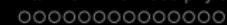


PHYLOGENETICS

An introductory lecture

Alejandra Vergara-Lope

December 8th, 2023



Outline

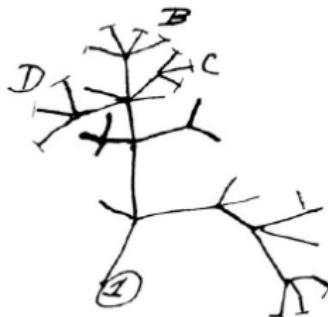
- 1** Introduction to phylogenetics
- 2** Sequence alignment
- 3** Phylogenetic inference
- 4** Maximum likelihood phylogenetics
- 5** Bootstrap
- 6** Bayesian phylogenetics inference
- 7** Some phylogenetic applications



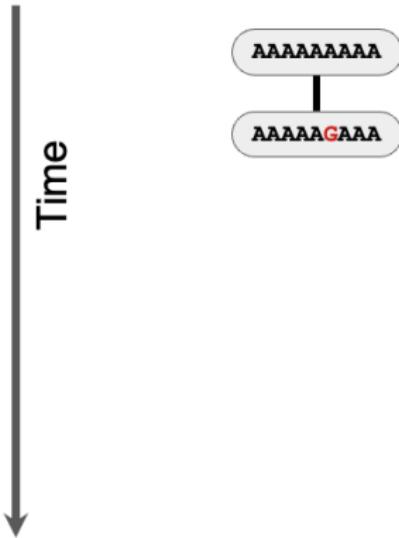
Introduction to phylogenetics

Origins of phylogenetics

- Since ancient times : Living things are similar and can be grouped
 - Several suggestions that organisms could be descended from others
 - Darwin proposes natural selection and shows how this could work
 - Naturalists, zoologists, ecologists, embryologists initially use morphological and anatomical traits to group organisms
 - Mathematical and statistical studies provide sophisticated methods
 - Molecular data adds a powerful tool for evolutionary comparisons
 - Genomic technologies vastly increase the amount of data available



What is a phylogeny?

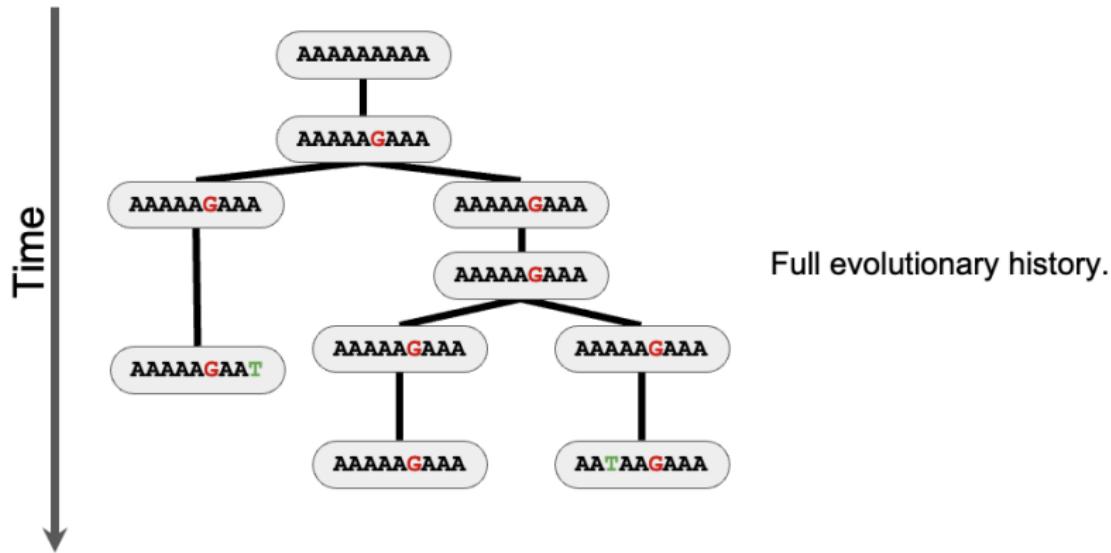


Sequence evolution: mutations alter the genomes being passed on from generation to generation.

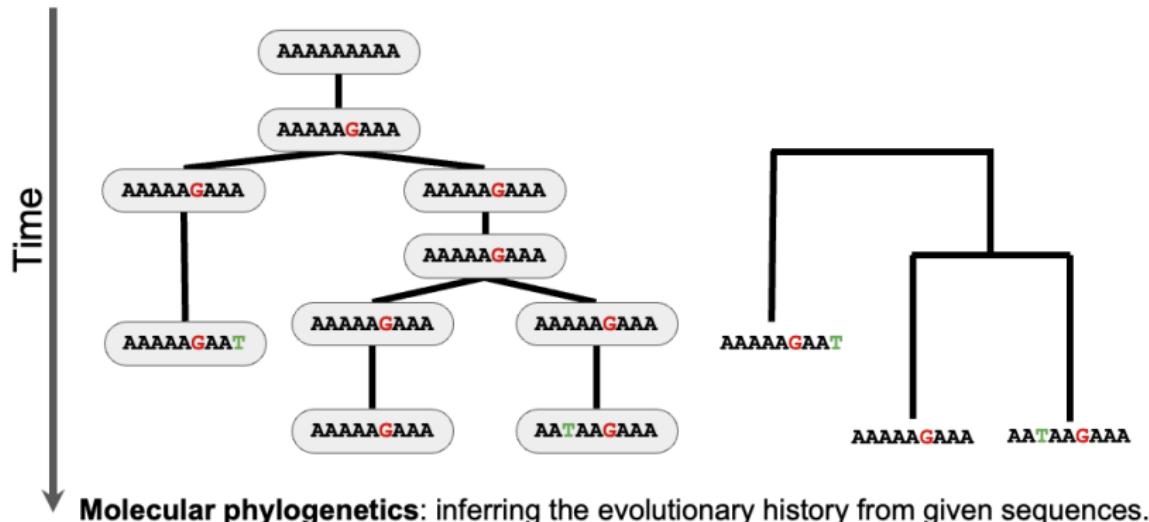
What is a phylogeny?



What is a phylogeny?



What is a phylogeny?



Basic Concepts and Terms

Phylogeny and phylogenetic tree

- A phylogenetic tree is a diagrammatic representation of the evolutionary relationships or phylogeny of a set of entities (organisms, molecules etc)
- Phylogeny reconstruction or inference is deducted from similarities and differences based on evolutionary traits or characters

Traits and characters

- Phenotypic : Morphology (e.g. Size and shape of body parts) Anatomy (Presence/Absence of anatomical features)
- Molecular : Nucleotide (Genes and other conserved elements) and Proteins

Examples of phylogenies

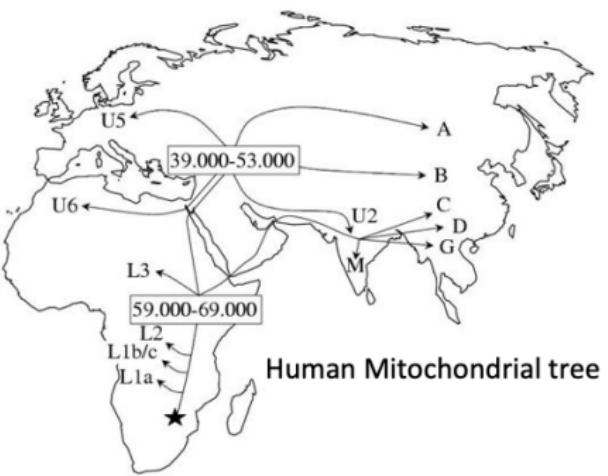
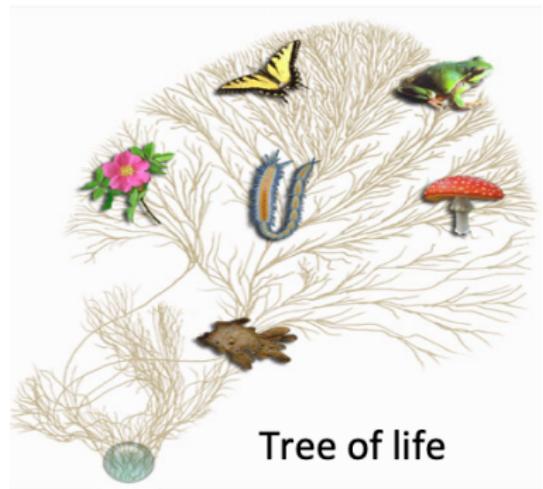


Figure – Taken from <http://tolweb.org> and Maca-Meyer et al 2001

Examples of phylogenies

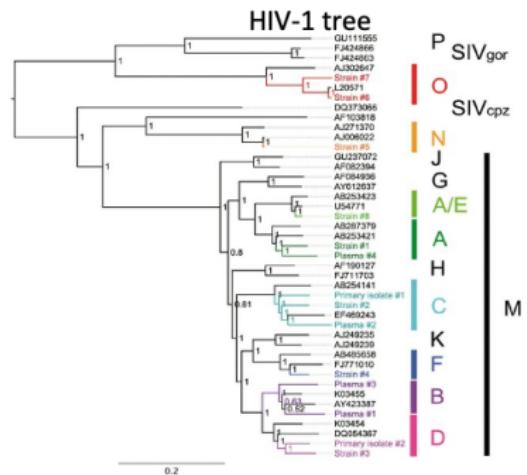
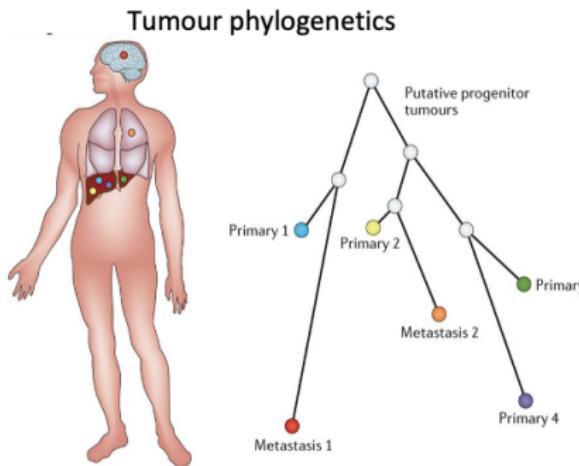


