A Primer into Phylogenetics

Julio Diaz Caballero
University of Oxford

wellcome connecting science

ACORN

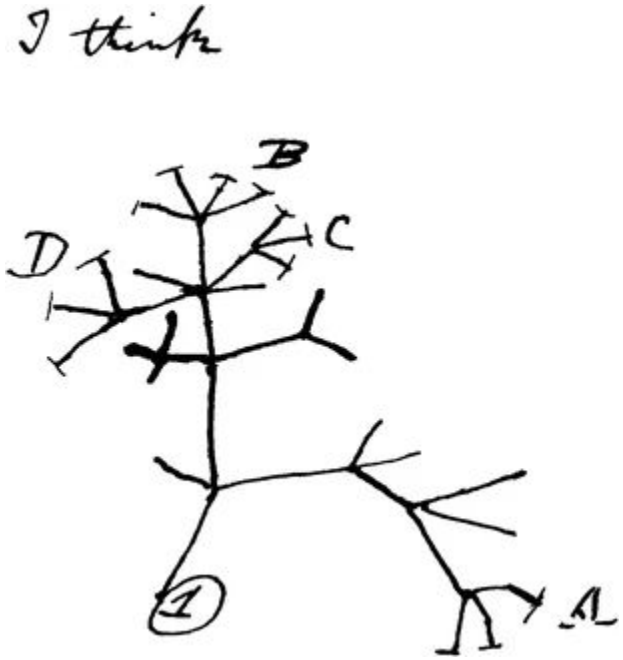
Centre for Genomic Pathogen Surveillance

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD

Phylogenetics is the scientific study of the **evolutionary relationships** among biological entities—often species, individuals, or genes.

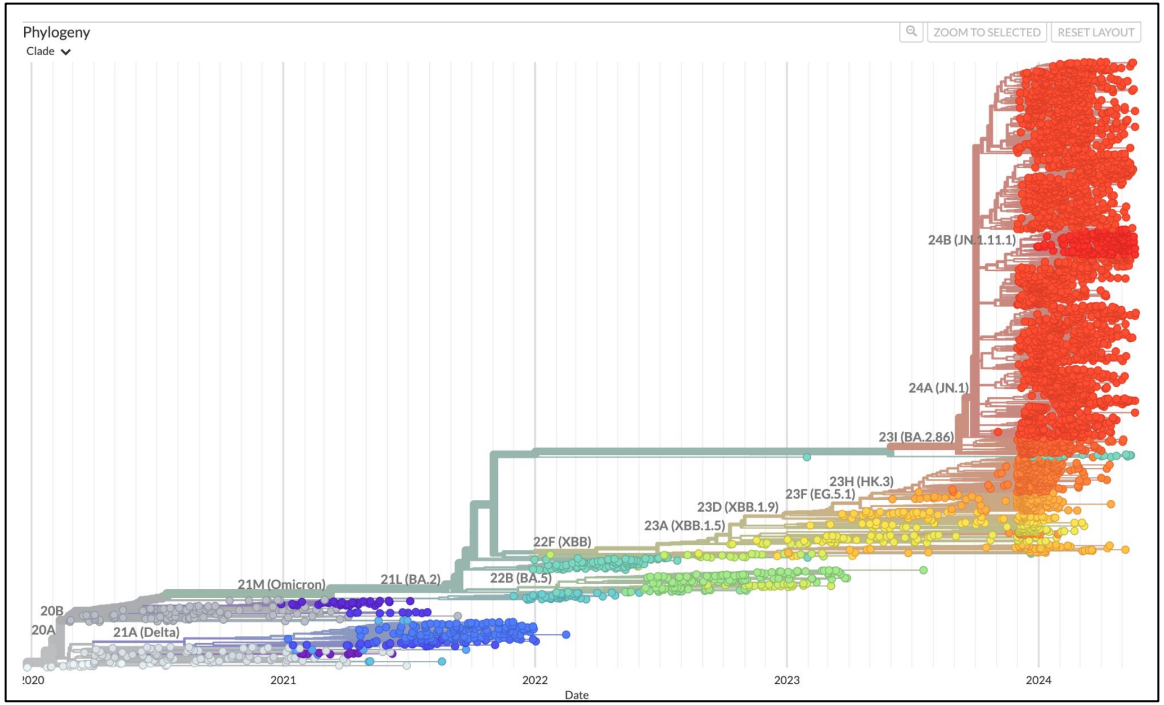
Evolutionary Tree sketched by Charles Darwin



