Technical Report Series
Center for Data and Simulation Science

Thomas Wiehe

Counting, grafting and evolving binary trees

Technical Report ID: CDS-2020-8
Available at `https://kups.ub.uni-koeln.de/id/eprint/12199`
Submitted on October 1, 2020

Contents

CHAPTER 1

Counting, grafting and evolving binary trees

Thomas Wiehe

Binary trees are fundamental objects in models of evolutionary biology and population genetics. Here, we discuss some of their combinatorial and structural properties as they depend on the tree class considered. Furthermore, the process by which trees are generated determines the probability distribution in tree space. *Yule trees*, for instance, are generated by a pure birth process. When considered as unordered, they have neither a closed-form enumeration nor a simple probability distribution. But their ordered siblings have both. They present the object of choice when studying tree structure in the framework of evolving genealogies.

1.1. Introduction

Trees appear in different contexts and with different properties. In graph theory, they are defined as connected, acyclic graphs: any pair of vertices (*nodes*) is connected by exactly one concatenated sequence of edges (*branches*). Tagging one node, called *root* of the tree, implicitly establishes a directionality of the graph. In theoretical biology, trees are used to describe genealogies of cells, genes, individuals or species. Depending on the biological context, planarity of the tree, degree and labelling of nodes, directionality and length of branches may or may not be of interest. Cardinality and probability distribution depend strongly on these properties.

The study of trees as mathematical objects reaches back at least to the 1850s, when Cayley [9] derived recursion formulas for the enumeration of trees with a finite number of nodes, and also recognised the link to isomer chemistry. As an alternative to recursions, bijections between trees and permutations can help to solve certain counting problems [2, 16]. More generally, and yielding insight into asymptotic behaviour for large trees, the tools of analytic combinatorics are particularly powerful. Comprehensive treatments are found in the classical textbook by Flajolet and Sedgwick [23] and, focusing on random trees only, in the textbook by Drmota [17]. With a view from computer science, where they appear primarily as data structures, trees are covered in the epitomic opus by Knuth [30, Vol. 1].

The link of 'tree theory' with biology has been established by Yule's seminal paper of 1925 [53], when seeking to explain the distribution of the number of species within genera. It initiated a long tradition of research in phylogenetics and macro-evolution on enumeration, topology and distribution of trees generated by random processes [8, 20, 44, 34, 29, 36, 46, 5, 31]. The border between macro- and micro-evolution is fuzzy, but intensely investigated in the context of gene tree embeddings in species trees [41, 13, 31, 15]. Perhaps the most genuine application of Yule's original model, and with most ramifications, lies in population genetics as a model of individual gene genealogies and their statistical properties. Kingman's [28] coalescent is its backward-in-time analogue and — in the guise of its evolved descendants — features in several chapters of this volume. The genetic operation of recombination translates into subtree-prune and -regraft operations, opening a field of active theoretical research on tree transformations [45], in part also covered in this volume. Standard references on the coalescent are the textbooks by Wakeley [49] and Durett [18]. Aldous [1] offers a view on Yule's paper from a modern perspective.

Given that trees are treated in different disciplines, and with different degree of mathematical rigour, it is not surprising to find oneself confronted with a non-unified, sometimes even inconsistent, terminology and nomenclature, which alone can make it difficult to identify the relevant theoretical features of some tree class for a specific biological application. Without claiming to authoritatively clarify this problem, we start the section below with an (incomplete) catalogue of tree classes and their enumerations (Section 1.2). We will then devote special attention to Yule trees and explore some of their structural properties (Sections 1.3 and 1.4). Since they represent the scaffold of the widely used coalescent model in population genetics, we will consider two such applications (Sections 1.5 and 1.6).

1.2. Counting trees

1.2.1. Preliminaries

We consider rooted, binary, finite trees: there is a unique node, the root, defining a directionality for all branches. Each branch is delimited by a *parent* and a *child* node. The root is *ancestor* of all other nodes. They are subdivided into $n < \infty$ *external* and $m = n-1$ *internal* nodes, including the root. All internal nodes have exactly two children. External nodes have no *descendants* and are also called *leaves*. The *size* of a tree is the number of its leaves. A *subtree* is a tree that is rooted at some node of the original tree. Subtrees of size 2 are also called *cherries*, subtrees of size 3 *pitchforks*. A *caterpillar* is a tree for which at least one of the subtrees at each of its internal nodes has size 1. Slightly more generally, a *c-caterpillar* is a (sub-)tree of size c that is a caterpillar. Thus, a cherry is a 2-caterpillar, and a pitchfork is a 3-caterpillar. Since trees here are binary, all

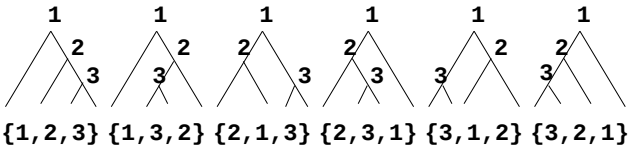


FIGURE 1.2.1. The six ordered ranked trees of size $n = 4$ and the corresponding six permutations of $\{1, 2, 3\}$ obtained by reading out internal labels during *in-order* tree traversal [30]. Note, for example, the difference between $\{2, 1, 3\}$ and $\{3, 1, 2\}$.

internal nodes have a *left* and a *right subtree*, which are rooted at the left and right child. Trees are *ordered* (plane), if left and right can be distinguished, otherwise they are *un-ordered* (non-plane).

1.2.2. Classification of binary trees

Tree enumerations depend crucially on the presence and the kind of node labels. Among the many possibilities, we restrict ourselves to the following cases: presence or absence of alphanumeric labels at external nodes, and presence or absence of totally ordered numeric labels at internal nodes. Trees without any node labels are called *shape trees* or *topologies* [8, 40]. We call a tree *ranked* or a *history* [25, 47, 13], if the internal nodes are labelled with integers $1, \ldots, n-1$ such that (i) the root has label 1, (ii) distinct nodes have distinct labels and (iii) every child has a larger label than its parent. We call a tree *labelled*, if the leaves carry labels. Labelled trees can be thought of as phylogenies with species names as leaf labels. Without internal labels, they are also called *cladograms*, with internal labels they are *ranked phylogenies* or *labelled histories* [47]. Their cardinality follows, for instance, from a coalsecent-like construction: randomly selecting two out of k labelled lineages to coalesce, there are $\binom{k}{2}$ possibilities [31]. The product is $\prod_{k=2}^{n} \binom{k}{2} = n!(n-1)!/2^{n-1}$.

When shape trees have a left/right orientation, they are called *Catalan* trees, because they are enumerated by the Catalan numbers $C_m = \binom{2m}{m}/(m+1)$, [43, A000108], where $m = n - 1$ is the number of internal nodes of such trees. Finally, *ordered histories* are ordered ranked trees. Since they map bijectively to permutations of $m = n-1$ integers, we also call them *permutation trees*. They are enumerated by the factorials $m!$. To see this, one can read the labels of all ordered ranked trees of a given size in an *in-order* [30] tree traversal, observing that all subtrees, except cherries, have a distinguishable left-right order (Figure 1.2.1).

We denote ordered trees of size n by $\overset{\circ}{\Lambda}{}^{\cdot\cdot}_n$ and un-ordered trees by $\Lambda^{\cdot\cdot}_n$. The exponent is a placeholder to indicate presence or absence of internal or external labels. The tree classes mentioned above are summarised in Table 1.2.1. Note

TABLE 1.2.1. Classes of un-ordered (Λ) and ordered ($\mathring{\Lambda}$) trees of size n. Presence ($+$) or absence ($-$) of internal or external labels is indicated by superscripts. Cardinalities are $|\Lambda_n|$ and $|\mathring{\Lambda}_n|$.

name	alias	int. lab.	ext. lab.	symbol	cardinality	OEIS[1] ID
unordered trees						
shape trees	topologies[2]	$-$	$-$	Λ_n^{--}	Eq. (1.2.1)	A001190
ranked trees	histories[3]	$+$	$-$	Λ_n^{+-}	Eq. (1.2.2)	A000111
labelled trees	phylogenies[4]	$-$	$+$	Λ_n^{-+}	$\frac{(2n-3)!}{2^{n-2}(n-2)!}$	A001147
labelled ranked trees[5]	ranked phylogenies	$+$	$+$	Λ_n^{++}	$\frac{n!(n-1)!}{2^{n-1}}$	A006472
ordered trees						
Catalan trees[6]	ordered topologies	$-$	$-$	$\mathring{\Lambda}_n^{--}$	$\frac{1}{n}\binom{2(n-1)}{(n-1)}$	A000108
permutation trees	ordered histories[7]	$+$	$-$	$\mathring{\Lambda}_n^{+-}$	$(n-1)!$	A000142

[1] `www.oeis.org`, [43]

[2] [8, 40]; called *topological types* in [44]

[3] [25, 47, 13]

[4] [1, 44]; called *rooted* phylogeny in [20] or *tree form* in [8]

[5] cf. [31], there in the context of Kingman's coalescent

[6] [17, p. 5]

[7] called *shapes* in [25]

that these classes represent only a subset of the possibilities. For instance, Felsenstein [20] discusses phylogenies with non-numeric labels at internal nodes. This constitutes a class that is different from Λ_n^{++} and that has a different cardinality: it leads to Cayley's formula [10], enumerating non-binary trees (cf. [43, A000169] and [25]). Not all tree classes have closed form enumerations. Often, ordered trees do, while un-ordered trees do not [23, p. 87]. In our list (Table 1.2.1), the cardinalities of un-ordered shape and ranked trees are given only implicitly via generating functions, but their ordered versions have closed formulae.