Figure – Taken from Schwartz, Schaffer 2017 and Gall et al 2012

Examples of phylogenies

Indo-European language family tree

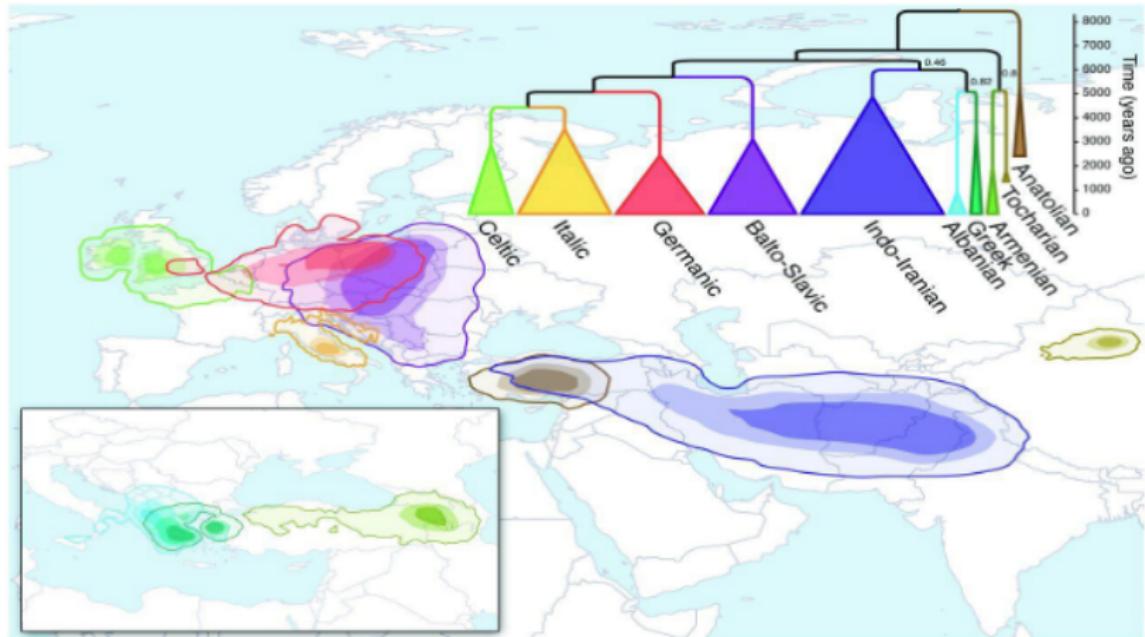
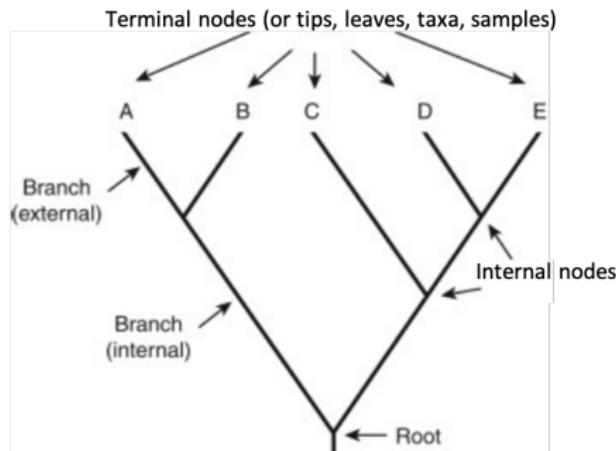


Figure – Taken from Bouckaert et al 2012

Phylogenetic Tree Characteristics, Topology and Types

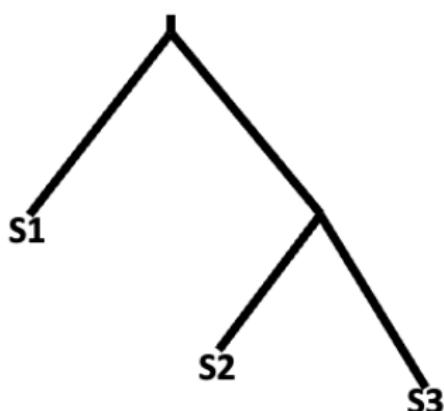
A tree graph - no cycles

- Parts of a tree (nodes and branches)
- Terminal nodes represent the species or entities studied
- Internal nodes represent the last shared ancestor of any two or more species
- A tree may or may not have a root, representing the last common ancestor of everything included in the phylogeny
- Branches connect all nodes showing their relationships and their length may represent change or time
- Basic Types of trees : Dendrogram, Cladogram, Phylogram, Chronogram



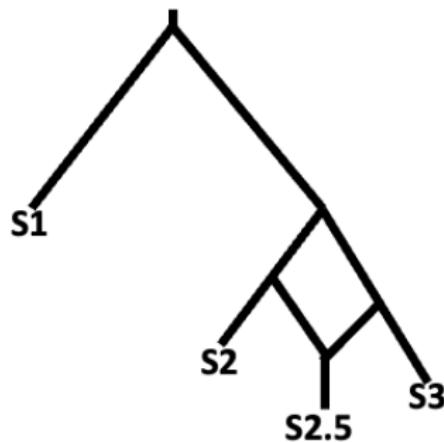
Phylogeny

A tree graph - no cycles



No cycles

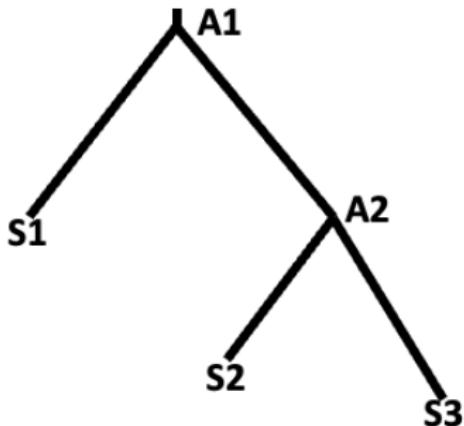
Network



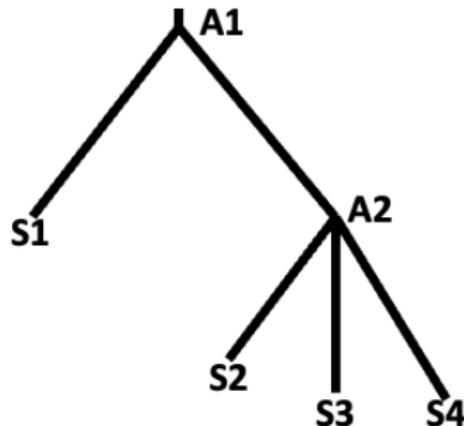
Cycle (e.g. hybridization): not a tree!

Phylogeny

A tree graph - no cycles



Binary tree

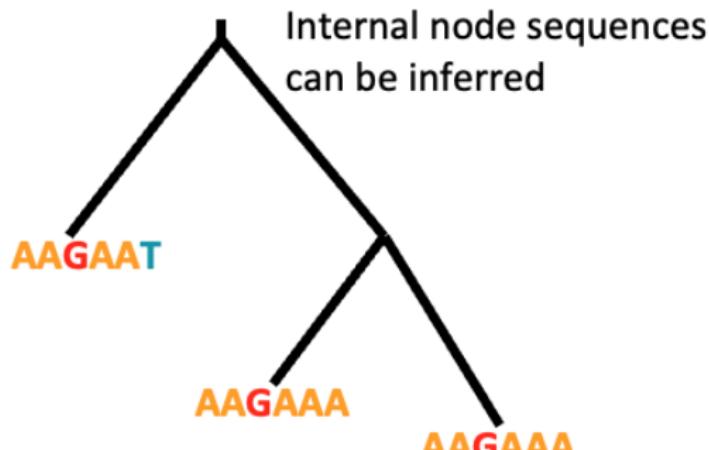


Tree with multifurcation

Phylogeny

A tree graph - no cycles.

Terminal nodes associated with data



Terminal node sequences are observed

Phylogeny

A tree graph - no cycles.

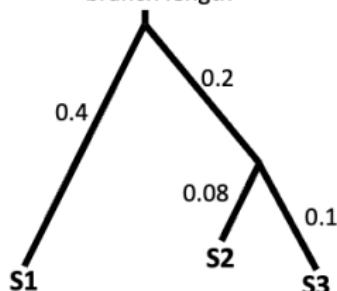
Terminal nodes associated with data.

Branch lengths represent time or divergence.

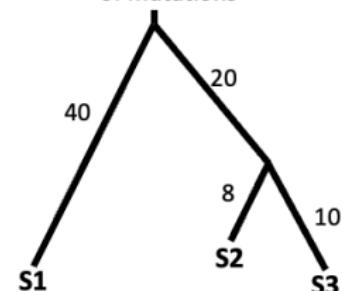
Time-unit branch lengths



Substitutions per site
branch length



Branch lengths in number
of mutations

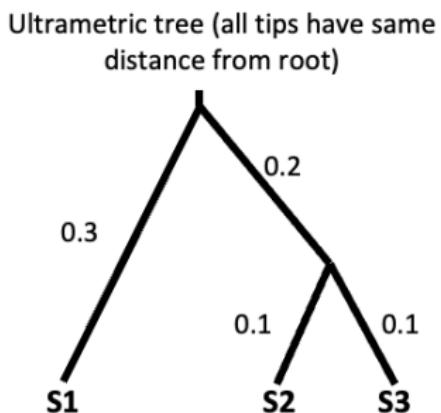
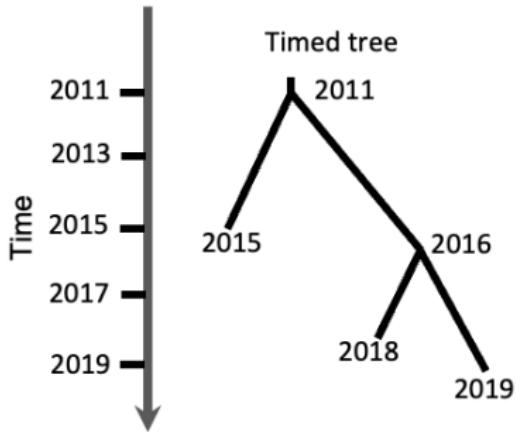


Phylogeny

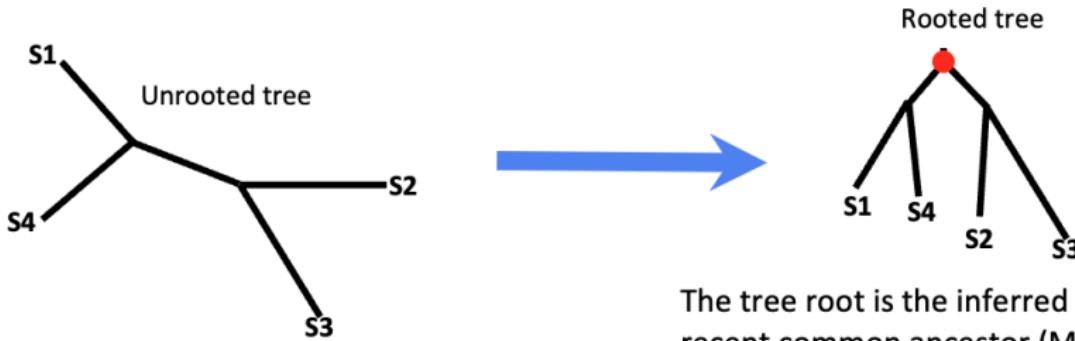
A tree graph - no cycles.

Terminal nodes associated with data.

Branch lengths represent time or divergence.



