Some simple varieties of trees arising in permutation analysis

Mathilde Bouvel[1†] Marni Mishna[2‡] Cyril Nicaud[3†]

[1] *LaBRI/CNRS, Université Bordeaux, France*
[2] *Dept. Mathematics, Simon Fraser University, Burnaby, Canada*
[3] *Laboratoire d'Informatique Gaspard Monge (LIGM), Université Paris-Est, Marne-la-Vallée, France*

Abstract After extending classical results on simple varieties of trees to trees counted by their number of leaves, we describe a filtration of the set of permutations based on their strong interval trees. For each subclass we provide asymptotic formulas for number of trees (by leaves), average number of nodes of fixed arity, average subtree size sum, and average number of internal nodes. The filtration is motivated by genome comparison of related species.

Résumé Nous commençons par étendre les résultats classiques sur les variétés simples d'arbres aux arbres comptés selon leur nombre de feuilles, puis nous décrivons une filtration de l'ensemble des permutations qui repose sur leurs arbres des intervalles communs. Pour toute sous-classe, nous donnons des formules asymptotiques pour le nombre d'arbres (comptés selon les feuilles), le nombre moyen de nœuds d'arité fixée, la moyenne de la somme des tailles des sous-arbres, et le nombre moyen de nœuds internes. Cette filtration est motivée par des problématiques de comparaison de génomes.

Keywords: permutations, simple varieties of trees, random generation, tree parameters, asymptotic formulas

This short paper is an extended abstract of [7], where details of the proofs are provided.

1 Introduction

The idea of viewing permutations as enriched trees has been around for several decades in different research communities. For example, the recent enumerative study [1] of pattern avoiding permutations, in which *(substitution) decomposition trees* play a crucial role. Also, the analysis of sorting algorithms is very linked to tree representations of permutations: *PQ trees* [5] appear in the context of graph algorithms and *strong interval trees* arise in comparative genomics [6, and references therein, for instance].

In each case it is of interest to understand the typical shape and structure of the trees that arise. For example, a cursory examination of permutations that arise in the comparison of mammalian genomes strongly suggests that not all permutations are equally likely, and in fact this is quite an understatement. Trees coming from permutations under the uniform distribution are somehow degenerate [6], and do not

†This work was partially supported by ANR project MAGNUM (2010-BLAN-0204).
‡This work was also supported in part by NSERC Discovery grant 31-611453, and funding by Université Paris-Est.

adequately represent the trees that arise in genomic comparisons. This has important consequences for algorithm analysis. Specifically, in [6], Bouvel *et al.* considered a subclass of strong interval trees – selected because they represent what is known as *commuting scenarios* [3]– that correspond to the class of *separable permutations*. This is a first step towards a more relevant model of permutations which arise in genome comparison. By studying asymptotic enumeration and parameter formulas for separable permutations, they proved that the complexity of the algorithm of [3] solving the *perfect sorting by reversals* problem is polynomial time on separable permutations, whereas this problem is NP-complete in general. Furthermore they were also able to describe some average-case properties of the perfect sorting scenarios for separable permutations.

Ultimately, a clear understanding of the properties possessed by the strong interval trees that represent the comparison of actual genomes might tell us something about the evolutionary process. Bouvel *et al.* [6] conclude their study on separable permutations with a suggestion for the next step: strong interval trees with degree restrictions on certain internal nodes. It is a very controlled way to introduce bias in the distribution of strong interval trees. This is precisely what we do in this work; namely, we study strong interval trees where the prime nodes have a bounded number of children. This is a class of trees that can be completely understood combinatorially and analytically, and so we have immediate access to enumeration and analysis of some tree parameters that are ultimately related to the complexity of computing perfect sorting scenarios, or to properties of these scenarios.

In this work, we focus on the combinatorial analysis of these restricted sets of trees. This study reveals a very lush substructure of permutations that is certainly of independent interest. We define nested simple varieties of trees whose limit is the set of all strong interval trees, recalling they form a class in a size preserving bijection with permutations. The components are families of trees, hence we are able to apply a very complete set of tools to all the components: asymptotic analysis, random generation– these tools are inaccessible to the full class without working through permutations. Thus, we decompose a transcendental and non-analytic class into neat, algebraic portions, each of which is easily understood.

The organization of this abstract is as follows: First, in Section 2 we present some very general theorems for asymptotic enumeration and parameter analysis that are widely applicable. Then in Section 3 we describe strong interval trees as a decomposable combinatorial class. Finally, we describe the class of prime-degree restricted trees in Section 4, and give tight bounds on values which control the asymptotic enumeration and the tree parameters.

2 When the size of a tree is the number of leaves

There are many works which consider the study of average case parameters of trees where the size is the number of internal nodes or of both internal nodes and leaves. The generating functions of these trees satisfy a functional equation of the form $T(z) = z \cdot \Phi(T(z))$, and when Φ satisfies certain conditions, such as analyticity, then there are formulas for inversion, resulting in explicit enumerative results. A class of trees amenable to this treatment is said to be a *simple variety of trees*. The subject is exhaustively treated in Section VII.3 of [10]. If, instead, we define size as the number of leaves, the generating function satisfies a relation of the form $T(z) = z + \Lambda(T(z))$. The same general theorems on inversion still work, and it suffices to apply them and unravel the results. Even though they are less frequent, these have also been well studied in the literature, and the applicability of the inversion lemmas is noted in Example VII.13 of [10]. In this section we do this explicitly.

