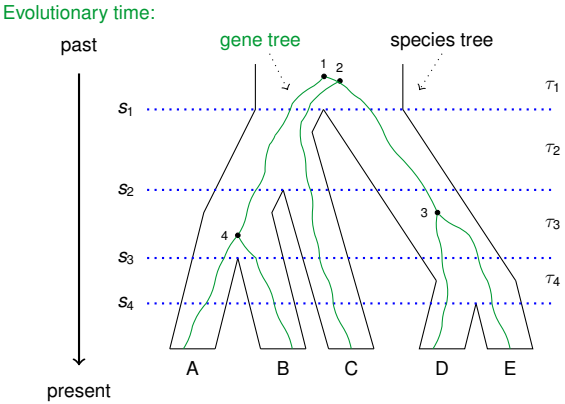
Heuristic approaches for ranked, unranked, and unrooted anomaly zones

Anastasiia Kim, James Degnan

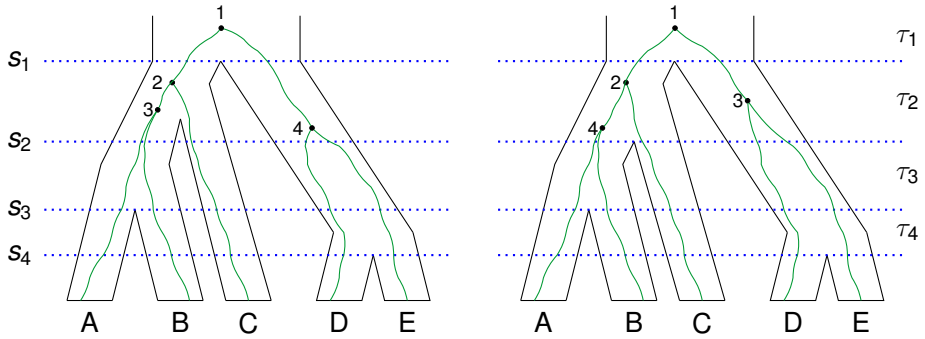
Department of Mathematics and Statistics, University of New Mexico
Funded by NIH R01 GM117590

Evolution 2019

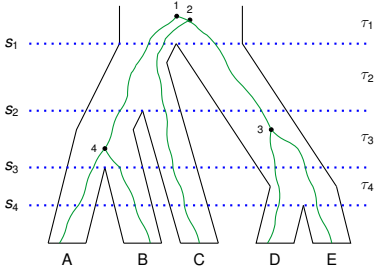
- The coalescent is the genealogical process of joining lineages when one traces the genealogy of the sample backwards in time.
- Several processes can lead to discordance between species and gene trees.



- Unranked trees depict the topological relationships among gene lineages.
- Ranked trees also depict the sequence in which the lineages coalesce.
- Gene trees have different ranked topologies but share the same unranked topology $((AB)C)(DE)$.



- 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.
- Let $P(\mathcal{G}_{\tau_i}, x|\mathcal{T})$ be the probability in interval τ_i for ranked history x.
- The probability of a ranked gene tree topology $\mathcal{G}$ with ranked history set $\mathcal{Y}$ given a species tree $\mathcal{T}$ is



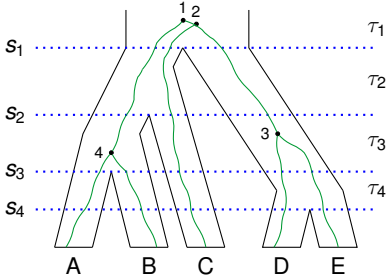
ranked history (1, 1, 3, 3)

$$\underbrace{\sum_{x \in \mathcal{Y}} H(x) \quad \underbrace{\prod_{i=2}^{n-1} P(G_{\tau_i}, x|\mathcal{T})}_{\text{product over speciation intervals } \tau_i}}_{\text{sum over all ranked histories}}$$

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.