The *(ordinary) generating function* and the *exponential generating function* of an integer sequence $(a_n)_n$ are given by the formal power series

$$f(x) = \sum_{n \geqslant 0} a_n x^n \qquad \text{and} \qquad F(x) = \sum_{n \geqslant 0} a_n \frac{x^n}{n!}\,,$$

respectively. If f or F are holomorphic functions defined in a neighbourhood around $x = 0$, the series can be interpreted as their Taylor expansions and, for instance, their asymptotic properties can be studied by analytic means.

In 1922, Wedderburn [50] showed that the cardinalities of shape trees can be implicitly represented via a functional equation of a generating function. De Bruijn and Klarner derived the somewhat simpler representation

$$f(x) = x + 1/2\left(f^2(x) + f(x^2)\right) \tag{1.2.1}$$

and showed [7] that its solution f generates the cardinalities of shape trees of size n, via

$$f(x) = \sum_n |\Lambda_n^{--}|\, x^n\,.$$

For $1 \leqslant n \leqslant 10$, the coefficients are $1, 1, 1, 2, 3, 6, 11, 23, 46, 98$.

For unordered ranked trees (histories), the cardinalities are identical with the *Euler numbers* and are given by the coefficients of the exponential generating function

$$F(x) = \sec(x) + \tan(x) = \sum_n |\Lambda_{n+1}^{+-}|\,\frac{x^n}{n!}\,, \tag{1.2.2}$$

for $1 \leqslant n \leqslant 10$, they are $1, 1, 1, 2, 5, 16, 61, 272, 1385, 7936$.

A natural way to construct unordered ranked trees of any finite size is by recursion: given a ranked tree of size $m = n - 1$, construct a tree of size n by randomly choosing one of the m leaves to give rise to two children and label the chosen leaf with the integer n. Following other authors [47, 11], we call trees generated in this way *Yule trees* and the underlying model (process) the *Yule model* (*Yule process*). In the equivalent backward process, one starts from n leaves and their n parental branches. One randomly, and iteratively, selects two branches to coalesce into a single one until all are coalesced. When, in addition, a time axis for the coalescent times is introduced, and when these times are exponentially distributed with a parameter proportional to $\binom{k}{2}$, where k is the current number of branches, Yule trees are called *coalescent trees*, generated by the (Kingman-) coalescent process [28]. They are the basis of a plethora of genealogical models in population genetics.

1.3. Properties of ranked trees

Note that the Yule process does not generate uniformly distributed trees in Λ_n^{+-}. For instance, in Fig 1.2.1 the 4-caterpillar is generated with probability $2/3$ and the balanced tree, corresponding to the permutations $\{2, 1, 3\}$ and $\{3, 1, 2\}$, with probability $1/3$. Only when considered as trees in $\mathring{\Lambda}_n^{+-}$, they become uniformly distributed under the Yule process, each with probability $1/(n-1)!$. Other tree generating processes may lead to still other probability distributions [34].

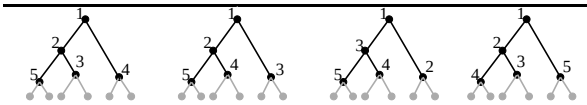
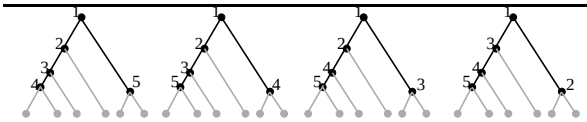
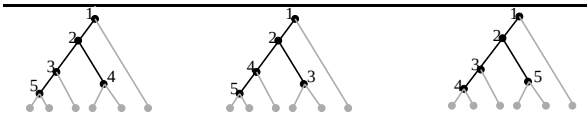
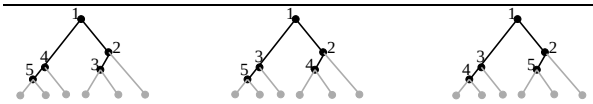
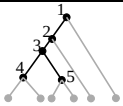
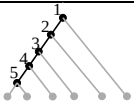
ranked trees	$\lvert\mathcal{C}_2\rvert$	$\lvert\mathcal{C}_3\rvert$	$\lvert\Lambda\rvert$	factor	$\lvert\mathring{\Lambda}\rvert$
	3	0	4	2^{5-3}	16
	2	1	4	2^{5-2}	32
	2	1	3	2^{5-2}	24
	2	2	3	2^{5-2}	24
	2	0	1	2^{5-2}	8
	1	1	1	2^{5-1}	16

FIGURE 1.2.2. The sixteen possible un-ordered ranked trees of size $n = 6$, classified by shape. Within each class, all admissible orderings of the internal nodes are displayed. Number of cherries ($|\mathcal{C}_2|$) and pitchforks ($|\mathcal{C}_3|$) are indicated. The number of all ordered ranked trees, classified by shape, is obtained by multiplying with the factor $2^{m-|\mathcal{C}_2|}$. The total number is $5! = 120$. Branch lengths are without meaning; position of an internal node in a tree is given by the node label, not by the actual drawing of its position. External nodes and branches are shown in grey. Removing them leads to the *reduced trees* of size 5. They can be uniquely identified with the original trees of size 6.

Since ordered and un-ordered trees are identical up to left/right order of subtrees that are not cherries, there are exactly 2^{n-1-o} different ordered trees for each unordered one with o cherries. Thus, given a ranked tree, one also knows the probability with which it is generated, by simply counting its cherries (cf. [48]). With $\mathcal{O}$ denoting the random variable for the number of cherries, we have

$$\text{Prob(given ranked tree of size } n \text{ with } \mathcal{O} = o \text{ cherries)} = \frac{2^{n-1-o}}{(n-1)!}\,. \tag{1.3.1}$$

To explore the unconditional distribution of Yule trees, we remark that all external nodes and branches (shown in grey in Figure 1.2.2) may be stripped from a ranked tree of size n without loss of information. Such stripping leads to a *reduced* tree with $m = n - 1$ nodes with ordered labels, all of out-degree 0, 1 or 2 [14]. Nodes of out-degree 0 represent cherries in the original tree. Sometimes, reduced trees are called *pruned* trees [23], a term which we avoid, to not confuse it with 'tree pruning' discussed later. Reduced trees with m nodes can be constructed recursively, starting from a reduced tree with one node, according to the following production rule

$$(o, m) \longrightarrow (o, m+1)^o (o+1, m+1)^{m-2o+1} ,$$

where o is the number of cherries and m the total number of nodes in the current tree. The exponent counts how many new trees with o (or $o+1$) cherries and $m+1$ nodes are produced. Note that in each step m is increased by one and the number of cherries may either remain unchanged or also increase by one. The former happens when the new branch and node are appended at a node of out-degree 0, the latter, when appended at a node of out-degree 1. At nodes of out-degree 2 (true internal nodes) nothing can be appended. For instance, starting with $(1, 1)$, the production rule generates the sequence

$$(1,2)^1(2,2)^0,\ (1,3)^1(2,3)^1,\ (1,4)^1(2,4)^2 \text{ and } (2,4)^2(3,4)^0, \ldots .$$

Consider now the bivariate exponential generating function

$$F(x, z) = \sum_{\substack{\text{reduced trees with } o \\ \text{cherries and } m \text{ nodes}}} x^o \frac{z^m}{m!} . \tag{1.3.2}$$