Asymptotic number of trees with n leaves	$\sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}$
The average number of nodes of arity κ in trees with n leaves	$\frac{\lambda_\kappa \tau^\kappa}{\rho} \cdot n$
The average number of internal nodes in trees with n leaves	$\frac{\Lambda(\tau)}{\rho} \cdot n = \frac{\tau-\rho}{\rho} \cdot n$
The average subtree size sum in trees with n leaves	$\sqrt{\frac{\pi}{2\rho\Lambda''(\tau)}} \cdot n^{3/2}$

Tab. 1: A summary of parameters of trees given by $T = z + \Lambda(T)$. The value τ is the unique solution to $\Lambda'(\tau) = 1$ between 0 and $R_\Lambda < 1$, and $\rho = \tau - \Lambda(\tau)$.

Consider the analytic solutions $T(z)$ of the equation

$$T(z) = z + \Lambda(T(z)), \tag{1}$$

where $\Lambda(z) = \sum_{n\geq 2} \lambda_n z^n$ is analytic with radius of convergence R_Λ, and such that $\lambda_n \geq 0$ for any $n \geq 2$. Furthermore, assume that Λ is not the null function. Let $\Psi(z) := z - \Lambda(z)$. Equation (1) rewrites as $\Psi(T(z)) = z$, so what we are looking for is precisely an analytic inversion of Ψ.

The Table 1 summarizes the results of this section. We determine asymptotic formulas for number of trees, and several key parameters. The shape of the formulas are, unsurprisingly, not unlike those that arise in the study of trees counted by internal nodes.

2.1 Asymptotic number of trees

Our entire analysis is roughly a consequence of the analytic inversion lemma and transfer theorems. The version to which we appeal is given and proved in [10]. Citations to original sources may be found therein. The following theorem is a slight adaptation of Proposition IV.5 and Theorem VI.6 to combinatorial equations of the form $\mathcal{T} = \mathcal{Z} + \Lambda(\mathcal{T})$ instead of $\mathcal{T} = \mathcal{Z} \cdot \Lambda(\mathcal{T})$.

Theorem 1 *Let Λ be a function analytic at 0, with non-negative Taylor coefficients, and such that, near 0,*

$$\Lambda(z) = \sum_{n\geq 2} \lambda_n z^n.$$

Let R_Λ be the radius of convergence of this series. Under the condition $\lim_{x\to R_\Lambda^-} \Lambda'(x) > 1$, there exists a unique solution $\tau \in (0, R_\Lambda)$ of the equation $\Lambda'(\tau) = 1$.

Then, the formal solution $T(z)$ of the equation $T(z) = z + \Lambda(T(z))$ is analytic at 0, its unique dominant singularity is at $\rho = \tau - \Lambda(\tau)$ and its expansion near ρ is

$$T(z) = \tau - \sqrt{\frac{2\rho}{\Lambda''(\tau)}}\sqrt{1 - z/\rho} + O(1 - z/\rho). \tag{2}$$

Moreover, if T is aperiodic, then one has

$$[z^n]T(z) \sim \sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}. \tag{3}$$

2.2 Parameter Analysis

In the case of trees counted by internal nodes, the study of recursively defined parameters is very straightforward, starting from generating function equations. We can describe analogous versions for trees counted by leaves. In particular, we consider additive parameters, and describe a Modified Iteration Lemma, adapted to our notion of size. We illustrate the lemma on number of internal nodes, subtree size sum and number of nodes of a given arity.

Our focus is on tree parameters that can be computed additively by parameters of subtrees. More precisely, given a parameter $\xi(t)$ for trees $t \in \mathcal{T}$ which satisfy the relation

$$\xi(t) = \eta(t) + \sum_{j=1}^{\deg(t)} \sigma(t_j),$$

where $\deg(t)$ is the arity of the root and t_j are its children. Let $\Xi(z)$, $H(z)$ and $\Sigma(z)$ be the associated cumulative functions of ξ, η and σ. That is, $\Xi(z) = \sum_{t\in\mathcal{T}} \xi(t)z^{|t|}$, $H(z) = \sum_{t\in\mathcal{T}} \eta(t)z^{|t|}$ and $\Sigma(z) = \sum_{t\in\mathcal{T}} \sigma(t)z^{|t|}$.

Lemma VII.1 in [10] has an analogue for trees counted by their leaves, and it is proved in a very similar way.

Lemma 2 (Iteration lemma for trees counted by their leaves) *Let $\mathcal{T}$ be a class of trees satisfying $\mathcal{T} = \mathcal{Z} + \Lambda(\mathcal{T})$. The cumulative generating functions are related by*

$$\Xi(z) = H(z) + \Lambda'(T(z))\,\Sigma(z).$$

In particular, if $\sigma \equiv \xi$, one has $\Xi(z) = \frac{H(z)}{1-\Lambda'(T(z))} = H(z) \cdot T'(z)$.

The last equality is a consequence of $T'(z)(1 - \Lambda'(T(z))) = 1$, which is obtained by differentiating $T(z) = z + \Lambda(T(z))$ with respect to z.