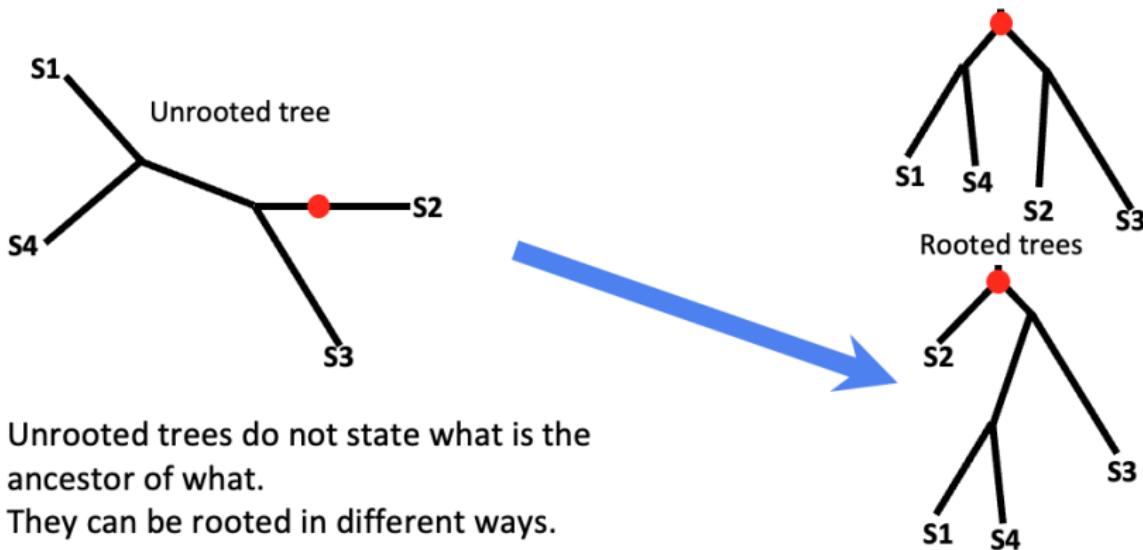
Tree rooting



Unrooted trees do not state what is the ancestor of what.
They can be rooted in different ways.

The tree root is the inferred most recent common ancestor (MRCA) of the considered sequences.

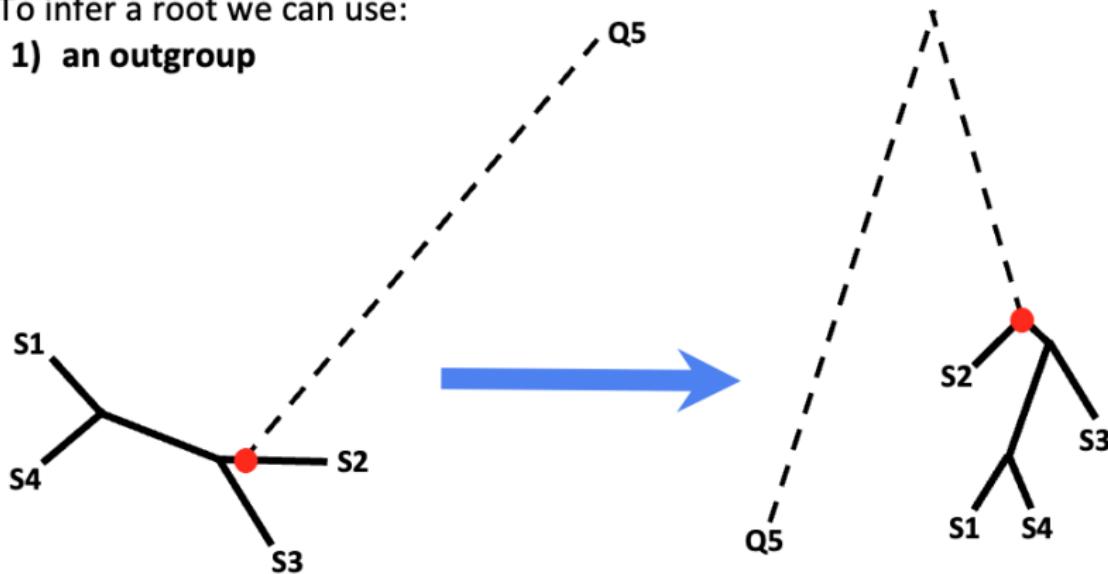
Tree rooting in practice



Tree rooting in practice

To infer a root we can use:

- 1) an outgroup



Tree rooting in practice

To infer a root we can use:

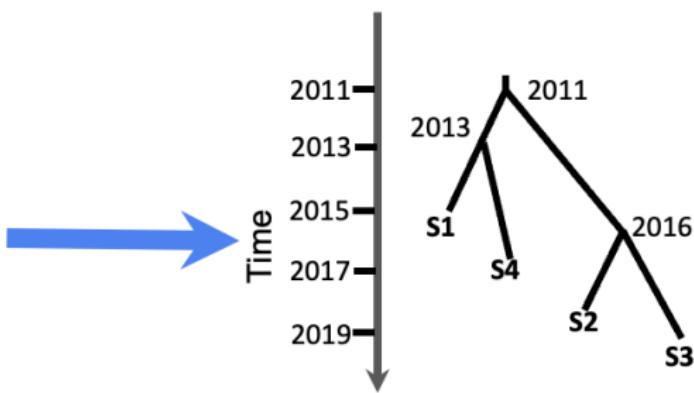
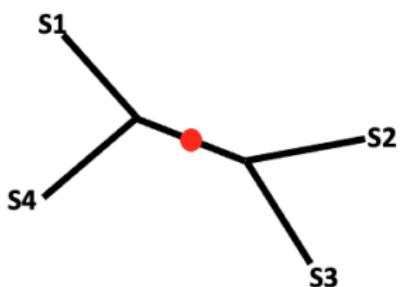
- 1) an outgroup
- 2) **constant evolution rate (midpoint rooting, ultrametric tree)**



Tree rooting in practice

To infer a root we can use:

- 1) an outgroup
- 2) constant evolution rate (midpoint rooting, ultrametric tree)
- 3) Time information



Tree rooting in practice

To infer a root we can use:

- 1) an outgroup
- 2) constant evolution rate (midpoint rooting, ultrametric tree)
- 3) Time information
- 4) A non-reversible model

Often genome composition (e.g. GC content) changes with time.
We can use this to root a tree.

Software | [Open Access](#) | [Published: 01 May 2021](#)

Root Digger: a root placement program for phylogenetic trees

[Ben Bettsworth](#) & [Alexandros Stamatakis](#)

[BMC Bioinformatics](#) 22, Article number: 225 (2021) | [Cite this article](#)

Tree rooting in practice

To root a tree we can use:

- 1) an outgroup
- 2) constant evolution rate (midpoint rooting, ultrametric tree)
- 3) Time information
- 4) A non-reversible model
- 5) All of the above!

Science

The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic

MICHAEL WORBEY, JOSHUA I. LEVY, LORENA MALPICA SERRANO, ALEXANDER CRITS-CHRISTOPH, JONATHAN E. PEKAR, STEPHEN A. GOLDSTEIN, ANGELA L. RASMUSSEN, MORITZ U. G. KRAEMER, CHRIS NEWMAN, I.-I. KRISTIAN G. ANDERSEN +9 authors

[Authors Info & Affiliations](#)

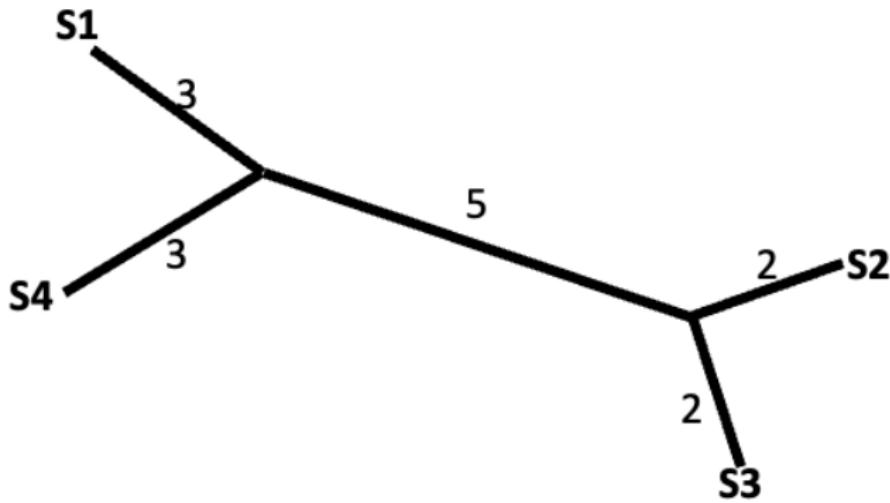
The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2

JONATHAN E. PEKAR, ANDREW MAGEE, EDYTH PARKER, NEEMA MOGHRI, KATHERINE IZHKEVICH, JENNIFER L. HAVENS, KARTHIK GANGAVARAPU, LORENA MARINA MALPICA SERRANO, ALEXANDER CRITS-CHRISTOPH, JOEL O. WERTHEIM +20 authors

[Authors Info & Affiliations](#)

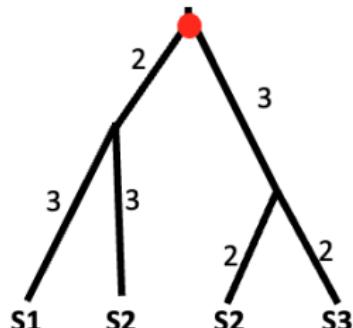
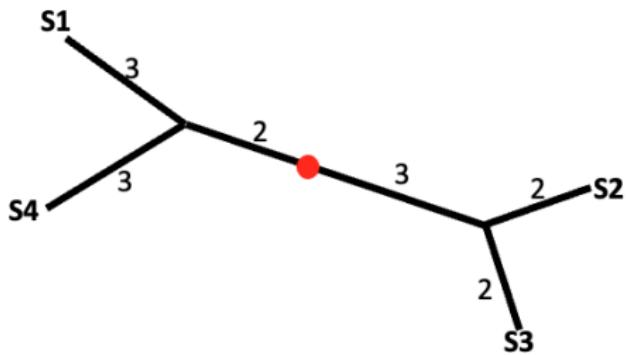
Practical

Root this tree by making it ultrametric (all terminal nodes with same distance from the root)



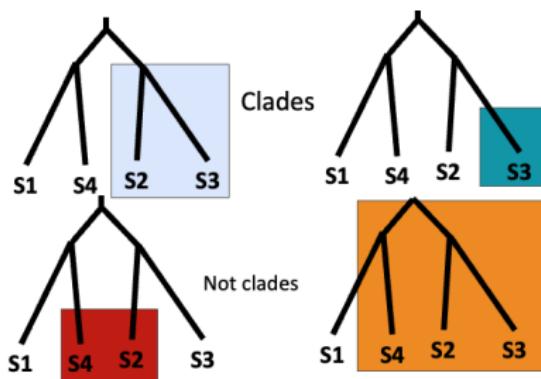
Answer

Assuming : Root this tree by making it ultrametric (all terminal nodes with same distance from the root)

