Bayesian Inference of Phylogenetic Trees

David Maxen [1] [2] Supervisor: Georgios Aliatimis [1]

[1]STOR-i Lancaster University [2]University of Warwick

1. Introduction

Since Darwin published his theory of evolution, humans have been trying to understand how different species evolved. This is both important both for understanding the origins of life, and predicting the evolution of fast evolving species such as HIV. Tree diagrams called **phylogenetic trees** are commonly used to represent the branching lineages of different species.

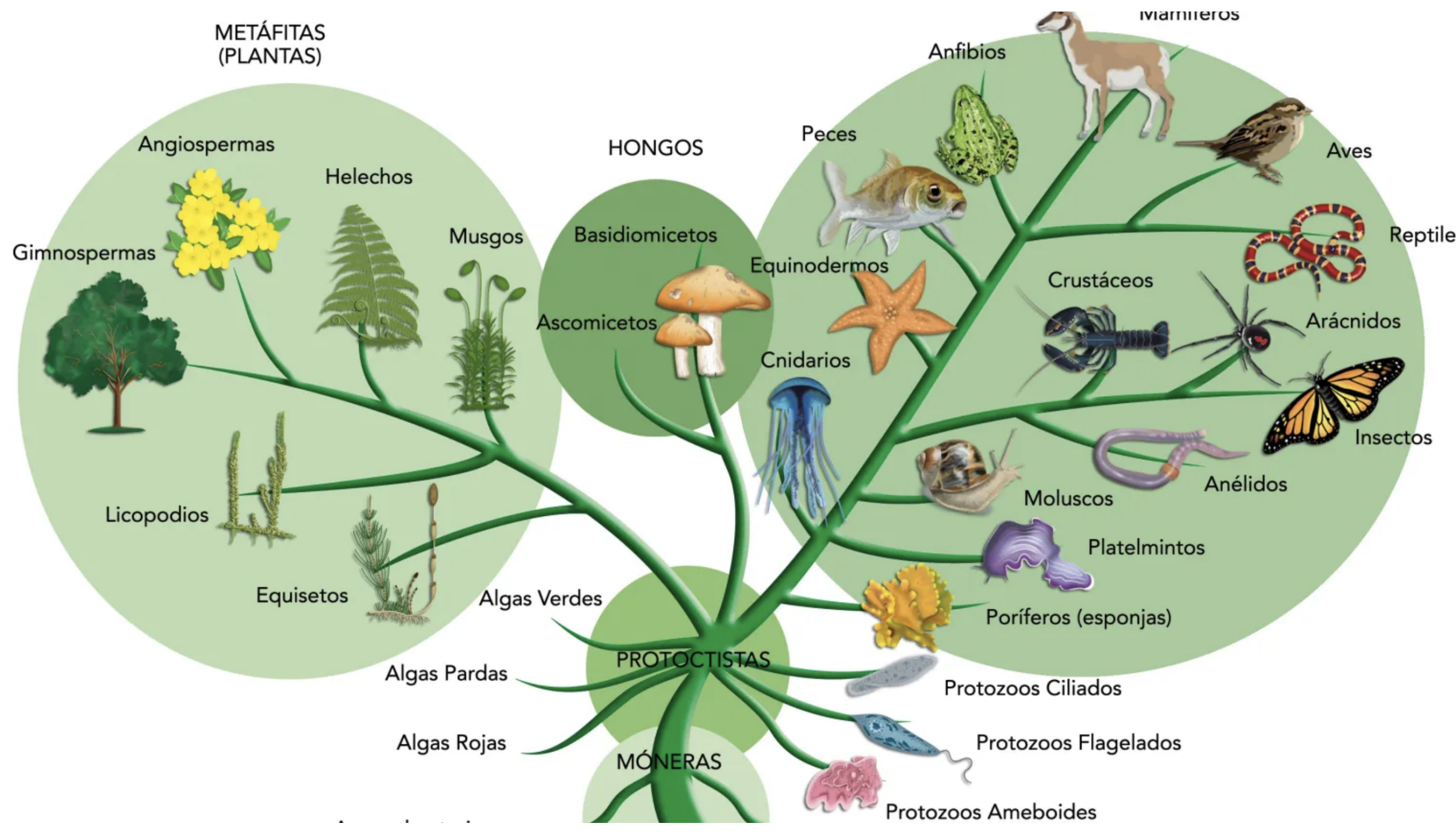


Figure 1. An example phylogenetic tree. Credit: fabelacorrea / Adobe Sto

Originally, phylogenetic trees were reconstructed using the physical features of animals, such as bone structure. The guiding principle is that the more similar species are, the more recently their lineage split. However, advances in science have let to an abundance of **DNA sequence data** that can be used for more accurate **inference of phylogenetic trees**.

2. Problem Statement

Our aim is to find the distribution of different phylogenetic trees given sequence data from different species. In particular, a tree is specified by both the structure, or **topology**, of the tree, and its **branch lengths**. This can be interpreted as a posterior distribution, and we use Bayes' theorem to rewrite it as the product of the likelihood and prior.

$$p(tree|data) \propto p(data|tree)\, p(tree)$$

There are a few simplifying assumptions we make when calculating the likelihood, such as different branches and nucleotides evolving independently. With these assumptions, our likelihood model reduces to specifying how a single nucleotide evolves over time.

3. Models of DNA Evolution

We consider the evolution of a nucleotide as a continuous time Markov chain on the state space $\{A, T, C, G\}$. The instantaneous rate matrix, or Q-matrix, specifies the rates at which each state mutates to another state. This gives us up to twelve parameters to infer during our computation, corresponding to the six double-sided arrows in the diagram below.

