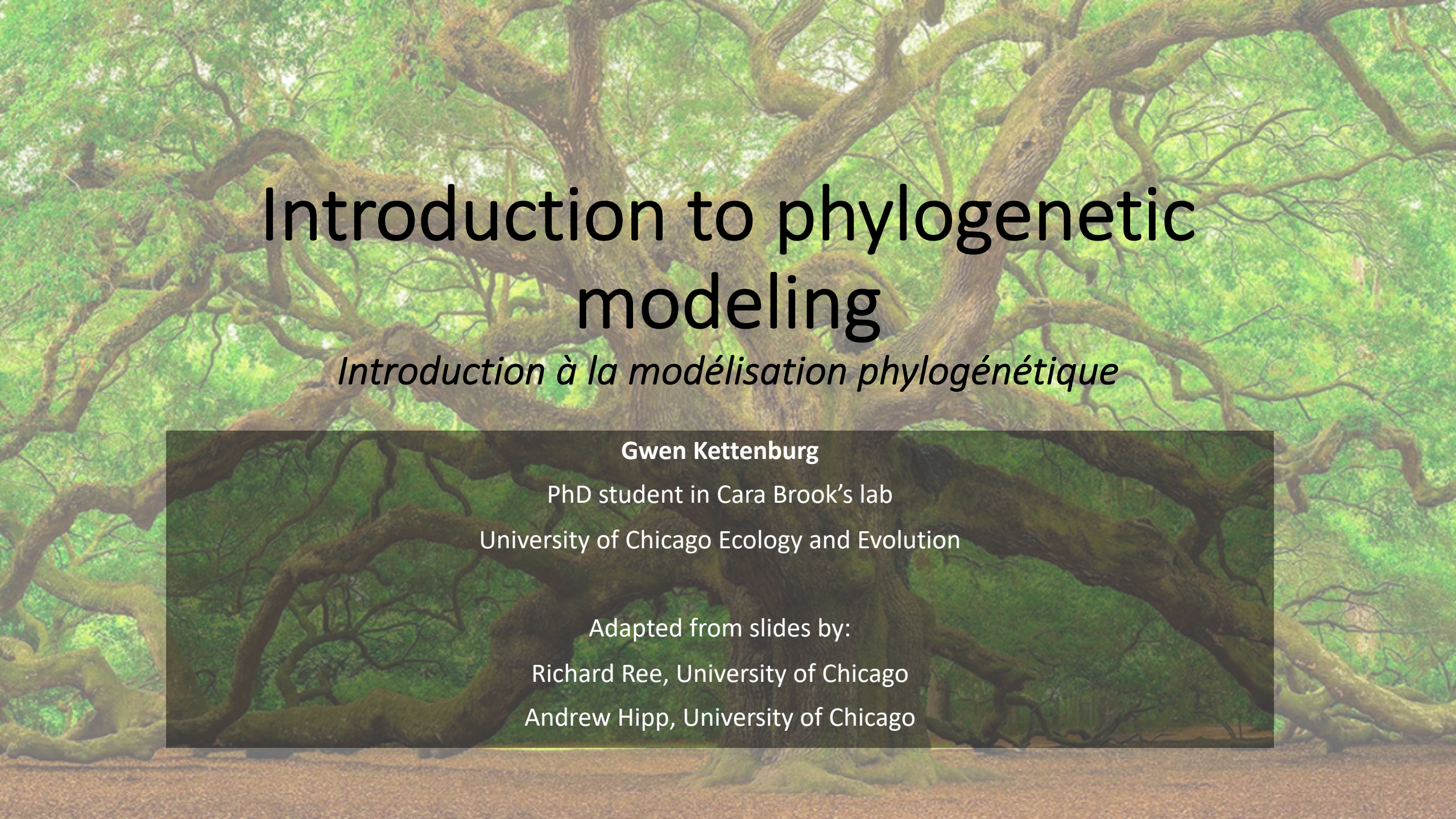


Introduction to phylogenetic modeling

Introduction à la modélisation phylogénétique

Gwen Kettenburg

PhD student in Cara Brook's lab

University of Chicago Ecology and Evolution

Adapted from slides by:

Richard Ree, University of Chicago

Andrew Hipp, University of Chicago

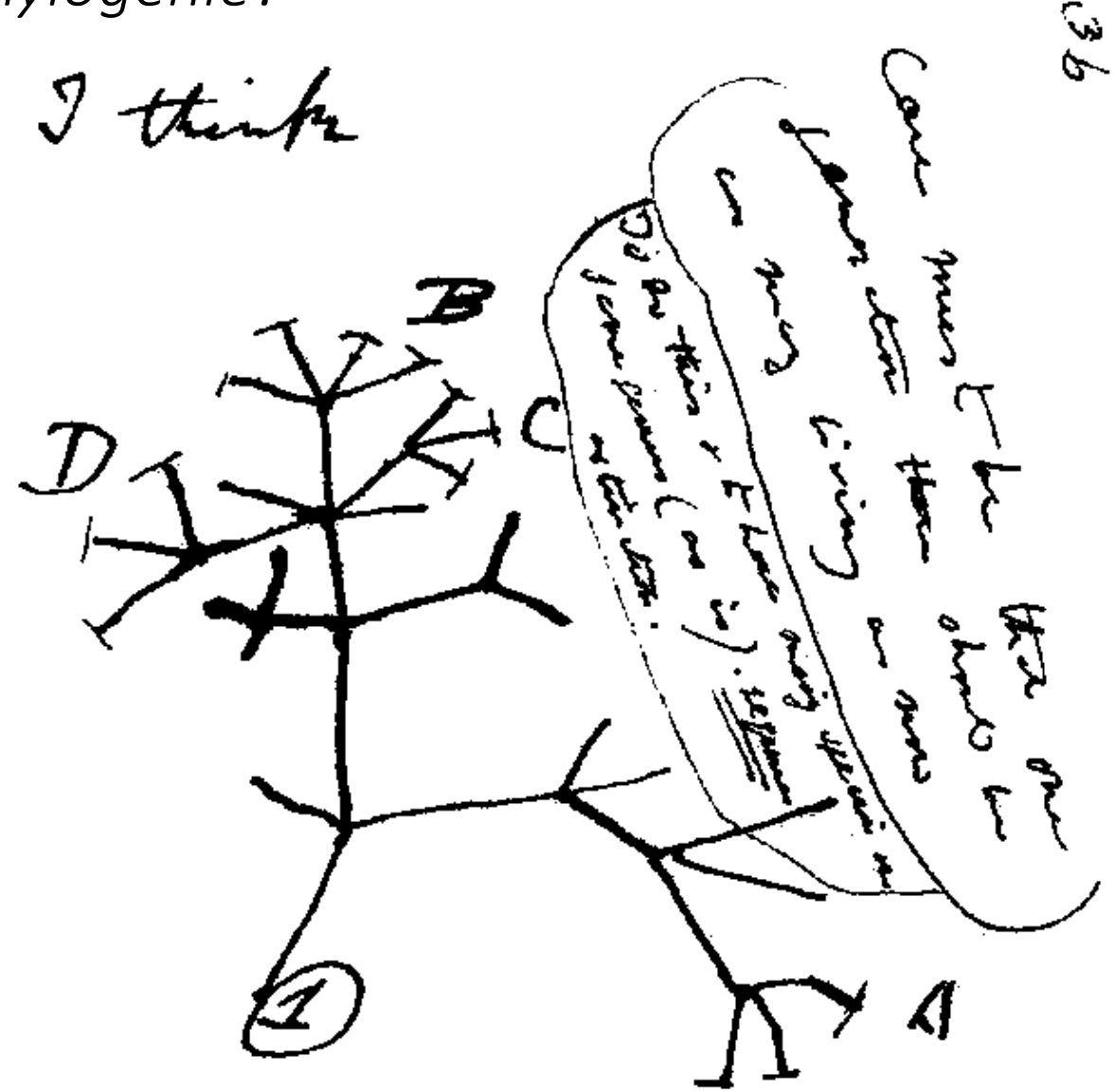
Reminders:

Rappels:

- If you need something clarified or have a question at any time, interrupt!
  - *Si vous avez besoin de clarifier quelque chose ou si vous avez une question à tout moment, interrompez !*
- If you need something translated to Malagasy/French so it’s clearer, please let me know
  - *Si vous avez besoin de quelque chose traduit en malgache/français pour que ce soit plus clair, faites-le moi savoir*
- Remember to say your name when you raise your hand
  - *N'oubliez pas de dire votre nom lorsque vous levez la main*

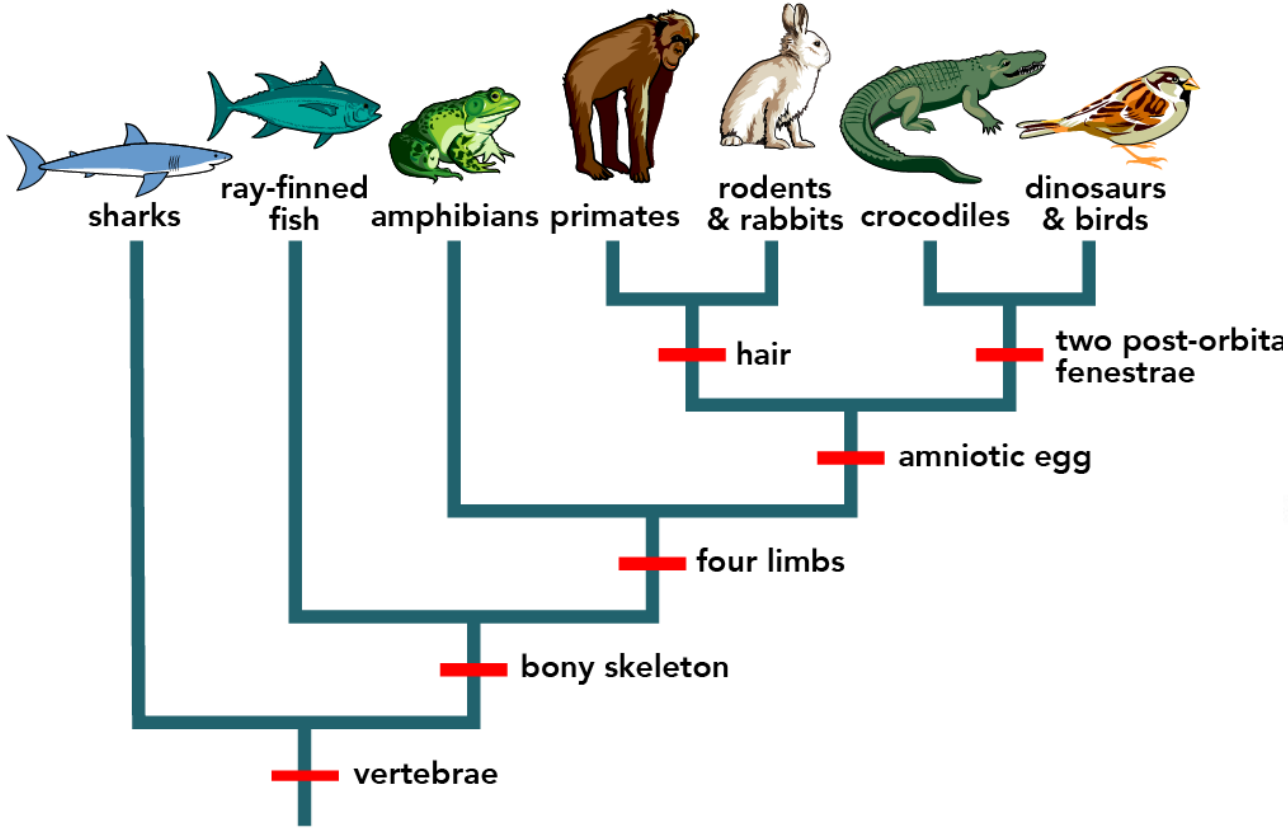
What is a phylogeny?

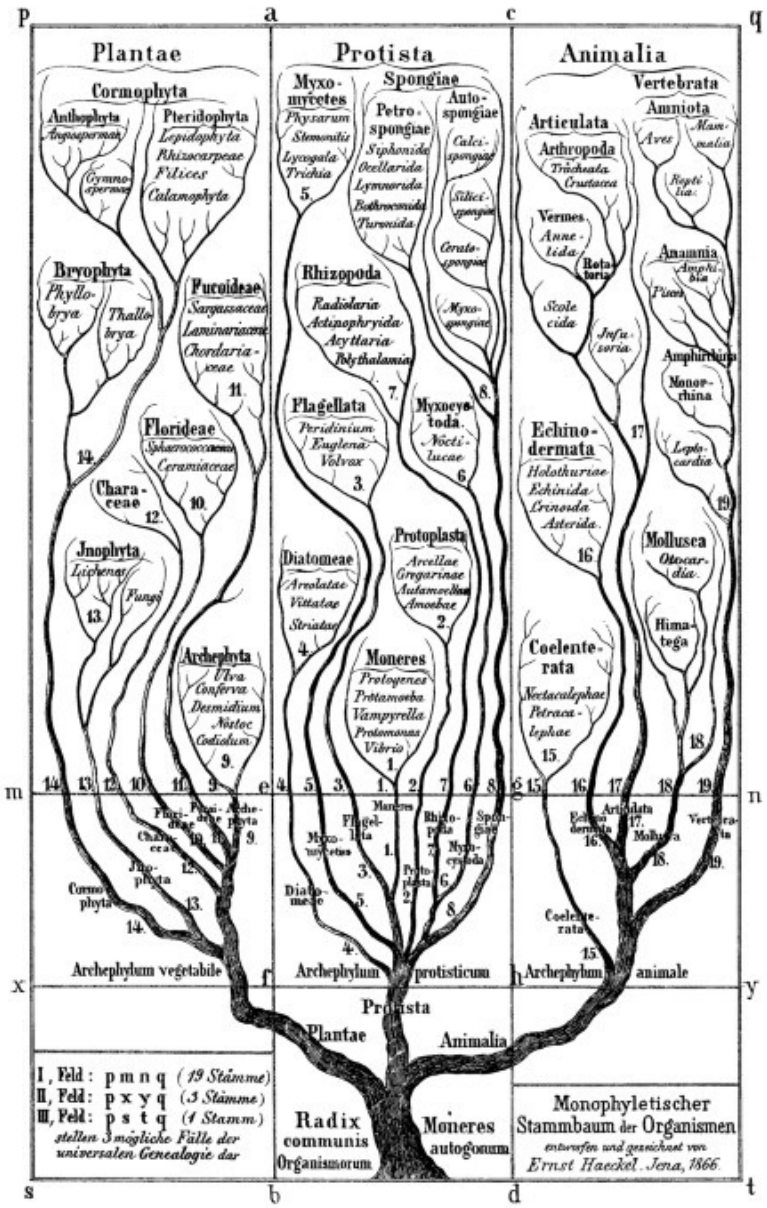
Qu'est-ce qu'une phylogénie?



What is a phylogeny?

Qu'est-ce qu'une phylogénie?

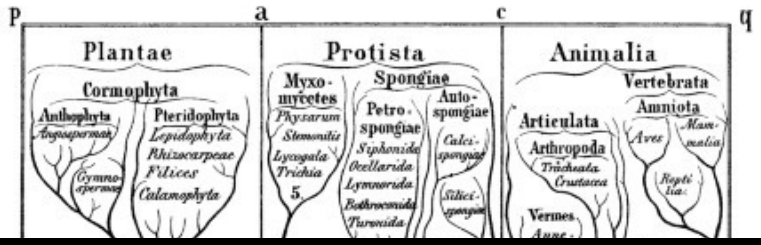


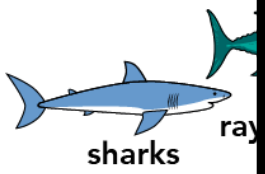


Hossfeld and Levit, Nature, 2016

What is a phylogeny?

Qu'est-ce qu'une phylogénie?





"A phylogenetic tree, or a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor."

"Un arbre phylogénétique, ou une phylogénie, est un diagramme qui décrit les lignes de descendance évolutive de différentes espèces, organismes ou gènes d'un ancêtre commun."

Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

How is this useful to the E's in E2M2?

En quoi est-ce utile aux E dans E2M2?

- Epidemiology and disease research uses phylogenies a lot
- Ecology is increasingly using phylogenetic methods to demonstrate relationships among species
- Evolutionary ecology focuses on ID'ing adaptive values of traits under different conditions
- Maybe it should be E3M2 in the future!

- *L'épidémiologie et la recherche sur les maladies utilisent beaucoup les phylogenies*
- *L'écologie utilise de plus en plus des méthodes phylogénétiques pour démontrer les relations entre les espèces*
- *L'écologie évolutive se concentre sur l'identification des valeurs adaptatives des traits dans différentes conditions*
- *Peut-être que ça devrait être E3M2 à l'avenir!*

Goals:

Buts:

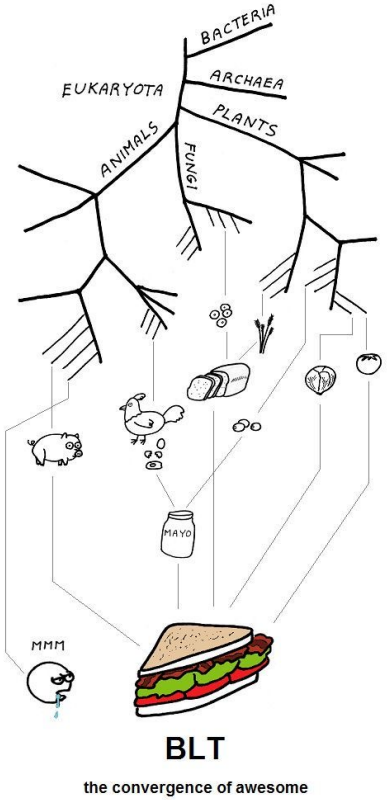
- Lecture component
  - Learn basics of what a phylogeny is
  - Learn how to read phylogenies
  - Basics of phylogenetic modeling
- Tutorial component
  - Learn how to make a phylogenetic tree from sequencing data
  - Using lemur sequences in MEGA software
  - Edit and visualize tree in R and FigTree

Composante conference

- *Apprendre les bases de ce qu'est une phylogénie*
- *Apprendre à lire les phylogénies*
- *Bases de la modélisation phylogénétique*