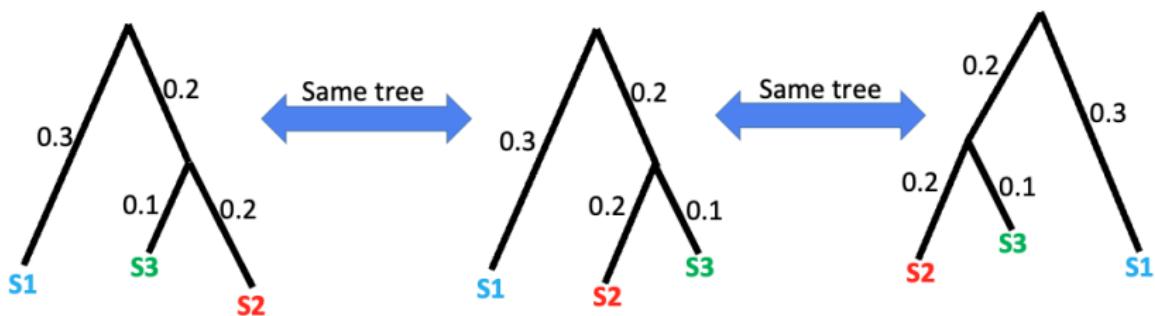


Tree topology

- A clade is a set of all sequences descending from node/ancestor
- Each node and each branch in a rooted tree correspond to one clade
- If two trees have the same clades, we say they have the same topology
- If two trees have the same topology and branch lengths, they are the same

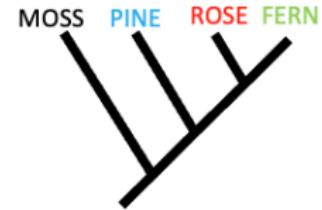
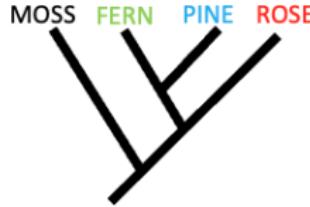
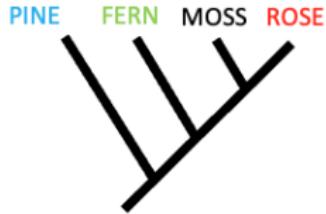
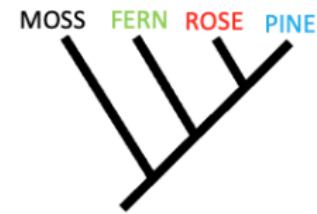
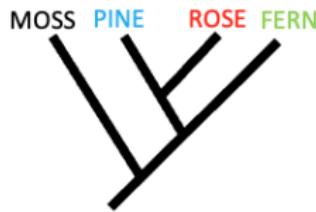
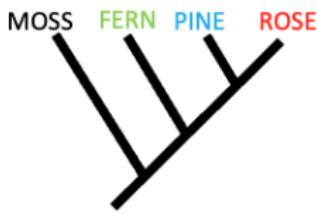


Tree topology

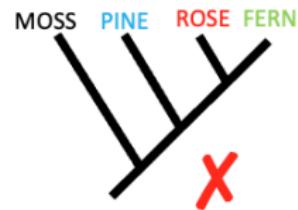
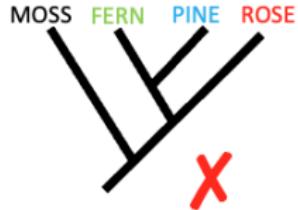
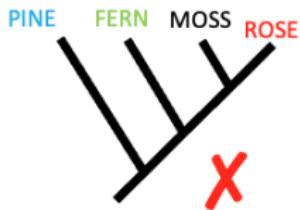
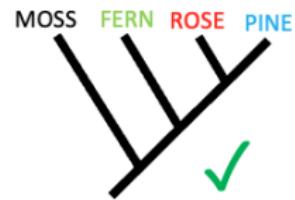
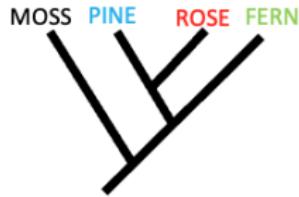
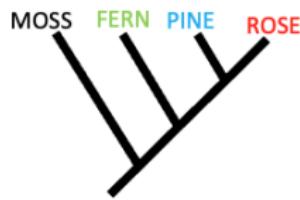


Practical

Which trees have the same topology ?



Answer

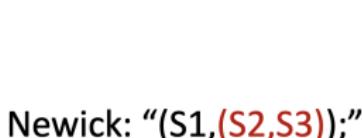


Newick format

Newick format The Newick format represents phylogenetic trees as text Usual for input/output in phylogenetic software



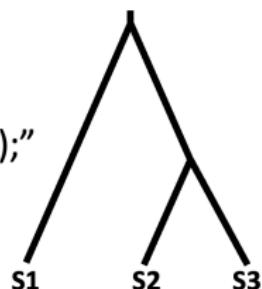
Newick: “**(S2,S3);**”



Newick: “**(S1,(S2,S3));**”



Newick with branch lengths:
“**(S2:0.1,S3:0.2);**”



Tree terminology

Gene duplication, loss, and horizontal transfer

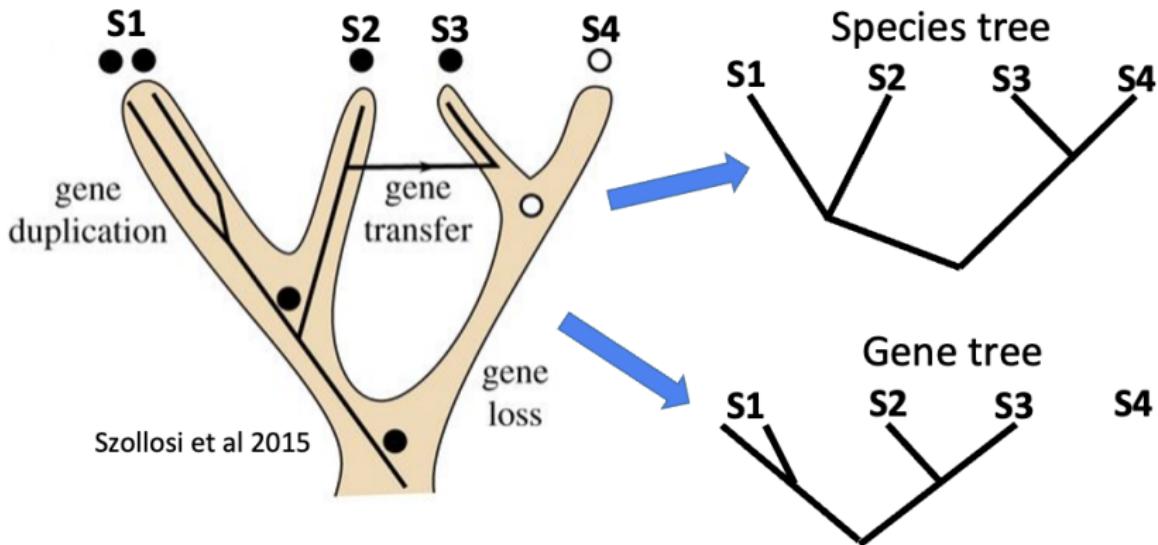
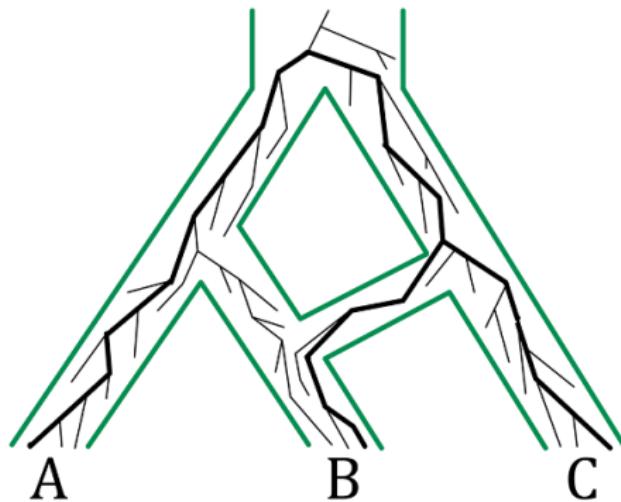


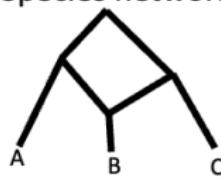
Figure – Software : ALE, PHYLDOLG

Tree terminology

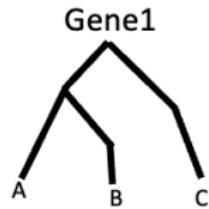
Hybridization



Species network



Gene1



Gene2

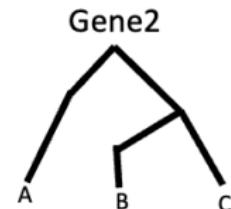


Figure – Taken from Wen and Nakhleh 2016, Software : phyloNet

Tree terminology

Due to recombination, different loci are within-species trees :

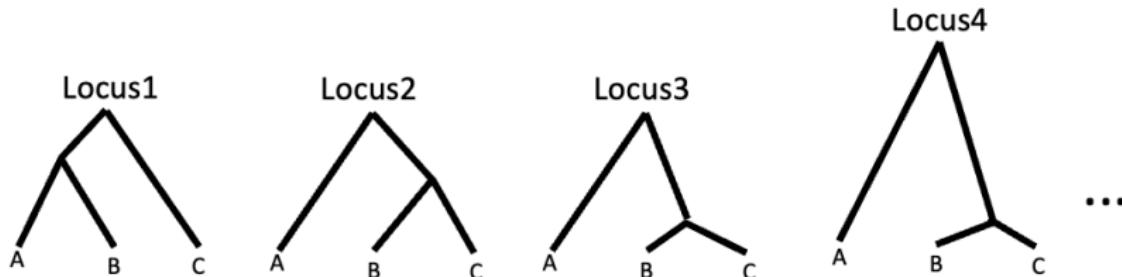


Figure – Software :ClonalFrameML, Gubbins

Tree terminology

Incomplete lineage sorting

Ancestral Polymorphism

Incomplete Sorting

Dmel

Dere

Dyak

Dana

Tree 2 Genealogy

Dmel Dere Dyak Dana

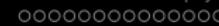
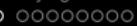
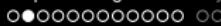
Polymorphisms Maintained Btwn Speciation Events

Figure – Taken from Pollard et al 2006, Some software : star-BEAST, SNAPP, BP&P

Summary

- A phylogeny is made of topology and branch lengths. It can be rooted or unrooted, binary or with multifurcations, ultrameric and/or timed. A phylogeny represents an evolutionary history.
- The Newick format is used for input and output by phylogenetic software.
- Sometimes there is not a single tree. Beware of hybridization, recombination, lateral gene transfer, gene duplications or losses, and incomplete lineage sorting.

