Calculating the probability of a ranked gene tree given a species tree under the multispecies coalescent model

Anastasiia Kim

University of New Mexico

ASA Albuquerque chapter meeting
April 13, 2018

Terminology for trees

- A species tree represents the evolutionary relationships among various species.
- Gene trees which are contained within the branches of the species phylogeny represent the evolutionary history of the sampled genes.

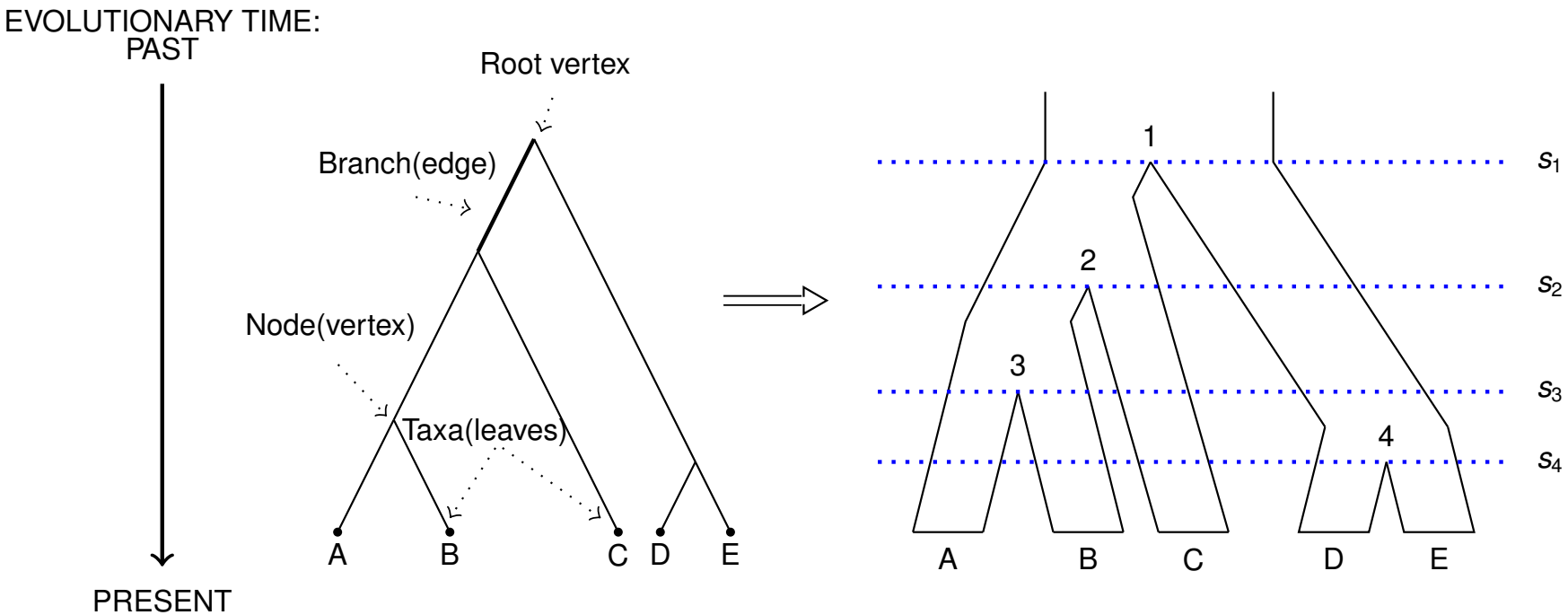


Figure 1: A rooted binary five taxon phylogeny where s_i is the time of the interior vertex of rank i.

