

Introduction to phylogenetic modeling

Gwen Kettenburg

PhD student in Cara Brook's lab

University of Chicago Ecology and Evolution

Adapted from slides by:

Richard Ree, University of Chicago

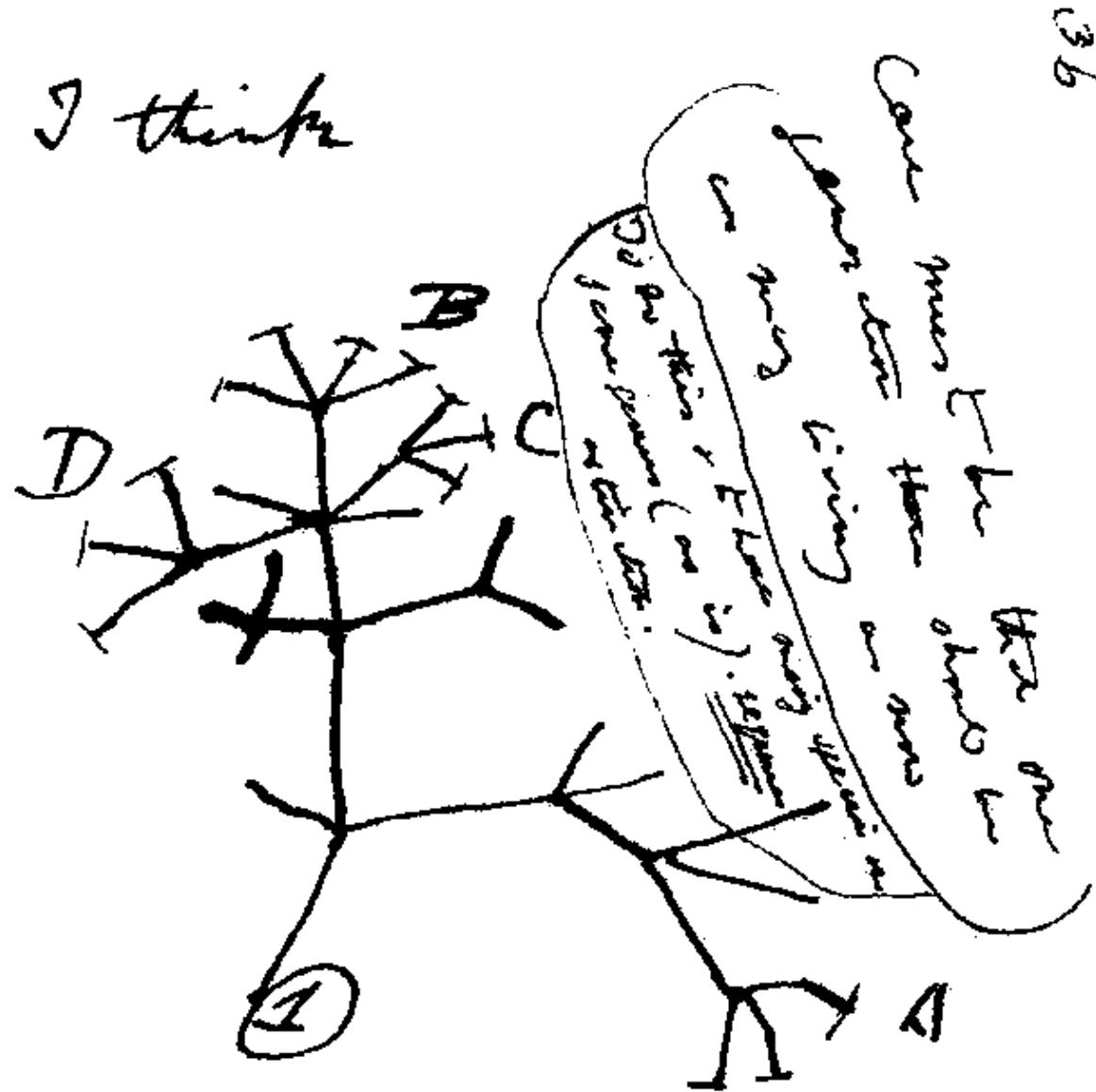
Andrew Hipp, University of Chicago

Reminders:

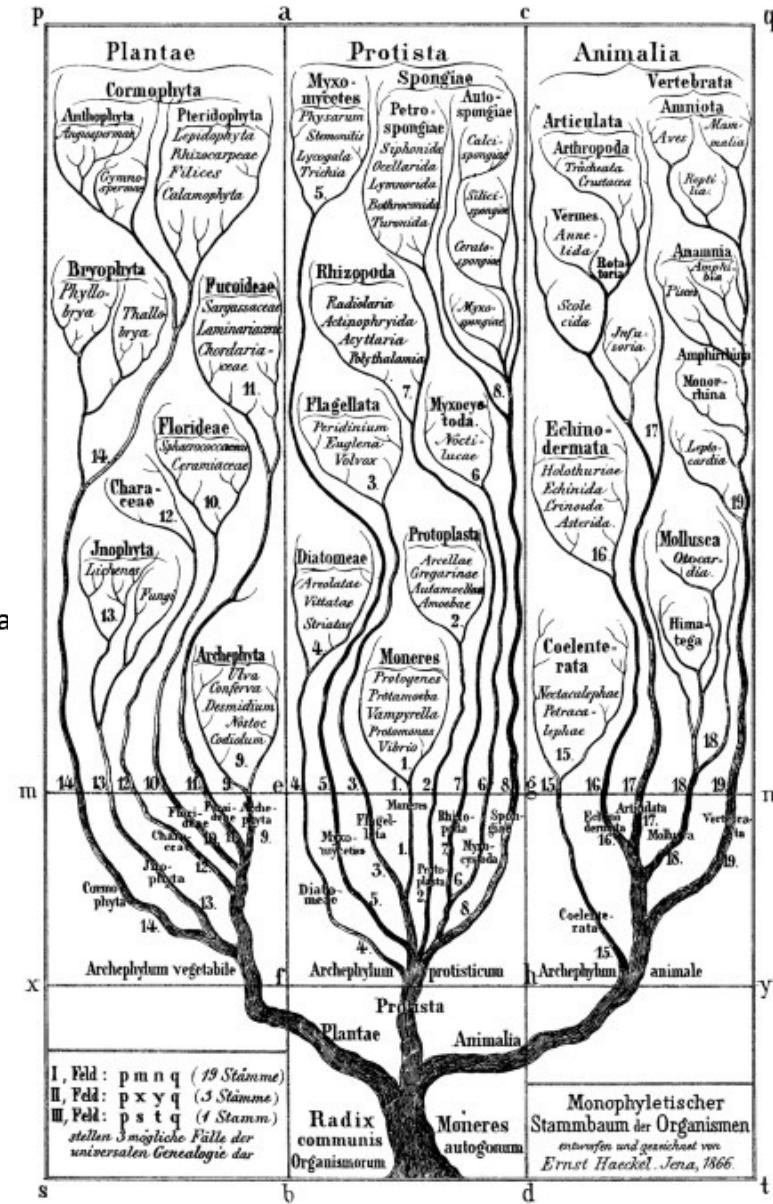
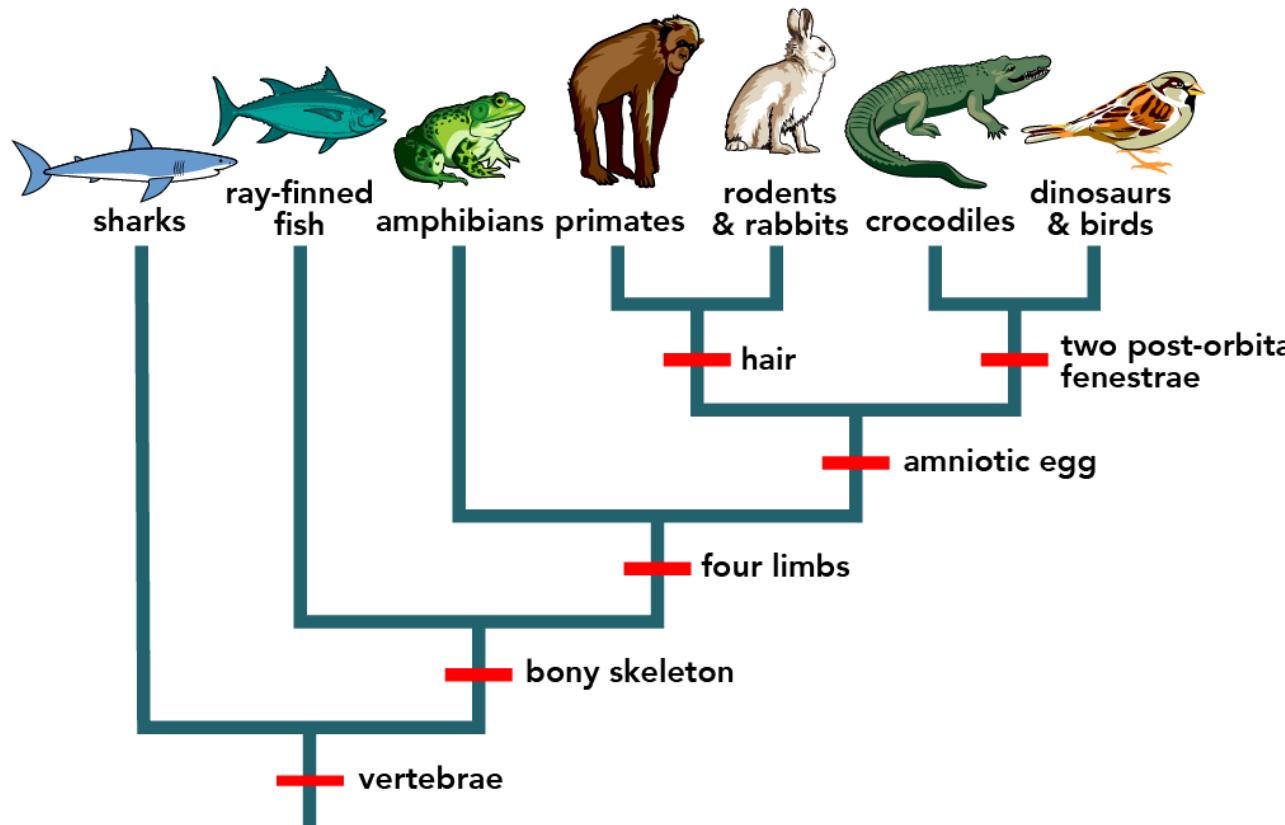
- If you need something clarified or have a question at any time, interrupt!
- If you need something translated to Malagasy/French so it's clearer, please let me know
- Remember to say your name when you raise your hand

What is a phylogeny?

I think



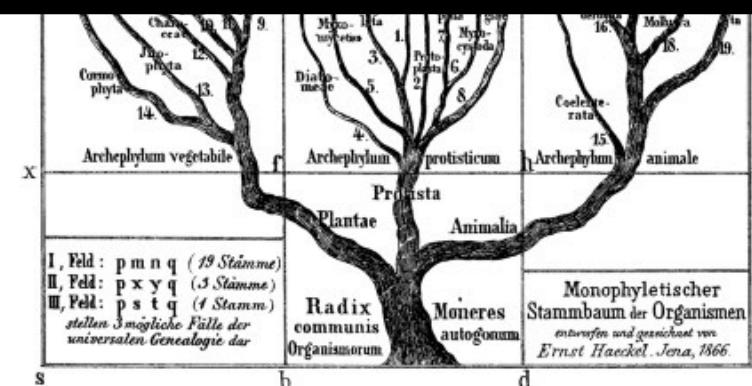
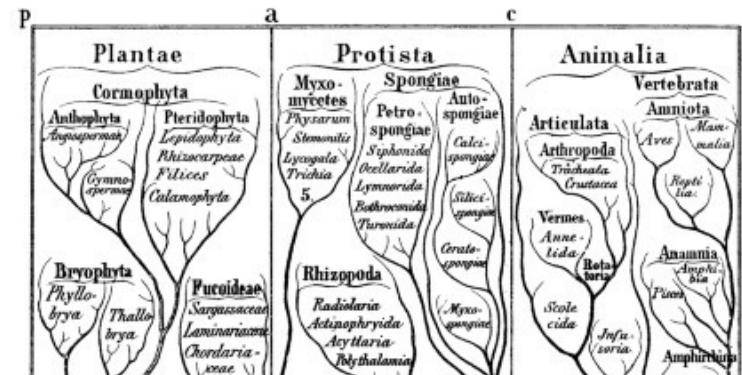
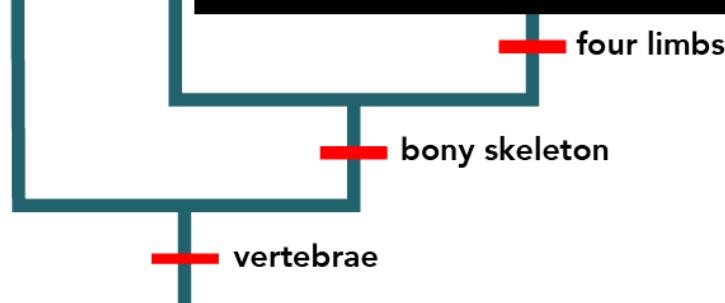
What is a phylogeny?