Sequence alignment



Molecular data : alignment

- Some phylogenetic methods are alignment-free (e.g. based on k-mers)
- However, most phylogenetic methods need sequence alignments
- Aligning means putting sequence data in a matrix
- We do not consider genome alignment (which might also be needed). Here, we focus on small scale gene alignment

Sequence 1 **AGCAAGTATCAT**

Sequence 2 **AGCACAT**

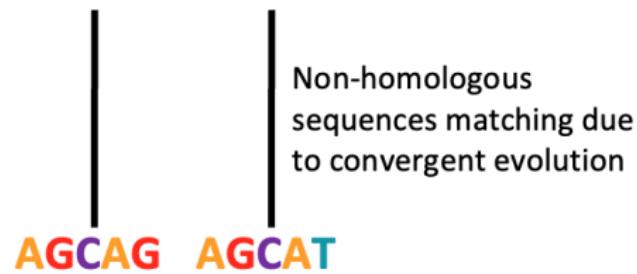
 **AGCAAGTATCAT**
AGCA - - - **CAT**

 **AGCAAGTATCAT**
- **AGC** - **A** - **CAT** - -

Molecular data : alignment

There are 2 types of alignment :

- Evolutionary alignment, defined by homology
- Structural alignment, exploring functional and structural similarity
- In phylogenetics we need evolutionary alignments



Which alignment is correct?

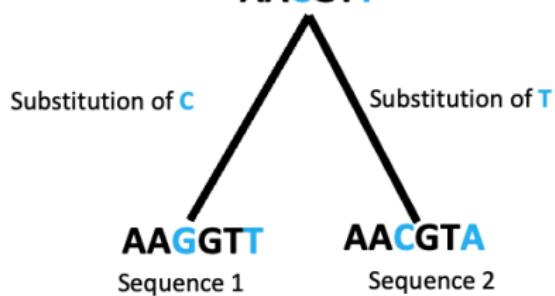
- 1) AGCAG - - - - AGCAT
- 2) AGCAG
 AGCAT

Molecular data : alignment

Pairwise alignment for two sequences

Ancestral sequence (unobserved)

AACGTT



Correct pairwise alignment:

Sequence 1 A A G G T T

Sequence 2 A A C G T A

Molecular data : alignment

Pairwise alignment for two sequences

Ancestral sequence (unobserved)

AACGTT

Deletion of C

AAGTT

Sequence 1

Insertion of G

AACGTGT

Sequence 2

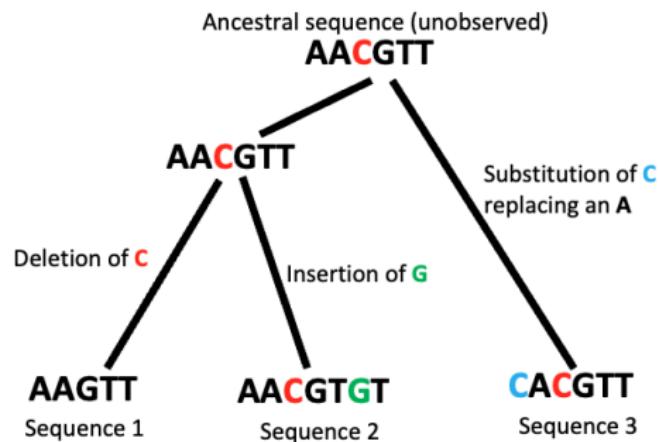
Correct pairwise alignment:

Sequence 1 AA - GT - T

Sequence 2 AA C GT GT

Molecular data : alignment

Multiple sequence alignment (MSA) is the alignment of multiple (more than 2) sequences



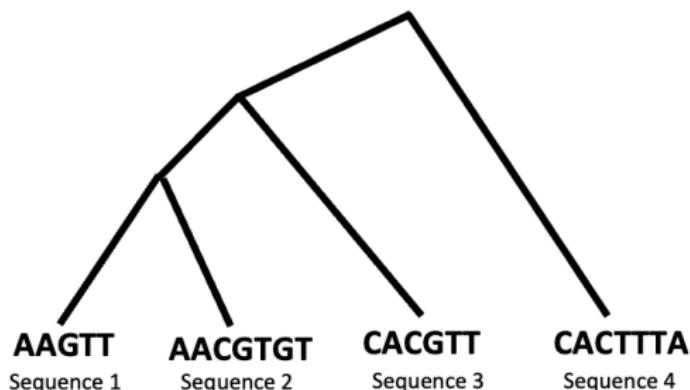
Correct MSA:

Sequence 1 **A A - G T - T**
Sequence 2 **A A C G T G T**
Sequence 3 **C A C G T - T**

Multiple sequence alignment (MSA)

Progressive alignment

- Align the two most closely related sequences
- Grow the MSA while keeping the previous sub-alignments constant
- Repeat until all sequences are aligned
- Note :** Early errors in alignment can snowball



Sequence 1	A A - G T - T -
Sequence 2	A A C G T G T -
Sequence 3	C A C G T - T -
Sequence 4	C A C T T - T A

Multiple sequence alignment (MSA)

Iterative alignment

- There are many MSA methods : Clustal, MAFFT, MUSCLE, PRANK, T-Coffee, ProbAlign, ...
- When sequences are similar, alignment is easy, and aligners will infer similar alignments
- At high divergence, different aligners or parameter values can give very different alignments

1) Start with an MSA

Sequence 1 **A A - G T - T -**
Sequence 2 **A A C G T G T -**
Sequence 3 **C A C G T - T -**
Sequence 4 **C A C T T - T A**



Sequence 1 **AAGTT**

2) Remove one sequence from MSA

Re-align it to rest of MSA

Sequence 1 **- A A G T - T -**
Sequence 2 **A A C G T G T -**
Sequence 3 **C A C G T - T -**
Sequence 4 **C A C T T - T A**



Sequence 2 **A A C G T G T -**
Sequence 3 **C A C G T - T -**
Sequence 4 **C A C T T - T A**

Multiple sequence alignment (MSA)

Recommendations

- MAFFT provides good accuracy and very good efficiency in most practical applications
- With high-divergence small datasets, PRANK can be more robust
- With very small datasets, BaliPhy can bypass alignments altogether

Nucleotide, amino acid, or codon ?

Nucleotide, amino acid, or codon ?

- Nucleotide alignments are highly informative at short divergence, but can be misleading at higher divergence due to saturation
- Amino acid alignments are preferred at high divergence ; they can also be used as a first step to create a codon alignment
- Codon alignments can be useful for coding sequences in particular for positive selection scans



Summary

- Usually, molecular phylogenetic methods require a multiple sequence alignment (MSA)
- Alignments can impact analyses of high-divergence data
- Alignment is easy at short divergence, but is hard and uncertain at high divergence
- Choice of nucleotide vs amino acid alignment mostly depends on divergence

Practical 2

Phylogenetic inference

Molecular Tree construction methods

Molecular Tree construction methods

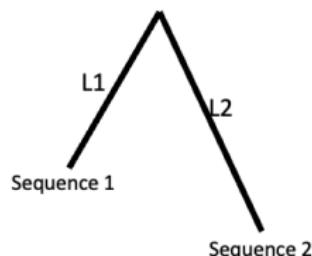
- Distance based methods
- Maximum Parsimony
- Maximum Likelihood
- Bayesian Inference

Important concepts

- Substitution Model selection
- Obtaining statistical support (bootstrapping)

Distance-based phylogenetics

Distance-based phylogenetics



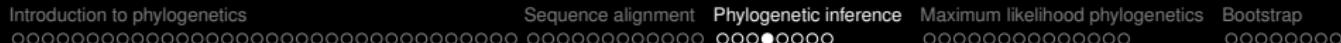
Evolutionary distance = L1 + L2

Sequence 1 AA - GT - T A A T G T G T A A G T
Sequence 2 A A C G T G T T A C G T G T A A G T

2 substitutions out of 16 match columns: $S=2/16$

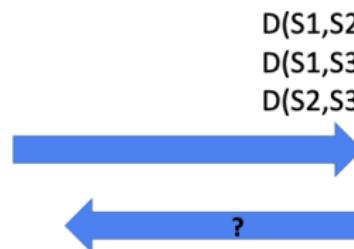
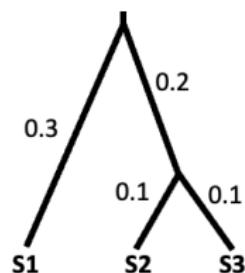
More distance, more substitutions. If S is the proportions of substitutions, the evolutionary distance D between sequences is estimated as:

$$D = -\frac{3}{4} \ln\left(1 - \frac{4}{3}S\right)$$

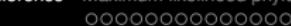


Distance-based phylogenetics

Distance-based phylogenetics Starting from the distances between all pairs of sequences (the distance matrix) we infer a phylogeny using Neighbor-Joining (NJ) or UPGMA. NJ is popular for fast tree inference, and as initial tree for more complex phylogenetic and alignment inference.



	S1	S2	S3
S1	0	0.6	0.6
S2	0.6	0	0.2
S3	0.6	0.2	0

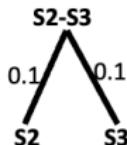


Distance-based phylogenetics

Neighbor-Joining

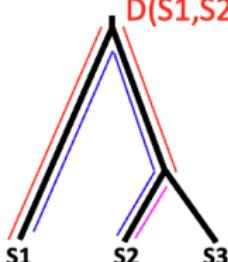
	S1	S2	S3
S1	0	0.6	0.6
S2	0.6	0	0.2
S3	0.6	0.2	0

S2 and S3 are closest to each other, so create clade (S2,S3) and internal node S2-S3



Replace S2 and S3 with S2-S3

$$D(S1, S2-S3) = D(S2, S1) - D(S2, S2-S3)$$



	S1	S2-S3
S1	0	0.5
S2-S3	0.5	0



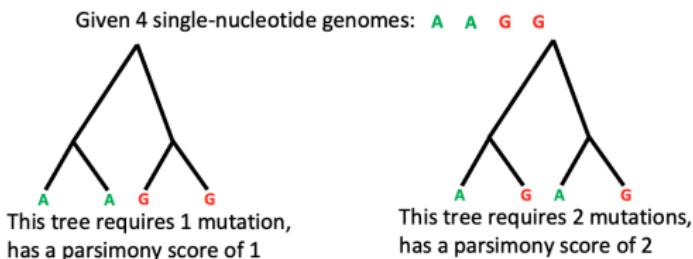
Distance-based phylogenetics

Summary

- Distance-based phylogenetic reconstruction uses pairwise evolutionary distances (the distance matrix)
- There are different methods to build a phylogeny from the distance matrix, like NJ
- Distance-based methods are fast but are considered less reliable. They are used for guide trees in MSA inference

