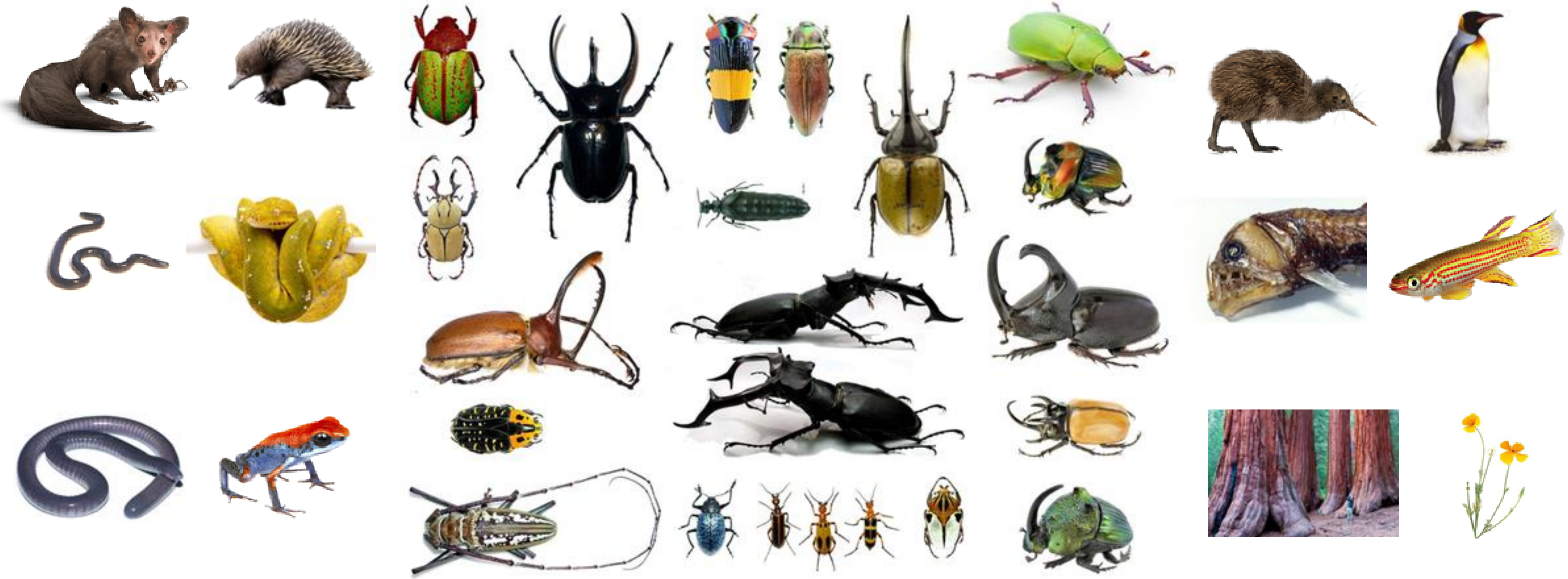
Introduction to Phylogenetics

Heath Blackmon

19 December 2024



Conceptual Foundations 1 hour

- What is a phylogeny
- Why perform phylogenetic inference
- Data for phylogenetics
- Overview of models of evolution
- Methods of inference
- Interpreting trees

Hands-On Tools and Methods 1 hour

- Exercise 1: Running a Simple RAxML Analysis
- Exercise 2: Setting up a Basic BEAST Run
- Exercise 3: Visualizing and evaluating trees in R

Mini-Project and Discussion 1 hour

Conceptual Foundations 1 hour

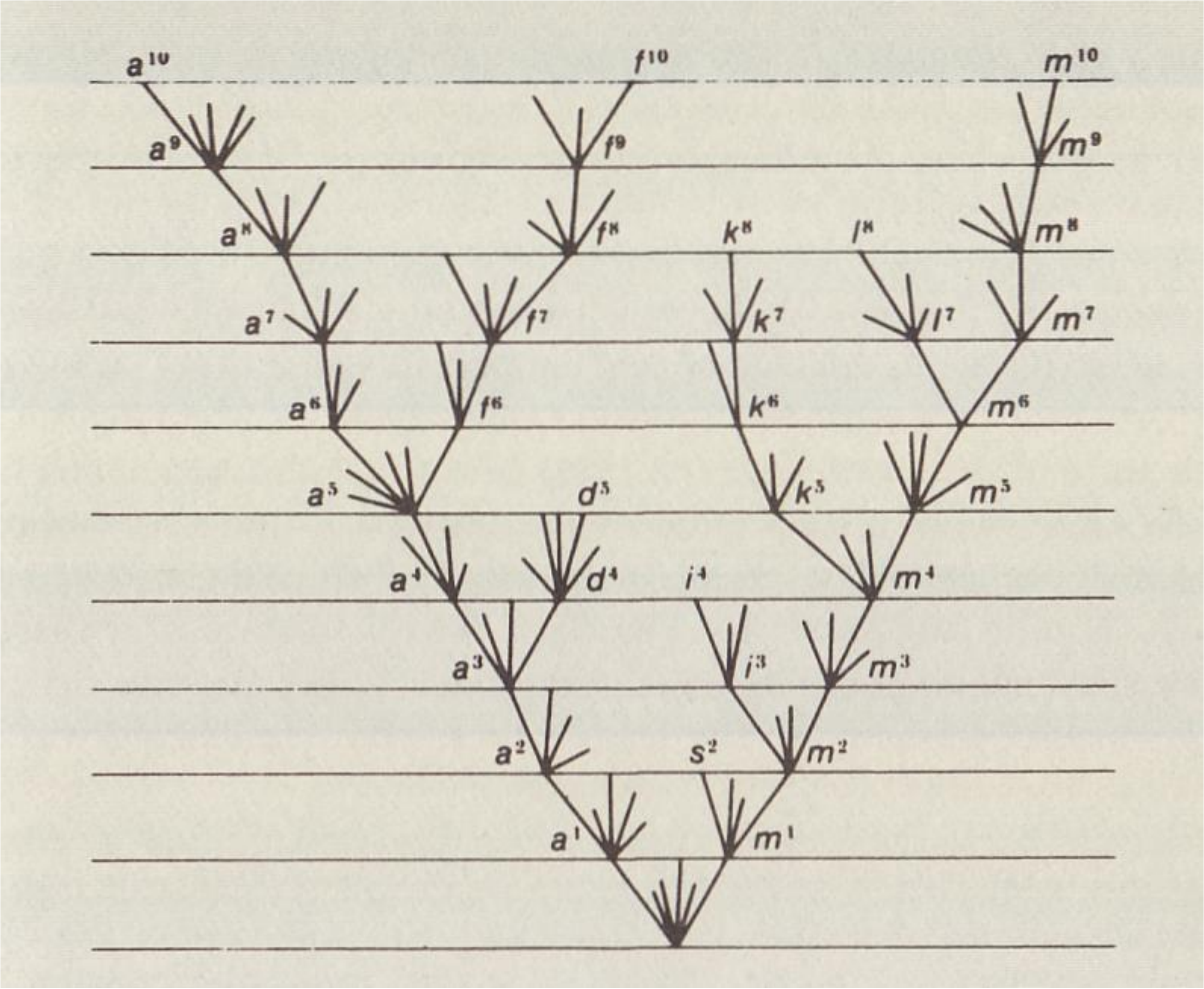
- What is a phylogeny
- Why perform phylogenetic inference
- Data for phylogenetics
- Overview of models of evolution
- Methods of inference
- Interpreting trees

Hands-On Tools and Methods 1 hour

- Exercise 1: Running a Simple RAxML Analysis
- Exercise 2: Setting up a Basic BEAST Run
- Exercise 3: Visualizing and evaluating trees in R