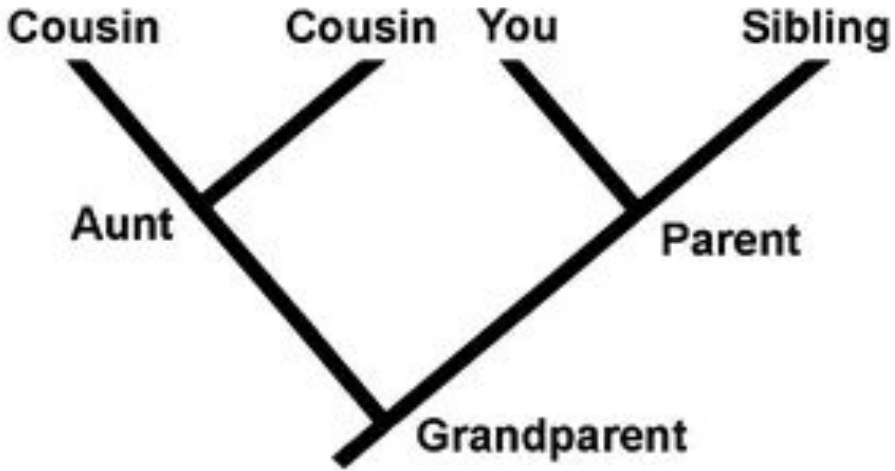
On The Origin of the Species (1859)

SARS-CoV-2 over the last 6 months

https://nextstrain.org/ncov/gisaid/global/6m

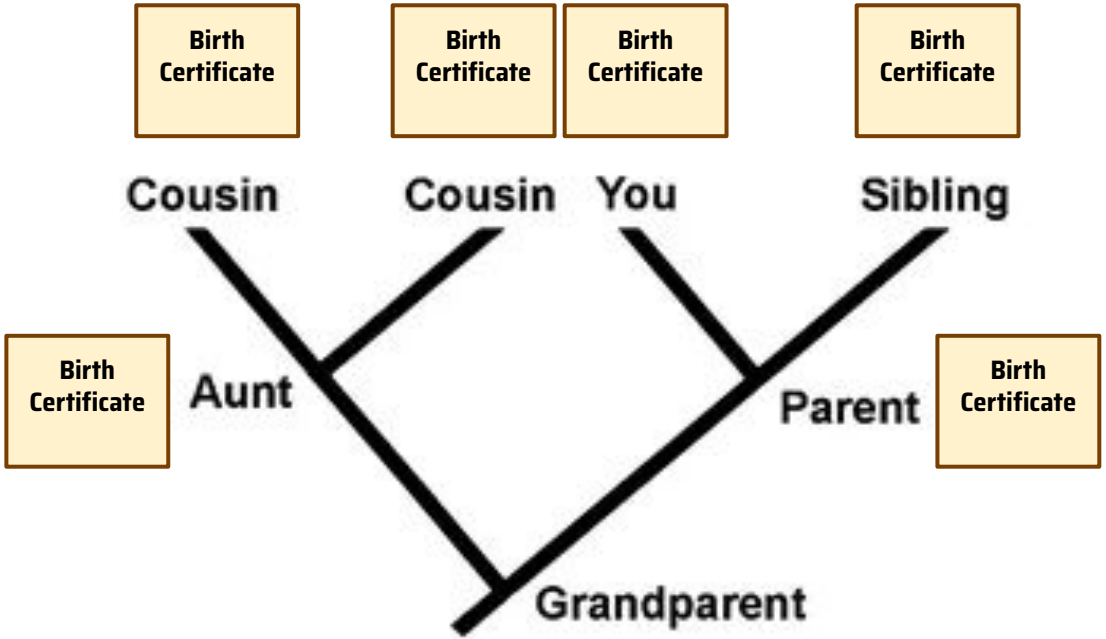


Genealogical Relationships



Ryan Gregory (2008)

Genealogical Relationships



Ryan Gregory (2008)