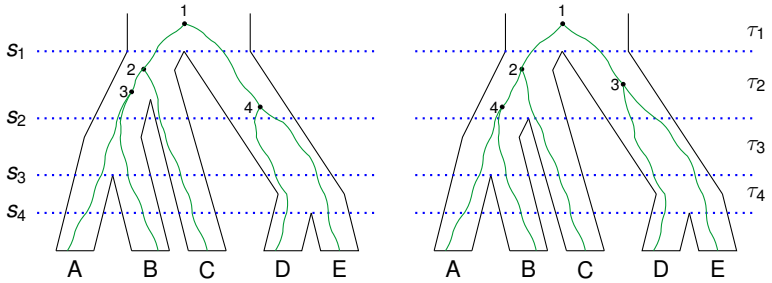
$$P(\mathcal{G}, (1,1,3,3)|\mathcal{T}) = \frac{1}{6} e^{-t_2} \left(1 - e^{-t_3}\right)^2 e^{-t_4},$$

.

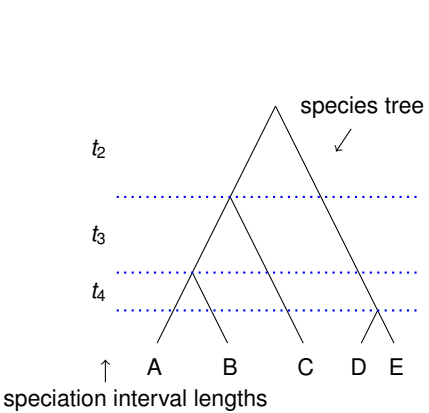


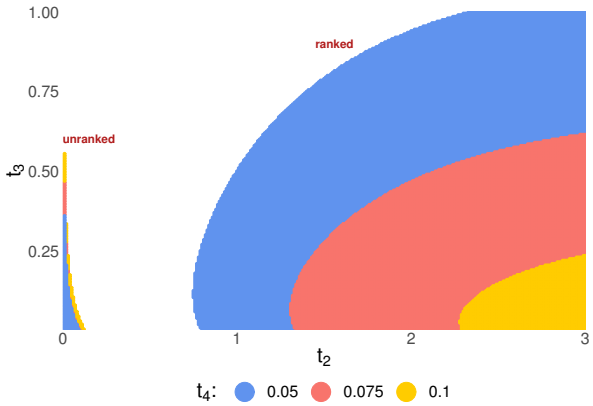
$$i = 1 : \frac{1}{3},$$
$$i = 2 : e^{-\binom{2}{2}t_2},$$
$$i = 3 : \left(1 - e^{-\binom{2}{2}t_3}\right)\left(1 - e^{-\binom{2}{2}t_3}\right)\frac{1}{2},$$
$$i = 4 : e^{-\binom{2}{2}t_4}.$$

- The unranked (ranked) gene tree topology that is more probable than the unranked (ranked) topology matching the species tree is called *anomalous unranked (ranked) gene tree*.
- Species trees that can generate anomalous gene trees are said to be in the *anomaly zone*.

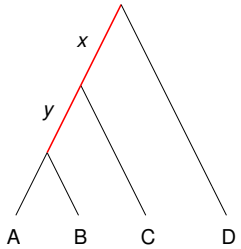


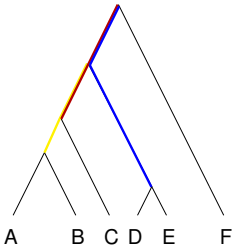
- We compute an entire distribution of gene trees and check if there is a nonmatching gene tree topology that is more probable than the matching tree.