Mini-Project and Discussion 1 hour

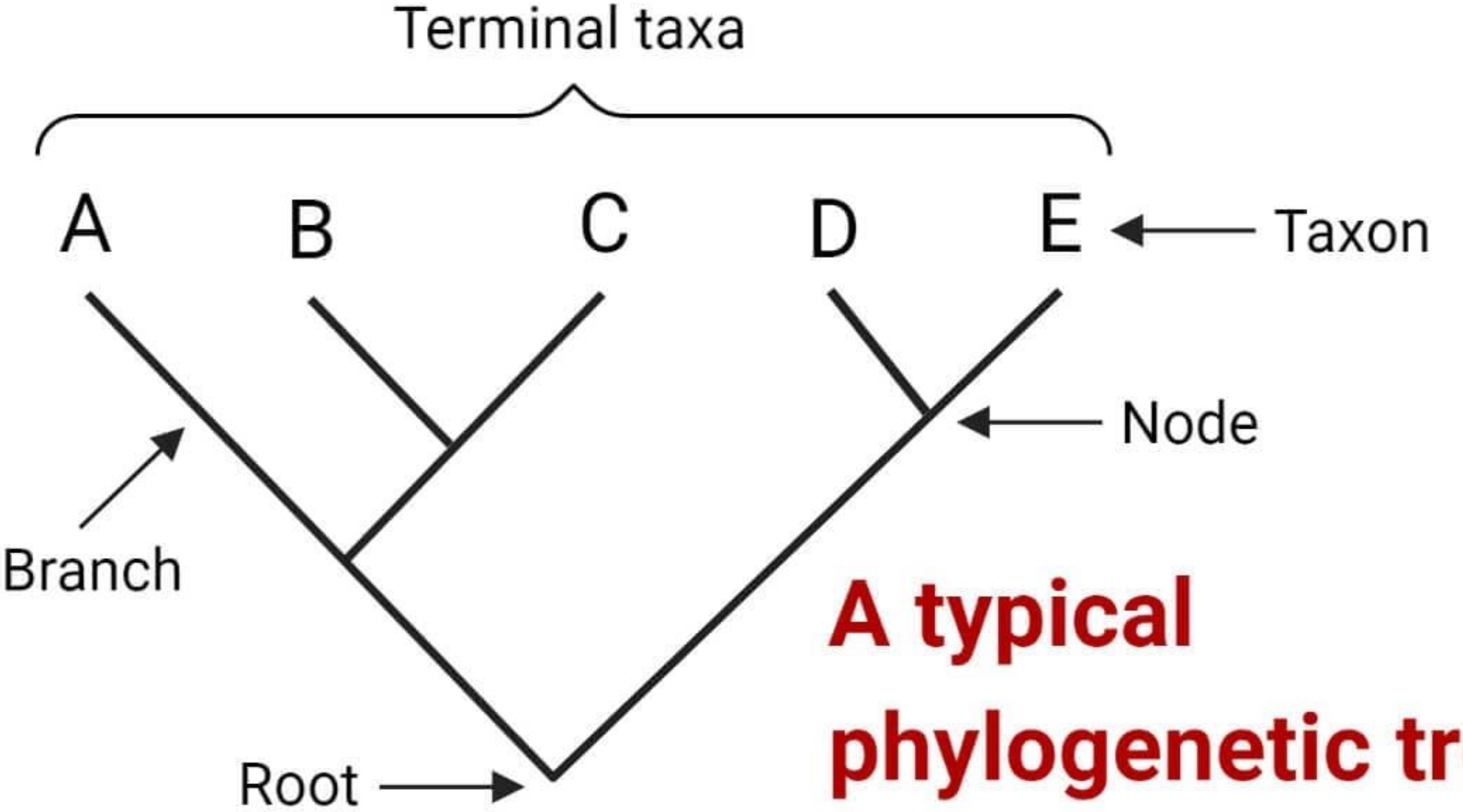
There is a beauty in phylogenies



There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

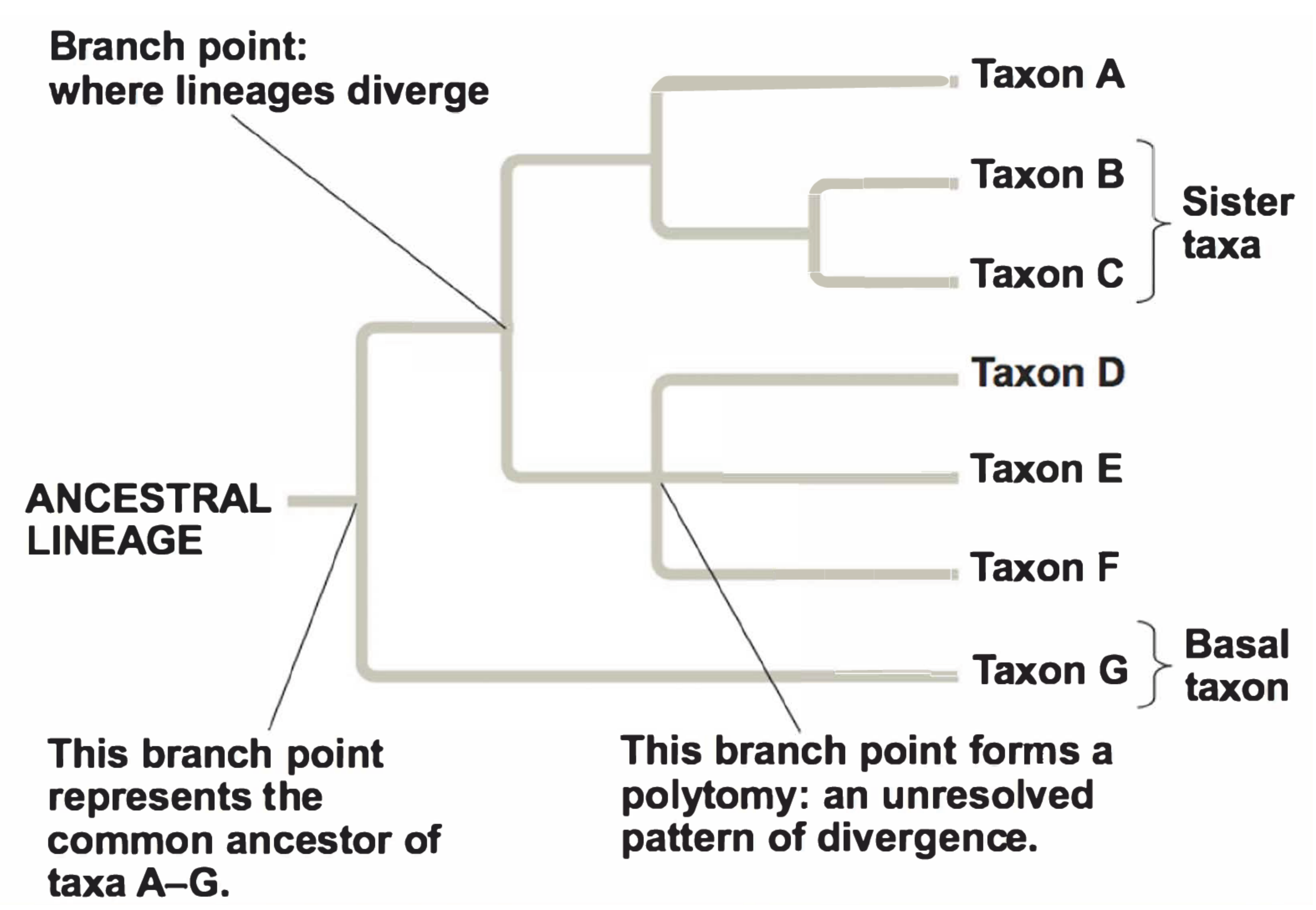
Darwin 1859

What is a phylogeny

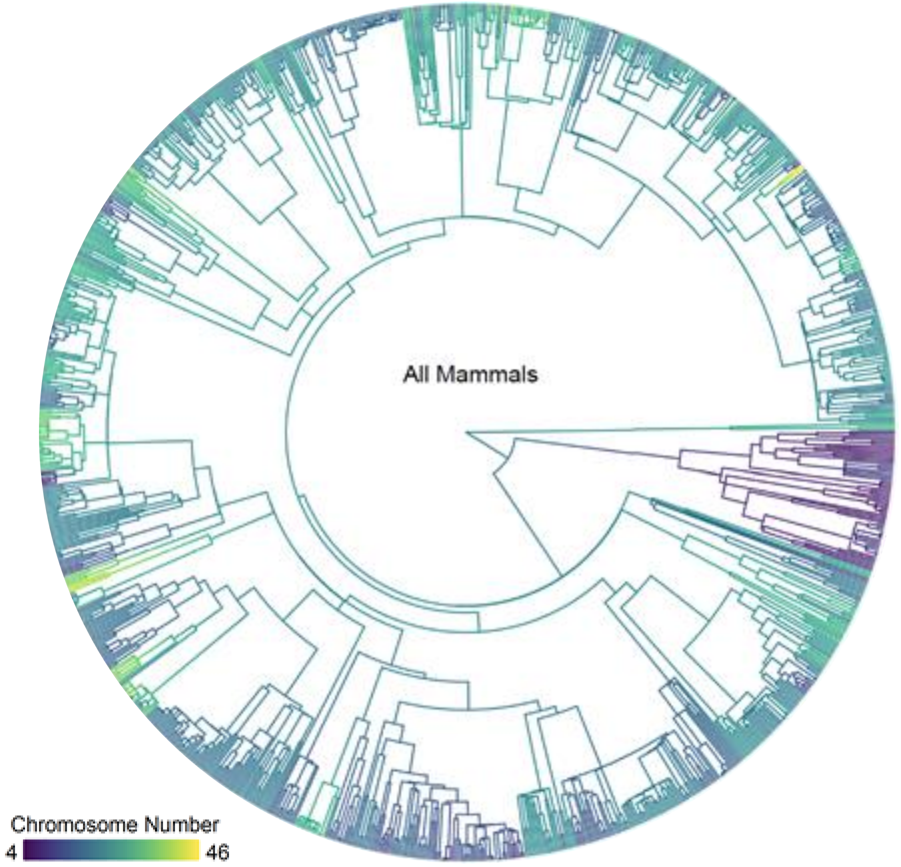


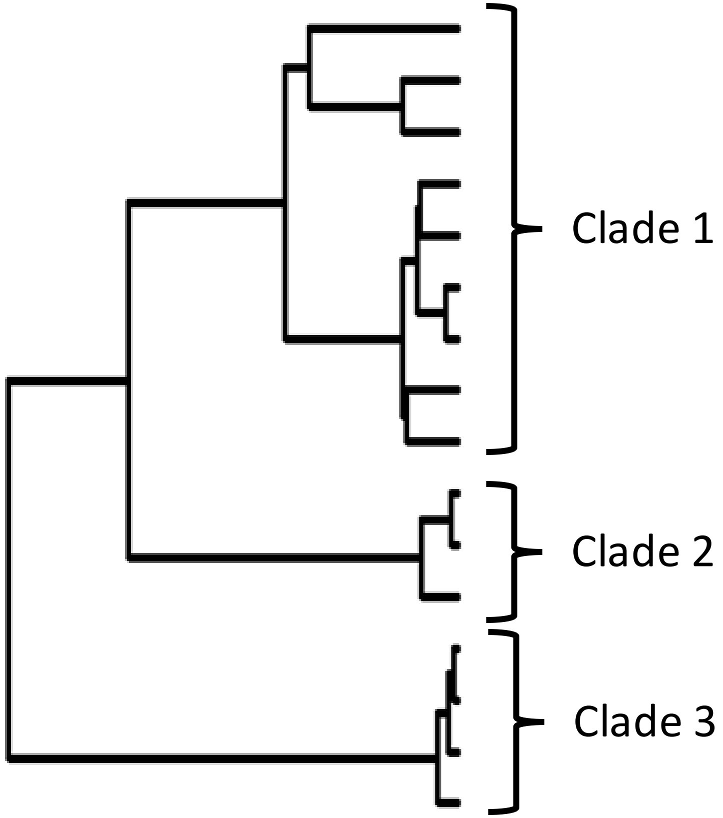
A typical phylogenetic tree

What is a phylogeny


Branch point:
where lineages diverge
Taxon A
Taxon B
Taxon C
Sister taxa
Taxon D
Taxon E
Taxon F
Taxon G
Basal taxon
ANCESTRAL LINEAGE
This branch point represents the common ancestor of taxa A–G.
This branch point forms a polytomy: an unresolved pattern of divergence.

