WISCERS Report

Zhirui Guo

August 24, 2023

1 Introduction

In this report, we aim to construct an unrooted four-species phylogenetic tree with one long branch by producing best estimates of the location for the fourth branch (i.e. the long branch) given a three-species tree. As changing the branch lengths of the three-species tree, we find how the fractions of five separate cases (i.e. attaches to branch 1, attaches to branch 2, attaches to branch 3, attaches to the intersection of three branches, branch 4 is infinitely long) changes. With the visualized statistics over how often each situation arises, we observed that the long branch shows tendency to be attached to the longest branch among the three branches, which shows bias that can be resulted from Long Branch Attraction.

2 Definitions and Notations

2.1 The True Model

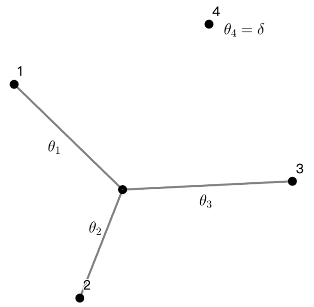


Figure 1: The true species tree in our model, under which data is generated

We consider an unrooted species tree with four leaves, denoted as 1, 2, 3, and 4. Leaf 4 is characterized by an infinitely long branch length.

We define parameter set H = $\{\theta_1, \theta_2, \theta_3, \delta\}$, which measure the amount of correlation between two nucleotides at the ends of the edge.

2.2 The Data

Given a quartet of binary nucleotides, there are eight possible site patterns of the form $xxxx, xxxy, xxyx, xyxx, xxyy, xyxy, xyyx, xyyy$, where x and y take different values in {-1, 1}.

$$n_{xxxx} = f_{++++} + f_{----}, n_{xxxy} = f_{+++-} + f_{---+}$$
$$n_{xxyx} = f_{++-+} + f_{--+-}, n_{xyxx} = f_{+-++} + f_{-+--}$$
$$n_{xxyy} = f_{++--} + f_{--++}, n_{xyxy} = f_{+-+-} + f_{-+-+}$$
$$n_{xyyx} = f_{+--+} + f_{-++-}, n_{xyyy} = f_{+---} + f_{-+++}$$

That is, n_{xxxx} is the number of observations taking the values (+1, +1, +1, +1) or (-1, -1, -1, -1), and similarly for the remaining notations.

3 Likelihood Calculation

3.1 Case 1. Leaf 4 is attached to branch 1

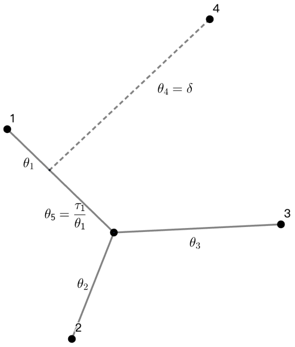


Figure 2: Leaf 4 is attached to branch 1

Using probability mass function of X from [1], we have the following probability:

$$\begin{aligned}\mathbb{P}[X = \sigma] =& \frac{1}{16}(1 + \sigma_1\sigma_2\theta_1\theta_5\theta_2 + \sigma_1\sigma_3\theta_1\theta_5\theta_3 + \sigma_1\sigma_4\theta_1\theta_4 + \sigma_2\sigma_3\theta_2\theta_3 + \sigma_2\sigma_4\theta_2\theta_5\theta_4 \\ &+ \sigma_3\sigma_4\theta_3\theta_5\theta_4 + \sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_2\theta_3\theta_4)\end{aligned} \quad (1)$$

Upon plugging in the probabilities into the likelihood function as outlined in [2], we derive the log likelihood function for case 1:

$$\begin{aligned}\log(L) &= \sum_{i=1}^{k} \log P(x_i = \sigma^{(i)}) \\ &= n_{xxxx} \log(1 + \theta_2\tau_1 + \theta_2\theta_3 + \theta_2\theta_5\delta + \theta_3\tau_1 + \theta_1\delta + \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\ &+ n_{xxxy} \log(1 + \theta_2\tau_1 + \theta_2\theta_3 - \theta_2\theta_5\delta + \theta_3\tau_1 - \theta_1\delta - \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\ &+ n_{xxyx} \log(1 + \theta_2\tau_1 - \theta_2\theta_3 + \theta_2\theta_5\delta - \theta_3\tau_1 + \theta_1\delta - \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\ &+ n_{xyxx} \log(1 - \theta_2\tau_1 - \theta_2\theta_3 - \theta_2\theta_5\delta + \theta_3\tau_1 + \theta_1\delta + \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\ &+ n_{xxyy} \log(1 + \theta_2\tau_1 - \theta_2\theta_3 - \theta_2\theta_5\delta - \theta_3\tau_1 - \theta_1\delta + \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\ &+ n_{xyxy} \log(1 - \theta_2\tau_1 - \theta_2\theta_3 + \theta_2\theta_5\delta + \theta_3\tau_1 - \theta_1\delta - \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\ &+ n_{xyyx} \log(1 - \theta_2\tau_1 + \theta_2\theta_3 - \theta_2\theta_5\delta - \theta_3\tau_1 + \theta_1\delta - \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\ &+ n_{xyyy} \log(1 - \theta_2\tau_1 + \theta_2\theta_3 + \theta_2\theta_5\delta - \theta_3\tau_1 - \theta_1\delta + \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta)\end{aligned}$$