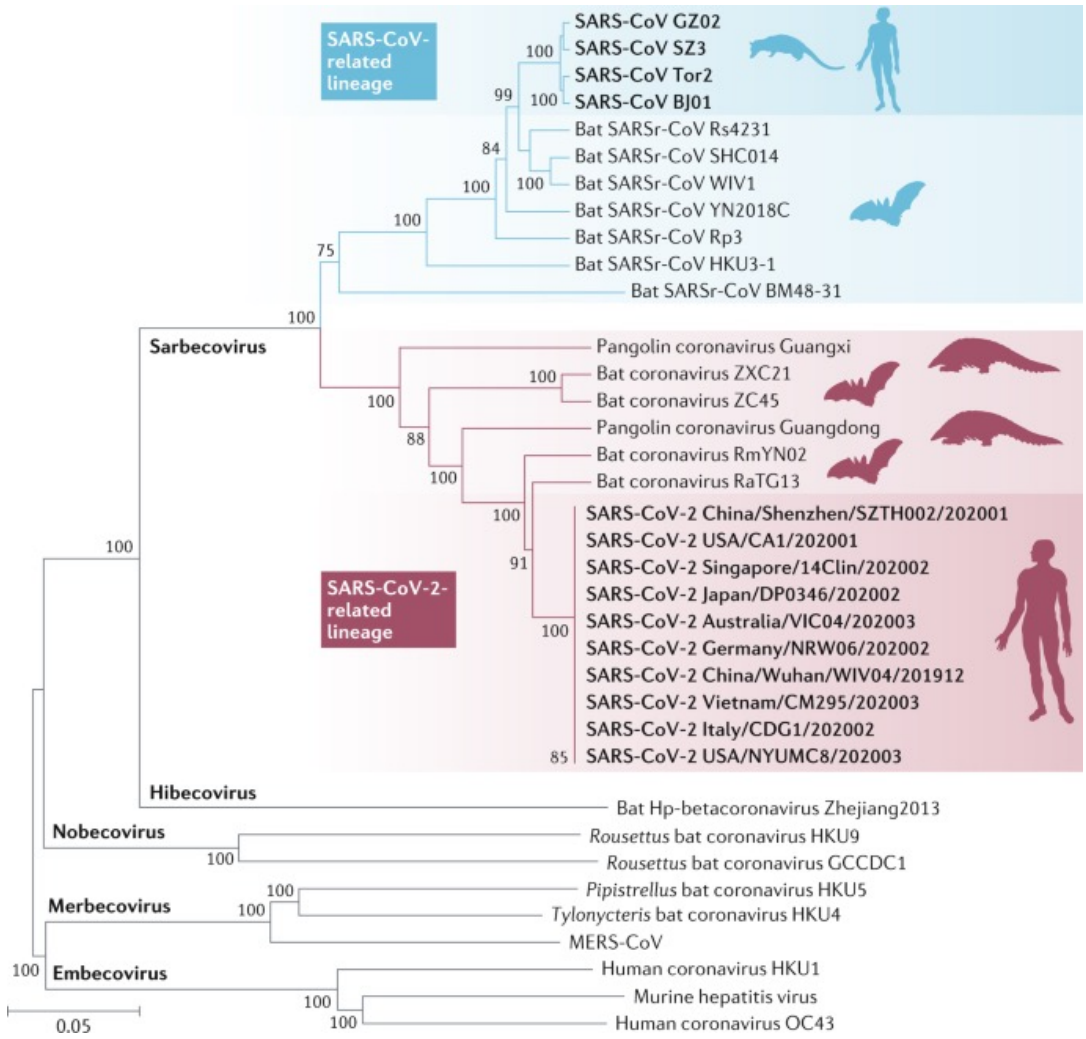
Composant de didacticiel

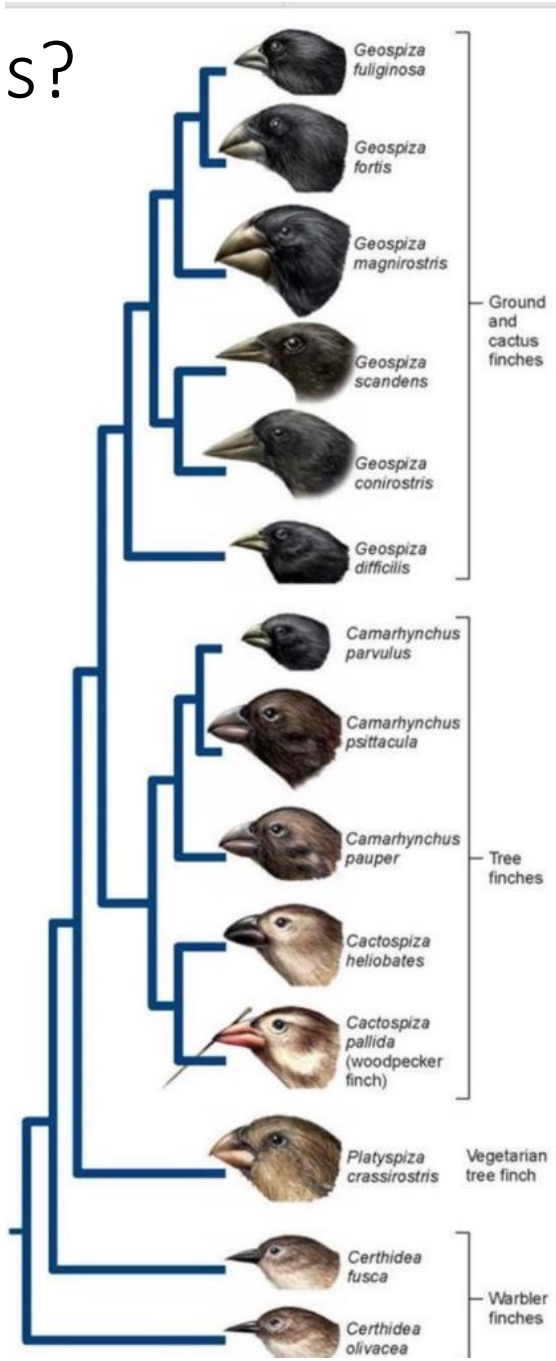
- *Apprendre à faire un arbre phylogénétique à partir de données de séquençage*
- *Utilisation des séquences de lémuriens dans le logiciel MEGA*
- *Modifier et visualiser l'arborescence dans R et FigTree*

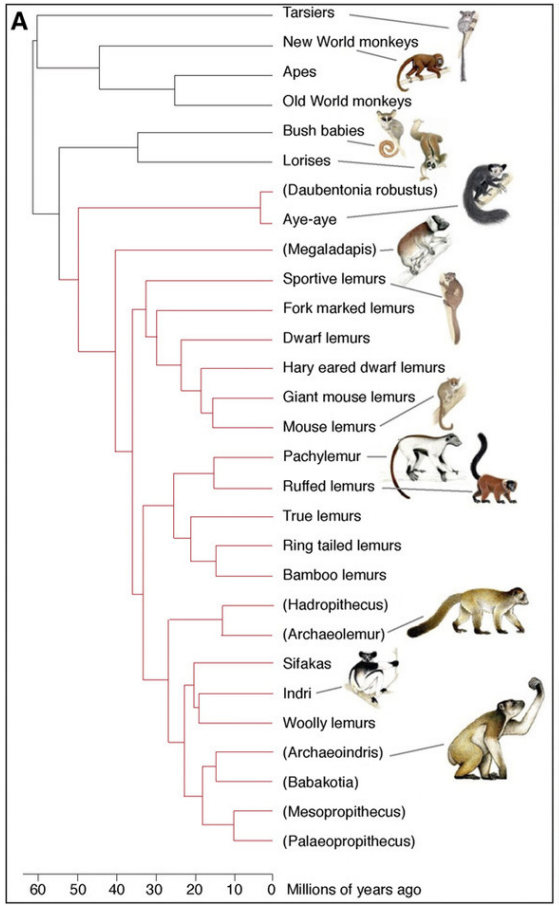


What can you do with phylogenies?

Que faire des phylogénies?





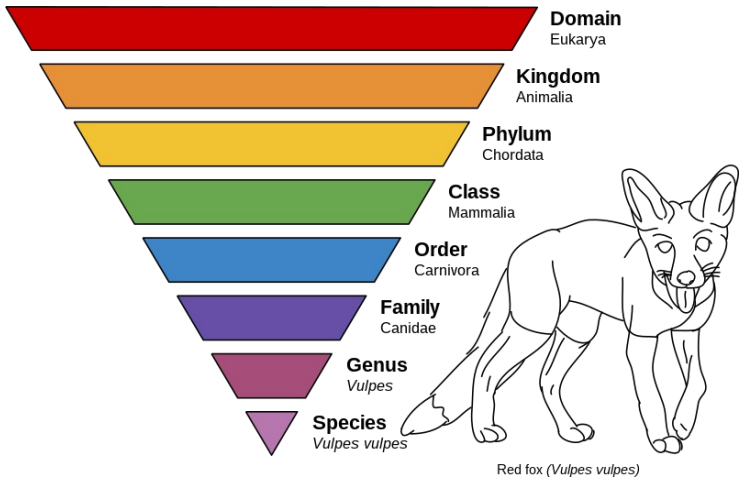


Ezran et al., Genetics, 2022

Hu et al., Nature reviews, 2022

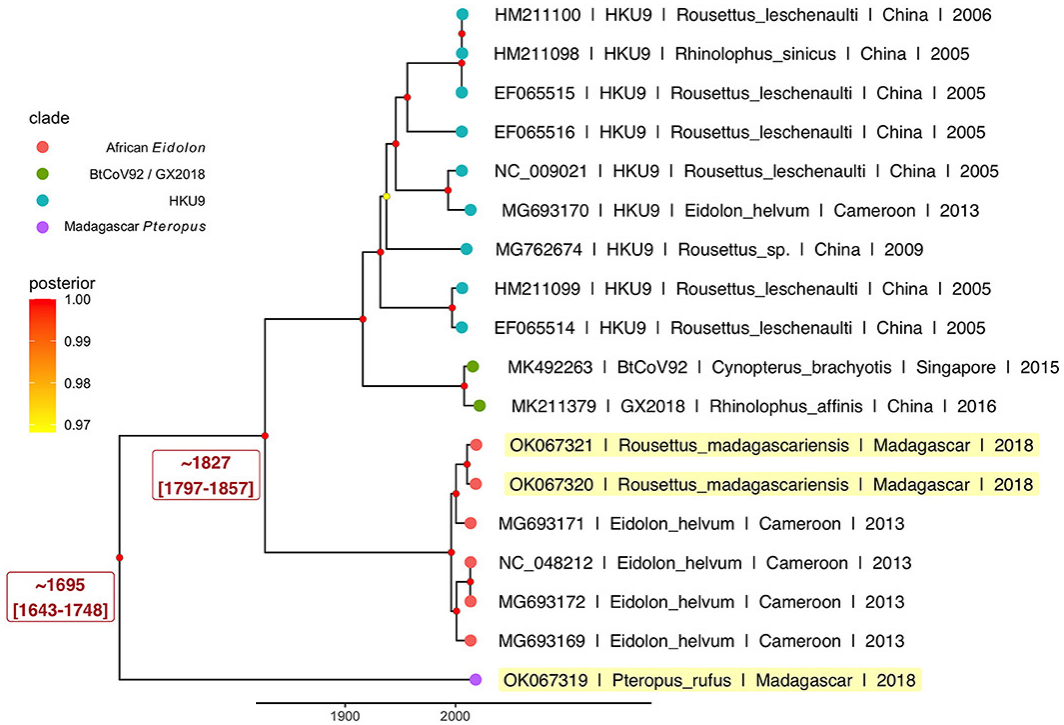
Taxonomy/nomenclature

Taxonomie/nomenclature



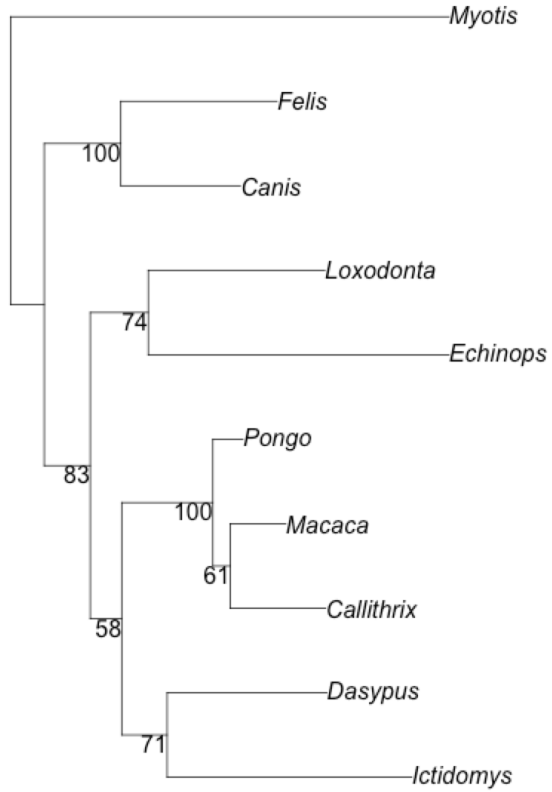
Bayesian trees

Arbres bayésiens

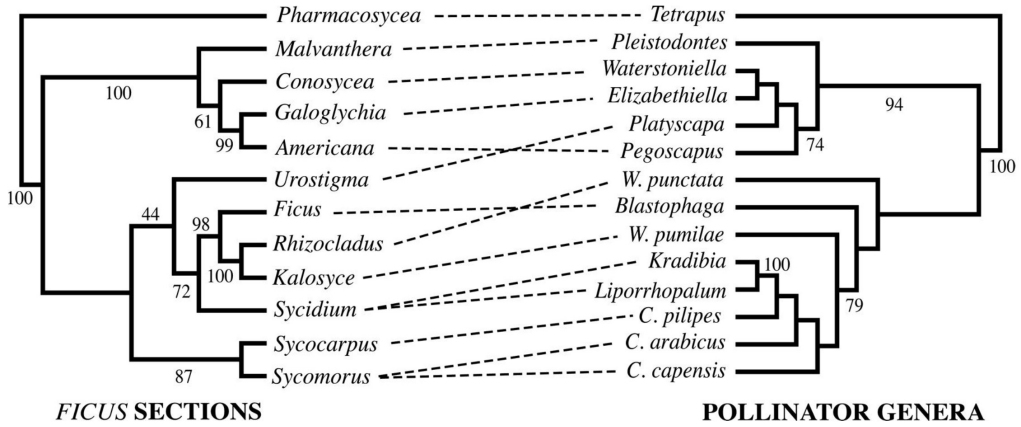


Maximum likelihood

Plausibilité maximum



Figs and fig wasps



plant and pollinator phylogenies show limited congruence; host switching and hybridization has been common in their coevolutinary history

Adapted from Ree and Hipp, UChicago, 2021

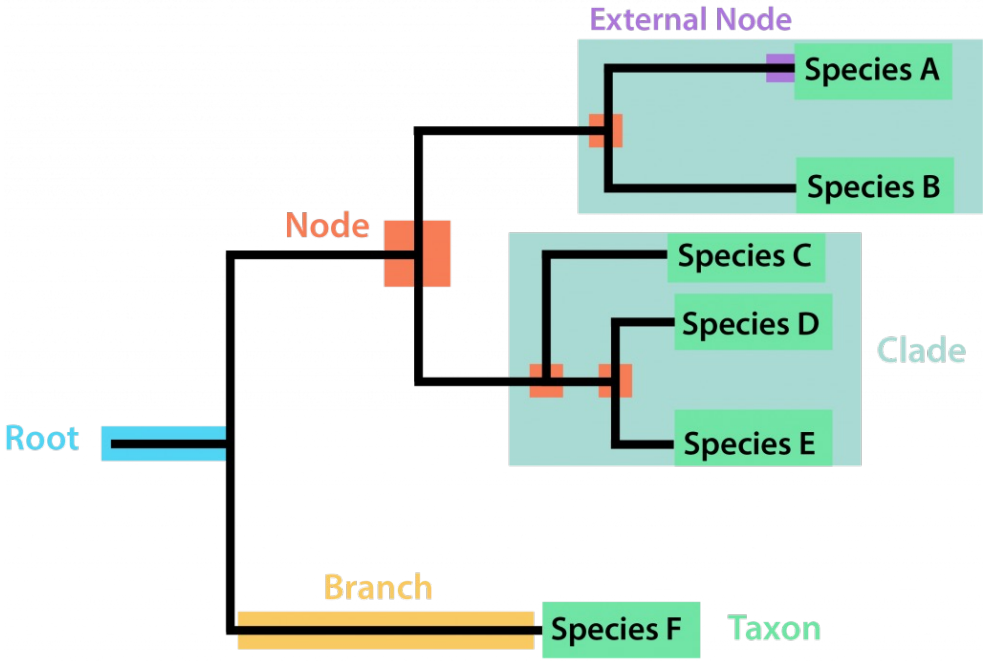
Wikipedia, 2022

Kettenburg et al., Frontiers in Public Health, 2022

Quick and dirty tree building in R, 2016

Anatomy of a phylogeny

Anatomie d'une phylogénie

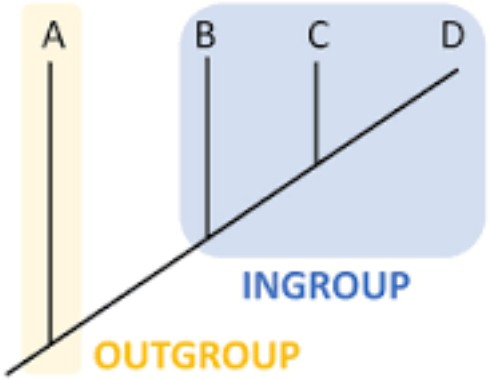


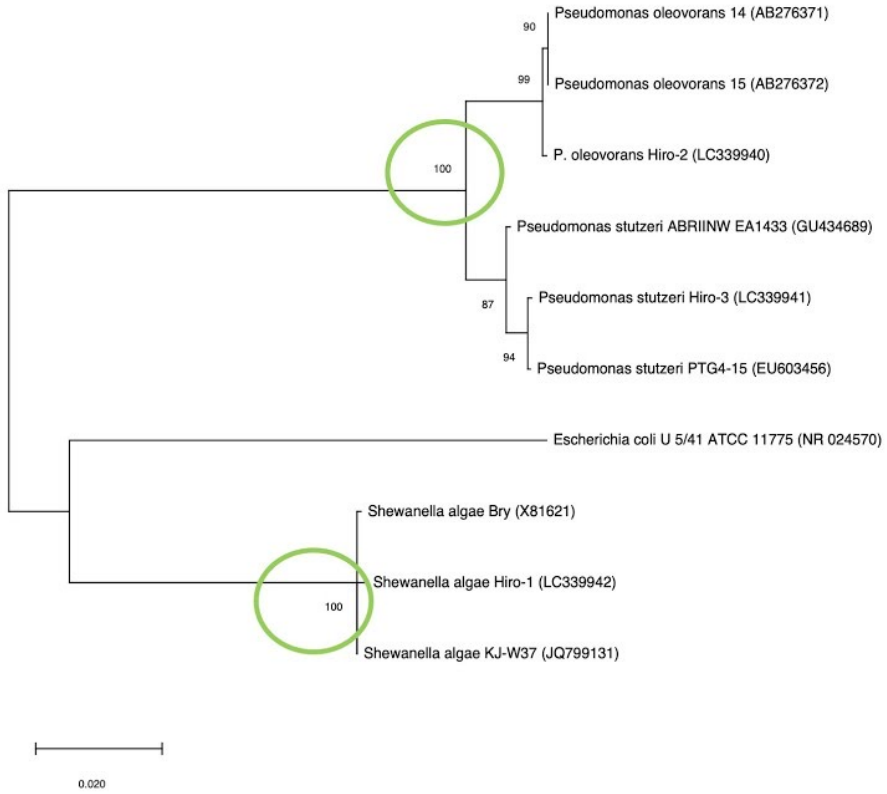
Phylogenetic Tree Structure

CONFIDENCE

BOOTSTRAP VALUE

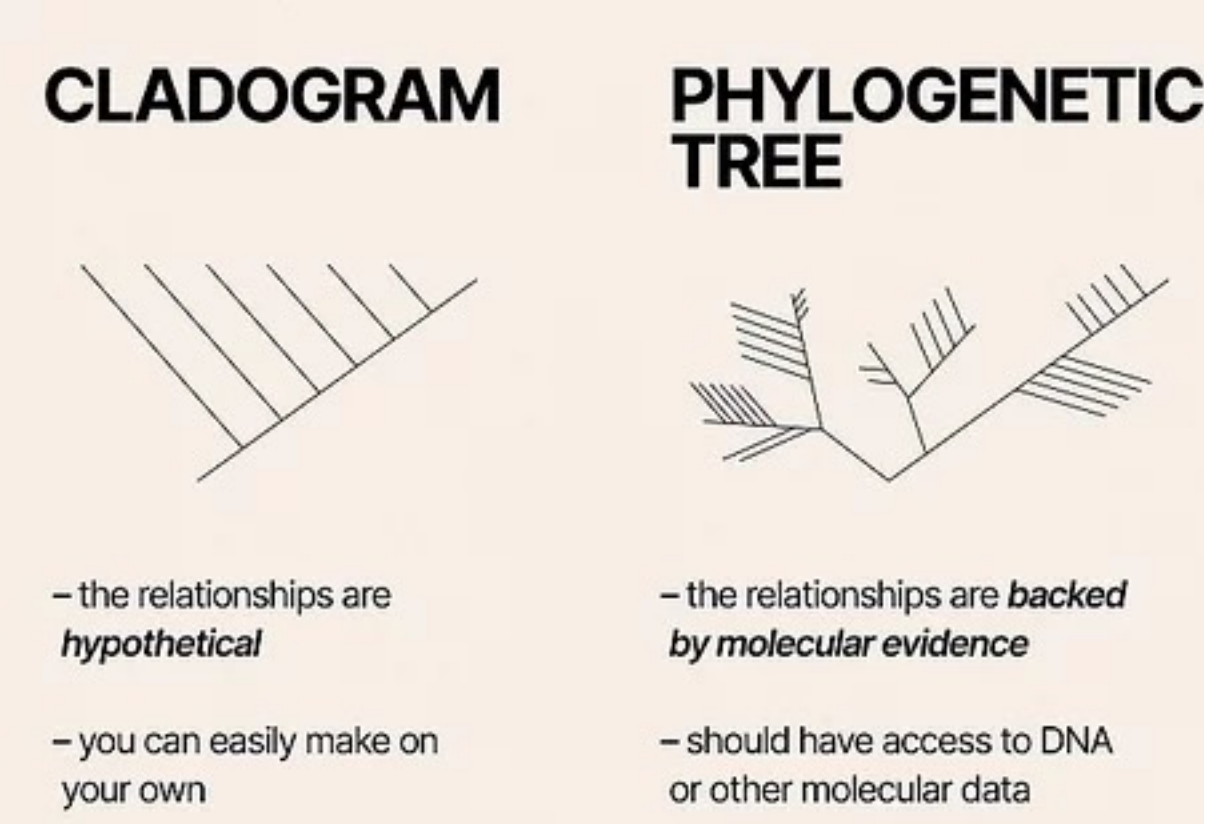
STRONGLY SUPPORTED	>90%
WELL SUPPORTED	70%-90%
WEAKLY SUPPORTED	50%-70%
NOT SUPPORTED	<50%