Evolutionary tree

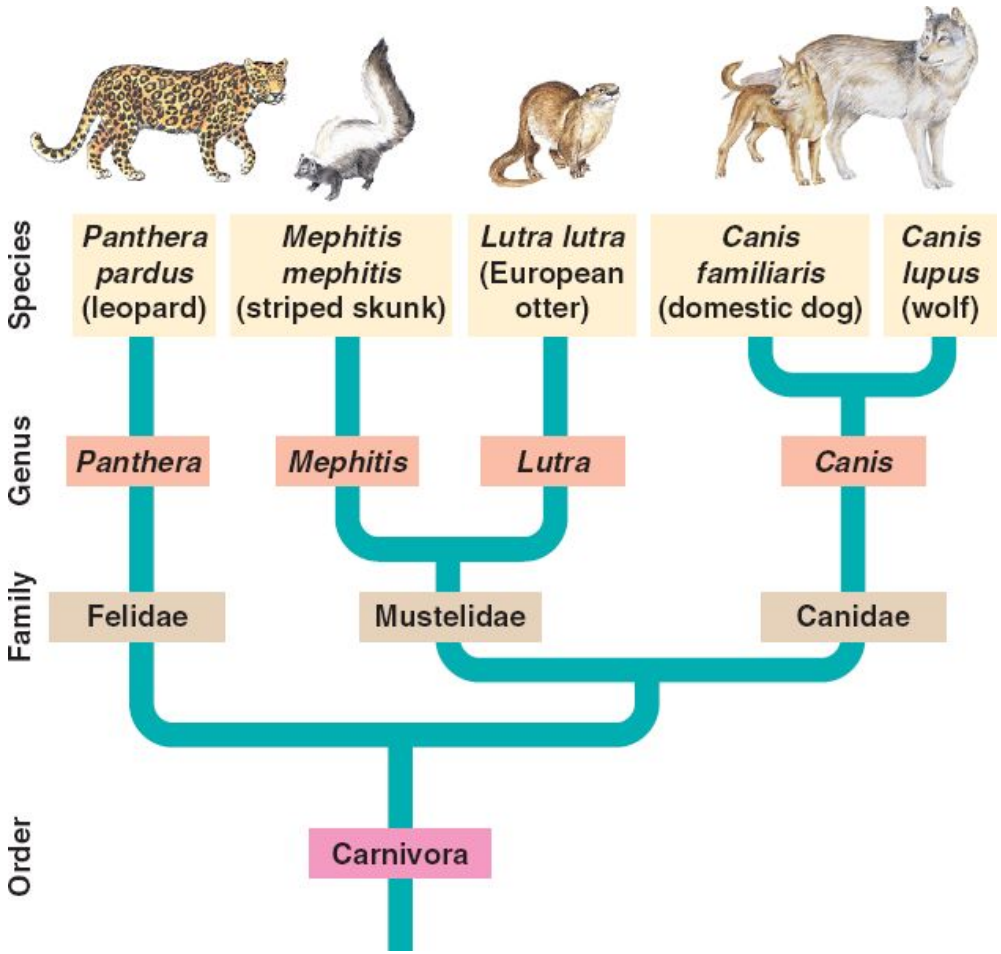
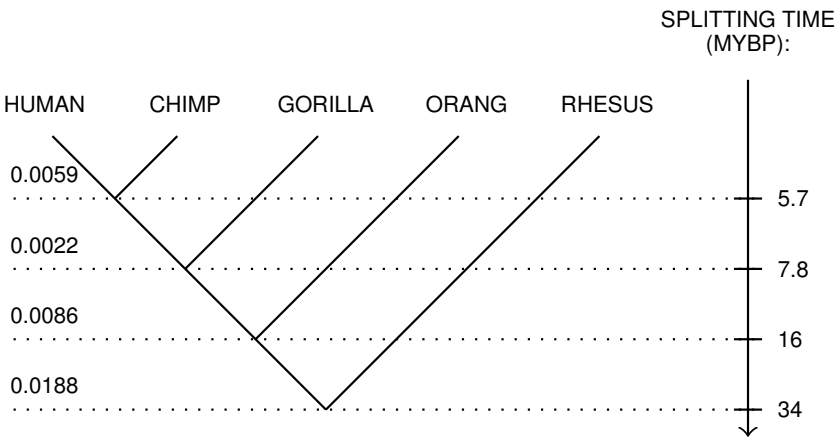


Figure 2: Phylogenetic tree showing inferred evolutionary relationships among species. Copyright © 2005 Pearson Education, Inc. publishing as Benjamin Cummings.

