Gene-Tree Statistics: Moments and Limit Laws for Ancestral Configurations

(joint with F. Disanto, C.-Y. Huang, A. R. Paningbatan, and N. A. Rosenberg)

Michael Fuchs

Department of Mathematical Sciences
National Chengchi University

NATIONAL CHENGCHI UNIVERSITY

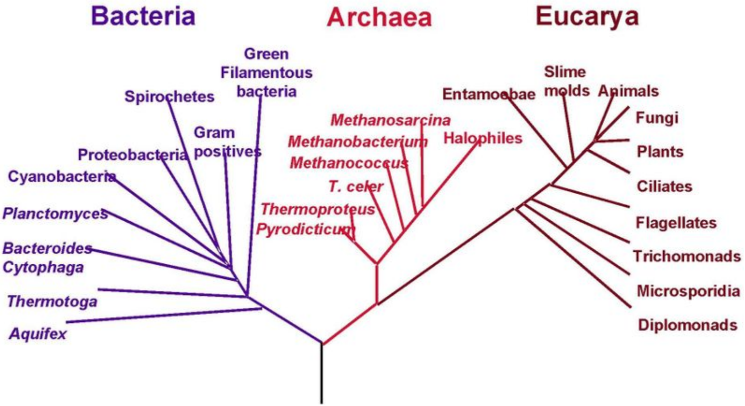
January 7th, 2024

What is a Labeled Topology (or Phylogenetic Tree)?

X . . . a finite set.

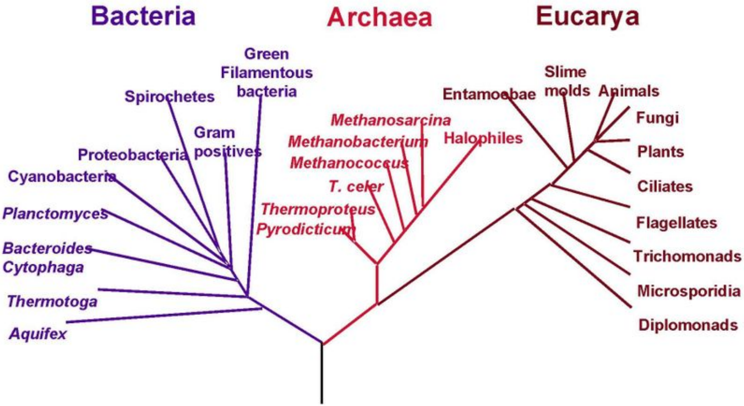
What is a Labeled Topology (or Phylogenetic Tree)?

X . . . a finite set.



What is a Labeled Topology (or Phylogenetic Tree)?

X . . . a finite set.



A labeled topology is a rooted, non-plane, binary tree with leaves labeled by X.

Species and Gene Trees

Species tree: tree of evolutionary relationship between species.

Gene tree: evolutionary relationship at a genomic site.

Species and Gene Trees

Species tree: tree of evolutionary relationship between species.

Gene tree: evolutionary relationship at a genomic site.

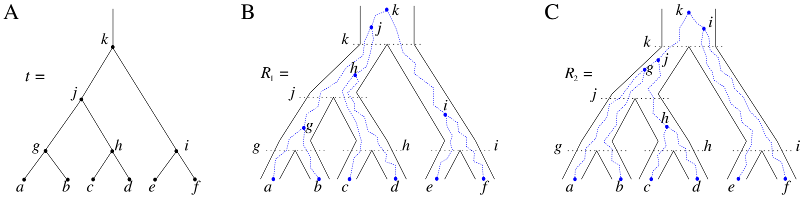


Figure: Two realization of gene trees in the same species tree

Species and Gene Trees

Species tree: tree of evolutionary relationship between species.

Gene tree: evolutionary relationship at a genomic site.