Bacteria do not have birth certificates

people

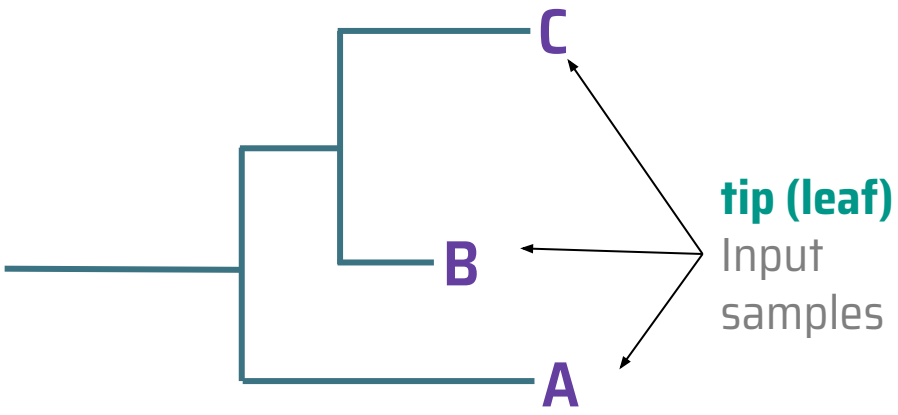
Birth Certificate

Name:
Date ot birth:
Parent 1 name:
Parent 2 name:
Address:

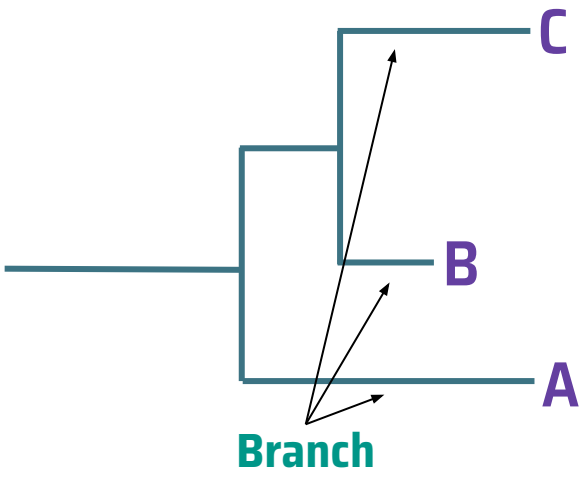
bacteria

> DNA
TAACCGGTAAAACACTTTGCTCGTGCACA
TAGGCGTGACACCTAAGACTGGAACAAG
CTCAGAAGAGTAGTAGGCGAGCATTTTTT
GACCGAGTCCGCTCCTTTCTAACTCACTG
ACTTCTCGCGGGCCGTATCCTGCACGCT
CAACAGCCAGCGGTGTCCCGTTACCCTT
CAAGCTCATCTTCCTCGAGGTCTGTTGTA
GTACCACACGCCTCTCCCGGCATTAGCC
CGCACTCCTCGACGGGACATTATGTGCC
TTCAGTTCCCGATCTCGGTCGCGCCCAG
CCGGAATCCCTTAGACACCAGGGCCGCG
TGAGCGAGAAGCGGGGGGAGAACTTTAT
AGGGCTGTGGCTCATACAATAGGGTAAG
GTT…

