

A phylogenetic tree is a visual representation of the evolutionary relationships among organisms, inferred from their biological sequences. The process of creating/constructing a phylogenetic tree has many steps:

1. **Sequence Data Collection:** Gather RNA, DNA, or protein sequences from multiple organisms.
2. **Sequence Alignment:** Align the sequences to identify similarities & differences.
3. **Model Selection:** Choose a suitable evolutionary model to account for the sequence variations.
4. **Tree Construction:** Use algorithms like Maximum Parsimony (MP), Maximum Likelihood (ML), or Bayesian Inference (BI) to construct the phylogenetic tree.
5. **Tree Evaluation:** Assess the tree's reliability & accuracy by using metrics such as bootstrap values or posterior probabilities.

One method of tree construction, Bayesian inference shall be discussed in much more detail here.

In a Bayesian analysis, inferences of phylogeny are based upon the posterior probabilities of phylogenetic trees. The posterior probabilities of the  $i$ th phylogenetic trees ( $T_i$ ) conditional on an alignment of DNA sequences ( $X$ ) can be calculated using Bayes' Theorem:

Ⓟ

$$f(T_i|X) = \frac{f(X|T_i) \cdot f(T_i)}{\sum_{j=1}^{B(B)} f(X|T_j) \cdot f(T_j)}$$

where

$$f(X|T_i) = \int \int f(X|T_i, v, \theta) f(v, \theta) dv d\theta$$



this summation is over all  $B(s)$  trees that are possible for a species  $[B(s) = \frac{(2s-5)!}{2^{s-3}(s-3)!}]$  {unrooted} &

$B(s) = \frac{(2s-3)!}{2^{s-2}(s-2)!}$  {rooted}], & the integral is over

all combinations of branch lengths ( $v$ ) & substitution parameters.

The prior for phylogenetic trees is  $f(T_i)$  & is usually set to

$$f(T_i) = \frac{1}{B(s)}$$

the prior on ~~the~~ branch lengths ( $v$ ) &  $\theta$  ~~are~~ is denoted by  $f(v, \theta)$ .

Typically, the likelihood function  $[f(X|T_i, v, \theta)]$  is calculated under the assumption that substitution occur acc. to a time-homogeneous Poisson process.

The summation & integrals ~~are~~<sup>are</sup> in a Bayesian analysis cannot typically be evaluated analytically. MRBAYES, the popular software that popularised BI uses Markov chain Monte Carlo (MCMC) to approximate the posterior probabilities of trees. ~~MCMC is~~

The basic MCMC algorithm works as follows:

(1) A new state for the chain is proposed using a stochastic mechanism.

(2) the acceptance probability for this new state is calculated. The acceptance probability is equal to the minimum of 'one' or the likelihood ratio [likelihood of new state to old state] times the prior ratio [prior



probability of new state ~~to~~ to old state] times the proposal ratio (probability of proposing the old state to that of the new state).

3. A uniform random variable is drawn b/w (0,1). If the no. is  $<$  acceptance probability, new state is accepted, else rejected.

This process is repeated thousands, or even millions of times. The proportion of the time any single tree is visited during the course of the chain is a valid approximation of its posterior probability. Metropolis coupled MCMC (MCMC<sup>3</sup>) is also a common alternative used in BI.

To summarise, there are 3 key concepts in BI:

- ① Prior Distribution: A probability distribution assigned to the phylogenetic tree, reflecting our initial beliefs about ~~the~~ <sup>particular tree</sup>.
- ② Likelihood Function: Describes the probability of observing the molecular sequence data given a topology.
- ③ Posterior Distribution: Updated probability distribution

Its advantages are:

- ① Robustness to Model Violations, such as relative branch-length differences, & model violations.
- ② Quantifies ~~the~~ uncertainty in tree topology & branch lengths.
- ③ It provides flexibility to accommodate various models of evolutionary processes.

CITATION: "MRBAYES: Bayesian inference of phylogenetic trees" by John P. Huelsenbeck & Frederick Ronquist (2001)