The production rule can then be translated into algebraic terms as

$$\begin{aligned} F(x,z) &= xz + \sum \frac{o x^o z^{m+1}}{(m+1)\,!} + \sum \frac{(m-2o+1)(x^{o+1} z^{m+1})}{(m+1)\,!} \\ &= xz + (1-2x) \sum \frac{o x^o z^{m+1}}{(m+1)\,!} + xz \sum \frac{x^o z^m}{m\,!} , \end{aligned}$$

where the summations are over all reduced trees with o cherries and m nodes and the first summand represents a tree of size $m = 1$. Differentiating both sides with respect to the variable z, one obtains a partial differential equation for F

$$x(1-2x)\frac{\partial F}{\partial x}(x,z) + (xz-1)\frac{\partial F}{\partial z}(x,z) = -xF(x,z) - x ,$$

which admits a solution in closed form [14] as

$$F(x,z) = \frac{2\big(x\, \exp(z\sqrt{-2x+1}) - x\big)}{(\sqrt{-2x+1}-1)\exp(z\sqrt{-2x+1}) + \sqrt{-2x+1}+1}. \tag{1.3.3}$$

TABLE 1.3.1. Partitions $e_{m,o}$ of Euler numbers [43, A000111]. $\mathcal{O}$: number of cherries. Column sums $\sum_o e_{m,o} = e_m$. For instance, for $m = 5$ (i.e., $n = 6$) there are one ranked tree with one cherry (the caterpillar), 11 trees with two cherries and 4 trees with three cherries.

$\mathcal{O}$	m									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	0	0	1	4	11	26	57	120	247	502
3	0	0	0	0	4	34	180	768	2904	10194
4	0	0	0	0	0	0	34	496	4288	28768
5	0	0	0	0	0	0	0	0	496	11056
$\sum$	1	1	2	5	16	61	272	1385	7936	50521

One direct application of F is to determine the probability that two randomly generated Yule trees are identical ([14, Thm. 1], with F replaced by Y). Furthermore, F can be used to find a *partition* of the Euler numbers e_m in such a way that $e_{m,o}$ represents the number of (unreduced) ranked trees of size $n = m + 1$ with o cherries. As shown in [14],

$$e_{m,o} = m! \cdot [x^o z^m] F \,,$$

where the brackets $[\cdot]$ denote coefficient extraction. The partitions of e_m for $m = 1, \ldots, 10$ and $o = 1, \ldots, 5$ are shown in Table 1.3.1. Other applications involve simple transformations of F. For instance, with

$$\tilde{F}(x, z) = zF\big(\tfrac{x}{2}, 2z\big)$$

one obtains the weighted (ordinary) generating function

$$\tilde{F}(x, z) = \frac{zx \exp\big(2z\sqrt{-x+1}\big) - zx}{\big(\sqrt{-x+1} - 1\big) \exp\big(2z\sqrt{-x+1}\big) + 1 + \sqrt{-x+1}} \,, \tag{1.3.4}$$

for the coefficients of $x^o z^n$, such that

$$\tilde{F}(x, z) = \sum_{\text{ranked trees of size } n} \frac{2^{n-1-o}}{(n-1)!} x^o z^n \,,$$

leading to the following [14] consequence.

Result 1.3.1. The probability that a Yule tree of size n has o cherries is given by the coefficient of $x^o z^n$ in the Taylor expansion of $\tilde{F}$ around $z = 0$, i.e.,

$$P_n(\mathcal{O} = o) = [x^o z^n] \tilde{F}(x, z) \,.$$

By differentiating $\tilde{F}$, one can easily derive the moments of $\mathcal{O}$. For instance, the mean number of cherries in ranked trees of size n is

$$\mathbb{E}(\mathcal{O}) = [z^n]\left.\frac{\partial \tilde{F}}{\partial x}(x,z)\right|_{x=1} = [z^n]\frac{z^4-3z^3+3z^2}{3(z-1)^2}\,.$$

If $n>2$, this simplifies to

$$\mathbb{E}(\mathcal{O}) = \frac{n}{3}\,.$$

The second moment is

$$\begin{aligned}\mathbb{E}(\mathcal{O}^2) &= [z^n]\left.\frac{\partial(x\frac{\partial \tilde{F}(x,z)}{\partial x})}{\partial x}\right|_{x=1} = [z^n]\left.\frac{\partial^2 \tilde{F}(x,z)}{\partial x^2}\right|_{x=1} + [z^n]\left.\frac{\partial \tilde{F}(x,z)}{\partial x}\right|_{x=1}\\ &= [z^n]\left(\frac{2}{(z-1)^3}\left(\frac{z^7}{45}-\frac{2z^6}{15}+\frac{z^5}{3}-\frac{z^4}{3}\right)\right)+\mathbb{E}(\mathcal{O})\,.\end{aligned}$$

If $n>6$, and using $\mathbb{V}(\mathcal{O}) = \mathbb{E}(\mathcal{O}^2)-\mathbb{E}^2(\mathcal{O})$, one obtains

$$\mathbb{V}(\mathcal{O}) = \frac{2n}{45}\,.$$

The distribution of $\mathcal{O}$ [35], and mean and variance of c-caterpillars [40], have been derived before, however with different methods not employing generating functions. The latter represent a powerful tool to handle the recursive production rules of binary trees, and readily offer a somewhat deeper look into tree structure. Focusing on general c-caterpillars, let

$$F(x_2,x_3,x_4,\ldots,x_k,z) = \sum_{\text{trees of size } n>1} x_2^o x_3^{c_3} x_4^{c_4}\ldots x_k^{c_k}\frac{z^{n-1}}{(n-1)!}$$

be a multi-variate exponential generating function, where c_i is the number of caterpillars of size $i>2$, and o the number of cherries. This function satisfies the partial differential equation

$$\begin{aligned}\frac{\partial F}{\partial z} &= x_2 + x_2F + x_2z\frac{\partial F}{\partial z} + (x_2x_3-2x_2^2)\frac{\partial F}{\partial x_2}\\ &\quad+\sum_{i=3}^{k-1}\Big(x_ix_{i+1}-x_i^2+x_2(1-x_i)\Big(1+\sum_{j=1}^{i-3}\frac{1}{x_{i-1}x_{i-2}\ldots x_{i-j}}\Big)\Big)\frac{\partial F}{\partial x_i}\\ &\quad+\Big(x_k-x_k^2+x_2(1-x_k)\Big(1+\sum_{j=1}^{k-3}\frac{1}{x_{k-1}x_{k-2}\ldots x_{k-j}}\Big)\Big)\frac{\partial F}{\partial x_k}\,,\end{aligned}$$

which leads to a recursively determined family of polynomials $(F_m)_{m\geqslant 1}$ with

$$F_m = \sum_{\text{trees } t \text{ of size } n=m+1}\frac{x_2^{o(t)}x_3^{c_3(t)}x_4^{c_4(t)}\ldots x_k^{c_k(t)}z^{n-1}}{(n-1)!}\,.$$

Defining the operator

$$\mathcal{G}(F) = \frac{\partial F}{\partial z} - x_2\,,$$

the recursion for $(F_m)_{m\geqslant 1}$ is given by

$$\begin{aligned} F_1 &= x_2 z\,,\\ F_{m+1} &= \int \mathcal{G}(F_m)\,\mathrm{d}z\,. \end{aligned} \tag{1.3.5}$$

As an example, fix $k=5$. Then, for $m=1,2,3,4,5$, one has

$$\begin{aligned} F_1 &= x_2 z\,,\\ F_2 &= \frac{1}{2}x_3 x_2 z^2\,,\\ F_3 &= \frac{1}{6}x_3 x_4 x_2 z^3 + \frac{1}{6}x_2^2 z^3\,,\\ F_4 &= \frac{1}{24}x_3 x_4 x_5 x_2 z^4 + \frac{1}{24}x_2^2 z^4 + \frac{1}{8}x_3 x_2^2 z^4\,,\\ F_5 &= \frac{1}{120}x_3 x_4 x_5 x_2 z^5 + \frac{1}{120}x_2^2 z^5 + \frac{1}{40}x_3 x_2^2 z^5 + \\ &\quad \frac{1}{40}{x_3}^2 x_2^2 z^5 + \frac{1}{30}x_3 x_4 x_2^2 z^5 + \frac{1}{30}x_2^3 z^5. \end{aligned}$$

Recursion (1.3.5) yields both the joint distribution of cherries and caterpillars of different sizes and the conditional distribution of caterpillars, conditioned on the number of cherries. Summarising, one can state the following result (cf. [14]).

