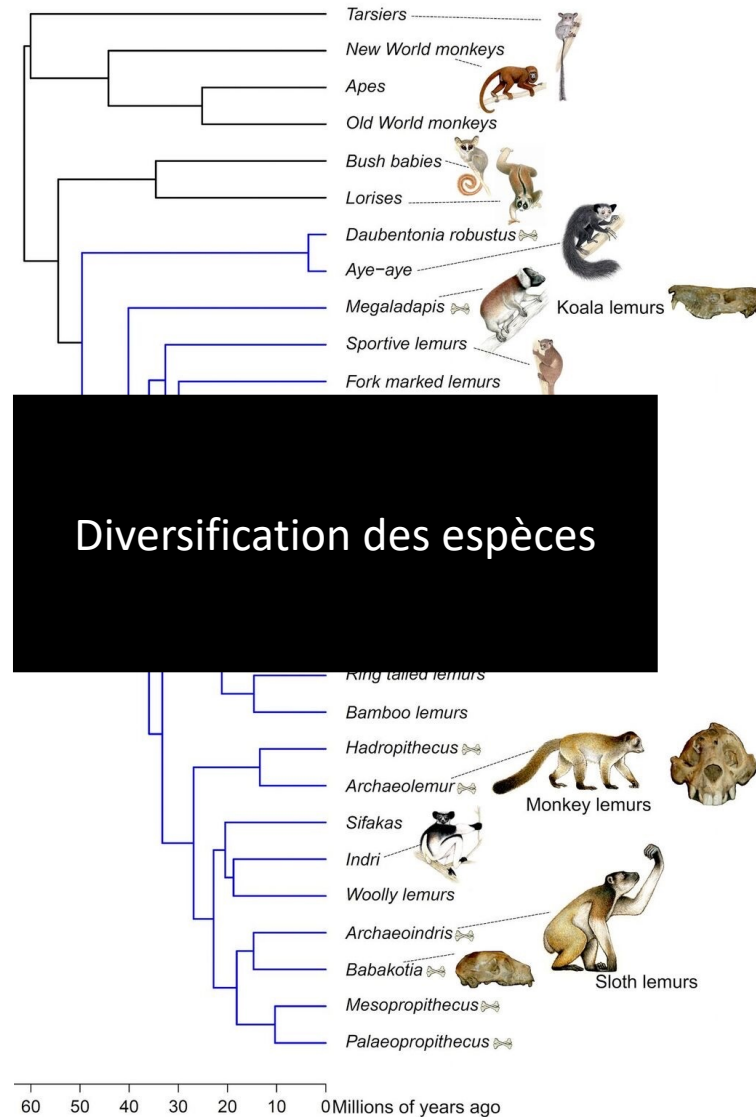
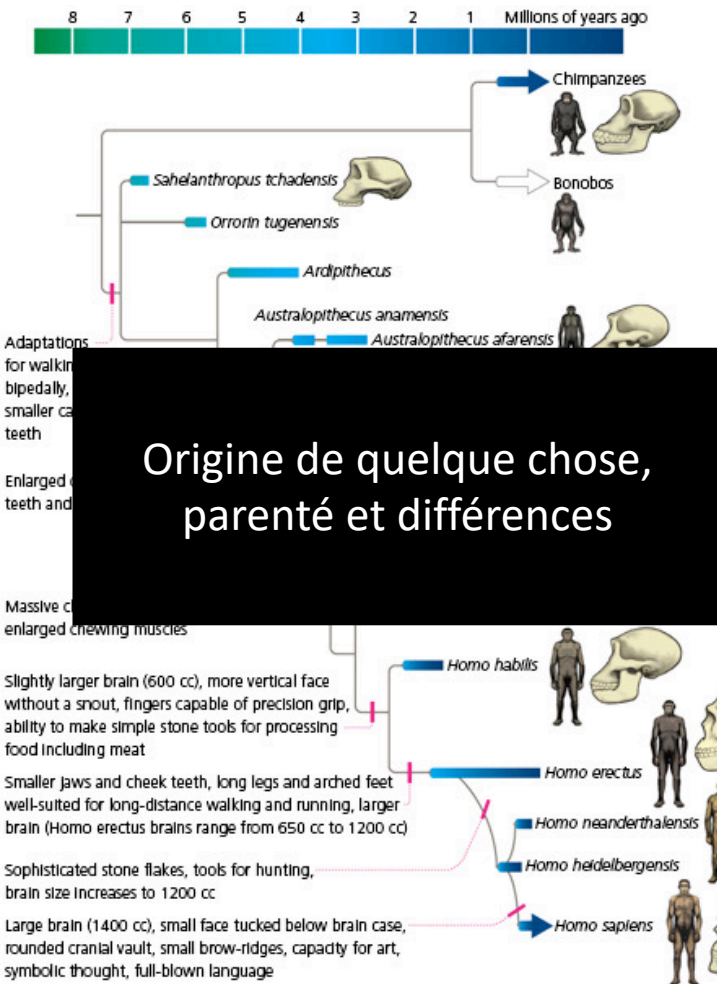


Molecular Evolutionary
Genetics Analysis

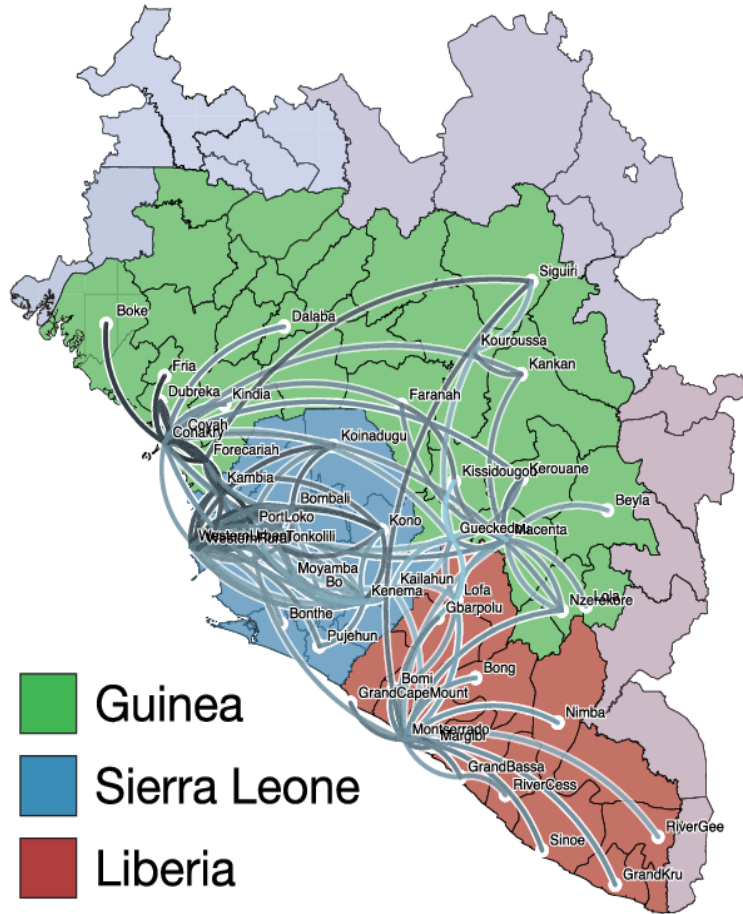
Que pouvez-vous faire avec les phylogénies ?



Que pouvez-vous faire avec les phylogénies ?

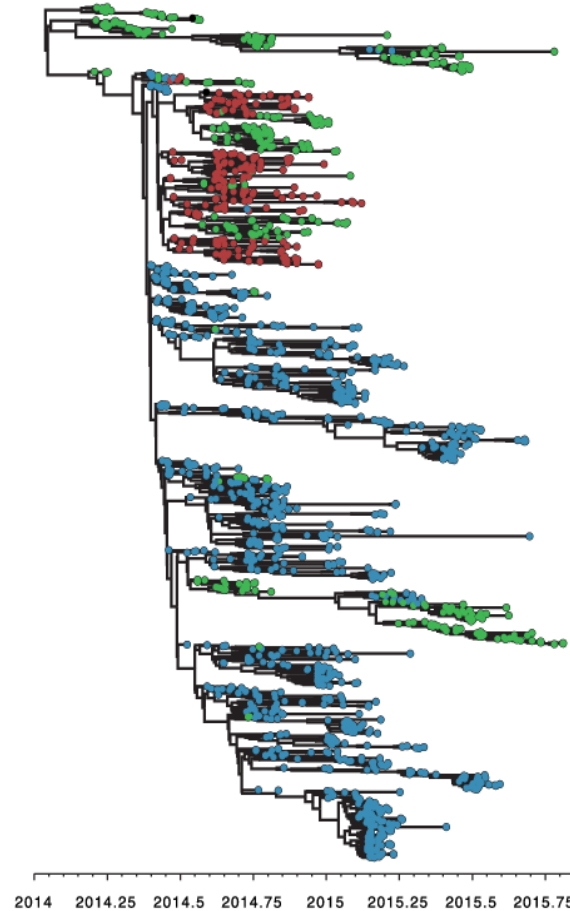


Phylodynamique



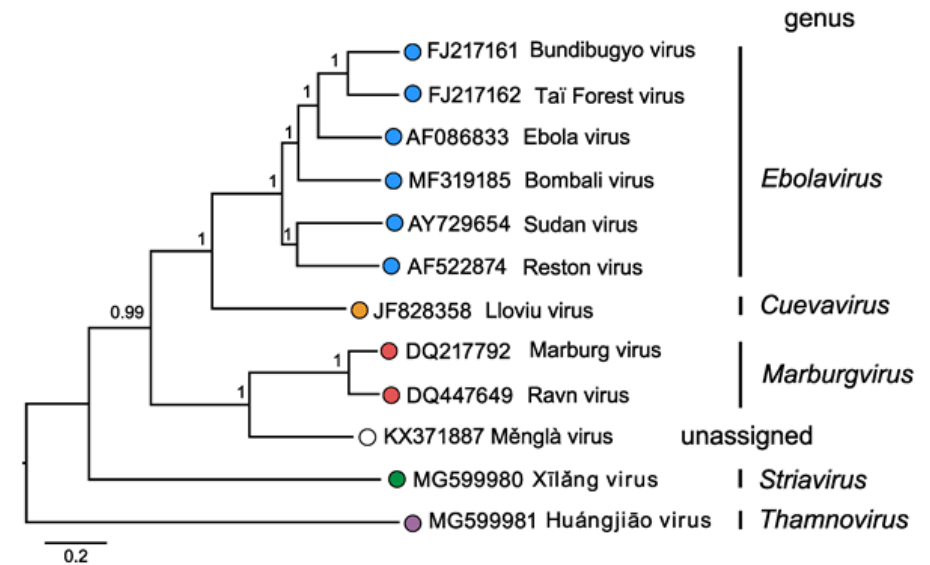
OÙ va-t-il ?

Arbres bayésiens



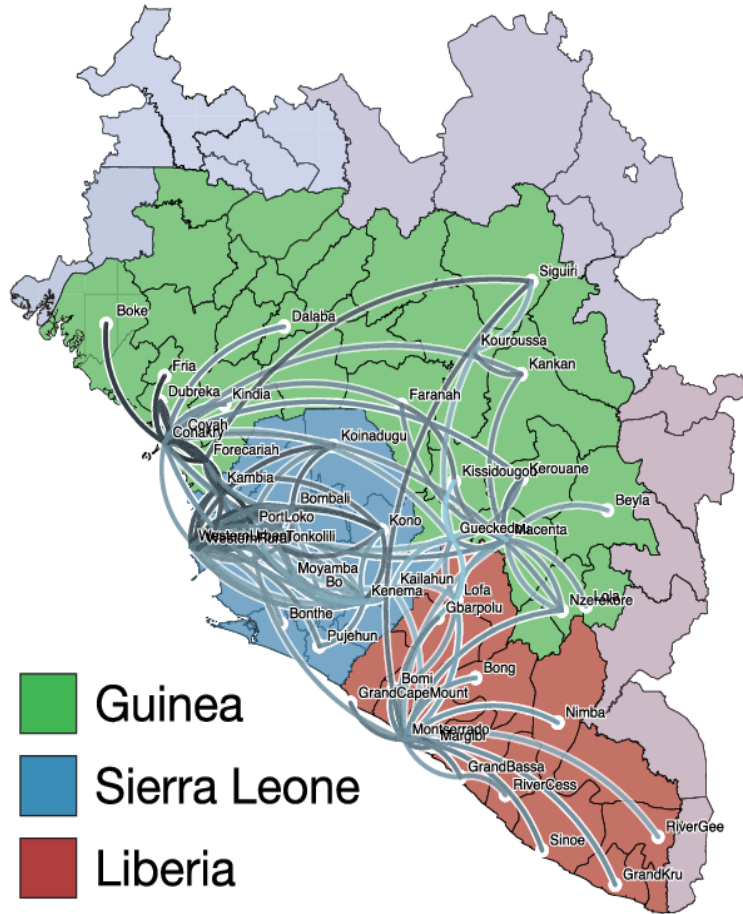
QUAND est l'ancêtre commun le plus récent ?

Maximum de vraisemblance



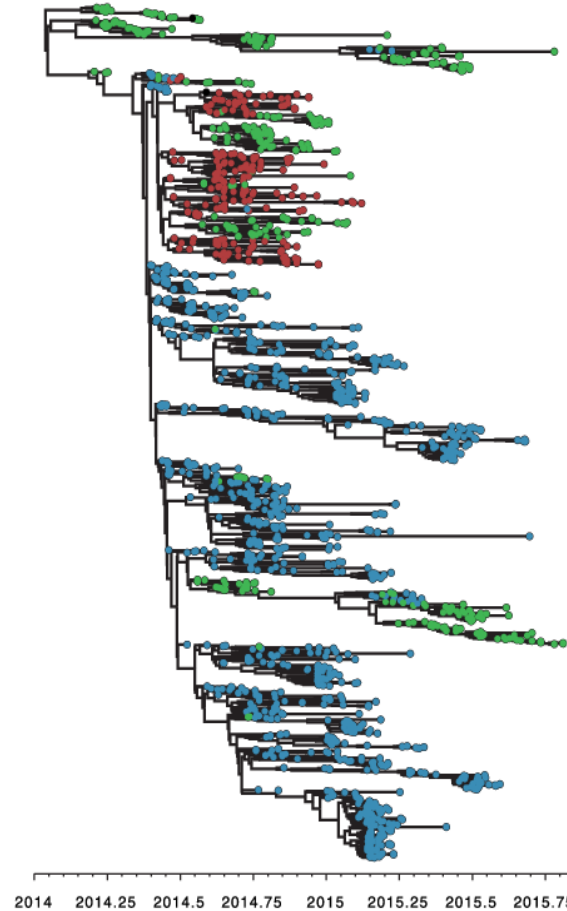
À quel point est-ce différent de ce que l'on connaît ?

Phylodynamique



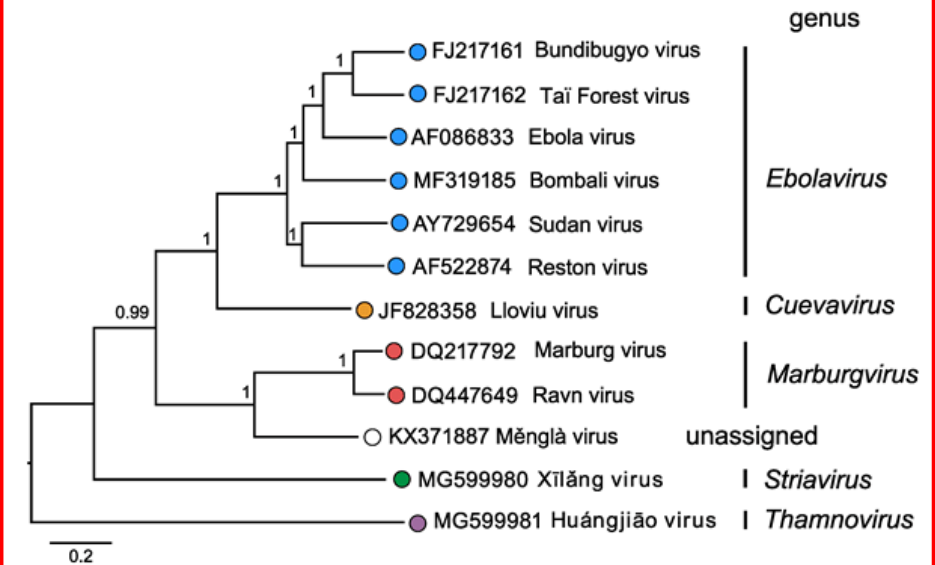
OÙ va-t-il ?

Arbres bayésiens



QUAND est l'ancêtre commun le plus récent ?

