Phylogenetic inference using maximum likelihood

Comparative Genomic Analyses

Rui Borges
Vetmeduni Vienna

Phylogeny inference

Sequences evolve according to an unknown tree, so within the phylogenetic framework, we want to estimate it.

- tree structure or topology
- branch lengths
- model of sequence evolution (JC, GTR ...) and respective parameters

Maximum likelihood principle

The likelihood function tells us the probability of the data given a set of parameters.

$$L = p(D|\tau, \theta)$$

- D is a set of aligned sequences
- τ represents the tree: branch lengths and topology
- θ represents the model of evolution parameters: substitution rates

Maximum likelihood principle

Some parameters produce the sequences with higher probability than others

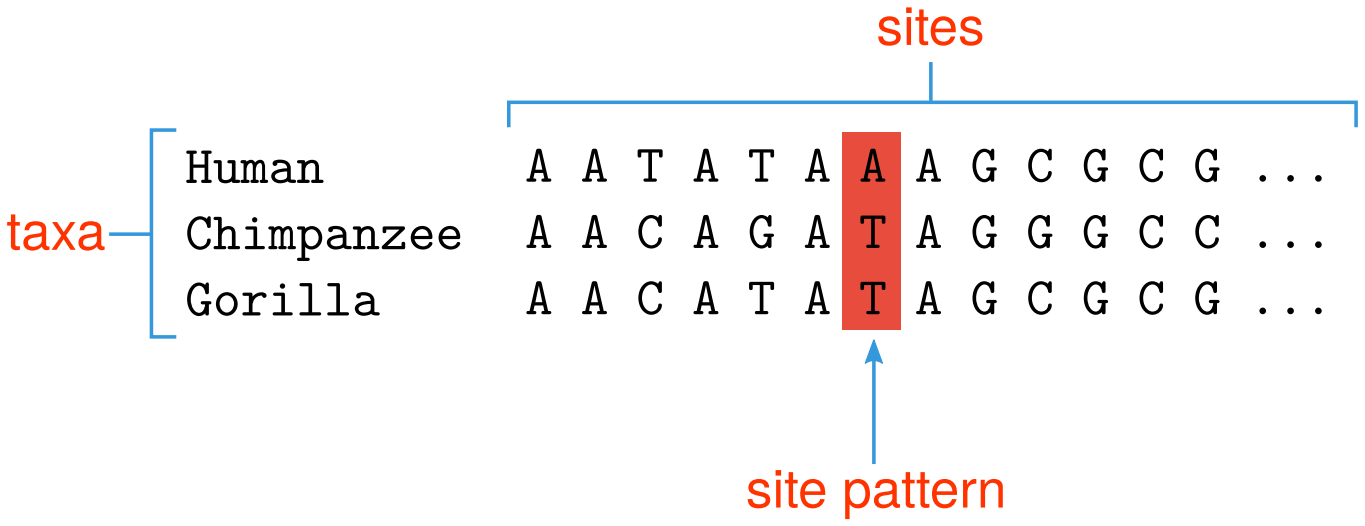
- we want the tree topology, branch lengths, and model parameters that best explain the observed the sequences: i.e., $\hat{\tau}$ and $\hat{\theta}$ that maximize the **likelihood function**

$$L = p(D|\hat{\tau}, \hat{\theta})$$

- $\hat{\tau}$ and $\hat{\theta}$ are the **maximum likelihood estimates**

Alignment and site patterns

A sequence alignment includes information from N taxa and S sites.



- a **site pattern** includes information from a single site an alignment is a collection of site patterns

Alignment and site patterns

The probability of the whole alignment can be obtained from the probability of each site pattern.

$$p(D|\tau,\theta) = \prod_{i=1}^{S} p(d_i|\tau,\theta)$$

- D is the whole alignment and d_i the i-th site pattern
- assume independent evolution at each site

Probability of a site pattern

The goal is to compute the probability of each and every observed site patterns in an alignment of N sequences.