Result 1.3.2. Given an (unordered) ranked tree T of size $n = m+1$. Then,

i) the probability that T contains c-caterpillars of size k is

$$P_m(\mathcal{C}_k = c) = [x_k^c]F_m\Big(\frac{1}{2}, 1, 1, \ldots, x_k, 2\Big);$$

ii) the joint probability that T contains o cherries and c caterpillars of size k is

$$P_m(\mathcal{O} = o,\, \mathcal{C}_k = c) = [x_2^o x_k^c]F_m\left(\frac{x_2}{2}, 1, 1, \ldots, x_k, 2\right);$$

iii) the conditional probability that T contains c caterpillars of size k, given it has o cherries is

$$P_m(\mathcal{C}_k = c \,|\, \mathcal{O} = o) = \frac{P_m(\mathcal{O} = o, \mathcal{C}_k = c)}{P_m(\mathcal{O} = o)} = \frac{[x_2^o x_k^c]F_m\left(\frac{x_2}{2}, 1, 1, \ldots, x_k, 2\right)}{[x_2^o]F_m\left(\frac{x_2}{2}, 1, 1, \ldots, 1, 2\right)};$$

iv) the probability that T contains c' caterpillars of size i, with $3 \leqslant i < k$, and c caterpillars of size k is

$$P_m(\mathcal{C}_i = c',\, \mathcal{C}_k = c) = [x_i^{c'} x_k^c]F_m\Big(\frac{1}{2}, 1, \ldots 1, x_i, 1, \ldots, x_k, 2\Big);$$

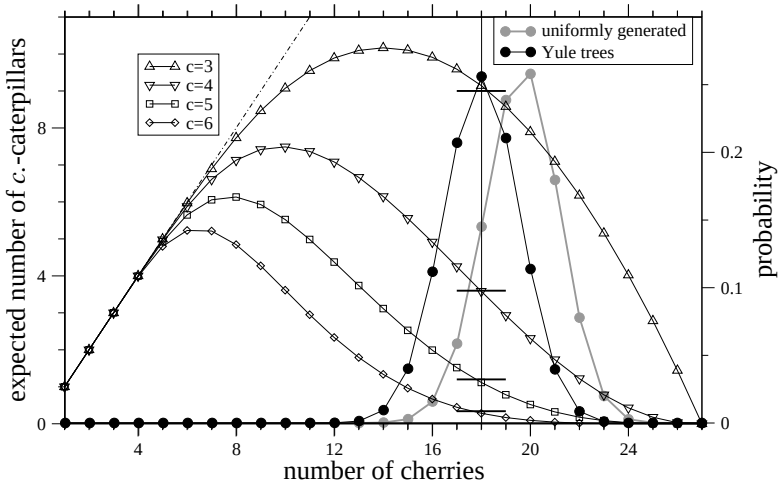


FIGURE 1.3.1. Ranked trees of size $n = 54$. Conditional expectation of the number of c-caterpillars (left y-axis, $c = 3, 4, 5, 6$), given the number of cherries (curves with triangles, diamonds and squares). Vertical black line at $x = 18$: expected number of cherries in unconstrained trees; horizontal black bars: unconditional expected number of c-caterpillars. Curves with filled circles: fraction of trees (right y-axis) with given number of cherries generated under the Yule process (black) and in uniformly generated trees (grey). Equivalently, this is the distribution of cherries ($\mathcal{O}$) in ranked trees. $\mathbb{V}(\mathcal{O})/\mathbb{E}(\mathcal{O}) \approx 0.13$. Dotted line: diagonal $x = y$.

v) the conditional probability that T contains c caterpillars of size k, given it has c' caterpillars of size i, with $3 \leqslant i < k$, is

$$\begin{aligned} P_m(\mathcal{C}_k = c \,|\, \mathcal{C}_i = c') &= \frac{P_m(\mathcal{C}_i = c', \mathcal{C}_k = c)}{P_m(\mathcal{C}_i = c')} \\ &= \frac{[x_i^{c'} x_k^c] F_m\left(\frac{1}{2}, 1, \ldots, 1, x_i, 1, \ldots, x_k, 2\right)}{[x_i^{c'}] F_m\left(\frac{1}{2}, 1, \ldots, 1, x_i, 1, \ldots, 1, 2\right)}. \end{aligned}$$

The distribution of $\mathcal{O}$, both under the Yule process and when trees are generated uniformly, as well as the conditional expectations for some c-caterpillars, are shown in Figure 1.3.1 for the example of size $n = 54$.

1.4. Induced subtrees

Induced subtrees occur as embedded genealogies of a subset of the leaves of a tree [42]. Let T_n be a ranked, labelled tree of size n with leaf labels $L = \{l_1, l_2, \ldots, l_n\}$. Choose $n' \leqslant n$, and select labels $L' = \{l'_1, l'_2, \ldots, l'_{n'}\}$, such that for

each $1 \leqslant i \leqslant n'$ there is exactly one j with $l'_i = l_j$. Then, the *induced subtree* T' is the tree that is obtained from T by maintaining only the branches connecting a leaf l'_i with the most recent common ancestor of all leaves L'. We write $T' \lhd T$ for short. Note that the root of T' is not necessarily identical with the root of T and that the topologies of different induced subtrees of the same supertree T may be different. There are $\binom{n}{n'}$ possible subsets of size n'. When conditioned on a fixed tree T, number and distribution of induced subtrees are obviously different from independently generated trees. There is no general enumeration formula for induced subtrees since the number depends on the topology of T. For instance, take a caterpillar of size n. Then all induced subtrees are caterpillars. Only when averaging over all Yule super-trees of size n, induced subtrees and independently generated trees are identical in number and distribution. We introduce now the notion of *node balance*.

Definition 1.4.1. For an internal node ν_i of a binary rooted tree T let $T_i(L)$ and $T_i(R)$ be the left and right subtrees at node ν_i. We call the minimum

$$\omega_i = \min\{|T_i(L)|, |T_i(R)|\}$$

node balance at node ν_i. In particular, ω_1 is the *root balance*.

It is a standard exercise to calculate the probability that T and T' have the same root (ν_1). Given T and fixing ω_1, one has

$$\mathrm{Prob}(\nu'_1 = \nu_1 \,|\, T, \omega_1) = \sum_{i=1}^{n'-1} \frac{\binom{\omega_1}{i}\binom{n-\omega_1}{n'-i}}{\binom{n}{n'}} = 1 - \frac{\binom{\omega_1}{n'} + \binom{n-\omega_1}{n'}}{\binom{n}{n'}}\,.$$

When n is large, one may replace the hypergeometric terms by binomials and get

$$\mathrm{Prob}(\nu'_1 = \nu_1 \,|\, T, \omega_1) \approx \sum_{i=1}^{n'-1} \binom{n'}{i} p^i (1-p)^{n'-i} = 1 - (1-p)^{n'} - p^{n'}\,, \tag{1.4.1}$$

where $p = \omega_1/n$, $0 < p \leqslant 1/2$. For trees generated by the Yule process, node balance is (nearly) uniformly distributed on $1, \ldots, \lfloor n/2 \rfloor$, hence p is uniform on $]0, 1/2[$. Integrating Eq. (1.4.1) with respect to p and multiplying with uniform weights, one obtains the well known result (cf. [42])

$$\mathrm{Prob}(\nu'_1 = \nu_1) \approx 2 \int_0^{1/2} \left(1 - (1-p)^{n'} - p^{n'}\right) \mathrm{d}p = \frac{n'-1}{n'+1}\,.$$

We now consider node balance in induced subtrees. Let the random variable Ω_1 be root balance in a Yule tree of size n. One has

$$\mathrm{Prob}(\Omega_1 = \omega_1) = \frac{2 - \delta_{\omega_1, n/2}}{n-1}\,.$$

Fixing T and selecting an arbitrary induced subtree $T' \lhd T$, consider the random variable $\Omega'_1 \mid \Omega_1$. To calculate the conditional distribution, one may use the auxiliary terms

$$p(\omega'_1 \mid \omega_1) \approx \mathrm{Prob}(\upsilon_1 = \upsilon'_1) \cdot \left(\frac{\binom{\omega_1}{\omega'_1}\binom{n-\omega_1}{n'-\omega'_1} + \binom{n-\omega_1}{\omega'_1}\binom{\omega_1}{n'-\omega'_1}}{\binom{n}{n'} - \binom{\omega_1}{n'} - \binom{n-\omega_1}{n'}} \right) \left(\frac{1}{1+\delta_{\omega'_1,n'/2}} \right) + \mathrm{Prob}(\upsilon_1 \neq \upsilon'_1) \cdot \left(\frac{2-\delta_{\omega'_1,n'/2}}{n'-1} \right),$$

assuming that the induced subtree T' is a random tree of size n' when roots of T and T' are different. Normalising, one obtains

$$\mathrm{Prob}(\omega'_1 \mid \omega_1) = \left(\sum_{\omega'_1=1}^{\lfloor n'/2 \rfloor} p(\omega'_1 \mid \omega_1) \right)^{-1} \cdot p(\omega'_1 \mid \omega_1). \tag{1.4.2}$$

Different roots, and the ensuing 'approximation', are likely to occur when ω_1 is small. Analytical, however lengthy, expressions of the conditional expectation $E(\Omega'_1 \mid \Omega_1)$ are then easily derived with software for symbolic algebra.