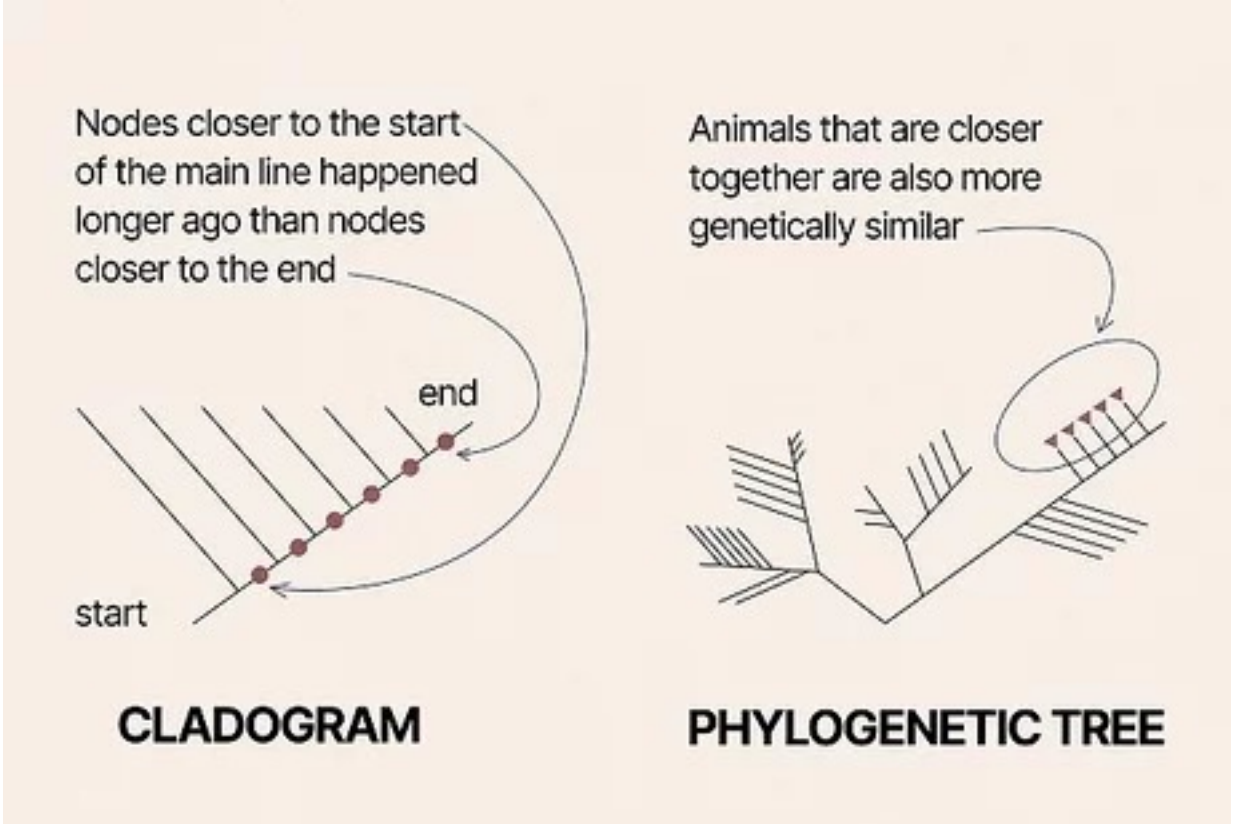




Cladogram versus phylogenetic tree

Cladogramme vs arbre phylogénétique





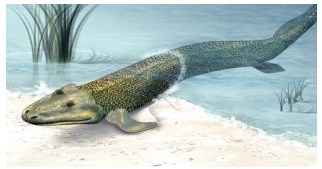
Wikipedia, 2022

Parsimony versus likelihood

Parcimonie vs vraisemblance

- Parsimony: minimum number of changes
  - *Parcimonie: nombre minimum de changements*
- Likelihood: maximum probability of the data having evolved on the tree
  - *Vraisemblance : probabilité maximale que les données aient évolué sur l'arbre*

Ancestor



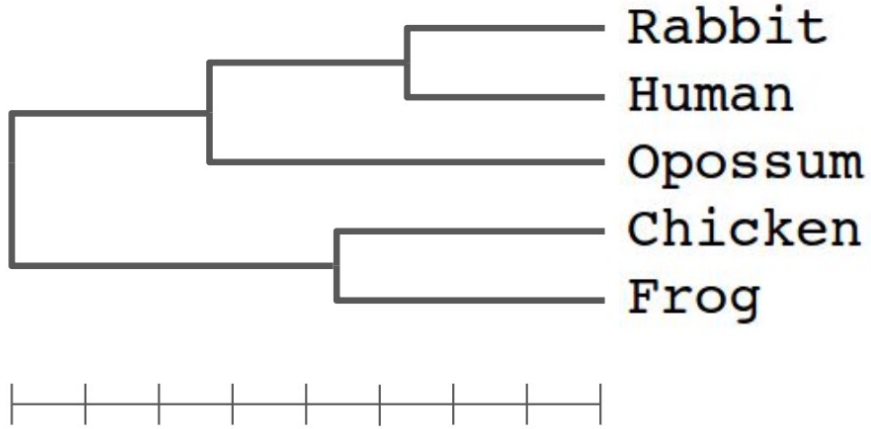
Descendant

branch length can mean different things:

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

La longueur de la branche peut signifier différentes choses

- *Nombre minimum de changements (parcimonie)*
- *Temps; opportunité de changement*
- *Nombre de changements attendus, compte tenu du modèle d'évolution*



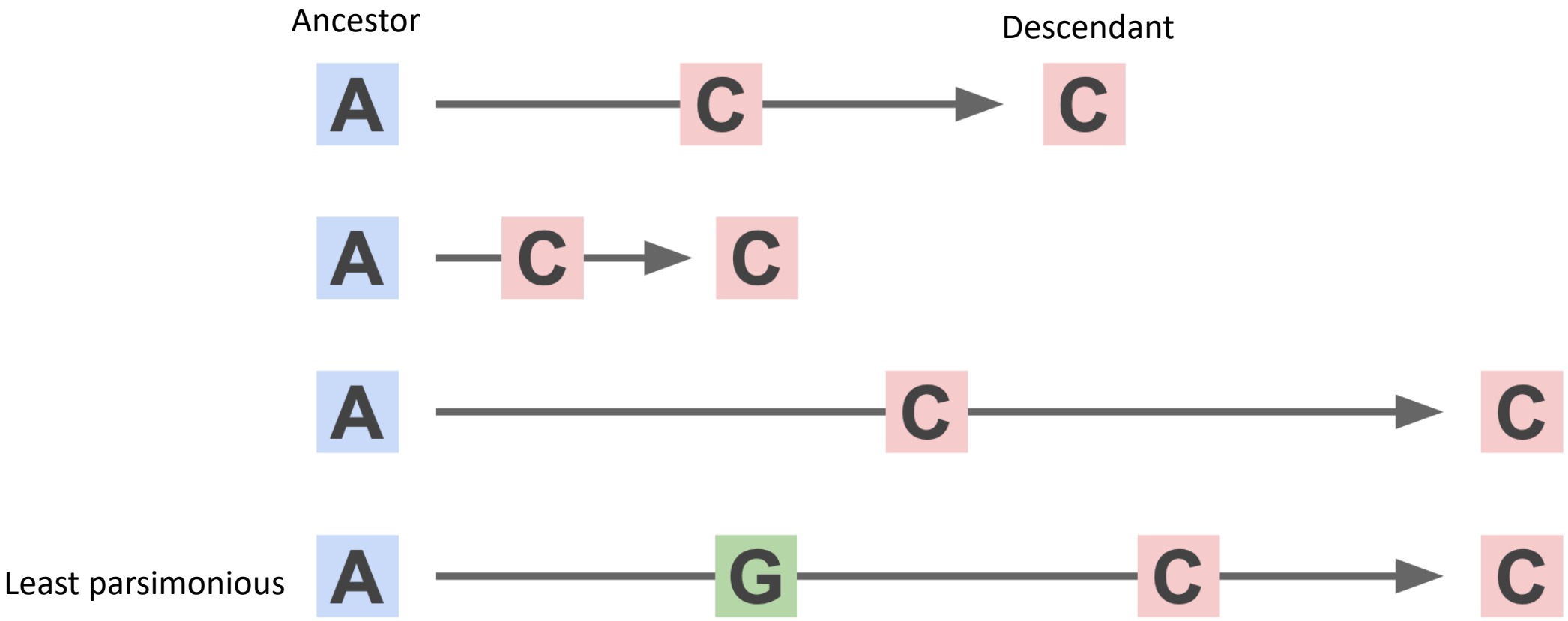
proposed tree has **branch lengths** in units of expected number of changes per site

L'arbre proposé a des longueurs de branche en unités du nombre prévu de changements par site

Adapted from Ree and Hipp, UChicago, 2021

Parsimony: minimum number of changes regardless of time/opportunity

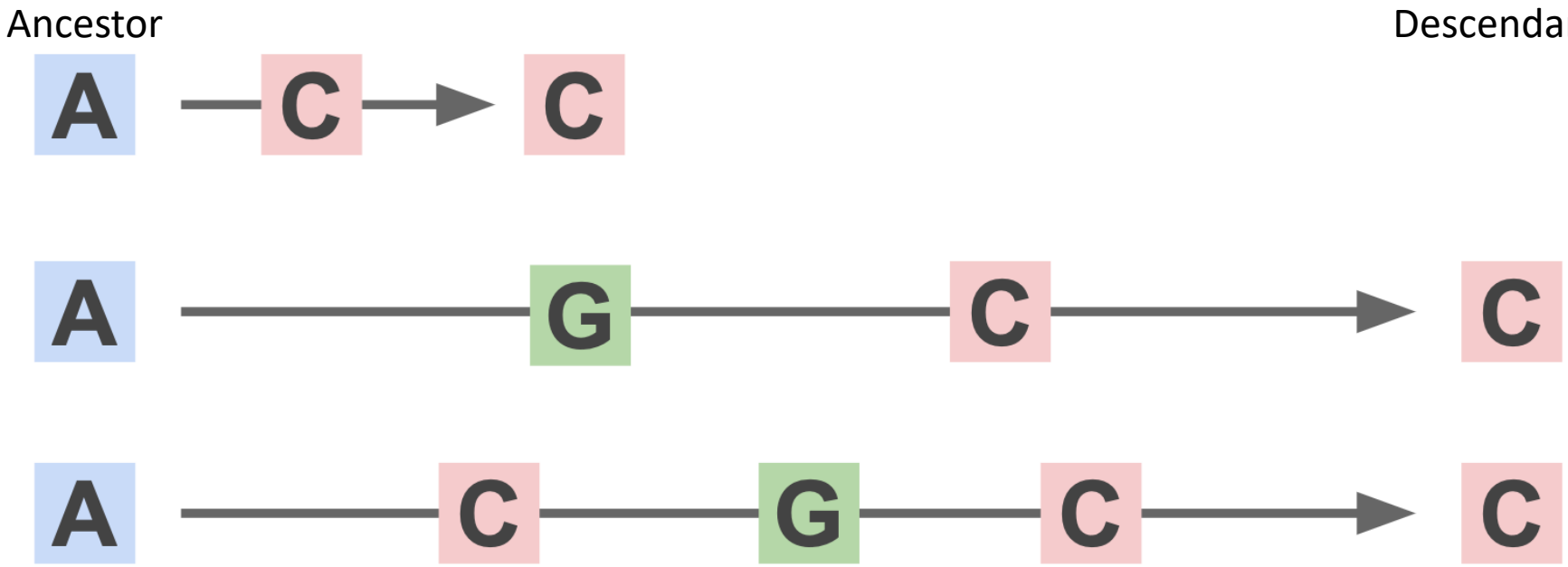
Parcimonie : nombre minimum de modifications quel que soit le moment/l'opportunité



Adapted from Ree and Hipp, UChicago, 2021

Likelihood: probability of ancestral and descendant status is a function of time (branch length)

Probabilité : la probabilité du statut ancestral et descendant est fonction du temps (longueur de la branche)



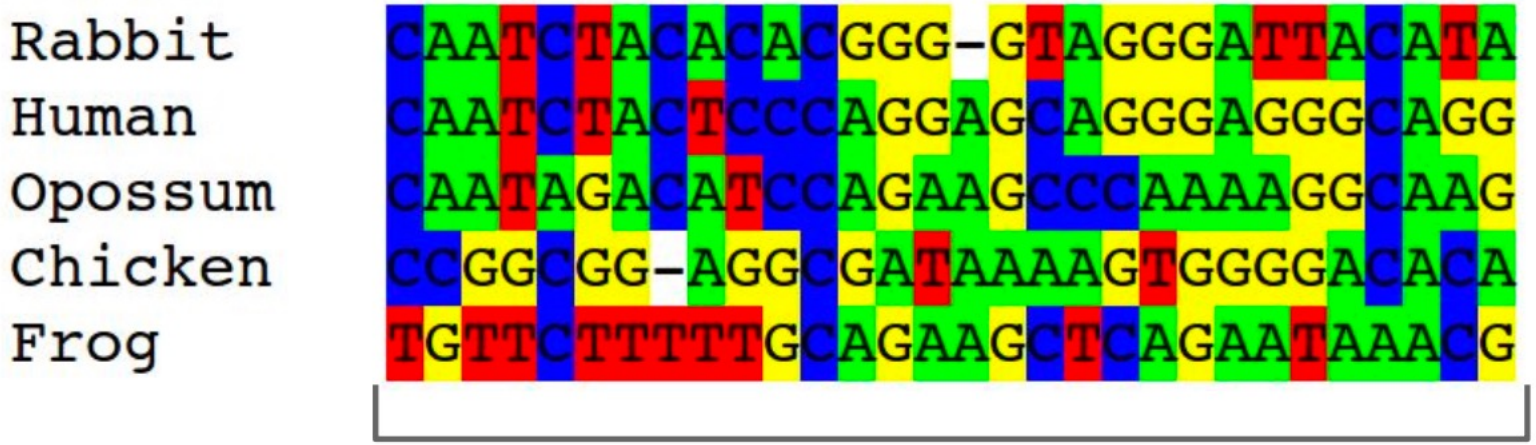
We don’t know what the actual history of the change is, so use a model of evolution to consider all possible histories

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

Adapted from Ree and Hipp, UChicago, 2021

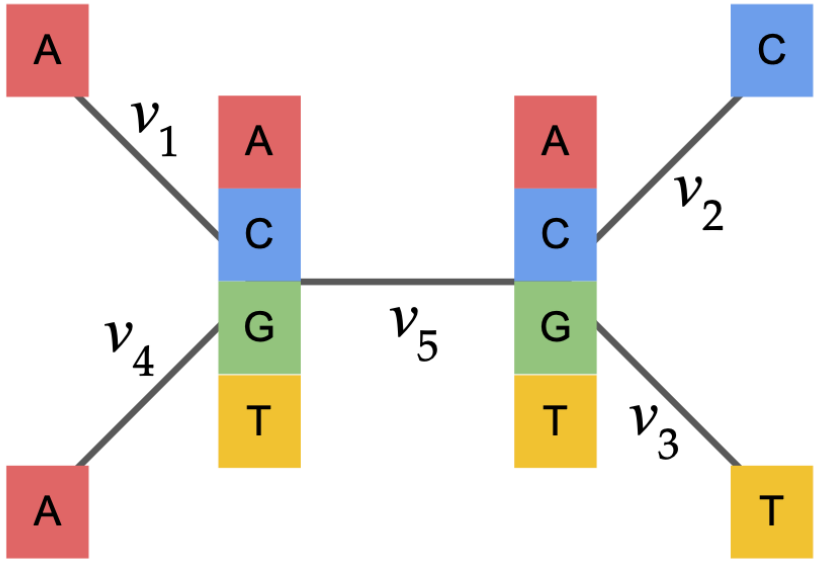
Likelihood cont'd.

Probabilité suite



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

La vraisemblance globale est le produit des vraisemblances entre les caractères (sites)



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

Considérez tous les états ancestraux possibles aux nœuds internes et calculez leur contribution à la probabilité globale

Parameters: tree topology, branch lengths, substitution rates estimated to maximize likelihood of data

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

Adapted from Ree and Hipp, UChicago, 2021

Models of DNA evolution

Modèles d'évolution de l'ADN

- Markov models that describe relative rates of different changes

  Modèles de Markov qui décrivent les taux relatifs de différents changements

  - JC69 (Jukes and Cantor 1969)
  - K80 model (Kimura 1980)
  - K81 model (Kimura 1981)
  - F81 (Felsenstein 1981)
  - HKY85 model (Hasegawa, Kishino and Yano 1985)
  - T92 model (Tamura 1992)
  - TN93 model (Tamura and Nei 1993)
  - GTR model (Tavaré 1986)
  - Yep there’s a lot of them!

Good news, most people don’t need to know the mathematical specifics of these models

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.

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

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

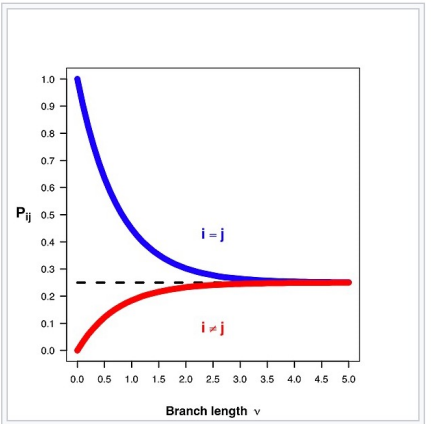
When branch length, ν, is measured in the expected number of changes per site then:

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

It is worth noticing that $\nu = \frac{3}{4}t\mu = (\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4})t$ 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

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



Probability P_{ij} of changing from initial state i to final state j as a function of the branch length (ν) for JC69. Red curve: nucleotide states i and j are different. Blue curve: initial and final states are the same. After a long time, probabilities tend to the nucleotide equilibrium frequencies (0.25: dashed line).

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

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.