What is a phylogeny?



“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.”



Baum et. al, Nature, 2008

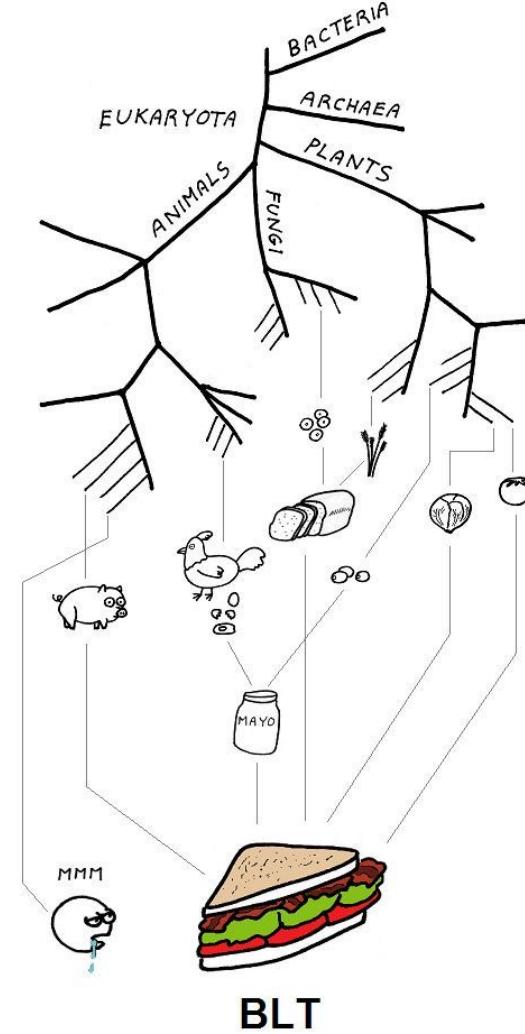
Hossfeld and Levit, *Nature*, 2016

How is this useful to the E's in 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!

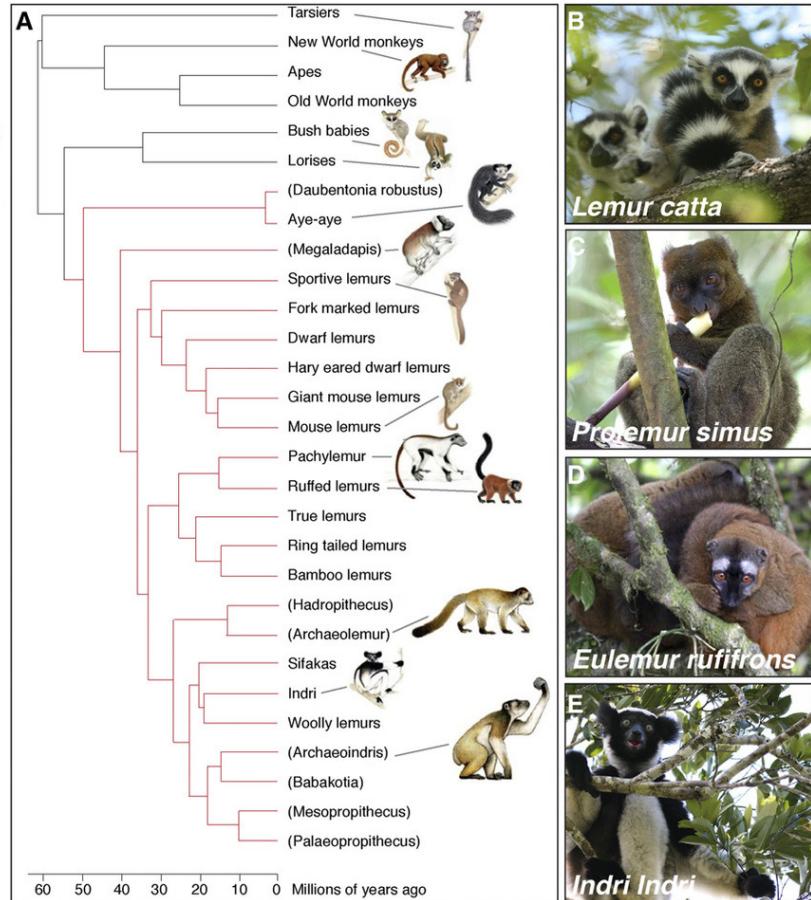
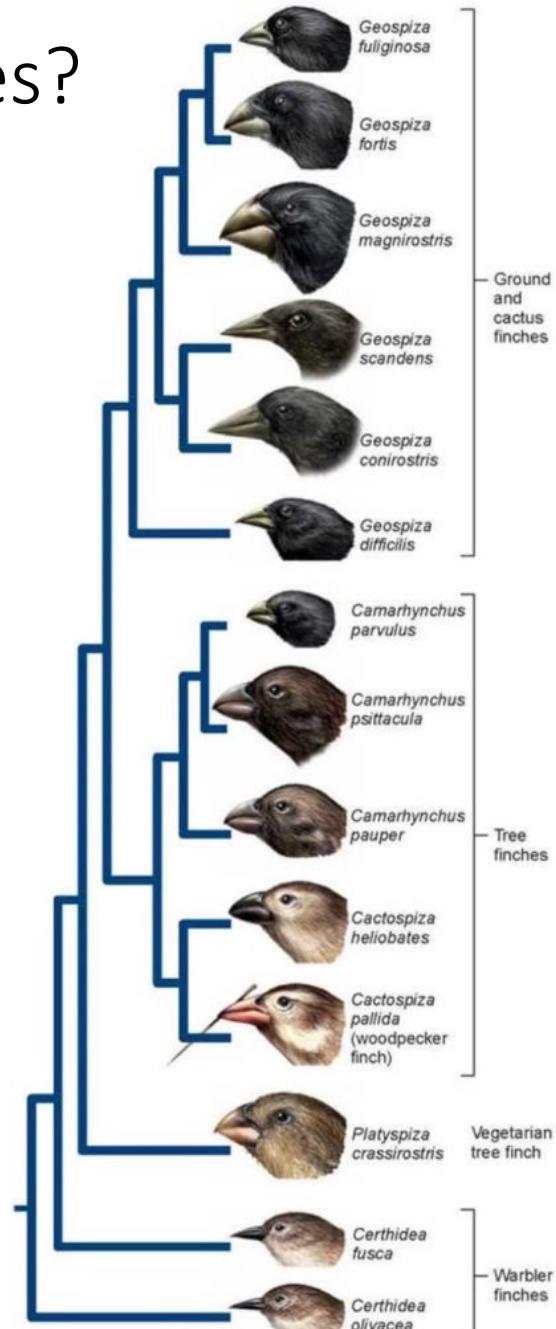
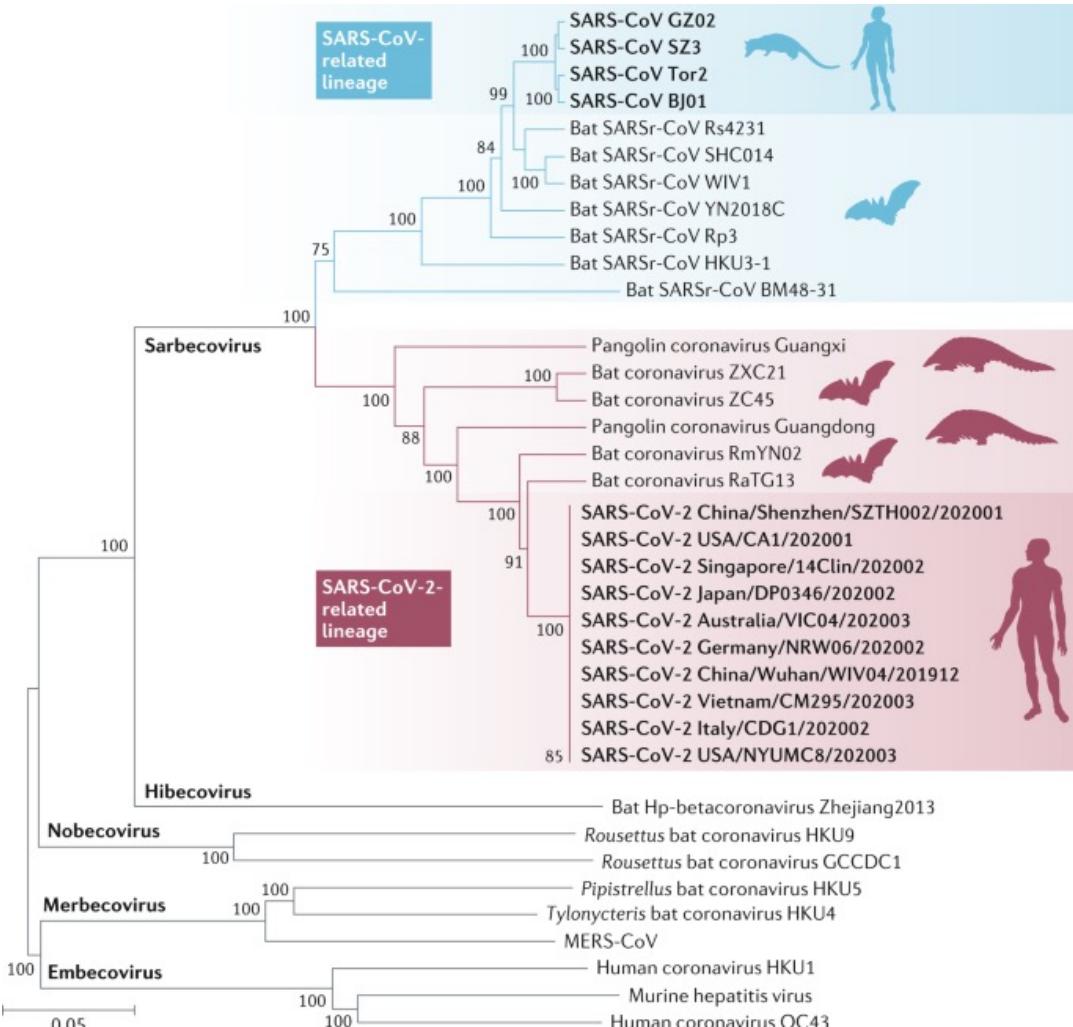
Goals:

- 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



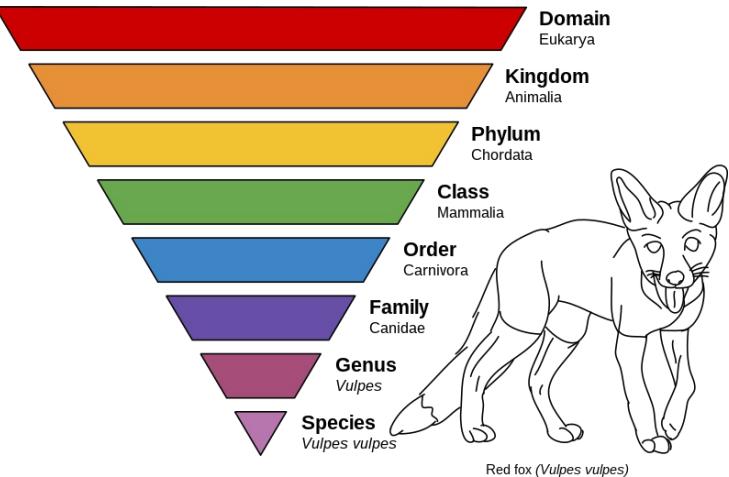
Molecular Evolutionary
Genetics Analysis

What can you do with phylogenies?