Maximum de vraisemblance

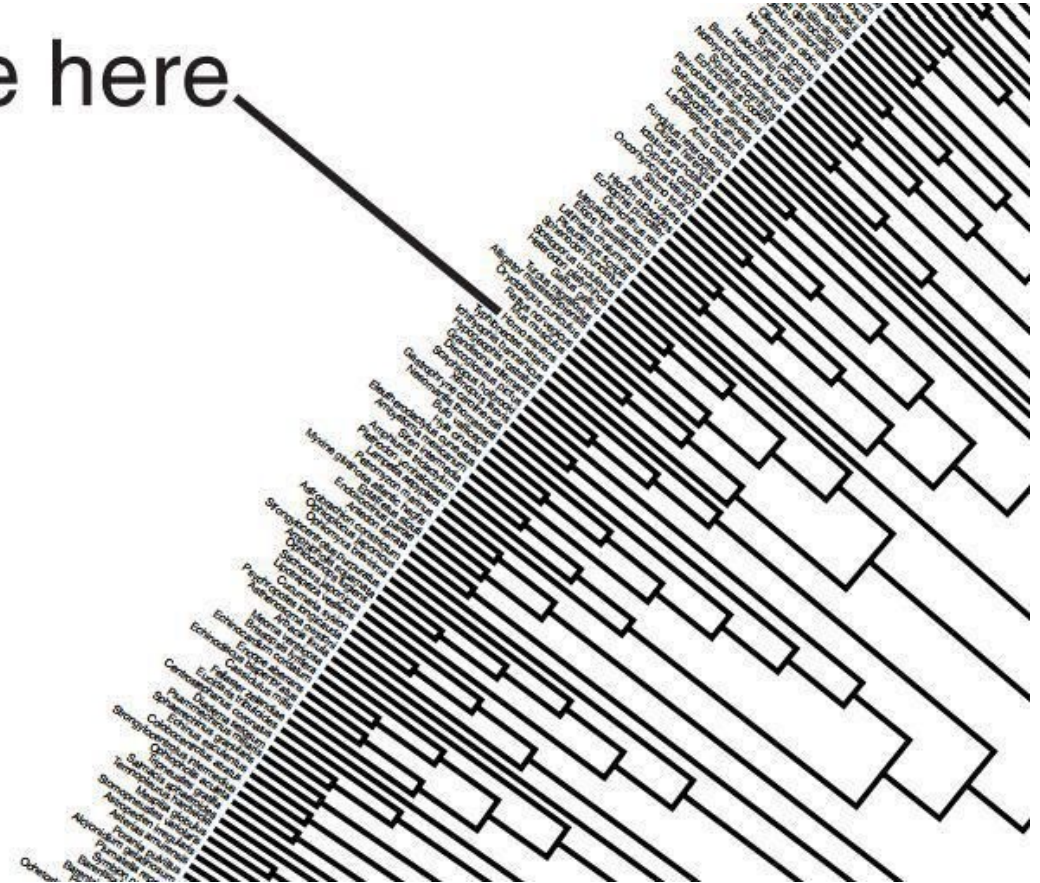


À quel point est-ce différent de ce que l'on connaît ?

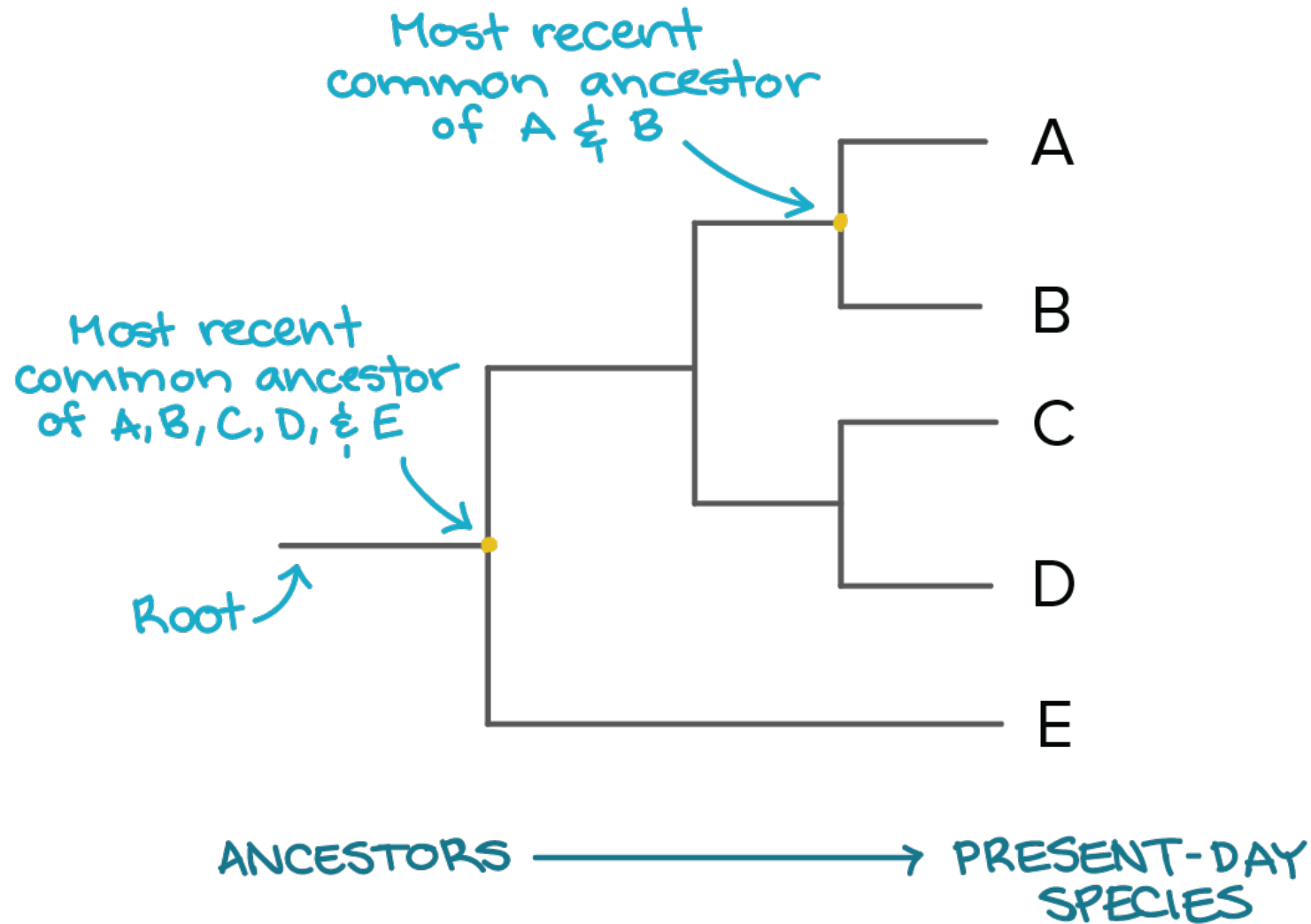
Point de contrôle!

- Qu'est-ce qu'une phylogénie ?
- Peut-on utiliser une analyse phylogénétique avec des données temporelles ?
- Pouvez-vous utiliser une analyse phylogénétique pour voir à quel point quelque chose est similaire à un autre ?

You are here



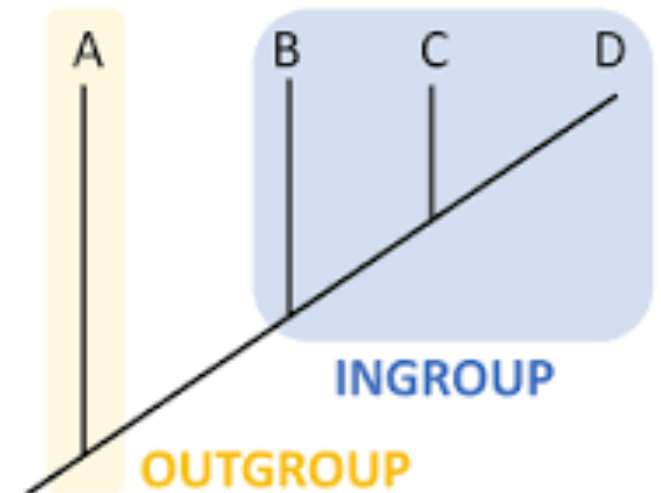
Anatomie d'une phylogénie



CONFIDENCE

BOOTSTRAP VALUE

STRONGLY SUPPORTED	>90%
WELL SUPPORTED	70%-90%
WEAKLY SUPPORTED	50%-70%
NOT SUPPORTED	<50%



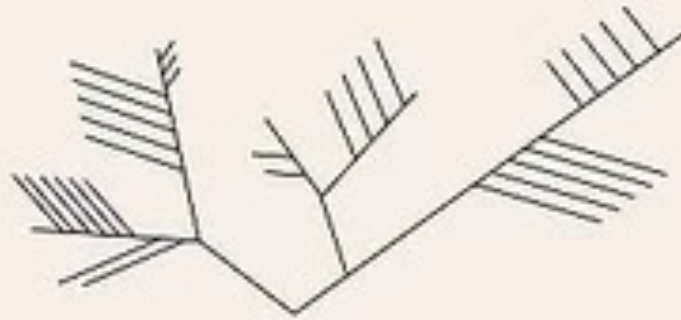
Cladogramme versus arbre phylogénétique

CLADOGRAM



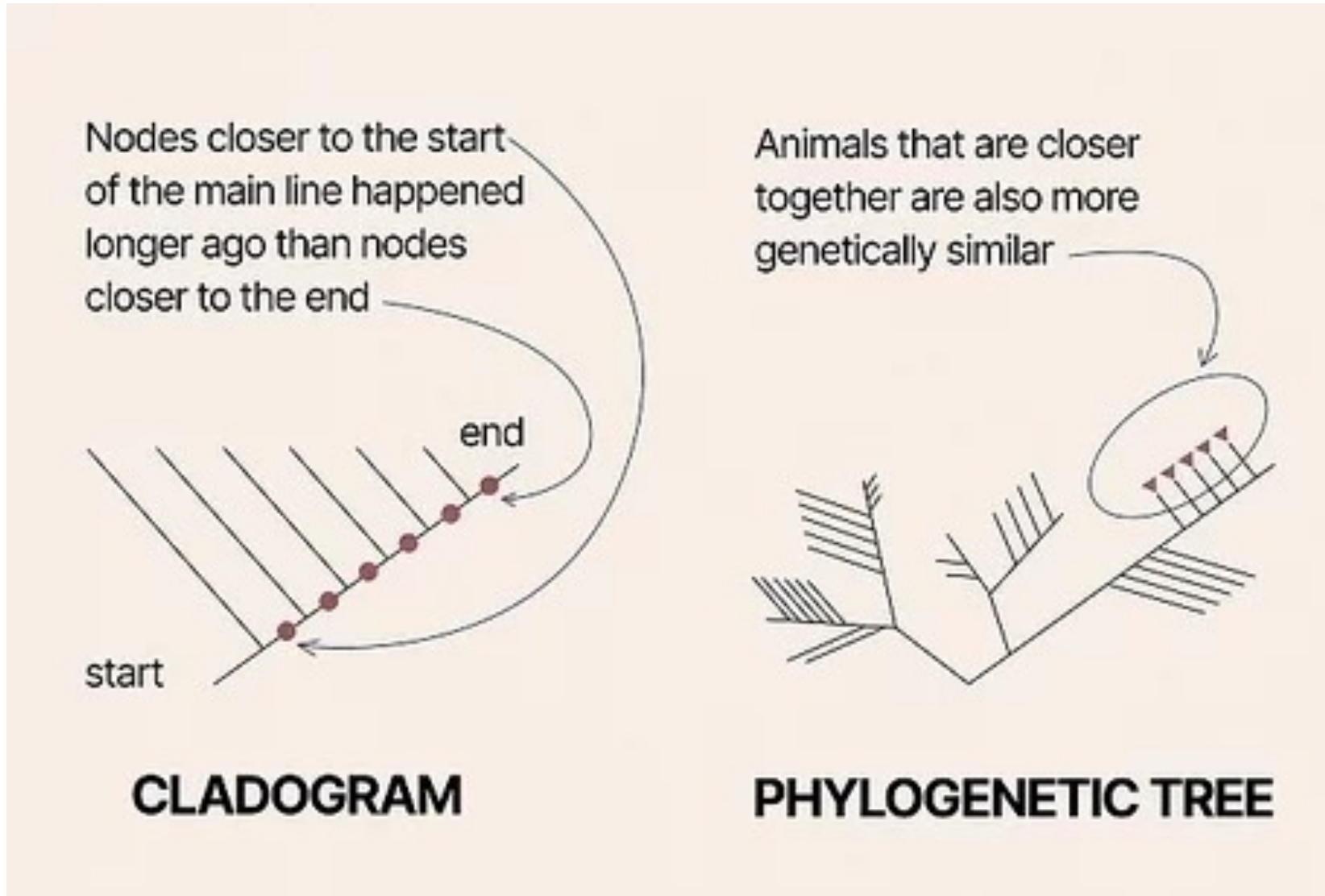
- the relationships are *hypothetical*
- you can easily make on your own

PHYLOGENETIC TREE



- the relationships are *backed by molecular evidence*
- should have access to DNA or other molecular data

Cladogram versus phylogenetic tree

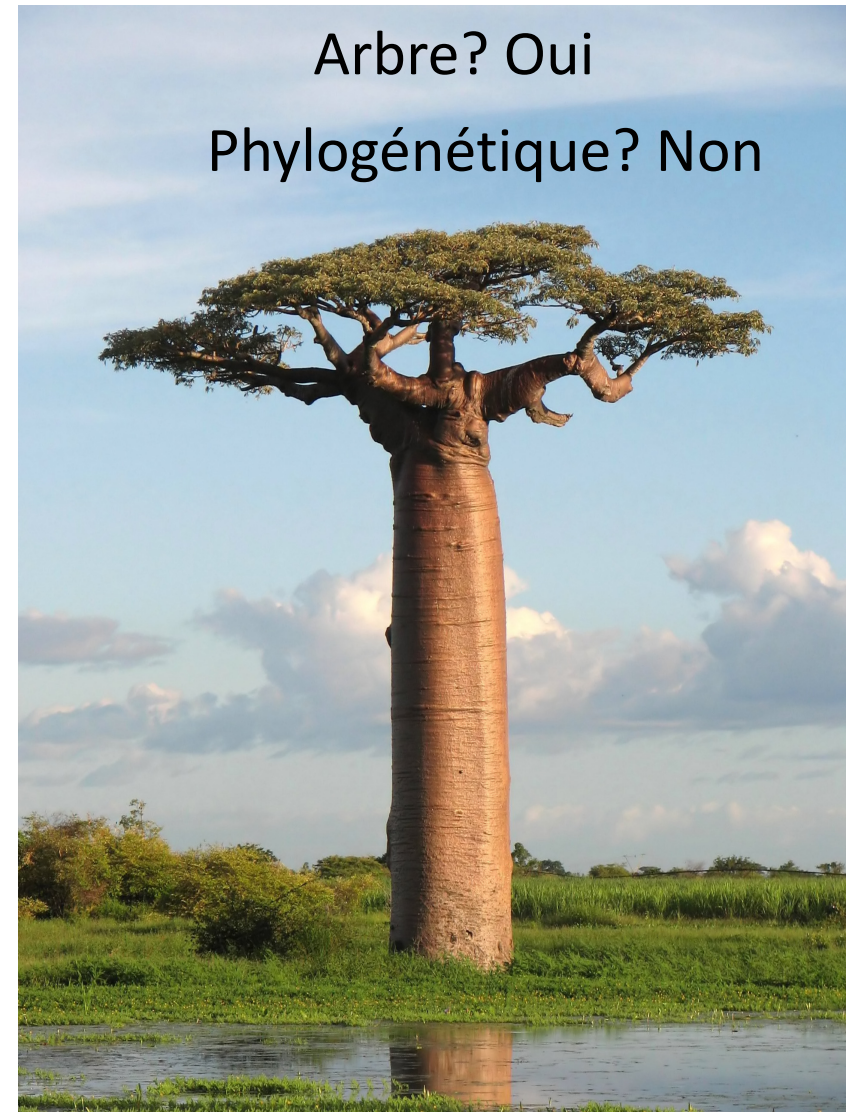


Utiliseriez-vous un cladogramme ou un arbre phylogénétique pour :

- **Hypothèses sur les ancêtres de l'homme**
- **Traquer un nouvel agent pathogène**

Point de contrôle!

- Les cladogrammes sont bons pour générer des hypothèses, les phylogénies montrent des similitudes génétiques
- Que montrent les valeurs d'amorçage ?
- Qu'est-ce qu'une racine ?