We make a remark, that if $\sigma \equiv \xi$, we say the parameter is *recursive*; most basic parameters are recursive, and in what follows we shall use this case only. Note also that when analytic treatment applies, $T(z)$ has a square-root singularity, so that $T'(z)$ has an inverse square-root singularity (by analytic derivation). Therefore, whenever $H(z)$ tends to a positive real when $z \to \rho$ (under some analytic conditions), then transfer yields an asymptotic equivalent of the mean value of the parameter of the form $c \cdot n$. This is for instance the case for the number of nodes of fixed arity and the number of internal nodes.

Number of nodes with exactly κ children We "mark" nodes of arity κ by setting

$$\eta(t) = \begin{cases} 1 & \text{if the root of } t \text{ is of arity } \kappa, \\ 0 & \text{otherwise.} \end{cases}$$

Hence if $\kappa \geq 2$, $H(z) = \sum_{t\in\mathcal{T}} \eta(t)z^{|t|} = \sum_{t_1,\ldots,t_\kappa\in\mathcal{T}} \lambda_\kappa z^{|t_1|+|t_2|+\ldots+|t_\kappa|}$ so that $H(z) = \lambda_\kappa\, T(z)^\kappa$. And[(i)] if $\kappa = 0$, $H(z) = z$ which is not interesting since it is counting the number of leaves *i.e.* the size of the tree.

[(i)] This is the only other possibility since there can be no unary nodes in a proper specification.

By Lemma 2, for any $\kappa \geq 2$ one has $\Xi(z) = \lambda_\kappa T(z)^\kappa \cdot T'(z)$. Since the singular expansion of $T(z)$ near ρ is

$$T(z) = \tau - \gamma\sqrt{1 - z/\rho} + o\left(\sqrt{1 - z/\rho}\right), \text{with } \gamma = \sqrt{\frac{2\rho}{\Lambda''(\tau)}} \tag{4}$$

then near ρ, one has $T(z)^\kappa = \tau^\kappa + O\left(\sqrt{1 - z/\rho}\right)$. Using the singular differentiation theorem we have

$$T'(z) = \frac{\gamma}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right), \text{ so that } \Xi(z) = \frac{\lambda_\kappa \gamma \tau^\kappa}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right),$$

from which we get the asymptotics of the cumulative generating function

$$[z^n]\Xi(z) \sim \frac{\lambda_\kappa \gamma \tau^\kappa \rho^{-n-1}}{2\sqrt{\pi n}}.$$

The asymptotics of the average value across all trees of size n is reported in Table 1.

Number of internal nodes For this parameter, just take the following definition for η:

$$\eta(t) = \begin{cases} 0 & \text{if } t \text{ is just one leaf,} \\ 1 & \text{otherwise.} \end{cases}$$

One has $H(z) = \sum_{t \in \mathfrak{T}} \eta(t) z^{|t|} = T(z) - z$, and therefore (with the γ of Equation (4))

$$\Xi(z) = (T(z) - z)\ T'(z) = \frac{\gamma(\tau - \rho)}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right) \quad \text{and} \quad [z^n]\Xi(z) \sim \frac{\gamma(\tau - \rho)\rho^{-n-1}}{2\sqrt{\pi n}}.$$

Subtree size sum We are interested in the subtree size sum parameter, defined by $\eta(t) = |t|$, hence $H(z) = zT'(z)$. So that

$$\Xi(z) = zT'(z)^2 = \frac{\gamma^2}{4\rho(1 - z/\rho)} + o\left(\frac{1}{1 - z/\rho}\right) \quad \text{and} \quad [z^n]\Xi(z) \sim \frac{\gamma^2}{4\rho} \cdot \rho^{-n}.$$

It is not an inverse of square-root singularity, and we find an asymptotic equivalent in $n^{\frac{3}{2}}$ for the average value of the subtree size sum (see Table 1), which is typical for path length related parameters.

There are many other tree parameters that we could consider in a similar fashion.

3 Strong Interval Trees

Our interest in trees counted by leaves is spawned by *strong interval trees*. They are in a size preserving bijection with permutations. This particular representation of permutations is a very effective data structure for algorithms in reconstruction of genome evolution scenarios, as we briefly mentioned in Section 1. Our analysis builds subclasses that are in fact each a simple variety of trees, and hence are very well understood, particularly given the generic analysis we have completed in Section 2.

3.1 Definition and examples

A description of the bijective correspondence between strong interval trees (sometimes also called (substitution) decomposition trees) and permutations is given in [6]. Truly, it could be viewed as a tree representation of the block decomposition of permutations described by Albert and Atkinson [1], the modular decomposition of permutation graphs of Bérard *et al.* in [4] and even has origins in the PQ-trees of Booth and Lueker [5]. The bijection is completely constructive, and can be computed in linear time, although this is quite difficult to achieve, see [4]. We do not describe the bijection in this work.

The class is a set of trees where some internal nodes are enriched with a simple permutation. A permutation is said to be *simple* if the only intervals $i, i+1, \dots, k$ mapped to an interval are the singletons, and $1, 2, \dots, n$. Because we take the convention that 1 2 and 2 1 are not simple permutations, the shortest ones are of size 4 and are $3\ 1\ 4\ 2$ and $2\ 4\ 1\ 3$. An enumerative study is done by Albert *et al.* [2], and we make use of their asymptotic enumeration formulas. Let s_n be the number of simple permutations of size n. This is sequence A111111 in the On-Line Encyclopedia of Integer Sequences [13]. The sequence is not P-recursive, but it does satisfy a simple functional inversion formula, and we have calculated exact values of for s_n for $n < 800$. Albert *et al.* determined the following bounds:

$$\frac{n!}{e^2}\left(1-\frac{4}{n}\right) \leq s_n \leq \frac{n!}{e^2}\left(1-\frac{4}{n}+\frac{2}{n(n-1)}\right). \tag{5}$$