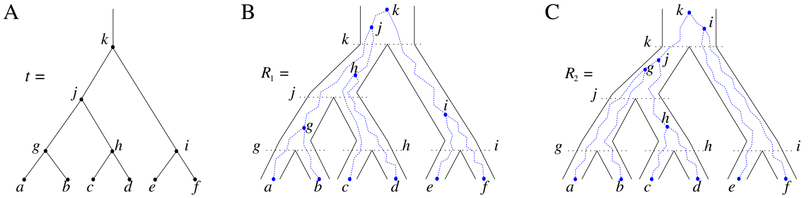


Figure: Two realization of gene trees in the same species tree

We assume throughout the talk that the specific tree and gene tree have the same labeled topology!

Ancestral Configurations

An ancestral configuration is a set of gene lineages present at a vertex of a species tree for a realization of gene tree.

Ancestral Configurations

An ancestral configuration is a set of gene lineages present at a vertex of a species tree for a realization of gene tree.

A root configuration is an ancestral configuration at the root.

Ancestral Configurations

An ancestral configuration is a set of gene lineages present at a vertex of a species tree for a realization of gene tree.

A root configuration is an ancestral configuration at the root.

$c_r(t)$ ... # of root configurations over all gene trees.

Lemma

$$c_r(t) = (c_{r_L}(t_L) + 1)(c_{r_R}(t_R) + 1),$$

where t_L and t_R are the trees rooted at the children of the root of t.

Ancestral Configurations

An ancestral configuration is a set of gene lineages present at a vertex of a species tree for a realization of gene tree.

A root configuration is an ancestral configuration at the root.

$c_r(t)$ ... # of root configurations over all gene trees.

Lemma

$$c_r(t) = (c_{r_L}(t_L) + 1)(c_{r_R}(t_R) + 1),$$

where t_L and t_R are the trees rooted at the children of the root of t.

$c(t)$ ... total number of ancestral configurations.

Then,

$$c(t) = \sum_v c(t_v).$$

Tree Classes

Tree Classes

(i) Labeled topologies: non-plane, leaf-labeled.

$$T_n = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}.$$

Tree Classes

(i) Labeled topologies: non-plane, leaf-labeled.

$$T_n = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}.$$

(ii) Ordered unlabeled topologies: plane, no labels.

$$U_n = C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}.$$

Tree Classes

(i) Labeled topologies: non-plane, leaf-labeled.

$$T_n = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}.$$

(ii) Ordered unlabeled topologies: plane, no labels.

$$U_n = C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}.$$

(iii) Labeled histories: non-plane, leaf-labeled, internal vertices labeled by increasing sequences.

$$H_n = \frac{n!(n-1)!}{2^{n-1}}.$$

Tree Classes

(i) Labeled topologies: non-plane, leaf-labeled.

$$T_n = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}.$$

(ii) Ordered unlabeled topologies: plane, no labels.

$$U_n = C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}.$$

(iii) Labeled histories: non-plane, leaf-labeled, internal vertices labeled by increasing sequences.

$$H_n = \frac{n!(n-1)!}{2^{n-1}}.$$

(iv) Ordered unlabeled histories: plane, no labels, internal vertices labeled by increasing sequences.

$$F_n = (n-1)!.$$

Random Labeled Topologies

Random Labeled Topologies

(i) Uniform model (or PDA model):

Labeled topologies with n leaves are picked uniformly at random, i.e.,

$$P_{\mathsf{uni}}(t) = \frac{1}{T_n} = \frac{1}{(2n-3)!!}.$$

Random Labeled Topologies

(i) Uniform model (or PDA model):

Labeled topologies with n leaves are picked uniformly at random, i.e.,

$$P_{\text{uni}}(t) = \frac{1}{T_n} = \frac{1}{(2n-3)!!}.$$

(ii) Yule-Harding model:

Random model induced on the set of labeled topologies by the uniform model on the set of labeled histories.