To find δ that maximize the likelihood, we can take the first derivative w.r.t. δ and find all of its critical points.

$$
\begin{aligned}
0 = \frac{d\log(L)}{d\delta} &= \frac{n_{xxxx}A}{P + A\delta} - \frac{n_{xxxy}A}{P - A\delta} \\
&+ \frac{n_{xxyx}B}{Q + B\delta} - \frac{n_{xxyy}B}{Q - B\delta} \\
&+ \frac{n_{xyxy}C}{R + C\delta} - \frac{n_{xyxx}C}{R - C\delta} \\
&+ \frac{n_{xyyy}D}{S + D\delta} - \frac{n_{xyyx}D}{S - D\delta}
\end{aligned} \tag{2}
$$

where

$$
\begin{aligned}
&A = \theta_2\theta_5 + \theta_1 + \theta_3\theta_5 + \theta_1\theta_2\theta_3, B = \theta_2\theta_5 + \theta_1 - \theta_3\theta_5 - \theta_1\theta_2\theta_3, \\
&C = \theta_2\theta_5 - \theta_1 - \theta_3\theta_5 + \theta_1\theta_2\theta_3, D = \theta_2\theta_5 - \theta_1 + \theta_3\theta_5 - \theta_1\theta_2\theta_3, \\
&P = 1 + \theta_2\tau_1 + \theta_2\theta_3 + \tau_1\theta_3, Q = 1 + \theta_2\tau_1 - \theta_2\theta_3 - \tau_1\theta_3, \\
&R = 1 - \theta_2\tau_1 - \theta_2\theta_3 + \tau_1\theta_3, S = 1 - \theta_2\tau_1 + \theta_2\theta_3 - \tau_1\theta_3
\end{aligned}
$$

Since the second derivative w.r.t. δ is negative, $\log(L)$ is concave. Thus, global maximum is achieved at the critical point.

3.2 Case 2. Leaf 4 is attached to branch 2

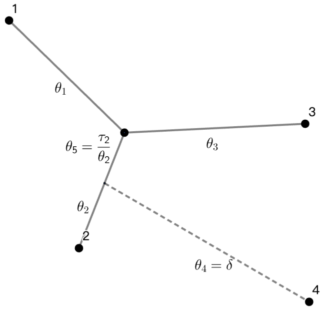


Figure 3: Leaf 4 is attached to branch 2

Using the probability mass function of X from [1], we have the following probability:

$$
\begin{aligned}
\mathbb{P}[X = \sigma] =& \frac{1}{16}(1 + \sigma_1\sigma_2\theta_1\theta_5\theta_2 + \sigma_1\sigma_3\theta_1\theta_3 + \sigma_1\sigma_4\theta_1\theta_5\theta_4 + \sigma_2\sigma_3\theta_2\theta_5\theta_3 + \sigma_2\sigma_4\theta_2\theta_4 \\
&+ \sigma_3\sigma_4\theta_3\theta_5\theta_4 + \sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_2\theta_3\theta_4)
\end{aligned} \tag{3}
$$