Here are the first few terms in the generating function for simple permutations:

$$S(z) = 2z^4 + 6z^5 + 46z^6 + 338z^7 + 2926z^8 + 28146z^9 + 298526z^{10} + 3454434z^{11} + \dots \tag{6}$$

Theorem 3 (Reformulated [1]) *The class of permutations is in a size-preserving bijection with the combinatorial class $\mathcal{P}$ of enriched trees defined by the following relations, where size is given by the number of leaves. The class $\mathcal{Z}$ is an atomic class with a single element of size 1, and the $\mathcal{N}$ classes are all epsilon classes containing a single element of size 0, marking internal nodes:*

$$\begin{aligned}
\mathcal{P} &= \mathcal{Z}_\square + \mathcal{N}_\oplus \cdot \mathrm{Seq}_{\geq 2}\, \mathcal{U}_\oplus + \mathcal{N}_\ominus \cdot \mathrm{Seq}_{\geq 2}\, \mathcal{U}_\ominus + \mathcal{N}_\bullet \cdot S(\mathcal{P}),\\
\mathcal{U}_\oplus &= \mathcal{Z}_\square + \mathcal{N}_\ominus \cdot \mathrm{Seq}_{\geq 2}\, \mathcal{U}_\ominus + \mathcal{N}_\bullet \cdot S(\mathcal{P}),\\
\mathcal{U}_\ominus &= \mathcal{Z}_\square + \mathcal{N}_\oplus \cdot \mathrm{Seq}_{\geq 2}\, \mathcal{U}_\oplus + \mathcal{N}_\bullet \cdot S(\mathcal{P}).
\end{aligned} \tag{7}$$

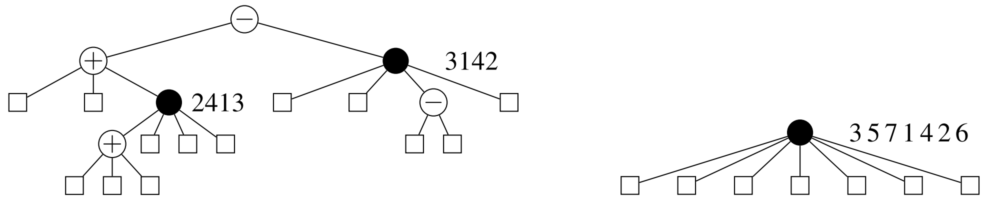
The internal nodes $\mathcal{N}_\bullet$ are called prime nodes *and the internal nodes $\mathcal{N}_\oplus$ and $\mathcal{N}_\ominus$ are called* linear *nodes. The function $S(z)$ is the generating function for simple permutations from Equation* (6)*.*

Figure 1 contains two examples. Figure 1(b) represents a simple permutation. We note that the trees corresponding to simple permutations contain only a single prime node with n children. The root is labeled by the permutation itself.

Notice that $\mathcal{U}_\oplus$ and $\mathcal{U}_\ominus$ define combinatorial classes which are in size-preserving bijection. In the following, in order to deal with one class instead of two, we replace them by the equivalent class $\mathcal{U} = \mathcal{Z}_\square + \mathcal{N}_\circ \cdot \mathrm{Seq}_{\geq 2}\, \mathcal{U} + \mathcal{N}_\bullet \cdot S(\mathcal{P})$. Doing so, we change the labels of the linear nodes having a linear parent (replacing them by $\circ$). This does not affect the enumeration of the class. Indeed, these labels are determined since a linear node and its linear parent have different labels.

Corollary 4 *The following combinatorial equivalences are true:*

$$\mathcal{P} \equiv \mathrm{Seq}_{\geq 1}\, \mathcal{U} \qquad \textit{and} \qquad \mathcal{U} \equiv \mathcal{Z} + \mathrm{Seq}_{\geq 2}\, \mathcal{U} + S(\mathrm{Seq}_{\geq 1}\, \mathcal{U}). \tag{8}$$



(a) $\sigma_1 = 6\ 7\ 9\ 10\ 11\ 13\ 8\ 12\ \ 3\ 1\ 5\ 4\ 2$ (b) $\sigma_2 = 3\ 5\ 7\ 1\ 4\ 2\ 6$

Fig. 1: Two permutations and their associated strong interval trees

Consequently, $\mathcal{U}$ is in bijection with a class of Λ-trees for $\Lambda(x) = \frac{x^2}{1-x} + \sum_{j\geq 4} s_j \left(\frac{x}{1-x}\right)^j$, where s_j is the number of simple permutations of size j.

Proof: This equivalence is derived from Equation (7), the fact that $\mathcal{U} \equiv \mathcal{U}_\oplus \equiv \mathcal{U}_\ominus$, and the intermediary equivalence $\mathcal{P} \equiv \mathcal{U} + \mathrm{Seq}_{\geq 2}\,\mathcal{U}$. □

Now, neither $\mathcal{P}$ nor $\mathcal{U}$ are simple varieties of trees because $S(z)$, and hence $\Lambda(x)$, are not analytic at the origin. In this case, we can, of course, use the bijection to permutations to have access to enumeration and random generation tools. However, we propose a different strategy: generate a sequence of analytic Λ_k such that as formal power series, $\lim_{k\to\infty} \Lambda_k = \Lambda$, and consider the set of Λ_k-trees. Can we describe conditions so that the limit of the asymptotics of the subclasses tends to the asymptotics of the whole class? To which extent are the parameter formulas valid under the limit? The example we have in hand is a particularly instructive one, since the limit is known by other means, and allows us to test the limits of analytic inversion.