IQ-TREE, 2022

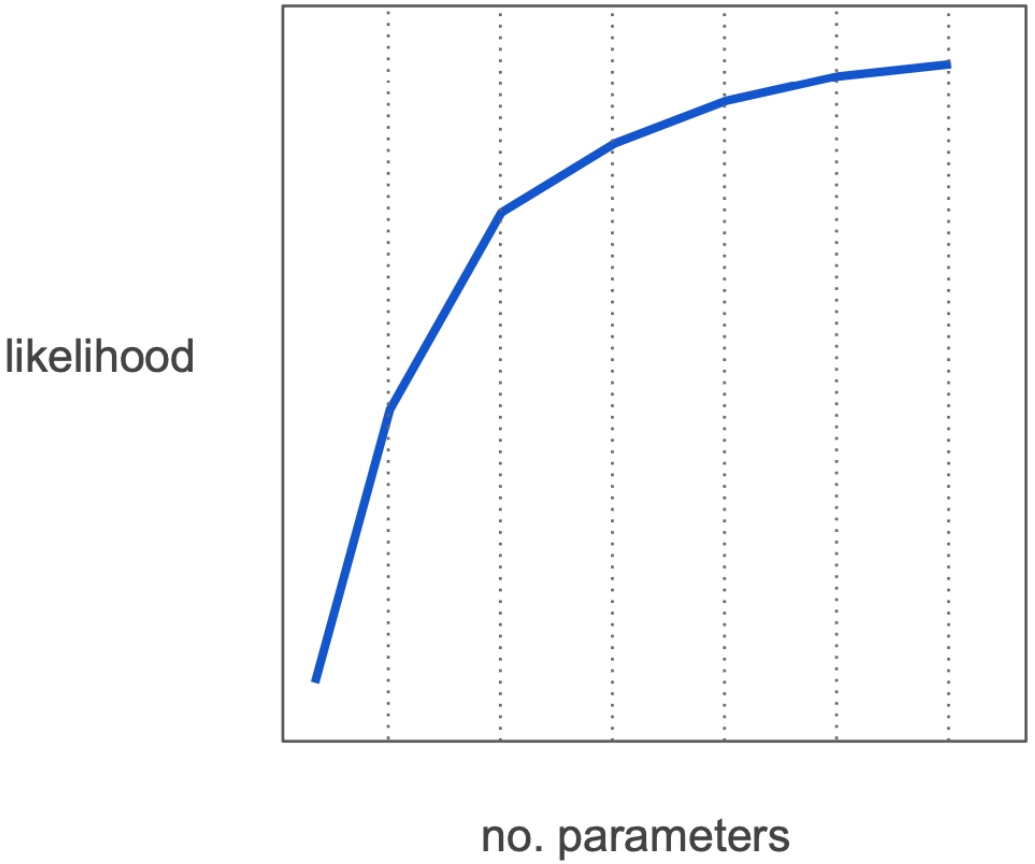
Model selection
Sélection du modèle

More parameters means higher likelihood, but is the increase in likelihood necessary? Adds much more complexity

- Programs will use statistical methods to answer this question using Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC), likelihood ratio tests, etc.

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

- *Les programmes utiliseront des méthodes statistiques pour répondre à cette question en utilisant les critères d'information d'Akaike (AIC), le critère d'information bayésien (BIC), les tests de rapport de vraisemblance, etc*



Model testing will give you BIC and AIC score

- AIC score: tries to select the model that most adequately describes an unknown, high dimensional reality
- BIC score: tries to find the TRUE model among the set of candidates

Les tests de modèles vous donneront un score BIC et AIC

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

Adapted from Ree and Hipp, UChicago, 2021

Rate heterogeneity across sites

Taux d'hétérogénéité entre les sites

- Do we expect all sites in an alignment to evolve at the same rate? What kind events would affect this?

Attendons-nous à ce que tous les sites d'un alignement évoluent au même rythme ? Quel genre d'événements affecterait cela?

Rate heterogeneity across sites

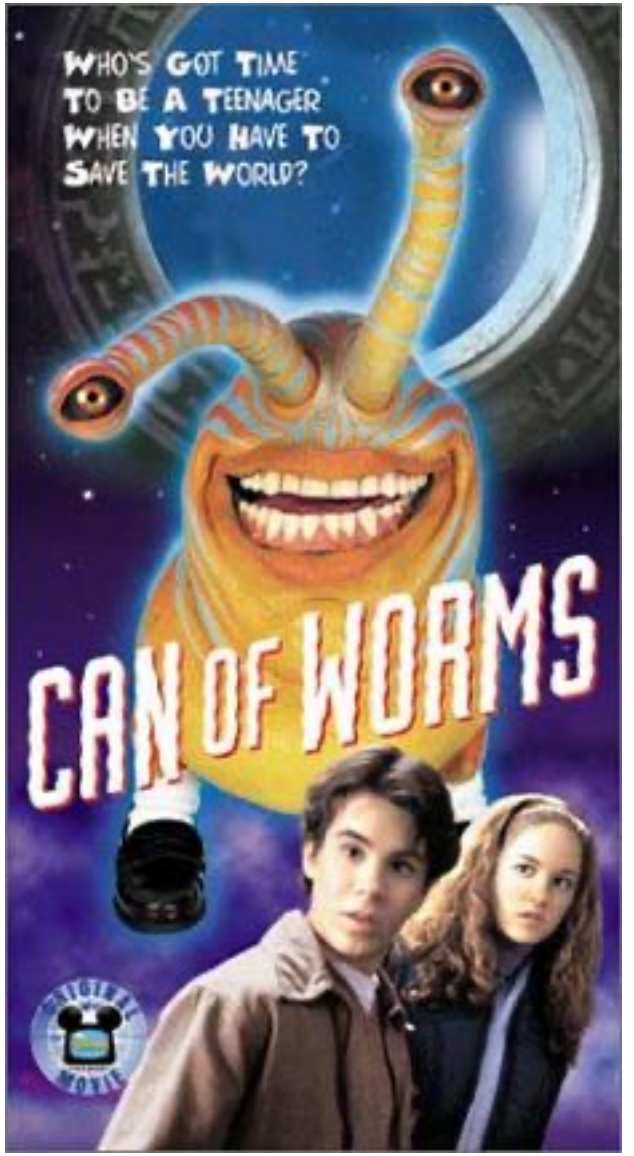
Taux d'hétérogénéité entre les sites

- Changes in rate heterogeneity:
  - Codon positions
  - Exons (coding regions) versus introns (non-coding regions)
  - Housekeeping genes versus non-functional genes
  - Structure in RNA (stems vs. loops)

  We can make inference about selection from these values, but that's another can of worms

- *Évolution de l'hétérogénéité des taux :*
  - *Positions des codons*
  - *Exons (régions codantes) versus introns (régions non codantes)*
  - *Gènes domestiques versus gènes non fonctionnels*
  - *Structure dans l'ARN (tiges vs boucles)*

  Nous pouvons faire des déductions sur la sélection à partir de ces valeurs, mais c'est une autre boîte de Pandore



Adapted from Ree and Hipp, UChicago, 2021

Bootstrapping

Amorçage

- Specify number of replicates: how many times does the test replicate the original sequence alignment?

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

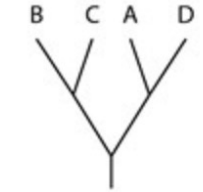
- Standard in MEGA is 500 replicates, 1000 is better but takes longer

La norme dans MEGA est de 500 répétitions, 1000 est mieux mais prend plus de temps

Original
sequence alignment

Site number

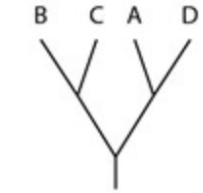
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Species A	A	A	T	G	C	T	A	G	T	G	G	T	G	A	T
Species B	A	A	G	C	T	A	T	G	G	T	G	A	T	C	G
Species C	A	G	C	C	T	A	T	G	T	G	G	A	A	C	G
Species D	A	A	C	C	C	A	T	T	G	G	G	T	G	A	T



Original alignment
tree

Bootstrap
pseudo-replicate #1

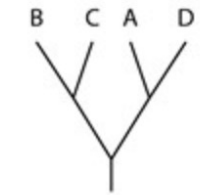
	5	3	3	1	12	9	2	4	11	13	10	14	8	11	13
Species A	C	T	T	A	T	T	A	G	G	G	G	A	G	G	G
Species B	T	G	G	A	A	G	A	C	G	T	T	C	G	G	T
Species C	T	C	C	A	A	T	G	C	G	A	G	C	G	G	A
Species D	C	C	C	A	T	T	A	C	G	G	G	A	T	G	G



Bootstrap tree #1

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

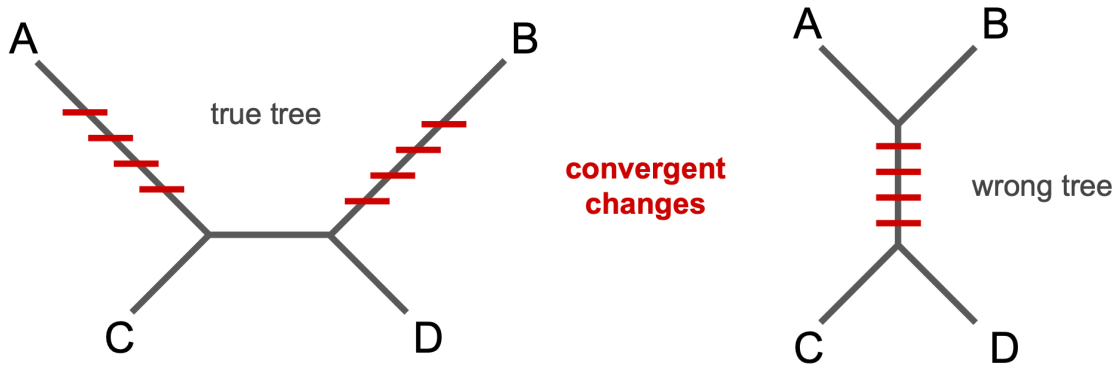


Bootstrap tree #2

Felsenstein zone

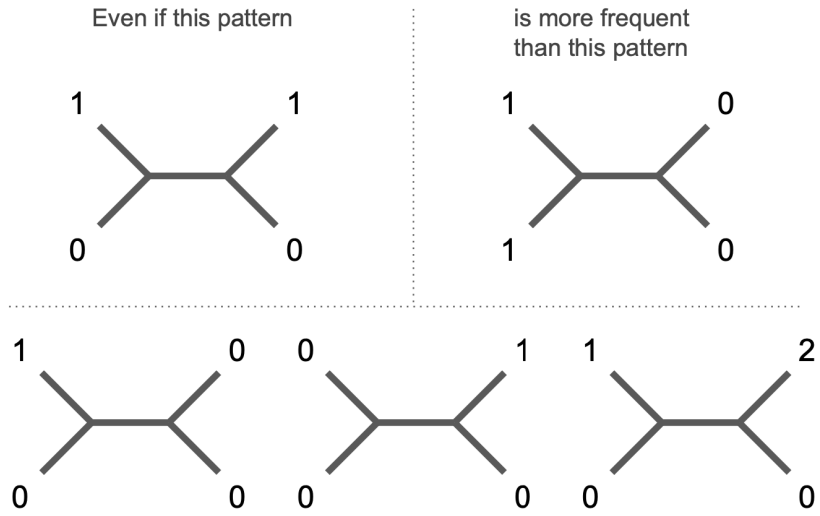
Zone de Felsenstein

- Branch lengths for which parsimony confidently infers the wrong topology, these can affect bootstrap values
  - *Longueurs de branche pour lesquelles la parcimonie déduit en toute confiance la mauvaise topologie, 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

la vraisemblance est un estimateur cohérent de la topologie arborescente car elle converge vers la valeur correcte avec l'augmentation des données



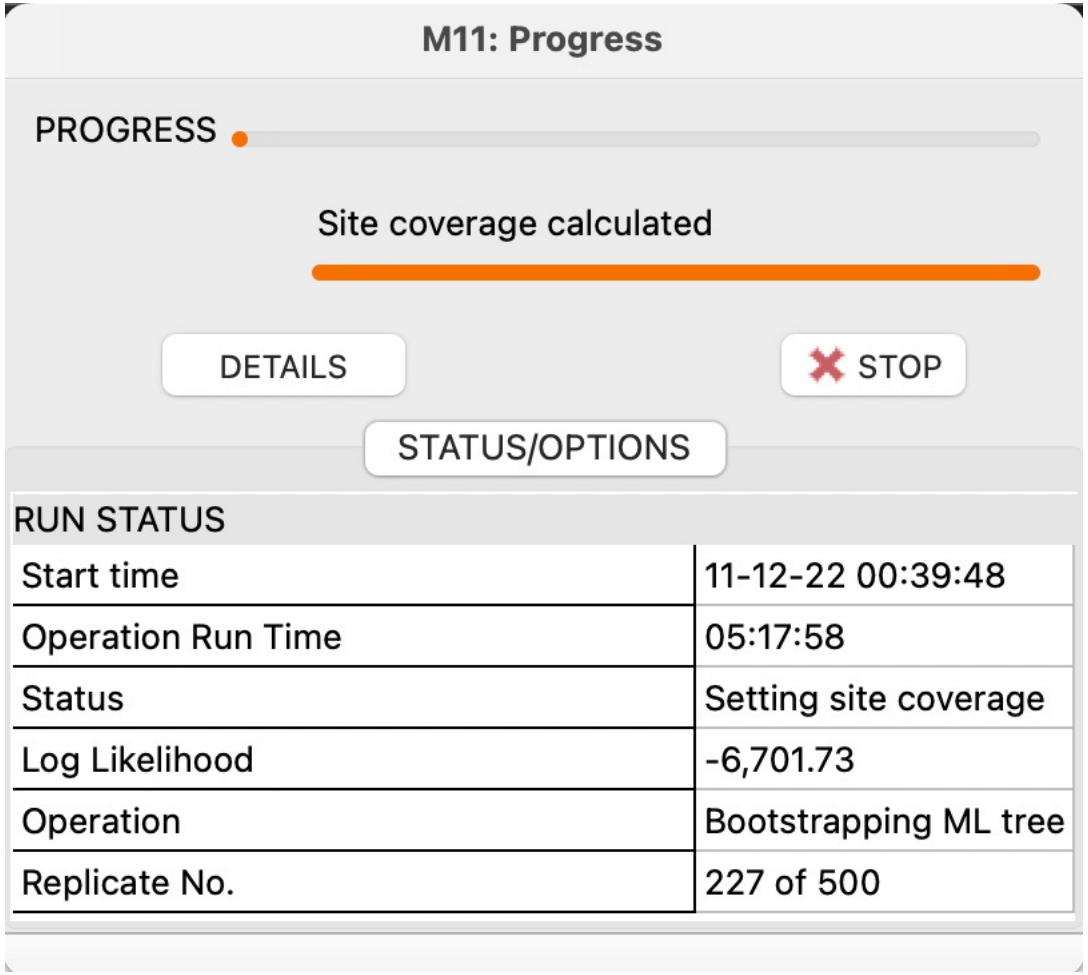
likelihood will correctly infer the true tree if these patterns are sufficiently frequent to allow accurate branch length estimation

Adapted from Ree and Hipp, UChicago, 2021

Warnings and limitations

Avertissements et limitations

- Building phylogenies takes a LONG time
  - Construire des phylogénies prend beaucoup de temps
- Without a proper outgroup or root, a phylogeny doesn't tell you much about order of descent
  - Sans groupe externe ou racine approprié, une phylogénie ne vous dit pas grand-chose sur l'ordre de descendance
- Does anyone know why it was so hard to make phylogenies for COVID-19?
  - Est-ce que quelqu'un sait pourquoi il était si difficile de faire des phylogénies pour COVID-19?



So you have your sequences, now what?

Alors vous avez vos séquences, et maintenant ?

- Get some reference sequences from NCBI
  Obtenez des séquences de référence du NCBI
- Get an outgroup from NCBI
  Obtenez un groupe externe de NCBI
- Align them (use a software like MEGA or online like MAFFT)
  Alignez-les (utilisez un logiciel comme MEGA ou en ligne comme MAFFT)
- Pick the best model (use a software like MEGA or ModelTest-NG)
  Choisissez le meilleur modèle (utilisez un logiciel comme MEGA ou ModelTest-NG)
- Run the phylogeny using your aligned sequences and chosen model (use a software like MEGA or RAxML)
  Exécutez la phylogénie en utilisant vos séquences alignées et le modèle choisi (utilisez un logiciel comme MEGA ou RAxML)
- Visualize/edit tree in either R or FigTree
  Visualiser/modifier l'arborescence dans R ou FigTree

All of this listed is free to use ☺

Tout ce qui est listé est libre d'utilisation

TUTORIAL

DIDACTICIEL

- Please make sure MEGA opens on your computer and you have the sequences downloaded from the syllabus online, let me know if you don’t have either of these!

  Veuillez vous assurer que MEGA s'ouvre sur votre ordinateur et que vous avez téléchargé les séquences à partir du programme en ligne, faites-moi savoir si vous n'avez ni l'un ni l'autre !

- Also have FigTree downloaded, or just follow along by watching’

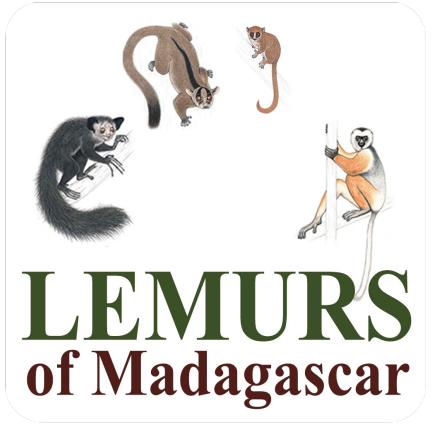
  Téléchargez également FigTree ou suivez simplement en regardant

- Additionally, open internet and go to NCBI.gov...if this does not open just follow along by watching

  De plus, ouvrez Internet et allez sur NCBI.gov... si cela ne s'ouvre pas, suivez simplement en regardant

Lemurs of Ranomafana national park

Parc national des lémuriens de Ranomafana



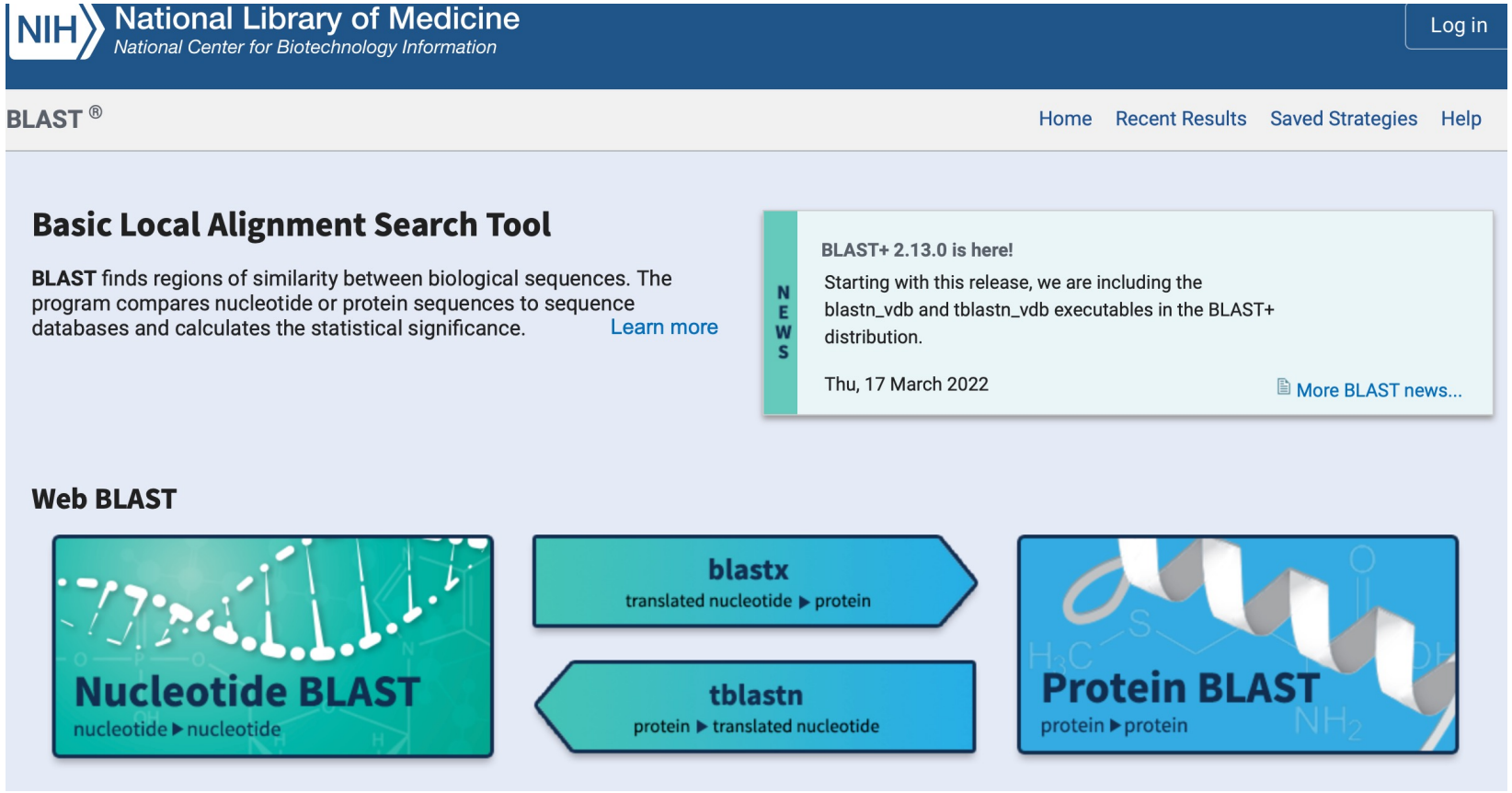
- Cytochrome B
  - Used a lot in species identification, limited variability within and much greater variation between species