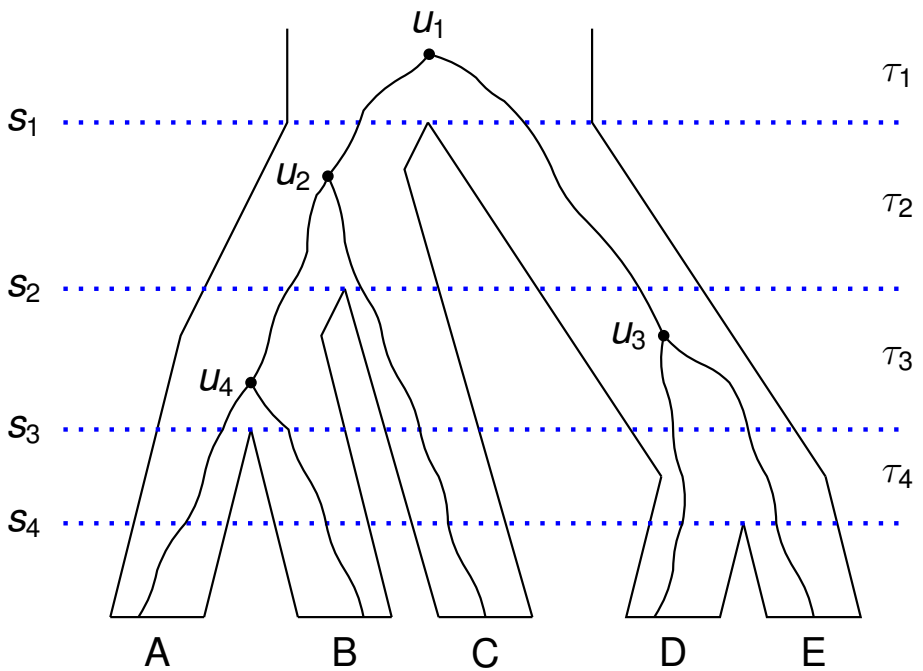


Topology	All (%)	Gene[a] (%)	Exon[b] (%)
H C G O R	20 (0.17)	8 (0.17)	2 (0.32)
H C G O R	9,148 (76.58)	3,814 (78.85)	487 (78.93)
H C O G R	19 (0.16)	10 (0.21)	2 (0.32)
G O C H R	0	0	0
G O H C R	1 (0.01)	0	0
H O C G R	5 (0.04)	2 (0.04)	0
H O C G R	0	0	0
H O G C R	0	0	0
C G O H R	4 (0.03)	1 (0.02)	0
C G H O R	1,369 (11.46)	504 (10.42)	63 (10.21)
H G C O R	13 (0.11)	6 (0.12)	1 (0.16)
H G O C R	5 (0.04)	0	0
H G C O R	1,361 (11.39)	492 (10.17)	62 (10.05)

Figure 3: Number of DNA sequence alignments in support of the sequence tree topologies and reconstructed sequence tree of the 5 species. Copyright © 2007 Mapping Human Genetic Ancestry by I. Ebersberger et al.

Enumeration of ranked histories

- To *coalesce* means to *join*, *merge*, or *grow together*.
- Let's define a ranked history of the gene tree as $x = (x_1, x_2, ..., x_{n-1})$, where for $i = 1, 2, ..., n-1$, $x_i = j$ if the *ith* coalescence occurs in species tree interval τ_j.



In this case nine ranked histories exist
(1, 1, 1, 1),
(1, 1, 1, 2),
.....
(1, 2, 2, 3),
(1, 2, 3, 3).

Figure 4: Gene tree with ranked history (1, 2, 3, 3) evolving on the species tree.

Probability of a ranked gene tree on a given species tree

The probability of a ranked gene tree topology $\mathcal{G}$ with ranked history set $\mathcal{Y}$ given a species tree $\mathcal{T}$ is

$$P(\mathcal{G}|\mathcal{T}) = \sum_{x \in \mathcal{Y}} H_\ell(x) \prod_{i=2}^{n-1} P(G_{\tau_i}, x | T),$$

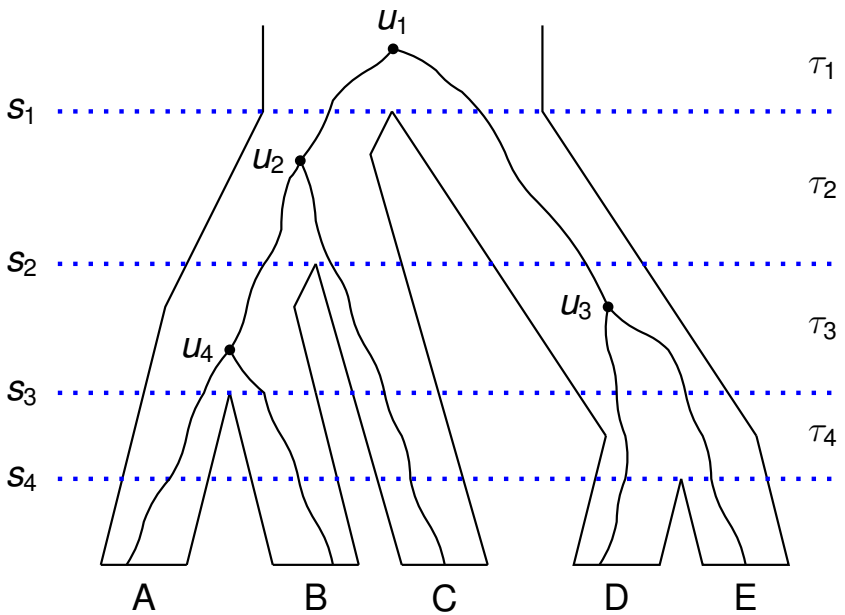
where $H_\ell(x)$ is the probability for the coalescence above the root appearing in the right order. If the number of lineages above the root is ℓ, then

$$H_\ell(x) = \frac{2^{\ell-1}}{\ell!(\ell-1)!}.$$

Since the probability that i lineages fail to coalesce in a time interval of length t_i is $e^{-\binom{i}{2}t_i}$, we can easily calculate the probability in each interval.

$$P(\mathcal{G}, (1,2,3,3)|\mathcal{T}) = \left(1 - e^{-t_2}\right)\frac{1}{2}\left(1 - e^{-t_3}\right)^2 e^{-t_4}$$

.



$$i = 2 : 1 - e^{-\binom{2}{2}t_2},$$
$$i = 3 : \left(1 - e^{-\binom{2}{2}t_3}\right)\left(1 - e^{-\binom{2}{2}t_3}\right)\frac{1}{2},$$
$$i = 4 : e^{-\binom{2}{2}t_4}.$$

Figure 5: Gene tree with ranked history (1, 2, 3, 3) evolving on the species tree.