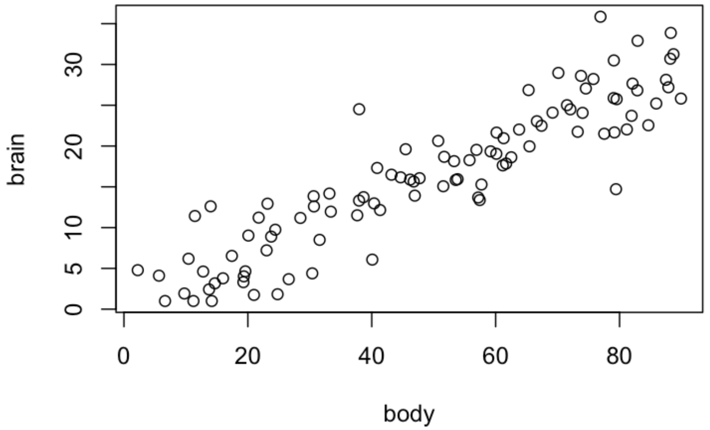
Why infer a phylogeny



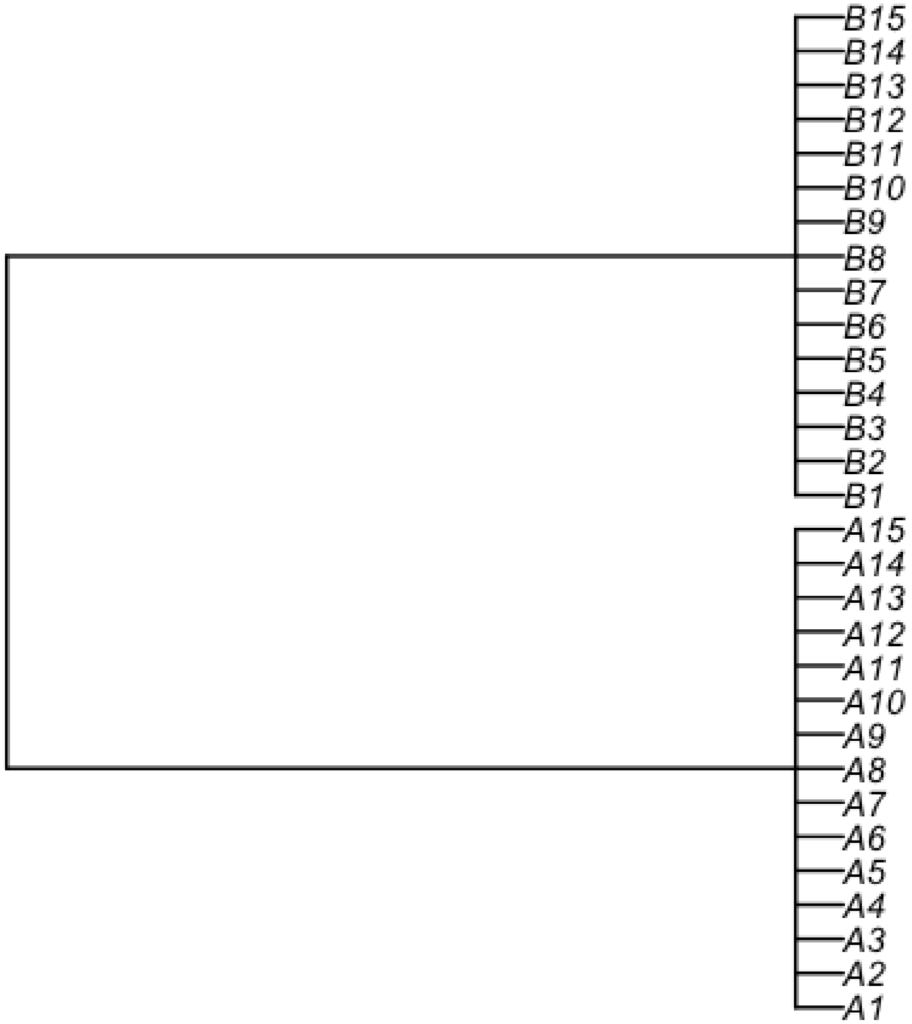

Clade 1

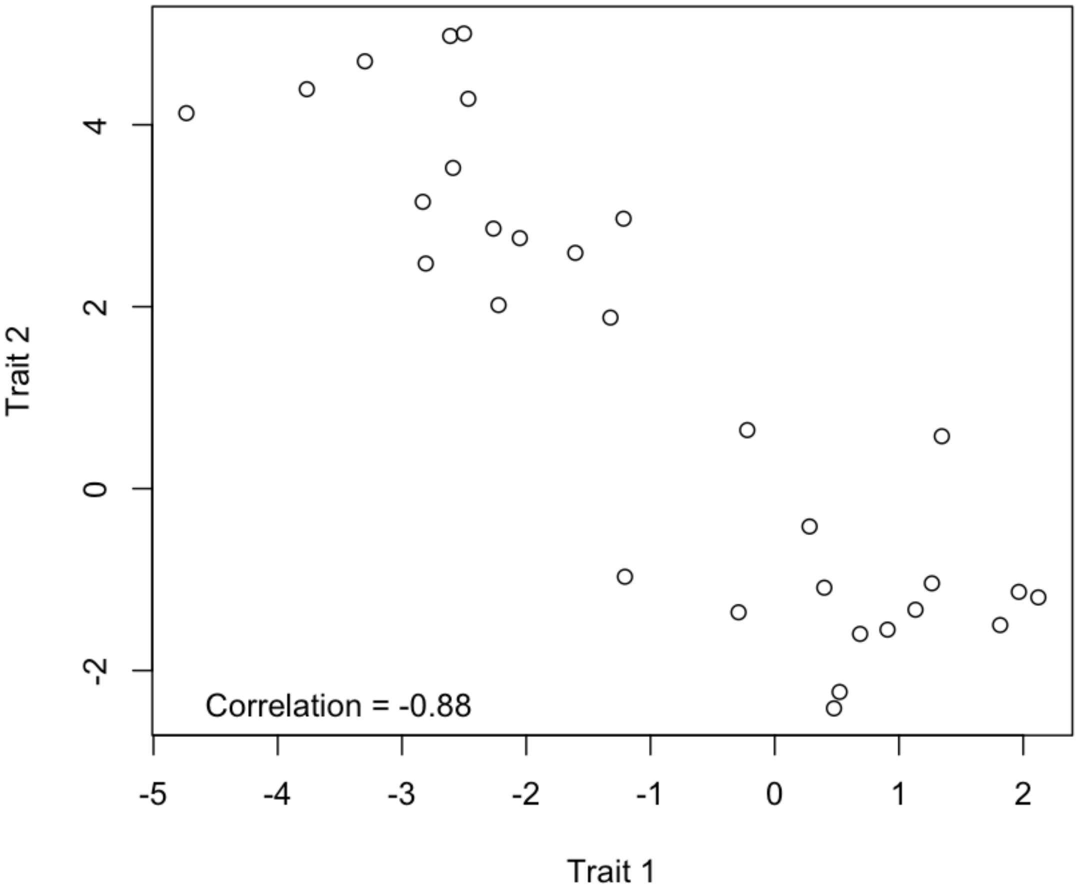
Clade 2

Clade 3

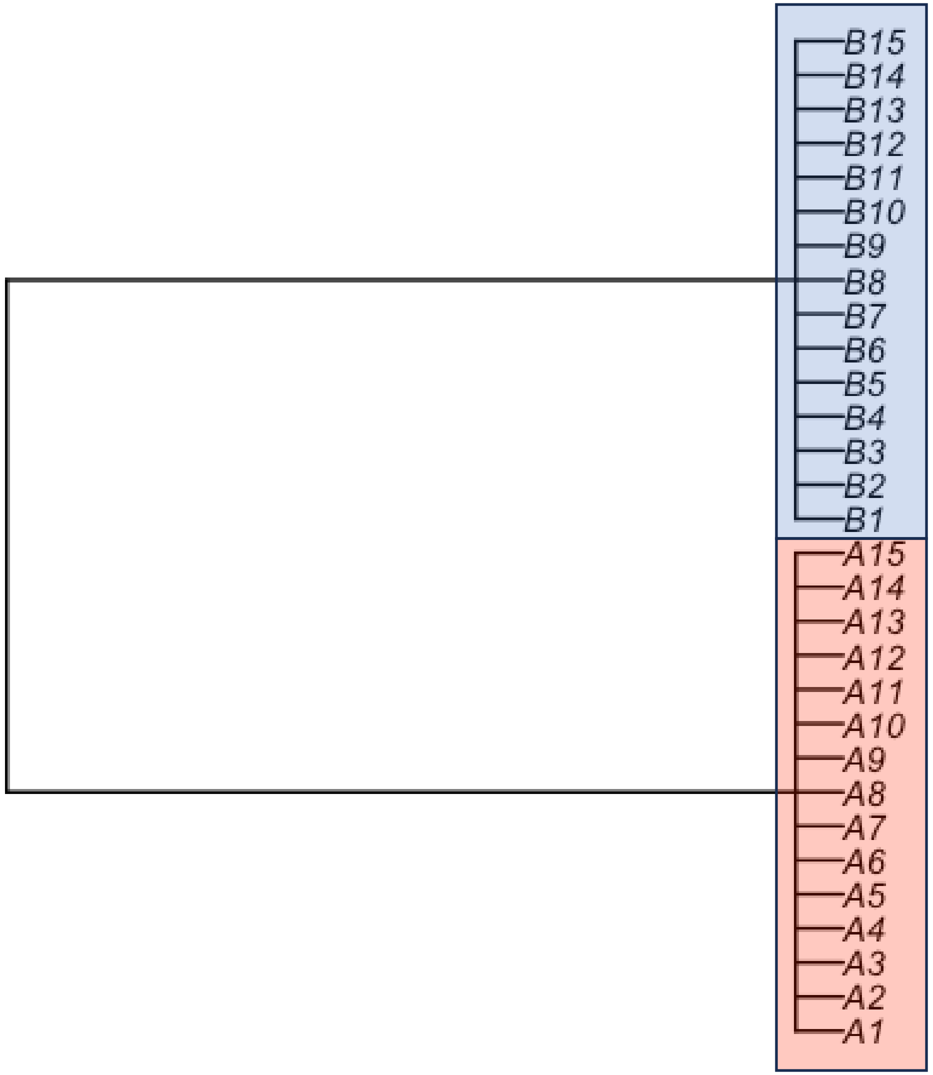


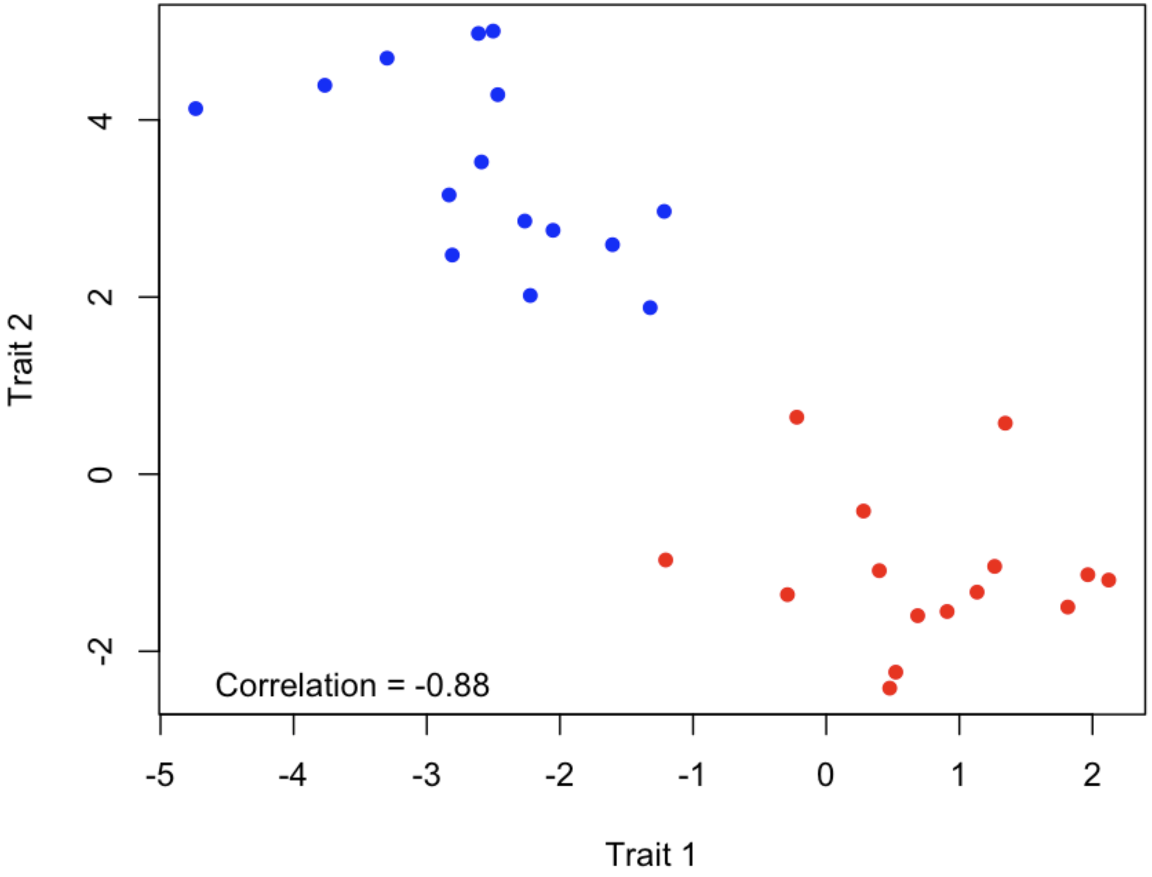
Why infer a phylogeny



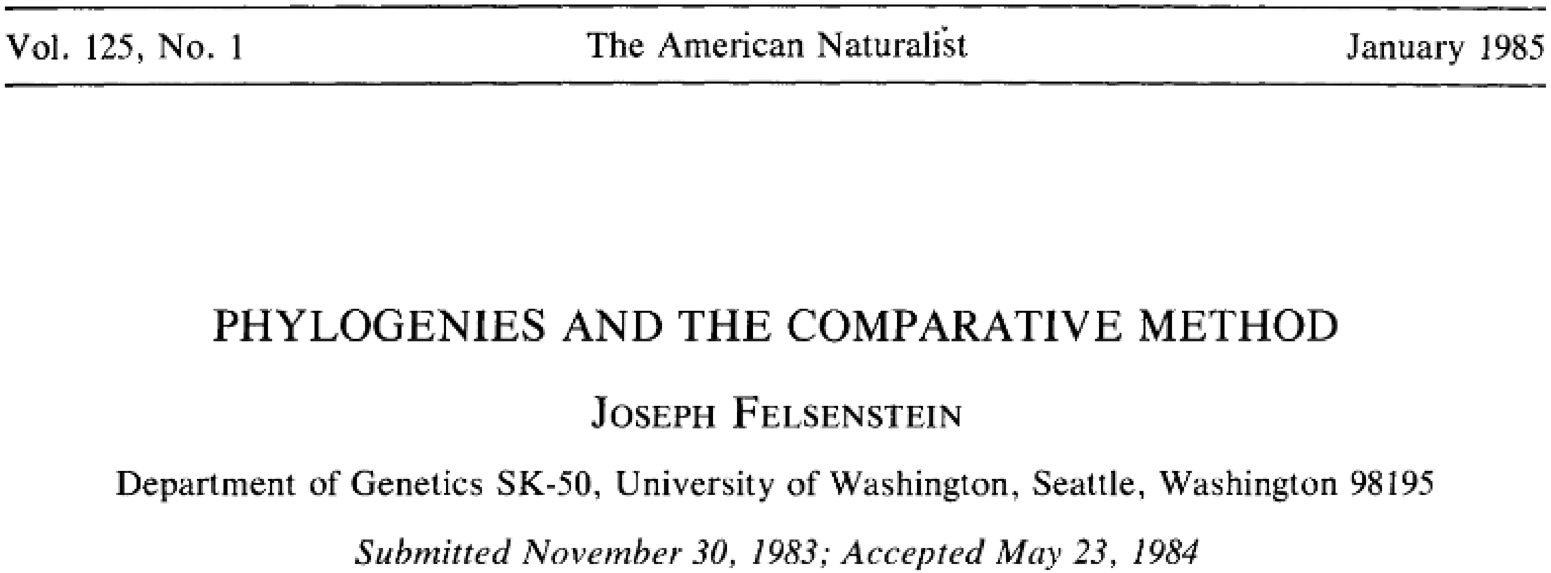


Why infer a phylogeny





Why infer a phylogeny



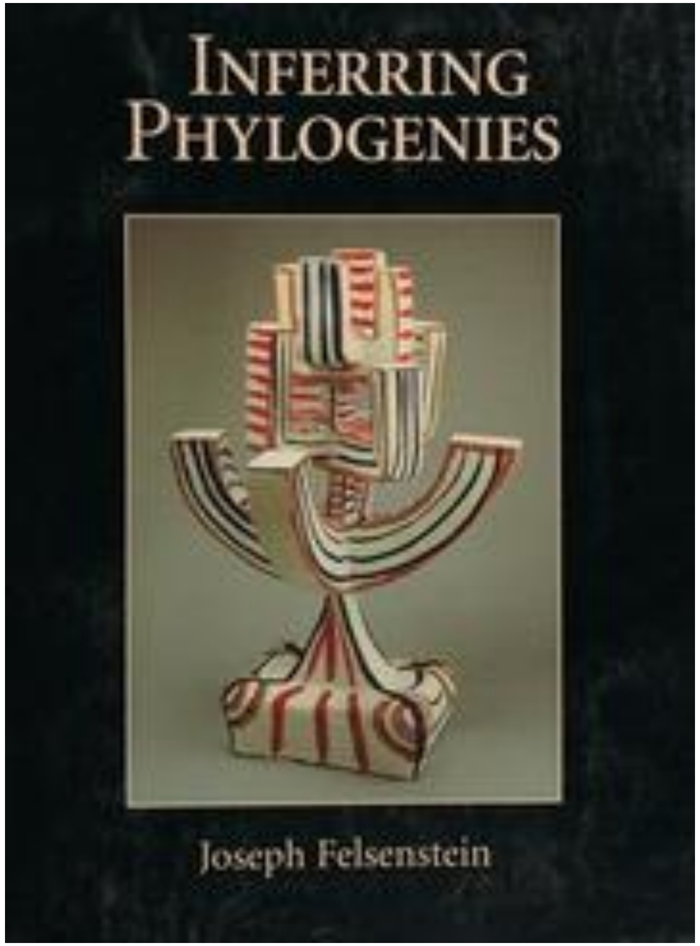
Vol. 125, No. 1 The American Naturalist January 1985