This computation can be extended to the balance Ω_2 of the root of the largest root subtree, to obtain the conditional expectation of $\Omega'_2 \mid (\Omega_1, \Omega_2)$ and of $\Omega'_2 \mid \Omega_2$ (Disanto and Wiehe, unpublished results). In Fig. 1.4.1, we show $E(\Omega'_1 \mid \omega_1)$ and $E(\Omega'_2 \mid \omega_2)$ as functions of ω_1 and ω_2 and compare them to simulated values. Shown are averages across arbitrary trees of fixed size n and arbitrary induced subtrees of fixed size n'. Note that induced subtrees, when conditioned on a fixed supertree, reflect node balance of the supertree only when the latter is not extremal. In principle, these calculations could be continued to further internal nodes. However, a full probabilistic treatment and the involved expressions become very clumsy.

Application: Neutrality test using node balance

Tree balance statistics [12, 29, 5] have traditionally been used to investigate evolutionary hypotheses in the context of phylogenetic species trees. However, they can also be defined and examined for gene genealogies modelled by the coalescent process and be integrated into powerful tests of the neutral evolution hypothesis [32, 33, 22]. Published versions of such tests, however, are typically a mixture of tree shape and branch length statistics. Relying, in contrast, only on node balance, one may define the statistic (cf. [33])

$$\mathcal{T}_3 = 2 \sum_{i=1}^{3} \left(2\tfrac{\Omega_i}{n_i} - \tfrac{1}{2} \right), \tag{1.4.3}$$

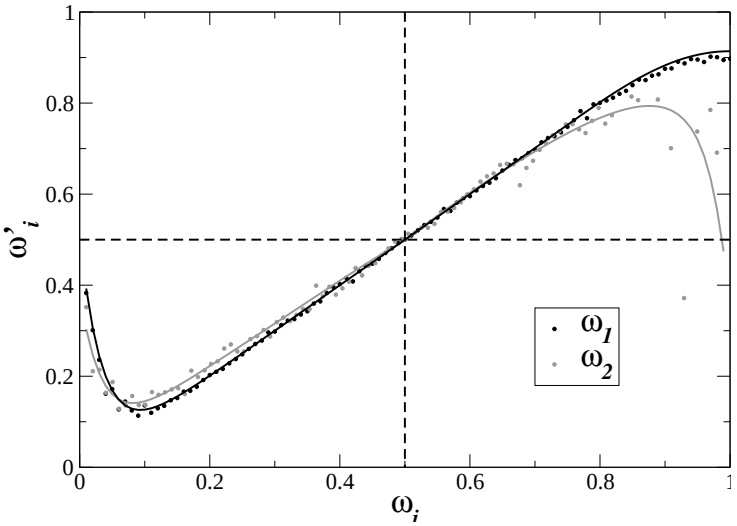


FIGURE 1.4.1. Standardised (i.e., scaled to $[0, 1]$) values of $\mathbb{E}(\omega_1' \mid \omega_1)$ (black) and $\mathbb{E}(\omega_2' \mid \omega_2)$ (grey) for $n = 200$ and $n' = 50$. Theoretical results (solid lines) according to Eq. (1.4.2) and simulation results (dots), obtained with `ms` [26].

where $n_1 = n$, $n_2 = n - \Omega_1$ and $n_3 = n - \Omega_1 - \Omega_2$. Since $2\Omega_i/n_i$ is approximately uniform on the interval $[2/n_i, 1]$, $\mathcal{T}_3$ is close to standard normal [33]. Small values of $\mathcal{T}_3$ are obtained for highly unbalanced trees, i.e., when ω_i are small, produced for instance by caterpillars, and large values for highly balanced trees. In the context of population genetics, a locally unbalanced genealogy of a sample of n genes can be produced by the rapid fixation of a favourable allele. Hence, an estimate of $\mathcal{T}_3$, based on observed genetic variability, provides a statistic with which the hypothesis of neutral evolution can be tested. The results on induced subtrees can be integrated into a nested test-strategy where samples and sub-samples are tested jointly. More details are described in [39].

1.5. Transformations I: Pruning, grafting and recombination

Let $T \in \Lambda_n^{+-}$ be a ranked tree. The *layer* l_j $(1 \leqslant j \leqslant n)$ of T is the 'interval' in which T has j branches. Layer l_1 can be imagined as the infinitely long layer above the root, which makes T a *planted* tree [17, p.6]. An internal node ν_j $(1 \leqslant j < n)$ marks the border between layers l_j and l_{j+1} and layers subdivide any branch b between two nodes into branch *segments* $s(b)_1, \ldots, s(b)_k$, where k depends on b. The *size* of a branch is the number of leaves below the branch. By extension, the size of a segment is the size of the branch to which the segment belongs. A tree T may be transformed into another tree $\tilde{T}$ by a prune and re-graft operation: (i) randomly select branch segments s_p in layer l_p for pruning and s_g

in layer l_g for re-grafting, such that $l_g \leqslant l_p$; (ii) prune the subtree spanned by s_p and re-graft it to segment s_g. This prune and re-graft operation is a model of genetic recombination. Recombination can also be thought of as a segmentation process, which subdivides a linear chromosome into (genomic) segments, such that all sites within one segment have the same genealogical history, or ranked tree; see the contributions of Baake and Baake [3], Birkner and Blath [4] and Dutheil [19] in this volume. Here, we ask two questions: (i) what is the probability that recombination changes the root of the tree and (ii) how is root balance affected by recombination? First note that only some recombination events affect tree topology. One way to change the root is by a re-graft operation to a segment in layer l_1 above the root. Such events may also change root balance ω_1. Re-grafting below the root may change root height or balance only if s_p and s_g belong to different root subtrees. On average, this happens with probability one third (see below).

So far, we ignored branch lengths, but for applications in population genetics it is of interest to assign branch lengths according to the coalescent process: the length of each layer l_j $(j > 1)$ is scaled by a factor proportional to $1/\binom{j}{2}$. Let $\tilde{P}_\uparrow(i)$ be the probability that a pruned branch in such a coalescent tree has size i and that re-grafting is above the current root, i.e., tree height increases. Averaging over coalescent trees of size n, this probability is [21]

$$\tilde{P}_\uparrow(i) = \frac{2}{a_n} \sum_{k=2}^{n} P_{n,k}(i) \frac{1}{k(k-1)(k+1)}\,, \tag{1.5.1}$$

where a_n is the n-th harmonic number and

$$P_{n,k}(i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}$$

is the probability that a branch of layer k has size i. Since re-grafting is above the root, one of the root-subtrees will have size i after re-grafting and Ω_1 will take the value $\omega_1 = \min(i, n-i)$ with probability

$$P_\uparrow(\omega_1) = \frac{\tilde{P}_\uparrow(\omega_1) + \tilde{P}_\uparrow(n-\omega_1)}{(1+\delta_{2\omega_1,n})}\,.$$

Similarly, one can also obtain the transition probabilities from ω_1^0 before recombination to ω_1 after recombination when tree height is increasing. Let $P_{n,j}(i \mid \omega_0)$ be the probability that a branch at level j has size i in a tree of total size n, given that the size of the root branches are ω_1^0 and $n - \omega_1^0$. Then [21],

$$\tilde{P}_\uparrow(i \mid \omega_1^0) = \frac{2}{a_n} \sum_{j=2}^{n} P_{n,j}(i \mid \omega_1^0) \frac{1}{j(j-1)(j+1)}$$

and

$$P_{\uparrow}(\omega_1 \mid \omega_1^0) = \frac{\tilde{P}_{\uparrow}(\omega_1 \mid \omega_1^0) + \tilde{P}_{\uparrow}(n-\omega_1 \mid \omega_1^0)}{1+\delta_{2\omega_1,n}}\,.$$

Similar calculations lead also to the transition probabilities of root balance under recombination events that do not change tree height, and to estimates of the 'correlation length' of root balance under multiple recombination events. These help to explore the speed with which genealogical trees and shapes change along a recombining chromosome. Considering only recombination events that change root height, we estimated the physical distance between such recombination events as [21, Eq. (51)]

$$\frac{1}{2(10-\pi^2)\rho} \sim \frac{3.83}{\rho}\,,$$

where ρ is the scaled recombination rate per nucleotide site. In other words, about every 4th recombination event affects tree height. For example, if $\rho = 10^{-3}$, the genomic distance between such events is about $4,000$ nucleotides. Recombination events that affect root balance are slightly more common, since more branches are available for re-grafting. The distance between such events can be estimated by the average run-length $(1 - P(\omega_1 \mid \omega_1))^{-1}$, i.e., the average size of a genomic fragment within which root balance ω_1 does not change. The run-length depends on n, is longer for more unbalanced trees (small ω_1) and is on the order of a few recombination events (about 2 to 6, for a typical sample size of $n = 100$) [21].