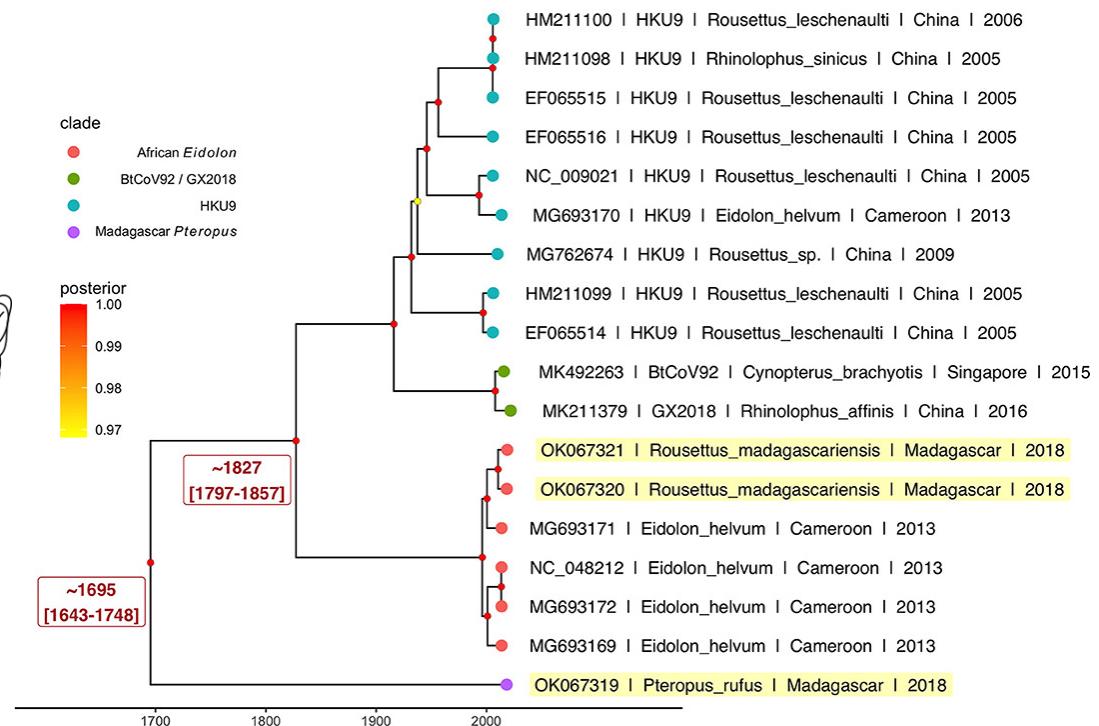


Ezran et al., Genetics, 2022
Hu et al., Nature reviews, 2022

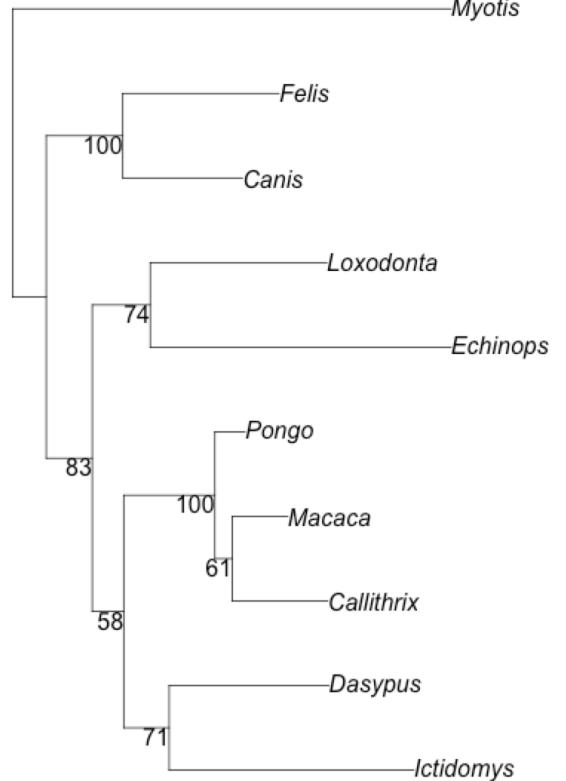
Taxonomy/nomenclature



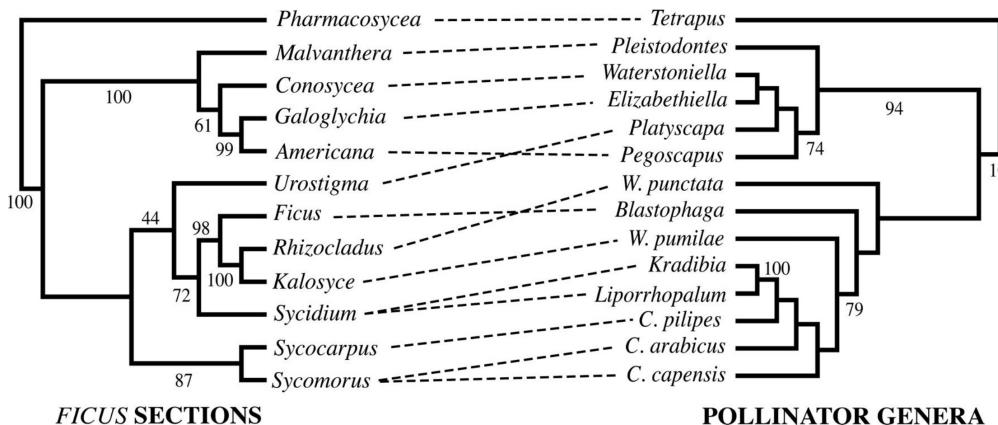
Bayesian trees



Maximum likelihood



Figs and fig wasps

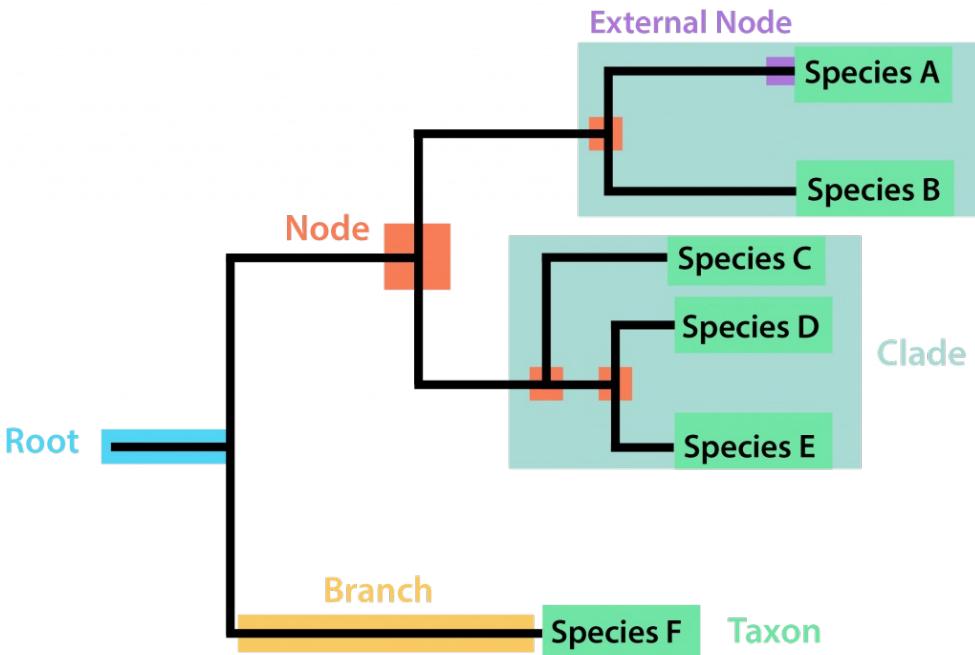


plant and pollinator phylogenies show limited congruence; host switching and hybridization has been common in their coevolutionary history

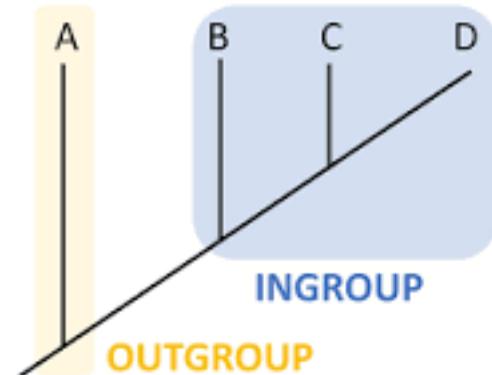
Adapted from Ree and Hipp, UChicago, 2021
Wikipedia, 2022

Kettenburg et al., Frontiers in Public Health, 2022
Quick and dirty tree building in R, 2016

Anatomy of a phylogeny



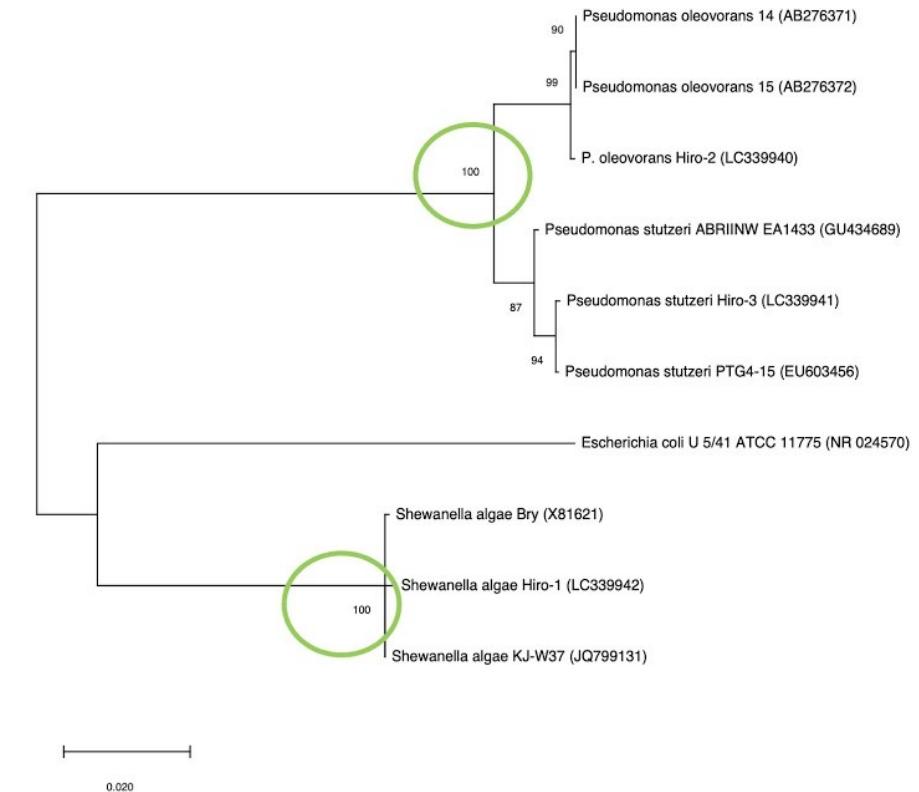
Phylogenetic Tree Structure



CONFIDENCE

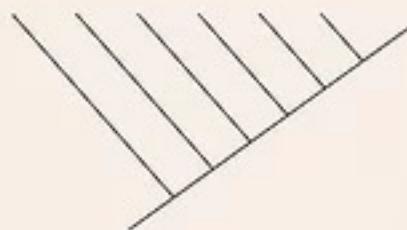
BOOTSTRAP VALUE

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



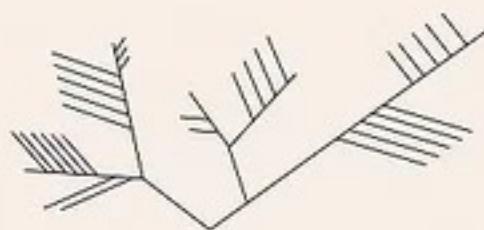
Cladogram versus phylogenetic tree

CLADOGRAM



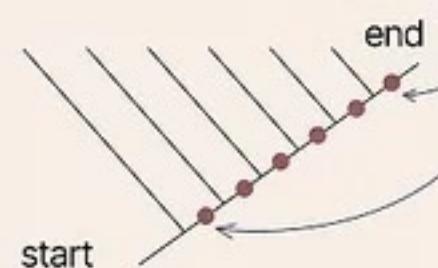
- the relationships are *hypothetical*
- you can easily make on your own

PHYLOGENETIC TREE



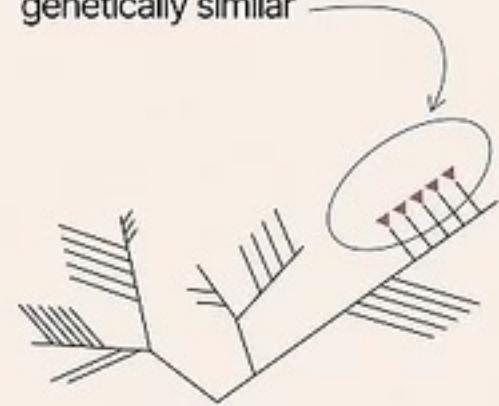
- the relationships are *backed by molecular evidence*
- should have access to DNA or other molecular data

Nodes closer to the start of the main line happened longer ago than nodes closer to the end