- Denote the number of lineages available for coalescence in population z just after the jth coalescence in interval τ_i by k_{ijz}.
- The waiting time until the next coalescent event (going backwards in time) has rate $\lambda_{i,j} = \sum_{z=1}^{i} \binom{k_{ijz}}{2}$.
- The density for the coalescent events in the interval τ_i is

$$f_i(v_0, v_1, ..., v_{m_i}) = e^{-\sum_{j=0}^{m_i} \lambda_{i,j} v_j},$$

where v_j is the time between the jth and (j+1)st coalescent events with v_0 being the time between s_{i-1} and the least recent coalescent event in τ_i and with v_{m_i} being the time between s_i and coalescent event m_i.

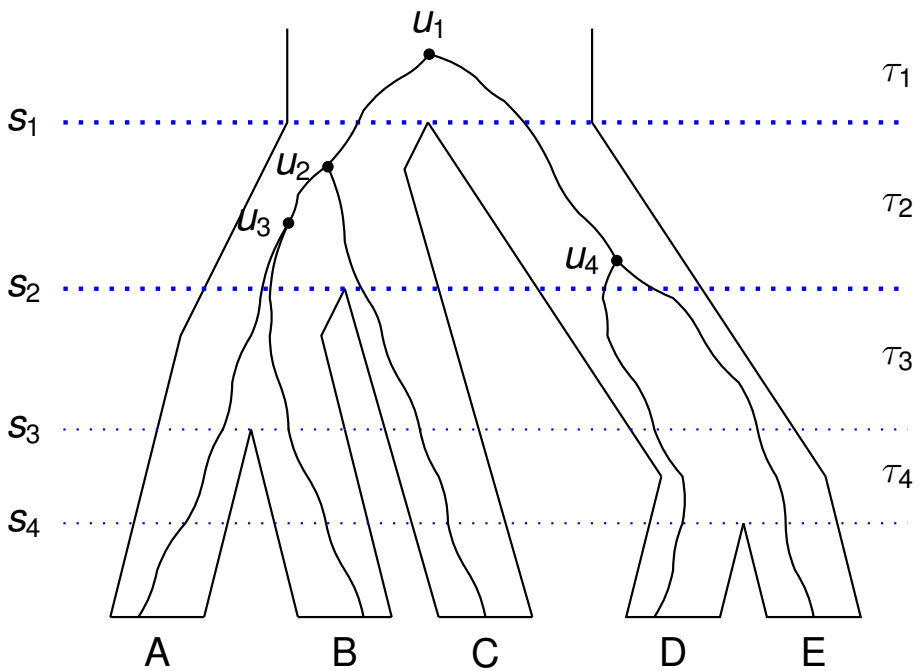
For $\lambda_{i,0} > 0$ we can rewrite f_i as

$$f_i(v_0, v_1, ..., v_{m_i}) = \frac{\prod_{j=0}^{m_i} \lambda_{i,j} e^{-\lambda_{i,j} v_j}}{\prod_{j=0}^{m_i} \lambda_{i,j}},$$

Using the fact that the integral of the numerator is the convolution of $m_i + 1$ exponential random variables with rates $\lambda_{i,0}, ..., \lambda_{i,m_i}$, which is hypoexponential distribution, the probability of the coalescent events in interval τ_i can be written as

$$P(\mathcal{G}_{\tau_i}|\mathcal{T}) = \int_v f_i(v_0, ..., v_{m_i}) dv = \sum_{j=0}^{m_i} \frac{e^{-\lambda i,j(s_{i-1} - s_i)}}{\prod_{k=0, k\neq j}^{m_i} (\lambda_{i,k} - \lambda_{i,j})}.$$

Calculation of k_{ijz}



k_{ijz} is the number of lineages available to coalesce in interval τ_i in population z at the jth coalescence in population.

The values of k_{ijz} in interval τ_2 are

$k_{2,0,1} = 1$, $k_{2,0,2} = 1$,
$k_{2,1,1} = 2$, $k_{2,1,2} = 1$,
$k_{2,2,1} = 3$, $k_{2,2,2} = 1$,
$k_{2,3,1} = 3$, $k_{2,3,2} = 2$.

Figure 6: Matching gene tree with ranked history (1, 2, 2, 2) evolving on the species tree.

Let's calculate the probability of the coalescent events in interval τ_2 in which $m_2 = 3$ coalescences occur.

$$P(\mathcal{G}_{\tau_2}|\mathcal{T}) = \sum_{j=0}^{3} \frac{e^{-\sum_{z=1}^{2}\binom{k_{2jz}}{2}(s_1 - s_2)}}{\prod_{k=0,k\neq j}^{3}\left(\sum_{z=1}^{2}\binom{k_{2kz}}{2} - \sum_{z=1}^{2}\binom{k_{2jz}}{2}\right)},$$

$$P(\mathcal{G}_{\tau_2}|\mathcal{T}) = \frac{1}{12} - \frac{1}{6}e^{-t_2} + \frac{1}{6}e^{-3t_2} - \frac{1}{12}e^{-4t_2},$$

where $t_i = s_{i-1} - s_i$.