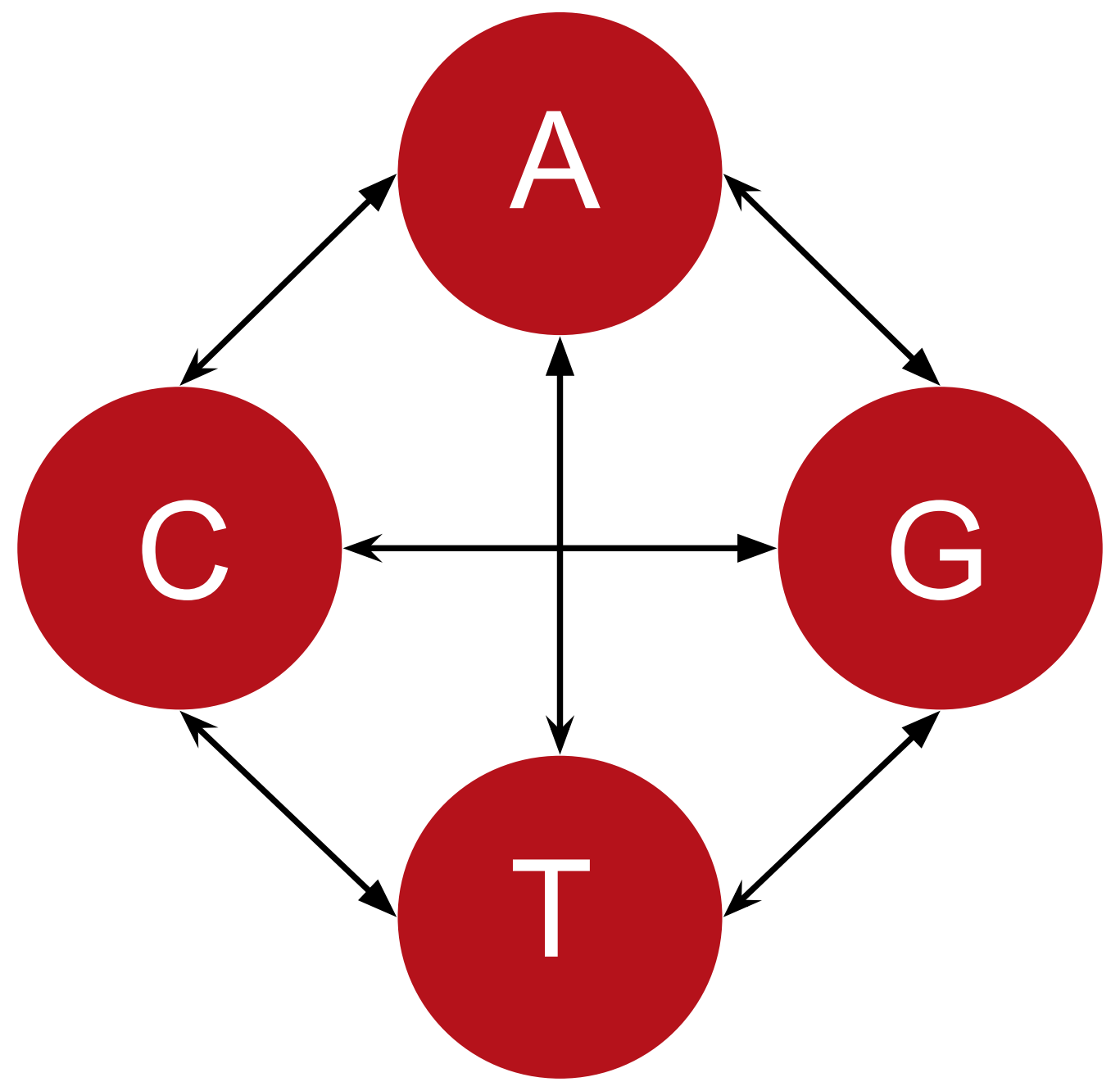


Figure 2. Diagram of the continuous Markov chain used to model DNA mutations.