3.2 A filtration for permutations

Next we describe the central filtration on the class of trees $\mathcal{P}$. The limit of the filtration is the entire class, and each subclass is a simple variety of trees that is very straightforward to analyze. We define the class $\mathcal{P}^{(k)}$ as follows, where $S^{\leq k}(z) = \sum_{j=4}^{k} s_j z^j$:

$$\mathcal{P}^{(k)} = \mathcal{Z} + 2\,\mathrm{Seq}_{\geq 2}\,\mathcal{U}^{(k)} + S^{\leq k}(\mathcal{P}^{(k)}) \qquad \text{and} \qquad \mathcal{U}^{(k)} = \mathcal{Z} + \mathrm{Seq}_{\geq 2}\,\mathcal{U}^{(k)} + S^{\leq k}(\mathcal{P}^{(k)}). \tag{9}$$

That is, we restrict the degree of the prime nodes. The containment $\mathcal{P}^{(k)} \subset \mathcal{P}^{(k+1)}$ is straightforward, and since $\mathcal{P}_n^{(k)} = \mathcal{P}_n$ when $k \geq n$, we can derive the limit of combinatorial classes $\lim_{k\to\infty} \mathcal{P}^{(k)} = \mathcal{P}$.

Furthermore, by the same manipulations as for the full class, we derive:

$$\mathcal{P}^{(k)} \equiv \mathrm{Seq}_{\geq 1}\,\mathcal{U}^{(k)} \qquad \text{and} \qquad \mathcal{U}^{(k)} \equiv \mathcal{Z} + \mathrm{Seq}_{\geq 2}\,\mathcal{U}^{(k)} + S^{\leq k}(\mathrm{Seq}_{\geq 1}\,\mathcal{U}^{(k)}). \tag{10}$$

Remark that $\mathcal{U}^{(k)}$ is isomorphic to a Λ_k-tree with $\Lambda_k(x) = \frac{x^2}{1-x} + \sum_{j=4}^{k} s_j \left(\frac{x}{1-x}\right)^j$. This class is certainly algebraic. It is easy to generate many terms in the enumerative sequence using this algebraic equation. We call the class denoted by $\mathcal{P}^{(k)}$, as *prime-degree restricted* strong interval trees.

More generally, one goal of this work is to illustrate a strategy for the analysis of classes of trees $\mathcal{C}$ that fail to be a simple variety of trees because the series governing the number of children available is not analytic. In such cases, one may look for a parameter such that each subclass of trees $\mathcal{C}^{(k)}$, for which that parameter take value at most k, is algebraic. We can then study the classes $\mathcal{C}^{(k)}$ at fixed k, and hopefully develop techniques to obtain information on $\mathcal{C}$ by letting k go to infinity. We study an example of such a class in the present work, and illustrate some of the challenges of sending the limits of both the parameter value k and the size n to infinity at the same time.

4 Enumerating Prime-Degree Restricted Strong Interval Trees

The enumerative analysis of Section 2 applies directly to these families of trees. Ideally, we would like to preserve k as much as possible in the formulas.

4.1 Asymptotic enumeration

The equations (10) allow us to directly apply Theorem 1 to determine asymptotic formulas for the coefficients of the generating functions.

Theorem 5 *For fixed k, the number of prime-degree restricted strong interval trees of size n, denoted $P_n^{(k)}$ grows asymptotically like*

$$P_n^{(k)} \sim \gamma_k \rho_k^{-n} n^{-3/2} \quad \textit{where} \quad \gamma_k = \sqrt{\frac{\rho_k}{2\pi\Lambda_k''(\tau_k)}} \quad \textit{as } n \to \infty. \tag{11}$$

Here, $\Lambda_k(x) = \frac{x^2}{1-x} + \sum_{j=4}^{k} s_j(\frac{x}{1-x})^j$, τ_k *satisfies* $1 - \Lambda_k'(\tau_k) = 0$ *and* $\rho_k = \tau_k - \Lambda_k(\tau_k)$.

Proof: First, we note that since $\sum_{j=4}^{k} s_j(\frac{x}{1-x})^j$ is a polynomial in $\frac{x}{1-x}$, $\Lambda_k(x)$ is certainly analytic at 0. Hence, the enumerative formulas of the first section apply, yielding the asymptotic estimate $U_n^{(k)} \sim \gamma_k \rho_k^{-n} n^{-3/2}$ where $\gamma_k = \sqrt{\frac{\rho_k}{2\pi\Lambda''(\tau_k)}}$.

Next, we note that by the second relation in Equation (10), $P^{(k)}(z) = \frac{U^{(k)}(z)}{1-U^{(k)}(z)}$. This is a subcritical composition, since the value of $U^{(k)}(z)$ at dominant singularity ρ_k is τ_k, which is less than 1 by Theorem 1. Consequently, $P_n^{(k)} \sim \frac{U_n^{(k)}}{1-U_n^{(k)}}$ for large n, hence the approximation stated holds. □

Table 2 contains numeric approximations for τ_k and ρ_k in the range $k = 4 \dots 13$. Using these estimates gives good asymptotic approximations and the enumerative formulas given in Equation (11) converge quickly for fixed k. Next we apply some refined analysis to bound the asymptotic estimate of Equation (11) – see Equation (16) below.