Similarly, we can calculate the probabilities for the other two intervals

$$P(\mathcal{G}_{\tau_3}|\mathcal{T}) = e^{-(s_2-s_3)}e^{-(s_2-s_3)},$$

$$P(\mathcal{G}_{\tau_4}|\mathcal{T}) = e^{-(s_3-s_4)}.$$

Since only two lineages available above the root, $H_2(1,2,2,2) = 1$ and the probability of ranked history (1,2,2,2) can be calculated as

$$P(\mathcal{G},(1,2,2,2)|\mathcal{T}) = \prod_{i=2}^{4} P(G_{\tau_i},(1,2,2,2)|T) =$$

$$\left(\frac{1}{12} - \frac{1}{6}e^{-t_2} + \frac{1}{6}e^{-3t_2} - \frac{1}{12}e^{-4t_2}\right)e^{-t_4-2t_3}$$

Anomalous gene trees

Under the multispecies coalescent, a species tree can produce *anomalous gene trees* - gene tree topologies that do not match the species tree topology, and whose probabilities exceed that of the gene tree topology that does match the species tree. Such species tree is said to be in the *anomaly zone*.

$$P(\mathcal{G}_{left}, (1,2,2,2)|\mathcal{T}) = \left(\frac{1}{12} - \frac{1}{6}e^{-t_2} + \frac{1}{6}e^{-3t_2} - \frac{1}{12}e^{-4t_2}\right)e^{-t_4-2t_3},$$

$$P(\mathcal{G}_{right}, (1,2,2,2)|\mathcal{T}) = \left(\frac{1}{8} - \frac{1}{3}e^{-t_2} + \frac{1}{4}e^{-2t_2} - \frac{1}{24}e^{-4t_2}\right)e^{-t_4-2t_3}$$

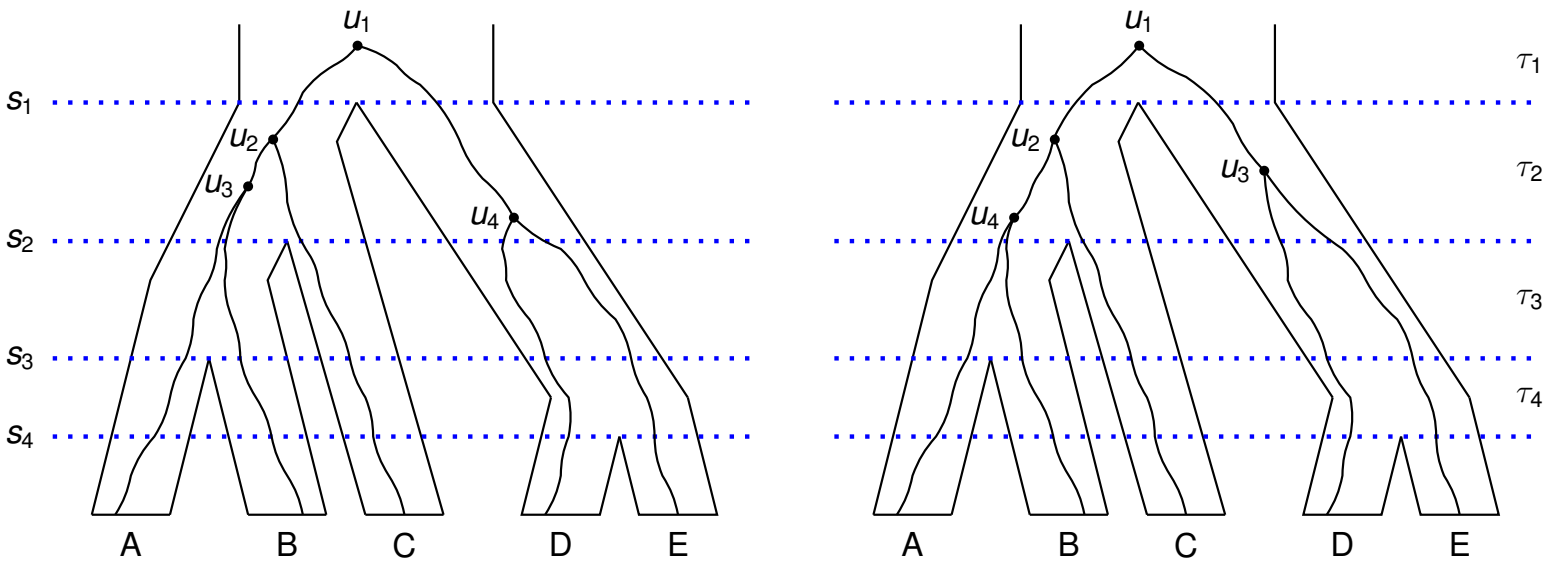
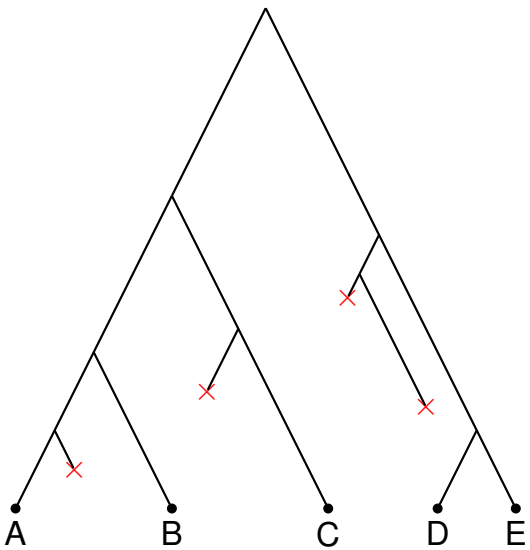