CLADOGRAM

Animals that are closer together are also more genetically similar



PHYLOGENETIC TREE

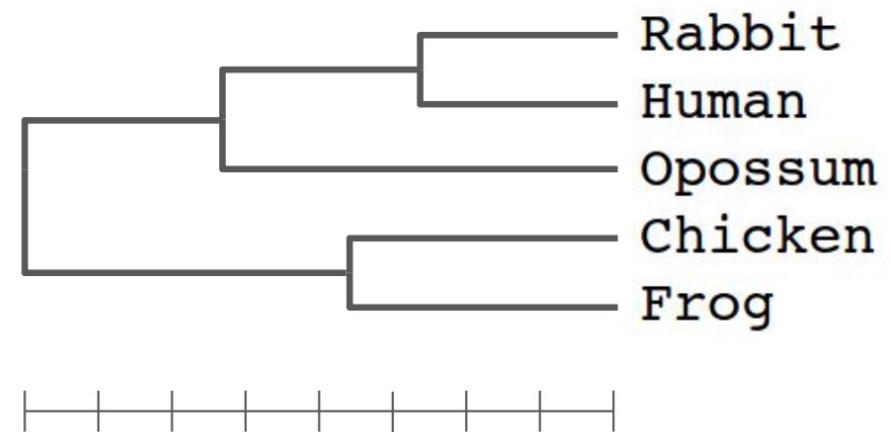
Parsimony versus likelihood

- Parsimony: minimum number of changes
- Likelihood: maximum probability of the data having evolved on the tree



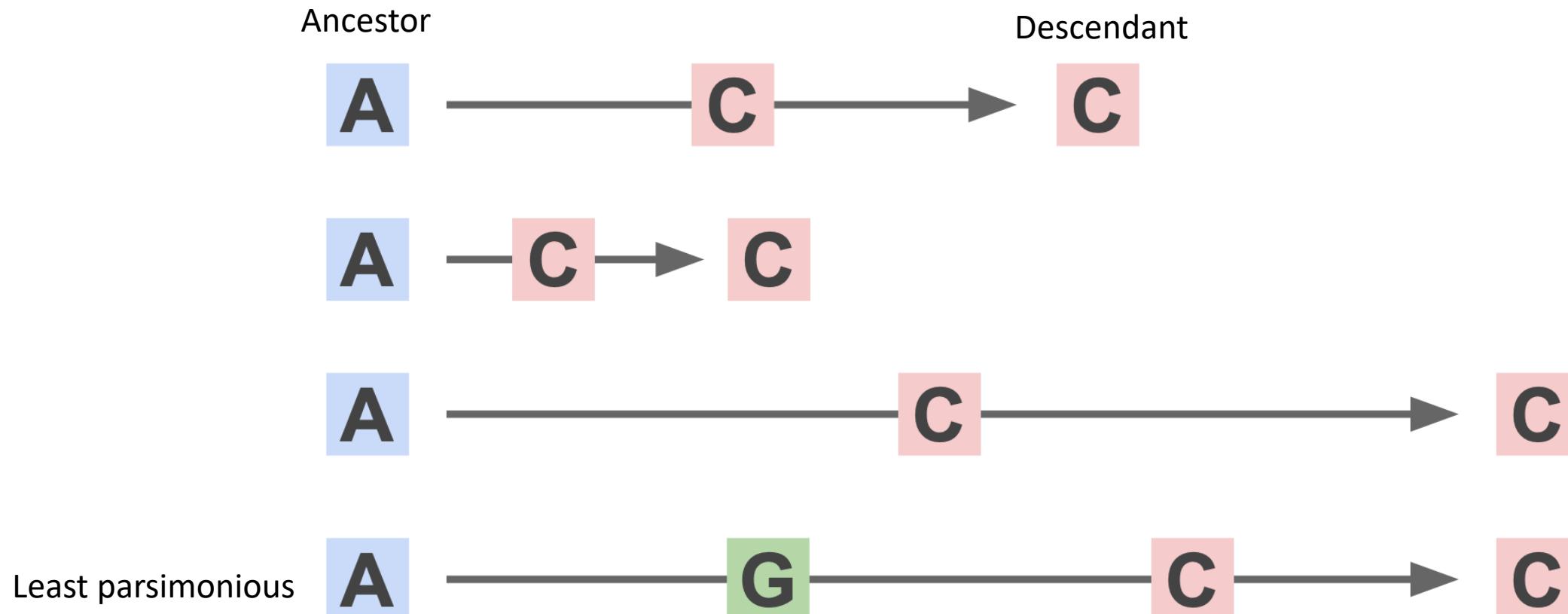
branch length can mean different things:

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

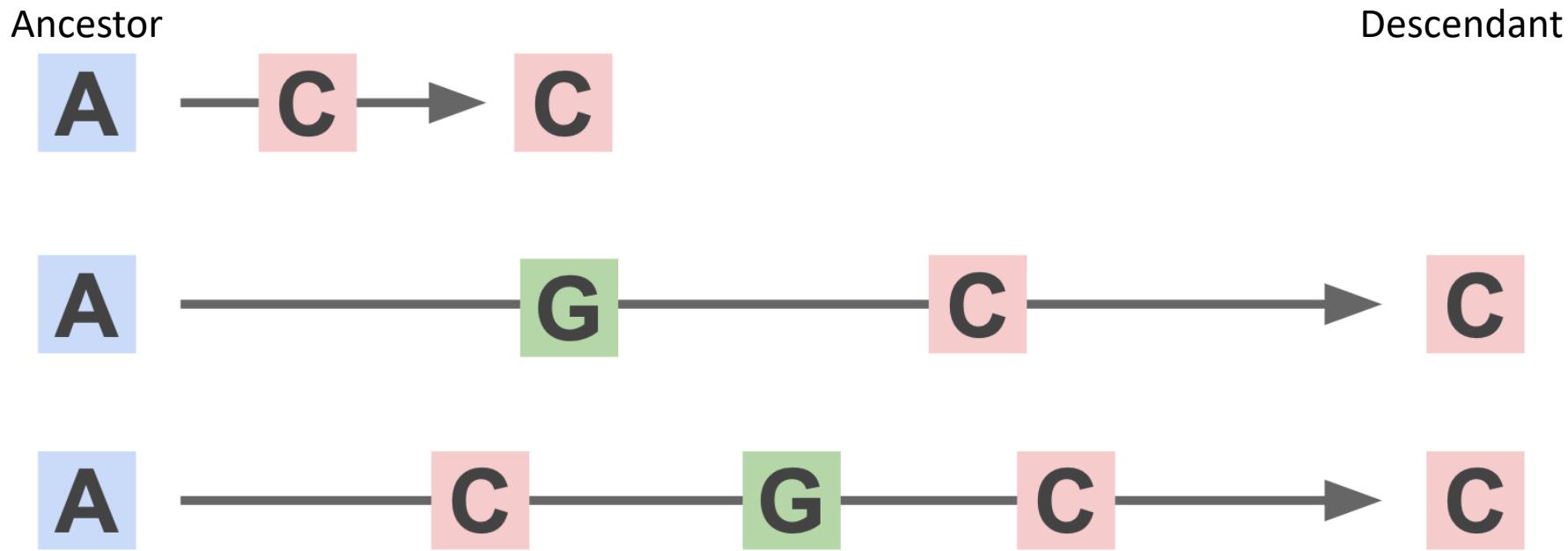


proposed tree has **branch lengths** in units of expected number of changes per site

Parsimony: minimum number of changes
regardless of time/opportunity



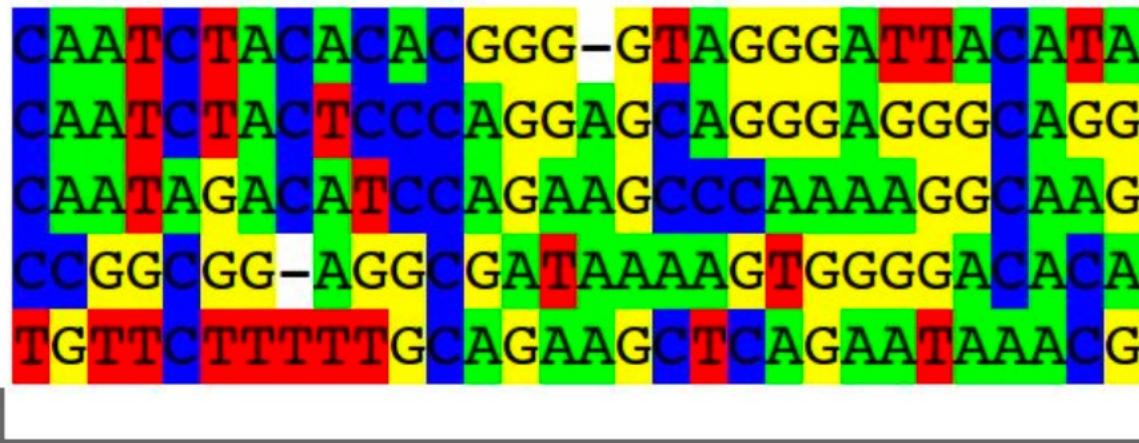
Likelihood: probability of ancestral and descendant status is a function of time (branch length)



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

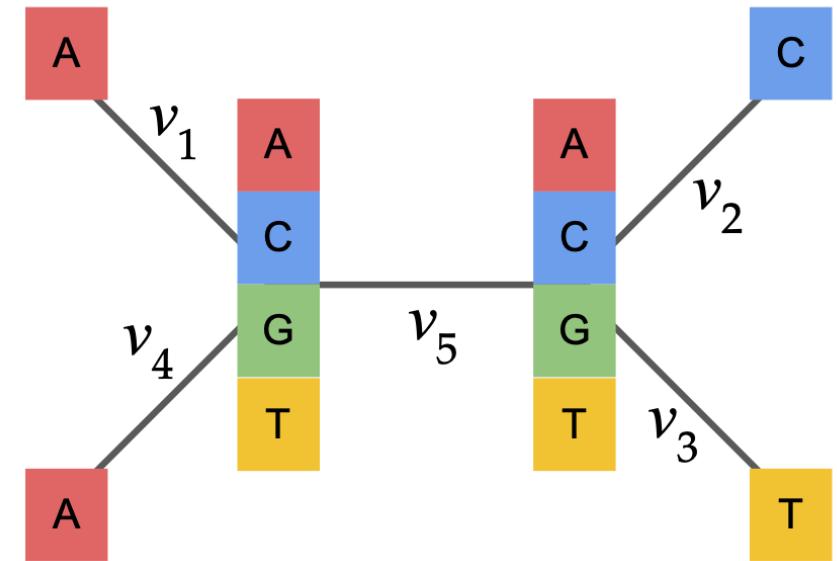
Likelihood cont'd.

Rabbit
Human
Opossum
Chicken
Frog



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

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



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

Models of DNA evolution

- Markov models that describe relative rates of different changes
 - 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

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

$(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4})$ 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:

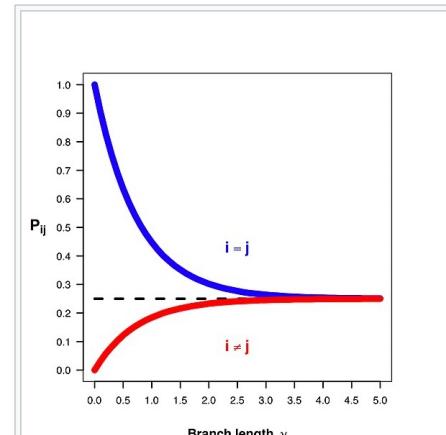
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

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. <code>+G8</code> .
+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 <code>+G</code> model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. <code>+R6</code> (default 4 categories if not specified). The FreeRate model typically fits data better than the <code>+G</code> model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

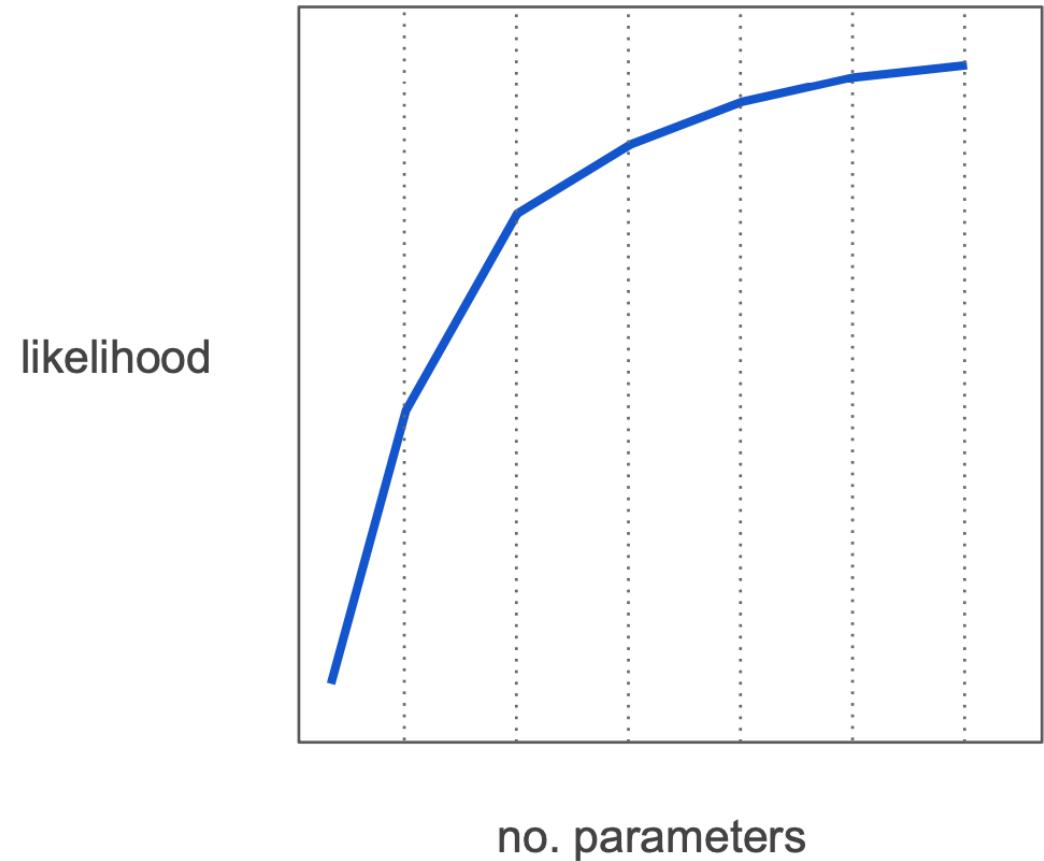
Model selection

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.