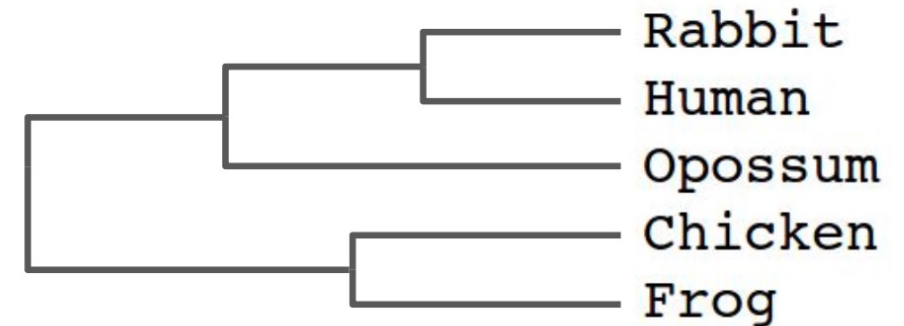
Parcimonie vs. Vraisemblance

- Parcimonie : nombre minimum de modifications
- Vraisemblance : probabilité maximale que les données aient évolué sur l'arbre



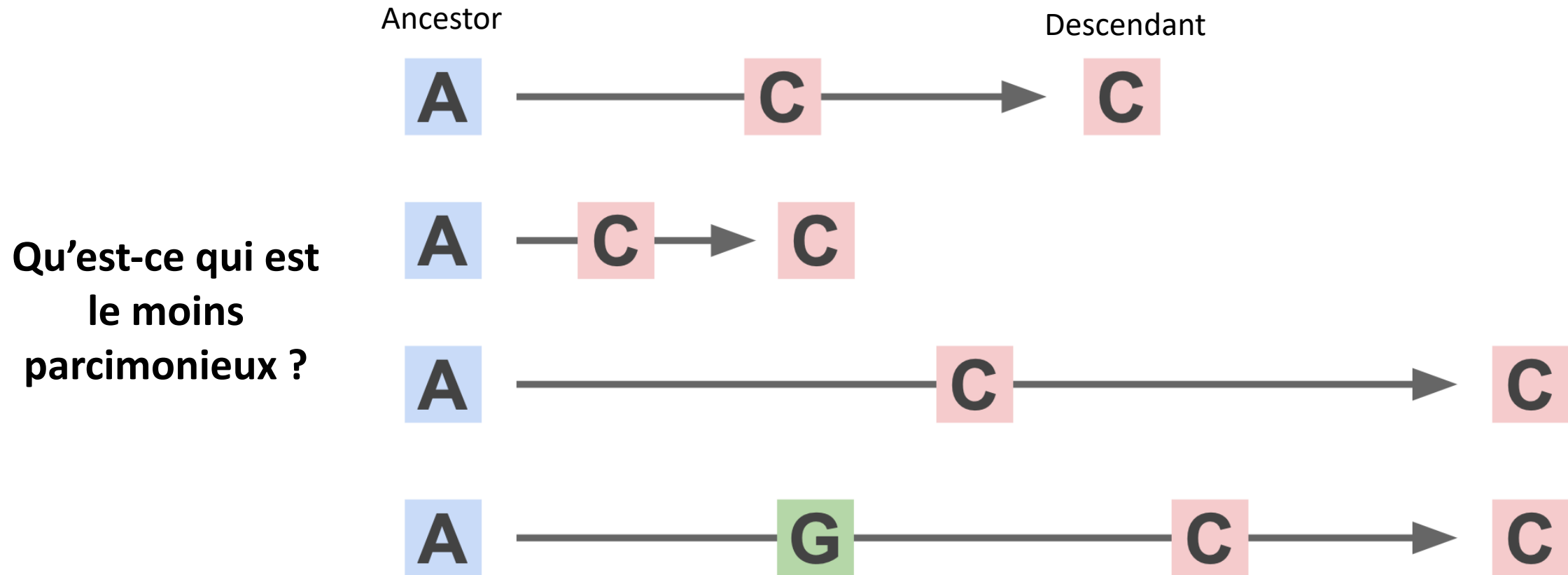
branch length can mean different things:

- minimum number of changes (parsimony)
- time; opportunity for change
- expected number of changes, given a model of evolution

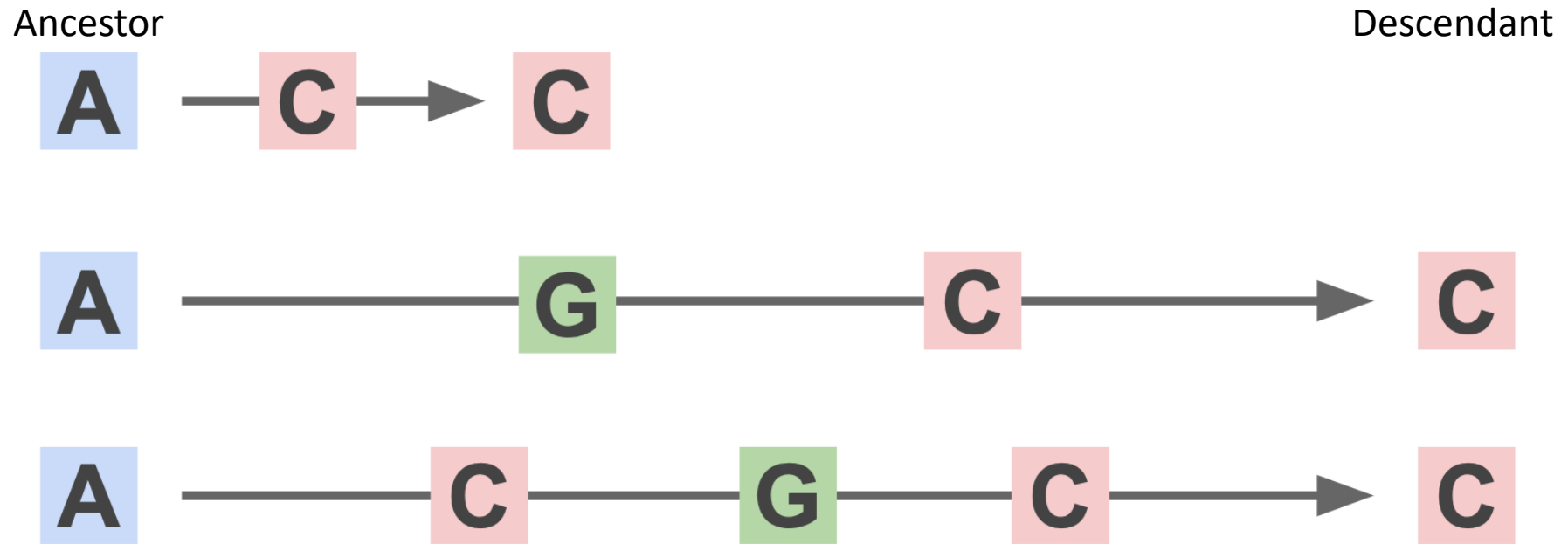


proposed tree has **branch lengths** in units of expected number of changes per site

Parcimonie : nombre minimum de changements,
quel que soit le moment ou l'opportunité.



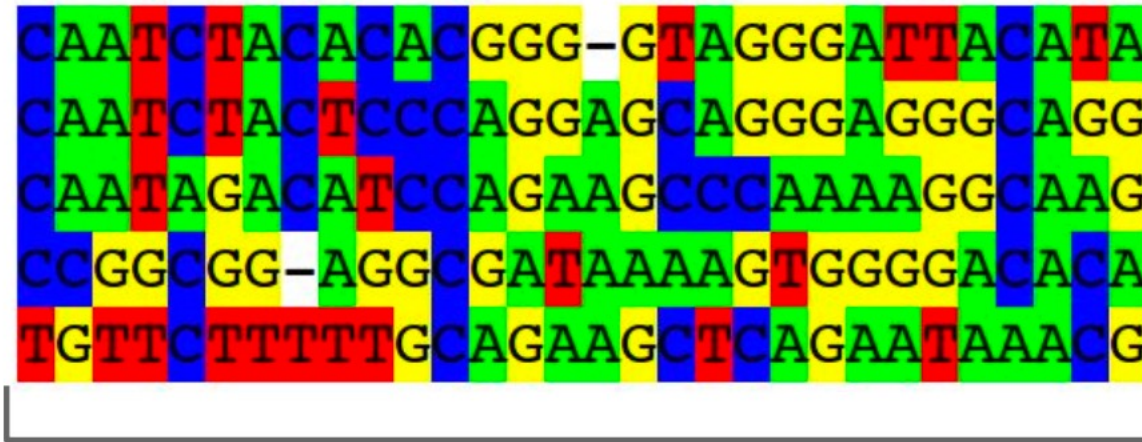
Vraisemblance : la probabilité d'un statut d'ancêtre et de descendant est fonction du temps (longueur de la branche)



Nous ne savons pas quelle est l'histoire réelle du changement, alors utilisez un modèle d'évolution pour considérer toutes les histoires possibles (probabilité maximale)

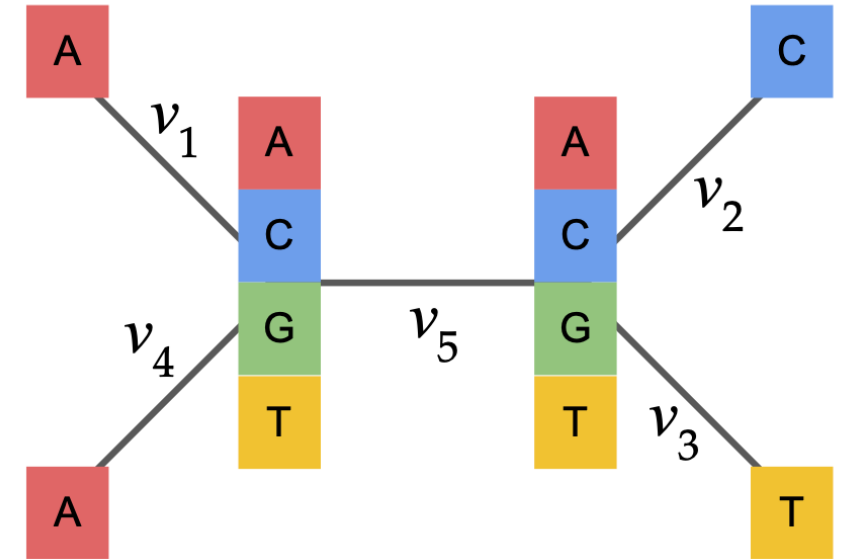
Vraisemblance (suite).

Rabbit
Human
Opossum
Chicken
Frog



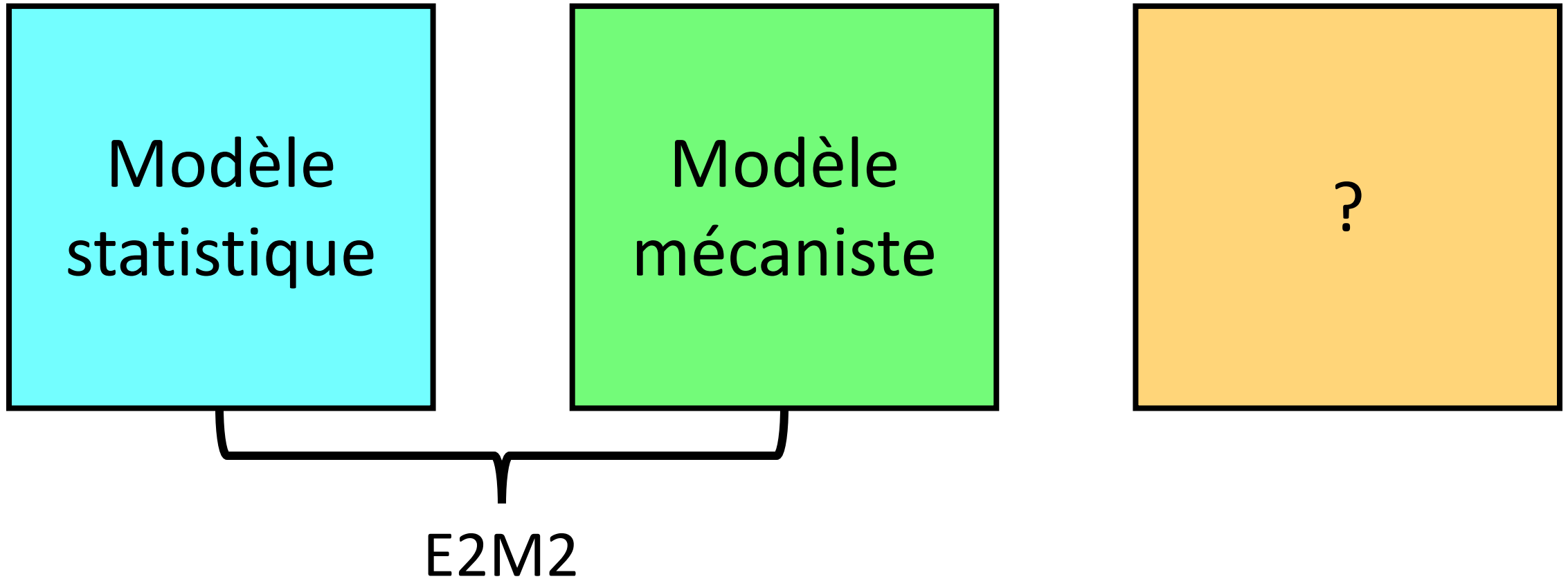
overall likelihood is the product of
likelihoods across characters (sites)

Paramètres : topologie arborescente,
longueurs des branches, taux de
substitution estimés pour maximiser la
vraisemblance des données



Consider *all possible ancestral states* at internal nodes, and calculate their contribution to the overall likelihood.*

Jusqu'à présent...



Aujourd'hui...

Modèle
statistique

Modèle
mécaniste

Modèle
évolutif



E3M2

Modèles de l'évolution de l'ADN

- Modèles de Markov qui décrivent les taux relatifs de différents changements
 - JC69 (Jukes et Cantor 1969)
 - Modèle K80 (Kimura 1980)
 - Modèle K81 (Kimura 1981)
 - F81 (Felsenstein 1981)
 - Modèle HKY85 (Hasegawa, Kishino et Yano 1985)
 - Modèle T92 (Tamura 1992)
 - Modèle TN93 (Tamura et Nei, 1993)
 - Modèle GTR (Tavaré 1986)
 - Oui, il y en a beaucoup ! Comment puis-je savoir ce qui est le mieux pour mes données ?

Bonne nouvelle, la plupart des gens n'ont pas besoin de connaître les spécificités mathématiques de ces modèles

JC69 model (Jukes and Cantor 1969) [\[edit \]](#)

JC69, the [Jukes](#) and [Cantor](#) 1969 model,^[2] is the simplest [substitution model](#). There are several assumptions. It assumes equal base frequencies

$\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4} \right)$ and equal [mutation rates](#). The only parameter of this model is therefore μ , the overall substitution rate. As previously

mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

When branch length, ν , is measured in the expected number of changes per site then:

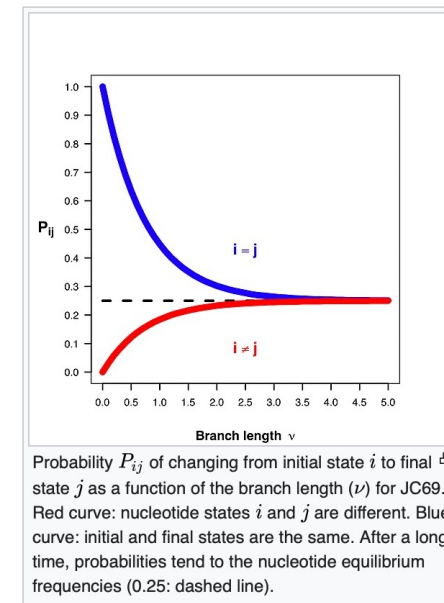
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

It is worth noticing that $\nu = \frac{3}{4}t\mu = \left(\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4} \right)t$ what stands for sum of any column (or row) of matrix

Q multiplied by time and thus means expected number of substitutions in time t (branch duration) for each particular site (per site) when the rate of substitution equals μ .

