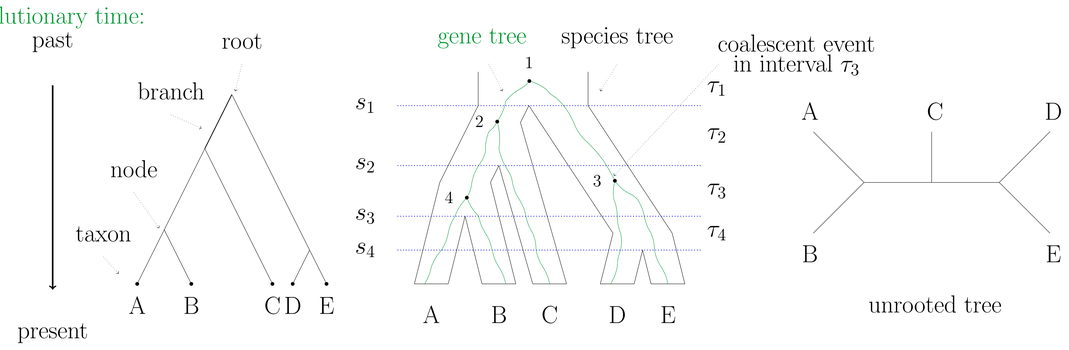
Using ranked gene tree distributions for detecting anomaly zone in a species tree and maximum likelihood inference

Anastasiia Kim, James Degnan

Department of Mathematics and Statistics, University of New Mexico, Albuquerque

Understanding phylogenies

- A species tree represents the evolutionary relationships among various species.
- Gene trees represent the genealogical relationships among the gene sequences sampled from the species.



A rooted 5 taxon phylogeny where s_i is the time of the interior node of rank i

Ranked vs Unranked gene trees

- Unranked trees depict the topological relationships among gene lineages.
- Ranked trees also depict the sequence in which the lineages coalesce (join).

Calculating the probability of a ranked gene tree topology $\mathcal{G}$ given a species tree $\mathcal{T}$

$$P(\mathcal{G}|\mathcal{T}) = \underbrace{\sum_{x\in\mathcal{Y}} H_{\ell_1}(x) \underbrace{\prod_{i=2}^{n-1} P(G_{\tau_i}, x|T)}_{\text{product over speciation intervals } \tau_i}}_{\text{sum over all ranked histories}} = \sum_{j=0}^{m_i} \frac{e^{-\lambda i,j(s_{i-1}-s_i)}}{\prod_{k=0,k\neq j}^{m_i} (\lambda_{i,k} - \lambda_{i,j})}$$

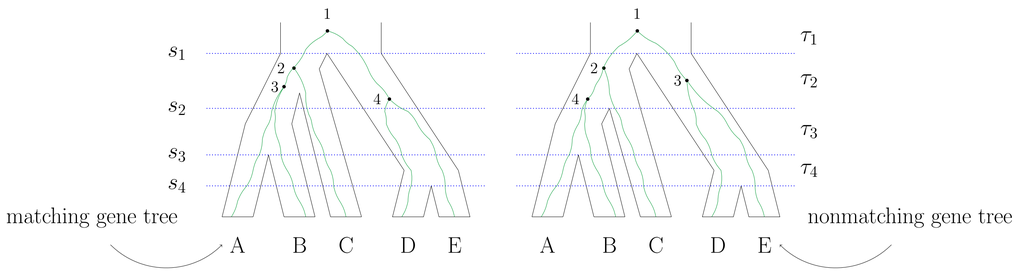
- $P(\mathcal{G}_{\tau_i}, x|\mathcal{T})$ is the probability in interval τ_i for ranked history x.

New software

- *PRANC* is a software written in *C++* that computes the *P*robabilities of *RAN*ked and unranked gene tree topologies under the *C*oalescent process (github.com/anastasiiakim/PRANC).
- *PRANC* has an option to compute maximum likelihood estimates for species trees given a sample of gene trees under the coalescent model (anastasiiakim@unm.edu).

Anomalous gene trees

- Gene trees that have different ranked topologies but share the same unranked topology.
- Both gene trees have the ranked history of $(1, 2, 2, 2)$.

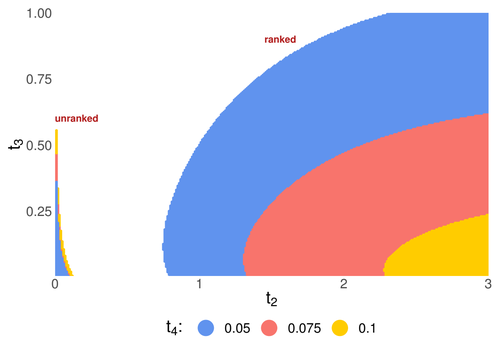


- The incorrect gene tree topology (one that does not match the species tree) that is more probable than the correct one is termed **anomalous gene tree** [1].
- Species trees that can generate anomalous gene trees are said to be in the **anomaly zone**.
- The method of choosing the most common gene tree as the estimate of the species tree in the anomaly zone will be statistically inconsistent.

Anomaly zones

How do we determine if the species tree is in the anomaly zone?

- Compute an entire distribution of gene trees and check each one to see if it is more probable than the matching gene tree.