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



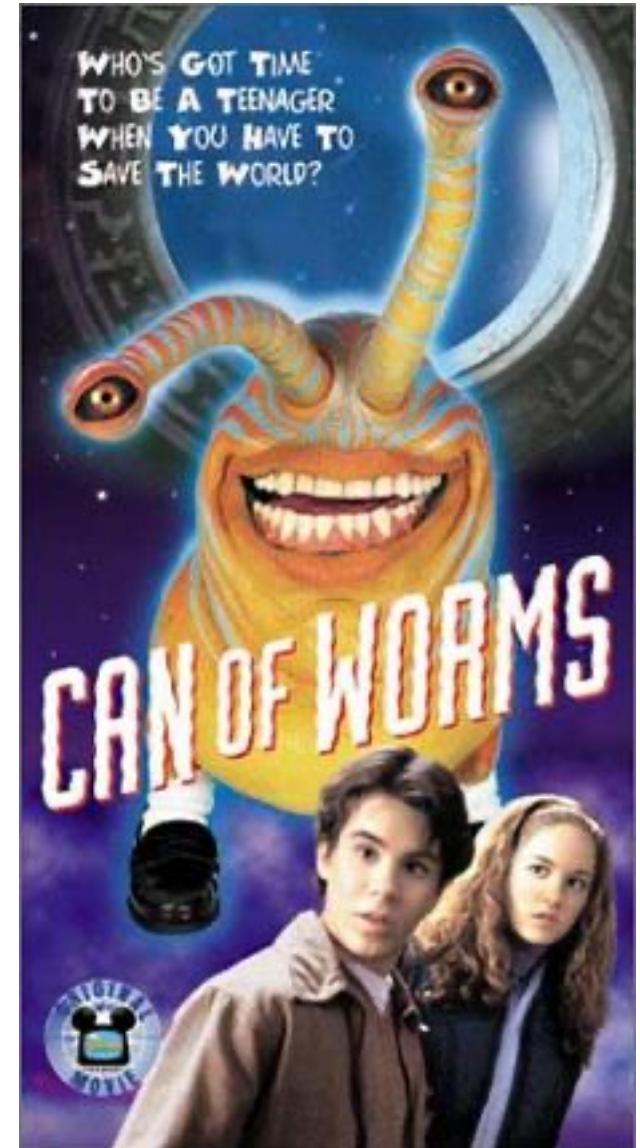
Rate heterogeneity across sites

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

Rate heterogeneity across 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



Bootstrapping

- Specify number of replicates: how many times does the test replicate the original sequence alignment?
- Standard in MEGA is 500 replicates, 1000 is better but takes longer



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	G	T	G	G	A	T	C	G	T
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

```
graph TD; Root --- B; Root --- C; C --- A; C --- D;
```

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	G	G	G	G
Species B	T	G	G	A	G	A	C	G	T	T	C	G	G	T	T
Species C	T	C	C	A	T	G	C	G	A	G	C	G	G	A	G
Species D	C	C	C	A	T	T	A	C	G	G	G	T	G	A	T

Bootstrap tree #1

```
graph TD; Root --- B; Root --- D; D --- C; D --- A;
```

Bootstrap pseudo-replicate #2

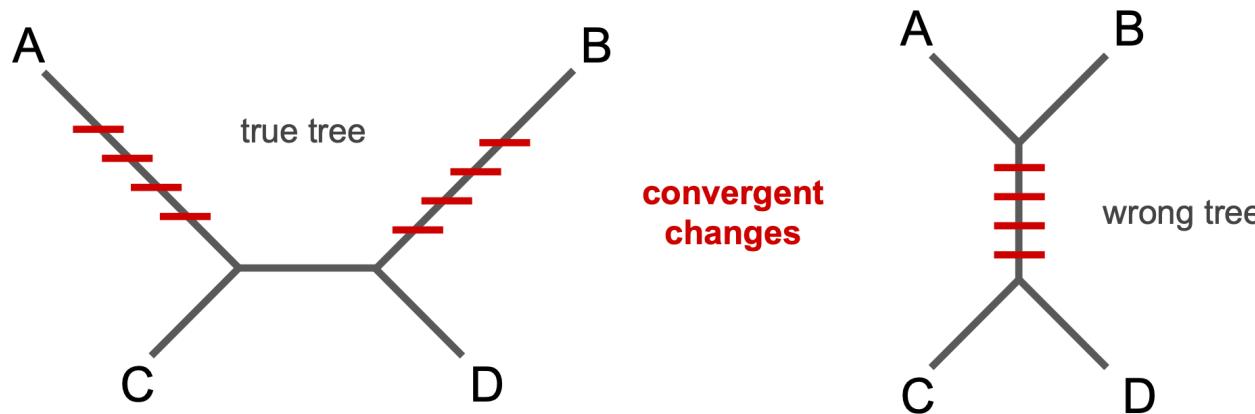
	9	7	12	5	2	4	2	6	14	9	4	9	7	2	1
Species A	T	A	T	C	A	G	A	T	T	G	T	A	A	T	A
Species B	G	T	A	T	C	A	C	G	C	G	T	A	A	T	A
Species C	T	T	A	T	G	C	G	A	C	T	T	G	A	A	G
Species D	T	T	C	A	C	A	A	T	C	T	T	A	A	T	A

Bootstrap tree #2

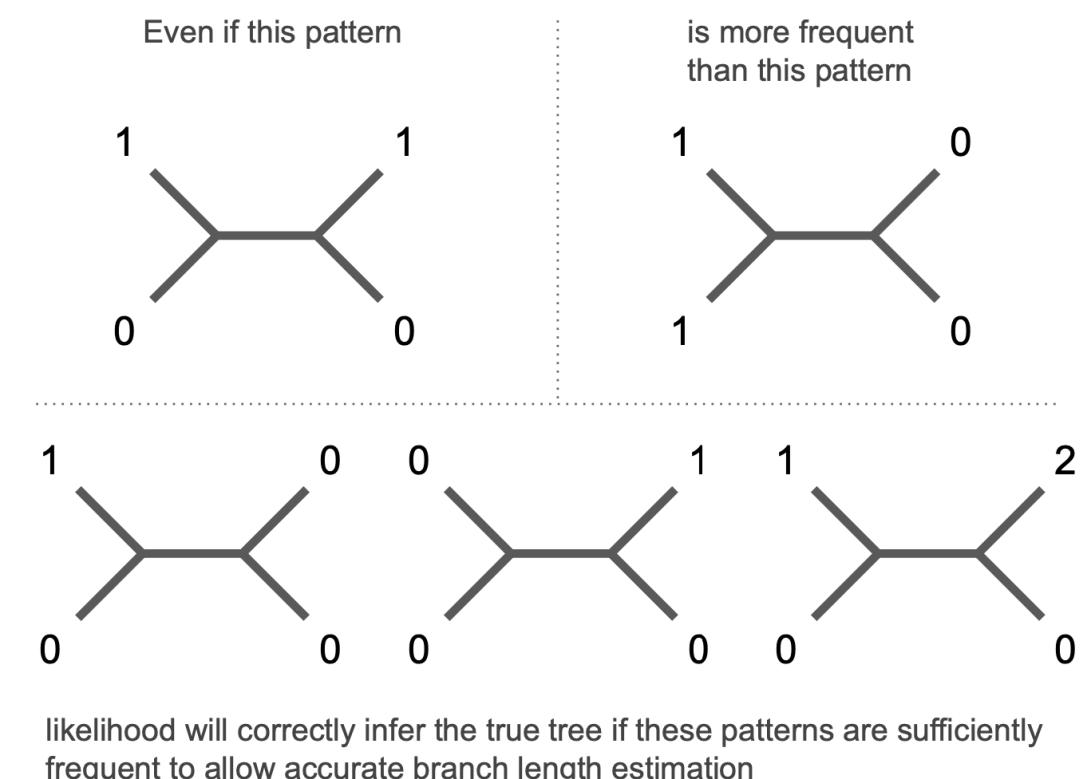
```
graph TD; Root --- B; Root --- C; C --- A; C --- D;
```

Felsenstein zone

- Branch lengths for which parsimony confidently infers the wrong topology, these can affect bootstrap values



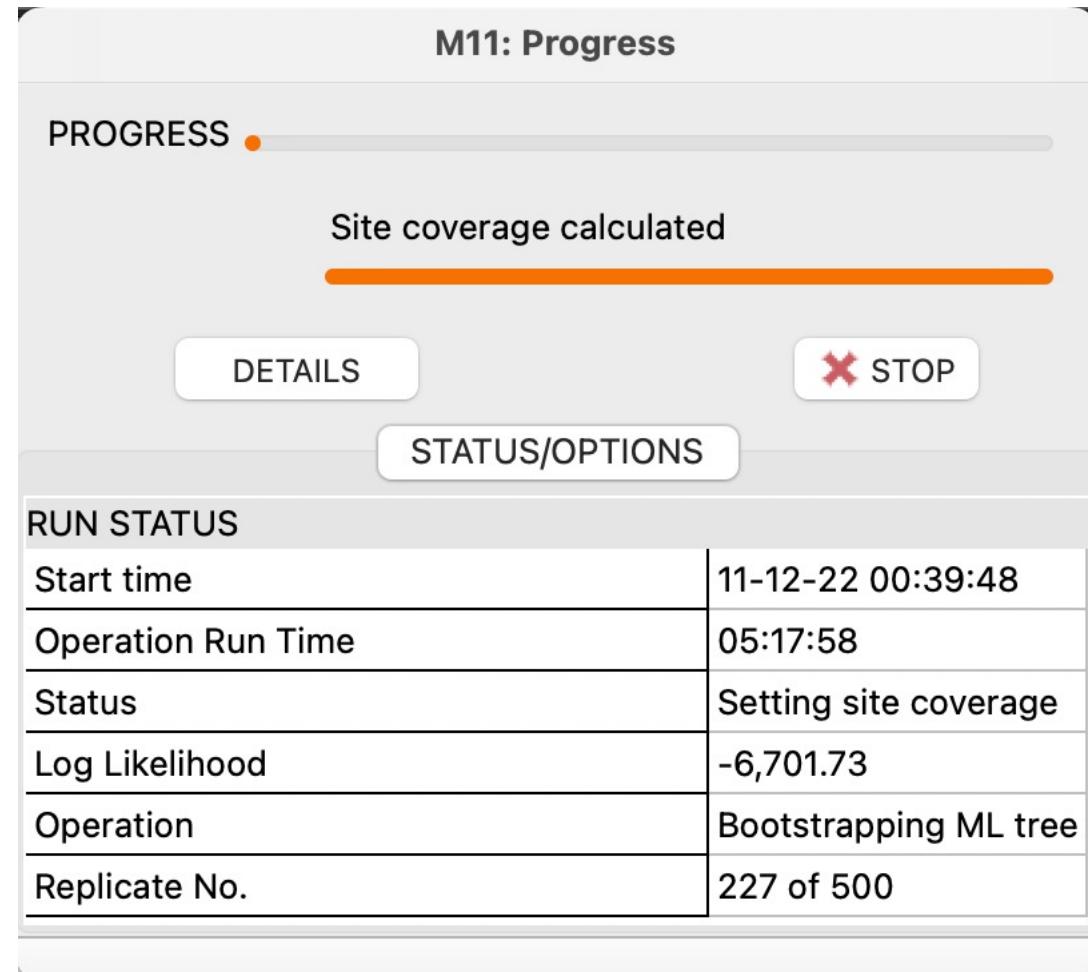
likelihood is a **consistent estimator** of tree topology because it converges on the correct value with increasing data



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

Warnings and limitations

- Building phylogenies takes a LONG time, larger ones can take up to a week to run and Bayesian phylogenies can run for months, so a computing cluster is almost necessary for this
- Without a proper outgroup or root, a phylogeny doesn't tell you much about order of descent
- Does anyone know why it was so hard to make phylogenies for COVID-19?