Given the proportion p of sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance (in terms of the expected number of changes) between two sequences is given by

$$\hat{d} = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right) = \hat{\nu}$$



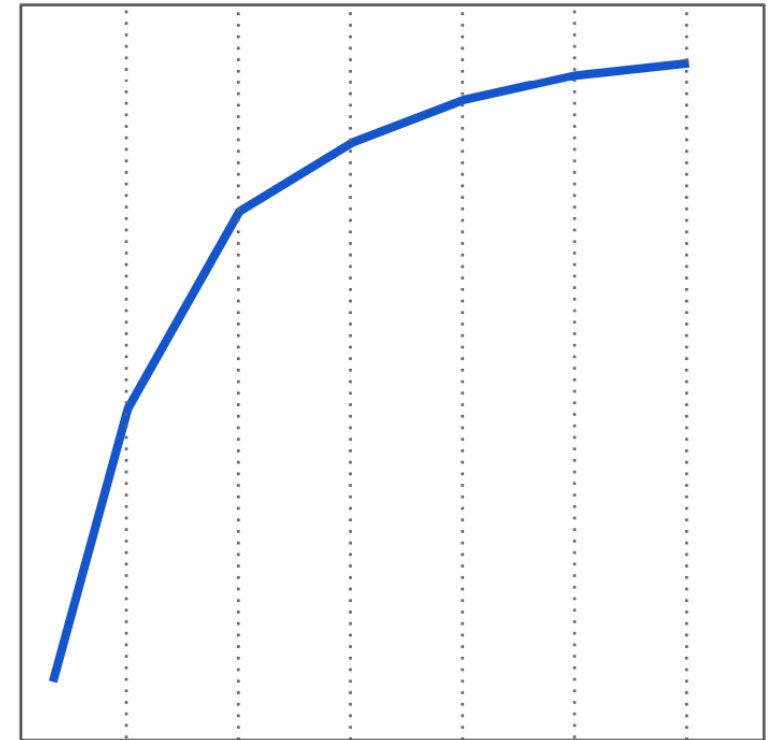
Sélection du modèle

Plus il y a de paramètres, plus la probabilité est élevée, mais l'augmentation de la probabilité est-elle nécessaire? Ajoute beaucoup plus de complexité

Les tests de modèle vous donneront deux scores

- Score AIC : tente de sélectionner le modèle qui décrit le mieux une réalité inconnue et de haute dimension.
- Score BIC : tente de trouver le modèle VRAI parmi l'ensemble des candidats

likelihood



no. parameters

DNA models

Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

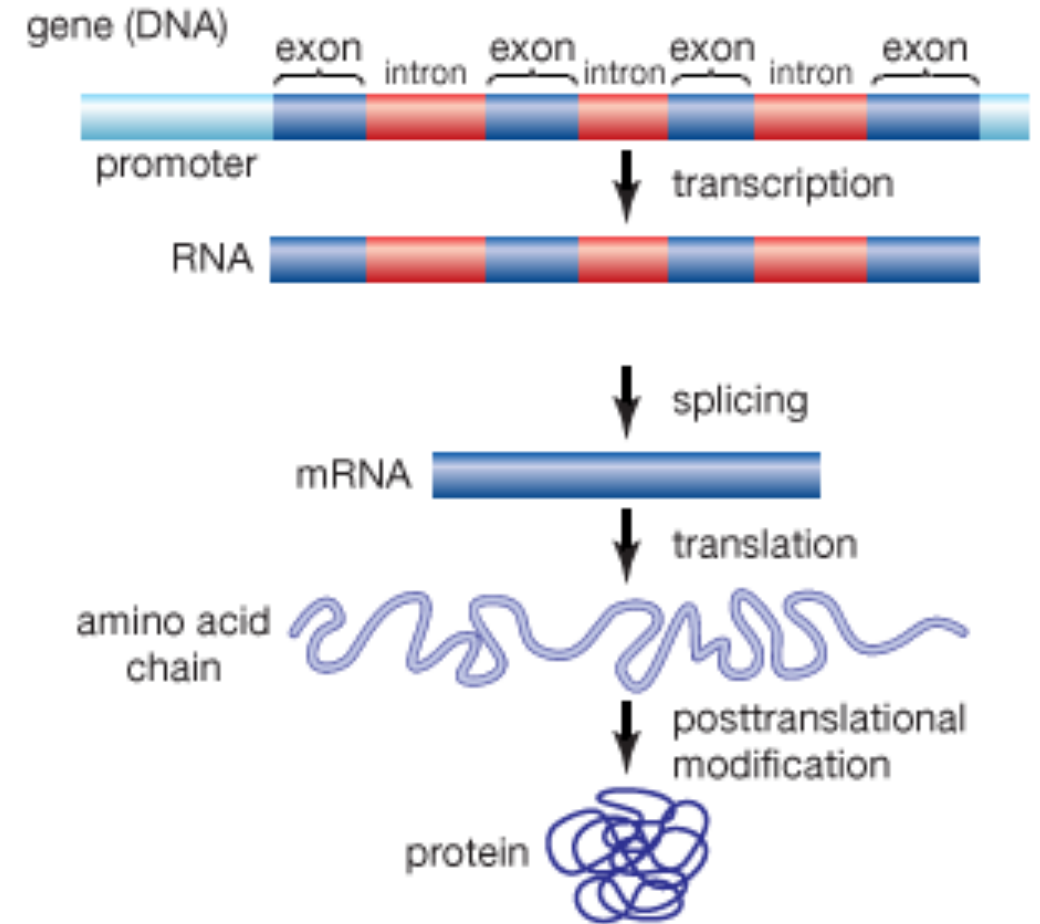
Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like HKY but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like TN but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like K81 but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like TPM2 but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like TPM3 but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIME	3	Like TIM but equal base freq.	012230

TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like TIM2 but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like TIM3 but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like TVM but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

Évaluer l'hétérogénéité entre les sites

- Évolution de l'hétérogénéité des taux :
- Positions des codons
- Exons (régions codantes) et introns (régions non codantes)
- Gènes d'entretien ménager et gènes non fonctionnels
- Structure de l'ARN (tiges et boucles)

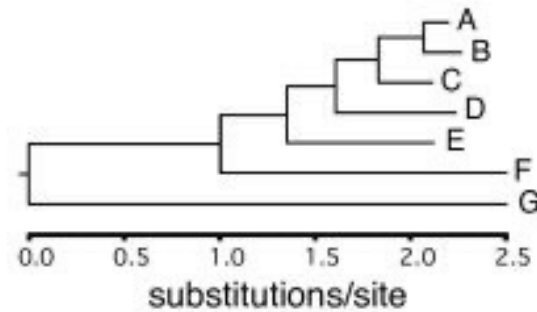
Nous pouvons faire des déductions sur la sélection à partir de ces valeurs, mais cela rend les choses beaucoup plus compliquées



Évaluer l'hétérogénéité entre les sites

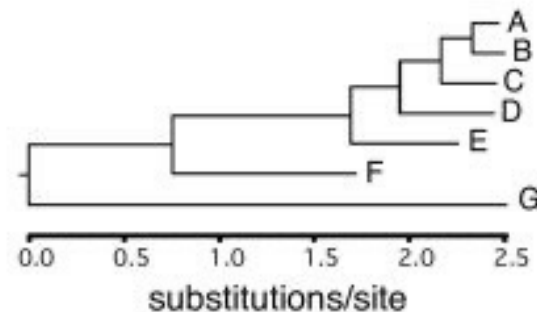
AVEC hétérogénéité de taux

```
A TTEQGIKSSSTSLPAPQLPNWSGQYHEWVLKS---FQNEVK---KTLHCALSQGTATQSVLDELHADVWALLASSEVCYAKPCGOVKPELAFLRKRA
B TACAGEKTGTSLPAPNLPNWSGQYHEWVLKS---ERADVI---KTMHCALSDGITATQSVLDELHFNVWALLASSEVCYARPCGDQKPELIYMKKQQA
C VAGGCEKAGTSLPAPYLPN--SGQYGEWVLKS---LSTHVI---KHMICEDLSDSDTTTQSVLDELHGERWALLOSSEVCYAKPCGOEKRVLEHECYKRA
D CAGQAEKTGTSLPALHLPNWA--QYGEWVLKS---FPSQPV---MPIQCVPLSDARTAAQSVLDELHVESDALLDSSEVCYAAPCGA--RHDLEKFCVYSKA
E DEGLTQKTGTSLPALALPNWSGQYFEWVLKS---YG---FGQGGGAATCKPLSGDKTSIHVSVDLDELHAVLAALLMSGVCYALPCGAYKKALEFKCYLKA
F GEGFIKKTGTSPAPVALPDFAEQYDEWPLKSTLAYGRVNF---AAVPGAYLSDFGTGSSHVSVDLDELHONHAALLLSSEVCFAAPCGESKGALVVVCYSHA
G NDGPHIKKGTSGPAALPNQPIQYDEWVLKS---CEAKSI---NGSNWKPLSGKYTGLDYLDELHVMKDALLHATEVCLAPPCGY--TADLKAALNGPA
```



SANS hétérogénéité de taux

```
A H----GENYFC--SQVAKYLAF-YSHNYL--EALLRHATLEIQH--KSNNAEHGTGLEGPESA--EDPR--VPAGNEKLLGKVVNNEFSAPGL----IKKP
B Q----GENYFC--GGVAKYLAW-VGHNYL--EALLRHATMEINR--KSKKEEQNGLDGPESA--EDPR--IPAGGEKLLGMHNNMFGAAGL----VKMP
C S----GENYFC--PQVAKYLAW-MSNNYL--HAFETQAKLEIER--KRNQVEHGCGLDGPNGD--EDPR--IKNSGOXLLGGY--KELKNPGL----FVKP
D Q----AHQYPC--SHIGKYFAW-VANAYH--HVLLRYAKLEVER--KRTADHSTDLVAPNGA--KFSV--LLPGPDALL--RMHAKFISTPLA----FIKT
E E----EDQYPC--KENYKYLAW-VGHGELRAHALSKHAKLATEK--KTEADHNTKLETAESP--LVEC--IPPLPDTRVAIVANTFFSAQHL----FIKT
F Q----GKKDLC--ENLN----TW-MQNRWL--QALHK-TITVVQHDGKSSMGDHCCKAIDSKAS--LSPC--VSSGGGYLQKSNQIDFFVSNVTV----YLKS
G NNDFSKPFLFCNYTGIL---ILQCAG-----YLDGETMIGRFQ--STQVGLYSTRFDFRYKCMGPTHKATNNTDTFGDRKAFKKRVSVKAFKQQTAPQ
```



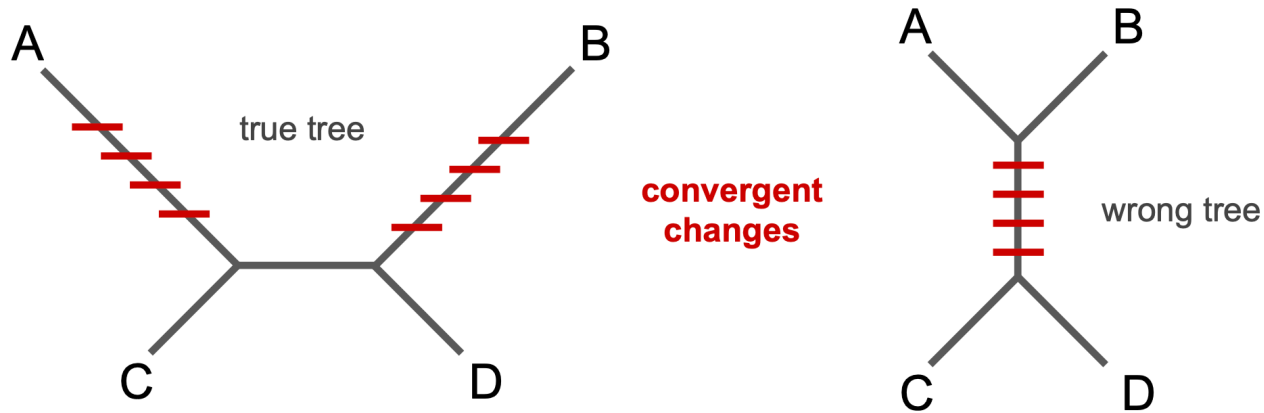
Rate heterogeneity across sites

IQ-TREE supports all common rate heterogeneity across sites models:

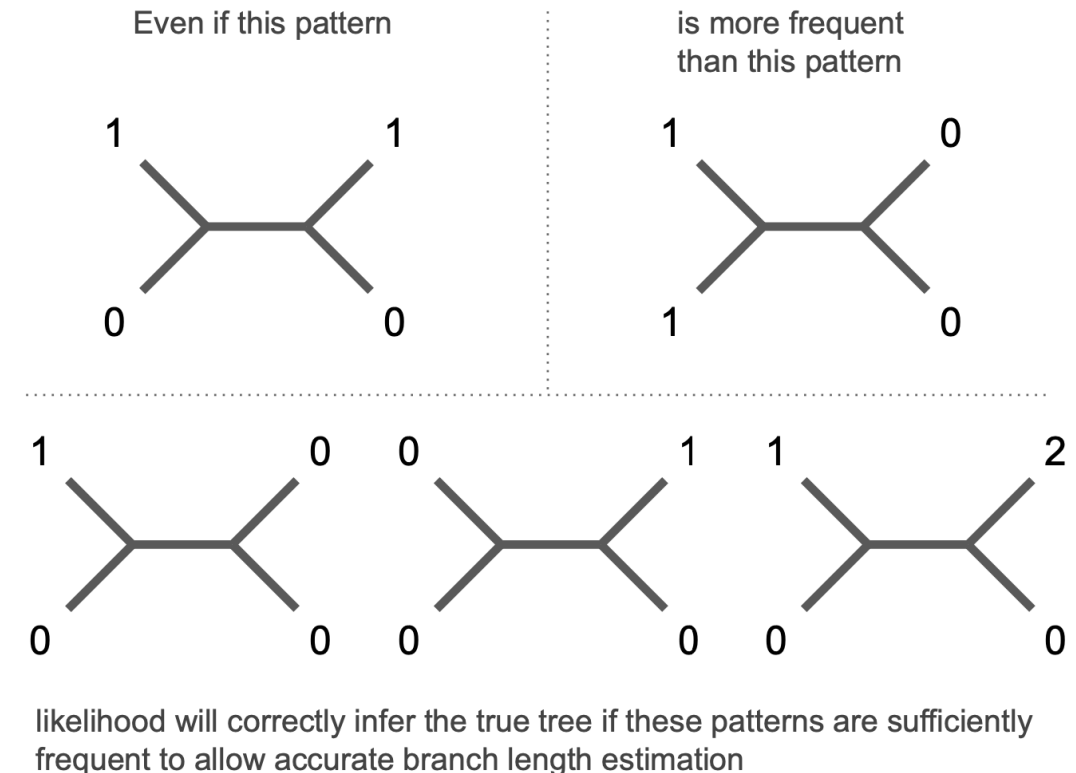
RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. +G8 .
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (Gu et al., 1995).
+R	FreeRate model (Yang, 1995; Soubrier et al., 2012) that generalizes the +G model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. +R6 (default 4 categories if not specified). The FreeRate model typically fits data better than the +G model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

Zone de Felsenstein

- Les longueurs de branche pour lesquelles la parcimonie déduit avec certitude une topologie incorrecte, celles-ci peuvent affecter les valeurs d'amorçage

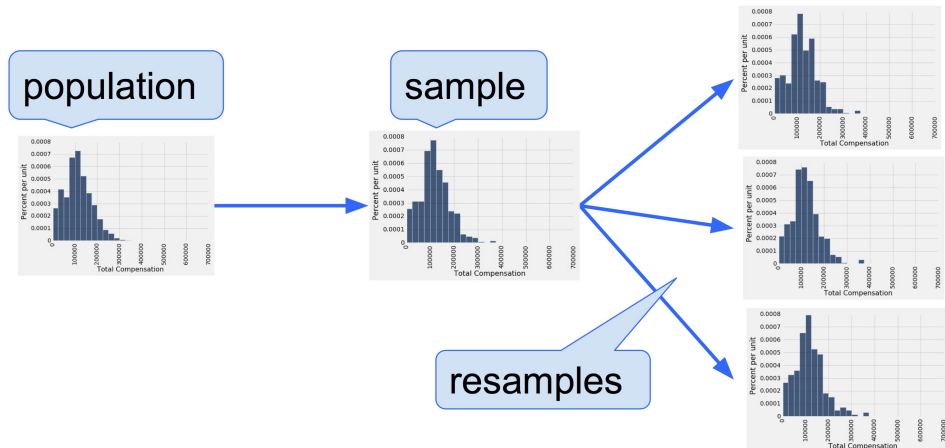


likelihood is a **consistent estimator** of tree topology because it converges on the correct value with increasing data



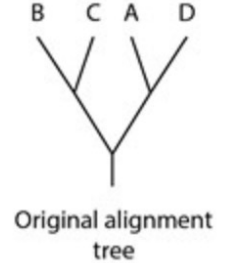
Bootstrapping

- Spécifier le nombre de répétitions : combien de fois le test réplique-t-il l'alignement de la séquence d'origine?