Random Labeled Topologies

(i) Uniform model (or PDA model):

Labeled topologies with n leaves are picked uniformly at random, i.e.,

$$P_{\mathsf{uni}}(t) = \frac{1}{T_n} = \frac{1}{(2n-3)!!}.$$

(ii) Yule-Harding model:

Random model induced on the set of labeled topologies by the uniform model on the set of labeled histories.

Thus,

$$P_{\mathsf{YH}}(t) = \frac{2^{n-1}}{n! \prod_{r=3}^{n} (r-1)^{d_r(t)}},$$

where $d_r(t)$ is the number of internal nodes with r leaves below them.

Known Results (Disanto & Rosenberg; 2017)

Known Results (Disanto & Rosenberg; 2017)

(i) Maximally balanced labeled topologies have the largest number of root configurations; caterpillars have the minimal number.

Known Results (Disanto & Rosenberg; 2017)

(i) Maximally balanced labeled topologies have the largest number of root configurations; caterpillars have the minimal number.

(ii) For the uniform model:

$$\mathbb{E}_n[c_r(t)] \sim \sqrt{\frac{3}{2}} \left(\frac{4}{3}\right)^n,$$
$$\mathbb{E}_n[c(t)] \bowtie \left(\frac{4}{3}\right)^n$$

and

$$\mathbb{V}_n[c_r(t)] \sim \sqrt{\frac{7(11-\sqrt{2})}{34}} \left(\frac{4}{7(8\sqrt{2}-11)}\right)^n,$$
$$\mathbb{V}_n[c(t)] \bowtie \left(\frac{4}{7(8\sqrt{2}-11)}\right)^n.$$

Uniform Model

Lemma (Disanto, F., Paningbatan, Rosenberg; 2022)

The distribution of the number of root configurations under the uniform model is the same as under uniformly ordered unlabeled topologies.

Uniform Model

Lemma (Disanto, F., Paningbatan, Rosenberg; 2022)

The distribution of the number of root configurations under the uniform model is the same as under uniformly ordered unlabeled topologies.

Proposition (Disanto, F., Paningbatan, Rosenberg; 2022)

For the number R_n of root configurations under the uniform model:

$$R_n \stackrel{d}{=} (R_{I_n} + 1)(R^*_{n-I_n} + 1),$$

where R^*_n is an independent copy of R_n and

$$P(I_n = j) = \frac{C_{j-1}C_{n-j-1}}{C_{n-1}}, \qquad (1 \le j \le n-1).$$

Limit Law under the Uniform Model

Additive tree functional: a function $F(t)$ which satisfies

$$F(t) = F(t_L) + F(t_R) + f(t),$$

where $f(t)$ is a given toll function.

Limit Law under the Uniform Model

Additive tree functional: a function $F(t)$ which satisfies

$$F(t) = F(t_L) + F(t_R) + f(t),$$

where $f(t)$ is a given toll function.

Wagner (2015) gives a CLT under mild conditions for $f(t)$.

Limit Law under the Uniform Model

Additive tree functional: a function $F(t)$ which satisfies

$$F(t) = F(t_L) + F(t_R) + f(t),$$

where $f(t)$ is a given toll function.

Wagner (2015) gives a CLT under mild conditions for $f(t)$.

Theorem (Disanto, F., Paningbatan, Rosenberg; 2022)

Under the uniform model, $c_r(t)$ is asymptotically lognormal distributed. Moreover,

$$\mathbb{E}_n[\log c_r(t)] \sim \mu n, \qquad \mathbb{V}_n[\log c_r(t)] \sim \sigma^2 n,$$

where $(\mu, \sigma^2) \approx (0.272, 0.034)$.

Yule-Harding Model

Lemma (Disanto, F., Paningbatan, Rosenberg; 2022)

The distribution of the number of root configurations under the Yule-Harding model is the same as under uniformly ordered unlabeled histories.

Yule-Harding Model

Lemma (Disanto, F., Paningbatan, Rosenberg; 2022)

The distribution of the number of root configurations under the Yule-Harding model is the same as under uniformly ordered unlabeled histories.