Figure 7: Matching and non-matching gene trees with ranked history (1, 2, 2, 2) evolving on the species tree.

Constant rate birth-death model



- We use a birth-death model with birth rate λ and death rate μ to simulate species trees.
- The expected waiting time to the next event (speciation or extinction) is $Exp(\lambda + \mu)$.
- One species splits over time to form two new species with probability of $\frac{\lambda}{\lambda+\mu}$ or dies with probability of $\frac{\mu}{\lambda+\mu}$.

Figure 8: A birth-death tree including all extinct and extant species.

Simulations

- Program for computing ranked gene tree probabilities under the coalescent process was developed.

INPUT:		OUTPUT:
Species tree and Collection of gene trees	⟹	Gene tree topologies along with their probabilities ACD-2-BE-3-CD-4 0.0167033592992527 CE-2-ABD-3-AB-4 0.0042975151468479 ABCE-2-BCE-3-BC-4 0.0006859326069307 ...

- We simulated 5000 species trees under a birth-death process to compute whole probability distribution for 5, 6, 7, and 8-taxon species trees.

Number of taxa:		Number of unique ranked gene tree topologies:
5		180
6	⟹	2,700
7		56,700
8		1,587,600

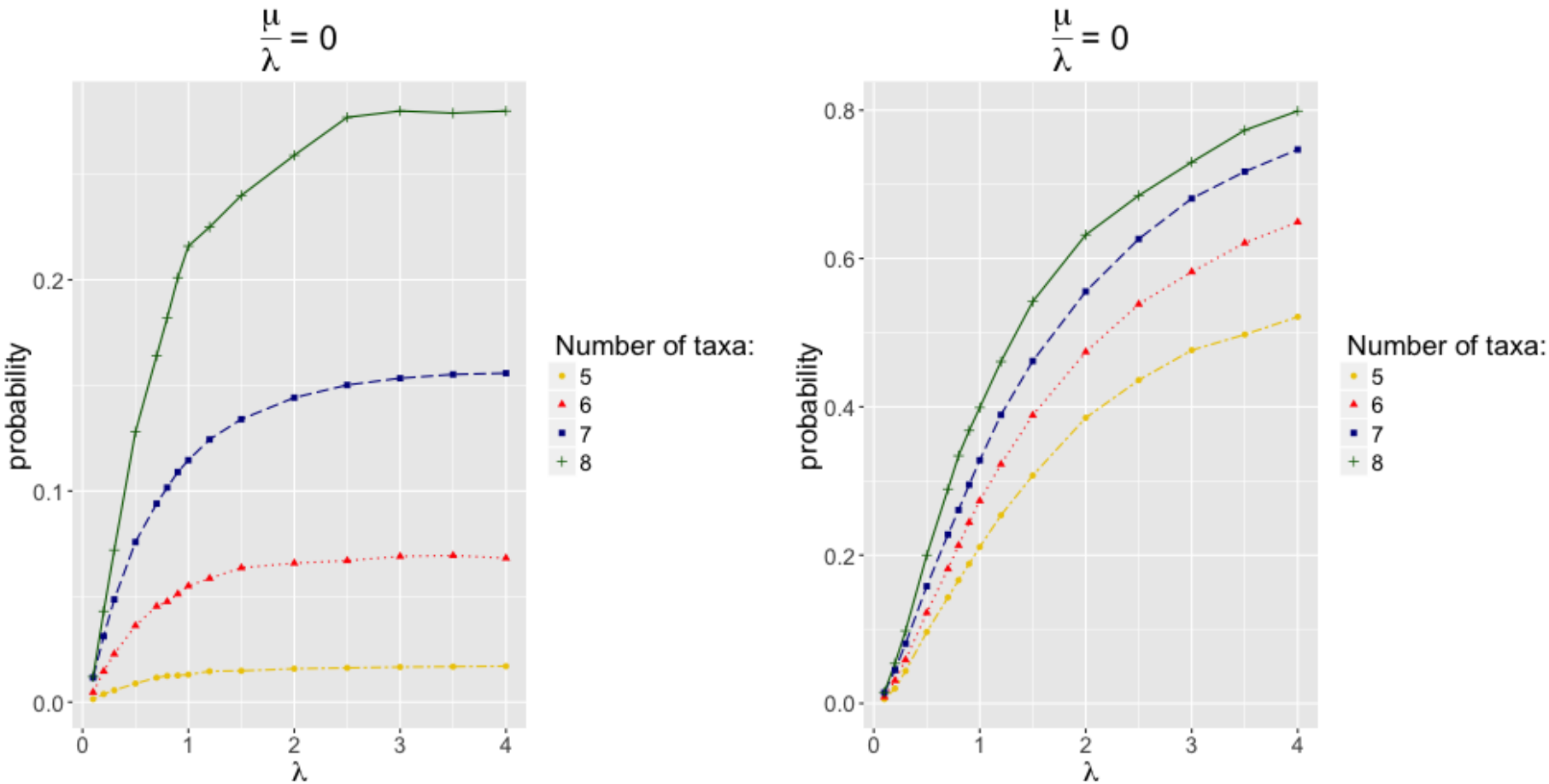


Figure 9: The impact of species tree parameters on the existence of ranked and unranked anomaly zones.

Unranked and Ranked anomaly zones for 5-taxon trees

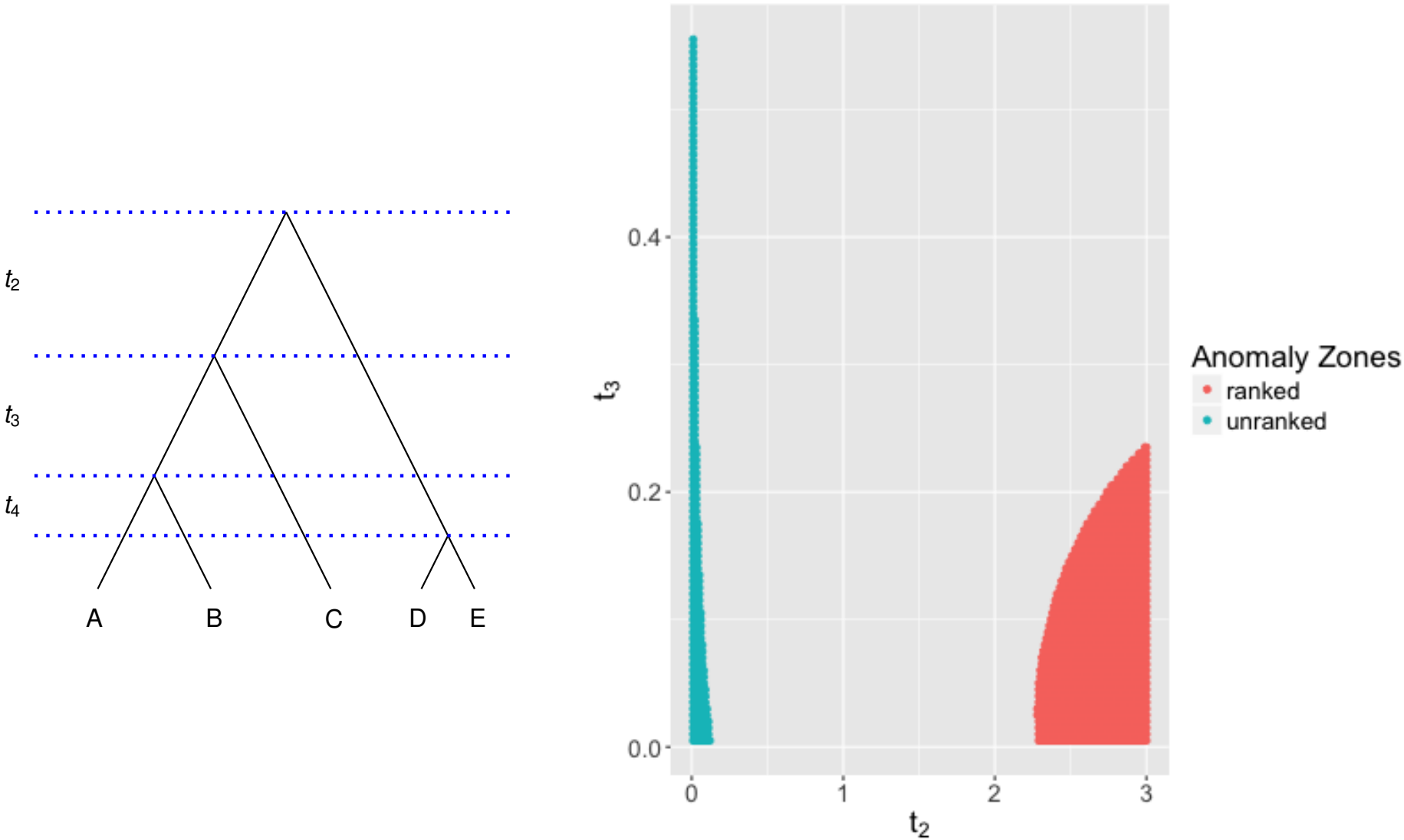


Figure 10: Unranked and ranked anomaly zones for the given five-taxon species tree topology when $t_4 = 0.1$.