4.2 Bounding the asymptotic estimate of $P_n^{(k)}$

We can produce an asymptotic estimate for $P_n^{(k)}$ in terms of k from Equation (11) by bounding ρ_k and $\Lambda_k''(\tau_k)$. The first ingredient is a more explicit bound for s_n, the number of simple permutations.

Lemma 6 *For every $n \geq 4$, $s_n \leq \sqrt{2\pi}\, n^{n+1/2}\, e^{-n-2}$.*

Proof: This inequality is a consequence of applying the Stirling bound to the bounds of Equation (5). In particular, we use $n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}\, e^{\frac{1}{12n}}$ and the inequality $(1 - \frac{4}{n} + \frac{2}{n(n-1)})e^{\frac{1}{12n}} \leq 1$ for $n \geq 4$, which can be proved by simple computations. □

k	τ_k	ρ_k	k	τ_k	ρ_k
4	0.2258458016	0.1454726242	9	0.1463252500	0.1102193554
5	0.2043553556	0.1364583031	10	0.1375961304	0.1057725121
6	0.1841224072	0.1277948168	11	0.1300393555	0.1017629085
7	0.1689470150	0.1210046262	12	0.1234001218	0.09810173382
8	0.1565912704	0.1152312243	13	0.1174959122	0.09472586497

Tab. 2: Computed values for ρ_k and τ_k for small values of k. For these values, the bounds of Subsection 4.2 on ρ_k and τ_k are not tight. However, we do note that in the limit, the sequences (τ_k), and (ρ_k) tend to 0, which is consistent with the fact that the ordinary generating function for permutations has zero radius of convergence.

From this estimate, the derivations of the bounds on τ_k and ρ_k are straighforward, but technical. Working with the value $\tilde{\tau}_k = \frac{\tau_k}{1-\tau_k}$ simplifies the expressions. Much of the bounds are then consequences of the inequalities $0 < \rho_k < \tau_k < \tilde{\tau}_k < 1$.

Proposition 7 (Bounds for $\tilde{\tau}_k$) *For any $\alpha < \frac{e-2}{e-1}$, there exists $k(\alpha)$ such that for $k > k(\alpha)$*

$$\left(\frac{\alpha}{ks_k}\right)^{\frac{1}{k-1}} < \tilde{\tau}_k < \left(\frac{1}{ks_k}\right)^{\frac{1}{k-1}}. \tag{12}$$

Consequently,

$$\frac{e}{k}\left(\frac{\alpha e^3}{\sqrt{2\pi}\,k^{5/2}}\right)^{\frac{1}{k-1}} < \tilde{\tau}_k < \frac{e}{k}\left(\frac{e^3}{\sqrt{2\pi}\,k^{3/2}(k-4)}\right)^{\frac{1}{k-1}} < \frac{e}{k}. \tag{13}$$

Computational evidence suggests that $k(\alpha) = 4$, for all α near $\frac{e-2}{e-1}$.

Proof: (sketch) The starting point is the equation $1 = \Lambda_k'(x)$, under the change of variables $y = \frac{x}{1-x} \iff x = \frac{y}{1+y}$. We first remark that the equation $1 = \Lambda_k'(\frac{y}{1+y})$ can be rewritten as

$$1 = (1+y)^2 - 1 + (1+y)^2 \sum_{j=4}^{k} js_j y^{j-1} \quad \text{which implies} \quad \frac{2-(1+y)^2}{(1+y)^2} = \sum_{j=4}^{k} js_j y^{j-1}. \tag{14}$$

The next step towards proving the stated inequalities is the fact that for $0 < y < 1$, $1-5y \le \frac{2-(1+y)^2}{(1+y)^2} \le 1$. Indeed, Equation (14) is satisfied at $y = \tilde{\tau}_k$, and consequently these inequalities yield an upper and a lower bound for $\sum_{j=4}^{k} js_j\tilde{\tau}_k^{j-1}$.

The announced upper bound on $\tilde{\tau}_k$ is easily derived from $ks_k\tilde{\tau}_k^{k-1} \le \sum_{j=4}^{k} js_j\tilde{\tau}_k^{j-1} \le 1$.

The lower bound is derived from the upper bound via the inequality $1-5\tilde{\tau}_k - \sum_{j=4}^{k-1} js_j\tilde{\tau}_k^{j-1} \le ks_k\tilde{\tau}_k^{k-1}$. For this purpose, we also need an upper bound on $\sum_{j=4}^{k-1} js_j\tilde{\tau}_k^{j-1}$. It is obtained splitting the sum into two parts, which can be bounded separately. More precisely, setting $\lambda_k = \lfloor k^{1/3} \rfloor$, we can show that

$$\sum_{j=4}^{k-\lambda_k-1} js_j\tilde{\tau}_k^{j-1} = O\left(\frac{1}{k^3}\right) \qquad \text{and that} \qquad \sum_{k-\lambda_k}^{k-1} js_j\tilde{\tau}_k^{j-1} = \frac{1}{e-1}\left(1+o(1)\right).$$

Full details are available in the long version [7] of this abstract. □

Theorem 8 (Bounds for ρ_k) *For any $\alpha < \frac{e-2}{e-1}$, there exist $\beta(\alpha)$ and $k(\alpha)$ such that for any $k \geq k(\alpha)$,*

$$\frac{e}{k}\left(\frac{e^3\alpha}{\sqrt{2\pi}\,k^{5/2}}\right)^{\frac{1}{k-1}}\left(1-\frac{\beta(\alpha)}{k}\right) < \rho_k < \frac{e}{k}\left(\frac{e^3}{\sqrt{2\pi}\,k^{3/2}(k-4)}\right)^{\frac{1}{k-1}}.$$

Consequently, $\rho_k = \frac{e}{k}\left(1-\frac{5}{2}\frac{\log k}{k}+\Theta(\frac{1}{k})\right)$.