Proposition (Disanto, F., Paningbatan, Rosenberg; 2022)

For the number R_n of root configurations under the Yule-Harding model:

$$R_n \stackrel{d}{=} (R_{I_n} + 1)(R^*_{n-I_n} + 1),$$

where R^*_n is an independent copy of R_n and

$$P(I_n = j) = \frac{1}{n-1}, \qquad (1 \leq j \leq n-1).$$

Mean under the Yule-Harding Model

Let $e_n := \mathbb{E}[R_n]$.

Mean under the Yule-Harding Model

Let $e_n := \mathbb{E}[R_n]$. Then,

$$e_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} e_j.$$

Mean under the Yule-Harding Model

Let $e_n := \mathbb{E}[R_n]$. Then,

$$e_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} e_j.$$

Set:

$$E(z) := \sum_{n\geq 1} e_n z^n.$$

Then, $E(z)$ satisfies the Riccati DE

$$zE'(z) = E(z)^2 + \frac{1+z}{1-z}E(z) + \frac{z^2}{(1-z)^2}$$

Mean under the Yule-Harding Model

Let $e_n := \mathbb{E}[R_n]$. Then,

$$e_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} e_j.$$

Set:

$$E(z) := \sum_{n\geq 1} e_n z^n.$$

Then, $E(z)$ satisfies the Riccati DE

$$zE'(z) = E(z)^2 + \frac{1+z}{1-z}E(z) + \frac{z^2}{(1-z)^2}$$

with solution

$$E(z) = \frac{2z\sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)}{(z-1)\left[\sqrt{3}\cos\left(\frac{\sqrt{3}}{2}\log(1-z)\right) + \sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)\right]}.$$

Mean and Variance under the Yule-Harding Model

From $E(z)$ we obtain the asymptotics of $[z^n]E(z)$ by singularity analysis.

Mean and Variance under the Yule-Harding Model

From $E(z)$ we obtain the asymptotics of $[z^n]E(z)$ by singularity analysis.

Theorem (Disanto, F., Paningbatan, Rosenberg; 2022)

Under the Yule-Harding model,

$$\mathbb{E}_n[c_r(t)] \sim \left(1 - e^{-2\pi\sqrt{3}/9}\right)^{-n}.$$

Mean and Variance under the Yule-Harding Model

From $E(z)$ we obtain the asymptotics of $[z^n]E(z)$ by singularity analysis.

Theorem (Disanto, F., Paningbatan, Rosenberg; 2022)

Under the Yule-Harding model,

$$\mathbb{E}_n[c_r(t)] \sim \left(1 - e^{-2\pi\sqrt{3}/9}\right)^{-n}.$$

Similarly, but with a more involved analysis, we obtain the variance.

Theorem (Disanto, F., Paningbatan, Rosenberg; 2022)

Under the Yule-Harding model,

$$\mathbb{V}_n[c_r(t)] \sim (2.0449954\cdots)^n.$$

Variance under the Yule-Harding Model (i)

Let $s_n := \mathbb{E}[R_n^2]$.

Variance under the Yule-Harding Model (i)

Let $s_n := \mathbb{E}[R_n^2]$. Then,

$$s_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} s_j s_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} s_j + \frac{4}{n-1}\sum_{j=1}^{n-1} s_j e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j.$$

Variance under the Yule-Harding Model (i)

Let $s_n := \mathbb{E}[R_n^2]$. Then,

$$s_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} s_j s_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} s_j + \frac{4}{n-1}\sum_{j=1}^{n-1} s_j e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j.$$

Set

$$S(z) = \sum_{n\geq 1} s_n z^n.$$

Then,

$$zS'(z) = S(z)^2 + \left[\frac{1+z}{1-z} + 4E(z)\right] S(z) + \frac{(z + 2(1-z)E(z))^2}{(1-z)^2}.$$

This is again a Riccati DE.