    Beaucoup utilisé dans l'identification des espèces, variabilité limitée au sein et variation beaucoup plus grande entre les espèces
- Prompt: You are a lemur researcher sampling feces to see if there is a new mouse lemur species that lives in the park but has not been spotted…you have a sequence and want to see how genetically related it is to other lemur species that reside in the park

  Invite : Vous êtes un chercheur sur les lémuriens qui prélève des excréments pour voir s'il existe une nouvelle espèce de lémurien souris qui vit dans le parc mais qui n'a pas été repérée... vous avez une séquence et souhaitez voir à quel point elle est génétiquement liée à d'autres espèces de lémuriens qui résident dans le parc

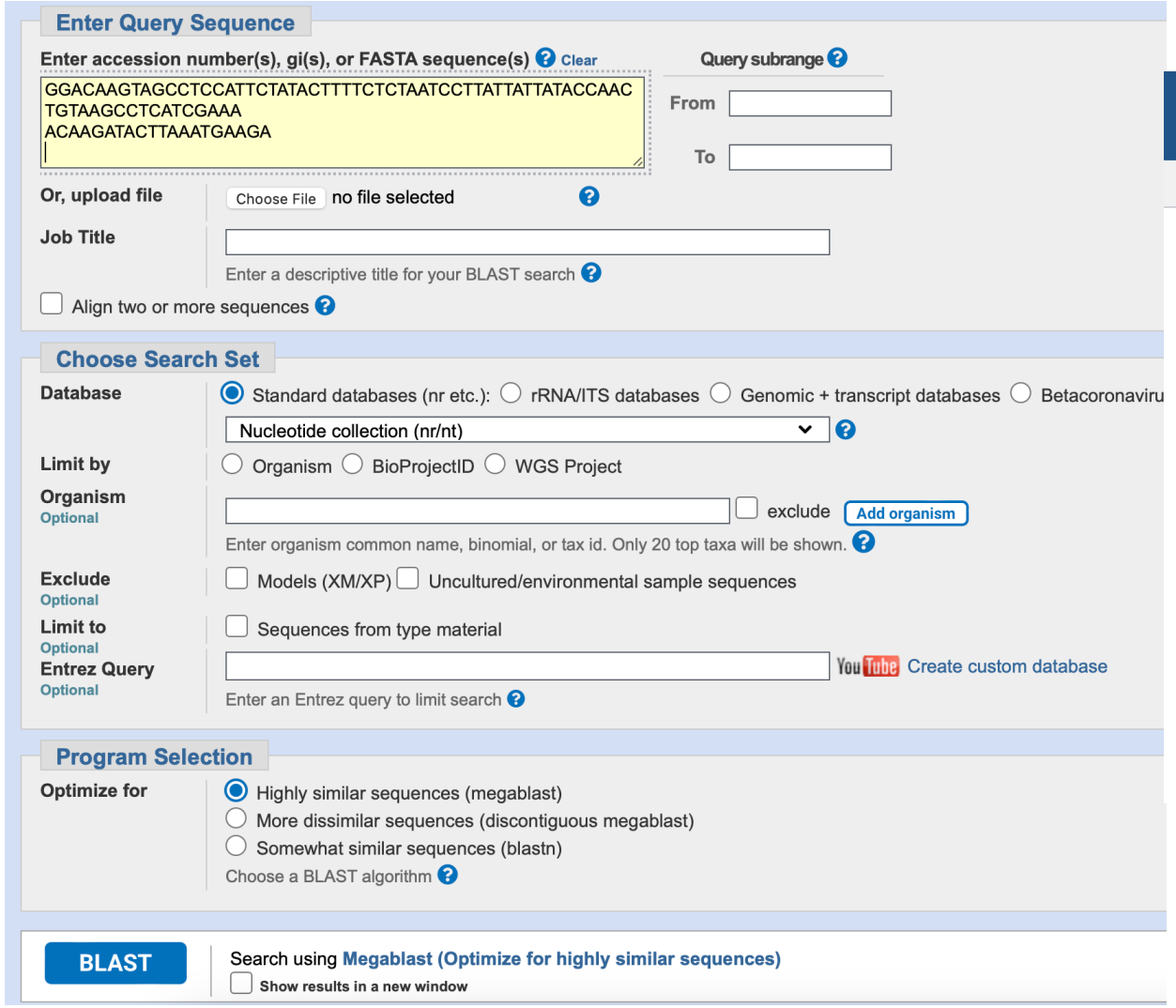
Steps to revisit later

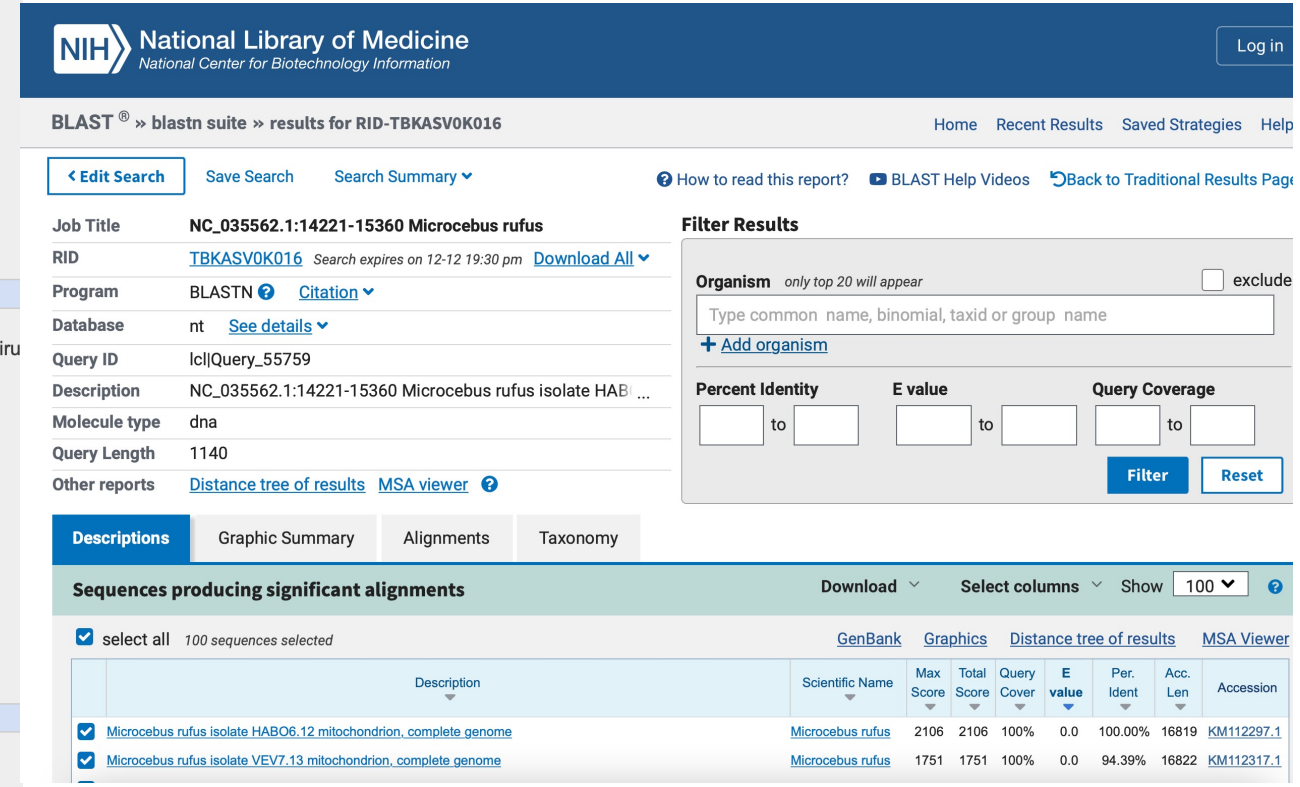
Étapes à revoir plus tard

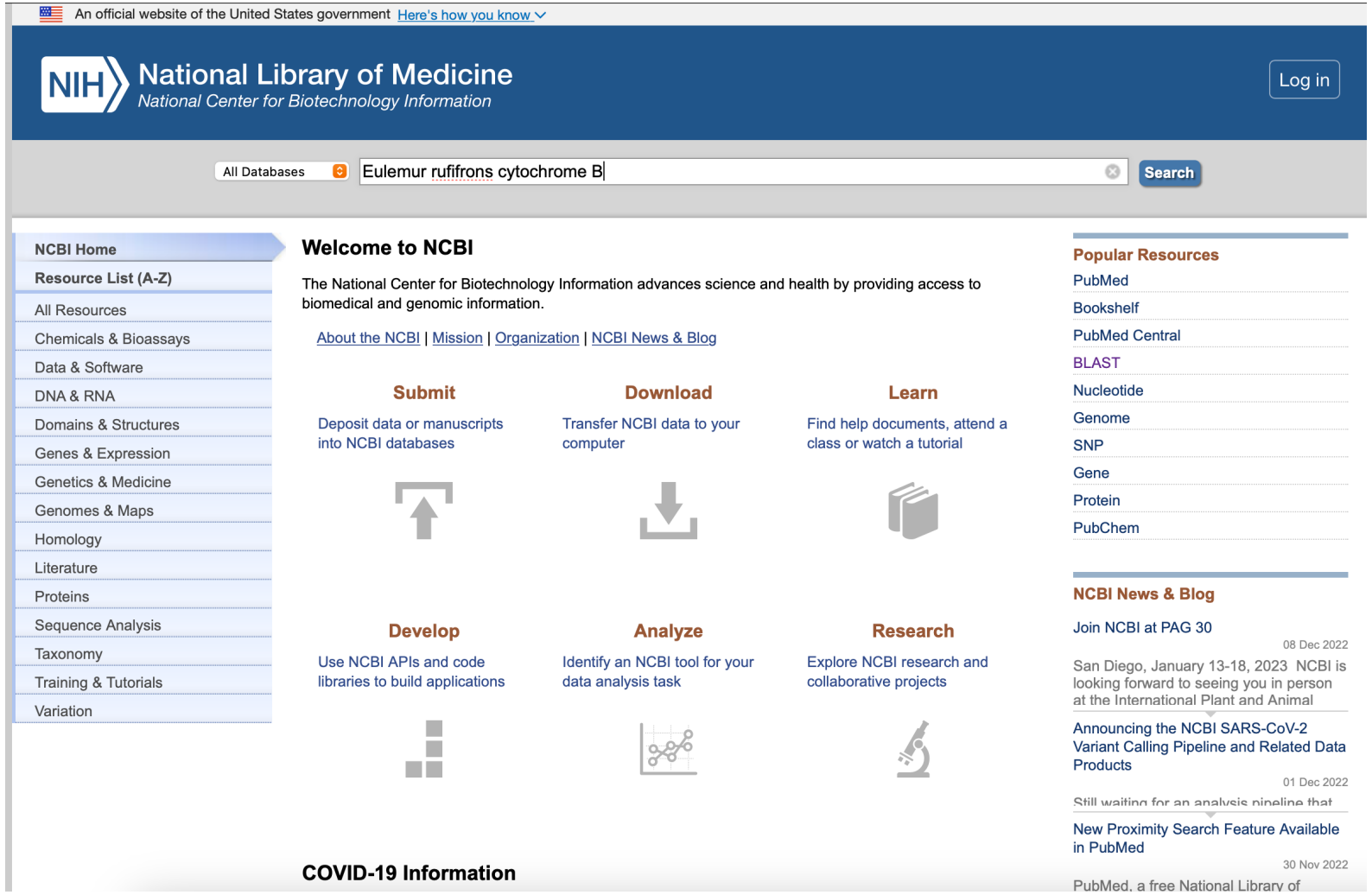


Check what kind of sequence you are dealing with by doing a BLAST search

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

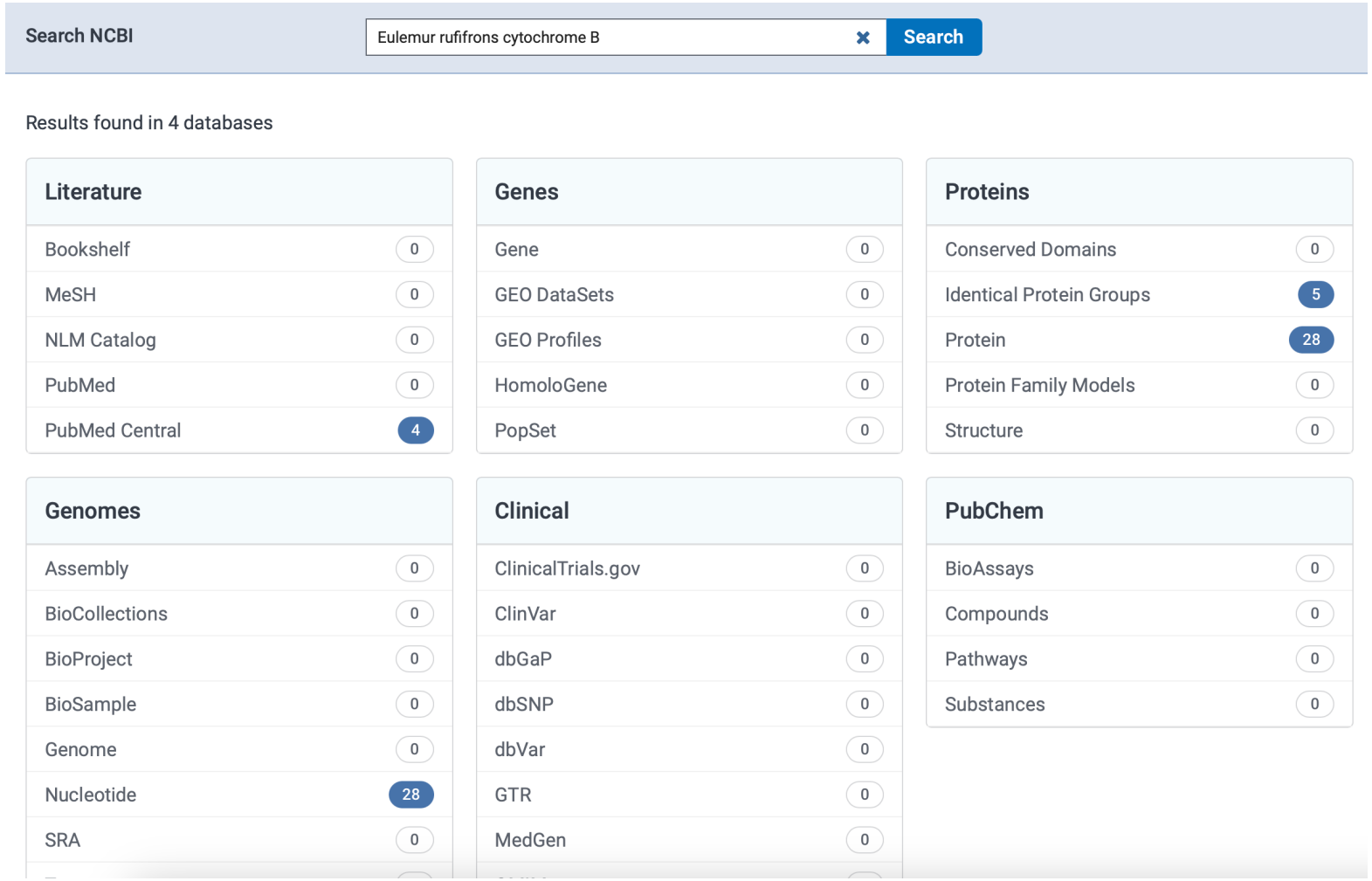

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA
Query subrange
From
To
Or, upload file
Choose File no file selected
Job Title
Enter a descriptive title for your BLAST search
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
Organism
Optional
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP) Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
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


National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastn suite » results for RID-TBKASV0K016
Home Recent Results Saved Strategies Help
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
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
Microcebus rufus isolate HABO6.12 mitochondrion, complete genome
Microcebus rufus
2106 2106 100% 0.0 100.00% 16819 KM112297.1
Microcebus rufus isolate VEV7.13 mitochondrion, complete genome
Microcebus rufus
1751 1751 100% 0.0 94.39% 16822 KM112317.1



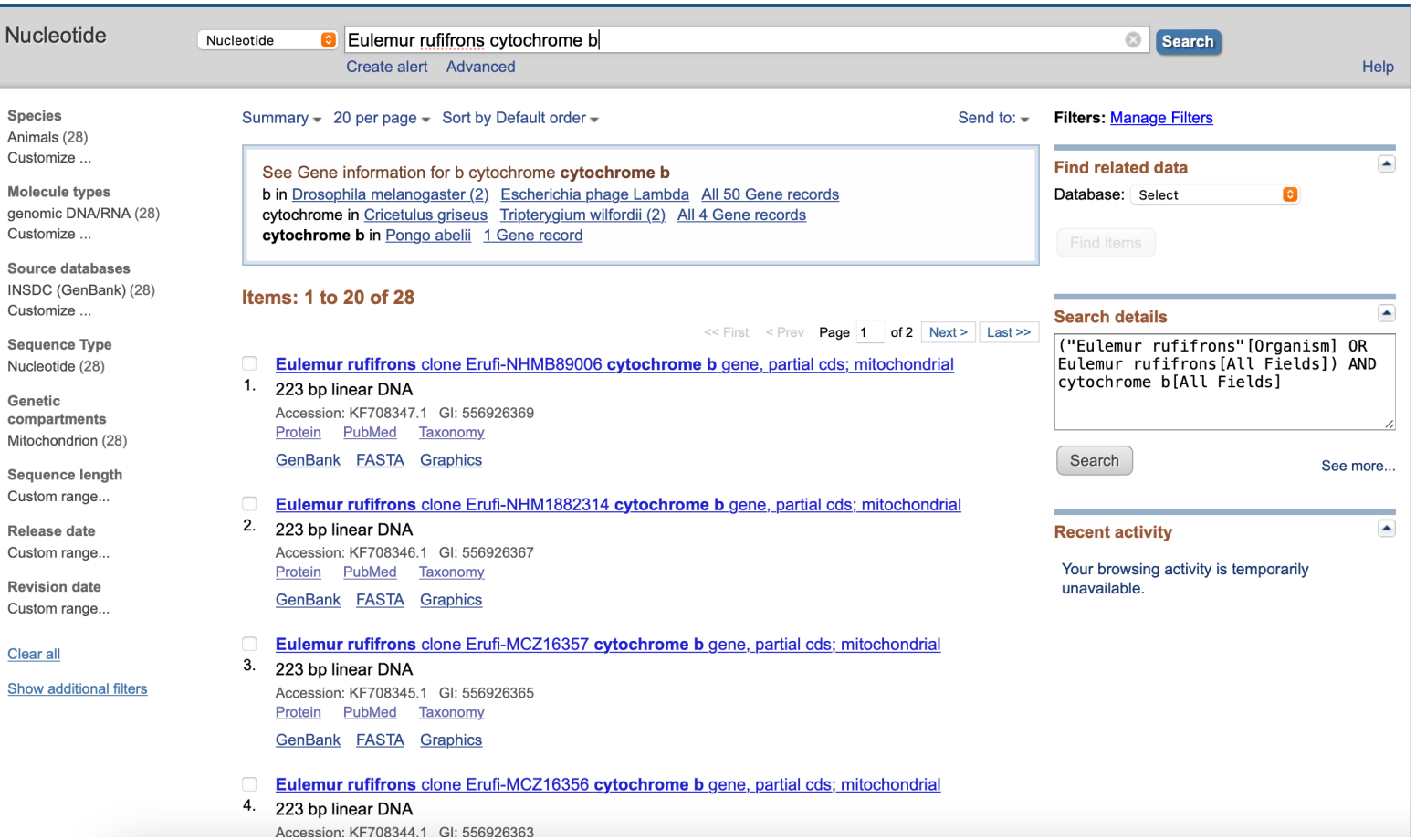
1. Go to NCBI, and search for the thing you want to build a phylogeny for, in our case cytochrome B of lemurs in Ranomafana national park