Proof: The upper bound is immediate from the bound $\rho_k < \tilde{\tau}_k$ and Proposition 7.
The lower bound is derived by showing that $\rho_k = \tau_k - \Lambda_k(\tau_k) = \tilde{\tau}_k\left(1-\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}-\sum_{j=4}^k s_j\tilde{\tau}_k^{j-1}\right) = \tilde{\tau}_k(1+\Theta(\frac{1}{k}))$. In much the same fashion as the previous proposition, we leverage upper bounds on $\tilde{\tau}_k$ to build a lower bound. In this case, we use $\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k} \leq 2\tilde{\tau}_k \leq 2\frac{e}{k}$, and the summation can be bounded by splitting the sum at the same place:

$$\sum_{j=4}^{k} s_j\,\tilde{\tau}_k^{j-1} = \sum_{j=4}^{k-\lambda_k-1} s_j\,\tilde{\tau}_k^{j-1} + \sum_{j=k-\lambda_k}^{k-1} s_j\,\tilde{\tau}_k^{j-1} + s_k\,\tilde{\tau}_k^{k-1}.$$

Even though it is not the same summation, we nonetheless re-use the same bounding process on the partial summations to recover

$$\sum_{j=4}^{k-\lambda_k-1} s_j\,\tilde{\tau}_k^{j-1} = O\left(\frac{1}{k^3}\right) \quad \text{and} \quad \sum_{j=k-\lambda_k}^{k-1} s_j\,\tilde{\tau}_k^{j-1} = \frac{1}{k-\lambda_k}\sum_{j=k-\lambda_k}^{k-1} js_j\,\tilde{\tau}_k^{j-1} = \Theta\left(\frac{1}{k}\right). \qquad (15)$$

Finally, since $k\,s_k\,\tilde{\tau}_k^{k-1} \leq 1$, we have that $\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}+\sum_{j=4}^k s_j\tilde{\tau}_k^{j-1} = \Theta\left(\frac{1}{k}\right)$, from which it follows that $\rho_k = \tilde{\tau}_k(1+\Theta(\frac{1}{k}))$. The remaining expressions arise from substituting the lower bounds for $\tilde{\tau}_k$, bounds for s_k, followed by some basic manipulations. □

It was known in [8] that $\rho_k = \frac{e}{k}(1+o(1))$, but we are able to produce a more precise estimate. We require this precision when we consider the limit as $k \to \infty$.

From the series expansion of $\Lambda''(x)$, we have $\Lambda''(\tau_k) \geq 2+6\tilde{\tau}_k$. We could expand this expression further, and use lower bounds on $\tilde{\tau}_k$, but it turns out that for our purposes, the bound $\Lambda''(\tau_k) \geq 2$ is sufficient.

Upper bound for the asymptotic estimate of $P_n^{(k)}$ Finally, we have all of the elements to determine an asymptotic estimate of $P_n^{(k)}$. We substitute the upper and lower bounds for ρ_k, and the bound $\Lambda''(\tau_k) \geq 2$ to obtain:

$$\gamma_k\rho_k^{-n}n^{-3/2} \leq \sqrt{\frac{e}{4k\pi}}\left(\frac{k}{e}\right)^n\left(1+\frac{5}{2}\frac{\log k}{k}+\Theta\left(\frac{1}{k}\right)\right)^n n^{-3/2}. \qquad (16)$$

In the limit, Stirling's approximation Our analysis of $\mathcal{P}$ brings together two classic asymptotic facts. The asymptotic growth of a simple variety of trees $\mathcal{T}$ is always of the form $T_n \sim \gamma\rho^{-n}n^{-3/2}$ for some real valued ρ and γ but the classic Stirling's approximation of $n!$ gives $P_n \sim \left(\frac{n}{e}\right)^n\sqrt{2\pi n}$. Subtle analysis is required to reconcile these two estimates.

The trees for all permutations of size n have prime nodes of arity at most n. Thus, if $k \geq n$, $\mathcal{P}_n^{(k)}$ contains all of them, and hence $P_n^{(k)} = n!$ for $k \geq n$. Now, consider Equation (16) with $k = n$. The upper bound is a constant times Stirling's formula(ii). However, when we consider $P_n^{(2n)}$, which is also $n!$, the upper bound gains an unwanted factor of 2^n. This does not contradict the correctness of our asymptotic form, for fixed k, and it rather emphasizes that it is an open problem to develop asymptotic formulas when k is a function of n, and they go to infinity together. This will require a return to the analytic inversion and transfer theorems to study how the error terms depend on k(iii).

4.3 Parameter analysis

From Equation (5), simple permutations make up about 1/9 of all permutations, and consequently the average case analysis of parameters is dominated by their very flat shape. However, the prime-degree restricted trees are much more rich and parameter analysis follows from Section 2.