PHYLOGENIES AND THE COMPARATIVE METHOD

JOSEPH FELSENSTEIN

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195

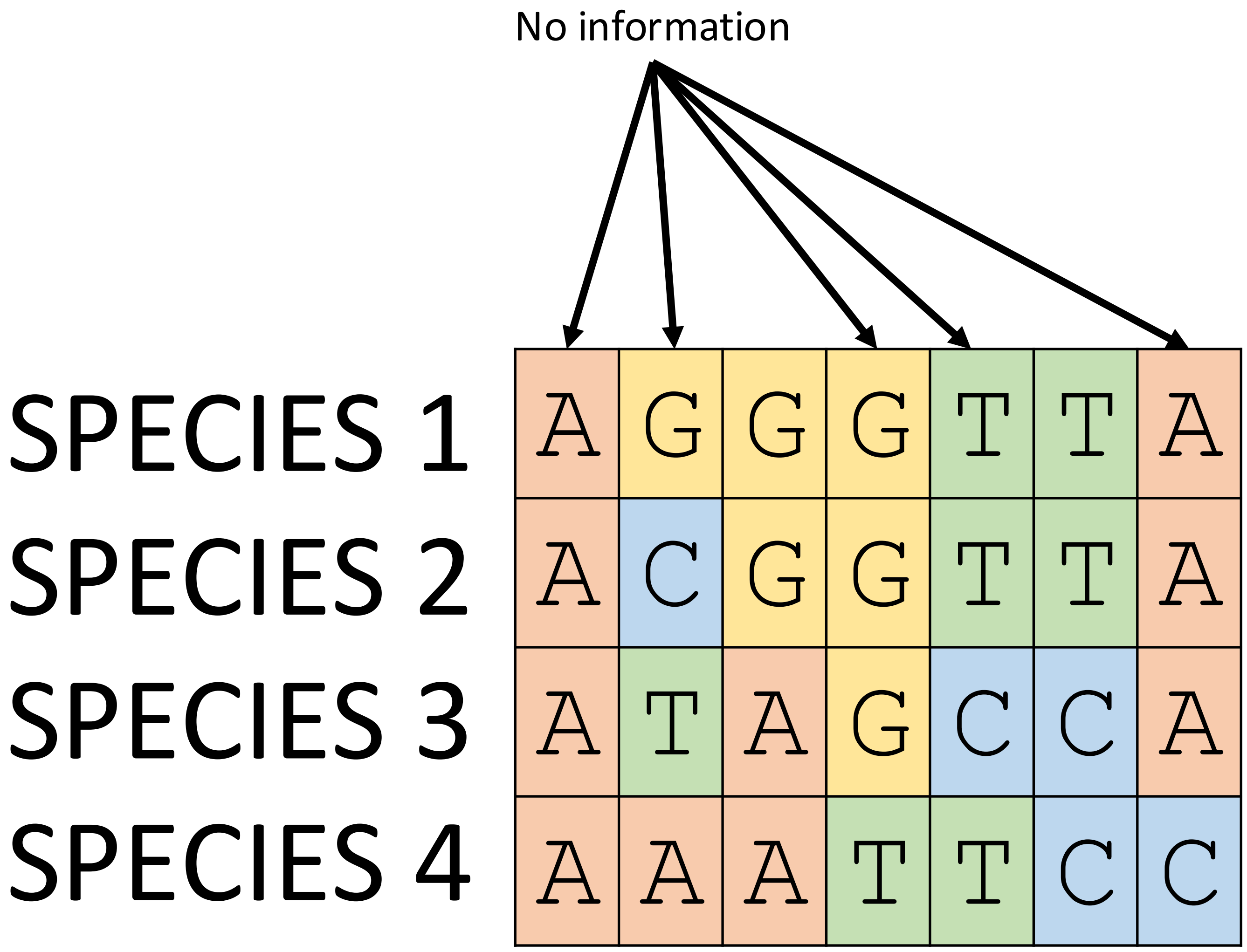
Submitted November 30, 1983; Accepted May 23, 1984



INFERRING PHYLOGENIES

Joseph Felsenstein

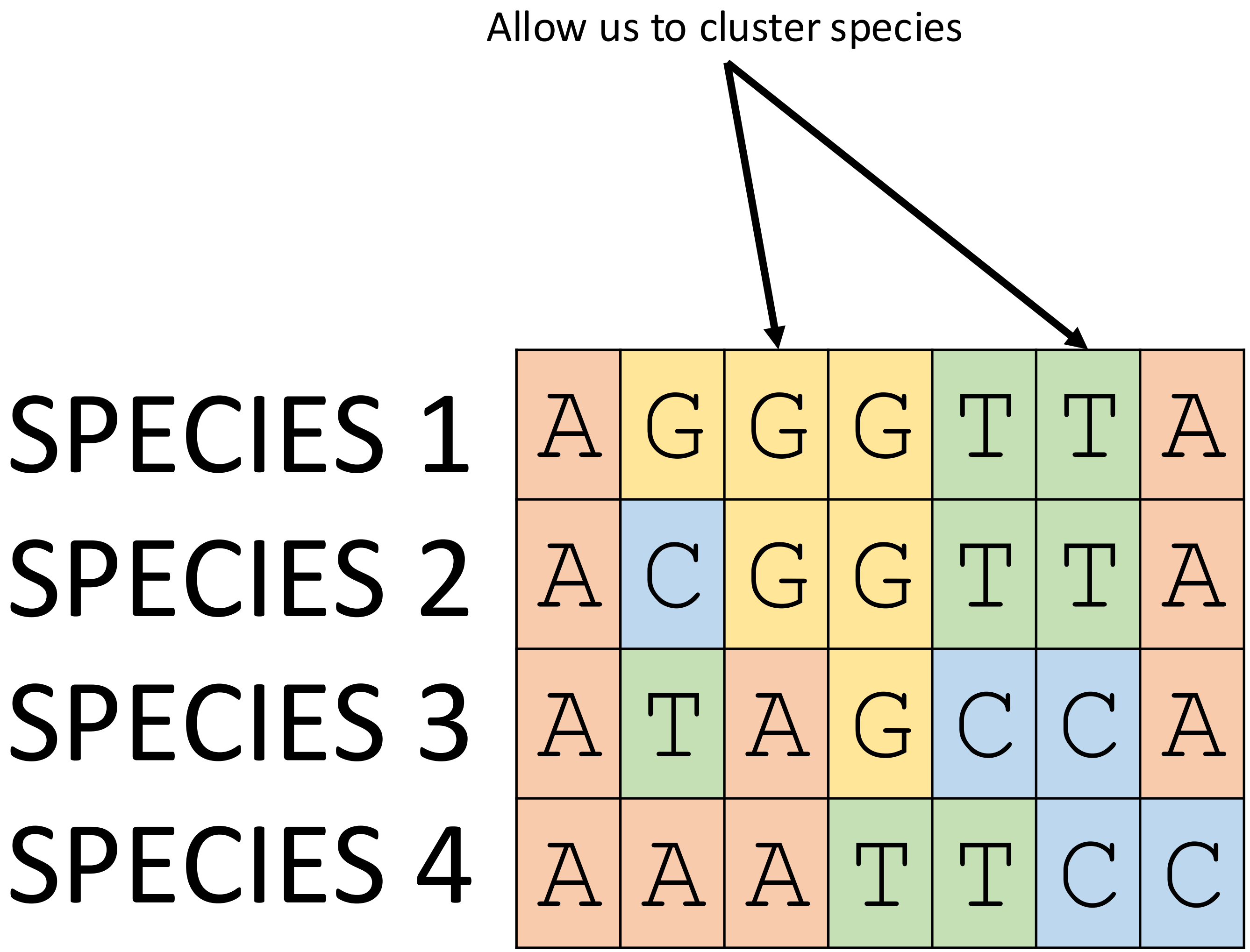
Data to infer a phylogeny



No information

SPECIES 1	A	G	G	G	T	T	A
SPECIES 2	A	C	G	G	T	T	A
SPECIES 3	A	T	A	G	C	C	A
SPECIES 4	A	A	A	T	T	C	C

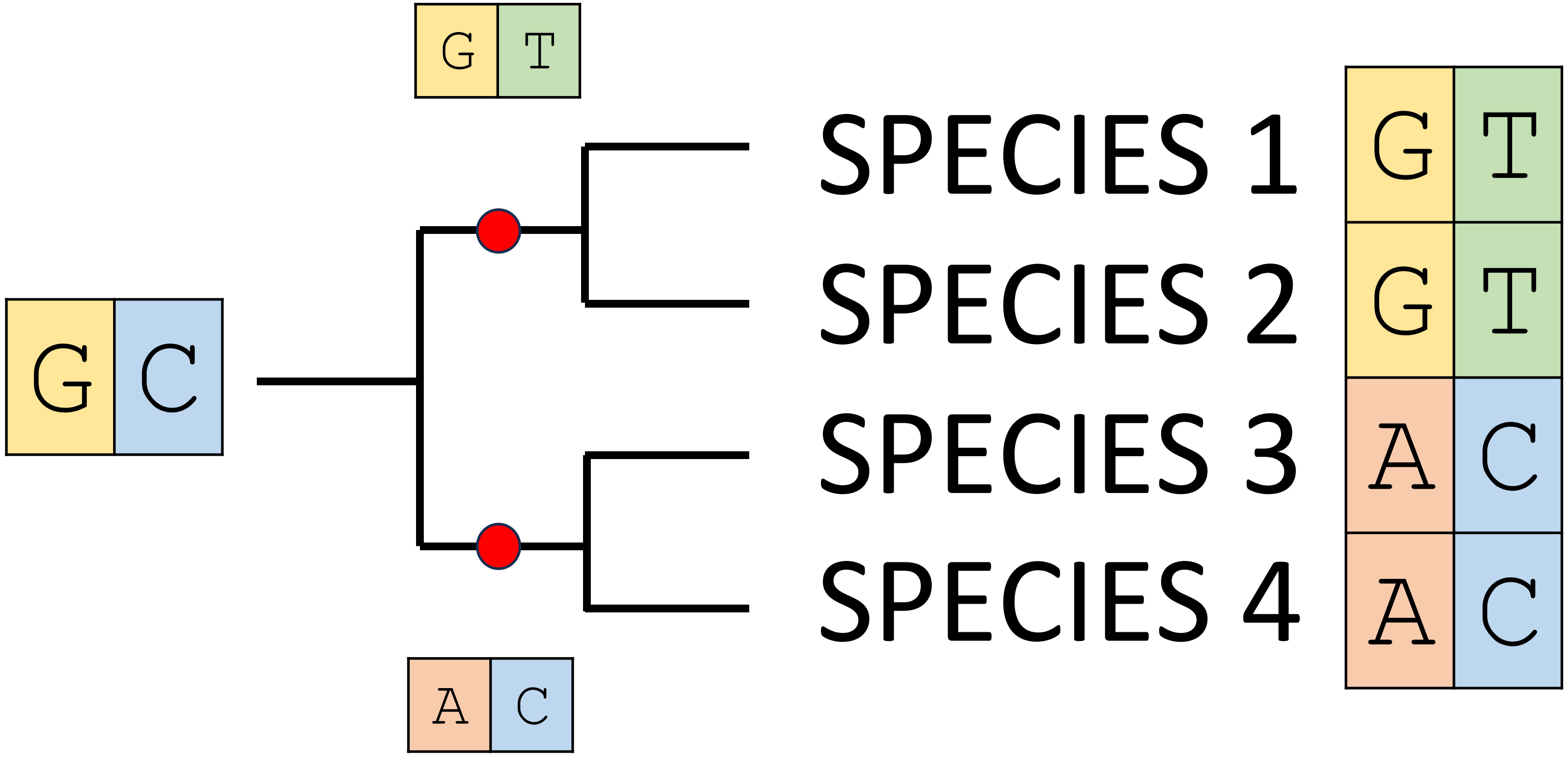
Data to infer a phylogeny



Allow us to cluster species

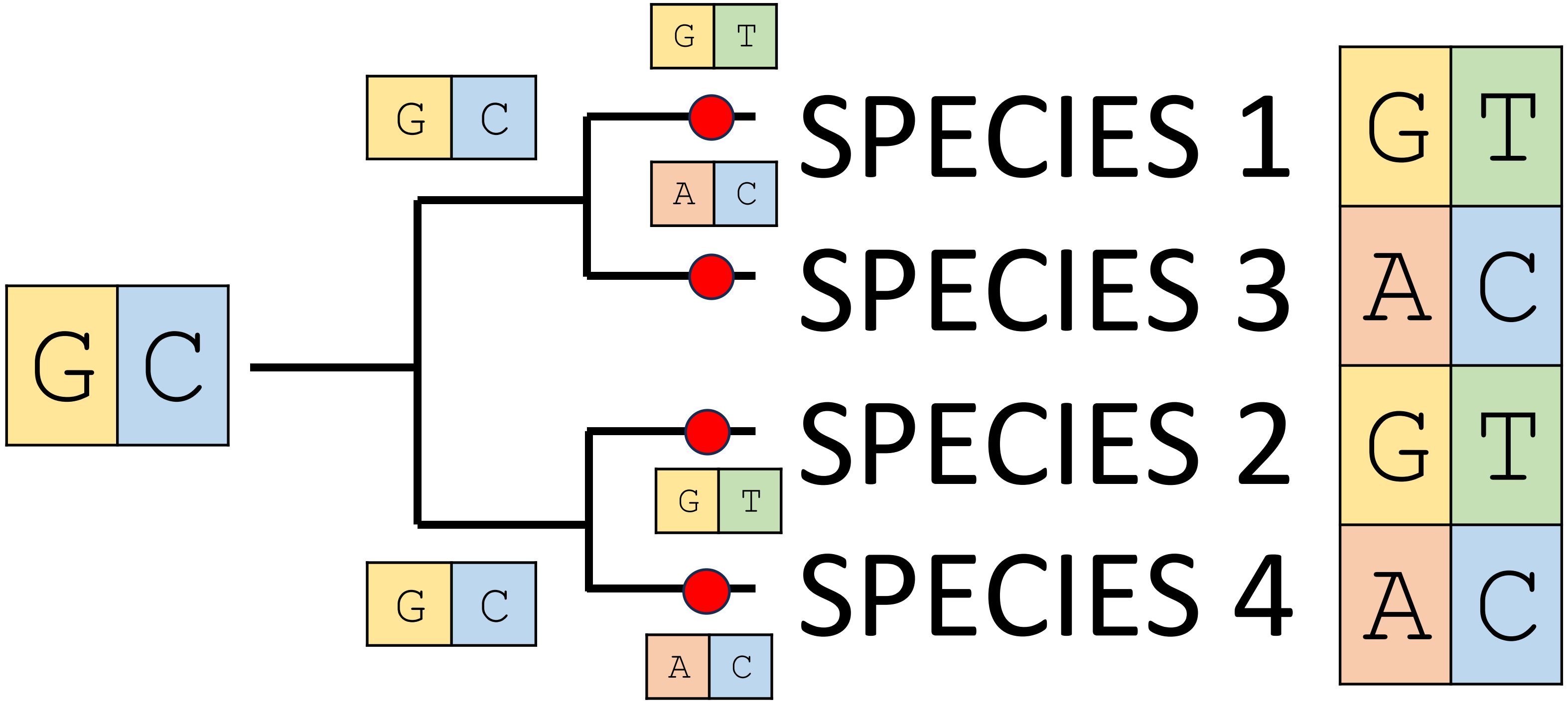
SPECIES 1	A	G	G	G	T	T	A
SPECIES 2	A	C	G	G	T	T	A
SPECIES 3	A	T	A	G	C	C	A
SPECIES 4	A	A	A	T	T	C	C