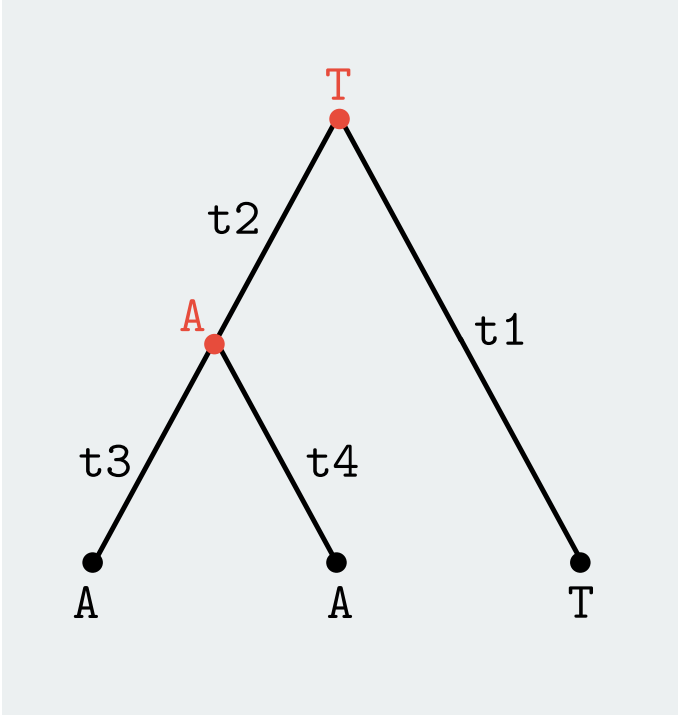
```
Human       A A T A T A A A G C G C G ...
Chimpanzee  A A C A G A T A G G G C C ...
Gorilla     A A C A T A T A G C G C G ...
```

$$p(D|\tau,\theta) = p(\{AAA\}|\tau,\theta) \times p(\{AAA\}|\tau,\theta) \times p(\{TCC\}|\tau,\theta)...$$

- calculating $p(d_i|\tau,\theta)$ is a small likelihood problem

Probability of a site pattern

The probability of a site pattern given the tree, the model of evolution and the ancestral states

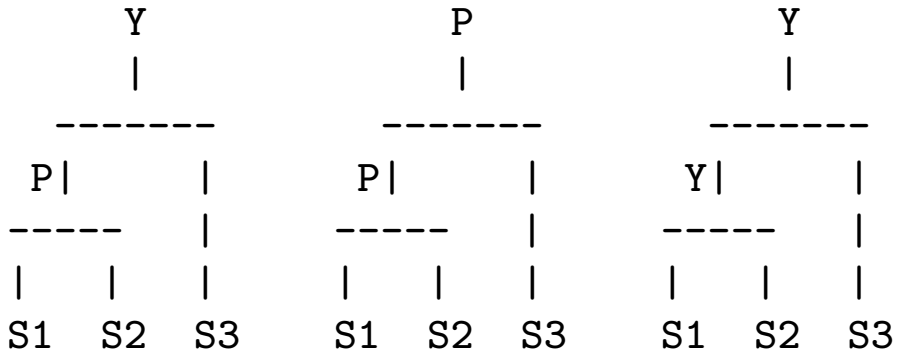


$$p(\{A, T, T\}|\tau, \theta) = p(T \to T|t_1) \times p(T \to A|t_2) \times p(A \to A|t_3) \times p(A \to A|t_4)$$

Probability of a site pattern

Exercise

The pyrimidine $Y = \{C, T\}$ and purine $P = \{A, G\}$ content in a certain site of a protein coding genes, was observed for three species: $\{S_1, S_2, S_3\} = \{P, P, Y\}$. Despite one does not know the species tree and the ancestral states, six scenarios for the evolution of this site were proposed:



Assuming that transitions between pyrimidines and purines occur with probability

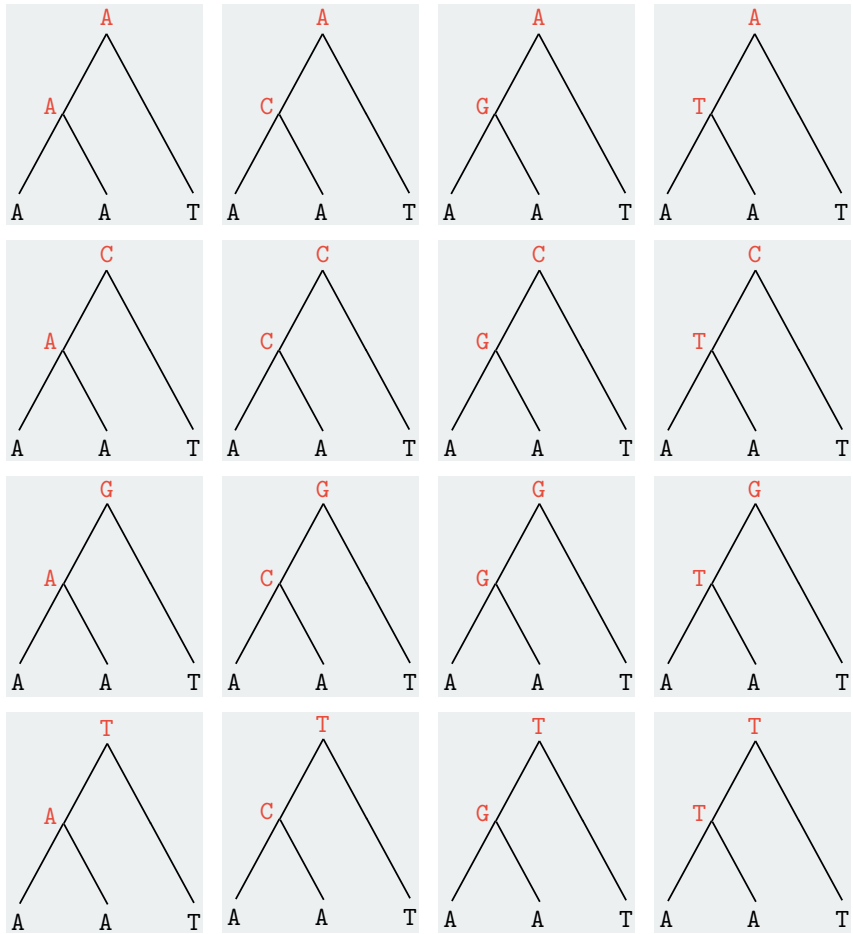
$$\begin{array}{c} \\ P \\ Y \end{array} \begin{array}{c} \begin{array}{cc} P & Y \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \end{array},$$

what is the most likely scenario?

Probability of a site pattern

We know how to calculate the likelihood of a site patterns when the ancestral states and the topology are both known.

- the ancestral states are unknown
- there are $n-1$ internal nodes in a tree with n taxa meaning that there are 4^{n-1} possible sets of ancestral states
- efficient algorithms exist: Felsenstein's pruning algorithm
  Felsenstein (1981)



Maximum likelihood tree

A possible approach to finding the maximum likelihood tree.

- maximizing the likelihood of an alignment for a given tree and model parameters is feasible, but we want the tree that best described the data
- try out several trees and find the one that maximizes the likelihood function

Maximum likelihood tree

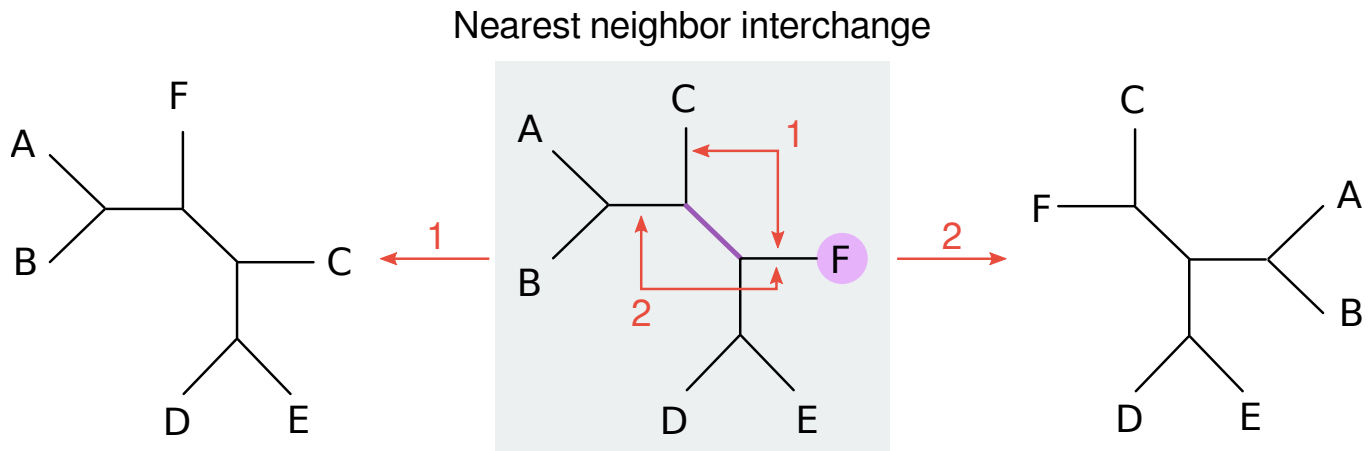
Finding the maximum likelihood tree has some computational limitations.