Variance under the Yule-Harding Model (ii)

Solving it gives $S(z) = -zU'(z)/U(z)$, where

$$U''(z) - \left(g_1(z) + \frac{g_2'(z)}{g_2(z)}\right)U'(z) + g_2(z)g_0(z)U(z) = 0$$

with

$$(g_2(z), g_1(z), g_0(z)) = \left(\frac{1}{z}, \frac{1}{z}\left(\frac{1+z}{1-z} + 4E(z)\right), \frac{(z+2(1-z)E(z))^2}{z(1-z)^2}\right).$$

Variance under the Yule-Harding Model (ii)

Solving it gives $S(z) = -zU'(z)/U(z)$, where

$$U''(z) - \left(g_1(z) + \frac{g_2'(z)}{g_2(z)}\right)U'(z) + g_2(z)g_0(z)U(z) = 0$$

with

$$(g_2(z), g_1(z), g_0(z)) = \left(\frac{1}{z}, \frac{1}{z}\left(\frac{1+z}{1-z} + 4E(z)\right), \frac{(z+2(1-z)E(z))^2}{z(1-z)^2}\right).$$

Lemma (Disanto, F., Paningbatan, Rosenberg; 2022)

$U(z)$ is analytic in $D(0;1/2)$ and has a unique, simple root β with

$$\beta \approx 0.4889986317.$$

Summary (# of Root Configurations)

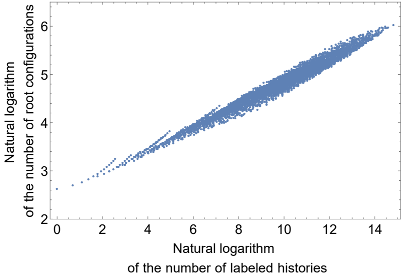
Summary (# of Root Configurations)

quantity	uniform model	Yule-Harding model
mean	$\mathbb{E}_n[c_r] \sim 1.225 \cdot 1.333^n$	$\mathbb{E}_n[c_r] \sim 1.425^n$
variance	$\mathbb{V}_n[c_r] \sim 1.405 \cdot 1.822^n$	$\mathbb{V}_n[c_r] \sim 2.045^n$
log-mean	$\mathbb{E}_n[\log c_r] \sim 0.272 \cdot n$	$\mathbb{E}_n[\log c_r] \sim 0.351 \cdot n$
log-variance	$\mathbb{V}_n[\log c_r] \sim 0.034 \cdot n$	$\mathbb{V}_n[\log c_r] \sim 0.008 \cdot n$

Summary (# of Root Configurations)

quantity	uniform model	Yule-Harding model
mean	$\mathbb{E}_n[c_r] \sim 1.225 \cdot 1.333^n$	$\mathbb{E}_n[c_r] \sim 1.425^n$
variance	$\mathbb{V}_n[c_r] \sim 1.405 \cdot 1.822^n$	$\mathbb{V}_n[c_r] \sim 2.045^n$
log-mean	$\mathbb{E}_n[\log c_r] \sim 0.272 \cdot n$	$\mathbb{E}_n[\log c_r] \sim 0.351 \cdot n$
log-variance	$\mathbb{V}_n[\log c_r] \sim 0.034 \cdot n$	$\mathbb{V}_n[\log c_r] \sim 0.008 \cdot n$

"Balanced" labeled topologies tend to have more root configurations.



Total Number of Ancestral Configurations

R_n ... # of root configurations;

T_n ... # total number of ancestral configurations.

Total Number of Ancestral Configurations

R_n ... # of root configurations;

T_n ... # total number of ancestral configurations.

Then,

$$R_n \stackrel{d}{=} R_{I_n} + R^*_{n-I_n} + R_{I_n} + R^*_{n-I_n} + 1,$$
$$T_n \stackrel{d}{=} T_{I_n} + T^*_{n-I_n} + R_n,$$

where R^*_n and T^*_n are independent copies of R_n and T_n and

$$P(I_n = j) = \begin{cases} C_{j-1}C_{n-1-j}/C_{n-1}, & \text{uniform model;} \\ 1/(n-1), & \text{Yule-Harding model.} \end{cases}$$

Total Number of Ancestral Configurations

R_n ... # of root configurations;

T_n ... # total number of ancestral configurations.