Linkage disequilibrium

Change in tree topology along a recombining chromosome can also be interpreted as a reduction of *linkage disequilibrium*. Two-loci linkage disequilibrium, LD, is the non-random association of two alleles (genetic variants) from two linked genetic loci or sites (alleles A, a at the first locus and alleles B, b at the second, say). Let X_A (X_B) be the indicator variable of allele A (B). Then, one standard way to express LD is by Pearson's correlation coefficient (e.g. [54]) of the indicator variables

$$r^2 = \frac{\mathrm{Cov}^2(X_A, X_B)}{\mathbb{V}(X_A)\mathbb{V}(X_B)}\,.$$

Alleles A and B are often interpreted as being *derived* from their ancestral forms a and b, respectively, by two independent mutation events that occurred some time ago in the genealogical history of each locus, i.e., by events that 'fall on' some branches of their genealogical trees. As such, a mutation event can be thought of as a 'subtree marker', marking the subtree below the branch on which it occurred. Thus, the frequency of the new mutation in the current population(-sample) is identical to the size of the marked subtree. Focusing on this property, one arrives at a slightly modified concept of linkage disequilibrium [51]: considering two, not necessarily adjacent, genomic segments, S and U, with *labelled* ranked trees $T(S)$

and $T(U)$, and left root subtrees $T(S)_L$ and $T(U)_L$, the leaf labels can be partitioned into four sets: (i) labels that belong to both left subtrees, (ii) both right subtrees, (iii) to either the left subtree in segment S and right subtree in segment U, or (iv) vice versa. With the indicator variables $X_{T(S)_L}$ and $X_{T(U)_L}$ one can calculate r^2 in exactly the same way as before and formulate

Definition 1.5.1. The quantity

$$r^2_{S,U} = \frac{\mathrm{Cov}^2(X_{T(S)_L}, X_{T(U)_L})}{\mathbb{V}(X_{T(S)_L})\mathbb{V}(X_{T(U)_L})}$$

is called *topological linkage disequilibrium* (*tLD*) of the segments S and U.

Here, a segment takes the role of a gene locus, and left/right take the roles of two alleles. The assignment of left and right is arbitrary, as much as the naming of two alleles in the context of conventional *LD*, and does not affect the value $r^2_{S,U}$. Let S_L, S_R, U_L and U_R denote the leaf labels in the left and right root subtrees at segments S and U. Note that $r^2_{S,U} = 1$, if and only if $S_L = U_L$ or $S_L = U_R$. This implies that subtrees are not only identical in size but also contain identically labelled leaves at both segments.

In contrast to conventional *LD*, a configuration of complete topological linkage, $r^2_{S,U} = 1$, can be broken only by recombination events that do change tree topology. Since only about every third recombination event changes tree topology, average decay of *tLD* with distance between segments is slower than decay of conventional *LD* [51]. A simple argument is the following: consider a pruning and a re-grafting event and the relative size p of the left root subtree. The probability that both events take place on opposite sides of the tree, i.e., on different root subtrees, is $2p(1-p)$. Integrating with uniform density over all left subtree sizes yields $\int_{p=0}^{1} 2p(1-p)\mathrm{d}p = 1/3$. Furthermore, *tLD* has an about 3-times higher signal-to-noise ratio (the inverse of the coefficient of variation) than conventional *LD* [51]. The limit of expected *tLD* at large distances between segments is

$$\lim_{\rho\to\infty} \mathbb{E}\big(r^2_{S,U}(\rho)\big) = \frac{1}{n-1}\,,$$

which is in agreement with a classical result by Haldane [24].

Generally, compared to conventional *LD*, *tLD* shows a sharper contrast among genomic regions that are in low versus high linkage disequilibrium. This is a welcome property when searching in whole genome scans for signatures of potential gene-gene interactions using patterns of linkage disequilibrium [51].

1.6. Transformations II: Pruning, grafting and evolving trees

1.6.1. The evolving Moran genealogy

The Yule process is a pure birth process. Augmented by a death process, such that each split of a leaf is compensated by removal of a uniformly chosen leaf and its parental branch, size $n < \infty$ remains constant in time and the Yule process becomes a *Moran process*. Following the Moran process over time τ naturally leads to the *evolving Moran genealogy* $(\mathrm{EMG}_\tau)_{\tau \geqslant 0}$ (see the contribution of Kersting and Wakolbinger [27] for a related class of evolving genealogies). Conversely, for any time $\tau = \tau *$, a tree $T(\tau *)$ of size n can be extracted from the sequence $(\mathrm{EMG}_\tau)_\tau$. In the following, we consider ordered, rather than un-ordered, trees and keep track of left/right when choosing a leaf for splitting.

The evolving Moran genealogy, *EMG* for short, induces a discrete Markov process on the set $\overset{\circ}{\Lambda}{}_n^{+-}$. This process is recurrent and aperiodic [52] and therefore has a stationary distribution P^* on $\overset{\circ}{\Lambda}{}_n^{+-}$. Since we may interpret the genealogy $T(\tau)$ for any given τ as a result of a Yule process, and since all T are uniformly distributed, P^* must be the uniform distribution as well, i.e., $P^*(T) = 1/(n-1)!$ (see Table 1.2.1).

Following the process of tree balance in an *EMG*, let $|T(\tau)_L|$ be the size of the left root subtree of $T(\tau)$ extracted from $(\mathrm{EMG}_\tau)_\tau$. The sequence $(|T(\tau)_L|)_\tau$ is subject to the same transition law as the frequency of a newly arising allele in a Moran model. A new allele arising at time τ^* can be imagined as 'marking' an external branch of $T(\tau^*)$ and the evolving subtree under this branch in $(\mathrm{EMG}_\tau)_{\tau > \tau^*}$. Only at the boundary, there is an exception: whenever the left (or right) root subtree is of size 1, this remaining branch may be killed with positive probability. This leads to loss or fixation of the allele and consequently to a *root jump* with a uniform 'entrance' law. After a root jump the new left root subtree has uniformly distributed size, and not necessarily size 1. We call the time interval between successive root jumps an *episode* of the evolving Moran process.

Result 1.6.1. For $2 \leqslant |T(\tau)_L| \leqslant n-2$, the transition probability of the tree balance process $(|T(\tau)_L|)_\tau$ is given by [52]

$$\mathrm{Prob}\Big(|T(\tau+1)_L| = \omega \mid |T(\tau)_L|\Big) = \begin{cases} \frac{|T(\tau)_L|(n-|T(\tau)_L|)}{n^2}, & \omega = |T(\tau)_L| + 1\,, \\ \frac{|T(\tau)_L|^2 + (n-|T(\tau)_L|)^2}{n^2}, & \omega = |T(\tau)_L|\,, \\ \frac{|T(\tau)_L|(n-|T(\tau)_L|)}{n^2}, & \omega = |T(\tau)_L| - 1\,. \end{cases}$$

At the boundary $|T(\tau)_L| = 1$, one has

$$\mathrm{Prob}\left(|T(\tau+1)_L| = \omega\right) = \begin{cases} \frac{1}{n}, & \omega = 2\,, \\ \frac{(n-1)^2+2}{n^2}, & \omega = 1\,, \\ \frac{1}{n^2}, & \text{otherwise}\,, \end{cases}$$

and at the boundary $|T(\tau)_L| = n-1$, one has

$$\mathrm{Prob}\left(|T(\tau+1)_L| = \omega\right) = \begin{cases} \frac{1}{n}, & \omega = n-2\,, \\ \frac{(n-1)^2+2}{n^2}, & \omega = n-1\,, \\ \frac{1}{n^2}, & \text{otherwise}\,. \end{cases}$$

The result is proved by simple enumeration of the discretely many admissible events and calculation of their probabilities.