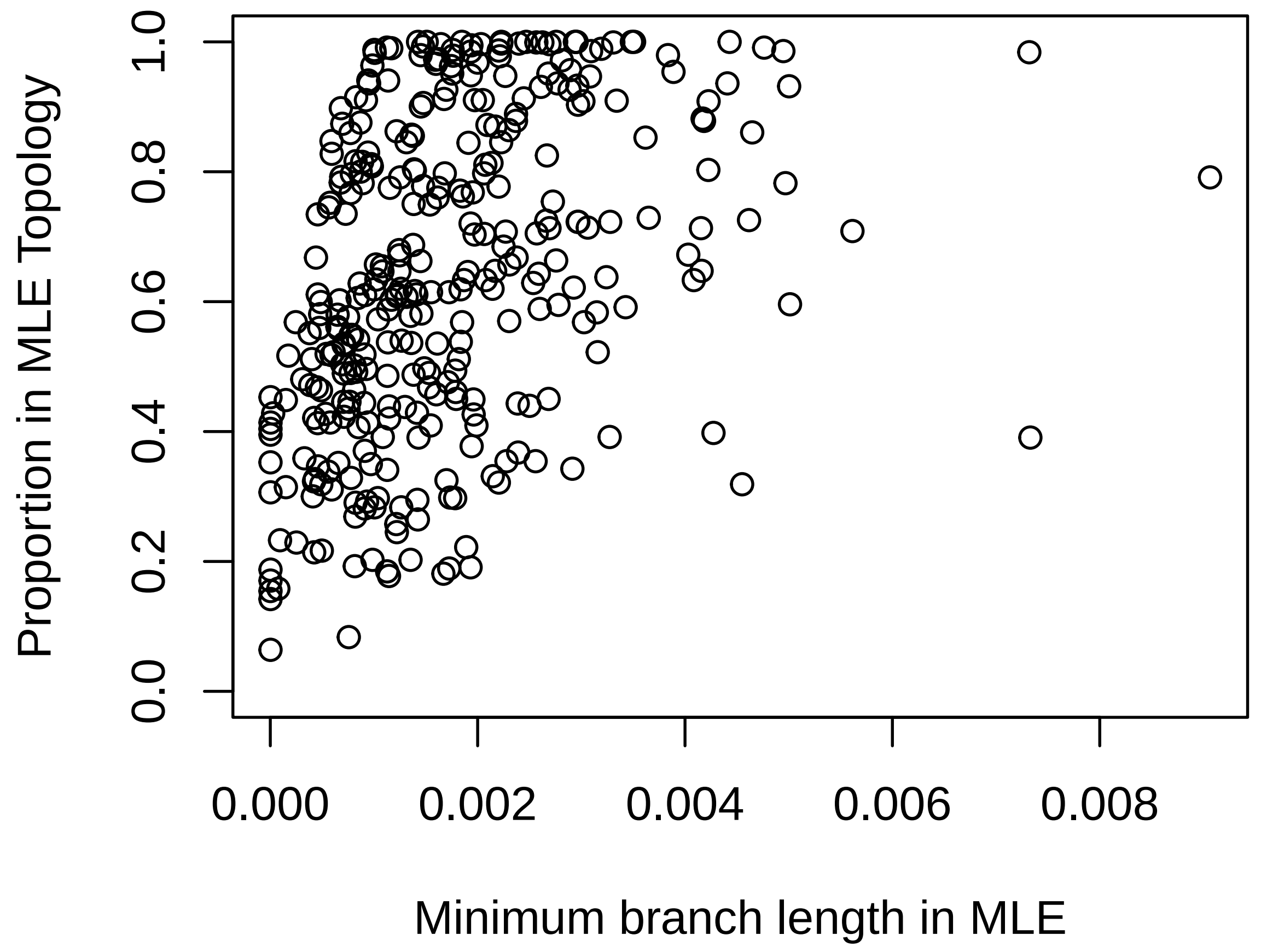
There are a wide variety of different models used in the literature, each corresponding to a different rate matrix of the Markov chain.

- JC Model (Jukes and Cantor 1969)
  This is the most simplistic model. Each state $\{A, T, C, G\}$ mutates to another state at a constant rate.
- K80 model (Kimura 1980)
  This model expands on the JC model by accounting for chemical similarities between the nucleotides A, G (purines) and C, T (pyrimidines). Hence, an A mutating to a G is more likely than it becoming a C or T. This model has two different rates of mutation, to model this similarity or difference in chemical structure.
- GTR Model (Tavaré 1986)
  One extremely useful property of a Markov chain is time-reversibility. The GTR (generalised time-reversible) model allows us to specify the most general model possible that is time reversible. This allows the underlying nucleotide frequencies, and all six rates of mutation, to be different. We used an adaptation of this model to compute our results.

Armed with these models, we can now compute the likelihood of a tree. This allows us to analyse DNA sequences using maximum likelihood estimators (MLEs) and using Markov Chain Monte Carlo (MCMC).

4. Results

One question we asked is whether there is a correlation between the structure of the MLE tree and the distribution of the samples from the MCMC. We used MrBayes3 [2] for the MCMC, and IQ-Tree [1] for the MLE.