Original sequence alignment

	Site number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Species A	A	A	T	G	C	T	A	G	T	G	G	T	G	A	T
Species B	A	A	G	C	T	A	T	G	G	T	G	A	T	C	G
Species C	A	G	C	C	T	A	T	G	T	G	G	A	A	C	G
Species D	A	A	C	C	C	A	T	T	G	G	G	T	G	A	T



Bootstrap pseudo-replicate #1

	5	3	3	1	12	9	2	4	11	13	10	14	8	11	13
Species A	C	T	T	A	T	T	A	G	G	G	G	A	G	G	G
Species B	T	G	G	A	A	G	A	C	G	T	T	C	G	G	T
Species C	T	C	C	A	A	T	G	C	G	A	G	C	G	G	A
Species D	C	C	C	A	T	T	A	C	G	G	G	A	T	G	G



Bootstrap pseudo-replicate #2

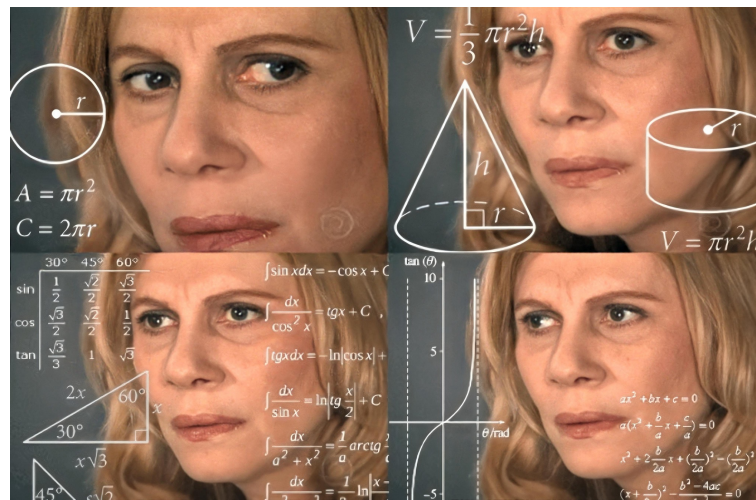
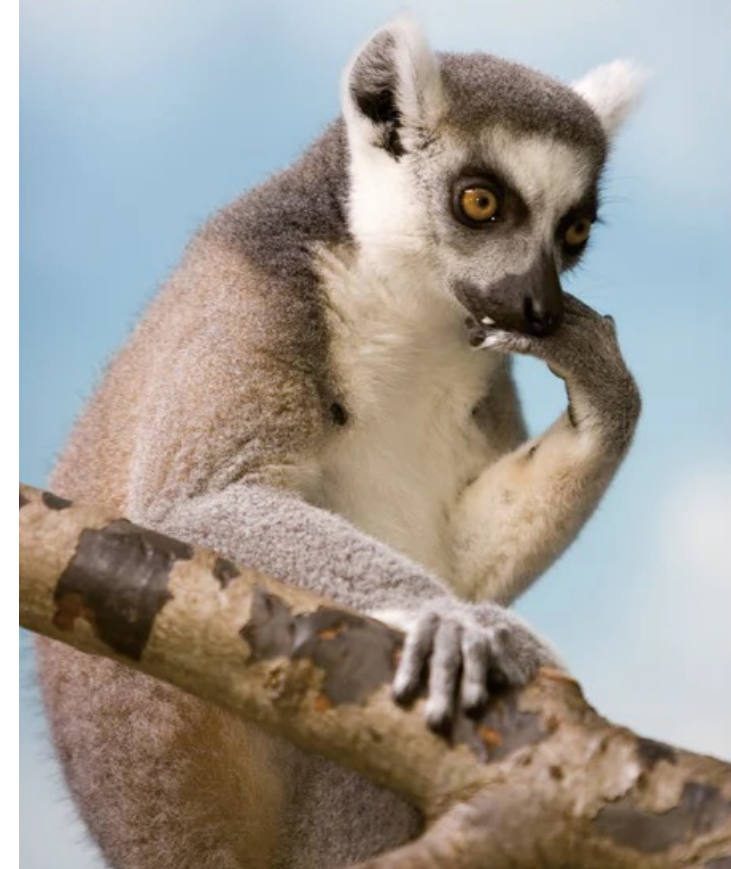
	9	7	12	5	2	4	2	6	14	9	4	9	7	2	1
Species A	T	A	T	C	A	G	A	T	A	T	G	T	A	A	A
Species B	G	T	A	T	A	C	A	A	C	G	C	G	T	A	A
Species C	T	T	A	T	G	C	G	A	C	T	C	T	T	G	A
Species D	T	T	T	C	A	C	A	A	A	T	C	T	T	A	A



Point de contrôle!

- Plus de paramètres = plus ou moins de probabilité?
- Des valeurs d'amorçage faibles signifient une faible confiance dans la topologie de l'arborescence
- Toutes les séquences ne subissent pas l'évolution de la même manière, les modèles rendent compte de ce changement

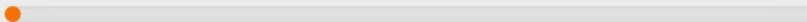
Shh I'm thinking!




Avertissements et limitations

- La construction de phylogénies prend beaucoup de temps, les plus grandes peuvent prendre jusqu'à une semaine pour s'exécuter et les phylogénies bayésiennes peuvent fonctionner pendant des mois, donc un cluster de calcul est presque nécessaire pour cela
- En l'absence d'un exogroupe ou d'une racine appropriée, une phylogénie ne vous dit pas grand-chose sur l'ordre de descendance

M11: Progress

PROGRESS 

Site coverage calculated 

[DETAILS](#) [✖ STOP](#)

[STATUS/OPTIONS](#)

RUN STATUS	
Start time	11-12-22 00:39:48
Operation Run Time	05:17:58
Status	Setting site coverage
Log Likelihood	-6,701.73
Operation	Bootstrapping ML tree
Replicate No.	227 of 500

Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny

Lenore Pipes, Hongru Wang, John P Huelsenbeck , Rasmus Nielsen 

Molecular Biology and Evolution, Volume 38, Issue 4, April 2021, Pages 1537–1543,

<https://doi.org/10.1093/molbev/msaa316>

Published: 09 December 2020

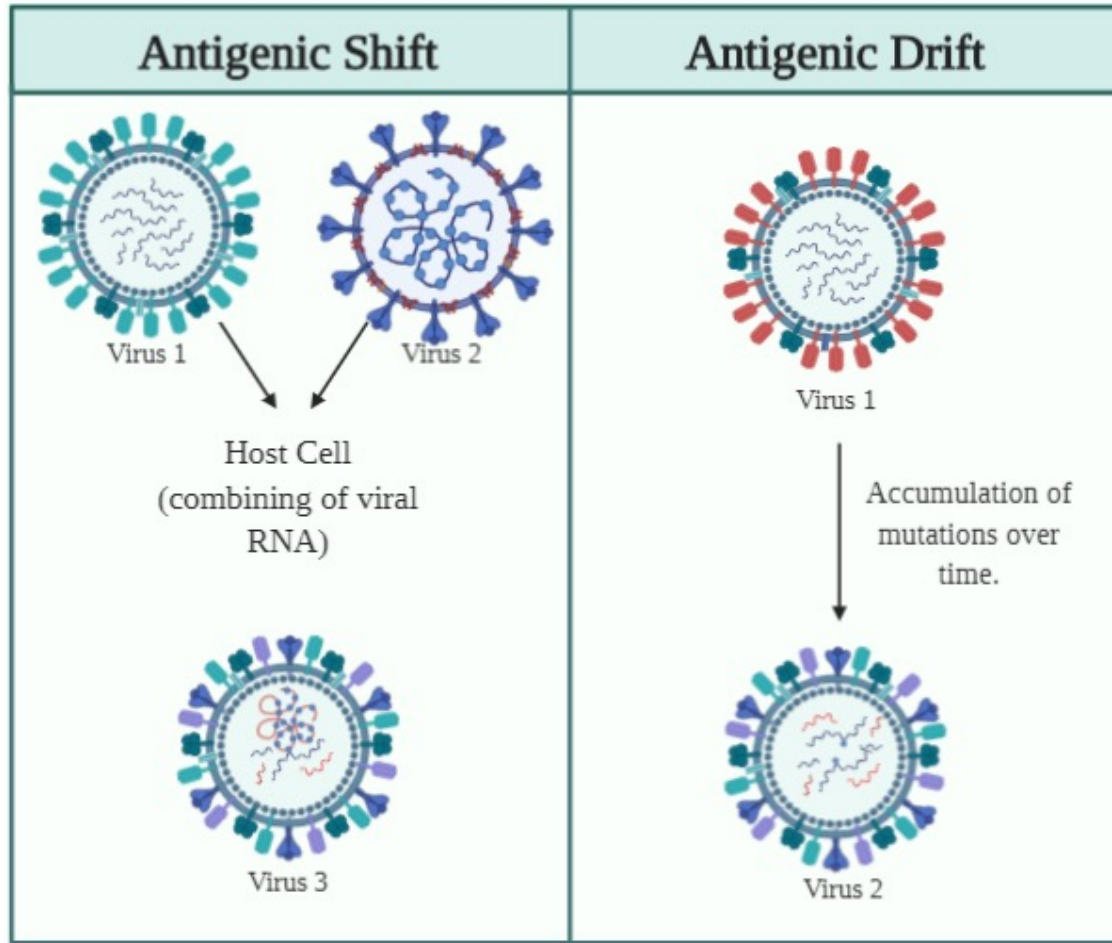
 PDF  Split View  Cite  Permissions  Share ▼

Abstract

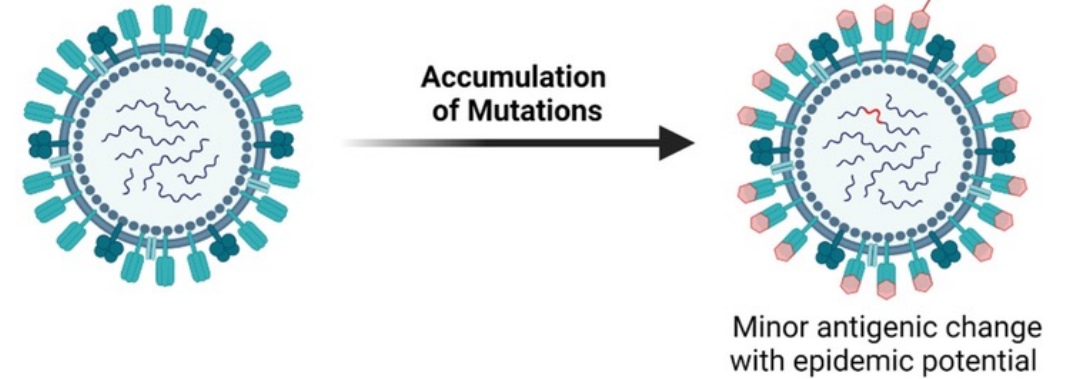
The rooting of the SARS-CoV-2 phylogeny is important for understanding the origin and early spread of the virus. Previously published phylogenies have used different rootings that do not always provide consistent results. We investigate several different strategies for rooting the SARS-CoV-2 tree and provide measures of statistical uncertainty for all methods. We show that methods based on the molecular clock tend to place the root in the B clade, whereas methods based on outgroup rooting tend to place the root in the A clade. The results from the two approaches are statistically incompatible, possibly as a consequence of deviations from a molecular clock or excess back-mutations. We also show that none of the methods provide strong statistical support for the placement of the root in any particular edge of the tree. These results suggest that phylogenetic evidence alone is unlikely to identify the origin of the SARS-CoV-2 virus and we caution against strong inferences regarding the early spread of the virus based solely on such evidence.

**Il est important de
replacer votre
arbre dans son
contexte, sans
contrôle, pouvez-
vous vraiment en
déduire quoi que
ce soit?**

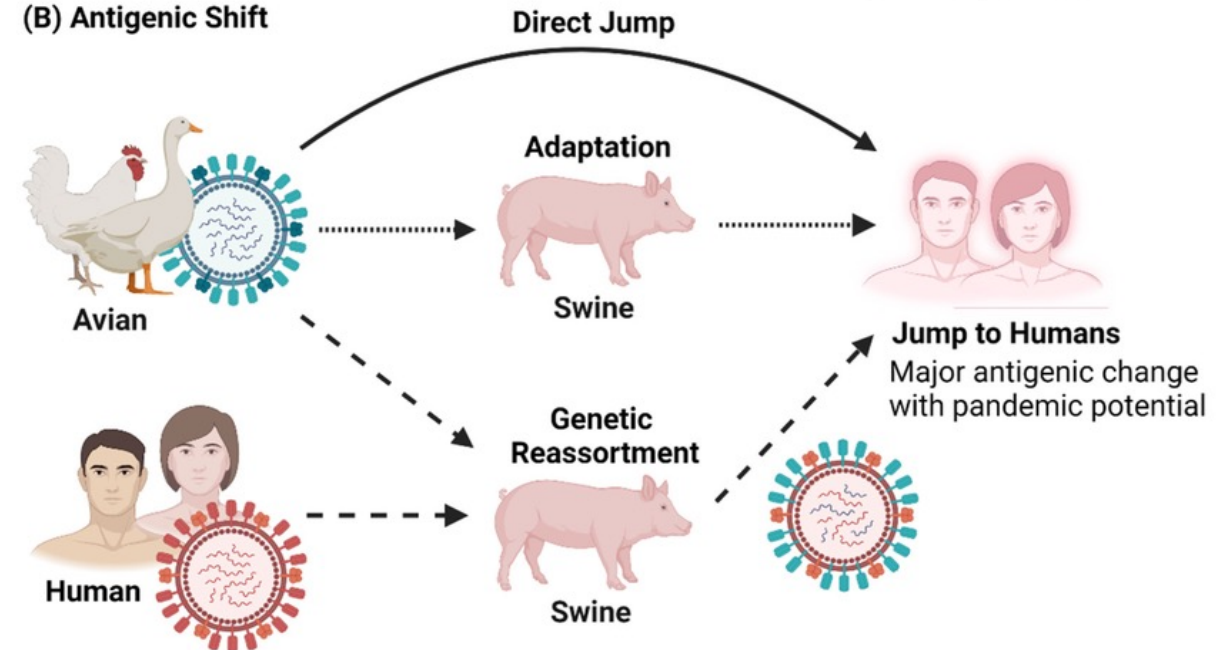
Étude de cas : grippe – décalage et dérive antigéniques