There is an alternative procedure of constructing an ordered ranked tree of size n: by random grafting of a new external branch onto any branch segment of an existing tree of size $n-1$. Random grafts onto existing segments can take place in two orientations, left- and right-oriented. Both constructions are equivalent and yield identical distributions. More precisely, we state the following result.

Result 1.6.2. The distributions of ordered Yule trees of size n, and those generated by successive random graftings are identical. Thus, for $T \in \overset{\circ}{\Lambda}{}_n^{+-}$ generated by successive random graft operations from trees of size $n-1, n-2, \ldots$, one has

$$\mathrm{Prob}(T) = \frac{1}{(n-1)!}$$

The proof goes by induction on tree size and using a Lemma derived in [52].

1.6.2. Time reversal of the *EMG*

The Moran process can be imagined as a forward-in-time processes. Reversing time, and starting with a planted tree $T \in \overset{\circ}{\Lambda}{}_n^{+-}$ of size n, consider now the following *merge-graft* operation that generates a tree $T' \in \overset{\circ}{\Lambda}{}_n^{+-}$: (i) including the branch segment parental to the root, there are in total $\binom{n+1}{2}$ segments in T; choose one branch segment $s*$; this is done with probability $1/n^2$ for segments ending in a leaf (n possibilities) and with probability $2/n^2$ for all other segments ($\binom{n}{2}$ possibilities); (ii) if a leaf segment was chosen then assign $T' \leftarrow T$. Otherwise, choose an orientation $\chi \in \{L, R\}$ (left/right) with equal probability, remove the n-th layer from T, re-graft a new branch in orientation χ at $s*$, and update the labels of all nodes. The resulting tree is returned as T' (see Figure 1.6.1).

Iterating the merge-graft operation one obtains a backward-in-time process that is dual to the Moran process. We call the genealogy generated by this process

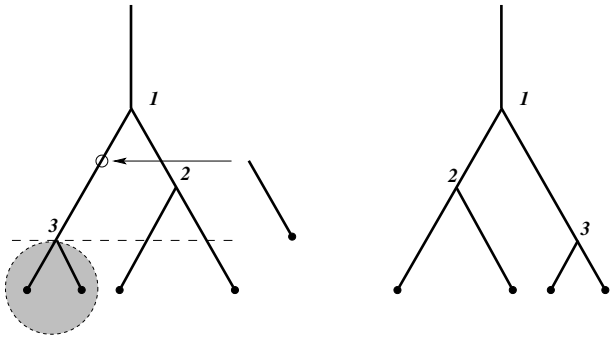


FIGURE 1.6.1. Tree transformation by a merge-graft operation on the tree shown left. Removing the lowest layer (below the dashed line) leads to removal of the shaded cherry. A new branch is grafted ('resurrected') on a segment marked by the open circle. Labels are updated, resulting in the tree shown on the right. Both trees belong to $\mathring{\Lambda}_4^{+-}$.

the *evolving Moran genealogy backward in time*, $EMG^\flat$ for short, and note [52] the following result.

Result 1.6.3. For all $T, T' \in \mathring{\Lambda}_n^{+-}$:

(1.6.1)

$$\mathrm{Prob}_{EMG}\big(T(\tau+1)=T' \mid T(\tau)=T\big) = \mathrm{Prob}_{EMG^\flat}\big(T(\tau)=T \mid T(\tau+1)=T'\big)\,,$$

with Prob_{EMG} ($\mathrm{Prob}_{EMG^\flat}$) denoting the transition probability of the *EMG*- ($EMG^\flat$-) process, respectively.

1.6.3. The root jump process

The $EMG^\flat$ is interesting theoretically as well as practically. While transitions in the *EMG* depend on two random events, splitting and killing, in the $EMG^\flat$ there is only one random operation, grafting. This fact simplifies some analytic approaches. Consider the root jump process. An obvious question to ask is how often do root jumps occur? Working in the framework of the $EMG^\flat$, one can immediately state the following.

Result 1.6.4. Root jumps in the *EMG* and in the $EMG^\flat$ occur according to a geometric jump process of intensity $\frac{2}{n^2}$.

PROOF. In the $EMG^\flat$, a root jump occurs if and only if the segment parental to the root is chosen for re-grafting. This happens with probability $\frac{2}{n^2}$. The same holds for the forward process due to duality. □

This result agrees with the one derived in [37], where the jump process in the infinite-population limit is identified as a Poisson process of intensity 1. This is the limit of the geometric jump process as $n \to \infty$ with time sped up by $n^2/2$, which is the average number of Moran steps before a root jump occurs. Also implied by Result 1.6.4, the number of steps needed to observe any number $k > 0$ of root jumps follows a negative binomial distribution with parameters k and $2/n^2$.

Finally, using the simple structure of the $EMG^\flat$, one can calculate the number of root jumps during fixation of a new allele. More precisely, consider time τ_0 when a new allele x^* is born (a subtree marker on some branch of T) and — conditional on fixation — time τ_1 when x^* becomes fixed, i.e., when all leaves of $T(\tau_1)$ are descendants of x^*. We have the following result.

Result 1.6.5. In an *EMG* of size $n \geqslant 2$, one expects $2(1 - \frac{1}{n})$ root jumps during the time interval $[\tau_0, \tau_1]$.

The proof goes by considering events in the backward process, where one finds that the expected total number of root jumps along the $EMG^\flat$-path is

$$\sum_{k=2}^{n-1} \frac{2}{k(k+1)} = \frac{n-2}{n}.$$

Adding one additional jump, which necessarily happens at the moment of fixation, one obtains the stated expectation. Hence, in an infinitely large sample ($n = \infty$) one expects two root jumps per one fixation, a result obtained with different means before [37].

In the framework of the $EMG^\flat$ one can calculate the exact distribution of root jumps during a fixation recursively for any n, and show that these distributions quickly converge as $n \to \infty$. For $n \geqslant 2$, let $\mathrm{Prob}_n(k)$ denote the probability of observing k root jumps during fixation of a new allele in an *EMG* of size n, and $\mathrm{Prob}_\infty(k)$ the same probability in the infinite-population limit. Then,

$$\mathrm{Prob}_n(k) = \sum_{2 \leqslant i_1, \ldots, i_{k-1} \leqslant n-1} \prod_1^k \frac{2}{i_k(i_k+1)} \prod_{j \neq i_1, \ldots, i_{k-1}} \left(1 - \frac{2}{j(j+1)}\right).$$

For small k, using software for symbolic algebra, one can easily write down closed-form expressions for $\mathrm{Prob}_n(k)$. For $n = 2$, $\mathrm{Prob}_2(1) = 1$. $\mathrm{Prob}_n(1)$ decreases monotonically in n, with $\lim_{n\to\infty} \mathrm{Prob}_n(1) = 1/3$. In Figure 1.6.2 root jump distributions are shown for some small values of n and for $n = \infty$, illustrating the fast convergence for $n \to \infty$.

Any root jump is tantamount to loss of some 'genetic memory'. In the future, it will be interesting to explore the root jump process in more detail, in particular under non-equilibrium and non-neutral population genetic scenarios, and with regard to the speed of loss of genetic memory.

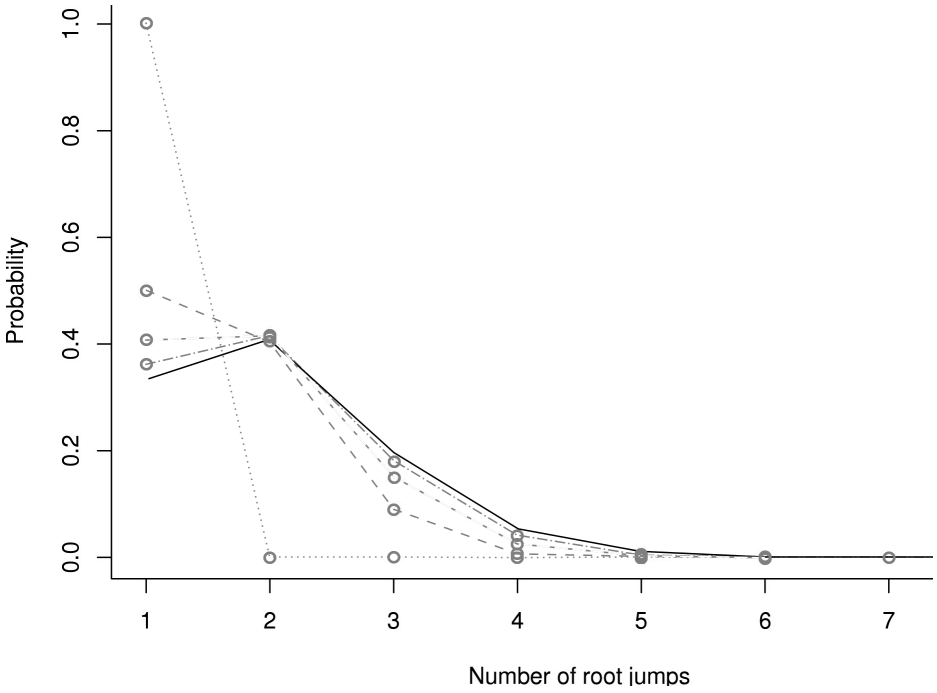


FIGURE 1.6.2. The distributions of $P_n(k)$, $k = 1, \ldots, 7$; $n = 2$ (dotted), $n = 5$ (dashed), $n = 10$ (short dashes), $n = 25$ (dot-dashed) and $n = \infty$ (black).