Maximum Parsimony

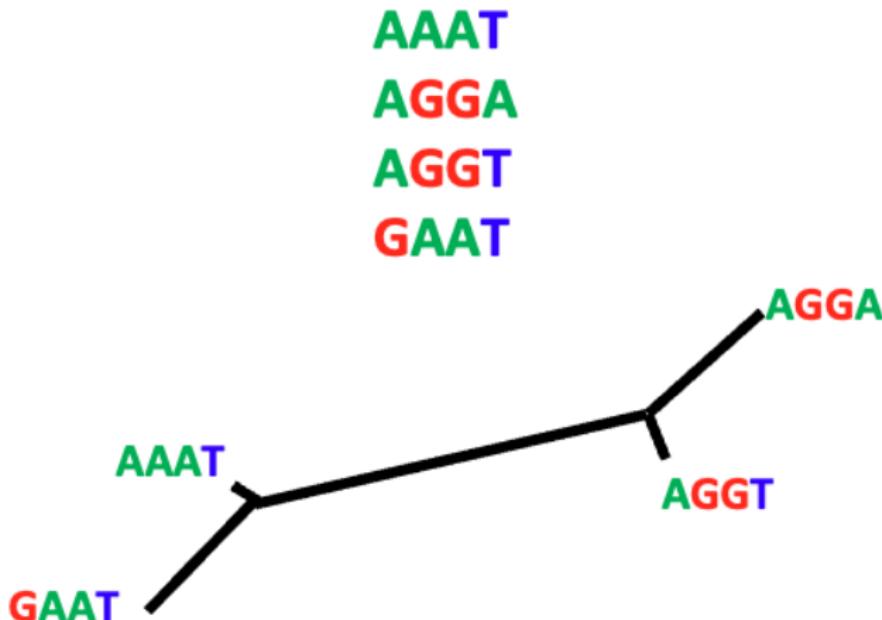
- Maximum parsimony methods assume the best tree requires the fewest mutations to explain the alignment
- What is the smallest number of evolutionary steps that lead to the observed state
- Simple enumeration is the simplest way, but not feasible for large data
- Heuristic algorithms needed for large datasets
- Often underestimates evolutionary change
- Long branch attraction
- More commonly used for morphological data

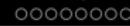
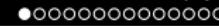


Maximum Parsimony

Example

- Maximum Parsimony tree relating these genomes





Maximum likelihood phylogenetics

Maximum likelihood phylogenetics

Conditional probability

- The probability of A conditional on B, or $P(A|B)$, is the probability of A assuming that B is true. If I work from home 50% of the time when it rains, and 80% when it doesn't :

$$P(\text{House} \mid \text{Rain}) = 0.5$$

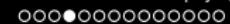
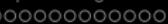
$$P(\text{House} \mid \text{Sunset}) = 0.8$$

Maximum likelihood phylogenetics

Likelihood - Given data (an alignment) and a hypothesis (a tree) the likelihood is the probability of the data conditional on the hypothesis

$$P(\begin{array}{c} \text{CGAC} \\ \text{CGAC} \\ \text{CGAT} \end{array} \mid \text{Tree})$$

- A hypothesis with a higher likelihood explains the data better - Maximum Likelihood aims at finding the tree with the highest likelihood



Maximum likelihood phylogenetics

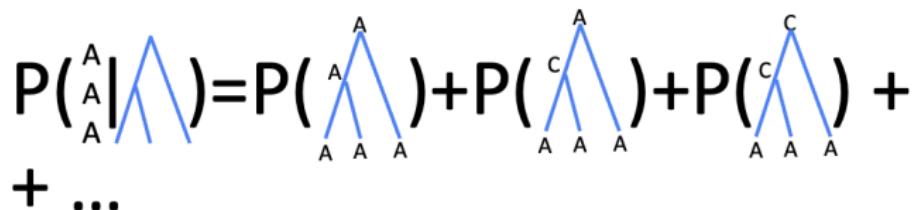
Phylogenetic Likelihood The likelihood of a tree is the product of the likelihood of the alignment columns (columns are assumed to evolve independently)

$$\begin{aligned} P(\text{CGAC} \mid \text{tree}) &= \\ P(\text{C} \mid \text{leaf}) * P(\text{G} \mid \text{leaf}) * P(\text{A} \mid \text{leaf}) * P(\text{C} \mid \text{leaf}) \end{aligned}$$

$$L = L_1 L_2 \cdots L_N = \prod_{j=1}^N L_j$$

Maximum likelihood phylogenetics

Phylogenetic Likelihood The likelihood of one alignment column is calculated by considering all possible states of the ancestors

$$P(A|A) = P(A|A) + P(C|A) + P(C|C) + \dots$$


This is done efficiently using dynamic programming (Felsenstein 1981 pruning algorithm)

Maximum likelihood phylogenetics

Phylogenetic Likelihood The probability of one nucleotide history can be calculated as the product of the probabilities of each branch

$$P(A) = P(A) * P(A) * P(A) * \dots$$

These probabilities are calculated using substitution matrices and matrix exponentiation

Maximum likelihood phylogenetics

Phylogenetic Likelihood

$$P(C \mid t \mid A)$$

Probabilities depend on the branch length and the two nucleotides A substitution matrix Q describes how often a nucleotide evolves into another

For short branch lengths: $P(C \rightarrow A) \approx tq_{CA}$

In general we use matrix exponentiation: e^{Qt}

Maximum likelihood phylogenetics

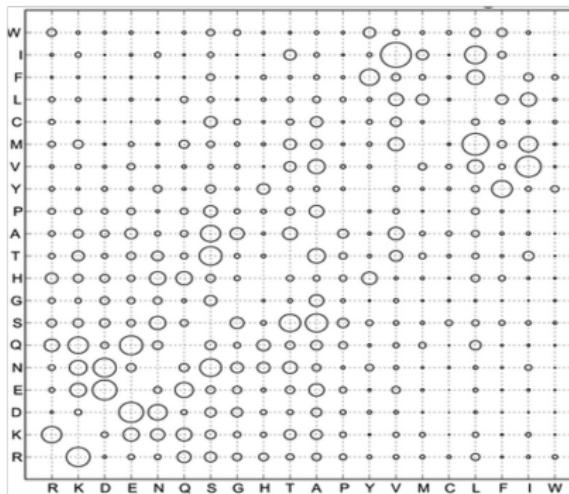
Substitution Models Substitution models can be parametric or empirical. Nucleotide models are usually parametric : defined from a few parameters estimated from the current data. For example :

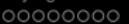
$$\begin{array}{c}
 \textbf{JC69} \\
 \left(\begin{array}{cccc}
 * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\
 \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\
 \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\
 \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & *
 \end{array} \right) \\
 \\
 \textbf{HKY85} \\
 \left(\begin{array}{cccc}
 * & \kappa\pi_G & \pi_C & \pi_T \\
 \kappa\pi_A & * & \pi_C & \pi_T \\
 \pi_A & \pi_G & * & \kappa\pi_T \\
 \pi_A & \pi_G & \kappa\pi_C & *
 \end{array} \right)
 \end{array}$$



Maximum likelihood phylogenetics

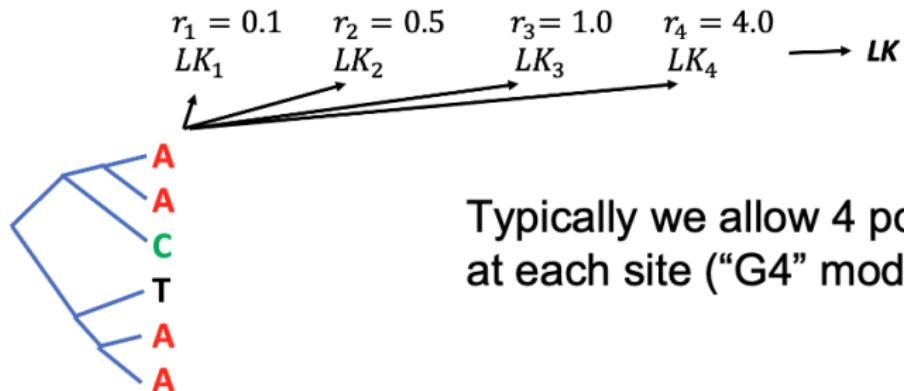
Substitution Models Amino acid models are usually empirical : estimated from large datasets and then applied to smaller ones without changes E.g., WAG (Whelan and Goldman 2001)





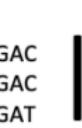
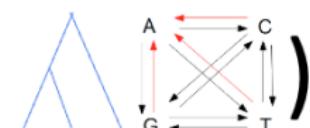
Maximum likelihood phylogenetics

Rate variation By default, we assume all genome positions and species evolve in the same way. This is not realistic, and rate variation along the genome can be modelled to improve inference.



Maximum likelihood phylogenetics

Maximum Likelihood

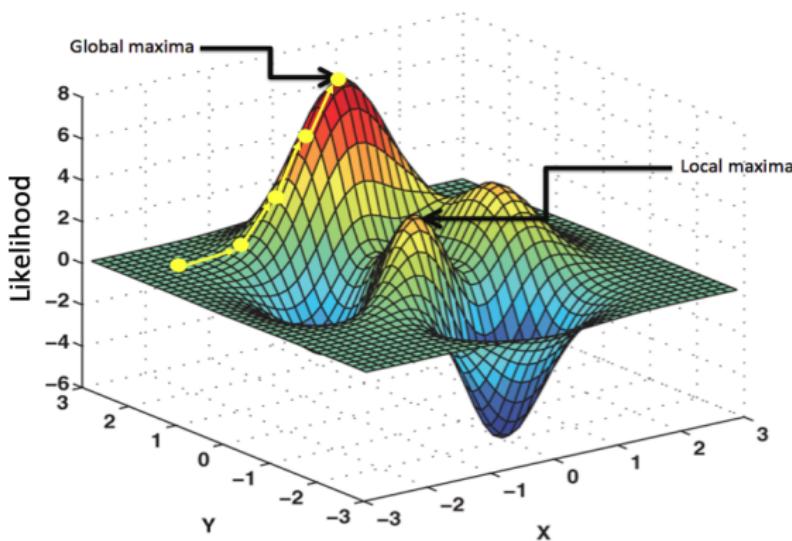
Likelihood: $P($ |   $)$

- Maximum likelihood infers not only the tree, but also model parameters in case of parametric models
- The number of possible trees is huge
- Instead of trying all possible trees, we heuristically modify a starting tree until no improvements can be found

Maximum likelihood phylogenetics

Maximum Likelihood

- The likelihood landscape
- We start from an initial random or NJ tree
- We propose new, similar trees, and accept them if they have higher likelihood
- When we reach an unimprovable tree, we stop



Maximum likelihood phylogenetics

Maximum Likelihood : Summary

- Maximum Likelihood (ML) searches for the tree with the highest likelihood as the best explanation for the data
- It infers one tree ; bootstrap can be used to assess uncertainty
- Slower than distance or parsimony methods, but still adequate for datasets with thousands of sequences
- It allows many models/features/analyses/tests, and is the most popular approach for tree inference due to its robustness

Practical 3

Bootstrap

Bootstrap

Bootstrap

- Maximum likelihood, parsimony, and distance methods reconstruct one tree
- How to measure confidence in the tree/branches ?
- A popular approach is bootstrap (Felsenstein, 1985)
- Each bootstrap replicate resamples alignment columns (with replacement) to create a new alignment of the same size as the original

Original alignment

The diagram illustrates the process of generating a bootstrap replicate. On the left, the text "Original alignment" is centered above a vertical sequence of four DNA-like strings. Each string consists of four positions: the first and third positions are green ('A' or 'G'), the second position is red ('G' or 'T'), and the fourth position is blue ('T' or 'A'). The strings are: "AAAT", "AGGA", "AGGT", and "GAAT". A vertical line of dots connects the bottom of the original alignment to the top of the bootstrap replicate. A horizontal line of dots connects the right side of the original alignment to the left side of the bootstrap replicate. A bracket on the right side of the original alignment encloses the second and third columns (the red and green positions). An arrow points from this bracket down to a smaller vertical stack of three boxes on the right, labeled "Bootstrap replicate #1". This smaller stack contains three boxes, each with one letter: the top box is green ('A'), the middle box is red ('G'), and the bottom box is red ('G').