Parts of a Phylogenetic Tree



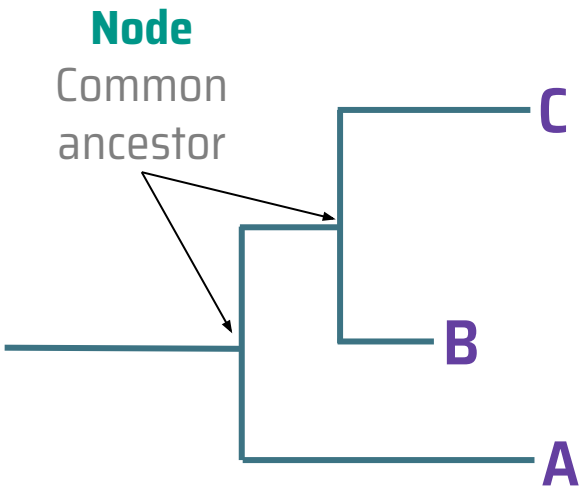
Parts of a Phylogenetic Tree



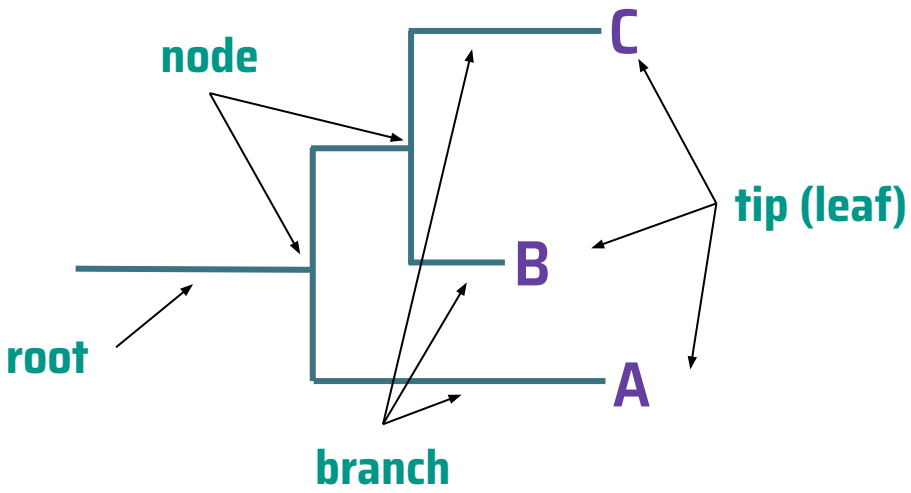
Branch

Evolutionary distance from ancestor to tip

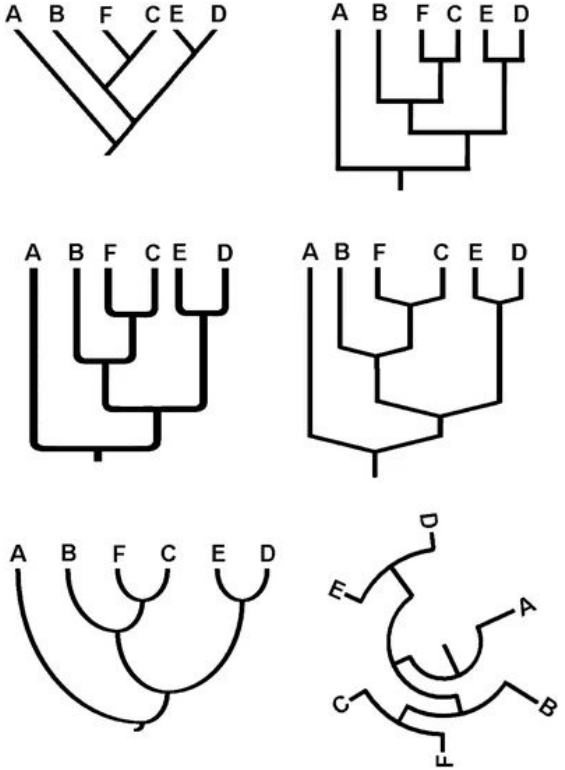
Parts of a Phylogenetic Tree



Parts of a Phylogenetic Tree

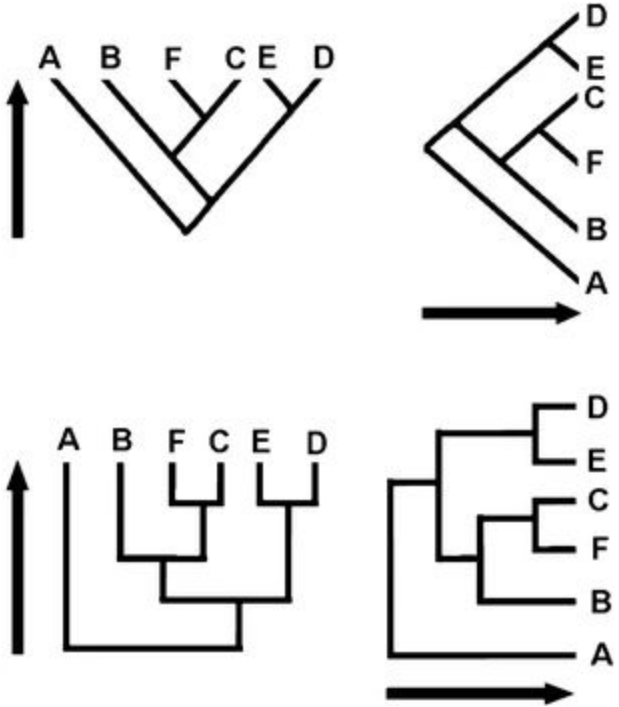


Many ways to the same tree



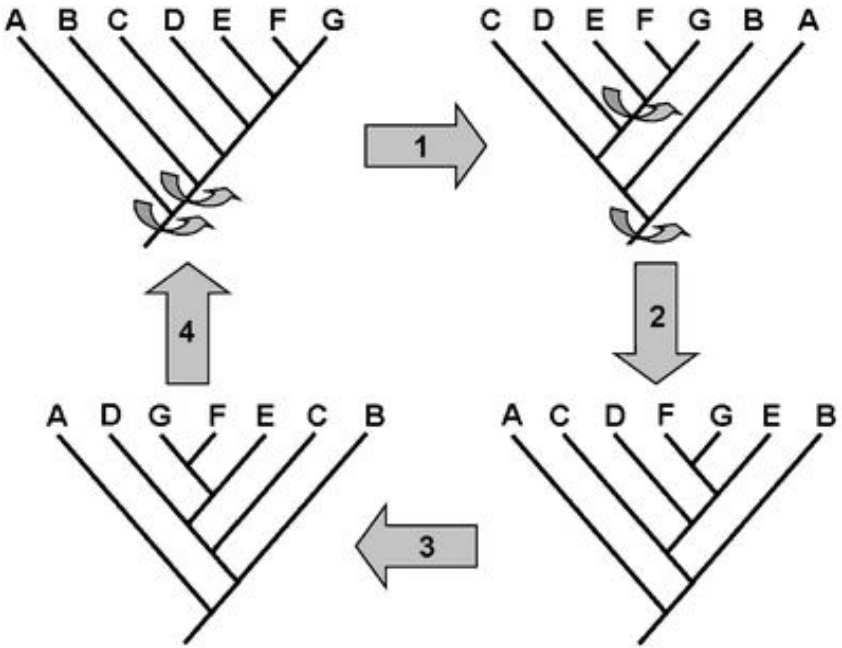
Ryan Gregory (2008)

Many ways to the same tree



Ryan Gregory (2008)

Many ways to the same tree



Ryan Gregory (2008)

How do we build a tree?

Distance-based methods

- Based on the dissimilarity (the distance) between sequences to construct trees
- UPGMA, Neighbour-Joining

Character-based methods

- Based on all the individual characters from each sequence
- Maximum-likelihood, Maximum-parsimony, Bayesian Analysis