Then,

$$R_n \stackrel{d}{=} R_{I_n} + R^*_{n-I_n} + R_{I_n} + R^*_{n-I_n} + 1,$$
$$T_n \stackrel{d}{=} T_{I_n} + T^*_{n-I_n} + R_n,$$

where R_n^* and T_n^* are independent copies of R_n and T_n and

$$P(I_n = j) = \begin{cases} C_{j-1}C_{n-1-j}/C_{n-1}, & \text{uniform model;} \\ 1/(n-1), & \text{Yule-Harding model.} \end{cases}$$

Also,

$$R_n \leq T_n \leq (2n-1)R_n.$$

Results under Uniform Model

Theorem (Disanto, F., Paningbatan, Rosenberg; 2024)

We have,

$$\mathbb{E}_n[c(t)] \sim \sqrt{6}\left(\frac{4}{3}\right)^n,$$

$$\mathbb{V}_n[c(t)] \sim \frac{2(15+11\sqrt{2})}{17}\sqrt{\frac{7(11-\sqrt{2})}{34}}\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n.$$

In addition,

$$\frac{\log c(t) - \mathbb{E}_n[\log c(t)]}{\sqrt{\mathbb{V}_n[\log c(t)]}} \xrightarrow{d} N(0,1)$$

with

$$\mathbb{E}_n[\log c(t)] \sim 0.272 \cdot n, \qquad \mathbb{V}_n[\log c(t)] \sim 0.034 \cdot n.$$

Results under Yule-Harding Model

Theorem (Disanto, F., Paningbatan, Rosenberg; 2024)

We have,

$$\mathbb{E}_n[c(t)] \sim \left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n,$$
$$\mathbb{V}_n[c(t)] \sim (2.0449954\cdots)^n.$$

In addition,

$$\frac{\log c(t) - \mathbb{E}_n[\log c(t)]}{\sqrt{\mathbb{V}_n[\log c(t)]}} \xrightarrow{d} N(0,1)$$

with

$$\mathbb{E}_n[\log c(t)] \sim 0.351 \cdot n, \qquad \mathbb{V}_n[\log c(t)] \sim 0.008 \cdot n.$$

Summary (# of Ancestral Configurations)

quantity	uniform model	Yule-Harding model
$\mathbb{E}_n[c]$	$\sqrt{6}\left(\frac{4}{3}\right)^n$	$\left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n$
$\mathbb{E}_n[c^2]$	$\frac{2(15+11\sqrt{2})}{17}\sqrt{\frac{7(11-\sqrt{2})}{34}}\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n$	$(2.0449954\cdots)^n$
$\mathbb{V}_n[c]$	$\frac{2(15+11\sqrt{2})}{17}\sqrt{\frac{7(11-\sqrt{2})}{34}}\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n$	$(2.0449954\cdots)^n$
$\mathbb{E}_n[c_r c]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n$	$(2.0449954\cdots)^n$
$\mathrm{Cov}_n[c_r, c]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n$	$(2.0449954\cdots)^n$
$\rho_n[c_r, c]$	$\frac{1+\frac{\sqrt{2}}{2}}{\sqrt{\frac{2(15+11\sqrt{2})}{17}}}$	1

References

1. *F. Disanto, M. Fuchs, A. R. Paningbatan, N. A. Rosenberg (2022). The distribution under two species-tree models of the number of root configurations for matching gene trees and species trees, Ann. Appl. Probab., 32:6, 4426–4458.*

2. *F. Disanto, M. Fuchs, C.-Y. Huang, A. R. Paningbatan, N. A. Rosenberg (2024). The distribution under two species-tree models of the total number of ancestral configurations for matching gene trees and species trees, Adv. Appl. Math., 152, 102594.*