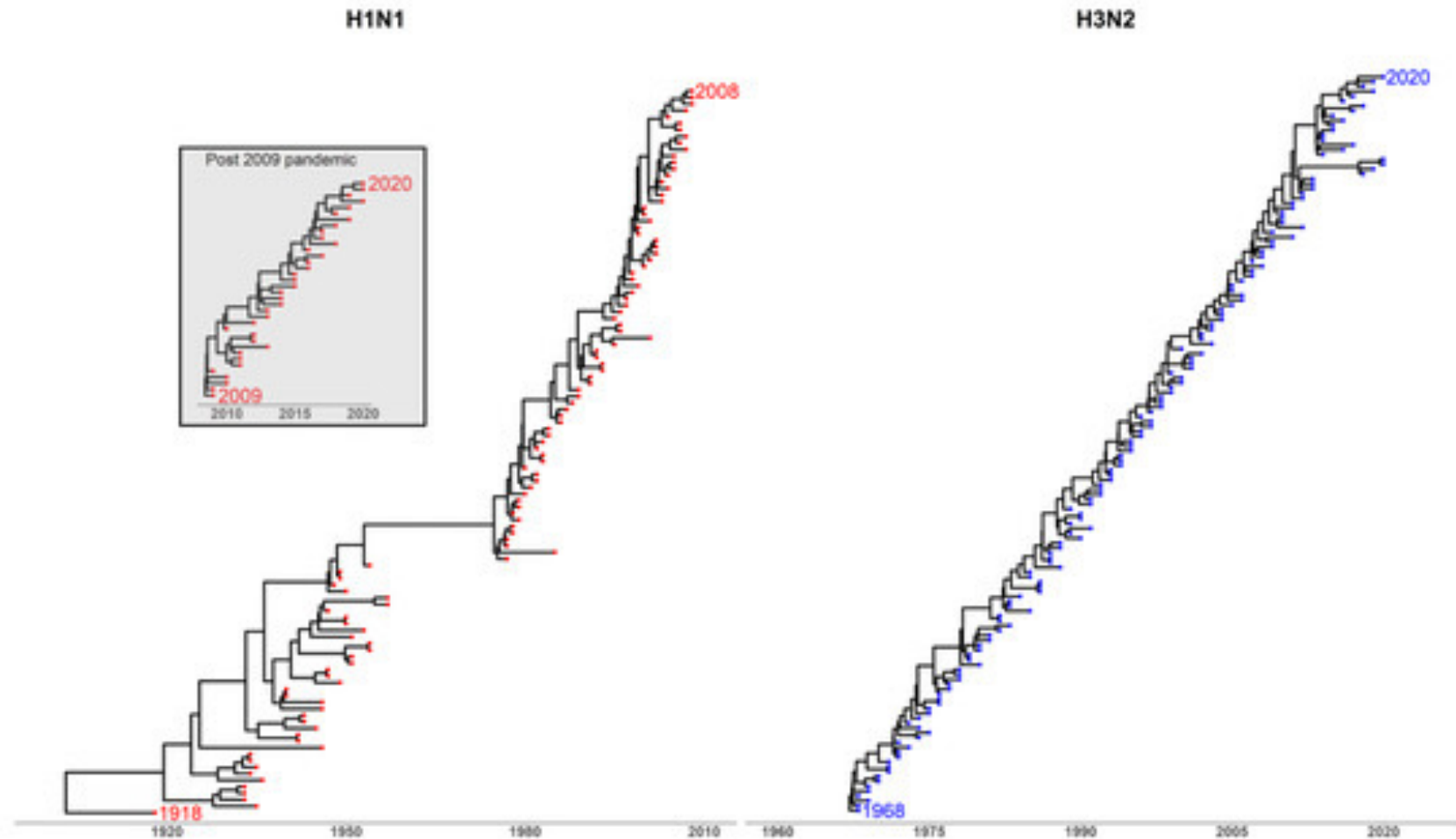
(A) Antigenic Drift



(B) Antigenic Shift

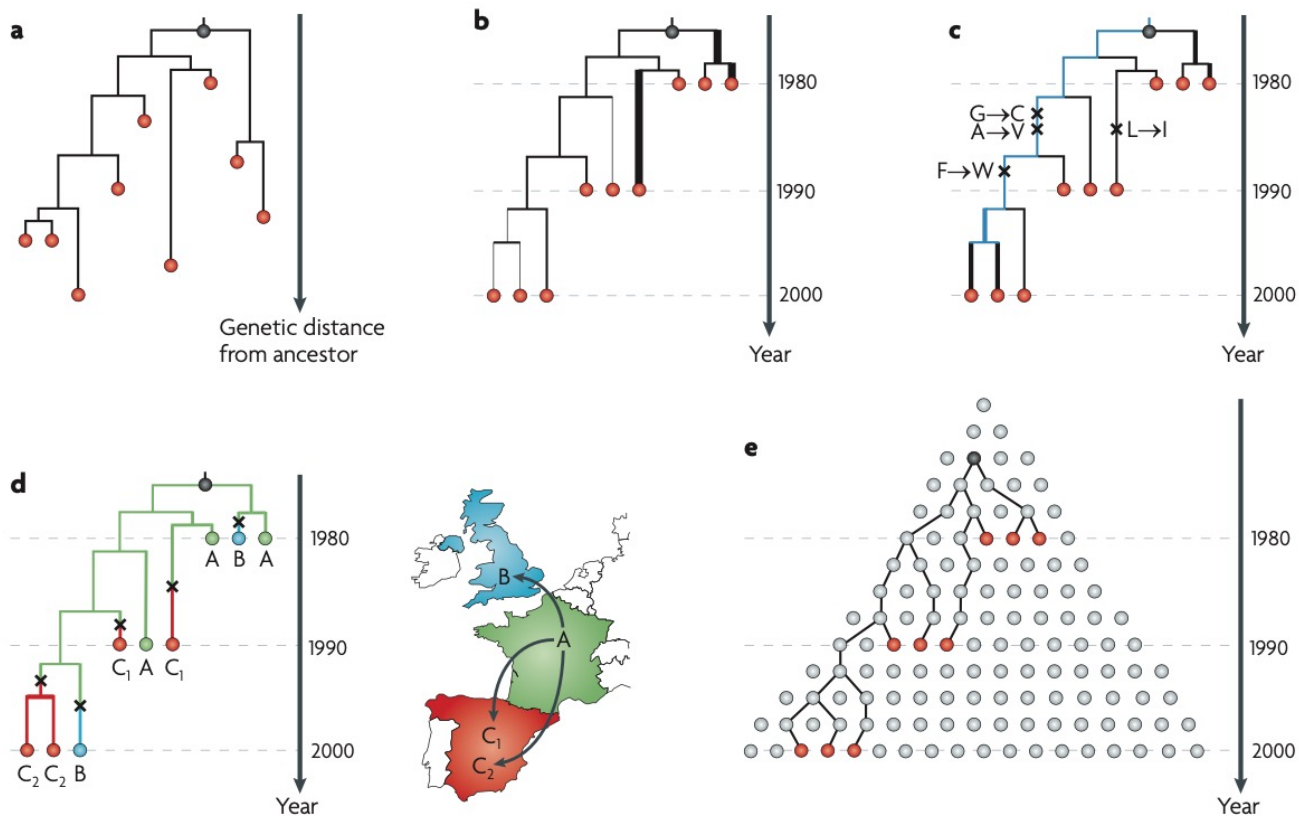


Étude de cas : grippe – décalage et dérive antigéniques



**Y a-t-il une dérive ou un décalage ici ?
Quel impact cela a-t-il sur la conception des vaccins ?**

Box 1 | Phylodynamic techniques



L'importance des formes

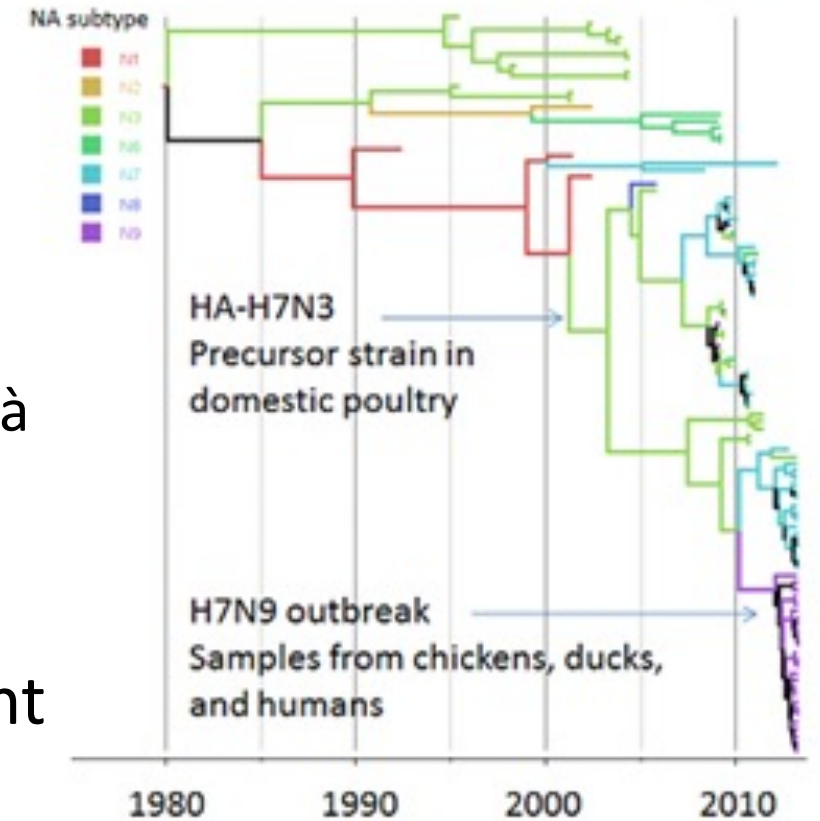
Idealized Phylogeny Shapes	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
		Population size dynamics	Spatial dynamics
Time →		Exponential growth	Strong spatial structure
		Constant size	Weak spatial structure
Examples	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Tree Inferences	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

L'importance des unités

On peut faire beaucoup de choses avec les phylogénies...

- Phylodynamique
- Peut examiner la phylogénie dans le contexte d'autres facteurs
 - Temps (depuis combien de temps ce virus a-t-il divergé)
 - Emplacement (comment un virus a-t-il changé au fur et à mesure qu'il se propage ?)
 - Hôte (comment un virus a-t-il changé dans différents hôtes ?)
- Les phylogénies du maximum de vraisemblance sont bonnes pour :
 - À quel point une chose est différente/similaire par rapport aux choses connues

Phylogeny of Hemagglutinin subtype H7 and reassortments with different Neuraminidases

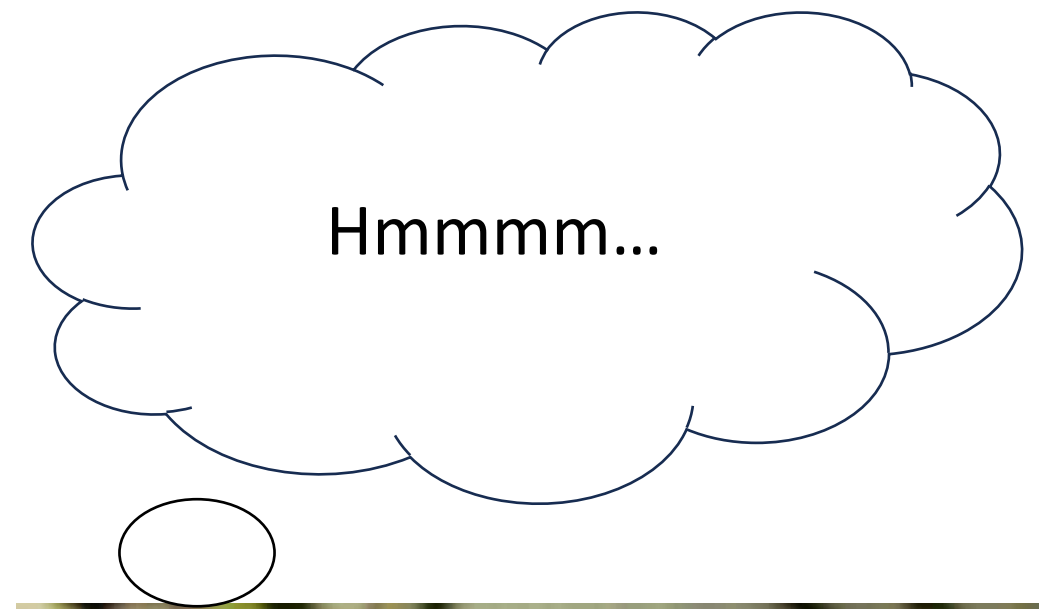


BEAST

Bayesian Evolutionary Analysis Sampling Trees

Point de contrôle!

- La phylogénétique est utile, mais elle demande beaucoup de calculs
- Pourquoi avons-nous besoin d'une racine dans les arbres phylogénétiques ?
- La forme d'une phylogénie peut-elle nous dire quelque chose sur le sujet ?



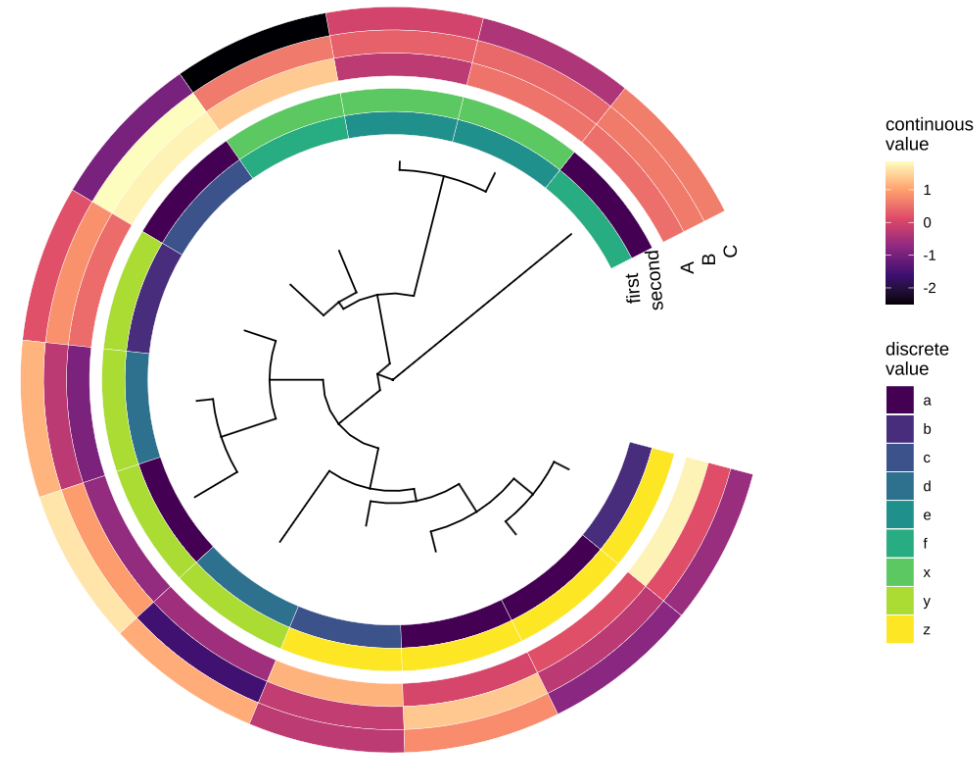
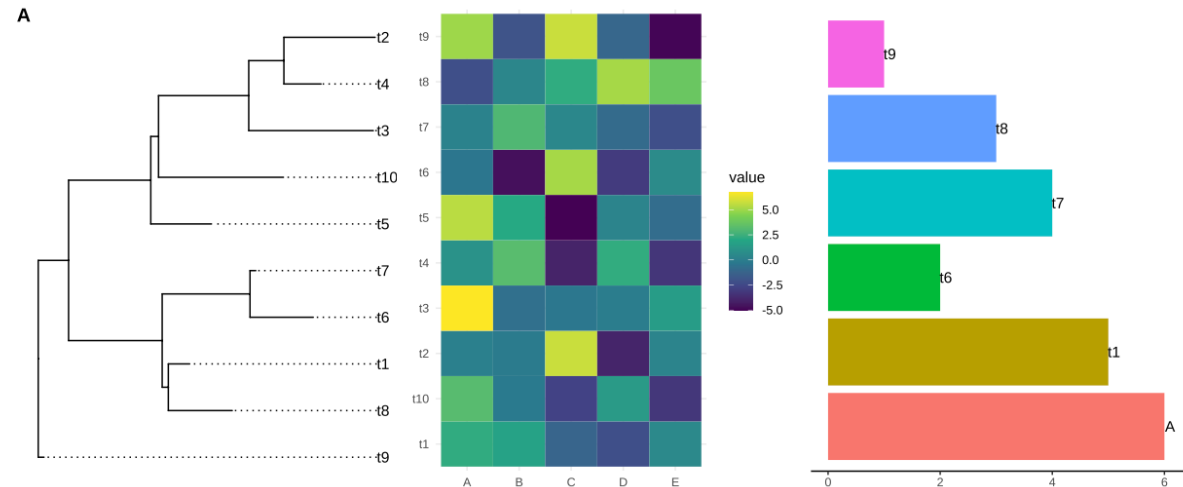
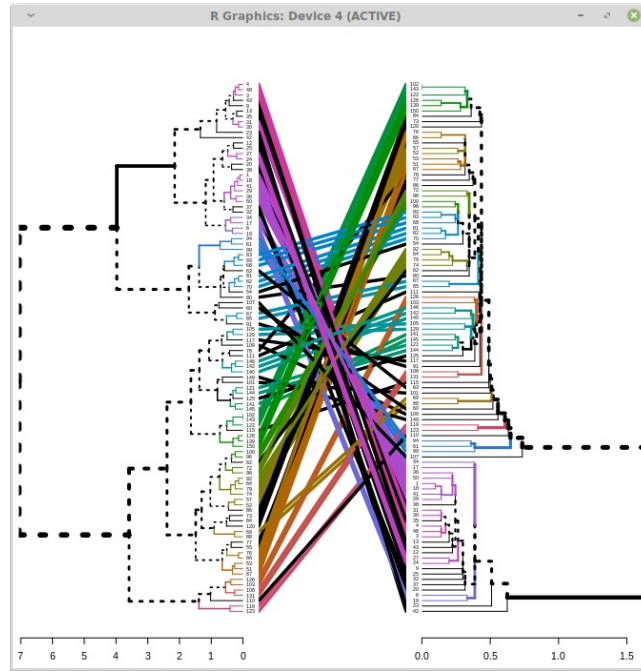
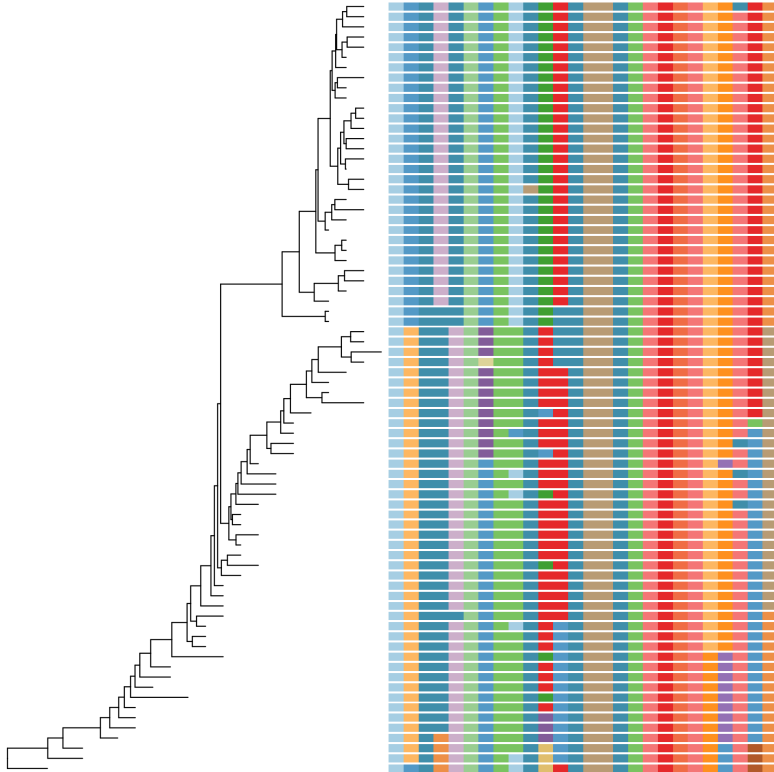
Vous avez donc vos séquences, et maintenant?

- Obtenir des séquences de référence du NCBI
- Obtenir un groupe externe de NCBI
- Alignez-les (utilisez un logiciel comme MEGA ou en ligne, vous avez donc vos séquences, et maintenant ? comme MAFFT)
- Choisissez le meilleur modèle (utilisez un logiciel comme MEGA ou ModelTest-NG)
- Exécutez la phylogénie en utilisant les séquences alignées et le modèle choisi (utilisez un logiciel comme MEGA ou RAxML)
- Visualiser/modifier l'arborescence dans R ou MEGA

Tout ce qui est répertorié est gratuit à utiliser

Potentiel pour de jolies figurines

ggtree: an R package for visualization of tree and annotation data



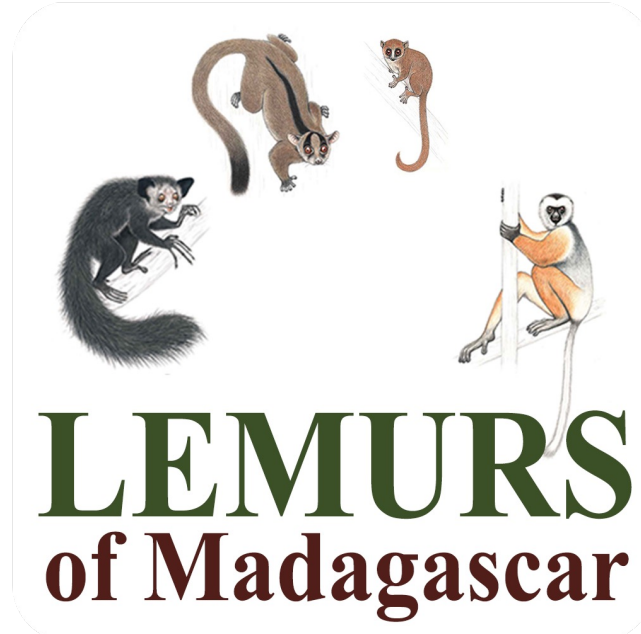
Point de contrôle final!