Data to infer a phylogeny

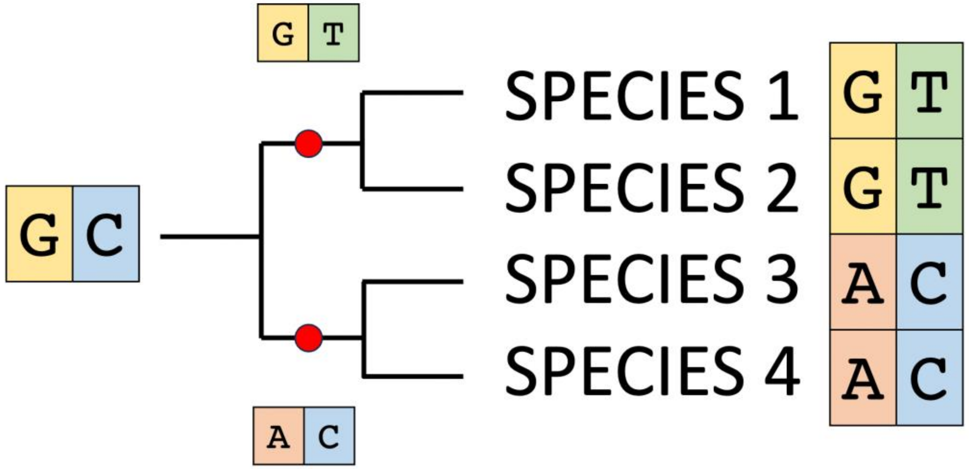


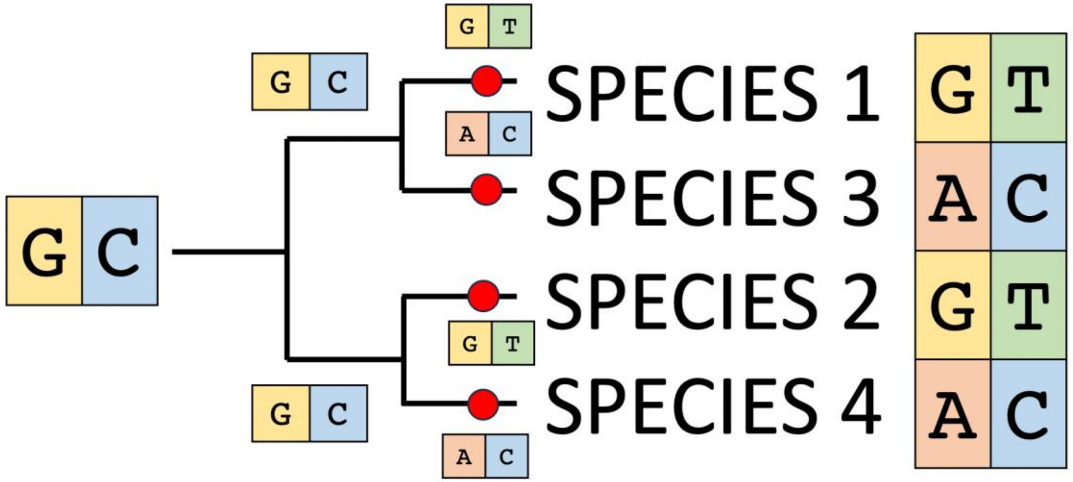
SPECIES 1	G	T
SPECIES 2	G	T
SPECIES 3	A	C
SPECIES 4	A	C