Distance-based methods

Sequence Alignment

```
A CGATGCTAGA
B GGAAGCACCA
C GCAAGCACGT
```

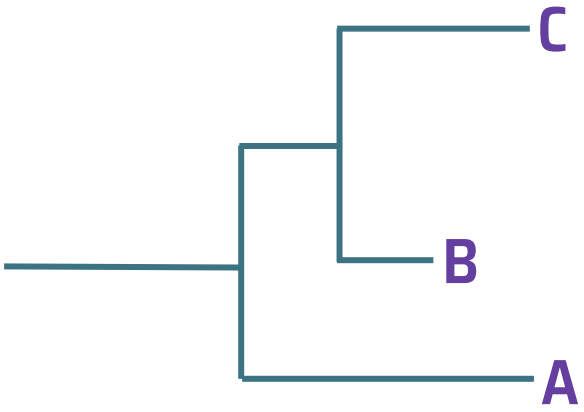
Number of nucleotide differences

	A	B	C
A		5	6
B			3
C			

Interpreting phylogenetic trees

A CGATGCTAGA

B GGAAGCACCA

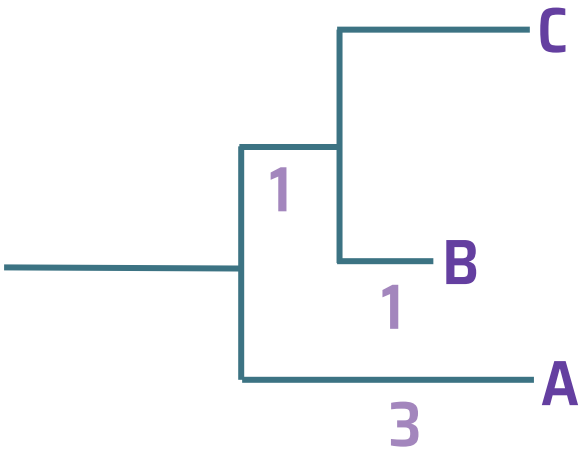


Here, sequences of A and B differ by 5 nucleotides

Interpreting phylogenetic trees

A CGATGCTAGA

B GGAAGCACCA

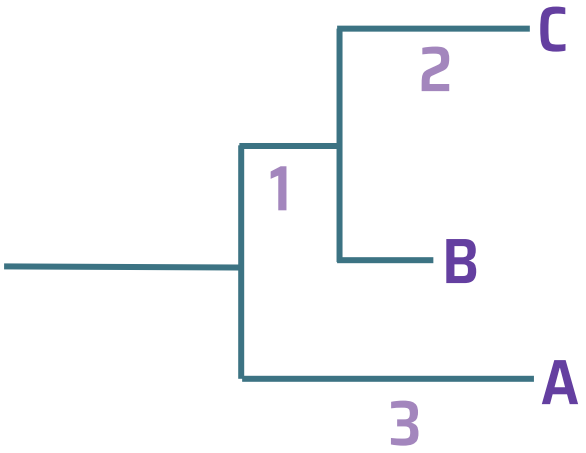


The length of the branches indicate this genetic distance

Interpreting phylogenetic trees

A CGATGCTAGA

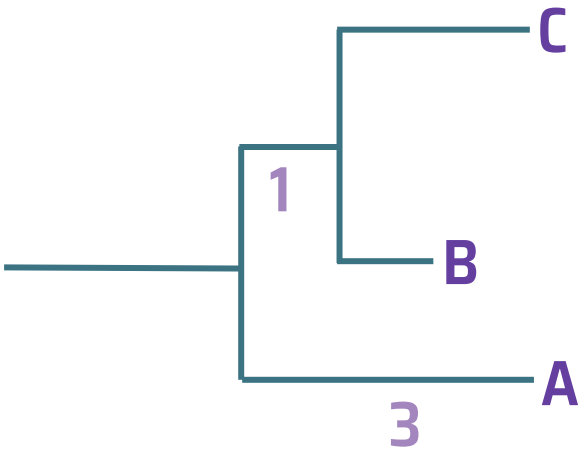
C GCAAGCACGT



A and C differ by 6 nucleotide differences

Interpreting phylogenetic trees

```
A CGATGCTAGA
B GGAAGCACCA
C GCAAGCACGT
```



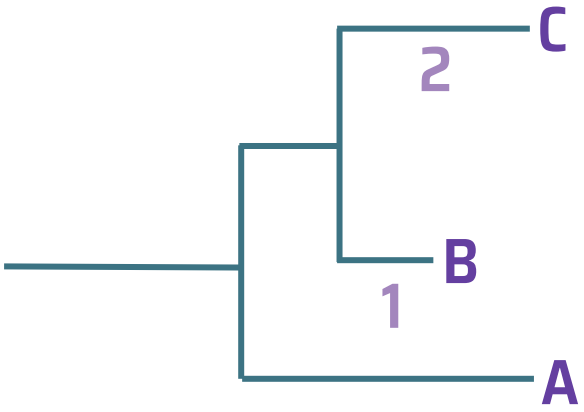
4 of these are shared with B

Interpreting phylogenetic trees

A CGATGCTAGA

B GGAAGCACCA

C GCAAGCACGT

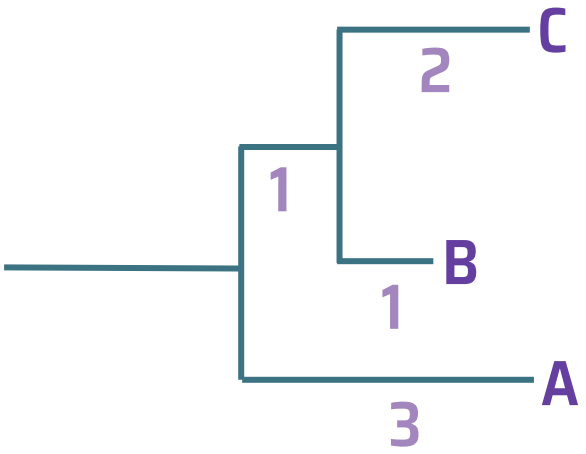


But B and C differ from each other by 3

Distance-based methods

Number of nucleotide differences

	A	B	C
A		5	6
B			3
C			



Distance Matrix

Character-based methods

Maximum-Parsimony

Identify the tree that can explain the data by identifying the minimum number of changes

