# Unexplained Behavior of Phylogeny Estimation Methods

Achutha Balaji †, Yuheng Cai †, Ruoran Huang †, Megan Lolling †, Mengwei Sun †, Max Bacharack †, Sebastien Roch †

†Department of Mathematics, University of Wisconsin-Madison

## Introduction

A ***Phylogeny*** is a tree diagram that depicts the lines of evolutionary descent of different species that share a common ancestor. The goal of the ***Phylogeny Estimation*** is to assemble a tree representing a hypothesis about the evolutionary ancestry of a set of genes, species, or other taxa.

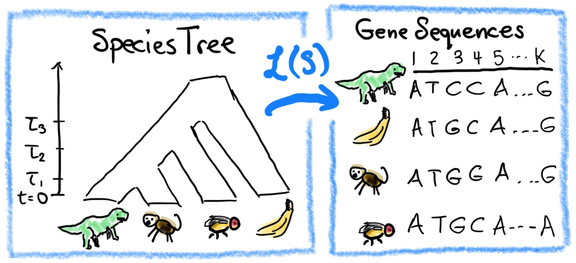


Fig. 1: From Gene Sequence Data to Species Tree

**Problem:** Does the ***Long Branch Attraction (LBA)*** (long branches being incorrectly placed together on a phylogentic tree [2]) introduce bias to the ***Maximum Likelihood Estimation (MLE)*** in phylogenetic tree reconstruction even for simple evolutionary models and small-scale trees?

**Objective:** Analytically calculate the optimality condition (parameters that give the maximum likelihood) of all possible configurations of the ***unrooted 4-leaf tree with two long branches*** through a combination of the ***MLE method*** and the ***Homotopy Continuation Method***.

## Definitions & Our Setting

We consider a phylogentic tree $S$ with three taxa, namely, 1,2,3 and 4. Of interest is the asymptotic case in which the length of two of the edges tends to infinity.

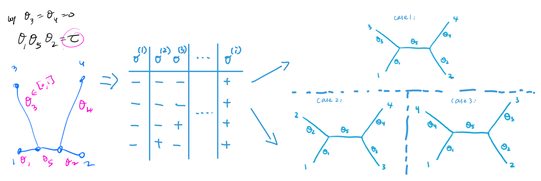

Fig. 2: Three Possible Tree Configurations

We assume a binary symmetric site substitution process (as in [3]; this is the JC69 model [1] in which purines ($+1$) and pyrimidines ($-1$) are not distinguished). It may be helpful to think of the elements of $S$ as the possible outcomes of running the site substitution process once on the tree, which generates a single binary nucleotide (either $+1$ or $-1$) for each of the three leaves of the species tree.
Our observations consist of $k$ i.i.d random variables $X_1, \ldots, X_k$ taking values in the 16-element set $S$ defined by

$$S := \{(\sigma_1, \sigma_2, \sigma_3, \sigma_4) : \sigma_1, \sigma_2, \sigma_3, \sigma_4 \in \{-1, +1\}\},$$

We define the likelihood estimation as the product of the probability of seeing all the observed site patterns in the dataset, i.e.

$$L_{x_1,\ldots,x_n} = \prod_{k=1}^{n} P[X_k = \sigma^{x_k}]$$

In our project, we are investigating into the cases where $\theta_3 = \theta_4 = 0$ and $\theta_1\theta_2\theta_5 = \tau$ or $\theta_1\theta_2 = \tau$ for $\tau \in [0, 1]$.

## Simulation of Sequence Data

1. **Data**

   The probability mass function for $X_i$ (observations of the tree, i.e., the quadruplets of nucleotides obtained by running the following simulation process), according to the truth configuration described in our setting, is given by the following formula:

   $$q_j := \mathbb{P}\left[X_i = \sigma^{(j)}\right] = \frac{1}{16}\left(1 + \sigma_1^{(j)}\sigma_2^{(j)}\theta_1\theta_5\theta_2\right) \tag{1}$$

   Suppose we have $k$ observations, we define $Y_i = e_j$ if $X_i = \sigma^{(j)}$ for $j = 1, \ldots, 16$, where $e_j$ are the standard basis of $\mathbb{R}^{16}$. From this, we introduce the frequency vector $F_k$ (which follows the multinomial distribution) of the 16 possible site patterns:

   $$F_k = [f_1, f_2, \ldots, f_{16}]^T = Y_1 + \cdots + Y_k$$

2. **Approximation of Multinomial by Multivariate Normal**

   Multivariate Central Limit Theorem: Suppose that $X_i \in \mathbb{R}^k$ are i.i.d random vectors with mean $\mu$ and covariance $\Sigma$, with finite variance for each component, then we have:

   $$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}(X_i - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma) \tag{2}$$