Unranked and Ranked anomaly zones for 6-taxon trees

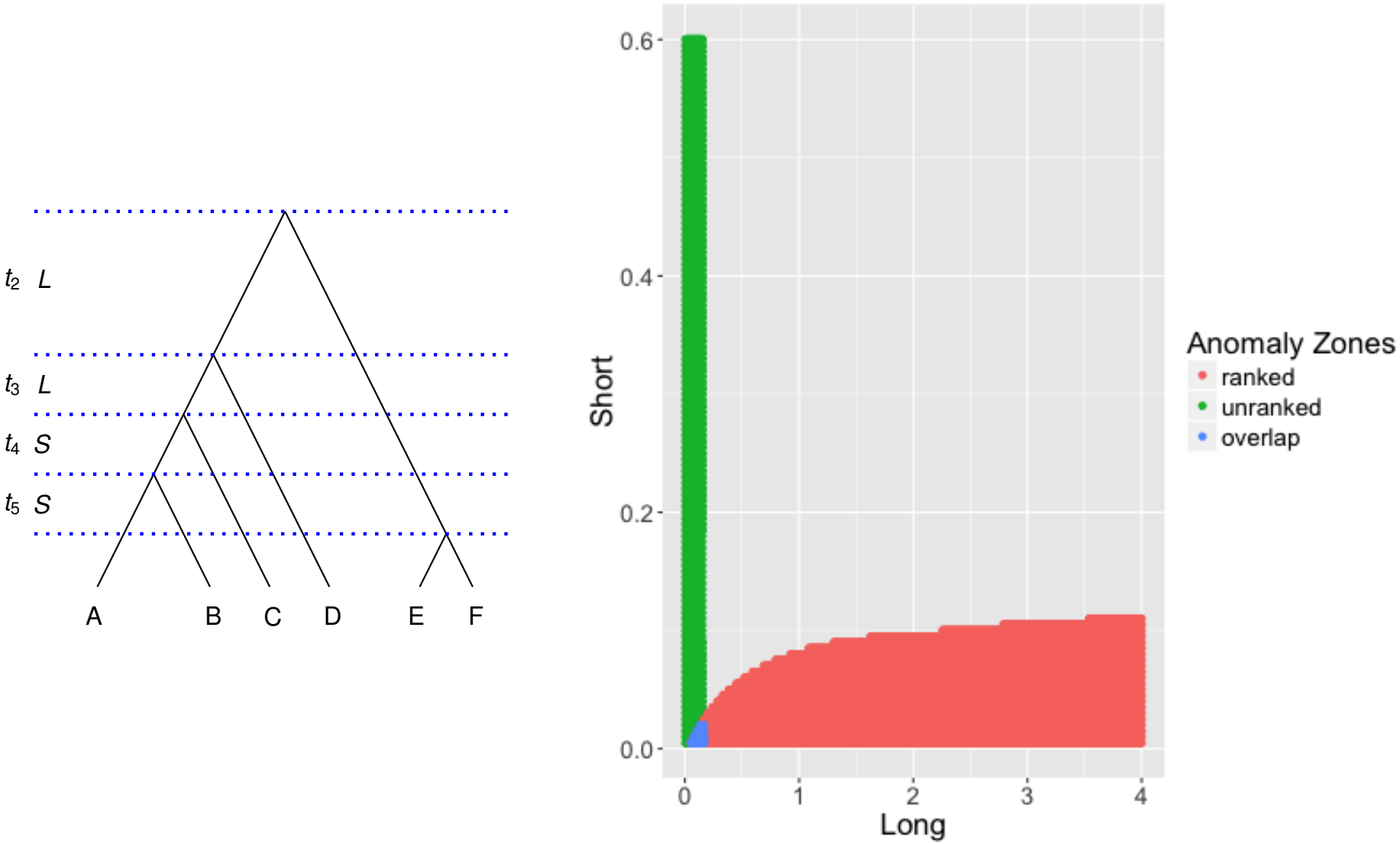


Figure 11: Cross-section of anomaly zones for the given six-taxon species tree topology.

Future work

- Further explore the existence of an overlap in ranked, unranked and unrooted anomaly zones.
- Study if the most common ranked tree has a different unranked or unrooted topology from that of the species tree.
- Infer a species tree from a collection of ranked gene trees using MLE technique.