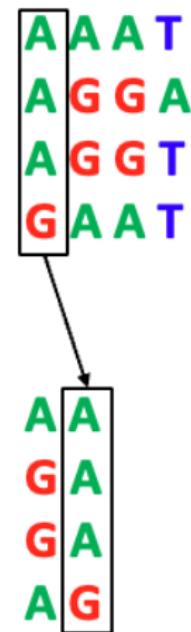
AAAT
AGGA
AGGT
GAAT

Bootstrap replicate #1

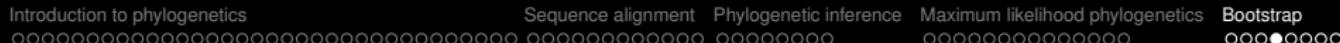
Bootstrap

Bootstrap

Original alignment



Bootstrap replicate #1



Bootstrap

Bootstrap

Original alignment

A	A	A	T
A	G	G	A
A	G	G	T
G	A	A	T

Bootstrap replicate #1

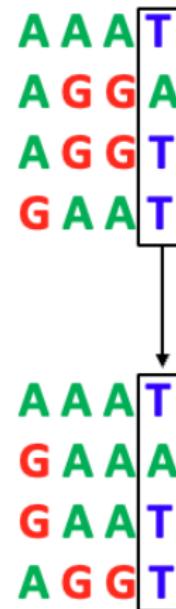
A	A	A
G	A	A
G	A	A
A	G	G

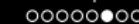
Bootstrap

Bootstrap

Original alignment

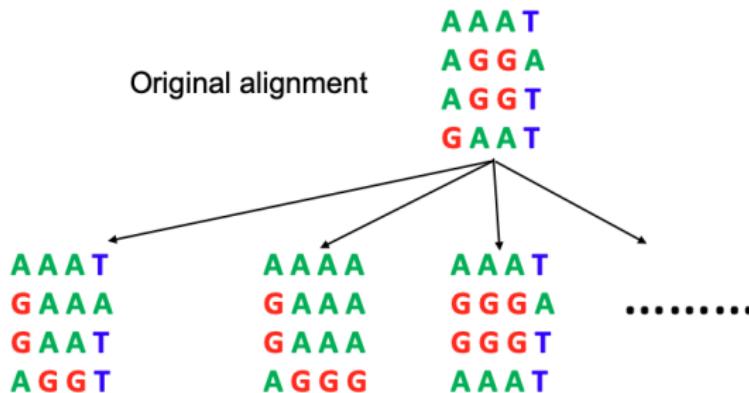
Bootstrap replicate #1





Bootstrap

Bootstrap

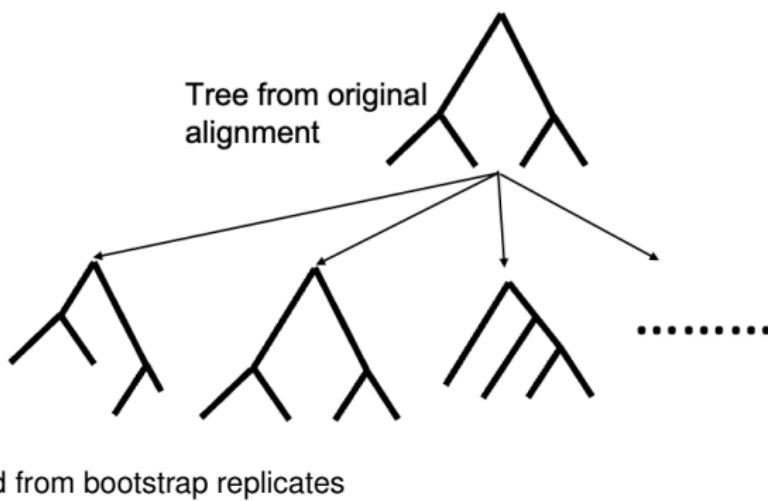


We create hundreds of such bootstrap replicate alignments

Bootstrap

Bootstrap

- From each alignment we then infer a bootstrap tree. The support of a branch in the initial tree is the proportion of times it is present in the bootstrap trees
- Typically 1000 replicates : it can be extremely slow
- There are faster approximations (e.g. UFBoot in IQ-TREE)



Trees inferred from bootstrap replicates

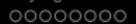
Bootstrap

Bootstrap : Summary

- The bootstrap creates replicate alignments from the true one and counts how often each clade occurs in the replicate trees
- Bootstrap is slow (e.g. 100x or 1000x slower than standard inference). Faster approximations are available such as UFBoot
- Other branch supports have different meaning : TBE represents the proportion of a clade that is stable. aLRT measures the confidence in a specific branching, rather than a clade

Practical 4

Bayesian phylogenetics inference



Bayesian inference

Bayesian inference

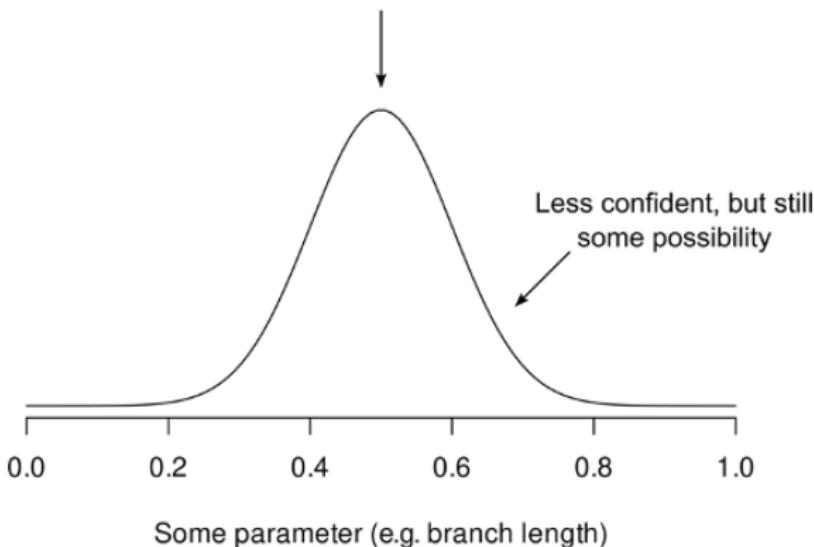
- The principle of likelihood asks the question : how probable are our data given our tree
- Many feel that this is the wrong question - generally it seems simpler to ask how probable your hypothesis (tree) is, given some data
- To ask this question, we must use Bayesian inference

Bayesian inference

Bayesian inference

- The Bayesian view is that probability represents a way of encoding beliefs about a hypothesis
- High probabilities mean that I am pretty certain of this hypothesis, but I express my uncertainty about it with a distribution

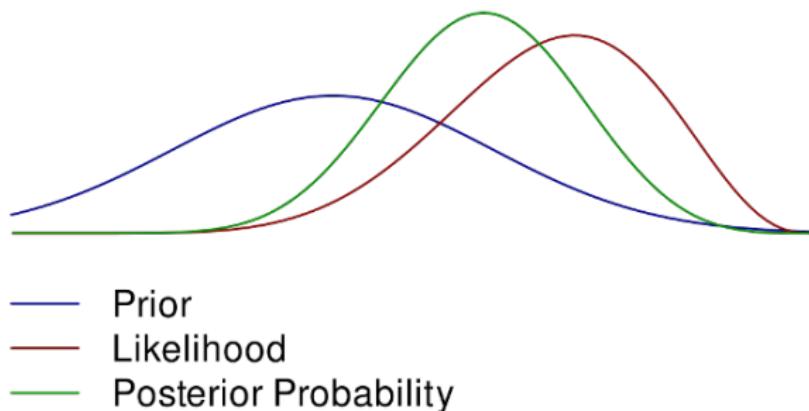
Most confident the value is around here

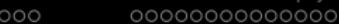


Bayesian inference

Bayesian inference Conceptually, it combines two sources of information to calculate Posterior Probabilities :

- Information from the data, encoded by the likelihood
- Information from our prior beliefs, the prior distribution





Bayesian inference

Bayesian inference

Likelihood: $P($

CGAC
CGAC
CGAT

|

)

Maximum likelihood looks for the tree and parameters that maximize the likelihood

Posterior: $P($

CGAC
CGAC
CGAT

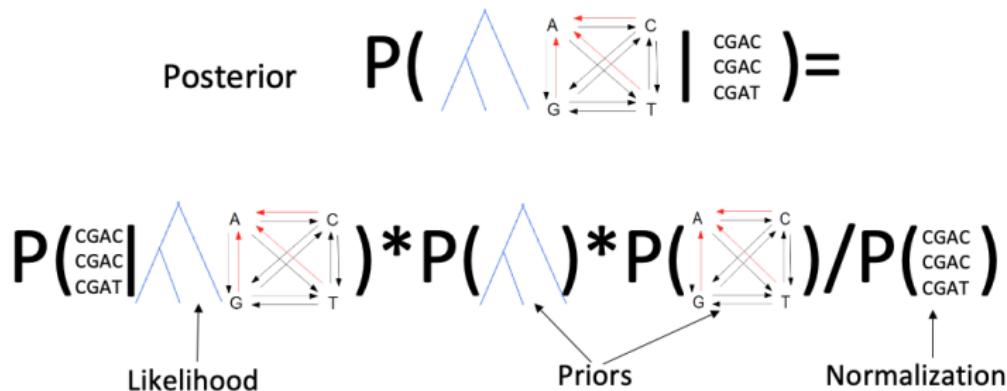
|

)