   Applying the theorem, we get

   $$\frac{1}{\sqrt{k}}(Y_1 + \ldots + Y_k) \sim \mathcal{N}\left(\sqrt{k}\mu, \Sigma\right) \tag{3}$$

   $$F_k \sim \mathcal{N}(k\mu, k\Sigma) \tag{4}$$

   where $k$ is the number of observations. It follows from the distribution of $Y_i$ that

   $$k\underline{\mu} = k\underline{q} \in \mathbb{R}^{16} \text{ and } k\Sigma = k\left(\text{diag}(\underline{q}) - \underline{q}\underline{q}^\top\right) \in \mathbb{R}^{16\times 16} \tag{5}$$

   Therefore, we can simulate the frequency vector by multivariate normal distribution with specified parameters (mean vector and covariance matrix) given the truth configuration.

## Analytic Methods

One of the way to find the tree that best fit the given data is through maximum likelihood estimation. For convenience, we use the negative log likelihood function

$$-\log(L) = -\sum_{i=1}^{n}\log P(X_i = \sigma^{(j)}) = -\sum_{i=1}^{n}\log\left(\frac{1}{16}(1 + \sigma_1^{(j)}\sigma_2^{(j)}\tau)\right)$$

We then simplified the maximum likelihood function into a least square problem by applying Taylor Expansion to the negative log of the likehood function:

$$\begin{aligned} -\log(L) &\approx -\sum_{\underline{\sigma}} k(f_{\underline{\sigma}} + \frac{z_{\underline{\sigma}}}{\sqrt{k}})\ln(f_{\underline{\sigma}} + \frac{\hat{\epsilon}_{\underline{\sigma}}}{\sqrt{k}}) \\ &= -\sum_{\underline{\sigma}} ka\ln a + \sum_{\underline{\sigma}}\sqrt{k}z_{\underline{\sigma}} + \frac{1}{2}\sum_{\underline{\sigma}}(a^{-1})(\hat{\epsilon}_{\underline{\sigma}} - z_{\underline{\sigma}})^2 + O(\frac{1}{\sqrt{k}}) \end{aligned}$$

where we did another Taylor Expansion for $a = f_{\underline{\sigma}} + \frac{z_{\underline{\sigma}}}{\sqrt{k}}$ to get the least square term in next section.

## Minimization Software

Our goal is to minimize the least squares formula:

$$\frac{1}{2}\sum_{\sigma} f_\sigma^{-1}(z_\sigma - \epsilon_\sigma)^2$$

where $f$ is the probability of observing a certain site pattern, $z_\sigma$ is our data, and $\epsilon_\sigma$ is a function of our branch lengths that differs based on the tree configuration.

1. **Homotopy Continuation**

   We form a system of equations with constraints specific to each tree configuration; that is, $\theta_1\theta_2\theta_5 = \frac{1}{2}$ for cases 1 and 3, and $\theta_1\theta_2 = \frac{1}{2}$ for case 2 (additional constraints are added to cover boundary cases as well). Using Lagrange multipliers, we set up our Lagrangian as:

   $$\frac{1}{2}\sum_{\sigma} f_\sigma^{-1}(z_\sigma - \epsilon_\sigma)^2 + \lambda * g$$

   After differentiating with respect to our branch lengths, we solve the system with a set of simulated data to output the real solutions of our branch length parameters as critical points.

2. **Julia**

   We are using the JuMP and nonlinear optimization packages to solve the optimization problem. We directly bring in the variables from the Least Square optimization problem and optimize directly through the nonlinear objective function and constraint.

   $$\min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\alpha_5}} \quad \frac{1}{2}\sum_{j=1}^{16} q_j^{-1}(z_j - \hat{\epsilon}_j)^2$$

## Challenges

- The two minimization software do not always yield consistent results. For certain boundary cases, i.e., when branches 3 and/or 4 are infinitely long, Homotopy Continuation produces no real solutions, while the Julia software still finds a result. Even when starting from the given solution, Homotopy still cannot compute a start system due to the added constraint variables. Further research is needed to determine why this occurs.

## References

[1] Thomas H Jukes, Charles R Cantor, et al. “Evolution of protein molecules”. In: *Mammalian protein metabolism* 3 (1969), pp. 21–132.

[2] Sarah L Parks and Nick Goldman. “Maximum likelihood inference of small trees in the presence of long branches”. In: *Systematic Biology* 63.5 (2014), pp. 798–811.

[3] Ziheng Yang. “Complexity of the simplest phylogenetic estimation problem”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1439 (2000), pp. 109–116.