We remark that the perfect sorting scenarios for σ are directly related to the number of internal nodes, and in particular the distribution among prime and linear nodes. The average subtree size is related to the average reversal size. These two parameters give important insight into the average case analysis of perfect sorting by reversals. A more elaborate discussion on the links between these parameters and algorithm analysis is presented in [6].

4.4 Random generation

Since our initial interest is the shape of the trees, and not the particulars of the internal nodes, we have produced a Boltzmann generator which generates trees of size approximately 10000 for k up to 800 without generating the simple permutation labels. Figure 2 illustrates a randomly generated tree from $\mathcal{P}^{(7)}$ with approximately 1000 leaves. Remark that the structure is dominated by prime nodes of arity 7.

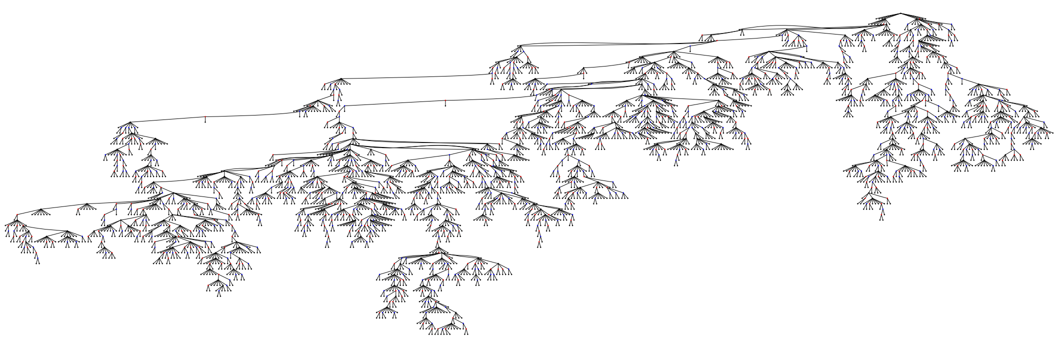


Fig. 2: A tree from $\mathcal{P}^{(7)}$ generated uniformly at random

5 Conclusion

On the biological side, our long term goal is to understand random permutations in order to identify the very specific traits which arise in permutations which encode mammalian genome comparisons. Chauve,

(ii) This constant is $\sqrt{\frac{e}{8\pi^2}}$ obtained replacing k by n in Equation (16).

(iii) The difficulty here lies in Λ being not analytic. Notice however that the same filtration by truncations at order k may also be defined when Λ is analytic: in this case, it is not difficult to prove that we obtain the correct asymptotic formula when taking the limit as k tends to infinity, *i.e.* that limits in n and k commute.

McCloskey and Mishna [9] have taken some preliminary steps in this direction.

On the analytical side, we would like to describe the parameters as functions of k. This will require a very delicate treatment of the bounds, and a much stronger understanding of how to take the limit as $k \to \infty$. This is a much larger undertaking, as essentially we are no longer guided by the inversion theorems.

Finally, one can ask other permutations properties with respect to this filtration. In particular, the model we investigate has a strong connection with the pattern avoiding permutation classes that contain a finite number of simple permutations [1].

Acknowledgements

We are indebted to Cedric Chauve for his guidance and access to the mammalian genome data set, prepared by Bradley Jones and Rosemary McCloskey. Furthermore, Ms. McCloskey wrote the code for the Boltzmann generator, amongst other extremely useful things. We thank Carine Pivoteau for demonstrating her interest in our project at several stages, and for her careful proofreading. We also thank the referees for their usefull suggestions, and for pointing out reference [8] to our attention.

References

[1] M.H. Albert and M.D. Atkinson. Simple permutations and pattern restricted permutations. *Discrete Math.*, 300:1–15, 2005.

[2] M.H. Albert, M.D. Atkinson, and M. Klazar. The enumeration of simple permutations. *J. Integer Seq.*, 6:03.4.4, 2003.

[3] S. Bérard, A. Bergeron, C. Chauve, and C. Paul. Perfect sorting by reversals is not always difficult. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 4:4–16, 2007.

[4] A. Bergeron, C. Chauve, F. de Montgolfier, and M. Raffinot. Computing common intervals of K permutations, with applications to modular decomposition of graphs. *Proc. 13th Annual European Symposium on Algorithms, in Lecture Notes in Comput. Sci.*, 3669:779–790, 2005.

[5] K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. System Sci.*, 13(3):335–379, 1976.

[6] M. Bouvel, C. Chauve, M. Mishna, and D. Rossin. Average-case analysis of perfect sorting by reversals. *Discrete Math. Algorithms Appl.*, 3(3):369–392, 2011.

[7] M. Bouvel, M. Mishna, and C. Nicaud. Some simple varieties of trees arising in permutation analysis. Full version in preparation.

[8] G. Chapuy, A. Pierrot, D. Rossin. On growth rate of wreath-closed permutation classes. Talk at the conference *Permutation Patterns*, 2011.

[9] C. Chauve, R. McCloskey, and M. Mishna. Personal communication, 2011.

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

[11] I. Gessel. Symmetric functions and P-recursiveness. *J. Combin. Theory Ser. A*, 53(2):257–285, 1990.

[12] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[13] The On-Line Encyclopedia of Integer Sequences. Published electronically at http://oeis.org.