Bayesian methods seek the posterior probability of tree and parameters

Bayesian inference

Bayesian inference



- Priors do need to be specified by the user
- We do not need to calculate the normalization factor if we use Monte Carlo Markov Chain (MCMC) methods

Bayesian inference

Priors

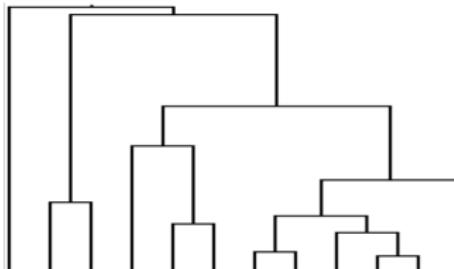


Figure – Yule and Birth-death priors model speciation/extinction and epidemics

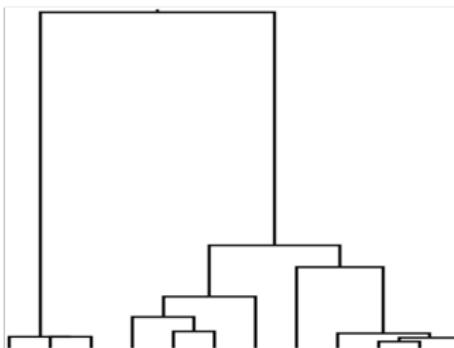


Figure – Coalescent priors are used for within-population data

Bayesian inference

Bayesian inference

- The difference between likelihood and posterior matters the most when data is limited there otherwise uncertainty (e.g., high or low divergence)
- Given this alignment of two species, what is the maximum likelihood divergence ?
- Is this a reasonable estimate ?
- The maximum likelihood divergence is 0
- This would mean the 2 species are the same, which is usually wrong
- A Bayesian method takes into account the uncertainty that comes from small data and low divergence

Monte Carlo Markov Chain

Monte Carlo Markov Chain Similar moves to Maximum Likelihood tree search, but :

- ML (usually) accepts only moves that increase the likelihood
- MCMC can accept moves that decrease the posterior

$$A(x'|x) = \min \left(1, \frac{P(x')}{P(x)} \frac{g(x|x')}{g(x'|x)} \right)$$

acceptance probability

Posterior

proposal rate

Posterior

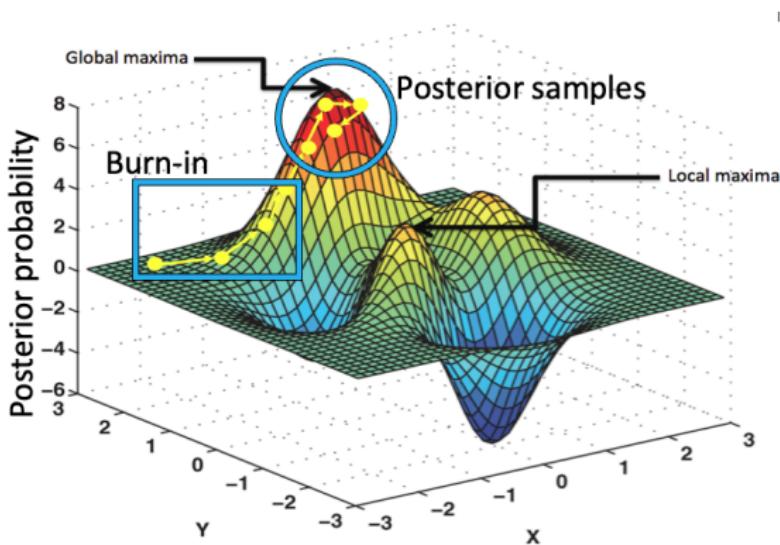
proposal rate

Figure – Metropolis-Hastings acceptance

Monte Carlo Markov Chain

Monte Carlo Markov Chain

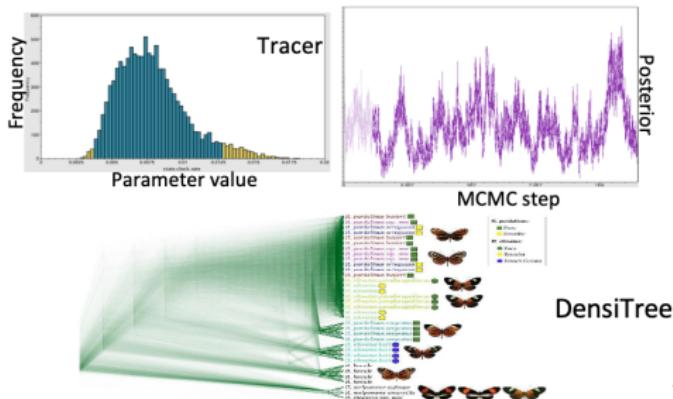
- The start of MCMC looks similar to ML
- MCMC moves can however also decrease the posterior
- Initial samples (“burn-in”) are then discarded
- In the final sample trees are represented proportionally with their probability



Bayesian inference

Bayesian inference

- Methods we have seen so far (distance, parsimony and maximum likelihood) estimate one tree
- Bayesian methods instead assess the probabilities of different trees : the output is a collection of trees
- Bayesian analyses provide multiple trees and parameter values as output



Bayesian inference

Bayesian inference : Summary

- Bayesian methods use prior probabilities of trees and parameters, and infer posterior distributions
- They are generally slower than ML
- They are good at measuring uncertainty
- They can use broad ranges of models and data

Practical 5

Some phylogenetic applications

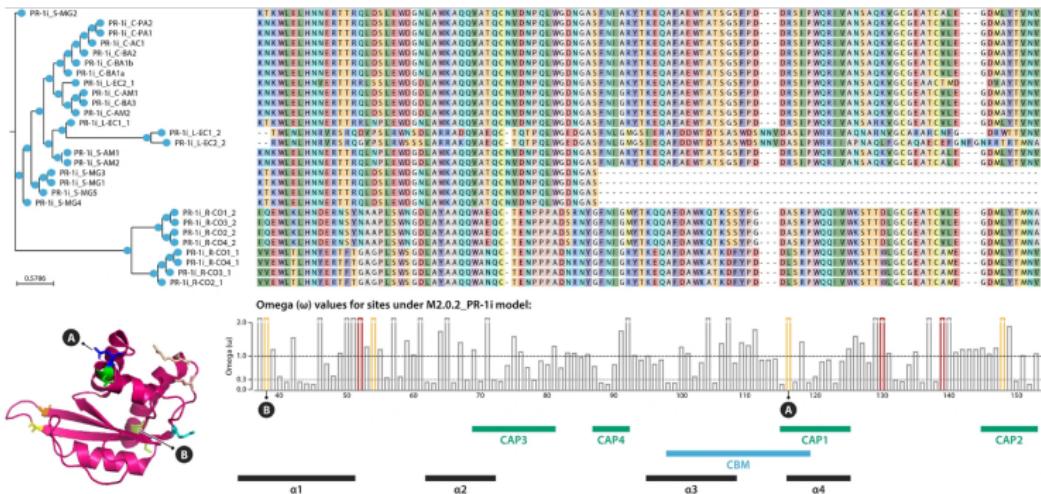
Inference of phylogeny from Molecular data

Inference of phylogeny from Molecular data : Recap

Distance/parsimony	Maximum Likelihood	Bayesian
MEGA	<u>FastTree</u>	
ape (R)	RaxML	MrBayes
bioPython	<u>IQTree</u>	BEAST
UShER	PhyML	
Fast and simple but without complex models	Often more accurate, more computational demand	Slow but accurately measure uncertainty. Broadest range of models and applications.

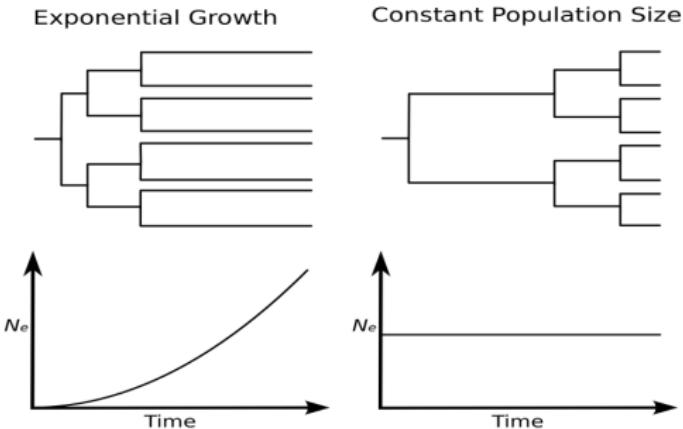
Some phylogenetic applications

Positive selection scans Statistical tests can highlight genes, gene positions, and branches under selection. Software : PAML, HyPhy



Some phylogenetic applications

Phylodynamics The shape of a phylogenetic tree can reveal past population/epidemiological dynamics. This is because the probability of branching/coalescing depends on the population size.



Some phylogenetic applications

Phylodynamics From the tree shape we can reconstruct the history of prevalence or population size changes. Software : BEAST

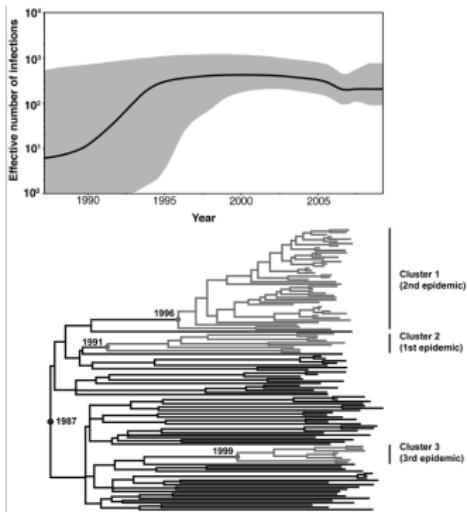
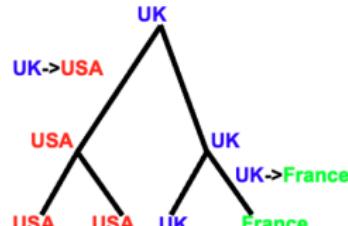
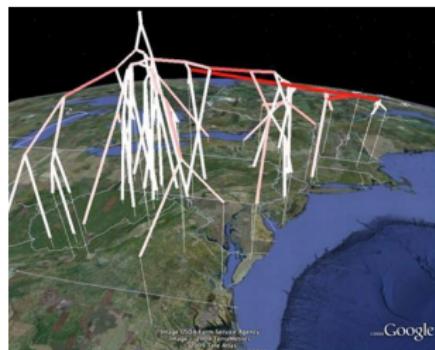


Figure – Hon-Kwan Chen et al 2011

Some phylogenetic applications

Viral Phylogeography

- We can integrate genetic sequences with other data like time, phenotype, or geographic information
- When we use time and geographic information we can do phylogeography : the study of spread in space and time using phylogenies
- Phylogeography can be done in discrete space (e.g. between countries) or continuous space using coordinate data. Software : BEAST
- Statistical tests can highlight genes, gene positions, and branches under selection. Software : PAML, HyPhy



Practical 6