- Les arbres phylogénétiques ont un large éventail d'applications à de nombreux sujets de recherche!
- Utilise la modélisation et les statistiques "en coulisses"
- **Quelqu'un peut-il donner un exemple de la façon dont il pourrait utiliser une phylogénie pour ses propres recherches? 😊**




Lémuriens du parc national de Ranomafana

- Cytochrome B
 - Utilisé beaucoup dans l'identification des espèces, variabilité limitée à l'intérieur et variation beaucoup plus grande entre les espèces
- Question : à quel point les lémuriens que l'on peut trouver dans le parc national de Ranomafana sont-ils similaires les uns aux autres ?



DIAPOSITIVES À REVOIR À VOTRE RYTHME

Étapes à revoir plus tard

 **National Library of Medicine**
National Center for Biotechnology Information

Log in

BLAST®

HomeRecent ResultsSaved StrategiesHelp

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS


BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 March 2022

[More BLAST news...](#)

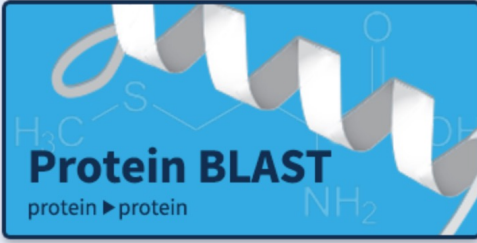
Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

Vérifiez à quel type de séquence vous avez affaire en effectuant une recherche BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA

Or, upload file

Choose File no file selected ?

Job Title

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Database

☒ Standard databases (nr etc.): ☐ rRNA/ITS databases ☐ Genomic + transcript databases ☐ Betacoronaviru

Nucleotide collection (nr/nt)

?

Limit by

Organism ☐ BioProjectID ☐ WGS Project

Optional

Exclude

Optional

Limit to

Optional

Entrez Query

Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

☐ Sequences from type material

Enter an Entrez query to limit search ?

YouTube Create custom database

Program Selection

Optimize for

☒ Highly similar sequences (megablast)
☐ More dissimilar sequences (discontiguous megablast)
☐ Somewhat similar sequences (blastn)
Choose a BLAST algorithm ?

BLAST

Search using Megablast (Optimize for highly similar sequences)
☐ Show results in a new window

NIH

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® » blastn suite » results for RID-TBKASV0K016

HomeRecent ResultsSaved StrategiesHelp

< Edit Search

Save Search

Search Summary ▾

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title

NC_035562.1:14221-15360 Microcebus rufus

RID

TBKASV0K016 Search expires on 12-12 19:30 pm Download All ▾

Program

BLASTN ? Citation ▾

Database

nt See details ▾

Query ID

lcl|Query_55759

Description

NC_035562.1:14221-15360 Microcebus rufus isolate HAB...

Molecule type

dna

Query Length

1140

Other reports

Distance tree of results MSA viewer ?

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity

E value

Query Coverage

to

to

to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾

Select columns ▾

Show 100 ▾ ?

☒ select all 100 sequences selected



GenBank


Graphics

Distance tree of results



MSA Viewer

	Description ▾	Scientific Name ▾	Max Score ▾	Total Score ▾	Query Cover ▾	E value ▾	Per. Ident	Acc. Len ▾	Accession
<input checked="" type="checkbox"/>	Microcebus rufus isolate HAB06.12 mitochondrion, complete genome	Microcebus rufus	2106	2106	100%	0.0	100.00%	16819	KM112297.1
<input checked="" type="checkbox"/>	Microcebus rufus isolate VEV7.13 mitochondrion, complete genome	Microcebus rufus	1751	1751	100%	0.0	94.39%	16822	KM112317.1

 An official website of the United States government [Here's how you know](#) 

 **National Library of Medicine**
National Center for Biotechnology Information

Log in


All Databases  Eulemur ruffrons cytochrome B  **Search**


NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation


Welcome to NCBI


The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.


[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)


Submit
Deposit data or manuscripts into NCBI databases


Download
Transfer NCBI data to your computer


Learn
Find help documents, attend a class or watch a tutorial


Develop
Use NCBI APIs and code libraries to build applications


Analyze
Identify an NCBI tool for your data analysis task


Research
Explore NCBI research and collaborative projects


COVID-19 Information

Popular Resources
[PubMed](#)
[Bookshelf](#)
[PubMed Central](#)
[BLAST](#)
[Nucleotide](#)
[Genome](#)
[SNP](#)
[Gene](#)
[Protein](#)
[PubChem](#)

NCBI News & Blog
[Join NCBI at PAG 30](#) 08 Dec 2022
San Diego, January 13-18, 2023 NCBI is looking forward to seeing you in person at the International Plant and Animal
[Announcing the NCBI SARS-CoV-2 Variant Calling Pipeline and Related Data Products](#) 01 Dec 2022
[Still waiting for an analysis pipeline that](#)
[New Proximity Search Feature Available in PubMed](#) 30 Nov 2022
PubMed, a free National Library of

Allez au NCBI, et recherchez la chose pour laquelle vous voulez construire une phylogénie, dans notre cas le cytochrome B des lémuriens dans le parc national de Ranomafana

Search NCBI

Eulemur rufifrons cytochrome B

×

Search

Results found in 4 databases

<div>Literature</div> <div><div>Bookshelf</div><div>0</div></div> <div><div>MeSH</div><div>0</div></div> <div><div>NLM Catalog</div><div>0</div></div> <div><div>PubMed</div><div>0</div></div> <div><div>PubMed Central</div><div>4</div></div>	<div>Genes</div> <div><div>Gene</div><div>0</div></div> <div><div>GEO DataSets</div><div>0</div></div> <div><div>GEO Profiles</div><div>0</div></div> <div><div>HomoloGene</div><div>0</div></div> <div><div>PopSet</div><div>0</div></div>	<div>Proteins</div> <div><div>Conserved Domains</div><div>0</div></div> <div><div>Identical Protein Groups</div><div>5</div></div> <div><div>Protein</div><div>28</div></div> <div><div>Protein Family Models</div><div>0</div></div> <div><div>Structure</div><div>0</div></div>
<div>Genomes</div> <div><div>Assembly</div><div>0</div></div> <div><div>BioCollections</div><div>0</div></div> <div><div>BioProject</div><div>0</div></div> <div><div>BioSample</div><div>0</div></div> <div><div>Genome</div><div>0</div></div> <div><div>Nucleotide</div><div>28</div></div> <div><div>SRA</div><div>0</div></div>	<div>Clinical</div> <div><div>ClinicalTrials.gov</div><div>0</div></div> <div><div>ClinVar</div><div>0</div></div> <div><div>dbGaP</div><div>0</div></div> <div><div>dbSNP</div><div>0</div></div> <div><div>dbVar</div><div>0</div></div> <div><div>GTR</div><div>0</div></div> <div><div>MedGen</div><div>0</div></div>	<div>PubChem</div> <div><div>BioAssays</div><div>0</div></div> <div><div>Compounds</div><div>0</div></div> <div><div>Pathways</div><div>0</div></div> <div><div>Substances</div><div>0</div></div>

Voici à quoi cela ressemblera, vous pouvez aller à Nucléotide dans la catégorie génome et cliquer dessus

Nucleotide

Nucleotide

Eulemur rufifrons cytochrome b

Search

Create alertAdvanced

Help

Species

Animals (28)

Customize ...

Molecule types

genomic DNA/RNA (28)

Customize ...

Source databases

INSDC (GenBank) (28)

Customize ...

Sequence Type

Nucleotide (28)

Genetic compartments

Mitochondrion (28)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

Clear all

Show additional filters

Summary

20 per page

Sort by Default order

Send to:

Filters: [Manage Filters](#)

See Gene information for b cytochrome **cytochrome b**

b in [Drosophila melanogaster \(2\)](#) [Escherichia phage Lambda](#) [All 50 Gene records](#)

cytochrome in [Cricetulus griseus](#) [Tripterygium wilfordii \(2\)](#) [All 4 Gene records](#)

cytochrome b in [Pongo abelii](#) [1 Gene record](#)

Items: 1 to 20 of 28

<< First < Prev Page 1 of 2 Next > Last >>

☐ [Eulemur rufifrons clone Erufi-NHMB89006 cytochrome b gene, partial cds; mitochondrial](#)

1. 223 bp linear DNA

Accession: KF708347.1 GI: 556926369

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-NHM1882314 cytochrome b gene, partial cds; mitochondrial](#)

2. 223 bp linear DNA

Accession: KF708346.1 GI: 556926367

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-MCZ16357 cytochrome b gene, partial cds; mitochondrial](#)

3. 223 bp linear DNA

Accession: KF708345.1 GI: 556926365

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)

☐ [Eulemur rufifrons clone Erufi-MCZ16356 cytochrome b gene, partial cds; mitochondrial](#)

4. 223 bp linear DNA

Accession: KF708344.1 GI: 556926363

Choisissez la séquence de ce qui vous intéresse, dans notre cas nous voulons un cds complet

Nous pourrions vouloir des cds partiels si nous avons une séquence partielle d'intérêt, mais pour l'instant, nous construisons simplement un arbre avec des données connues, donc les cds complets sont les meilleurs

Cds : séquence codante pour la protéine

Ensuite, téléchargez le fastas

☐ [Eulemur rufifrons clone Erufi-MM-448 cytochrome b gene, complete cds; mitochondrial](#)

7. 1,140 bp linear DNA

Accession: KF708293.1 GI: 556926260

[Protein](#) [PubMed](#) [Taxonomy](#)

[GenBank](#) [FASTA](#) [Graphics](#)



Lorsque vous avez toutes vos séquences d'intérêt et votre exogroupe, vous devez concaténer les séquences en un seul fichier, vous pouvez le faire en créant un fichier texte/édition et en collant chaque séquence, sinon suivez les instructions sur la ligne de commande (mac) ou powershell (windows) pour le faire

```
lemur_cytochrome_b_fastas — -bash — 121x27
Last login: Wed Nov 23 12:34:19 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Gwenddolens-MacBook-air:~ gwenddolenkettenburg$ cd Desktop
(base) Gwenddolens-MacBook-air:Desktop gwenddolenkettenburg$ cd Intro_phylogenetic_modeling_Kettenburg
(base) Gwenddolens-MacBook-air:Intro_phylogenetic_modeling_Kettenburg gwenddolenkettenburg$ cd lemur_cytochrome_b_fastas
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$ cat *.fasta>lemur_cytB_concatenated
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$
```

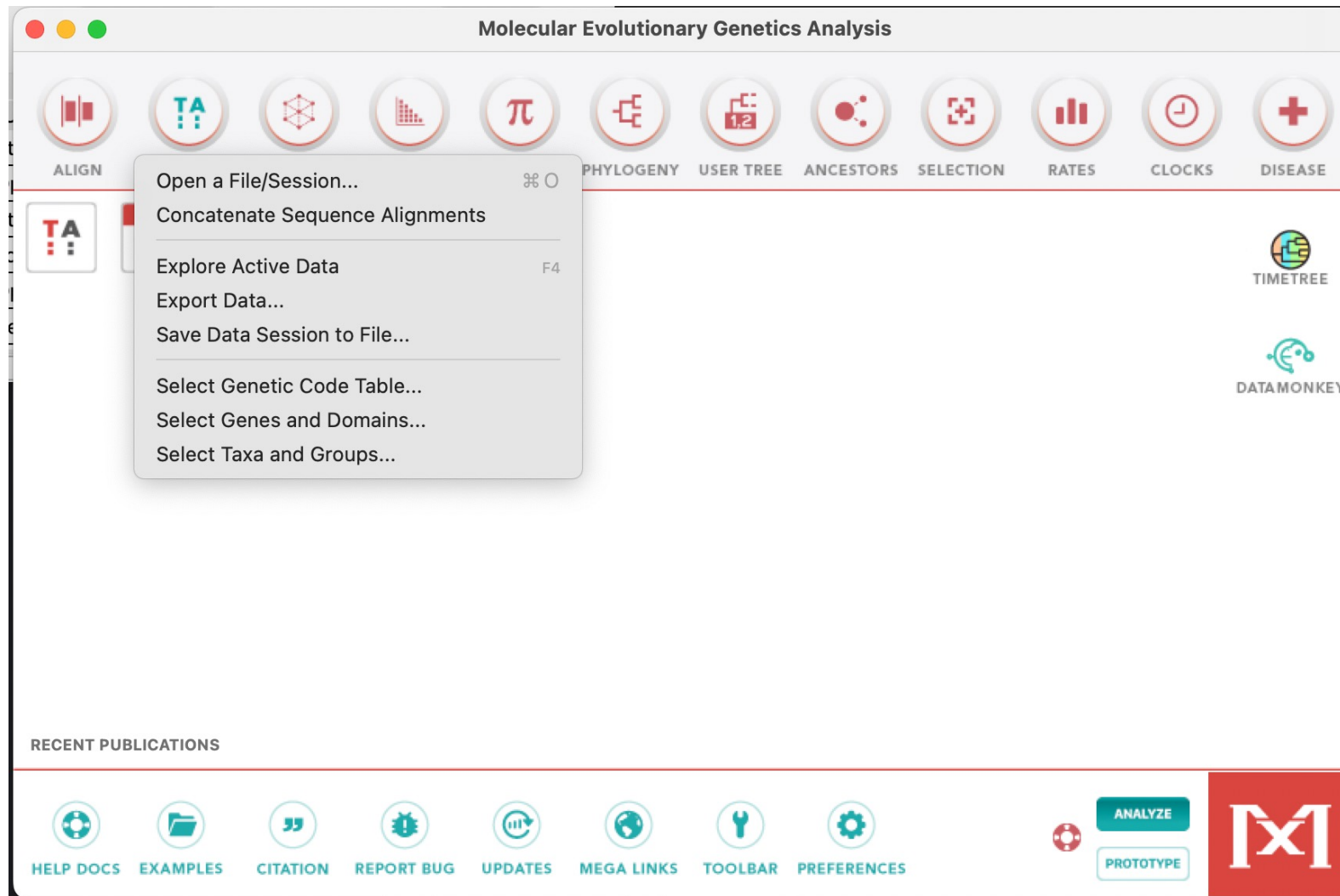
Example 1: Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
type file1.csv file2.csv > concat.csv
```

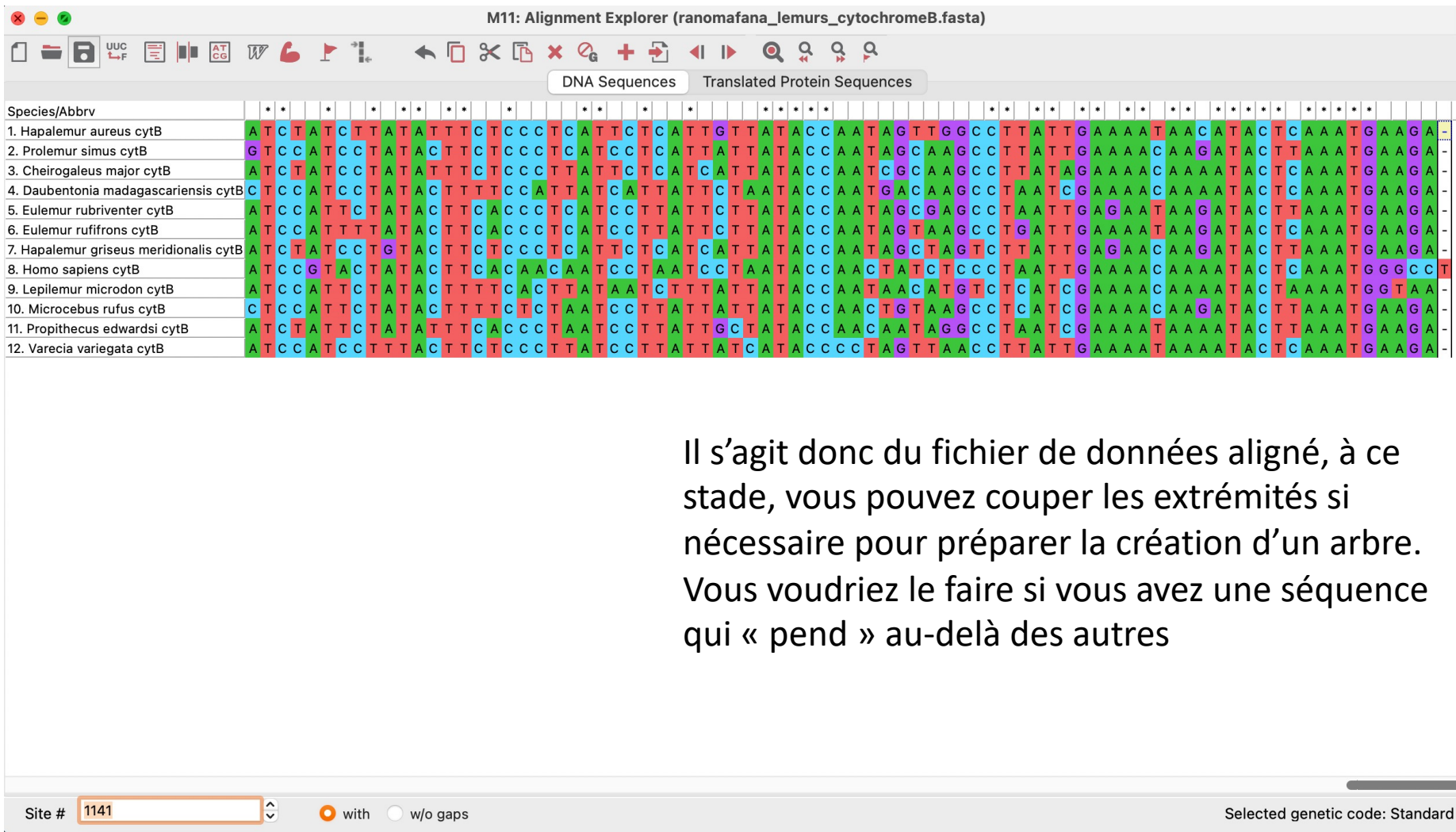
Example 2: Merge files with pattern (This will merge all files with csv extension and create concat.csv)

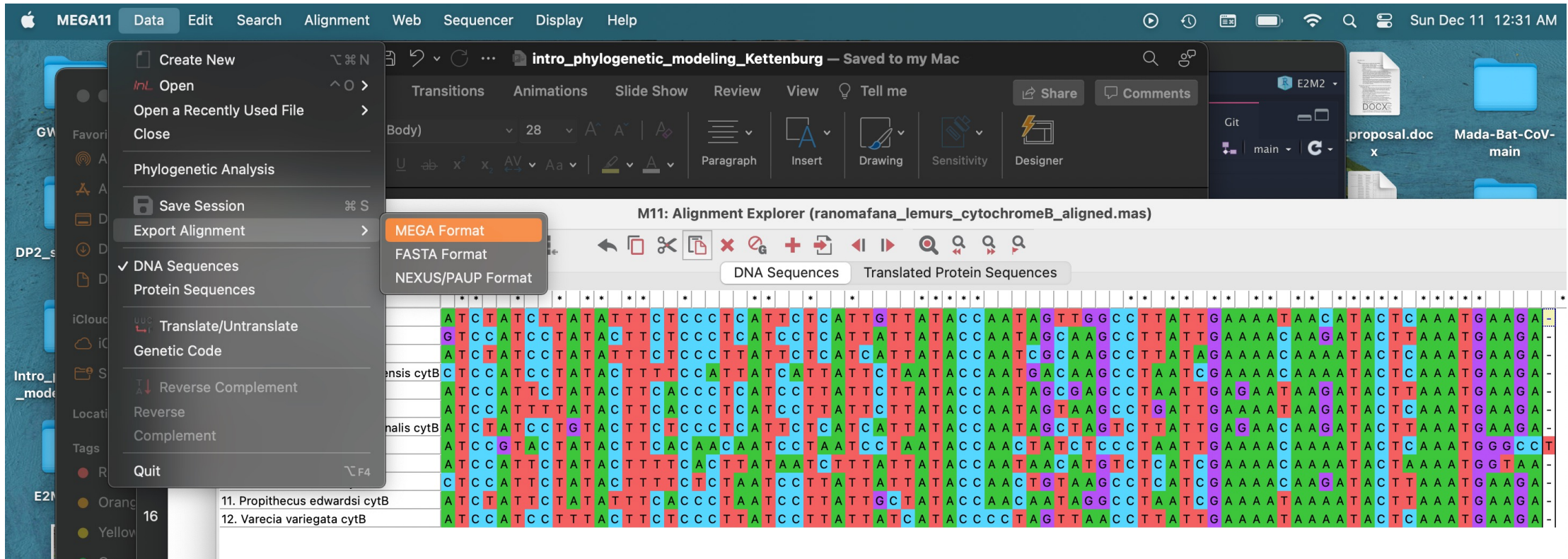
When using asterisk(*) to concatenate all files. Please DON'T use same extension for target file(Eg. .csv). There should be some difference in pattern else target file will also be considered in concatenation