ACKNOWLEDGEMENTS

I would like to express my gratitude to Filippo Disanto and Johannes Wirtz for their intellectual input to the projects pursued as part of SPP 1590. I am very grateful also to Luca Ferretti for his enthusiastic discussions, sharing of ideas and his contributions to tree transformations under recombination. Finally, I would like to thank two reviewers for their critical and constructive comments on an earlier version of this chapter.

References

[1] D.J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Stat. Sci.* **16** (2001), 23–34.

[2] D. André, Mémoire sur les permutations alternées, *J. Math. Pures Appl., 3^e série*, **7** (1881), 167–184.

[3] E. Baake and M. Baake, Ancestral lines under recombination, *this volume*.

[4] M. Birkner and J. Blath, Genalogies and inference for populations with highly skewed offspring distributions, *this volume*.

[5] M.G.B. Blum and O. François, On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited, *Math. Biosci.* **195** (2005), 141–153.

[6] N. Bortolussi, E. Durand, M.G.B. Blum, and O. François, apTreeshape: statistical analysis of phylogenetic tree shape, *Bioinformatics* **22** (2006), 363–364.

[7] N.G. de Bruijn and D.A. Klarner, Multisets of aperiodic cycles, *SIAM J. Discr. Math.* **3** (1982), 359–368.

[8] L.L. Cavalli-Sforza and A.W. Edwards, Phylogenetic analysis. Models and estimation procedures, *Am. J. Hum. Genet.* **19** (1967), 233–257.

[9] A. Cayley, XXVIII. On the theory of the analytical forms called trees, *Lond. Edinb. Dubl. Phil. Mag.* **13** (1857), 172–176.

[10] A. Cayley, A theorem on trees, *Quart. J. Math.* **23** (1889), 376–378.

[11] H. Chang and M. Fuchs, Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.* **60** (2010), 481–512.

[12] D. H. Colless, Review of 'Phylogenetics: The Theory and Practice of Phylogenetic Systematics', by E.O. Wiley, *Syst. Zool.* **31** (1982), 100–104.

[13] J.H. Degnan, N.A. Rosenberg, and T. Stadler, The probability distribution of ranked gene trees on a species tree, *Math. Biosci.* **235** (2012), 45–55.

[14] F. Disanto and T. Wiehe, Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model, *Math. Biosci.* **242** (2013), 195–200.

[15] F. Disanto and N.A. Rosenberg, Enumeration of ancestral configurations for matching gene trees and species trees, *J. Comput. Biol.* **24** (2017), 831–850.

[16] R. Donaghey, Alternating permutations and binary increasing trees, *J. Combin. Theory A* **18** (1975), 141–148.

[17] M. Drmota, *Random Trees: An Interplay Between Combinatorics and Probability*, Springer, Wien, 2009.

[18] R. Durrett, Probability Models for DNA Sequence Evolution, 2nd ed., Springer, New York, 2008.

[19] J.Y. Dutheil, Towards more realistic models of genomes in populations: the Markov-modulated sequentially Markov coalescent, *this volume*.

[20] J. Felsenstein, The number of evolutionary trees, *Syst. Zool.* **27** (1978), 27–33.

[21] L. Ferretti, F. Disanto and T. Wiehe, The effect of single recombination events on coalescent tree height and shape, *PLoS One* **8** (2013), e60123: 15p.

[22] L. Ferretti, A. Ledda, T. Wiehe, G. Achaz, and S.E. Ramos-Onsins, Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests, *Genetics* **207** (2017), 229–240.

[23] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2009.

[24] J.B.S. Haldane, The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small, *Biometrika* **31** (1940), 346–355.

[25] E.F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation, *Adv. Appl. Prob.* **3** (1971), 44–77.

[26] R.R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation, *Bioinformatics* **18** (2002), 337–338.

[27] G. Kersting and A. Wakolbinger, Probabilistic aspects of Λ-coalescents in equilibrium and in evolution, *this volume*.

[28] J.F.C. Kingman, The coalescent, *Stochastic Processes Appl.* **13** (1982), 235–248.

[29] M. Kirkpatrick and M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, *Evolution* **47** (1993), 1171–1181.

[30] D. Knuth, *The Art of Computer Programming, Vol. 1 (Fundamental Algorithms)*, 3rd ed., Addison-Wesley, Boston, 2004.

[31] A. Lambert and T. Stadler, Birth-death models and coalescent point processes: the shape and probability of reconstructed phylogenies, *Theor. Popul. Biol.* **90** (2013), 113–128.

[32] H. Li, A new test for detecting recent positive selection that is free from the confounding impacts of demography, *Mol. Biol. Evol.* **28** (2011), 365–375.

[33] H. Li and T. Wiehe, Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation, *PLoS Comp. Biol.* **9** (2013), e1003060: 14p.

[34] W.P. Maddison and M. Slatkin, Null models for the number of evolutionary steps in a character on a phylogenetic tree, *Evolution* **45** (1991), 1184–1197.

[35] A. McKenzie and M. Steel, Distributions of cherries for two models of trees, *Math. Biosci.* **164** (2000), 81–92.

[36] A.Ø. Mooers and S.B. Heard, Inferring evolutionary process from phylogenetic tree shape, *Q. Rev. Biol.* **72** (1997), 31–54.

[37] P. Pfaffelhuber and A. Wakolbinger, The process of most recent common ancestors in an evolving coalescent, *Stochastic Processes Appl.* **116** (2006), 1836–1859.

[38] G. Pólya, Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen, *Acta Mathem.* **68** (1937), 145–254.

[39] M. Rauscher, *Topology of genealogical trees — theory and applications*, Doctoral thesis, University of Cologne, `urn:nbn:de:hbz:38-90303`, 2018.

[40] N.A. Rosenberg, The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees, *Ann. Comb.* **10** (2006), 129–146.

[41] N.A. Rosenberg, Counting coalescent histories, *J. Comput. Biol.* **14** (2007), 360–377.

[42] I.W. Saunders, S. Tavaré, and G.A. Watterson, On the genealogy of nested subsamples from a haploid population, *Adv. Appl. Prob.* **16** (1984), 471–491.

[43] N.J.A. Sloane, *The Online Encyclopedia of Integer Sequences*, available online at `https://oeis.org`

[44] J.B. Slowinski and C. Guyer, Testing the stochasticity of patterns of organismal diversity: an improved null model, *Amer. Nat.* **134** (1989), 907–921.

[45] Y.S. Song, Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees, *Ann. Comb.* **10** (2006), 147–163.

[46] E. Stam, Does imbalance in phylogenetics reflect only bias? *Evolution* **56** (2002), 1292–1295.

[47] M. Steel and A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, *Math. Biosci.* **170** (2001), 91–112.

[48] F. Tajima, Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105** (1983), 437–460.

[49] J. Wakeley, *Coalescent Theory — an Introduction*, Roberts, Greenwood Village, CO, 2009.

[50] J.H.M. Wedderburn, The functional equation $g(x^2) = 2\alpha x + [g(x)]^2$, *Ann. Math.* **24** (1922), 121–140.

[51] J. Wirtz, M. Rauscher, and T. Wiehe, Topological linkage disequilibrium calculated from coalescent genealogies, *Theor. Popul. Biol.* **124** (2018), 41–50.

[52] J. Wirtz and T. Wiehe, The evolving Moran genealogy, *Theor. Popul. Biol.* **130** (2019), 94–105.

[53] G.U. Yule, A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S., *Phil. Trans. R. Soc.* **213** (1925), 21–87.

[54] D.V. Zaykin, A. Pudovkin, and B.S. Weir, Correlation-based inference for linkage disequilibrium with multiple alleles, *Genetics* **180** (2008), 533–545.

Index