So you have your sequences, now what?

- Get some reference sequences from NCBI
- Get an outgroup from NCBI
- Align them (use a software like MEGA or online like MAFFT)
- Pick the best model (use a software like MEGA or ModelTest-NG)
- Run the phylogeny using your aligned sequences and chosen model (use a software like MEGA or RAxML)
- Visualize/edit tree in either R or FigTree

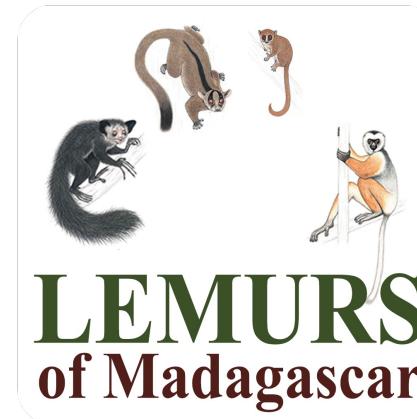
All of this listed is free to use ☺

TUTORIAL

- 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!
- Also have FigTree downloaded, or just follow along by watching
- Additionally, open internet and go to NCBI.gov...if this does not open just follow along by watching

Lemurs of Ranomafana national park

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



Steps to revisit later

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

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

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

Enter Query Sequence

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

```
GGACAAGTAGCCTCCATTCTATCTTCTAATCCTTATTATACCAAC
TGTAAAGCTCATCGAAA
ACAAAGATACTTAAATGAAGA
```

Query subrange [?](#)

From
To

Or, upload 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 Betacoronavirus

Nucleotide collection (nr/nt) [?](#)

Limit by Organism BioProjectID WGS Project

exclude [Add organism](#) [?](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

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

Limit to Sequences from type material

Entrez Query [YouTube](#) [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
<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>	<input type="text"/> to <input type="text"/>

[Filter](#) [Reset](#)

[Descriptions](#) [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments [Download](#) [Select columns](#) [Show 100](#) [?](#)

<input checked="" type="checkbox"/> 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



Log in

All Databases

Eulemur rufifrons 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



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

COVID-19 Information

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

Search NCBI

Eulemur rufifrons cytochrome B



Search

Results found in 4 databases

Literature	
Bookshelf	0
MeSH	0
NLM Catalog	0
PubMed	0
PubMed Central	4

Genes	
Gene	0
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	0

Proteins	
Conserved Domains	0
Identical Protein Groups	5
Protein	28
Protein Family Models	0
Structure	0

Genomes	
Assembly	0
BioCollections	0
BioProject	0
BioSample	0
Genome	0
Nucleotide	28
SRA	0

Clinical	
ClinicalTrials.gov	0
ClinVar	0
dbGaP	0
dbSNP	0
dbVar	0
GTR	0
MedGen	0

PubChem	
BioAssays	0
Compounds	0
Pathways	0
Substances	0

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

Nucleotide Nucleotide Eulemur rufifrons cytochrome b| Search Help

Create alert Advanced

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 [Cricetus 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-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](#)

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



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

```
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$ █
```

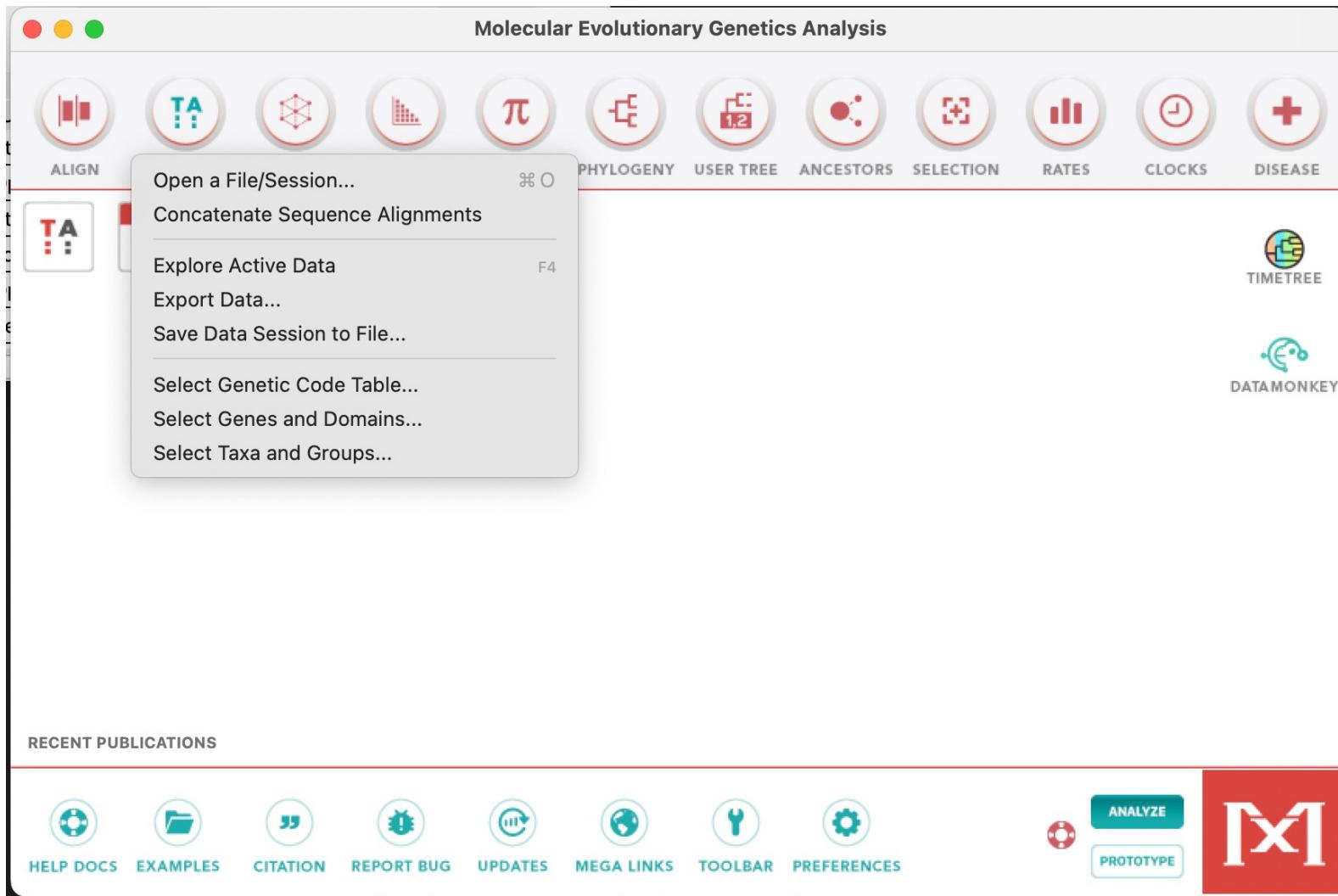
Example 1: Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
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)

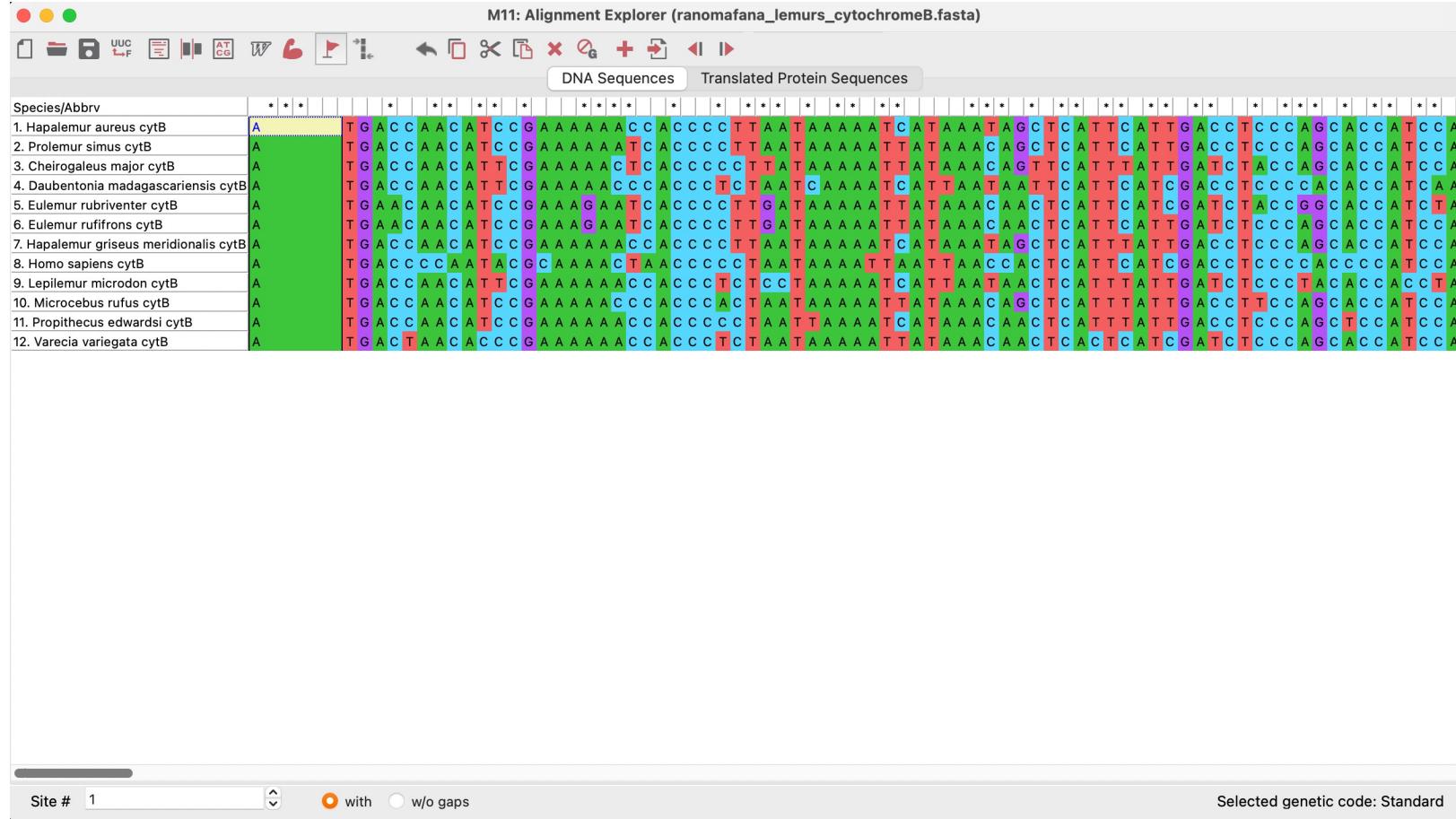
When using asterisk(*) to concatenate all files. Please DON'T use same extension for target file(Eg. .csv). There should be some difference in pattern else target file will also be considered in concatenation

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



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



So the sequences are loaded into MEGA like this:

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

M11: Alignment Explorer (ranomafana_lemurs_cytochromeB.fasta)

DNA Sequences Translated Protein Sequences

Species/Abbr

1. Hapalemur aureus cytB	A T G A C C A A C A T C C G A A A A A C C A C C C C T T A A T A A A A T C A T A A T A G C T C A T T C A T T G A C C T C C C A G G C A C C A T C C A A C A T C
2. Prolemur simus cytB	A T G A C C A A C A C T C C G A A A A A T C A C C C C C T T A A T A A A A T T A A A C A G G C T C A T T C A T T G A C C T C C C A G G C A C C A T C C A A C A T C
3. Cheirogaleus major cytB	A T G A C C A A C A C T C G A A A A A C T C A C C C C C T T A A T A A A A T T A A A C A G G T C A T T C A T T G A T C T A C C A G G C A C C A T C C A A C A T C
4. Daubentonia madagascariensis cytB	A T G A C C A A C A C T C G A A A A A C C C C C C T T A A T A A A A T T A A A C A G G T C A T T C A T T G A T C T A C C A G G C A C C A T C C A A C A T C
5. Eulemur rubriventer cytB	A T G A A C A A C A C T C C G A A A G A A T C A C C C C
6. Eulemur rufifrons cytB	A T G A A C A A C A C T C C G A A A G A A T C A C C C C
7. Hapalemur griseus meridionalis cytB	A T G A C C A A C A C T C C G A A A A A C C C C C C
8. Homo sapiens cytB	A T G A C C C C A A T A C G C A A A A C T A A C C C C
9. Lepilemur microdon cytB	A T G A C C A A C A C T C G A A A A A C C C C C C
10. Microcebus rufus cytB	A T G A C C A A C A C T C C G A A A A A C C C C C C
11. Propithecus edwardsi cytB	A T G A C T A A C A C C C G A A A A A C C C C C C
12. Varecia variegata cytB	A T G A C T A A C A C C C G A A A A A C C C C C C