Data to infer a phylogeny

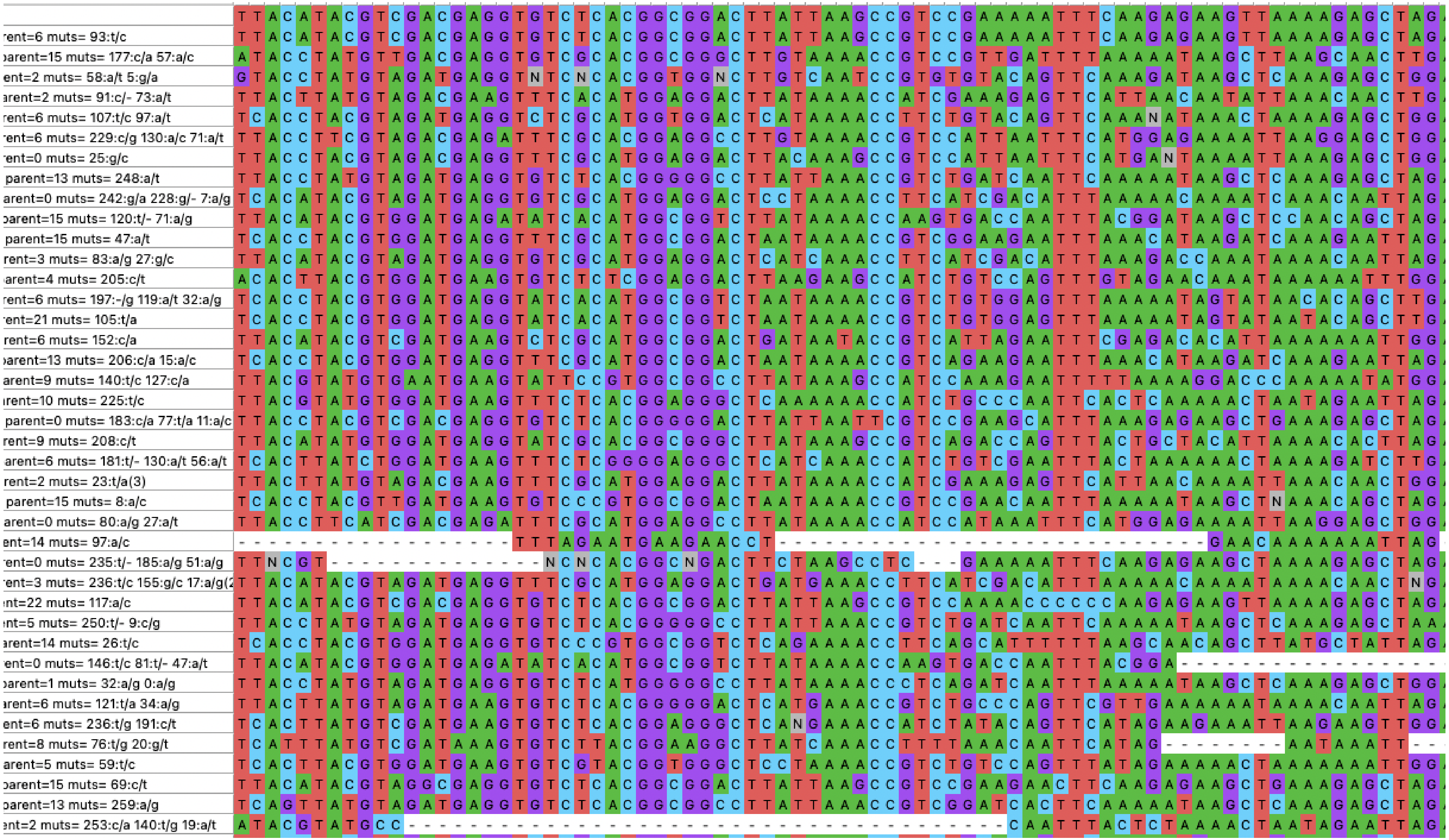


Data to infer a phylogeny



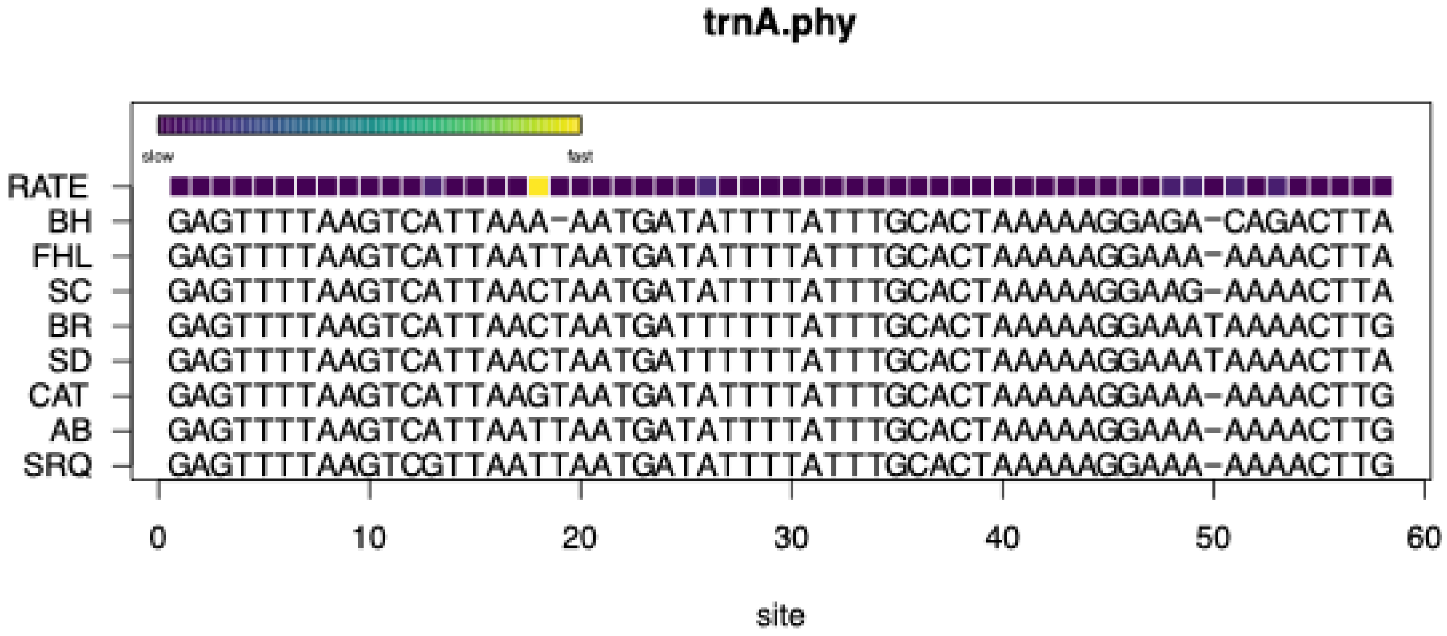


Real alignments can be messy

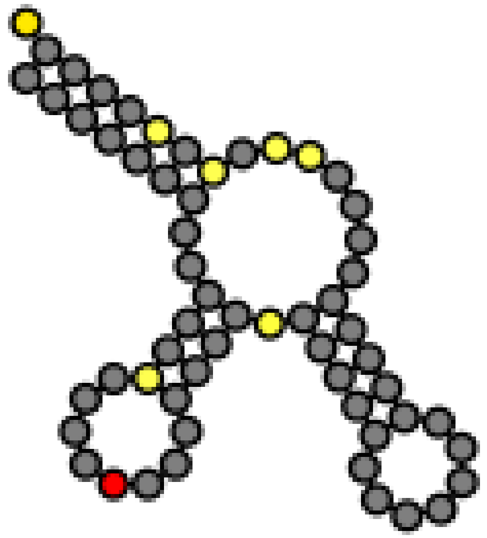


```
                                               TTACATACGTCGACGAGGTGTCTCACGGCGGACTTATTAAGCCGTCCGAAAAATTTCAAGAGAAGTTAAAAGAGCTAG
rent=6 muts= 93:t/c                            TTACATACGTCGACGAGGTGTCTCACGGCGGACTTATTAAGCCGTCCGAAAAATTTCAAGAGAAGTTAAAAGAGCTAG
parent=15 muts= 177:c/a 57:a/c                 ATACCTATGTTGACGAGGTGTCGCACGGCGGGCTTGTAAAACCGTCCGTTGATTTTAAAAATAAGCTTAAGCAACTTG
ent=2 muts= 58:a/t 5:g/a                       GTACCTATGTAGATGAGGTNTCNCACGGTGGNCTTGTCAATCCGTGTGTACAGTTCAAAGATAAGCTCAAAGAGCTGG
arent=2 muts= 91:c/- 73:a/t                    TTACTTATGTAGACGAAGTTTCACATGGAGGACTTATAAAACCATCGAAAGAGTTCATTAACAATATTAAACAACTTG
rent=6 muts= 107:t/c 97:a/t                    TCACCTACGTAGATGAGGTCTCGCATGGTGGACTCATAAAACCTTCTGTACAGTTCAAANATAAACTAAAAGAGCTGG
rent=6 muts= 229:c/g 130:a/c 71:a/t            TTACCTTCGTAGACGAGATTTCGCACGGAGGCCTTGTAAAACCGTCCATTAATTTCATGGAGAAAATTAAGGAGCTGG
rent=0 muts= 25:g/c                            TTACCTACGTAGACGAGGTTTCGCATGGAGGACTTACAAAGCCGTCCATTAATTTCATGANTAAAATTAAAGAGCTGG
parent=13 muts= 248:a/t                        TTACCTATGTAGATGAGGTGTCTCACGGGGGCCTTATTAAACCGTCTGATCAATTCAAAAATAAGCTCAAAGAGCTAG
arent=0 muts= 242:g/a 228:g/- 7:a/g            TCACATACGTAGATGAGGTGTCGCATGGAGGACTCCTAAAACCTTCATCGACATTTAAAAACAAAATCAAACAATTAG
parent=15 muts= 120:t/- 71:a/g                 TTACATACGTGGATGAGATATCACATGGCGGTCTTATAAAACCAAGTGACCAATTTACGGATAAGCTCCAACAGCTAG
parent=15 muts= 47:a/t                         TCACCTACGTGGATGAGGTTTCGCATGGCGGACTAATAAAACCGTCGGAAGAATTTAAACATAAGATCAAAGAATTAG
rent=3 muts= 83:a/g 27:g/c                     TTACATACGTAGATGAGGTGTCGCATGGAGGACTCATCAAACCTTCATCGACATTTAAAGACCAAATAAAACAATTAG
arent=4 muts= 205:c/t                          ACACTTACGTGGATGAAGTGTCTCTCGGAGGACTTAAGAAGCCATCTGTCCAGTTTGTAGAACAAATAAAAAATTTGG
rent=6 muts= 197:-/g 119:a/t 32:a/g            TCACCTACGTGGATGAGGTATCACATGGCGGTCTAATAAAACCGTCTGTGGAGTTTAAAAATAGTATAACACAGCTTG
rent=21 muts= 105:t/a                          TCACCTACGTGGATGAGGTATCACATGGCGGTCTAATAAAACCGTCTGTGGAGTTTAAAAATAGTATAATACAGCTTG
rent=6 muts= 152:c/a                           TTACATACGTCGATGAAGTCTCGCATGGCGGACTGATAATACCGTCATTAGAATTTCGAGACACATTAAAAAAATTGG
parent=13 muts= 206:c/a 15:a/c                 TCACCTACGTGGATGAGGTTTCGCATGGCGGACTAATAAAACCGTCAGAAGAATTTAAACATAAGATCAAAGAATTAG
arent=9 muts= 140:t/c 127:c/a                  TTACGTATGTGAATGAAGTATTCCGTGGCGGCCTTATAAAGCCATCCAAAGAATTTTTAAAAGGACCCAAAAATATGG
rent=10 muts= 225:t/c                          TTACGTATGTGGATGAAGTTTCTCACGGAGGGCTCAAAAAACCATCTGCCCAATTCACTCAAAAACTAATAGAATTAG
parent=0 muts= 183:c/a 77:t/a 11:a/c           TTACCTACGTCGACGAGGTGTCTCACGGGGGACTTATTAATTCGTCCGAAGCATTTAAAGAGAAGCTGAAAGAGCTAG
rent=9 muts= 208:c/t                           TTACATATGTGGATGAGGTATCGCACGGCGGGCTTATAAAGCCGTCAGACCAGTTTACTGCTACATTAAAACACTTAG
arent=6 muts= 181:t/- 130:a/t 56:a/t           TCACTTATCTGGATGAAGTTTCTCGGGGAGGGCTCATCAAACCATCTGTCGAATTTACTAAAAAACTAAAAGATCTTG
rent=2 muts= 23:t/a(3)                         TTACTTATGTAGACGAAGTTTCGCATGGAGGACTTATAAAACCATCGAAAGAGTTCATTAACAAAATTAAACAACTGG
parent=15 muts= 8:a/c                          TCACCTACGTTGATGAAGTGTCCCGTGGCGGACTAATAAAACCGTCCGAACAATTTAAAAATAAGCTNAAACAGCTAG
arent=0 muts= 80:a/g 27:a/t                    TTACCTTCATCGACGAGATTTCGCATGGAGGCCTTATAAAACCATCCATAAATTTCATGGAGAAAATTAAGGAGCTGG
ent=14 muts= 97:a/c                            ------------------TTTAGAATGAAGAACCT----------------------------GAACAAAAAAATTAG
rent=0 muts= 235:t/- 185:a/g 51:a/g            TTNCGT--------------NCNCACGGCNGACTTCTAAGCCTC---GAAAAATTTCAAGAGAAGCTAAAAGAGCTAG
rent=3 muts= 236:t/c 155:g/c 17:a/g(2          TTACATACGTAGATGAGGTTTCGCATGGAGGACTGATGAAACCTTCATCGACATTTAAAAACAAAATAAAACAACTNG
nt=22 muts= 117:a/c                            TTACATACGTCGACGAGGTGTCTCACGGCGGACTTATTAAGCCGTCCAAAACCCCCCAAGAGAAGTTAAAAGAGCTAG
nt=5 muts= 250:t/- 9:c/g                       TTACCTATGTAGATGAGGTGTCTCACGGGGGCCTTATTAAACCGTCTGATCAATTCAAAAATAAGCTCAAAGAGCTAA
arent=14 muts= 26:t/c                          TCACCTACGTGGATGAGGTGTCCCGTGGCGGTCTCAGAAAACCTTCAGCATTTTTTAAGCAACAGCTTATGCTATTAG
ent=0 muts= 146:t/c 81:t/- 47:a/t              TTACATACGTGGATGAGATATCACATGGCGGTCTTATAAAACCAAGTGACCAATTTACGGA-----------------
parent=1 muts= 32:a/g 0:a/g                    TTACCTATGTAGATGAGGTGTCTCATGGGGGCCTTATAAAACCCTCAGATCAATTTAAAAATAAGCTCAAAGAGCTGG
arent=6 muts= 121:t/a 34:a/g                   TTACTTATGTAGATGAAGTGTCTCACGGGGGACTCATGAAACCGTCTGCCCAGTTCGTTGAAAAAATAAAACAATTAG
ent=6 muts= 236:t/g 191:c/t                    TCACTTATGTCGATGAAGTGTCTCACGGAGGGCTCANGAAACCATCTATACAGTTCATAGAAGAAATTAAGAAGTTGG
rent=8 muts= 76:t/g 20:g/t                     TCATTTATGTCGATAAAGTGTCTTACGGAAGGCTTATCAAACCTTTTAAACAATTCATAG--------AATAAATT--
arent=5 muts= 59:t/c                           TCACTTACGTGGATGAAGTGTCGTACGGTGGGCTCCTAAAACCGTCTGTCCAGTTTATAGAAAAACTAAAAAAATTGG
parent=15 muts= 69:c/t                         TTACATACGTAGGCGAGGTGTCTCACGGCGGACTTATTAAGCCGTCCGAAGAACTTCAAGAGAAGCTGAAAGAGCTAG
parent=13 muts= 259:a/g                        TCAGTTATGTAGATGAGGTGTCTCACGGCGGCCTTATTAAACCGTCGGATCACTTCAAAAATAAGCTCAAAGAGCTAG
nt=2 muts= 253:c/a 140:t/g 19:a/t              ATACGTATGCC---------------------------------------CAATTTACTCTAAAACTAATAGAATTAG
```

Alignments tell us a story that can span eons!



```
RATE
  BH  GAGTTTTAAGTCATTAAA-AATGATATTTTATTTGCACTAAAAAGGAGA-CAGACTTA
 FHL  GAGTTTTAAGTCATTAATTAATGATATTTTATTTGCACTAAAAAGGAAA-AAAACTTA
  SC  GAGTTTTAAGTCATTAACTAATGATATTTTATTTGCACTAAAAAGGAAG-AAAACTTA
  BR  GAGTTTTAAGTCATTAACTAATGATTTTTTATTTGCACTAAAAAGGAAATAAAACTTG
  SD  GAGTTTTAAGTCATTAACTAATGATTTTTTATTTGCACTAAAAAGGAAATAAAACTTA
 CAT  GAGTTTTAAGTCATTAAGTAATGATATTTTATTTGCACTAAAAAGGAAA-AAAACTTG
  AB  GAGTTTTAAGTCATTAATTAATGATATTTTATTTGCACTAAAAAGGAAA-AAAACTTG
 SRQ  GAGTTTTAAGTCGTTAATTAATGATATTTTATTTGCACTAAAAAGGAAA-AAAACTTG
```



Real alignments can be messy

Our tree will only be as informative as our alignment and thus our alignments must be assessed/adjusted/QCd before we use them!

- Alignment tool: MAFFT, T-Coffee, Clustal, Muscle, PRANK,

- Assessment: Visually inspect alignments (MEGA, Geneious, Jalview, Clustal X)

- Algorithmically prune sites: Gblocks, TrimAI, T-Coffee + TCS

Models of sequence evolution describe how DNA changes?

	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

Kimura’s two-parameter model

α represents transitions A<->G or C<->T (purine to purine or pyrimidine to pyrimidine)

β represents transversions transitions between purines and pyrimidines

	A	G	C	T
A	-	r_1	r_2	r_3
G	r_1	-	r_4	r_5
C	r_2	r_4	-	r_6
T	r_3	r_5	r_6	-

General Time Reversible Model

Each type of transition has its own rate but reverse rates equal forward rates

How do we handle models?

- Just use GTR and forget about it.*
- Pick the best model (jModel Test, Model Test NG, IQ-Tree)
- Average across models based on probability (MrBayes, RevBayes)

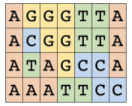
* Abadi, S., Azouri, D., Pupko, T. and Mayrose, I., 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature communications*, *10*(1), p.934.

How do we actually choose among trees and branch lengths

Parsimony	Maximum Likelihood	Bayesian

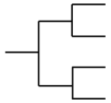
Maximum likelihood algorithm

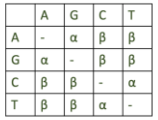
$\boldsymbol{P}($



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

$|$



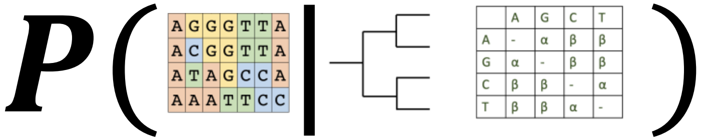


	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

$)$

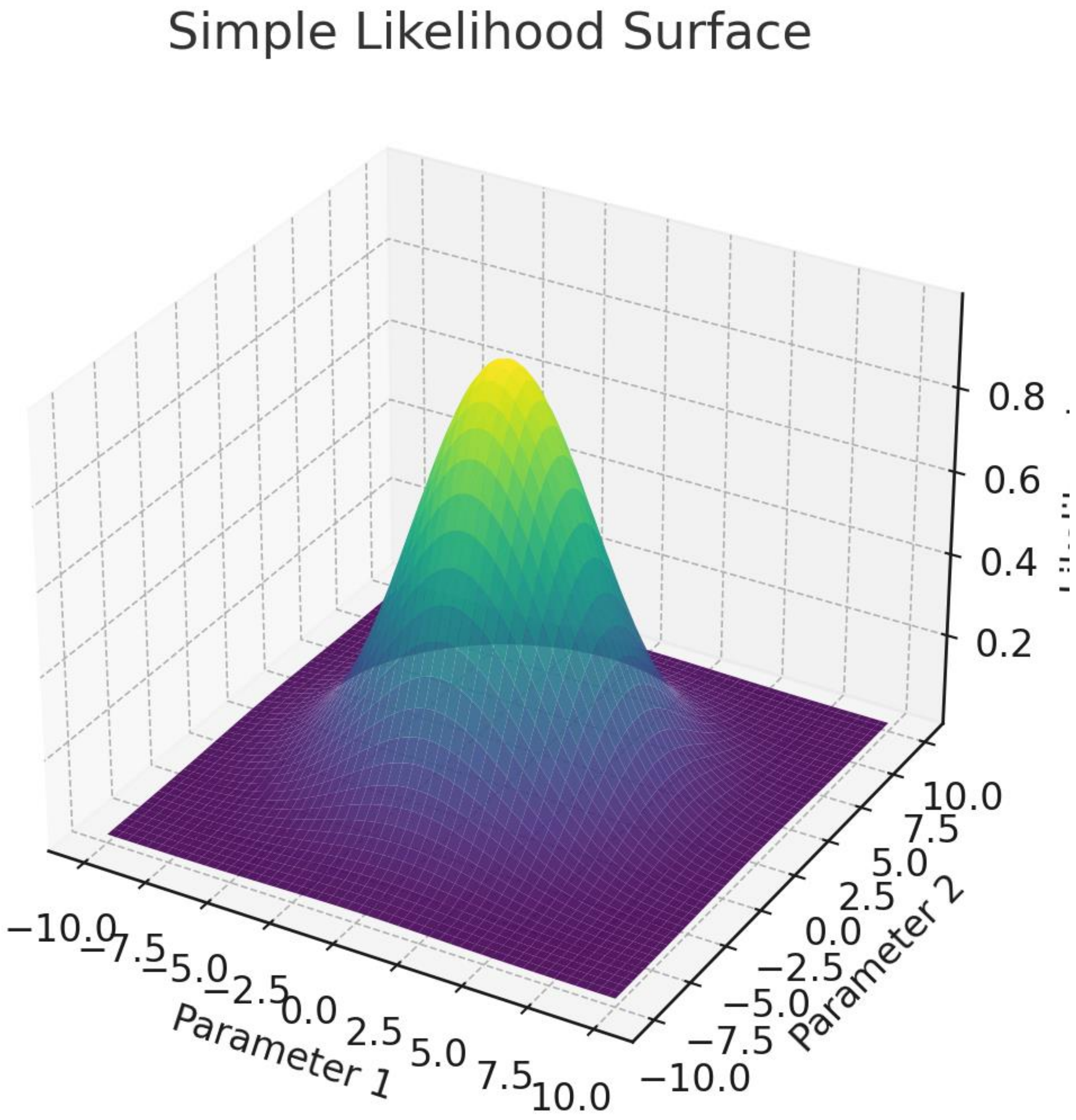
1) Pick a random starting value for all parameters
2) Calculate the likelihood above
3) Make a small change to one of the parameters
4) Keep that change if the likelihood of the alignment improved
5) Repeat until you get no further improvement in likelihood