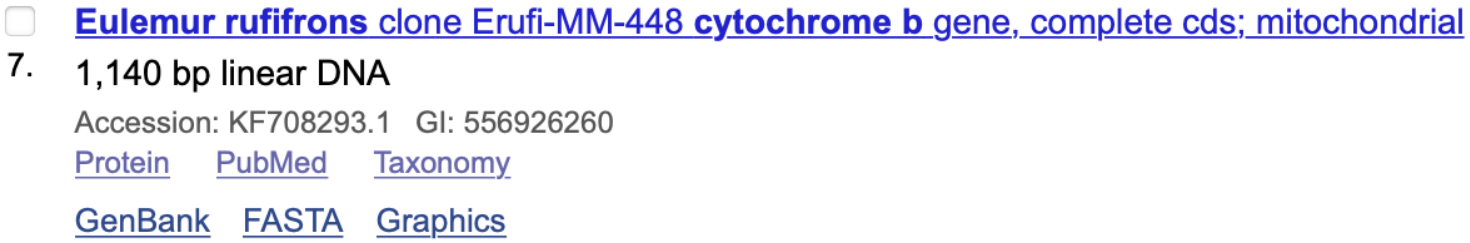
1. Allez sur NCBI, et recherchez la chose pour laquelle vous voulez construire une phylogénie, dans notre cas le cytochrome B des lémuriens du parc national de Ranomafana



This is what it will look like, you can go to Nucleotide under the genome category and click on that

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





Pick the sequence of what you're interested in, in our case we want a complete cds

We might want partial cds if we have a partial sequence of interest, but right now we're just building a tree with known data, so complete cds is best

Cds: protein coding sequence

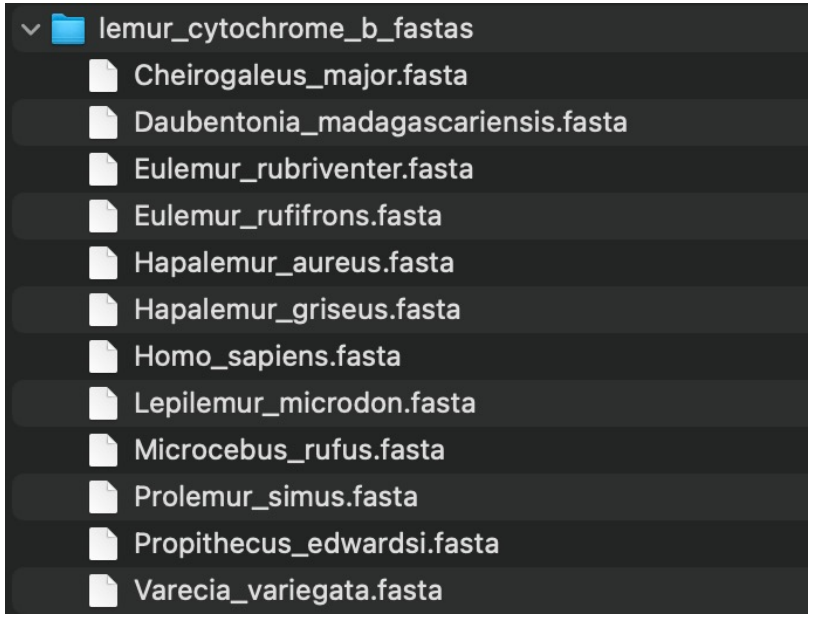
Then download the fastas

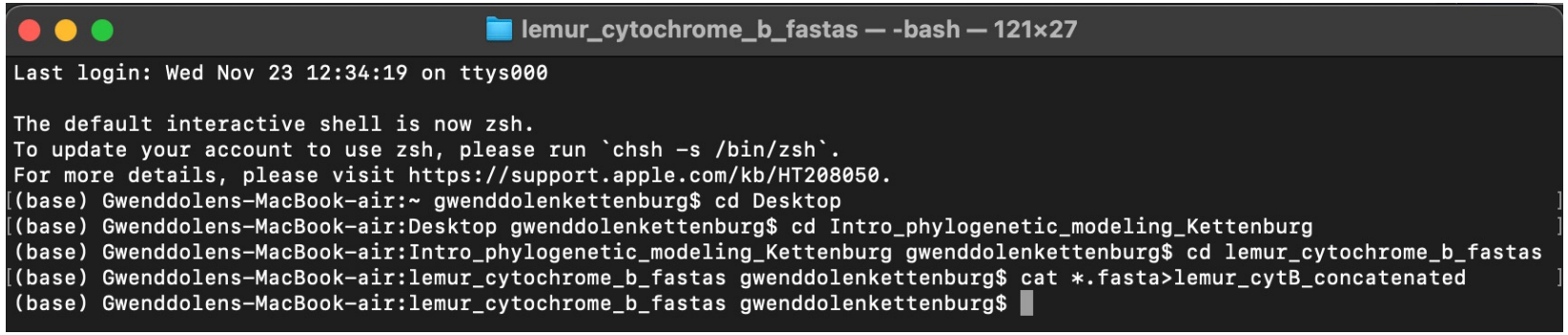
Choisissez la séquence de ce qui vous intéresse, dans notre cas nous voulons un cd complet

Nous pourrions vouloir des CD partiels si nous avons une séquence partielle d'intérêt, mais pour le moment, nous construisons simplement un arbre avec des données connues, donc les CD complets sont préférables

Cds : séquence codant pour la protéine

Téléchargez ensuite les fastas





Step 2: when you have all your sequences of interest and your outgroup, you need to concatenate the sequences into one file, you can do this by making a text/edit file and pasting each sequence in, otherwise follow instructions on command line (mac) or powershell (windows) to do this

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

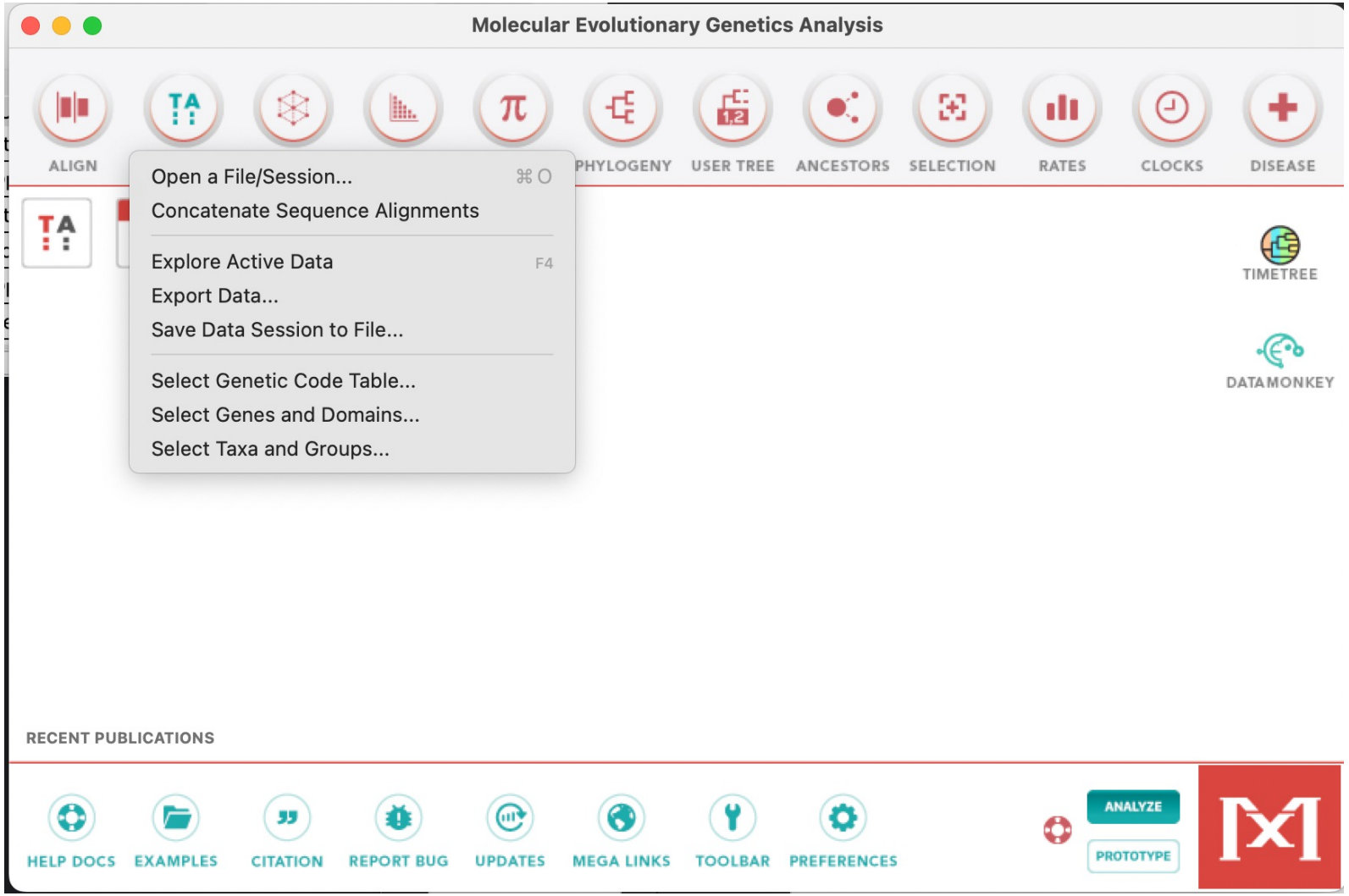
```
type file1.csv file2.csv > concat.csv
```