Upon plugging in the probabilities into the likelihood function as outlined in [2], we derive the log likelihood function for case 2:

$$
\begin{aligned}
\log(L) &= \sum_{i=1}^{k} \log P(x_i = \sigma^{(i)}) \\
&= n_{xxxx} \log(1 + \theta_2\tau_1 + \theta_2\theta_3 + \theta_2\theta_5\delta + \theta_3\tau_1 + \theta_1\delta + \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\
&+ n_{xxxy} \log(1 + \theta_2\tau_1 + \theta_2\theta_3 - \theta_2\theta_5\delta + \theta_3\tau_1 - \theta_1\delta - \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\
&+ n_{xxyx} \log(1 + \theta_2\tau_1 - \theta_2\theta_3 + \theta_2\theta_5\delta - \theta_3\tau_1 + \theta_1\delta - \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\
&+ n_{xyxx} \log(1 - \theta_2\tau_1 - \theta_2\theta_3 - \theta_2\theta_5\delta + \theta_3\tau_1 + \theta_1\delta + \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta) \\
&+ n_{xxyy} \log(1 + \theta_2\tau_1 - \theta_2\theta_3 - \theta_2\theta_5\delta - \theta_3\tau_1 - \theta_1\delta + \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\
&+ n_{xyxy} \log(1 - \theta_2\tau_1 - \theta_2\theta_3 + \theta_2\theta_5\delta + \theta_3\tau_1 - \theta_1\delta - \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\
&+ n_{xyyx} \log(1 - \theta_2\tau_1 + \theta_2\theta_3 - \theta_2\theta_5\delta - \theta_3\tau_1 + \theta_1\delta - \theta_3\theta_5\delta + \theta_1\theta_2\theta_3\delta) \\
&+ n_{xyyy} \log(1 - \theta_2\tau_1 + \theta_2\theta_3 + \theta_2\theta_5\delta - \theta_3\tau_1 - \theta_1\delta + \theta_3\theta_5\delta - \theta_1\theta_2\theta_3\delta)
\end{aligned}
$$

To find δ that maximize the likelihood, we can take the first derivative w.r.t. δ and find all of its critical points.

$$
\begin{aligned}
0 = \frac{d\log(L)}{d\delta} &= \frac{n_{xxxx}A}{P + A\delta} - \frac{n_{xxxy}A}{P - A\delta} \\
&+ \frac{n_{xxyx}B}{Q + B\delta} - \frac{n_{xxyy}B}{Q - B\delta} \\
&+ \frac{n_{xyxx}C}{R + C\delta} - \frac{n_{xyxy}C}{R - C\delta} \\
&+ \frac{n_{xyyx}D}{S + D\delta} - \frac{n_{xyyy}D}{S - D\delta}
\end{aligned} \tag{4}
$$

where

$$
\begin{aligned}
&A = \theta_1\theta_5 + \theta_2 + \theta_3\theta_5 + \theta_1\theta_2\theta_3, B = \theta_1\theta_5 + \theta_2 - \theta_3\theta_5 - \theta_1\theta_2\theta_3, \\
&C = \theta_1\theta_5 - \theta_2 + \theta_3\theta_5 - \theta_1\theta_2\theta_3, D = \theta_1\theta_5 - \theta_2 - \theta_3\theta_5 + \theta_1\theta_2\theta_3, \\
&P = 1 + \theta_1\tau_2 + \theta_2\theta_3 + \tau_2\theta_3, Q = 1 + \theta_1\tau_2 - \theta_2\theta_3 - \tau_2\theta_3, \\
&R = 1 - \theta_1\tau_2 + \theta_2\theta_3 - \tau_2\theta_3, S = 1 - \theta_1\tau_2 - \theta_2\theta_3 + \tau_2\theta_3
\end{aligned}
$$

Since the second derivative w.r.t. δ is negative, $\log(L)$ is concave. Thus, global maximum is achieved at the critical point.

3.3 Case 3. Leaf 4 is attached to branch 3

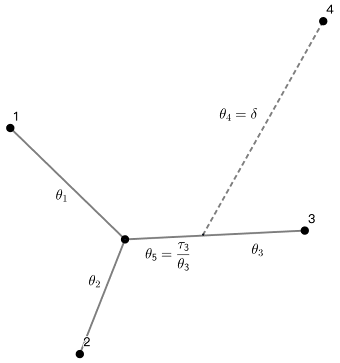


Figure 4: Leaf 4 is attached to branch 3

Using the probability mass function of X from [1], we have the following probability:

$$\begin{aligned}\mathbb{P}[X=\sigma] =& \frac{1}{16}(1+\sigma_1\sigma_2\theta_1\theta_2+\sigma_1\sigma_3\theta_1\theta_5\theta_3+\sigma_1\sigma_4\theta_1\theta_5\theta_4+\sigma_2\sigma_3\theta_2\theta_5\theta_3+\sigma_2\sigma_4\theta_2\theta_5\theta_4 \\ &+\sigma_3\sigma_4\theta_3\theta_4+\sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_2\theta_3\theta_4)\end{aligned} \tag{5}$$