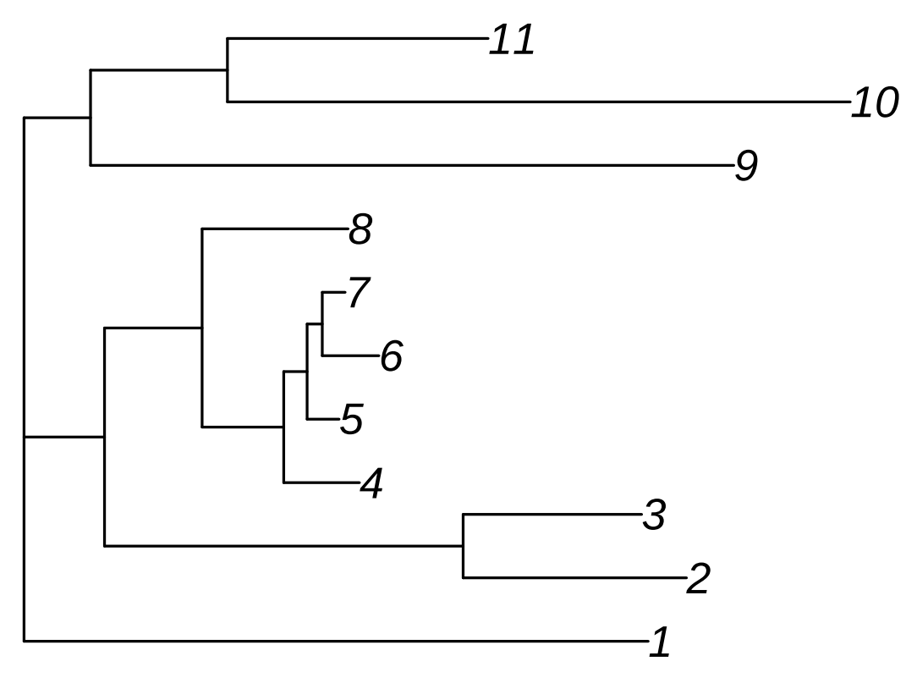


Figure 3. MLE Tree for gene 262

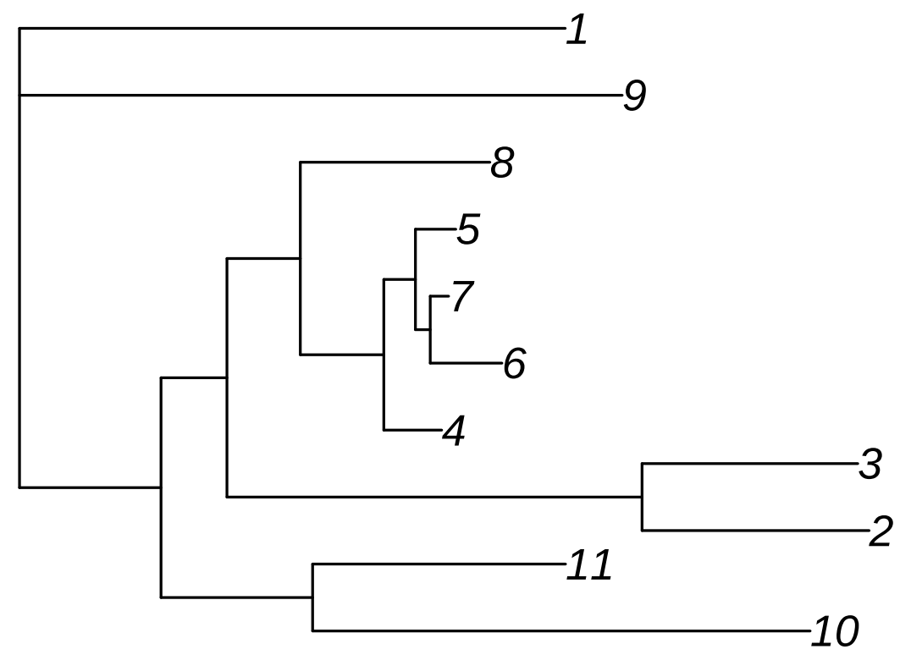


Figure 4. Tree in MCMC majority topology for gene 262

References

[1] O. Chernomor D. Schrempf M.D. Woodhams A. von Haeseler R. Lanfear B.Q. Minh, H.A. Schmidt.
IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era.
2020.

[2] Fredrik Ronquist and John P. Huelsenbeck.
MrBayes 3: Bayesian phylogenetic inference under mixed models.
2003.