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
```

Étape 2 : lorsque vous avez toutes vos séquences d'intérêt et votre groupe externe, vous devez concaténer les séquences dans 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 ce faire



Step 3: open MEGA, and open a file/session, select your concatenated fasta file

MEGA will ask if you want to align or analyze, click on align

Étape 3 : ouvrez MEGA, et ouvrez un fichier/session, sélectionnez votre fichier fasta concaténé

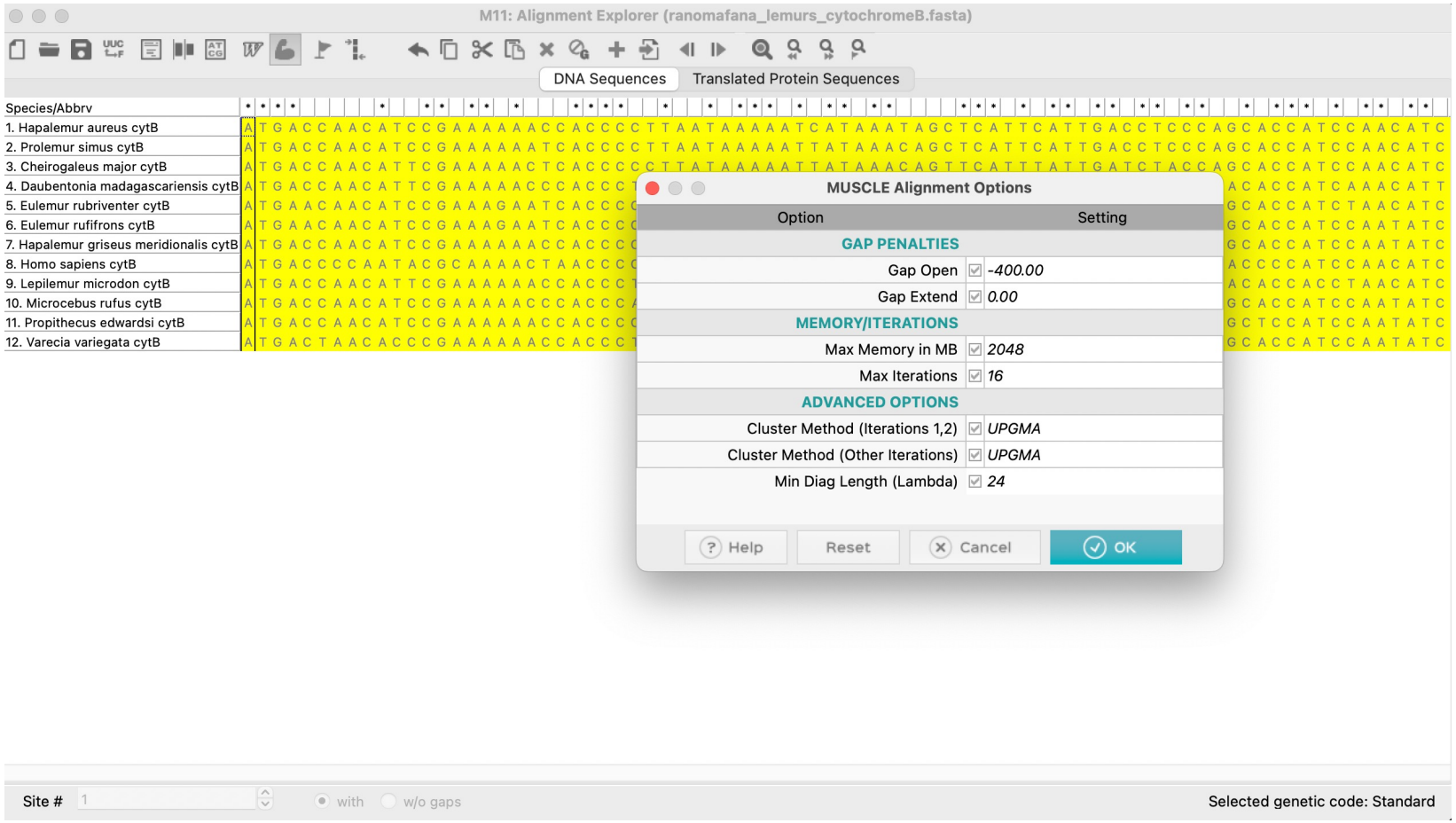
MEGA vous demandera si vous voulez aligner ou analyser, cliquez sur aligner

So the sequences are loaded into MEGA like this:

Step 4: click on the muscle arm to align with MUSCLE program

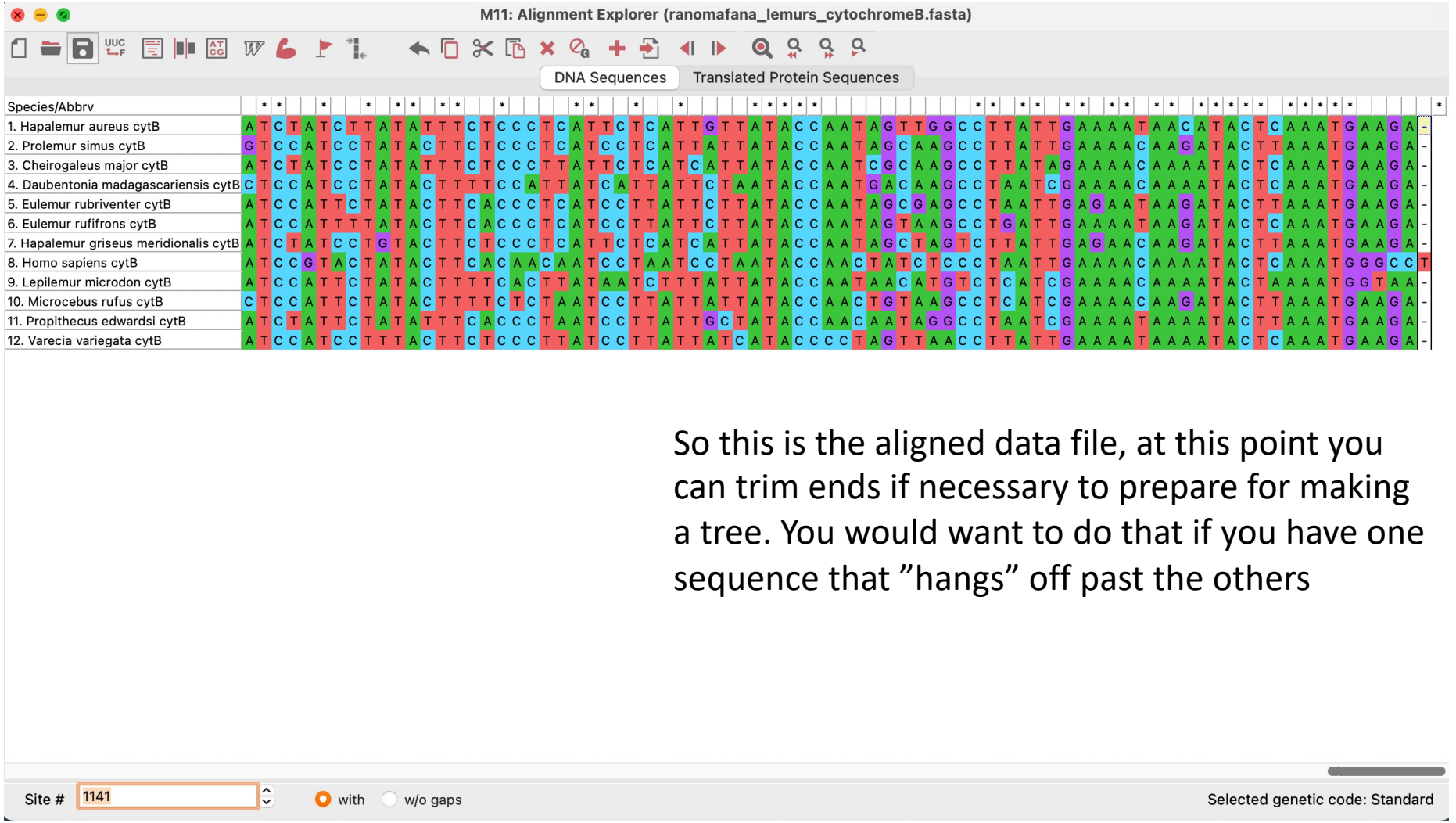
Ainsi les séquences sont chargées dans MEGA comme ceci :

Étape 4 : cliquez sur le bras musculaire pour l'aligner avec le programme MUSCLE



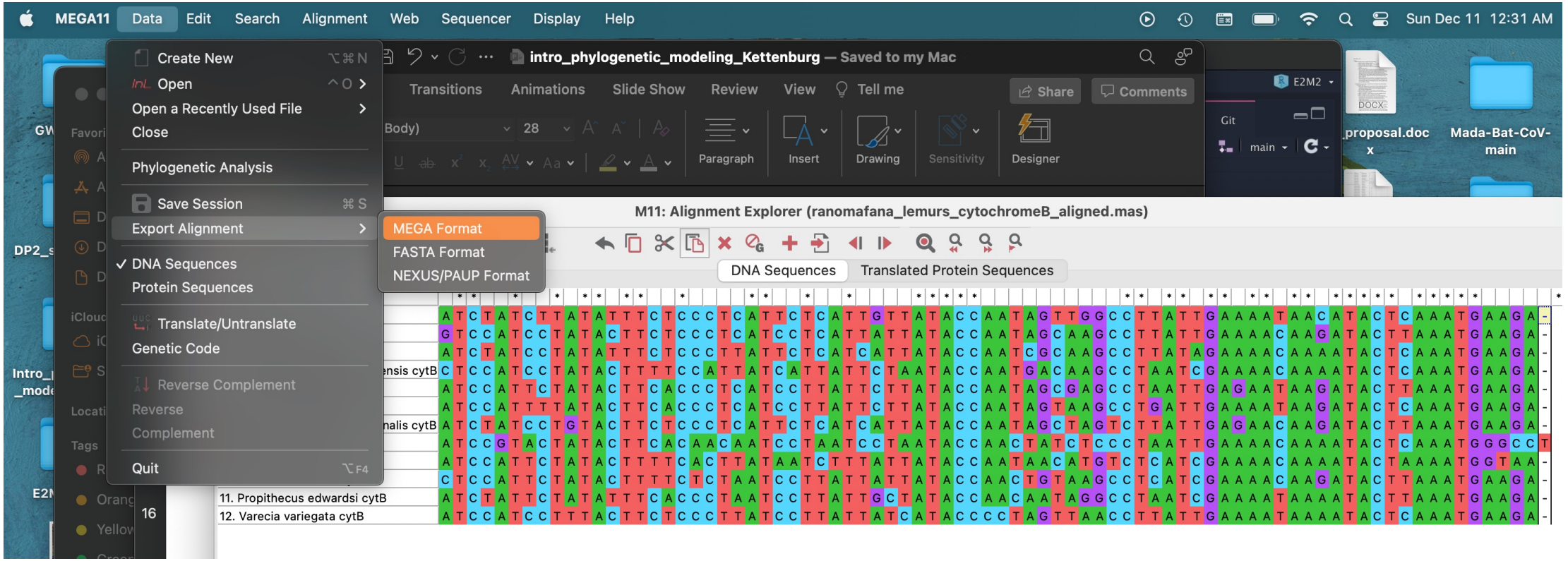
Go with suggested options, then hit okay

Allez avec les options suggérées, puis appuyez sur OK



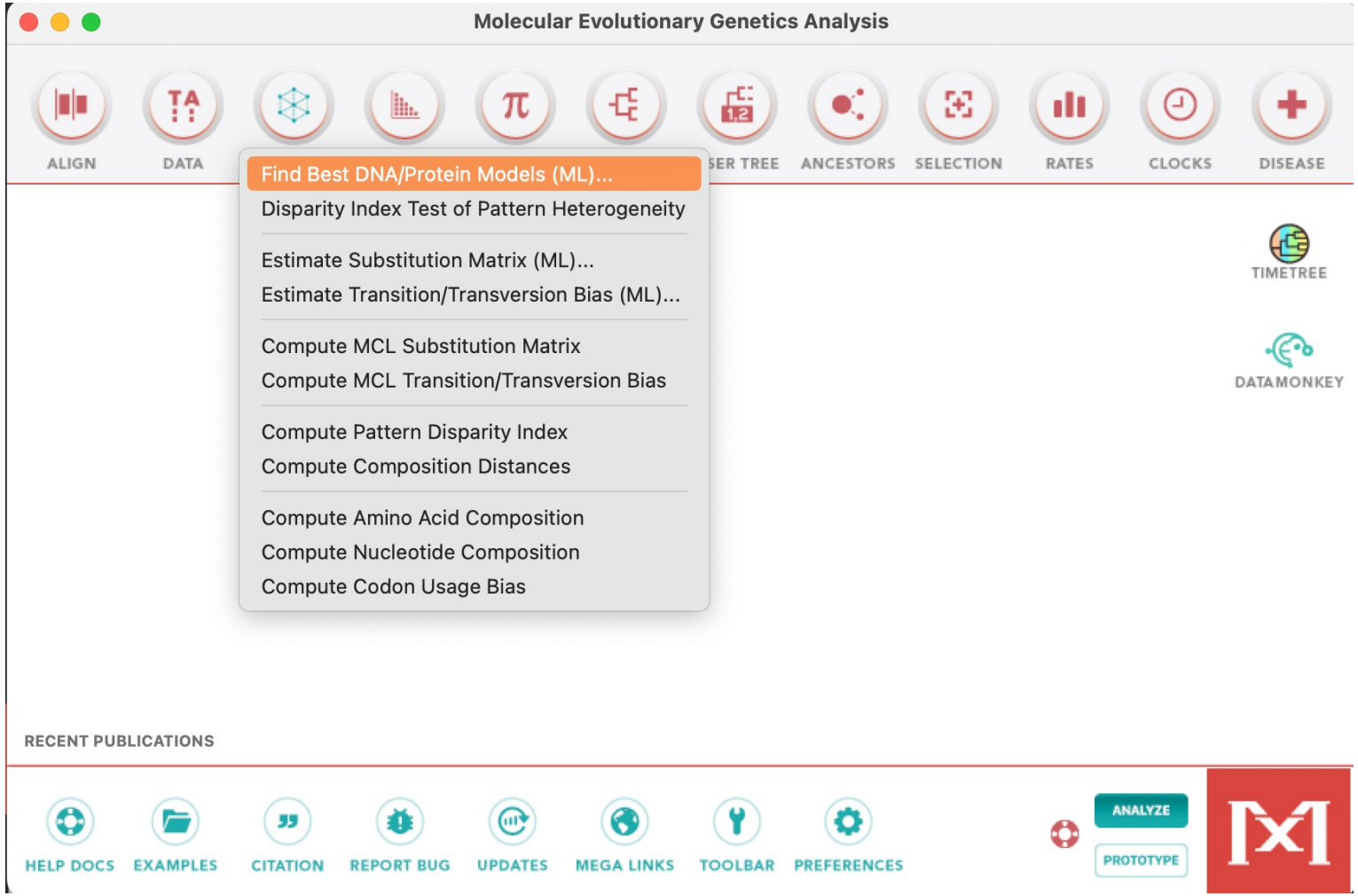
So this is the aligned data file, at this point you can trim ends if necessary to prepare for making a tree. You would want to do that if you have one sequence that ”hangs” off past the others

Voici donc le fichier de données aligné, à ce stade, vous pouvez couper les extrémités si nécessaire pour préparer la création d'un arbre. Vous voudriez faire cela si vous avez une séquence qui "se bloque" après les autres



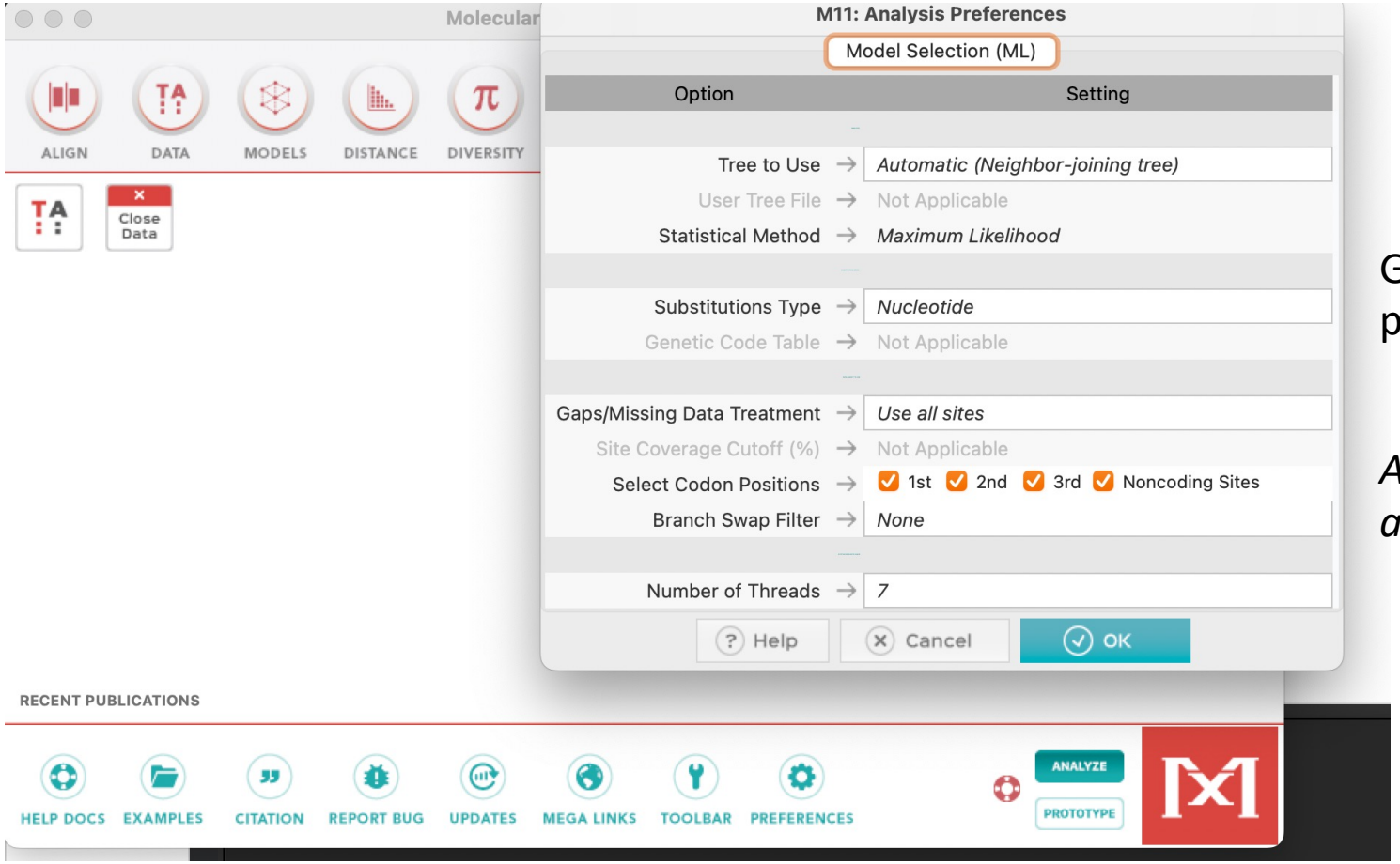
Save the aligned file, then we will proceed to model selection

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



Step 5: click on ”Models” and select “Find best DNA/Protein models ML)”

Étape 5 : cliquez sur « Modèles » et sélectionnez « Trouver les meilleurs modèles d'ADN/protéines ML) »



Go with the suggested parameters, then click okay

Allez avec les options suggérées, puis appuyez sur OK

MEGA Caption Expert: Find Best-Fit Substitution Model (ML)

Results

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)
GTR+G+I	31	14046.342	13813.251	-6875.553	0.37	0.81	5.67	0.292	0.287	0.295	0.125	0.025	0.041	0.048	0.025	0.352	0.002	0.041
GTR+G	30	14054.041	13828.464	-6884.164	n/a	0.28	5.81	0.292	0.287	0.295	0.125	0.023	0.041	0.048	0.023	0.354	0.002	0.041
TN93+G+I	28	14080.384	13869.838	-6906.859	0.37	0.78	5.80	0.292	0.287	0.295	0.125	0.023	0.024	0.048	0.023	0.344	0.010	0.023
TN93+G	27	14086.626	13883.595	-6914.742	n/a	0.27	6.02	0.292	0.287	0.295	0.125	0.022	0.023	0.047	0.023	0.348	0.010	0.023
HKY+G+I	27	14131.860	13928.829	-6937.359	0.40	0.91	4.52	0.292	0.287	0.295	0.125	0.026	0.027	0.103	0.026	0.242	0.011	0.026
HKY+G	26	14137.705	13942.190	-6945.044	n/a	0.28	4.58	0.292	0.287	0.295	0.125	0.026	0.026	0.103	0.026	0.243	0.011	0.026
GTR+I	30	14160.325	13934.749	-6937.306	0.49	n/a	3.67	0.292	0.287	0.295	0.125	0.037	0.053	0.065	0.037	0.295	0.005	0.052
TN93+I	27	14200.028	13996.998	-6971.443	0.49	n/a	3.56	0.292	0.287	0.295	0.125	0.033	0.034	0.064	0.034	0.282	0.014	0.034
HKY+I	26	14249.626	14054.111	-7001.004	0.50	n/a	2.53	0.292	0.287	0.295	0.125	0.041	0.042	0.090	0.041	0.212	0.018	0.041
T92+G+I	25	14526.440	14338.441	-7144.173	0.39	1.17	3.50	0.290	0.290	0.210	0.210	0.032	0.023	0.164	0.032	0.164	0.023	0.032
T92+G	24	14530.821	14350.339	-7151.125	n/a	0.34	3.58	0.290	0.290	0.210	0.210	0.031	0.022	0.165	0.031	0.165	0.022	0.031
K2+G	23	14587.330	14414.364	-7184.142	n/a	0.33	3.71	0.250	0.250	0.250	0.250	0.027	0.027	0.197	0.027	0.197	0.027	0.027