Upon plugging in the probabilities into the likelihood function as outlined in [2], we derive the log likelihood function for case 2:

$$\begin{aligned}\log(L) &= \sum_{i=1}^{k} \log P(X_i=\sigma^{(i)}) \\ &= n_{xxxx}\log(1+\theta_1\tau_3+\theta_1\theta_2+\theta_1\theta_5\delta+\theta_2\tau_3+\theta_3\delta+\theta_2\theta_5\delta+\theta_1\theta_2\theta_3\delta) \\ &+ n_{xxxy}\log(1+\theta_1\tau_3+\theta_1\theta_2-\theta_1\theta_5\delta+\theta_2\tau_3-\theta_3\delta-\theta_2\theta_5\delta-\theta_1\theta_2\theta_3\delta) \\ &+ n_{xxyx}\log(1-\theta_1\tau_3+\theta_1\theta_2+\theta_1\theta_5\delta-\theta_2\tau_3-\theta_3\delta+\theta_2\theta_5\delta-\theta_1\theta_2\theta_3\delta) \\ &+ n_{xyxx}\log(1+\theta_1\tau_3-\theta_1\theta_2+\theta_1\theta_5\delta-\theta_2\tau_3+\theta_3\delta-\theta_2\theta_5\delta-\theta_1\theta_2\theta_3\delta) \\ &+ n_{xxyy}\log(1-\theta_1\tau_3+\theta_1\theta_2-\theta_1\theta_5\delta-\theta_2\tau_3+\theta_3\delta-\theta_2\theta_5\delta+\theta_1\theta_2\theta_3\delta) \\ &+ n_{xyxy}\log(1+\theta_1\tau_3-\theta_1\theta_2-\theta_1\theta_5\delta-\theta_2\tau_3-\theta_3\delta+\theta_2\theta_5\delta+\theta_1\theta_2\theta_3\delta) \\ &+ n_{xyyx}\log(1-\theta_1\tau_3-\theta_1\theta_2+\theta_1\theta_5\delta+\theta_2\tau_3-\theta_3\delta-\theta_2\theta_5\delta+\theta_1\theta_2\theta_3\delta) \\ &+ n_{xyyy}\log(1-\theta_1\tau_3-\theta_1\theta_2-\theta_1\theta_5\delta+\theta_2\tau_3+\theta_3\delta+\theta_2\theta_5\delta-\theta_1\theta_2\theta_3\delta)\end{aligned}$$

To find δ that maximize the likelihood, we can take the first derivative w.r.t. δ and find all of its critical points.

$$\begin{aligned}0=\frac{d\log(L)}{d\delta} &= \frac{n_{xxxx}A}{P+A\delta}-\frac{n_{xxxy}A}{P-A\delta} \\ &+\frac{n_{xxyx}B}{Q+B\delta}-\frac{n_{xxyy}B}{Q-B\delta} \\ &+\frac{n_{xyxx}C}{R+C\delta}-\frac{n_{xyxy}C}{R-C\delta} \\ &+\frac{n_{xyyx}D}{S+D\delta}-\frac{n_{xyyy}D}{S-D\delta}\end{aligned} \tag{6}$$

where

$$\begin{aligned}&A=\theta_1\theta_5+\theta_3+\theta_2\theta_5+\theta_1\theta_2\theta_3, B=\theta_1\theta_5-\theta_3+\theta_2\theta_5-\theta_1\theta_2\theta_3, \\ &C=\theta_1\theta_5+\theta_3-\theta_2\theta_5-\theta_1\theta_2\theta_3, D=\theta_1\theta_5-\theta_3-\theta_2\theta_5+\theta_1\theta_2\theta_3, \\ &P=1+\theta_1\tau_3+\theta_1\theta_2+\tau_3\theta_2, Q=1-\theta_1\tau_3+\theta_1\theta_2-\tau_3\theta_2, \\ &R=1+\theta_1\tau_3-\theta_1\theta_2-\tau_3\theta_2, S=1-\theta_1\tau_3-\theta_1\theta_2+\tau_3\theta_2\end{aligned}$$

Since the second derivative w.r.t. δ is negative, $\log(L)$ is concave. Thus, global maximum is achieved at the critical point.

4 Software Approaches to Optimization

Given a 3-species tree S, the positioning of the fourth branch can result in five distinct scenarios: attach to branch 1, attach to branch 2, attach to branch 3, attach at the intersection of three branches, or attach at an infinitely remote distance from the tree.