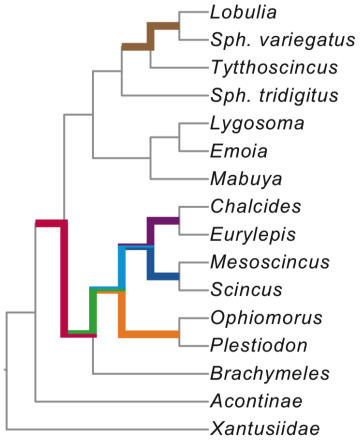


- Computing an entire distribution of gene trees for detecting anomalous trees is not practical for larger trees.
- The 4-taxon species tree is in unranked anomaly zone if $y < a(x)$, where $a(x) = \log\left[\frac{2}{3} + \frac{3e^{2x}-2}{18(e^{3x}-e^{2x})}\right]$.
- Pairs of any two internal consecutive branches in larger tree can be checked for anomaly zone condition $y < a(x)$ (Linkem et al. 2016).





- Anomaly zone calculations were done for each pair of internodes in the extended MRC.



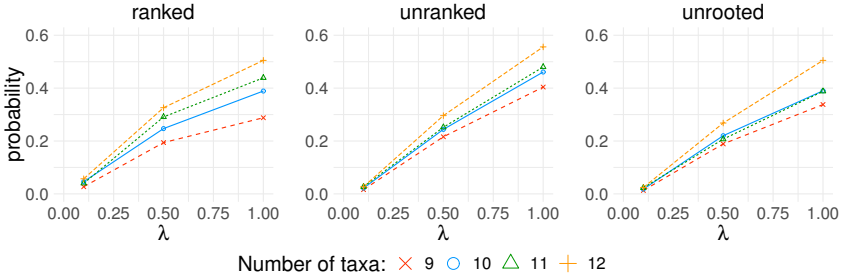
- The majority of relationships in Scincinae have internode lengths that are expected to produce AGTs.
- Strong conflict between species trees and concatenated gene trees.
- Parts of the tree in conflict correspond with areas of the tree that are also estimated to be in the anomaly zone.

- All pairs of internode branch lengths were used to check if at least one pair satisfying the anomaly zone limit condition $y < a(x)$.

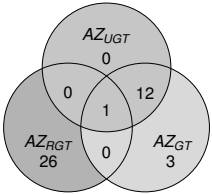
		$y < a(x)$			one step NNI		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
True positive	5	96.55	94.19	92.80	100.00	100.00	99.91
	6	97.73	95.26	94.51	100.00	99.67	99.27
	7	93.15	95.95	95.24	100.00	99.75	99.76
	8	92.94	95.61	95.45	100.00	99.89	99.84
False positive	5	0.00	2.16	7.06	0.00	0.00	0.00
	6	0.18	4.08	9.72	0.00	0.00	0.00
	7	0.12	4.60	11.76	0.00	0.00	0.00
	8	0.18	5.82	13.30	0.00	0.00	0.00

- In the cases where the species tree produces anomalous gene trees, the majority of the most probable gene tree topologies are not too far from the species tree topology.
- Considering unranked and unrooted gene tree topologies within one nearest neighbour interchange from the species tree topology is a reasonable heuristic to infer the existence of anomalous trees.
- Anomalous ranked gene trees tended to have the same unranked topology as the species tree or at least are tied for that.

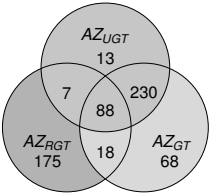
- We simulated 1000 species phylogenies under a birth-death model.
- The probability of being in an anomaly zone increases with the number of taxa n and with speciation rate λ.