K2+G+I	24	14589.035	14408.553	-7180.232	0.38	1.04	3.64	0.250	0.250	0.250	0.250	0.027	0.027	0.196	0.027	0.196	0.027	0.027
T92+I	24	14596.501	14416.019	-7183.965	0.49	n/a	2.99	0.290	0.290	0.210	0.210	0.036	0.026	0.159	0.036	0.159	0.026	0.036
K2+I	23	14670.097	14497.132	-7225.525	0.50	n/a	3.07	0.250	0.250	0.250	0.250	0.031	0.031	0.189	0.031	0.189	0.031	0.031
GTR	29	15043.026	14824.965	-7383.419	n/a	n/a	2.60	0.292	0.287	0.295	0.125	0.039	0.078	0.062	0.039	0.270	0.004	0.077
TN93	26	15153.141	14957.626	-7452.762	n/a	n/a	2.58	0.292	0.287	0.295	0.125	0.042	0.043	0.059	0.043	0.260	0.018	0.043
HKY	25	15208.117	15020.119	-7485.012	n/a	n/a	2.55	0.292	0.287	0.295	0.125	0.040	0.042	0.090	0.041	0.212	0.018	0.041
T92	23	15453.262	15280.296	-7617.108	n/a	n/a	2.50	0.290	0.290	0.210	0.210	0.041	0.029	0.151	0.041	0.151	0.029	0.041
JC+G+I	23	15473.601	15300.635	-7627.277	0.42	1.94	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	22	15478.114	15312.665	-7634.296	n/a	0.40	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	22	15514.443	15348.995	-7652.460	0.49	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083
K2	22	15552.407	15386.958	-7671.442	n/a	n/a	2.51	0.250	0.250	0.250	0.250	0.036	0.036	0.179	0.036	0.179	0.036	0.036
JC	21	16320.513	16162.581	-8060.257	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 12 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. There were a total of 1141 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 [2][3]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

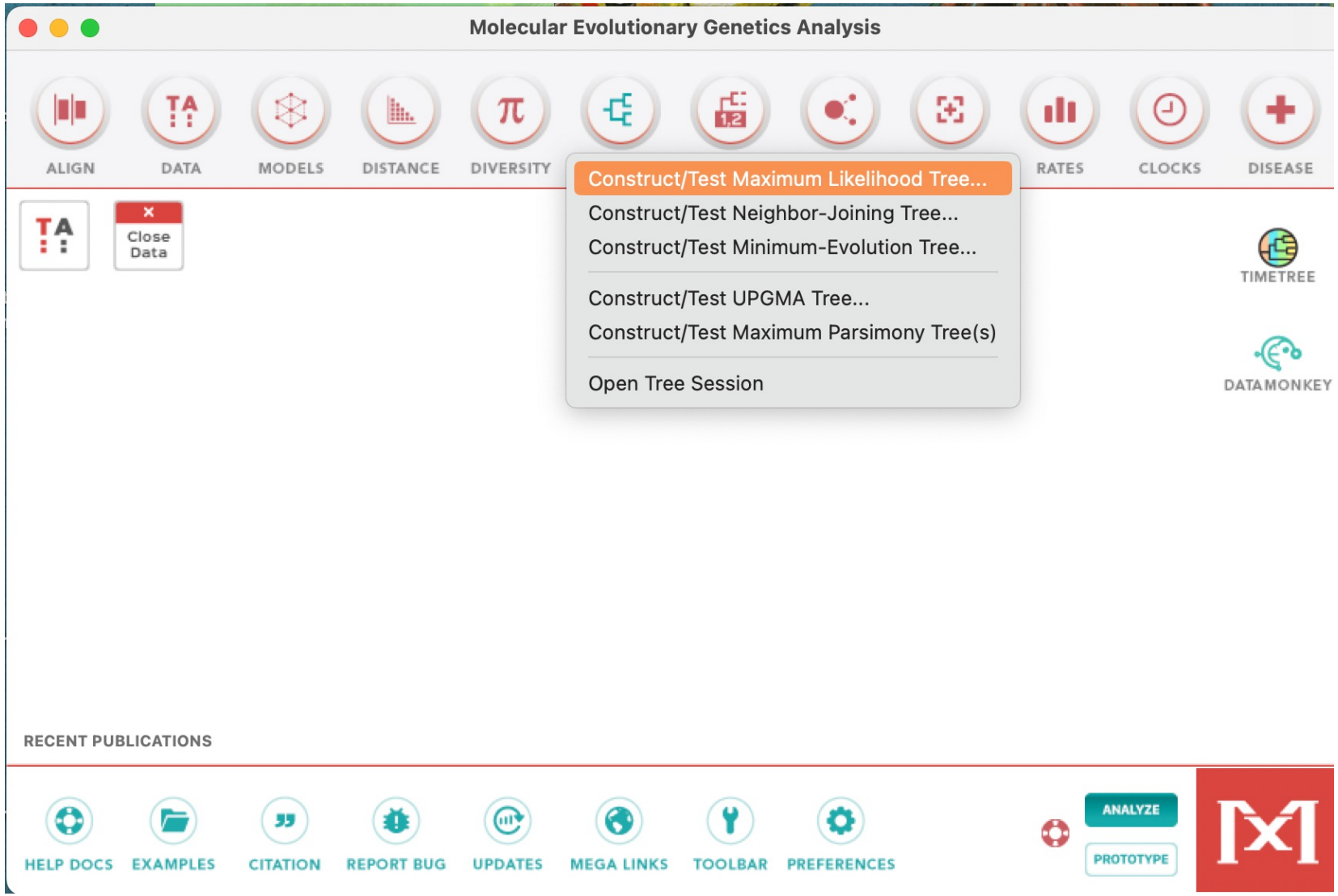
1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Tamura K., Stecher G., and Kumar S. (**2021**). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* https://doi.org/10.1093/molbev/msab120.
3. Stecher G., Tamura K., and Kumar S. (**2020**). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution* 37:1237-1239.

These are the results, according to the description, the lowest BIC score is the best DNA model to use.

We will use the GTR+G+I model

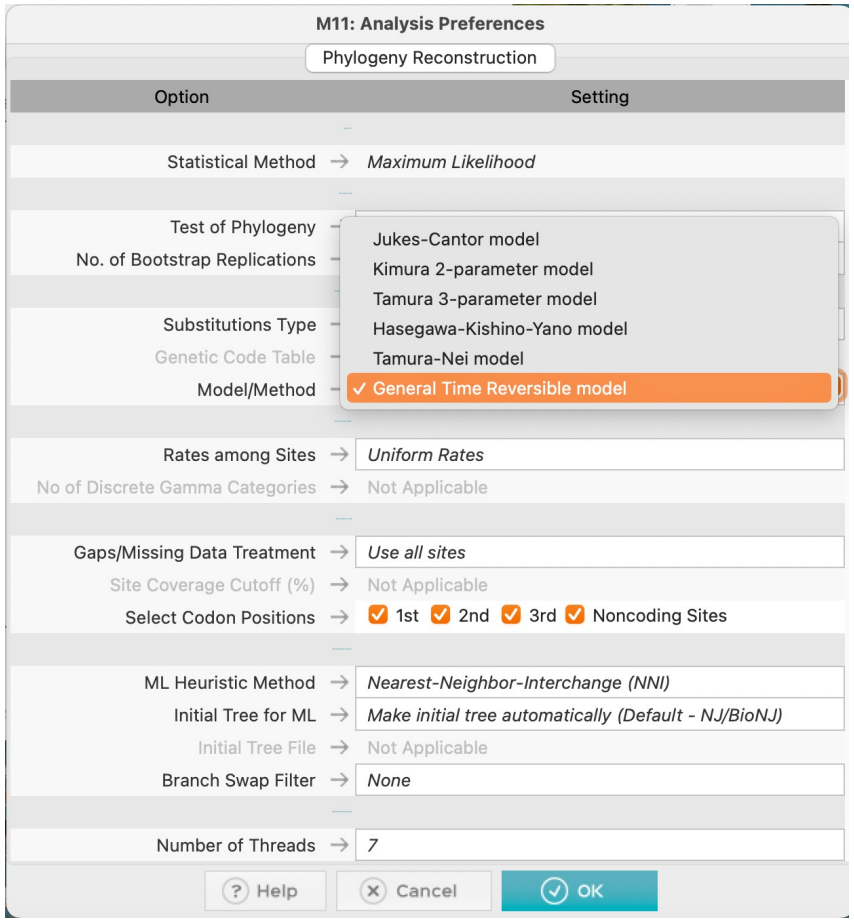
Ce sont les résultats, selon la description, le score BIC le plus bas est le meilleur modèle d'ADN à utiliser.

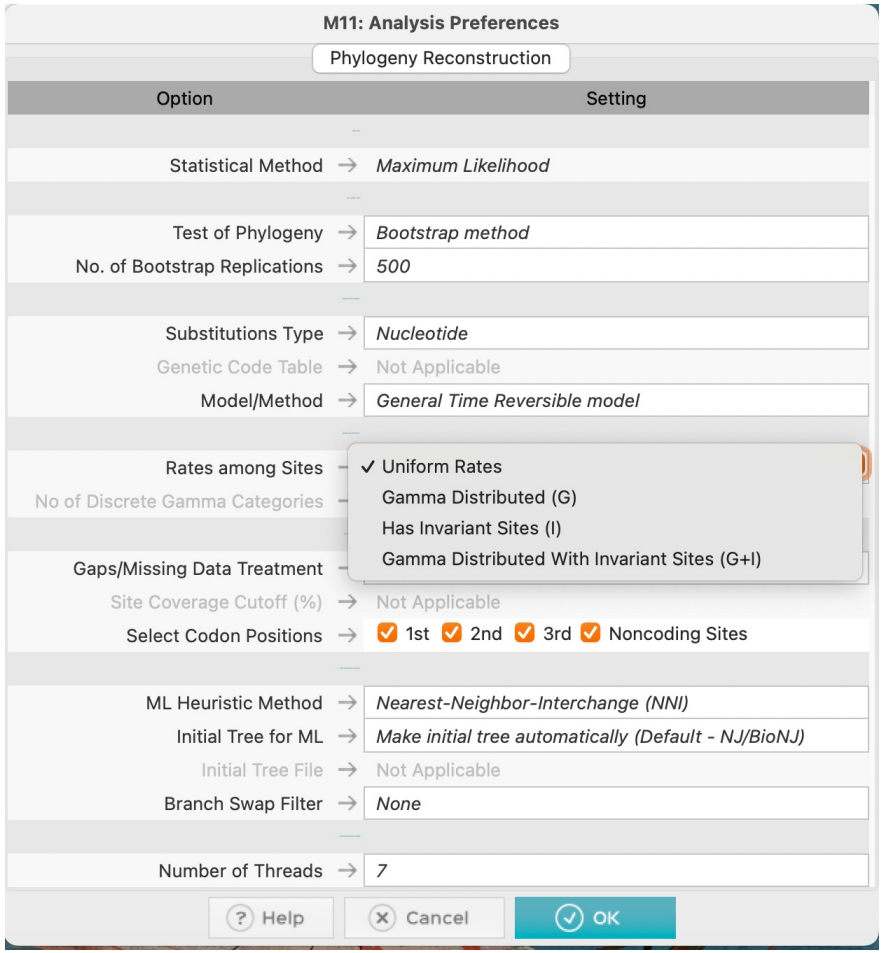
Nous utiliserons le modèle GTR+G+I



Step 6: click on “Phylogeny” then select construct/test maximum likelihood tree

Étape 6 : cliquez sur « Phylogénie » puis sélectionnez construire/tester l'arbre de maximum de vraisemblance



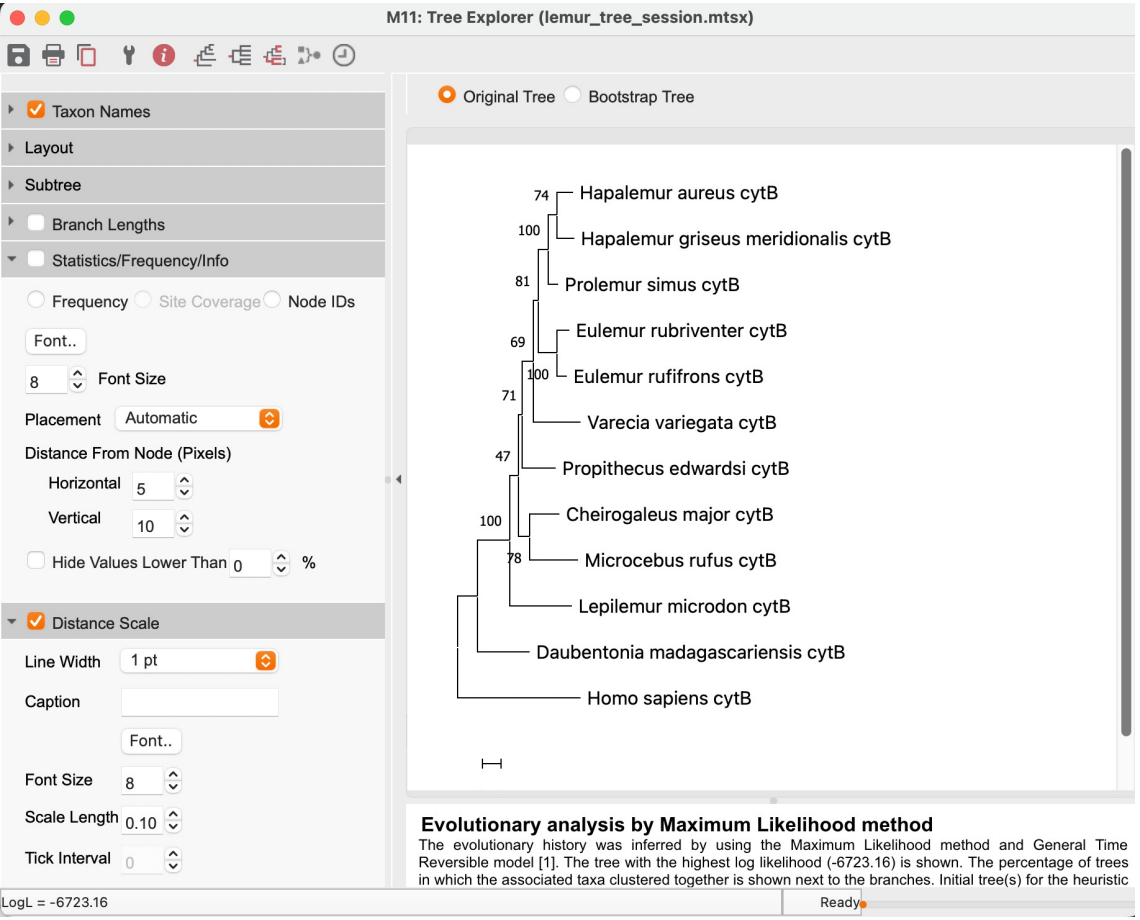


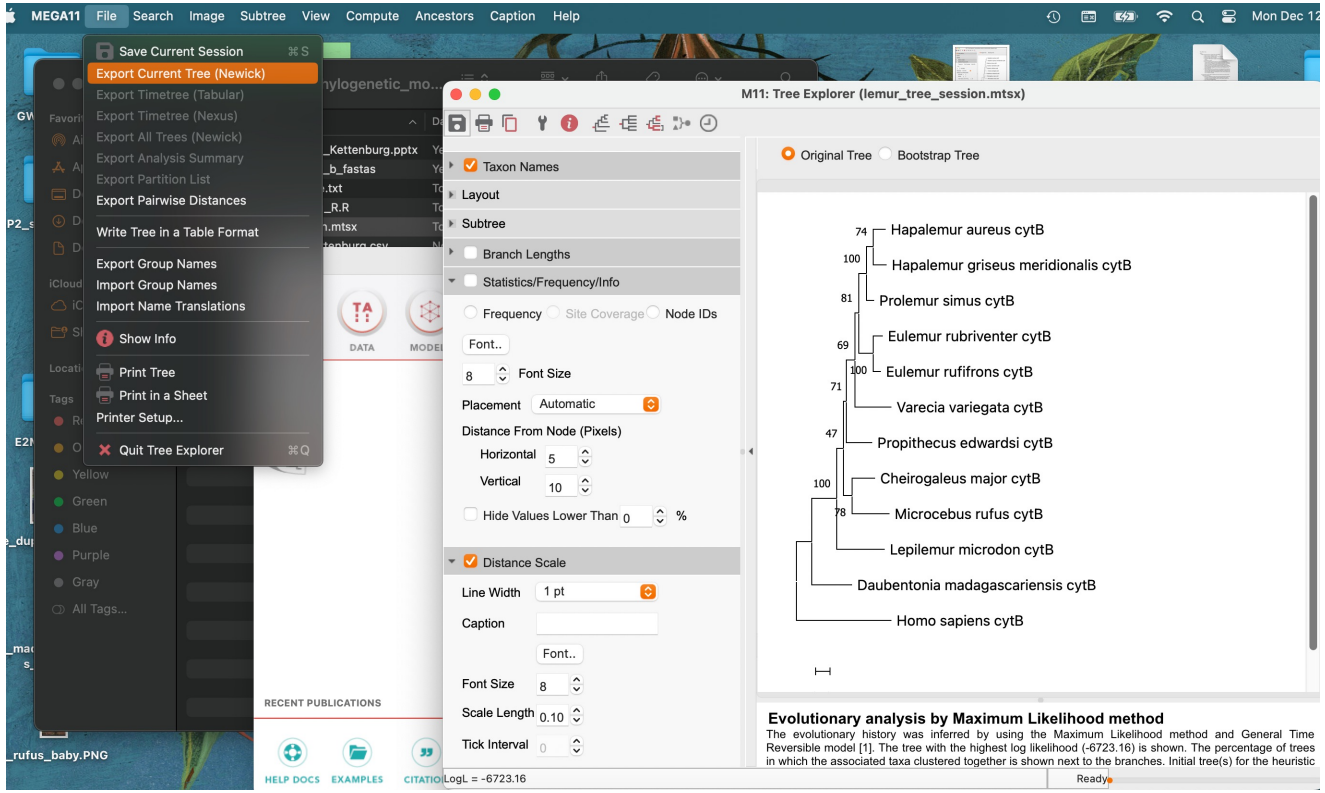
Select general time reversible model and for rates among sites, select Gamma distributed with invariant sites (G+I), hit okay, then wait a bit!

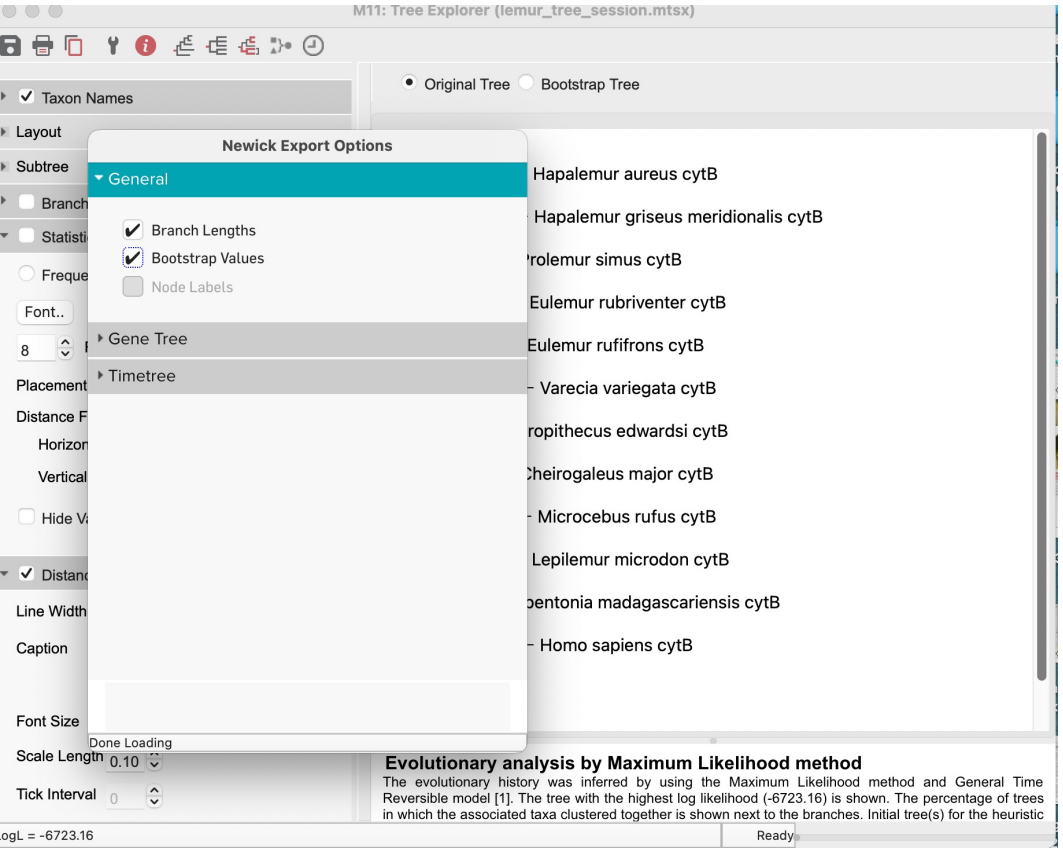
Sélectionnez le modèle général réversible en temps et pour les taux entre les sites, sélectionnez Gamma distribué avec des sites invariants (G+I), appuyez sur OK, puis attendez un peu !

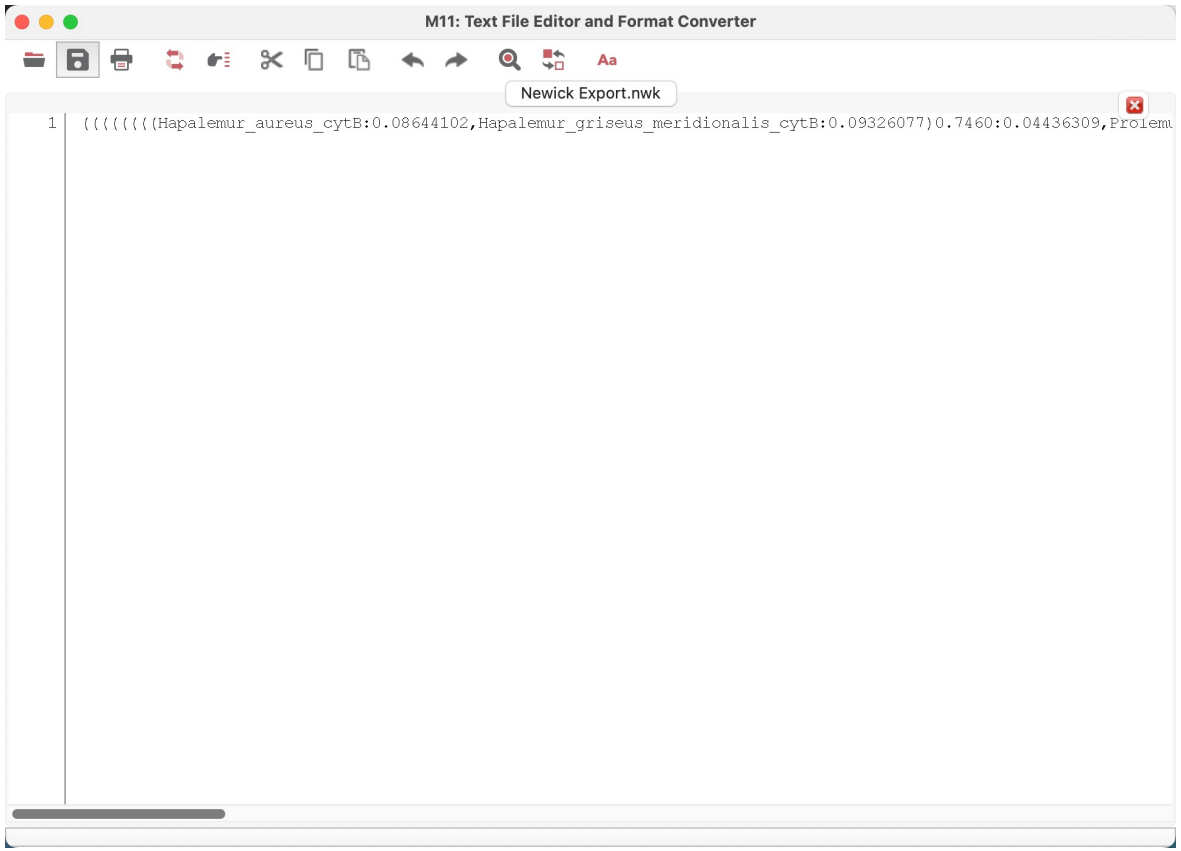
File -> export current tree (Newick)

Fichier -> exporter l'arborescence actuelle (Newick)







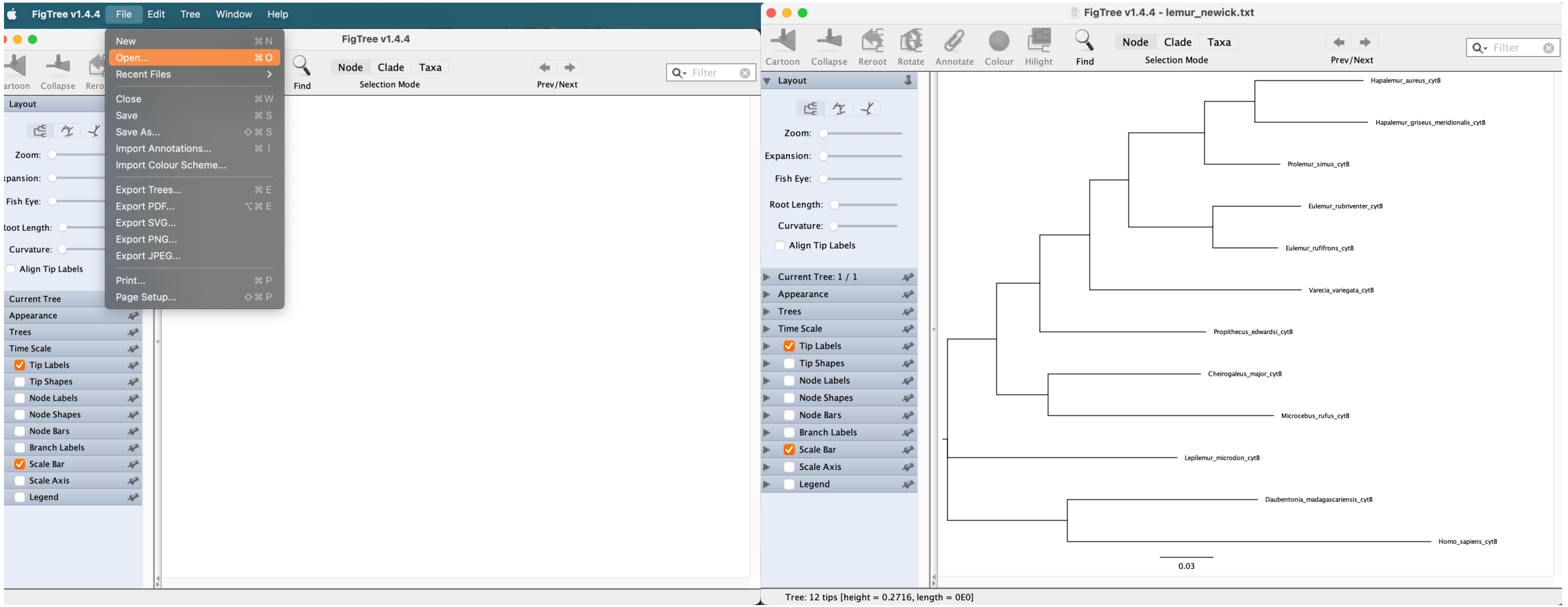


Copy the Newick string and past into a text/edit file…this is what we will import into R

Copiez la chaîne Newick et collez-la dans un fichier texte/édition… c'est ce que nous allons importer dans R

Good practice to check tree in FigTree first

Bonne pratique pour vérifier d'abord l'arbre dans FigTree



Then make pretty in R!
Alors fais joli en R !

- Follow instructions in lemur_tree_editing_R.R file
- *Suivez les instructions dans le fichier lemur_tree_editing_R.R*