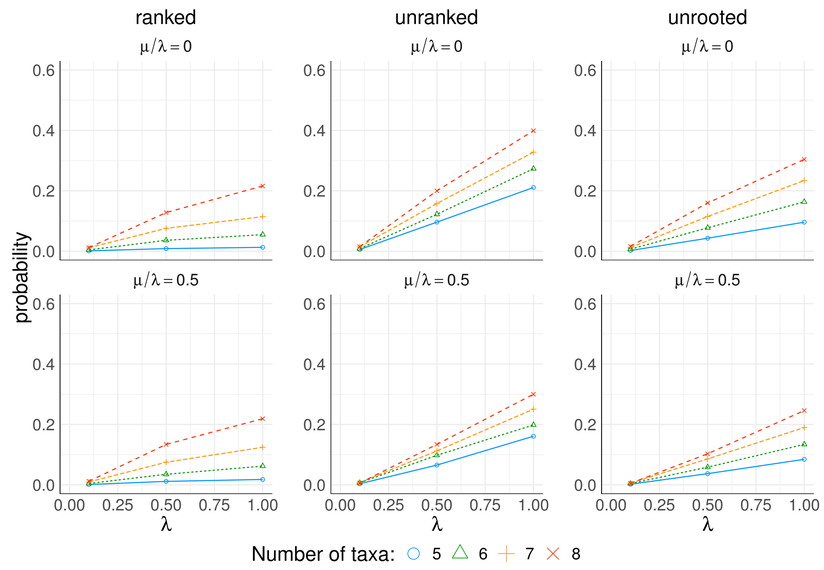
References

[1] J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. PLoS Genetics, 2006.

[2] J. H. Degnan, N. A. Rosenberg, and T. Stadler. The probability distribution of ranked gene trees on a species tree. Mathematical Biosciences, 2012.

Anomaly zones

How the speciation λ and extinction μ rates of a species tree simulated under a constant rate birth-death process can affect the probability that the species tree lies in the anomaly zone?

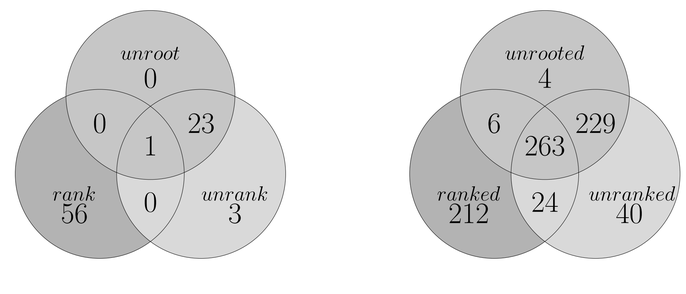


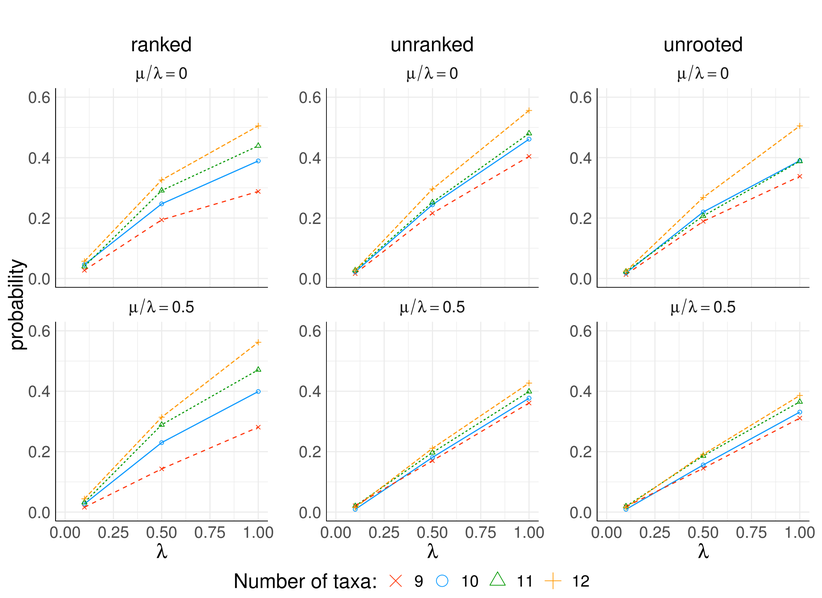
Heuristics for larger trees

- Consider unranked and unrooted gene tree topologies within one nearest neighbour interchange from the species tree topology to infer the existence of anomalous trees in larger trees.
- Use only those ranked gene trees which topologies match the unranked species tree topology to make an inference with larger trees.

Can a species tree simultaneously be in different types of anomaly zones?

$n = 12$, $\lambda = 0.1$ $\quad$ $n = 12$, $\lambda = 1$





Maximum likelihood

- The probability of ranked gene trees can be used to determine the ML species tree.
- The maximum likelihood species tree $\mathcal{T}_{ML}$ given the observed $\mathcal{N}$ ranked gene trees is

$$\mathcal{T}_{ML} = \underset{\mathcal{T}}{\operatorname{argmax}}\, P[\mathcal{G}_1, ..., \mathcal{G}_N|\mathcal{T}] = \underset{\mathcal{T}}{\operatorname{argmax}} \prod_{i=1}^{N} P[\mathcal{G}_i|\mathcal{T}]$$

- Measure the accuracy of the methods by looking at the proportion of false or missing splits in the inferred tree compared to the true tree.