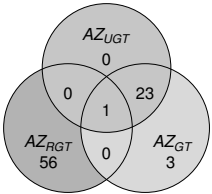
(a) $n = 9$, $\lambda = 0.1$



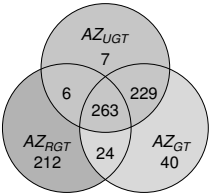
(b) $n = 9$, $\lambda = 1$



(c) $n = 12$, $\lambda = 0.1$



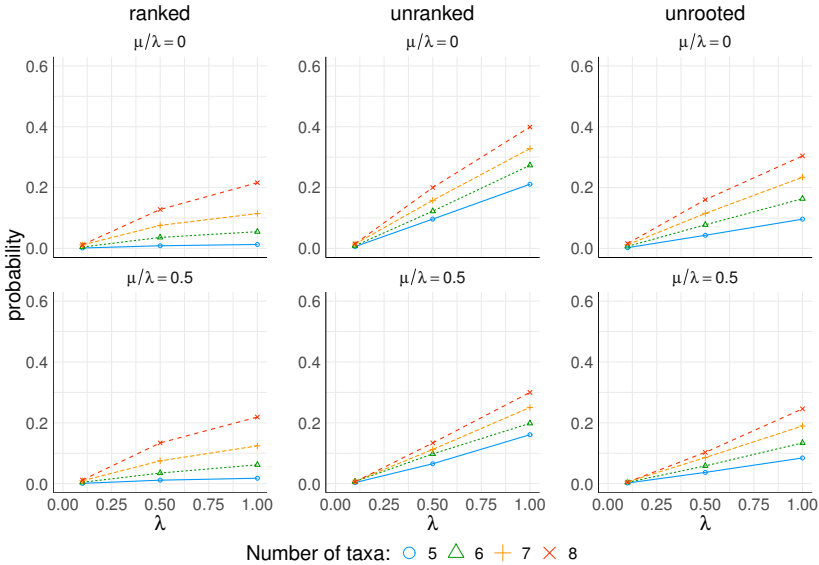
(d) $n = 12$, $\lambda = 1$



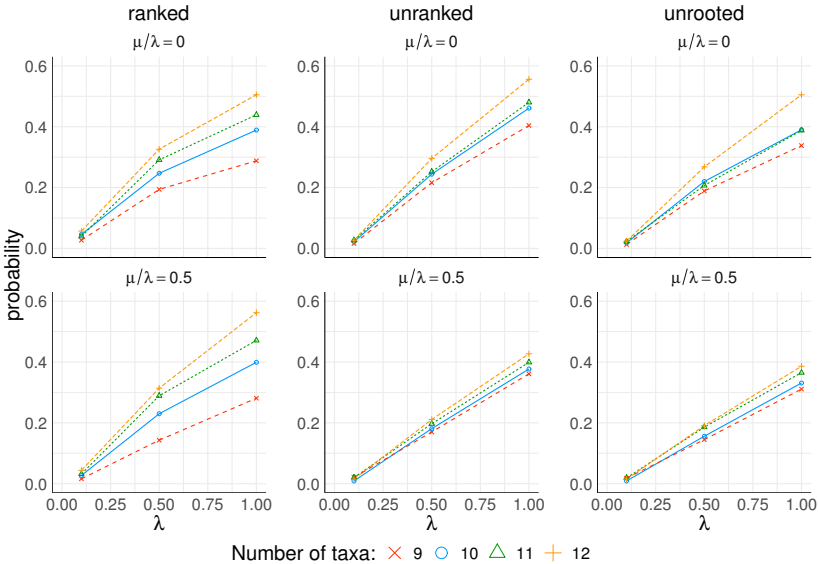
- Probability of being in the all types of anomaly zones increases with the number of taxa and speciation rate λ.
- Probabilities of unranked and unrooted anomaly zones grow much faster than that of the ranked anomaly zone as the speciation rate increases.
- Heuristic approaches are generally useful for anomaly zone calculation with high true positive rates and low false positives.

- J. H. Degnan and N.A. Rosenberg. Discordance of species trees with their most likely gene trees. PLoS Genet., 2006.
- J. H. Degnan, N.A. Rosenberg, and T. Stadler. The probability distribution of ranked gene trees on a species tree. Math. Biosci., 2012.
- C. W. Linkem, V. N. Minin, and A. D. Leache. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). Syst. Biol., 2016
- N. A. Rosenberg. Discordance of species trees with their most likely gene trees: A unifying principle. Mol. Biol. Evol., 2013.

Probabilities of the 5-8-taxon species trees being in the anomaly zones



Probabilities of the 9-12-taxon species tree being in the anomaly zones



Probabilities of the 9-taxon species trees being in the anomaly zones