We will perform global optimization in Python to determine which scenario wins and what the tree parameters θ and δ are in the maximum likelihood tree. The code can be found at https://github.com/Carinaguo/WISCERS.

4.1 Generate Data as $n_{xxxx}, n_{xxxy}, n_{xxyx}, n_{xyxx}, n_{xxyy}, n_{xyxy}, n_{xyyx}, n_{xyyy}$

Under the model described in Section 2.1, it holds that $\delta = 0$ therefore by Eq.(1), (3), (5)

$$\begin{aligned}\mathbb{P}[X = \sigma] &= \frac{1}{16}(1 + \sigma_1\sigma_2\theta_1\theta_5\theta_2 + \sigma_1\sigma_3\theta_1\theta_3 + \sigma_2\sigma_3\theta_2\theta_5\theta_3) \\ &= \frac{1}{16}(1 + \sigma_1\sigma_2\theta_1\theta_2 + \sigma_1\sigma_3\theta_1\theta_5\theta_3 + \sigma_2\sigma_3\theta_2\theta_5\theta_3) \\ &= \frac{1}{16}(1 + \sigma_1\sigma_2\theta_1\theta_5\theta_2 + \sigma_1\sigma_3\theta_1\theta_5\theta_3 + \sigma_2\sigma_3\theta_2\theta_3)\end{aligned}$$

We assume $\theta_5 = 1, \theta_2 = \theta_3$. As θ_1 goes from 0 to 1, we generate samples from the multinomial distribution using numpy.random.multinomial.

4.2 Global Optimization

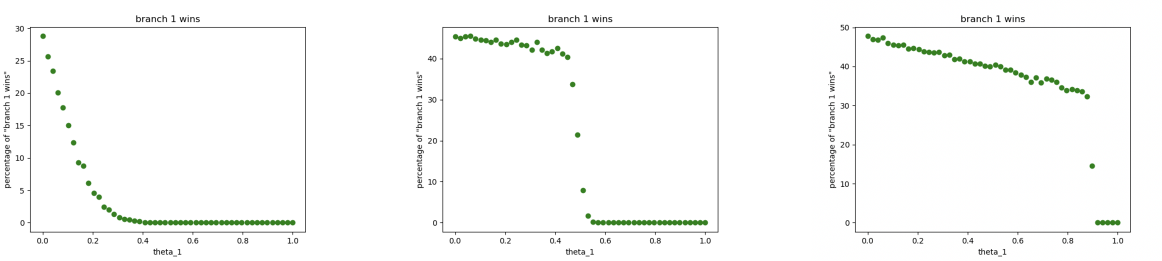
As mentioned previously in Section 3, we can find δ that achieves the maximum likelihood by identifying critical points in Eq. (2), (4), (6). As the parameter θ_i, where i denotes the branch to which leaf 4 attaches, varies across the interval $[\tau, 1]$, we apply Newton's Method to solve Eq. (2), (4), and (6) to determine the solution of δ. In this manner, for each θ_i in $[\tau, 1]$, a corresponding δ is obtained which achieves the maximum likelihood.

Then from the set of all possible pairs of θ_i and δ, we find the specific one that maximize the likelihood. Therefore, we can determine which branch the fourth branch attaches to. However, in scenarios where the parameter $\delta = 0$ or $\theta_i = \tau$, no branch wins. These cases respectively correspond to situations where "the fourth branch is infinitely distant from the tree" and "the fourth branch attaches at the intersection of the three branches." Note that the scenario $\delta = 0$ aligns with the true model.

4.3 Discussion

As θ_1 goes from 0 to 1, we plot how the fractions of each scenario winning change.

- When $\theta_1 = \theta_2 = \theta_3$, all branches win roughly the same number of times. In Fig. 5, the percentage of "branch 1 wins" demonstrates a sharp decline at $\theta_1 = \theta_2$. The fourth branch tends to attach to the longer branch. Note that the decline in Fig. 5(a) is milder than the other two.