MSCLE Alignment Options

Option	Setting
GAP PENALTIES	
Gap Open	<input checked="" type="checkbox"/> -400.00
Gap Extend	<input checked="" type="checkbox"/> 0.00
MEMORY/ITERATIONS	
Max Memory in MB	<input checked="" type="checkbox"/> 2048
Max Iterations	<input checked="" type="checkbox"/> 16
ADVANCED OPTIONS	
Cluster Method (Iterations 1,2)	<input checked="" type="checkbox"/> UPGMA
Cluster Method (Other Iterations)	<input checked="" type="checkbox"/> UPGMA
Min Diag Length (Lambda)	<input checked="" type="checkbox"/> 24

Help Reset Cancel OK

Site # 1 with w/o gaps Selected genetic code: Standard

Go with suggested options, then hit okay

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

MEGA11 Data Edit Search Alignment Web Sequencer Display Help Sun Dec 11 12:31 AM

Create New
InL Open
Open a Recently Used File
Close
Phylogenetic Analysis
Save Session
Export Alignment
✓ DNA Sequences
Protein Sequences
Translate/Untranslate
Genetic Code
Reverse Complement
Reverse
Complement
Quit

MEGA Format
FASTA Format
NEXUS/PAUP Format

M11: Alignment Explorer (ranomafana_lemurs_cytochromeB_aligned.mas)

DNA Sequences Translated Protein Sequences

11. Propithecus edwardsi cytB
12. Varecia variegata cytB

16

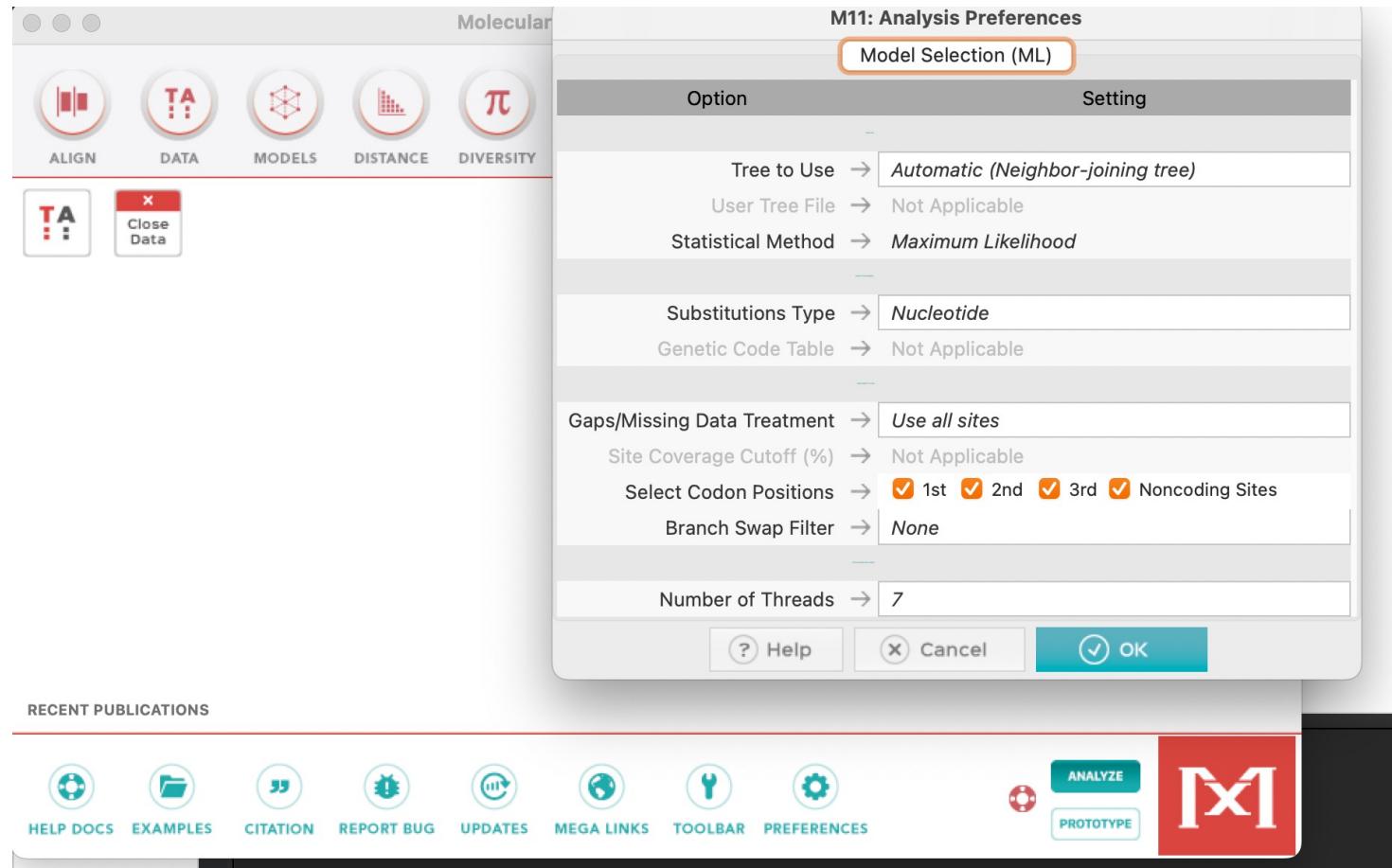
E2M2

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

Molecular Evolutionary Genetics Analysis

The screenshot shows the MEGA software interface. At the top, there is a toolbar with various icons: ALIGN, DATA, TA, SER TREE, ANCESTORS, SELECTION, RATES, CLOCKS, and DISEASE. Below the toolbar, a vertical menu is open under the 'DATA' icon. The menu items are: Find Best DNA/Protein Models (ML)... (highlighted with an orange box), Disparity Index Test of Pattern Heterogeneity, Estimate Substitution Matrix (ML)..., Estimate Transition/Transversion Bias (ML)..., Compute MCL Substitution Matrix, Compute MCL Transition/Transversion Bias, Compute Pattern Disparity Index, Compute Composition Distances, Compute Amino Acid Composition, Compute Nucleotide Composition, and Compute Codon Usage Bias. To the right of the menu, there are two sections: 'TIMETREE' with its logo and 'DATAMONKEY' with its logo. At the bottom, there is a 'RECENT PUBLICATIONS' section and a footer with links: HELP DOCS, EXAMPLES, CITATION, REPORT BUG, UPDATES, MEGA LINKS, TOOLBAR, PREFERENCES, ANALYZE, and PROTOTYPE. The 'ANALYZE' button is highlighted with a red box.

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



Go with the suggested parameters, then click okay



Results

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	InL	(+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 (*InL*), 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.

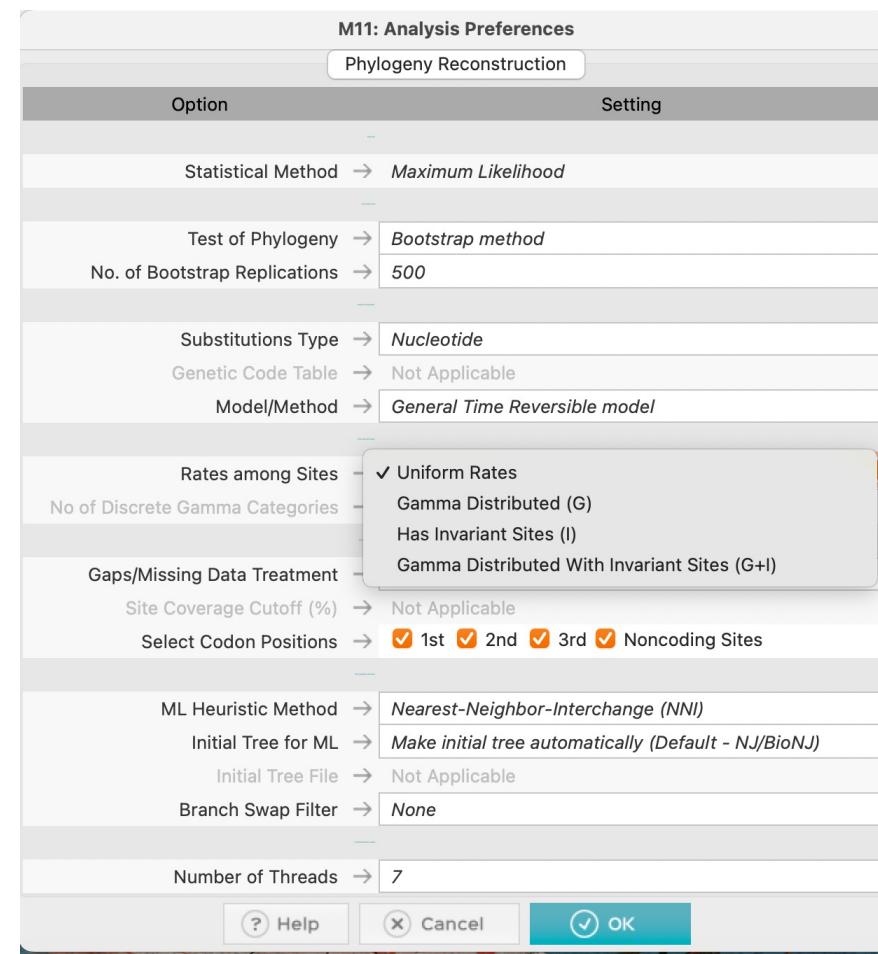
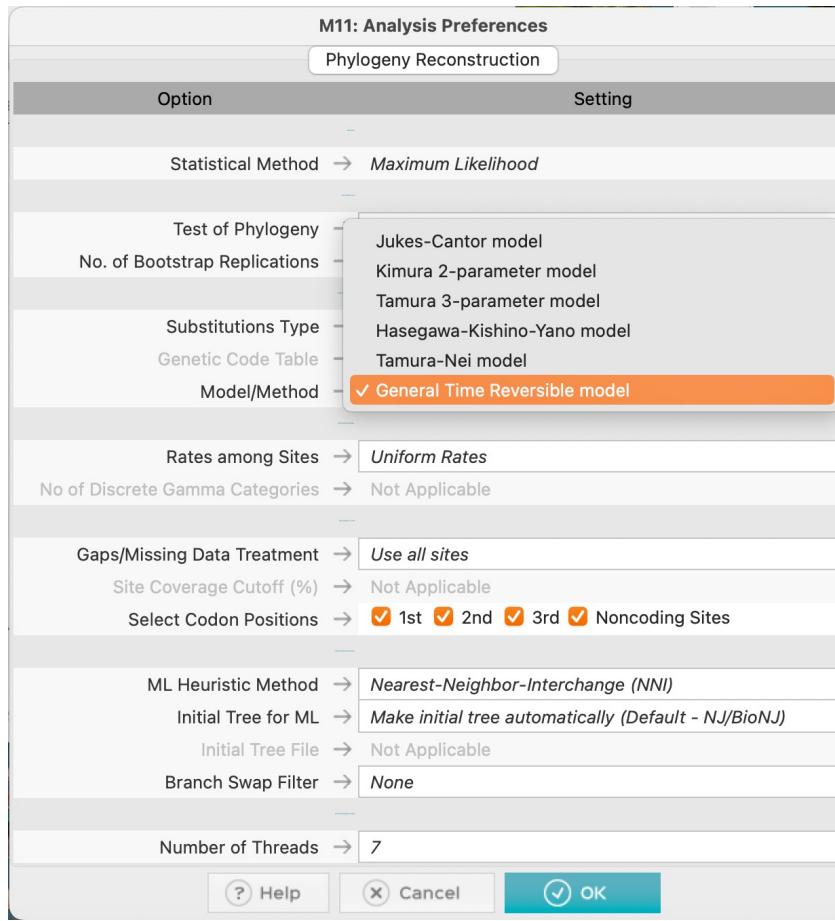
These are the results, according to the description, the lowest BIC score is the best DNA model to use.

We will use the GTR+G+I model

Disclaimer: Although utmost care has been taken to ensure the correctness of the caption, the caption text is provided "as is" without any warranty of any kind. Authors advise the user to carefully check the caption prior to its use for any purpose and report any errors or problems to the authors immediately (www.megasoftware.net). In no event shall the authors and their employers be liable for any damages, including but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties expressed or implied, including but not limited to the determination of suitability of this caption text for a specific purpose, use, or application.