Maximum likelihood



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

	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

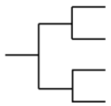


Maximum likelihood

P (

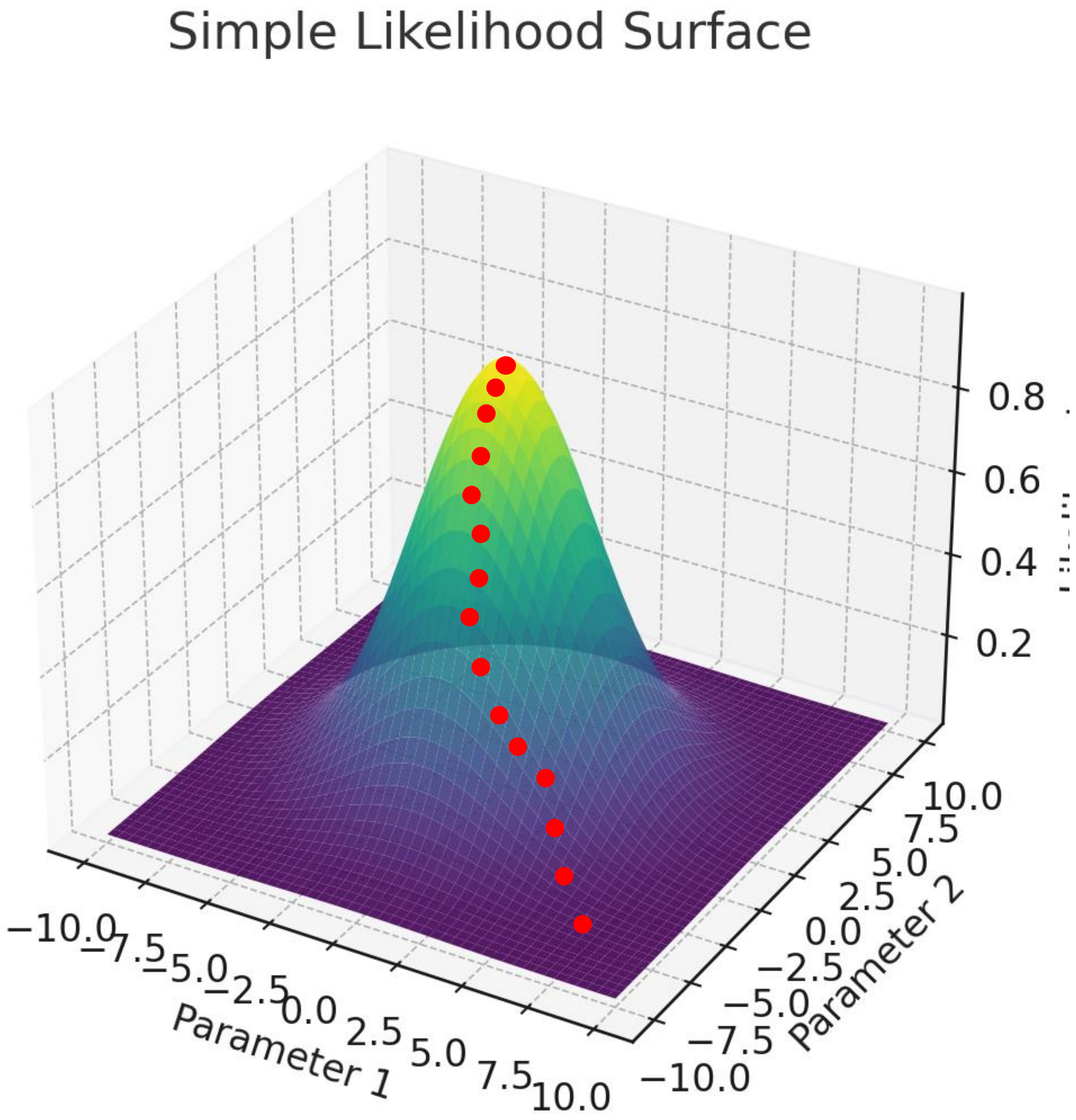
A	G	G	G	T	T	A
A	C	G	G	T	T	A
A	T	A	G	C	C	A
A	A	A	T	T	C	C

|

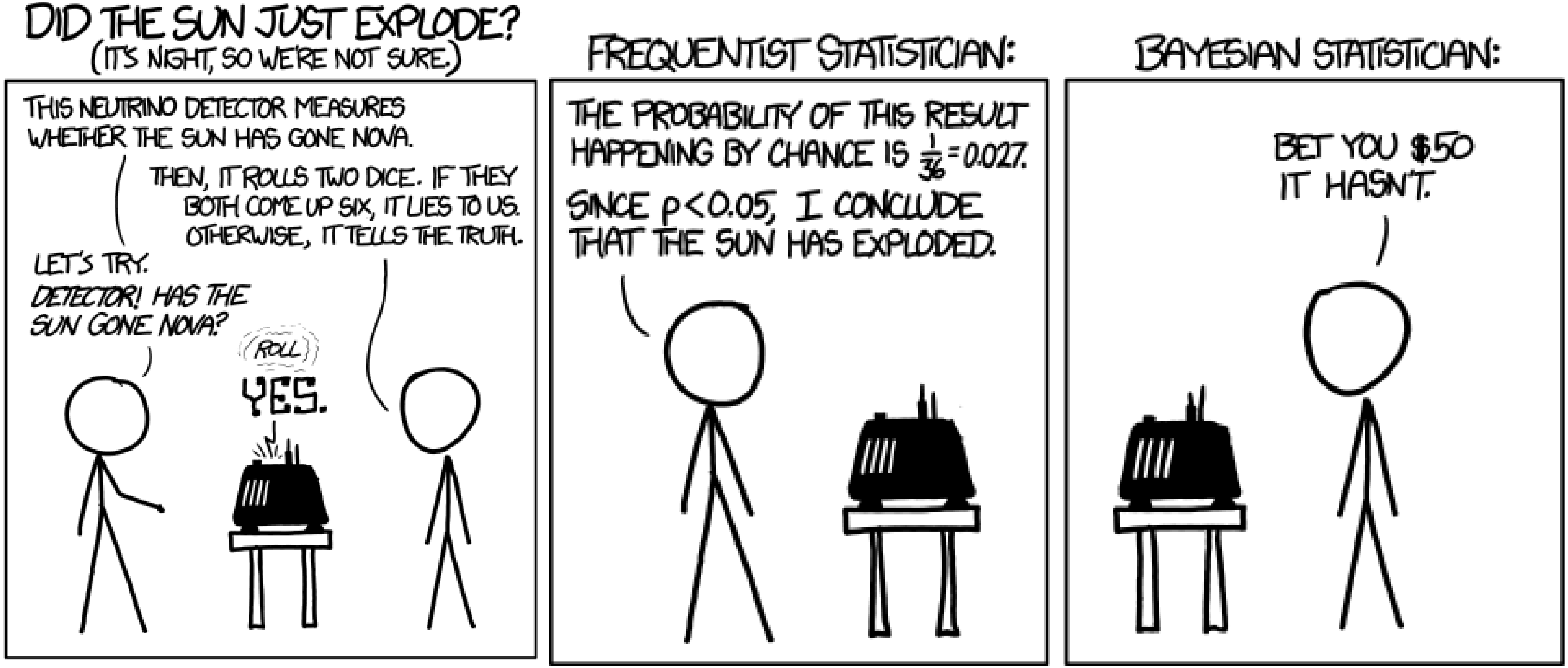


	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

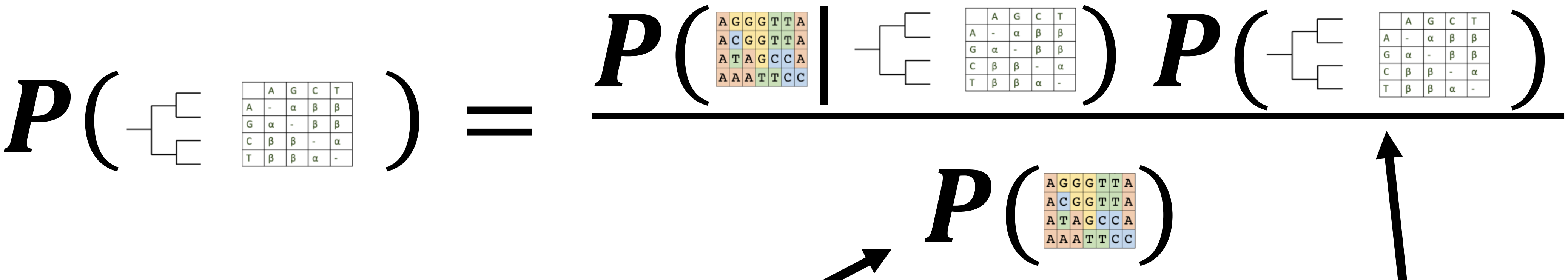
)



What is Bayesian Statistics?


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bayes’ Theorem



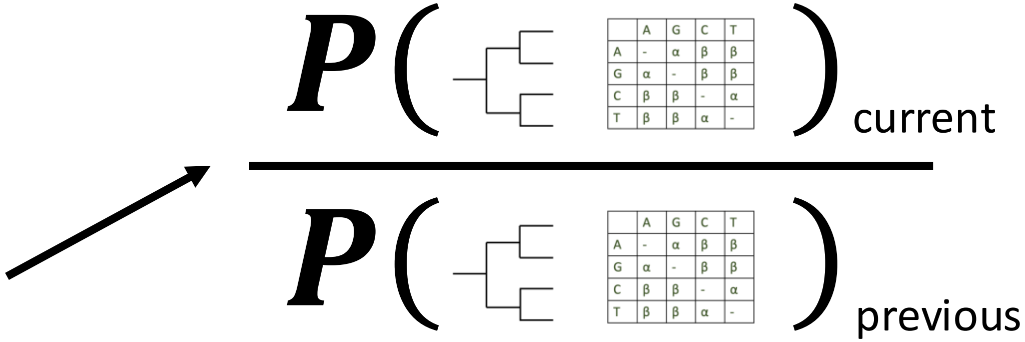
Impractical to calculate for non-trivial problems but can be avoided via MCMC

These are the priors

Reverend Thomas Bayes (1701–1761) was an English statistician, philosopher, and Presbyterian minister best known for Bayes' Theorem, which provides a mathematical framework for updating probabilities based on new evidence. His posthumously published work laid the foundation for Bayesian inference, a cornerstone of modern statistics, machine learning, and phylogenetics.