(a) $\theta_2 = \theta_3 = 0.1$

(b) $\theta_2 = \theta_3 = 0.5$

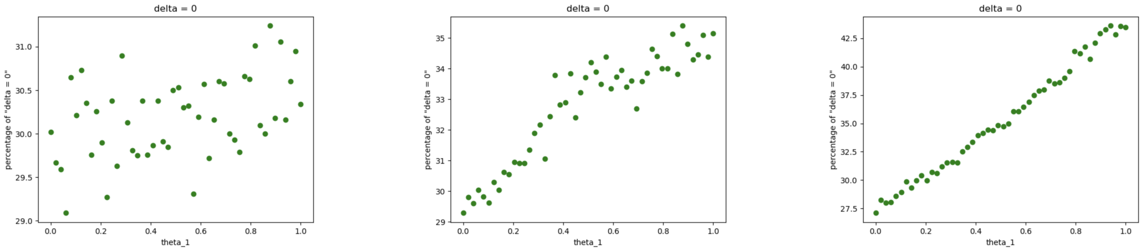
(c) $\theta_2 = \theta_3 = 0.9$

Figure 5: Plots of the percentage of scenario "the fourth branch attaches to branch 1" winning

- In Fig. 6, as θ_2 and θ_3 increase, the percentage of "$\delta = 0$" exhibits a more and more clear upward trend. Specifically, Fig. 6(c) shows linearity.

- In Fig. 7, the percentage of "the fourth branch attaches at the intersection of three branches" exhibits no distinct pattern of change. The data range all over the place from 24 to 26.

- When $\theta_1 = 1$, the fourth branch is attached at the end of branch 1. Fig. 8 shows how the fractions of "$\theta_1 = 1$" changes. As the percentage of "branch 1 wins" decreases shown in Fig 5, the fractions of "$\theta_1 = 1$" increases. Until branch 1 is never selected as the attachment point for the fourth branch (i.e., the percentage of "branch 1 wins" $= 0$), the fraction also becomes 0.

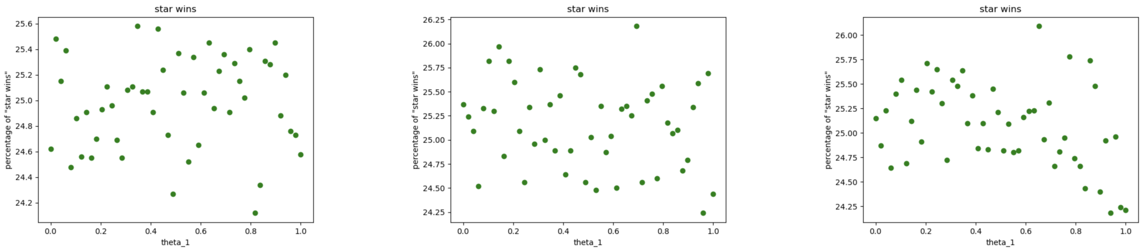


(a) $\theta_2 = \theta_3 = 0.1$

(b) $\theta_2 = \theta_3 = 0.5$

(c) $\theta_2 = \theta_3 = 0.9$

Figure 6: Plots of the percentage of scenario "$\delta = 0$" winning

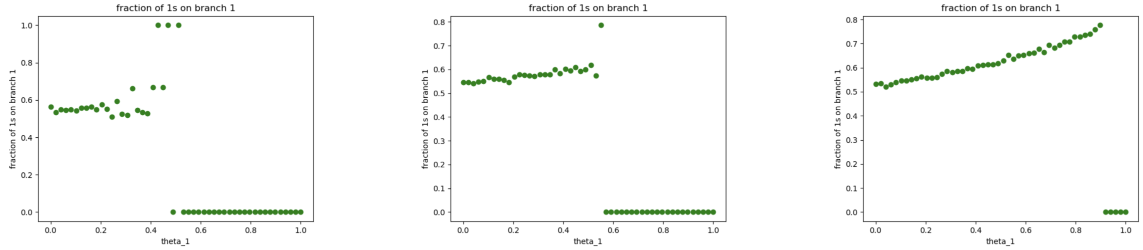


(a) $\theta_2 = \theta_3 = 0.1$

(b) $\theta_2 = \theta_3 = 0.5$

(c) $\theta_2 = \theta_3 = 0.9$

Figure 7: Plots of the percentage of scenario "the fourth branch attaches at the intersection of three branches" winning



(a) $\theta_2 = \theta_3 = 0.1$

(b) $\theta_2 = \theta_3 = 0.5$

(c) $\theta_2 = \theta_3 = 0.9$

Figure 8: Plots of the fractions of "$\theta_1 = 1$"

References

[1] A. Balaji, Y. Cai, R. Huang, M. Lolling, M. Sun, M. Bacharach, and S. Roch. Unexplained behavior of phylogeny estimation methods. 2022.

[2] A. Balaji, B. Chen, N. Jongsawatsataporn, J. Wan, M. Bacharach, and S. Roch. Unexplained behavior of phylogeny estimation methods. 2022.