The screenshot shows the MEGA software interface. At the top, there is a toolbar with various icons and labels: ALIGN, DATA, MODELS, DISTANCE, DIVERSITY, RATES, CLOCKS, and DISEASE. Below the toolbar, there is a data panel with a 'Close Data' button. A context menu is open over the 'DATA' icon, listing options: Construct/Test Maximum Likelihood Tree..., Construct/Test Neighbor-Joining Tree..., Construct/Test Minimum-Evolution Tree..., Construct/Test UPGMA Tree..., Construct/Test Maximum Parsimony Tree(s), and Open Tree Session. The 'Construct/Test Maximum Likelihood Tree...' option is highlighted with a red border. To the right of the menu, there are two sections: 'TIMETREE' with a green icon and 'DATAMONKEY' with a blue icon. At the bottom, there is a 'RECENT PUBLICATIONS' section and a footer with links: HELP DOCS, EXAMPLES, CITATION, REPORT BUG, UPDATES, MEGA LINKS, TOOLBAR, PREFERENCES, ANALYZE, PROTOTYPE, and a large red MEGA logo.

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



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

File -> export current tree (Newick)

M11: Tree Explorer (lemur_tree_session.mtsx)

Original Tree Bootstrap Tree

```

graph TD
    Root --- Hapalemur_aureus_cytB[74 Hapalemur aureus cytB]
    Root --- Hapalemur_griseus_meridionalis_cytB[100 Hapalemur griseus meridionalis cytB]
    Hapalemur_aureus_cytB --- Prolemur_simus_cytB[81 Prolemur simus cytB]
    Hapalemur_griseus_meridionalis_cytB --- Eulemur_rubriventer_cytB[69 Eulemur rubriventer cytB]
    Prolemur_simus_cytB --- Eulemur_rufifrons_cytB[100 Eulemur rufifrons cytB]
    Eulemur_rubriventer_cytB --- Varecia_variegata_cytB[71 Varecia variegata cytB]
    Eulemur_rubriventer_cytB --- Propithecus_edwardsi_cytB[47 Propithecus edwardsi cytB]
    Eulemur_rufifrons_cytB --- Cheirogaleus_major_cytB[100 Cheirogaleus major cytB]
    Varecia_variegata_cytB --- Microcebus_rufus_cytB[78 Microcebus rufus cytB]
    Propithecus_edwardsi_cytB --- Cheirogaleus_major_cytB
    Cheirogaleus_major_cytB --- Lepilemur_microdon_cytB[78 Lepilemur microdon cytB]
    Microcebus_rufus_cytB --- Lepilemur_microdon_cytB
    Lepilemur_microdon_cytB --- Daubentonia_madagascariensis_cytB[100 Daubentonia madagascariensis cytB]
    Lepilemur_microdon_cytB --- Homo_sapiens_cytB[Homo sapiens cytB]
  
```

Evolutionary analysis by Maximum Likelihood method
The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model [1]. The tree with the highest log likelihood (-6723.16) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search

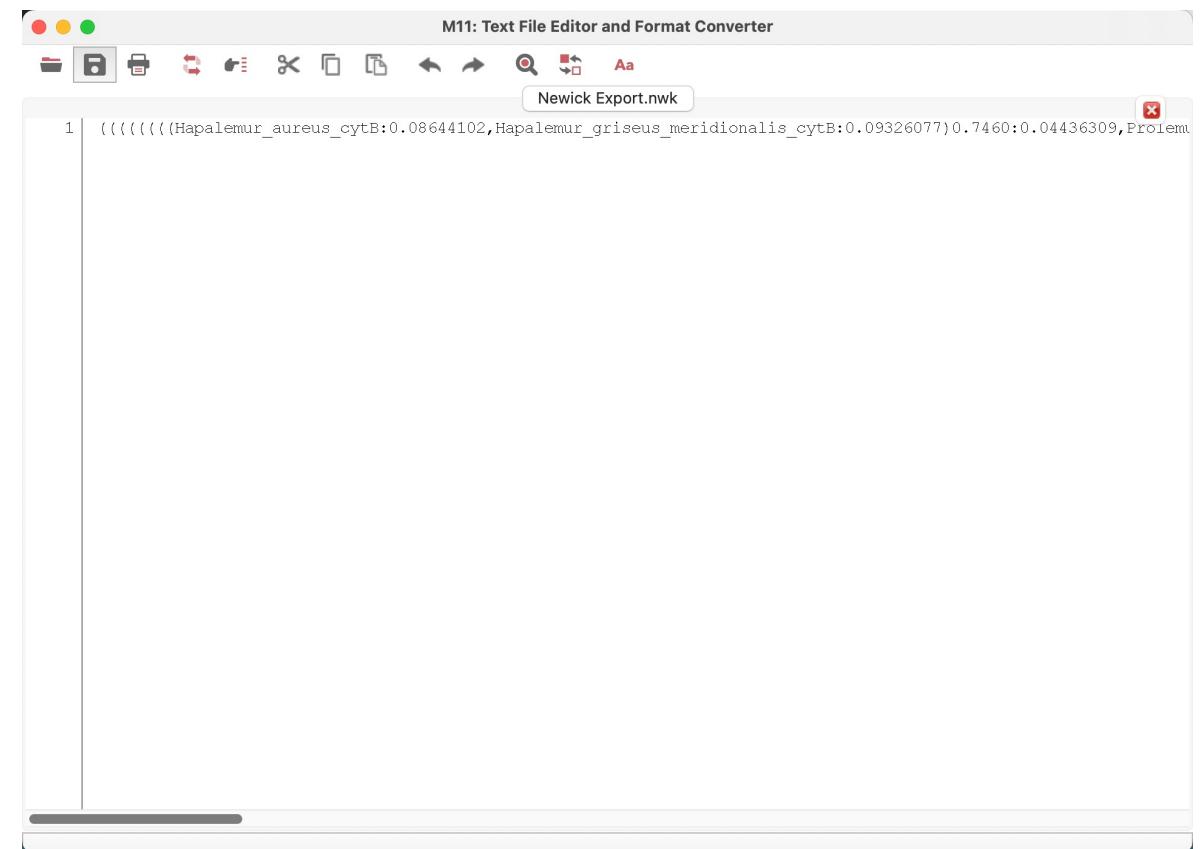
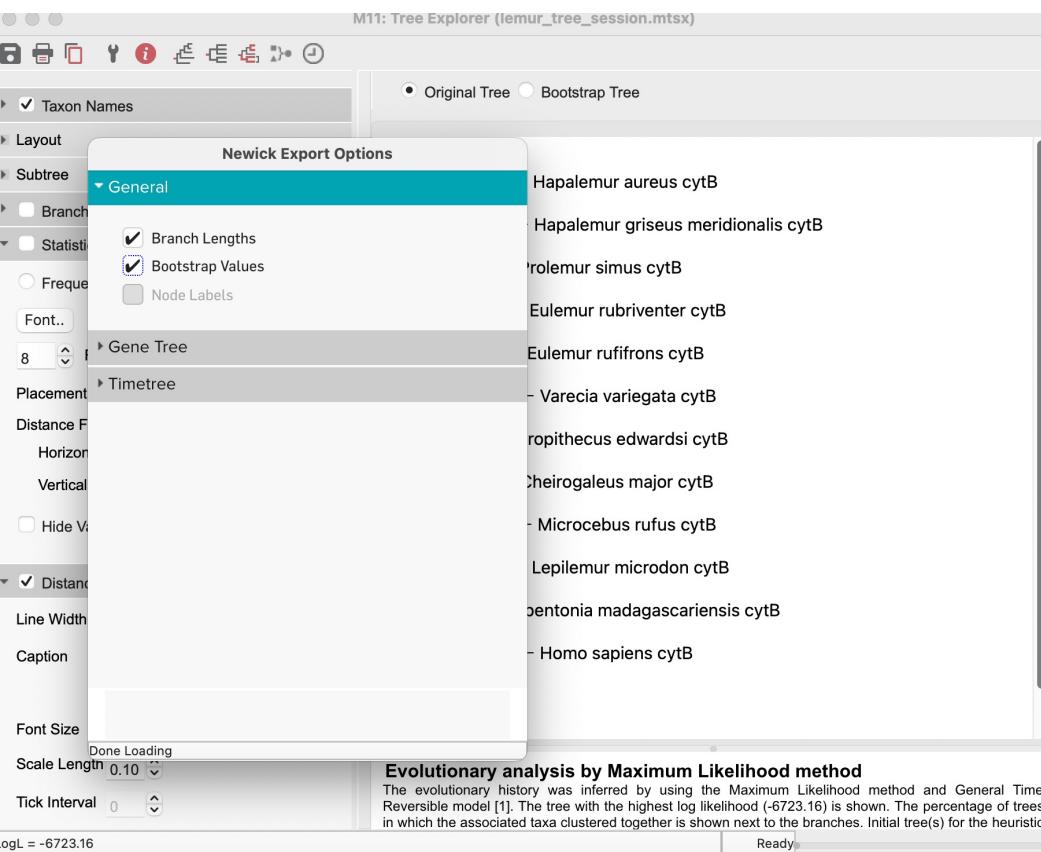
MEGA11 File Search Image Subtree View Compute Ancestors Caption Help

Save Current Session Export Current Tree (Newick) Export Timetree (Tabular) Export Timetree (Nexus) Export All Trees (Newick) Export Analysis Summary Export Partition List Export Pairwise Distances Write Tree in a Table Format Export Group Names Import Group Names Import Name Translations Show Info Print Tree Print in a Sheet Printer Setup... Quit Tree Explorer

Taxon Names Layout Subtree Branch Lengths Statistics/Frequency/Info Frequency Site Coverage Node IDs Font... 8 Font Size Placement Automatic Distance From Node (Pixels) Horizontal 5 Vertical 10 Hide Values Lower Than 0 % Distance Scale Line Width 1 pt Caption Font... Font Size 8 Scale Length 0.10 Tick Interval 0 LogL = -6723.16

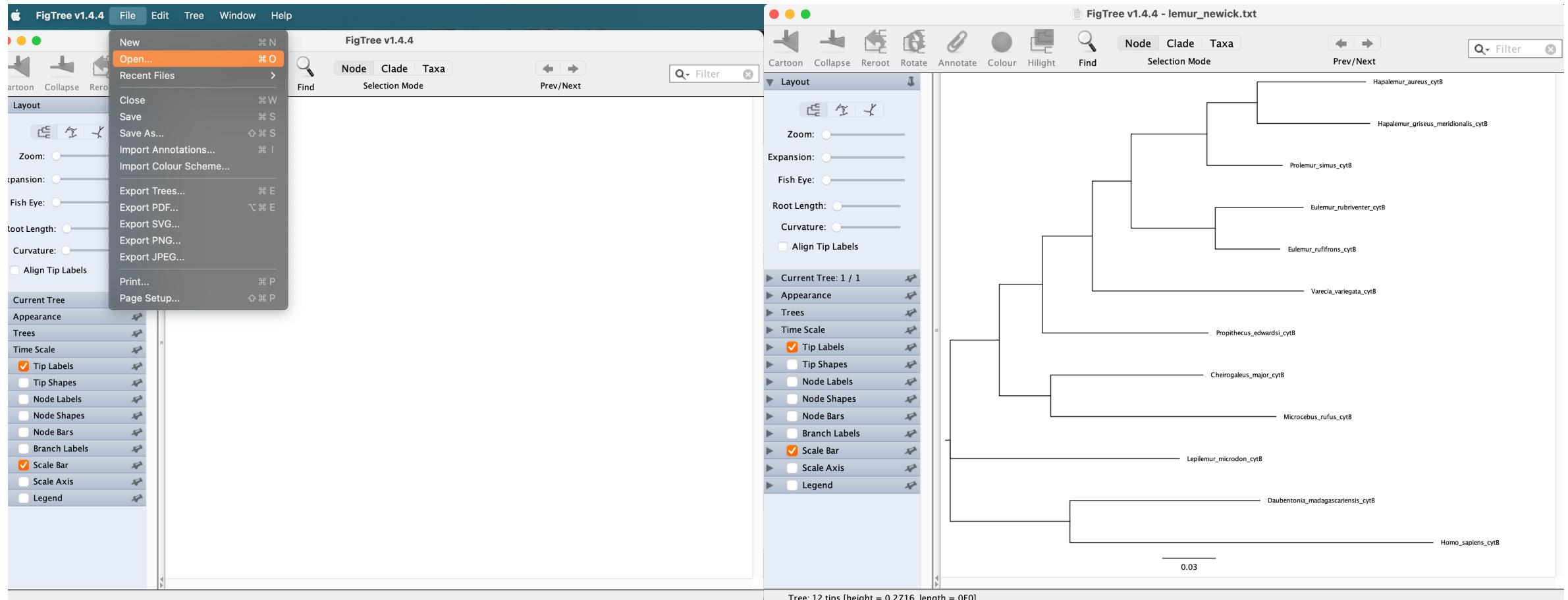
Original Tree Bootstrap Tree

Evolutionary analysis by Maximum Likelihood method
The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model [1]. The tree with the highest log likelihood (-6723.16) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search



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

Good practice to check tree in FigTree first



Then make pretty in R!

- Follow instructions in `lemur_tree_editing_R.R` file