```
type *.csv > concat_csv.txt
```



Ouvrez MEGA, et ouvrez un fichier/session, sélectionnez votre fichier fasta concaténé MEGA vous demandera si vous souhaitez aligner ou analyser, cliquez sur aligner





Enregistrez le fichier aligné, puis nous procéderons à la sélection du modèle

Ouvrir l'interface graphique RAxML

I
N
P
U
T

LOAD ALIGNMENT

A
N
A
L
Y
S
I
S

ML + transfer bootstrap expectation + consensus ▾
Analysis

1 ▾ 100 ▾
Runs Reps.

628076
Seed

O
U
T
P
U
T

Select output directory

output
Select output name

raxmlGUI 2.0

R
A
X
M
L

raxml-ng-ARM64 ▾
Binary

RUN

raxml-ng-ARM64 --all --msa RAxML_output_concat.txt --model --prefix output
seed 628076 --bs-metric tbe --tree rand{1} --bs-trees 100
Command

C
O
N
S
O
L
E

raxmlGUI 2.0.10 raxml-ng-ARM64 1.1.0

[How to cite?](#) [For questions or suggestions contact us!](#)

nucleotide

ranomafana_lemurs_cytochromeB_align.fas

12 sequences of length 1141



GTR



none



none



Substitution model

Stationary frequencies

Proportion of invariant sites

none



Rate heterogeneity

RUN MODELTEST

Partition 1/1:

	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

Chargez le fichier aligné et effectuez modeltest, pour la sélection du modèle, utilisez le score BIC, il crachera un rapport qui s'enregistrera dans vos fichiers

nucleotide

ranomafana_lemurs_cytochromeB_align.fas

12 sequences of length 1141

TIM2

none

+I (ML estimate)

Substitution model

Stationary frequencies

Proportion of invariant sites

+GAMMA (mean)

RUN MODELTEST

Rate heterogeneity

ML + transfer bootstrap expectation + consensus

1

100

<none>

Hapalemur_aureus_cytB

Prolemur_simus_cytB

Cheirogaleus_major_cytB

Daubentonia_madagascariensis_cytB

Eulemur_rubriventer_cytB

Eulemur_rufifrons_cytB

Hapalemur_griseus_meridionalis_cytB

Homo_sapiens_cytB

Lepilemur_microdon_cytB

Microcebus_rufus_cytB

Propithecus_edwardsi_cytB

Varecia_variegata_cytB

cytochromeB_align'

_cytochromeB_align.ckp

_cytochromeB_align.log

_cytochromeB_align.out

_cytochromeB_align.tree

RAXML

raxml-ng-ARM64

RUN

Binary

raxml-ng-ARM64 --all --msa /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytochr --model TIM2+I+G --prefix /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytochr seed 177748 --bs-metric tbe --tree rand{1} --bs-trees 100

Command

P.Inv: 0.4945
Alpha: 0.2560
Alpha-P.Inv: 0.7313
P.Inv-Alpha: 0.3876
Frequencies: 0.3075 0.3419 0.1142 0.2364

Commands:
> phylml -i /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cyto as -m 010232 -f m -v e -a e -c 4 -o tlr
> raxmlHPC-SSE3 -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lem align.fas -m GTRGAMMAIX -n EXEC_NAME -p PARSIMONY_SEED
> raxml-ng --msa /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs ign.fas --model TIM2+I+G4
> paup -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cytoch as -m TIM2+I+G4
> iqtree -s /Users/gwenddolenkettenburg/Desktop/RAXML/ranomafana_lemurs_cyto as -m TIM2+I+G4

Summary:

Partition 1/1:

	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

Execution results written to /Users/gwenddolenkettenburg/Desktop/RAXML/RAXML_C nomafana_lemurs_cytochromeB_align.out
Starting tree written to /Users/gwenddolenkettenburg/Desktop/RAXML/RAXML_GUI_M fana_lemurs_cytochromeB_align.tree

raxmlGUI 2.0.10

raxml-ng-ARM64 1.1.0

How to cite? For questions or suggestions contact i

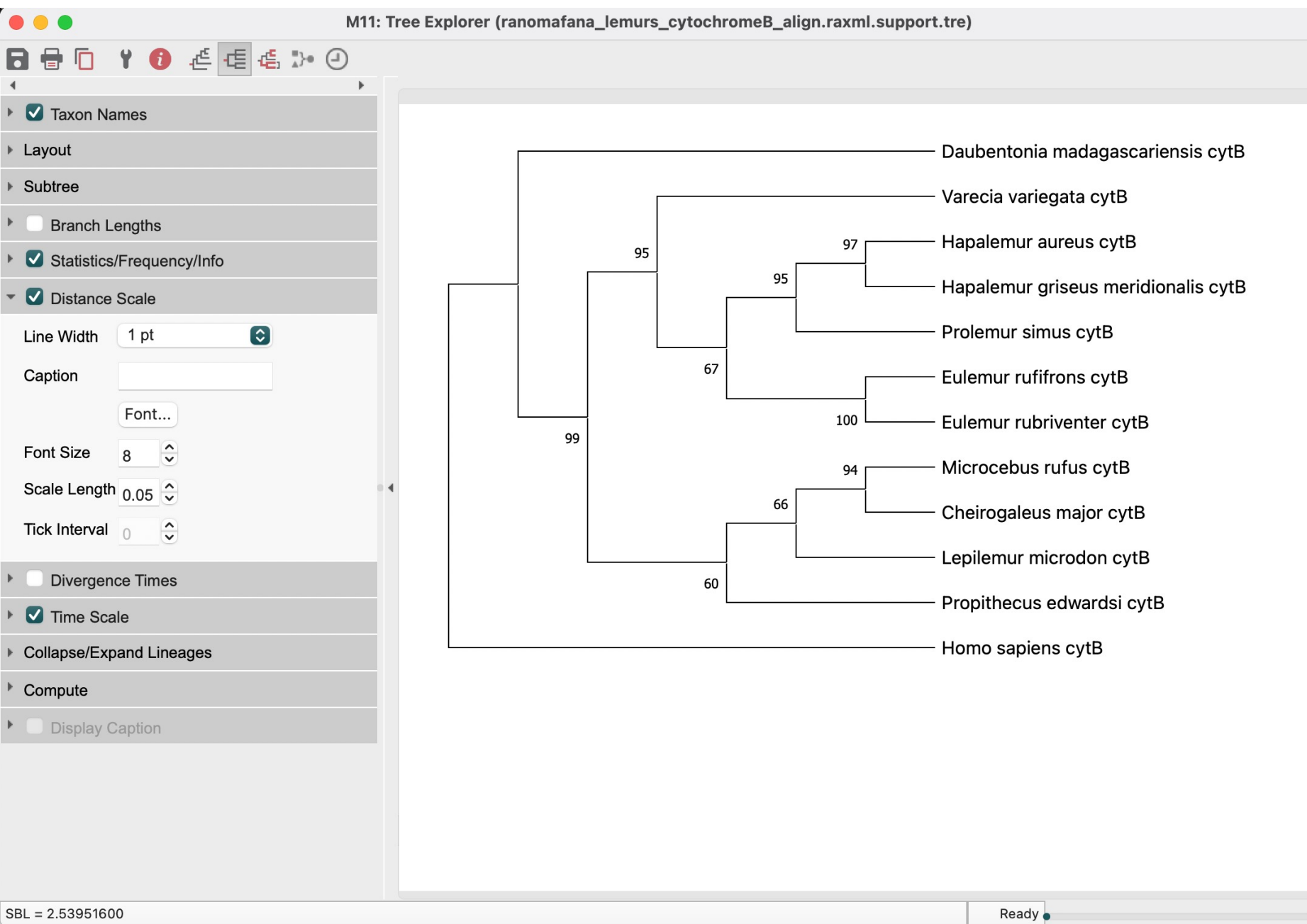
À l'aide du modèle de modeltest, modifiez selon vos besoins dans la zone de saisie, puis dans l'analyse, définissez l'outgroup, puis appuyez sur exécuter dans la section RAXML

Il va aussi recracher quelques fichiers...

>	modeltest	Today, 1:24 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.fas	Today, 1:22 PM	14 KB	De
▼	raxml	Today, 1:26 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.raxml.bestModel.txt	Today, 1:26 PM	90 bytes	Pl
	ranomafana_lemurs_cytochromeB_align.raxml.bestTree.tre	Today, 1:26 PM	518 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.bootstraps.tre	Today, 1:26 PM	52 KB	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.log.txt	Today, 1:26 PM	11 KB	Pl
	ranomafana_lemurs_cytochromeB_align.raxml.rba	Today, 1:26 PM	8 KB	De
	ranomafana_lemurs_cytochromeB_align.raxml.startTree.tre	Today, 1:26 PM	507 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.support.tre	Today, 1:26 PM	590 bytes	Fi
	RAxML_GUI_Settings_ranomafana_lemurs_cytochromeB_align.txt	Today, 1:26 PM	566 bytes	Pl

Nous sommes intéressés par le fichier .support.tre

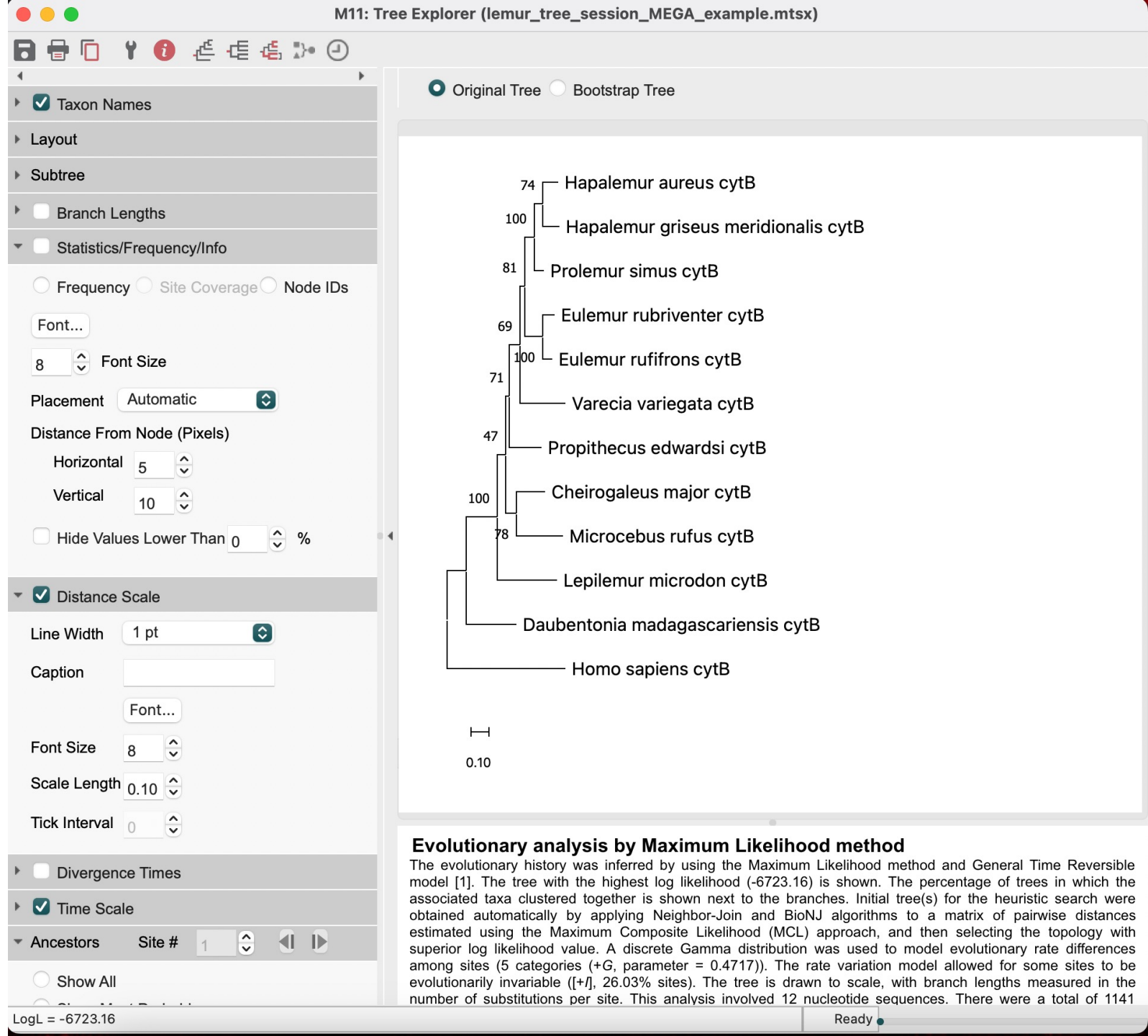
Cela inclura la topologie de l'arborescence et les valeurs de prise en charge des amorçages, ouvertes dans MEGA



Nous voulons un fichier de l'arbre enregistré qui peut être lu dans R, donc newick. Cliquez sur Exporter les arbres... choisissez le format newick et personnalisez-le dans R

Vous pouvez également personnaliser dans MEGA... c'est juste plus limité

Nous aurions
pu faire du
modeltest et du
RAxML dans
MEGA aussi...
mais prend une
éternité !



Alors faites joli en R !

- Suivez les instructions de `lemur_tree_editing`. Fichier R