Maximum-Likelihood

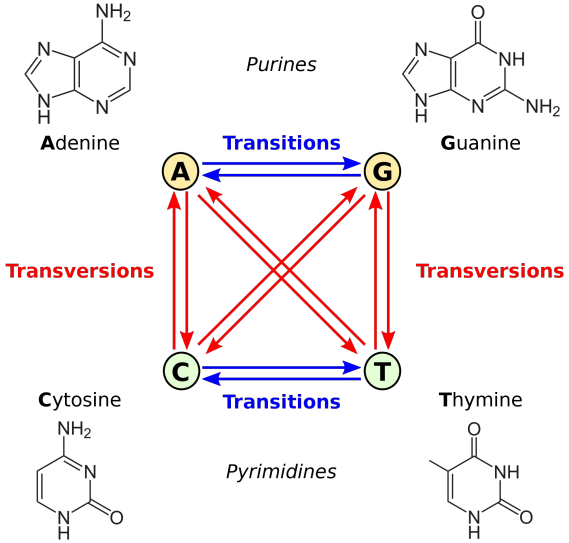
What is the likelihood of the data given a model with some parameters?

Bayesian

What is the likelihood of the model given the data?

Maximum-Likelihood

- Model:
  - Substitution model



https://www.wikipedia.org

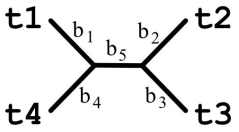
Maximum-Likelihood

- Model:
  - Substitution model
  - Rate Variation Among Sites

t1	G	C	T	T	C	T	G	A	T	T	A	A	C	C	T	G	C	T
t2	G	C	T	T	C	T	G	A	T	T	T	C	T	C	T	G	C	C
t3	G	C	T	T	C	T	G	A	T	T	A	C	T	C	T	G	C	C
t4	G	C	T	T	C	T	G	A	C	T	A	G	T	C	T	G	C	T

Site pattern	Count
GGGG	3
CCCC	4
TTTT	5
AAAA	1
TTTC	1
ATAA	1*
ACCG	1*
CTTT	1*
TCCT	1*

$$L(H|R) = \gamma P(R|H)$$



The tree (with branch lengths b_i) is the hypothesis (H) and the site patterns are the data (R)

$$d_{12} = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{12})$$

https://www.wikipedia.org

Maximum-Likelihood

- Model
  - Substitution model
  - Rate Variation Among Sites
  - Tree topology

Maximum-Likelihood

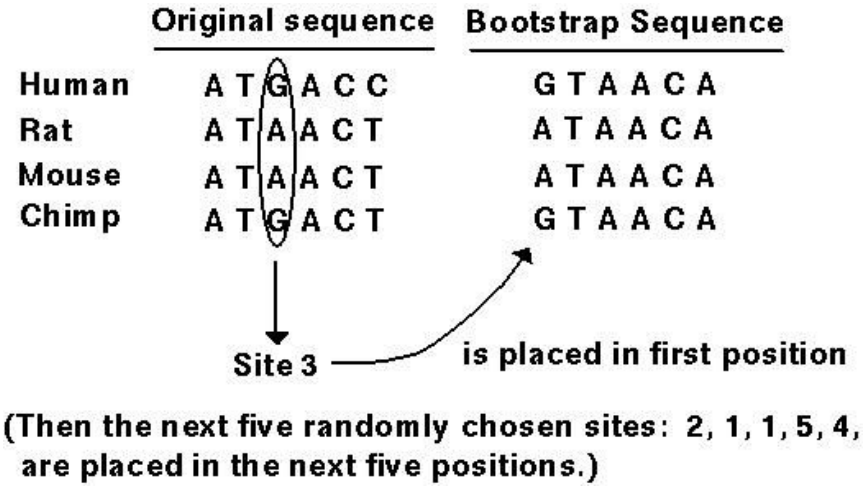
- Model
  - Substitution model
  - Rate Variation Among Sites
  - Tree topology

- Resampling
  - Bootstrap or Jacknife
  - Bootstrap: every character is replaced at random with another character a.k.a resample with replacement
  - Increases confidence