- there is a huge space of possible topologies
- testing all possible trees is just impossible, even for moderately sized data sets

number of taxa	possible unrooted trees
3	1
5	15
10	2 027 025
50	8.5×10^{74}
100	5.1×10^{182}

Maximum likelihood tree

Because testing all the possible trees is not computationally feasible, several algorithms are used to suggest reasonable trees.



- **full-tree arrangement operations**: change the structure of a given tree within its neighborhood

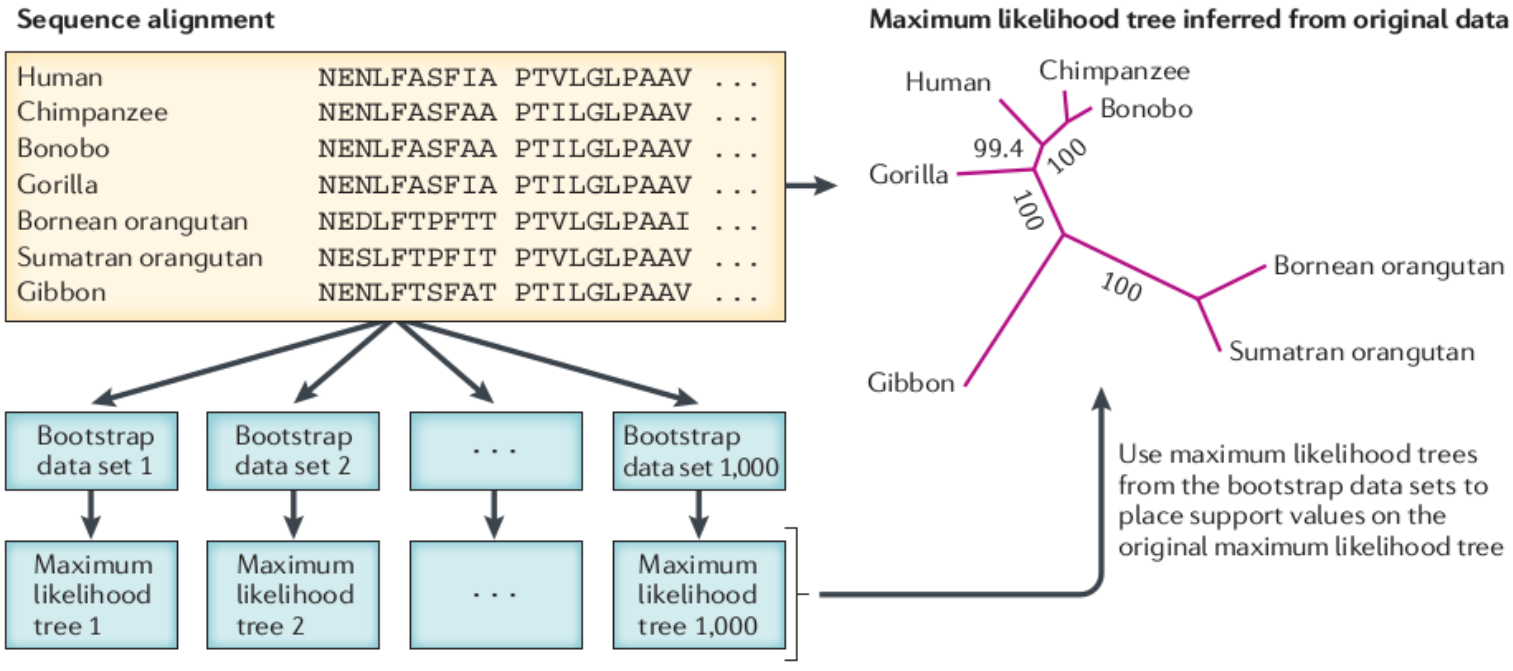
Maximum likelihood tree

Several measures are used to assess the certainty of a tree or its clades: the most widely used approach is the **bootstrapping**.

Efron (1979) and Felsenstein (1985)

- pseudo-alignments are created by subsetting the alignment
- pseudo-trees are inferred for each pseudo-alignment
- bootstraps represent the number of times a certain clade is present in the pseudo-trees

Maximum likelihood tree



Yang & Rannala (2012)

Literature

The Phylogenetic Handbook by Lemey, Salemi and Vandamme (2009) Cambridge University Press

- Chapter 6: sections 6.1, 6.2 and 6.3, 6.4 and 6.5