MCMC algorithm

1) Pick starting values
2) Calculate the probability
3) Make a small change to one parameter
4) Calculate the new probability
5) Accept the changed parameter with this probability
6) Return to step 3



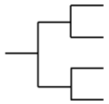
	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

Repeat steps 3-6 1,000,0000s of times

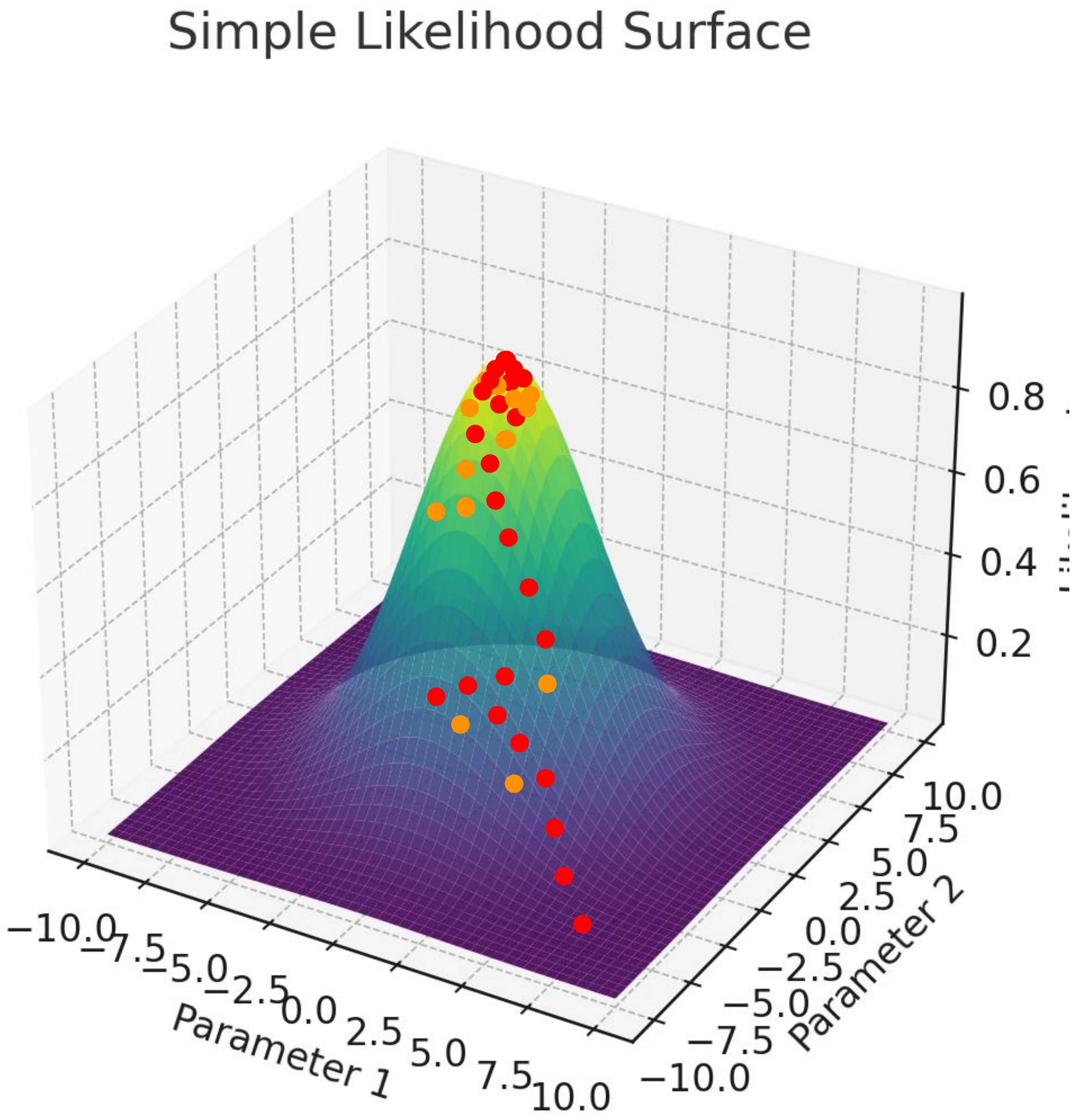
MCMC Run

P(

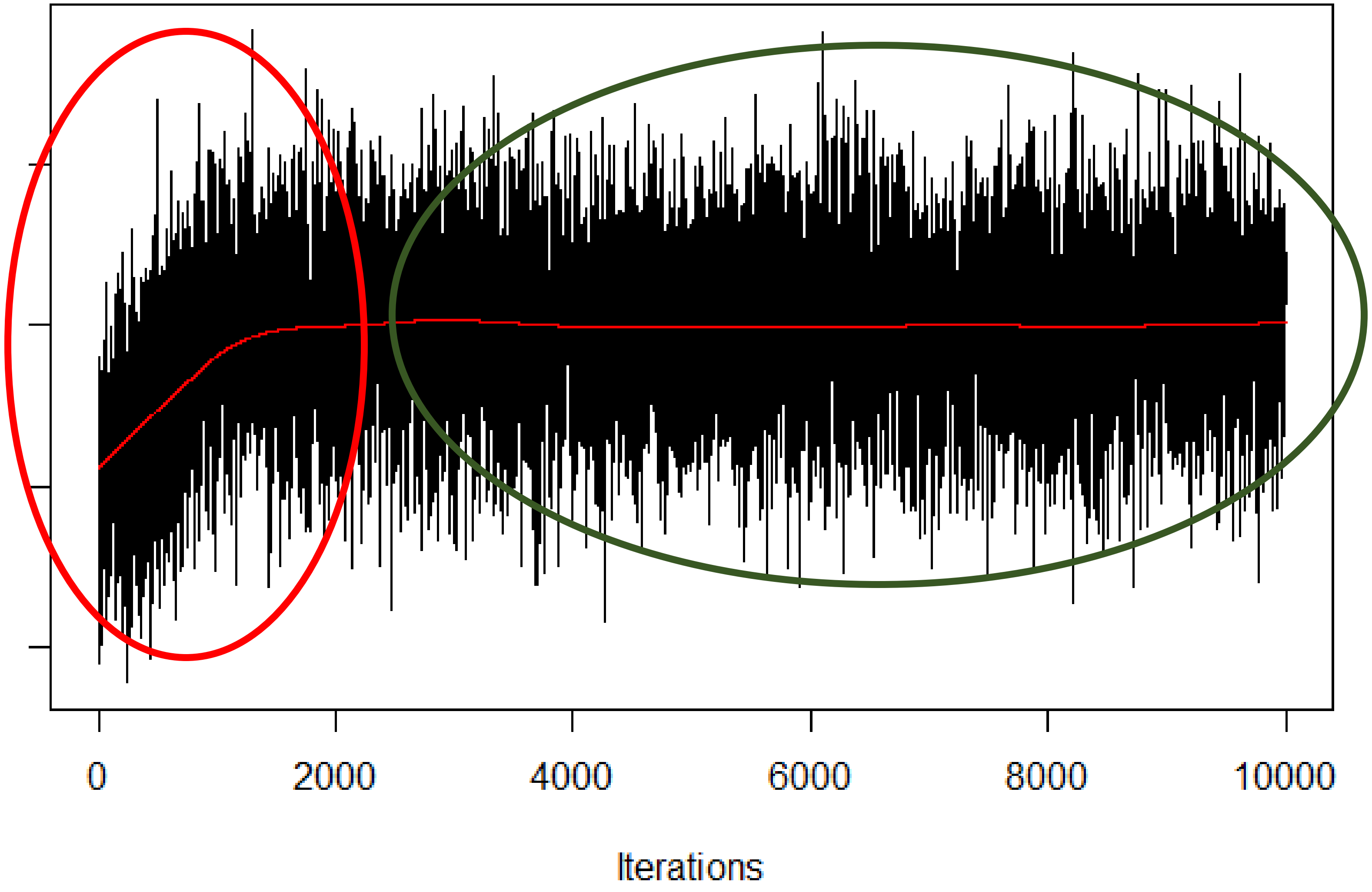


	A	G	C	T
A	-	α	β	β
G	α	-	β	β
C	β	β	-	α
T	β	β	α	-

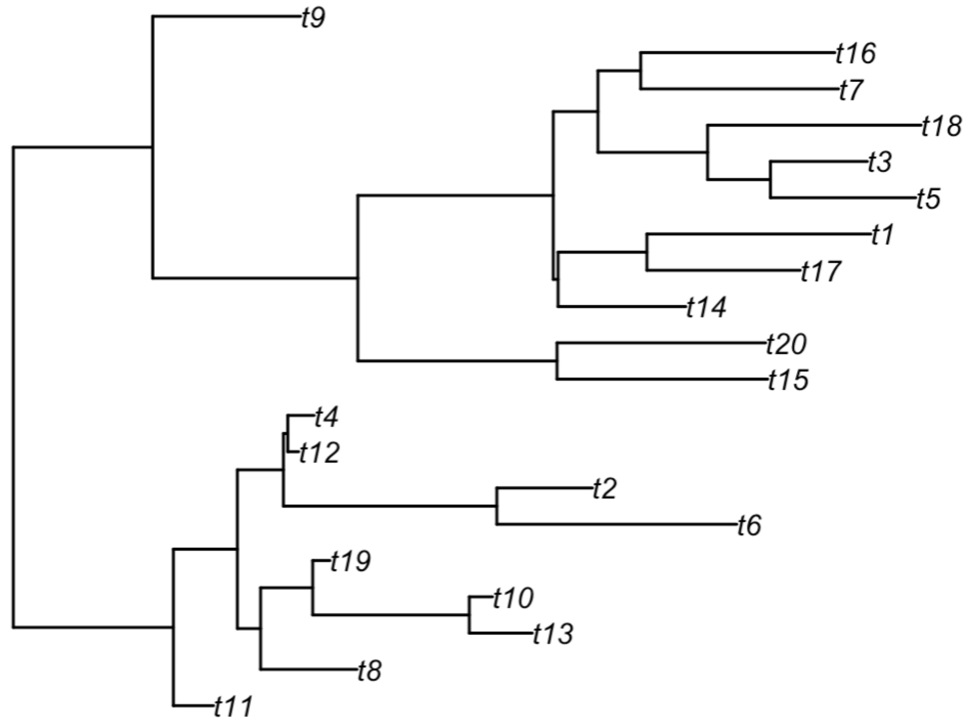
)

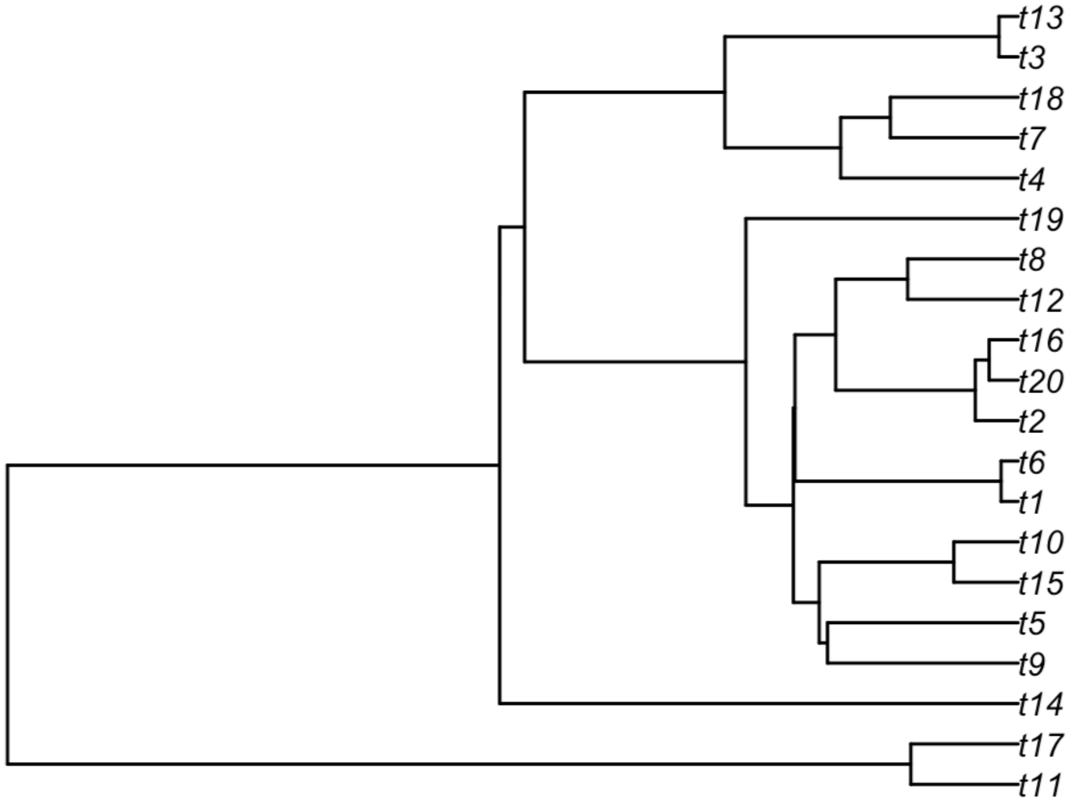


MCMC Run


This is the burnin
This is the sample of the posterior
0
2000
4000
6000
8000
10000
Iterations

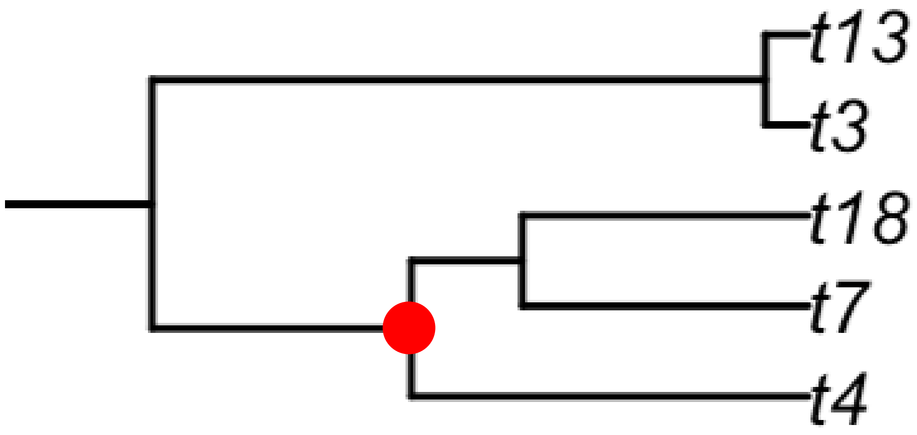
What do branch lengths mean?





Typically, branch lengths will either be expected substitutions per site or time (MY)

Questions



Name two species that are sister.

Which species is most closely related to t18?

What does the red dot indicate?

What are the tradeoffs between ML and Bayesian

Why do we need phylogenies?

If this is your future what next?

- Read inferring phylogenies!
- Talk to your mentors about good labs that might interest you.
- As you start grad school look for workshops (MBL Mol. Evol. Course)
- Amazing youtube vidoes

If this is your future what next?

- Break
- Guided walkthrough of running raxml
- Visualizing trees