Maximum-Likelihood

- Model
  - Substitution model
  - Rate Variation Among Sites
  - Tree topology

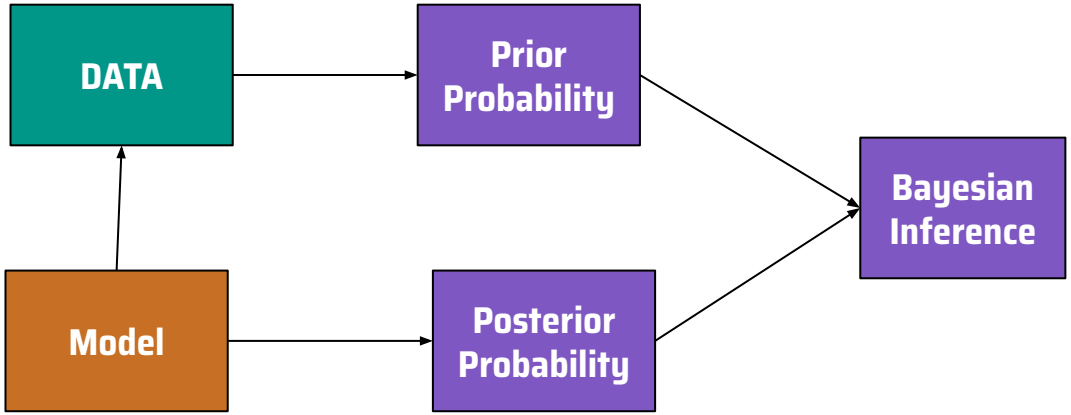
- Resampling (bootstrap)



https://www.zoology.ubc.ca/~bio336/Bio336/Lectures/Lecture14/Overheads.html

Bayesian Approach

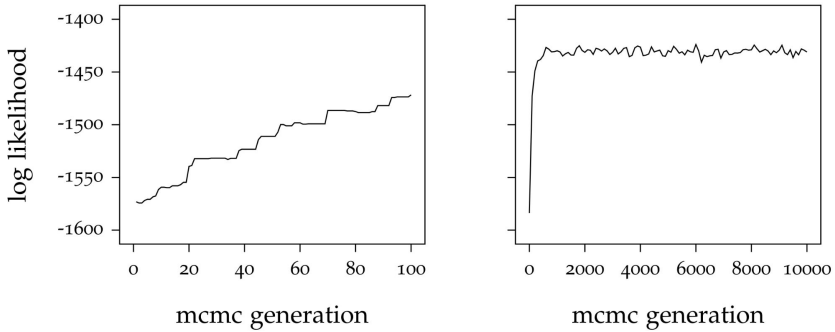
- Bayesian Inference
  - Combine **prior probabilities** with the likelihood of the model given the data (**posterior probabilities**)



Bayesian Approach

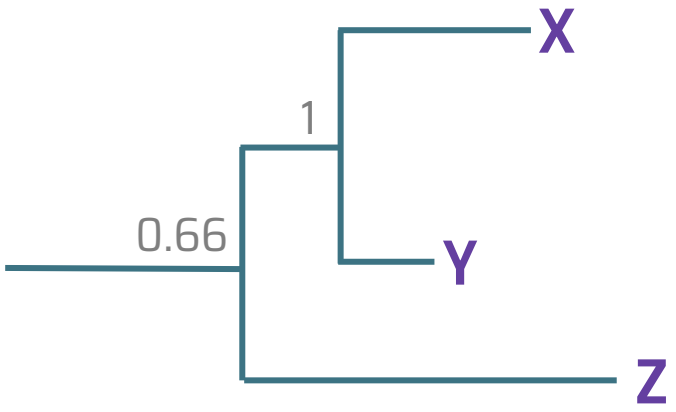
- Bayesian Inference
  - Combine **prior probabilities** with the likelihood of the model given the data (**posterior probabilities**)

- Monte Carlo Markov Chain (MCMC)
  - Simulates trees sampled from the posterior



Confidence in Phylogenies

- Proportion of trees supporting node
- Likelihood of Node



Tools

- General
  - MEGA
- Neighbour-Joining
  - RapidNJ, NINJA
- Maximum-Likelihood
  - RAxML, IQ-TREE
- Bayesian
  - Mr Bayes, RevBayes
- Visualisation
  - iTOL, Microreact, ape (R package)

Challenges

- **Computationally expensive**
  - Specially character-based methods
  - Not easily scalable
- **Requires recombination disentanglement**
  - Identification of clonal frame
  - Tools that take recombination (e.g. Gubbins) into account
- **Does not include gain/loss evolutionary events**
  -

Thank you!